Blue text on a black background

Description automatically generated

***Σχολή Μηχανικών***

***Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών***

*Διαχείριση Δεδομένων Μεγάλης Κλίμακας*

***Αναφορά Ανάλυσης Δεδομένων***

*Εργασία: Best-Value Smartphone Analysis*

*ΕΥΦΡΟΣΥΝΗ ΒΑΡΣΟΥ 21390021*

*ΑΓΓΕΛΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΕΝΤΖΕΛΟΣ 21390132*

# 1. Εισαγωγή

Αυτό το έγγραφο αποτελεί την συγγραφή της αναφοράς για την ανάλυση δεδομένων που πραγματοποιήσαμε σε ένα dataset που αφορά για τα διάφορα κινητά που υπάρχουν στην αγορά και τα αντίστοιχα χαρακτηριστικά τους. Θα επικεντρωθούμε στην ανάλυση σχετικά με την τιμή του κάθε κινητού σε αντιπαράθεση με τον επεξεργαστή που έχουν και την RAM.

Αρχικά, θα περιγράψουμε το πρόβλημα που επιχειρούμε να λύσουμε με την ανάλυση αυτή, έπειτα θα περιγράψουμε εν συντομία το dataset που χρησιμοποιήσαμε και τέλος θα επικεντρωθούμε στην περιγραφή του κώδικα που υλοποιήσαμε καθώς και την περιγραφή/αποτίμηση των αποτελεσμάτων που συμπεραίνουμε από αυτά.

Σε αυτήν την αναφορά δεν θα συμπεριλάβουμε όλα τα περιεχόμενα των πινάκων καθώς είναι πολλά για τις δύο αναλύσεις που υλοποιήσαμε (Processor, RAM) ωστόσο υπάρχουν όλοι αναλυτικά στον κώδικα .ipynb. Ο κώδικας .py περιέχει και αυτός όλη την ανάλυση αλλά συμπεριλήφθηκε για πρακτικούς λόγους, συνεπώς η ανάλυση μαζί με τις πηγές που χρησιμοποιήσαμε βρίσκεται αναλυτικά στον κώδικα .ipynb (θα ήταν προτιμότερο δηλαδή να προβληθεί από εκεί αφού είναι ενσωματωμένοι όλοι οι πίνακες και τα σχήματα).

# 2. Ορισμός Προβλήματος

Επιλέξαμε αυτό το πρόβλημα γιατί θέλαμε να προσεγγίσουμε το θέμα ανάλυσης των πιο value for money smartphones. Γενικά, στις μέρες μας ειδικότερα στο tech world του διαδικτύου όλοι επιχειρούν να απαντήσουν στο ερώτημα ποιο κινητό ή λειτουργικό είναι καλύτερο. Όπως για παράδειγμα, συχνό debate είναι η επιλογή μεταξύ Apple ή Android κινητού. Με τους υποστηρικτές της Apple να υποστηρίζουν ότι αν και πιο ακριβά είναι πιο ποιοτικά και με τους υποστηρικτές των Android να υποστηρίζουν ότι είναι το ίδιο ποιοτικά (αν όχι και περισσότερο) και σε πιο προσιτή τιμή. Στην ανάλυση που θα κάνουμε θα προσπαθήσουμε να δείξουμε σε πιο εύρος τιμών το hardware(processor, RAM) ισοδυναμεί με την τιμή. Αν δηλαδή αυτά που πληρώνεις είναι η «μάρκα» ή όντως η αξία του κινητού σε σύγκριση με τις δυνατότητες του. Και γενικά, ποιο εύρος τιμών είναι το καλύτερο για μια αξιοπρεπής αγορά ενός κινητού. Συνεπώς, ποιο είναι το πιο best value smartphone στην αγορά σήμερα.

# 3. Περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε περιέχει 11 στήλες και 1021 γραμμές που αφορούν σε γενικές πληροφορίες σχετικά με κινητά τηλέφωνα. Έχει γενικά πληροφορίες για κινητά από πολύ φθηνά έως πολύ ακριβά και όλων των λειτουργικών. Από τις 11 στήλες του dataset εμείς χρησιμοποιήσαμε την στήλη Price (την τιμή σε ινδικά ρούπια την οποία μετατρέψαμε σε ευρώ), την Processor (που έχει το όνομα του επεξεργαστή , τους πυρήνες του και στην συχνότητα του) και την RAM (που έχει την μνήμη RAM και τον αποθηκευτικό builitin χώρο). Οι υπόλοιπες στήλες που υπάρχουν στο αρχείο αλλά δεν χρησιμοποιήσαμε είναι γενικά χαρακτηριστικά που αφορούν στο όνομα του κινητού, το Rating του, την Sim που δέχεται (δηλαδή τον αριθμό των καρτών), το μέγεθος της μπαταρίας του σε mAh, την ανάλυση της οθόνης, την κάμερα που έχει, αν δέχεται έξτρα κάρτα αποθηκευτικού χώρου και το λειτουργικό σύστημα του κάθε κινητού τηλεφώνου.

# 4. Περιγραφή των μεθόδων ανάλυσης δεδομένων

*Η υποενότητα της ανάλυσης των δεδομένων θα πρέπει να περιγράφει τις τεχνικές που χρησιμοποιήσατε και μια εξήγηση γιατί! Πολύ σημαντικό είναι να προσπαθήσετε να πείσετε τον αναγνώστη ότι μια συγκεκριμένη τεχνική που χρησιμοποιείται είναι αυτή που ταιριάζει στο πρόβλημα. Να είστε σαφείς και περιεκτικοί*

Για την υλοποίηση της ανάλυσης χρησιμοποιήσαμε τρείς τεχνικές. Αρχικά, κάναμε την τεχνική One Hot Encoding στις κατηγορικές τιμές του Processor (και αντίστοιχα της RΑΜ) στην συνέχεια κάναμε Clustering (K-Means) για το grouping των κινητών και τέλος κάναμε την τεχνική του Association Rule Mining (Apriori function) για την εξαγωγή των συσχετίσεων. Ο λόγος που χρησιμοποιήσαμε αυτές τις τεχνικές θα τις παρουσιάσουμε παρακάτω εξηγώντας εν συντομία τι κάνει γενικά η κάθε τεχνική και γιατί την χρησιμοποιούμε στην ανάλυση μας.

Η τεχνική One Hot Encoding χρησιμοποιείται γενικά στη μηχανική μάθηση για τη μετατροπή κατηγορικών δεδομένων σε αριθμητικά δεδομένα. Τα κατηγορικά δεδομένα δηλαδή μετατρέπονται σε στήλες (ανάλογα πόσα είναι) και για κάθε στοιχείο η τιμή της στήλης γίνεται “Hot” (1) ή “Cold” (0) αν η συγκεκριμένη γραμμή έχει ή όχι το στοιχείο αυτό. Αυτή είναι η μορφή που θέλουμε n-dimensions με κάθε κατηγορία να χωρίζεται σε στήλες. Υπάρχει και άλλη μορφή που γίνεται η κωδικοποίηση σε μια στήλη (1-dimension) αναθέτοντας ένα unique value (int) ανάλογα με όλες τις κατηγορίες που έχει μέσα και αν το συναντήσει ξανά θέτει τον ίδιο αριθμό. Αυτήν την μορφή χρησιμοποιήσαμε για να κάνουμε το scatter plot μεταξύ του column Price και του column Processor/RAM οι οποίες στήλες έχουν κατηγορικά δεδομένα. Με αυτόν τον τρόπο, στο σχήμα για τον ίδιο τύπο Processor/RAM (με συγκεκριμένη τιμή int) αντιστοιχίζονται διαφορετικά Prices και βγαίνει σωστό το γράφημα μεταξύ τιμής και Processor/RAM.

Η τεχνική Clustering ….

Η τεχνική Association Rule Mining ….

# 5. Πειραματικά Αποτελέσματα

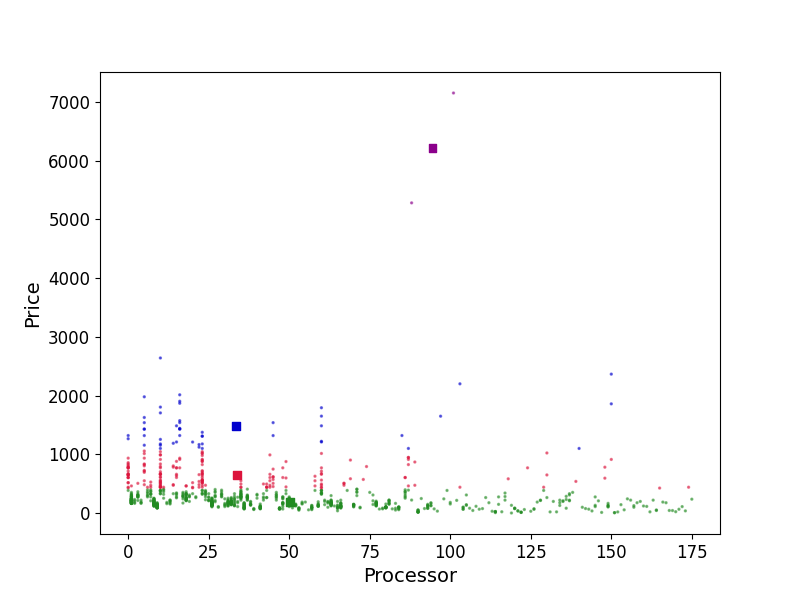
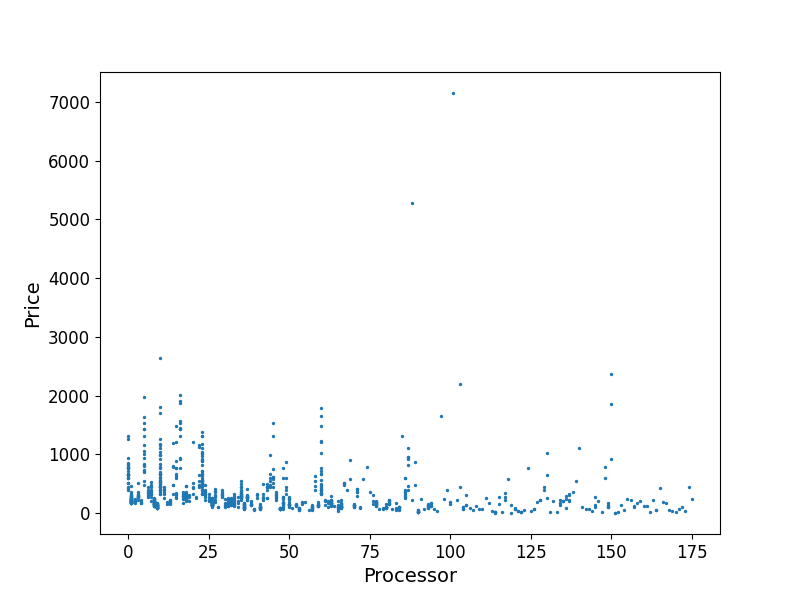
*Η ενότητα των πειραματικών αποτελεσμάτων θα πρέπει να περιλαμβάνει όλα τα πειράματα που χρησιμοποιήσατε. Συζητήστε τις παραμέτρους και ιδιαίτερα τους χρόνους εκτέλεσης, καθώς και τυχόν μέτρα αξιολόγησης που χρησιμοποιήθηκαν. Συμπεριλάβετε πίνακες/σχήματα όπως κρίνετε απαραίτητο (τα περισσότερα έγγραφα ανάλυσης δεδομένων τα διαθέτουν).*

Οι παράμετροι που πειραματιστήκαμε ήταν για τον αλγόριθμο K-means «ποιος είναι ο ιδανικός αριθμός clusters που ταιριάζουν στο πρόβλημά μας» και για το Association Rule Mining και στην τιμή του minimum support και confidence για

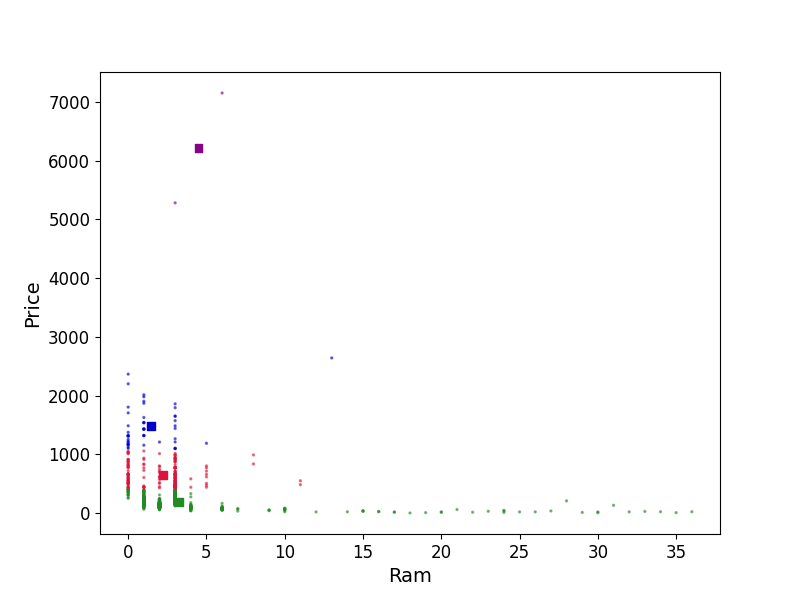
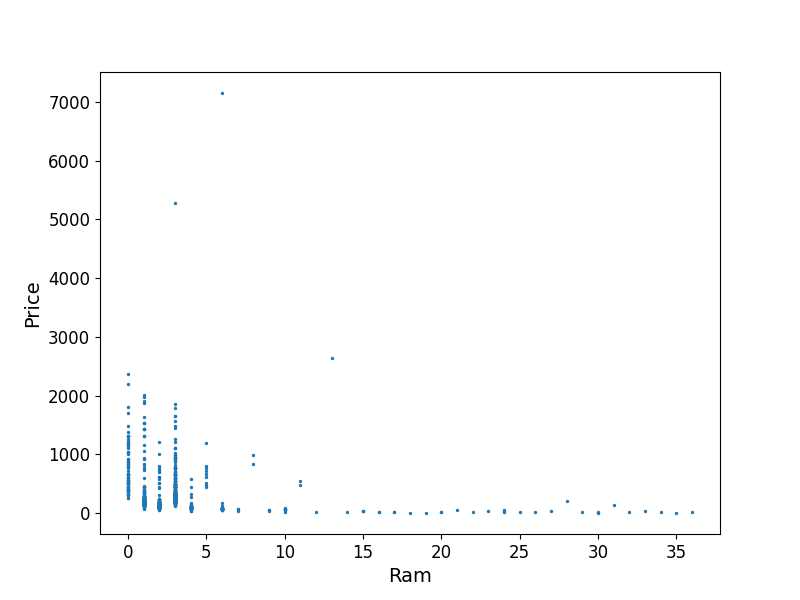
*MIN SUPPORT, MIN CONFIDENCE -> Discord τι σημαίνει έχουμε μεγαλύτερο ή μικρότερο min\_sup κτλ.*

*ΧΡΟΝΟΙ ΕΚΤΕΛΕΣΕΙΣ ΕΙΔΙΚΑ ΓΙΑ CLUSTERING (το γράφει το .ipynb, 33.4 sec και για τα δύο έβγαλα) -> πολυπλοκότητα clustering?*

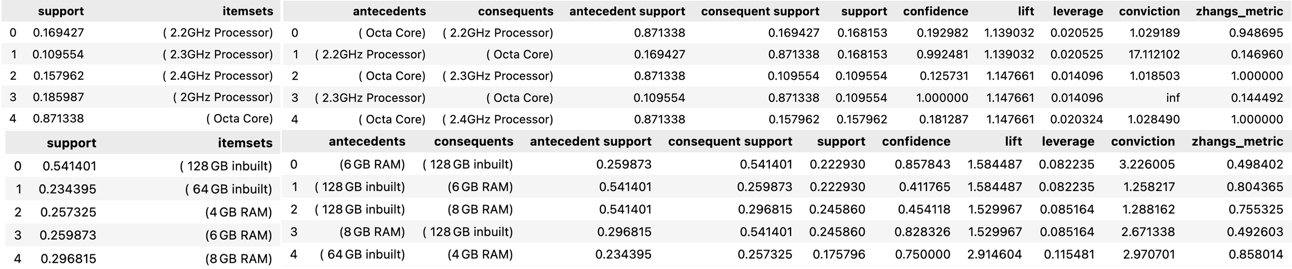
*ΣΧΗΜΑΤΑ CLUSTERING -> 4 clusters (δοκιμή αρχικά με 3, αλλά υπήρχαν κάποια πολύ έξω οπότε 4 για καλύτερο grouping)*

**

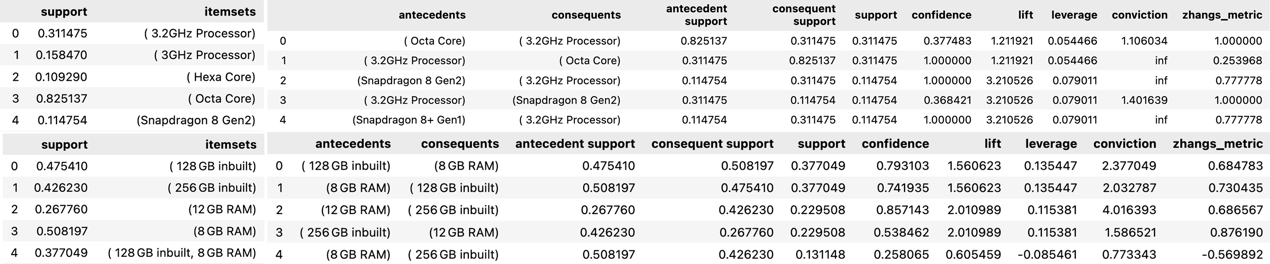
*Εικόνα 1: Απλό scatter γράφημα (Price – Processor) Εικόνα 2: Γράφημα Cluster (Price – Processor)*

**

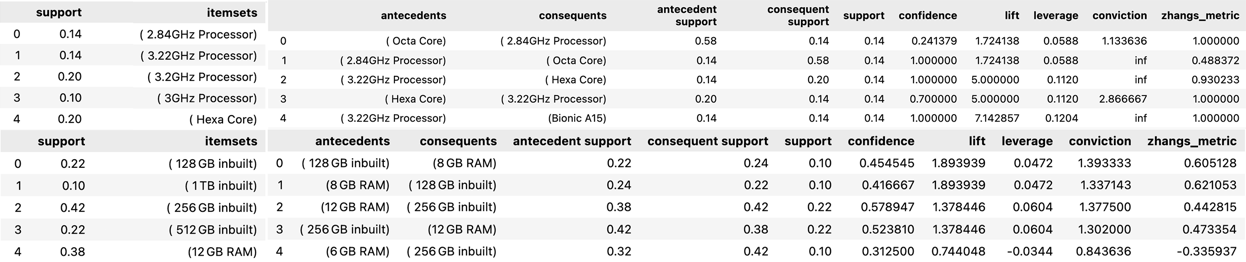
*Εικόνα 3: Απλό scatter γράφημα (Price – RAM) Εικόνα 4: Γράφημα Cluster (Price – RAM)*

**

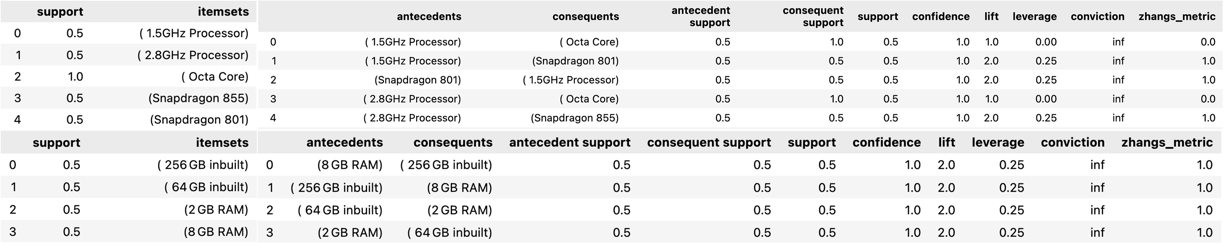
*Εικόνα 5: Association Rule Mining (Processor, RAM) για Low-End smartphones*

**

*Εικόνα 6: Association Rule Mining (Processor, RAM) για Middle Range smartphones*

**

*Εικόνα 7: Association Rule Mining (Processor, RAM) για High-End smartphones*

**

*Εικόνα 8: Association Rule Mining (Processor, RAM) για Luxury smartphones*

# 6. Κριτική αποτίμηση αποτελεσμάτων

ΦΡΟΣΩ για price processor

ΑΓΓΕΛΟΣ: για price ram (συνοπτικά για κάθε cluster, ποια είναι τα frequent items και οι συσχετίσεις του (συνοπτικές αυτές) και αν βγάζει νόημα να έχουν αυτά από προσωπική άποψη)

# 7. Συμπεράσματα

Τι συμπεράσματα βγάζουμε για τα διαφορετικά είδους κινητά (low-end, middle-end, higher-end, luxury). Απο τα αποτελέσματα ποιο είναι το καλύτερο? Σε σύγκριση με την τιμή δηλαδή πάντα. Γενικά αν εμπίπτουν τα αποτελέσματα με την πραγματικότητα… Αρκετά συνοπτικά!