ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ
ΙΝΣΤΙΤΟΥΤΟ ΧΗΜΙΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΒΙΟΕΠΙΧΕΙΡΕΙΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη μοντέλου μηχανικής μάθησης για την ανίχνευση νεφρικών λίθων με την πλατφόρμα Isalos Analytics Platform»

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΓΕΩΡΓΙΟΣ ΚΟΝΤΟΠΙΔΗΣ, ΔΙΕΥΘΥΝΤΗΣ ΕΡΓΑΣΤΗΡΙΟΥ ΒΙΟΧΗΜΕΙΑΣ, ΤΜΗΜΑ ΚΤΗΝΙΑΤΡΙΚΗΣ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΘΕΣΣΑΛΙΑΣ

ΤΕΧΝΙΚΟΣ ΣΥΜΒΟΥΛΟΣ: ΑΝΤΡΕΑΣ ΑΦΑΝΤΙΤΗΣ, MANAGING DIRECTOR, NOVAMECHANICS LTD.

ΜΙΧΑΗΛ-ΠΑΝΑΓΙΩΤΗΣ ΠΑΥΛΙΔΗΣ
Α.Μ. 00119
ΛΑΡΙΣΑ, 2023

UNIVERSITY OF THESSALY
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF BIOCHEMISTRY AND BIOTECHNOLOGY

NATIONAL HELLENIC RESEARCH FOUNDATION
INSTITUTE OF CHEMICAL BIOLOGY

INTERINSTITUTIONAL PROGRAM OF POSTGRADUATE STUDIES
IN
BIOENTREPRENEURSHIP

MASTER THESIS

«Development of a machine learning model capable of detecting the presence of kidney stones using Isalos Analytics Platform»

SUPERVISOR: GEORGIOS KONTOPIDIS, HEAD OF BIOCHEMISTRY LAB, VETERINARY SCHOOL, UNIVERSITY OF THESSALY

TECHNICAL ADVISOR: ANTREAS AFANTITIS, MANAGING DIRECTOR, NOVAMECHANICS LTD.

MICHAIL-PANAGIOTIS PAVLIDIS
A.M. 00119
LARISSA,GREECE 2023

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στο

ΒΙΟΕΠΙΧΕΙΡΕΙΝ

που απονέμει το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, σε συνεργασία με την εταιρεία Nova-mechanics LTD.

Εγκρίθηκε την Παρασκευή, 30/06/2023, από την τριμελή εξεταστική επιτροπή:

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	**ΒΑΘΜΙΔΑ**	**ΥΠΟΓΡΑΦΗ**
Γεώργιος Κοντοπίδης	Καθηγητής	
Σωτήριος Βασιλειάδης	Επίκουρος Καθηγητής	
Σπυρίδων Ζωγράφος	Ερευνητής Α	

Ευχαριστίες

Καταρχήν, θα ήθελα να ευχαριστήσω τον υπεύθυνο καθηγητή μου κ. Γεώργιο Κοντοπίδη για την άριστη συνεργασία μας καθ' όλη τη διάρκεια συγγραφής της μεταπτυχιακής μου διατριβής. Ιδιαίτερο ευχαριστώ για την άμεση ανταπόκριση του στα γραφειοκρατικά ζητήματα που ανέκυψαν κατά την επαφή μου με τα διαγνωστικά εργαστήρια για λήψη δεδομένων, που βοήθησαν σημαντικά στην εκπλήρωση της διπλωματικής μου.

Θα ήθελα επίσης να ευχαριστήσω των κ. Αντρέα Αφαντίτη και την εταιρεία Novamechanics Ltd. για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα, το οποίο πέραν από εξαιρετικά ενδιαφέρον, μου έδωσε τη δυνατότητα να αναπτύξω δεξιότητες και να αποκτήσω γνώση, που δίχως την ενασχόληση μου με το παρόν θέμα δε θα είχα την ευκαιρία να λάβω.

Ένα πολύ μεγάλο ευχαριστώ ανήκει στην κ. Δήμητρα Βάρσου, μέλος της εταιρείας Novamechanics, δίχως την πολύτιμη βοήθεια της οποίας δε θα μπορούσε να πραγματοποιηθεί η συγκεκριμένη διπλωματική. Η Δήμητρα διέθεσε τον πολύτιμο χρόνο της ώστε να με εκπαιδεύσει στην πλατφόρμα Isalos και να απαντήσει στα χιλιάδες ερωτήματα μου που ανέκυπταν όσο προχωρούσα βαθύτερα στον κόσμο του machine learning. Δήμητρα, θα είμαι πάντα ευγνώμων.

Ευχαριστώ τους αγαπημένους μου φίλους και την κοπέλα μου, οι οποίοι ήταν πάντα εκεί για να με ακούσουν και να με στηρίξουν, καθ΄ όλη τη διάρκεια του προηγούμενου, πολύ δύσκολου και απαιτητικού για εμένα έτους.

Κλείνοντας, δε θα μπορούσα να μην ευχαριστήσω εκείνους που με έμαθαν να κυνηγώ τα όνειρα μου και ακούραστα αγωνίζονται ώστε να με βοηθήσουν να τα κάνω πραγματικότητα. Ένα τεράστιο ευχαριστώ στον πατέρα μου Κώστα, τη μητέρα μου Ζέτα και τον μικρό μου αδερφό, Γιώργο.

Στον πατέρα μου, Κώστα

Στη μητέρα μου, Ζέτα

Στον αδερφό μου, Γιώργο

Περιεχόμενα

Περίληψη

Στόχος της παρούσας μεταπτυχιακής διατριβής είναι η δημιουργία και η εκπαίδευση ενός μοντέλου μηχανικής μάθησης, το οποίο θα είναι σε θέση να προβλέπει την ύπαρξη ή μη νεφρικών λίθων σε έναν άνθρωπο, χρησιμοποιώντας παραμέτρους που μετρούνται κατά τη διάρκεια μιας τυπικής εξέτασης ούρων.

Για τη δημιουργία του συγκεκριμένου μοντέλου χρησιμοποιείται η πλατφόρμα Isalos Analytics, την οποία έχει αναπτύξει η εταιρεία Novamechanics LTD. Η συγκεκριμένη πλατφόρμα χρησιμοποιεί μερικούς από τους πιο σύγχρονους αλγορίθμους επεξεργασίας δεδομένων και δημιουργίας μοντέλων μηχανικής μάθησης, χωρίς την ανάγκη ο χρήστης της να έχει γνώσεις προγραμματισμού.

Για τη δημιουργία του μοντέλου χρησιμοποιήθηκε ένα σετ δεδομένων από το διαδίκτυο, ενώ για τον έλεγχο της αποτελεσματικότητας του, συλλέχθηκαν πραγματικά δεδομένα από διαγνωστικό εργαστήριο της ελληνικής αγοράς. Τα αποτελέσματα αυτού του ελέγχου είναι άκρως ενθαρρυντικά και πιθανά να πυροδοτήσουν περαιτέρω ανάπτυξη του μοντέλου στο εγγύς μέλλον.

Το μοντέλο αυτό δύναται να χρησιμοποιηθεί μελλοντικά από διαγνωστικά εργαστήρια, με στόχο την άμεση πρόβλεψη ύπαρξης νεφρικών λίθων από τα υπό εξέταση δείγματα ούρων των ασθενών.

Λέξεις-Κλειδιά

Machine Learning, Isalos Analytics Platform, Μοντελοποίηση, Νεφρικοί Λίθοι

Abstract

The main purpose of this thesis is to create and train a machine learning model, which will be able to predict the presence or absence of kidney stone (or stones) in a human being, by utilizing some simple parameters which are measured during a typical urine sample testing.

To create this model, the Isalos Analytics Platfrom by Novamechanics Ltd. was used. This particular platform offers some of the most used data processing and machine learning algorithms, without the precondition that its user knows programming.

In order to create this model, an initial dataset that was found online was used. To test its applicability, real world data from a Greek diagnostics lab were collected. The results of this testing are very encouraging and they could possibly lead to further development of this model in the not distant future.

This model could be used in the real world by diagnostic labs in the future, so they can provide their patients with a much faster prediction about having or not a kidney stone.

Keywords

Machine learning, Isalos Analytics Platform, Modelling, Kidney stones

Σκοπός-Σύνοψη Διπλωματικής Εργασίας

Ο σκοπός της συγκεκριμένης μεταπτυχιακής διατριβής είναι η χρήση των πανίσχυρων εργαλείων που προσφέρει η πλατφόρμα Isalos για την ανάπτυξη μοντέλων μηχανικής μάθησης. Πιο συγκεκριμένα, το μοντέλο το οποίο επιλέχθηκε να δημιουργηθεί, θα βασίζεται σε δεδομένα ούρων, ασθενών και μη ατόμων, με στόχο την ανίχνευση ύπαρξης νεφρικών λίθων.

Πολλοί ήταν οι λόγοι που συντέλεσαν στην επιλογή ανάπτυξης του μοντέλου για το συγκεκριμένο λόγο. Πρώτος εξ αυτών ήταν η εύρεση ενός πολύ ικανοποιητικού σετ δεδομένων, που όπως θα δούμε παρακάτω, βοήθησε τόσο στην ανάπτυξη ενός μοντέλου με πολύ ενθαρρυντικά αποτελέσματα, όσο και στην ανάδειξη της χρησιμότητας της πλατφόρμας ως ένα εργαλείο που μπορεί να βοηθήσει ουσιαστικά, στην περαιτέρω εξέλιξη διαφόρων επιστημονικών κλάδων, πέραν αυτών στους οποίους αξιοποιείται ήδη.

Πολύ σημαντικό ρόλο στην επιλογή του μοντέλου ανάπτυξης αποτέλεσε η συμπτωματολογία της συγκεκριμένης πάθησης. Ως γνωστόν, η ύπαρξη νεφρικών λίθων και συγκεκριμένα, η μετακίνηση τους κατά μήκους του ουροποιητικού συστήματος, προκαλεί οξύ πόνο στον ασθενή. Επομένως, είναι μεγάλη η ανάγκη για εύρεση νέων, αξιόπιστων λύσεων στο συγκεκριμένο ιατρικό ζήτημα, με στόχο τη βελτίωση της ποιότητας ζωής των ασθενών.

Όπως θα δούμε παρακάτω, η δημιουργία αυτών των λίθων οφείλεται σε πολλούς παράγοντες, εξωγενείς και μη. Οι λίθοι μπορούν να παραμείνουν προσκολλημένοι για χρόνια, χωρίς να προκαλούν συμπτώματα στο φορέα τους. Κατά τον ετήσιο προληπτικό έλεγχο που οι περισσότεροι πλέον πραγματοποιούν στην εποχή μας, συνήθως ο ιατρός συνταγογραφεί, μεταξύ άλλων, και ουρολογικές εξετάσεις. Τα δεδομένα αυτών των εξετάσεων επιχειρεί να εκμεταλλευτεί η συγκεκριμένη μεταπτυχιακή διατριβή, με στόχο τη δημιουργία πρόβλεψης της ύπαρξης νεφρικών λίθων, από ένα απλό τεστ ούρων.

Τέλος, σημαντικό ρόλο στην τελική επιλογή, έπαιξε και η πιθανότητα εκμετάλλευσης ενός τέτοιου μοντέλου επιχειρηματικά. Αν τα αποτελέσματα είναι τα επιθυμητά, πολλά διαγνωστικά κέντρα και εργαστήρια πιθανόν να θελήσουν να χρησιμοποιήσουν το συγκεκριμένο μοντέλο, προκειμένου να προσφέρουν μια επιπλέον υπηρεσία στους ασθενείς τους.

Κεφάλαιο 1: Εισαγωγή

Στην εποχή μας, η συμβολή της τεχνητής νοημοσύνης γίνεται όλο και περισσότερο έκδηλη σε διάφορους τομείς της καθημερινότητας μας. Από τις προτεινόμενες διαφημίσεις που όλοι συναντάμε καθημερινά κατά την περιήγηση μας στο διαδίκτυο, οι οποίες βασίζονται στις αναζητήσεις που πραγματοποιούμε και στο συμπέρασμα στο οποίο διάφοροι αλγόριθμοι καταλήγουν για τα ενδιαφέροντα μας, μέχρι τις πολλά υποσχόμενες τεχνολογίες αυτόματης οδήγησης, που στόχο έχουν να αλλάξουν μια για πάντα τις παγκόσμιες μεταφορές στο εγγύς μέλλον.

Η μηχανική μάθηση αποτελεί έναν εκ των σημαντικότερων τομέων τεχνητής νοημοσύνης (για συντομία AI: Artificial Intelligence). Μοντέλα μηχανικής μάθησης κάνουν την εμφάνιση τους σε πολυάριθμους τομείς της καθημερινότητας, με πολύ ενθαρρυντικά αποτελέσματα. Ο τομέας της υγείας ίσως να αποτελεί τον σημαντικότερο εξ αυτών, καθώς μοντέλα μηχανικής μάθησης έχουν αρχίσει να εφαρμόζονται τόσο για πρόβλεψη εμφάνισης ασθενειών, όσο και για την πρόγνωση της εξέλιξης αυτών. Εξαιρετικά σημαντική φαίνεται να είναι η συμβολή των μοντέλων μηχανικής μάθησης σε πολυπαραγοντικές ασθένειες, όπως ο καρκίνος [1], όπου οι ασθενείς μπορούν να κατηγοριοποιηθούν σε ομάδες (clusters) [2], με βάση δεδομένα που αναλύει ένα τέτοιο μοντέλο. Η κατηγοριοποίηση αυτή βοηθά σημαντικά στην πρόβλεψη της εξέλιξης διαφόρων ασθενειών, με βάση την πρότερη γνώση που ανακτήθηκε για την εξέλιξη αυτών σε άτομα του ίδιου cluster. Επίσης, η ομαδοποίηση των ατόμων σε clusters με βάση το προφίλ προδιάθεσης τους, αναμένεται να επιταχύνει σημαντικά τις προσπάθειες έλευσης της εποχής της εξατομικευμένης ιατρικής [3], καθώς τα μοντέλα θα είναι σε θέση, μεταξύ άλλων, να προβλέπουν τις φαρμακοκινητικές ιδιότητες των ατόμων, με βάση το cluster στο οποίο ανήκουν.

Τα παραπάνω είναι μονάχα ένα μικρό δείγμα όσων η τεχνητή νοημοσύνη θα συνεισφέρει στον τομέα της ανθρώπινης υγείας, και πιθανά όχι μόνο.

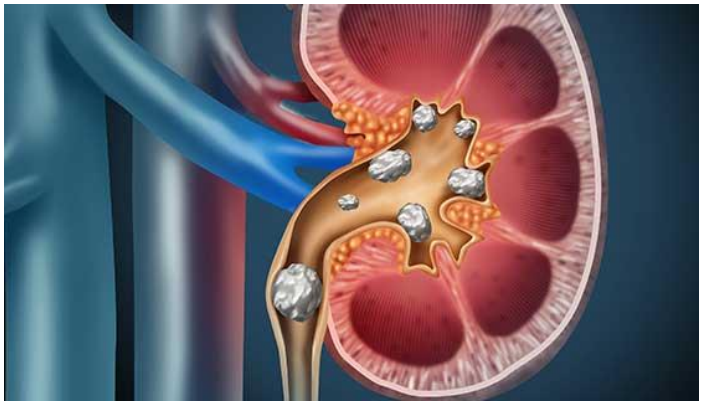
Έως τώρα, για την ανάπτυξη μοντέλων τεχνητής νοημοσύνης, απαιτούνταν εξειδικευμένες γνώσεις πληροφορικής και προγραμματισμού. Έτσι, πολλοί επιστήμονες δε μπορούσαν να επωφεληθούν από τέτοιου είδους μοντέλα, καθώς λίγοι ήταν αυτοί που συνδύαζαν τόσο τις βασικές επιστημονικές γνώσεις που θα απαιτούνταν, όσο και την ικανότητα προγραμματισμού με τη χρήση κάποιας γλώσσας. Η εταιρεία **NovaMechanics Ltd.** Παρέχει πρόσβαση στην ανάπτυξη τέτοιων μοντέλων μηχανικής μάθησης, με τη χρήση ευρέως διαδεδομένων στο σχετικό πεδίο αλγορίθμων, σε επιστήμονες και όχι μόνο, με την πλατφόρμα **Isalos®** (βλ. https://isalos.novamechanics.com/).

Η πλατφόρμα Isalos προσφέρει τη δυνατότητα ανάπτυξης μοντέλων μηχανικής μάθησης σε ανθρώπους χωρίς καμία απολύτως γνώση προγραμματισμού, οι οποίοι μέσα από ένα απλό και φιλικό προς το χρήστη περιβάλλον, μπορούν να επεξεργάζονται, και στη συνέχεια να αξιοποιούν τα δεδομένα τους, με πανίσχυρα υπολογιστικά εργαλεία, μόλις με το πάτημα λίγων πλήκτρων στο πληκτρολόγιο και το ποντίκι του υπολογιστή τους.

Όπως αναφέρθηκε κατά την περιγραφή του σκοπού, στη συγκεκριμένη μεταπτυχιακή διατριβή θα χρησιμοποιηθεί η πλατφόρμα Isalos, με στόχο τη δημιουργία ενός μοντέλου μηχανικής μάθησης, το οποίο θα είναι σε θέση να προβλέπει, από ένα απλό τεστ ούρων, την ύπαρξη ή μη νεφρικών λίθων στον εκάστοτε εξεταζόμενο ασθενή.

1.1 Νεφρικοί Λίθοι-Ορισμός και τύποι λίθων

Τα ούρα περιέχουν διαλυμένα μέταλλα και διάφορα άλατα, τα οποία καταλήγουν σε αυτά μετά τον καθαρισμό του αίματος από τους νεφρούς. Όταν η συγκέντρωση τους είναι πολύ μεγάλη, προάγεται η δημιουργία λίθων [5]. Κατά την έναρξη του σχηματισμού τους, οι λίθοι είναι πολύ μικροί σε μέγεθος. Με την πάροδο όμως του χρόνου διογκώνονται, καθώς συσσωρεύουν όλο και περισσότερα μέταλλα και λοιπά άλατα. Κάποιες φορές, οι λίθοι παραμένουν προσκολλημένοι στους νεφρούς, χωρίς να δημιουργούν προβλήματα ή ενοχλήσεις στο άτομο που τους φέρει. Πολλές φορές όμως, αποκολλώνται από τους νεφρούς και μετακινούνται, μέσω του ουρητήρα, προς την ουροδόχο κύστη.



Εικόνα 1: Οπτική απεικόνιση της αποκόλλησης των λίθων από τους νεφρούς, με κατεύθυνση τον ουρητήρα. Μικρού μεγέθους λίθοι είναι πιθανό να περάσουν από τον ουρητήρα, χωρίς να προκαλέσουν συμπτώματα. Μεγαλύτερου μεγέθους λίθοι, ακουμπάνε στα τοιχώματα του ουρητήρα και προκαλούν οξύ πόνο στον ασθενή.
Πηγή: Johns Hopkins Medicine, Kindey Stones Overview: https://www.hopkinsmedicine.org/health/conditions-and-diseases/kidney-stones

Ο ουρητήρας αποτελεί ένα κανάλι, μέσω του οποίου τα ούρα που προκύπτουν από το φιλτράρισμα του αίματος, ταξιδεύουν από τους νεφρούς προς την ουροδόχο κύστη για αποθήκευση, και εν συνεχεία για αποβολή.

Στην περίπτωση όπου το μέγεθος του λίθου είναι τέτοιο, όπου δε μπορεί να μεταφερθεί με ευκολία μέσω του ουρητήρα, το «ταξίδι» της πέτρας προς την ουροδόχο κύστη είναι εξαιρετικά επώδυνο για το φορέα. Σε κάποιες ακραίες περιπτώσεις, ο λίθος μπορεί ακόμα και να φράξει εντελώς τον ουρητήρα, εμποδίζοντας έτσι τη διέλευση των ούρων.

Υπάρχουν τέσσερις τύποι νεφρικών λίθων, η δημιουργία καθενός εκ των οποίων παρατηρείται με διαφορετική συχνότητα. Οι τέσσερις αυτοί τύποι είναι οι λίθοι ασβεστίου, οι λίθοι στρουβίτη, οι λίθοι ουρικού οξέος και τέλος, οι πέτρες κυστίνης (εξαιρετικά σπάνιες, εμφανίζονται σε άτομα που πάσχουν από κυστινουρία [32].

Η συντριπτική πλειοψηφία των λίθων που δημιουργούνται στους ανθρώπινους νεφρούς είναι με βάση το ασβέστιο, σε ποσοστό 80% [5]. Οι λίθοι αυτοί, χωρίζονται με τη σειρά τους σε 2 υποκατηγορίες, τους λίθους οξαλικού ασβεστίου (CaC_2O_4), που είναι και η πιο κοινή μορφή, και τους λίθους φωσφορικού ασβεστίου ($Ca_3(PO_4)_2$).

Το οξαλικό οξύ ($C_2H_2O_4$) είναι ένα φυσικό συστατικό διαφόρων τροφών [6]. Συνεπώς, εισέρχεται στον ανθρώπινο οργανισμό μέσα από τη διατροφή. Εν συνεχεία, μεταφέρεται στο ήπαρ, προκειμένου να αποβληθεί φυσιολογικά με τα ούρα. Οι λίθοι οξαλικού ασβεστίου σχηματίζονται όταν τουλάχιστον μία από τις ακόλουθες συνθήκες ισχύει:

- Διατροφή πλούσια σε οξαλικό
- Χαμηλή πρόσληψη νερού και διαφόρων υγρών

Φυσιολογικά, το οξαλικό προσκολλάται στο διαθέσιμο ασβέστιο, δημιουργώντας μία διαλυτή δομή, η οποία αποβάλλεται από τον οργανισμό μέσω της ούρησης [6]. Όταν η πυκνότητα των ούρων είναι πολύ μεγάλη ή η παραγόμενη ποσότητα ούρων είναι μικρή λόγω μειωμένης πρόσληψης υγρών, το σύμπλοκο οξαλικού και ασβεστίου σχηματίζει κρυστάλλους, οι οποίοι οδηγούν στη δημιουργία των νεφρικών λίθων.

Εικόνα 2: Λίθος Οξαλικού Ασβεστίου
Πηγή: Medical News Today - What Do Kidney Stones Look Like?
https://www.medicalnewstoday.com/articles/kidney-stones-pictures

Διάφοροι παράγοντες αυξάνουν τον κίνδυνο σχηματισμού λίθων οξαλικού ασβεστίου. Μερικοί εξ αυτών είναι:

- Μειωμένη πρόσληψη νερού, που οδηγεί σε σταδιακή αφυδάτωση του οργανισμού
- Διατροφή πλούσια σε τροφές που περιέχουν υψηλές συγκεντρώσεις οξαλικού
- Παχυσαρκία
- Νόσος Dent (επηρεάζει τα εγγύς σωληνάρια του ήπατος, εμφανίζεται μόνο σε άντρες)
- Δυσλειτουργίες του θυρεοειδή αδένα (υπερθυρεοειδισμός)
- Φλεγμονώδης Νόσοι του Εντέρου (ελκώδης κολίτιδα, Νόσος του Crohn)
- Μετεγχειρητική περίοδος γαστρικού bypass (Γαστρική Παράκαμψη-συνήθως πραγματοποιείται για την αντιμετώπιση ακραίων περιπτώσεων παχυσαρκίας)

Αναφορικά με τους νεφρικούς λίθους φωσφορικού ασβεστίου, ο τρόπος σχηματισμού τους είναι παρόμοιος με τους λίθους οξαλικού ασβεστίου. Σχηματίζονται λόγω μεγάλης συγκέντρωσης φωσφόρου στα ούρα και υποβοηθούνται όταν τα ούρα έχουν αλκαλικό ph [18]. Όπως και στη δημιουργία των λίθων οξαλικού ασβεστίου, ισχύουν οι προαναφερθέντες παράγοντες αυξημένου κινδύνου.

Ο δεύτερος συχνότερος τύπος νεφρικού λίθου που σχηματίζεται στον ανθρώπινο οργανισμό είναι με βάση το ουρικό οξύ ($C_5H_4N_4O_3$). Οι λίθοι ουρικού οξέος σχηματίζονται όταν η συγκέντρωση ουρικού οξέος στα ούρα είναι πολύ μεγάλη για μεγάλο χρονικό διάστημα [34].

Το ουρικό οξύ αποτελεί παραπροϊόν του μεταβολισμού των τροφών και φυσιολογικά αποβάλλεται μέσω των ούρων από τον οργανισμό. Παράγεται από το μεταβολισμό των πουρινών ($C_5H_4N_4$), ενός ετεροκυκλικού αρωματικού δακτυλίου που περιέχεται σε τροφές πλούσιες σε πρωτεΐνη, όπως το κρέας και όλα τα παράγωγα του, καθώς και σε ορισμένα όσπρια και λαχανικά, όπως τα φασόλια, η φακή, το σπανάκι, το κουνουπίδι και τα σπαράγγια [35]. Πουρίνες περιέχονται επίσης και στα μανιτάρια.

Η αυξημένη συγκέντρωση ουρικού οξέος στα ούρα ενός ασθενή, αποτελεί ένδειξη για την ύπαρξη διαφόρων θεμάτων υγείας, όπως μεταβολικά σύνδρομα, σακχαρώδης διαβήτης, υπέρταση και ρευματοειδής αρθρίτιδα [34]. Αυξημένη συγκέντρωση ουρικού οξέος παρατηρείται συχνά υπέρβαρα άτομα.

Οι λίθοι ουρικού οξέος αποτελούν το 10% των συνολικών λίθων που εμφανίζονται στο ανθρώπινο είδος. Όπως και οι λίθοι οξαλικού ασβεστίου και σε αντίθεση με τους λίθους φωσφορικού ασβεστίου, οι λίθοι ουρικού οξέος σχηματίζονται συχνότερα σε όξινο περιβάλλον.

Από τα παραπάνω αναδεικνύεται η σημαντικότητα του pH στο υπό εξέταση ζήτημα του σχηματισμού νεφρικών λίθων. Όξινο pH προάγει το σχηματισμό λίθων οξαλικού ασβεστίου και ουρικού οξέος, ενώ αλκαλικό pH προάγει το σχηματισμό λίθων φωσφορικού ασβεστίου.

1.2 Τρόποι Αντιμετώπισης

Όπως ισχύει και στις περισσότερες ασθένειες, η καλύτερη θεραπεία είναι η πρόληψη. Ο καλύτερος τρόπος ώστε να προστατευθεί κανείς από την εμφάνιση νεφρικών λίθων είναι η ημερήσια κατανάλωση της απαιτούμενης ποσότητας νερού [7]. Φυσικά, δεν είναι εγγυημένη η αποφυγή δημιουργίας λίθων, καθώς σε αυτήν παίζουν ρόλο και άλλοι, περιβαλλοντικοί και γενετικοί παράγοντες.

Από τη στιγμή που θα γίνει η διάγνωση της ύπαρξης του λίθου, το πιθανότερο είναι πως ο θεράπων ιατρός θα προτείνει κάποιες μικρές, αλλά σημαντικές αλλαγές στην καθημερινότητα του πάσχοντα. Η πρώτη και κύρια αλλαγή που πρέπει κανείς να κάνει είναι να αυξήσει την ημερήσια πρόσληψη υγρών που λαμβάνει, κυρίως νερό. Το νερό κρατά τα ούρα όσο το δυνατόν πιο αραιωμένα, ώστε να αποφευχθεί η δημιουργία περαιτέρω λίθων.

Επιπλέον, συνιστάται η μειωμένη κατανάλωση αλατιού [8]. Η σύσταση αυτή δίνεται διότι, προκειμένου να αποβληθεί από τον οργανισμό το νάτριο, μεταφέρεται στους νεφρούς, από όπου εν συνεχεία προωθείται στα ούρα. Η μεταφορά αυτή υποβοηθάτε από το ασβέστιο. Έτσι, η αυξημένη πρόσληψη άλατος οδηγεί σε αυξημένη συγκέντρωση ασβεστίου στα ούρα, προωθώντας έτσι τη δημιουργία περαιτέρω λίθων. Αλλαγές συνιστώνται επίσης και στη συνολική διατροφή του πάσχοντα, καθώς προτείνεται να μειώσει την κατανάλωση σε τροφές πλούσιες σε πρωτεΐνη, όπως το κρέας, καθώς αυτές περιέχουν μεγάλη ποσότητα ασβεστίου και πουρινών, προωθώντας έτσι τη δημιουργία περαιτέρω λίθων [9].

Επίσης, είναι πολύ πιθανό ο θεράπων ιατρός να προτείνει τη μείωση της πρόσληψης τροφών με μεγάλη συγκέντρωση οξαλικού οξέος [10]. Τέτοιες τροφές είναι το σπανάκι, ο μαϊντανός, τα φύλλα παντζαριού, το μαρούλι κ.α. Όπως γίνεται εύκολα αντιληπτό, οι συγκεκριμένες τροφές, παρόλη τη μεγάλη συγκέντρωση τους σε οξαλικό, είναι απαραίτητες για μια σωστή διατροφή. Η επεξεργασία πριν την κατανάλωση τους έχει αποδειχθεί πως μειώνει, χωρίς φυσικά να εξαλείφει, τη συγκέντρωση οξαλικού. Πρακτικές όπως το μούλιασμα ή το βράσιμο των συγκεκριμένων τροφών, συνιστώνται, προκειμένου να μειωθεί η περιεκτικότητα του οξαλικού σε αυτές.

Φυσικά, υπάρχουν και φαρμακευτικά σκευάσματα, τα οποία βοηθούν στην ανακούφιση από τα συμπτώματα. Συμπληρώματα με κιτρικό κάλιο συνιστώνται, καθώς ανεβάζουν το pH των ούρων, δημιουργώντας έτσι μη ευνοϊκό περιβάλλον για το σχηματισμό λίθων οξαλικού ασβεστίου και ουρικού οξέος [11].

Θειαζιδικά διουρητικά, όπως η υδροχλωροθειαζίδη, περιορίζουν το ασβέστιο που εκκρίνεται στα ούρα, με στόχο την αποφυγή δημιουργίας περαιτέρω λίθων [12].

Άλφα-αναστολείς συνιστώνται από τους θεράποντες ιατρούς, όταν πλέον ο λίθος έχει αποκολληθεί από τα τοιχώματα και κατευθύνεται μέσω του ουρητήρα προς αποβολή [13]. Οι α-αναστολείς βοηθούν στη χαλάρωση των μυών του ουρητήρα, διευκολύνοντας έτσι τη διέλευση του λίθου και περιορίζοντας, όσο είναι δυνατόν, τα συμπτώματα πόνου του πάσχοντα. Χρήση α-αναστολέων γίνεται συνήθως για λίθους με διάμετρο όχι πάνω από 5 χιλιοστά.

Πολλές φορές και ανάλογα την εκάστοτε περίσταση, μπορεί να χρειαστεί να εφαρμοστούν συγκεκριμένες ιατρικές διαδικασίες, επεμβατικές ή μη, προκειμένου να διαλυθούν ή να αφαιρεθούν πλήρως οι λίθοι [14]. Η πιο συχνή εξ αυτών είναι η λιθοτριψία, κατά την οποία χρησιμοποιούνται κρουστικά κύματα, στόχος των οποίων είναι η διάλυση του λίθου και η κατάτμηση του σε μικρότερα κομμάτια, ώστε να αποβληθεί φυσικά από τον οργανισμό. Λιθοτριψία χρησιμοποιείται συνήθως για λίθους με μέγιστη διάμετρο τα 2 εκατοστά, που εδράζονται στους νεφρούς ή στο άνω μέρος του ουρητήρα.

Άλλη συχνή, επεμβατική μέθοδος, που χρησιμοποιείται για την αντιμετώπιση των λίθων είναι η ουρητηροσκοπία [15]. Κατά τη διαδικασία αυτή, ο χειρούργος εισάγει ένα μικρό “τηλεσκόπιο” μέσω του ουρητήρα του πάσχοντα, με στόχο να φτάσει στο σημείο που εδράζεται ο λίθος και, είτε να τον διασπάσει με τη χρήση ειδικού laser, είτε να τον αφαιρέσει εντελώς. Η απόφαση αυτή λαμβάνεται κατά περίπτωση.

Στις πολύ δύσκολες περιπτώσεις, επιλέγεται η μέθοδος της διαδερμικής νεφρολιθοτομής [16]. Η μέθοδος αυτή είναι καθαρά επεμβατική και περιλαμβάνει τη δημιουργία μια τομής στην πλάτη του ασθενούς και τη χρήση ενός νεφροσκοπίου, για την ακριβή εξακρίβωση της θέσης του λίθου και την πλήρη αφαίρεση του. Η μέθοδος αυτή χρησιμοποιείται κυρίως για μεγάλους λίθους, όταν αυτοί δε μπορούν να αποβληθούν με οποιονδήποτε άλλο τρόπο.

1.3 Το σετ δεδομένων

Το σετ δεδομένων το οποίο χρησιμοποιήθηκε για την αρχική δημιουργία του μοντέλου προήλθε από δεδομένα του εργαστηρίου του James S. Elliot M.D. της ουρολογικής κλινικής του πανεπιστημίου του Stanford [4], ενώ η λήψη του έγινε από τη σελίδα Kaggle, μία online πλατφόρμα με χιλιάδες σετ και εργαλείων ανάλυσης δεδομένων, τα οποία διατίθενται δωρεάν προς οποιονδήποτε ενδιαφερόμενο (Kaggle: Your Machine Learning and Data Science Community: https://www.kaggle.com/).

Το αρχικό σετ δεδομένων περιλάμβανε τις εξής παραμέτρους:

- Ειδικό βάρος ούρων (Specific Gravity of Urine), μέτρο της συγκέντρωσης διαλυτών ουσιών στα ούρα, όπως άλατα, πρωτεΐνες και διάφορα άλλα στοιχεία [17]. Μετριέται πάντα κατά τις γενικές εξετάσεις ούρων με στόχο τη διερεύνηση της λειτουργίας των νεφρών, τα επίπεδα ενυδάτωσης, αλλά και τη γενικότερη υγεία του οργανισμού. Η συνήθης τιμή του ειδικού βάρους ούρων είναι από 1,005 έως 1,030, αποτέλεσμα που εκφράζεται ως η συγκέντρωση των ούρων προς τη συγκέντρωση του καθαρού νερού. Υψηλότερες τιμές υποδεικνύουν αφυδάτωση ή γενικότερα προβλήματα στη νεφρική λειτουργία. Μειωμένες τιμές υποδεικνύουν υπερυδάτωση ή αδυναμία των νεφρών να συμπυκνώσει ως θα έπρεπε τα ούρα.

- pH, πολύ σημαντική παράμετρος που εξετάζεται κατά τις γενικές εξετάσεις ούρων [18]. Φυσιολογικές τιμές θεωρούνται όσες είναι ανάμεσα στο 5 και το 8, με τις πιο συνηθισμένες τιμές να είναι κοντά στο 6. Πολλοί είναι οι παράγοντες που μπορεί να επηρεάσουν το pH των ούρων, με τους συχνότερους να είναι η διατροφή του ατόμου,

όπως και διάφορες φαρμακευτικές αγωγές. Τιμές pH κάτω του 5 είναι υπόδειξη για την ύπαρξη μεταβολικών νόσων, ενώ pH άνω του 8 μπορεί να σημαίνει λοίμωξη του ουροποιητικού συστήματος. Λίθοι φωσφορικού ασβεστίου τείνουν να δημιουργούνται ευκολότερα σε αλκαλικό περιβάλλον, δηλαδή τιμές pH άνω του 7, ενώ λίθοι οξαλικού ασβεστίου και ουρικού οξέος σε όξινο pH.

- Ωσμομοριακότητα των ούρων (Urine Osmolarity), όπου εκφράζει τον αριθμό σωματιδίων διαλυμένης ουσίας ανά λίτρο διαλύτη [19]. Η μέτρηση της ωσμομοριακότητας είναι πολύ σημαντική, προκειμένου να εξακριβωθεί η δυνατότητα των νεφρών να κατανέμουν ορθώς τη συγκέντρωση των διαλυμένων ουσιών, στο σώμα και τα ούρα. Το φυσιολογικό εύρος τιμών για τα ούρα είναι τυπικά μεταξύ 150 και 1400 mOsm/lt, με μέσες τιμές περίπου μεταξύ 500 έως 800 mOsm/lt. Τιμές άνω του 1400 mOsm/lt υποδεικνύουν ανικανότητα των νεφρών να κατανέμουν σωστά τις συγκεντρώσεις διαφόρων στοιχείων στο σώμα, ενώ πολύ χαμηλές τιμές, δηλαδή κάτω των 150 mOsm/lt είναι ένδειξη υπερυδάτωσης ή αδυναμία των νεφρών να συγκεντρώσουν τα ούρα.

- Αγωγιμότητα ούρων (Urine Conductivity), που αποτελεί μέτρηση της συγκέντρωσης διάφορων ηλεκτρολυτών στα ούρα, όπως το κάλιο, το νάτριο και το χλώριο [20]. Συνήθης μονάδα μέτρησης είναι τα microsiemens/cm, μονάδα ηλεκτρικής αγωγιμότητας. Όσο μεγαλύτερη είναι η συγκέντρωση των ηλεκτρολυτών στα ούρα, τόσο μεγαλύτερη είναι η τιμή αγωγιμότητας. Διάφοροι παράγοντες όπως η διατροφή και φαρμακευτικές αγωγές επηρεάζουν τη συγκέντρωση των ηλεκτρολυτών στα ούρα. Δεν είναι μια μέτρηση που γίνεται συχνά, αλλά είναι εξαιρετικά χρήσιμη για τον θεράπων ιατρό όταν θέλει να ελέγξει την επίδραση μιας υπάρχουσας φαρμακευτικής αγωγής ή τη γενικότερη λειτουργία των νεφρών.

- Συγκέντρωση ουρίας στα ούρα (Urea Concentration in Urine), δείκτης της ορθής λειτουργίας των νεφρών [21]. Η ουρία αποτελεί παραπροϊόν του μεταβολισμού που παράγεται από τη διάσπαση των πρωτεϊνών και αποβάλλεται από τον οργανισμό διαλυμένη στα ούρα. Μονάδα μέτρησης της ουρίας είναι τα mmol/lt και οι σύνηθες τιμές είναι από 250 έως 650 mmol/lt. Τιμές μικρότερες του ανωτέρω εύρους αποτελούν ένδειξη υπερυδάτωσης ή δυσλειτουργίας των νεφρών, ενώ οι τιμές άνω των 650 mmol/lt μπορεί να αποτελούν ένδειξη για νεφρικά προβλήματα ή είναι αποτέλεσμα διατροφής υψηλής σε πρωτεΐνη. Η ουρία διαδραματίζει σπουδαίο ρόλο στη διατήρηση της ισορροπίας των αζωτούχων ενώσεων στο ανθρώπινο σώμα.

- Συγκέντρωση ασβεστίου στα ούρα (Calcium Concentration in Urine), δείκτης της συνολικής ποσότητας ασβεστίου που αποβάλλεται από το σώμα [22]. Όπως γίνεται εύκολα κατανοητό, αποτελεί εξαιρετικά σημαντικό δείκτη, μιας και η συντριπτική πλειοψηφία των νεφρικών λίθων που δημιουργούνται στον ανθρώπινο οργανισμό, έχουν ως βάση το ασβέστιο. Όπως και η ουρία, έτσι και το ασβέστιο μετράται σε mmol/L, με τις φυσιολογικές τιμές να κυμαίνονται ανάμεσα σε 2,5 και 7,5. Τιμές εκτός του παραπάνω εύρους αποτελούν ένδειξη για ύπαρξη διαφόρων νόσων, όπως υπο- και παρά- θυρεοειδισμός, νεφρικές νόσους κ.α.

- TRUE/FALSE κατηγοριοποίηση. Τα άτομα εκείνα του σετ, τα οποία έφεραν όντως λίθο ή λίθους, χαρακτηρίζονται ως TRUE, ενώ τα άτομα εκείνα στα οποία δεν ανιχνεύθηκε κανένας λίθος ως FALSE. Η συγκεκριμένη στήλη του σετ δεδομένων αποτελεί τη μεταβλητή την οποία θα κληθεί να προβλέψει στη συνέχεια το μοντέλο που θα δημιουργηθεί.

Παρόλο που οι παραπάνω παράμετροι είναι πολύ σημαντικοί για την ανίχνευση νεφρικών προβλημάτων αλλά και πιθανών λίθων, η ωσμομοριακότητα και η αγωγιμότητα δεν

αποτελούν μέρος μιας τυπικής ανάλυσης ούρων από τα ελληνικά διαγνωστικά εργαστήρια. Έτσι, αφαιρέθηκαν από τον τελικό σχεδιασμό και την ανάπτυξη του μοντέλου πρόβλεψης, ώστε αυτό να έχει πιθανότητες να αξιοποιηθεί μελλοντικά. Φυσικά, προκειμένου να γίνει η συγκεκριμένη αφαίρεση των δύο αυτών παραμέτρων, εξετάστηκε η επιρροή τους στη μοντελοποίηση. Πιθανή επιρροή των συγκεκριμένων μεταβλητών στο σχεδιασμό του μοντέλου θα εξεταστεί στην ενότητα «Αποτελέσματα».

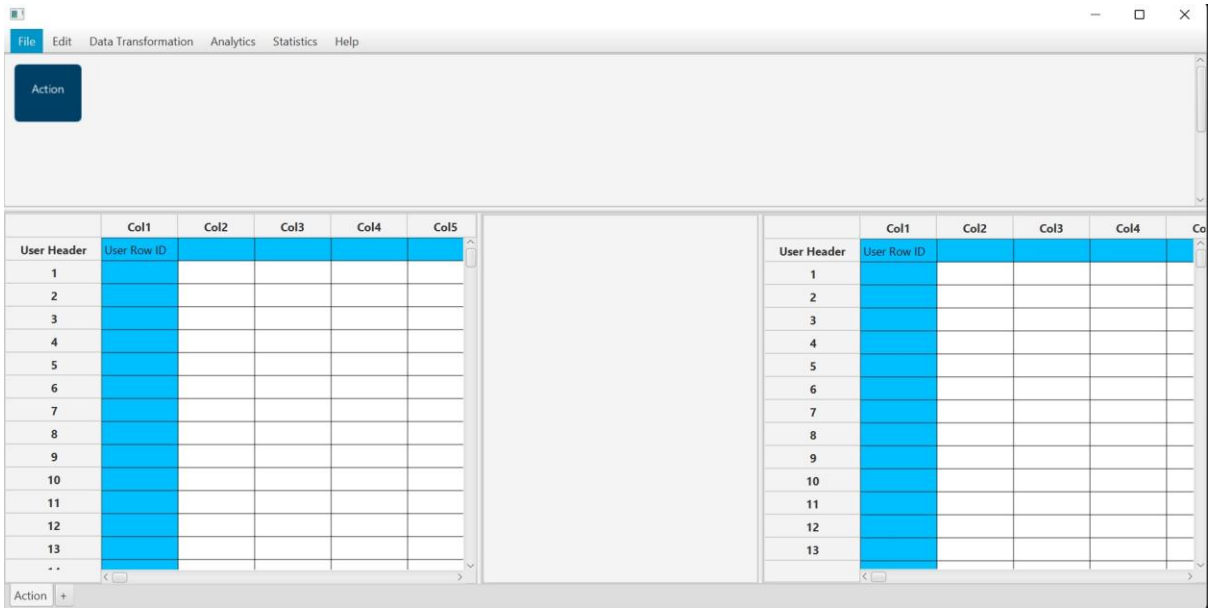
1.4 Η πλατφόρμα Isalos Analytics

Όπως αναφέρθηκε και παραπάνω, η μοντελοποίηση πραγματοποιήθηκε με τη χρήση των εργαλείων της πλατφόρμας Isalos, προϊόν της Novamechanics Ltd., εταιρείας με έδρα την Κύπρο, που προσφέρει λύσεις σε διάφορους επιστημονικούς κλάδους, με κυριότερο εξ αυτών τη νανοτεχνολογία και τη χημειοπληροφορική, κυρίως σε θέματα που αφορούν την ανάλυση και την αξιολόγηση δεδομένων.

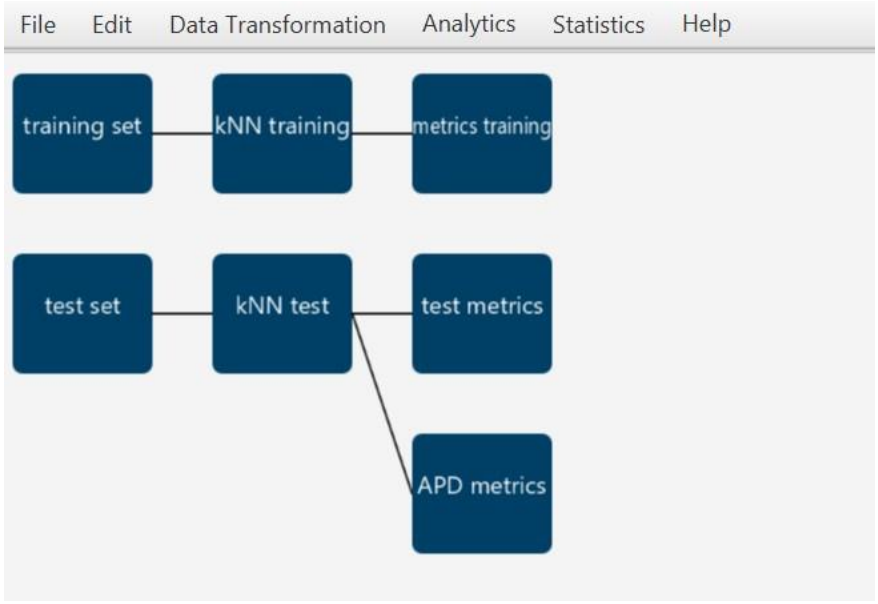
Η πλατφόρμα Isalos προσφέρει μία πληθώρα εργαλείων επεξεργασίας και μοντελοποίησης δεδομένων, στα οποία ένας απλός χρήστης χωρίς γνώσεις προγραμματισμού δεν είχε πρόσβαση μέχρι σήμερα [23]. Στο Isalos θα βρει κανείς αλγορίθμους μηχανικής μάθησης, στατιστικής ανάλυσης και οπτικοποίησης των δεδομένων του, χωρίς την ανάγκη συγγραφής κώδικα, παρά μόνο με το πάτημα λίγων πλήκτρων.

Το Isalos επιτρέπει την ανάπτυξη γραφημάτων ροής και επεξεργασίας δεδομένων τα οποία μετακινούνται μεταξύ των διαφορετικών καρτελών (tabs). Κάθε καρτέλα είναι ένας κόμβος (node) στον οποίο εισέρχονται δεδομένα, μετασχηματίζονται με σαφή και συγκεκριμένο τρόπο και στη συνέχεια εξέρχονται ώστε να προχωρήσουν προς μια νέα καρτέλα-κόμβο. Κατά την εκκίνηση του Isalos εμφανίζεται το αρχικό παράθυρο διαλόγου στο οποίο απεικονίζεται η πρώτη καρτέλα με το προκαθορισμένο όνομα Action (Εικόνα 3). Κάθε καρτέλα αποτελείται από τέσσερα επιμέρους πεδία: το πεδίο όπου εμφανίζεται αυτόματα το διάγραμμα ροής της ανάλυσης, το πεδίο εισαγωγής δεδομένων, το πεδίο παραμέτρων της εκάστοτε συνάρτησης και το πεδίο εξόδου δεδομένων ή αποτελεσμάτων. Αναλυτικότερα, το πεδίο εισόδου αποτελείται από ένα λογιστικό φύλλο (spreadsheet) μέσω του οποίου εισάγονται τα δεδομένα στον κόμβο και το πεδίο εξόδου από ένα λογιστικό φύλλο στο οποίο παρουσιάζονται τα αποτελέσματα ή τα μετασχηματισμένα δεδομένα. Το λογιστικό φύλλο εξόδου μπορεί να αποτελέσει την είσοδο για κάποια επόμενη καρτέλα.

Καθώς προστίθενται νέες καρτέλες στην πορεία ανάλυσης ή μοντελοποίησης των δεδομένων (πατώντας στο σύμβολο «+»), προστίθενται αυτόματα οι νέοι κόμβοι στο διάγραμμα ροής (workflow) στο αντίστοιχο πεδίο. Τα ονόματα των καρτελών δίνονται από τον χρήστη και μπορούν αντιστοιχούν στα ονόματα των βημάτων (Εικόνα 4).



Εικόνα 3: Το παράθυρο διαλόγου του Isalos Analytics Platform



Εικόνα 4: Το “workflow display” του Isalos

Η ανάλυση των δεδομένων με τη χρήση των παραπάνω διαδραματίζει σπουδαίο ρόλο στην πραγματοποίηση προβλέψεων για πιθανά ενδεχόμενα αποτελέσματα, καθώς επίσης και στη λήψη κρίσιμων αποφάσεων, με βάση αυτές τις προβλέψεις.

Όπως γίνεται εύκολα κατανοητό, η χρήση του Isalos για την πραγματοποίηση τέτοιων προβλέψεων, μπορεί να οδηγήσει σε δραματική μείωση του λειτουργικού κόστους σε διάφορους κλάδους, μειώνοντας ταυτόχρονα και το ενδεχόμενο ρίσκο που λαμβάνουν διάφορες επιχειρήσεις.

Παρακάτω, θα αναλυθούν τα εργαλεία εκείνα του Isalos, τα οποία χρησιμοποιήθηκαν για το σκοπό της συγκεκριμένης μεταπτυχιακής διατριβής, και οδήγησαν στο τελικό αποτέλεσμα της μοντελοποίησης.

1.5 Μετασχηματισμός τιμών σε τιμές z

Η μέθοδος που χρησιμοποιήθηκε για την κανονικοποίηση του σετ δεδομένων είναι η χρήση του Z-score. Η μέθοδος αυτή χρησιμοποιείται ευρέως σε εφαρμογές ανάλυσης δεδομένων και σε τεχνολογίες μηχανικής μάθησης και μοντελοποίησης [24]. Είναι μία τεχνική που μετατρέπει το σύνολο των δεδομένων, ώστε ο συνολικός μέσος όρος τους να είναι 0 και η τυπική απόκλιση 1.

Η διαδικασία της κανονικοποίησης Z-score πραγματοποιείται ακολουθώντας τα εξής βήματα:

1) Υπολογισμός του μέσου όρου και της τυπικής απόκλισης για κάθε μεταβλητή του σετ δεδομένων
2) Αφαίρεση του μέσου όρου από κάθε τιμή του σετ
3) Διαίρεση του αποτελέσματος της αφαίρεσης με την τυπική απόκλιση

Τα παραπάνω βήματα περιγράφονται από την εξίσωση:

(χ-μ)/σ (Εξίσωση 1), όπου:

χ=αρχική τιμή, μ=μέσος όρος δείγματος, σ=τυπική απόκλιση του δείγματος.

Οι τιμές που προκύπτουν από την παραπάνω διαδικασία ονομάζονται Z-scores και μπορούν να πάρουν θετικές, αλλά και αρνητικές τιμές. Αρνητικό Z-score δείχνει ότι η αρχική τιμή είναι κάτω του μέσου όρου του δείγματος για τη συγκεκριμένη μεταβλητή, ενώ θετικό Z-score δείχνει ότι η αρχική τιμή είναι πάνω από το μέσο όρο του δείγματος. Δεν αποκλείεται το Z-score για μια τιμή να είναι 0, αν αυτή η τιμή είναι ακριβώς ο μέσος όρος του δείγματος.

Η κανονικοποίηση με τη χρήση Z-score είναι ένα πολύ σημαντικό βήμα για τη δημιουργία του μοντέλου μηχανικής μάθησης, καθώς φροντίζει ώστε όλες οι τιμές να βρίσκονται σε παρόμοια επίπεδα, ώστε να μπορεί να γίνει σωστά η εκπαίδευση του μοντέλου.

1.6 Kennard-Stone Split

Όπως θα δούμε πιο αναλυτικά στην ενότητα «Μεθοδολογία», αμέσως μετά την κανονικοποίηση ακολουθεί ένα βήμα διαχωρισμού του σετ δεδομένων σε δύο νέα σετ. Το ένα εξ αυτών χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το δεύτερο για τον έλεγχο της εγκυρότητας του. Ο διαχωρισμός αυτός μπορεί να γίνει είτε τυχαία, είτε με τη χρήση του αλγορίθμου Kennard-Stone [25]. Ο αλγόριθμος οφείλει το όνομα του στους 2 επιστήμονες που τον πρωτοεφάρμοσαν.

Ο χρήστης του Isalos επιλέγει το ποσοστό του αρχικού σετ το οποίο θέλει να χρησιμοποιήσει ως training set. Στη συνέχεια, ο αλγόριθμος Kennard-Stone διαιρεί το αρχικό σετ σε 2 νέα. Αρχικά, ο αλγόριθμος βρίσκει τα 2 δείγματα του σετ (άτομα) που απέχουν το ένα περισσότερο από το άλλο, αναφορικά με τις τιμές στις διάφορες μεταβλητές τους. Αφού εντοπίσει τα 2 πιο απομακρυσμένα δείγματα, σχηματίζει δύο σετ, στα οποία τα 2 αυτά δείγματα είναι οι πρώτες τιμές των 2 σετ. Εν συνεχεία, βρίσκει το δείγμα εκείνο που είναι πιο κοντά σε ένα από τα 2 αρχικά δείγματα, και το κατατάσσει στο σετ που εμπεριέχει το άλλο, μακρινό δείγμα.

Η παραπάνω διαδικασία συνεχίζεται, έως ότου δημιουργηθούν τα τελικά σετ. Ακολουθώντας

αυτή τη λογική, ο αλγόριθμος δημιουργεί 2 νέα σετ, τα οποία είναι αντιπροσωπευτικά του αρχικού σετ, διατηρώντας την ποικιλομορφία των δειγμάτων.

Ο συγκεκριμένος αλγόριθμος χρησιμοποιείται ευρέως κατά τη δημιουργία μοντέλων μηχανικής μάθησης, καθώς είναι πολύ σημαντικό τα 2 σετ που προκύπτουν από το αρχικό, να είναι αντιπροσωπευτικά των 2 κλάσεων, ώστε το μοντέλο να εκπαιδευτεί όσο το δυνατόν καλύτερα και να κάνει όσο το δυνατόν, ασφαλέστερες προβλέψεις.

1.7 Προσεγγίσεις μηχανικής μάθησης

Για την επίλυση προβλημάτων μηχανικής μάθησης, υπάρχουν 3 βασικές διαφορετικές προσεγγίσεις: Επιβλεπόμενη μάθηση (Supervised Learning), Μη επιβλεπόμενη μάθηση (Unsupervised Learning) και ενισχυμένη μάθηση (Reinforced Learning) [33].

Κατά την επιβλεπόμενη μάθηση, τα δεδομένα που θα εκπαιδεύσουν το μοντέλο είναι κατηγοριοποιημένα, και κάθε αριθμητική τιμή αποτελεί την ένδειξη για το άτομο στη συγκεκριμένη κατηγορία. Στόχος της επιβλεπόμενης μάθησης είναι η εκπαίδευση ενός μοντέλου βάσει των σχέσεων ανάμεσα στις κατηγορίες και τις αριθμητικές τιμές τους, ώστε αυτό να είναι σε θέση να προβλέπει πιθανά αποτελέσματα σε νέα, εντελώς καινούργια δεδομένα. Το μοντέλο εκπαιδεύεται να κάνει αυτές τις προβλέψεις μέσα από διάφορους αλγορίθμους.

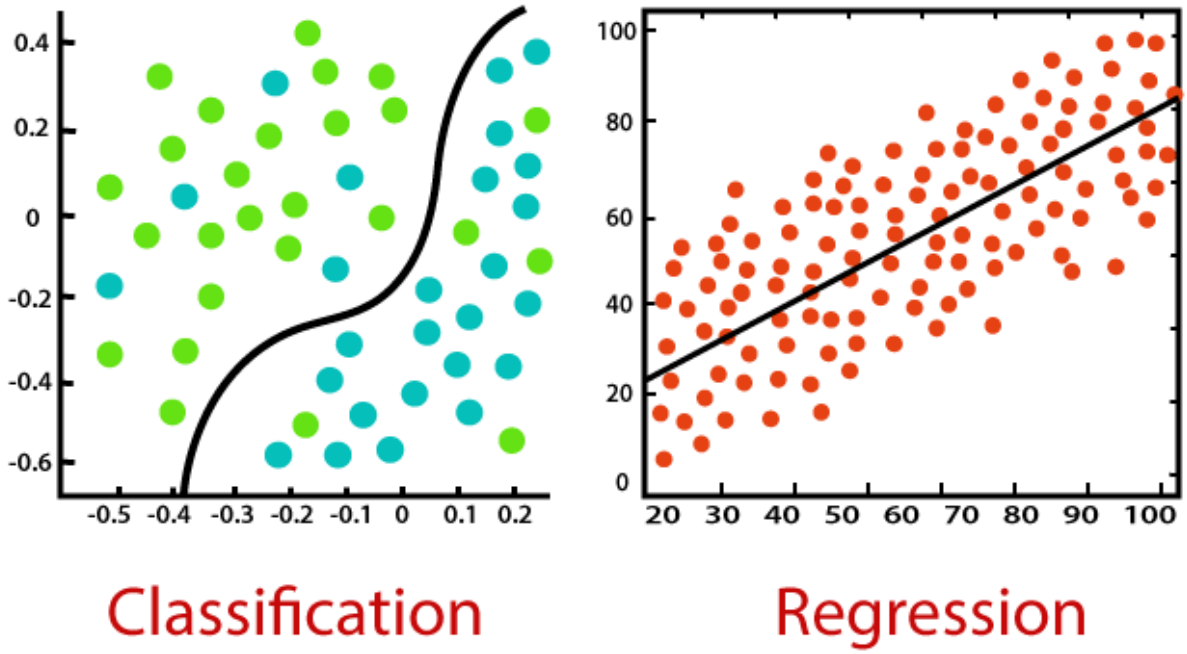
Ο όρος μη επιβλεπόμενη μάθηση περιλαμβάνει τη δημιουργία μοντέλων από μη κατηγοριοποιημένα δεδομένα. Στο συγκεκριμένο τύπο μάθησης, τα μοντέλα προσπαθούν να διακρίνουν σχέσεις και συσχετισμούς μεταξύ των δεδομένων, με στόχο την κατηγοριοποίηση των ατόμων και τελικώς, τη δημιουργία προβλέψεων.

Κατά την ενισχυμένη μάθηση, τα μοντέλα εκπαιδεύονται ώστε να αλληλεπιδρούν με το περιβάλλον και μαθαίνουν από αυτό, με στόχο την όσο το δυνατόν καλύτερη πρόβλεψη. Το μοντέλο αλληλεπιδρά με το περιβάλλον, λαμβάνει ανατροφοδότηση για αυτήν την αλληλεπίδραση και προσαρμόζεται αναλόγως αυτής, ώστε να παράγει τα καλύτερα δυνατά αποτελέσματα σε βάθος χρόνου.

Καθένας από αυτούς τους τρόπους μηχανικής μάθησης είναι ιδιαίτερα χρήσιμος στα διαφορετικά προβλήματα που ανακύπτουν. Για τις ανάγκες της συγκεκριμένης διπλωματικής, ο αλγόριθμος που θα χρησιμοποιηθεί ανήκει στην κατηγορία της επιβλεπόμενης μάθησης, καθώς τα δεδομένα εκπαίδευσης του μοντέλου ανήκουν σε σαφείς κατηγορίες.

1.8 Classification vs Regression

Τα προβλήματα μηχανικής μάθησης συνήθως χωρίζονται σε 2 κατηγορίες [26]. Η πρώτη κατηγορία έχει ως στόχο την πρόβλεψη κάποιας συγκεκριμένης αριθμητικής τιμής, όπως για παράδειγμα η θερμοκρασία ή η τιμή των ακινήτων σε μια περιοχή μελλοντικά. Τα συγκεκριμένα προβλήματα ανήκουν στην κατηγορία Regression, όπου στόχος είναι η αξιοποίηση σχετικών ή μη τιμών, για τη δημιουργία μια πρόβλεψης για μια συγκεκριμένη αριθμητική τιμή.



Εικόνα 5: Σχηματική απεικόνιση των διαφορών ανάμεσα στις 2 κατηγορίες
Πηγή: https://www.javatpoint.com/regression-vs-classification-in-machine-learning

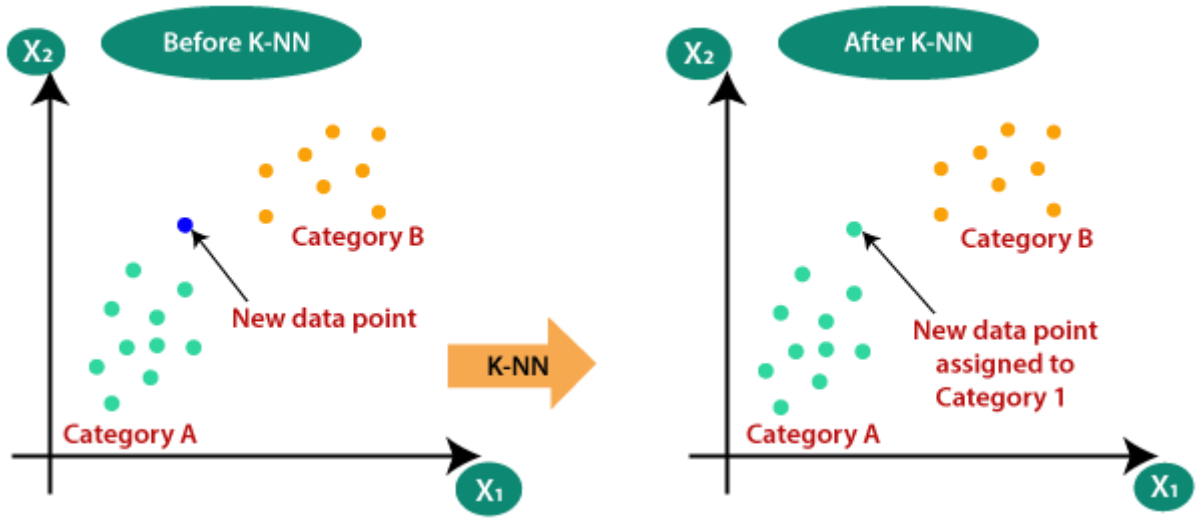
Η δεύτερη κατηγορία προβλημάτων έχει ως στόχο την αξιοποίηση των αρχικών δεδομένων για την κατηγοριοποίηση των δειγμάτων σε 2 ή περισσότερες τάξεις. Τα συγκεκριμένα προβλήματα καλούνται Classification και όπως εύκολα γίνεται αντιληπτό, σε αυτή την κατηγορία ανήκει και το πρόβλημα με το οποίο ασχολείται η συγκεκριμένη μεταπτυχιακή διατριβή.

1.9 Ο αλγόριθμος k-Nearest Neighbor

Η πλατφόρμα Isalos παρέχει διάφορους, ευρέως διαδεδομένους αλγορίθμους για τη δημιουργία μοντέλων μηχανικής μάθησης. Στην ενότητα αυτή, θα συζητηθεί ο αλγόριθμος ο οποίος χρησιμοποιήθηκε για τη δημιουργία του μοντέλου ενδιαφέροντος, ο k-Nearest Neighbor (kNN για συντομία) [27,28,29].

Ο αλγόριθμος βασίζεται στον υπολογισμό των ευκλείδιων αποστάσεων ανάμεσα σε δύο δείγματα, όπως αυτές ορίζονται με βάση τις τιμές τους στις εκάστοτε μεταβλητές, ώστε να διευκρινίσει τη μεταξύ τους σχετικότητα. Στην προκειμένη περίπτωση, στόχος του είναι να ομαδοποιήσει τα άτομα του δείγματος σε 2 κατηγορίες, αυτούς που θεωρεί πως έχουν νεφρικούς λίθους και αυτούς που δεν έχουν.

Μία από τις σημαντικότερες αποφάσεις που πρέπει να λάβει ο χρήστης κατά τη χρήση του συγκεκριμένου αλγορίθμου, είναι ο αριθμός των “γειτόνων” που θα επιλέξει για τη δημιουργία του μοντέλου. Για να το κάνει αυτό, ο χρήστης δοκιμάζει διαφορετικούς αριθμούς γειτόνων, ώστε να διευκρινίσει ποιος αριθμός δίνει το καλύτερο αποτέλεσμα. Συνήθως, μία τιμή ανάμεσα στο 3 και στο 5 είναι η ιδανική, καθώς σε αυτό το εύρος δεν υπερεκτιμούνται, αλλά ούτε και υποτιμούνται, οι σχέσεις μεταξύ των δειγμάτων. Ιδιαίτερα για τα classification προβλήματα, καλό είναι να χρησιμοποιείται μονός αριθμός γειτόνων, δηλαδή 3 ή 5, ώστε να αποφεύγεται η πιθανότητα “ισοπαλίας” και να γίνεται ορθότερα η κατηγοριοποίηση του προς εξέταση δείγματος [27].



Εικόνα 6: Σχηματική απεικόνιση του τρόπου λειτουργίας του αλγορίθμου kNN
Πηγή: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Αφού ο αλγόριθμος υπολογίσει τις αποστάσεις μεταξύ των σημείων και τις σχέσεις που αυτά έχουν μεταξύ τους, προβλέπει το αποτέλεσμα για το συγκεκριμένο πρόβλημα. Στην προκειμένη περίπτωση, αν για ένα συγκεκριμένο δείγμα, ο χρήστης επιλέξει k=3 και οι 2 ή 3 από αυτούς τους γείτονες έχουν νεφρικό λίθο (ή λίθους), ο αλγόριθμος προβλέπει ότι και αυτό το δείγμα θα ανήκει στη συγκεκριμένη κατηγορία, δηλαδή και αυτό το άτομο θα έχει λίθο. Στην πρόβλεψη της κλάσης των υπό εξέταση δειγμάτων συμμετέχει και ένας σταθμικός παράγοντας, ίσος με την αντίστροφη τιμή της απόστασης του κάθε γείτονα από το δείγμα προς πρόβλεψη. Με πιο απλά λόγια, όσο πιο μακρινό το δείγμα, τόσο λιγότερο συμμετέχει στην υπό εξέταση πρόβλεψη.

Μετά τη χρήση του αλγορίθμου kNN στο Isalos για τη μοντελοποίηση, ο χρήστης μπορεί να δει τους κοντινότερους γείτονες κάθε ατόμου. Χάρη σε αυτό, μπορεί να καταλάβει ποιοι γείτονες συνεισέφεραν περισσότερο στην τελική απόφαση της κατάταξης του δείγματος σε μία εκ των δύο (ή περισσότερων) κατηγοριών.

Η συγκεκριμένη μέθοδος, αν και απλή, είναι πολύ αποτελεσματική και αποδεικνύει τη σημαντικότητα των παραμέτρων που μετρούνται κατά τις διαγνωστικές εξετάσεις στην ανίχνευση της ύπαρξης νεφρικών λίθων, όπως θα δούμε στην ενότητα «Αποτελέσματα-Συζήτηση».

1.10 Έλεγχος εγκυρότητας του μοντέλου

Αφού δημιουργηθεί το μοντέλο μηχανικής μάθησης, πρέπει να αξιολογηθεί ως προς την αποτελεσματικότητα του. Ένας από τους πιο κοινούς αλλά και πιο ουσιώδης τρόπους αξιολόγησης είναι ο υπολογισμός της ευαισθησίας και της εξειδίκευσης του μοντέλου [30].

Η ευαισθησία εκφράζει το ποσοστό των αληθώς θετικών αποτελεσμάτων σε ένα μοντέλο. Στη συγκεκριμένη δηλαδή περίπτωση, η ευαισθησία είναι ένας δείκτης πληροφόρησης για το πόσα από τα άτομα που είχαν όντως νεφρικούς λίθους αναγνώρισε ως θετικά το μοντέλο. Η ευαισθησία υπολογίζεται από τον τύπο:

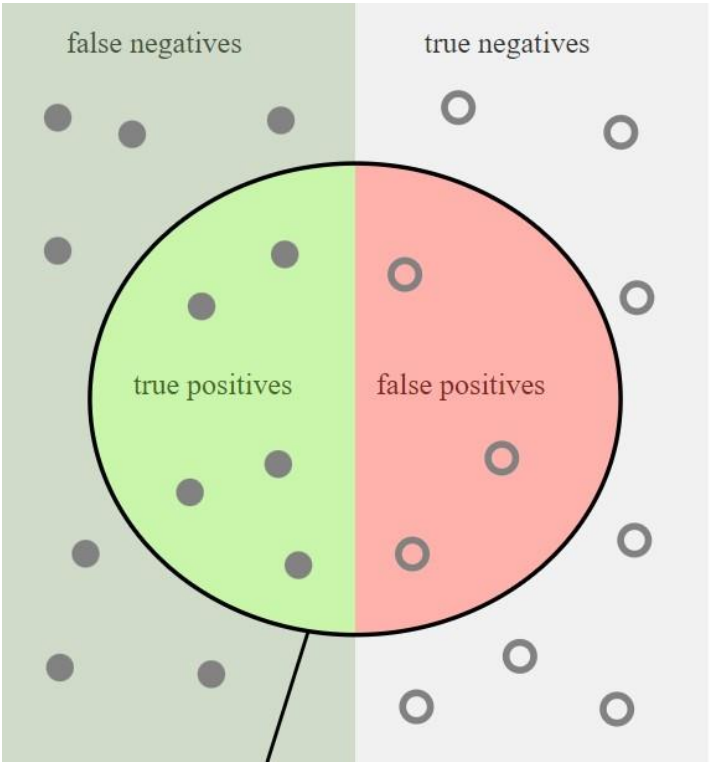
Ευαισθησία = TP / (TP + FN) (Εξίσωση 2), όπου:

TP=αριθμός των αληθώς θετικών δειγμάτων και FN=αριθμός των ψευδώς αρνητικών δειγμάτων.

Η εξειδίκευση από την άλλη, είναι το ακριβώς αντίθετο. Δηλαδή, εκφράζει το ποσοστό των αληθώς αρνητικών δειγμάτων που προβλέπει το μοντέλο ως αρνητικά. Συνεπώς, πόσα από τα άτομα που εξετάστηκαν και δεν είχαν νεφρικούς λίθους, αναγνώρισε όντως το μοντέλο ως αρνητικά. Η εξειδίκευση υπολογίζεται από τον τύπο:

Εξειδίκευση = TN / (TN + FP) (Εξίσωση 3), όπου:

TN=αριθμός των αληθώς αρνητικών δειγμάτων και FP=αριθμός των ψευδώς θετικών δειγμάτων.



Εικόνα 7: Σχηματική απεικόνιση ενός συνόλου δεδομένων μετά τη μοντελοποίηση. Εντός του κύκλου είναι τα άτομα τα οποία προβλέφθηκαν ορθώς, ενώ εκτός τα άτομα στα οποία η πρόβλεψη ήταν λανθασμένη. Από τα δεδομένα αυτά προκύπτει η ευαισθησία και η εξειδίκευση.
Πηγή: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Η ευαισθησία και η εξειδίκευση είναι πολύ σημαντικές παράμετροι για την αξιολόγηση ενός Classification μοντέλου και δη, του συγκεκριμένου, καθώς αποτελούν ένδειξη της ικανότητας του μοντέλου να προβλέπει ορθώς την ύπαρξη ή μη νεφρικών λίθων. Ένα καλό μοντέλο πρέπει να έχει υψηλό βαθμό ευαισθησίας και εξειδίκευσης.

Στη συνέχεια, υπολογίζεται το F1-score του μοντέλου. Το F1-score αξιοποιεί την ευαισθησία και την εξειδίκευση του μοντέλου, ώστε να αξιολογήσει το μοντέλο, ως προς την αποτελεσματικότητα του να κάνει προβλέψεις [31]. Το F1-score υπολογίζεται από τον τύπο:

F1-score = 2 * (εξειδίκευση * ευαισθησία) / (εξειδίκευση + ευαισθησία) (Εξίσωση 4)

Το F1-score παίρνει τιμές από 0 έως 1, με το 1 να δείχνει 100% εξειδίκευση και ευαισθησία, ενώ το 0 είναι το χειρότερο δυνατό αποτέλεσμα, κοινώς 0% εξειδίκευση και ευαισθησία του μοντέλου.

Τέλος, πολύ σημαντική παράμετρος για την απόδειξη της χρηστικότητας ενός μοντέλου είναι ο υπολογισμός της ακρίβειας πρόβλεψης αυτού. Η ακρίβεια (Accuracy) προβλέπεται μέσω της ακόλουθης εξίσωσης:

Ακρίβεια= TP+TN / TP+TF+FP+FN (Εξίσωση 5), όπου:

TP=αριθμός των αληθώς θετικών δειγμάτων, FN=αριθμός των ψευδώς αρνητικών δειγμάτων, TN=αριθμός των αληθώς αρνητικών δειγμάτων, FP=αριθμός των ψευδώς θετικών δειγμάτων.

Φυσικά, όλα τα παραπάνω τεστ γίνονται με τη βοήθεια του Isalos, καθώς παρέχει μία πληθώρα επιλογών για τη στατιστική ανάλυση των διαφόρων μοντέλων.

Τα αποτελέσματα των στατιστικών αυτών τεστ θα αναλυθούν περαιτέρω στην αντίστοιχη ενότητα της εργασίας.

Κεφάλαιο 2: Μεθοδολογία

Πρώτο μέλημα είναι να ελεγχθεί η αξιοπιστία του σετ δεδομένων από το Kaggle. Το αρχικό dataset θα παρατεθεί στο παράρτημα της παρούσας εργασίας. Τα δεδομένα αντιγράφονται από τον διαδικτυακό τόπο σε ένα φύλλο του Microsoft Excel (Εικόνα 8). Στη συνέχεια, στη στήλη Target, όλα τα 0 αντικαθίστανται από την λέξη FALSE, ενώ όλα τα 1 αντικαθίστανται από τη λέξη TRUE. Το σετ δεδομένων εμφανίζει τα αποτελέσματα απουσίας και παρουσίας νεφρικών λίθων σε δυαδική μορφή. Η αντικατάσταση αυτή γίνεται για ευκολία του χρήστη κατά τη μοντελοποίηση στην πλατφόρμα Isalos.

	A	B	C	D	E	F	G	H
1	User Row	Gravity	pH	Osmolarit	Conductiv	Urea	Calcium	Target
2	1	1.021	4.91	725	14	443	2.45	FALSE
3	2	1.017	5.74	577	20	296	4.49	FALSE
4	3	1.008	7.2	321	14.9	101	2.36	FALSE
5	4	1.011	5.51	408	12.6	224	2.15	FALSE
6	5	1.005	6.52	187	7.5	91	1.16	FALSE
7	6	1.02	5.27	668	25.3	252	3.34	FALSE
8	7	1.012	5.62	461	17.4	195	1.4	FALSE
9	8	1.029	5.67	1107	35.9	550	8.48	FALSE
10	9	1.015	5.41	543	21.9	170	1.16	FALSE
11	10	1.021	6.13	779	25.7	382	2.21	FALSE
12	11	1.011	6.19	345	11.5	152	1.93	FALSE
13	12	1.025	5.53	907	28.4	448	1.27	FALSE
14	13	1.006	7.12	242	11.3	64	1.03	FALSE
15	14	1.007	5.35	283	9.9	147	1.47	FALSE
16	15	1.011	5.21	450	17.9	161	1.53	FALSE
17	16	1.018	4.9	684	26.1	284	5.09	FALSE
18	17	1.007	6.63	253	8.4	133	1.05	FALSE
19	18	1.025	6.81	947	32.6	395	2.03	FALSE
20	19	1.008	6.88	395	26.1	95	7.68	FALSE

Εικόνα 8: Τα δεδομένα στο Excel

Εν συνεχεία, τα δεδομένα αντιγράφονται εκ νέου και επικολλώνται στο πρώτο φύλλο εργασίας της πλατφόρμας Isalos (Εικόνα 9).

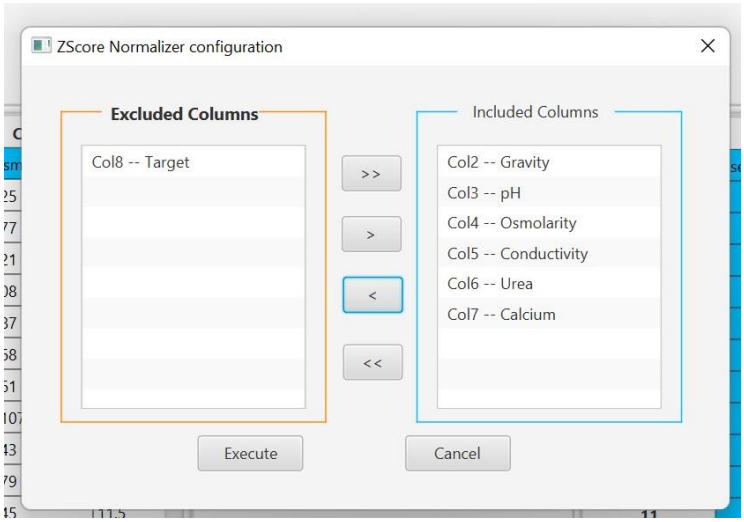
	Col1	Col2 (D)	Col3 (D)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (D)	Col8 (S)	Col9
User Header	User Row ID	Gravity	pH	Osmolarity	Conductivity	Urea	Calcium	Target	
1	1	1.021	4.91	725	14	443	2.45	FALSE	
2	2	1.017	5.74	577	20	296	4.49	FALSE	
3	3	1.008	7.2	321	14.9	101	2.36	FALSE	
4	4	1.011	5.51	408	12.6	224	2.15	FALSE	
5	5	1.005	6.52	187	7.5	91	1.16	FALSE	
6	6	1.02	5.27	668	25.3	252	3.34	FALSE	
7	7	1.012	5.62	461	17.4	195	1.4	FALSE	
8	8	1.029	5.67	1107	35.9	550	8.48	FALSE	
9	9	1.015	5.41	543	21.9	170	1.16	FALSE	
10	10	1.021	6.13	779	25.7	382	2.21	FALSE	
11	11	1.011	6.19	345	11.5	152	1.93	FALSE	
12	12	1.025	5.53	907	28.4	448	1.27	FALSE	
13	13	1.006	7.12	242	11.3	64	1.03	FALSE	
14	14	1.007	5.35	283	9.9	147	1.47	FALSE	

Εικόνα 9: Η επικόλληση των δεδομένων στην πλατφόρμα Isalos

Αφού τα δεδομένα έχουν φορτωθεί στην πλατφόρμα, ακολουθεί η κανονικοποίηση τους με τη μέθοδο του Z-score. Η κανονικοποίηση γίνεται μέσω του μενού του Isalos, και συγκεκριμένα επιλέγοντας:

Data Transformation >> Normalizers >> Z Score

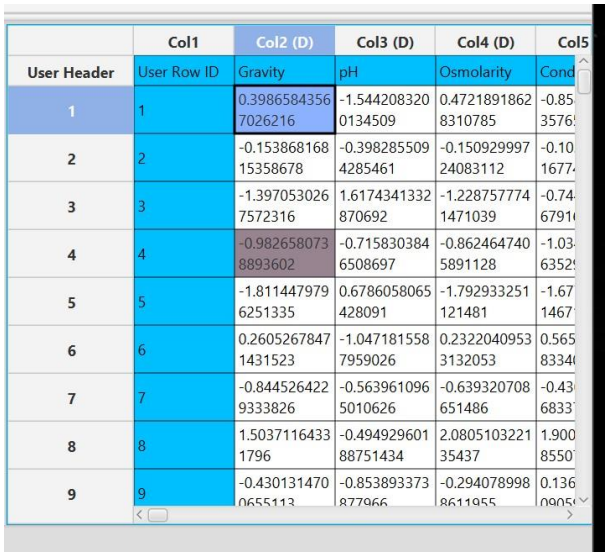
Μόλις ο χρήστης ακολουθήσει την παραπάνω διαδρομή, στην οθόνη του θα δει το ακόλουθο παράθυρο διαλόγου:



Εικόνα 10: Μετατροπή σε τιμές z

Στο παράθυρο αυτό, ο χρήστης καλείται να επιλέξει ποιες στήλες δεδομένων θα κανονικοποιηθούν, από το σύνολο των δεδομένων. Στη συγκεκριμένη περίπτωση, μιας και όλα τα δεδομένα έχουν αξία, εξαιρείται μόνο η στήλη Target, όπου δεν έχει αριθμητική τιμή, αλλά απεικονίζει μόνο την απουσία ή παρουσία νεφρικών λίθων.

Μετά την ολοκλήρωση της κανονικοποίησης, ο χρήστης θα δει στο δεξί μέρος της οθόνης το αποτέλεσμα της παραπάνω διαδικασίας (Εικόνα 11).

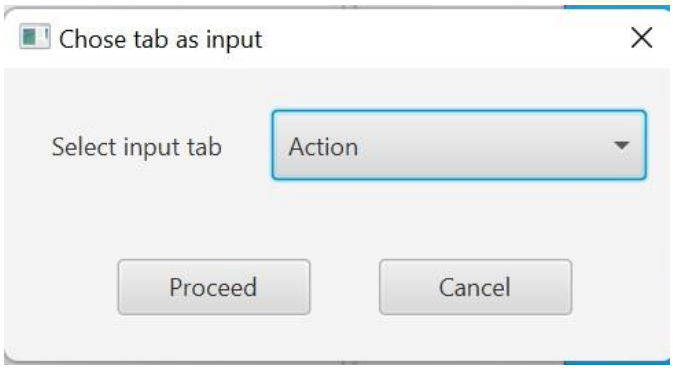


	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5
User Header	User Row ID	Gravity	pH	Osmolarity	Cond
1	1	0.39865843567026216	-1.5442083200134509	0.47218918628310785	-0.85 3576
2	2	-0.15386816815358678	-0.3982855094285461	-0.15092999724083112	-0.10 1677
3	3	-1.3970530267572316	1.6174341332870692	-1.2287577741471039	-0.74 6791
4	4	-0.9826580738893602	-0.7158303846508697	-0.8624647405891128	-1.03 6352
5	5	-1.8114479796251335	0.6786058065428091	-1.792933251121481	-1.67 1467
6	6	0.26052678471431523	-1.0471815587959026	0.23220409533132053	0.565 8334
7	7	-0.8445264229333826	-0.5639610965010626	-0.6393207085651486	-0.43 6833
8	8	1.503711643317796	-0.4949296018875143 4	2.080510322135437	1.900 8550
9	9	-0.4301314700655113	-0.8538933738779 66	-0.2940789988611955	0.136 0905

Εικόνα 11: Τα κανονικοποιημένα δεδομένα

Αφού έχει ολοκληρωθεί και η κανονικοποίηση, ο χρήστης δημιουργεί ένα νέο φύλλο στο Isalos, στο οποίο εισάγει τα κανονικοποιημένα δεδομένα. Κάνοντας δεξί κλικ στο νέο φύλλο, ο

χρήστης επιλέγει «Import from Spreadsheet», για να του εμφανιστεί ένα παράθυρο διαλόγου, στο οποίο καλείται να επιλέξει από που θα φορτωθούν τα δεδομένα στο φύλλο αυτό (Εικόνα 12).



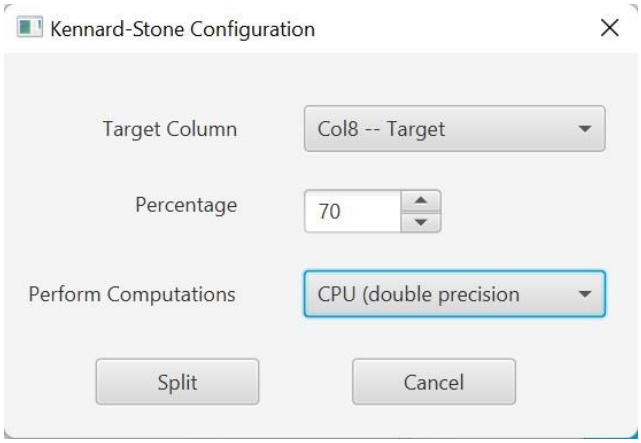
Εικόνα 12: Το παράθυρο διαλόγου για Import δεδομένων σε νέο φύλλο

Αφού τα κανονικοποιημένα δεδομένα φορτωθούν στο νέο φύλλο, ακολουθεί μία διαδικασία χωρίσματος του σετ δεδομένων σε δύο μέρη. Το πρώτο μέρος θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και καλείται «Training Set», ενώ το δεύτερο μέρος θα χρησιμεύσει για τον έλεγχο του μοντέλου, ώστε να αποσαφηνιστεί η προβλεπτική του ικανότητα. Το δεύτερο μέρος του σετ καλείται «Test Set».

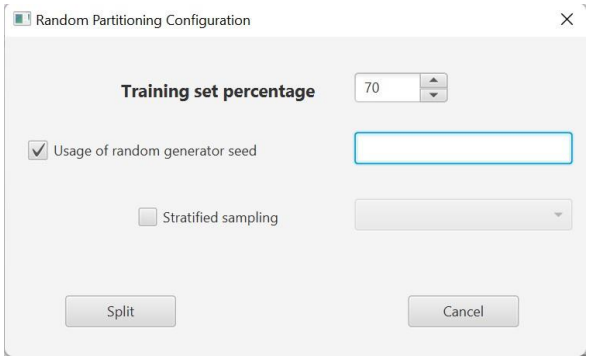
Στο Isalos, ο διαχωρισμός των διαφόρων σετ δεδομένων σε Training και Test, μπορεί να γίνει με δύο μεθοδολογίες. Η πρώτη επιλογή είναι η χρήση του αλγορίθμου Kennard-Stone, για τον οποίο μιλήσαμε στην εισαγωγή, ενώ η δεύτερη επιλογή είναι ο τυχαίος διαχωρισμός του δείγματος σε 2 σετ (Random Partitioning).

Στο συγκεκριμένο βήμα, δεν υπάρχει συγκεκριμένη επιλογή. Κατά το διαχωρισμό του σετ δεδομένων σε Training και Test, έγιναν δεκάδες δοκιμές, και με τις δύο μεθοδολογίες, ώστε να βρεθεί η φόρμουλα που δίνει τελικώς τα καλύτερα αποτελέσματα κατά τη μοντελοποίηση.

Ακολουθούν τα δύο παράθυρα διαλόγου που βλέπει ο χρήστης, όταν επιλέγει τις αντίστοιχες επιλογές (Εικόνα 13, Εικόνα 14).



Εικόνα 13: Kennard-Stone Split



Εικόνα 14: Random Partitioning

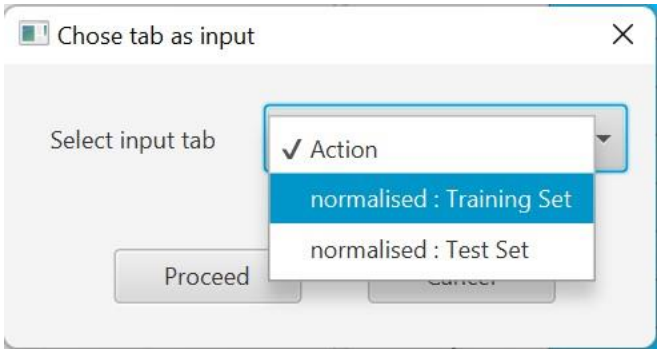
Στο βήμα του διαχωρισμού, συνήθως επιλέγεται ένα ποσοστό ανάμεσα σε 70-80% ως training set και 20-30% ως test set.

Ο διαχωρισμός του δείγματος φαίνεται στο δεξί μέρος της οθόνης, στο παράθυρο του outcome της πλατφόρμας Isalos. Ανάμεσα στα δύο δείγματα, εισάγεται μία κενή γραμμή, η οποία συμβολίζει το διαχωρισμό του αρχικού σετ σε training και test (Εικόνα 15).

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)
60	27	-0.153868168 15358678	2.1834923891 18167	-0.361443234 91783753	0.6280488449 824989	-1.458273738 1927942
61	3	-1.397053026 7572316	1.6174341332 870692	-1.228757774 1471039	-0.744921097 6791644	-1.260185367 1691847
62	75	0.2605267847 1431523	-0.481123302 96480493	1.1079391640 676672	1.8876542969 656758	0.3169028175 1878324
63	45	0.3986584356 7026216	-0.688217786 8054496	0.6827024239 601143	1.3082357890 534144	0.2711901165 133349
64						
65	6	0.2605267847 1431523	-1.047181558 7959026	0.2322040953 3132053	0.5650685723 833401	-0.109749058 53206798
66	7	-0.844526422 9333826	-0.563961096 5010626	-0.639320708 651486	-0.430019734 6833703	-0.544019718 0838272
67	11	-0.982658073 8893602	0.2229979420 9339036	-1.127711420 0621408	-1.173186951 3534448	-0.871627408 6228737
68	16	-0.015736517 197609223	-1.558014618 9361601	0.2995683313 8796257	0.6658370085 419943	0.1340520134 9698986
69	17	-1.535184677 7132092	0.8304750946 926162	-1.515055777 3878326	-1.563664641 4682296	-1.016384295 1401267

Εικόνα 15: Set Split Indication – Γραμμή 64

Σε ένα νέο φύλλο του Isalos, γίνεται φόρτωση του Training Set (Εικόνα 16).



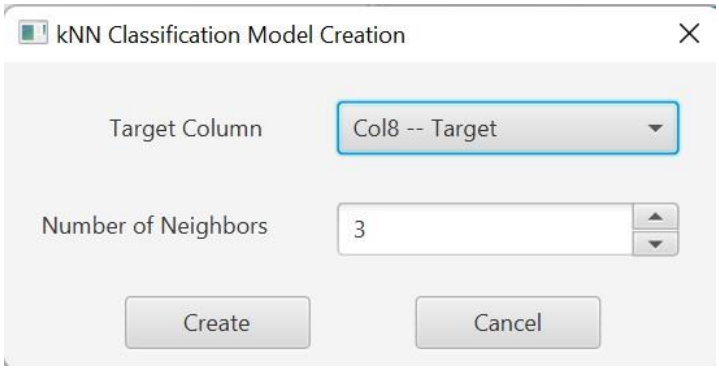
Εικόνα 16: Φόρτωση του Training Set

Σε αυτό το φύλλο, θα γίνει η δημιουργία του μοντέλου πρόβλεψης ύπαρξης νεφρικών λίθων.

Η πλατφόρμα Isalos προσφέρει μια πληθώρα επιλογών αλγορίθμων μηχανικής μάθησης. Το σετ αυτό αποτελεί ένα κλασικό πρόβλημα Classification, όπως έχει αναφερθεί και στην εισαγωγή. Από τους αλγορίθμους που προσφέρει το Isalos για μοντελοποίηση Classification προβλημάτων, επιλέχθηκε η χρήση του αλγορίθμου kNN, για τον οποίο έχει γίνει εκτενής

αναφορά στην εισαγωγή. Ο λόγος της επιλογής του συγκεκριμένου αλγορίθμου είναι η χρησιμότητα του στο συγκεκριμένο πρόβλημα και η απλότητα στη χρήση του.

Ο χρήστης επιλέγει τη διαδρομή Analytics > Classification > kNN και εμφανίζεται το ακόλουθο παράθυρο διαλόγου (Εικόνα 17):



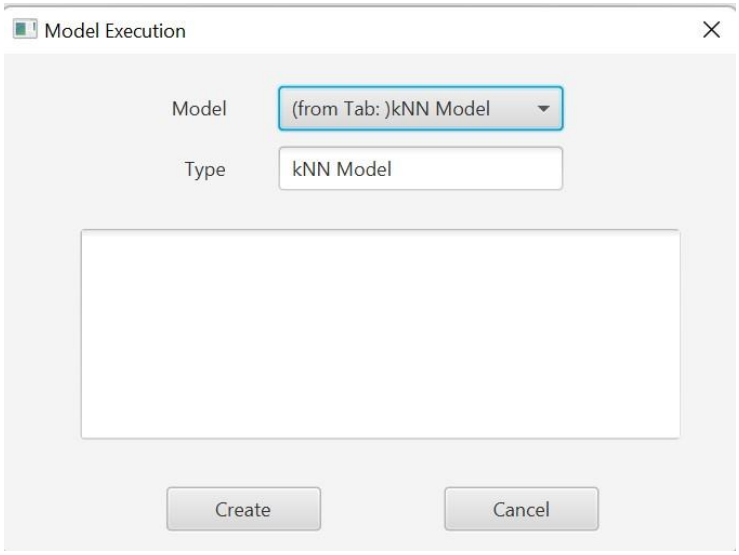
Εικόνα 17: Χρήση του αλγορίθμου kNN

Φυσικά, έγιναν διάφορες δοκιμές με διαφορετικούς αριθμούς “γειτόνων” για τα διάφορα training set που προέκυψαν κατά τη διαδικασία του splitting.

Πλέον, το προβλεπτικό μοντέλο είναι έτοιμο προς χρήση.

Επόμενο βήμα είναι η χρήση του μοντέλου που δημιουργήθηκε, στα δεδομένα του Test set. Για να συμβεί αυτό, ο χρήστης ανοίγει εκ νέου ένα ακόμα φύλλο στην πλατφόρμα Isalos, στο οποίο και εισάγει τα κανονικοποιημένα δεδομένα του Test Set, με τον ίδιο ακριβώς τρόπο που εισήγαγε τα δεδομένα του Training Set προηγουμένως.

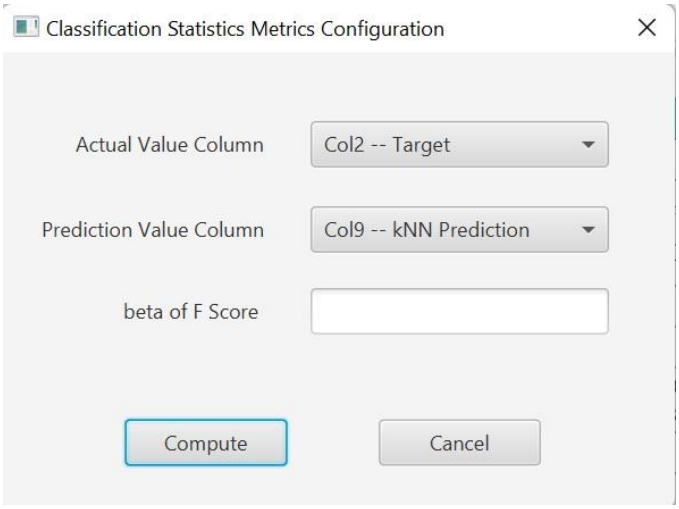
Για να χρησιμοποιήσει το μοντέλο που δημιούργησε, ο χρήστης ακολουθεί τη διαδρομή Analytics > Existing Model Utilisation και του εμφανίζεται το ακόλουθο παράθυρο διαλόγου (Εικόνα 18):



Εικόνα 18: Existing Model Utilisation

Προκειμένου να γίνει έλεγχος της αποτελεσματικότητας του μοντέλου, τα δεδομένα που προκύπτουν από τη χρήση του υποβάλλονται σε στατιστικά τεστ, και πάλι μέσα από την πλατφόρμα Isalos.

Για τη διεκπεραίωση αυτών των Test, γίνεται εκ νέου φόρτωση, σε νέο φύλλο εργασίας, των δεδομένων που προέκυψαν από τη χρήση του προβλεπτικού μοντέλου. Στη συνέχεια, ο χρήστης επιλέγει από το μενού του Isalos τη διαδρομή Statistics > Model Metrics > Classification Metrics και εμφανίζεται το ακόλουθο παράθυρο διαλόγου (Εικόνα 19):

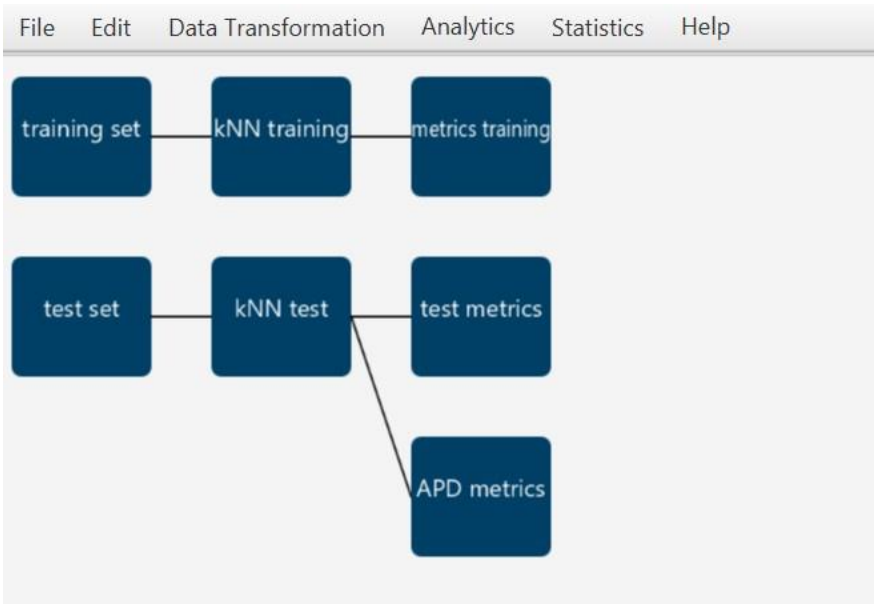


Εικόνα 19: Classification Metrics

Στο παράθυρο αυτό, ο χρήστης καλείται να επιλέξει ποια στήλη του δείγματος απεικονίζει την «πραγματικότητα» και ποια στήλη απεικονίζει την πρόβλεψη του μοντέλου βάσει του αλγορίθμου kNN.

Με τον ίδιο ακριβώς τρόπο διεξάγεται το στατιστικό τεστ και στα δεδομένα που προέκυψαν κατά τη δημιουργία του μοντέλου πρόβλεψης.

Κάθε βήμα της παραπάνω διαδικασίας, απεικονίζεται στο workflow του Isalos, στον επάνω μέρος της εφαρμογής (Εικόνα 20).



Εικόνα 20: Workflow Display

Τα στατιστικά αποτελέσματα που προκύπτουν θα παρουσιαστούν και θα αναλυθούν στην ενότητα «αποτελέσματα».

Όπως προαναφέρθηκε, η παραπάνω μεθοδολογία ακολουθήθηκε δεκάδες φορές, αλλάζοντας διάφορες παραμέτρους, προκειμένου να διασφαλισθεί το καλύτερο δυνατό αποτέλεσμα με τα συγκεκριμένα δεδομένα. Από τα διάφορα μοντέλα που δημιουργήθηκαν, προέκυψαν ορισμένα συμπεράσματα, τα οποία βοήθησαν πολύ στη βελτίωση του τελικού προβλεπτικού μοντέλου.

Κύριο συμπέρασμα σχετικά με το αρχικό σετ δεδομένων, είναι ότι σε αυτό περιλαμβάνονται άτομα, των οποίων οι μετρήσεις δημιουργούν “θόρυβο” στο μοντέλο. Τα άτομα αυτά αλλοίωναν την προβλεπτική ικανότητα του μοντέλου, και για αυτό το λόγο αποφασίσθηκε η αφαίρεση τους από το σετ. Τα άτομα που αφαιρέθηκαν ήταν συνολικά 11 και είναι τα ακόλουθα: 8, 21, 49, 50, 52, 56, 61, 62, 65, 67, 75. Η εξ ολοκλήρου αφαίρεση των ατόμων αυτών από το σετ βελτίωσε αισθητά την προβλεπτική ικανότητα του μοντέλου, που είναι το πραγματικό ζητούμενο της παρούσας εργασίας.

Εκτός όμως από την αφαίρεση ατόμων από το σετ, έγινε και αφαίρεση δύο μεταβλητών. Η απόφαση για την αφαίρεση δύο στηλών από τα αρχικά δεδομένα πάρθηκε για δύο πολύ σημαντικούς λόγους. Ο πρώτος έχει να κάνει με τη βελτιστοποίηση του μοντέλου, καθώς παρατηρήθηκε ότι η αφαίρεση αυτών των μεταβλητών βελτίωνε την προβλεπτική του ικανότητα. Ο δεύτερος, και σημαντικότερος, λόγος για την αφαίρεση των συγκεκριμένων μεταβλητών από το αρχικό σετ είναι η προσαρμογή του μοντέλου στα δεδομένα της ελληνικής, και πιθανά όχι μόνο, πραγματικότητας. Ερχόμενος σε επικοινωνία με διάφορα μικροβιολογικά εργαστήρια, τα όποια θα είναι και οι κύριοι ενδιαφερόμενοι για το συγκεκριμένο μοντέλο, συμπεράθηκε πως οι δύο παράμετροι που τελικώς αφαιρέθηκαν, δε μετρούνται κατά τη διαδικασία μιας τυπικής, γενικής εξέταση ούρων. Συνεπώς, η ύπαρξη των παραμέτρων αυτών θα καταστούσε το μοντέλο πρακτικά “άχρηστο” για τα μικροβιολογικά εργαστήρια, εκτός αν ξεκινούσαν να μετράνε αυτές τις παραμέτρους στις γενικές εξετάσεις ούρων, πράγμα αδύνατον λόγω των μεθόδων που χρησιμοποιούνται για τη μέτρηση των συγκεκριμένων παραμέτρων. Όπως γίνεται εύκολα αντιληπτό, αυτό είναι πρακτικά αδύνατο να συμβεί, οπότε η δημιουργία του μοντέλου προσαρμόστηκε στις ανάγκες των εργαστηρίων, ώστε αυτό να έχει πιθανότητες να βρει πρακτική εφαρμογή στο μέλλον.

Αφού εξακριβώθηκε η βέλτιστη μορφή του αρχικού σετ για την ανάπτυξη του μοντέλου, έπρεπε να δοκιμαστεί σε πραγματικές συνθήκες. Έτσι, σε συνεργασία με ένα μεγάλο διαγνωστικό κέντρο της ελληνικής αγοράς, συλλέχθηκαν δεδομένα από εξετάσεις ασθενών, προκειμένου να χρησιμοποιηθούν τόσο για τη δοκιμή του υπάρχοντος μοντέλου, όσο και για την περαιτέρω βελτίωση του. Τα δεδομένα αυτά υποβλήθηκαν στην ίδια ακριβώς διαδικασία που αναλύθηκε διεξοδικώς παραπάνω, και έδωσαν κάποια στατιστικά αποτελέσματα, τα οποία θα αναλυθούν στη συνέχεια της εργασίας. Οφείλει να αναφερθεί πως διασφαλίσθηκε η πλήρη ανωνυμία των ασθενών, από τις εξετάσεις των οποίων προήλθαν τα δεδομένα που χρησιμοποιήθηκαν.

Το τελικό σετ, το οποίο αποτελεί συνδυασμό του αρχικού σετ με τα δεδομένα που συλλέχθηκαν από το διαγνωστικό κέντρο, θα παρατεθεί στο παράρτημα της παρούσας εργασίας.

Κεφάλαιο 3: Αποτελέσματα

Μέσα από τα βήματα της μεθοδολογίας που περιεγράφηκε διεξοδικώς παραπάνω, είναι πλέον διαθέσιμα δεκάδες μοντέλα μηχανικής μάθησης, που έχουν εκπαιδευτεί προκειμένου να προβλέπουν την ύπαρξη ή μη νεφρικών λίθων από ορισμένους διαγνωστικούς δείκτες που μετρούνται κατά τη διενέργεια μίας τυπικής εξέτασης ούρων.

Ορισμένα από αυτά τα μοντέλα είχαν απογοητευτικά αποτελέσματα. Όπως αναφέρθηκε στη μεθοδολογία, υπήρχαν κάποια άτομα εντός του αρχικού σετ δεδομένων, τα οποία επιδρούσαν συνεχώς αρνητικά, στην προβλεπτική ικανότητα των μοντέλων που κατασκευάζονταν.

3.1 Δημιουργία μοντέλων με τα δεδομένα του αρχικού set

Παρακάτω, παρουσιάζεται η στατιστική ανάλυση 3 μοντέλων που δημιουργήθηκαν με τον αλγόριθμο kNN για 3,4 και 5 αριθμούς γειτόνων. Τα μοντέλα αυτά δημιουργήθηκαν από αυτούσιο το αρχικό σετ δεδομένων. Αυτό σημαίνει πως σε αυτές τις πρώτες δοκιμές, συμπεριλαμβάνονται τόσο τα άτομα που στη συνέχεια αφαιρέθηκαν, όσο και οι μεταβλητές «ωσμομοριακότητα» και «αγωγιμότητα», που επίσης αφαιρέθηκαν. Να σημειωθεί επίσης πως για το διαχωρισμό του Set στα επιμέρους training και test sets στα αποτελέσματα που παρουσιάζονται, χρησιμοποιήθηκε ο αλγόριθμος Kennard-Stone με 70% Split Percentage. Φυσικά, έγιναν δοκιμές για διάφορα split percentage, ανάμεσα στο 70 και το 80%, με το 70% να δίνει τα καλύτερα αποτελέσματα, για αυτό και παρουσιάζεται. Πολύ σημαντικό κατά τη διαδικασία του splitting είναι οι δύο κλάσεις να εκπροσωπούνται επαρκώς και στα δύο νέα σετ, ώστε τα στατιστικά που προκύπτουν να δίνουν ασφαλή συμπεράσματα.

Το πρώτο μοντέλο δημιουργήθηκε για k=3 (Μοντέλο 1) και έδωσε τα εξής αποτελέσματα αναφορικά με τα στατιστικά του:

Πίνακας 1: kNN model για k=3 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	30 (TN)	24 (FN)
Predicted True	0 (FP)	1 (TP)

Τα αντίστοιχα στατιστικά για το test set ήταν:

Πίνακας 2: kNN model για k=3 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	3 (FN)
Predicted True	2 (FP)	6 (TP)

Από τα αποτελέσματα αυτά παρατηρείται ένα εμφανές παράδοξο. Ενώ το μοντέλο που εκπαιδεύτηκε είναι ξεκάθαρο πως έχει σχεδόν μηδενική ευαισθησία, παρόλα αυτά είναι σε θέση να προβλέψει με επιτυχία 66,7% τα αληθώς θετικά δείγματα και με 87% επιτυχία τα αληθώς αρνητικά δείγματα του test set, με τη συνολική ακρίβεια να ανέρχεται σε 79% (Εξίσωση 5).

Μία πιθανή εξήγηση για αυτό το γεγονός θα μπορούσε να είναι η δυσανάλογη κατανομή των θετικών και αρνητικών δειγμάτων για τη δημιουργία του μοντέλου. Όμως, όπως φαίνεται και από τους παραπάνω πίνακες, κάτι τέτοιο δεν υφίσταται.

Για τα ίδια training και test set, δηλαδή μετά από χρήση του αλγορίθμου Kennard-Stone και για 70% split percentage, για k=4 (Μοντέλο 2), τα στατιστικά που προέκυψαν είναι τα εξής:

Πίνακας 3: kNN model για k=4 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	29 (TN)	23 (FN)
Predicted True	1 (FP)	2 (TP)

Τα αντίστοιχα στατιστικά για το test set ήταν:

Πίνακας 4: kNN model για k=3 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	2 (FN)
Predicted True	2 (FP)	7 (TP)

Παρατηρείται αμέσως ότι η απόκλιση σε σχέση με το πρώτο μοντέλο είναι πολύ μικρή, με μια μικρή βελτίωση στην προβλεπτική του ικανότητα στο τεστ σετ, όπου προβλέπει ως ορθώς θετικά τα 7 από τα 9 όντως θετικά άτομα.

Τέλος, τα στατιστικά για k=5 (Μοντέλο 3) είναι τα εξής:

Πίνακας 5: kNN model για k=5 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	30 (TN)	22 (FN)
Predicted True	0 (FP)	3 (TP)

Και για το test set:

Πίνακας 6: kNN model για k=5 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	5 (FN)
Predicted True	2 (FP)	4 (TP)

Σε αυτή τη μοντελοποίηση, η προβλεπτική ικανότητα για τα θετικά άτομα πέφτει ακόμα περισσότερο από τις προηγούμενες δύο δοκιμές.

Από τα παραπάνω δε μπορεί να εξαχθεί κάποιο συμπέρασμα, παρά μόνο ότι ίσως ο αλγόριθμος Kennard-Stone να μην είναι ο ιδανικός για το συγκεκριμένο σετ δεδομένων, καθώς στα training και test set που δημιουργούνται, δεν εκπροσωπούνται επαρκώς οι δύο κλάσεις.

Για τον έλεγχο αυτής της διαπίστωσης, δημιουργήθηκαν εκ νέου μοντέλα για k=3, k=4 και k=5 με τη μέθοδο του Random Partitioning να χρησιμοποιείται για το διαχωρισμό του αρχικού σετ σε training και test. Πολύ σημαντικό κατά τη χρήση του random partitioning ως μέθοδο διαχωρισμού είναι να διασφαλιστεί ότι οι δύο κλάσεις εκπροσωπούνται επαρκώς τόσο στο training, όσο και στο test set. Τα αποτελέσματα των μοντελοποιήσεων με την παρούσα

μέθοδο διαχωρισμού για k=3 και k=5 ήταν απογοητευτικά. Στους πίνακες 7 και 8 παρουσιάζονται τα αποτελέσματα της μοντελοποίησης μετά από τυχαίο διαχωρισμό του σετ για k=4.

Τα στατιστικά που προέκυψαν από αυτή τη μοντελοποίηση (Μοντέλο 4) αναγράφονται παρακάτω:

Πίνακας 7: kNN model για k=4 μετά από Split Random Partitioning για 70% Split Percentage: Training Set

	False	True
Predicted False	27 (TN)	18 (FN)
Predicted True	1 (FP)	9 (TP)

Και για το test set:

Πίνακας 8: kNN model για k=4 μετά από Split με Random Partitioning για 70% Split Percentage: Test Set

	False	True
Predicted False	14 (TN)	4 (FN)
Predicted True	3 (FP)	3 (TP)

Όπως αναφέρθηκε και σε προηγούμενες ενότητες, η ωσμομοριακότητα και η αγωγιμότητα δεν αποτελούν παραμέτρους που μετρούνται από τα ελληνικά διαγνωστικά εργαστήρια κατά τη διάρκεια μιας τυπικής εξέτασης ούρων. Συνεπώς, οι δύο αυτές παράμετροι αφαιρούνται από το αρχικό σετ δεδομένων, προκειμένου αυτό να μπορεί μελλοντικώς να χρησιμοποιηθεί από αυτά.

Δημιουργήθηκαν εκ νέου διάφορα μοντέλα, με το καλύτερο εξ αυτών να δημιουργείται μετά από χρήση του αλγορίθμου Kennard-Stone με 70% Split Percentage και για k=3. Τα στατιστικά αποτελέσματα για την προβλεπτική ικανότητα του μοντέλου αυτού στο test set ακολουθούν στον παρακάτω πίνακα (Μοντέλο 5):

Πίνακας 9: kNN model για k=3 μετά από Split με Random Partitioning για 70% Split Percentage: Test Set

	False	True
Predicted False	14 (TN)	3 (FN)
Predicted True	3 (FP)	4 (TP)

Από τα παραπάνω εξάγεται το συμπέρασμα ότι το αρχικό σετ δεδομένων χρήζει βελτίωσης, προκειμένου να είναι σε θέση να καταστεί κάποιο από τα παραγόμενα μοντέλα χρηστικό.

3.2 Δημιουργία μοντέλων μετά από τροποποίηση του αρχικού σετ

Κατά τις δοκιμές που προηγήθηκαν στην ενότητα 3.1, παρατηρήθηκε πως υπάρχουν κάποια άτομα τα οποία προβλέπονται σχεδόν σε όλες τις περιπτώσεις λάθος κατά τη μοντελοποίηση. Τα άτομα αυτά, όπως προαναφέρθηκε και στην ενότητα «Μεθοδολογία» είναι τα εξής: 8, 21, 49, 50, 52, 56, 61, 62, 65, 67, 75. Τα άτομα αυτά αφαιρούνται από το αρχικό set, με στόχο τη βελτίωση των μοντέλων που δημιουργούνται.

Ακολουθούν τα στατιστικά αποτελέσματα 3 μοντελοποιήσεων, για 3,4 και 5 γείτονες αντίστοιχα, όπου χρησιμοποιήθηκε ο αλγόριθμος Kennard-Stone (70% split percentage) για τη δημιουργία των training και test sets.

Για k=3 (Μοντέλο 6):

Πίνακας 10: kNN model για k=3 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	24 (TN)	6 (FN)
Predicted True	4 (FP)	14 (TP)

Και για το test set:

Πίνακας 11: kNN model για k=3 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	2 (FN)
Predicted True	1 (FP)	4 (TP)

Για k=4 (Μοντέλο 7):

Πίνακας 12: kNN model για k=4 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	24 (TN)	7 (FN)
Predicted True	4 (FP)	13 (TP)

Και για το test set:

Πίνακας 13: kNN model για k=4 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	2 (FN)
Predicted True	1 (FP)	4 (TP)

Για k=5 (Μοντέλο 8):

Πίνακας 14: kNN model για k=5 μετά από Split με Kennard-Stone για 70% Split Percentage: Training Set

	False	True
Predicted False	25 (TN)	10 (FN)
Predicted True	3 (FP)	10 (TP)

Και για το test set:

Πίνακας 15: kNN model για k=5 μετά από Split με Kennard-Stone για 70% Split Percentage: Test Set

	False	True
Predicted False	13 (TN)	1 (FN)
Predicted True	1 (FP)	5 (TP)

Από τα παραπάνω αποτελέσματα παρατηρείται μεγάλη βελτίωση της προβλεπτικής ικανότητας του μοντέλου, ιδιαίτερα όσον αφορά την ευαισθησία του. Πιο συγκεκριμένα, τα αποτελέσματα για k=3 και k=4 είναι σχεδόν πανομοιότυπα. Για k=5, φαίνεται από τους παραπάνω πίνακες πως ενώ η ευαισθησία του μοντέλου στο training set πέφτει (Εξίσωση 2), στα άτομα του test set αυξάνεται. Φυσικά, ο αριθμός των δειγμάτων είναι τόσο μικρός, που χρειάζεται

περαιτέρω διερεύνηση, ώστε να διαπιστωθεί αν κάποιο από τα παραπάνω μοντέλα σε αυτήν τους τη μορφή παρουσιάζει κάποια χρηστικότητα.

3.3 Χρήση δεδομένων από ελληνικό εργαστήριο για βελτίωση της προβλεπτικής ικανότητας των μοντέλων

Σε αυτό το σημείο έρχονται στο προσκήνιο τα δεδομένα που συλλέχθηκαν από το διαγνωστικό εργαστήριο. Είναι πολύ σημαντικό να αναφερθεί ότι διασφαλίστηκε η πλήρης ανωνυμία των δεδομένων, καθώς αυτά συλλέχθηκαν σύμφωνα με το πρωτόκολλο ασφαλείας που ακολουθεί το διαγνωστικό εργαστήριο. Στον παρακάτω πίνακα παρατίθενται τα δεδομένα που συλλέχθηκαν:

Πίνακας 16: Λίστα των δεδομένων από το διαγνωστικό εργαστήριο

User Row ID	Gravity	pH	Urea	Calcium	Target
80	1.026	6.23	250	7.33	TRUE
81	1.016	5.7	210	9.86	TRUE
82	1.021	6.02	351	6.82	TRUE
83	1.017	6.33	370	5.93	TRUE
84	1.015	6.91	563	10.0	TRUE
85	1.013	6.11	100	6.02	TRUE
86	1.020	5.88	136	6.91	TRUE
87	1.014	6.24	267	7.3	TRUE
88	1.020	5.91	254	6.42	TRUE
89	1.010	7.37	333	7.3	TRUE
90	1.015	7.13	170	3.03	FALSE
91	1.010	6.11	192	2.92	FALSE
92	1.009	7.22	105	2.14	FALSE
93	1.012	6.32	316	4.05	FALSE
94	1.020	6.25	241	3.77	FALSE
95	1.017	6.92	175	2.62	FALSE
96	1.013	7.15	155	2.49	FALSE
97	1.015	6.87	230	2.17	FALSE
98	1.010	7.63	25	2.02	FALSE
99	1.025	6.06	106	2.5	FALSE
100	1.025	6.43	146	2.61	FALSE
101	1.011	6.97	210	2.52	FALSE

Στα δεδομένα δόθηκαν ID από το 80 και έπειτα, σε συνέχεια της αρίθμησης του αρχικού σετ.

Το νέο αυτό set δεδομένων μπορεί να χρησιμοποιηθεί με δύο τρόπους. Ο πρώτος τρόπος είναι να χρησιμοποιηθεί αυτούσιο το σετ αυτό ως test set, προκειμένου να εξακριβωθεί εάν κάποιο από τα μοντέλα που δημιουργήθηκαν προηγουμένως μπορεί να προβλέψει σωστά την ύπαρξη ή μη λίθων στα παραπάνω άτομα. Ο δεύτερος τρόπος χρήσης αυτών των δεδομένων είναι η συγχώνευση τους με το αρχικό set και η εκ νέου δημιουργία δύο training και test set, με στόχο τη δημιουργία ενός νέου, βελτιωμένου μοντέλου.

Πρώτα, θα δοκιμαστούν τα 2 μοντέλα που δημιουργήθηκαν και παρουσιάστηκαν στην ενότητα 3.2, για k=4 και k=5.

Έτσι, μετά την εφαρμογή του μοντέλου για k=4 στα παραπάνω δεδομένα, προέκυψαν τα ακόλουθα στατιστικά στοιχεία:

Πίνακας 17: Εφαρμογή του μοντέλου 7 στα δεδομένα του ελληνικού εργαστηρίου

	False	True
Predicted False	12 (TN)	1 (FN)
Predicted True	0 (FP)	9 (TP)

Τα παραπάνω αποτελέσματα είναι κάτι παραπάνω από ενθαρρυντικά για τη χρηστικότητα του συγκεκριμένου μοντέλου, καθώς η ευαισθησία είναι 0,9 [(Εξίσωση 2) (ή 90% σωστές προβλέψεις παρουσίας λίθων)] και η εξειδίκευση 1 [(Εξίδωση 3) (100% σωστές προβλέψεις απουσίας λίθων)]. Συνεπώς, το ανωτέρω μοντέλο προβλέπει με περίπου 95% επιτυχία την παρουσία ή απουσία λίθων στα παραπάνω δεδομένα (Εξίσωση 5).

Η ίδια διαδικασία επαναλήφθηκε και για το μοντέλο με k=5, και τα αποτελέσματα που προέκυψαν είναι τα παρακάτω:

Πίνακας 18: Εφαρμογή του μοντέλου 8 στα δεδομένα του ελληνικού εργαστηρίου

	False	True
Predicted False	12 (TN)	0 (FN)
Predicted True	0 (FP)	10 (TP)

Σε αυτόν τον πίνακα φαίνεται πως το μοντέλο που δημιουργήθηκε με το 70% του αρχικού σετ, μετά από split με τον αλγόριθμο Kennard-Stone και με χρήση του αλγορίθμου kNN για k=5, προέβλεψε με επιτυχία 100% την παρουσία ή απουσία νεφρικών λίθων στους 22 ασθενείς του σετ δεδομένων που ανακτήθηκε από το διαγνωστικό εργαστήριο. Τα παραπάνω αποτελέσματα, αν και κάτι παραπάνω από ενθαρρυντικά, χρήζουν ιδιαίτερης προσοχής. Περαιτέρω μελέτες σε ακόμα περισσότερα δεδομένα και ακόμα περισσότερους ασθενείς, και σε δεδομένα από διαφορετικά διαγνωστικά εργαστήρια, είναι απαραίτητο να γίνουν, προκειμένου να εξασφαλιστεί η αξιοπιστία του συγκεκριμένου μοντέλου για χρήση.

3.4 Χρήση των δεδομένων του ελληνικού εργαστηρίου για τη δημιουργία νέων μοντέλων

Ακολούθως, δημιουργήθηκαν τρία νέα μοντέλα, τα οποία βασίστηκαν σε ένα νέο, ενιαίο σετ, που προέκυψε από τη συγχώνευση του αρχικού με τα δεδομένα από το διαγνωστικό εργαστήριο. Όπως ακριβώς και στις μοντελοποιήσεις που προηγήθηκαν, έτσι και σε αυτή, έγιναν δοκιμές για διαφορετικά split percentage, είτε με τον αλγόριθμο Kennard-Stone, είτε με τη μέθοδο της τυχαίας διάσπασης του σετ (Random Partitioning).

Μετά από χρήση του αλγορίθμου Kennard-Stone για το splitting του σετ, για ένα εύρος μεταξύ 70 και 80% split percentage, παρατηρήθηκε ότι ο συγκεκριμένος αλγόριθμος δεν εφαρμό-

ζεται ορθώς στα συγκεκριμένα δεδομένα, καθώς οι δύο κλάσεις TRUE και FALSE, που αντιπροσωπεύουν την παρουσία ή την απουσία νεφρικών λίθων αντίστοιχα, δεν εκπροσωπούνται παρομοίως στα training και test σετ που προκύπτουν. Έτσι, η μοντελοποίηση μετά τη χρήση του συγκεκριμένου αλγορίθμου απορρίφθηκε.

Ακολούθησαν δοκιμές για την εύρεση του καλύτερου Split Percentage με τη μέθοδο Random Partitioning. Ιδιαίτερα σημαντικό κατά τη χρήση της συγκεκριμένης μεθόδου διαχωρισμού είναι η χρήση των ίδιων set που θα προκύψουν για τη μοντελοποίηση για διαφορετικό αριθμό γειτόνων με τον αλγόριθμο kNN, καθώς το σετ διαχωρίζεται με διαφορετικό κάθε φορά τρόπο, εν αντιθέσει με τον αλγόριθμο Kennard-Stone, που διαχωρίζει το σετ με βάση τις ευκλείδειες αποστάσεις των δειγμάτων, όπως αυτό περιεγράφηκε στην «Εισαγωγή» της παρούσας εργασίας.

Μετά τον τυχαίο διαχωρισμό των αρχικών δεδομένων, προέκυψε ένα training set που περιείχε 66/90 δείγματα και ένα test set που περιείχε 24/90 δείγματα, με 70% Split Percentage. Τα δύο αυτά set θα παρατεθούν στο παράρτημα της παρούσας διπλωματικής. Από τις δοκιμές που ακολούθησαν προέκυψαν μοντέλα με χαμηλή εξειδίκευση και ευαισθησία, τα περισσότερα με μικρότερη του 50%. Το μοναδικό μοντέλο το οποίο φαίνεται πως έχει επαρκή προβλεπτική ικανότητα ήταν το μοντέλο που προέκυψε για Split Percentage 70% για k=5 (Μοντέλο 9). Στον παρακάτω πίνακα παρατίθενται τα στατιστικά του αποτελέσματα:

Για το training set:

Πίνακας 19: kNN model για k=5 μετά από Split με Random Partitioning για 70% Split Percentage: Training Set

	False	True
Predicted False	37 (TN)	8 (FN)
Predicted True	8 (FP)	17 (TP)

Και για το test set:

Πίνακας 20: kNN model για k=5 μετά από Split με Random Partitioning για 70% Split Percentage: Test Set

	False	True
Predicted False	14 (TN)	3 (FN)
Predicted True	2 (FP)	8 (TP)

Από τα παραπάνω στατιστικά που προέκυψαν από την εφαρμογή του συγκεκριμένου μοντέλου στο test set, προκύπτει αποτελεσματικότητα πρόβλεψης κοντά στο 81% (Εξίσωση 5) για την ύπαρξη ή μη νεφρικών λίθων, με την ευαισθησία να είναι 72,7% (Εξίσωση 2) και την εξειδίκευση 87,5% (Εξίσωση 3).

3.5 Συγκεντρωτικά αποτελέσματα των ανωτέρω μοντελοποιήσεων

Στον παρακάτω πίνακα (Πίνακας 20), θα παρουσιαστούν συγκεντρωτικά τα αποτελέσματα όλων των μοντελοποιήσεων που προηγήθηκαν. Συγκεκριμένα, θα παρατεθούν η ευαισθησία, η εξειδίκευση, η ακρίβεια πρόβλεψης, καθώς και το F1 score του κάθε μοντέλου, αναφορικά με την πρόβλεψη αυτού για τα άτομα των διαφόρων Test Sets.

Ο υπολογισμός αυτών των παραμέτρων πραγματοποιείται αυτόματα από το Isalos, κατά τη χρήση της διαδρομής Statistics > Model Metrics > Classification Metrics, με βάση τις εξισώσεις 2, 3 , 5 και 4 αντίστοιχα.

Πίνακας 21: Συγκεντρωτικά αποτελέσματα μοντελοποιήσεων-ικανότητα πρόβλεψης των ατόμων των Test Sets

Μοντέλα	Ευαισθησία (Εξίσωση 2)	Εξειδίκευση (Εξίσωση 3)	F1 Score (Εξίσωση 4)	Ακρίβεια (Εξίσωση 5)
Αρχικό Set k=3 Kennard-Stone Split 70% (Μοντέλο 1)	0,75	0,8125	0,78	79%
Αρχικό Set k=4 Kennard-Stone Split 70% (Μοντέλο 2)	0,78	0,87	0,82	83%
Αρχικό Set k=5 Kennard-Stone Split 70% (Μοντέλο 3)	0,67	0,72	0,69	71%
Αρχικό Set k=4 Random Partitioning 70% (Μοντέλο 4)	0,5	0,78	0,61	71%
Τροποποιημένο σετ μετά την αφαίρεση των 2 παραμέτρων k=3, Kennatd-Stone Split 70% (Μοντέλο 5)	0,57	0,82	0,67	75%
Τροποποιημένο Set μετά την αφαίρεση 11 ατόμων και 2 παραμέτρων k=3 Kennard-Stone Split 70% (Μοντέλο 6)	0,8	0,87	0,83	85%
Τροποποιημένο Set μετά την αφαίρεση 11 ατόμων και 2 παραμέτρων k=4 Kennard-Stone Split 70% (Μοντέλο 7)	0,8	0,87	0,83	85%
Τροποποιημένο Set μετά την αφαίρεση 11 ατόμων και 2 παραμέτρων k=5 Kennard-Stone Split 70% (Μοντέλο 8)	0,83	0,93	0,88	90%
Εφαρμογή του μοντέλου 7 στα δεδομένα του ελληνικού εργαστηρίου	1	0,92	0,96	95%
Εφαρμογή του μοντέλου 8 στα δεδομένα του ελληνικού εργαστηρίου	1	1	1	100%

Μοντέλο 9	0,8	0,82	0,81	81%

Από τον πίνακα αυτό μπορούν πλέον να εξαχθούν ασφαλή συμπεράσματα για τη χρηστικότητα των μοντέλων που δημιουργήθηκαν στα πλαίσια της παρούσας διπλωματικής. Καταρχήν, φαίνεται πλέον ξεκάθαρα πόσο βοήθησε στην προβλεπτική ικανότητα η απομάκρυνση από το αρχικό set, τόσο των 11 ατόμων, όσο και των παραμέτρων της ωσμομοριακότητας και της αγωγιμότητας των ούρων των υπό εξέταση ασθενών. Με την απομάκρυνση από τα δεδομένα, η αποτελεσματικότητα πρόβλεψης των μοντέλων που προέκυψαν στη συνέχεια, δεν έπεσε ποτέ κάτω από 80%.

Φαίνεται λοιπόν επίσης, πως το μοντέλο 8, το οποίο προέβλεψε με 100% επιτυχία την ύπαρξη ή μη νεφρικών λίθων στους ασθενείς του ελληνικού διαγνωστικού κέντρου, είναι το πιο ακριβές από όλα τα δεδομένα που δημιουργήθηκαν. Συνεπώς, είναι και αυτό που έχει τις περισσότερες πιθανότητες να βρει πραγματική εφαρμογή μελλοντικά.

Κεφάλαιο 4: Συμπεράσματα-Συζήτηση

Στόχος της συγκεκριμένης μεταπτυχιακής διατριβής ήταν η δημιουργία ενός μοντέλου μηχανικής μάθησης με τη χρήση της πλατφόρμας Isalos, το οποίο να είναι σε θέση να προβλέπει, με όσο δυνατόν μεγαλύτερη ακρίβεια, την παρουσία ή μη νεφρικών λίθων σε ασθενείς, χρησιμοποιώντας τιμές που μετρούνται σε μία απλή γενική εξέταση ούρων καθημερινά.

Για το σκοπό αυτό, χρησιμοποιήθηκαν διάφορες προσεγγίσεις. Με τη χρήση ενός αρχικού σετ δεδομένων από το διαδίκτυο, δημιουργήθηκαν κάποια πρώτα μοντέλα, των οποίων η προβλεπτική ικανότητα, η ευαισθησία και η εξειδίκευση δεν ήταν επαρκής, ώστε αυτά να χαρακτηριστούν ως χρηστικά.

Μετά την αφαίρεση 2 εκ των 6 παραμέτρων του αρχικού σετ και την απομάκρυνση 11 εκ των 79 δειγμάτων που δημιουργούσαν “θόρυβο” κατά τη μοντελοποίηση, παρατηρήθηκε σημαντική βελτίωση της προβλεπτικής ικανότητας των εξαγόμενων μοντελοποιήσεων.

Τα μοντέλα που προέκυψαν εξετάστηκαν σε πραγματικά δεδομένα που συλλέχθηκαν από διαγνωστικό εργαστήριο, με ένα εξ αυτών (Μοντέλο 8) να προβλέπει με 100% αποτελεσματικότητα την παρουσία ή απουσία νεφρικών λίθων στα 22 επιβεβαιωμένα δείγματα που εφαρμόστηκε.

Έγιναν επίσης δοκιμές χρήσης των νέων αυτών δεδομένων, σε συνδυασμό με τα αρχικά, για νέα μοντελοποίηση, χωρίς αυτές να έχουν τα επιθυμητά αποτελέσματα, τουλάχιστον όχι στο βαθμό των προηγούμενων μοντέλων με την απόλυτη αποτελεσματικότητα πρόβλεψης.

Είναι πολύ σημαντικό να μην αγνοηθεί η πολυμορφικότητα που χαρακτηρίζει τα δεδομένα που χρησιμοποιήθηκαν για τη δημιουργία όλων των μοντέλων στην παρούσα διπλωματική. Από τη μία, τα δεδομένα του αρχικού set δεδομένων προέρχονται από Αμερικανική ουρολογική κλινική, ενώ τα δεδομένα που χρησιμοποιήθηκαν προκειμένου να ελεγχθεί η προβλεπτική ικανότητα των μοντέλων, από ελληνικό διαγνωστικό κέντρο. Όπως γίνεται εύκολα αντιληπτό, τα υλικά και οι μέθοδοι που χρησιμοποιούν τα δύο αυτά απομακρυσμένα εργαστήρια, είναι πολύ πιθανό να διαφέρουν. Συνεπώς, χρήζει περαιτέρω μελέτης αν η διαφορά αυτή στα υλικά και τις μεθόδους των δύο διαγνωστικών εργαστηρίων, διαδραματίζει κάποιο ρόλο στην ικανότητα του μοντέλου να προβλέπει ορθώς την ύπαρξη ή μη νεφρικών λίθων στους διάφορους ασθενείς.

Επιπλέον, οφείλει να διερευνηθεί η σημασία της δημογραφικής πολυπλοκότητας των ατόμων των δύο σετ, μιας και αποτελούν μέρη δύο εντελώς ξεχωριστών και απομακρυσμένων πληθυσμιακών ομάδων. Χρήζει περαιτέρω έρευνας για το αν υπάρχουν γενετικοί ή περιβαλλοντικοί παράγοντες, οι οποίοι είναι σε θέση να επηρεάζουν την προβλεπτική ικανότητα του μοντέλου 8.

Από τα παραπάνω αποκαλύπτεται το πόσο πραγματικά ρεαλιστικό είναι το σενάριο της πρόβλεψης ύπαρξης νεφρικών λίθων από ένα απλό τεστ ούρων. Φυσικά, θα χρειαστούν περαιτέρω μελέτες, σε πολύ μεγαλύτερο εύρος δεδομένων και δειγμάτων, προκειμένου να επιβεβαιωθεί, και πιθανά να βελτιωθεί ακόμα περισσότερο, η προβλεπτική ικανότητα ενός τέτοιου μοντέλου.

Αποκαλύπτεται όμως, και μία επιχειρηματική ευκαιρία, για την εταιρεία που έχει στην κατοχή της, ένα τέτοιο μοντέλο. Δε θα είναι λίγα τα διαγνωστικά εργαστήρια, τα οποία θα θελήσουν να χρησιμοποιήσουν μια τέτοια υπηρεσία, προκειμένου να προσφέρουν την καλύτερη δυνατή φροντίδα στους ασθενείς τους.

Παράλληλα, η χρήση ενός τέτοιου μοντέλου από ένα διαγνωστικό κέντρο, μπορεί να αποτελέσει στρατηγικό πλεονέκτημα έναντι των ανταγωνιστών του. Ένα εργαστήριο το οποίο θα είναι σε θέση να χρησιμοποιεί ένα μοντέλο με επιβεβαιωμένη την προβλεπτική του ικανότητα σε υψηλό ποσοστό, θα μπορεί να δίνει απαντήσεις στους ασθενείς του σε λιγότερο χρόνο από ένα άλλο και πιθανώς, με μικρότερο κόστος.

Η εταιρεία Novamechanics Ltd. φιλοξενεί μοντέλα μηχανικής μάθησης στην πλατφόρμα Enalos Cloud, των οποίων η χρήση γίνεται δωρεάν από κάθε ενδιαφερόμενο. Η εταιρεία θα μπορούσε να φιλοξενήσει στη συγκεκριμένη πλατφόρμα και το συγκεκριμένο μοντέλο δωρεάν, τουλάχιστον σε πρώτη φάση, και να ενθαρρύνει διάφορα διαγνωστικά κέντρα, της Ελλάδας, της Κύπρου, ακόμα και του εξωτερικού, να το χρησιμοποιήσουν.

Τα κίνητρα για τα εργαστήρια μπορούν να είναι οικονομικής φύσεως. Για παράδειγμα, θα μπορούσαν να υπογραφούν συμβάσεις με τα εργαστήρια, οι οποίες θα προβλέπουν μειωμένο κόστος χρήσης του προβλεπτικού μοντέλου, όταν αυτό ξεκινήσει να χρησιμοποιείται ευρέως στην αγορά. Επίσης, μια τέτοια σύμβαση θα μπορούσε να προβλέπει, ειδικά για περιοχές με μικρό αριθμό διαγνωστικών, αποκλειστική χρήση του μοντέλου, αν και εφόσον αυτό διατεθεί προς ευρεία χρήση στην αγορά. Τα παραπάνω κίνητρα πρέπει να είναι αρκετά, ώστε να πείσει τα διαγνωστικά εργαστήρια να συνεργαστούν.

Η συνεργασία αυτή με τα εργαστήρια είναι αναγκαία για πολλούς λόγους. Ο πρώτος έχει να κάνει με την επιβεβαίωση της χρηστικότητας του μοντέλου σε πραγματικά δεδομένα πολλών, διαφορετικών εργαστηρίων, τα οποία πιθανώς να χρησιμοποιούν διαφορετικές μεθόδους με τις οποίες διενεργούν τις μετρήσεις τους στα ούρα των ασθενών. Έτσι, είναι πιθανόν να εξαχθούν αποτελέσματα σχετικά με τις μεθόδους, οι οποίες λειτουργούν αρμονικότερα με το μοντέλο. Αυτό θα δώσει ένα στρατηγικό πλεονέκτημα στην εταιρεία, καθώς πλέον θα μπορεί να απευθύνεται πρώτα στα εργαστήρια που χρησιμοποιούν τις συγκεκριμένες μεθόδους, ώστε οι προβλέψεις του μοντέλου να έχουν την καλύτερη δυνατή ακρίβεια, δημιουργώντας θετικό κλίμα γύρω από τη χρήση της συγκεκριμένης τεχνολογίας.

Άλλος πολύ σημαντικός λόγος, είναι η συλλογή ακόμα περισσότερων δεδομένων, τα οποία να μπορούν πιθανά να βελτιώσουν ακόμα περισσότερο την προβλεπτική ικανότητα του μοντέλου. Τα δεδομένα που εισάγει ένα εργαστήριο για πρόβλεψη, μπορούν να αποθηκεύονται από την εταιρεία, για πιθανή χρήση τους στη μοντελοποίηση, με στόχο τη βελτίωση της. Φυσικά, η εταιρεία οφείλει να διασφαλίζει την ανωνυμία των δεδομένων, καθώς επίσης και την αδυναμία διασταύρωσης τους με φυσικά πρόσωπα, από οποιονδήποτε και για οποιονδήποτε λόγο.

Τέλος, αυτή η συνεργασία με τα διαγνωστικά εργαστήρια, δημιουργεί ένα έτοιμο πελατολόγιο, το οποίο θα έχει χρησιμοποιήσει το μοντέλο και δε θα χρειάζεται να πειστεί για την αποτελεσματικότητα του.

Η χρήση του μοντέλου από τα διαγνωστικά εργαστήρια μπορεί να τιμολογείται σαν μια συνδρομητική υπηρεσία. Τα οφέλη μιας τέτοιας τιμολόγησης για μια επιχείρηση είναι πολλαπλά. Πρώτο και κύριο, η ικανότητα υπολογισμού με μεγάλη ακρίβεια των εσόδων του επόμενου μήνα, τριμήνου, εξαμήνου ή έτους (εξαρτάται από τον τύπο της συνδρομής). Μάλιστα, τα συγκεκριμένα έσοδα είναι επαναλαμβανόμενα, πράγμα που προσδίδει οικονομική σταθερότητα και μεγάλη οικονομική προβλεψιμότητα στην επιχείρηση.

Τα συνδρομητικά μοντέλα δημιουργούν σχέσεις εμπιστοσύνης ανάμεσα στην επιχείρηση και τους πελάτες της. Ένας πελάτης ο οποίος χρησιμοποιεί μια συνδρομητική υπηρεσία για καιρό, έχει πολύ μικρές πιθανότητες μετάβασης σε μια παρόμοια, ανταγωνιστική υπηρεσία μελλοντικά. Επιπλέον, ένας ευχαριστημένος από την υπηρεσία πελάτης, θα συνεχίσει να τη χρησιμοποιεί για καιρό, και εν τέλει θα πληρώσει περισσότερα για αυτήν, απ' ότι θα πλήρωνε

για την αγορά της μία και μόνο φορά. Αυτό από μόνο του οδηγεί, σε βάθος χρόνου, σε αυξημένα έσοδα την εταιρεία.

Επιπρόσθετα, μία συνδρομητική υπηρεσία με δυνατότητα ακύρωσης οποιαδήποτε στιγμή, είναι πιθανότερο να προσελκύσει περισσότερους πελάτες, απ' ότι εάν προτιμηθεί η τιμολόγηση ως ένα προϊόν που κανείς αγοράζει μία φορά, καθώς το κόστος σε αυτήν την περίπτωση, θα ήταν εμφανώς μεγαλύτερο και πιθανά αποτρεπτικό για το διαγνωστικό εργαστήριο.

Ακόμα, η συνεχής επαφή της επιχείρησης με τους πελάτες τις λόγω της συνδρομητικής δόμησης της, προσφέρει μεγάλες ευκαιρίες για συμπληρωματικές πωλήσεις στο μέλλον. Για παράδειγμα, ένας πελάτης ο οποίος χρησιμοποιεί για καιρό το μοντέλο πρόβλεψης ύπαρξης νεφρικών λίθων, είναι πολύ πιθανό να θελήσει να χρησιμοποιήσει και κάποιο άλλο μοντέλο της εταιρείας. Οι συνδρομές δημιουργούν ένα χρήσιμο κανάλι επικοινωνίας για τη μελλοντική προώθηση νέων προϊόντων και υπηρεσιών.

Φυσικά, όντας σε έναν κλάδο του οποίου πρώτο μέλημα δεν είναι το κέρδος, αλλά η βελτίωση της ανθρώπινης ζωής, οφείλουν να αναφερθούν τα πλεονεκτήματα χρήσης ενός τέτοιου προβλεπτικού μοντέλου για τον ασθενή. Με τη χρήση του μοντέλου, ο εκάστοτε ιατρός θα είναι σε θέση να προτείνει πολύ γρηγορότερα, τις απαραίτητες αλλαγές στην καθημερινότητα του ασθενούς, απ' ότι το έκανε έως τώρα, με τις κλασικές μεθόδους ανίχνευσης λίθων.

Συνήθως, ο ιατρός προτείνει πρόσληψη όσο το δυνατόν περισσότερων υγρών, αλλαγές στη διατροφή του ασθενούς για μείωση της ποσότητας πρόσληψης οξαλικού, μείωση της πρόσληψης νατρίου (ειδικά από το αλάτι), καθώς επίσης και μια συνολική βελτίωση του τρόπου ζωής του ασθενούς, ιδιαίτερα με ένταξη γυμναστικής στην καθημερινότητα του. Η υιοθέτηση ενός υγιούς συνολικά τρόπου ζωής μειώνει αποδεδειγμένα την πιθανότητα εμφάνισης νεφρικών λίθων. Φυσικά, μπορεί να χρειαστεί και η λήψη φαρμάκων, ιδιαίτερα για ασθενείς με έντονους πόνους ή συγκεκριμένους τύπους λίθων, όπως αυτά περιεγράφηκαν στην «Εισαγωγή».

Επίσης, ο ασθενής δε θα είναι αναγκαίο να επιβαρύνεται με περισσότερες διαγνωστικές εξετάσεις, αφού ο ιατρός θα είναι σε θέση να προβλέπει μονάχα από την εξέταση των ούρων του, αν έχει ή όχι νεφρικό λίθο. Πέραν της ταλαιπωρίας, το μοντέλο συμφέρει οικονομικά και τον ίδιο τον ασθενή, και εν τέλει, το σύστημα υγείας στο σύνολο του.

Τέλος, είναι πολύ σημαντικό να καταστεί σαφές πως το συγκεκριμένο μοντέλο είναι σε θέση να προβλέπει την ύπαρξη νεφρικών λίθων που σχηματίζονται με βάση το ασβέστιο αλλά και το ουρικό οξύ, καλύπτοντας το 80% των συνολικών λίθων που εμφανίζονται στο ανθρώπινο είδος. Θα ήταν πολύ χρήσιμο για τη χρηστικότητα του μοντέλου, η διεύρυνση της χρήσης του και για άλλους τύπους λίθων μελλοντικά, ώστε να αποτελεί μία πλήρης και αξιόπιστη λύση, που θα προβλέπει με ακρίβεια την ύπαρξη κάθε τύπου λίθου.

Η μηχανική μάθηση και το AI έχει μπει για τα καλά στις ζωές όλων μας και η χρήση του σε διάφορους τομείς της ανθρώπινης δραστηριότητας μονάχα θα αυξάνεται. Ο τομέας της υγείας είναι ένας από τους τομείς που έχουν να κερδίσουν πολλά από την υιοθέτηση μοντέλων σαν και το παρών. Όπως συνηθίζεται πάντα, οι καινοτόμοι κάθε νέας τεχνολογίας, που είναι σε θέση να δίνουν αξιόπιστες λύσεις στους καταναλωτές με τα προϊόντα ή τις υπηρεσίες που παράγουν, κερδίζουν πάντα τη "μερίδα του λέοντος" της παγκόσμιας αγοράς. Έτσι, όσοι εκμεταλλευτούν πρώτοι τα όσα έχει να προσφέρει το AI στον τομέα της υγείας, θα ορίσουν και την αγορά του μέλλοντος στο συγκεκριμένο κλάδο.

Βιβλιογραφία

1. Cruz, J. A., Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics (Vol. 2)
2. Curtis, J. R., Weinblatt, M., Saag, K., Bykerk, V. P., Furst, D. E., Fiore, S., St John, G., Kimura, T., Zheng, S., Bingham, C. O., Wright, G., Bergman, M., Nola, K., Charles-Schoeman, C., Shadick, N. (2021). Data-Driven Patient Clustering and Differential Clinical Outcomes in the Brigham and Women's Rheumatoid Arthritis Sequential Study Registry. Arthritis Care and Research, 4, 471–480.
3. Maceachern, S. J., Forkert, N. D. (2021). Machine learning for precision medicine. Genome, 64(4), 416–425
4. Andrews, D.F., Herzberg, A.M. (1985). Physical Characteristics of Urines With and Without Crystals. In: Data. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-5098-2_45
5. InformedHealth.org [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2006-. Kidney stones: Overview. 2016 Feb 25 [Updated 2019 Feb 28]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK348937/
6. Mitchell T, Kumar P, Reddy T, Wood KD, Knight J, Assimos DG, Holmes RP. (2019). Dietary oxalate and kidney stone formation. The American Journal of Physiology-Renal Physiology. 316(3), 409-413.
7. Fontenelle LF, Sarti TD. (2019). Kidney Stones: Treatment and Prevention. American Family Physician. 15;99(8). 490-496.
8. Massey LK, Whiting SJ. (1995). Dietary salt, urinary calcium, and kidney stone risk. Nutrition Reviews. 53(5). 131-139.
9. Siener R. (2021). Nutrition and Kidney Stone Disease. Nutrients. 13(6). 1917.
10. Heilberg IP, Goldfarb DS. (2013). Optimum nutrition for kidney stone disease. Chronic Kidney Disease. 20(2). 165-174.
11. Krieger NS, Asplin JR, Frick KK, Granja I, Culbertson CD, Ng A, Grynpas MD, Bushinsky DA. (2015). Effect of Potassium Citrate on Calcium Phosphate Stones in a Model of Hypercalciuria. Journal of the American Society of Nephrology. 26(12). 3001-3008.
12. Cunha TDS, Gomes SA, Heilberg IP. (2021). Thiazide and thiazide-like diuretics in nephrolithiasis. Brazilian Journal of Nephrology (BJN). 43(1). 103-109.
13. Campschroer T, Zhu X, Vernooij RW, Lock MT. (2018). Alpha-blockers as medical expulsive therapy for ureteral stones. Cochrane Library: Cochrane Reviews. 4(4). CD008509.

14. Chung VY, Turney BW. (2016). The success of shock wave lithotripsy (SWL) in treating moderate-sized (10-20 mm) renal stones. Urolithiasis. 44(5). 441-444.
15. Wason SE, Monfared S, Ionson A, Klett DE, Leslie SW. (2022). Ureteroscopy. StatPearls Publishing, Treasure Island (FL). PMID: 32809391.
16. Knoll T, Daels F, Desai J, Hoznek A, Knudsen B, Montanari E, Scoffone C, Skolarikos A, Tozawa K. (2017). Percutaneous nephrolithotomy: technique. World Journal of Urology. 35(9). 1361-1368.
17. Flasar C. (2008). What is urine specific gravity? Nursing. 38(7). 14.
18. Wagner CA, Mohebbi N. (2010). Urinary pH and stone formation. Journal of Nephrology. 23(16). 165-169.
19. Kamel KS, Ethier JH, Richardson RM, Bear RA, Halperin ML. (1990). Urine electrolytes and osmolality: when and how to use them. Journal of Nephrology. 10(2). 89-102.
20. Fazil Marickar YM. (2010). Electrical conductivity and total dissolved solids in urine. Urological Research. 38(4). 233-235.
21. Yang B, Bankir L. (2005). Urea and urine concentrating ability: new insights from studies in mice. American Journal of Physiology-Renal Physiology. 288(5). 881-896.
22. Lam GS, Asplin JR, Halperin ML. (2000). Does a high concentration of calcium in the urine cause an important renal concentrating defect in human subjects? Clinical science (London). 98(3). 313-319.
23. Varsou, DD., Tsoumanis, A., Papadiamantis, A.G., Melagraki, G., Afantitis, A. (2023). Isalos Predictive Analytics Platform: Cheminformatics, Nanoinformatics, and Data Mining Applications. In: Hong, H. (eds) Machine Learning and Deep Learning in Computational Toxicology. Computational Methods in Engineering & the Sciences. Springer, Cham. https://doi.org/10.1007/978-3-031-20730-3_9
24. Patro, S.G., & Sahu, K.K. (2015). Normalization: A Preprocessing Stage. ArXiv, abs/1503.06462.
25. Daniel Pelliccia. (2022). The Kennard-Stone algorithm. Nirpy Research. Available from: https://nirpyresearch.com/kennard-stone-algorithm/
26. Jason Brownlee. (2017). Difference Between Classification and Regression in Machine Learning. Available from: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/
27. Varun Jain. (2022). Introduction to kNN algorithms. Analytics Vidhya. Available from: https://www.analyticsvidhya.com/blog/2022/01/introduction-to-knn-algorithms/
28. IBM. What is the k-nearest neighbors algorithm?. IBM. Available from: https://www.ibm.com/topics/knn

29. JavaTpoint. K-Nearest Neighbor (KNN) Algorithm for Machine Learning. JavaTpoint. Available from: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
30. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology. 56(1). 45-50.
31. Expert.ai. F-score (F-measure, F1 measure). Expert.ai. Available from: https://www.expert.ai/glossary-of-ai-terms/f-score-f-measure-f1-measure/
32. Khan SR, Pearle MS, Robertson WG, Gambaro G, Canales BK, Doizi S, Traxer O, Tiselius HG. (2016). Kidney stones. Nature Reviews Disease Primers. 16008(2).
33. Jovel J., Greiner R. (2021). An Introduction to Machine Learning Approaches for Biomedical Research.. Frontiers in Medicine. 771607(8).
34. Manish K. C., Leslie W. S. (2022). Uric Acid Nephrolithiasis. StatPearls Publishing.
35. Aihemaitijiang S., Zhang Y., Zhang L., Yang J., Ye C., Halimulati M., Zhang W., Zhang Z. (2020). The Association between Purine-Rich Food Intake and Hyperuricemia: A Cross-Sectional Study in Chinese Adult Residents. Nutrients. 15;12(12):3835.

Χρήσιμοι Διαδικτυακοί Τόποι

Kaggle: Your Machine Learning and Data Science Community: https://www.kaggle.com/

Novamechanics LTD.: https://novamechanics.com/

Isalos Analytics Platform: https://isalos.novamechanics.com/

Enalos Cloud: http://www.enaloscloud.novamechanics.com/

Παράρτημα

Παρακάτω παρατίθενται τα 3 σημαντικότερα σετ δεδομένων που αναλύθηκαν κατά τη διάρκεια της παρούσας εργασίας.

Το πρώτο σετ είναι αυτούσιο το αρχικό, όπως αυτό χρησιμοποιήθηκε για τη δημιουργία των πρώτων μοντελοποιήσεων, τα αποτελέσματα των οποίων παρουσιάστηκαν στην ενότητα 3.1.

User Row ID	Gravity	pH	Osmolarity	Conductivity	Urea	Calcium	Target
1	1.021	4.91	725	14	443	2.45	FALSE
2	1.017	5.74	577	20	296	4.49	FALSE
3	1.008	7.2	321	14.9	101	2.36	FALSE
4	1.011	5.51	408	12.6	224	2.15	FALSE
5	1.005	6.52	187	7.5	91	1.16	FALSE
6	1.02	5.27	668	25.3	252	3.34	FALSE
7	1.012	5.62	461	17.4	195	1.4	FALSE
8	1.029	5.67	1107	35.9	550	8.48	FALSE
9	1.015	5.41	543	21.9	170	1.16	FALSE
10	1.021	6.13	779	25.7	382	2.21	FALSE
11	1.011	6.19	345	11.5	152	1.93	FALSE
12	1.025	5.53	907	28.4	448	1.27	FALSE
13	1.006	7.12	242	11.3	64	1.03	FALSE
14	1.007	5.35	283	9.9	147	1.47	FALSE
15	1.011	5.21	450	17.9	161	1.53	FALSE
16	1.018	4.9	684	26.1	284	5.09	FALSE
17	1.007	6.63	253	8.4	133	1.05	FALSE
18	1.025	6.81	947	32.6	395	2.03	FALSE
19	1.008	6.88	395	26.1	95	7.68	FALSE
20	1.014	6.14	565	23.6	214	1.45	FALSE
21	1.024	6.3	874	29.9	380	5.16	FALSE
22	1.019	5.47	760	33.8	199	0.81	FALSE
23	1.014	7.38	577	30.1	87	1.32	FALSE
24	1.02	5.96	631	11.2	422	1.55	FALSE
25	1.023	5.68	749	29	239	1.52	FALSE
26	1.017	6.76	455	8.8	270	0.77	FALSE
27	1.017	7.61	527	25.8	75	2.17	FALSE
28	1.01	6.61	225	9.8	72	0.17	FALSE
29	1.008	5.87	241	5.1	159	0.83	FALSE
30	1.02	5.44	781	29	349	3.04	FALSE
31	1.017	7.92	680	25.3	282	1.06	FALSE
32	1.019	5.98	579	15.5	297	3.93	FALSE
33	1.017	6.56	559	15.8	317	5.38	FALSE
34	1.008	5.94	256	8.1	130	3.53	FALSE
35	1.023	5.85	970	38	362	4.54	FALSE

36	1.02	5.66	702	23.6	330	3.98	FALSE
37	1.008	6.4	341	14.6	125	1.02	FALSE
38	1.02	6.35	704	24.5	260	3.46	FALSE
39	1.009	6.37	325	12.2	97	1.19	FALSE
40	1.018	6.18	694	23.3	311	5.64	FALSE
41	1.021	5.33	815	26	385	2.66	FALSE
42	1.021	5.94	774	27.9	325	6.96	TRUE
43	1.024	5.77	698	19.5	354	13	TRUE
44	1.024	5.6	866	29.5	360	5.54	TRUE
45	1.021	5.53	775	31.2	302	6.19	TRUE
46	1.024	5.36	853	27.6	364	7.31	TRUE
47	1.026	5.16	822	26	301	14.34	TRUE
48	1.013	5.86	531	21.4	197	4.74	TRUE
49	1.01	6.27	371	11.2	188	2.5	TRUE
50	1.011	7.01	443	21.4	124	1.27	TRUE
51	1.022	6.21	442	20.6	398	4.18	TRUE
52	1.011	6.13	364	10.9	159	3.1	TRUE
53	1.031	5.73	874	17.4	516	3.01	TRUE
54	1.02	7.94	567	19.7	212	6.81	TRUE
55	1.04	6.28	838	14.3	486	8.28	TRUE
56	1.021	5.56	658	23.6	224	2.33	TRUE
57	1.025	5.71	854	27	385	7.18	TRUE
58	1.026	6.19	956	27.6	473	5.67	TRUE
59	1.034	5.24	1236	27.3	620	12.68	TRUE
60	1.033	5.58	1032	29.1	430	8.94	TRUE
61	1.015	5.98	487	14.8	198	3.16	TRUE
62	1.013	5.58	516	20.8	184	3.3	TRUE
63	1.014	5.9	456	17.8	164	6.99	TRUE
64	1.012	6.75	251	5.1	141	0.65	TRUE
65	1.025	6.9	945	33.6	396	4.18	TRUE
66	1.026	6.29	833	22.2	457	4.45	TRUE
67	1.028	4.76	312	12.4	10	0.27	TRUE
68	1.027	5.4	840	24.5	395	7.64	TRUE
69	1.018	5.14	703	29	272	6.63	TRUE
70	1.022	5.09	736	19.8	418	8.53	TRUE
71	1.025	7.9	721	23.6	301	9.04	TRUE
72	1.009	5.64	386	17.7	104	1.22	FALSE
73	1.015	6.79	541	20.9	187	2.64	FALSE
74	1.01	5.97	343	13.4	126	2.31	FALSE
75	1.02	5.68	876	35.8	308	4.49	FALSE
76	1.017	4.81	410	13.3	195	0.58	TRUE
77	1.024	5.4	803	21.8	394	7.82	TRUE
78	1.016	6.81	594	21.4	255	12.2	TRUE
79	1.015	6.03	416	12.8	178	9.39	TRUE

Ακολουθεί το σετ από το οποίο, όπως συζητήθηκε στο κυρίως μέρος της εργασίας, αφαιρέθηκαν οι παράμετροι ωσμομοριακότητα και αγωγιμότητα, καθώς επίσης και τα 11 άτομα που μείωναν την αποτελεσματικότητα των μοντελοποιήσεων (Ενότητα 3.2). Σε αυτό το σετ, προστίθενται και τα δεδομένα που συλλέχθηκαν από το διαγνωστικό εργαστήριο. Τα νέα αυτά δείγματα, με αρίθμηση που ξεκινά από το 80 και έπειτα, διαχωρίζονται με μία κίτρινη γραμμή, προς ευκολία οποιουδήποτε ενδιαφερόμενου.

User Row

ID	Gravity	pH	Urea	Calcium	Target
1	1.021	4.91	443	2.45	FALSE
2	1.017	5.74	296	4.49	FALSE
3	1.008	7.2	101	2.36	FALSE
4	1.011	5.51	224	2.15	FALSE
5	1.005	6.52	91	1.16	FALSE
6	1.020	5.27	252	3.34	FALSE
7	1.012	5.62	195	1.4	FALSE
9	1.015	5.41	170	1.16	FALSE
10	1.021	6.13	382	2.21	FALSE
11	1.011	6.19	152	1.93	FALSE
12	1.025	5.53	448	1.27	FALSE
13	1.006	7.12	64	1.03	FALSE
14	1.007	5.35	147	1.47	FALSE
15	1.011	5.21	161	1.53	FALSE
16	1.018	4.9	284	5.09	FALSE
17	1.007	6.63	133	1.05	FALSE
18	1.025	6.81	395	2.03	FALSE
19	1.008	6.88	95	7.68	FALSE
20	1.014	6.14	214	1.45	FALSE
22	1.019	5.47	199	0.81	FALSE
23	1.014	7.38	87	1.32	FALSE
24	1.020	5.96	422	1.55	FALSE
25	1.023	5.68	239	1.52	FALSE
26	1.017	6.76	270	0.77	FALSE
27	1.017	7.61	75	2.17	FALSE
28	1.010	6.61	72	0.17	FALSE
29	1.008	5.87	159	0.83	FALSE
30	1.020	5.44	349	3.04	FALSE
31	1.017	7.92	282	1.06	FALSE
32	1.019	5.98	297	3.93	FALSE
33	1.017	6.56	317	5.38	FALSE
34	1.008	5.94	130	3.53	FALSE
35	1.023	5.85	362	4.54	FALSE
36	1.020	5.66	330	3.98	FALSE
37	1.008	6.4	125	1.02	FALSE
38	1.020	6.35	260	3.46	FALSE

39	1.009	6.37	97	1.19	FALSE
40	1.018	6.18	311	5.64	FALSE
41	1.021	5.33	385	2.66	FALSE
42	1.021	5.94	325	6.96	TRUE
43	1.024	5.77	354	13.0	TRUE
44	1.024	5.6	360	5.54	TRUE
45	1.021	5.53	302	6.19	TRUE
46	1.024	5.36	364	7.31	TRUE
47	1.026	5.16	301	14.34	TRUE
48	1.013	5.86	197	4.74	TRUE
51	1.022	6.21	398	4.18	TRUE
53	1.031	5.73	516	3.01	TRUE
54	1.020	7.94	212	6.81	TRUE
55	1.040	6.28	486	8.28	TRUE
57	1.025	5.71	385	7.18	TRUE
58	1.026	6.19	473	5.67	TRUE
59	1.034	5.24	620	12.68	TRUE
60	1.033	5.58	430	8.94	TRUE
63	1.014	5.9	164	6.99	TRUE
64	1.012	6.75	141	0.65	TRUE
66	1.026	6.29	457	4.45	TRUE
68	1.027	5.4	395	7.64	TRUE
69	1.018	5.14	272	6.63	TRUE
70	1.022	5.09	418	8.53	TRUE
71	1.025	7.9	301	9.04	TRUE
72	1.009	5.64	104	1.22	FALSE
73	1.015	6.79	187	2.64	FALSE
74	1.010	5.97	126	2.31	FALSE
76	1.017	4.81	195	0.58	TRUE
77	1.024	5.4	394	7.82	TRUE
78	1.016	6.81	255	12.2	TRUE
79	1.024	5.91	218	2.37	TRUE
80	1.026	6.23	250	7.33	TRUE
81	1.016	5.7	210	9.86	TRUE
82	1.021	6.02	351	6.82	TRUE
83	1.017	6.33	370	5.93	TRUE
84	1.015	6.91	563	10.0	TRUE
85	1.013	6.11	100	6.02	TRUE
86	1.020	5.88	136	6.91	TRUE
87	1.014	6.24	267	7.3	TRUE
88	1.020	5.91	254	6.42	TRUE
89	1.010	7.37	333	7.3	TRUE
90	1.015	7.13	170	3.03	FALSE
91	1.010	6.11	192	2.92	FALSE
92	1.009	7.22	105	2.14	FALSE

93	1.012	6.32	316	4.05	FALSE
94	1.020	6.25	241	3.77	FALSE
95	1.017	6.92	175	2.62	FALSE
96	1.013	7.15	155	2.49	FALSE
97	1.015	6.87	230	2.17	FALSE
98	1.010	7.63	25	2.02	FALSE
99	1.025	6.06	106	2.5	FALSE
100	1.025	6.43	146	2.61	FALSE
101	1.011	6.97	210	2.52	FALSE

Τέλος, παρατίθονται τα training και test set που προέκυψαν μετά τη χρήση του random partitioning για τη δημιουργία του μοντέλου (Ενότητα 3.4), προκειμένου η διαδικασία να μπορεί να αναπαραχθεί επακριβώς, από οποιονδήποτε ενδιαφερόμενο. Είναι σημαντικό να τονιστεί ότι το βήμα της κανονικοποίησης πρέπει να γίνει πριν το διαχωρισμό των δύο σετ, ούτως ώστε οι τιμές να αντικατοπτρίζουν επακριβώς την απόκλιση από το μέσο όρο του συνολικού δείγματος.

Πρώτα το Training Set:

User Row ID	Gravity	pH	Urea	Calcium	Target
3	1.008	7.2	101	2.36	FALSE
4	1.011	5.51	224	2.15	FALSE
5	1.005	6.52	91	1.16	FALSE
6	1.020	5.27	252	3.34	FALSE
9	1.015	5.41	170	1.16	FALSE
11	1.011	6.19	152	1.93	FALSE
13	1.006	7.12	64	1.03	FALSE
15	1.011	5.21	161	1.53	FALSE
16	1.018	4.9	284	5.09	FALSE
17	1.007	6.63	133	1.05	FALSE
20	1.014	6.14	214	1.45	FALSE
23	1.014	7.38	87	1.32	FALSE
24	1.020	5.96	422	1.55	FALSE
25	1.023	5.68	239	1.52	FALSE
26	1.017	6.76	270	0.77	FALSE
27	1.017	7.61	75	2.17	FALSE
28	1.010	6.61	72	0.17	FALSE
31	1.017	7.92	282	1.06	FALSE
33	1.017	6.56	317	5.38	FALSE
34	1.008	5.94	130	3.53	FALSE
35	1.023	5.85	362	4.54	FALSE
36	1.020	5.66	330	3.98	FALSE
39	1.009	6.37	97	1.19	FALSE
40	1.018	6.18	311	5.64	FALSE
41	1.021	5.33	385	2.66	FALSE
42	1.021	5.94	325	6.96	TRUE

43	1.024	5.77	354	13.0	TRUE
44	1.024	5.6	360	5.54	TRUE
45	1.021	5.53	302	6.19	TRUE
46	1.024	5.36	364	7.31	TRUE
47	1.026	5.16	301	14.34	TRUE
48	1.013	5.86	197	4.74	TRUE
51	1.022	6.21	398	4.18	TRUE
55	1.040	6.28	486	8.28	TRUE
57	1.025	5.71	385	7.18	TRUE
58	1.026	6.19	473	5.67	TRUE
59	1.034	5.24	620	12.68	TRUE
63	1.014	5.9	164	6.99	TRUE
68	1.027	5.4	395	7.64	TRUE
69	1.018	5.14	272	6.63	TRUE
71	1.025	7.9	301	9.04	TRUE
72	1.009	5.64	104	1.22	FALSE
73	1.015	6.79	187	2.64	FALSE
74	1.010	5.97	126	2.31	FALSE
77	1.024	5.4	394	7.82	TRUE
78	1.016	6.81	255	12.2	TRUE
80	1.026	6.23	250	7.33	TRUE
81	1.016	5.7	210	9.86	TRUE
82	1.021	6.02	351	6.82	TRUE
84	1.015	6.91	563	10.0	TRUE
87	1.014	6.24	267	7.3	TRUE
88	1.020	5.91	254	6.42	TRUE
89	1.010	7.37	333	7.3	TRUE
90	1.015	7.13	170	3.03	FALSE
91	1.010	6.11	192	2.92	FALSE
92	1.009	7.22	105	2.14	FALSE
93	1.012	6.32	316	4.05	FALSE
94	1.020	6.25	241	3.77	FALSE
95	1.017	6.92	175	2.62	FALSE
97	1.015	6.87	230	2.17	FALSE
98	1.010	7.63	25	2.02	FALSE
100	1.025	6.43	146	2.61	FALSE
101	1.011	6.97	210	2.52	FALSE

Ακολούθως, το Test Set:

User Row ID	Gravity	pH	Urea	Calcium	Target
1	1.021	4.91	443	2.45	FALSE
2	1.017	5.74	296	4.49	FALSE
7	1.012	5.62	195	1.4	FALSE
10	1.021	6.13	382	2.21	FALSE
12	1.025	5.53	448	1.27	FALSE
14	1.007	5.35	147	1.47	FALSE

18	1.025	6.81	395	2.03	FALSE
19	1.008	6.88	95	7.68	FALSE
22	1.019	5.47	199	0.81	FALSE
29	1.008	5.87	159	0.83	FALSE
30	1.020	5.44	349	3.04	FALSE
32	1.019	5.98	297	3.93	FALSE
37	1.008	6.4	125	1.02	FALSE
38	1.020	6.35	260	3.46	FALSE
53	1.031	5.73	516	3.01	TRUE
54	1.020	7.94	212	6.81	TRUE
60	1.033	5.58	430	8.94	TRUE
64	1.012	6.75	141	0.65	TRUE
66	1.026	6.29	457	4.45	TRUE
70	1.022	5.09	418	8.53	TRUE
76	1.017	4.81	195	0.58	TRUE
79	1.024	5.91	218	2.37	TRUE
83	1.017	6.33	370	5.93	TRUE
85	1.013	6.11	100	6.02	TRUE
86	1.020	5.88	136	6.91	TRUE
96	1.013	7.15	155	2.49	FALSE
99	1.025	6.06	106	2.5	FALSE