ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΧΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΥΔΑΤΙΚΩΝ ΠΟΡΩΝ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΟΣ

Πρόβλεψη καταναλώσεων Νερού – Ενέργειας με χρήση εξελιγμένων μοντέλων μηχανικής μάθησης

Χρήστος Μιχαλόπουλος

Επιβλέποντες καθηγητές: Χρήστος Μακρόπουλος Καθηγητής, ΕΜΠ

Ευστρατιάδης Ανδρέας, Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Φεβρουάριος 2022

Σύλληψη θέματος και πρακτική αξία

- Όπως κάθε σύγχρονη εταιρία παροχής και διαχείρισης νερού έτσι και η ΕΥΔΑΠ θέλει να γνωρίζει την συνολική ποσότητα νερού που χάνεται κατά την διάρκεια της μεταφοράς στο καταναλωτή. Η διαδικασία υπολογισμού δεν είναι απλή λόγο της μεγάλης κλίμακας (πάνω από 2 εκατομμύρια συνδέσεις), και εμπεριέχει πολλές παραδοχές οι οποίες μπορούν να προβούν σε σφάλματα.

- Λόγω του πολύ μεγάλου πλήθους συνέσεων είναι αδύνατη πλήρη καταγραφή. Πάνω από 800 χιλιάδες μη πραγματοποιημένες μετρήσεις ετησίως στην Αττική.

- Ακόμα και οι πιο απλές μεθοδολογίες απαιτούν την πληρότητα των δεδομένων για να προβούν σε σωστά συμπεράσματα.

Στόχοι της έρευνας

- Στόχος της έρευνας αυτής είναι η απόπειρα πρόβλεψης των καταναλώσεων των πελατών οι οποίοι δεν πρόλαβαν να καταμετρηθούν.

- Η πρόβλεψη αυτή έχει διττή σημασία.
  - Ολοκλήρωση δεδομένων για το κλείσιμο του ετήσιου ισοζυγίου νερού.
  - Καλύτερη κοστολόγηση έναντι λογαριασμών.

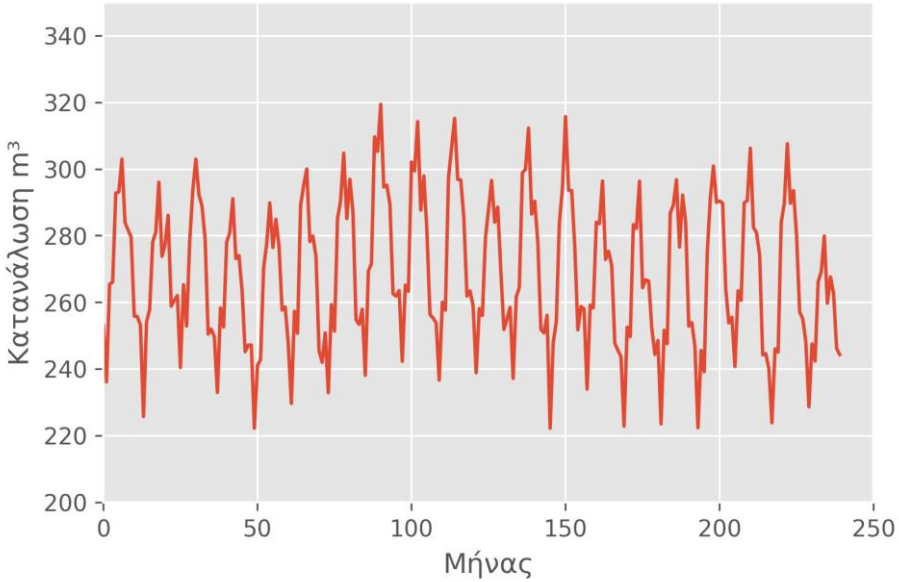
- Για την πραγματοποίηση των προβλέψεων γίνεται χρήση εξελιγμένων μοντέλων Στατιστικής, Μηχανικής Μάθησης, Ομαδοποίησης καθώς και συνδυασμό αυτών σε υπολογιστικό περιβάλλον Python.

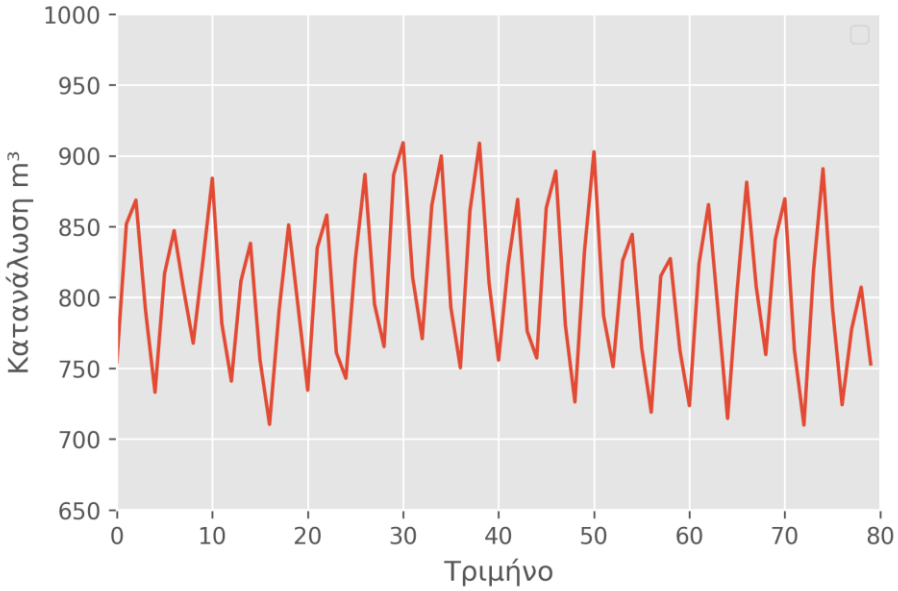
Δεδομένα

Πραγματικά δεδομένα ✗ → Συνθετικά δεδομένα (Tsoukalas et al,2020)

Χαρακτηριστικά συνθετικών δεδομένων

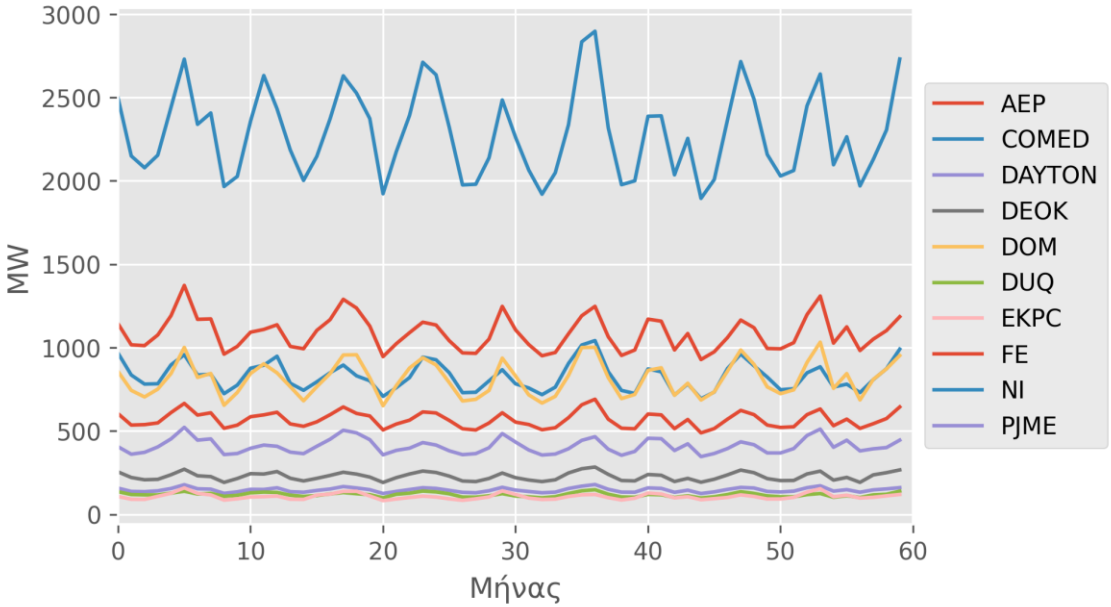
- Κάθε πελάτης αποτελείται από 1 χρονοσειρά 240 μηνών δηλαδή 20 χρόνων.
- Τα συνολικά δεδομένα αφορούν 100 χιλιάδες καταναλωτές.





Δεδομένα

Λόγω έλλειψης πραγματικών δεδομένων οι μέθοδοι δοκιμάζονται σε δεδομένα ενέργειας.



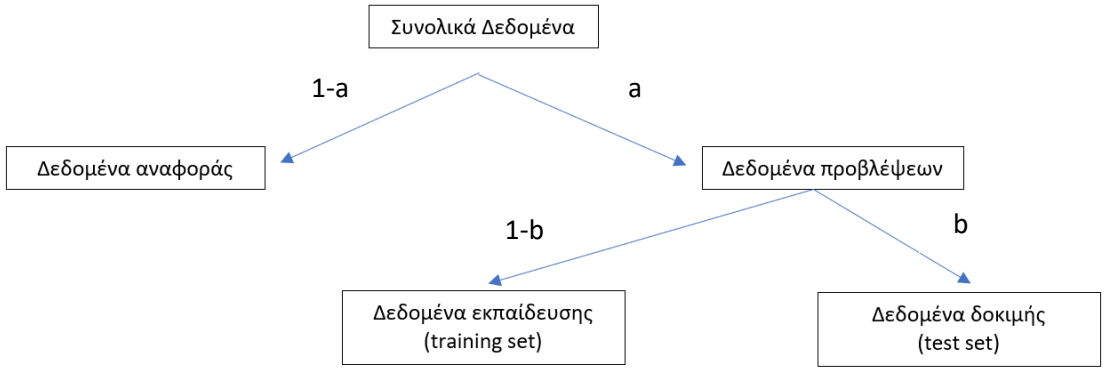
Δεδομένα από:

Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia και της Περιφέρειας της Columbia

Λόγω μικρού μήκους χρονοσειράς επιλέγεται μηνιαία κλίμακα μελέτης σε αντίθεση με τα συνθετικά δεδομένα.

Δεδομένα

Κατά την προ-επεξεργασία τα δεδομένα χωρίζονται σε:

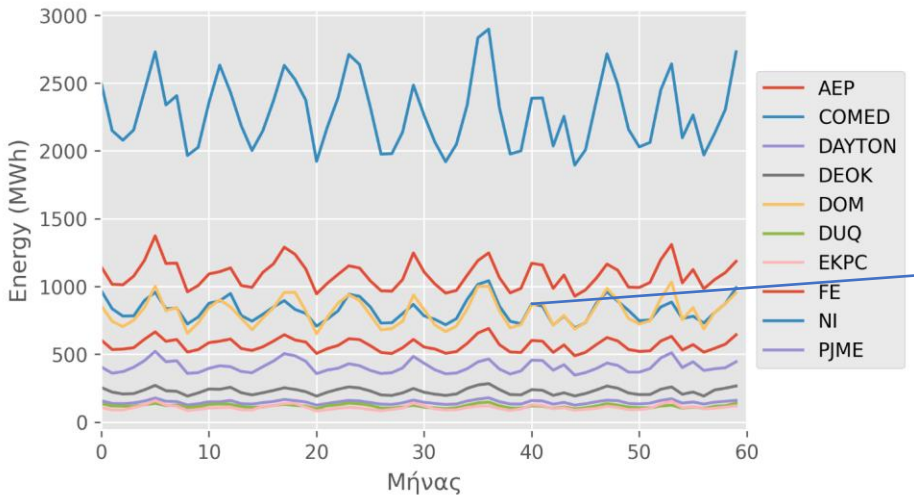


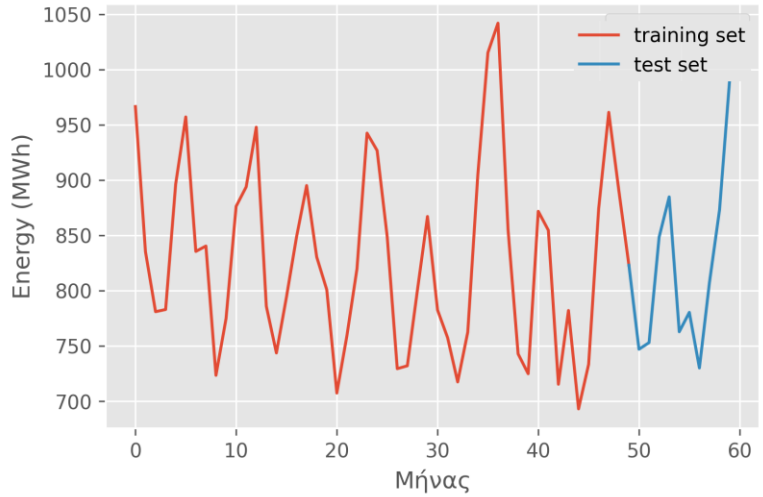
Οι συντελεστές για την κάθε περίπτωση δεδομένων αλλάζουν:

- Για συνθετικά δεδομένα a=5% και b=20%
- Για δεδομένα ενέργειας a=20% και b=$\frac{1}{6}$%

Δεδομένα

Παράδειγμα:





- Ο λόγος που δεν επιλέγονται όλες οι χρονοσειρές για δοκιμή είναι ότι ορισμένες μεθοδολογίες θα τροφοδοτούνται από χρονοσειρές που πρέπει να ξέρουμε την τελευταία τιμή!

Αξιολόγηση μοντέλων

Για την αξιολόγηση των μοντέλων και την σύγκριση των αποδόσεων τους θέτουμε τρείς συναρτήσεις σφάλματος:

- Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error-MAPE)

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|$$

- Συνολικό Άθροισμα (Total Sum-TS)

$$TS = \sum_{i=1}^{n} A_i - F_i$$

- Μέγιστη Διαφορά (Max Difference -MD)

$$MD = max(A_i - F_i)$$

- Κριτήριο πληροφορίας (Bayesian Information Criterion-BIC)

$$\mathrm{BIC} = k \ln(n) - 2 \ln(\widehat{L})$$

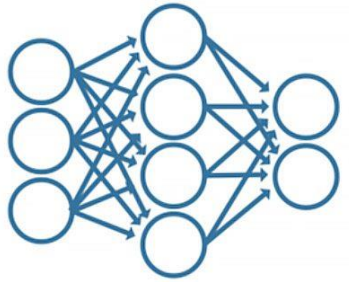
Όπου:

- A_i: πραγματική τιμή
- F_i: τιμή πρόβλεψης

Όπου:

- $\widehat{L}$: πιθανοφάνεια
- k: αριθμός ανεξάρτητων μεταβλητών
- n: αριθμός των δεδομένων

Μοντέλα



- **Naive**

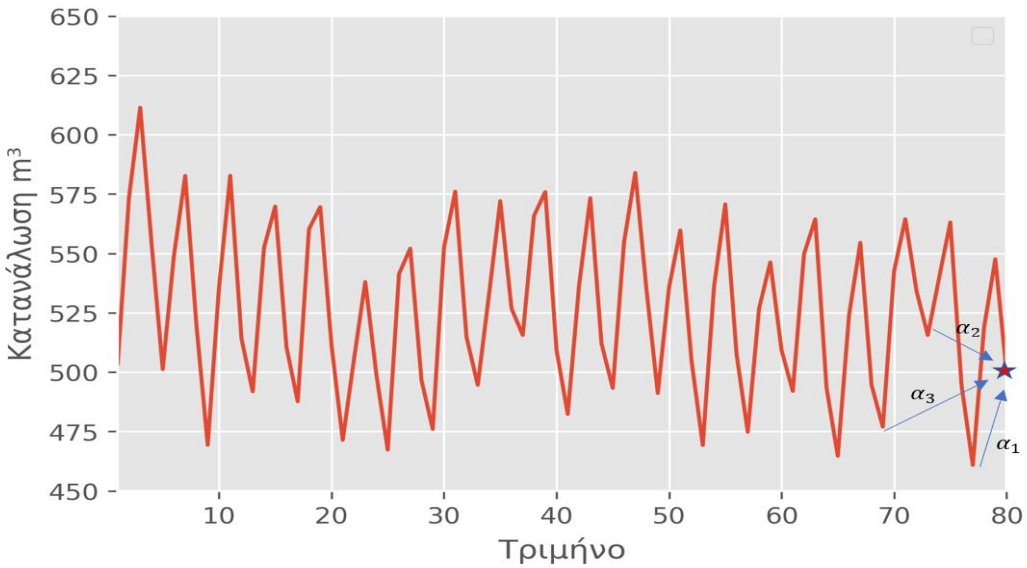
Η πιο απλή μέθοδος αφού η τιμή πρόβλεψης είναι ίση με τιμή της χρονοσειράς μιας περιόδου στο παρελθόν.

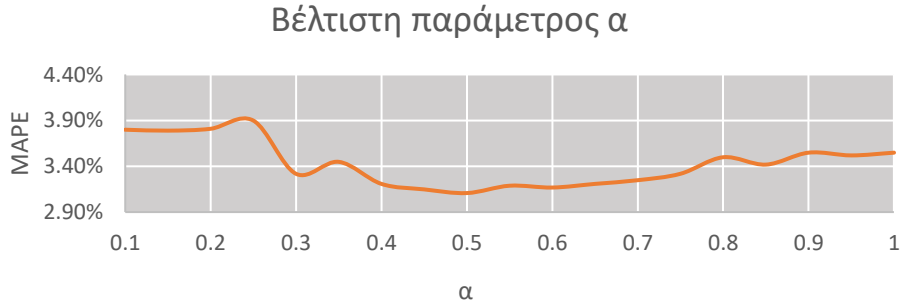
$$F_i = A_{i,t-T}$$

Για μεγαλύτερη ακρίβεια κατά τον υπολογισμό και αποφυγή χονδροειδών σφαλμάτων στις μετρήσης λαμβάνονται υπόψη τιμές μεγαλύτερης μιας περιόδου με τον ανάλογο συντελεστή βαρύτητας.

$$F_i = a_1 * A_{i,t-T} + a_2 * A_{i,t-2T} + a_3 * A_{i,t-3T} \dots$$

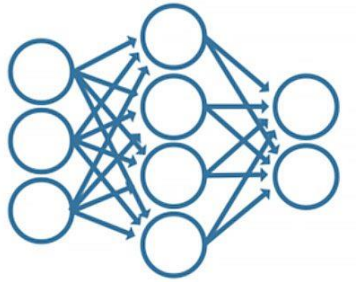
Όπου: $\sum_{i=1}^{n} a_i = 1$





Ο συνδυασμός που έδωσε τα καλύτερα αποτελέσματα είναι: $\vec{\alpha} = (0.50, 0.25, 0.25)$

Μοντέλα



- **ARIMA**(p,d,q) - AutoRegressive Integrated Moving Average
  - p: Παράμετρος Lag (Πόσες τιμές του παρελθόντος θα ληφθούν υπόψη)
  - d: Παράμετρος διαφοροποίησης (Βαθμός αφαίρεσης 2 διαδοχικών χρονικών στιγμών)
  - q: Παράμετρος κινητού μέσου (Πόσες τιμές του σφάλματος θα ληφθούν υπόψη)

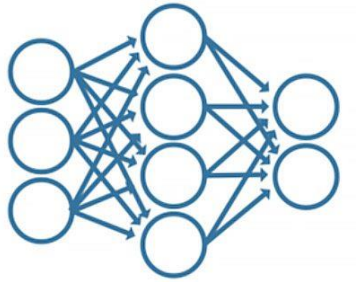
  Είναι ο συνδυασμός τριών μοντέλων AR(p), I(d) και MA(q).

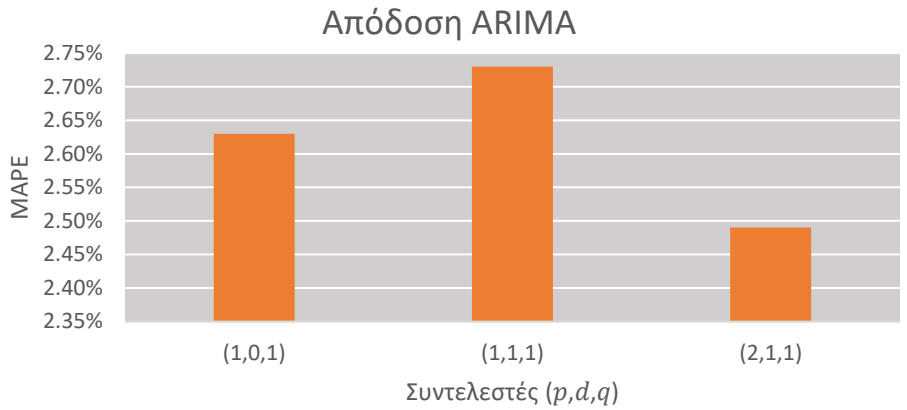
$$X_t - a_1 X_{t-1} - \cdots - a_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \rightarrow \left(1 - \sum_{i=1}^{p'} a_i L^i\right) X_i = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

  Κατά την διαδικασία διαφοροποίησης της χρονοσειράς μπορεί να επιτευχθεί αφαίρεση της εποχικότητας. Το μοντέλο αυτό ονομάζεται εποχική ARIMA ή SARIMA.

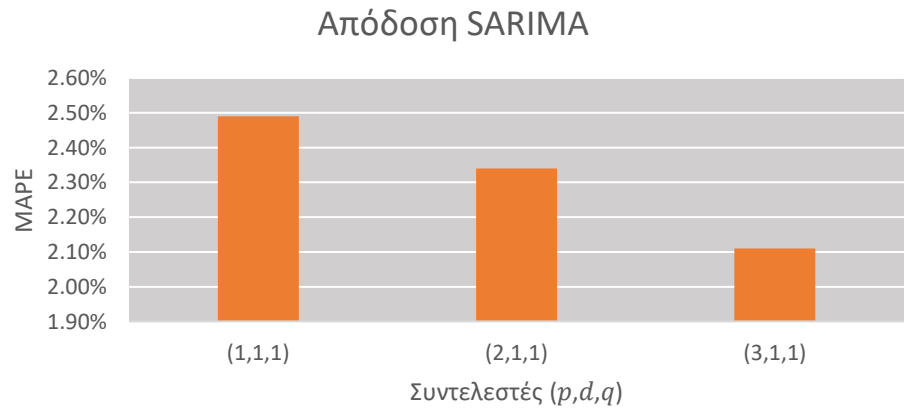
- **SARIMA**(p,d,q),m -Seasonal AutoRegressive Integrated Moving Average
  - Οι παράμετροι είναι οι ίδιοι με το μοντέλο ARIMA το μόνο που αλλάζει είναι η παράμετρος m όπου δηλώνει την εποχικότητα του φαινομένου.

Μοντέλα



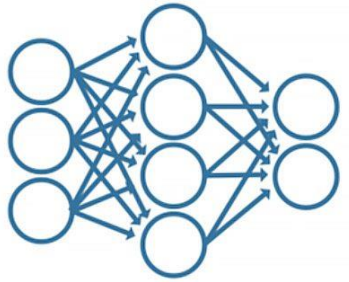


Βέλτιστοι παράμετροι ARIMA: (2,1,1)



Βέλτιστοι παράμετροι SARIMA: (3,1,1)

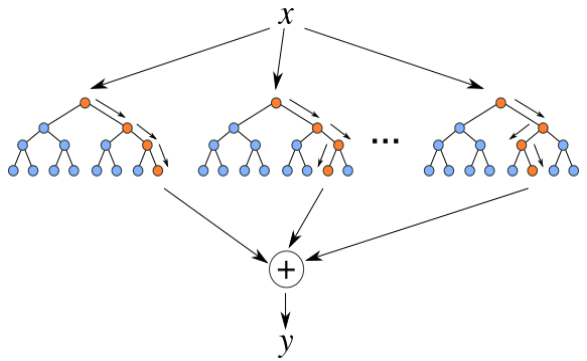
Μοντέλα

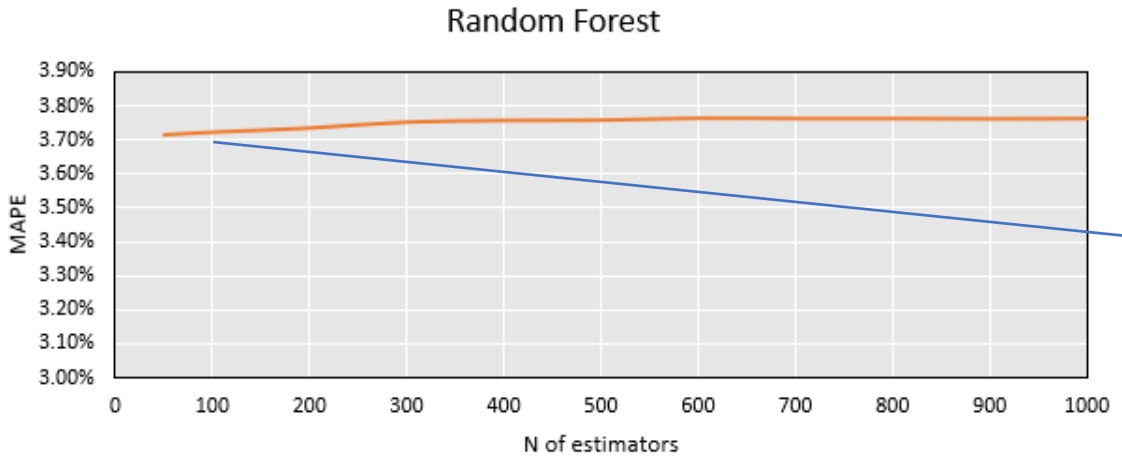


- **Τυχαίο Δάσος (Random Forest-RF)**

Η μέθοδος αυτή βασίζεται στην λειτουργία πολλαπλών τυχαίων δέντρων αποφάσεων. Το αποτέλεσμα κάθε δέντρου συμψηφίζεται για την τελική εκτίμηση.

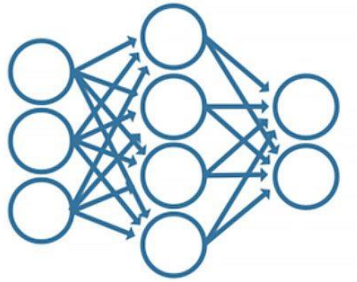
Ορίσματα εισόδου ορίζονται 8 διαδοχικές τιμές της χρονοσειράς.





Επιλέγεται αριθμός τυχαίων δέντρων = 50

Μοντέλα



- **Νευρωνικά Δίκτυα (Neural Networks-NN)**

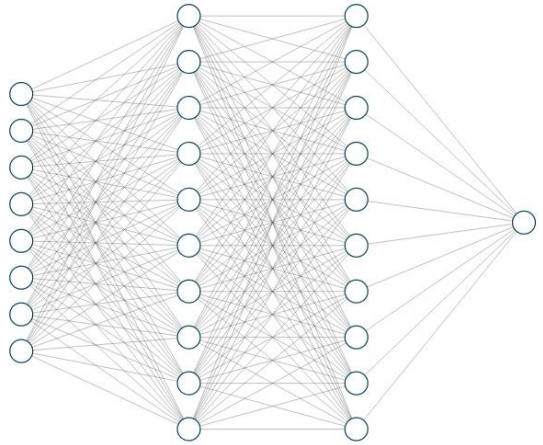
Ένα νευρωνικό δίκτυο αποτελείται από πολλούς απλούς νευρώνες. Το αποτέλεσμα κάθε νευρώνα προκύπτει από κάποια μη-γραμμική συνάρτηση του συνόλου των ορισμάτων εισόδου.

Οι υπερπαράμετροι που χαρακτηρίζουν το μοντέλο είναι ο αριθμός τον ενδιάμεσων στρωμάτων (Layers) και ο αριθμός των κόμβων που απαρτίζουν κάθε στρώμα (Nodes).

Ορίσματα εισόδου ορίζονται 8 διαδοχικές τιμές της χρονοσειράς.

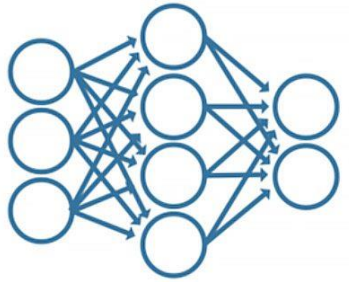
Δοκιμάστηκαν δύο σενάρια με 10 και 30 κόμβους ανά στρώμα

Στην περίπτωση παλινδρόμησης το τελευταίο στρώμα θα πρέπει να περιέχει μόνο ένα κόμβο όπου θα προκύπτει το αποτέλεσμα του μοντέλου.



Αρχιτεκτονική νευρωνικού δικτύου 2 στρωμάτων με 10 κόμβους το καθένα

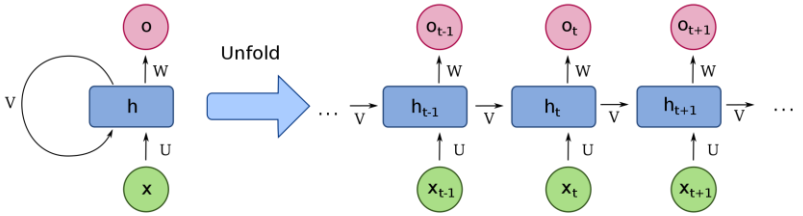
Μοντέλα



- **Δίκτυα Μακράς Βραχύχρονης Μνήμης (Long Short Term Memory – LSTM)**

Το LSTM είναι μορφή τεχνητών ανατροφοδοτούμενων νευρωνικών δικτύων -RNN

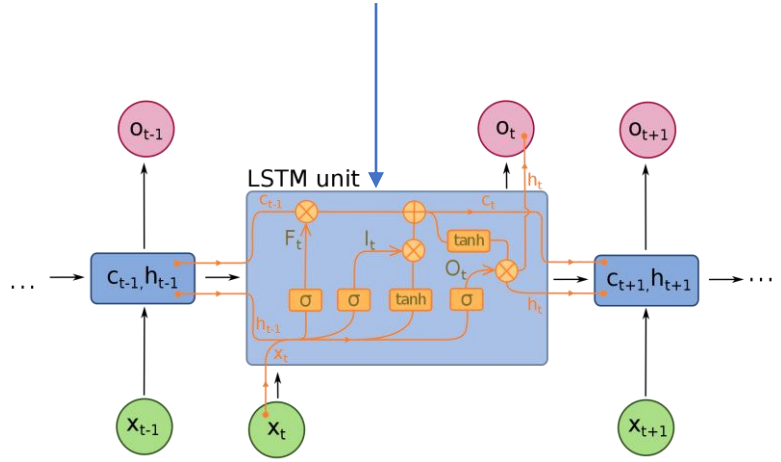
Σε αντίθεση με τα κοινά RNN τα LSTM δεν έχουν το πρόβλημα **εξαφάνισης κλίσεων** καθιστώντας τα ιδανικά για την μοντελοποίηση μεγάλων χρονοσειρών.



Τυπική διάταξη ανατροφοδοτούμενου δικτύου

Το LSTM αποτελείτε από τις πύλες:

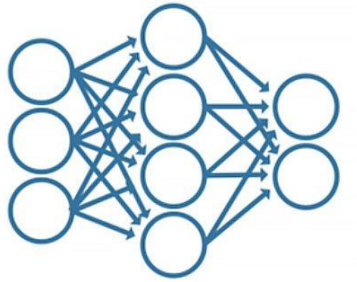
- Εισόδου-Input (I_t)
- Εξόδου-Output (O_t)
- Λήθης-Forget (F_t)



Εσωτερική Αρχιτεκτονική LSTM

Δοκιμάστηκαν τα μοντέλα με 100 και 30 κόμβους σε κάθε στρώμα.

Μοντέλα



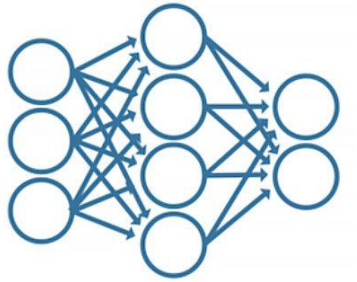
Τα μοντέλα που εξετάστηκαν μπορούν να χωριστούν σε 2 κατηγορίες ανάλογα με τον τρόπο διαχείρισης της πληροφορίας.

- Οι οριζόντιες μεθοδολογίες όπου η πληροφορία ρέει κατά μήκος της χρονοσειράς και απαρτίζονται από τα μοντέλα που αναλύθηκαν παραπάνω
- Οι κατακόρυφες μεθοδολογίες όπου η πρόβλεψή βγαίνει από τα αποτελέσματα άλλων χρονοσειρών που γνωρίζουμε την τελευταία τιμή.

Ο τρόπος λειτουργείας των κατακόρυφων μεθοδολογίων είναι εμπνευσμένος από την μορφή των δεδομένων που έχουμε στην διάθεσή μας. Δηλαδή γνωρίζουμε από ένα μεγάλο μέρος του πληθυσμού και καλούμαστε να προβλέψουμε τις μετρήσεις για του καταναλωτές που δεν γνωρίζουμε.

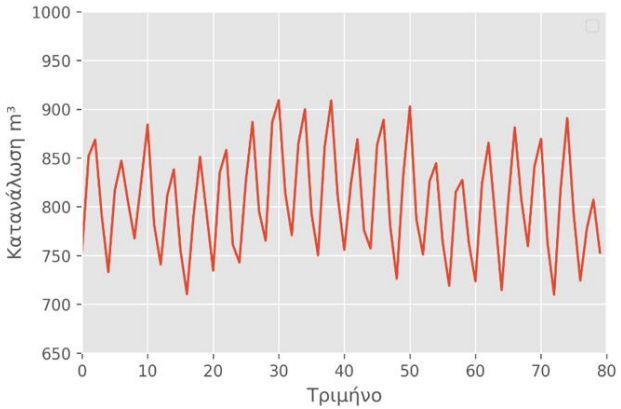
Η ορολογία οριζόντιας και κάθετης μεθοδολογίας βγαίνει από την κατεύθυνση της πρόβλεψης. Δηλαδή στην οριζόντια το αποτέλεσμα εξαρτάται από το παρελθόν του κάθε πελάτη, η πληροφορία ρέει «οριζόντια», ενώ στην κάθετη η πρόβλεψη προκύπτει από άλλες χρονοσειρές, δηλαδή κάθετα.

Μοντέλα



Αρχικά για την επίτευξη των μεθοδολογίων αυτών θα πρέπει να γίνει τροποποίηση των δεδομένων.

Για την διευκόλυνση των πράξεων καθώς μείωση της πολυπλοκότητας κάθε χρονοσειρά ανάγεται σε ένα σημείο στον $\mathbb{R}^4$.



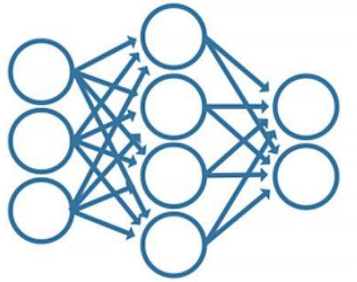
(x_1, x_2, x_3, x_4)

$$x_i = \sum_{n=0}^{N} a_n \, y_{k-nT}$$

$$\hat{y}_i = (x_1, x_2, x_3)$$

Κάθε διάσταση σημείου υποδηλώνει την σχετική τιμή της κάθε εποχής της χρονοσειράς. Για τις χρονοσειρές που γνωρίζουμε την τελευταία τιμή ξέρουμε όλες τις διαστάσεις του σημείου $y_i = (x_1, x_2, x_3, x_4)$, ενώ για τις χρονοσειρές που δεν γνωρίζουμε την τελευταία τιμή ξέρουμε τις τρείς πρώτες διαστάσεις $\widehat{y_i} = (x_1, x_2, x_3)$.

Μοντέλα



- **Κ- Πλησιέστεροι Γείτονες (K-Nearest Neighbors- KNN)**

Όπως και στην κλασσική μεθοδολογία η πρόβλεψη είναι ο μέσος όρος τον γειτόνων.

$$F = \frac{1}{K}\sum_{k=1}^{K} {x_4}^{(k)}$$

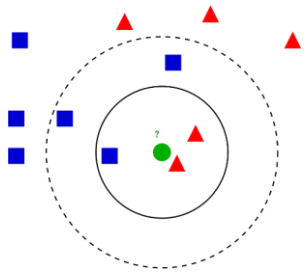
Στην περίπτωση που θέλουμε να προσδώσουμε συντελεστή βαρύτητα στους γείτονες που είναι πιο κοντά θα εφαρμόσουμε την μεθοδολογία των ανάστροφων αποστάσεων.

$$F = \sum_{k=1}^{K} \frac{{r_k}^{-n}}{R} {x_4}^{(k)}, \text{όπου } R = \sum_{k=1}^{K} {r_k}^{-n}$$

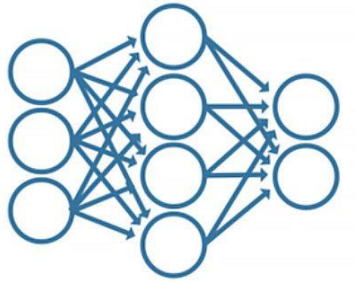
όπου:

- ${x_4}^{(k)}$ η τιμή της 4ης διάστασης του γείτονα.
- r_k: Απόσταση από τον γείτονα.
- n : τάξη των ανάστροφων αποστάσεων.

Οι αποστάσεις μεταξύ των γνωστών και αγνώστων σημείων γίνεται από τις τρείς πρώτες διαστάσεις όπου είναι γνωστές.

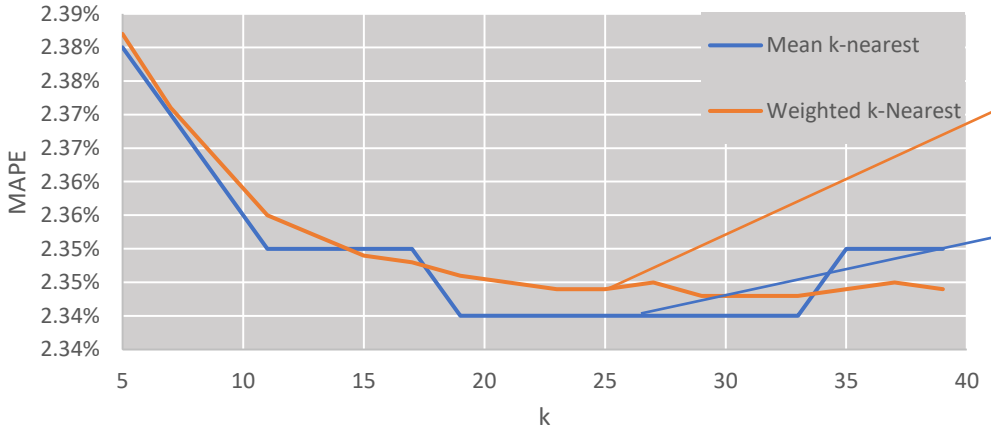


Μοντέλα



- **Κ- Πλησιέστεροι Γείτονες (K-Nearest Neighbors- KNN)**

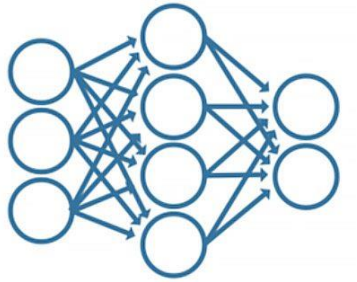
Η βέλτιστη παράμετρος k υπολογίζεται εξαντλητικά.



k= 21 για γείτονες με συντελεστή βάρους

k= 19 για μέσο όρο γειτόνων

Μοντέλα



- **Μίξη γκαουσιανών κατανομών (Gaussian Mixture Models-GMM)**

Με τον όρο μίξη αναφερόμαστε στην κατανομή όπου αποτελείται από πολλές Γκαουσιανές κατανομές.

$$p(\theta|x) = \sum_{i=1}^{K} \pi_i N(x|\mu_i, \sigma_i)$$

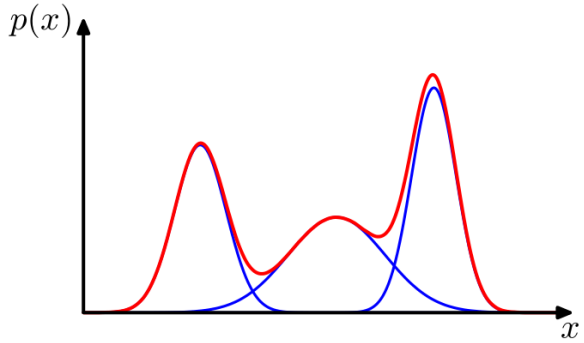
Η μίξη γκαουσιανών κατανομών είναι ένα πιθανοτικό μοντέλο ικανό να περιγράψει φαινόμενα υποομάδων σε ένα πληθυσμό.

π_i : ποσοστό υποομάδας ως προς τον συνολικό πληθυσμό
μ_i : Μέση τιμή υποομάδας
σ_i : Μητρώο τυπικής απόκλισης

Στόχος είναι η εύρεση των π_i, μ_i,σ_i ώστε η συνάρτηση $p(\theta|x)$ να ταιριάζει περισσότερα στα δεδομένα

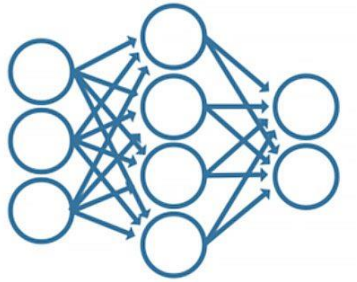
⇩

Αλγόριθμος E-M



Δημιουργία περίπλοκης κατανομής με μίξη απλών

Μοντέλα



- **Μίξη γκαουσιανών κατανομών (Gaussian Mixture Models-GMM)**

Το μοντέλο εκπαιδεύεται με τα γνωστά δεδομένα ⟹ Βέλτιστα π_i , μ_i σ_i για τα δεδομένα

Αφού γνωρίζουμε τους συντελεστές του μοντέλου μπορούμε να υπολογίσουμε την πιθανότητα του σημείο i να ανήκει στην υποομάδα j δεδομένου μόνο των 3 πρώτων διαστάσεων.

$$p(j|i) = \frac{\pi_j N\left(x_{C_u}^{(i)}, \mu_{C_u}^{(j)}, \sigma_{C_u}^{(j)}\right)}{\sum_{j=1}^{K} \pi_j N\left(x_{C_u}^{(i)}, \mu_{C_u}^{(j)}, \sigma_{C_u}^{(j)}\right)}$$

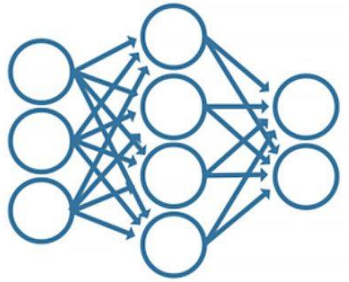
- π_j το βάρος την υποομάδας
- $\mu_{C_u}^{(j)}$ η μέση τιμή αφαιρώντας την διάσταση την οποία δεν γνωρίζουμε.
- $\sigma_{C_u}^{(j)}$ το μητρώο διασποράς αφαιρώντας την διάσταση που δεν γνωρίζουμε.

Η πρόβλεψη υπολογίζεται:

$$F_i = \sum_{j=1}^{K} p(j|i) * \mu_4^{(j)}$$

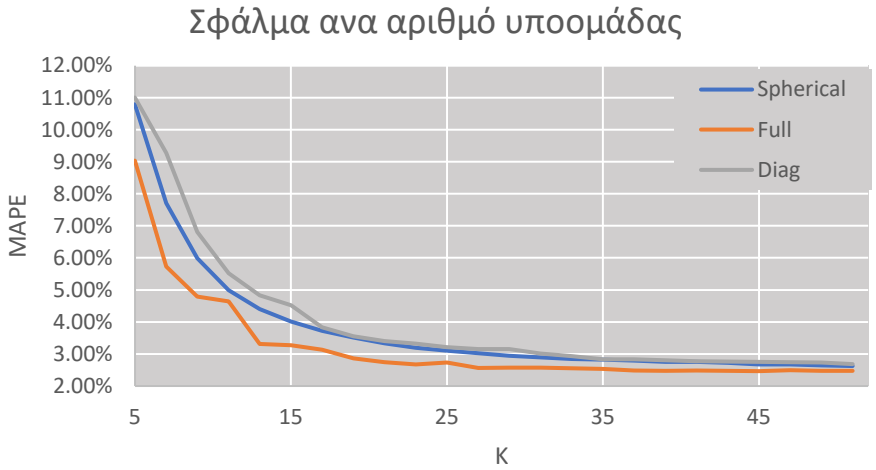
- $p(j|i)$ η πιθανότητα του στοιχείου i να ανήκει στην υποομάδα j.
- $\mu_4^{(j)}$ η μέση τιμής της 4ης διάστασης της υποομάδας j.

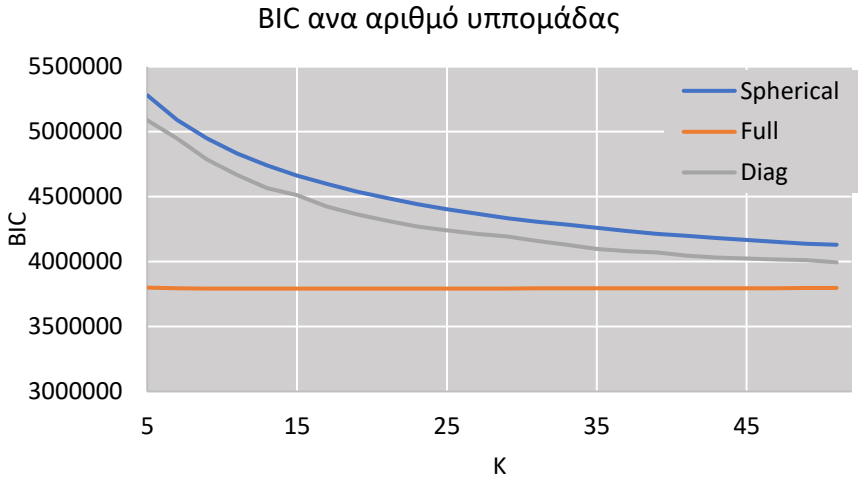
Μοντέλα



- **Μίξη γκαουσιανών κατανομών (Gaussian Mixture Models-GMM)**

Η εύρεση του αριθμού υποομάδων γίνεται με την εξαντλητική μέθοδο.





Βέλτιστοι παράμετροι για 29 υποομάδες και μητρώο πλήρες.

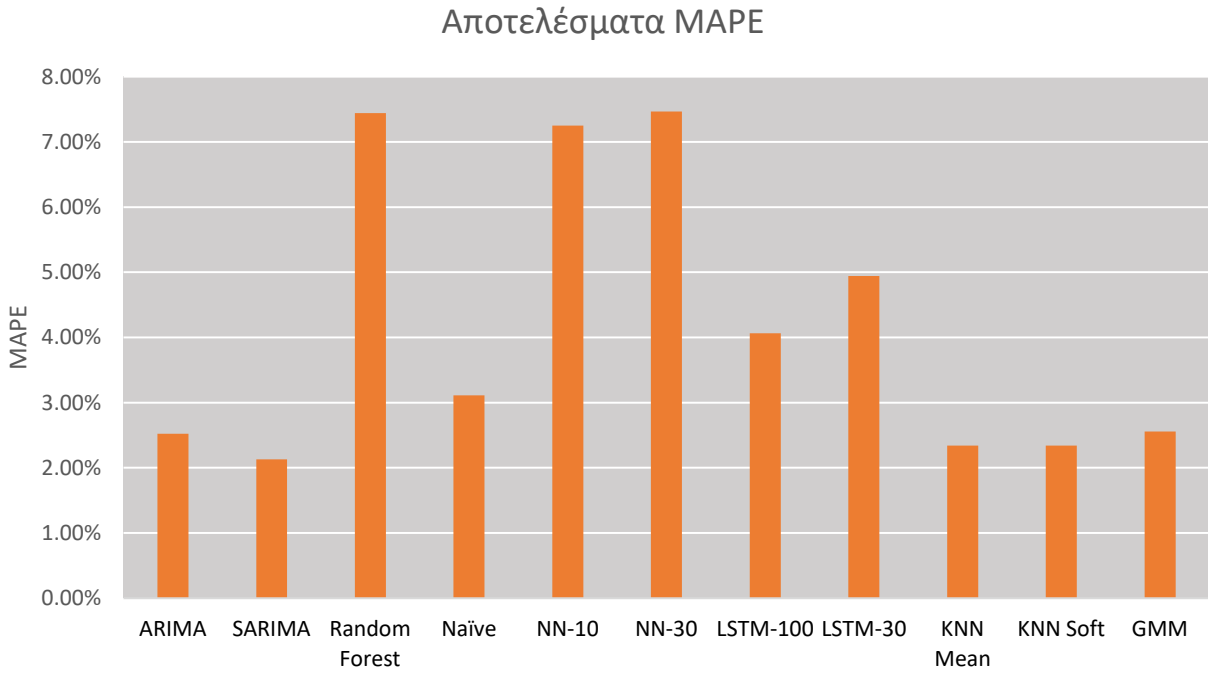
Spherical $\Sigma = \begin{bmatrix} {\sigma_1}^2 & 0 & \cdots & 0 \\ 0 & {\sigma_1}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\sigma_1}^2 \end{bmatrix}$

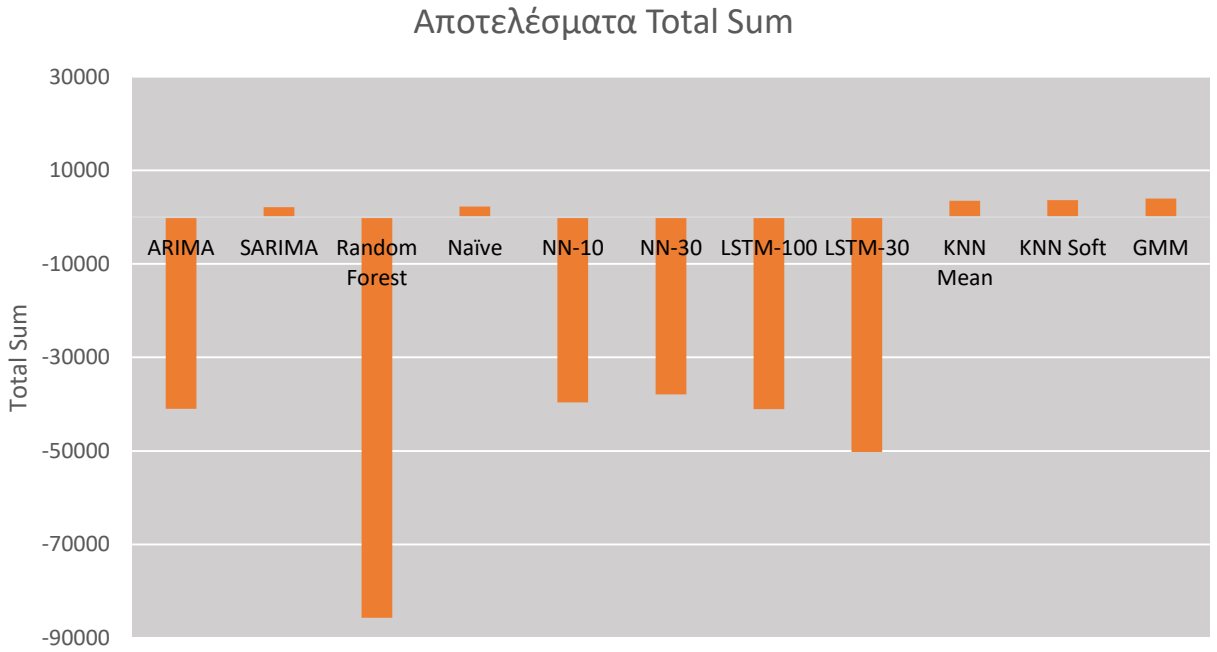
Diag $\Sigma = \begin{bmatrix} {\sigma_1}^2 & 0 & \cdots & 0 \\ 0 & {\sigma_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\sigma_n}^2 \end{bmatrix}$

Full $\Sigma = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & {\sigma_2}^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & {\sigma_n}^2 \end{bmatrix}$

Αποτελέσματα

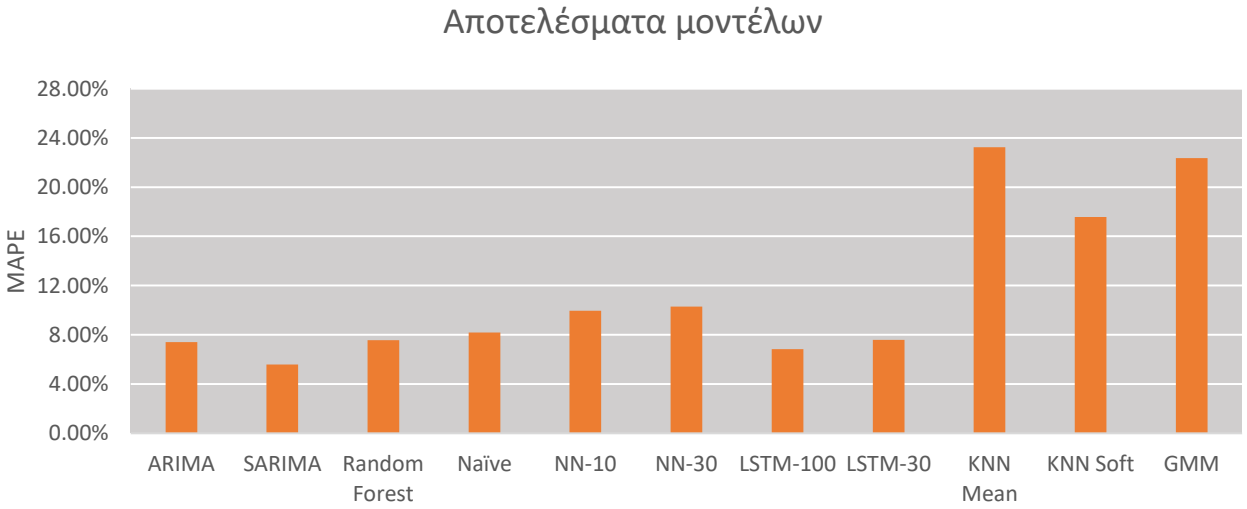
- **Για τα συνθετικά δεδομένα**

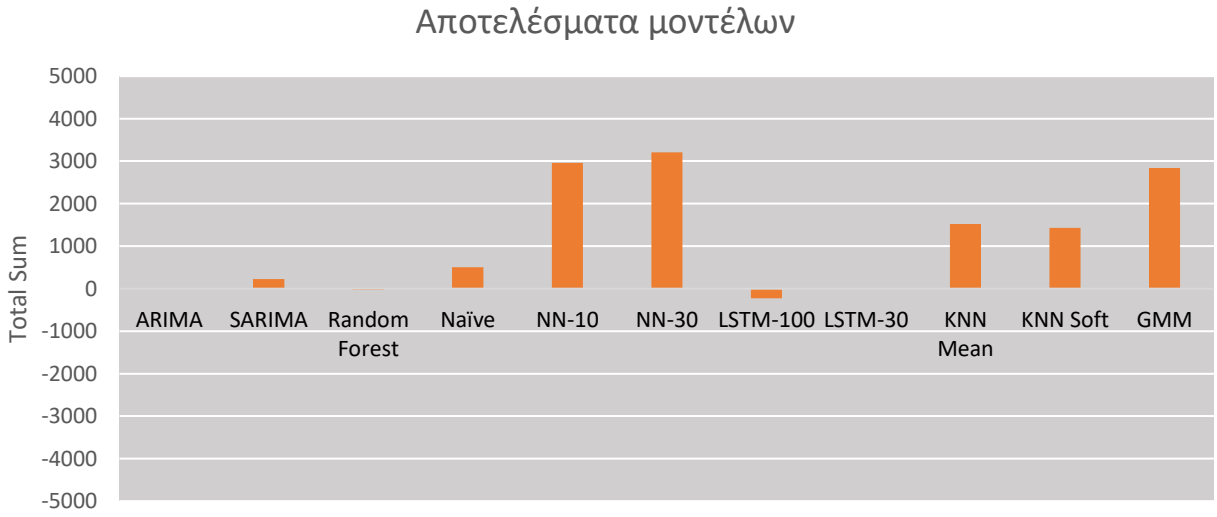




Αποτελέσματα

- **Δεδομένα Ενέργειας**





Συμπεράσματα

- Το μοντέλο SARIMA πέτυχε της καλύτερες αποδόσεις σε όλους τους δείκτες και σε όλα τα δεδομένα
- Τα μοντέλα κάθετης μεθοδολογίας πέτυχαν πολύ καλές αποδόσεις στα συνθετικά δεδομένα (2η θέση) αλλά δεν τα πήγαν καλά στα πραγματικά δεδομένα σημειώνοντας τις χαμηλότερες επιδόσεις. Ο λόγος μπορεί να οφείλεται στην μικρή ποσότητα δεδομένων γνωστών καταναλωτών στην περίπτωση των πραγματικών δεδομένων.
- Η απόδοση του μοντέλου Naive είναι πολύ καλή και για τα 2 είδη δεδομένων, σε συνδυασμό με την χαμηλή πολυπλοκότητα, την καθιστά την πιο αποτελεσματική μέθοδο, στην περίπτωση που χρειαζόμαστε άμεσα αποτελέσματα.
- Το μοντέλο LSTM είχε εξαιρετική απόδοση ειδικά στην περίπτωση των δεδομένων ενέργειας όπου σημείωσαν την 2η καλύτερη απόδοση δεδομένου του μικρού πλήθους των δεδομένων που είχαμε στην διάθεσή μας.
- Το μοντέλο νευρωνικού δικτύου είναι ακατάλληλο για προβλέψεις χρονοσειρών.
- Το μοντέλο τυχαίου δάσους και στις δύο περιπτώσεις πέτυχε ίδια απόδοση, παρ’ όλα αυτά χαμηλή.

Ευχαριστώ πολύ!