Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων & Φυσικων Επιστημων

Κατευθυνση Φυσικου Εφαρμογων

Ανάπτυξη μοντέλου πρόβλεψης επίδρασης Ιοντίζουσας Ακτινοβολίας (X, γ) σε ανθρώπινα καρκινικά κύτταρα μέσω αλγορίθμων Μηχανικής Μάθησης

Διπλωματικη Εργασια

Δωροθεα Μανετα

Επιβλέπων: Δρ. Αλέξανδρος Γεωργακίλας,
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων & Φυσικων Επιστημων
Κατευθυνση Φυσικου Εφαρμογων

Ανάπτυξη μοντέλου πρόβλεψης επίδρασης Ιοντίζουσας Ακτινοβολίας (X, γ) σε ανθρώπινα καρκινικά κύτταρα μέσω αλγορίθμων Μηχανικής Μάθησης

Διπλωματικη Εργασια

Δωροθεα Μανετα

Επιβλέπων: Δρ. Αλέξανδρος Γεωργακίλας,
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23η Σεπτεμβρίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Γεωργακίλας	Κωνσταντίνος Αναγνωστόπουλος	Κωνσταντίνος Κουσουρής
Αναπ. Καθηγητής Ε.Μ.Π.	Αναπ. Καθηγητής Ε.Μ.Π.	Επίκ. Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

(Υπογραφή)

.....

Δωροθεα Μανετα

Διπλωματούχος Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Ε.Μ.Π.

Αντικείμενο Διπλωματικής Εργασίας

Με την εξέλιξη των πειραματικών τεχνικών η βιολογική έρευνα δίνει ολοένα και περισσότερα δεδομένα τα οποία ψηφιοποιούνται, αποθηκεύονται και μπορούν να μελετηθούν από ερευνητές, καθηγητές και φοιτητές παγκοσμίως. Αυτός ο όγκος δεδομένων μπορεί να αξιοποιηθεί και να λειτουργήσει ως εργαλείο πιθανής συσχέτισης, πρόβλεψης και πρόληψης για αντιμετώπιση μελλοντικών καταστάσεων και εξοικονόμηση χρόνου. Συγκεκριμένα, ο τομέας της ραδιοβιολογίας που μελετά τη δράση της ιοντίζουσας ακτινοβολίας σε ζωντανούς οργανισμούς έχει προσφέρει πλήθος πειραματικών δεδομένων σχετικά με την επίδραση της ακτινοβολίας στα καρκινικά κύτταρα μέσω της μεθόδου *clonogenic assay*.

Η διπλωματική εργασία πραγματεύεται την ανάπτυξη ενός μοντέλου πρόβλεψης της επιβίωσης ανθρωπίνων καρκινικών κυττάρων μετά από έκθεση σε ιοντίζουσα ακτινοβολία. Για να επιτευχθεί αυτό, επιστρατεύονται τεχνικές μηχανικές μάθησης που είναι σε θέση να αντιμετωπίζουν μεγάλους όγκους δεδομένων που δεν έχουν απαραίτητα γραμμικές σχέσεις μεταξύ τους. Ειδικότερα, αντλούνται ραδιοβιολογικά δεδομένα από εύρος 35 - 40 χρόνων, σχετικά με 8 καρκινικές κυτταρικές σειρές που χρησιμοποιούνται ευρέως πειραματικά. Τα δεδομένα δίνουν πληροφορίες ως προς την κυτταρική σειρά, το χρονικό καλλιέργειας των κυττάρων, τον τύπο της ιοντίζουσας ακτινοβολίας, το ρυθμό δόσης και το ποσοστό επιβίωσης για διάφορες δόσεις από 2 έως 8 Gy. Σκοπός της εργασίας είναι η ανάπτυξη ενός μοντέλου που θα είναι σε θέση να προσδιορίσει το ποσοστό επιβίωσης συναρτήσει αυτών των παραμέτρων προσθέτοντας μερικές επιπλέον στην πορεία.

Ξεκινώντας από την προ-επεξεργασία, το σύνολο των δεδομένων υφίσταται μετατροπές για να είναι κατάλληλο να χρησιμοποιηθεί από τους αλγορίθμους. Εφαρμόζονται κατά κύριο λόγο δύο προσεγγίσεις με τη βοήθεια των μη παραμετρικών αλγορίθμων Random Forest και Gradient Boosting. Αναπτύσσονται δύο μοντέλα για τον καθένα, ένα βασικό κι ένα νέο, αφού πραγματοποιηθεί αναζήτηση των παραμέτρων που βελτιστοποιούν τα αποτελέσματα. Καταγράφονται η απόδοση και τα αποτελέσματά τους, ενώ στο τέλος πραγματοποιείται σύγκριση για να αναδειχθεί το καλύτερο για τα υπάρχοντα δεδομένα.

Λέξεις κλειδιά

Ραδιοβιολογία, Ιοντίζουσα ακτινοβολία, Ακτίνες Χ,γ Κλάσμα κυτταρικής επιβίωσης, Καρκίνος, Μηχανική μάθηση, Παλινδρόμηση, Τυχαίο δάσος, Ενδυναμωμένα δέντρα απόφασης

Abstract

With the evolution of experimental techniques, biological research provides more and more data that are digitalized, saved on clouds and can be studied by other researchers, professors and students globally. This volume of data can be utilized as a correlation, prediction and prevention tool in order to cope with situations that may arise in the future and to save time while doing so. In particular, radiobiology which studies the action of ionizing radiation on living organisms has offered significant amount of data that contributes to the study of the interactions between radiation and cancer cells by applying clonogenic assays.

The thesis deals with the development of a predictive model regarding the survival of human cancer cells after their exposure to ionizing radiation. To achieve this, machine learning techniques are employed because they are able to deal with large volumes of data, whose variables do not necessarily have linear relationships between them. More specifically radiobiological data are obtained from a range of 35-40 years on 8 cancer cell lines that are widely used experimentally. The data provide intel on the cell line, timing of cell seeding, type of ionizing radiation used, dose rate and the survival fraction for various doses from 2 to 8 Gy. The aim is to develop a model that will be able to determine the survival fraction as a function of these variables and a few more that are added along the way.

Starting with preprocessing, all data is converted to be suitable for use by algorithms. Two approaches are studied with the aid of non-parametric algorithms, Random Forest and Gradient Boosting. For each one, two models are developed, one using the default parameters and one customized after a search of the parameters that optimize the procedure. Their performance and results are noted, while at the end a comparison is made in order to indicate which one performs best for the existing data.

Key words

Radiobiology, Ionizing radiation, X,γ rays, Survival fraction, Cancer, Machine learning, Regression, Random Forest, Gradient Boosting

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε με την υποστήριξη, την αντοχή και την ανοχή που έδειξαν μία σειρά ατόμων τα οποία και ευχαριστώ.

Πρωτίστως, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Δρ. Αλέξανδρο Γεωργακίλα για την ανάθεση ενός θέματος που ξέφευγε από τις γνώσεις μου σε όλους τους τομείς και με προκάλεσε να εκπαιδευτώ, να διαβάσω και να αποκτήσω γνώσεις για ένα σύγχρονο και συνεχώς αναπτυσσόμενο τομέα, τη Μηχανική Μάθηση. Η συνεργασία μας κι οι συμβουλές που μου έδωσε ήταν απαραίτητες τόσο για την εκπόνηση της εργασίας όσο και για την ανάπτυξη των επικοινωνιακών και τεχνικών μου δυνατοτήτων.

Στη συνέχεια, ευχαριστώ τη Βάσω με την οποία ξεκινήσαμε μαζί να δουλεύουμε κι ήταν εκεί για να αντιμετωπίζουμε στιγμές άγνοιας κι ερωτηματικών με γέλιο κι ελπίδα. Φυσικά, θερμές ευχαριστίες οφείλονται στο Δημήτρη που από την αρχή μέχρι το τέλος προσέφερε καθοδήγηση και υποστήριξη αλλά κυρίως έδειξε υπομονή σε κάθε σχόλιο που είχα να κάνω για κάθε βήμα της εργασίας.

Ένα μέρος της εργασίας υλοποιήθηκε κατά τη διάρκεια της κοινωνικής απομόνωσης. Σε εκείνην την περίοδο, πολύτιμη ήταν η επικοινωνία κι η βοήθεια που έλαβα από τις «ξαδέρφες» μου, οι οποίες καθημερινά αντιμετώπιζαν την κατάσταση με χιούμορ. Επιπλέον, το άτομο που έζησε όλη τη διαδικασία από πρώτο χέρι κι έδινε νότα εξέλιξης, ήταν η Μυρόπη, την οποία ευχαριστώ για την αμέριστη υπομονή που έκανε μέχρι το τέλος.

Η εργασία αφιερώνεται στους γονείς μου οι οποίοι με τον τρόπο τους άλλοτε ήσυχο και άλλοτε πιο έντονο, με παρότρυναν να προχωράω βήμα-βήμα στη ζωή. Με αυτόν τον τρόπο συνέβαλαν στην περάτωση της εργασίας και μαζί με αυτήν στο πολυπόθητο πτυχίο.

Πίνακας περιεχομένων

Ευρετήριο εικόνων

1 Εισαγωγή

Στην εισαγωγή της διπλωματικής εργασίας γίνεται αναφορά και μερική ανάλυση στο φυσικό, χημικό, βιολογικό κι υπολογιστικό υπόβαθρο που χρειάζεται ο αναγνώστης για να εξοικειωθεί με τα δεδομένα που χρησιμοποιούνται. Ξεκινώντας από την επιλογή της ακτινοβολίας, περνάει στις διαδικασίες παραγωγής της και στους τρόπους αλληλεπίδρασής της με την ύλη. Εν συνεχεία, γίνεται λόγος για τα είδη της δράσης που σχετίζονται άμεσα με τη βιολογική ύλη και τις πιθανές βλάβες που μπορεί να προκύψουν. Επεξηγείται η έννοια του κυτταρικού θανάτου και το γραμμικό – τετραγωνικό μοντέλο που είναι μία μέθοδος υπολογισμού της κυτταρικής επιβίωσης, η οποία είναι και το θέμα μελέτης της εργασίας. Η ενότητα κλείνει με αναφορά των θεμελιωδών εννοιών της Μηχανικής Μάθησης.

1.1 Ιοντίζουσα Ακτινοβολία

Η μελέτη της επίδρασης της ακτινοβολίας στον ανθρώπινο οργανισμό γίνεται μέσω της ραδιοβιολογίας. Πρόκειται για τη μελέτη του αποτελέσματος που μπορεί να επέλθει σε ένα βιολογικό ιστό, μέσω της ενέργειας που απορροφά κατά την ακτινοβόληση με *ιοντίζουσα ακτινοβολία*. Το είδος αυτής της ακτινοβολίας μεταφέρει ενέργεια ικανή να ιοντίσει ένα άτομο ή μόριο δηλαδή να απομακρύνει ένα ή περισσότερα τροχιακά ηλεκτρόνια από αυτό. Λόγω αυτής της αλλοίωσης, θεωρείται επικίνδυνη για τους ζωντανούς οργανισμούς, καθώς μπορεί να έχει καταστρεπτικές συνέπειες στο μόριο του DNA. Επιπλέον, όμως χάρη σε αυτό το χαρακτηριστικό, η ιοντίζουσα ακτινοβολία μπορεί να χρησιμοποιηθεί στη θεραπεία του καρκίνου ώστε να εξαλείψει κακοήθη κύτταρα.

1.1.1 Είδη ακτινοβολίας

Η ιοντίζουσα ακτινοβολία χωρίζεται σε ηλεκτρομαγνητική και σωματιδιακή. Στην ηλεκτρομαγνητική ακτινοβολία ανήκουν οι ακτίνες Χ και γ ενώ στη σωματιδιακή, η ακτινοβολία α, β και νετρονίων με υψηλές ταχύτητες. Αντικείμενο της παρούσας εργασίας αποτελούν οι ακτινοβολίες ηλεκτρομαγνητικής προέλευσης. Όπως και στις υπόλοιπες μορφές ακτινοβολίας, οι βασικές φυσικές ιδιότητες είναι γνωστές. Χαρακτηρίζονται από κυματοσωματιδιακό δυϊσμό, καθώς μπορούν να περιγραφούν τόσο ως συζευγμένα ηλεκτρομαγνητικά κύματα όσο ως σωματίδια που λέγονται φωτόνια και μεταφέρουν διακριτές ποσότητες ενέργειας κι ορμής.

1.1.1.1 Ακτίνες X

Πρόκειται για ακτινοβολία που έχει μεγαλύτερη ενέργεια από την ακτινοβολία του υπεριώδους και κατά κύριο λόγο μικρότερη από την ακτινοβολία γάμμα. Ανακαλύφθηκαν το 1895 από το γερμανό φυσικό Wilhelm Konrad Röntgen, ο οποίος μελετούσε ακτίνες ηλεκτρονίων σε καθοδικούς σωλήνες. Συγκεκριμένα, είχε καλύψει το σωλήνα, ώστε να μην υπάρχει περαιτέρω διάδοση ακτινών αλλά διαπίστωσε ότι μία εξωτερική φθορίζουσα οθόνη έλαμπε παρόλα αυτά [4]. Αυτό σήμαινε ότι κάποια μορφή ακτινοβολίας μη ορατή από το ανθρώπινο μάτι, είχε διαφύγει του προστατευτικού καλύμματος κι επειδή δε γνώριζε τι είδος ήταν, τις ονόμασε ακτίνες X. Σήμερα, χωρίζονται σε δύο είδη [5]:

- *Μαλακές ακτίνες X:*
  Έχουν μήκος κύματος από 10 έως 0,1 nm κι ενέργεια από 0,10 έως 5 kilo electronvolts (keV).

- *Σκληρές ακτίνες X:*
  Έχουν μήκος κύματος μικρότερο από 0,2 nm κι ενέργεια από 5 έως 10 keV.

Με βάση τα όσα γνωρίζουμε σήμερα η διαδικασία παραγωγής των ακτίνων X είναι η παρακάτω [6]. Ηλεκτρόνια υψηλής ενέργειας παράγονται σε έναν *καθοδικό σωλήνα (X-ray tube)*, επιταχύνονται προς την άνοδο και χτυπούν το μεταλλικό στόχο έχοντας μεγάλη κινητική ενέργεια. Κατά τη σύγκρουση παρατηρούνται τρία φαινόμενα, μερική επιβράδυνση ή ολική ακινητοποίηση των αρχικών ηλεκτρονίων κι αποδέσμευση ηλεκτρονίων από τα άτομα του στόχου. Ολική ακινητοποίηση συμβαίνει όταν το ηλεκτρόνιο συγκρουστεί με πυρήνα ατόμου του στόχου με αποτέλεσμα όλη η ενέργειά του να μετατραπεί σε φωτόνιο X. Η ενέργεια του φωτονίου τότε, είναι η ίδια που είχε το ηλεκτρόνιο στη σύγκρουση. Αυτό το φαινόμενο συμβαίνει σπάνια. Η συνήθης διαδικασία είναι η μερική επιβράδυνση, στην οποία το ηλεκτρόνιο περνά κοντά από κάποιο πυρήνα, λόγω της έλξης χάνει μερική από την ταχύτητά του και κατ' επέκταση ελαττώνεται η κινητική του ενέργεια. Η απώλεια ενέργειας αποδίδεται στο περιβάλλον με τη μορφή φωτονίου X. Το ηλεκτρόνιο συνεχίζει να αλληλεπιδρά με άτομα μέχρι να μηδενιστεί η ενέργειά του. Αυτό το είδος αλληλεπιδράσεων δίνει τη λεγόμενη *ακτινοβολία πέδησης (Bremsstrahlung radiation).* Εν γένει, παράγονται φωτόνια X διαφόρων ενεργειών, οπότε λαμβάνεται ένα συνεχές φάσμα τιμών. Το είδος της ακτινοβολίας που σχετίζεται με την αποδέσμευση ηλεκτρονίων του στόχου καλείται *χαρακτηριστική ακτινοβολία (Characteristic radiation).* Κατά την αποδέσμευση, το άτομο ιοντίζεται και τότε άλλο ηλεκτρόνιο από εξωτερική στοιβάδα έλκεται από την οπή και την καλύπτει. Λόγω της διαφοράς ενέργειας των δύο τροχιακών, η περίσσεια ενέργειας αποδίδεται με τη μορφή φωτονίου X. Αυτές οι μεταπτώσεις έχουν χαρακτηριστικές, διακριτές τιμές οι οποίες είναι μεγαλύτερες της μέγιστης τιμής που μπορεί να λάβει ένα φωτόνιο ακτινοβολίας πέδησης. Στο 99% των περιπτώσεων η

ενέργεια μετατρέπεται σε θερμότητα και μόνο το 1% σε φωτόνιο ακτίνας Χ. Ο αριθμός των ακτίνων Χ είναι ανάλογος των ηλεκτρονίων που παράγονται στην κάθοδο.

Με την εξέλιξη της οργανολογίας και της τεχνολογίας είναι δυνατό να παραχθούν ακτίνες Χ της τάξης των mega electronvolts (MeV) με τη χρήση των *γραμμικών επιταχυντών (linear accelerator, linac)* [7]. Αυτές οι ενέργειες είναι επιθυμητές για τη δημιουργία βιολογικής βλάβης για τη θεραπεία του καρκίνου. Η αρχή λειτουργίας αυτών των διατάξεων είναι η επιτάχυνση των ηλεκτρονίων μέσω ενός κυματοδηγού με χρήση εναλλασσόμενων μικροκυματικών πεδίων. Ο κυματοδηγός επιτρέπει την πραγματοποίηση της επιτάχυνσης σε ευθείες τροχιές μέσω ηλεκτροδίων. Κάθε φορά που τα ηλεκτρόνια διέρχονται από ένα ηλεκτρόδιο, η τάση της ταλάντωσης αλλάζει πολικότητα ώστε όταν τα σωματίδια φτάσουν στο κενό μεταξύ των ηλεκτροδίων, το ηλεκτρικό πεδίο να είναι στη σωστή κατεύθυνση για να τα επιταχύνει. Εν συνεχεία, τα ηλεκτρόνια συγκρούονται με βαρύ μεταλλικό στόχο και παράγουν ακτίνες Χ υψηλών ενεργειών μέσω των φαινομένων που αναπτύχθηκαν προηγουμένως. Το εύρος των ενεργειών που μπορούν να παραχθούν είναι από 4 έως 25 MeV. Κάποιοι παρέχουν ακτίνες μόνο στη χαμηλή κλίμακα των 4-6 MeV. Πλέον ένας τυπικός γραμμικός επιταχυντής υψηλής ενέργειας μπορεί να δώσει 2-3 ενέργειες.

1.1.1.2 Ακτίνες γ

Ο διαχωρισμός μεταξύ ακτίνων Χ και γ είναι σχετικά αυθαίρετος. Ένας συνήθης τρόπος είναι ένα όριο των 10^{-11} m, όπου για χαμηλότερα μήκη κύματος η ακτινοβολία κατατάσσεται σε ακτίνες γ. Επίσης, άλλη διαφορά έγκειται στον τρόπο παραγωγής τους καθώς όπως προέκυψε από τα παραπάνω οι ακτίνες Χ παράγονται εξωπυρηνικά ενώ οι ακτίνες γ ενδοπυρηνικά [5]. Ειδικότερα, προέρχονται από τη ραδιενεργή διάσπαση πυρήνων, οι οποίοι συνήθως δίνουν ενέργειες μερικών εκατοντάδων keV. Τυπικά συνοδεύουν μία εκ των δύο άλλων μορφών ακτινοβολίας άλφα και βήτα. Πρόκειται για εξαιρετικά διεισδυτικές ακτίνες χωρίς μάζα και φορτίο και μπορούν να απορροφηθούν μόνο από υλικά υψηλού ατομικού αριθμού Z όπως ο μόλυβδος.

Ανακαλύφθηκαν το 1900 από το γάλλο φυσικοχημικό Paul Villard, ο οποίος μελετούσε ακτινοβολία προερχόμενη από το ράδιο (Ra). Κατά την άλφα διάσπαση ατόμου μεγάλου μαζικού αριθμού, αρχικά απελευθερώνεται ένας πυρήνας Ηλίου ($^{4}_{2}$He) και προκύπτει ένα ασταθές, διεγερμένο στοιχείο. Τα νουκλεόνιά του είναι σε αταξία και «επιδιώκουν» να βρουν τη βασική τους κατάσταση. Καθώς επαναταχτοποιούνται, απελευθερώνουν ένα τεράστιο ποσό ενέργειας με τη μορφή φωτονίου, που καλείται φωτόνιο γ. Στη βήτα διάσπαση, το άτομο διασπάται απελευθερώνοντας ένα ηλεκτρόνιο και ένα αντινετρίνο. Το νέο, διεγερμένο άτομο που προκύπτει προσπαθεί να επιστρέψει στη θεμελιώδη κατάσταση ελευθερώνοντας αντίστοιχα φωτόνιο γ.

Ένα υλικό που εκπέμπει ακτινοβολία γάμμα που χρησιμοποιείται για ακτινοβόληση ή απεικόνιση καλείται πηγή γάμμα, ραδιενεργός πηγή ή πηγή ισοτόπου. Οι φυσικές πηγές των ακτίνων γ στη Γη περιλαμβάνουν διάσπαση από φυσικά ραδιοϊσότοπα όπως το κάλιο

και το ράδιο κι ως δευτερεύουσα ακτινοβολία ποικίλες ατμοσφαιρικές αλληλεπιδράσεις με σωματίδια κοσμικών ακτίνων. Τα ραδιενεργά στοιχεία συναντώνται στον αέρα, στο έδαφος ακόμα και στις τροφές. Κάποια έχουν παραχθεί τεχνητά ως αποτέλεσμα πυρηνικής σχάσης και θερμοπυρηνικών εκρήξεων. Μέσω της πυρηνικής σχάσης δημιουργείται ευρύ φάσμα προϊόντων, τα περισσότερα από τα οποία είναι ραδιοϊσότοπα.

1.1.2 Πηγές ακτινοβολίας

Τα δεδομένα στα οποία βασίζεται η διπλωματική εργασία αναφέρονται σε ακτίνες Χ και γ. Οι πρώτες έχουν εύρος από 100 έως 320 kVp και από 4 έως 16 MeV που παράγονται είτε σε λυχνίες είτε σε ιατρικούς γραμμικούς επιταχυντές. Οι δεύτερες προέρχονται από το κοβάλτιο (^{60}Co) και το κέσιο (^{137}Cs) που εκπέμπουν ακτίνες χαρακτηριστικής ενέργειας.

Ιστορικά, οι ακτίνες Χ από τις λυχνίες είχαν πρώτες εφαρμογή στη ραδιοθεραπεία. Πριν την εποχή του Β' Παγκοσμίου Πολέμου, οι εφαρμογές πραγματοποιούνταν με διατάξεις που έφταναν το πολύ τα 150 kV [8]. Ωστόσο, γρήγορα έγινε αντιληπτό ότι για να υπάρξει αποτέλεσμα σε όγκους βαθιά στο σώμα, χρειαζόταν μεγαλύτερη ενέργεια ακτίνας της τάξης των MeV. Ενώ κατασκευάστηκαν τέτοια μηχανήματα τις επόμενες δεκαετίες, το πρόβλημά τους ήταν οι διαστάσεις αφού τα περισσότερα νοσοκομεία δεν είχαν τέτοιες προδιαγραφές. Απ' την άλλη πλευρά, το ράδιο που ήταν φυσικό ραδιοϊσότοπο και μπορούσε να παράξει ακτίνες γ σε MeV, ήταν σπάνιο και ακριβό.

Η εφεύρεση του πυρηνικού αντιδραστήρα επέτρεψε την παραγωγή τεχνητών ραδιοϊσοτόπων για ακτινοθεραπεία. Τότε, άρχισε να κερδίζει έδαφος η θεραπεία με τη χρήση του κοβαλτίου-60 το οποίο κατά τη διάσπασή του εκπέμπει δύο ακτίνες γ, στα 1,173 και 1,33 MeV. Άλλα προσόντα για την εποχή, ήταν ο χρόνος ημιζωής που είναι στα 5,27 χρόνια, η απλότητα στη χρήση της μηχανής και η προσιτή τιμή της.

Το κέσιο-137 παράγεται από πυρηνική σχάση και υπάρχει στο περιβάλλον από δοκιμές πυρηνικών όπλων κι ατυχήματα πυρηνικών αντιδραστήρων που συνέβησαν από το 1950 και μετά με αποκορύφωμα το ατύχημα στο Τσερνόμπιλ το 1986 και χρησιμοποιείται εκτός των άλλων και σε ιατρικές εφαρμογές [9]. Αντικατέστησε το ράδιο που χρησιμοποιούταν στη βραχυθεραπεία, χάρη στο σταθερό θυγατρικό του πυρήνα. Κατά τη διάσπασή του παράγει ακτίνα γ με ενέργεια 0,662 MeV. Προτέρημα του κεσίου είναι ο μεγάλος χρόνος ημιζωής που έχει υπολογιστεί στα 30,17 χρόνια πράγμα που σημαίνει ότι δε χρειάζεται συχνή αντικατάσταση.

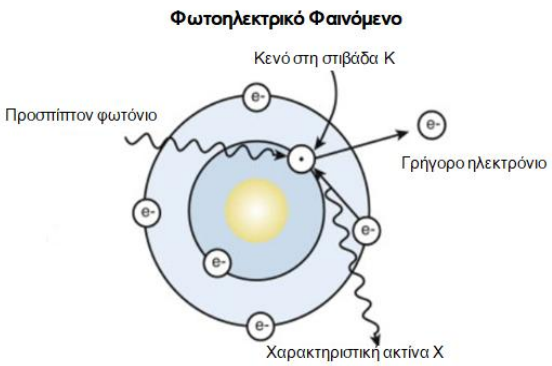
Σήμερα, ο ρόλος του κοβαλτίου-60 έχει αντικατασταθεί μερικώς από τους γραμμικούς επιταχυντές, οι οποίοι μπορούν να δημιουργήσουν ακτινοβολία υψηλής ενέργειας. Σημαντικό προσόν αποτελεί και το γεγονός ότι οι χρήστες δεν έχουν να διαχειριστούν ραδιενεργά απόβλητα και την απόρριψή τους. Επίσης, οι δέσμες που παράγονται είναι συγκεντρωτικές και έχουν μεγαλύτερη ακρίβεια. Ωστόσο, ακόμα πρόκειται για ακριβές διατάξεις, οπότε χρησιμοποιούνται από σχετικά περιορισμένο κοινό για την ώρα.

1.1.3 Αλληλεπίδραση με την ύλη

Όταν μία δέσμη φωτονίων εισέρχεται στην ύλη, συγκεκριμένα εδώ σε ένα κύτταρο, εξετάζεται η αλληλεπίδραση με αυτό και τα αποτελέσματά της. Κατά την ακτινοβόληση, θα πραγματοποιηθούν αλληλεπιδράσεις με τα άτομα των μορίων του υλικού. Η ηλεκτρομαγνητική ακτινοβολία έχει κατά κύριο λόγο *έμμεση δράση (indirect effect)*, διότι δεν προκαλεί η ίδια χημική ή βιολογική καταστροφή. Απορροφάται από το υλικό, επάγονται κάποιες αλληλεπιδράσεις με δευτερογενή παραγωγή σωματιδίων και τα προϊόντα αυτών των αλληλεπιδράσεων μπορούν να προκαλέσουν καταστροφή. Μελετάται κυρίως η επίδραση στα άτομα του μορίου του DNA ενός κυττάρου, καθώς η δική του αλλοίωση μπορεί να προκαλέσει ή όχι κυτταρικό θάνατο. Μία αλληλεπίδραση χαρακτηρίζεται από στοχαστικότητα, καθώς ένα φωτόνιο μπορεί να αλληλεπιδράσει ή όχι με το υλικό κι αυτό δεν μπορεί να προβλεφθεί. Η διαδικασία με την οποία απορροφούνται τα φωτόνια εξαρτάται από την ενέργειά τους και τη χημική σύσταση του απορροφούντος υλικού. Παρακάτω επεξηγούνται οι διάφορες αλληλεπιδράσεις που μπορεί να πραγματοποιηθούν μεταξύ φωτονίων και ύλης [10].

1.1.3.1 Φωτοηλεκτρικό Φαινόμενο

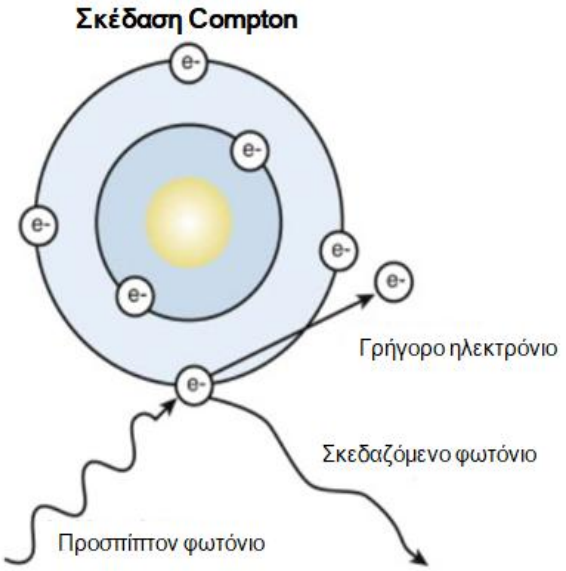
Το *φωτοηλεκτρικό φαινόμενο (photoelectric process)* υφίσταται κυρίως στις χαμηλές ενέργειες μέχρι τα 50 keV. Στη δεδομένη περίπτωση παρατηρείται αλληλεπίδραση μεταξύ ενός φωτονίου κι ενός εσωτερικού ηλεκτρονίου. Το φωτόνιο εναποθέτει όλη του την ενέργεια στο ηλεκτρόνιο. Ένα ποσοστό χρησιμοποιείται για να ξεπεραστεί η ενέργεια σύνδεσης του ηλεκτρονίου και να μπορεί να αποσπαστεί από το τροχιακό και το υπόλοιπο δίνεται ως κινητική ενέργεια. Το κενό που αφήνεται με τη διαφυγή του ηλεκτρονίου, πρέπει να κλείσει, με αποτέλεσμα ή να πέσει ένα ηλεκτρόνιο από εξωτερική στιβάδα ή να καλυφθεί από άλλο ελεύθερο ηλεκτρόνιο. Αυτό προκαλεί αλλαγή στην ενεργειακή κατάσταση, η οποία εξισορροπείται με εκπομπή φωτονίου χαρακτηριστικής ηλεκτρομαγνητικής ακτινοβολίας.



Εικόνα 1.1: Σχηματική απεικόνιση Φωτοηλεκτρικού Φαινομένου [2]

1.1.3.2 *Σκέδαση Compton*

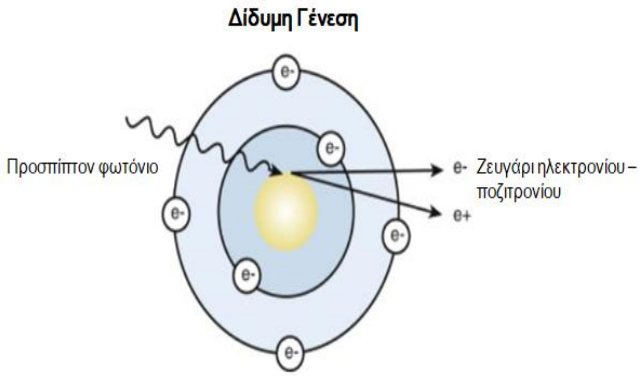
Για υψηλές ενέργειες, αντίστοιχες του κοβαλτίου-60 και του γραμμικού επιταχυντή δηλαδή από 0,1 έως 10 MeV, κυριαρχεί η *σκέδαση Compton (Compton scattering)*. Σε αυτό το φαινόμενο, το φωτόνιο αλληλεπιδρά με ένα εξωτερικό ηλεκτρόνιο, του οποίου η ενέργεια σύνδεσης είναι μικρή σε σχέση με την ενέργεια του φωτονίου. Τότε, μέρος της ενέργειας του φωτονίου θα δοθεί στο ηλεκτρόνιο ως κινητική ενέργεια (KE), ενώ το υπόλοιπο μένει στο φωτόνιο το οποίο σκεδάζεται και συνεχίζει σε άλλη πορεία. Επομένως, πλέον υπάρχει ένα ελεύθερο, γρήγορο ηλεκτρόνιο κι ένα φωτόνιο με μειωμένη ενέργεια το οποίο μπορεί να κάνει περαιτέρω αλληλεπιδράσεις μέχρις ότου αλληλεπιδράσει μέσω φωτοηλεκτρικού φαινομένου ή εξέλθει από το υλικό.



Εικόνα 1.2: Σχηματική απεικόνιση Σκέδασης Compton [2]

1.1.3.3 *Δίδυμη Γένεση*

Στην περίπτωση όπου υπάρξει αλληλεπίδραση με πυρήνα ατόμου του υλικού τότε πραγματοποιείται η *δίδυμη γένεση (pair production)*. Αυτό το φαινόμενο παρατηρείται αποκλειστικά σε μεγάλες ενέργειες άνω του 1,022 MeV. Καθώς το φωτόνιο περνά κοντά από έναν πυρήνα, εξαφανίζεται και στο σημείο αυτό δημιουργούνται ένα ηλεκτρόνιο κι ένα ποζιτρόνιο. Η ενέργεια ηρεμίας των δύο σωματιδίων βρίσκεται στα 0,511 MeV έκαστη, εξ' ου κι η αρχική απαίτηση για την ενέργεια του φωτονίου. Τα δύο σωματίδια που διαγράφουν δικές τους τροχιές, συνεχίζουν τις αλληλεπιδράσεις ως δευτερογενής πηγή. Το μεν ηλεκτρόνιο, καθώς χάνει



Εικόνα 1.3: Σχηματική απεικόνιση Δίδυμης Γένεσης [2]

ενέργεια είναι ευάλωτο στις έλξεις των πυρήνων των ιοντισμένων ατόμων που έχουν προκύψει από τις αλληλεπιδράσεις φωτοηλεκτρικού φαινομένου και σκέδασης Compton. Έτσι, συλλαμβάνεται από κάποιο ιόν, το οποίο καθίσταται πλέον ουδέτερο. Το δε ποζιτρόνιο, που έχει μικρό χρόνο ζωής μπορεί να συναντηθεί με άλλο ελεύθερο ηλεκτρόνιο με αποτέλεσμα την εξαΰλωση των δύο σωματιδίων και τη δημιουργία δύο αντιδιαμετρικών ακτινών γ με 0,511 MeV ενέργεια η καθεμία. Την εξαΰλωση εκμεταλλεύεται η απεικονιστική τεχνική ποζιτρονικής εκπομπής (Positron Emission Tomography, PET).

1.2 Επίδραση Ακτινοβολίας στη βιολογική ύλη

Το φυσικό στάδιο δράσης της ακτινοβολίας που περιγράφηκε παραπάνω και σχετίζεται με την απορρόφηση της ενέργειας, τις διεγέρσεις και τους ιονισμούς διαρκεί από 10^{-18} έως 10^{-15} δευτερόλεπτα. Το ποσοστό της ενέργειας που έχει από απορροφηθεί μέσω ιονισμών μπορεί να προκαλέσει ποικιλία χημικών μεταβολών. Χαρακτηριστικό μέγεθος που σχετίζεται με την κατανομή της ακτινοβολίας στο χώρο είναι η *γραμμική μεταφορά ενέργειας (linear energy transfer, LET =* $^{dE}/_{dx}$*)*. Έτσι, μπορεί να γίνει διάκριση του είδους ή της ποιότητας της ακτινοβολίας, αφού αυτό το μέγεθος περιγράφει την πυκνότητα ιοντισμών ανά μονάδα διαδρομής σωματιδίου σε ένα υλικό. Η ηλεκτρομαγνητική ακτινοβολία έχει χαμηλό LET, ενώ η σωματιδιακή φέρει μεγάλο LET.

Το αποτέλεσμα της επίδρασης της ιοντίζουσας ακτινοβολίας εξαρτάται από την απορροφούμενη δόση και το ρυθμό δόσης. Η απορροφούμενη δόση αντιστοιχεί στη μέση τιμή της ενέργειας που μεταφέρεται στο βιολογικό υλικό. Μονάδα μέτρησης της δόσης στο SI είναι το 1 Gy = 1 $^{\mathrm{J}}/_{\mathrm{kg}}$. Επόμενο είναι το χημικό στάδιο, στο οποίο λαμβάνει χώρα η αλληλεπίδραση των ηλεκτρονίων που έχουν που έχουν απελευθερωθεί, με μόρια του κυττάρου και το νερό. Το στάδιο αυτό έχει διάρκεια από 10^{-15} έως 10^{-7} περίπου δευτερόλεπτα. Εξετάζονται κυρίως οι χημικές και βιολογικές επιδράσεις στο DNA, καθώς αυτό είναι το σημαντικότερο μόριο του κυττάρου. Οι αλλοιώσεις σε αυτό μπορούν να προκαλέσουν καθολική αλλαγή στη λειτουργία και τη μορφή του κυττάρου.

1.2.1 Άμεση/Έμμεση δράση

Όταν κάποια μορφή ακτινοβολίας απορροφάται από ένα βιολογικό υλικό, υπάρχει η πιθανότητα να αλληλεπιδράσει άμεσα με το στόχο (DNA). Τα άτομα του στόχου ιοντίζονται κι έτσι αρχικοποιείται μία σειρά γεγονότων που οδηγούν σε κάποια βιολογική αλλαγή. Αυτή η κατάσταση καλείται *άμεση δράση (direct action)* κι είναι η διαδικασία που συμβαίνει συνήθως στις ακτινοβολίες με υψηλό LET [11]. Η άμεση δράση προκαλεί θραύση χημικών δεσμών. Κάτι τέτοιο έχει ως συνέπεια την απώλεια ενός ατόμου υδρογόνου ή ενός μεθυλίου κι άρα την παραγωγή ελεύθερων ριζών ή τη διάσπαση του μορίου σε μικρότερα. Με τον όρο ελεύθερη ρίζα εννοούμε ένα μόριο ή άτομο που ενώ είναι ηλεκτρικά ουδέτερο, φέρει ένα ασύζευκτο ηλεκτρόνιο στην εξωτερική στοιβάδα. Το γεγονός αυτό καθιστά την ουσία εξαιρετικά δραστική.

Ωστόσο, στις ηλεκτρομαγνητικές ακτινοβολίες που η ακτινοβολία πέφτει διάχυτη στο κύτταρο, συνήθως πραγματοποιείται αλληλεπίδραση με άλλα μόρια και το νερό. Έτσι, υφίσταται η παραγωγή ελευθέρων ριζών που σχηματίζονται κατά τη ραδιόλυση του νερού, οι οποίες είναι σε θέση να ταξιδέψουν έως και λίγα νανόμετρα βλάπτοντας γειτονικά μόρια. Αυτή είναι η *έμμεση δράση (indirect action)* της ακτινοβολίας. Χαρακτηριστικά, παρατίθεται η αλληλεπίδραση της ακτινοβολίας με ένα μόριο νερού.

Πρωτογενώς, παρατηρούνται οι αντιδράσεις που περιγράφουν τη διάσπαση ενός μορίου νερού προς το σχηματισμό των ελεύθερων ριζών:

$$H_2O \rightarrow H_2O^+ + e^-_{aq} \rightarrow H^+ + OH^\cdot + e^-_{aq}$$
$$H_2O \rightarrow H_2O^* \rightarrow H^\cdot + OH^\cdot$$

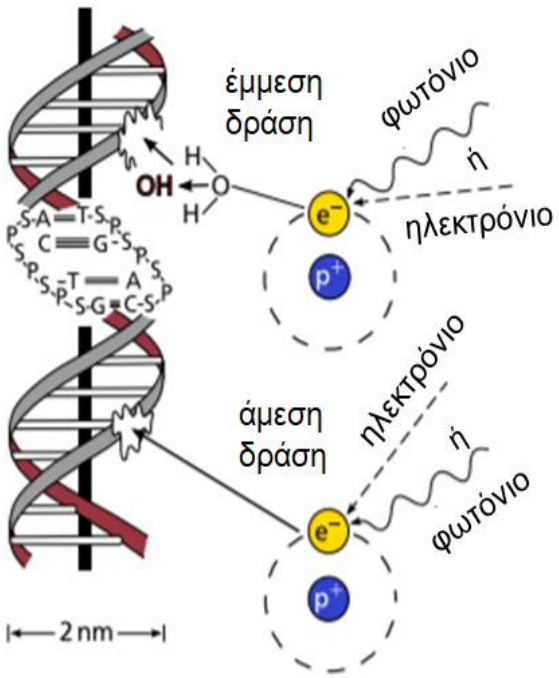
Δευτερογενώς, οι ελεύθερες ρίζες υδρογόνου κι υδροξυλίου, τα κατιόντα υδρογόνου και τα ενυδατωμένα ηλεκτρόνια διαχέονται και συμμετέχουν σε αντιδράσεις μεταξύ τους ή με άλλα μόρια του κυττάρου. Ενδεικτικά αναφέρονται κάποιες όπως:

- Η «αδρανοποίηση» ελευθέρων ριζών κατά την αντίδραση ριζών υδρογόνου μεταξύ τους ή την αντίδραση ρίζας υδρογόνου και ρίζας υδροξυλίου ή την αντίδραση κατιόντος υδρογόνου και ρίζας υδροξυλίου:

$$H^\cdot + H^\cdot \rightarrow H_2$$
$$H^\cdot + OH^\cdot \rightarrow H_2O$$
$$H^+ + OH^\cdot \rightarrow H_2O$$

- Η αντίδραση δύο ριζών υδροξυλίου κατά την οποία παράγεται υπεροξείδιο του υδρογόνου:

$$OH^\cdot + OH^\cdot \rightarrow H_2O_2$$



Εικόνα 1.4: Δράση ακτινοβολίας στο DNA [1]

το οποίο με τη σειρά του μπορεί να αντιδράσει με ρίζα υδροξυλίου και να οδηγήσει στο σχηματισμό μιας νέας δραστικής ελεύθερης ρίζας υπεροξυλίου:

$$H_2O_2 + OH^\cdot \rightarrow HO_2$$

- Ο σχηματισμός ρίζας υπεροξυλίου από την αντίδραση ρίζας υδρογόνου με μοριακό οξυγόνο:

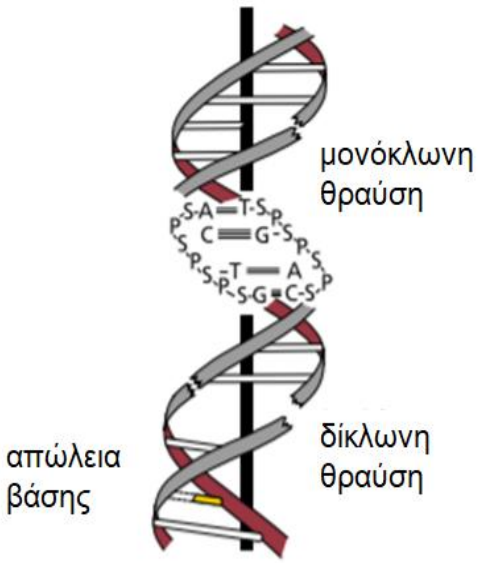
$$H^\cdot + O_2 \rightarrow HO_2$$

Εκτιμάται ότι περίπου τα δύο τρίτα της καταστροφής που προκαλείται στο DNA των κυττάρων των θηλαστικών από ακτίνες X προκύπτει λόγω της

ρίζας υδροξυλίου. Χάρη στην υπερίσχυση του οξειδωτικού χαρακτήρα, το βασικό αποτέλεσμα της ραδιόλυσης του νερού από την έκθεση σε ιοντίζουσα ακτινοβολία είναι η οξείδωση των μικρομορίων και μακρομορίων του κυττάρου.

1.2.2 Είδη βλαβών

Ως αποτέλεσμα της δράσης των μηχανισμών που αναφέρονται παραπάνω, προκαλούνται διάφορα είδη βλαβών στο DNA. Χαρακτηριστικές βλάβες είναι η αποσύνθεση βάσης κι η *μονόκλωνη (single strand break, SSB)* ή *δίκλωνη* θραύση *(double strand break, DSB)*. Ακόμα και χωρίς εξωγενείς παράγοντες όπως η ακτινοβολία, το κύτταρο αντιμετωπίζει καθημερινά πληθώρα βλαβών τις οποίες καλείται να επιδιορθώσει. Για το λόγο αυτό, διαθέτει ένα δίκτυο κυτταρικών διεργασιών που ανιχνεύει τις βλάβες, τις επισημαίνει ανάλογα με τη βαρύτητά τους κι ενεργοποιεί τους επιδιορθωτικούς μηχανισμούς. Έτσι, συμπληρώνονται βάσεις όπου απουσιάζουν, επιδιορθώνονται εσφαλμένως τοποθετημένες βάσεις κι ενώνονται άκρα. Με το πέρας του χημικού σταδίου, ενεργοποιούνται κατευθείαν οι επιδιορθωτικοί μηχανισμοί και γίνεται εκκίνηση του βιολογικού σταδίου, το οποίο μπορεί να έχει διάρκεια από 10^{-6} δευτερόλεπτα έως και μήνες ανάλογα τη σοβαρότητα της βλάβης και το αν θα κληρονομηθεί στα θυγατρικά κύτταρα κατά τη διαίρεση, όπου και θα επιφέρει σοβαρότερες συνέπειες στον οργανισμό. Έχει εκτιμηθεί πειραματικά, ότι οι βλάβες βάσεων κι οι SSBs επιδιορθώνονται στο μεγαλύτερο ποσοστό χωρίς να αφήσουν μεταλλάξεις στο γενετικό κώδικα του κυττάρου [12]. Οι DSBs καθώς κι οι *σύνθετες ομαδοποιημένες βλάβες (clustered DNA lesions)*, που αποτελούνται από παραπάνω από μία βλάβες σε απόσταση το πολύ 10 ζευγών βάσεων, είναι πιο απαιτητικές στην επιδιόρθωση. Επομένως, τα βιολογικά αποτελέσματα που παρατηρούνται οφείλονται σε κάποιο ποσοστό μη επιδιορθωμένων ή ανεπιτυχώς επιδιορθωμένων βλαβών DNA.



Εικόνα 1.5: Τύποι βλαβών στο DNA [1]

1.2.3 Κυτταρικός θάνατος και μελέτη μέσω clonogenic assay

Αν το βιολογικό αποτέλεσμα που επιφέρουν οι βλάβες είναι αδύνατο να επιδιορθωθεί, τότε επέρχεται ο θάνατος του κυττάρου. Ως θάνατος, ορίζεται η αδυναμία διαίρεσης και παραγωγής επ' αόριστον θυγατρικών κυττάρων. Συγκεκριμένα για τα καρκινικά κύτταρα που μελετώνται εδώ, ο όγκος είναι δυνατόν να εξαλειφθεί αν τα κύτταρα αποδειχθούν ανίκανα να διαιρεθούν και να επιτρέψουν περαιτέρω ανάπτυξη της κακοήθειας [12]. Η συχνότητα εμφάνισης μεταλλάξεων αυξάνει με τη δόση της ακτινοβολίας, με αποτέλεσμα περισσότερα κύτταρα να οδηγούνται στον κυτταρικό θάνατο. Στα περισσότερα κύτταρα πραγματοποιείται είτε *μιτωτικός θάνατος (mitotic death)* είτε *απόπτωση (apoptosis)* δηλαδή προγραμματισμένος κυτταρικός θάνατος.

Η ικανότητα ενός κυττάρου να διαιρείται και να δημιουργεί θυγατρικά κύτταρα μπορεί να οδηγήσει στη δημιουργία μιας αποικίας. Σε ένα σύνολο κυττάρων, αυτό μπορεί να γίνει αντιληπτό και να χαρακτηριστεί μέσω της μεθόδου *clonogenic assay.* Συγκεκριμένα, πραγματοποιείται καλλιέργεια κυττάρων με τη χρήση κάποιων ουσιών για δεδομένο χρονικό διάστημα, ενώ είτε πριν είτε μετά πραγματοποιείται η ακτινοβόληση σε δεδομένες συνθήκες. Έχοντας ως δεδομένα τον αριθμό των μεμονωμένων κυττάρων (cells seeded) που έχουν τοποθετηθεί στο πιάτο καλλιέργειας και τον αριθμό των αποικιών που καταμετρώνται (colonies counted), υπολογίζεται η απόδοση δημιουργίας αποικιών (plating efficiency). Εν συνεχεία, έχοντας υπολογίσει την απόδοση, μπορεί να υπολογιστεί το *κλάσμα επιβίωσης (survival fraction, SF)* για τα νέα δεδομένα της ακτινοβόλησης μέσω της σχέσης

$$Survival\ Fraction = \frac{Colonies\ counted}{Cells\ seeded * (Plating\ Efficiency/100)}$$

Αυτή η διαδικασία επαναλαμβάνεται ώστε να ληφθούν τα κλάσματα επιβίωσης για κάποιο εύρος δόσεων. Ο αριθμός των κυττάρων που καλλιεργούνται σε κάθε πιάτο προσαρμόζεται ανάλογα τη δόση, ώστε να προκύπτει ένας μετρήσιμος αριθμός αποικιών αλλά να μη συγχωνεύονται κιόλας μεταξύ τους. Έτσι, σχεδιάζεται η *καμπύλη επιβίωσης (survival curve)*, με τη δόση στον άξονα x σε γραμμική κλίμακα και το κλάσμα επιβίωσης στον άξονα y σε λογαριθμική κλίμακα. Για τις ακτινοβολίες χαμηλού LET όπως οι ακτίνες X και γ, η καμπύλη ξεκινά κατευθείαν με κάποια κλίση και ακολουθεί μία εκθετικά φθίνουσα πορεία συναρτήσει της δόσης. Σε υψηλές δόσεις (≥ 8 Gy) τείνει να γίνει ευθεία.

1.2.4 Γραμμικό – Τετραγωνικό μοντέλο (Linear – Quadratic, LQ Model)

Το μοντέλο αυτό είναι το πιο σύνηθες στην περιγραφή των παραπάνω καμπυλών επιβίωσης. Βασίζεται στην ιδέα ότι πολλαπλές βλάβες αλληλεπιδρούν για να προκληθεί ο θάνατος του κυττάρου. Οι βλάβες που αλληλεπιδρούν μπορεί να έχουν προκληθεί είτε από ένα γεγονός ακτινοβολίας δίνοντας μία άμεση εξάρτηση της θνησιμότητας από τη δόση. Μπορεί όμως να έχουν προκληθεί από διαφορετικά γεγονότα δίνοντας εξάρτηση από υψηλότερες δυνάμεις δόσης. Το μοντέλο λαμβάνει υπόψη δύο συνιστώσες στον κυτταρικό θάνατο από ακτινοβολία. Η μία είναι ανάλογη της δόσης ακτινοβόλησης κι άλλη ανάλογη του τετραγώνου της δόσης. Η υπόθεση ότι δύο βλάβες πρέπει να αλληλεπιδράσουν για να προκληθεί κυτταρικός θάνατος δίνει μία εξίσωση η οποία προσεγγίζει τις περισσότερες καμπύλες επιβίωσης. Οπότε η έκφραση που περιγράφει την καμπύλη επιβίωσης μέσω αυτού του μοντέλου όπως φαίνεται και στην εικόνα είναι

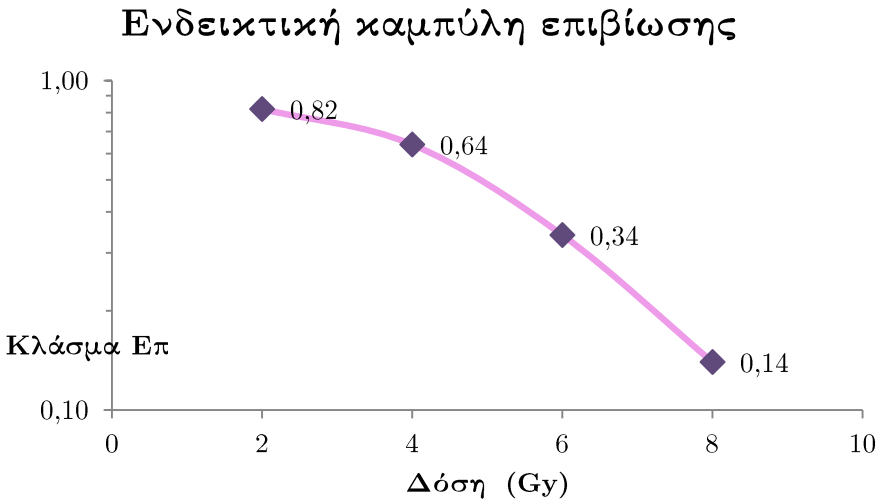
$$S = e^{-\alpha D - \beta D^2}$$

όπου S είναι το κλάσμα επιβίωσης στη δεδομένη δόση D, ενώ οι ποσότητες α, β είναι σταθερές. Ο συντελεστής α σχετίζεται με την πιθανότητα πρόκλησης βλάβης από την τροχιά ενός σωματιδίου κι είναι ανεξάρτητος του ρυθμού δόσης, ενώ ο συντελεστής β εκφράζει την πιθανότητα πρόκλησης βλάβης από την τροχιά δύο σωματιδίων οπότε κι εξαρτάται από το ρυθμό δόσης [11]. Σημαντικό μέγεθος στην κλινική διαδικασία είναι η δόση για την οποία οι δύο συνιστώσες που αναφέρθηκαν είναι ίσες δηλαδή

$$\alpha D = \beta D^2$$

ή

$$D = {}^{\alpha}/_{\beta}.$$



Εικόνα 1.6: Καμπύλη επιβίωσης από το σύνολο δεδομένων

1.3 Στοιχεία Μηχανικής Μάθησης

Η πρώτη αναφορά στον όρο Μηχανική Μάθηση (Machine Learning) γίνεται το 1959 από τον Arthur Samuel, ο οποίος περιγράφει αυτόν τον κλάδο ως ένα πεδίο μελέτης της επιστήμης των υπολογιστών, που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί. Άλλος ένας χρήσιμος ορισμός έρχεται από τον Tom Mitchell το 1997, "Ενα πρόγραμμα λέγεται πως μαθαίνει από την εμπειρία E σε σχέση με μία σειρά εργασιών T και ένα μέτρο απόδοσης P, αν η απόδοσή του σε σειρά εργασιών T, όπως μετράται με P, βελτιώνεται με την εμπειρία E" [13].

Επομένως, οι αλγόριθμοι μηχανικής μάθησης, μαθαίνουν επαναληπτικά από τα δεδομένα και επιτρέπουν στα υπολογιστικά συστήματα να βρουν μοτίβα και να κάνουν προβλέψεις για νέα δεδομένα με τα οποία δεν έχουν εκπαιδευτεί.

1.3.1 Είδη

Οι αλγόριθμοι μπορούν να κατηγοριοποιηθούν με διάφορους τρόπους. Οι τρεις πιο χρησιμοποιούμενες προσεγγίσεις είναι:

- *Επιτηρούμενη Μάθηση (Supervised Learning):*
  Η διαδικασία με την οποία ο αλγόριθμος εκπαιδεύεται έχοντας τόσο τα δεδομένα εισόδου όσο και τα δεδομένα εξόδου. Στόχος είναι η μάθηση ενός κανόνα που να έχει γενικευμένη εφαρμογή και να αντιστοιχεί τα δεδομένα εισόδου με εκείνα της εξόδου.

- *Μη Επιτηρούμενη Μάθηση (Unsupervised Learning):*
  Η διαδικασία με την οποία ο αλγόριθμος εκπαιδεύεται έχοντας μόνο δεδομένα εισόδου. Στόχος μπορεί να είναι είτε η αναγνώριση μοτίβων ως αυτοσκοπός για ομαδοποίηση είτε η ανίχνευση νέων χαρακτηριστικών – συσχετισμών μεταξύ των δεδομένων.

- *Ενισχυτική Μάθηση (Reinforcement Learning):*
  Η διαδικασία με την οποία ο αλγόριθμος αλληλεπιδρά με ένα δυναμικό περιβάλλον, στο οποίο πρέπει να επιτύχει έναν στόχο. Καθώς προχωράει προς το στόχο, λαμβάνει κριτική ανάλογη της ανταμοιβής, την οποία προσπαθεί να μεγιστοποιήσει.

Στην παρούσα εργασία γίνεται εφαρμογή Επιτηρούμενης Μάθησης. Η προσέγγιση αυτή χρησιμοποιεί τα δεδομένα (dataset) που παρέχει ο χρήστης και χτίζει με αυτά ένα μαθηματικό μοντέλο, που θα μπορεί να προβλέψει δεδομένα εξόδου προερχόμενα από νέα δεδομένα εισόδου. Το σύνολο των δεδομένων αποτελείται από παραδείγματα

(training examples), τα οποία έχουν ένα ή περισσότερα χαρακτηριστικά (features) και τη μεταβλητή εξόδου (target/label). Συνήθως και παρακάτω ακολουθείται αυτός ο συμβολισμός, σημειώνουμε τις μεταβλητές εισόδου που είναι ανεξάρτητες ως ένα διάνυσμα $\mathbf{X}$ και την εξαρτημένη μεταβλητή εξόδου ως y. Οι μεταβλητές – δεδομένα εξόδου μπορεί να ανήκουν σε ένα συνεχές φάσμα τιμών ή να λαμβάνουν συγκεκριμένες και διακριτές τιμές. Επομένως, η Επιτηρούμενη Μάθηση περιλαμβάνει δύο είδη αλγορίθμων:

- *Παλινδρόμηση (Regression):*
  Σ' ένα μοντέλο παλινδρόμησης, οι ανεξάρτητες μεταβλητές μπορεί να είναι οποιασδήποτε μορφής είτε κατηγορικές είτε αριθμητικές. Η εξαρτημένη μεταβλητή ωστόσο πρέπει να ανήκει σε φάσμα συνεχών τιμών. Παράδειγμα τέτοιας εφαρμογής είναι η πρόβλεψη της τιμής εκπομπής διοξειδίου του άνθρακα ενός αυτοκινήτου με βάση χαρακτηριστικά όπως το μέγεθος της μηχανής, ο αριθμός των κυλίνδρων και η κατανάλωση καυσίμου.

- *Ταξινόμηση (Classification):*
  Σ' ένα μοντέλο ταξινόμησης, οι ανεξάρτητες μεταβλητές πάλι μπορούν να έχουν οποιαδήποτε μορφή. Η εξαρτημένη όμως πρέπει να είναι διακριτή. Το σύνολο των πιθανών τιμών αποτελούν τις κλάσεις ή κατηγορίες. Έτσι, όταν γίνονται προβλέψεις, οι νέες εξαρτημένες τιμές θα ταξινομούνται στις κλάσεις. Παράδειγμα τέτοιας εφαρμογής είναι η πιθανότητα εξόφλησης ή όχι ενός δανείου. Με βάση χαρακτηριστικά όπως η ηλικία, το εισόδημα και η εκπαίδευση μία τράπεζα μπορεί να δει αν ένας πιθανός πελάτης θα είναι σε θέση να αποπληρώσει ένα δάνειο.

2 Μεθοδολογία

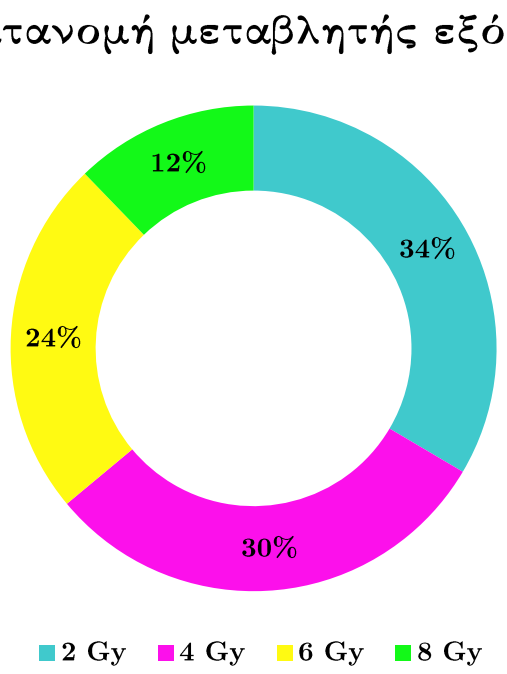
Στην παρούσα ενότητα, γίνεται αναφορά στα δεδομένα που μελετώνται και στον τρόπο που αυτά συλλέχθηκαν, επεξεργάστηκαν κι έφτασαν στην τελική μορφή για να μπορούν να αξιοποιηθούν από τους αλγορίθμους που έχουν επιλεχθεί. Περνώντας στο υπολογιστικό πλέον τμήμα, γίνεται περιγραφή των υπολογιστικών στοιχείων, αλγορίθμων και βοηθητικών εργαλείων, που χρησιμοποιούνται για την ανάπτυξη των μοντέλων. Περιγράφονται μαθηματικά οι αλγόριθμοι κι οι μετρικές αξιολόγησης που έχουν επιλεχθεί και συζητούνται μερικά υπολογιστικά ζητήματα όπως η υπερεκπαίδευση ενός μοντέλου κι η βελτιστοποίησή του.

2.1 Περιγραφή δεδομένων και προβλήματος

Μέσα στα χρόνια έχουν γίνει αρκετές μελέτες με *in vitro* πειράματα σε καρκινικά κύτταρα στα οποία η πηγή της ακτινοβολίας είναι η ηλεκτρομαγνητική ιοντίζουσα ακτινοβολία. Σκοπός των μελετών είναι ο χαρακτηρισμός της ραδιοευαισθησίας κυττάρων διαφόρων καρκινικών σειρών είτε ανθρώπινων είτε από ζώα. Αυτό μπορεί να γίνει μελετώντας την επίδραση της ακτινοβολίας σε αποικίες κυττάρων μέσω του clonogenic assay που περιγράφηκε παραπάνω. Επιπλέον, μπορεί να συνδυαστεί με την προσθήκη κάποιου φαρμάκου ή με τη σύλληψη του κυτταρικού κύκλου σε κάποια συγκεκριμένη φάση στην οποία τα κύτταρα είναι πιο ακτινοευαίσθητα. Τέλος, συνήθης είναι η σίγαση κάποιου γονιδίου που σχετίζεται με την ακτινοαντοχή ή την επιδιόρθωση των κυττάρων με αποτέλεσμα να καθίστανται εντέλει πιο ευαίσθητα.

Η ακτινοθεραπεία είναι η πλέον διαδεδομένη θεραπεία του καρκίνου. Όμως, ακόμα οι έρευνες είναι μεμονωμένες και δεν υπάρχουν ισχυροί προγνωστικοί βιοδείκτες ραδιοευαισθησίας ώστε να μπορεί σταδιακά να εξατομικευτεί η ακτινοθεραπεία. Παρόλο που υπάρχει μεγάλος αριθμός δημοσιεύσεων που περιγράφει τη χρήση του clonogenic assay για τη μέτρηση της ραδιοευαισθησίας καρκινικών κυττάρων, η δυναμική των αποτελεσμάτων μεταξύ μελετών είναι ασαφής. Σε αυτό εστίασαν οι συγγραφείς του άρθρου [14] στο οποίο βασίζεται η διπλωματική εργασία. Με μία εκτενή έρευνα, ξεκινώντας από 256 καρκινικές κυτταρικές σειρές και τη χρήση κριτηρίων που επεξηγούνται παρακάτω, συγκέντρωσαν 566 δημοσιεύσεις στις οποίες είχε πραγματοποιηθεί clonogenic assay σε 8 ανθρώπινες καρκινικές σειρές, είχαν σχεδιαστεί οι καμπύλες επιβίωσης κι είχαν ληφθεί τα *τελικά σημεία (endpoints)* των κλασμάτων επιβίωσης για δόσεις 2, 4, 6, και 8 Gy όπου αυτά ήταν διαθέσιμα. Στην εικόνα 2.1 φαίνεται η κατανομή των τιμών των κλασμάτων επιβίωσης που θα χρησιμοποιηθούν ως μεταβλητές εξόδου, στο πλαίσιο της διπλωματικής εργασίας. Εν συνεχεία, για τις

μελέτες όπου δίνονταν τα SF_2, SF_4 και SF_6 έγινε υπολογισμός των D_{10}, D_{50} και Đ μετά την προσαρμογή (fitting) του κλάσματος επιβίωσης στο LQ Model.



Εικόνα 2.1: Κατανομή τιμών που χρησιμοποιούνται ως μεταβλητή εξόδου

Ύστερα, στο ίδιο άρθρο πραγματοποιήθηκε μία στατιστική ανάλυση για να αναδειχθεί κάποιο στατιστικό στοιχείο που θα ξεχώριζε πιθανώς ένα από τα προαναφερθέντα μεγέθη. Η ακρίβεια μεταξύ των clonogenic assays εκτιμήθηκε υπολογίζοντας τους *συντελεστές διασποράς (coefficient of variation, CV)* των μεγεθών. Λήφθηκαν υπόψη οι συντελεστές με τιμή $\leq 30\%$. Έτσι, τα συμπεράσματά τους ήταν ότι τα μεγέθη SF_2 και D_{10}, των οποίων οι συντελεστές ήταν κάτω του 30%, εκτιμώνται με καλή ακρίβεια μεταξύ των μελετών κι αποτελούν αποδεκτούς βιοδείκτες για την ευαισθησία καρκινικών κυττάρων σε φωτονική ακτινοβόληση υπό διαφορετικές πειραματικές συνθήκες.

Στην εικόνα 2.2 απεικονίζεται η εξέλιξη του αριθμού των δημοσιεύσεων που σχετίζονται με την ιοντίζουσα ακτινοβολία μέσα στα χρόνια. Για την κατασκευή του χρησιμοποιήθηκε η πλατφόρμα αναζήτησης PubMed με τελευταία αναζήτηση τον Ιούλιο 2020. Τα κριτήρια αναζήτησης ήταν οι 8 κυτταρικές που επιλέχθηκαν από τους συγγραφείς του προαναφερθέντος άρθρου, με τον τρόπο που περιγράφεται στην επόμενη ενότητα. Παρατηρείται ότι από τις αρχές του 2000 και με την εξέλιξη της φυσικής τεχνολογίας που επιτρέπει την κατασκευή δυνατότερων μηχανημάτων, ο αριθμός μελετών αυξάνεται εκθετικά.

Εικόνα 2.2: Χρονική εξέλιξη αριθμού δημοσιεύσεων

Πηγή Δεδομένων: PubMed

2.1.1 Συλλογή δεδομένων

Οι μελέτες που χρησιμοποιήθηκαν προέκυψαν από την αναζήτηση συγκεκριμένων όρων. Το μοτίβο που ακολουθήθηκε ήταν

[ονομασία καρκινικής σειράς] AND (X-rays OR gamma rays OR radiation)

Οι 8 κυτταρικές σειρές που εντέλει χρησιμοποιήθηκαν είναι οι

- H1299, A549, H460 (non-small cell lung cancer, NSCLC – μη μικροκυτταρικός καρκίνος πνεύμονα)
- HT-29, SW480, HCT-116 (colorectal cancer – καρκίνος παχέος εντέρου)
- DU145, PC-3 (prostate cancer – καρκίνος του προστάτη)

Αυτή η επιλογή έγινε διότι οι συγκεκριμένες είναι οι πιο συχνά χρησιμοποιούμενες για clonogenic assays. Από τη συγκέντρωση των σχετικών μελετών καταγράφηκαν τα SF_2, SF_4, SF_6 και SF_8, όπου αυτά ήταν διαθέσιμα και δημιουργήθηκε ένα αρχείο excel το οποίο περιλάμβανε την ονομασία της καρκινικής σειράς (Cell Line), το PubMedID, το χρόνο καλλιέργειας των κυττάρων (πριν ή μετά την ακτινοβόληση), τον τύπο της ακτινοβολίας (X-rays ή γ-rays) και το ρυθμό δόσης μετρημένο σε $^{\mathrm{Gy}}/_{\mathrm{min}}$. Επιπλέον,

προστέθηκαν 3 στήλες με τις τιμές των D_{10}, D_{50} και Ð που υπολογίστηκαν εκ των υστέρων.

Εν συνεχεία, έγινε προσπάθεια εμπλουτισμού αυτού του αρχείου για τις ανάγκες της διπλωματικής εργασίας αλλά και για να μπορεί να συσχετιστεί με τη βάση δεδομένων [15] στην οποία έχουν συγκεντρωθεί στοιχεία από πειράματα επιβίωσης κυττάρων ακτινοβοληµένα με ιόντα δηλαδή σωματιδιακή ακτινοβολία. Πραγματοποιήθηκε μελέτη των άρθρων και προστέθηκαν οι στήλες Cell Class, Cell Origin, Cell Cycle και Photon Radiation. Οι δύο πρώτες συμπληρώθηκαν αμέσως με τα γράμματα t (=tumor) και h (=human), αφού ήταν γνωστό ότι μιλάμε για ανθρώπινες καρκινικές σειρές. Οι δύο τελευταίες ήθελαν περισσότερη προσοχή αφού έπρεπε να διαπιστωθεί αν και σε ποιο σημείο του κυτταρικού κύκλου είχαν συγχρονιστεί τα κύτταρα κατά τη διάρκεια του clonogenic assay αλλά και την ακριβή πηγή της ακτινοβολίας. Η στήλη Cell Cycle συμπληρώθηκε με τις τιμές a (=asychronous) και G1 από την αντίστοιχη φάση του κυτταρικού κύκλου. Η στήλη Photon Radiation έλαβε τις τιμές 137Cs και 60Co για ακτίνες γ κι ένα εύρος τιμών από 100 έως 320 kVp κι από 4 έως 16 MV για ακτίνες X. Για να αναχθούν όλες οι ενέργειες στην ίδια μονάδα προστέθηκε η στήλη Photon Rad (MeV), η οποία συμπληρώθηκε με τις τιμές 0,662 και 1,173 MeV για το 137Cs και το 60Co αντίστοιχα και τις ήδη υπάρχουσες των ακτίνων X.

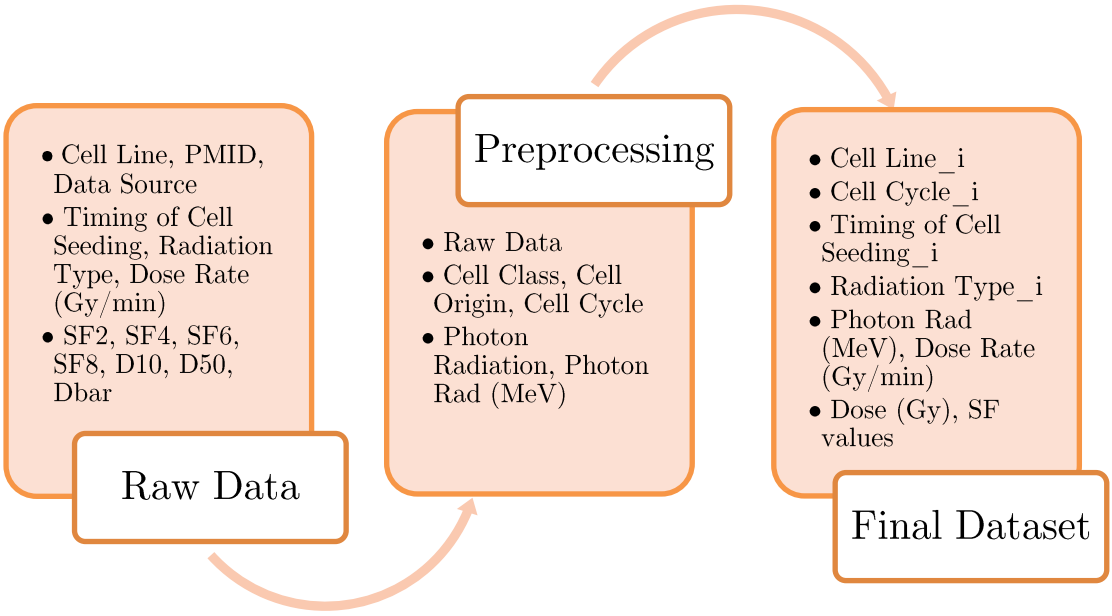
2.1.2 Περιγραφή του προβλήματος και διαχείριση αρχικών δεδομένων

Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη ενός μοντέλου, το οποίο θα μπορεί να προβλέψει την κυτταρική επιβίωση συναρτήσει των παραπάνω παραγόντων. Η υλοποίηση του μοντέλου θα γίνει με τη χρήση αλγορίθμων Μηχανικής Μάθησης. Τα αρχικά δεδομένα αποτελούνταν από 566 σειρές με παρατηρήσεις που δεν επαρκούν για ικανοποιητική εκπαίδευση ενός μοντέλου Μηχανικής Μάθησης. Έτσι, γίνεται μία τροποποίηση στην ιδέα η οποία σχετίζεται με τη δόση.

Το πρώτο βήμα του κώδικα είναι να πάρει τις 4 στήλες με τα Survival Fractions και να τις μετατρέψει σε μία, βάζοντας επομένως όλες τις τιμές επιβίωσης σε μία στήλη με την ονομασία SF values. Κάνοντάς το αυτό, αντιγράφει τις πληροφορίες που συνοδεύουν κάθε τιμή επιβίωσης σε μία καινούρια σειρά. Οπότε τώρα υπάρχουν 566 επί 4 δηλαδή 2264 σειρές δεδομένων. Επειδή όμως αρκετές τιμές επιβίωσης είναι null διότι δεν υπολογίστηκαν στα αντίστοιχα πειράματα, ο κώδικας διαγράφει ολόκληρη τη σειρά. Έτσι, το τελικό σύνολο δεδομένων μένει με 1595 σειρές. Προστίθεται δε μία ακόμα στήλη με το όνομα Dose (Gy), η οποία αντιστοιχεί κάθε κλάσμα επιβίωσης στη δόση 2, 4, 6 και 8 Gy αντίστοιχα. Τέλος, δε λαμβάνονται υπόψη οι στήλες D_{10}, D_{50} και Ð μιας και δεν περιλαμβάνουν πειραματικές πληροφορίες αλλά υπολογισμούς που πραγματοποιήθηκαν μετά από προσαρμογή/ βελτιστοποίηση των δεδομένων και θα προσδώσουν ένα μη ρεαλιστικό χαρακτήρα (bias) σε οποιαδήποτε μετα-ανάλυση.

Αρκετές στήλες περιέχουν μη καθαρά αριθμητικά στοιχεία. Αυτές οι μεταβλητές καλούνται κατηγορικές. Οπότε επόμενο βήμα είναι η τροποποίηση των στηλών αυτών, ώστε οι κατηγορικές μεταβλητές να αναπαριστώνται αριθμητικά με 0 και 1. Ο μετασχηματιστής ελέγχει όλες τις στήλες και τροποποιεί τις Cell Line, Cell Cycle, Timing of Cell Seeding και Radiation Type. Συγκεκριμένα, δημιουργούνται 8 στήλες που αναφέρονται στην εκάστοτε κυτταρική σειρά, 2 για τον κυτταρικό κύκλο, 3 για το χρόνο καλλιέργειας (πριν/μετά/άγνωστο ως προς την ακτινοβόληση) και 3 για το είδος της ακτινοβολίας (X-rays/γ-rays/άγνωστο). Οπότε το σύνολο πλέον περιλαμβάνει 1595 γραμμές και 20 στήλες, 16 από την κωδικοποίηση και τις στήλες PhotonRad (MeV), Dose Rate (Gy/min), Dose (Gy) και SF values. Όλη η διαδικασία απεικονίζεται συνοπτικά στην εικόνα 2.3.

Το σύνολο δεδομένων περιέχει τιμές που λείπουν και κωδικοποιούνται με το συμβολισμό NA ή με απλό κενό. Αυτό το γεγονός είναι ασύμβατο με τους εκτιμητές που χρησιμοποιούνται εδώ για τη μηχανική μάθηση, οι οποίοι προϋποθέτουν ότι όλες οι τιμές σε ένα σύνολο είναι με κάποιο τρόπο υπαρκτές. Οπότε τελικό βήμα προετοιμασίας των δεδομένων είναι η εκτίμηση των τιμών που λείπουν, μέσω ενός «γειτονικού», γνωστού και συσχετιζόμενου μέρος του σετ για κάθε σειρά παρατήρησης.



Εικόνα 2.3: Σχηματική απεικόνιση προ-επεξεργασίας δεδομένων

2.2 Αλγόριθμοι & Εκπαίδευση μοντέλων

Κατά τη δημιουργία ενός μοντέλου υπάρχουν πολλά ζητήματα που πρέπει να ληφθούν υπόψη όπως οι αλγόριθμοι που θα χρησιμοποιηθούν, τα δεδομένα που έχει ο χρήστης, ο τρόπος που θα αξιολογηθούν κι η εγκυρότητα των αποτελεσμάτων. Χάρη στην ανάπτυξη της μηχανικής μάθησης, υπάρχουν σε ένα βαθμό έτοιμα εργαλεία για να επιλέξει ο χρήστης αυτά που ταιριάζουν καλύτερα στις ανάγκες του προβλήματός του.

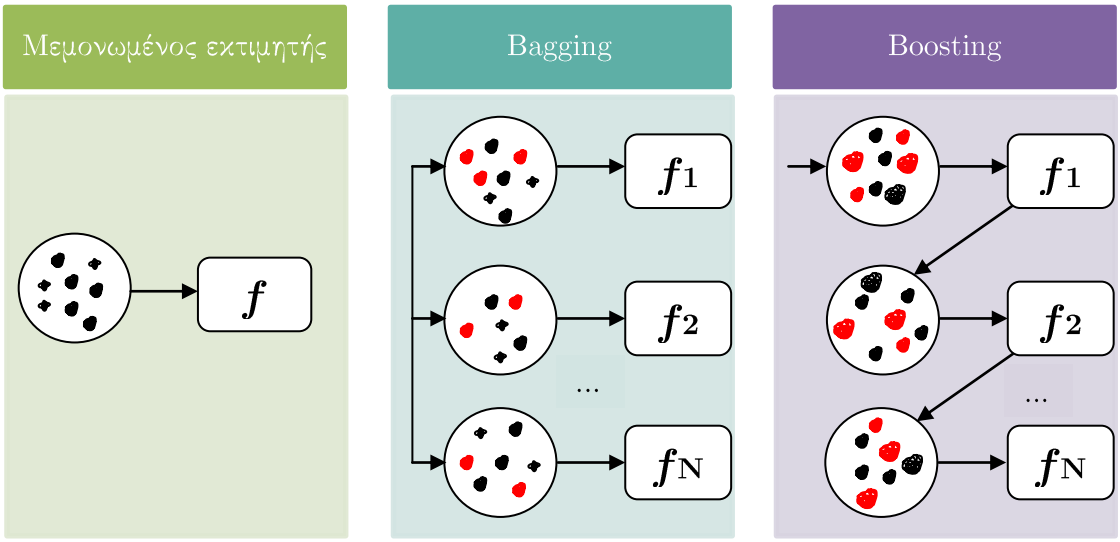
2.2.1 Επιλεγμένοι αλγόριθμοι

Στην παρούσα εργασία χρησιμοποιούνται *μέθοδοι συνόλου (ensemble methods)* [16]. Οι μέθοδοι αυτές βασίζονται στην ιδέα των πολλών υποθέσεων για την εύρεση της ιδανικής που θα βγάζει τα βέλτιστα αποτελέσματα. Δεν είναι απαραίτητο η τελική υπόθεση να περιλαμβάνεται στις αρχικές. Ο όρος χρησιμοποιείται για τις μεθόδους που παράγουν πολλαπλές υποθέσεις χρησιμοποιώντας τον ίδιο εκτιμητή ως βάση. Ο εκτιμητής χαρακτηρίζεται ως ένα αντικείμενο που διαχειρίζεται τον προσδιορισμό και την αποκωδικοποίηση του μοντέλου. Λειτουργεί με τον ορισμό παραμέτρων και μία *συνάρτηση προσαρμογής (fit function)* ώστε να μπορεί να αξιολογήσει τα εισερχόμενα δεδομένα, να εκπαιδευτεί και να χρησιμοποιηθεί σε νέα δεδομένα. Οι ensemble μέθοδοι χωρίζονται σε δύο κατηγορίες [17]:

- *Averaging (Bagging) Methods:*
  Στόχος αυτών των μεθόδων είναι να κατασκευάσουν ανεξάρτητους εκτιμητές και μετά να υπολογίσουν το μέσο όρο των προβλέψεων του καθενός. Η ιδέα είναι να δημιουργηθούν τυχαία διάφορα υποσύνολα από το σύνολο εκπαίδευσης, τα οποία θα εκπαιδεύσουν από ένα δέντρο απόφασης. Καταλήγοντας σε ένα σύνολο διαφορετικών μοντέλων, υπολογίζεται ο μέσος όρος των προβλέψεων των διαφόρων δέντρων. Ο συνδυασμένος πλέον εκτιμητής, έχει συνήθως καλύτερα αποτελέσματα λόγω της ελαττωμένης διασποράς του. Εφαρμογή αυτής της μεθόδου θα γίνει μέσω του αλγορίθμου ***Random Forest***.

- *Μέθοδοι Ενδυνάμωσης (Boosting Methods):*
  Σε αυτές τις μεθόδους, οι εκτιμητές κατασκευάζονται διαδοχικά κι ο καθένας προσπαθεί να ελαττώσει την προκατάληψη (bias) του επόμενου. Κάθε δέντρο απόφασης μαθαίνει από το προηγούμενο, δίνοντας βάση στις παρατηρήσεις που εμφανίζουν μεγάλη απόκλιση ανάμεσα στην προβλεπόμενη και την πραγματική τιμή. Οπότε το τελικό δέντρο, είναι το βεβαρυμμένο ως προς το μέσο όρο όλων των προηγούμενων. Στόχος είναι να συνδυαστούν μερικά αδύναμα μοντέλα ώστε να παραχθεί ένα δυνατό, συνδυαστικό μοντέλο

με μειωμένο σφάλμα. Εφαρμογή αυτής της μεθόδου θα γίνει μέσω του αλγορίθμου ***Gradient Boosting***.



Εικόνα 2.4: Οπτικοποίηση τεχνικών Bagging και Boosting

Στην εικόνα 2.4 φαίνεται σχηματικά ο τρόπος λειτουργίας των τεχνικών που αναφέρθηκαν παραπάνω και που θα χρησιμοποιηθούν στη διπλωματική εργασία. Γίνεται αντιληπτό, ότι τα μοντέλα που αναπτύσσονται μέσω αυτών των τεχνικών έχουν καλύτερη απόδοση σε σχέση με το αποτέλεσμα ενός μεμονωμένου εκτιμητή.

2.2.1.1 Decision Trees

Τα μοντέλα που θα υλοποιηθούν βασίζονται στο χειρισμό των *δέντρων απόφασης (Decision Trees)*. Γι' αυτό κρίνεται απαραίτητο να οριστούν τα βασικά στοιχεία κι ο τρόπος λειτουργίας τους. Πρόκειται για απλές δομές που ντύνουν μοντέλα επιβλεπόμενης μηχανικής μάθησης. Έχουν εφαρμογές και στην παλινδρόμηση καθώς μπορούν να διαχειριστούν τις μη-γραμμικές σχέσεις που υπάρχουν μεταξύ των μεταβλητών. Μετά την εκπαίδευσή τους είναι σε θέση να προβλέψουν την τιμή της μεταβλητής εξόδου μέσω κανόνων εκμάθησης από τις μεταβλητές εισόδου του προβλήματος.

Ο εκτιμητής αποτελείται από τριών ειδών κόμβους. Ο *αρχικός κόμβος (root node)* αντιπροσωπεύει όλο το σύνολο εκπαίδευσης και διαχωρίζεται σε επιμέρους κόμβους. Ένας *εσωτερικός κόμβος (interior node)* περιλαμβάνει μεταβλητές του συνόλου εκπαίδευσης. Απ' αυτόν φεύγουν *κλαδιά (branches)* τα οποία αντιπροσωπεύουν τους κανόνες εκμάθησης. Τέλος, ο *κόμβος φύλλου ή τερματικός κόμβος (leaf/ terminal node)* αντιπροσωπεύει το αποτέλεσμα [18].

Σκοπός του δέντρου είναι να βρει ποια μεταβλητή διαχωρίζει τις τιμές της μεταβλητής εξόδου πιο «αγνά». Ο διαχωρισμός πραγματοποιείται σε κάθε επίπεδο/κόμβο του δέντρου, για όσο πληρούνται κάποια κριτήρια και μετά αναπτύσσεται

ο τερματικός κόμβος. Η τιμή που δίνεται ως πρόβλεψη σε κάθε κόμβο είναι ο μέσος όρος των τιμών της μεταβλητής εξόδου των δειγμάτων που περιλαμβάνονται στον κόμβο. Η διαδικασία που ακολουθείται για σύνολο N παρατηρήσεων και p μεταβλητών περιγράφεται παρακάτω [19].

Ορίζεται το σύνολο εκπαίδευσης $\boldsymbol{\mathcal{X}} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$, $\boldsymbol{x_i} = (x_{i1}, \dots, x_{ip})$.
Για κάθε μεταβλητή j και σημείο διαχωρισμού s, ορίζονται δύο χώροι

$$R_<(j,s) = \{x_{ij} | x_{ij} \leq s\} \ \kappa\alpha\iota \ R_>(j,s) = \{x_{ij} | x_{ij} \geq s\}$$

Για κάθε χώρο υπολογίζεται η προβλεπόμενη τιμή ως

$$c_< = ave(y_i | x_{ij} \in R_<) \ \kappa\alpha\iota \ c_> = ave(y_i | x_{ij} \in R_>)$$

Εφαρμόζοντας το κριτήριο του μέσου τετραγωνικού σφάλματος υπολογίζεται το σφάλμα της μεταβλητής j

$$MSE_j = \sum_{i \in R_<} (y_i - c_<)^2 + \sum_{i \in R_>} (y_i - c_>)^2$$

Εντέλει, για το διαχωρισμό επιλέγεται η μεταβλητή που δίνει το χαμηλότερο μέσο τετραγωνικό σφάλμα μέχρις ότου ο κόμβος που προκύπτει έχει ένα μόνο δείγμα αξιολόγησης. Αυτός είναι ο κόμβος φύλλου.

2.2.1.2 Random Forest

Στο *τυχαίο δάσος (Random Forest)* κάθε δέντρο απόφασης κατασκευάζεται από ένα υποσύνολο του συνόλου εκπαίδευσης που επιλέγεται τυχαία με αντικατάσταση κάθε φορά από το σύνολο εκπαίδευσης [20, 21]. Το μέγεθος του κάθε υποσυνόλου είναι ίδιο με το σύνολο εκπαίδευσης οπότε κάποιες παρατηρήσεις μπορεί να επαναληφθούν. Κατά την κατασκευή του δέντρου και το χωρισμό των κόμβων, η καλύτερη διάσπαση πραγματοποιείται είτε από όλα τα input features είτε από ένα τυχαίο υποσύνολο του οποίου το μέγεθος ορίζεται από την παράμετρο max_features. Η τυχαιότητα που χαρακτηρίζει αυτές τις διαδικασίες έχει ως σκοπό την ελάττωση της διακύμανσης του εκτιμητή [22]. Η μειωμένη διακύμανση πραγματοποιείται συνδυάζοντας διαφορετικά δέντρα, έχοντας ως κόστος κάποιες φορές ελαφρά αύξηση στην προκατάληψη. Ο αλγόριθμος Random Forest χρησιμοποιείται ευρέως έναντι των τυπικών αλγορίθμων που σχετίζονται με μεμονωμένα *δέντρα αποφάσεων (Decision Trees)*, διότι δείχνει αντοχή έναντι της υπερεκπαίδευσης.

Στη συγκεκριμένη εργασία που μελετάται ένα πρόβλημα παλινδρόμησης, γίνεται υλοποίηση της μεθόδου **RandomForestRegressor**. Η θεωρία του αλγορίθμου για N δείγματα και M δέντρα έχει ως εξής:

Random Forest predicted value at **x**

Input: Training set $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$

m_{try}: number of features to be taken into account when learning a split

nodesize: number of examples to be included in each cell in order to split

for i=1 to M **do:**

Draw $\boldsymbol{X}^i$ subsets of size N by bootstrap sampling (parameter Θ) from $\boldsymbol{\mathcal{D}}$

Set $n_{nodes} = 1$ and

while $n_{nodes} < n_{min}$ **do:**

Select m_{try} features at random

Select the best split in A by optimizing the CART-split criterion

Split the node A to two daughter nodes A_L and A_R

$n_{nodes} = n_{nodes} + 1$

end

Compute the predicted value $m_n(\boldsymbol{x};\ \Theta_i, \boldsymbol{\mathcal{D}})$ at **x** equal to the average of y_i falling in the node of **x**

end

Output: Compute the random forest estimate $m_{M,n}$ at the query point **x**

$$m_{M,n}(\boldsymbol{x};\ \Theta_1, \ldots, \Theta_M, \boldsymbol{\mathcal{D}}) = \frac{1}{M}\sum_{i=1}^{M} m_n(\boldsymbol{x};\ \Theta_i, \boldsymbol{\mathcal{D}})$$

Οι παράμετροι που επηρεάζουν τα αποτελέσματα ενός εκτιμητή RandomForest είναι πολλές αφού βασίζεται στο συνδυασμό πολλών δέντρων απόφασης. Παρακάτω αναγράφονται οι παράμετροι που θα εξεταστούν στο πλαίσιο της εργασίας όπως αυτές αναγράφονται στο Scikit-learn.

- **n_estimators:**
  ο αριθμός των δέντρων απόφασης σε κάθε δάσος
- **criterion:**
  η συνάρτηση με την οποία μετράται η ποιότητα του split
- **max_depth:**
  το μέγιστο βάθος ενός δέντρου απόφασης
- **min_samples_split:**
  ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για να συμβεί το split σε έναν κόμβο
- **min_samples_leaf:**
  ο ελάχιστος αριθμός δειγμάτων που απαιτούνται σε κάθε κόμβο φύλλου
  Το split θα πραγματοποιηθεί αν αφήνει τουλάχιστον τόσα δείγματα σε κάθε αριστερό και δεξί κλαδί.

2.2.1.3 Gradient Boosting

Ο αλγόριθμος Gradient Boosting κατασκευάζει προσθετικά μοντέλα προσαρμόζοντας διαδοχικά μία παραμετροποιημένη συνάρτηση στα ψευδό-υπόλοιπα μέσω της μεθόδου των ελαχίστων τετραγώνων σε κάθε επανάληψη. Συνδυάζει τον αλγόριθμο Gradient Descent με στοιχεία ενδυνάμωσης [17] καταλήγοντας σε *ενδυναμωμένα δέντρα απόφασης*. Η ενδυνάμωση περιλαμβάνει την προσαρμογή μιας πρόσθετης επέκτασης σε ένα σύνολο στοιχειωδών συναρτήσεων. Σκοπός είναι να ελαχιστοποιηθεί η συνάρτηση κόστους ως προς τη συνάρτηση ενδυνάμωσης και να υπολογιστούν τα ψευδό-υπόλοιπα που σχετίζονται με την κλίση της συνάρτησης κόστους σε σχέση με τις τιμές του μοντέλου σε κάθε σημείο του συνόλου εκπαίδευσης. Όπως αναφέρθηκε και πριν, η ακρίβεια των προσεγγίσεων κατά τις επαναλήψεις κι η ταχύτητα της εκτέλεσης μπορούν να βελτιωθούν αν ενσωματωθεί η τυχαιότητα . Αυτό σημαίνει ότι σε κάθε επανάληψη, επιλέγεται ένα δείγμα του συνόλου εκπαίδευσης στην τύχη και πάνω σε αυτό προσαρμόζεται αρχικά ο εκτιμητής.

Στην εργασία γίνεται εφαρμογή της μεθόδου **GradientBoostingRegressor** για τη μελέτη του προβλήματος. Η μαθηματική θεωρία του κώδικα ξεκινάει με την επιλογή συνάρτησης κόστους – Loss Function $L\big(y, F(\boldsymbol{x})\big)$ η οποία είναι παραγωγίσιμη κι αξιολογεί την απόδοση του μοντέλου. Ως προκαθορισμένη παράμετρος επιλέγεται η συνάρτηση ελαχίστων τετραγώνων (least squares, ls function)

$$L = \frac{1}{N}\sum_{i=1}^{N}\big(y_{actual} - y_{pred}\big)^2$$

Ο αλγόριθμος Gradient Boosting χρησιμοποιεί ένα βασικό – ασθενή εκτιμητή (weak learner) $\boldsymbol{h(x; a)}$ που θα κάνει τις προβλέψεις καθώς κι ένα επιπρόσθετο μοντέλο το οποίο προσθέτει αδύναμους εκτιμητές για να ελαχιστοποιήσει τη συνάρτηση κόστους [23]. Ως εκτιμητές επιλέγονται τα δέντρα παλινδρόμησης, των οποίων το αποτέλεσμα μπορεί να προστεθεί, επιτρέποντας τη διόρθωση των ψευδο-υπολοίπων στις προβλέψεις. Με το επιπλέον μοντέλο προστίθενται δέντρα ένα τη φορά ελαχιστοποιώντας για κάθε ένα τις παραμέτρους $\{\boldsymbol{a}_m\}_0^M$ στην εξίσωση της παλινδρόμησης.

Η προσέγγιση Gradient Tree Boosting [19, 24] που χρησιμοποιείται, αναφέρεται σε έναν εκτιμητή που είναι ένα δέντρο παλινδρόμησης τερματικού κόμβου J. Σε κάθε επανάληψη m, ένα δέντρο παλινδρόμησης χωρίζει το χώρο των **x** μεταβλητών σε $\{R_{jm}\}_1^J$ περιοχές. Μία σταθερά γ_j ορίζεται για κάθε περιοχή ώστε να ισχύει ο κανόνας πρόβλεψης

$$x \in R_j \Rightarrow F(x) = \gamma_j.$$

Επομένως ο εκτιμητής παίρνει τη μορφή

$$h\left(\boldsymbol{x};\ \{R_{jm}\}_1^{J_m}\right) = \sum_{j=1}^{J_m} \gamma_{jm} I\left(\boldsymbol{x} \in R_{jm}\right)$$

ενώ για κάθε περιοχή R που ορίζεται από τον κόμβο j, υπολογίζονται οι σταθερές

$$\gamma_{jm} = arg\,\min_{\gamma} \sum_{\boldsymbol{x}_i \in R_{jm}} L(y_i, F_{m-1}(\boldsymbol{x}_i) + \gamma).$$

Έπειτα για κάθε περιοχή ανανεώνεται η προσέγγιση $F_{m-1}(\boldsymbol{x})$ με τη σχέση

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \nu * \gamma_{jm} I\left(\boldsymbol{x} \in R_{jm}\right)$$

Η μορφή του αλγορίθμου έχει ως εξής:

Gradient Boosting prediction for observation vector **x**

Input: Training set $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$

Cost (Loss) Function $L\left(y, F(\boldsymbol{x})\right)$

M number of iterations

Set $F_0(\boldsymbol{x}) = arg\,min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$, a single terminal node tree

for m=1 to M **do**:

for i=1 to N **do:**

Calculate the *pseudo* residuals $r_{im} = -\left[\frac{\partial L(y_i, F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)}\right]_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})}$

for j=1 to J **do:**

$\{R_{jm}\}_1^{J_m} = J - terminal\ node\ tree(\{r_{im}, \boldsymbol{x}_i\}_1^N)$

$\gamma_{jm} = arg\,min_{\gamma} \sum_{\boldsymbol{x}_i \in R_{km}} L(y_i, F_{m-1}(\boldsymbol{x}_i) + \gamma)$

end

end

Update the predicted value $F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \nu * \gamma_{jm} I\left(\boldsymbol{x} \in R_{jm}\right)$, ν refers to the learning rate

end

Output: Predicted value $F_M(\boldsymbol{x})$

Η παράμετρος loss που αντιστοιχεί στη συνάρτηση κόστους διατηρείται στην προκαθορισμένη τιμή ls. Οι υπόλοιπες παράμετροι που θα εξεταστούν στη διπλωματική είναι οι εξής όπως περιγράφονται στο Scikit-learn:

- **learning_rate:**
  ελέγχει το κατά πόσο ελαττώνεται η συμβολή κάθε δέντρου. Όσο πιο μικρή είναι η παράμετρος τόσο καλύτερο generalization error δίνει ο αλγόριθμος δηλαδή βελτιώνεται η ακρίβεια.

- **n_estimators:**
  ο αριθμός των σταδίων που θα γίνουν. Στο συγκεκριμένο αλγόριθμο μεγάλος αριθμός εκτιμητών συνήθως δίνει καλύτερη απόδοση.
- **subsample:**
  το κλάσμα των δειγμάτων που χρησιμοποιείται για fitting των μεμονωμένων εκτιμητών. Επιλέγοντας τιμή μικρότερη της μονάδας, ελαττώνεται η διακύμανση αλλά αυξάνεται πιθανώς το bias.
- **criterion:**
  η συνάρτηση που μετράει την ποιότητα του διαχωρισμού. Η προκαθορισμένη τιμή εδώ είναι η 'friedman_mse' συνάρτηση, η οποία περιλαμβάνει βελτιώσεις από τον Friedman.
- **min_samples_split:**
  ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για το διαχωρισμό ενός εσωτερικού κόμβου
- **min_samples_leaf:**
  ο ελάχιστος αριθμός δειγμάτων που χρειάζονται σε έναν κόμβο φύλλων
- **max_depth:**
  το μέγιστο βάθος των μεμονωμένων εκτιμητών παλινδρόμησης. Περιορίζει τον αριθμό κόμβων σε ένα δέντρο.

2.2.2 Μετρικές αξιολόγησης

Μετά την κατασκευή ενός μοντέλου πρέπει να αξιολογηθούν τα αποτελέσματα ώστε να δει ο χρήστης πόσο αποδοτικός είναι ο αλγόριθμος που κατασκεύασε κι αν μπορεί να γενικευτεί η χρήση του. Για να γίνει αυτό, χρησιμοποιούνται οι λεγόμενες μετρικές αξιολόγησης ανάλογα το είδος του αλγορίθμου. Με τις τιμές των μετρικών που λαμβάνει ο χρήστης, μπορεί να κάνει τροποποιήσεις ώστε να βελτιώσει το μοντέλο και να φτάσει σε μεγαλύτερη ακρίβεια. Παρακάτω παρουσιάζονται οι μετρικές που χρησιμοποιούνται στους αλγορίθμους παλινδρόμησης [25, 26]. Σημειώνεται ότι ως σφάλμα (error) θεωρείται η απόσταση της πραγματικής τιμής από την προβλεπόμενη.

- *Mean Square Error (MSE):*
  Ανάλογα το πρόβλημα μπορεί να είναι επιθυμητή άλλη προσέγγιση που να δίνει βάση στις απομακρυσμένες τιμές του dataset που χρησιμοποιείται. Σε τέτοια περίπτωση, μπορεί να χρησιμοποιηθεί αυτή η μετρική που είναι το άθροισμα των τετραγωνικών διαφορών μεταξύ πραγματικής και προβλεπόμενης τιμής διαιρούμενο με το σύνολο των παρατηρήσεων.

$$MSE = \frac{\sum_{i=1}^{n}\left(y_{actual} - y_{pred}\right)^2}{n}$$

Επειδή δεν μπορεί να γίνει σύγκριση μεταξύ των σφαλμάτων του ίδιου μοντέλου αλλά μεταξύ όμοιων σφαλμάτων ανταγωνιστικών μοντέλων, χρησιμοποιούνται κι άλλες μετρικές γενικευμένης φύσεως.

- *Root Mean Square Error (RMSE):*
  Πρόκειται για τη ρίζα του μέσου τετραγωνικού σφάλματος. Θεωρείται καλός εκτιμητής της τυπικής απόκλισης των σφαλμάτων διότι με την αύξηση του n, η μέτρηση προσαρμόζεται και βελτιώνεται η ακρίβειά της.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{actual} - y_{pred}\right)^2}{n}}$$

  Είναι χρήσιμη μετρική, διότι αποδίδει την ακρίβεια του μοντέλου στην κλίμακα της εξαρτημένης μεταβλητής y.

- *Mean Absolute Percentage Error (MAPE):*
  Αναφέρεται στο λόγο του σφάλματος της πρόβλεψης προς την πραγματική τιμή. Προσδιορίζει το πόσο μακριά είναι οι προβλέψεις από τις πραγματικές τιμές κατά μέσο όρο.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_{actual} - y_{pred}}{y_{actual}}\right|$$

  Δε φέρει μονάδες κι εκφράζεται σε μορφή ποσοστού οπότε μπορεί να χρησιμοποιηθεί για να συγκρίνει δεδομένων διαφορετικών τάξεων.

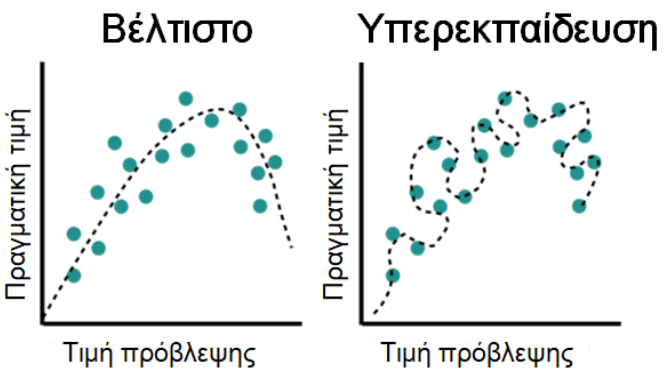
- *R-Squared (R^2):*
  Εκφράζει το κατά πόσο η διακύμανση στην εξαρτημένη μεταβλητή εξαρτάται από τη διακύμανση των ανεξάρτητων. Δίνεται σε ποσοστό και λαμβάνει τιμές μεταξύ 0 και 1. Μία μεγάλη τιμή θα σημαίνει ότι το μοντέλο δεν εξαρτάται από άλλους εξωτερικούς παράγοντες.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{actual} - y_{pred}\right)^2}{\sum_{i=1}^{n}(y_{actual} - \bar{y})^2}$$

  Ο αριθμητής αντιστοιχεί στο MSE, ενώ ο παρονομαστής στη διασπορά της μεταβλητής εξόδου. Οπότε, όσο μεγαλύτερο είναι το MSE, τόσο χαμηλότερη θα βγαίνει η συγκεκριμένη μετρική.

2.2.3 Υπερεκπαίδευση μοντέλου (Overfitting)

Κατά την εκπαίδευση ενός μοντέλου υπάρχουν κάποια ζητήματα που μπορεί να προκύψουν και να αλλοιώσουν την ακρίβεια του αλγορίθμου. Αυτά συνήθως σχετίζονται με το είδος και το πλήθος των δεδομένων.



Εικόνα 2.5: Διαγραμματική απεικόνιση υπερεκπαίδευσης στην παλινδρόμηση [3]

Ζητούμενο ενός μοντέλου μηχανικής μάθησης είναι ένα η ακρίβεια εκπαίδευσης, η οποία αναφέρεται στις σωστές προβλέψεις που κάνει το μοντέλο. Η υψηλή ακρίβεια στην εκπαίδευση όμως δεν είναι πάντα επιθυμητή. Για παράδειγμα, κάτι τέτοιο μπορεί να έχει προκύψει από *υπερβολική προσαρμογή* δεδομένων ή *υπερεκπαίδευση* (*overfitting*). Με την έννοια αυτή γίνεται λόγος για τη διαδικασία κατά την οποία το μοντέλο έχει μάθει κάθε λεπτομέρεια του συνόλου δεδομένων, μπορεί να συλλάβει θόρυβο και να επηρεαστεί η απόδοσή του σε νέα δεδομένα. Έχει υψηλή απόδοση στα δεδομένα στα οποία εκπαιδεύεται αλλά αδυνατεί να αποδώσει με παρόμοιο τρόπο σε άγνωστα δεδομένα. Ως αποτέλεσμα δημιουργείται ένα, εντέλει, μη γενικευμένο μοντέλο.

Επιπλέον, άλλο ζήτημα είναι η *ακρίβεια εκτός δείγματος (out-of-sample accuracy)*. Πρόκειται για το ποσοστό των σωστών προβλέψεων που κάνει το μοντέλο, σε δεδομένα πάνω στα οποία δεν έχει εκπαιδευτεί. Ζητούμενο για την ακρίβεια του μοντέλου είναι να έχει υψηλή out-of-sample accuracy, ώστε να μπορεί να χρησιμοποιηθεί σε επόμενα, άγνωστα δεδομένα.

Για να βελτιωθεί αυτού του είδους η ακρίβεια, χρησιμοποιείται μία προσέγγιση αξιολόγησης που καλείται train/test split. Πιο συγκεκριμένα, επιλέγεται ένα μέρος του dataset για *εκπαίδευση (train)* και το υπόλοιπο για *δοκιμή (test)*. Το μοντέλο αναπτύσσεται κι εκπαιδεύεται στο πρώτο σύνολο. Στη συνέχεια, τα features του συνόλου δοκιμής περνούν στο μοντέλο για εξαγωγή προβλέψεων και στο τέλος αυτές οι προβλέψεις συγκρίνονται με τις πραγματικές τιμές του συνόλου δοκιμής. Αυτή η πρακτική χρησιμοποιείται ευρέως διότι είναι πιο ρεαλιστική για πραγματικά προβλήματα.

Ο διαχωρισμός του συνόλου δεδομένων στα δύο υποσύνολα γίνεται με τυχαίο τρόπο. Στον κώδικα ωστόσο, έχει οριστεί το σύνολο δοκιμής να περιλαμβάνει το 25% των παρατηρήσεων, οπότε το σύνολο εκπαίδευσης θα περιλαμβάνει το υπόλοιπο 75%. Επομένως, το σύνολο εκπαίδευσης θα δουλέψει σε 1196 σειρές δεδομένων και το σύνολο δοκιμής στις υπόλοιπες 399, από τις 1595 που διατίθενται συνολικά.

Το ενδεχόμενο της υπερεκπαίδευσης μπορεί να διαπιστωθεί με τη βοήθεια διαγραμμάτων που αποδίδουν τα υπολειπόμενα σφάλματα των συνόλων εκπαίδευσης και δοκιμής. Στον άξονα x απεικονίζονται οι πραγματικές τιμές της μεταβλητής εξόδου ενώ στον y η διαφορά της πραγματικής από την προβλεπόμενη τιμή. Προστίθεται και μία ευθεία στο επίπεδο y=0. Αυτό συμβαίνει διότι ιδανικά, σκοπός είναι οι προβλέψεις να μην απέχουν από τις πραγματικές τιμές. Οπότε τα σημεία του άξονα y να βρίσκονται κοντά στην ευθεία y=0. Δεδομένου του ότι μελετάμε πληροφορίες από πειράματα που έχουν ποικιλομορφία στις συνθήκες υλοποίησής τους, αυτό δεν καθίσταται δυνατό. Ωστόσο αν στη συμπεριφορά των δύο συνόλων δε διαφαίνεται κάποια ομοιομορφία μπορεί να γίνει λόγος για πιθανή υπερεκπαίδευση.

2.2.4 Βελτιστοποίηση απόδοσης μοντέλου

Αφού έχει εκπαιδευτεί κι αξιολογηθεί το μοντέλο, μπορεί τα αποτελέσματα να μην είναι πλήρως ικανοποιητικά. Η αυθόρμητη σκέψη κάποιου θα είναι να βρει κάποιο άλλο ίσως πιο σύνθετο μοντέλο αλλά αυτή η λύση δεν είναι πάντα εύκολη στην υλοποίηση κι οδηγεί σε σύγχυση το χρήστη. Τότε υπάρχουν κάποιες πρακτικές που αφορούν το ίδιο, αρχικό μοντέλο και που ενδεχομένως να βελτιώνουν την απόδοσή του.

Από τη μία, μία συνήθως χρήσιμη σκέψη είναι να χρησιμοποιηθούν περισσότερα δεδομένα σχετικά με το πρόβλημα. Έχει συζητηθεί το γεγονός ότι η ποσότητα των δεδομένων μπορεί να είναι σημαντικότερη από τη συνθετότητα του μοντέλου που αναπτύσσεται. Εξ ορισμού, ένα μοντέλο μηχανικής μάθησης μαθαίνει μέσω της εμπειρίας δηλαδή καθώς βλέπει κάποια νέα σειρά παρατήρησης, μαθαίνει κάτι περισσότερο για τις σχέσεις μεταξύ των μεταβλητών εισόδου και της μεταβλητής εξόδου του προβλήματος. Παράλληλα, μπορεί να αυξηθεί ο αριθμός των μεταβλητών που θα λάβει υπόψη ο χρήστης. Έτσι, σε συνδυασμό με τη συσχέτιση μεταβλητών μπορεί να φανεί ποιες μεταβλητές είναι χρησιμότερες.

Από την άλλη, μπορεί να πραγματοποιηθεί μία διαδικασία που καλείται *συντονισμός υπερπαραμέτρων (hyperparameter tuning)*. Με απλά λόγια, σκοπός είναι η εύρεση συγκεκριμένων παραμέτρων ώστε το μοντέλο να δίνει τη βέλτιστη απόδοση [27]. Ο όρος hyperparameter χρησιμοποιείται για αποφυγή σύγχυσης σε σχέση με τις προκαθορισμένες παραμέτρους με τις οποίες εκπαιδεύεται αρχικά το μοντέλο. Τις υπερπαραμέτρους, πρέπει να τις ορίσει ο χρήστης συγκεκριμένα στην αρχή της σύνθεσης του εκτιμητή του μοντέλου. Για να δίνει αυτό, πρέπει να δοκιμάσει κάποιος όλους τους δυνατούς συνδυασμούς παραμέτρων, κάτι που με το χέρι μπορεί να πάρει πολλή ώρα. Πλέον έχουν αναπτυχθεί αυτοματοποιημένες μέθοδοι στο Scikit-learn, οι οποίες επιτρέπουν τον έλεγχο όλων των συνδυασμών δίνοντας κάποιο εύρος τιμών για κάθε παράμετρο που ενδιαφέρει το χρήστη. Συγκεκριμένα, δημιουργείται ένα *πλέγμα παραμέτρων (parameter grid)* για κάθε εκτιμητή κι επιλέγεται το μοντέλο GridSearchCV το οποίο υλοποιεί επίσης τις μεθόδους fit, predict και score. Οι

παράμετροι που χρησιμοποιούνται για την εφαρμογή των μεθόδων βελτιστοποιούνται με *διασταυρωμένη αναζήτηση (cross-validation)* στο πλέγμα με τις παραμέτρους.

Η έννοια του cross-validation σχετίζεται με την αξιολόγηση της απόδοσης του εκτιμητή. Όπως αναφέρθηκε παραπάνω, αλγόριθμος που εκπαιδεύεται κι αξιολογείται στο ίδιο σύνολο δεδομένων μπορεί να οδηγήσει σε overfitting. Για να αποφευχθεί αυτό, στις εφαρμογές επιτηρούμενης μάθησης στη βασική προσέγγιση του cross-validation, το σύνολο εκπαίδευσης χωρίζεται σε k υποσύνολα και κάθε υποσύνολο σε k επιμέρους τμήματα (k-folds). Η εκπαίδευση του μοντέλου σε κάθε υποσύνολο γίνεται χρησιμοποιώντας τα k-1 τμήματα κι η αξιολόγηση στο τμήμα που περισσεύει. Η απόδοση που προκύπτει από το k-fold cross-validation είναι ο μέσος όρος των τιμών από τα k υποσύνολα.

Εν γένει, ο συνδυασμός των παραμέτρων που δίνει την καλύτερη απόδοση στα υποσύνολα είναι αυτό στο οποίο καταλήγει το μοντέλο GridSearchCV. Μαζί με το τελική απόδοση, ο χρήστης μπορεί να δει την απόδοση κάθε συνδυασμού παραμέτρων, καθώς κι άλλα στατιστικά νούμερα που σχετίζονται με το χρόνο εκπαίδευσης και τα σφάλματα που δίνουν οι υπόλοιποι συνδυασμοί.

2.2.5 Κριτήρια επιλογής αλγορίθμων

Όπως έχει αναφερθεί κι οι δύο προς μελέτη αλγόριθμοι ανήκουν στις *ensemble* μεθόδους και συνδυάζουν το εξαγόμενο αποτέλεσμα μεμονωμένων δέντρων απόφασης. Πρόκειται για μη παραμετρικές μεθόδους οι οποίες διαφέρουν βασικά στον τρόπο που χτίζουν τα δέντρα, καθώς ο Random Forest χτίζει κάθε δέντρο ξεχωριστά διαλέγοντας ένα τυχαίο δείγμα του συνόλου δεδομένων, ενώ ο Gradient Boosting χτίζει ένα δέντρο τη φορά και το επόμενο παίρνει πληροφορίες από το προηγούμενο για να διορθώσει λάθη. Παρακάτω παρατίθεται ένας πίνακας με τις ιδιαιτερότητες κάθε αλγορίθμου και τους λόγους για τους οποίους επιλέχθηκε [28, 29].

Κριτήρια Επιλογής	Random Forest	Gradient Boosting
Ευκολία στο συντονισμό των υπερπαραμέτρων	✔	✘
Καλύτερα αποτελέσματα εφόσον γίνει καλός συντονισμός	✘	✔
Ανθεκτικότητα στην υπερεκπαίδευση	✔	✘
Bias υπέρ των κατηγορικών μεταβλητών	✔	✘
Καλύτερη απόδοση σε προβλήματα που σχετίζονται με την κλίση συνάρτησης	✘	✔

2.2.6 Ερμηνεία μοντέλου – Σημαντικότητα μεταβλητών

Σε κάθε μοντέλο είναι επιθυμητό να ποσοτικοποιηθεί η χρησιμότητα κάθε μεταβλητής. Πάλι μέσω του Scikit-learn, μπορεί να εκτιμηθεί η συμβολή κάθε χαρακτηριστικού ώστε να βελτιωθεί η προβλεψιμότητα του αλγορίθμου. Όσο υψηλότερη η τιμή, τόσο σημαντικότερη θεωρείται η μεταβλητή. Το άθροισμα της σημαντικότητας όλων των μεταβλητών κάνει μονάδα. Για να εξακριβωθεί η συμβολή κάθε μεταβλητής, η μελέτη γίνεται ανάποδα, αφαιρώντας τη μεταβλητή από την εκπαίδευση του μοντέλου και παρατηρώντας πως μεταβάλλονται οι τιμές πρόβλεψης. Στην πορεία, πραγματοποιείται αριθμητικός υπολογισμός, δημιουργείται μία λίστα που δείχνει κάθε μεταβλητή με το ποσοστό σημαντικότητας και στο τέλος γίνεται μία ταξινόμηση από την υψηλότερη προς τη χαμηλότερη συμβολή. Επομένως, διαφαίνονται οι μεταβλητές που παίζουν ενεργό ρόλο στην πρόβλεψη των εκάστοτε τιμών y. Αυτή η διαδικασία ακολουθείται στα σχετικά διαγράμματα σημαντικότητας των μεταβλητών που θα παρουσιαστούν στα αποτελέσματα στο επόμενο κεφάλαιο.

Σε μελλοντικές εκτελέσεις του μοντέλου, αν αφαιρεθούν οι μεταβλητές με χαμηλή σημαντικότητα, η απόδοση θα πρέπει να παραμείνει πάνω κάτω σταθερή [30]. Εναλλακτικά, αν χρησιμοποιηθεί άλλος, διαφορετικός αλγόριθμος, ο χρήστης μπορεί να καταχωρήσει μόνο τις μεταβλητές που έχουν ταυτοποιηθεί ως σημαντικές από το πρώτο μοντέλο. Συγκεκριμένα εδώ, θα δοκιμαστούν κι οι δύο τεχνικές, τα αποτελέσματα των οποίων θα παρουσιαστούν στη συνέχεια.

2.2.7 Προγραμματιστικά εργαλεία – Python

Τα χαρακτηριστικά της Python είναι πολλά κι είναι αυτά που την έχουν αναδείξει ως την καλύτερη γλώσσα δημιουργίας κι εφαρμογής αλγορίθμων μηχανικής μάθησης. Πρόκειται για μία γλώσσα σταθερή, ευέλικτη και με πληθώρα εργαλείων φιλικών προς το χρήστη. Μπορεί να χρησιμοποιηθεί για οποιαδήποτε διαδικασία που σχετίζεται με τεχνητή νοημοσύνη όπως η ανάλυση δεδομένων, η οπτικοποίηση κι η επεξεργασία φυσικής γλώσσας [31]. Στην εργασία χρησιμοποιήθηκε η Python 3.8.2.

Ένα βασικό προσόν της αποτελεί η ύπαρξη βιβλιοθηκών κι επεκτάσεων που διευκολύνουν το χρήστη κι εξοικονομούν χρόνο [32]. Η έννοια της βιβλιοθήκης αναφέρεται σε ήδη έτοιμο γραμμένο κώδικα που μπορεί να χρησιμοποιηθεί για επίλυση προγραμματιστικών ζητημάτων. Οι βιβλιοθήκες/ πακέτα/ τάξεις που χρησιμοποιούνται στην παρούσα εργασία παρατίθενται παρακάτω με την έκδοση που επιλέχτηκε:

- *Numpy (1.17.4):*
  Πακέτο γενικής χρήσης για επεξεργασία διανυσμάτων ή πινάκων. Παρέχει ένα πολυδιάστατο πίνακα ως αντικείμενο και τα αντίστοιχα εργαλεία για την επεξεργασία του.

- *Pandas (0.25.3):*
  Πακέτο που χρησιμοποιείται για διαχείριση κι ανάλυση δεδομένων διαφόρων τύπων όπως αρχεία excel ή csv. Παρέχει ένα DataFrame ως αντικείμενο και τα εργαλεία που το διατρέχουν κι επιτελούν διάφορες λειτουργίες στις σειρές και τις στήλες.

- *Matplotlib (3.1.2):*
  Πακέτο που παρέχει εργαλεία για οπτικοποίηση στατικών, κινούμενων και διαδραστικών διαγραμμάτων.

- *Category Encoders (2.2.2):*
  Πρόκειται για σύνολο μετασχηματιστών τύπου scikit-learning για κωδικοποίηση κατηγορικών μεταβλητών σε αριθμητικά με διαφορετικές τεχνικές. Συγκεκριμένα εδώ, χρησιμοποιείται η OneHot ή πλαστή κωδικοποίηση όπου για κάθε κατηγορία ενός χαρακτηριστικού, δημιουργείται ένα νέο χαρακτηριστικό, το οποίο είναι δυαδικό.

- *Scikit-learn (0.23.0):*
  Βιβλιοθήκη που περιέχει διαμορφωμένους, ποικίλους αλγορίθμους μηχανικής μάθησης. Έχει εργαλεία για προ-επεξεργασία δεδομένων, για πρόβλεψη κι αξιολόγηση των μοντέλων.

- *KNNImputer:*
  Τάξη του Scikit-learn που πραγματοποιεί συμπλήρωση τιμών που λείπουν χρησιμοποιώντας την προσέγγιση K πλησιέστερων γειτόνων. Για να βρεθούν οι γείτονες χρησιμοποιείται από προεπιλογή μία μετρική ευκλείδειας απόστασης. Κάθε τιμή στοιχείου που λείπει εκτιμάται από τις τιμές των κοντινών γειτόνων που έχουν βρεθεί στο προηγούμενο βήμα.

- *Seaborn (0.10.1):*
  Βιβλιοθήκη οπτικοποίησης δεδομένων Python που βασίζεται στο Matplotlib και παρέχει διαδραστικά κι ελκυστικά στατιστικά διαγράμματα.

3 Εκτέλεση - Αποτελέσματα

Αφού έχουν περιγραφεί τα εργαλεία που χρησιμοποιούνται, πλέον παρουσιάζεται η ροή εργασίας που ακολουθείται και τα αποτελέσματα που προκύπτουν. Γίνεται αναφορά στη συσχέτιση των μεταβλητών κι έπειτα πραγματοποιείται ξεχωριστή μελέτη για κάθε αλγόριθμο. Λόγω του μικρού μεγέθους του dataset, για να υπάρχει καλύτερη στατιστική εικόνα, κάθε μοντέλο εκτελέστηκε 10 φορές. Ενδεικτικά, παρουσιάζονται τα σχετικά διαγράμματα. Κατά την ανάπτυξη των μοντέλων που βασίζονται στους αλγορίθμους Random Forest και Gradient Boosting ακολουθείται η εξής διαδικασία:

- *Πρώτη προσέγγιση:*
  Ορισμός, εκπαίδευση και αξιολόγηση βασικού μοντέλου (χρησιμοποιώντας μόνο τις προκαθορισμένες παραμέτρους)
- Δημιουργία πλέγματος παραμέτρων δίνοντας έμφαση σε εκείνες που χαρακτηρίζουν περισσότερο κάθε μοντέλο ώστε να βρεθεί ο συνδυασμός που δίνει το βέλτιστο αποτέλεσμα
- *Δεύτερη προσέγγιση:*
  Ορισμός, εκπαίδευση και αξιολόγηση νέου μοντέλου (χρησιμοποιώντας τις παραμέτρους που επιλέχθηκαν από το πλέγμα)
- *Τρίτη προσέγγιση:*
  (μόνο στο μοντέλο Random Forest)
  Ορισμός, εκπαίδευση και αξιολόγηση εκ νέου βασικού και τροποποιημένου μοντέλου (τροποποιώντας τις τιμές της παραμέτρου βάθος δέντρου απόφασης)

Ταυτόχρονα για κάθε μοντέλο γίνεται ξεχωριστή μελέτη

- Ενδεχόμενης υπερεκπαίδευσης
- Σφαλμάτων
- Σημαντικότητας μεταβλητών

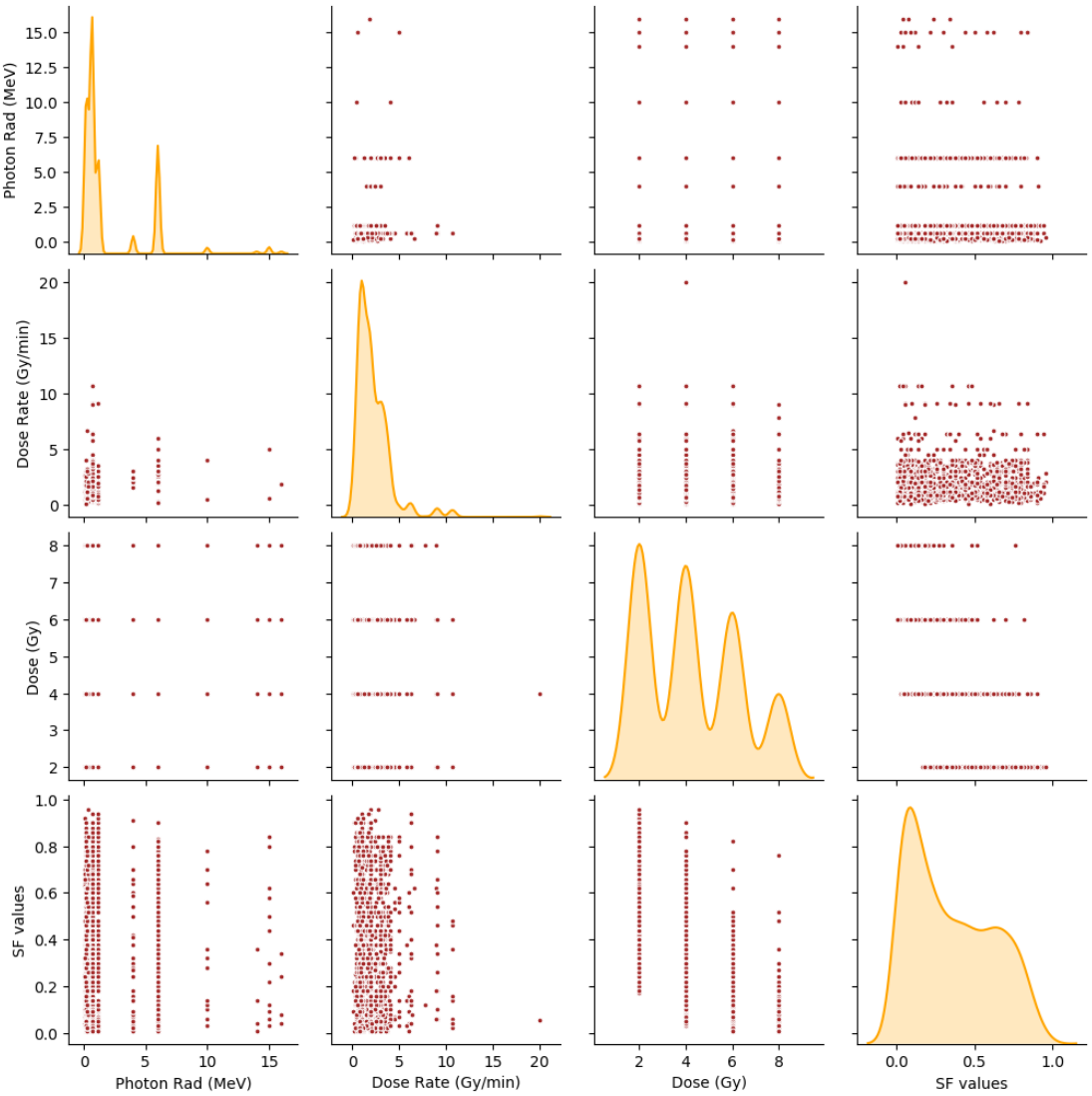
3.1 Συσχετίσεις Μεταβλητών

Όπως αναφέρθηκε στο Κεφάλαιο 2, πραγματοποιήθηκε η προ-επεξεργασία των δεδομένων ώστε να ετοιμαστεί το τελικό dataset που θα χρησιμοποιηθεί για την υλοποίηση των μοντέλων. Πριν γίνει η ανάπτυξη κάποιου μοντέλου μηχανικής μάθησης, είναι χρήσιμη η μελέτη των συσχετίσεων που υπάρχουν μεταξύ των μεταβλητών. Αυτό πραγματοποιείται μέσω της κατασκευής του πίνακα συσχετίσεων, ο οποίος περιλαμβάνει τους συντελεστές γραμμικής συσχέτισης μεταξύ των μεταβλητών. Οι συντελεστές παίρνουν αριθμητικές τιμές και για την εκτίμησή τους δε λαμβάνονται υπόψη οι τιμές του συνόλου των δεδομένων που μπορεί να είναι null.

Συγκεκριμένα για το πρόβλημα που αντιμετωπίζεται εδώ, στον πίνακα θα απεικονίζονται οι συντελεστές μεταξύ των στηλών Photon Rad (MeV), Dose Rate (Gy/min), Dose (Gy) και SF values. Επιλέγεται η μέθοδος Pearson, στην οποία το εύρος τιμών είναι από -1 έως +1. Η τιμή -1 αντιστοιχεί σε 100% αρνητική συσχέτιση, ενώ η τιμή +1 σε 100% θετική.

	Photon Rad	**Dose Rate**	**Dose**	**SF values**
Photon Rad	1	0,113715	0,054179	-0,029868
Dose Rate	0,113715	1	0,002066	-0,038433
Dose	0,054179	0,002066	1	-0,788845
SF values	-0,029868	-0,038433	-0,788845	1

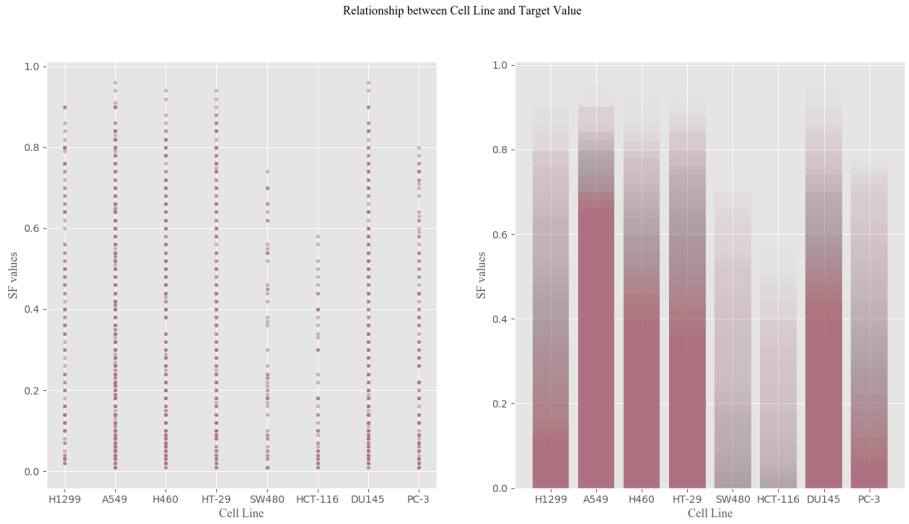
Διαπιστώνεται η αναμενόμενη αντίστροφη σχέση μεταξύ της δόσης και του κλάσματος επιβίωσης κατά ένα ποσοστό σχεδόν 80%. Αυτό σημαίνει ότι όσο αυξάνεται η δόση, θα ελαττώνεται η επιβίωση των καρκινικών κυττάρων. Για τις υπόλοιπες μεταβλητές δεν προκύπτουν έντονες συσχετίσεις. Για καλύτερη οπτικοποίηση των δεδομένων, παρατίθενται τα σχετικά διαγράμματα.



Εικόνα 3.1: Pairplot συσχετίσεων αριθμητικών μεταβλητών

Μέσω του pairplot στα διαγράμματα της διαγωνίου δίνεται η κατανομή των δεδομένων για την εκάστοτε μεταβλητή. Παρατηρείται ότι μεταξύ των περισσότερων μεταβλητών δεν υπάρχει γραμμική συσχέτιση μεταξύ των δεδομένων όπως φαίνεται άλλωστε κι αριθμητικά από τους συντελεστές. Συνδυάζοντας την πληροφορία από τον πίνακα και το αντίστοιχο διάγραμμα, γίνεται αντιληπτή η ανάστροφη γραμμική συσχέτιση μεταξύ της δόσης και του κλάσματος επιβίωσης, ενώ οι υπόλοιπες μεταβλητές φαίνονται να είναι ασυσχέτιστες μεταξύ τους.

Για την κατηγορική μεταβλητή Cell Line κατασκευάζονται τα παρακάτω διαγράμματα όπου παρουσιάζεται η συσχέτιση με την μεταβλητή εξόδου.



Εικόνα 3.2: Scatter plot και Barplot συσχέτισης κατηγορικής μεταβλητής Cell Line

Παρατηρείται ότι τα κλάσματα επιβίωσης κι η ανθεκτικότητα των κυτταρικών σειρών SW480 και HCT-116 ξεκινούν από χαμηλότερες τιμές και φθίνουν γρηγορότερα. Από τη μία, αυτό εξηγείται γιατί τα δεδομένα που υπήρχαν για τις δεδομένες κυτταρικές σειρές ήταν λιγότερα. Από την άλλη, φαίνεται ότι οι υψηλότερες τιμές επιβίωσης είναι έτσι κι αλλιώς χαμηλότερες από των υπολοίπων σειρών που σημαίνει ότι περισσότερα κύτταρα εξοντώθηκαν για παρόμοιες δόσεις ακτινοβολίας. Επιπλέον, ενδιαφέρον παρουσιάζει η κατανομή των τιμών στις σειρές DU145 και PC-3 που φθίνουν ομοιόμορφα και αγγίζουν χαμηλά ποσοστά επιβίωσης. Αυτό σημαίνει ότι η αντιμετώπιση του καρκίνου του προστάτη είναι ευκολότερη σε σχέση με άλλους τύπους καρκίνου πράγμα που επιβεβαιώνεται από πραγματικά στατιστικά στοιχεία [33, 34].

3.2 Random Forest

Για την ανάπτυξη του μοντέλου μέσω Random Forest καλείται η βιβλιοθήκη Scikit-learn που επιτρέπει τη χρήση των σχετικών μεθόδων του αλγορίθμου:

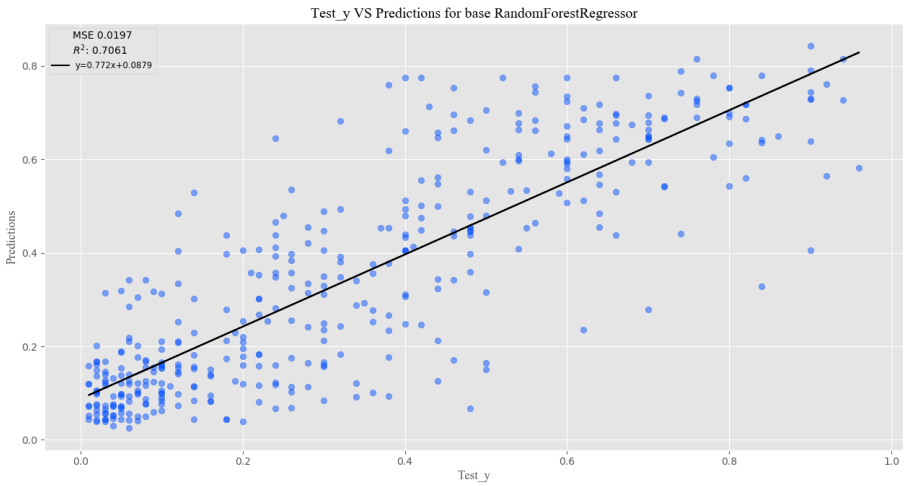
sklearn.ensemble.RandomForestRegressor

3.2.1 Πρώτη Προσέγγιση – Βασικό Μοντέλο

Αρχικά, κατασκευάζεται το βασικό μοντέλο με τις προκαθορισμένες παραμέτρους:

('criterion'= 'mse', 'max_depth'= None, 'min_samples_leaf'= 1,
'min_samples_split'= 2, 'n_estimators'= 100)

Στο παρακάτω διάγραμμα αναπαρίσταται η προσαρμογή του συνόλου δοκιμής κι η σχέση μεταξύ των πραγματικών τιμών στον οριζόντιο άξονα και των προβλεπόμενων στον κατακόρυφο. Σκοπός είναι να φανεί πόσο κοντά στην ευθεία y = x έρχεται η ευθεία που δημιουργούν προσεγγιστικά τα ζευγάρια πραγματικών και προβλεπόμενων τιμών.



Εικόνα 3.3: Προσαρμογή βασικού μοντέλου RF

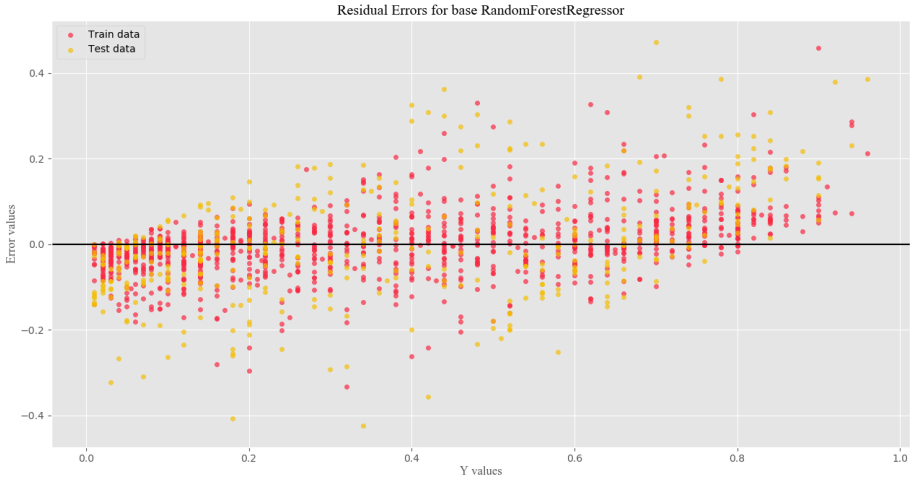
Υπάρχει ήδη μία σχετικά καλή προσαρμογή με την ευθεία που δημιουργείται μέσω γραμμικής παλινδρόμησης να είναι η

$$y = 0.772x + 0.0879.$$

Η αξιολόγηση των αποτελεσμάτων πραγματοποιείται με τη χρήση των μετρικών που αναφέρθηκαν στο προηγούμενο κεφάλαιο. Από την προσαρμογή προκύπτουν

$$MSE = 0.0197 \ \& \ R^2 = 0.7061.$$

Το παρακάτω διάγραμμα δίνει δεδομένα ως προς την απουσία της υπερεκπαίδευσης.

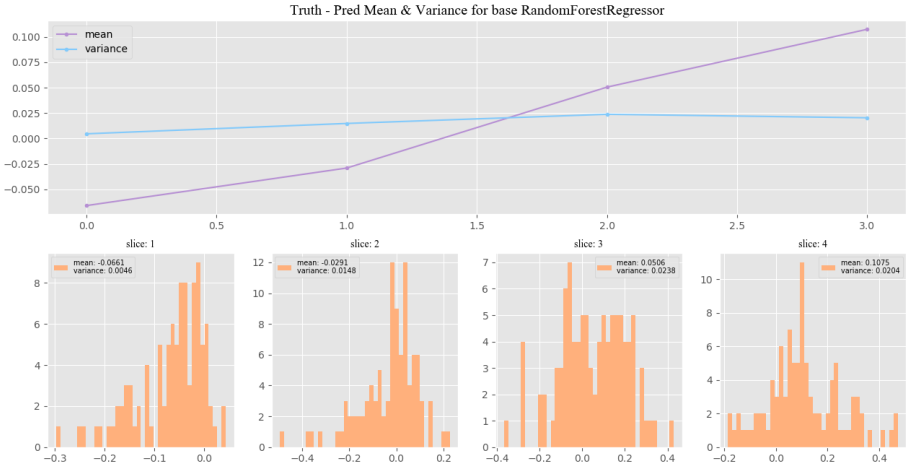


Εικόνα 3.4: Εκτίμηση σφαλμάτων βασικού μοντέλου RF

Οπτικά, φαίνεται ότι τα σφάλματα του συνόλου δοκιμής σε ένα βαθμό περικλείουν τα σφάλματα του συνόλου εκπαίδευσης, έχουν δηλαδή μεγαλύτερη διακύμανση. Ταυτόχρονα όμως, υπάρχει μία ομοιομορφία στο διάγραμμα, αφού δεν παρατηρείται συσσώρευση σφαλμάτων στην οριζόντια ευθεία που αντιστοιχεί σε μηδενική τιμή σφάλματος. Εντέλει φαίνεται πως ο αλγόριθμος προσαρμόζει αποδοτικά τον εκτιμητή του. Μιλώντας με νούμερα τα μέσα τετραγωνικά σφάλματα για κάθε σύνολο έχουν ως εξής:

$$MSE_{Train} = 0.0052 \ \& \ MSE_{Test} = 0.0182 \, .$$

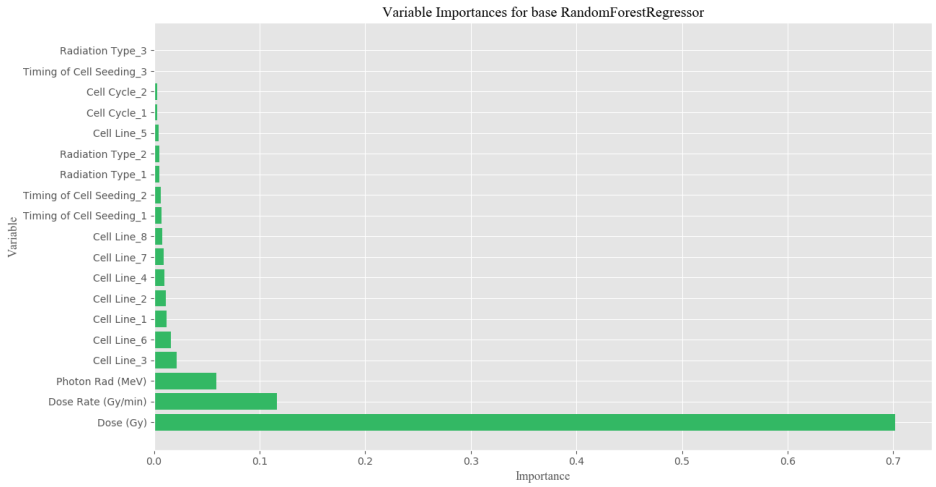
Σαν επιπλέον μέτρα απόδοσης χρησιμοποιούνται ο μέσος όρος κι η διακύμανση των σφαλμάτων. Οι πραγματικές κι οι προβλεπόμενες τιμές γίνονται ζευγάρια, ταξινομούνται σε αύξουσα σειρά και σπάνε σε 4 ισομεγέθεις λίστες. Έτσι, μπορεί να γίνει τμηματική μελέτη του μέσου όρου και της διακύμανσης, ώστε να φανεί η πιθανή ύπαρξη κάποιου μοτίβου στο σφάλμα της πρόβλεψης.



Εικόνα 3.5: Μελέτη κατανομής σφαλμάτων βασικού μοντέλου RF

Στην εικόνα 3.5 παρατηρείται ότι τα περισσότερα σφάλματα βρίσκονται συγκεντρωμένα στην περιοχή [-0.07, 0.1]. Αυτό δίνει μία σταθερή διακύμανση που φαίνεται στη γαλάζια γραμμή. Η μωβ γραμμή του μέσου όρου έχει αυξητική τάση ξεκινώντας από το -0.07 περίπου έως το 0.1. Αυτό σημαίνει ότι για μικρά ποσοστά επιβίωσης, η πρόβλεψη δίνει ελαφρώς υψηλότερες τιμές, ενώ όσο αυξάνονται οι τιμές επιβίωσης, η πρόβλεψη δίνει μικρότερες τιμές με αποτέλεσμα να αυξάνεται κι ο μέσος όρος του σφάλματος.

Τέλος, παρατίθεται το διάγραμμα που περιγράφει τη σημαντικότητα όλων των μεταβλητών με τον αντίστοιχο πίνακα.



Εικόνα 3.6: Σημαντικότητα μεταβλητών βασικού μοντέλου RF

Σημαντικότητα μεταβλητών για *βασικό* μοντέλο Random Forest			
Dose (Gy)	0.70	**Timing of Cell Seeding_1**	0.01
Dose Rate (Gy/min)	0.12	**Timing of Cell Seeding_2**	0.01
Photon Rad (MeV)	0.06	**Radiation Type_1**	0.01
Cell Line_3	0.02	**Radiation Type_2**	0.01
Cell Line_6	0.02	**Cell Line_5**	0.00
Cell Line_1	0.01	**Cell Cycle_1**	0.00
Cell Line_2	0.01	**Cell Cycle_2**	0.00
Cell Line_4	0.01	**Timing of Cell Seeding_3**	0.00
Cell Line_7	0.01	**Radiation Type_3**	0.00
Cell Line_8	0.01		

Προκύπτει ότι οι σημαντικότερες μεταβλητές στο πρόβλημα είναι η δόση, ο ρυθμός δόσης κι η ενέργεια με ποσά 0.70, 0.12 και 0.06 ενώ ακολουθούν οι κυτταρικές σειρές 3 κι 6 που αντιστοιχούν στον H460 μη- μικροκυτταρικό καρκίνο πνεύμονα και στον HCT-116 καρκίνο παχέος εντέρου.

3.2.2 Δεύτερη Προσέγγιση – Νέο Μοντέλο

Για να διαπιστωθεί αν μπορεί να βελτιωθεί η απόδοση του αλγορίθμου, δημιουργείται ένα πλέγμα παραμέτρων. Δοκιμάζονται όλοι οι πιθανοί συνδυασμοί με σκοπό να βρεθεί το σύνολο των παραμέτρων που βελτιστοποιεί τα αποτελέσματα. Η εισαγωγή του πλέγματος γίνεται μέσω του Scikit-learn:

sklearn.model_selection.GridSearchCV

Οι παράμετροι που εξετάζονται έχουν περιγραφεί στο Κεφάλαιο 2, οπότε το πλέγμα έχει ως εξής:

Κριτήριο διαχωρισμού	mae, mse
Βάθος δέντρου απόφασης	None, 5, 10
Αριθμός ελάχιστων δειγμάτων κόμβου φύλλου	1, 2, 3
Αριθμός ελάχιστων δειγμάτων για διαχωρισμό	2, 5, 10
Αριθμός δέντρων απόφασης	100, 500, 1000

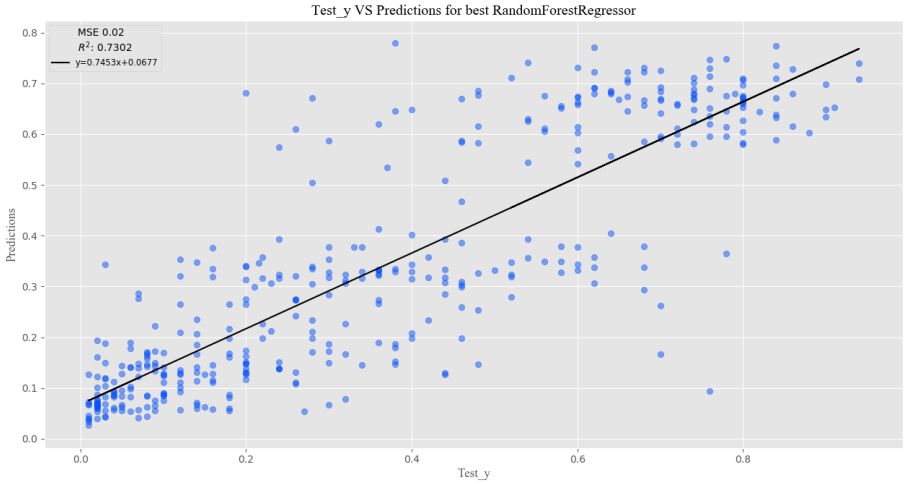
Δημιουργείται ένα αρχείο στο οποίο περιλαμβάνονται αποτελέσματα όλων των συνδυασμών των παραμέτρων. Τα αποτελέσματα σχετίζονται με την χρόνο της προσαρμογής και τα σκορς που δίνει καθένα από τα 5 σύνολα δοκιμής του cross validation. Ως βέλτιστος συνδυασμός παραμέτρων επιλέγεται αυτός που δίνει το

μεγαλύτερο μέσο σκορ. Για το μοντέλο Random Forest και το σύνολο των παρακάτω διαγραμμάτων, αυτό συμβαίνει με το συνδυασμό

('criterion'= 'mae', 'max_depth'= 10, 'min_samples_leaf'= 3,
'min_samples_split'= 10, 'n_estimators'= 500)

με τον οποίο επανεκτελείται το μοντέλο.

Κατασκευάζονται τα αντίστοιχα διαγράμματα με τη βασική προσέγγιση.



Εικόνα 3.7: Προσαρμογή νέου μοντέλου RF

Ενώ φαίνεται πως οι χαμηλές τιμές προσαρμόζονται καλύτερα, όσο αυξάνονται οι αριθμητικές τιμές του ζευγαριού πραγματική τιμή – πρόβλεψη οι τιμές απομακρύνονται από την ευθεία y = x, με την ευθεία που προκύπτει να είναι η

$$y = 0.7453x + 0.0677\,.$$

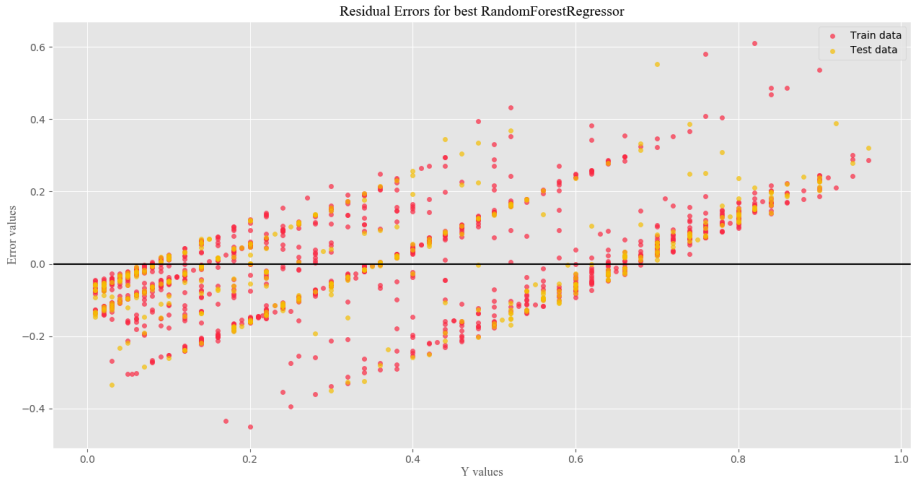
Οι μετρικές του διαγράμματος δίνουν

$$MSE = 0.02 \;\&\; R^2 = 0.7302\,.$$

Ουσιαστικά το σφάλμα μένει σταθερό και το R^2 αυξάνεται αλλά πολλές τιμές φαίνεται να ξεφεύγουν από την προσαρμογή της ευθείας. Μελετώντας την υπερεκπαίδευση επιβεβαιώνεται ότι η κατανομή των δεδομένων αλλάζει καθώς στην εικόνα 3.8 φαίνεται μία περίεργη δομή που ενώ έχει μικρά σφάλματα, δείχνει να είναι προβληματική. Πράγματι, τα μέσα τετραγωνικά σφάλματα είναι τα παρακάτω:

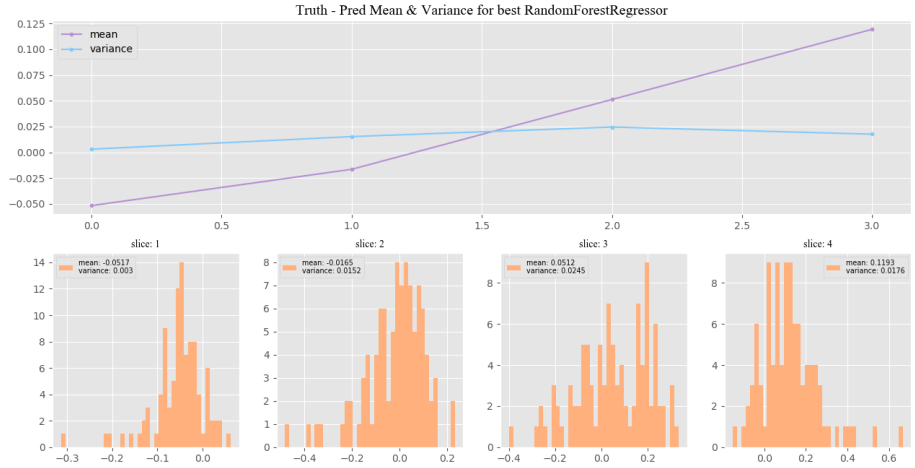
$$MSE_{Train} = 0.0179 \;\&\; MSE_{Test} = 0.0176\,.$$

Σε προσπάθεια άλλων εκτελέσεων το διάγραμμα που επεστράφη δεν είχε ιδιαίτερες διαφορές.



Εικόνα 3.8: Εκτίμηση σφαλμάτων νέου μοντέλου RF

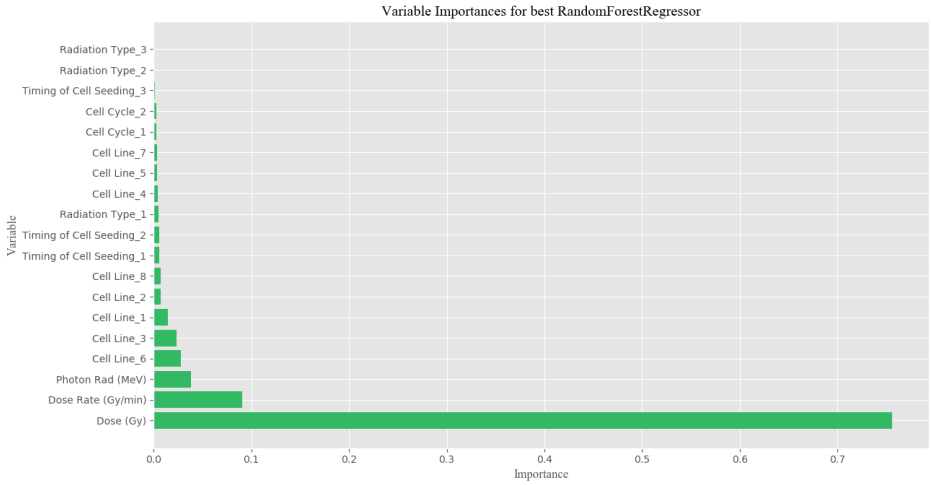
Παρατίθενται επίσης τα διαγράμματα μελέτης των σφαλμάτων για το νέο μοντέλο καθώς και της σημαντικότητας των μεταβλητών.



Εικόνα 3.9: Μελέτη κατανομής σφαλμάτων νέου μοντέλου RF

Εδώ πλέον γίνεται ξεκάθαρο ότι η επιλεγμένη προσέγγιση με τις νέες παραμέτρους στην ουσία δε βελτιώνει το μοντέλο, αφού το εύρος του μέσου όρου των σφαλμάτων μένει περίπου το ίδιο με πριν, με τιμές από το -0.05 στο 0.125. Στα τμήματα δε, που περιλαμβάνουν τις μεγαλύτερες αριθμητικές τιμές των ζευγαριών πραγματικής και

προβλεπόμενης τιμής, υπάρχουν σφάλματα που ξεπερνούν το 0.2 που παρατηρήθηκε πριν και αγγίζουν το ±0.4 και το 0.6.



Εικόνα 3.10: Σημαντικότητα μεταβλητών νέου μοντέλου RF

Οι μεταβλητές παραμένουν ως επί το πλείστον σταθερές ως προς τη σημαντικότητα με κάποιες αριθμητικές διαφορές ως προς το βασικό μοντέλο.

Σημαντικότητα μεταβλητών για *νέο* μοντέλο Random Forest			
Dose (Gy)	0.76	**Radiation Type_1**	0.01
Dose Rate (Gy/min)	0.09	**Cell Line_4**	0.00
Photon Rad (MeV)	0.04	**Cell Line_5**	0.00
Cell Line_6	0.03	**Cell Line_7**	0.00
Cell Line_3	0.02	**Cell Cycle_1**	0.00
Cell Line_1	0.01	**Cell Cycle_2**	0.00
Cell Line_2	0.01	**Timing of Cell Seeding_3**	0.00
Cell Line_8	0.01	**Radiation Type_2**	0.00
Timing of Cell Seeding_1	0.01	**Radiation Type_3**	0.00
Timing of Cell Seeding_2	0.01		

3.2.3 Τρίτη Προσέγγιση – Τροποποίηση Βάθους ΔΑ

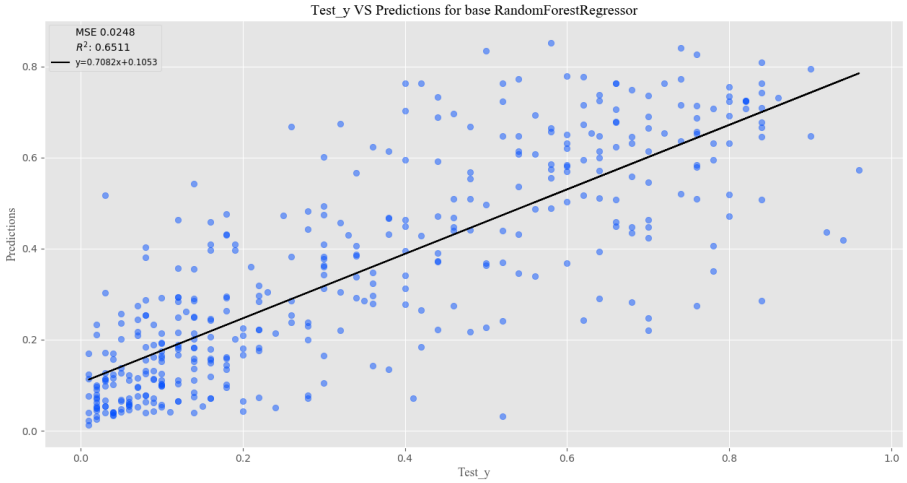
Δεδομένης της μορφής που έχουν τα διαγράμματα της προσαρμογής και της υπερεκπαίδευσης του νέου μοντέλου Random Forest, αποφασίζεται να τροποποιηθούν οι τιμές της παραμέτρου βάθος δέντρου απόφασης. Αυτό συμβαίνει διότι κρίνεται πως η μορφή των διαγραμμάτων σχετίζεται με τη μεγάλη τιμή βάθους (10) που έχει επιλεχθεί από την αναζήτηση που έγινε, με αποτέλεσμα να πραγματοποιείται μάλλον υπερεκπαίδευση. Οι νέες τιμές και το νέο πλέγμα παραμέτρων έχουν ως εξής:

Κριτήριο διαχωρισμού	mae, mse
Βάθος δέντρου απόφασης	None, 3, 4
Αριθμός ελάχιστων δειγμάτων κόμβου φύλλου	1, 2, 3
Αριθμός ελάχιστων δειγμάτων για διαχωρισμό	2, 5, 10
Αριθμός δέντρων απόφασης	100, 500, 1000

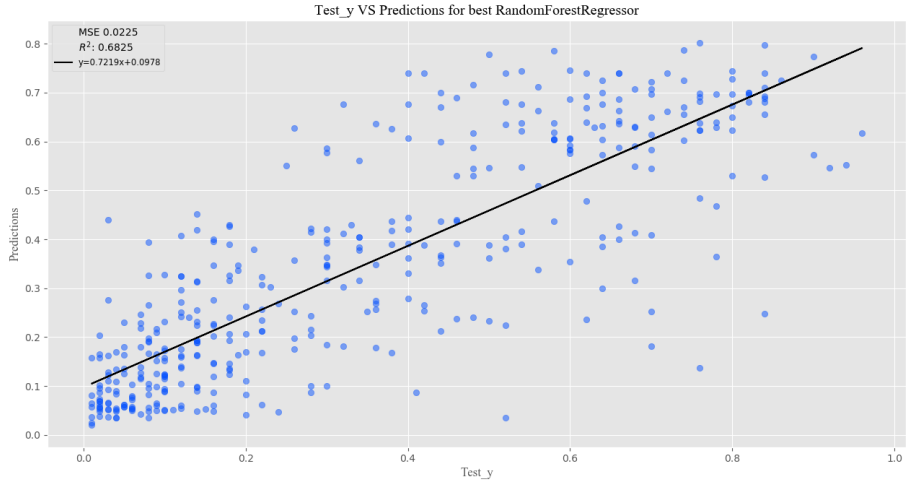
Εκτελώντας το μοντέλο ξανά, προκύπτει ο παρακάτω συνδυασμός των παραμέτρων που βελτιώνει το αποτέλεσμα

('criterion'= 'mae', 'max_depth'= None, 'min_samples_leaf'= 1, 'min_samples_split'= 10, 'n_estimators'= 500)

Επειδή πραγματοποιήθηκε εκ νέου εκπαίδευση του μοντέλου, παρακάτω συγκρίνονται τα διαγράμματα της βασικής και της τροποποιημένης προσέγγισης.



Εικόνα 3.11: Προσαρμογή νέας εκτέλεσης βασικού μοντέλου RF



Εικόνα 3.12: Προσαρμογή νέας εκτέλεσης τροποποιημένου μοντέλου RF

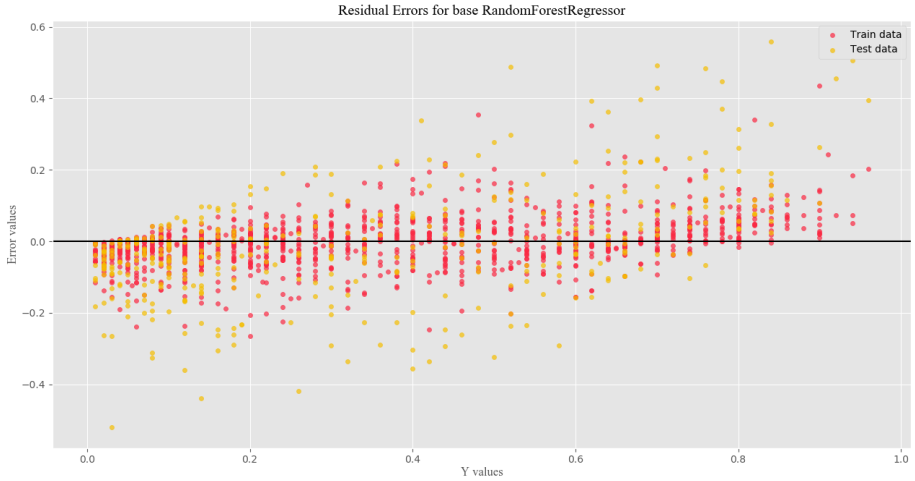
Από τα διαγράμματα της προσαρμογής προκύπτει μικρή βελτίωση του μοντέλου με τις μετρικές να δείχνουν ελαφρώς πιο σίγουρα αποτελέσματα και την κλίση της ευθείας να αυξάνει. Συγκεκριμένα, οι τιμές των μετρικών είναι

$$MSE_{base} = 0.0248 \ \& \ {R^2}_{base} = 0.6511$$
$$MSE_{best} = 0.0225 \ \& \ {R^2}_{best} = 0.6825$$

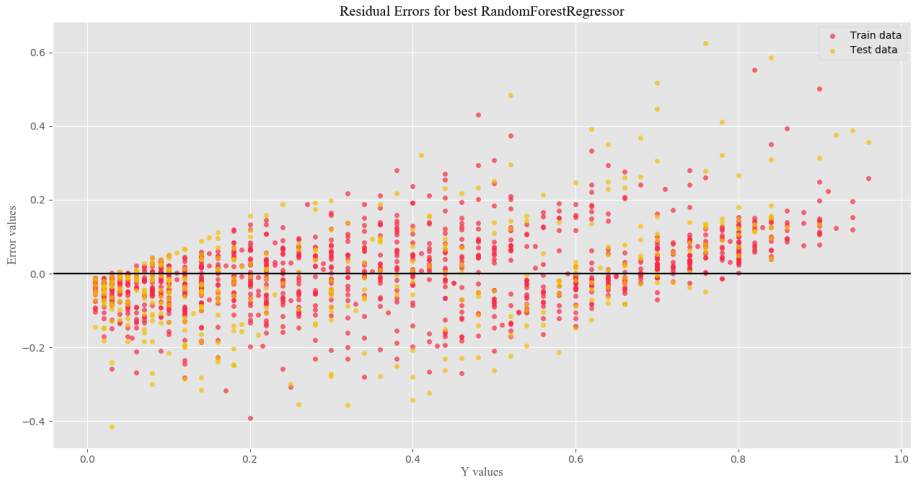
ενώ οι ευθείες είναι της μορφής

$$y_{base} = 0.7082x + 0.1053$$
$$y_{best} = 0.7219x + 0.0978\,.$$

Αντίστοιχα από τα διαγράμματα μελέτης της υπερεκπαίδευσης προκύπτουν οι παρακάτω δομές.

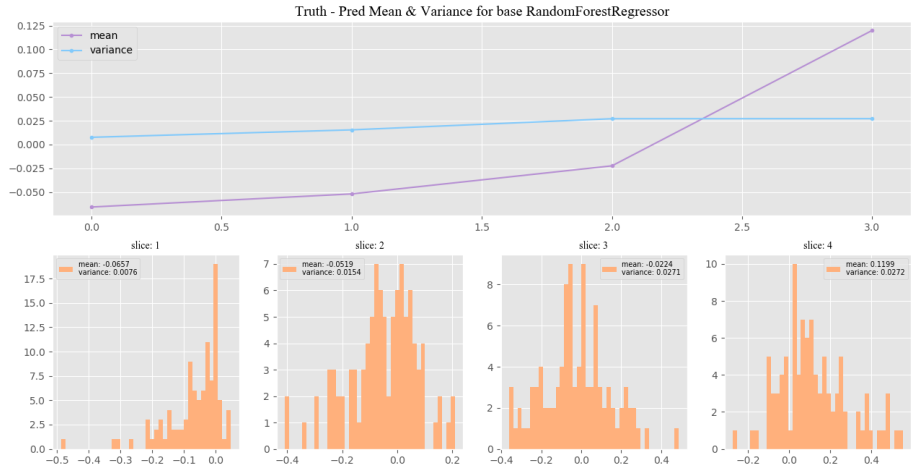


Εικόνα 3.13: Εκτίμηση σφαλμάτων νέας εκτέλεσης βασικού μοντέλου RF



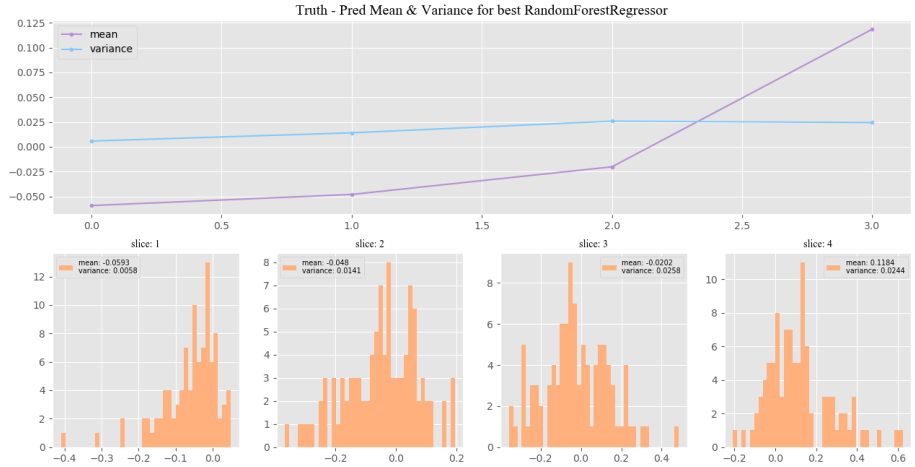
Εικόνα 3.14: Εκτίμηση σφαλμάτων νέας εκτέλεσης τροποποιημένου μοντέλου RF

Από τα παραπάνω διαγράμματα και σε σχέση με τη δεύτερη προσέγγιση, η εκπαίδευση του μοντέλου φαίνεται να είναι πιο επιτυχημένη. Δείχνει όμως, να έχει μία τάση προς το διαχωρισμό των τιμών όπως έγινε στη δεύτερη προσέγγιση.



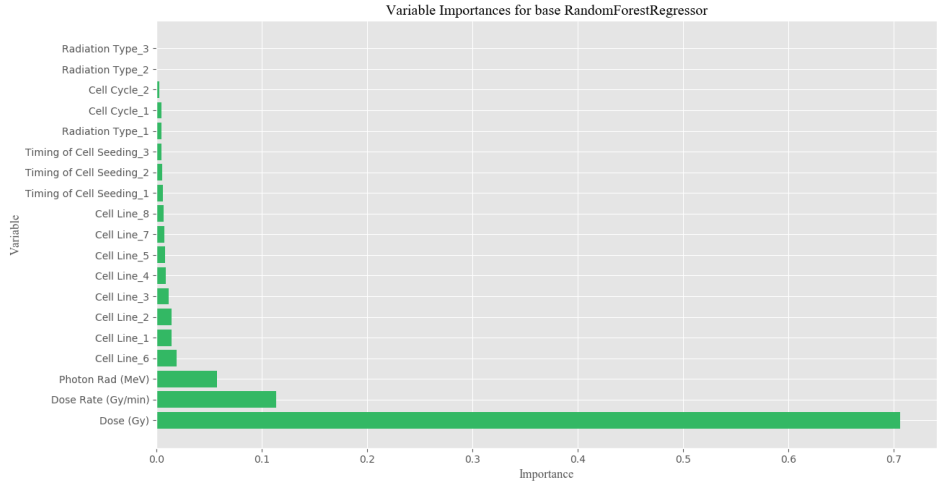
Εικόνα 3.15: Μελέτη κατανομής σφαλμάτων νέας εκτέλεσης βασικού μοντέλου RF

Στη στατιστική μελέτη που πραγματοποιείται στα διαγράμματα κατανομής σφαλμάτων, δεν παρατηρείται βελτίωση ή τροποποίηση στις τιμές του μέσου όρου των σφαλμάτων και της διακύμανσης.

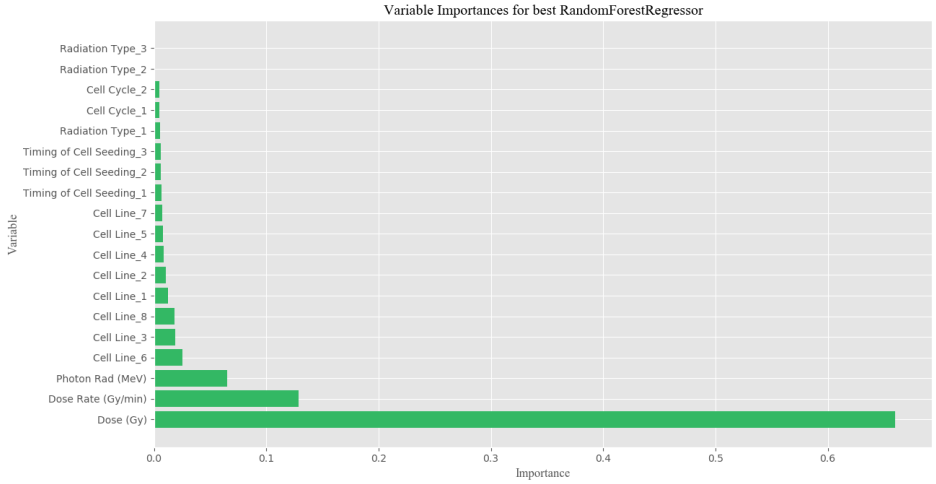


Εικόνα 3.16: Μελέτη κατανομής σφαλμάτων νέας εκτέλεσης τροποποιημένου μοντέλου RF

Μελετώντας τη σημαντικότητα των μεταβλητών διαπιστώνεται ελάττωση στην εξάρτηση από τη δόση που είναι η κυρίαρχη μεταβλητή κι αύξηση στο ποσοστό εξάρτησης του ρυθμού δόσης και της ενέργειας της ακτινοβολίας σύμφωνα με την εικόνα 3.18.



Εικόνα 3.17: Σημαντικότητα μεταβλητών νέας εκτέλεσης βασικού μοντέλου RF



Εικόνα 3.18: Σημαντικότητα μεταβλητών νέας εκτέλεσης τροποποιημένου μοντέλου RF

Σημαντικότητα μεταβλητών για *νέα εκτέλεση βασικού/ τροποποιημένου* μοντέλου Random Forest	
Dose (Gy)	0.71/ 0.66
Dose Rate (Gy/min)	0.11/ 0.13
Photon Rad (MeV)	0.06/ 0.07
Cell Line_6	0.02/ 0.03
Cell Line_1/ Cell Line_3	0.01/ 0.02
Cell Line_2/ Cell Line_8	0.01/ 0.02
Cell Line_3/ Cell Line_1	0.01
Cell Line_4/ Cell Line_2	0.01
Cell Line_5/ Cell Line_4	0.01
Cell Line_7/ Cell Line_5	0.01
Cell Line_8/ Cell Line_7	0.01
Timing of Cell Seeding_1	0.01
Timing of Cell Seeding_2	0.01
Timing of Cell Seeding_3	0.01
Radiation Type_1	0.01
Cell Cycle_1	0.00
Cell Cycle_2	0.00
Radiation Type_2	0.00
Radiation Type_3	0.00

3.3 Gradient Boosting

Η σχετική μελέτη πραγματοποιείται πάλι μέσω της βιβλιοθήκης Scikit-learn και συγκεκριμένα των μεθόδων που υλοποιεί ο αλγόριθμος Gradient Boosting με την εισαγωγή του στον κώδικα:

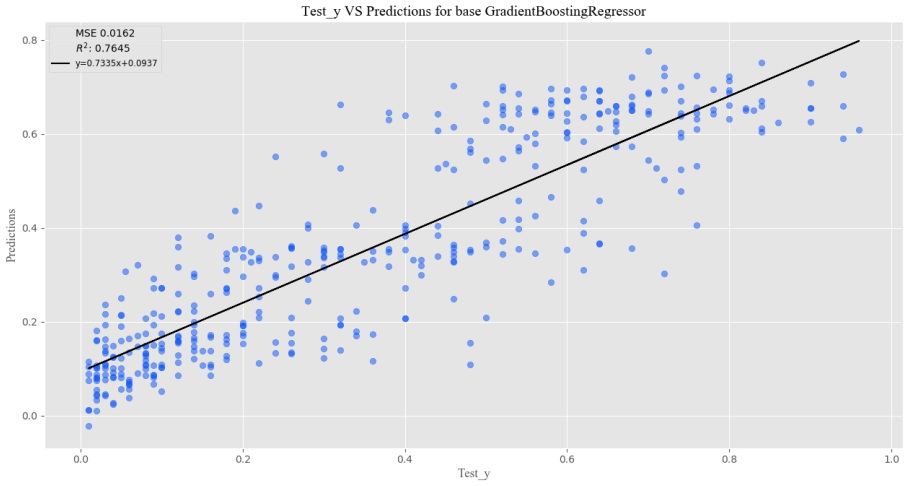
sklearn.ensemble.GradientBoostingRegressor

3.3.1 Πρώτη Προσέγγιση – Βασικό Μοντέλο

Ορίζεται το βασικό μοντέλο χρησιμοποιώντας τις προκαθορισμένους παραμέτρους:

('loss'= 'ls', 'criterion'= 'friedman_mse', 'learning_rate'= 0.1,
'max_depth'= 3, 'min_sample_leaf'= 1, 'min_samples_split'= 2,
'n_estimators'= 100, 'subsample'= 1.0)

Με την ίδια διαδικασία όπως πριν δημιουργείται το διάγραμμα προσαρμογής του βασικού μοντέλου με σκοπό να φανεί πόσο κοντά στην ευθεία y = x βρίσκονται τα ζευγάρια πραγματικών και προβλεπόμενων τιμών.



Εικόνα 3.19: Προσαρμογή βασικού μοντέλου GB

Οι τιμές από το 0.3 ως το 0.7 παρουσιάζουν μεγαλύτερη διασπορά από τις υπόλοιπες. Ωστόσο, η προσαρμογή συνολικά φαίνεται επιτυχής με την ευθεία της παλινδρόμησης να υπακούει στην εξίσωση

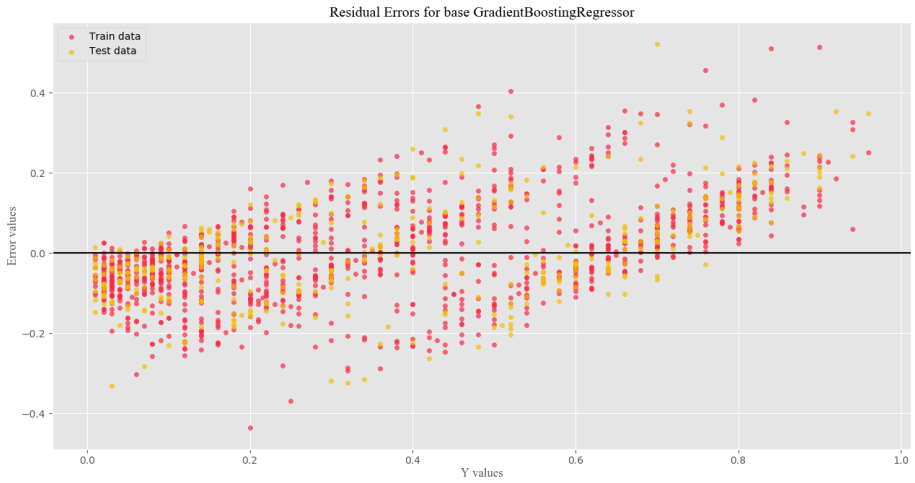
$$y = 0.7335x + 0.0937\,.$$

Οι μετρικές του διαγράμματος είναι

$$MSE = 0.0162 \ \& \ R^2 = 0.7645 .$$

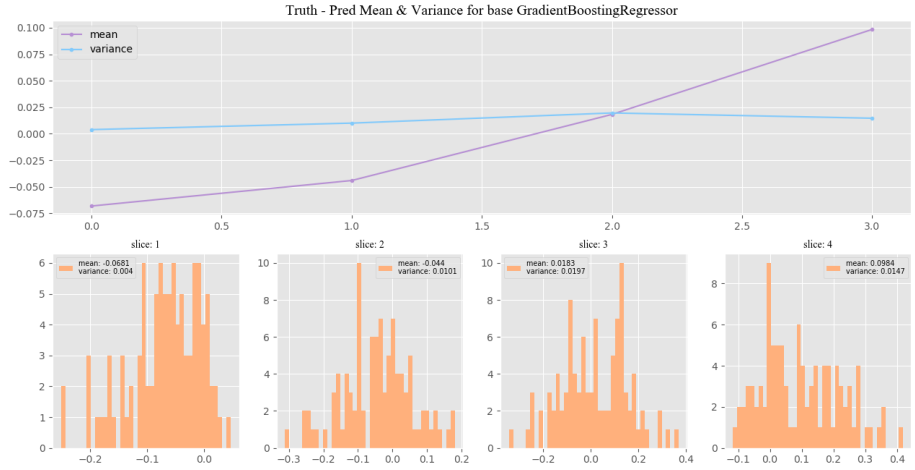
Η εικόνα 3.20 δίνει μία ιδέα για την υπερεκπαίδευση. Η δομή του ίσως θυμίζει εκείνη της δεύτερης προσέγγισης στο μοντέλο Random Forest αλλά είναι σίγουρα καλύτερη. Τα μέσα τετραγωνικά σφάλματα βρίσκονται στις τιμές

$$MSE_{Train} = 0.0141 \ \& \ MSE_{Test} = 0.0157 .$$



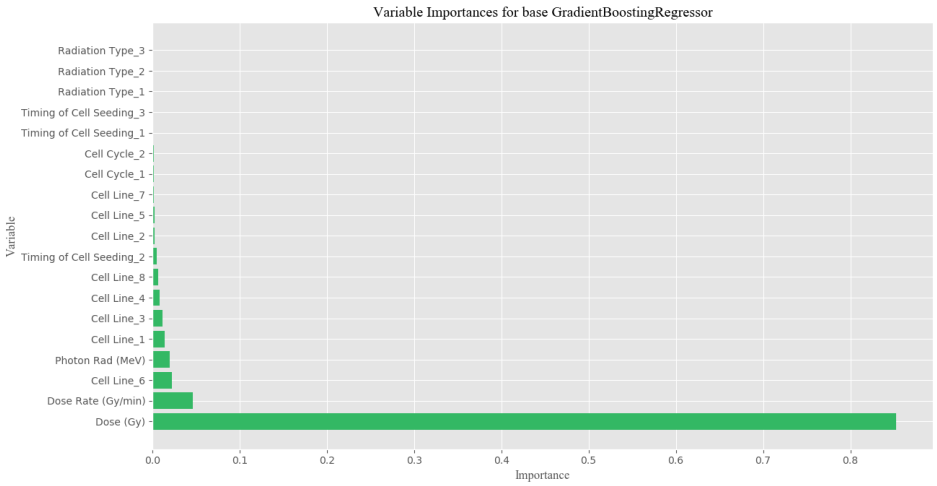
Εικόνα 3.20: Εκτίμηση σφαλμάτων βασικού μοντέλου GB

Υπάρχει ομοιομορφία στην κατανομή των σημείων των συνόλων εκπαίδευσης και δοκιμής αλλά ταυτόχρονα φαίνεται ένα μοτίβο στην κατανομή των σφαλμάτων, το οποίο επιβεβαιώνεται από το επόμενο διάγραμμα.



Εικόνα 3.21: Μελέτη κατανομής σφαλμάτων βασικού μοντέλου GB

Ο μέσος όρος των σφαλμάτων αποκτά μεγάλο εύρος αφού συναντάται περίπου στο ±0.1. Οι περισσότερες τιμές όμως απ' όσο φαίνεται στα επιμέρους διαγράμματα βρίσκονται γύρω από το ±0.05.



Εικόνα 3.22: Σημαντικότητα μεταβλητών βασικού μοντέλου GB

Η σημαντικότητα των μεταβλητών όπως φαίνεται στην εικόνα 3.22 παρατίθεται στον παρακάτω πίνακα.

Σημαντικότητα μεταβλητών για *βασικό* μοντέλο Gradient Boosting			
Dose (Gy)	0.85	**Cell Line_5**	0.00
Dose Rate (Gy/min)	0.05	**Cell Line_7**	0.00
Cell Line_6	0.02	**Cell Cycle_1**	0.00
Photon Rad (MeV)	0.02	**Cell Cycle_2**	0.00
Cell Line_1	0.01	**Timing of Cell Seeding_1**	0.00
Cell Line_3	0.01	**Timing of Cell Seeding_3**	0.00
Cell Line_4	0.01	**Radiation Type_1**	0.00
Cell Line_8	0.01	**Radiation Type_2**	0.00
Timing of Cell Seeding_2	0.01	**Radiation Type_3**	0.00
Cell Line_2	0.01		

Η συμβολή της δόσης φαίνεται να είναι πολύ ισχυρότερη σε σχέση με το Random Forest σε ποσοστό 85% ενώ ακολουθούν ο ρυθμός δόσης με 5% κι η κυτταρική σειρά 6 του καρκίνου παχέος εντέρου HCT-116.

3.3.2 Δεύτερη Προσέγγιση – Νέο Μοντέλο

Δοκιμάζεται να κατασκευαστεί πλέγμα παραμέτρων για το δεδομένο αλγόριθμο, για να διαπιστωθεί αν μπορεί να βελτιωθεί η απόδοσή του. Η εισαγωγή του πλέγματος γίνεται μέσω του Scikit-learn:

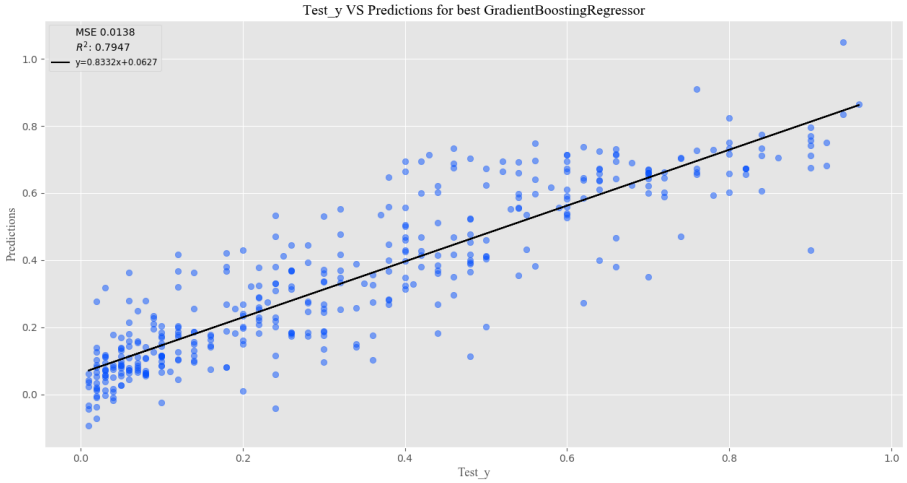
sklearn.model_selection.GridSearchCV

Οι παράμετροι που εξετάζονται έχουν περιγραφεί στο Κεφάλαιο 2, οπότε το πλέγμα έχει ως εξής:

Κριτήριο διαχωρισμού	friedman_mse, mae, mse
Ρυθμός εκμάθησης	0.01, 0.05, 0.09
Βάθος δέντρου απόφασης	None, 3, 6
Αριθμός ελάχιστων δειγμάτων κόμβου φύλλου	1, 2, 3
Αριθμός ελάχιστων δειγμάτων για διαχωρισμό	2, 5, 10
Αριθμός επαναλήψεων	100, 500, 1000
Κλάσμα δειγμάτων	0.9, 1.0

Όπως αναφέρθηκε και πριν, μέσω της διαδικασίας cross-validation που αξιολογεί τους συνδυασμούς των παραμέτρων 5 φορές σε διαφορετικό υποσύνολο των δεδομένων κάθε φορά, προκύπτει ο συνδυασμός των παραμέτρων που βελτιστοποιεί την ακρίβεια του αλγορίθμου. Ο συνδυασμός που δίνει τη βελτιωμένη εικόνα 3.23 είναι ο παρακάτω:

('loss'= 'ls', 'criterion'= 'mse', 'learning_rate'= 0.05,
'max_depth'= 3, 'min_samples_leaf'= 1, 'min_samples_split'= 2,
'n_estimators'= 1000, 'subsample'= 0.9)



Εικόνα 3.23: Προσαρμογή νέου μοντέλου GB

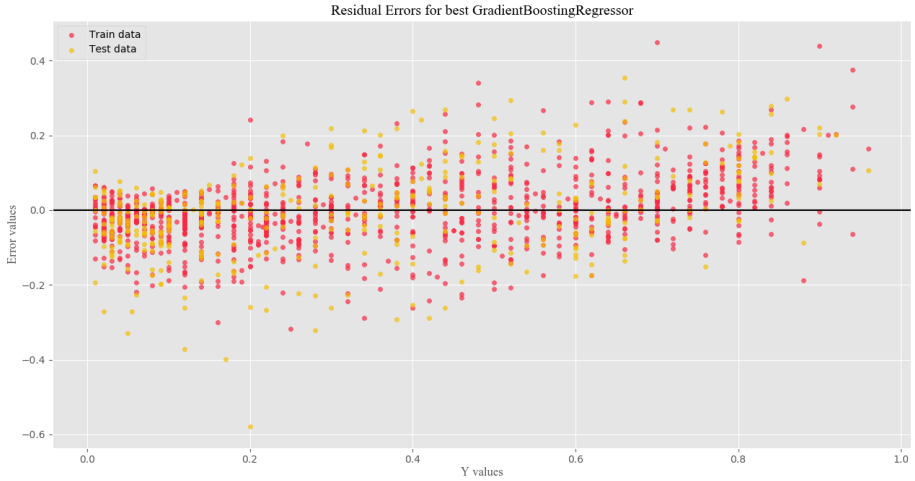
Σε σχέση με το βασικό μοντέλο, τα ζεύγη τιμών πλησιάζουν φανερά περισσότερο την ευθεία που έχει εξίσωση

$$y = 0.8332x + 0.0627$$

με αποτέλεσμα να τείνουν περισσότερο στην ευθεία y = x. Επίσης, από τις μετρικές του διαγράμματος που είναι

$$MSE = 0.0138 \ \& \ R^2 = 0.7947$$

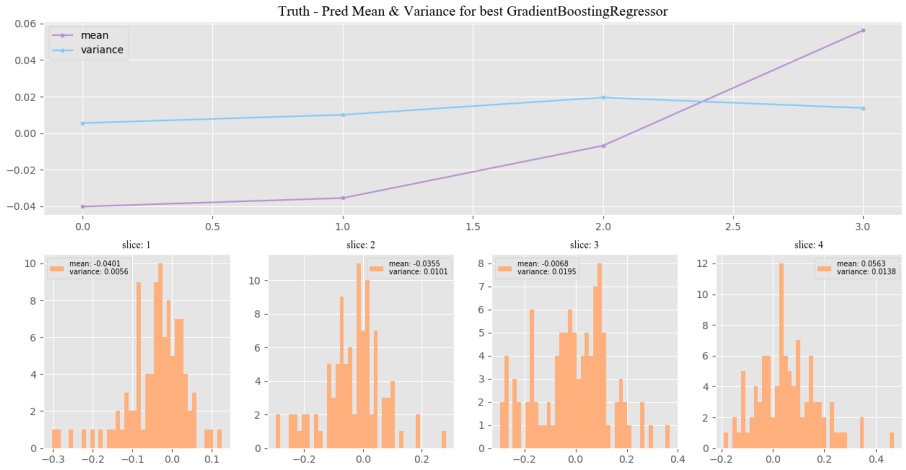
φαίνεται ότι το μέσο τετραγωνικό σφάλμα ελαττώνεται, ενώ το R^2 αυξάνεται όπως είναι το επιθυμητό.



Εικόνα 3.24: Εκτίμηση σφαλμάτων νέου μοντέλου GB

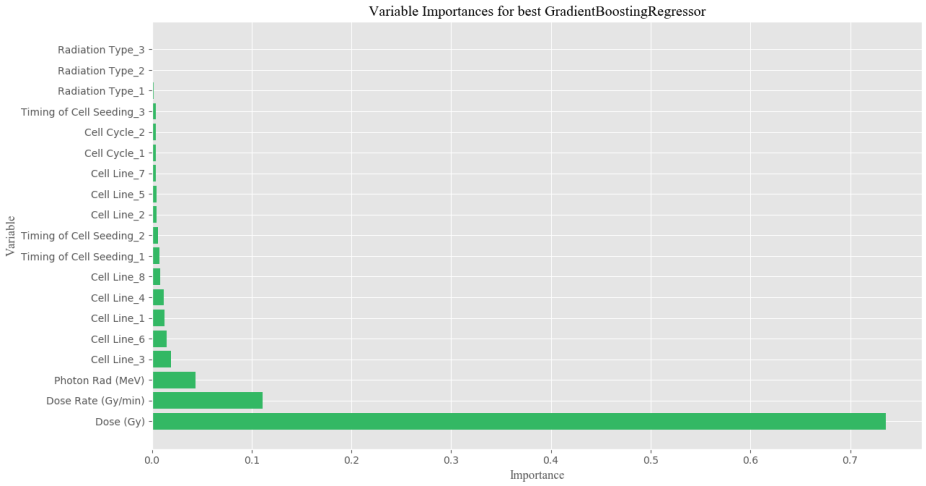
Η εικόνα των σφαλμάτων φαίνεται πιο ομοιόμορφη και βελτιωμένη σε σχέση με το βασικό μοντέλο με τις τιμές των μέσων τετραγωνικών σφαλμάτων να είναι:

$$MSE_{Train} = 0.0078 \ \& \ MSE_{Test} = 0.0152 .$$



Εικόνα 3.25: Μελέτη κατανομής σφαλμάτων νέου μοντέλου GB

Σύμφωνα με την εικόνα 3.25 το εύρος του μέσου όρου των σφαλμάτων ελαττώνεται και κυμαίνεται μεταξύ του -0.04 έως το 0.06, πράγμα που επιβεβαιώνεται κι από τα επιμέρους διαγράμματα.



Εικόνα 3.26: Σημαντικότητα μεταβλητών νέου μοντέλου GB

Σημαντικότητα μεταβλητών για *νέο* μοντέλο Gradient Boosting			
Dose (Gy)	0.74	**Cell Line_2**	0.00
Dose Rate (Gy/min)	0.11	**Cell Line_5**	0.00
Photon Rad (MeV)	0.04	**Cell Line_7**	0.00
Cell Line_3	0.02	**Cell Cycle_1**	0.00
Cell Line_6	0.02	**Cell Cycle_2**	0.00
Cell Line_1	0.01	**Timing of Cell Seeding_3**	0.00
Cell Line_4	0.01	**Radiation Type_1**	0.00
Cell Line_8	0.01	**Radiation Type_2**	0.00
Timing of Cell Seeding_1	0.01	**Radiation Type_3**	0.00
Timing of Cell Seeding_2	0.01		

Η σημαντικότητα των μεταβλητών τροποποιείται λίγο καθώς η σειρά τους μένει πάνω κάτω σταθερή αλλά η συμβολή της δόσης ελαττώνεται και διαμοιράζεται στις υπόλοιπες μεταβλητές.

3.4 Σύγκριση Μοντέλων

Μετά το πέρας της μελέτης κάθε μοντέλου, πραγματοποιείται σύγκριση των αποτελεσμάτων τους για να διαπιστωθεί ποιος αλγόριθμος εξυπηρετεί καλύτερα τις ανάγκες του προβλήματος. Καλύτερο θεωρείται το μοντέλο που παρουσιάζει τα χαμηλότερα σφάλματα κατά την παλινδρόμηση.

Η σύγκριση πραγματοποιείται μελετώντας τα σφάλματα Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) και τη μετρική R^2 της παλινδρόμησης. Υπενθυμίζεται ότι κάθε μοντέλο εκτελέστηκε 10 φορές για να καταγραφεί η συμπεριφορά του.

RMSE RF Evaluation

# of run	base	adjusted	improvement (%)
1.0	0,149	0,141	4,845
2.0	0,140	0,136	2,716
3.0	0,145	0,141	2,692
4.0	0,135	0,125	7,413
5.0	0,137	0,128	6,715
6.0	0,128	0,131	-2,259
7.0	0,148	0,141	4,459
8.0	0,144	0,141	1,739
9.0	0,138	0,131	5,278
10.0	0,140	0,135	3,438
Mean Value	**0,140**	**0,135**	**3,704**

RMSE GB Evaluation

# of run	base	adjusted	improvement (%)
1.0	0,136	0,131	3,965
2.0	0,127	0,117	7,820
3.0	0,131	0,122	7,246
4.0	0,118	0,115	2,620
5.0	0,121	0,118	2,805
6.0	0,124	0,114	7,977
7.0	0,134	0,124	7,485
8.0	0,129	0,123	4,570
9.0	0,125	0,118	5,449
10.0	0,128	0,117	9,048
Mean Value	**0,127**	**0,120**	**5,899**

Όπως έχει αναφερθεί το Root Mean Square Error βρίσκεται στις «μονάδες» της μεταβλητής εξόδου. Συγκρίνοντας τις μετρικές των δύο αλγορίθμων παρατηρούνται τα εξής:

- Ο αλγόριθμος Random Forest παρουσιάζει μερικώς μεγαλύτερο σφάλμα στην πρόβλεψη σε σχέση με το Gradient Boosting.
- Επίσης τα ποσοστά βελτίωσης έχουν μεγάλο εύρος, οπότε δεν είναι βέβαιο αν μπορεί κανείς να βασιστεί σε αυτά.

MAPE RF Evaluation

# of run	base	adjusted	improvement (%)
1.0	75,58	69,38	8,203
2.0	85,06	88,07	-3,539
3.0	90,72	94,11	-3,737
4.0	76,15	76,55	-0,525
5.0	81,11	80,45	0,814
6.0	80,83	88,38	-9,341
7.0	90,88	79,32	12,720
8.0	83,46	70,55	15,468
9.0	81,34	80,77	0,701
10.0	83,29	88,71	-6,507
Mean Value	**82,84**	**81,63**	**1,426**

MAPE GB Evaluation

# of run	base	adjusted	improvement (%)
1.0	75,81	71,82	5,263
2.0	75,07	67,62	9,924
3.0	78,21	66,27	15,267
4.0	75,24	68,46	9,011
5.0	68,10	63,28	7,078
6.0	79,04	70,68	10,577
7.0	82,96	69,63	16,068
8.0	76,05	71,02	6,614
9.0	72,64	57,71	20,553
10.0	78,08	64,74	17,085
Mean Value	**76,12**	**67,12**	**11,744**

Το Mean Absolute Percentage Error είναι καθαρό ποσοστό το οποίο εκφράζει την απόσταση μεταξύ πραγματικής και προβλεπόμενης τιμής κατ' απόλυτη τιμή. Για τους δύο αλγορίθμους παρατηρείται ότι:

- Αρχικά κι οι 2 φέρουν μεγάλο σφάλμα άρα πιθανώς να μην μπορούν να χρησιμοποιηθούν με σιγουριά για το δεδομένο πρόβλημα.
- Ο Random Forest έχει συστηματικά μεγαλύτερο σφάλμα από τον Gradient Boosting, ενώ στις μισές σχεδόν εκτελέσεις το MAPE αντί να μειώνεται, αυξάνεται. Ως αποτέλεσμα, ο Random Forest δε δείχνει συνέπεια, καθώς μέσα στις εκτελέσεις βγαίνουν αποτελέσματα που δείχνουν ξεκάθαρη χειροτέρευση του μοντέλου.

R^2 RF Evaluation

# of run	base	adjusted	improvement (%)
1.0	0,66	0,69	5,023
2.0	0,73	0,75	1,969
3.0	0,70	0,71	2,331
4.0	0,72	0,76	5,469
5.0	0,72	0,75	5,115
6.0	0,75	0,74	-1,507
7.0	0,70	0,73	3,748
8.0	0,72	0,73	1,345
9.0	0,72	0,75	3,964
10.0	0,71	0,73	2,779
Mean Value	**0,71**	**0,73**	**3,023**

R^2 GB Evaluation

# of run	base	adjusted	improvement (%)
1.0	0,71	0,73	3,210
2.0	0,78	0,81	4,256
3.0	0,75	0,79	4,638
4.0	0,79	0,80	1,424
5.0	0,78	0,79	1,541
6.0	0,78	0,80	3,298
7.0	0,75	0,79	4,691
8.0	0,75	0,77	3,064
9.0	0,77	0,80	3,143
10.0	0,75	0,80	5,605
Mean Value	**0,76**	**0,79**	**3,487**

Η μετρική R^2 αξιολογεί το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που επεξηγείται από τις ανεξάρτητες. Συγκρίνοντας τις μετρικές των δύο αλγορίθμων βγαίνουν τα εξής συμπεράσματα:

- Ο Gradient Boosting αποδίδει συστηματικά ελαφρώς μεγαλύτερες τιμές.
- Ο Random Forest και σε αυτή τη μετρική εμφανίζει διασπορά στις τιμές που δίνει σε κάθε εκτέλεση.
- Ως μέσος όρος, η συνολική βελτίωση που φαινομενικά παρατηρείται στους αλγορίθμους είναι παρόμοια, της τάξης του 3-3.5%.

4 Συμπεράσματα

Στην τελική ενότητα γίνεται σύνοψη των αποτελεσμάτων κι εξάγονται κάποια κατευθυντήρια συμπεράσματα. Επιπλέον, σχολιάζονται οι περιορισμοί που υπήρχαν και συζητιούνται μελλοντικές κατευθύνσεις που μπορούν να ακολουθηθούν.

4.1 Συζήτηση Αποτελεσμάτων

Λαμβάνοντας υπόψη μόνο τα στατιστικά στοιχεία φαίνεται ότι και τα δύο μοντέλα καταφέρνουν να βελτιώσουν τους εκτιμητές τους μέσω του συντονισμού των υπερπαραμέτρων έστω και λίγο. Κοιτάζοντας όμως τα διαγράμματα των σφαλμάτων και τη διασπορά των μετρικών, διαπιστώνεται ότι στην πραγματικότητα δε συμβαίνει αυτό.

Το μοντέλο που έχει ως βάση τον αλγόριθμο Random Forest δεν καταφέρνει να βελτιώσει στην πραγματικότητα την απόδοσή του μέσω των υπερπαραμέτρων. Αυτό φαίνεται και στη λήψη των διαγραμμάτων του, που εμφανίζουν ιδιάζουσες δομές αλλά και στα σφάλματά του, τα οποία εμφανίζουν πότε αύξουσα και πότε φθίνουσα εξέλιξη. Το φαινόμενο αυτό είναι ιδιαίτερα έντονο στον υπολογισμό του MAPE, ενώ στις άλλες μετρικές δεν είναι προφανές. Το βασικό μοντέλο που εκτελείται με τις θεμελιώδεις παραμέτρους δίνει καλύτερα αποτελέσματα έχοντας ίσως μια τάση προς την υπερεκπαίδευση αφού το μέσο τετραγωνικό σφάλμα του συνόλου εκπαίδευσης βγαίνει αρκετά μικρότερο απ' το αντίστοιχο του συνόλου δοκιμής. Ωστόσο, αυτό χρήζει περισσότερης μελέτης. Στην τρίτη προσέγγιση που τροποποιούνται οι τιμές της παραμέτρου βάθος δέντρου απόφασης, τα διαγράμματα βρίσκονται σε μία ενδιάμεση κατάσταση σε σχέση με τις δύο άλλες προσεγγίσεις. Το μοντέλο δείχνει να βελτιώνεται ελάχιστα σε σχέση με το βασικό ενώ δεν καταλήγει στη δημιουργία περίεργων δομών στη μελέτη της υπερεκπαίδευσης. Επομένως, το μοντέλο της πρώτης προσέγγισης μέσω Random Forest κρίνεται ότι μπορεί να ανταπεξέρθει στο πρόβλημα. Οι άλλες δύο προσεγγίσεις πρέπει να μελετηθούν περισσότερο ώστε να εξαχθούν πιο σίγουρα συμπεράσματα.

Ως προς το μοντέλο που έχει κατασκευαστεί μέσω του αλγορίθμου Gradient Boosting, τα αποτελέσματα δείχνουν πιο ιδανικά. Από τα διαγράμματα της προσαρμογής φαίνεται ότι καταφέρνει να βελτιώσει τις προβλέψεις έναντι των πραγματικών τιμών, ενώ βελτιώνεται κι η κατανομή των σφαλμάτων του. Όπως και στον Random Forest υπάρχει μία ελάττωση του μέσου τετραγωνικού σφάλματος του συνόλου εκπαίδευσης οπότε ίσως κι εδώ να τείνει να πραγματοποιηθεί υπερεκπαίδευση του αλγορίθμου. Κοιτάζοντας τα σφάλματα και τη μετρική R^2 διαπιστώνεται ότι τα σφάλματα έχουν μικρότερη διασπορά και συστηματικότητα στις τιμές τους, που σημαίνει ότι ο αλγόριθμος βγάζει σταθερά αποτελέσματα. Αν επιβεβαιωθεί λοιπόν ότι δεν

υπερεκπαιδεύεται το μοντέλο, τότε το μοντέλο Gradient Boosting μπορεί να χρησιμοποιηθεί για την αντιμετώπιση του δεδομένου προβλήματος.

Ως προς τη σημαντικότητα των μεταβλητών υπάρχει πλήρης σταθερότητα σε όλες τις προσεγγίσεις. Από τις 19 μεταβλητές που έχει το σύνολο δεδομένων μετά την προ-επεξεργασία, δύο ξεχωρίζουν ενώ οι υπόλοιπες έχουν ελάχιστη συνεισφορά. Η μεταβλητή της δόσης είναι μακράν η σημαντικότερη σε ποσοστό 70-80% ανάλογα την εκτέλεση. Επίσης, σταθερή είναι κι η μεταβλητή του ρυθμού δόσης που έρχεται μόνιμα στη δεύτερη θέση με σημαντικότητα 5-12%. Από κει και πέρα ακολουθούν κάποιες από τις καρκινικές σειρές ή η ενέργεια της ακτινοβολίας που έχει χρησιμοποιηθεί στα πειράματα. Τέλος, πρέπει να ληφθεί υπόψη ότι η σημαντικότητα της μεταβλητής της καρκινικής σειράς μοιράζεται, γιατί κάθε μία λογίζεται ως ξεχωριστή μεταβλητή. Αν οι σημαντικότητες αθροιστούν, προκύπτει η συμβολή της μεταβλητής σε ποσοστό γύρω στο 5-8%.

4.2 Περιορισμοί

Η διαδικασία συγκερασμού των βιολογικών κι υπολογιστικών στοιχείων φέρει κάποιους φυσικούς περιορισμούς, οι οποίοι έρχονται να προστεθούν στους ήδη υπάρχοντες του συγκεκριμένου προβλήματος. Από τη μία, σημαντικό είναι το γεγονός ότι η μεταβλητή εξόδου προέρχεται από πειραματικά δεδομένα τα οποία δεν είναι δυνατόν να αναπαραχθούν με πλήρη ακρίβεια ακόμα κι αν το πείραμα επαναληφθεί. Δηλαδή κατά τη διεξαγωγή του πειράματος από άλλο εργαστήριο άλλη μέρα, τα κλάσματα κυτταρικής επιβίωσης κατά πάσα πιθανότητα δε θα είναι τα ίδια, καθώς οι βιολογικές διαδικασίες χαρακτηρίζονται από στοχαστικότητα [35]. Εκτός αυτού, υπάρχουν μικροπαράγοντες που ακόμα κι αν το πρωτόκολλο του πειράματος ακολουθηθεί πλήρως, μπορεί να επηρεάσουν έστω και λίγο τις μετρήσεις που θα ληφθούν. Ακόμα και στις δημοσιεύσεις από τις οποίες προέρχονται τα δεδομένα, συχνά αναφερόταν ότι τα πειράματα πραγματοποιήθηκαν 2-3 φορές για να καταγραφεί η μέση συμπεριφορά της επίδρασης της ακτινοβολίας στα κύτταρα. Από την άλλη, ο υπολογιστικός παράγοντας σχετίζεται με τη δύναμη του υπολογιστή και το χρόνο που απαιτήθηκε για να γίνει μελέτη των υπερπαραμέτρων. Για κάθε εκτέλεση του ολικού τελικού αλγορίθμου χρειάστηκαν κατά μέσο όρο 6.30 ώρες όποτε δεν ήταν δυνατό να διερευνηθούν πιο συγκεκριμένες τιμές των υπερπαραμέτρων.

Οι περιορισμοί του προβλήματος προκύπτουν από το ίδιο το σύνολο των δεδομένων. Ο αριθμός των αρχικών παρατηρήσεων που διαχώριζαν τα κλάσματα επιβίωσης με τη δόση ανερχόταν μόλις στις 566. Με την ιδέα ενοποίησης όλων των κλασμάτων επιβίωσης και την τοποθέτηση της δόσης ως μεταβλητή, ο αριθμός των παρατηρήσεων πήγε στις 1595. Πιθανότατα, ούτε αυτός ο αριθμός παρατηρήσεων αρκεί για να εξαχθούν βέβαια και σημαντικά συμπεράσματα.

Παρά τους περιορισμούς που αναφέρθηκαν, εξακολουθεί να ισχύει το γεγονός ότι κάποια από τα μοντέλα είναι σε θέση να ερμηνεύσουν τη συμπεριφορά των κυττάρων μετά την ακτινοβόληση σε ικανοποιητικό βαθμό. Με πιθανές προσθήκες ή τροποποιήσεις στις προσεγγίσεις που χρησιμοποιούνται, τα αποτελέσματα μπορούν να βελτιωθούν.

4.3 Βελτιώσεις – Μελλοντικές Κατευθύνσεις

Η βελτίωση του υπάρχοντος μοντέλου σίγουρα μπορεί να ωφελήσει, καθώς θα μπορεί να πλαισιώσει καλύτερα τη μελέτη της κυτταρικής επιβίωσης από την ιοντίζουσα ακτινοβολία. Για να συμβεί αυτό, μπορούν μελλοντικά να γίνουν τα ακόλουθα:

- Αύξηση/ εμπλουτισμός του συνόλου δεδομένων ώστε οι αλγόριθμοι να μπορούν να εξάγουν σχέσεις εξάρτησης μεταξύ των μεταβλητών.
- Επιπλέον μελέτη των υψηλών δόσεων οι οποίες έδιναν τις λιγότερες μετρήσεις στο σύνολο δεδομένων.
- Διαφοροποίηση σε παράγοντες όπως ο κυτταρικός κύκλος για να παρατηρηθούν τυχόν διαφορές στην επιβίωση.
- Ενδελεχέστερη μελέτη ως προς τις παραμέτρους και την απόδοση των αλγορίθμων που χρησιμοποιούνται.
- Πιθανή επέκταση και στη μελέτη επίδρασης της σωματιδιακής ακτινοβολίας με κάποιες αλλαγές.

Βιβλιογραφία

1. Richter, D., *Treatment planning for tumors with residual motion in scanned ion beam therapy*. 2012, Technical University of Darmstadt.
2. Stewart, J., *X-ray Imaging*, Compton, Editor. 2019.
3. Rhys, H., *Machine Learning with R, the tidyverse and mlr*. 2020, Manning.
4. Stark, G. *X-ray radiation beam*.
5. *The Electromagnetic Spectrum*. Available from: https://courses.lumenlearning.com/boundless-physics/chapter/the-electromagnetic-spectrum/.
6. Goaz, P., *Production of X-rays and Interactions of X-rays with Matter*. p. 11-20.
7. Wikipedia. *Linear Particle Acceleration*. Available from: https://en.wikipedia.org/wiki/Linear_particle_accelerator.
8. Wikipedia. *Cobalt Therapy*. Available from: https://en.wikipedia.org/wiki/Cobalt_therapy.
9. *Radioisotope Brief: Cesium-137 (Cs-137)*. 2018; Available from: https://www.cdc.gov/nceh/radiation/emergencies/isotopes/cesium.htm.
10. *Radiation Biology: A Handbook for Teachers and Students*. 2010, Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY.
11. ΓΕΩΡΓΙΟΥ, Ε., *ΙΑΤΡΙΚΗ ΦΥΣΙΚΗ*. 2014: ΠΑΣΧΑΛΙΔΗΣ. 578.
12. Hall, E. and A. Giaccia, *Radiobiology for the Radiologist* 7th ed. 2012: Wolters Kluwer Health. 556.
13. Wikipedia. *Machine Learning*. Available from: https://en.wikipedia.org/wiki/Machine_learning.
14. Matsui, T., et al., *Robustness of Clonogenic Assays as a Biomarker for Cancer Cell Radiosensitivity*. Int J Mol Sci, 2019. **20**(17).
15. Friedrich, T., et al., *Systematic analysis of RBE and related quantities using a database of cell survival experiments with ion beam irradiation*. J Radiat Res, 2013. **54**(3): p. 494-514.
16. *Ensemble Methods*. Available from: https://scikit-learn.org/stable/modules/ensemble.html.
17. Nagpal, A. *Decision Tree Ensembles - Bagging and Boosting*. 2017.
18. Gurucharan, M.K. *Machine Learning Basics: Decision Tree Regression*. 2020.
19. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2nd ed. 2009: Springer.
20. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**: p. 5-32.
21. Biau, G. and E. Scornet, *A Random Forest Guided Tour*. TEST, 2016. **25**: p. 197-227.
22. Emtiyaz, K., *Random Forests*. 2015.
23. Brownlee, J. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. 2016.
24. Friedman, J., *Stochastic gradient boosting*. Computational Statistics & Data Analysis, 2002. **38**: p. 367-378.
25. Pascual, C., *Tutorial: Understanding Regression Error Metrics in Python*. 2018.
26. Hayes, A., *R-Squared Definition*. 2020.
27. Koehrsen, W. *Hyperparameter tuning the Random Forest in Python*. 2018.

28. Amatriain, X. *When would one use Random Forests over Gradient Boosted Machines (GBMs)?* 2015; Available from: https://www.quora.com/When-would-one-use-Random-Forests-over-Gradient-Boosted-Machines-GBMs.
29. Ravanshad, A. *Gradient Boosting vs Random Forest*. 2018.
30. Koehrsen, W. *Random Forest in Python*. 2017.
31. Nesmiyanova, A. *What is Python & Django and why are they considered a top choice for web development?* ; Available from: https://steelkiwi.com/blog/why-python-django-are-your-top-choice-for-web-development/.
32. Beklemysheva, A. *Why use Python for AI and Machine Learning?* ; Available from: https://steelkiwi.com/blog/python-for-ai-and-machine-learning/.
33. Borreli, L. *Testicular Cancer and 6 Other Curable Cancers with the best 5-Year Survival Rate*. 2016.
34. *5 Curable Cancers*.
35. Casadevall, A. and F.C. Fang, *Reproducible science.* Infect Immun, 2010. **78**(12): p. 4972-5.