# DA, CDL, ORCA, ORCB, Performance, Sport, κτλ.

Γιάννης Καρυωτάκης
S/Y Mimosa
Πέμπτη, 22 Αυγούστου 2024

## Περίληψη

Για πολύ καιρό προσπάθησα να καταλάβω όλες τις μεταβλητές που βρίσκουμε στα πιστοποιητικά ORC. Όχι τόσο τις βασικές, όπως το μήκος (LOA) ή το βύθισμα (Draft), αλλά τις σύνθετες, DA, CDL, πχ. Ιδιαίτερα τις δύο τελευταίες γιατί χρησιμοποιούνται από την ΕΑΘ αλλά και την ORC για να δημιουργηθούν κατηγορίες σκαφών σε κάποιο αγώνα. Έτσι με τον καιρό έγραψα τρία κείμενα. Το πρώτο εξηγεί τι είναι οι DA και CDL https://www.sailing-mimosa.eu/enimerotika/1999293_dynamic-allowance-da-kai-class-division-length-cdl. Στο δεύτερο αναπτύσσω μια ανορθόδοξη γνώμη όπως γράφω για να δημιουργήσουμε κατηγορίες στην Ελλάδα, διαφορετικές από τις σημερινές του Σαρωνικού, Performance και Sport. Και στο τρίτο χρησιμοποιώ μεθόδους Machine Learning για να δημιουργήσω αμερόληπτες κατηγορίες βασισμένες σε πολλά χαρακτηριστικά των σκαφών και όχι μόνο σε ένα ή δύο.

Είμαι πεπεισμένος ότι το κριτήριο κατηγοριοποίησης πρέπει να είναι η ταχύτητα των σκαφών όπως υπολογίζεται από το VPP, και εν μέρει εκφράζεται από την τιμή της CDL. Χρησιμοποιώντας την CDL μόνο, μπορούμε να δημιουργήσουμε κατηγορίες ορίζοντας διάφορα σταθερά όρια όπως πχ προτείνει η ORC αλλά και εμείς, ή να αφήσουμε κάποιον αλγόριθμο Machine Learning να μας πει ποιο σκάφος ανήκει σε ποια κατηγορία. Στις επόμενες σελίδες θα βρείτε τα τρία κείμενα ξεχωριστά.

*Σχόλια και παρατηρήσεις είναι ευπρόσδεκτα e-mail: SailingMimosa@orange.fr*

## Περιεχόμενο

# Performance έναντι Sport, μια ανορθόδοξη γνώμη

Γιάννης Καρυωτάκης
S/Y Mimosa
Πέμπτη, 22 Αυγούστου 2024

Σε κάθε αγώνα στον Σαρωνικό κοιτώντας τα αποτελέσματα θα δούμε πάντα δύο κατηγορίες σκαφών, Performance και Sport και εάν ο αριθμός των αγωνιζομένων είναι αρκετός θα δούμε και υποκατηγορίες. Ο διαχωρισμός Performance και Sport γίνεται βάσει της τιμής της Dynamic Allowance (DA) και ο διαχωρισμός σε υποκατηγορίες βάσει της τιμής της Class Division Length (CDL). DA και CDL είναι σύνθετες μεταβλητές, εξαρτώμενες από πολλές βασικές παραμέτρους του σκάφους, πχ βάρος, μήκος, επιφάνεια πανιών, σχήμα της γάστρας, κτλ. και υπολογίζονται κατά την έκδοση πιστοποιητικού από το πρόγραμμα Velocity Prediction Program (VPP).

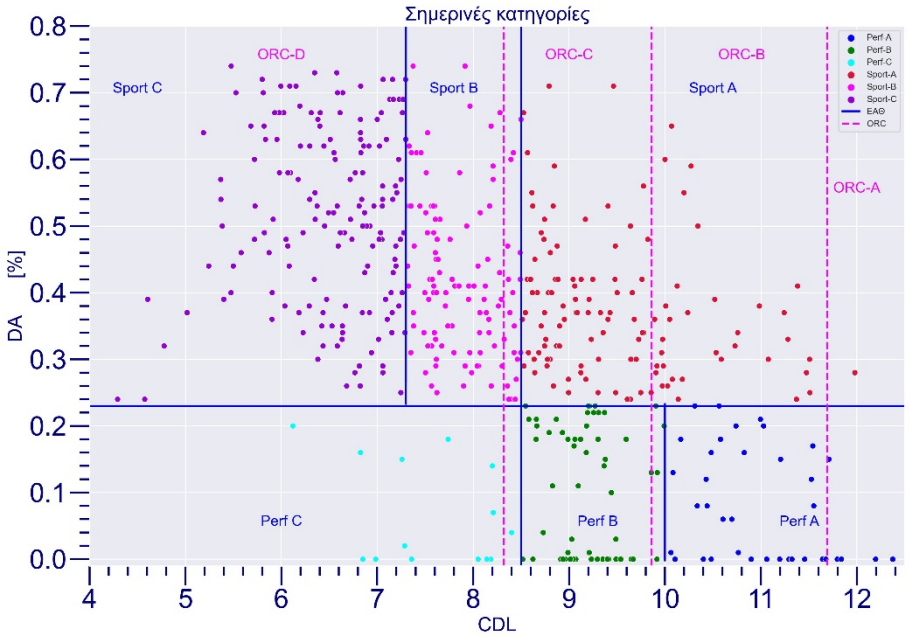
Ο σκοπός της διαδικασίας του ισοζυγισμού είναι να εξισώσει τις θεωρητικές επιδόσεις των σκαφών με την χρήση διορθωτικών συντελεστών στους χρόνους, δημιουργώντας έτσι έναν στόλο όμοιων σκαφών. Ο διαχωρισμός σε κατηγορίες δεν είναι θεωρητικά απαραίτητος, μια και τα σκάφη είναι ισοζυγισμένα, αλλά εξυπηρετεί κατά την γνώμη μου περισσότερο κοινωνικούς σκοπούς. Οι αθλητές έχουν το αίσθημα ότι συναγωνίζονται παρόμοια με το δικό τους σκάφη και μπορούν να ελπίζουν σε κάποια διάκριση. Αναδεικνύει επίσης πολλούς νικητές και είναι κίνητρο για κάποιον που ονειρεύεται ένα κύπελλο. Το ίδιο γίνεται και σε άλλα αθλήματα. Σε έναν μεγάλο αγώνα δρόμου, ένα μαραθώνιο, θα πάρει κύπελλο ο πρώτος άνδρας η πρώτη γυναίκα και ίσως ο πρώτος νέος κάτω από 20 χρονών πχ. Ψυχολογικά αυτή η επιβράβευση παίζει μεγάλο ρόλο για την συνέχεια του αθλητή αλλά και του αθλήματος. Νικητής είναι όμως μόνο ένας!

Στα νερά του Σαρωνικού τώρα, ένα σκάφος με DA ≤ 0,23% θα τρέξει στην κατηγορία Performance και ένα άλλο με DA > 0,23% στην κατηγορία Sport, εκτός και εάν οικειοθελώς ζητήσει να τρέξει στην Performance. Γιατί διαλέξαμε την DA και την τιμή 0,23% και όχι 0,24% ή 0,25% ; Η DA=0,23% είναι κάποια μαγική μεταβλητή και μαγική τιμή που ξεχωρίζει τα γρήγορα από τα αργά σκάφη; Επιβάλλεται από κάποιον νόμο της υδροδυναμικής ή αεροδυναμικής; Για να απαντήσει σε αυτή την ερώτηση κάποιος με εμπειρία στην στατιστική ανάλυση δεδομένων, θα κοιτάξει την κατανομή της DA όλων των σκαφών, ψάχνοντας αρχικά με το μάτι να δει εάν η τιμή των 0,23% χωρίζει την κατανομή σε δύο ευδιάκριτα σύνολα. Χρησιμοποιώντας τον κατάλογο τον πιστοποιητικών του 2023 που δημοσιεύει η ΕΑΘ, ένα σύνολο 677 σκαφών, με λίγο προγραμματισμό είναι δυνατόν να βρούμε αυτή την κατανομή, βλέπετε Εικόνα 9. Δεν χρειάζονται ιδιαίτερες γνώσεις για να δει κανείς ότι η κατανομή της DA είναι συνεχής και δεν παρουσιάζει εύκολα ταυτοποιήσιμα υποσύνολα. Η δε τιμή των 0,23% (κόκκινη γραμμή) δεν ξεχωρίζει κανένα σύνολο από ένα άλλο και κάλλιστα θα μπορούσε να ήταν διαφορετική. Τα ίδια θα μπορούσα να έγραφα και για την CDL. Οι μεταβλητές DA και CDL σίγουρα χαρακτηρίζουν τις επιδόσεις των σκαφών, αλλά εάν χρησιμοποιήσουμε κάποια τιμή αυτών για να ορίσουμε κατηγορίες, αυτή θα είναι αυθαίρετη, δηλαδή δεν θα μπορούμε να την δικαιολογήσουμε με αντικειμενικά επιχειρήματα.

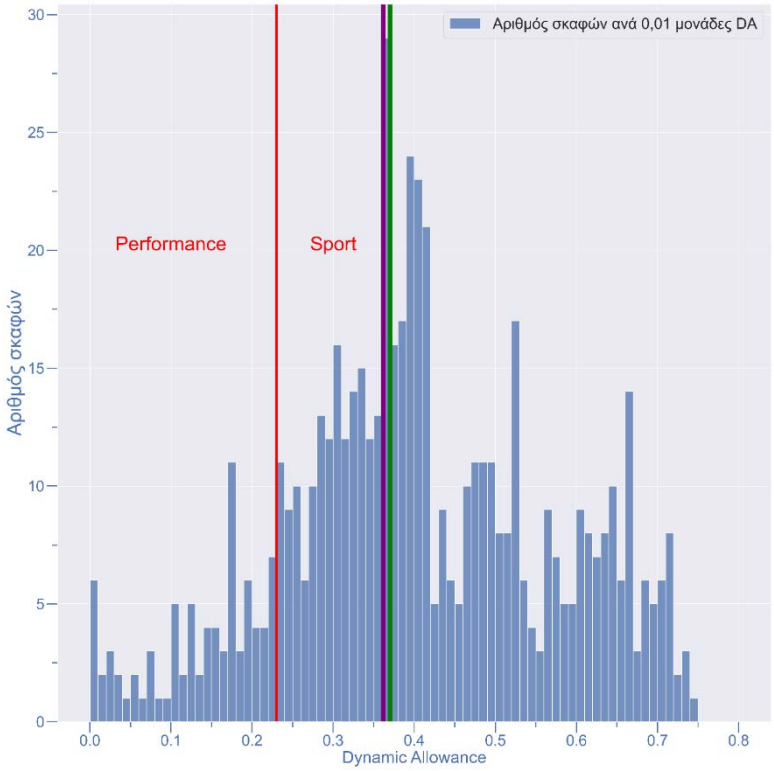
Ένα διαφορετικό τρόπο ορισμού κατηγοριών, προτείνει ο οδηγός διαχείρισης αγώνων της ORC https://orc.org/uploads/files/ORC-Race-Management-Guide-2023.pdf ο οποίος ορίζει 3 κατηγορίες για τα παγκόσμια και εθνικά πρωταθλήματα, βασισμένες στην τιμή του CDL.

- Class A: 16.400 ≥ CDL > 11.690
- Class B: 11.690 ≥ CDL > 9.860
- Class C: 9.860 ≥ CDL > 8.320

Στην Ιταλία προσθέτουν και μια τέταρτη κατηγορία Class D: 8.320 ≥ CDL. Για το πρωτάθλημα τους ορίζουν ομάδα 1 τις κατηγορίες A και B που θα λέγαμε Performance και ομάδα 2 τις κατηγορίες C και D που θα λέγαμε Sport. Στις **Εικόνα 8** και **Εικόνα 9** βλέπουμε τις κατανομές των DA vs CDL και DA.



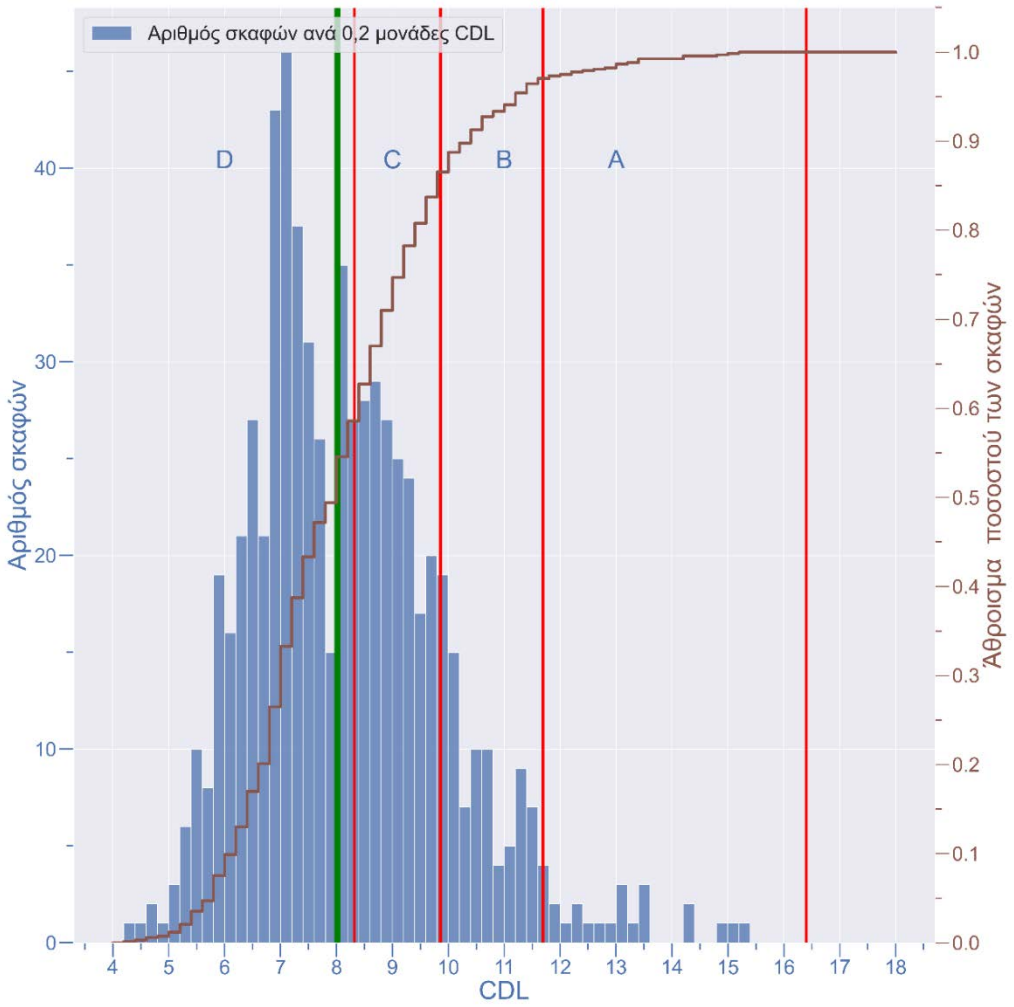

*Εικόνα 1: Οι σημερινές κατηγορίες Performance και Sport. Κάθε σκάφος είναι ένα σημείο*




*Εικόνα 2 : Κατανομή της DA, στόλος του 2023. Έχουν αφαιρεθεί τα σκάφη με DA=0. Η μωβ γραμμή είναι η μέση τιμή, η πράσινη είναι η διάμεσος και η κόκκινη χωρίζει τον στόλο, αριστερά Performance, δεξιά Sport.*

Έτσι λοιπόν από την μια μεριά θέλουμε κατηγορίες και από την άλλη δεν έχουμε κάποια διακριτική μεταβλητή[1] για να τις ορίσουμε αμερόληπτα και χωρίς a fortiori επιχειρήματα του τύπου, 'Μα αυτό το σκάφος που κερδίζει τα πάντα δεν μπορεί να είναι Sport' ή 'αυτό πρέπει σίγουρα να είναι στην κατηγορία Performance'. Οπότε τι κάνουμε;

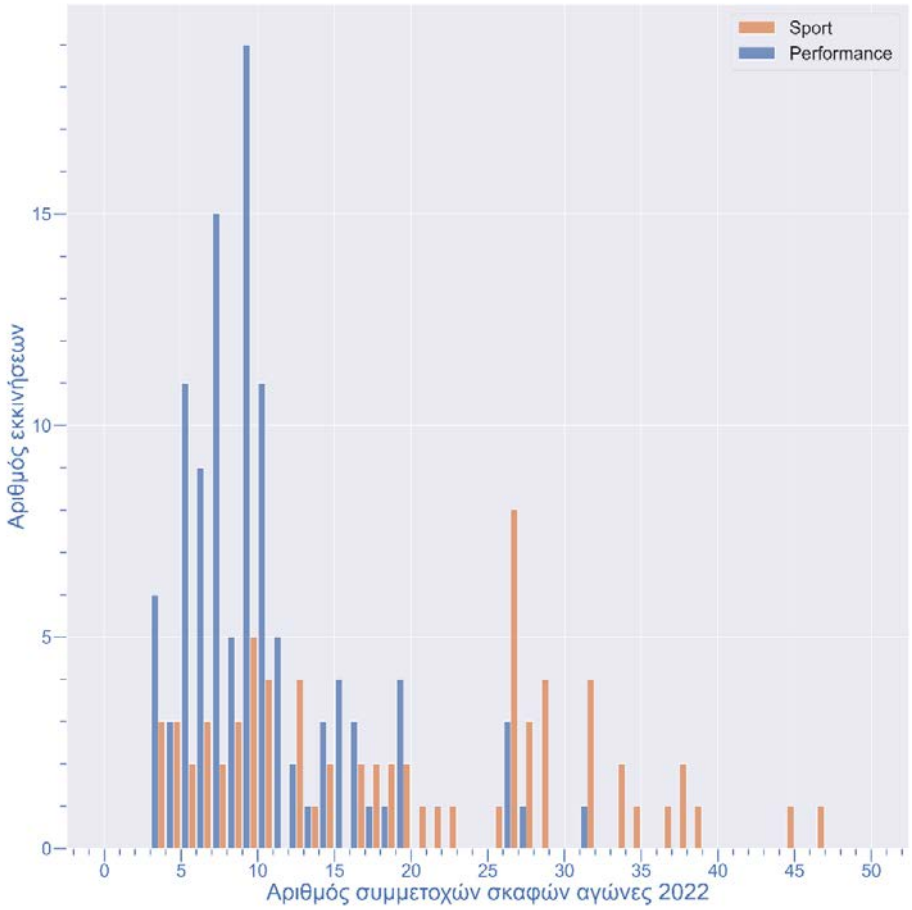
a) Τίποτα! μας πήρε αρκετό καιρό και ατελείωτες διαπραγματεύσεις για να ορίσουμε τις δύο κατηγορίες Performance και Sport, και δεν θέλουμε να ξαναρχίσουμε ανοίγοντας τον ασκό του Αιόλου.
b) Η τιμή της DA=0,23% δεν μας αρέσει και βαράμε με ένα βελάκι σε μια ευθεία από το 0 μέχρι το 1, και διαλέγουμε την τιμή που θα πάει ! Θα έχει την ίδια αξία και νομιμότητα ως προς την επιλογή της, με την σημερινή τιμή!
c) Εάν κάναμε αυτό που ορίζει η ORC και οι Ιταλοί, θα είχαμε 85% των σκαφών στην ομάδα 1 και 15% στην ομάδα 2, Εικόνα 10.




***Εικόνα 3 : CDL, στόλος του 2023 και οι κατηγορίες A,B,C, κόκκινες γραμμές βάσει ORC και D από την Ιταλία. Αριστερά η κλίμακα του αθροίσματος των ποσοστών, καφέ γραμμή, ανά 0,2 μονάδες CDL.***

[1] Χρησιμοποιώντας μεθόδους Machine Learning, συγκρίνοντας καθαρά αγωνιστικά σκάφη, με τον υπόλοιπο στόλο, ίσως να μπορούσαμε να ορίσουμε ένα κριτήριο αγωνιστικότητας και να χωρίσουμε τον στόλο βάσει αυτού. Η πρώτη προσπάθεια είναι σε εξέλιξη.

Μιας και οποιαδήποτε τιμή για την DA ή άλλη μεταβλητή διαλέξουμε θα είναι πάντα αυθαίρετη, θα προτείνω ένα άλλο κριτήριο. Περισσότερο κοινωνικό. Στην κατηγορία Performance σήμερα έχουμε το 23% του στόλου, με αποτέλεσμα σε πολλούς αγώνες να έχουμε σχετικά λίγα σκάφη Performance και πολύ περισσότερα Sport. Στην Εικόνα 11 βλέπουμε ότι στη μεγάλη πλειοψηφία των αγώνων στην κατηγορία Performance τρέχουν λιγότερα από 10 σκάφη. Δύσκολα γίνονται υποκατηγορίες σε αυτή την κατηγορία. Επιπλέον, σκάφη με παραπλήσιες επιδόσεις ταχύτητας, και δίπλα δίπλα στους αγώνες δεν συναγωνίζονται, γιατί τρέχουν σε δύο διαφορετικές κατηγορίες λόγω της τιμής του DA τους. Πιστεύω ότι ένας αγώνας είναι πιο ενδιαφέρον όταν έχουμε και πολλούς παρόμοιους αντιπάλους. Σημειωτέων αυτή η διαφορά στον αριθμό των σκαφών επηρεάζει την γενική κατάταξη ανοιχτής θαλάσσης στην οποία ανακατεύουμε και τις δύο κατηγορίες, δίνοντας περισσότερο βάρος στα σκάφη Sport[2]. Γιατί λοιπόν να μην χωρίσουμε τον στόλο στα δύο χρησιμοποιώντας την διάμεσο της κατανομής της DA (πράσινη γραμμή στην **Εικόνα 9,** DA=0,37%), αν επιμένουμε να χρησιμοποιήσουμε αυτήν την μεταβλητή; Με αυτόν τον τρόπο στατιστικά σε κάθε αγώνα θα έχουμε ισάριθμα σκάφη ανά κατηγορία.



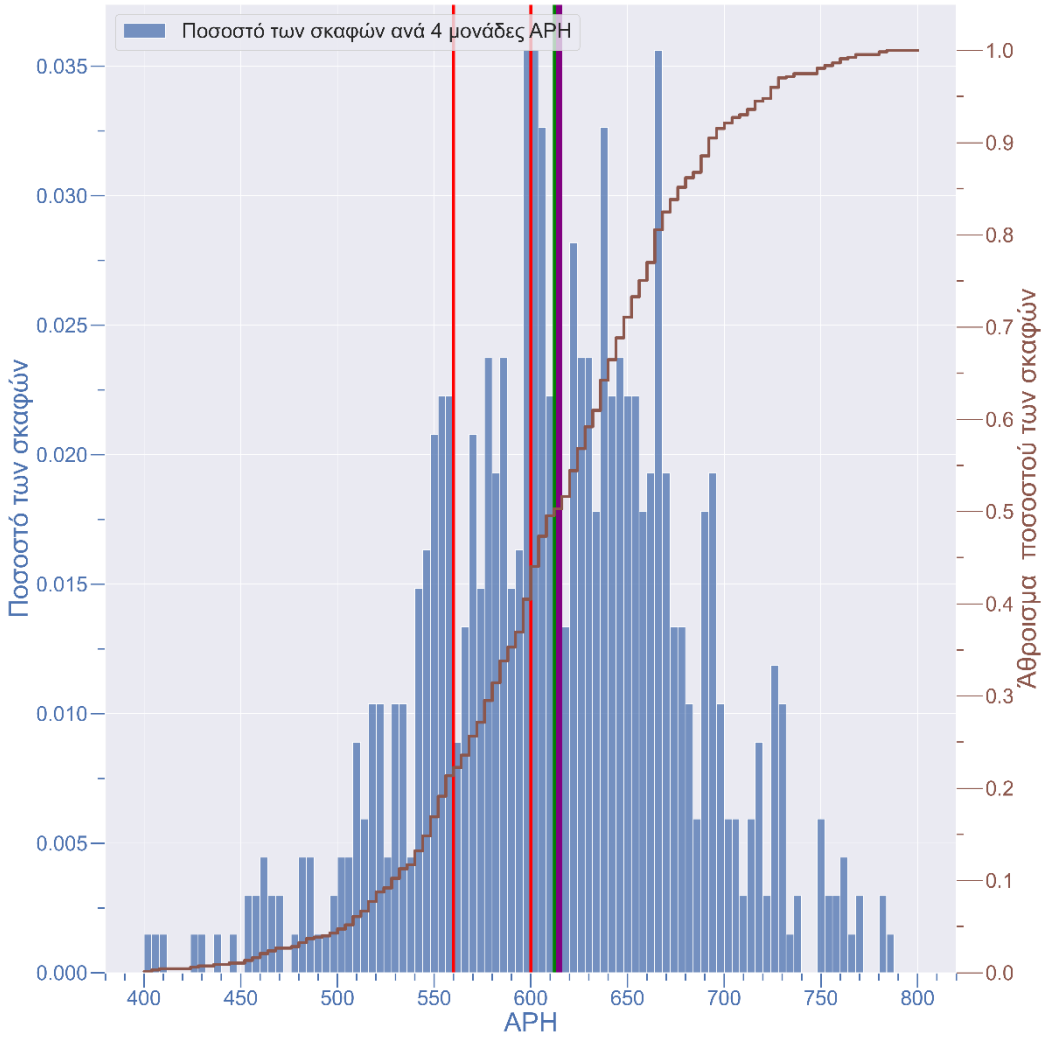

***Εικόνα 4: Αριθμός συμμετοχών, στην κατηγορία Performance. Τρέχουν συνήθως λιγότερα από 10 σκάφη***

Ίσως όμως να μπορούσαμε να κάνουμε κάτι καλύτερο. Κοιτώντας την κατανομή του APH (All-Purpose Handicap), ***Εικόνα 12***, βλέπουμε ότι είναι πιο ομαλή από αυτή της DA, μοιάζει περισσότερο με μια Gaussian (μια καμπάνα!) και αρέσει πολύ περισσότερο στους στατιστικολόγους! Η διάμεσος και η

[2] Ο πρώτος στην κατηγορία Performance που τρέχουν 10 σκάφη πχ, θα πάρει λιγότερους πόντους από τον πρώτο στην κατηγορία Sport που τρέχουν 30 σκάφη

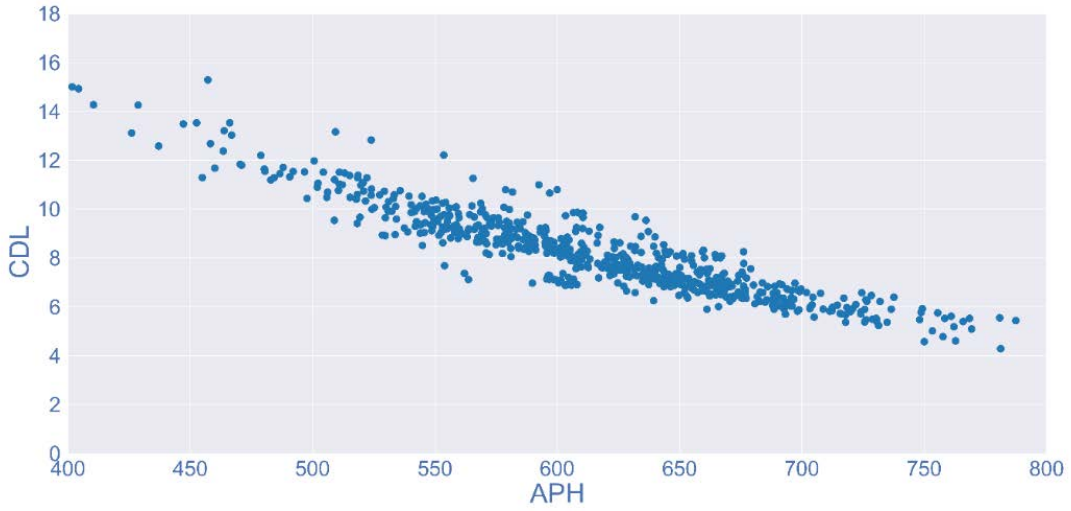
μέση τιμή συμπίπτουνε (μωβ και πράσινη γραμμή στην ***Εικόνα 12***, APH = 613) και έτσι είναι πιο ενδιαφέρουσα μεταβλητή για να χωρίσουμε τον στόλο στα δύο ή και περισσότερα ισάριθμα σύνολα. Αυτό έκανε και ο ΣΕΑΝΑΤΚ που οργάνωσε το τελευταίο Πανελλήνιο διμελούς πληρώματος, ξεχνώντας Performance και Sport. Χώρισε τον στόλο σε τρείς κατηγορίες με ισάριθμα σκάφη, με βάση το APH (APH<560, 560≤APH≤600 και APH>600, κόκκινες γραμμές στην ***Εικόνα 12***). Ανακηρύχτηκε ένας νικητής και νικητές ανά κατηγορία.

Θα είχαμε το ίδιο αποτέλεσμα, δύο κατηγορίες με ισάριθμα σκάφη, εάν χρησιμοποιούσαμε την CDL και ορίζαμε Sport την κατηγορία D των Ιταλών (CDL ≤ 8.) και Performance τις κατηγορίες A+B+C (CDL>8.). Επιπλέον θα εφαρμόζαμε ότι γίνεται και αλλού, συμπεριλαμβανομένων και των άλλων ελληνικών περιφερειών εκτός Σαρωνικού, χωρίς να ανακαλύπτουμε τον τροχό. Επιπλέον η CDL είναι συνάρτηση της ταχύτητας της γάστρας του σκάφους. Έτσι σκάφη με παρόμοιες ταχύτητες βρίσκουν τον ίδιο αέρα σε ένα αγώνα, και συναγωνίζονται στον στίβο γάστρα με γάστρα όπως συχνά συμβαίνει σήμερα. Δεν συναγωνίζονται όμως για την τελική κατάταξη, όταν λόγω του DA τους είναι σε διαφορετικές κατηγορίες.



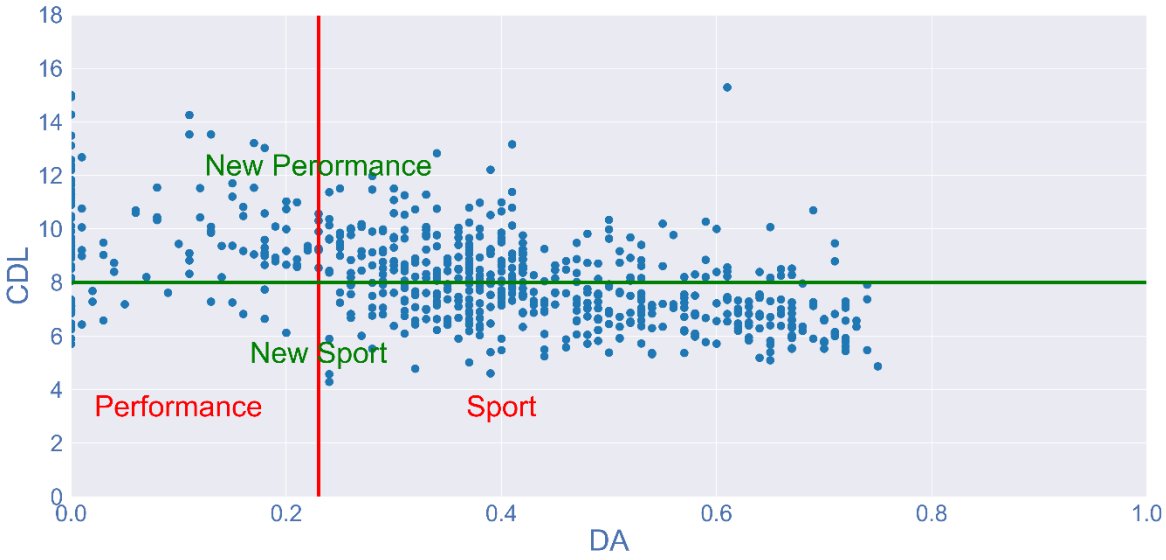

**Εικόνα 5 : Κατανομή του APH, στόλος του 2023. Οι κόκκινες γραμμές ορίζουν τις τρεις υποκατηγορίες στο Πανελλήνιο διμελούς πληρώματος το 2023**

Σημειώστε ότι είτε το APH χρησιμοποιήσουμε είτε την CDL, είναι το ίδιο πράγμα σε μια πρώτη προσέγγιση, μιας και οι δύο μεταβλητές αντί συσχετίζονται. Όσο πιο μεγάλο είναι το APH τόσο πιο μικρή είναι η CDL όπως δείχνει η **Εικόνα 13**.




***Εικόνα 6 : APH σε σχέση με CDL, στόλος του 2023. Οι δύο μεταβλητές αντί συσχετίζονται***

Στην **Εικόνα 14** βλέπουμε την DA σε συνάρτηση με την CDL, και τις σημερινές κατηγορίες (με κόκκινο) και τις νέες προτεινόμενες κατηγορίες βασιζόμενες στην CDL (με πράσινο).




***Εικόνα 7 : DA versus CDL, η κόκκινη γραμμή χωρίζει τον στόλο κάθετα σε δύο κατηγορίες με την DA και η πράσινη γραμμή τον χωρίζει οριζόντια, με την CDL.***

Συμπέρασμα. Οποιαδήποτε τιμή για την DA ή άλλη μεταβλητή διαλέξουμε για να χωρίσουμε τον στόλο σε κατηγορίες, θα είναι πάντα αυθαίρετη. Ο διαχωρισμός θα πρέπει να είναι αμερόληπτος και να βασίζεται σε ένα κριτήριο που εύκολα μπορεί κάποιος να δικαιολογήσει και να εξηγήσει στους αθλητές. Πιστεύω ότι η μέθοδος που ακολούθησε ο οργανωτικός όμιλος στο τελευταίο Πανελλήνιο διμελούς πληρώματος, εξυπηρετεί περισσότερο την αγωνιστική ιστιοπλοΐα. Είναι πολύ πιο ενδιαφέρον, να τρέχουμε σε μια κατηγορία με πολλούς παρόμοιους συναγωνιστές, δίνοντας και την

δυνατότητα για την δημιουργία υπό κατηγοριών, παρά με τρείς και κούκο!! Ο συναγωνισμός και ανταγωνισμός είναι το κίνητρο και γινόμαστε καλύτεροι. Έτσι λοιπόν η πρόταση μου είναι να χωριστεί ο στόλος σε δύο ή περισσότερες ισάριθμες κατηγορίες με βάση το ΑΡΗ ή το CDL, που αντικατοπτρίζουν την ταχύτητα της γάστρας του σκάφους. Οι γρήγοροι θα τρέχουν με τους γρήγορους και οι αργοί με τους αργούς. Το κριτήριο διαχωρισμού του στόλου, είναι η δημιουργία ισάριθμων κατηγοριών. Για τους αγώνες ανοιχτής θαλάσσης κατά την κατάταξη της ΕΑΘ, προτείνω να ανακηρύσσεται ένας νικητής overall και επιμέρους νικητές ανά κατηγορία. Έτσι θα ξέρουμε ποιος κέρδισε την Ύδρα ή το Ράλλυ Αιγαίου!

Κλείνοντας ας θυμηθούμε λίγο ιστορία. Οι Ίσθμιοι Αγώνες, τα Ίσθμια, διεξάγονταν το έτος πριν και το έτος μετά τους Ολυμπιακούς, προς τιμήν του Μελικέρτη-Παλαίμωνα, που πνίγηκε στη θάλασσα μαζί με τη μητέρα του Ινώ. Οι νικητές έπαιρναν ένα στεφάνι από πεύκο και όχι αγριελιάς! Έτσι οι ιστιοπλόοι δανείστηκαν το όνομα, και έφτιαξαν μια κατηγορία σκαφών, Κορινθιακή, Corinthian, που τρέχουν οικογενειάρχες συν γυναιξί και τέκνοις! Τρέχουν σε αυτή την κατηγορία, αυτοί που θέλουν να δουν, να μάθουν, να συναγωνιστούν με παρόμοιους. Κοινώς ερασιτέχνες ! Γιατί λοιπόν παράλληλα με τις αγωνιστικές κατηγορίες, σε με μίαν Ύδρα ή έναν Πόρο να μην έχουμε και μια Κορινθιακή, ανοιχτή σε όποιον θέλει, με πιστοποιητικό ή όχι, με αθλητικές ταυτότητες ή όχι, απαλλαγμένη από τις γραφειοκρατικές διαδικασίες; Μια κατηγορία, σκαλοπάτι για τις αγωνιστικές, μύηση στην αγωνιστική ιστιοπλοΐα. Ίσως είναι ένας τρόπος να τραβήξουμε περισσότερο κόσμο στους αγώνες. Τρώγοντας έρχεται η όρεξη. Τώρα αν οι Ίσθμιοι αγώνες, θα μετρούν για το ιερό δισκοπότηρο της βεβαίωσης ναυταθλητικού, αφήνω άλλους να αποφασίσουν!

ΥΓ-α. Η MIMOSA τρέχει σήμερα στην κατηγορία Performance (DA=0,21%) και εάν αλλάζαμε τις κατηγορίες χρησιμοποιώντας την CDL πάλι στην κατηγορία Performance θα έτρεχε (CDL=8,867), και δεν σκοπεύουμε με κανένα τρόπο να αλλάξουμε την παρέα μας.

ΥΓ-β. Ευχαριστώ τους διάφορους φίλους που με τα σχόλια τους εμπλούτισαν το κείμενο μου.

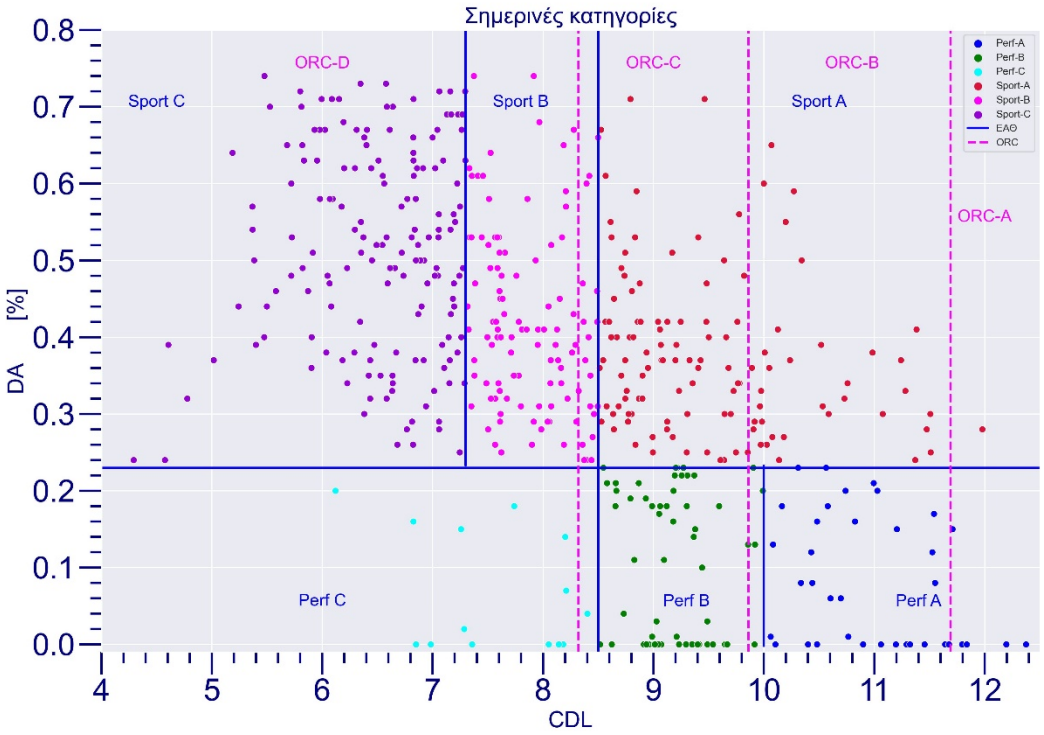
## Βιβλιογραφία


1. Παρουσίαση συστημάτων καταμέτρησης και ισοζυγισμού https://www.offshore.org.gr/docs/RatingSystems2015.pdf Γιάννης Καλατζης και για το 2023 στα αγγλικά https://orc.org/uploads/files/Rules-Regulations/2023/ORC-Rating-Systems-2023-PDF.pdf
2. VPP ORC Documentation https://orc.org/uploads/files/ORC-VPP-Documentation-2023.pdf Στην σελίδα 20 παράγραφο 3.6 θα βρείτε πως ορίζεται η DA, στην σελίδα 83 και παράγραφο 8.2.4 θα βρείτε πως ορίζεται η CDL.
3. Πανελλήνια συνάντηση καταμετρητών ΕΑΘ 2015 https://offshore.org.gr/news/nw193/HOC_Measurers_Meeting_Athens_2015.pdf Στην σελίδα 37 θα βρείτε πληροφορίες για την DA
4. Όλα τα δεδομένα που χρησιμοποίησα προέρχονται από την σελίδα της ΕΑΘ https://www.offshore.org.gr/

# Δημιουργούμε αγωνιστικές κατηγορίες με ML

Γιάννης Καρυωτάκης
S/Y Mimosa
Πέμπτη, 22 Αυγούστου 2024

Σήμερα στον Σαρωνικό διαχωρίζουμε τον στόλο των σκαφών που λαμβάνουν μέρος σε αγώνες, σε 6 κατηγορίες χρησιμοποιώντας σταθερές τιμές των DA και CDL, Εικόνα 15. Η ORC προτείνει τρεις κατηγορίες βασισμένες στην τιμή του CDL και στην Ιταλία προσθέτουν και μια τέταρτη. Σε άλλες περιφέρειες στην Ελλάδα χρησιμοποιούμε μόνο την CDL. Μπορεί η τεχνητή νοημοσύνη (AI) ή μάλλον Machine Learning (ML), να μας βοηθήσει να δημιουργήσουμε αμερόληπτα, κατηγορίες σκαφών με παρόμοια χαρακτηριστικά; Η διαφορετικά, αντί να χρησιμοποιούμε μόνο δύο χαρακτηριστικά των σκαφών μας, DA και CDL, μπορούμε να κάνουμε ομάδες σκαφών με παρόμοια χαρακτηριστικά πχ, LOA, εκτόπισμα, DA, APH, CDL, εμβαδόν πανιών, κτλ; Και αντί να κατατάσσουμε σε δύο διαστάσεις, να κατατάσσουμε σε ν διαστάσεις, όπου ν είναι ο αριθμός χαρακτηριστικών που θέλουμε;

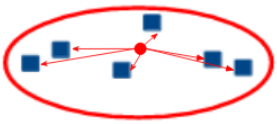



***Εικόνα 8**: Σημερινές κατηγορίες, κάθε σημείο είναι και ένα σκάφος*

Το πρόβλημα μας είναι γνωστό στην ML, και λέγεται clusterisation, ομαδοποίηση[3]. Πχ μια τράπεζα θέλει να φτιάξει συστάδες (clusters) πελατών ανάλογα με το εισόδημα τους, την ηλικία τους και την περιουσία τους, για να δει ποιοι θα μπορούσαν να πάρουν ποια πιστωτική κάρτα. Για να το λύσουμε έχουμε στην διάθεση μας ελεύθερο λογισμικό (software), και ο αλγόριθμος ή το μοντέλο όπως λέμε,

[3] Ένας απλός ορισμός για την ομαδοποίηση ή συσταδοποίηση ή συσσωμάτωση (clustering): "Ομαδοποίηση ονομάζεται η διαδικασία που οργανώνει πρότυπα (παρατηρήσεις, δεδομένα) σε ομάδες (συστάδες-clusters), όπου τα μέλη μιας ομάδας είναι παρόμοια μεταξύ τους σύμφωνα με κάποιο κριτήριο".

που δημιουργεί ομάδες σε ένα σύνολο δεδομένων, ένα από τα πιο δημοφιλή λέγεται K-mean[4]. Βασίζεται σε σοβαρή μαθηματική θεωρία και προσπαθεί να βρει κέντρα (centroids) στα δεδομένα μας γύρω από τα οποία η 'απόσταση' μιας ομάδας ή συστάδας (cluster) του συνόλου μας είναι ελάχιστη, βλέπε την εικόνα δίπλα.



Intra cluster distance

Πως εφαρμόζουμε αυτό το μοντέλο στον στόλο μας; Διαβάζοντας από την ΕΑΘ όλα τα πιστοποιητικά, δημιουργούμε ένα σύνολο σκαφών που για το κάθε ένα ξέρουμε όλες τις τιμές που αναγράφονται στο πιστοποιητικό του: όνομα, βάρος του πληρώματος, μήκος, πλάτος, βύθισμα, εκτόπισμα, GPH, DA, CDL, εμβαδά πανιών, που θα πω βασικά χαρακτηριστικά και εισάγουμε στο μοντέλο. Αποφεύγουμε να συμπεριλάβουμε δύο μεταβλητές άκρως συσχετισμένες όπως πχ η CDL και APH. Κατόπιν έχουμε όλο το απαραίτητο λογισμικό να δημιουργήσουμε τις ομάδες ή συστάδες ή τις κατηγορίες μας. Επίσης έχουμε όλα τα εργαλεία να ελέγξουμε εάν το αποτέλεσμα έχει ή δεν έχει νόημα. Έτσι ενδεικτικά χρησιμοποιώντας όλα τα ελληνικά πιστοποιητικά του 2023, ο K-mean μας λέει ότι ο βέλτιστος αριθμός κατηγοριών για τα περίπου 700 σκάφη μας, είναι από τρεις μέχρι έξι. Το μοντέλο, K-mean, στην περίπτωση μας, αφού εκτελεσθεί, θα καταλογίσει σε κάθε σκάφος μια κατηγορία ανάλογα με την απόσταση του σκάφους μας, σε ν διαστάσεις, από ένα κέντρο. Διαλέγοντας να δημιουργήσουμε τρεις κατηγορίες, στην **Εικόνα 16** βλέπουμε με διαφορετικά χρώματα τα σκάφη κάθε κατηγορίας σε ένα επίπεδο δύο μεταβλητών που διαλέγουμε, πχ DA versus CDL. Κάθε σκάφος είναι ένα σημείο και το χρώμα του υποδεικνύει την κατηγορία του. Στην **Εικόνα 17** έχουμε το ίδιο διάγραμμα αλλά δημιουργώντας πέντε κατηγορίες. Δεν είναι δύσκολο να δούμε ότι η διαίρεση σε κατηγορίες που επιτυγχάνει το μοντέλο K-mean βασίζεται περισσότερο στην CDL παρά την DA. Υπενθυμίζω ότι η CDL εξαρτάται από το μήκος του σκάφους και την ταχύτητα του, Beat VMG, σε αέρα 12 κόμβων. Φυσικά ο διαχωρισμός δεν εξαρτάται μόνο από τις δύο αυτές μεταβλητές που δείχνουν οι εικόνες αλλά από όλες και δεν διαχωρίζουμε το σύνολο χρησιμοποιώντας σταθερές τιμές μόνο δύο χαρακτηριστικών όπως σήμερα, αλλά όλα τα βασικά χαρακτηριστικά του σκάφους. Στην **Εικόνα 18** έχουμε ενδεικτικά το διάγραμμα CDL versus APH. Ο αλγόριθμος K-mean δημιουργεί καθαρά ευδιάκριτες κατηγορίες στο επίπεδο CDL versus APH ενώ οι σημερινές κατηγορίες επικαλύπτονται, που σημαίνει ότι σκάφη ίδιων ταχυτήτων τρέχουν σε διαφορετικές κατηγορίες. Στην **Εικόνα 19** πάνω έχουμε την κατανομή της CDL μετά από την ομαδοποίηση λαμβάνοντας υπόψιν τα βασικά χαρακτηριστικά. Βλέπουμε ότι η επικάλυψη είναι μικρή. Και εάν αποφασίζαμε να δημιουργήσουμε κατηγορίες εισάγοντας στον αλγόριθμο μόνο την τιμή της CDL, δεν θα είχαμε φυσικά καμμιά επικάλυψη, **Εικόνα 19** κάτω. Στην **Εικόνα 20** βλέπουμε την κατανομή της DA με τις ίδιες συνθήκες όπως την CDL.

Για να εκτελέσουμε τον αλγόριθμο K-mean πρέπει να του δώσουμε τον αριθμό των ομάδων που θέλουμε να δημιουργήσουμε. Πως τον διαλέγουμε; Έχουμε στην διάθεση μας, δύο μαθηματικούς δείκτες, Inertia και Silhouette Score[5], που μετράνε για όλα τα σημεία μας (τα σκάφη μας) την απόσταση τους σε ν διαστάσεις από το κέντρο της ομάδας και μεταξύ τους, με κάποια μαθηματική μέθοδο. Όσο πιο πολλές ομάδες δημιουργήσουμε τόσο η απόσταση μικραίνει, αλλά από έναν αριθμό ομάδων και πάνω παραμένει σταθερή. Η τιμή του αριθμού των ομάδων όπου οι δείκτες μας αρχίζουν να σταθεροποιούνται είναι η προτιμότερη τιμή των κατηγοριών που πρέπει να δημιουργήσουμε. Επιτυγχάνουμε έναν καλό διαχωρισμό σε κατηγορίες όταν η τιμή των δεικτών μας είναι μίνιμουμ (ή μάξιμουμ) και ο αριθμός των κατηγοριών μικρός. Από τις **Εικόνα 22** και **Εικόνα 23**, βλέπουμε ότι ο βέλτιστος αριθμός κατηγοριών είναι μεταξύ τρεις και έξι. Στην **Εικόνα 21**, έχουμε το αποτέλεσμα εάν δημιουργούσαμε έξι κατηγορίες, χρησιμοποιώντας όπως κάνουμε σήμερα μόνο δύο χαρακτηριστικά,

[4] Βλέπε πχ εδώ https://scikit-learn.org/stable/modules/clustering.html#k-means
[5] Βλέπε πχ εδώ https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

DA και CDL. Το μοντέλο ομαδοποιεί τελείως διαφορετικά τα σκάφη, από την σημερινή επιλογή, που δείχνουν οι μπλε γραμμές.

Αντί να εισάγουμε βασικά μεγέθη του σκάφους για να δημιουργήσουμε κατηγορίες, μπορούμε να χρησιμοποιήσουμε τις ταχύτητες των σκαφών ανάλογα με την δύναμη και κατεύθυνση του ανέμου όπως υπολογίζονται από το VPP. Τα δεδομένα του ελληνικού στόλου για το 2024 δεν είναι διαθέσιμα σήμερα (Φλεβάρης 2024) και για την άσκηση χρησιμοποιώ τα δεδομένα του ισπανικού στόλου. Μετά από ομαδοποίηση βάσει των ταχυτήτων κοιτάμε πως κατανέμονται τα σκάφη σε ένα επίπεδο CDL versus APH, δημιουργώντας 3 κατηγορίες. Ο αλγόριθμος δημιουργεί κατηγορίες σκαφών κοντινών ταχυτήτων και αυτό φαίνεται καθαρά στο γράφημα CDL versus APH. Στις **Εικόνα 24** και **Εικόνα 25** βλέπουμε την διαφορά των κατηγοριών εάν εισάγουμε στο μοντέλο όλα τα βασικά χαρακτηριστικά των σκαφών ή μόνο τις ταχύτητες. Οι κατηγορίες δεν έχουν σχεδόν καμία επικάλυψη πράγμα που δεν συμβαίνει με την σημερινή ομαδοποίηση βασισμένη στην CDL και DA.

Όλα είναι ρόδινα; Ο αλγόριθμος K-mean έχει ανάγκη για να ξεκινήσει τα αρχικά κέντρα (centroids) των συστάδων. Στην αρχή διαλέγονται στην τύχη και σιγά σιγά με αλλεπάλληλες επαναλήψεις συγκλίνει στα τελικά κέντρα, που γύρω τους οργανώνονται οι συστάδες. Όταν αυτές είναι μακριά η μια με την άλλη τα κέντρα δεν θα αλλάξουν όταν ξανατρέξουμε τον αλγόριθμο και θα διαλέξει διαφορετικά αρχικά κέντρα. Δυστυχώς δεν είναι η περίπτωση μας. Εάν τρέξουμε τον αλγόριθμο 2000 φορές τα κέντρα θα μετακινηθούν λίγο και ορισμένα σκάφη στα όρια των συστάδων μπορεί να αλλάξουν κατηγορία, **Εικόνα 26**.

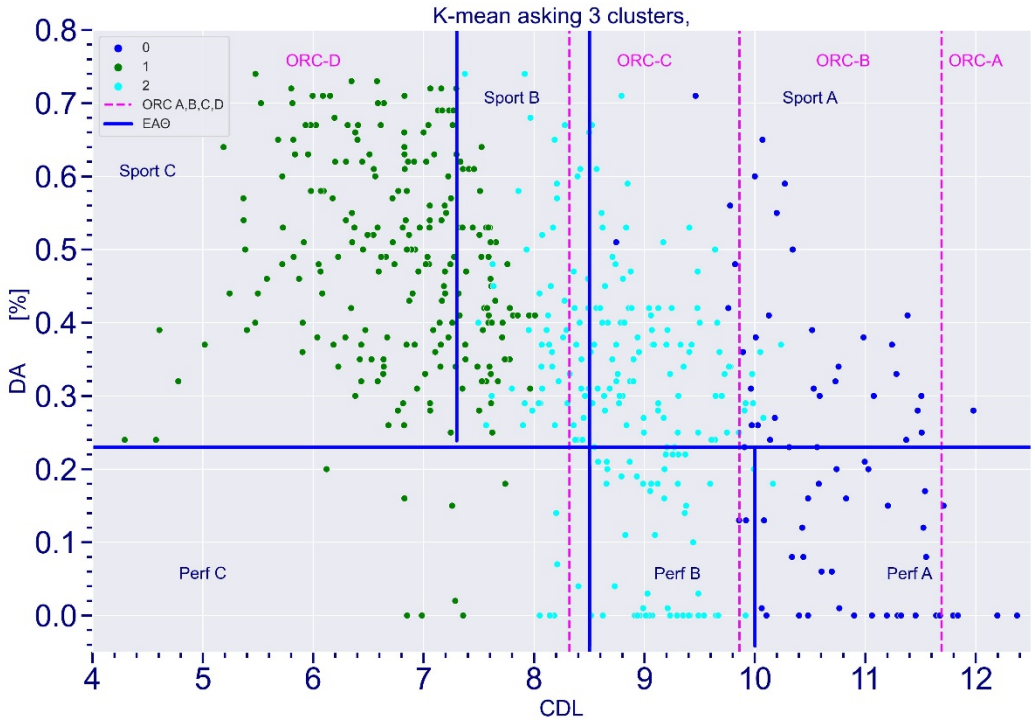
Η ομαδοποίηση δεν είναι ο μόνος τρόπος για να δημιουργήσουμε κατηγορίες με αλγόριθμους Machine Learning. Ένας δεύτερος είναι η κατηγοριοποίηση, Classification. Με αυτήν την μέθοδο ξεκινάμε από ένα σύνολο σκαφών που ξέρουμε καλά ποιο ανήκει σε ποια κατηγορία. Με αυτό το σύνολο εκπαιδεύουμε το μοντέλο μας, δηλαδή του μαθαίνουμε να αναγνωρίζει ποιος είναι ποιος, και μετά το χρησιμοποιούμε για να προβλέψουμε σε ποια κατηγορία ανήκουν τα σκάφη ενός διαφορετικού στόλου. Χρησιμοποιώντας τα πιστοποιητικά σκαφών ξένων χωρών[6] δημιουργούμε το πρώτο σύνολο, training sample. Η δυσκολία είναι να ορίσουμε σε ποια κατηγορία ανήκει το κάθε σκάφος του ξένου στόλου. Για την άσκηση διάλεξα να χρησιμοποιήσω τον ορισμό IMS, Racer/Cruiser, δύο κατηγορίες που ονόμασα Performance και Sport. Κατόπιν επέλεξα διάφορα μοντέλα, Random Forest, Gradient-Boosted Trees[7], κτλ. που τα εκπαίδευσα με τα σκάφη της Ολλανδίας, Ιταλίας, Γαλλίας, Αμερικής και Αυστραλίας, πάντα χρησιμοποιώντας όλα τα βασικά μεγέθη που αναφέρονται στο πιστοποιητικό τους. Ο ορισμός , Racer/Cruiser είναι επίσης διαθέσιμος για τα ελληνικά σκάφη και μπορούμε να συγκρίνουμε την απόφαση του μοντέλου για τα ελληνικά σκάφη με αυτή του IMS. Όλα τα μοντέλα πρόβλεψαν με ποσοστό επιτυχίας 92-96% σε ποια κατηγορία ανήκουν τα ελληνικά σκάφη, όπως μας δείχνει ο πίνακας σύγχυσης στην **Εικόνα 27**. Η άσκηση μας δείχνει ότι εάν είχαμε ένα σύνολο σκαφών που γνωρίζουμε με ακρίβεια σε ποια κατηγορία ανήκει ένα σκάφος, είναι δυνατόν να προβλέψουμε τις κατηγορίες για ένα άλλο σύνολο. Μη έχοντας ένα τέτοιο σύνολο για τα ξένα σκάφη, προτίμησα την πρώτη μέθοδο, την ομαδοποίηση.

Η άσκηση που έκανα σε αυτό το κείμενο **είναι ενδεικτική** και θα χρειαζόταν αρκετή δουλειά και πιο συστηματική μελέτη από κάποιον με καλό υπόβαθρο σε προβλήματα ML και ιστιοπλόο με γνώση των σκαφών, πριν βρούμε την καλύτερη λύση. Ποιες είναι οι καλύτερες και οι πιο ευαίσθητες μεταβλητές που δίνουμε στο μοντέλο; Πόσο σταθερό είναι το αποτέλεσμα με διαφορετικά σύνολα μεταβλητών ή μεταβάλλοντας τα αρχικά κέντρα, K-mean είναι το καλύτερο μοντέλο; και τόσες άλλες ερωτήσεις.

[6] Όλα τα δεδομένα που χρησιμοποιώ εδώ προέρχονται από την ORC https://orc.org/race-managment/rms-files

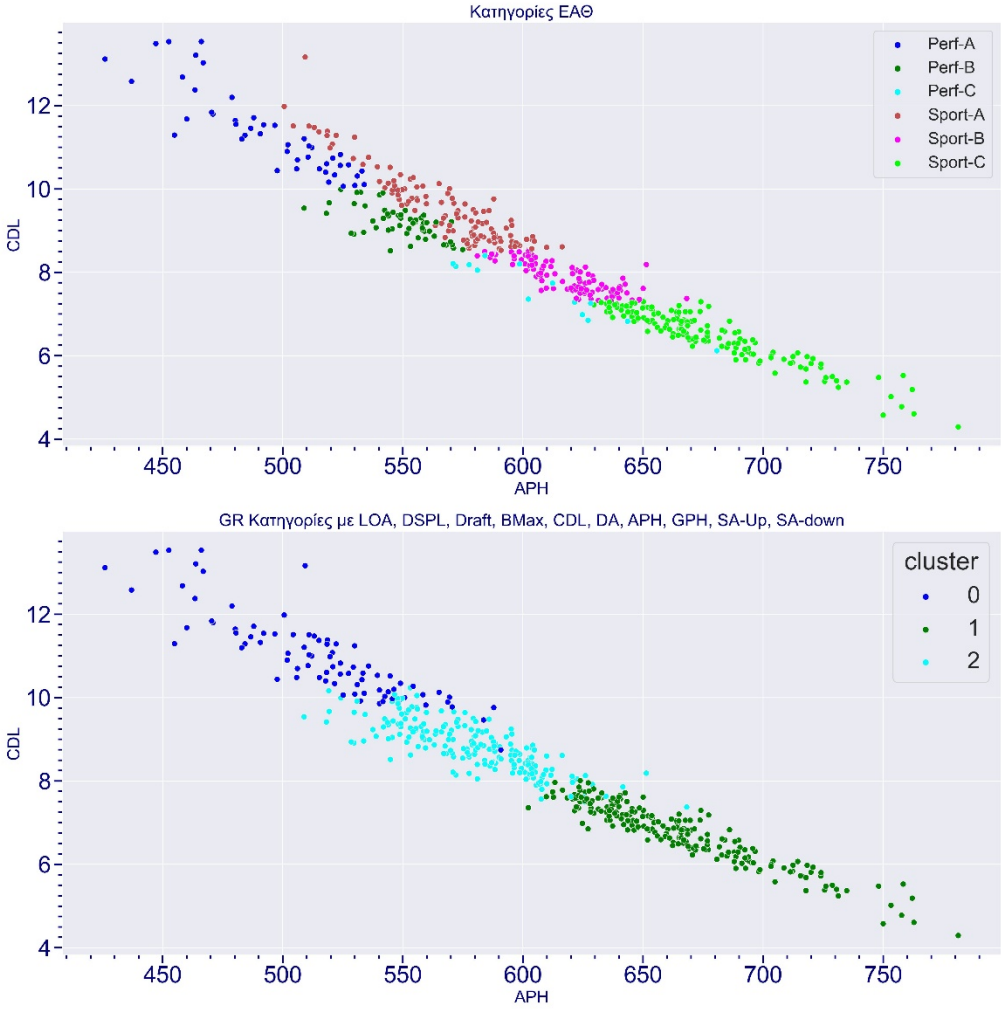
[7] Βλέπε πχ εδώ https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting

Δείχνει όμως ότι το πρόβλημα έχει μάλλον λύση, ίσως όχι μοναδική! Δημιουργώντας κατηγορίες με ένα τέτοιο μοντέλο ML, είμαστε σίγουροι ότι είναι αμερόληπτες. Επιπλέον κρατώντας το μοντέλο ‘μυστικό’, δεν θα μπορεί ο καθένας να μεταβάλει την ιστιοφορία του πχ με σκοπό να τρέχει σε άλλη κατηγορία!

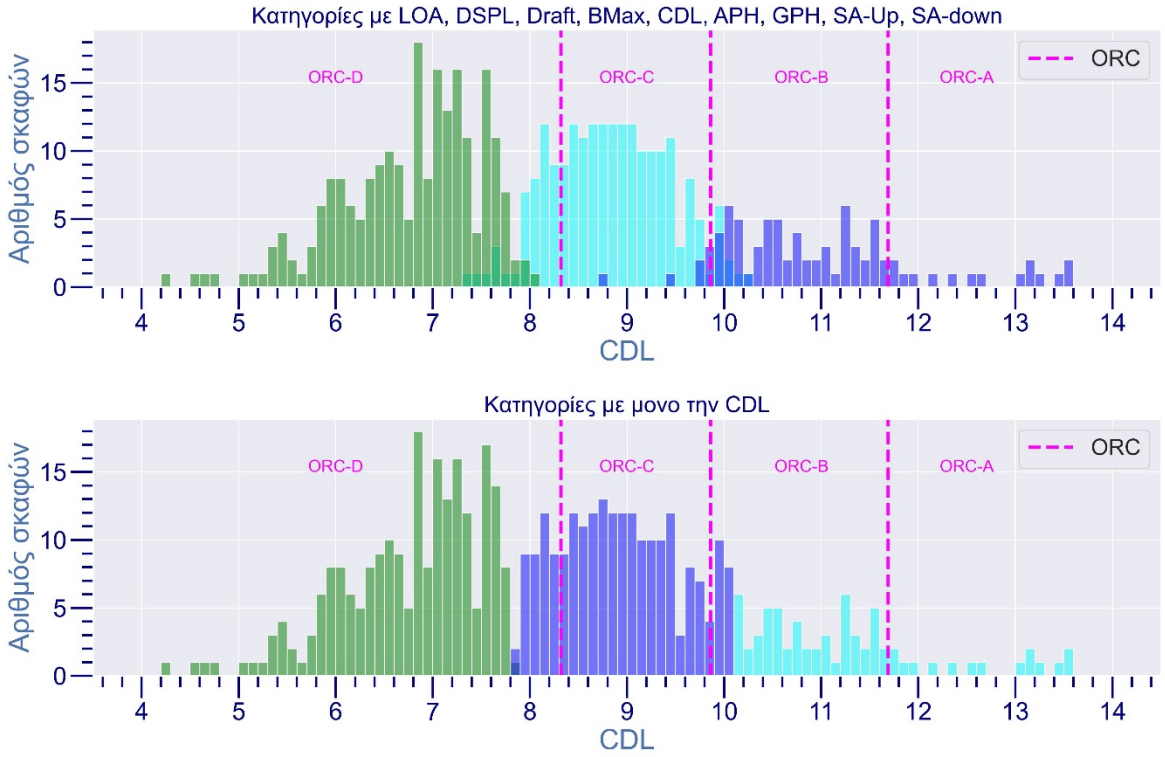



***Εικόνα 9*** *: Συσσωμάτωση K-Mean με 3 κατηγορίες*

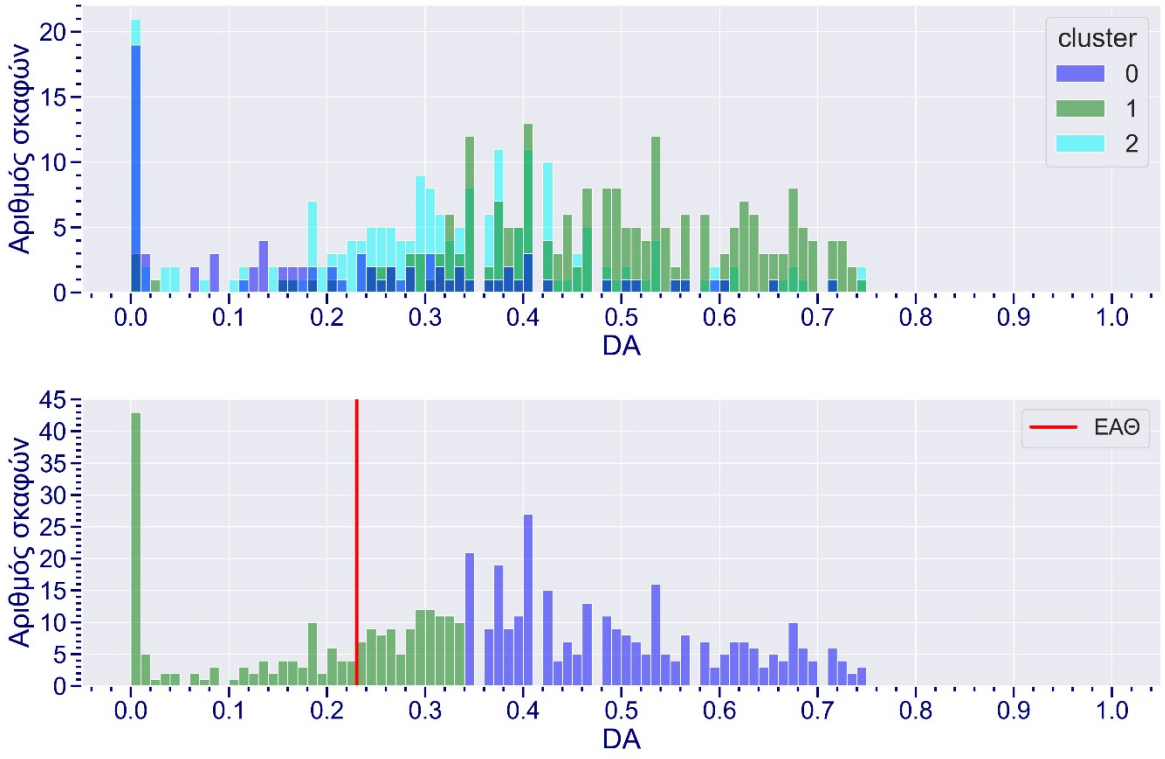

***Εικόνα 10*** *: Συσσωμάτωση K-Mean με 5 κατηγορίες*



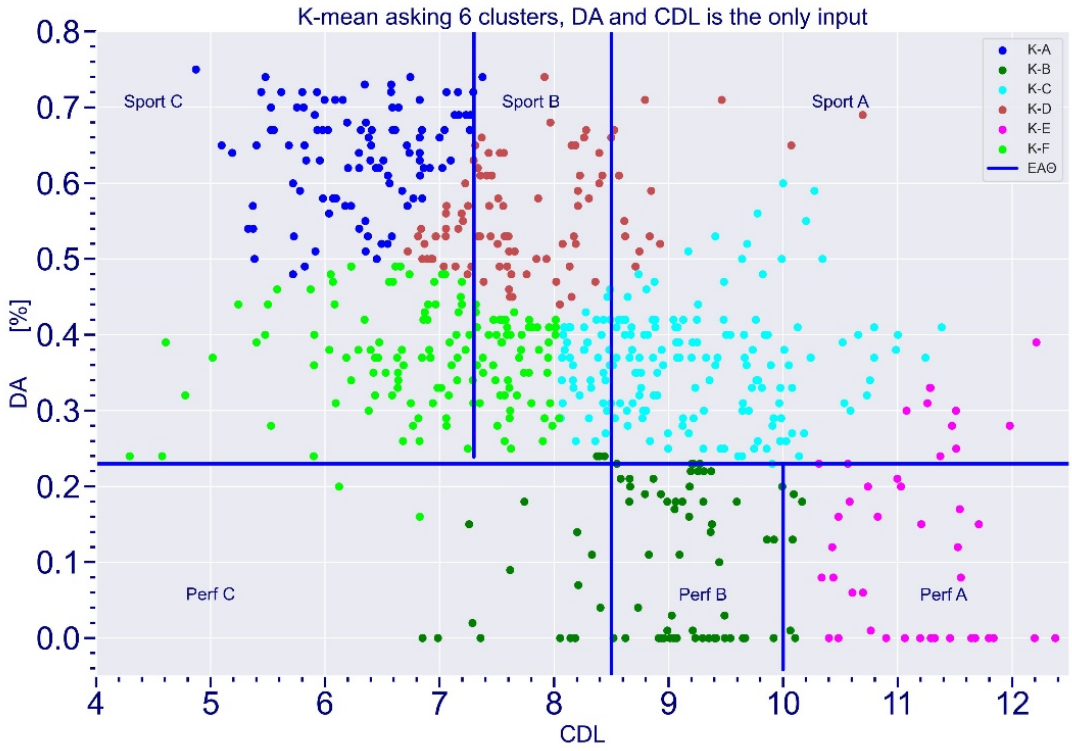

***Εικόνα 11*** *: Συσσωμάτωση K-Mean με 3 κατηγορίες, κάτω. Οι σημερινές κατηγορίες, επάνω, επικαλύπτονται. APH versus CDL*




***Εικόνα 12*** *: Συσσωμάτωση K-Mean με 3 κατηγορίες με όλα τα χαρακτηριστικά πάνω και μόνο την CDL κάτω, κατανομή της CDL*

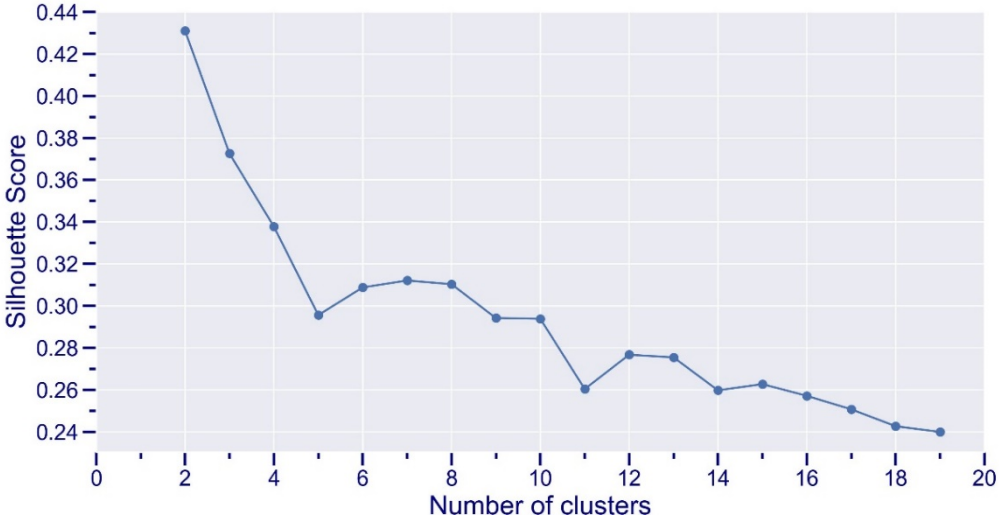



***Εικόνα 13*** *: Συσσωμάτωση K-Mean με όλα τα χαρακτηριστικά και 3 κατηγορίες πάνω, και μόνο την DA και 2 κατηγορίες κάτω, κατανομή της DA*

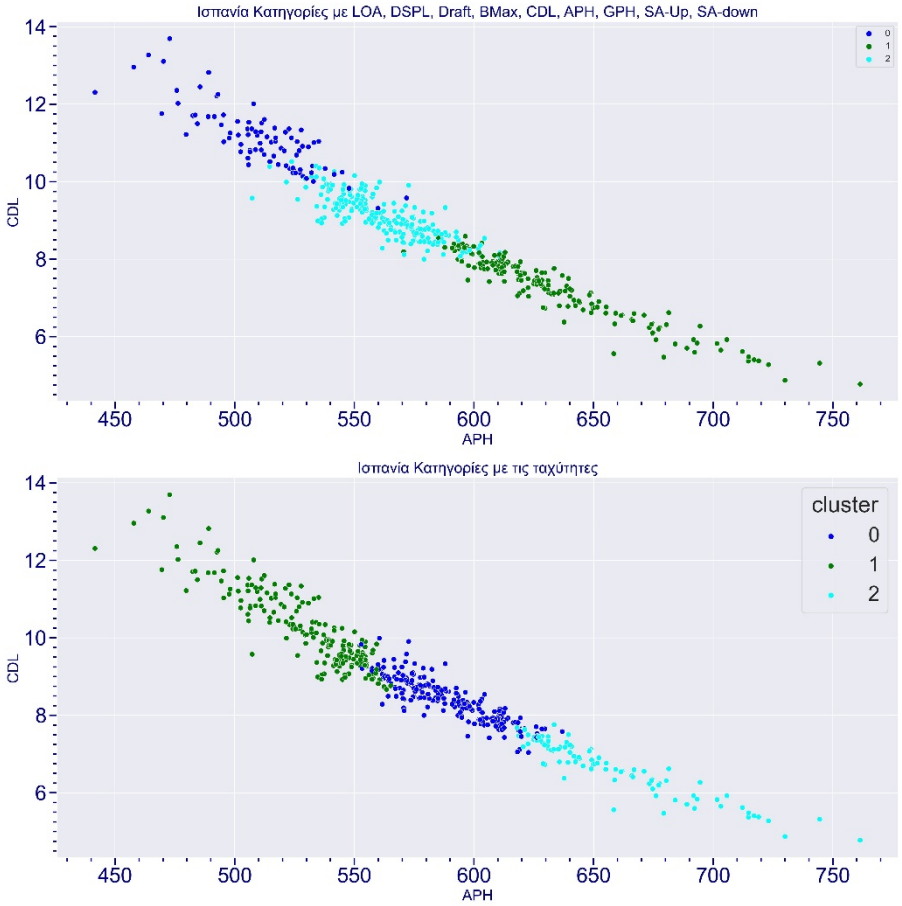



***Εικόνα 14*** *: Συσσωμάτωση K-Mean με 6 κατηγορίες και μόνο 2 χαρακτηριστικά DA, CDL. Οι μπλε γραμμές αντιστοιχούν στις σημερινές κατηγορίες, διαφορετικές από το μοντέλο με τις ίδιες συνθήκες.*

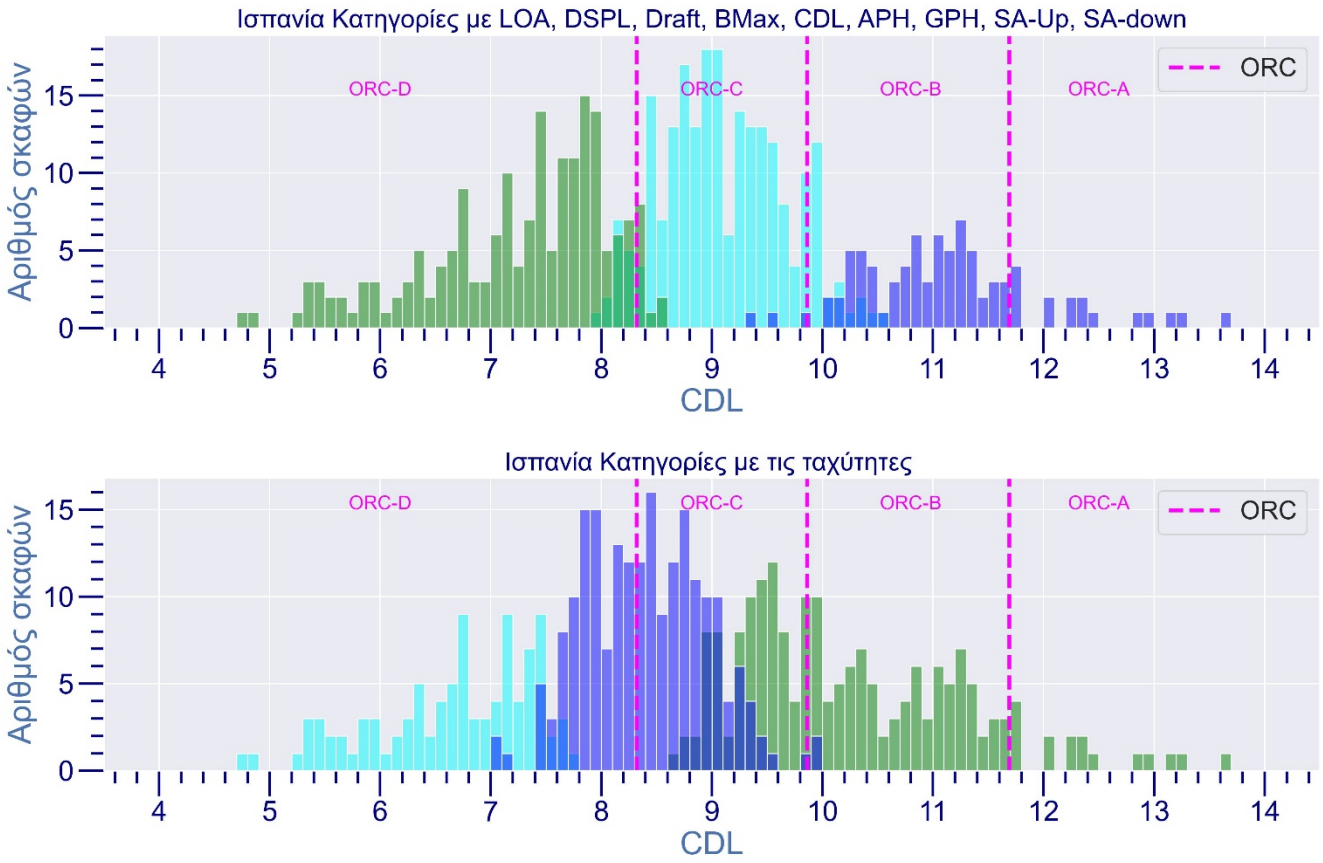

***Εικόνα 15*** *: Μεταβολή της Inertia με τον αριθμό των ομάδων. Βλέπουμε ότι αρχίζει να σταθεροποιείται μεταξύ 4-6 ομάδων*




***Εικόνα 16** : Μεταβολή του Silhouette Score με τον αριθμό των ομάδων.*

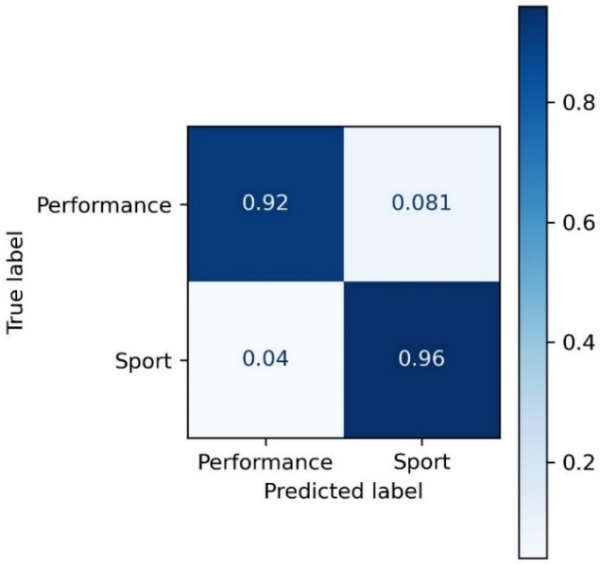



***Εικόνα 17** : Ισπανικός στόλος, Input Time Allowance @ 52°, 90°, 120°, 150°, 6 και 3 κατηγορίες CDL vs APH*




***Εικόνα 18** : Ισπανικός στόλος, Input Time Allowance @ 52°, 90°, 120°, 150°, 3 κατηγορίες όλα τα χαρακτηριστικά πάνω και μόνο τις ταχύτητες κάτω, κατανομή της CDL*


***Εικόνα 19** : Μεταβολή των κέντρων μετά από 2000 εκτελέσεις του αλγόριθμου*




***Εικόνα 20**: Confusion matrix, πίνακας σύγχυσης μοντέλο Gradient-Boosted Tree*