# ΑΝΑΦΟΡΑ

της εξαμηνιαίας εργασίας με τίτλο:

# «ΜΕΛΕΤΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΚΑΡΚΙΝΟΥ ΤΟΥ ΜΑΣΤΟΥ»

για το μάθημα

Προσομοίωση Φυσιολογικών Συστημάτων


Μέλη της ομάδας:

1) Δωροθέα Κουμίδου 03119712
2) Χριστόφορος Κυπριανού 03119711
3) Γιώργος Χαραλάμπους 03119706

## Περίληψη

Στην παρούσα μελέτη εξάγονται συμπεράσματα σχετικά με την εύρεση υποσυνόλου χαρακτηριστικών που οδηγεί σε καλύτερη επίδοση μοντέλων ταξινόμησης, και ταυτόχρονα τον ταχύτερο χαρακτηρισμό των δειγμάτων,από όγκους του μαστού, ως καλοήθη ή κακοήθη. Επίσης πραγματεύεται την επιρροή των ανισόρροπων δεδομένων στην επίδοση τέτοιων μοντέλων ταξινόμησης.



**Λέξεις Κλειδιά:** breast cancer database, ανισόρροπα δεδομένα, σημαντικότητα χαρακτηριστικών, μοντέλα ταξινόμησης, statistical tests.

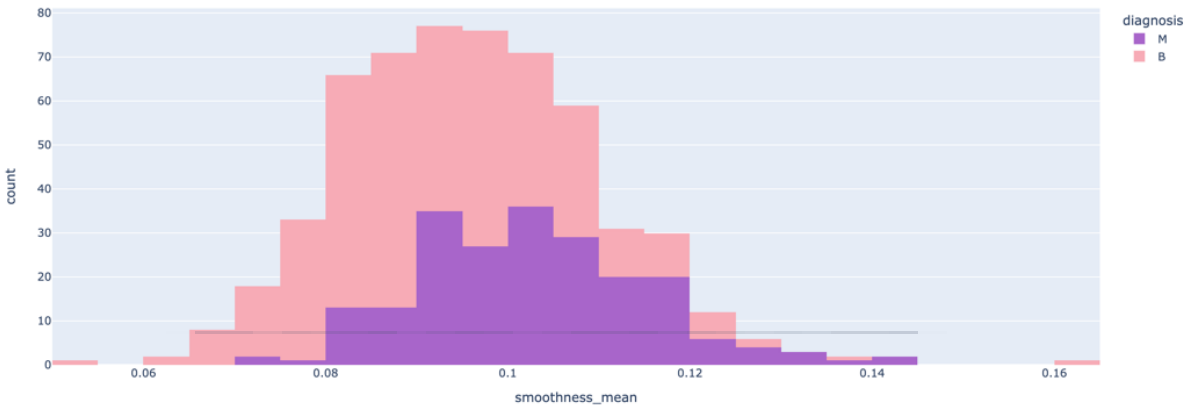
## Περιγραφη συνολου δεδομενων

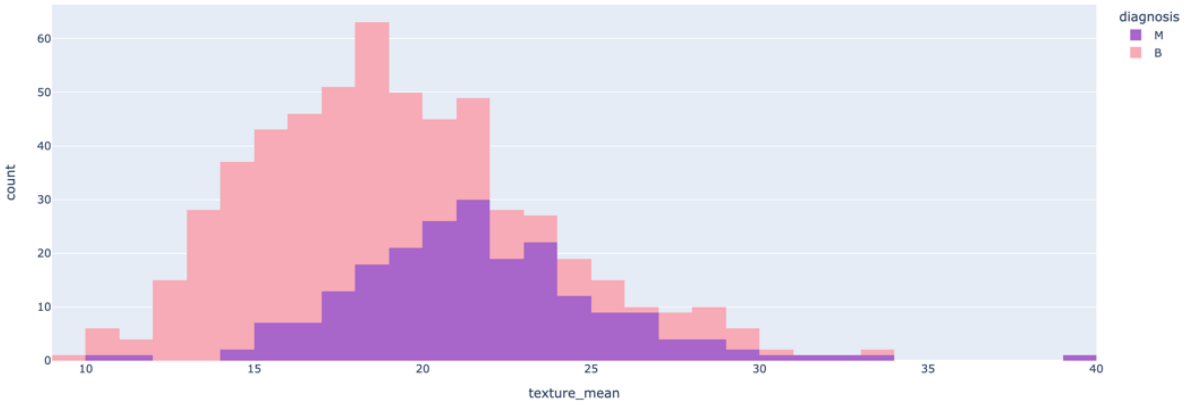
Στην παρούσα μελέτη έχει χρησιμοποιηθεί βάση δεδομένων με χαρακτηριστικά που έχουν εξαχθεί από επεξεργασία MRI εικόνων με όγκων στο μαστό. Συγκεκριμένα, η βάση δεδομένων είναι η «Breast Cancer Wisconsin (diagnostic) Dataset»[1], με πλήθος 569 labeled δείγματα, και 30 χαρακτηριστικά. Τα δεδομένα είναι αριθμητικά εκτός της ετικέτας διάγνωσης του εκάστοτε δείγματος, η οποία έχει τιμές: M- Malignant (Κακοήθης) ή B-Benign (Καλοήθης). Τα χαρακτηριστικά του προς μελέτη όγκου είναι η περίμετρος(perimeter), ακτίνα(radius), υφή(texture), ομαλότητα(smoothness), συμμετρία(symmetry), πλήθος κοίλων σημείων(concave points), μορφοκλασματική διάσταση(fractal dimension), συμπάγεια(compactness) , εμβαδόν (area) και κοίλος χαρακτήρας του όγκου (concavity). Η βάση δεδομένων έχει για κάθε χαρακτηριστικό τον μέσο όρο του (mean), το πειραματικό σφάλμα (se) και το σημείο που αποκλίνει περισσότερο από τα υπόλοιπα δείγματα του εκάστοτε όγκου(worst).

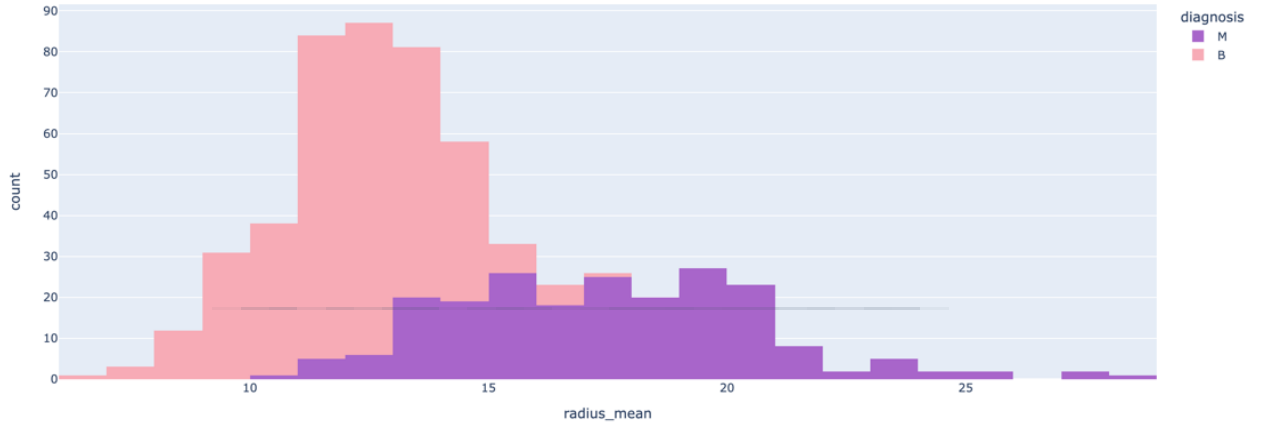
Τα δεδομένα αυτά μπορούν να αξιοποιηθούν για διάγνωση και πρόγνωση καρκίνου του μαστού, επιλογή βέλτιστου σχεδίου θεραπείας όπως στην ακτινοθεραπεία και στην χειρουργική επέμβαση, έρευνα και ανάλυση για ταξινόμηση και βελτιστοποίηση μεθόδων μελέτης όγκων με αποτέλεσμα την μείωση χρόνου και της ανθρώπινης επιρροής [2].

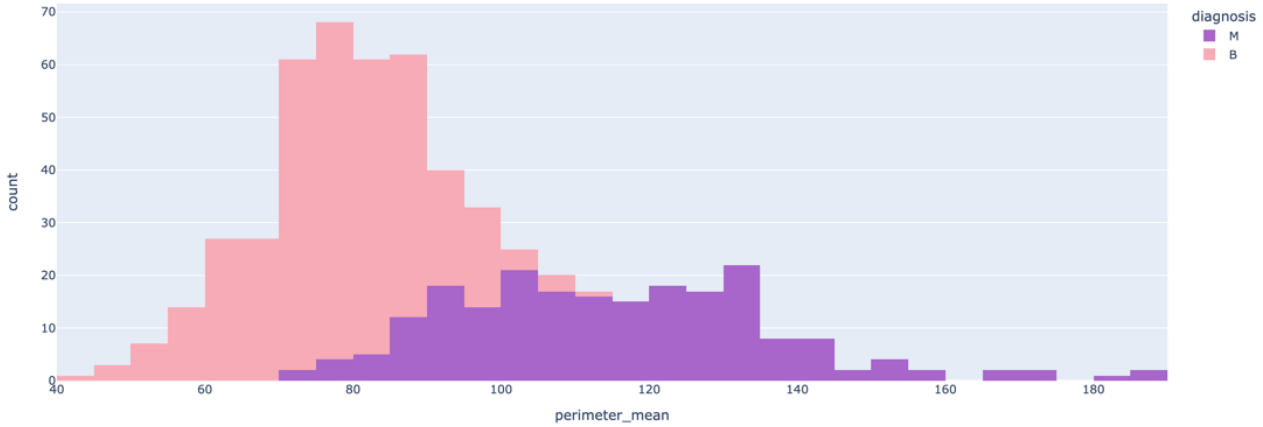
Με την εφαρμογή στατιστικής ανάλυσης στα δεδομένα της βάσης, με χρήση γραφημάτων, τον δείκτη συσχέτισης και στατιστικά τεστ, συμπεραίνεται ότι το πειραματικό σφάλμα και η μορφοκλασματική διάσταση  κάθε χαρακτηριστικού δεν αποτελούν κριτήριο ταξινόμησης του όγκου σε καλοήθη ή κακοήθη.

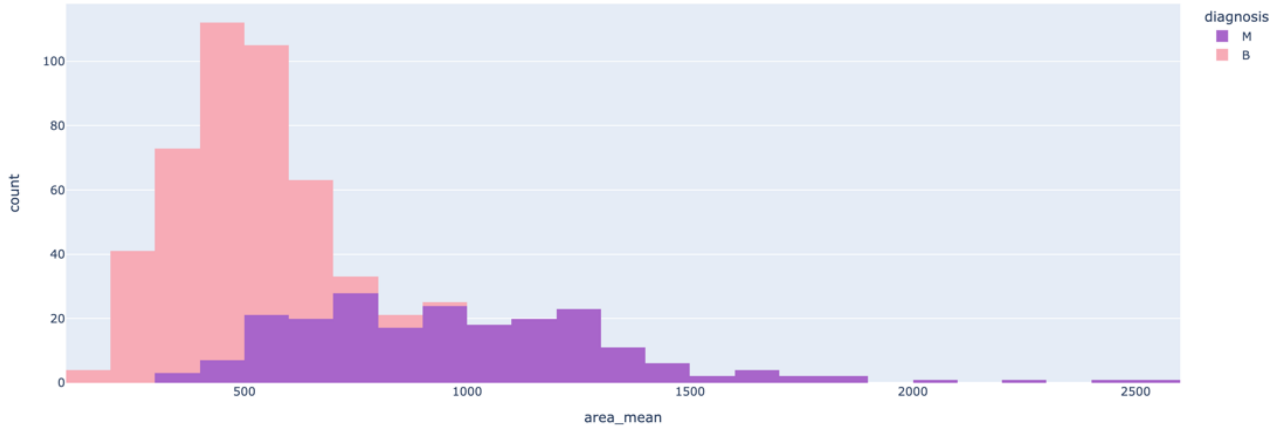
- Παρακάτω φαίνονται τα ιστογράμματα των χαρακτηριστικών συναρτήσει του πλήθους των δειγμάτων στην κάθε κατηγορία:



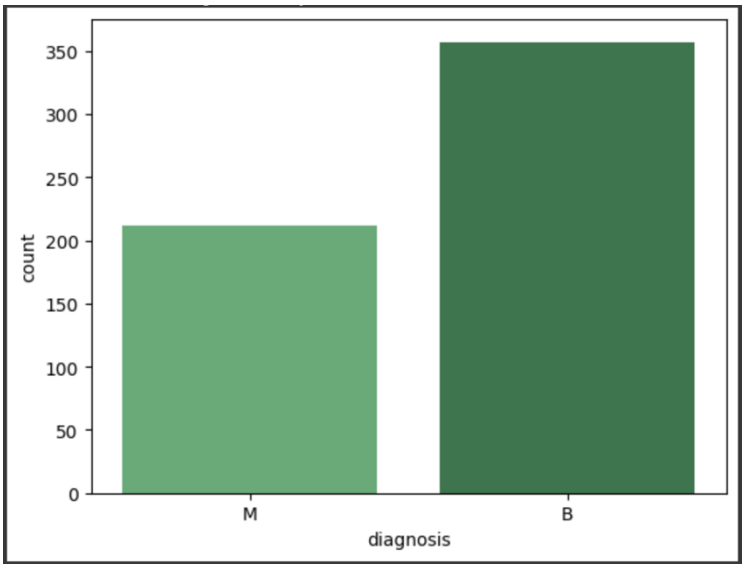



*Σύνολο Πινάκων 1: Μελέτη Δεδομένων*

- Παρουσίαση των άνισα κατανεμημένων κλάσεων (Malignant Benign):




*Πίνακας 1*

- *Πλήθος κενών μεταβλητών:*

```
Empty data:
id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
```

## ΔΙΑΤΥΠΩΣΗ ΕΡΕΥΝΗΤΙΚΟΥ ΕΡΩΤΗΜΑΤΟΣ

Το ερευνητικό ερώτημα που αποτέλεσε έναυσμα της παρούσας μελέτης είναι αν είναι δυνατή η εύρεση ενός υποσυνόλου χαρακτηριστικών που να οδηγεί σε καλύτερη επίδοση μοντέλων ταξινόμησης, και ταυτόχρονα τον ταχύτερο διαχωρισμό των δειγμάτων. Κατά την διαδικασία της μελέτης του πιο πάνω ερωτήματος, αφού τα δείγματά μας ήταν ανισόρροπα, οδηγηθήκαμε σε ένα δεύτερο ερώτημα πως επηρεάζει η ανισορροπία κλάσεων την επίδοση των μοντέλων ταξινόμησης.

## ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΔΟΛΟΓΙΑΣ

Αρχικά μέσω του Kaggle βρέθηκε η βάση δεδομένων «Breast cancer Wisconsin (diagnostic) Dataset»[1] η οποία αποτελεί την ραχοκοκαλιά της μελέτης. Στην συνέχεια γίνεται ανασκόπηση της βάσης για κενές παραμέτρους ώστε να διορθωθούν πριν την επεξεργασία των χαρακτηριστικών. Μετά γίνεται καθορισμός των κατηγοριών και το πλήθος των δειγμάτων ανά κλάση. Παρατηρούνται δύο κατηγορίες, καλοήθης και κακοήθης όγκος, με 357 δείγματα και 212 αντίστοιχα, δηλαδή οι κλάσεις μας είναι άνισα κατανεμημένες. Με βιβλιογραφική

ανασκόπηση διαπιστώθηκε ότι, η ανισορροπία των κλάσεων, αποτελεί επιβαρυντικό παράγοντα στην επίδοση του μοντέλου, έτσι κατά την εκπαίδευση του, εφαρμόζεται υπερδειγματοληψία στην κλάση μειονότητας (κακοήθης όγκος)[3]. Το test set παραμένει αναλλοίωτο ώστε να αξιολογηθεί σωστά το μοντέλο χωρίς την ύπαρξη «προκαταλήψεων» (bias). Αυτό υλοποιείται με την χρήση της συνάρτησης «Random Over Sampler» η οποία επιλέγει τυχαίο διπλασιασμό δειγμάτων της μειονάζουζας κλάσης.

Με πρότυπο την συγκεκριμένη βάση, έγινε ένας πρώτος οπτικός διαχωρισμός των χαρακτηριστικών που αποτελούν σημαντικό κριτήριο για την ταξινόμηση στις δύο κλάσεις. Οι γραφικές παραστάσεις που χρησιμοποιήθηκαν ώστε να βοηθήσουν τον οπτικό διαχωρισμό των χαρακτηριστικών, παρουσιάζουν τον μέσο όρο κάθε χαρακτηριστικό του δείγματος συναρτήσει του μέγιστου αποκλίνον σημείο του εκάστοτε όγκου.

Στην συνέχεια γίνεται επαλήθευση των αρχικών μας υποθέσεων με χρήση στατιστικών test, συγκεκριμένα με το «mann-Whitney-u», το οποίο είναι μη παραμετρικό και έτσι έχει την ικανότητα να συγκρίνει της κατανομές των ανεξάρτητων δειγμάτων στις 2 μη ισορροπημένες κλάσεις συγκρίνοντας το p-value που προκύπτει. Η στατιστική αυτή δοκιμή αποσκοπεί στην μελέτη ύπαρξης στατιστικά σημαντικών διαφορών ανάμεσα στα χαρακτηριστικά των όγκων ώστε να εξαχθεί συμπέρασμα αν το κάθε χαρακτηριστικό αποτελεί κριτήριο για τον διαχωρισμό σε καλοήθη ή κακοήθη όγκο[4]. Η παράμετρος "p-value" η οποία δείχνει την πιθανότητα οποιαδήποτε παρατηρούμενη διαφορά να είναι τυχαία μεταξύ των δειγμάτων, καθορίζει τις στατιστικά σημαντικές παραμέτρους. Αν αυτή η τιμή είναι κάτω από 0.05, υποδηλώνεται ότι υπάρχει σημαντική διαφορά στον τρόπο με τον οποίο διαφέρει το κάθε χαρακτηριστικό μεταξύ των καλοηθών και κακοηθών όγκων, άρα το εκάστοτε χαρακτηριστικό αποτελεί κριτήριο[5].

Στη συνέχεια αφαιρείται η στήλη "id" από το σύνολο δεδομένων και με τη χρήση της συνάρτησης LabelEncoder μετασχηματίζεται η στήλη διάγνωσης από M και B σε 1 και 0, αντίστοιχα, έτσι ώστε να γίνει χρήση του κριτηρίου συσχέτισης και η εκπαίδευση του μοντέλου. Με βάση το κριτήριο αυτό, δημιουργείται ο πίνακας συσχέτισης(correlation matrix) των χαρακτηριστικών, που υποδηλώνει τις σχέσεις μεταξύ τους. Επιλέγονται τα χαρακτηριστικά, όπου ο δείκτης συσχέτισης τους με την στήλη διάγνωσης είναι μεγαλύτερη από 0.2[6].

Έπειτα, αφαιρούνται από τη βάση δεδομένων τα χαρακτηριστικά που δεν αποτελούν κριτήριο διαχωρισμού, όπως το fractal_dimension_mean και όλα τα πειραματικά λάθη. Αφού έγινε η μελέτη των δεδομένων και ο καθορισμός των σημαντικών χαρακτηριστικών έγινε η διάκριση σε test και train σετ με ποσοστό 20% και 80% αντίστοιχα. Με την χρήση του grid search with cross validation, βρίσκεται το βέλτιστο μοντέλο με τον καλύτερο συνδυασμό υπερπαραμέτρων μεταξύ τεσσάρων μοντέλων ταξινόμησης. Ο χώρος αναζήτησης αποτελείται από Logistic regression, Random Forest classifier, KNeighbors classifier, MLP classifier. Μετά τον καθορισμό της στρατηγικής δειγματοληψίας, ορίστηκε cross validation με folds=5.

Αφού εκτελέστηκε η εκπαίδευση, προκύπτει ως βέλτιστη εκπαίδευση το Random Forest Classifier με υπερ-παραμέτρους max depth=7, number estimator=610. Η αξιολόγηση της πρόβλεψης

του test set γίνεται με χρήση μετρικής f1 score, για τρεις περιπτώσεις. Η πρώτη περίπτωση είναι με την χρήση της συνάρτησης εξισορρόπησης δεδομένων στο επιλεγμένο υποσύνολο χαρακτηριστικών, η δεύτερη περίπτωση χωρίς αυτήν και η τρίτη πάλι χωρίς εξισορρόπηση αλλά με χρήση ολόκληρου του αρχικού συνόλου δεδομένων για την εκπαίδευση του μοντέλου.

## ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

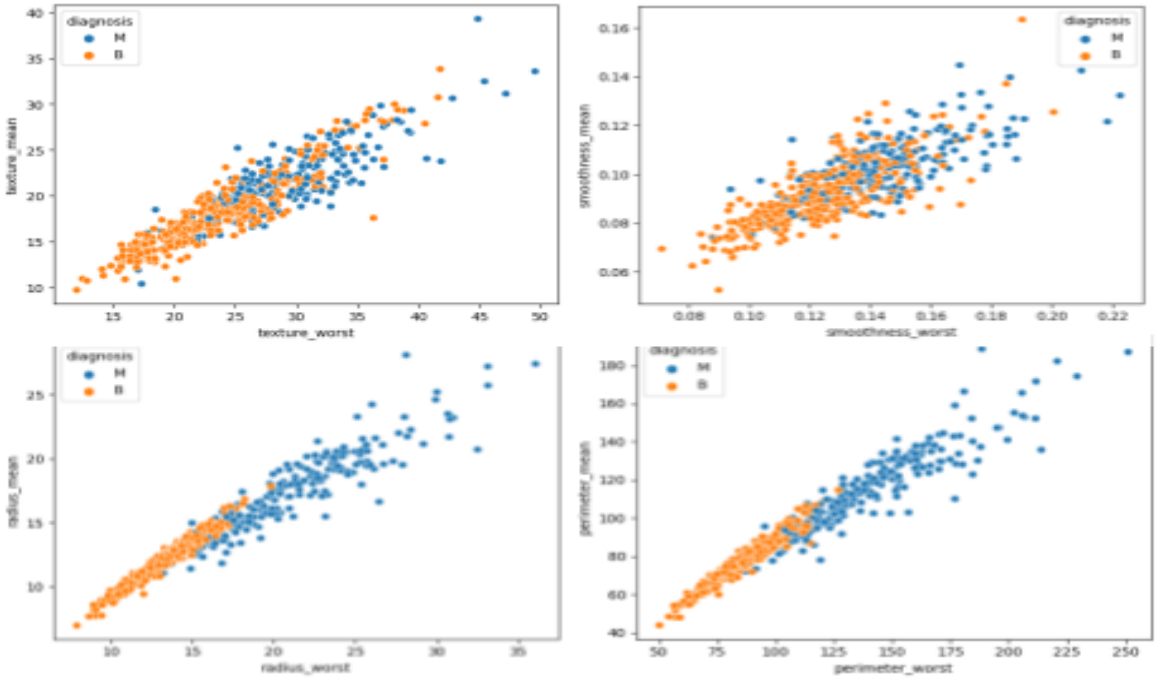
Με μια πρώτη ματιά φαίνεται η ανισορροπία των δειγμάτων ως προς την κατηγορία ταξινόμησης με περισσότερα δείγματα στη καλοήθη κλάση *[Πίνακας 1]*. Στην ανασκόπηση της βάσης βρέθηκε ότι δεν υπήρχαν κενές παράμετροι έτσι συνεχίζεται η επεξεργασία χωρίς κάποια αλλαγή. Στον οπτικό διαχωρισμό των χαρακτηριστικών σε σημαντικά και μη, επιλέχθηκαν τα fractal dimension worst, fractal dimension mean, symmetry worst, symmetry mean, concavity worst, concavity mean, texture worst, texture mean και όλα τα πειραματικά σφάλματα *[Σύνολο Πινάκων 2, Πίνακας 2]* ως μη σημαντικά χαρακτηριστικά κατηγοριοποίησης. Στην συνέχεια στην προσπάθεια επαλήθευσης των αρχικών υποθέσεων συμπεραίνεται, από τον πίνακα συσχέτισης *[Πίνακας 3]* και τα στατιστικά test *[Πίνακας 4]*, ότι τα χαρακτηριστικά με λιγότερη σημαντικότητα στον διαχωρισμό των δύο κλάσεων, τελικά είναι μόνο το fractal dimension mean και τα πειραματικά λάθη, όπως φαίνεται και πιο κάτω:

```
These features have correlation<0.2=> are unrelevant for our purpose:
fractal_dimension_mean    0.012838
texture_se                0.008303
smoothness_se             0.067016
symmetry_se               0.006522
fractal_dimension_se      0.077972
```

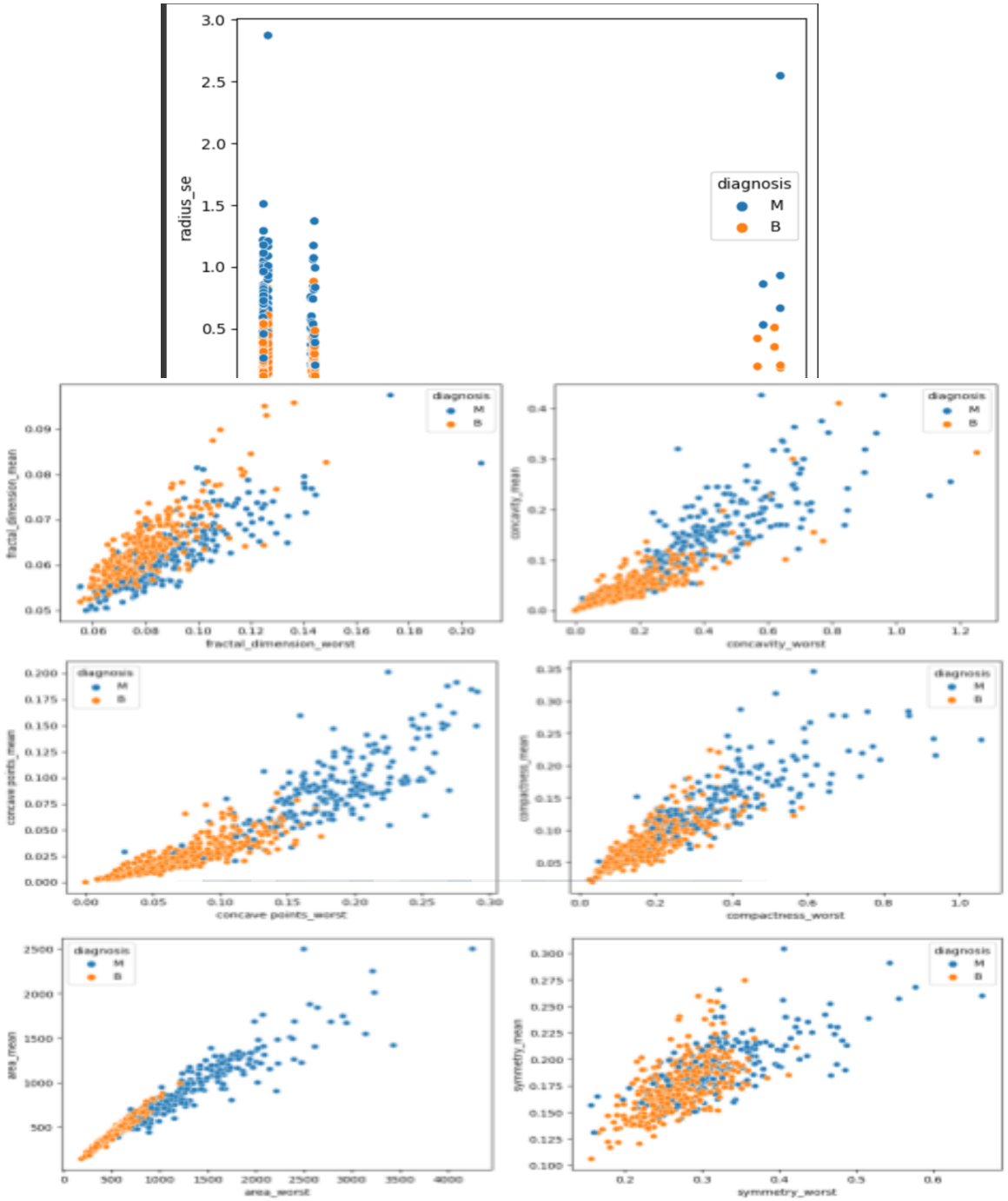
```
The result of the Mann-Whitney U-statistic :
[['perimeter_mean' '3.553870225963875e-71']
 ['perimeter_worst' '2.5830037182989858e-80']
 ['radius_mean' '2.6929427727965647e-68']
 ['radius_worst' '1.1356300904893913e-78']
 ['texture_mean' '3.428626504744227e-28']
 ['texture_worst' '6.517717977951487e-30']
 ['smoothness_mean' '7.793006595586556e-19']
 ['smoothness_worst' '3.637942156482749e-24']
 ['fractal_dimension_worst' '1.1442398346150665e-13']
 ['symmetry_mean' '2.2680501067477204e-15']
 ['symmetry_worst' '3.1512369934706657e-21']
 ['concave points_mean' '1.0063237037340002e-76']
 ['concave points_worst' '1.8639972354360316e-77']
 ['compactness_mean' '8.951992005223344e-48']
 ['compactness_worst' '2.1155252552555298e-47']
 ['concavity_mean' '2.164548790621846e-68']
 ['concavity_worst' '1.7617231681140704e-63']
 ['area_mean' '1.539780362858885e-68']
 ['area_worst' '1.8033090105551777e-78']
 ['fractal_dimension_mean' '0.537185602135624']]
.................................................
These features have p-value>0.05=> are unrelevant for our purpose:
[['fractal_dimension_mean' '0.537185602135624']]
```

Έπειτα από την εκτέλεση του grid search with cross-validation στη βάση διαπυστώθηκε πως ο καλύτερος ταξινομητής για την ταξινόμηση αυτών των δειγμάτων ειναι ο RandomForestClassifier(max_depth=7, n_estimators=610). Να σημειωθεί πως το GridSearchCV χρησιμοποιεί ως μετρική την ακρίβεια (accuracy) για την επιλογή του βέλτιστου μοντέλου. Στη συνέχεια από τις τρεις διαφορετικές εκπαιδεύσεις διαπιστώνεται πως η εξισορρόπηση των δεδομένων στις δύο κλάσεις(με RandomOverSampler(sampling_strategy='minority')) βελτιώνει την απόδοση του μοντέλου (f1_score), αλλά η επιλογή ή όχι των σημαντικών χαρακτηριστικών για τον διαχωρισμό των δεδομένων ως σύνολο δεδομένων της εκπαίδευσης δεν διαφοροποιεί την απόδοση του μοντέλου, τα f1 scores παραμένουν τα ίδια *[Πίνακας 5]* .

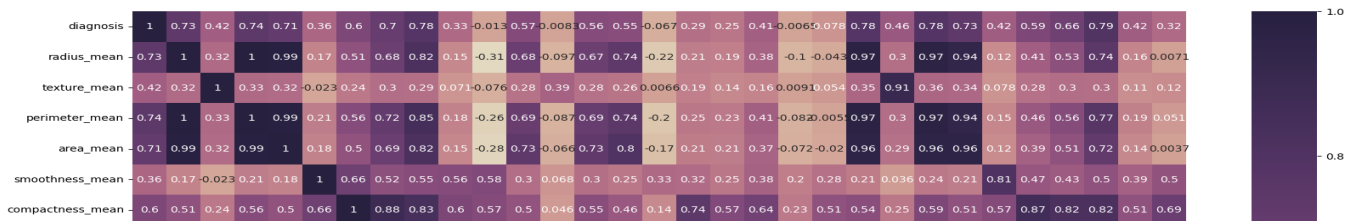
- Παρακάτω παρατίθενται οι γραφικές παραστάσεις των αποτελεσμάτων:



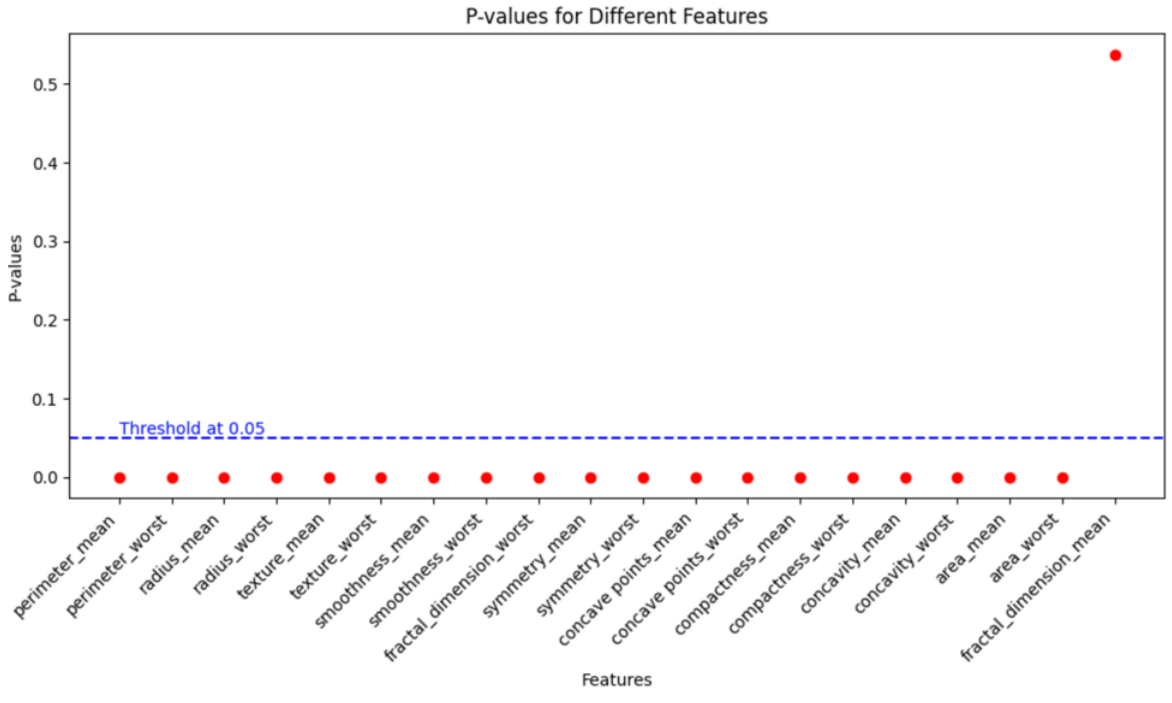
*Σύνολο Πινάκων 2: Οπτικός διαχωρισμός Χαρακτηριστικών*



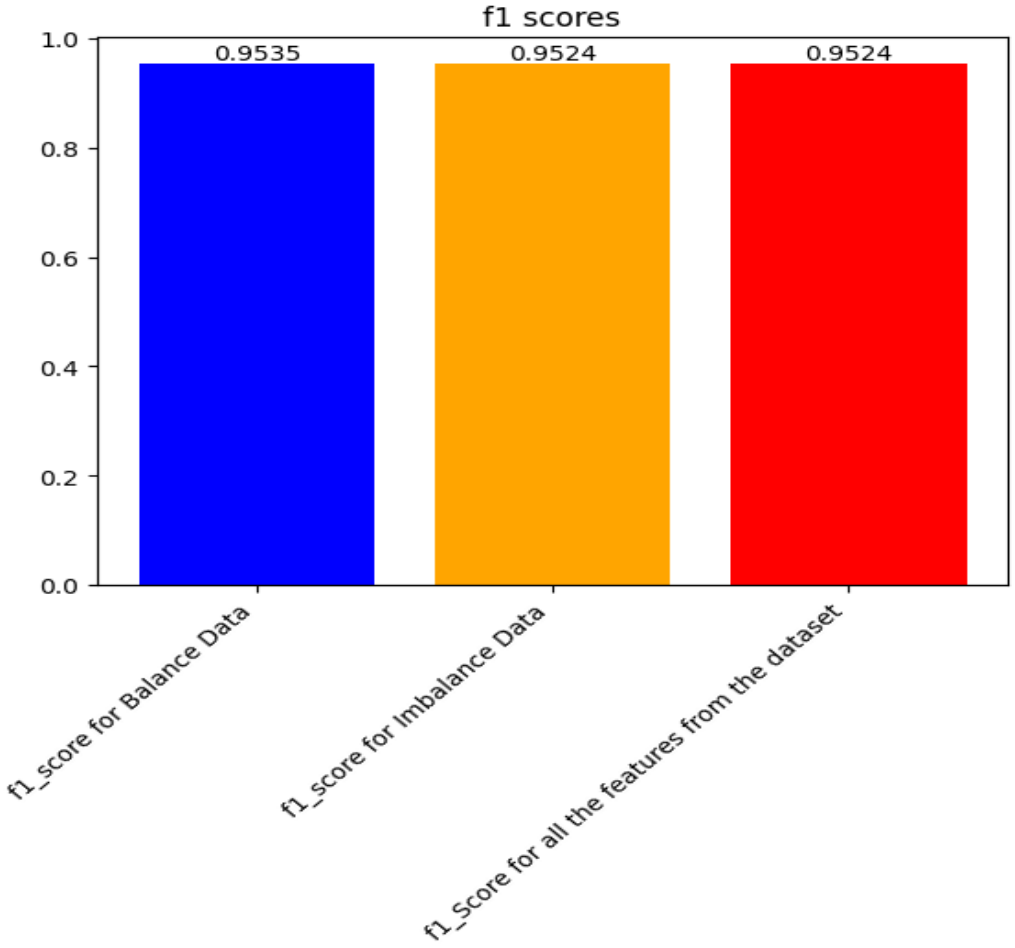

*Πίνακας 2: Μελέτη πειραματικών σφαλμάτων*



*Πίνακας 3: Πίνακας Συσχέτισης*



*Πίνακας 4: P-Values των χαρακτηριστικών και threshold= 0.05*




*Πίνακας 5: Επιδόσεις του μοντέλου με διαφορετική επεξεργασία των δεδομένων*

## ΣΥΖΗΤΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Αναφορικά με τα προαναφερθέντα αποτελέσματα σχετικά με την μελέτη και επεξεργασία των δεδομένων, συμπεραίνεται, αρχικά ότι ο οπτικός διαχωρισμός χαρακτηριστικών σημαντικότητας δεν είναι έγκυρος αφού αφαιρέθηκαν περισσότερα χαρακτηριστικά από όσα έπρεπε όπως αποδείχθηκε από το στατιστικό τεστ και τον πίνακα συσχέτισης. Δεύτερο, με την χρήση της παραμέτρου p-value παρατηρείται αφαίρεση μόνο του χαρακτηριστικού «fractal dimension mean» όπως φαίνεται και στον *Πίνακα 4* , μόνο αυτό ξεπερνά το όριο 0.05. Τρίτον, ο πίνακας συσχέτισης εξάγει μεγαλύτερο πλήθος χαρακτηριστικών προς αφαίρεση για την ταξινόμηση τους σε σημαντικά και μη με στόχο την μελέτη της ακρίβεια του μοντέλου*[Πίνακας 3]*. Επομένως αποδεικνύεται ότι δεν επαρκεί μόνο μία μέθοδος για τον καθορισμό της σημαντικότητας των χαρακτηριστικών.

Όσον αφορά την εκπαίδευση και επαλήθευση του μοντέλου, επιλέχθηκε ως βέλτιστο μοντέλο το Random Forest. Επίσης το ποσοστό της f1 score είναι μεγαλύτερο στην περίπτωση των ισορροπημένων χαρακτηριστικών, πράγμα που αποδεικνύει την ανάγκη της εξισορρόπησης των δεδομένων σε διάφορες μελέτες με αντίστοιχο περιεχόμενο και πως η επιλογή της RandomOverSampler(sampling_strategy='minority') ως στρατηγική υπερδειγματοληψίας συνεισφέρει στην ακρίβεια της ταξινόμησης . Αλλά όπως φαίνεται από την τρίτη περίπτωση με χρήση αναλλοίωτου του συνόλου δεδομένου, η επιλογή των στατιστικά σημαντικών δεδομένων και αυτών με την μεγαλύτερη συσχέτιση δεν επηρεάζει την επίδοση του μοντέλου ούτε θετικά αλλά ούτε αρνητικά. Συμπερασματικά είναι δυνατή η εύρεση υποσυνόλου δεδομένων ώστε να οδηγήσει σε καλύτερη επίδοση μοντέλων ταξινόμησης, και ταυτόχρονα τον ταχύτερο διαχωρισμό των δειγμάτων αλλά στην συγκεκριμένη περίπτωση αυτή η επιλογή υποσυνόλου δεν διαφοροποιεί το αποτέλεσμα του ταξινομητή. Επίσης όπως προαναφέρθηκε ότι η ανισορροπία κλάσεων επηρεάζει αρνητικά την επίδοση των μοντέλων ταξινόμησης.

## ΒΙΒΛΙΟΓΡΑΦΙΑ


[1] H, M. Y. (2021, December 29). *Breast cancer dataset*. Kaggle. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data

[2] *Chatgpt*. ChatGPT. (n.d.). https://openai.com/chatgpt

[3] Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on Imbalanced Data. *Journal of Biomedical Informatics*, *90*, 103089. https://doi.org/10.1016/j.jbi.2018.12.003

[4] Vrbin, C. M. (2022). Parametric or nonparametric statistical tests: Considerations when choosing the most appropriate option for your data. *Cytopathology*, *33*(6), 663–667. https://doi.org/10.1111/cyt.13174

[5] Dahiru, T. (2011). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*, *6*(1). https://doi.org/10.4314/aipm.v6i1.64038