Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Στατιστικής

Αριθμός Μητρώου:

Κωνσταντίνα Ζήκου

Ιανουάριος 2023

Contents

[1. Εγκατάσταση της Python και των απαραίτητων πακέτων 5](#_Toc124632718)

[2. Πρόκληση και στόχος της παρούσας εργασίας 6](#_Toc124632719)

[3. Συσχέτιση 7](#_Toc124632720)

[4. Stratified Data 8](#_Toc124632721)

[5. Εκτιμητές 10](#_Toc124632722)

[6. Εκπαίδευση του συνόλου δεδομένων μας 11](#_Toc124632723)

[7. Οδηγίες 12](#_Toc124632724)

Εικόνες

[Εικόνα 1Σχήμα συσχέτισης των μεταβλητών 6](#_Toc124626889)

[Εικόνα 2Ιστόγραμμα για τις μεταβλητές Survived, Pclass 8](#_Toc124626890)

Πίνακες

Table 1Ο πίνακας δείχνει τον αιρθμών των διαθέσιμων δεδομένων

Table 2 Αποτέλεσμα ακρίβειας μοντέλου μηχανικής μάθησης

# Εγκατάσταση της Python και των απαραίτητων πακέτων

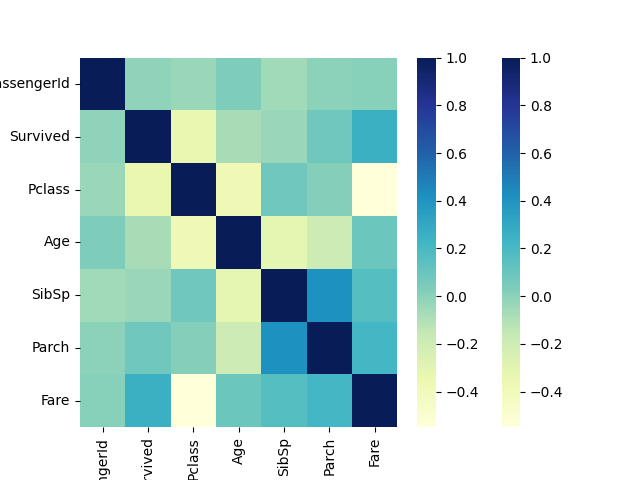
Αρχικά, έγινε εγκατάσταση του περιβάλλοντος της Python 3.11. σε λειτουργικό σύστημα Windows. Μετέπειτα, έγινε εγκατάσταση των πακέτων numpy, matplotlib, pandas, scikit-learn, seaborn. Όσον αφορά το numpy, έγινε εγκατάσταση για να καταφέρουμε να κάνουμε διαχείριση των πινάκων και γενικότερα του συνόλου δεδομένων στην Python. Όσον αφορά το matplotlib έγινε εγκατάσταση για να αναπαραστήσουμε τα δεδομένα μας και να κάνουμε περιγραφική στατιστική. Όσον αφορά το pandas έγινε εγκατάσταση για να φορτώσουμε το σύνολο δεδομένων μας στη δομή δεδομένων όπου παρέχει το pandas. Όσον αφορά το scikit-learn ;έγινε εγκατάσταση για να θέσουμε τυχόν κανονικοποιητές στο μοντέλο μας όπως το StandardScale. Όσον αφορά το seasborn είναι μια βιβλιοθήκη η οποία έχει απώτερο σκοπό την αναπαράσταση των δεδομένων.

# Πρόκληση και στόχος της παρούσας εργασίας

Η βύθιση του Τιτανικού το 1912 ήταν ένας από τα πιο κραυγαλέα και εντυπωσιακά γεγονότα στην ιστορία. Εντυπωσιακά με την έννοια όταν άφησε ιστορία τόσο στο γεγονός ότι βυθίστηκε για την τότε εποχή μεγαλύτερο κρουζιερόπλοιο όσο για τον θάνατο χιλιάδων ανθρώπων. Σύμφωνα με τα σημερινά δεδομένα έχασαν τη ζωή τους 1502 επιβάτες από τους 2224. Αυτός ο αριθμός είναι αξιοσημείωτος μιας και που χάθηκες τόσες ζωές από τη σύγκρουση του πλοίου με το παγόβουνο. Σύμφωνα με τα στοιχεία που έχουμε μέχρι σήμερα, ορισμένες ομάδες ανθρώπων ήταν περισσότερο τυχερές να επιβιώσουν με βάση το κύρος και το κοινωνικό στρώμα στο οποίο ανήκαν. Σε αυτή την πρόκληση λοιπόν στόχος είναι η δημιουργία ενός μοντέλου μηχανικής μάθησης το οποίο θα έχει ως στόχο να μας απαντά στην ερώτηση ποια είδη ανθρώπων ήταν πιο πιθανό να επιβιώσουν. Έτσι, η λέξη επιβίωση είναι η πιο σημαντική σε αυτή την ενότητα γιατί θα παίξει καθοριστικό ρόλο στη διαχείριση των δεδομένων μας αλλά και στην εξαγωγή των συμπερασμάτων μας. Γι’αυτό το λόγο λοιπόν, κάνουμε αναλυτική περιγραφή των βημάτων υλοποίησης σε κείμενο όπως αυτό που διαβάζεται αυτή τη στιγμή για εξηγήσουμε βήμα προς βήμα την περιγραφική στατιστική, αλλά και την ντετερμινιστική στατιστική που ακολουθήθηκε. Όπως δόθηκαν υπάρχουν δύο αρχεία όπου χρησιμοποιήσαμε και τα δύο για την πρόβλεψη της συμβίωσης η μη των ανθρώπων σε βάση κάποιες μεταβλητές-χαρακτηριστικά.

# Συσχέτιση

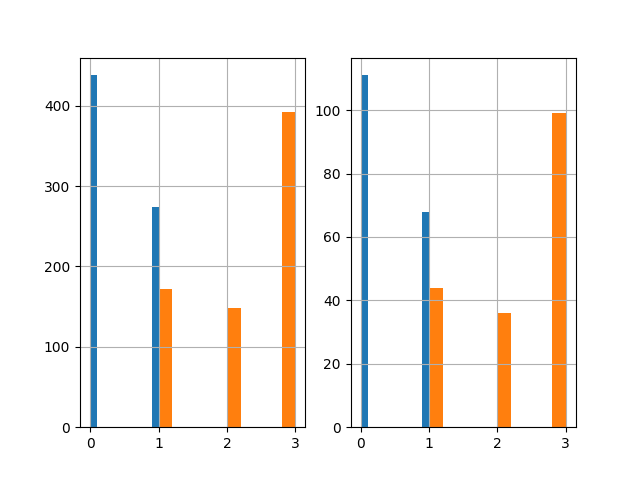
Στο πρώτο βήμα της ανάλυσης των δεδομένων μας αυτό που κάναμε με περιγραφική στατιστική είναι ο εντοπισμός της συσχέτισης των δεδομένων με βάση τις αριθμητικές μεταβλητές αφήνοντας από έξω τις κατηγορηματικές μεταβλητές. Έτσι λοιπόν, προέκυψε το εξής συμπέρασμα. Το συμπέρασμα που προέκυψε είναι το γεγονός ότι υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών Survived – PassengerId, Pclass – PassengerId, Age-PassengerId, SibSp-PassengerId, Fare-PassengerId. Μικρή είναι η συσχέτιση των μεταβλητών μεταξύ των μεταβλητών Pclass-Survived, Age-Pclass, SibSp-Age, Parch-Age, Fare-Pclass. Το συσχετισμό των αντιλαμβανόμαστε από το χρώμα σύμφωνα με την κλίμακα που εμφανίζεται δεξιά της οθόνης. Όσο πιο σκούρο το χρώμα τόσο μεγαλύτερη η συσχέτιση των μεταβλητών. Όσο πιο ανοιχτό το χρώμα τόσο μικρότερη η συσχέτιση μεταξύ των μεταβλητών.



Εικόνα 1Σχήμα συσχέτισης των μεταβλητών

# Stratified Data

Όταν έχουμε να αντιμετωπίσουμε τέτοιου είδους προβλήματα απώτερος στόχος είναι η επιλογή του συνόλου δεδομένων για εκπαίδευση και η επιλογή του συνόλου δεδομένων για έλεγχο. Σε αυτή την περίπτωση, δουλεύουμε αρκετά με το σύνολο δεδομένων που είναι για εκπαίδευση μέχρι να σιγουρευτούμε ότι το μοντέλο μηχανικής μας μάθησης είναι αρκετά έμπιστο και έτοιμο για την αγορά έτσι ώστε να παράγει τα κατάλληλα αποτελέσματα. Στη συνέχεια, ο διαχωρισμός του συνόλου δεδομένων μας γίνεται συνήθως με 80% για το σύνολο εκπαίδευσης και 20% για το σύνολο ελέγχου. Ένας εναλλακτικός διαχωρισμός είναι ότι γίνεται 70% για το σύνολο εκπαίδευσης και ένα 30% για το σύνολο ελέγχου. Λέγοντας διαχωρισμό εννοούμε να επιλέξουμε τον αντίστοιχο αριθμό παρατηρήσεων για τα αντίστοιχα σύνολα δεδομένων. Αυτό γίνεται γιατί θέλουμε να εκπαιδεύσουμε το σύνολο δεδομένων και να τα ελέγξουμε ότι οι προβλέψεις που προκύπτουν είναι σχετικές για τα αντίστοιχα αποτελέσματα και ότι η ακρίβεια που παίρνουμε από το μοντέλο μηχανικής μάθησης επικυρώνει ότι έχουμε εκπαιδεύσει μέχρι στιγμής. Αρχίζοντας λοιπόν την ανάλυση των δεδομένων μας, κάνουμε αρχικά stratify τα δεδομένα μας δηλαδή θέλουμε αρχικά να ανακατέψουμε τα δεδομένα μας έτσι ώστε να είναι τυχαία και να μην συσχετίζεται η μια παρατήρηση με την άλλη. Να υπάρχει δηλαδή ποικιλομορφία στα δεδομένα μας. Οπότε σε αυτή την περίπτωση η τυχαιότητα είναι τυχαιότητα στη στατιστική. Στόχος μας είναι να έχουμε ίδιες κατανομές στις τυχαίες μεταβλητές τόσο στο σύνολο εκπαίδευσης όσο και το σύνολο ελέγχου μας. Αυτό ακριβώς είναι το stratify δεδομένων μας. Το stratify των δεδομένων μας δηλαδή είναι ένας τρόπος επιλογής του συνόλου εκπαίδευσης και του συνόλου ελέγχου. Παραδείγματος χάριν, στο σύνολο δεδομένων μας όπου η target μεταβλητή μας είναι το survived δεν θα θέλαμε να έχουμε όλους τους επιζώντες το σύνολο ελέγχου και όλους τους αποθανόντες στο σύνολο εκπαίδευσης, αλλά θα θέλαμε να υπάρχει μια τυχαιότητα και στα δύο σύνολο δεδομένων μας. Στη συνέχεια, παρουσιάζονται δύο ιστογράμματα τα οποία παρουσιάζουν τις μεταβλητές Survived και Pclass όπως φαίνονται στην Εικόνα 2. Τα ιστογράμματα κρύβουν μια πολύ σημαντική πληροφορίας. Λοιπόν, το πρώτο ιστόγραμμα είναι ένα ιστόγραμμα το οποίο λήφθηκε από το σύνολο δεδομένων, ενώ το δεύτερο ιστόγραμμα είναι ένα ιστόγραμμα που λήφθηκε από το σύνολο ελέγχου. Αυτά τα δύο ιστογράμματα διακρίνουμε ότι είναι παρόμοια άρα η κατανομή από την οποία προέρχονται είναι η ίδια. Οπότε αυτό μας χαροποιεί ιδιαιτέρως διότι υπάρχει τυχαιότητα στα δεδομένα μας. Στη συνέχεια, ο Table 1παρουσιάζει πέντε στήλες. Η πρώτη στήλη παρουσιάζει τον μοναδικό κωδικό της στήλης η δεύτερη στήλη παρουσιάζει το όνομα της στήλης, η Τρίτη στήλη παρουσιάζει τον αριθμό των διαθέσιμων τιμών όπου είναι η πιο σημαντική στήλη και ο λόγος που δημιουργήθηκε ο πίνακας. Στην τέταρτη στήλη εμφανίζεται το non-null όπου δείχνει ότι η στήλη δεν είναι κενή και η Πέμπτη στήλη παρουσιάζει την κωδικοποίηση των δεδομένων. Στη γραμμή Age έχουμε 512 εγγραφές οπότε σημαίνει ότι κάποιες εγγραφές ή παρατηρήσεις λείπουνε. Αυτό μπορούμε να το χειριστούμε εκ των πρώτης όψεως διαγράφοντας τις κενές εγγραφές και να μην τις λάβουμε καθόλου υπόψιν. Όμως αυτό σημαίνει ότι θα χάσουμε πολλά δεδομένα. Οπότε θα ακολουθήσουμε την τακτική Imputing για να μην χάσουμε δεδομένα. Επιπρόσθετα, στη μεταβλητή Cabin έχουμε 165 εγγραφές όπου σημαίνει ότι λείπουν αρκετές. Σε αυτή την περίπτωση επειδή λείπουν πολλές και η φύση της μεταβλητής Cabin είναι ανούσια δεν θα μας απασχολήσει πολύ η μεταβλητή αυτή, οπότε μπορούμε να την αποσύρουμε.



Εικόνα 2Ιστόγραμμα για τις μεταβλητές Survived, Pclass

Table 1Ο πίνακας δείχνει τον αιρθμών των διαθέσιμων δεδομένων

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | PassengerId | 712 | Non-null | Int64 |
| 1 | Survived | 712 | Non-null | Int64 |
| 2 | Pclass | 712 | Non-null | Int64 |
| 3 | Name | 712 | Non-null | Object |
| 4 | Sex | 712 | Non-null | Object |
| 5 | Age | 577 | Non-null | Float64 |
| 6 | SibSp | 712 | Non-null | Int64 |
| 7 | Parch | 712 | Non-null | Int64 |
| 8 | Ticket | 712 | Non-null | Object |
| 9 | Fare | 712 | Non-null | Float64 |
| 10 | Cabin | 158 | Non-null | Object |
| 11 | Embarked | 711 | Non-null | Object |

# Εκτιμητές

Σε αυτή την ενότητα θα επεκταθούμε αρκετά με τους εκτιμητές μας στο σύνολο δεδομένων. Η υλοποίηση αυτή μπορεί να βρεθεί στον κώδικα στο κομμάτι όπου έγινε υλοποίηση με αντικειμενοστρεφή προγραμματισμό στην Python. Έτσι λοιπόν, σε συνδυασμό με το Imputing μπορούμε να γεμίσουμε τις μεταβλητές οι οποίες έχουν ανελλιπείς γραμμές/εγγραφές/παρατηρήσεις. Στη δική μας περίπτωση θα τις γεμίσουμε με τη μέση τιμή της κάθε μεταβλητές στις μεταβλητές που λείπουν. Δημιουργήθηκε η κλάση Age Impute όπου είναι επέκταση της Base Estimator και TransformerMixin. Στην κλάση AgeImputer δημιουργήθηκε η συνάρτηση transform η οποία έχει ως στόχο να γεμίσει τις παρατηρήσεις οι οποίες είναι κενές με τη μέση τιμή όλων των εγγραφών της μεταβλητής Age. Μετέπειτα, στόχος είναι να κωδικοποιηθούν οι μεταβλητές η οποίες είναι κατηγορηματικές και δεν έχουν αριθμητική τιμή. Αυτό διότι κατά της εισαγωγή των δεδομένων μας στο μοντέλο της μηχανικής μάθησης ο υπολογιστής καταλαβαίνει αριθμούς και όχι γράμματα. Θέλουμε δηλαδή στήλες όπου να αποτελούνται από 0, 1,2 3 κλπ. binary τιμές. Αυτό γίνεται καλώντας και κάνοντας import OneHoteEncoders στην κλάση FeatureEncoder επεκτείνοντας την Base Estimator και την TransformMixin. Μέσα στην κλάση δημιουργούμε την συνάρτηση transform όπου δημιουργούμε πίνακα και μετέπειτα, καλούμε τις στήλες όπου έχουν γράμματα όπως τις C, S, Q, N όπου για κάθε μια μεταβλητή καλώ την μέθοδο OneHotEncoder έτσι ώστε να κάνω binary τιμές. Δηλαδή, πιο απλά, η μεταβλητή Sex γίνεται 0 για τον άνδρα και 1 για την γυναίκα. Στη συνέχεια, δημιουργούμε την κλάση FeatureDropper έτσι ώστε να διαγράψει και να αποσύρει από το σύνολο δεδομένων μας τυχόν μεταβλητές οι οποίες δεν είναι χρήσιμες για την εκπαίδευση των δεδομένων μας. Σε αυτή την κλάση έχουμε τη συνάρτηση/μέθοδο transform η οποία αποσύρει τις εξής μεταβλητές: Embarked, Name, Ticket, Cabin, Sex, N. Οπότε αυτοί είναι οι τρεις εκτιμητές οι οποίοι θα χρησιμοποιήσουμε σε αυτό το πείραμα για να αποσύρουμε κάθε φορά μεταβλητές ή να αλλάξουμε σε binary τιμές τα δεδομένα μας. Μετέπειτα, δημιουργούμε Pipeline έτσι ώστε να καλούμε αυτούς τους τρεις εκτιμητές αυτόματα μέσα στο κομμάτι κώδικά μας. Έτσι αν τρέξουμε αυτό το κώδικα μέχρι στιγμής τότε οι εκτιμητές μας δουλεύουν χωρίς καμία ευλάβια. Έτσι παρατηρούμε ότι πλέον στην μεταβλητή Sex έγινε 0 και 1 τιμές. Άρα, μπορούμε να συνεχίσουμε με ευχέρεια στην εκπαίδευση του συνόλου δεδομένων μας.

# Εκπαίδευση του συνόλου δεδομένων μας

Για την εκπαίδευση του συνόλου δεδομένων μας καλούμε το Standard Scaler. Το Standard Scaler είναι ένα είδος μορφής κανονικοποίησης των δεδμοένων λίγο ξεχωριστό όπως έχουμε την κανονικοποίηση στο μυαλό μας. Λοιπόν, το Standard Scaler κανονικοποιεί τα δεδομένα μας αφαιρώντας τη μέση τιμή και κάνοντας αντιστάθμιση με τη διακύμανση. Έτσι, λοιπόν κεντράροντας και κάνοντας scaling ανεξάρτητα σε κάθε μεταβλητή/χαρακτηριστικό υπολογίζοντας τα ανάλογα στατιστικά στα δείγματα μας στο σύνολο δεδομένων όπου είναι για εκπαίδευση. Η μέση τιμή και τη τυπική απόκλιση τότε αποθηκεύονται για μεταγενέστερη χρήση χρησιμοποιώντας την μέθοδο transform. Η Equation 1δείχνει τον τύπο της κανονικοποίησης των δεδομένων.

Equation 1 Τύπος κανονικοποίησης

Η κανονικοποίηση των δεδομένων είναι απαραίτητη προϋπόθεση έτσι ώστε τα δεδομένων να είναι ομοιόμορφα. Είναι βασικό κριτήριο για τους εκτιμητές των μοντέλων της μηχανικής μάθησης. ¨ετσι λοιπόν σε αυτό το πείραμα θα χρησιμοποιήσουμε το Random Forest Classifier. Κάνοντας χρήση φυσικά του Grid Search CV όπου το CV αντιστοιχεί στο cross validation. Αυτό σημαίνει ότι τρέχει το μοντέλο πολλές φορές έτσι ώστε να βρει τους καλύτερους εκτιμητές και συνεπώς να εκπαιδεύσει με αυτούς το μοντέλο μας. Κάνει δηλαδή hyper parameter tuning. Έτσι λοιπόν αρχίζουμε πραγματικά με την εκπαίδευση του μοντέλου της μηχανικής μάθησης. ¨όπως προαναφέρθηκε επιλέξαμε το Random Forest Classifier γιατί χρησιμοποιεί πολλαπλά δέντρα έτσι ώστε να έρθει στην τελική απόφαση μετά από δέντρα και φύλλα όπου εξελίσσει σύμφωνα με τα δεδομένα μας και τις αποφάσεις που παίρνει. Επιλέξαμε το συγκεκριμένο μοντέλο γιατί η φύση των δεδομένων μας μάς το επιτρέπει. Στη συνέχεια, στόχος είναι να βρεθούν οι καλύτεροι εκτιμητές για το μοντέλο μας με το Grid Search. Αν βάλουμε για cross validation 10 τότε αυτό σημαίνει ότι θα ξεδιπλώσει 9 φορές το σύνολο δεδομένων μας και θα κάνει τον έλεγχο στο 1 που απέμεινε. Έτσι λοιπόν στο μοντέλο μηχανικής μάθησης μας, δηλαδή στο Random Forest Classifier έχουμε βάλει ως παραμέτρους το max\_depth=5, n\_estimator=500. Και αυτό έγινε διότι αυτοί είναι οι καλύτεροι εκτιμητές για το μοντέλο μηχανικής μάθησης μας σύμφωνα με το Grid Search. Αυτό δεν είναι όμως, απόλυτο. Μπορεί αν τρέξει πάλι ο αλγόριθμος οι εκτιμητές να αλλάξουν ελαφρώς έως δραματικά. Κατόπιν πειράματος εκτιμήθηκε ότι οι προαναφερθέντες εκτιμητές είναι οι πιο αποτελεσματικού για το μοντέλο μηχανικής μάθησης μας. Το αποτέλεσμα της ακρίβειας του μοντέλου μηχανικής μάθησης μας απεικονίζεται στον Table 1. Σαφώς, αν τρέξει πολλές θα φέρει αποτέλεσμα παρόμοιο με το προαναφερθέν. Το συγκεκριμένο αποτέλεσμα είναι ένα αξιοσημείωτο και αποτελεσματικό για τα δεδομένα μας αποτέλεσμα.

Table 2 Αποτέλεσμα ακρίβειας μοντέλου μηχανικής μάθησης

|  |
| --- |
| Ακρίβεια |
| 0.8156424581005587 |

Μετέπειτα, στόχος είναι να συνδυάσουμε τα δεδομένα του συνόλου της εκπαίδευσης με αυτά του ελέγχου για να κάνουμε εκτιμήσεις, αλλά και προβλέψεις για την μεταβλητή στόχο όπου είναι η Survived. Στη συνέχεια, δημιουργούμε ένα Data Frame έτσι ώστε να περάσουμε όλες τις προβλέψεις που κάναμε στο σύνολο ελέγχου από το σύνολο εκπαίδευσης μας. Μετέπειτα, το αποθηκεύουμε σε ένα αρχείο csv έτσι ώστε να εμφανιστούν οι προβλέψεις.

# Οδηγίες

Για να τρέξει ο κώδικας απαραίτητη προϋπόθεση είναι να εισαχθεί το μονοπάτι (path)του αρχείου train.csv στην αρχή του κώδικα script στην μεταβλητή titanic\_data στην γραμμή 5 και να εισαχθεί το σωστό μονοπάτι για το σύνολο ελέγχου στο κώδικα στην γραμμή 120.