

Δράση 2: Ανάπτυξη του προϊόντος

Δ.2.2: Τεχνικός Σχεδιασμός - Υλοποίηση ΠΑΡΑΔΟΤΕΟ: Π.Δ 2.2.1

**ΤΙΤΛΟΣ: Π.Δ. 2.2.1 Τεχνητή Βάση Δεδομένων Στοιχείων
Προσωπικού Φακέλου Υγείας**

ΑΚΡΩΝΥΜΙΟ: MediBRAIN



ΚΩΔΙΚΟΣ ΕΡΓΟΥ

116911

Ιούνιος 2024



«ΑΝΑΠΤΥΞΗ ΨΗΦΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ», ΑΞΟΝΑΣ 2.3- “Ψηφιοποίηση των Επιχειρήσεων”.
Δράση 16706 - “Ψηφιακός Μετασχηματισμός των Μικρομεσαίων Επιχειρήσεων”.

για το έργο: «**MediBRAIN - Ανάπτυξη Συστήματος υποστήριξης προσωπικών αποφάσεων, με χρήση των εργαλείων της Τεχνητής Νοημοσύνης**» με κωδικό **116911** στο πλαίσιο της δράσης «**ΑΝΑΠΤΥΞΗ ΨΗΦΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ**» ΤΟΥ ΑΞΟΝΑ 2.3 “Ψηφιοποίηση των Επιχειρήσεων”, της Δράσης 16706 “Ψηφιακός Μετασχηματισμός των Μικρομεσαίων Επιχειρήσεων”, του Εθνικού Σχεδίου Ανάκαμψης και Ανθεκτικότητας Ελλάδα 2.0.

Ημερομηνία Έναρξης του Έργου: 1η Νοεμβρίου 2023

Ημερομηνία Παράδοσης του Παραδοτέου Π.Δ 2.2.1: 30 Ιουνίου 2024

Ημερομηνία Λήξης του έργου: 31 Οκτωβρίου 2025

Δράσεις:

Δ 2.2 [Τεχνικός Σχεδιασμός - Υλοποίηση] (Υπεύθυνος Φορέας: CSSA)

Παραδοτέο: Π.Δ 2.2.1

Τίτλος Παραδοτέου: [Τεχνητή Βάση Δεδομένων Στοιχείων Προσωπικού Φακέλου Υγείας]

Υπεύθυνος φορέας: CSSA

Συμμετέχοντες Φορείς στο Έργο:



«ΑΝΑΠΤΥΞΗ ΨΗΦΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ», ΑΞΟΝΑΣ 2.3- “Ψηφιοποίηση των Επιχειρήσεων”.
Δράση 16706 - “Ψηφιακός Μετασχηματισμός των Μικρομεσαίων Επιχειρήσεων”.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. Εισαγωγή.....	4
2. Πηγές Δεδομένων.....	5
OHIO-T1DM dataset.....	5
Πεδία βάσης δεδομένων OHIO-T1DM.....	6
Παράδειγμα Αρχείου.....	7
3. Επεξεργασία Δεδομένων για την Εκπαίδευση και Επαλήθευση των Αλγορίθμων.....	9
Βήματα Επεξεργασίας Δεδομένων.....	9
Εποπτευόμενη Μάθηση (Supervised Learning) – Μετασχηματισμός Ανεξάρτητων Μεταβλητών.....	11
4. Εκπαίδευση και Επαλήθευση.....	25
One-Hot-Encoding.....	26
Επιλογή Παραμέτρων (Feature Selection).....	27
5. Συμπεράσματα.....	27
Βιβλιογραφία.....	29

1. ΕΙΣΑΓΩΓΗ

Η ραγδαία πρόοδος της τεχνολογίας στον τομέα της ψηφιακής υγείας και η συνεχώς αυξανόμενη ανάγκη για προσωποποιημένη ιατρική φροντίδα έχουν καταστήσει επιτακτική την ανάπτυξη σύγχρονων εργαλείων συλλογής, επεξεργασίας και αξιοποίησης δεδομένων υγείας. Στο πλαίσιο αυτό, το έργο **MediBRAIN** στοχεύει στη δημιουργία μιας καινοτόμου πλατφόρμας η οποία θα αξιοποιεί προηγμένους αλγόριθμους μηχανικής μάθησης, δεδομένα από φορετές συσκευές (wearables), καθώς και τεχνολογίες διαλειτουργικότητας, προκειμένου να προσφέρει βελτιωμένες υπηρεσίες διαχείρισης για ασθενείς με χρόνια νοσήματα, όπως ο σακχαρώδης διαβήτης.

Βασικό προαπαιτούμενο για την ανάπτυξη, αξιολόγηση και δοκιμή των υπολογιστικών μοντέλων που θα ενσωματωθούν στο MediBRAIN αποτελεί η ύπαρξη μιας **τεχνητής βάσης δεδομένων προσωπικού φακέλου υγείας**. Η βάση αυτή πρέπει να προσομοιώνει με ακρίβεια τα χαρακτηριστικά πραγματικών κλινικών δεδομένων, διασφαλίζοντας την εγκυρότητα των πειραμάτων και την αποτελεσματικότητα των αλγορίθμων. Για τον σκοπό αυτό, στο παρόν παραδοτέο αναπτύχθηκε και τεκμηριώνεται μια τεχνητή βάση δεδομένων, βασισμένη στο διεθνώς αναγνωρισμένο dataset **OHIO-T1DM¹**, το οποίο περιλαμβάνει λεπτομερή δεδομένα συνεχούς παρακολούθησης γλυκόζης (CGM), ινσουλίνης, διατροφής και φυσικής δραστηριότητας από ασθενείς με διαβήτη τύπου 1.

Η επιλογή του OHIO-T1DM dataset κρίθηκε ιδανική, καθώς προσφέρει υψηλής ποιότητας διαχρονικά δεδομένα που καλύπτουν ένα ευρύ φάσμα παραμέτρων που σχετίζονται με τη διαχείριση του διαβήτη. Το dataset αυτό χρησιμοποιείται ευρέως στη διεθνή ερευνητική κοινότητα και παρέχει μια σταθερή βάση για την προσομοίωση της λειτουργίας του MediBRAIN σε ρεαλιστικές συνθήκες. Η ενσωμάτωσή του ως πυρήνα της τεχνητής βάσης συμβάλλει καθοριστικά στον σχεδιασμό, την εκπαίδευση και την επαλήθευση των υπολογιστικών μεθόδων της πλατφόρμας.

Η παρούσα έκθεση περιγράφει αναλυτικά τα βήματα για τη διαμόρφωση της τεχνητής βάσης, τη σύνθεσή της, τη δομή των δεδομένων που περιλαμβάνει, και την αντιστοίχιση αυτών με τα λειτουργικά σενάρια χρήσης του MediBRAIN. Η τεκμηρίωση αυτή αποτελεί το θεμέλιο για τα επόμενα στάδια ανάπτυξης του έργου, διασφαλίζοντας ότι οι τεχνολογικές λύσεις που θα υλοποιηθούν βασίζονται σε αξιόπιστα και ρεαλιστικά δεδομένα.

¹ <https://pubmed.ncbi.nlm.nih.gov/33584164/>

2. ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ

OHIO-T1DM DATASET

Για τους σκοπούς της παρούσας έρευνας, μία από τις βασικές παραδοχές, η οποία δεν επηρεάζει την αξία των προσεγγίσεων και των αποτελεσμάτων της προβλεπτικής μοντελοποίησης, είναι πως η χρονοσειρά που προκύπτει από τις μετρήσεις CGM είναι άμεσα αντιπροσωπευτική της δυναμικής της συγκέντρωσης της γλυκόζης στο πλάσμα του αίματος, ακριβής όσον αφορά στα επίπεδα της εκάστοτε τιμής και με κλινικά ασήμαντη χρονική καθυστέρηση όσον αφορά στο χρόνο της μέτρησης.

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήθηκε το OHIO-T1DM dataset. Το **OHIO-T1DM Dataset** αποτελεί ένα από τα πιο αναλυτικά και αξιόπιστα δημόσια διαθέσιμα σύνολα δεδομένων που αφορούν τη διαχείριση του **Σακχαρώδη Διαβήτη Τύπου 1** με τη χρήση φορετών συσκευών (wearables), αισθητήρων συνεχούς παρακολούθησης γλυκόζης (CGM) και άλλων ψηφιακών πηγών. Το σύνολο δεδομένων δημιουργήθηκε από ερευνητές του *Ohio University* και χρησιμοποιείται ευρέως σε μελέτες τεχνητής νοημοσύνης, μηχανικής μάθησης και προγνωστικών αναλύσεων για τη διαχείριση του διαβήτη.

Η μοναδικότητα του OHIO-T1DM έγκειται στη συνδυασμένη, πολυτροπική φύση των δεδομένων, που επιτρέπει τη χρήση σύνθετων αλγορίθμων πρόβλεψης, ταξινόμησης και συστάσεων. Η ποιότητα και πληρότητα του dataset το καθιστούν ιδανικό για την ανάπτυξη και την επαλήθευση έξυπνων εφαρμογών διαχείρισης του διαβήτη, όπως η πλατφόρμα MediBRAIN.

Επιπλέον, το OHIO-T1DM είναι δημόσια διαθέσιμο μέσω της πλατφόρμας IEEE Dataport, προσφέροντας πλήρη διαφάνεια, εύκολη πρόσβαση και συμβατότητα με εργαλεία προεπεξεργασίας, όπως Python pandas και NumPy. Το γεγονός αυτό ευνοεί σημαντικά την αναπαραγωγικότητα των πειραμάτων, καθώς και την αξιολόγηση της απόδοσης των αλγορίθμων σε ρεαλιστικά σενάρια.

Το MediBRAIN αξιοποιεί το OHIO-T1DM ως βάση για τη δημιουργία τεχνητών σεναρίων εκπαίδευσης και δοκιμής των μοντέλων μηχανικής μάθησης. Παράλληλα, η δυνατότητα προσαρμογής των δεδομένων για την προσομοίωση πραγματικών χρηστών συμβάλλει στη βελτίωση της γενίκευσης και της ακρίβειας πρόβλεψης σε πραγματικές συνθήκες χρήσης.

Τέλος, η χρήση του OHIO-T1DM dataset εξασφαλίζει συμβατότητα με διεθνή πρότυπα, ενώ δεν απαιτεί επιπλέον διαδικασίες αδειοδότησης ή προστασίας προσωπικών δεδομένων, καθώς είναι ψευδωνυμοποιημένο και διατίθεται για ερευνητικούς σκοπούς.

Το σύνολο δεδομένων OHIO-T1DM [1] περιλαμβάνει **δεδομένα από 12 ασθενείς** με διαβήτη τύπου 1, οι οποίοι παρακολούθηθηκαν εντατικά για χρονικό διάστημα έως και 8 εβδομάδων. Οι συμμετέχοντες χρησιμοποίησαν σύγχρονες συσκευές όπως Dexcom G4 (συνεχής παρακολούθηση γλυκόζης) και Insulet Omnipod (αντλία ινσουλίνης), ενώ



παράλληλα κατέγραφαν με χειροκίνητο τρόπο στοιχεία διατροφής, φυσικής δραστηριότητας, λήψης ινσουλίνης και άλλα σημαντικά συμβάντα.

Βασικά χαρακτηριστικά του συνόλου δεδομένων:

- Συχνότητα καταγραφής γλυκόζης: κάθε 5 λεπτά, μέσω CGM.
- Δεδομένα χορήγησης ινσουλίνης: βασική και bolus, με χρονικές σημάνσεις.
- Καταγραφή γευμάτων: ώρα, τύπος και ποσότητα υδατανθράκων.
- Φυσική δραστηριότητα: χρόνος, διάρκεια και ένταση άσκησης.
- Υποκειμενικά δεδομένα: συμβάντα υπογλυκαιμίας, άγχους, αρρώστιας ή κατανάλωσης αλκοόλ.
- Συνολικός όγκος: περίπου 500.000 σημεία δεδομένων ανά ασθενή για όλη την περίοδο.

Πεδία βάσης δεδομένων OHIO-T1DM

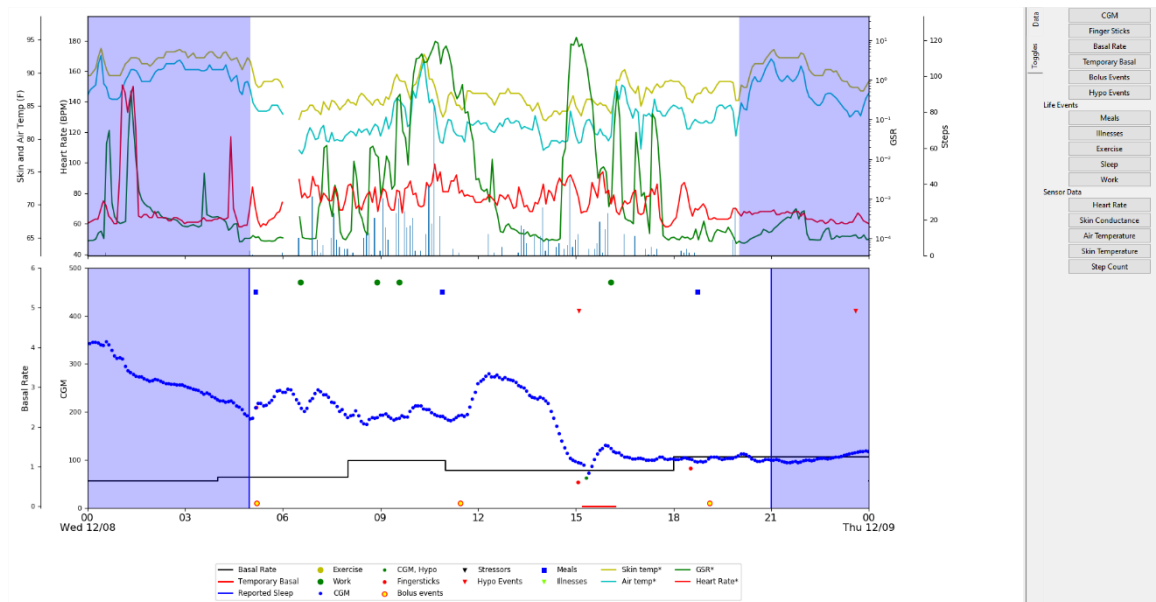
Τα πεδία που περιέχονται στα XML αρχεία παρατίθενται στον Πίνακα 1 που ακολουθεί:

Πίνακας 1: Πεδία OHIO-T1DM Dataset

Πεδίο	Περιγραφή
<patient>	Ο αριθμός ταυτότητας ασθενούς και ο τύπος ινσουλίνης. Το βάρος έχει οριστεί σε 99 ως τιμή υποκατάστασης (placeholder), καθώς τα πραγματικά βάρη των ασθενών δεν είναι διαθέσιμα.
<glucose level>	Δεδομένα συνεχούς παρακολούθησης γλυκόζης (CGM), που καταγράφονται κάθε 5 λεπτά.
<finger stick>	Τιμές γλυκόζης αίματος που λαμβάνονται μέσω αυτό-παρακολούθησης από τον ασθενή.
<basal>	Ο ρυθμός με τον οποίο η βασική ινσουλίνη εγχέεται συνεχώς. Ο βασικός ρυθμός αρχίζει στην καθορισμένη χρονική σήμανση ts και συνεχίζεται μέχρι να οριστεί ένας άλλος βασικός ρυθμός.
<temp basal>	Προσωρινός ρυθμός βασικής ινσουλίνης που αντικαθιστά τον φυσιολογικό βασικό ρυθμό του ασθενούς. Όταν η τιμή είναι 0, αυτό δείχνει ότι η βασική ροή ινσουλίνης έχει ανασταλεί. Στο τέλος του προσωρινού ρυθμού, ο βασικός ρυθμός επιστρέφει στον κανονικό βασικό ρυθμό, <basical>.
<bolus>	Ινσουλίνη που χορηγείται στον ασθενή, συνήθως πριν από το γεύμα ή όταν ο ασθενής είναι υπεργλυκαιμικός. Ο πιο κοινός τύπος bolus, φυσιολογικός, παρέχει όλη την ινσουλίνη ταυτόχρονα. Άλλοι τύποι bolus μπορούν να καταναείμουν χρονικά τη δόση ινσουλίνης κατά την περίοδο μεταξύ ts αρχής και ts τέλους.
<meal>	Ο αυτό-αναφερόμενος χρόνος και τύπος γεύματος, καθώς και η εκτίμηση υδατανθράκων του ασθενούς για το γεύμα.
<sleep>	Οι ώρες του αυτό-αναφερόμενου ύπνου, καθώς και η υποκειμενική αξιολόγηση του ασθενούς για την ποιότητα του ύπνου: 1-Κακός, 2-Μέτριος, 3-Καλός.
<work>	Αυτό-αναφερόμενες ώρες μετάβασης από και προς την εργασία. Η ένταση είναι η υποκειμενική αξιολόγηση της σωματικής άσκησης από τον ασθενή, σε μια κλίμακα από το 1 έως το 10, με το 10 να είναι πιο σωματικά ενεργό.
<stressors>	Χρόνος αυτό-αναφερόμενου άγχους.
<hypo event>	Χρόνος αυτό-αναφερόμενου υπογλυκαιμικού επεισοδίου.
<illness>	Χρόνος αυτό-αναφερόμενης ασθένειας.
<exercise>	Χρόνος και διάρκεια, σε λεπτά, αυτό-αναφερόμενης άσκησης. Η ένταση είναι η υποκειμενική αξιολόγηση της σωματικής άσκησης από τον ασθενή, σε μια κλίμακα από το 1 έως το 10, με το 10 να είναι πιο σωματικά ενεργό.
<basis heart rate>	Καρδιακός ρυθμός, αθροιστικά κάθε 5 λεπτά.
<basis gsr>	Η γαλβανική απόκριση του δέρματος, γνωστή και ως αγωγιμότητα του δέρματος, αθροιστικά κάθε 5 λεπτά.

<basis temperature>	skin	Θερμοκρασία δέρματος, σε βαθμούς Φαρενάιτ, αθροιστικά κάθε 5 λεπτά.
<basis air temperature>		Η θερμοκρασία του αέρα, σε βαθμούς Φαρενάιτ, αθροιστικά κάθε 5 λεπτά.
<basis steps>		Αριθμός βημάτων, αθροιστικά κάθε 5 λεπτά.
<basis sleep>		Φορές που η συσκευή Basis ανέφερε ότι το άτομο κοιμόταν, μαζί με την εκτίμησή της για την ποιότητα του ύπνου.

Σημειώνεται ότι, κατά την ανωνυμοποίηση των δεδομένων, όλες οι ημερομηνίες έχουν μετατοπιστεί κατά το ίδιο τυχαίο χρονικό διάστημα στο μέλλον [2].



Εικόνα 1: Παράδειγμα εμφάνισης δεδομένων του OHIO-T1DM dataset μέσω του ενσωματωμένου εργαλείου

Παράδειγμα Αρχείου

```
<?xml version="1.0" encoding="UTF-8"?>
<patient id="559" weight="99" insulin_type="Novalog">
  <glucose_level>
    <event ts="07-12-2021 01:17:00" value="101"/>
    <event ts="07-12-2021 01:22:00" value="98"/>
    <event ts="07-12-2021 01:27:00" value="104"/>
    <event ts="07-12-2021 01:32:00" value="112"/>
    <event ts="07-12-2021 01:37:00" value="120"/>
    <event ts="07-12-2021 01:42:00" value="127"/>
    <event ts="07-12-2021 01:47:00" value="135"/>
    ...
    <event ts="17-01-2022 23:51:00" value="172"/>
    <event ts="17-01-2022 23:56:00" value="176"/>
  </glucose_level>
  <finger_stick>
    <event ts="07-12-2021 01:03:20" value="107"/>
    <event ts="07-12-2021 01:08:04" value="107"/>
    <event ts="07-12-2021 05:08:14" value="191"/>
    <event ts="07-12-2021 11:21:49" value="586"/>
    <event ts="07-12-2021 12:22:21" value="439"/>
    <event ts="07-12-2021 13:22:30" value="263"/>
    <event ts="07-12-2021 16:54:32" value="71"/>
    <event ts="07-12-2021 18:07:28" value="84"/>
  </finger_stick>

```

```

...
<event ts="17-01-2022 11:07:11" value="64"/>
<event ts="17-01-2022 18:33:13" value="205"/>
</finger_stick>
<basal>
<event ts="07-12-2021 00:00:00" value="0.65"/>
<event ts="07-12-2021 04:00:00" value="0.73"/>
<event ts="07-12-2021 08:00:00" value="1.15"/>
<event ts="07-12-2021 11:00:00" value="0.9"/>
...
<event ts="17-01-2022 05:30:00" value="1.15"/>
<event ts="17-01-2022 16:00:00" value="0.83"/>
</basal>
<temp_basal>
<event ts_begin="07-12-2021 04:49:28" ts_end="07-12-2021 05:08:39" value="0.0"/>
<event ts_begin="07-12-2021 17:01:44" ts_end="07-12-2021 18:03:20" value="0.0"/>
<event ts_begin="08-12-2021 15:12:19" ts_end="08-12-2021 16:12:27" value="0.0"/>
<event ts_begin="09-12-2021 11:11:10" ts_end="09-12-2021 12:19:17" value="0.0"/>
...
<event ts_begin="16-01-2022 07:04:40" ts_end="16-01-2022 07:18:11" value="0.0"/>
<event ts_begin="17-01-2022 09:57:49" ts_end="17-01-2022 12:55:30" value="0.0"/>
</temp_basal>
<bolus>
<event ts_begin="07-12-2021 01:08:04" ts_end="07-12-2021 01:08:04" type="normal" dose="1.6" bwz_carb_input="25"/>
<event ts_begin="07-12-2021 11:21:49" ts_end="07-12-2021 11:21:49" type="normal" dose="9.3" bwz_carb_input="0"/>
<event ts_begin="07-12-2021 23:50:34" ts_end="07-12-2021 23:50:34" type="normal" dose="3.8" bwz_carb_input="0"/>
<event ts_begin="08-12-2021 05:11:38" ts_end="08-12-2021 05:11:38" type="normal" dose="4.1" bwz_carb_input="45"/>
...
<event ts_begin="17-01-2022 12:55:54" ts_end="17-01-2022 12:55:54" type="normal" dose="4.4" bwz_carb_input="40"/>
<event ts_begin="17-01-2022 18:05:28" ts_end="17-01-2022 18:05:28" type="normal" dose="4.4" bwz_carb_input="40"/>
</bolus>
<meal>
<event ts="07-12-2021 15:25:00" type="Snack" carbs="18"/>
<event ts="07-12-2021 16:54:00" type="HypoCorrection" carbs="8"/>
<event ts="07-12-2021 18:08:00" type="Dinner" carbs="75"/>
<event ts="08-12-2021 05:10:00" type="Breakfast" carbs="45"/>
...
<event ts="17-01-2022 04:30:00" type="Breakfast" carbs="36"/>
<event ts="17-01-2022 12:00:00" type="Lunch" carbs="40"/>
<event ts="17-01-2022 17:30:00" type="Dinner" carbs="40"/>
</meal>
<sleep>
<event ts_begin="08-12-2021 04:58:00" ts_end="07-12-2021 21:16:00" quality="2"/>
<event ts_begin="09-12-2021 05:30:00" ts_end="08-12-2021 21:00:00" quality="2"/>
<event ts_begin="10-12-2021 05:37:00" ts_end="09-12-2021 22:09:00" quality="2"/>
...
<event ts_begin="17-01-2022 04:15:00" ts_end="16-01-2022 21:00:00" quality="3"/>
<event ts_begin="18-01-2022 05:30:00" ts_end="17-01-2022 21:30:00" quality="3"/>
</sleep>
<work>
<event ts_begin="08-12-2021 06:32:00" ts_end="08-12-2021 08:53:00" intensity="3"/>
<event ts_begin="08-12-2021 09:34:00" ts_end="08-12-2021 16:05:00" intensity="4"/>
...
<event ts_begin="14-01-2022 06:39:00" ts_end="14-01-2022 10:00:00" intensity="5"/>
<event ts_begin="16-01-2022 18:35:00" ts_end="17-01-2022 06:35:00" intensity="5"/>
</work>
<stressors>
<event ts="22-12-2021 17:03:00" type=" " description=" "/>
</stressors>
<hypo_event>
<event ts="07-12-2021 16:54:00">
  <symptom name=""/>
</event>
<event ts="08-12-2021 15:06:00">
  <symptom name=""/>
</event>
...
<event ts="22-12-2021 19:45:00">
  <symptom name=""/>
</event>

```

```

</hypo_event>
<illness>
  <event ts_begin="12-12-2021 10:12:00" ts_end="" type="" description="" />
  <event ts_begin="15-12-2021 06:41:00" ts_end="" type="" description="" />
  <event ts_begin="25-12-2021 14:30:00" ts_end="" type="" description="" />
  <event ts_begin="25-12-2021 14:40:00" ts_end="" type="" description="" />
</illness>
<exercise>
  <event ts="13-12-2021 13:55:00" intensity="3" type="" duration="150" competitive="" />
  <event ts="23-12-2021 08:30:00" intensity="6" type="" duration="30" competitive="" />
  <event ts="16-01-2022 05:50:00" intensity="6" type="" duration="85" competitive="" />
  <event ts="17-01-2022 05:45:00" intensity="6" type="" duration="50" competitive="" />
</exercise>
<basis_heart_rate>
  <event ts="07-12-2021 12:57:00" value="103"/>
  <event ts="07-12-2021 13:02:00" value="68"/>
  ...
  <event ts="17-01-2022 23:50:00" value="57"/>
  <event ts="17-01-2022 23:55:00" value="58"/>
</basis_heart_rate>
<basis_gsr>
  <event ts="07-12-2021 12:59:00" value="8.7E-5"/>
  <event ts="07-12-2021 13:04:00" value="1.01E-4"/>
  <event ts="07-12-2021 13:09:00" value="1.47E-4"/>
  ...
  <event ts="17-01-2022 23:51:00" value="1.92E-4"/>
  <event ts="17-01-2022 23:56:00" value="1.88E-4"/>
</basis_gsr>
<basis_skin_temperature>
  <event ts="07-12-2021 12:59:00" value="84.92"/>
  <event ts="07-12-2021 13:04:00" value="87.98"/>
  ...
  <event ts="17-01-2022 23:50:00" value="89.6"/>
  <event ts="17-01-2022 23:55:00" value="89.6"/>
</basis_skin_temperature>
<basis_steps>
  <event ts="07-12-2021 12:57:00" value="0"/>
  <event ts="07-12-2021 13:02:00" value="0"/>
  <event ts="07-12-2021 13:07:00" value="0"/>
  ...
  <event ts="17-01-2022 23:54:00" value="0"/>
  <event ts="17-01-2022 23:59:00" value="0"/>
</basis_steps>
<basis_sleep>
  <event tbegin="07-12-2021 21:23:00" tend="07-12-2021 21:25:00" quality="92" type="" />
  <event tbegin="07-12-2021 21:25:00" tend="07-12-2021 21:34:00" quality="92" type="" />
  <event tbegin="07-12-2021 21:34:00" tend="07-12-2021 21:43:00" quality="92" type="" />
  ...
  <event tbegin="17-01-2022 23:54:00" tend="17-01-2022 23:58:00" quality="94" type="" />
  <event tbegin="17-01-2022 23:58:00" tend="18-01-2022 00:12:00" quality="94" type="" />
</basis_sleep>
</patient>

```

3. ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΕΠΑΛΗΘΕΥΣΗ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ

ΒΗΜΑΤΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

Βήμα 1: Δημιουργία πίνακα χρονοσειράς.



«ΑΝΑΠΤΥΞΗ ΨΗΦΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ», ΑΞΟΝΑΣ 2.3- "Ψηφιοποίηση των Επιχειρήσεων".
Δράση 16706 - "Ψηφιακός Μετασχηματισμός των Μικρομεσαίων Επιχειρήσεων".

Σε πρώτη φάση χρησιμοποιείται η κλάση² που έχει αναπτυχθεί για τις ανάγκες επεξεργασίας, διαχείρισης και παροχής προγραμματιστικών διεπαφών σχετικά με τα πρωτογενή δεδομένα του OHIO T1DM Dataset (Βλέπε Παράδειγμα αρχείου ανωτέρω) και συγκεκριμένα ειδική μέθοδος, η οποία δημιουργεί το βασικό πίνακα της χρονοσειράς των μετρήσεων συγκέντρωσης γλυκόζης με πρόσθετες στήλες/παραμέτρους: Ημερομηνία, Ώρα, Χρονική Διαφορά από την Έναρξη της Χρονοσειράς και Τμήμα της Ημέρας κατά τη μέτρηση³. Ειδικά για την τελευταία παράμετρο, η οποία είναι αυτή που κυρίως χρησιμοποιείται στη συγκεκριμένη ανάλυση ορίστηκαν οι εξής κλάσεις:

Πίνακας 2: Χρησιμοποιούμενα διαστήματα εντός της ημέρας

	Κλάση	Διάστημα
Πρωί	1. Morning	07:00πμ – 11:59πμ
Μεσημέρι	2. Afternoon	12:00μμ – 15:59μμ
Απόγευμα	3. Evening	16:00μμ – 20:59μμ
Βράδυ	4. Night	21:00μμ – 23:59μμ
Νύχτα	5. Late Night	12:00πμ – 06:59πμ

Η έξοδος της εν λόγω μεθόδου μας δίνει πίνακα όπως στο παράδειγμα που ακολουθεί (πρώτες 20 διαθέσιμες μετρήσεις του ασθενή 559):

Πίνακας 3: Δείγμα Πίνακα Χρονοσειράς

	date_time	mock_date	time_of_day	part_of_day	time	bg_value
0	2021-12-07 01:17:00	2021-12-07	01:17:00	late night	0	101
1	2021-12-07 01:22:00	2021-12-07	01:22:00	late night	0.08	98
2	2021-12-07 01:27:00	2021-12-07	01:27:00	late night	0.17	104
3	2021-12-07 01:32:00	2021-12-07	01:32:00	late night	0.25	112
4	2021-12-07 01:37:00	2021-12-07	01:37:00	late night	0.33	120
5	2021-12-07 01:42:00	2021-12-07	01:42:00	late night	0.42	127
6	2021-12-07 01:47:00	2021-12-07	01:47:00	late night	0.5	135
7	2021-12-07 01:52:00	2021-12-07	01:52:00	late night	0.58	142
8	2021-12-07 01:57:00	2021-12-07	01:57:00	late night	0.67	140
9	2021-12-07 02:02:00	2021-12-07	02:02:00	late night	0.75	145
10	2021-12-07 02:07:00	2021-12-07	02:07:00	late night	0.83	148
11	2021-12-07 02:12:00	2021-12-07	02:12:00	late night	0.92	151
12	2021-12-07 02:17:00	2021-12-07	02:17:00	late night	1	150
13	2021-12-07 02:22:00	2021-12-07	02:22:00	late night	1.08	124
14	2021-12-07 02:27:00	2021-12-07	02:27:00	late night	1.17	130
15	2021-12-07 02:32:00	2021-12-07	02:32:00	late night	1.25	127
16	2021-12-07 02:37:00	2021-12-07	02:37:00	late night	1.33	121
17	2021-12-07 02:42:00	2021-12-07	02:42:00	late night	1.42	115
18	2021-12-07 02:47:00	2021-12-07	02:47:00	late night	1.5	111
19	2021-12-07 02:52:00	2021-12-07	02:52:00	late night	1.58	109

² https://github.com/spitoglou/diabetes/blob/thesis/src/bgc_providers/ohio_bgc_provider.py

³ https://github.com/spitoglou/diabetes/blob/thesis/src/bgc_providers/ohio_bgc_provider.py#L65

Βήμα 2: Υπολογισμός Ανεξάρτητων Μεταβλητών.

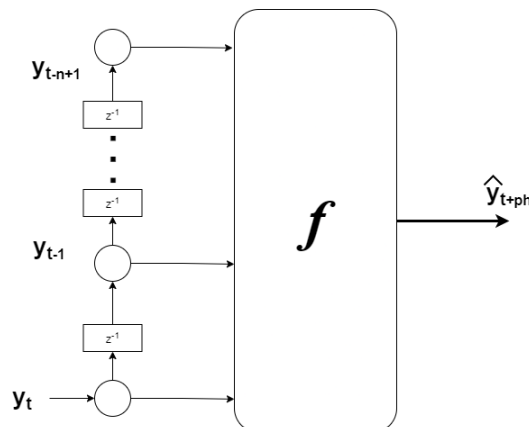
Για κάθε έναν από τους πίνακες που προέκυψαν από το Βήμα 1, οι οποίοι αποτελούν πίνακες χρονοσειρών, έλαβε χώρα η επεξεργασία τους κατά τρόπο ώστε να προκύψουν τα δεδομένα εισόδου για την μετατροπή του αρχικού προβλήματος πρόβλεψης χρονοσειράς σε πρόβλημα επιτηρούμενης μηχανικής μάθησης. Η επεξεργασία αυτή περιλαμβάνει: i) τη δημιουργία μικρότερων χρονοσειρών με μήκος ίσο με το εκάστοτε επιλεγμένο χρονικό παράθυρο, και στη συνέχεια ii) τον υπολογισμό του συνόλου των στατιστικών χαρακτηριστικών των επιμέρους χρονοσειρών, οι οποίες θα χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές κατά τη δημιουργία του προβλεπτικού μοντέλου, όπως αναλύεται παρακάτω:

Εποπτευόμενη Μάθηση (Supervised Learning) – Μετασχηματισμός Ανεξάρτητων Μεταβλητών

Από την Εξ. 1 προκύπτει πως από τη στιγμή που διαθέτουμε μια ιστορική καταγραφή S , το πρόβλημα της πρόβλεψης μελλοντικών τιμών μπορεί να αντιμετωπιστεί ως πρόβλημα Εποπτευόμενης Μάθησης.

$$y_t = f(y_{t-d}, y_{t-d-1}, \dots, y_{t-d-n+1}) + w(t) \quad (\text{Εξ. 1})$$

Η Εποπτευόμενη Μάθηση συνίσταται στη μοντελοποίηση, χρησιμοποιώντας ως βάση ένα πεπερασμένο σύνολο μετρήσεων-παρατηρήσεων, της σχέσης μεταξύ μιας σειράς μεταβλητών εισόδου (ανεξάρτητες μεταβλητές) και μιας ή περισσότερων μεταβλητών εξόδου (εξαρτημένες μεταβλητές), οι οποίες θεωρείται πως συνδέονται και εξαρτώνται από τις πρώτες. Από τη στιγμή που θα κατασκευαστεί ένα μοντέλο της συνάρτησης f , μπορεί να χρησιμοποιηθεί για προβλεπτικούς σκοπούς. Στην Εικόνα 2 δείχνεται σχηματικά η αποτύπωση του προβλήματος της πρόβλεψης με διαθέσιμες n προηγούμενες τιμές της σειράς σε μορφή γενικού προβλήματος παλινδρόμησης (generic regression problem).



Εικόνα 2: Σχηματική αποτύπωση πρόβλεψης μετά τον μετασχηματισμό του χώρου των ανεξάρτητων μεταβλητών.
 $stat_k = g_k(y_t, \dots, y_{t-n+1})$ όπου g_k στατιστικές συναρτήσεις/διαδικασίες μετασχηματισμού

Στα πλαίσια της πειραματικής διερεύνησης της παρούσας διατριβής, χρησιμοποιήθηκε ένας πρωτότυπος μετασχηματισμός του χώρου των ανεξάρτητων μεταβλητών, ο οποίος, με βάση τη συγκεκριμένη προσέγγιση, δεν περιλαμβάνει πλέον τις τιμές της χρονοσειράς για χρονικό διάστημα που αντιστοιχεί σε n βήματα-μετρήσεις, αλλά μια σειρά από στατιστικά χαρακτηριστικά ($stat_k = g_k(y_b, \dots, y_{t-n+1})$) όπου g_k οι στατιστικές συναρτήσεις και διαδικασίες μετασχηματισμού που περιλαμβάνει ο Πίνακας 4) που προκύπτουν από την επεξεργασία των τιμών αυτών για το ίδιο χρονικό διάστημα.

Πίνακας 4: Χαρακτηριστικά Χρονοσειρών ως Υπολογιζόμενες Ανεξάρτητες Μεταβλητές (Χι οι τιμές της χρονοσειράς, n το μήκος της χρονοσειράς, σ^2 διακύμανση μ η μέση τιμή)

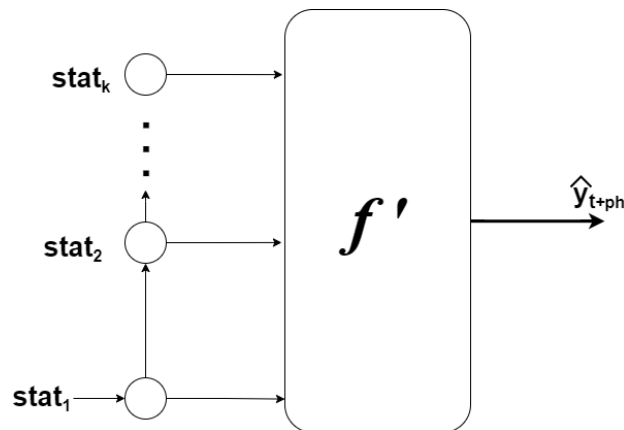
Παράμετρος	Περιγραφή - Παρατηρήσεις
abs_energy(x) [πηγή]	Η «απόλυτη ενέργεια» (absolute energy) της χρονοσειράς ήτοι το άθροισμα των τετραγώνων των επιμέρους τιμών: $E = \sum_{i=1, \dots, n} x_i^2$
absolute_sum_of_changes(x) [πηγή]	Το άθροισμα των απολύτων τιμών των διαδοχικών αλλαγών της χρονοσειράς. $\sum_{i=1, \dots, n-1} x_{i+1} - x_i $
agg_autocorrelation(x, param) [πηγή]	Υπολογίζει την τιμή μιας συνάρτησης συνάθροισης f_{agg} (π.χ. η διακύμανση ή ο μέσος όρος) επί της αυτοσυσχέτισης $R(l)$ για διάφορα διαστήματα [373]. Η αυτοσυσχέτιση $R(l)$ για διάστημα l ορίζεται ως: $R(l) = \frac{1}{(n-l)\sigma^2} \sum_{t=1}^{n-l} (X_t - \mu)(X_{t+l} - \mu)$
agg_linear_trend(x, param) [πηγή]	Υπολογίζει γραμμική παλινδρόμηση ελάχιστων τετραγώνων για τιμές της χρονοσειράς ομαδοποιημένες ανά τμήματα.
approximate_entropy(x, m, r) [πηγή]	Υλοποιεί ένα διανυσματικό αλγόριθμο προσεγγιστικής εντροπίας (Vectorized Approximate Entropy)[374]–[377].
augmented_dickey_fuller(x, param) [πηγή]	Η δοκιμή Augmented Dickey-Fuller είναι μια δοκιμή υπόθεσης που ελέγχει εάν υπάρχει μια μονάδα ρίζας (unit root) σε ένα δείγμα χρονοσειράς. Επιστρέφει την τιμή του αντίστοιχου στατιστικού ελέγχου. [378]–[380]
autocorrelation(x, lag) [πηγή]	Η αυτοσυσχέτιση ενός ορισμένου διαστήματος l σύμφωνα με τον μαθηματικό τύπο: $\frac{1}{(n-l)\sigma^2} \sum_{t=1}^{n-l} (X_t - \mu)(X_{t+l} - \mu)$
binned_entropy(x, max_bins) [πηγή]	Αρχικά χωρίζεται η σειρά x σε max_bins χωρία ίσως αποστάσεων. Στη συνέχεια υπολογίζεται το: $-\sum_{k=0}^{min(max_bins, len(x))} p_k \log(p_k) \cdot \mathbf{1}_{(p_k > 0)}$ όπου p_k το ποσοστό των δειγμάτων στο χωρίο k .
c3(x, lag) [πηγή]	Υπολογίζει την τιμή: $\frac{1}{n-2lag} \sum_{i=0}^{n-2lag} x_{i+2lag}^2 \cdot x_{i+lag} \cdot x_i$ Προτάθηκε στο [381] ως μέτρο της μη γραμμικότητας μιας χρονοσειράς.
change_quantiles(x, ql, qh, isabs, f_agg) [πηγή]	Δημιουργεί έναν «διάδρομο» (corridor), ο οποίος καθορίζεται από τα ποσοστημόρια ql and qh της κατανομής του x . Στη

	συνέχεια υπολογίζει το μέσο όρο της απόλυτης τιμής των διαδοχικών αλλαγών της σειράς x εντός αυτού του διαδρόμου.
cid_ce(x, normalize) [πηγή]	Υπολογίζει μια εκτίμηση της πολυπλοκότητας της χρονοσειράς [382], χρησιμοποιώντας το μαθηματικό τύπο: $\sqrt{\sum_{i=0}^{n-2lag} (x_i - x_{i+1})^2}$
count_above_mean(x) [πηγή]	Το πλήθος των τιμών που είναι μεγαλύτερες από τη μέση τιμή.
cwt_coefficients(x, param) [πηγή]	Υπολογίζει ένα συνεχή μετασχηματισμό χρησιμοποιώντας το κυματίδιο (wavelet) Ricker [383] (γνωστό επίσης και ως “Mexican hat wavelet”), το οποίο ορίζεται: $\frac{2}{\sqrt{3a\pi^{\frac{1}{4}}}} \left(1 - \frac{x^2}{a^2}\right) \exp\left(-\frac{x^2}{2a^2}\right)$ όπου a είναι η παράμετρος μήκους του κυματίδιο.
fft_aggregated(x, param) [πηγή]	Επιστρέφει το φασματικό κεντροειδές (μέση τιμή), τη διακύμανση, την κλίση και την κύρτωση του φάσματος απόλυτου μετασχηματισμού Fourier.
fft_coefficient(x, param) [πηγή]	Υπολογίζει τους συντελεστές Fourier του μονοδιάστατου διακριτού μετασχηματισμού Fourier για πραγματική είσοδο με αλγόριθμο γρήγορου μετασχηματισμού Fourier $A_k = \sum_{m=0}^{n-1} a_m \exp\left\{-2\pi i \frac{mk}{n}\right\}, \quad k = 0, \dots, n-1.$
first_location_of_maximum(x) [πηγή]	Επιστρέφει την πρώτη θέση της μέγιστης τιμής x . Η θέση υπολογίζεται σε σχέση με το μήκος του x .
first_location_of_minimum(x) [πηγή]	Επιστρέφει την πρώτη θέση της ελάχιστης τιμής του x . Η θέση υπολογίζεται σε σχέση με το μήκος του x .
friedrich_coefficients(x, param) [πηγή]	Συντελεστές πολωνόμου $h(x)$, το οποίο έχει προσαρμοστεί στην ντετερμινιστική δυναμική του μοντέλου Langevin $\dot{x}(t) = h(x(t)) + \mathcal{N}(0, R)$ όπως περιγράφεται στο [384]
has_duplicate(x) [πηγή]	Ελέγχει εάν οποιαδήποτε τιμή εμφανίζεται πάνω από μία φορά
has_duplicate_max(x) [πηγή]	Ελέγχει εάν η μέγιστη τιμή εμφανίζεται πάνω από μία φορά
has_duplicate_min(x) [πηγή]	Ελέγχει εάν η ελάχιστη τιμή εμφανίζεται πάνω από μία φορά
index_mass_quantile(x, param) [πηγή]	Υπολογίζεται ο σχετικός δείκτης i όπου $q\%$ της «μάζας» της χρονοσειράς x βρίσκεται αριστερά του i . Για παράδειγμα, για $q = 50\%$ θα επιστρέψει το κέντρο μάζας της χρονοσειράς
kurtosis(x) [πηγή]	Επιστρέφει την κύρτωση (kurtosis) του x , η οποία υπολογίζεται με τον προσαρμοσμένο τυποποιημένο συντελεστή ροπής G2 Fisher-Pearson (adjusted Fisher-Pearson standardized moment coefficient G2).
large_standard_deviation(x, r) [πηγή]	Boolean μεταβλητή που δηλώνει εάν η τυπική απόκλιση του x είναι μεγαλύτερη από r φορές το εύρος που ορίζει η διαφορά μεταξύ μέγιστου και ελάχιστου x . Ως εκ τούτου, ελέγχει εάν $std(x) > r * (max(X) - min(X))$ Σύμφωνα με έναν εμπειρικό κανόνα, η τυπική απόκλιση πρέπει να είναι ένα τέταρτο του εύρους των τιμών.
last_location_of_maximum(x) [πηγή]	Η σχετική τελευταία θέση της μέγιστης τιμής. Η θέση υπολογίζεται σε σχέση με το μήκος της χρονοσειράς.
last_location_of_minimum(x) [πηγή]	Η σχετική τελευταία θέση της ελάχιστης τιμής. Η θέση υπολογίζεται σε σχέση με το μήκος της χρονοσειράς.
length(x) [πηγή]	Το μήκος της χρονοσειράς.

linear_trend(x, param) [πηγή]	Υπολογίζει γραμμική παλινδρόμηση ελαχίστων τετραγώνων για τις τιμές των χρονοσειρών έναντι της ακολουθίας από το 0 έως το μήκος της χρονοσειράς μείον ένα. Προϋποθέτει ότι το σήμα δειγματίζεται ομοιόμορφα. Δεν χρησιμοποιεί τις χρονικές σφραγίδες κατά τη δημιουργία του μοντέλου.
linear_trend_timewise(x, param) [πηγή]	Υπολογίζει γραμμική παλινδρόμηση ελαχίστων τετραγώνων για τις τιμές των χρονοσειρών έναντι της ακολουθίας από το 0 έως το μήκος της χρονοσειράς μείον ένα. Χρησιμοποιεί το ευρετήριο της χρονοσειράς για τη δημιουργία του μοντέλου.
longest_strike_above_mean(x) [πηγή]	Το μήκος του μεγαλύτερου διαδοχικού τμήματος της χρονοσειράς με τιμές μεγαλύτερες από το μέσο όρο της.
longest_strike_below_mean(x) [πηγή]	Το μήκος του μεγαλύτερου διαδοχικού τμήματος της χρονοσειράς με τιμές μικρότερες από το μέσο όρο της.
maximum(x) [πηγή]	Η μέγιστη τιμή της χρονοσειράς.
mean(x) [πηγή]	Ο μέσος όρος της χρονοσειράς.
mean_abs_change(x) [πηγή]	Ο μέσος όρος των απόλυτων διαφορών μεταξύ διαδοχικών τιμών της χρονοσειράς: $\frac{1}{n} \sum_{i=1, \dots, n-1} x_{i+1} - x_i $
mean_change(x) [πηγή]	Ο μέσος όρος των διαφορών μεταξύ διαδοχικών τιμών της χρονοσειράς: $\frac{1}{n} \sum_{i=1, \dots, n-1} x_{i+1} - x_i$
mean_second_derivative_central(x) [πηγή]	Επιστρέφει τη μέση τιμή μιας κεντρικής προσέγγισης της δεύτερης παραγώγου $\frac{1}{n} \sum_{i=1, \dots, n-1} \frac{1}{2} (x_{i+2} - 2 \cdot x_{i+1} + x_i)$
median(x) [πηγή]	Ο διάμεσος της χρονοσειράς.
number_crossing_m(x, m) [πηγή]	Υπολογίζει τον αριθμό των διασταυρώσεων του x στο m . Μια διασταύρωση ορίζεται ως δύο διαδοχικές τιμές όπου η πρώτη τιμή είναι μικρότερη από m και η επόμενη είναι μεγαλύτερη ή αντίστροφα.
number_cwt_peaks(x, n) [πηγή]	Αναζητά διαφορετικές κορυφές στο x . Για να γίνει αυτό, το x εξομαλύνεται από ένα κυματίδιο Ricker [383] και για πλάτη που κυμαίνονται από 1 έως n . Επιστρέφει τον αριθμό των κορυφών που εμφανίζονται σε αρκετές κλίμακες πλάτους και με αρκετά υψηλό λόγο σήματος προς θόρυβο (SNR)
number_peaks(x, n) [πηγή]	Υπολογίζει τον αριθμό των κορυφών με υποστήριξη τουλάχιστον n στη χρονοσειρά x . Μια κορυφή υποστήριξης n ορίζεται ως μια υπό-ακολουθία του x όπου εμφανίζεται μια τιμή, η οποία είναι μεγαλύτερη από τους n γείτονές της στα αριστερά και στα δεξιά.
partial_autocorrelation(x, param) [πηγή]	Υπολογίζει την τιμή της συνάρτησης μερικής αυτοσυσχέτισης στη δεδομένη χρονική υστέρηση. Η μερική αυτοσυσχέτιση υστέρησης k μιας χρονοσειράς $\{x_t, t = 1 \dots T\}$ ισούται με τη μερική συσχέτιση των x_t και x_{t-k} , προσαρμοσμένο για τις ενδιάμεσες μεταβλητές $\{x_{t-1}, \dots, x_{t-k+1}\}$ [385]. Μπορεί να οριστεί ως: $\alpha_k = \frac{Cov(x_t, x_{t-k} x_{t-1}, \dots, x_{t-k+1})}{\sqrt{Var(x_t x_{t-1}, \dots, x_{t-k+1}) Var(x_{t-k} x_{t-1}, \dots, x_{t-k+1})}}$ με x_t και x_{t-k} μοντέλα αυτοσυσχέτισης $AR(k-1)$ που μπορούν να προκύψουν με τη μέθοδο Κοινών Ελάχιστων Τετραγώνων OLS.
percentage_of_reoccurring_datapoints_to_all_datapoints(x) [πηγή]	Επιστρέφει την αναλογία των μοναδικών τιμών που υπάρχουν στη χρονοσειρά περισσότερες από μία φορές. Η

	αναλογία κανονικοποιείται στον αριθμό των μοναδικών τιμών.
percentage_of_reoccurring_values_to_all_values(x) [πηγή]	Επιστρέφει την αναλογία μοναδικών τιμών που υπάρχουν στη χρονοσειρά περισσότερες από μία φορές. Η αναλογία κανονικοποιείται στον αριθμό των σημείων δεδομένων στη χρονοσειρά.
quantile(x, q) [πηγή]	Υπολογίζει το q ποσοστημόριο του x . Το πλήθος με τιμή μεγαλύτερη από $q\%$ των διατεταγμένων τιμών του x .
range_count(x, min, max) [πηγή]	Πλήθος των τιμών της χρονοσειράς εντός του διαστήματος $[min, max)$
ratio_beyond_r_sigma(x, r) [πηγή]	Λόγος τιμών που απέχουν περισσότερο από $r \cdot std(x)$ (άρα $r \cdot sigma$) από τον μέσο όρο του x .
ratio_value_number to time_series_length(x) [πηγή]	Αναλογία, η οποία παίρνει την τιμή 1 στην περίπτωση που όλες οι τιμές της χρονοσειράς εμφανίζονται μόνο μία φορά και τιμή <1 σε διαφορετική περίπτωση. Επί της αρχής, υπολογίζει: $\frac{\text{πλήθος μοναδικών τιμών}}{\text{πλήθος τιμών}}$
sample_entropy(x) [πηγή]	Η «δειγματική εντροπία» (sample entropy) [376], [386] της χρονοσειράς
skewness(x) [πηγή]	Επιστρέφει την ασυμμετρία (skewness) δείγματος x , υπολογιζόμενη με τον προσαρμοσμένο τυποποιημένο συντελεστή ροπής $G1$ Fisher-Pearson (adjusted Fisher-Pearson standardized moment coefficient $G1$).
spkt_welch_density(x, param) [πηγή]	Υπολογίζει τη φασματική πυκνότητα διασταυρούμενης ισχύος (cross power spectral density) των χρονοσειρών X σε διαφορετικές συχνότητες. Για να γίνει αυτό, οι χρονοσειρές μετατοπίζονται πρώτα από τον τομέα χρόνου στον τομέα συχνότητας.
standard_deviation(x) [πηγή]	Η τυπική απόκλιση των τιμών της χρονοσειράς.
sum_of_reoccurring_data_points(x) [πηγή]	Το άθροισμα των σημείων που εμφανίζονται στη χρονοσειρά πάνω από μία φορά.
sum_of_reoccurring_values(x) [πηγή]	Το άθροισμα των τιμών που εμφανίζονται στη χρονοσειρά πάνω από μία φορά.
sum_values(x) [πηγή]	Το άθροισμα των τιμών της χρονοσειράς.
symmetry_looking(x, param) [πηγή]	Δυαδική τιμή ενδεικτική για το εάν η κατανομή της χρονοσειράς «δείχνει» συμμετρική. Αυτό συμβαίνει όταν: $ mean(X) - median(X) < r * (max(X) - min(X))$
time_reversal_asymmetry_statistic(x, lag) [πηγή]	Υπολογίζει την τιμή: $\frac{1}{n - 2lag} \sum_{i=0}^{n-2lag} x_{i+2 \cdot lag}^2 \cdot x_{i+lag} - x_{i+lag} \cdot x_i^2$ Προτάθηκε στο [387] ως υποσχόμενη χαρακτηριστική παράμετρος μιας χρονοσειράς.
value_count(x, value) [πηγή]	Πόσες φορές η τιμή $value$ απαντάται στη χρονοσειρά.
variance(x) [πηγή]	Η διακύμανση (variance) της χρονοσειράς
variance_larger_than_standard_deviation(x) [πηγή]	Δυαδική μεταβλητή που δηλώνει εάν η διακύμανση του x είναι μεγαλύτερη από την τυπική απόκλιση της. Ισούται με τη διακύμανση του x που είναι μεγαλύτερη από 1.
variation_coefficient(x) [πηγή]	Επιστρέφει το συντελεστή διακύμανσης (τυπικό σφάλμα / μέσος όρος, ως σχετική τιμή της διακύμανσης γύρω από το μέσο όρο) του x .

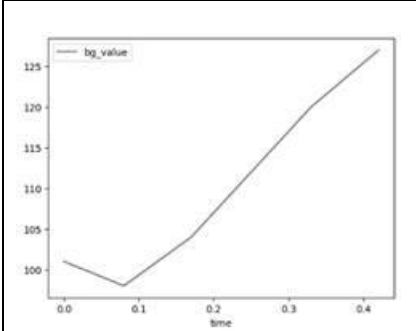
Για τους εν λόγω μετασχηματισμούς χρησιμοποιήθηκε η βιβλιοθήκη λογισμικού TsFresh [3], η οποία χρησιμοποιείται ευρέως για τη επεξεργασία χαρακτηριστικών παραμέτρων από χρονοσειρές και άλλα σειριακά δεδομένα. Έτσι το προς λύση πρόβλημα Μηχανικής Μάθησης μετασχηματίζεται στη μορφή που παρατίθεται στην Εικόνα 3.

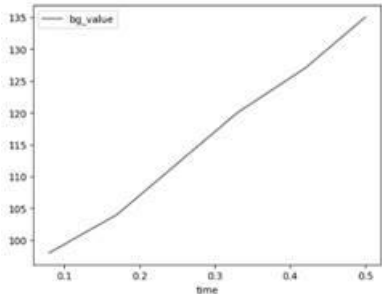


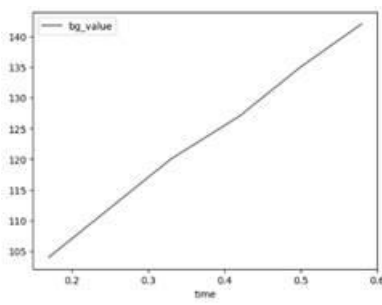
Εικόνα 3: Σχηματική αποτύπωση πρόβλεψης μετά τον μετασχηματισμό του χώρου των ανεξάρτητων μεταβλητών.
 $stat_k = g_k(y_1, \dots, y_{t-n+1})$ όπου g_k στατιστικές συναρτήσεις/διαδικασίες μετασχηματισμού

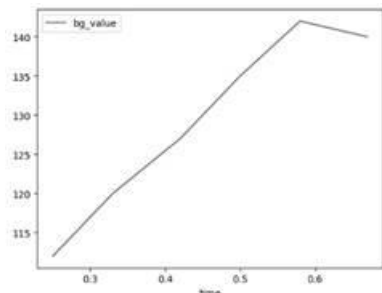
Στον πίνακα που ακολουθεί (Πίνακας 5), εμφανίζονται, χάριν παραδείγματος, οι εννέα πρώτες επιμέρους χρονοσειρές για το υποκείμενο με αριθμό 559 και χρονικό παράθυρο 6 βημάτων (ήτοι 30 min). Στην τελευταία στήλη εμφανίζονται χαρακτηριστικό υποσύνολο των υπολογισμένων χαρακτηριστικών που συμμετέχουν ως ανεξάρτητες μεταβλητές στην ανάπτυξη του προβλεπτικού μοντέλου.

Πίνακας 5: Επιμέρους Χρονοσειρές και Αντίστοιχες Ανεξάρτητες Μεταβλητές

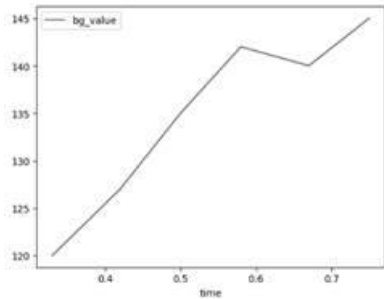
	<table> <tr> <th>time</th><th>bg_value</th><th>id</th></tr> <tr> <td>0 0.00</td><td>101</td><td>a</td></tr> <tr> <td>1 0.08</td><td>98</td><td>a</td></tr> <tr> <td>2 0.17</td><td>104</td><td>a</td></tr> <tr> <td>3 0.25</td><td>112</td><td>a</td></tr> <tr> <td>4 0.33</td><td>120</td><td>a</td></tr> <tr> <td>5 0.42</td><td>127</td><td>a</td></tr> </table>	time	bg_value	id	0 0.00	101	a	1 0.08	98	a	2 0.17	104	a	3 0.25	112	a	4 0.33	120	a	5 0.42	127	a	<table> <tr> <td>0</td><td>bg_value_sum_values</td><td>662.0</td></tr> <tr> <td>1</td><td>bg_value_median</td><td>108.0</td></tr> <tr> <td>2</td><td>bg_value_mean</td><td>110.333333</td></tr> <tr> <td>3</td><td>bg_value_length</td><td>6.0</td></tr> <tr> <td>4</td><td>bg_value_standard_deviation</td><td>10.434984</td></tr> <tr> <td>5</td><td>bg_value_variance</td><td>108.888889</td></tr> <tr> <td>6</td><td>bg_value_root_mean_square</td><td>110.825689</td></tr> <tr> <td>7</td><td>bg_value_maximum</td><td>127.0</td></tr> <tr> <td>8</td><td>bg_value_absolute_maximum</td><td>127.0</td></tr> <tr> <td>9</td><td>bg_value_minimum</td><td>98.0</td></tr> </table>	0	bg_value_sum_values	662.0	1	bg_value_median	108.0	2	bg_value_mean	110.333333	3	bg_value_length	6.0	4	bg_value_standard_deviation	10.434984	5	bg_value_variance	108.888889	6	bg_value_root_mean_square	110.825689	7	bg_value_maximum	127.0	8	bg_value_absolute_maximum	127.0	9	bg_value_minimum	98.0
time	bg_value	id																																																			
0 0.00	101	a																																																			
1 0.08	98	a																																																			
2 0.17	104	a																																																			
3 0.25	112	a																																																			
4 0.33	120	a																																																			
5 0.42	127	a																																																			
0	bg_value_sum_values	662.0																																																			
1	bg_value_median	108.0																																																			
2	bg_value_mean	110.333333																																																			
3	bg_value_length	6.0																																																			
4	bg_value_standard_deviation	10.434984																																																			
5	bg_value_variance	108.888889																																																			
6	bg_value_root_mean_square	110.825689																																																			
7	bg_value_maximum	127.0																																																			
8	bg_value_absolute_maximum	127.0																																																			
9	bg_value_minimum	98.0																																																			

		10 start 0 11 end 5 12 start_time 0.0 13 end_time 0.42 14 start_time_of_day 01:17:00 15 end_time_of_day 01:42:00 16 part_of_day late night
	time bg_value id 1 0.08 98 a 2 0.17 104 a 3 0.25 112 a 4 0.33 120 a 5 0.42 127 a 6 0.50 135 a	0 bg_value__sum_values 696.0 1 bg_value__median 116.0 2 bg_value__mean 116.0 3 bg_value__length 6.0 4 bg_value__standard_deviation 12.793227 5 bg_value__variance 163.666667 6 bg_value__root_mean_square 116.703328 7 bg_value__maximum 135.0 8 bg_value__absolute_maximum 135.0 9 bg_value__minimum 98.0 10 start 1 11 end 6 12 start_time 0.08 13 end_time 0.5 14 start_time_of_day 01:22:00

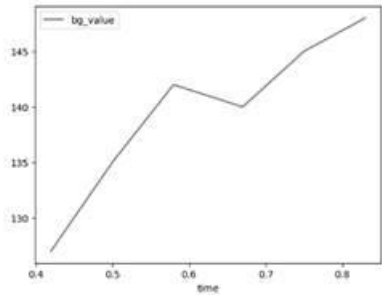
			15	end_time_of_day	01:47:00
			16	part_of_day	late night
	time	bg_value	id	0	bg_value__sum_values
	2	0.17	104	a	740.0
	3	0.25	112	a	1
	4	0.33	120	a	bg_value__median
	5	0.42	127	a	123.5
	6	0.50	135	a	2
	7	0.58	142	a	bg_value__mean
					123.333333
				3	bg_value__length
					6.0
				4	bg_value__standard_deviation
					12.982895
				5	bg_value__variance
					168.555556
				6	bg_value__root_mean_square
					124.014784
				7	bg_value__maximum
					142.0
				8	bg_value__absolute_maximum
					142.0
				9	bg_value__minimum
					104.0
				10	start
					2
				11	end
					7
				12	start_time
					0.17
				13	end_time
					0.58
				14	start_time_of_day
					01:27:00
				15	end_time_of_day
					01:52:00
				16	part_of_day
					late night



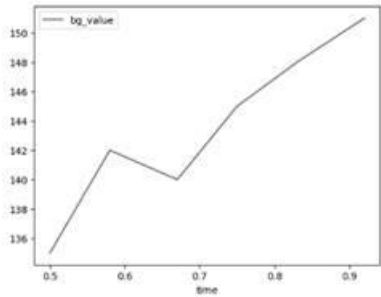
time	bg_value	id	0	bg_value__sum_values	776.0	
3	0.25	112	a	1	bg_value__median	131.0
4	0.33	120	a	2	bg_value__mean	129.333333
5	0.42	127	a	3	bg_value__length	6.0
6	0.50	135	a	4	bg_value__standard_deviation	10.79609
7	0.58	142	a	5	bg_value__variance	116.555556
8	0.67	140	a	6	bg_value__root_mean_square	129.783152
			7	bg_value__maximum	142.0	
			8	bg_value__absolute_maximum	142.0	
			9	bg_value__minimum	112.0	
			10	start	3	
			11	end	8	
			12	start_time	0.25	
			13	end_time	0.67	
			14	start_time_of_day	01:32:00	
			15	end_time_of_day	01:57:00	
			16	part_of_day	late night	



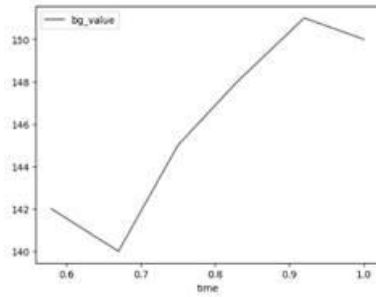
time	bg_value	id	
0	bg_value__sum_values	809.0	
1	bg_value__median	137.5	
2	bg_value__mean	134.833333	
3	bg_value__length	6.0	
4	bg_value__standard_deviation	8.782875	
5	bg_value__variance	77.138889	
6	bg_value__root_mean_square	135.119083	
7	bg_value__maximum	145.0	
8	bg_value__absolute_maximum	145.0	
9	bg_value__minimum	120.0	
10	start	4	
11	end	9	
12	start_time	0.33	
13	end_time	0.75	
14	start_time_of_day	01:37:00	
15	end_time_of_day	02:02:00	
16	part_of_day	late night	



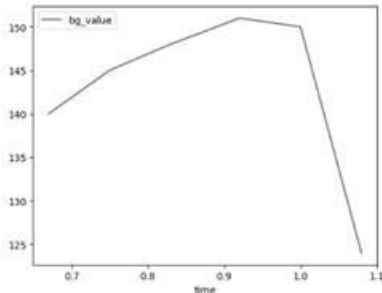
time	bg_value	id	0	bg_value__sum_values	837.0
5	0.42	127 a	1	bg_value__median	141.0
6	0.50	135 a	2	bg_value__mean	139.5
7	0.58	142 a	3	bg_value__length	6.0
8	0.67	140 a	4	bg_value__standard_deviation	6.898067
9	0.75	145 a	5	bg_value__variance	47.583333
10	0.83	148 a	6	bg_value__root_mean_square	139.670445
			7	bg_value__maximum	148.0
			8	bg_value__absolute_maximum	148.0
			9	bg_value__minimum	127.0
			10	start	5
			11	end	10
			12	start_time	0.42
			13	end_time	0.83
			14	start_time_of_day	01:42:00
			15	end_time_of_day	02:07:00
			16	part_of_day	late night



time	bg_value	id	0	bg_value__sum_values	861.0
6	0.50	135 a	1	bg_value__median	143.5
7	0.58	142 a	2	bg_value__mean	143.5
8	0.67	140 a	3	bg_value__length	6.0
9	0.75	145 a	4	bg_value__standard_deviation	5.251984
10	0.83	148 a	5	bg_value__variance	27.583333
11	0.92	151 a	6	bg_value__root_mean_square	143.596077
			7	bg_value__maximum	151.0
			8	bg_value__absolute_maximum	151.0
			9	bg_value__minimum	135.0
			10	start	6
			11	end	11
			12	start_time	0.5
			13	end_time	0.92
			14	start_time_of_day	01:47:00
			15	end_time_of_day	02:12:00
			16	part_of_day	late night



time	bg_value	id	0	bg_value__sum_values	876.0
7	0.58	142 a	1	bg_value__median	146.5
8	0.67	140 a	2	bg_value__mean	146.0
9	0.75	145 a	3	bg_value__length	6.0
10	0.83	148 a	4	bg_value__standard_deviation	4.041452
11	0.92	151 a	5	bg_value__variance	16.333333
12	1.00	150 a	6	bg_value__root_mean_square	146.055925
			7	bg_value__maximum	151.0
			8	bg_value__absolute_maximum	151.0
			9	bg_value__minimum	140.0
			10	start	7
			11	end	12
			12	start_time	0.58
			13	end_time	1.0
			14	start_time_of_day	01:52:00
			15	end_time_of_day	02:17:00
			16	part_of_day	late night

	<table> <tr> <th>time</th><th>bg_value</th><th>id</th></tr> <tr> <td>8</td><td>0.67</td><td>140 a</td></tr> <tr> <td>9</td><td>0.75</td><td>145 a</td></tr> <tr> <td>10</td><td>0.83</td><td>148 a</td></tr> <tr> <td>11</td><td>0.92</td><td>151 a</td></tr> <tr> <td>12</td><td>1.00</td><td>150 a</td></tr> <tr> <td>13</td><td>1.08</td><td>124 a</td></tr> </table>	time	bg_value	id	8	0.67	140 a	9	0.75	145 a	10	0.83	148 a	11	0.92	151 a	12	1.00	150 a	13	1.08	124 a	<table> <tr> <td>0</td><td>bg_value__sum_values</td><td>858.0</td></tr> <tr> <td>1</td><td>bg_value__median</td><td>146.5</td></tr> <tr> <td>2</td><td>bg_value__mean</td><td>143.0</td></tr> <tr> <td>3</td><td>bg_value__length</td><td>6.0</td></tr> <tr> <td>4</td><td>bg_value__standard_deviation</td><td>9.237604</td></tr> <tr> <td>5</td><td>bg_value__variance</td><td>85.333333</td></tr> <tr> <td>6</td><td>bg_value__root_mean_square</td><td>143.298058</td></tr> <tr> <td>7</td><td>bg_value__maximum</td><td>151.0</td></tr> <tr> <td>8</td><td>bg_value__absolute_maximum</td><td>151.0</td></tr> <tr> <td>9</td><td>bg_value__minimum</td><td>124.0</td></tr> <tr> <td>10</td><td>start</td><td>8</td></tr> <tr> <td>11</td><td>end</td><td>13</td></tr> <tr> <td>12</td><td>start_time</td><td>0.67</td></tr> <tr> <td>13</td><td>end_time</td><td>1.08</td></tr> <tr> <td>14</td><td>start_time_of_day</td><td>01:57:00</td></tr> <tr> <td>15</td><td>end_time_of_day</td><td>02:22:00</td></tr> <tr> <td>16</td><td>part_of_day</td><td>late night</td></tr> </table>	0	bg_value__sum_values	858.0	1	bg_value__median	146.5	2	bg_value__mean	143.0	3	bg_value__length	6.0	4	bg_value__standard_deviation	9.237604	5	bg_value__variance	85.333333	6	bg_value__root_mean_square	143.298058	7	bg_value__maximum	151.0	8	bg_value__absolute_maximum	151.0	9	bg_value__minimum	124.0	10	start	8	11	end	13	12	start_time	0.67	13	end_time	1.08	14	start_time_of_day	01:57:00	15	end_time_of_day	02:22:00	16	part_of_day	late night
time	bg_value	id																																																																								
8	0.67	140 a																																																																								
9	0.75	145 a																																																																								
10	0.83	148 a																																																																								
11	0.92	151 a																																																																								
12	1.00	150 a																																																																								
13	1.08	124 a																																																																								
0	bg_value__sum_values	858.0																																																																								
1	bg_value__median	146.5																																																																								
2	bg_value__mean	143.0																																																																								
3	bg_value__length	6.0																																																																								
4	bg_value__standard_deviation	9.237604																																																																								
5	bg_value__variance	85.333333																																																																								
6	bg_value__root_mean_square	143.298058																																																																								
7	bg_value__maximum	151.0																																																																								
8	bg_value__absolute_maximum	151.0																																																																								
9	bg_value__minimum	124.0																																																																								
10	start	8																																																																								
11	end	13																																																																								
12	start_time	0.67																																																																								
13	end_time	1.08																																																																								
14	start_time_of_day	01:57:00																																																																								
15	end_time_of_day	02:22:00																																																																								
16	part_of_day	late night																																																																								

Βήμα 3: Προσθήκη γνωστών τιμών της εξαρτημένης μεταβλητής.

Το συγκεκριμένο βήμα αφορά στην προσθήκη της γνωστής εξαρτημένης μεταβλητής σε κάθε στοιχείο εισόδου, ήτοι η γνωστή τιμή της αρχικής χρονοσειράς ανάλογα με τον εκάστοτε προβλεπτικό ορίζοντα.

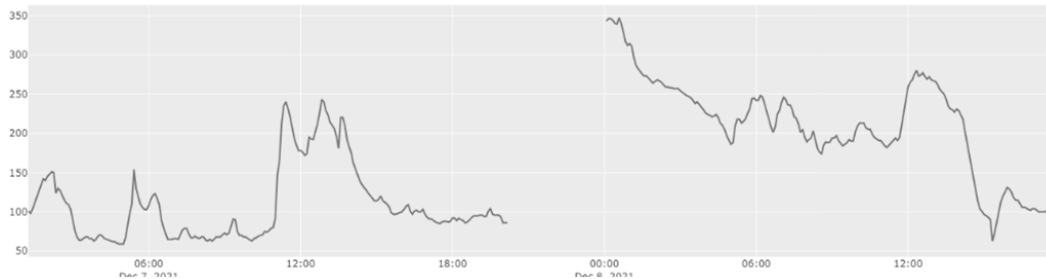
Απαλοιφή Παραμέτρων



«ΑΝΑΠΤΥΞΗ ΨΗΦΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ», ΑΞΟΝΑΣ 2.3- "Ψηφιοποίηση των Επιχειρήσεων".
Δράση 16706 - "Ψηφιακός Μετασχηματισμός των Μικρομεσαίων Επιχειρήσεων".

Για τη δημιουργία πεπερασμένου και σαφώς καθορισμένου πολυδιάστατου χώρου ανεξάρτητων μεταβλητών, κατά το βήμα αυτό απαλείφονται παράμετροι, οι οποίες, για οποιοδήποτε λόγο περιέχουν τιμές ακαθόριστες (*NaN*) ή απειρισμένες (*+Inf*).

Απαλοιφή Χρονικών Κενών



Εικόνα 4: Παράδειγμα χρονικού κενού μετρήσεων (δεδομένα ασθενή 559)

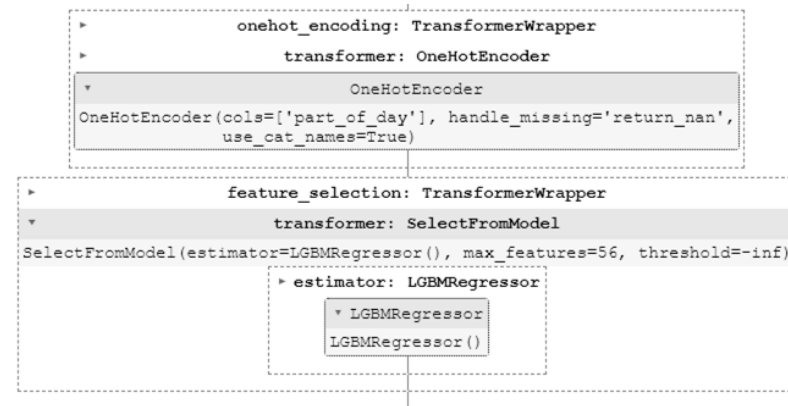
Στα δεδομένα που προέρχονται από το OHIO T1DM Dataset, παρατηρείται σε αρκετές περιπτώσεις η ύπαρξη διακοπών στις μετρήσεις CGM, όπως φαίνεται για παράδειγμα στην Εικόνα 4.

Για να αποφευχθεί ο κίνδυνος συμπερίληψης των οριακών σημείων ως χρονικά συνεχών και η συνεπάγηση νόθευση του συνόλου δεδομένων εκπαίδευσης και επαλήθευσης, στο βήμα αυτό της επεξεργασίας απαλείφονται οι εγγραφές εκατέρωθεν των σημείων αυτών με τρόπο ώστε να τηρείται η συνέπεια της πληροφίας.

4. ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΕΠΑΛΗΘΕΥΣΗ

Data Engineering

Η τελική προετοιμασία των δεδομένων για τους σκοπούς της εκπαίδευσης και επαλήθευσης των αλγορίθμων Μηχανικής Μάθησης, περιλαμβάνει: i) One-Hot-Encoding για ενδεχόμενες κατηγορικές παραμέτρους και ii) επιλογή παραμέτρων (Feature Selection), όπως παρατίθεται και αναλύεται στις επόμενες παραγράφους και την Εικόνα 5.



Εικόνα 5: Βήματα τελικής προετοιμασίας δεδομένων πριν από την εκπαίδευση και επαλήθευση

ONE-HOT-ENCODING

Παρότι υπάρχουν κάποιοι αλγόριθμοι, οι οποίοι μπορούν να λειτουργήσουν απ' ευθείας με κατηγορικά δεδομένα⁴, οι περισσότεροι απαιτούν παραμέτρους εισόδου σε αριθμητική μορφή, είτε ως αυστηρή προϋπόθεση λειτουργίας, είτε ως παράγοντα αποτελεσματικότητας. Αυτό σημαίνει πως τα κατηγορικά δεδομένα θα πρέπει να μετατραπούν σε αριθμητικές τιμές.

Η τεχνική One-Hot-Encoding [4] δημιουργεί νέες δυαδικές στήλες (παραμέτρους), οι οποίες παίρνουν τιμή 1 ή 0 ανάλογα με την παρουσία κάθε κατηγορικής τιμής στα αρχικά δεδομένα (π.χ. Εικόνα 6). Στο πλαίσιο των πειραματισμών της τρέχουσας εργασίας η στήλη για την οποία απαιτήθηκε ο εν λόγω μετασχηματισμός ήταν η «Part Of Day».

Color	Red Yellow Green		
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

Εικόνα 6: Παράδειγμα One-Hot-Encoding

⁴ Χαρακτηριστικό παράδειγμα αποτελούν τα δέντρα αποφάσεων, τα οποία, υπό προϋποθέσεις, δύνανται να λειτουργήσουν με κατηγορικά δεδομένα χωρίς χρεία μετασχηματισμού

ΕΠΙΛΟΓΗ ΠΑΡΑΜΕΤΡΩΝ (FEATURE SELECTION)

Οποιαδήποτε μέθοδος εποπτευόμενης μηχανικής μάθησης είθισται να χρησιμοποιεί κάποια τεχνική «διαίρει και βασίλευε» για τον εντοπισμό του σχετικού υποσυνόλου των διαθέσιμων ανεξάρτητων μεταβλητών, το οποίο ενδείκνυται για την πραγματοποίηση ακριβέστερων προβλέψεων (διά της αντιμετώπισης των προϋποθέσεων εμφάνισης προκατάληψης-bias και overfitting) καθώς επίσης και για τη βελτίωση της υπολογιστικής αποτελεσματικότητας της διαδικασίας (μέσω της μείωσης των διαστάσεων του χώρου των παραμέτρων και συνεπάγωση των απαιτούμενων πόρων που απαιτούνται για τους υπολογισμούς). Στην τυπική εποπτευόμενη μάθηση, αυτό είναι γνωστό ως πρόβλημα επιλογής παραμέτρων και υπάρχουν διάφορες στρατηγικές για την επίλυσή του. Η πλέον δόκιμη στρατηγική, γνωστή ως προσέγγιση wrapper [5], συνίσταται στη δημιουργία διάφορων υποσυνόλων παραμέτρων και την αξιολόγησή τους εκτελώντας αλγόριθμους εκμάθησης και μετρώντας την ακρίβεια του μοντέλου που προκύπτει.

Για τις ανάγκες της τρέχουσας εργασίας, στην παραμετροποίηση του μηχανισμού προετοιμασίας των δεδομένων εισόδου ζητήθηκε να επιλεγεί το 10% των πιο σημαντικών διαθέσιμων παραμέτρων και ως διαδικασία επιλογής προκρίθηκε μέθοδος βασισμένη στον αλγόριθμο LightGBM [6],[7].

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η δημιουργία της τεχνητής βάσης δεδομένων που παρουσιάστηκε στο παρόν παραδοτέο αποτέλεσε ένα καθοριστικό βήμα για την επιστημονικά τεκμηριωμένη ανάπτυξη και δοκιμή των υπολογιστικών μοντέλων του MediBRAIN. Με βάση το διεθνώς καταξιωμένο dataset **OHIO-T1DM**, συγκροτήθηκε ένα πολυδιάστατο, δυναμικό και αξιόπιστο περιβάλλον εκπαίδευσης για την εφαρμογή μοντέλων πρόβλεψης επιπέδων γλυκόζης και τη διαχείριση δεδομένων υγείας.

Η πολυτροπική φύση των δεδομένων, η υψηλή συχνότητα καταγραφής (ανά 5 λεπτά), καθώς και ο μεγάλος όγκος πληροφορίας για κάθε ασθενή, κατέστησαν δυνατή τη μοντελοποίηση πολύπλοκων σχέσεων μεταξύ βιοδεικτών, συνθηκών ζωής και θεραπευτικών παρεμβάσεων. Ιδιαίτερη έμφαση δόθηκε στην προεπεξεργασία των χρονοσειρών, στον υπολογισμό στατιστικών χαρακτηριστικών μέσω της βιβλιοθήκης TSFresh, καθώς και στη μετατροπή του προβλήματος πρόβλεψης σε πρόβλημα εποπτευόμενης μηχανικής μάθησης.

Η χρήση τεχνικών feature engineering και feature selection, καθώς και η εφαρμογή μεθόδων όπως το One-Hot-Encoding και η επιλογή βέλτιστων παραμέτρων μέσω LightGBM, βελτιστοποίησαν τον χώρο των εισόδων για τα μοντέλα μηχανικής μάθησης. Η αφαίρεση

ασαφών τιμών και διακοπών στις μετρήσεις CGM εξασφάλισε την εγκυρότητα των πειραμάτων, ενισχύοντας τη σταθερότητα και την αξιοπιστία των αποτελεσμάτων.

Η δομή της τεχνητής βάσης επιτρέπει όχι μόνο την επανάληψη και αναπαραγωγή των πειραμάτων αλλά και την περαιτέρω εμπλουτισμό της με επιπλέον σενάρια, καθιστώντας την ιδανική για προσομοιώσεις πραγματικών χρηστών και σεναρίων απόκρισης της πλατφόρμας MediBRAIN. Επίσης, η διαλειτουργικότητά της με διεθνή πρότυπα και η απουσία ευαίσθητων προσωπικών δεδομένων διασφαλίζουν τη συμμόρφωση με τις απαιτήσεις της ηθικής και νομικής διακυβέρνησης.

Συνολικά, η τεχνητή βάση δεδομένων αποτελεί ένα απαραίτητο επιστημονικό και τεχνολογικό θεμέλιο για την επόμενη φάση ανάπτυξης της πλατφόρμας MediBRAIN, προσφέροντας τα απαραίτητα εργαλεία για την προσαρμοστική εκπαίδευση, αξιολόγηση και ενίσχυση των αλγορίθμων πρόβλεψης και σύστασης. Η επιτυχής ολοκλήρωση αυτού του σταδίου ενισχύει την ετοιμότητα του έργου να προχωρήσει σε φάσεις πιλοτικής εφαρμογής και κλινικής αξιολόγησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. C. Marling and R. Bunescu, “The OhioT1DM dataset for blood glucose level prediction: Update 2020,” in *CEUR Workshop Proceedings*, 2020.
2. S. Pitoglou, “Machine Learning in Healthcare,” in *Quality Assurance in the Era of Individualized Medicine*, 2020, pp. 92–109. doi: 10.4018/978-1-7998-2390-2.ch004.
3. M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package),” *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018, doi: 10.1016/j.neucom.2018.03.067.
4. “Using Categorical Data with One Hot Encoding.” Accessed: Oct. 10, 2023. [Online]. Available: <https://kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding>
5. G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 30, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bd9eb6b76fa-Abstract.html
6. I. Bhattacharyya, “Feature Selection (Boruta /Light GBM/Chi Square)-Categorical Feature Selection,” Medium. Accessed: Aug. 30, 2023. [Online]. Available: <https://medium.com/@indreshbhattacharyya/feature-selection-categorical-feature-selection-boruta-light-gbm-chi-square-bf47e94e2558>
7. M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 1st ed. 2013, Corr. 2nd printing 2018 edition. New York: Springer, 2013.