ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

ΓΡΑΠΤΗ ΑΝΑΦΟΡΑ

Εισαγωγή

Στα πλαίσια του μαθήματος "Διαχείριση δεδομένων μεγάλης κλίμακας" εκπονήσαμε την συγκεκριμένη εργασία που έχει ως θέμα την ανάλυση των big data. Ως big data , αναφερόμαστε σε δεδομένα μεγάλης κλίμακας και ειδικότερα στην συλλογή, την προεπεξεργασία και την ανάλυση τους. Ως προς τον όρο big data, δεν υπάρχει σαφής ορισμός, καθώς εξαρτάται από τον τρόπο χρήσης και το περιβάλλον που ανήκουν τα δεδομένα. Παρ' όλα αυτά, η πιο ασφαλής προσέγγιση είναι τα 3 Vs(volume, velocity, variety). Τα τρία Vs αποτελούν τα βασικά χαρακτηριστικά των big data και μας βοηθούν να αντιληφθούμε τη διαφορά τους από τα «κανονικά» δεδομένα. Ο τεράστιος όγκος τους απαιτεί μια τεχνική παράλληλης επεξεργασίας δεδομένων όπως το MapReduce. Αποτελούνται από μεγάλη ποικιλία τύπων δεδομένων και έχει δημιουργηθεί η ανάγκη για τη γρήγορη επεξεργασία τους σε πραγματικό χρόνο.

Έπειτα από τη συλλογή, την αποθήκευση και την προεπεξεργασία των big data ακολουθεί η διαδικασία ανάλυσης τους. Σε αυτό το σημείο επιλέγεται μία τεχνική που ταιριάζει με το εκάστοτε πρόβλημα, για παράδειγμα κάποια μέθοδος Εξόρυξης Δεδομένων όπως η Ομαδοποίηση, και πιο συγκεκριμένα επιλέγονται συγκεκριμένοι αλγόριθμοι Ομαδοποίησης ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων και το σκοπό της ανάλυσης. Παράλληλα, επιλέγεται η αρχιτεκτονική αλλά και το εργαλείο στο οποίο θα πραγματοποιηθεί η συγκεκριμένη διαδικασία. Τα περισσότερα εργαλεία παρέχουν τη δυνατότητα οπτικοποίησης των αποτελεσμάτων της ανάλυσης των δεδομένων με διάφορους τρόπους. Συμπερασματικά, μέσω των παραπάνω βημάτων θα εξαχθούν τα αποτελέσματα της ανάλυσης, τα οποία και θα ερμηνεύσουν οι κατάλληλοι άνθρωποι, ώστε ανακαλυφθεί σημαντική γνώση και να βελτιωθεί η διαδικασία λήψης αποφάσεων.

Σύντομη περιγραφή του συνόλου δεδομένων - Ορισμός προβλήματος και κίνητρο

Για την εργασία χρησιμοποιήσαμε το σύνολο δεδομένων Billionaires Statistics Dataset (2023) που κατεβάσαμε μέσω της πλατφόρμας Kaggle(Statistics Billionaires Dataset.csv). Πιο συγκεκριμένα, περιέχει διάφορες πληροφορίες για τους 2640 πιο πλούσιους ανθρώπους στον κόσμο για το 2023. Αξίζει να τονιστεί ότι δεν έγινε κάποιος ιδιαίτερος έλεγχος για να επαληθευτεί ότι τα στοιχεία ανταποκρίνονται στην πραγματικότητα, καθώς αρχικά θα ήταν αδύνατο να έρθουμε σε επαφή με τέτοιου είδους πληροφορίες και δεύτερον δε θα επηρεάσει το σκοπό της συγκεκριμένης εργασίας. Το dataset αποτελείται από 35 χαρακτηριστικά/στήλες. Συνοπτικά, τα πιο σημαντικά χαρακτηριστικά είναι το όνομα του ανθρώπου, η συνολική περιουσία του, ηλικία, η κατηγορία στην οποία ανήκουν οι επιχειρήσεις του, διάφορα προσωπικά στοιχεία(όπως το εάν δημιούργησε ή κληρονόμησε την περιουσία του) και αρκετά οικονομικά στοιχεία της χώρας που κατοικεί ο εκάστοτε άνθρωπος όπως το ΑΕΠ της χώρας και το ποσοστό φορολόγησης. Το συγκεκριμένο σύνολο δεδομένων δηλαδή δεν ασχολείται αποκλειστικά με το μέγεθος της περιουσίας των πλουσιότερων ανθρώπων στον κόσμο, αλλά και με την κατανομή των χρημάτων στις χώρες, την παρατήρηση διάφορων οικονομικών δεικτών σε κάποια συγκεκριμένη χώρα σε σύγκριση με τον αριθμό των κατοίκων της που εμφανίζονται στο συγκεκριμένο dataset και πολλά άλλα.

Αναμφίβολα, πολλοί άνθρωποι θα ήθελαν να γνωρίζουν το μυστικό της επιτυχίας των επιχειρηματιών του dataset, κάτι το οποίο είναι αδύνατον να εξαχθεί από οποιαδήποτε ανάλυση. Ωστόσο, θα προσπαθήσουμε να ανακαλύψουμε κοινά σημεία μεταξύ των πλουσιότερων ανθρώπων στον κόσμο, θα δημιουργήσουμε αρκετά διαγράμματα με στόχο να δούμε αν υπάρχουν σημαντικά σημεία τα οποία είναι άξια σχολιασμού, όπως πόσοι άνθρωποι έχουν πολύ μεγαλύτερη περιουσία σε σχέση με το μέσο όρο του συνόλου δεδομένων ή αν υπάρχουν πολλοί νέοι δισεκατομμυριούχοι. Επίσης, θα προσπαθήσουμε να συλλέξουμε κρυφή γνώση και να την ερμηνεύσουμε.

Ειδικότερα, τα αποτελέσματα της ανάλυσης του συγκεκριμένου συνόλου δεδομένων είναι πιθανόν χρήσιμα για ανακάλυψη οικονομικών απατών και γενικότερα για οικονομικό έλεγχο σε φυσικά πρόσωπα και επιχειρήσεις, που έχει καταγραφεί ότι η συμπεριφορά τους διαφέρει αρκετά από τη γενικότερη κατανομή ή την κατανομή σε συγκεκριμένη χώρα που βρίσκεται σε οικονομική κρίση για παράδειγμα. Δηλαδή, η ενασχόληση με δεδομένα τέτοιου

τύπου αποτελούν αφορμή για κοινωνικό-οικονομική συζήτηση, μέσω της σύγκρισης της περιουσίας κάποιων ανθρώπων και της οικονομικής κατάστασης που βρίσκεται η χώρα τους. Παράλληλα, τα αποτελέσματα τέτοιας μορφής ανάλυσης θα μπορούσαν να δημοσιευθούν σε περιοδικά που ασχολούνται με την οικονομία και της επιχειρήσεις. Επίσης, είναι συχνό γεγονός η δημιουργία ρεπορτάζ στα ΜΜΕ σχετικά με το συγκεκριμένο θέμα. Από την άλλη πλευρά, όσον αφορά της επιχειρήσεις, μέσω της εστίασης σε μια συγκεκριμένη κατηγορία επιχειρήσεων(χαρακτηριστικό category του dataset), μπορεί να πραγματοποιηθεί ανάλυση του ανταγωνισμού. Τέλος, τα αποτελέσματα της ανάλυσης του συγκεκριμένου dataset είναι δυνατόν να αποτελέσουν την είσοδο για κάποιο άλλο σύστημα διεξοδικότερης ανάλυσης. Για παράδειγμα, γίνεται να χρησιμοποιηθούν για έρευνα σχετικά με την πορεία συγκεκριμένων δισεκατομμυριούχων του συνόλου δεδομένων, ώστε να αναλυθεί η στρατηγική τους και να αποτελέσει έμπνευση για νεαρούς επιχειρηματίες.

Περιγραφή της μεθόδου ανάλυσης των δεδομένων

Αρχικά, πρέπει να σημειωθεί ότι ξεκινήσαμε με το στάδιο του καθαρισμού και της προεπεξεργασίας των δεδομένων, πριν χρησιμοποιήσουμε κάποια μέθοδο ανάλυσης των δεδομένων, ώστε τα δεδομένα που περιέχονται στο dataset μας να είναι κατάλληλα για τα επόμενα βήματα. Η προεπεξεργασία των big data είναι μία πολύ σημαντική διαδικασία και αφιερώνεται σε αυτή μεγάλο χρονικό διάστημα. Αναλυτικότερα, εφαρμόσαμε έλεγχο για ελλιπείς τιμές(για παράδειγμα υπήρχε κενό στην ηλικία ενός δισεκατομμυριούχου) και για διπλότυπα. Διπλότυπα δεν υπήρχαν, ωστόσο συγκεκριμένες εγγραφές είχαν ελλιπείς τιμές σε κάποια χαρακτηριστικά του, και αφαιρέθηκαν από το dataset(θα μπορούσαμε για κάθε ελλιπή τιμή να τοποθετήσουμε το μέσο όρο της τιμής του χαρακτηριστικού, αλλά επιλέξαμε απλά να αφαιρέσουμε από το dataset τα συγκεκριμένα πρόσωπα). Στη συνέχεια, ασχοληθήκαμε με τα κατηγορικά δεδομένα, δηλαδή χαρακτηριστικά που περιέχουν κείμενο ως τιμές, όπως η κατηγορία που υπάγεται η εκάστοτε επιχείρηση. Μετατρέψαμε τα κατηγορικά δεδομένα σε αριθμητικά, δίνοντας τους έναν αριθμό για κάθε διαφορετική τιμή τους. Στη συνέχεια, μέσω διαγραμμάτων box plot αποτυπώσαμε τις διάφορες τιμές που λαμβάνουν κάποια χαρακτηριστικά του dataset και παρατηρήσαμε εάν και πόσες ακραίες τιμές περιέχουν. Ωστόσο, οι ακραίες τιμές δεν αφαιρέθηκαν από το dataset, καθώς είναι ιδιαίτερα σημαντικές για τον σκοπό της ανάλυσης μας. Ακολούθησε η ανάλυση συσχέτισης μεταξύ συγκεκριμένων χαρακτηριστικών του dataset, με τη βοήθεια του συντελεστή Pearson, που λαμβάνει τιμές από -1 μέχρι και 1. Χαρακτηριστικά που εμφανίζουν μεγάλη συσχέτιση, δε χρειάζονται και τα 2 στο dataset, παρά μόνο το 1 καθώς δεν κερδίζουμε κάποια πληροφορία με την ύπαρξη τους στο dataset(δηλαδή κοντά στο 1:θετική συσχέτιση, η αύξηση της μιας μεταβλητής συμβαδίζει με την ύπαρξη και της δεύτερης ή στο -1). Η συγκεκριμένη διαδικασία είχε ως αποτέλεσμα τη μείωση των χαρακτηριστικών του. Ο τελικός αριθμός των χαρακτηριστικών/στηλών που χρησιμοποιήθηκαν στα επόμενα βήματα της ανάλυσης είναι 13. Το τελευταίο στάδιο της προεπεξεργασίας των δεδομένων είναι η κανονικοποίηση, δηλαδή ο μετασχηματισμός των δεδομένων σε μια κοινή κλίμακα. Η συγκεκριμένη διαδικασία είναι απαραίτητη για τη συνέχεια της ανάλυσης, καθώς για παράδειγμα είναι αδύνατον να συνυπάρχουν χαρακτηριστικά με εντελώς διαφορετικές κλίμακες στο dataset, όπως η περιουσία, η ηλικία του ανθρώπου, και το ΑΕΠ της χώρας(3 διαφορετικές κλίμακες). Για την κανονικοποίηση των δεδομένων χρησιμοποιήσαμε τη μέθοδο Z-score. Το τελικό dataset που χρησιμοποιήσαμε στην ανάλυση των δεδομένων είναι το εξής και δημιουργήθηκε μέσω της προεπεξεργασίας: Billionaires Statistics Dataset_normalized.csv .

Για την ανάλυση των δεδομένων μας χρησιμοποιήσαμε μία από τις πιο δημοφιλείς μεθόδους Εξόρυξης Δεδομένων, την Ομαδοποίηση/Συσταδοποίηση. Η συγκεκριμένη μέθοδος είναι υπεύθυνη για την τοποθέτηση παρατηρήσεων σε ομάδες, με τέτοιον τρόπο ώστε τα αντικείμενα που βρίσκονται στην ίδια ομάδα να έχουν μεγάλη ομοιότητα. Ενώ, αντικείμενα που ανήκουν σε διαφορετικές ομάδες έχουν μικρή ομοιότητα. Έχουμε ήδη χρησιμοποιήσει τον όρο ομοιότητα, οπότε πρέπει να αναφέρουμε τα μέτρα ομοιότητας. Είναι μέτρα απόστασης που μετρούν πόσο όμοια είναι τα δεδομένα. Οι συστάδες εκπροσωπούνται από το κέντρο, δηλαδή το μέσο όρο όλων των σημείων της συστάδας. Στο πρακτικό μέρος της εργασίας μας χρησιμοποιήσαμε 3 τεχνικές Συσταδοποίησης: Ιεραρχική Συσταδοποίηση,

k-means και DBSCAN. Η κάθε μία τεχνική ανταποκρίνεται καλύτερα σε δεδομένα άλλης μορφής και χρησιμοποιεί διαφορετικό αλγόριθμο ομαδοποίησης. Η Ιεραρχική ανήκει στους παραδοσιακούς αλγορίθμους, όπως και η k-means(Διαμεριστικός αλγόριθμος), ενώ η DBSAN ανήκει στην κατηγορία των Μοντέρνων αλγόριθμων Πυκνότητας. Η επιλογή της τεχνικής Ομαδοποίησης εξαρτάται από τη μορφή των χαρακτηριστικών του dataset που θα χρησιμοποιήσουμε στη διαδικασία, αλλά και το στόχο που έχουμε θέσει. Στο πρακτικό μέρος της εργασίας γράψαμε κώδικα για 2 εκτενή παραδείγματα ομαδοποίησης χρησιμοποιώντας και τους 3 αλγορίθμους στα ίδια χαρακτηριστικά και κάναμε τις απαραίτητες συγκρίσεις ως προς την ποιότητα της συσταδοποίησης που παρέχουν.

Σε αυτό το σημείο θα περιγράψουμε συνοπτικά τις 3 μεθόδους Ομαδοποίησης που αναφέραμε παραπάνω. Όσον αφορά την Ιεραρχική Συσταδοποίηση, χτίζει μια ιεραρχία από συστάδες, που αναπαρίστανται μέσω του δενδρογράμματος, ενός σχήματος που οπτικοποιεί την ιεραρχία των συστάδων και μας δείχνει τη σειρά με την οποία ενώθηκαν οι διάφορες συστάδες. Δηλαδή, δημιουργείται ένα σύνολο από συστάδες, όπου κάθε συστάδα περιέχει μικρότερες συστάδες. Χρησιμοποιείται πίνακας και μέτρο απόστασης, ενώ σε κάθε βήμα συγχωνεύονται οι πιο όμοιες συστάδες. Ως μέτρα απόστασης χρησιμοποιούνται η Ευκλείδεια ή η Manhattan απόσταση. Αξίζει να τονιστεί ότι η συγκεκριμένη μέθοδος δυσκολεύεται σε προβλήματα που υπάρχουν ακραία δεδομένα(υπάρχει τέτοιο παράδειγμα σε Python στο πρακτικό μέρος της εργασίας). Υπάρχουν 3 υποκατηγορίες Ιεραρχικής Συσταδοποίησης: απλού, μέσου και πλήρους δεσμού, και η διαφορά τους έχει να κάνει με τον πίνακα αποστάσεων, ωστόσο δε θα επεκταθούμε παραπάνω. Για τη μέτρηση της ποιότητας της συσταδοποίησης χρησιμοποιείται το μέτρο Silhouette Coefficient. Παρέχει πληροφορίες για την τοποθέτηση ενός δείγματος στο χώρο, σε σχέση με τις συστάδες. Έχει εύρος τιμών από -1 έως και 1, για μεγάλες τιμές του συντελεστή, δηλαδή τιμές κοντά στο 1 καταλαβαίνουμε ότι έχει πραγματοποιηθεί καλή συσταδοποίηση.

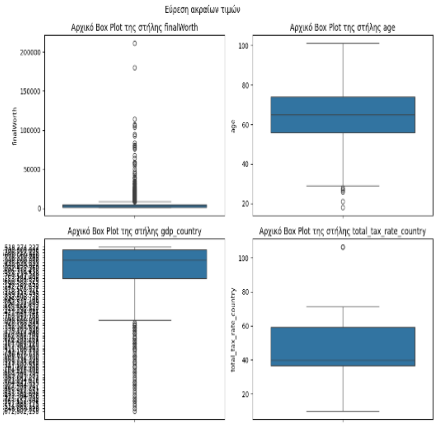
Στη Συσταδοποίηση k-means ο αριθμός των συστάδων k είναι εκ των προτέρων γνωστός. Αρχικά, τα κέντρα των συστάδων επιλέγονται τυχαία. Έπειτα, ξεκινάει μια επαναληπτική διαδικασία, όπου μετά τον υπολογισμό της απόστασης του κάθε σημείου από κάθε κέντρο, δημιουργούνται k συστάδες, και κάθε σημείο ανήκει στην κοντινότερη του συστάδα. Στη συνέχεια υπολογίζουμε το μέσο της συστάδας και μετακινούμε εκεί το κέντρο της. Η επαναληπτική διαδικασία σταματά όταν σταματήσουν να αλλάζουν θέση τα κέντρα. Ως μέτρα απόστασης χρησιμοποιούνται η Ευκλείδεια απόσταση, η Manhattan και η ομοιότητα συνημιτόνου(στην Python επιτρέπεται η χρήση μόνο της Ευκλείδειας). Η μέθοδος k-means είναι αρκετά δημοφιλής και απλή. Και εκείνη έχει ευαισθησία στα ακραία δεδομένα. Για τη μέτρηση της ποιότητας της συσταδοποίησης και την εύρεση του κατάλληλου k για το εκάστοτε πρόβλημα χρησιμοποιείται το SSE, δηλαδή το Συνολικό Τετραγωνικό Σφάλμα. Συνοπτικά, υπολογίζει το συνολικό άθροισμα των τετραγώνων των αποστάσεων μεταξύ κάθε δείγματος και του κέντρου της συστάδας στην οποία ανήκει το δείγμα. Όσο μικρότερο είναι το SSE αντιστοιχεί σε καλύτερη συσταδοποίηση. Επίσης, μπορεί να χρησιμοποιηθεί και το Silhouette Coefficient ή να γίνει και συνδυασμός τους.

Τέλος, ο αλγόριθμος DBSCAN είναι κατάλληλος για ομάδες με μεγάλη πυκνότητα, δηλαδή πολλά σημεία σε μια συγκεκριμένη ακτίνα. Ο αριθμός των σημείων(min_samples) και η ακτίνα(eps) είναι οι 2 σημαντικότερες παράμετροι του αλγορίθμου. Γενικά, τα σημεία χωρίζονται σε 3 κατηγορίες: κεντρικά, οριακά και θορύβου. Τα σημεία θορύβου ανήκουν σε περιοχές με χαμηλή πυκνότητα και εξαλείφονται. Επίσης, τα κεντρικά σημεία έχουν πυκνότητα μεγαλύτερη από min_samples. Η μέθοδος DBSCAN μπορεί να εφαρμόσει Ομαδοποίηση σε συστάδες με διαφορετικά σχήματα ή μεγέθη. Χρησιμοποιηθεί το Silhouette Coefficient ως μέτρο ποιότητας της συσταδοποίησης.

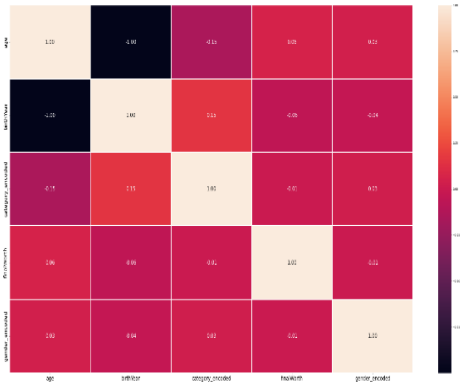
Αξίζει να σημειωθεί ότι σκεφτήκαμε να συγκρίνουμε το πως συμπεριφέρονται κάτω από τις ίδιες συνθήκες οι 3 τεχνικές Ομαδοποίησης που αναλύσαμε παραπάνω. Συνοπτικά, είναι πιθανό να ταιριάζουν στο σύνολο δεδομένων μας και γενικά στο πρόβλημα μας αρκετές μέθοδοι ανάλυσης δεδομένων. Αποφασίσαμε να χρησιμοποιήσουμε την Ομαδοποίηση, καθώς είναι απλή στην υλοποίηση, γρήγορη(όσον αφορά το τρέξιμο της σε Python), δημοφιλής και κατάλληλη για το σκοπό της ανάλυσης μας.

Πειραματικά Αποτελέσματα - Συζήτηση/Κριτική αποτίμηση αποτελεσμάτων - Συμπεράσματα

Αρχικά, αφού φορτώσαμε το dataset μας, τυπώσαμε κάποια στατιστικά και σημαντικά στοιχεία για εκείνο. Στη συνέχεια, ασχοληθήκαμε με τον καθαρισμό και την προεπεξεργασία του dataset, που αναλύθηκαν εκτενώς παραπάνω. Συνοπτικά, διενεργήσαμε έλεγχο για ελλιπείς τιμές, μετατρέψαμε ορισμένα σημαντικά κατηγορικά χαρακτηριστικά σε αριθμητικά, βρήκαμε ακραίες τιμές, κάναμε ανάλυση συσχέτισης και μειώσαμε αρκετά το μέγεθος των δεδομένων.

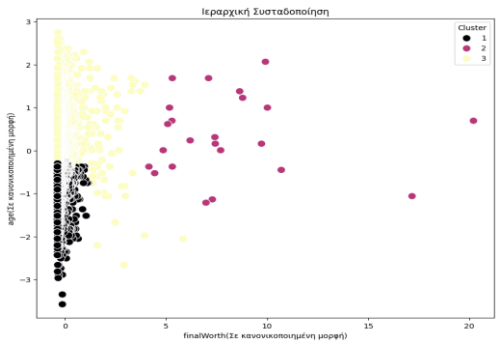


Box Plots(outliers)

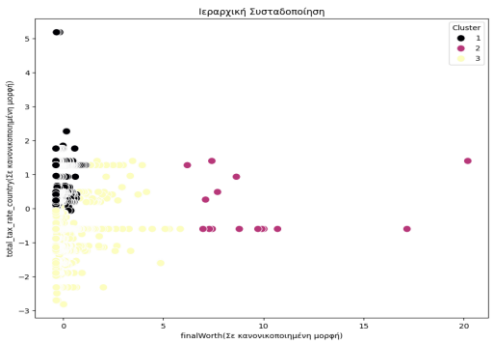


Ανάλυση συσχέτισης (Pearson)

Αξίζει να σημειωθεί ότι στην πορεία των πειραμάτων που κάναμε χρειάστηκε να δημιουργήσουμε από το αρχικό μας .csv αρχείο(dataset) κάποια νέα(είτε για backup, είτε για άλλους λόγους όπως η κανονικοποίηση κάποιων χαρακτηριστικών).
Μετά την προεπεξεργασία των δεδομένων μας, δημιουργήσαμε κώδικα που εφαρμόζει 3 τεχνικές Ομαδοποίησης: Ιεραρχική, k-means και DBSCAN. Για κάθε μία από τις 3 τεχνικές δημιουργήσαμε μία συνάρτηση, για λόγους οπτικούς και πρακτικούς. Ξεκινώντας με την **Ιεραρχική Συσταδοποίηση**(συνάρτηση hierarchical_clustering), χρησιμοποιήσαμε ως παραμέτρους τα χαρακτηριστικά που περιέχονται στην Ομαδοποίηση, τον αριθμό των συστάδων που θέλουμε να δημιουργηθούν(παίρνει την default τιμή 3) και τη μετρική απόστασης cityblock. Για τη δημιουργία του δενδρογράμματος είχαμε τη δυνατότητα να χρησιμοποιήσουμε τις μετρικές Ευκλείδεια ή Manhattan απόσταση, σε συνδυασμό με τη μέθοδο ward. Με τη βοήθεια της συνάρτησης fcluster εφαρμόσαμε Ιεραρχική Συσταδοποίηση(ως παράμετρο παίρνει το criterion=’maxclust’, αφού θέλουμε όταν δημιουργηθούν 3 συστάδες να σταματήσει η συσταδοποίηση). Έπειτα, με τη βοήθεια της εντολής metrics.silhouette_score μετρήσαμε το Silhouette Coefficient, που αποτελεί μέτρο αξιολόγησης της συσταδοποίησης, όπως αναφέραμε και παραπάνω. Τέλος, εμφανίσαμε το διάγραμμα της συσταδοποίησης, με κάθε συστάδα να είναι σχεδιασμένη με διαφορετικό χρώμα. Η συγκεκριμένη συνάρτηση(στην ουσία και οι 3 συναρτήσεις) επιστρέφει το χρόνο εκτέλεσης του αλγορίθμου συσταδοποίησης και την τιμή του Silhouette Coefficient.

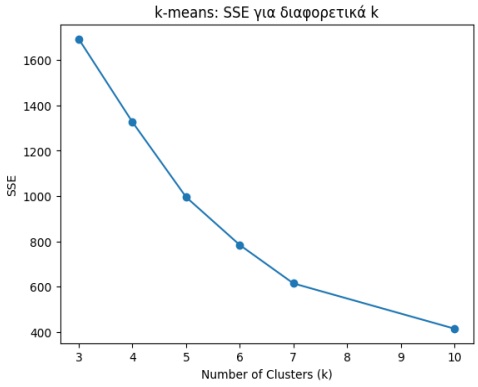


Ιεραρχική Συσταδοποίηση - Παράδειγμα 1

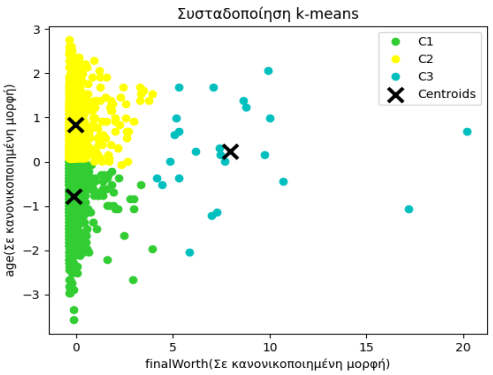


Ιεραρχική Συσταδοποίηση - Παράδειγμα 2

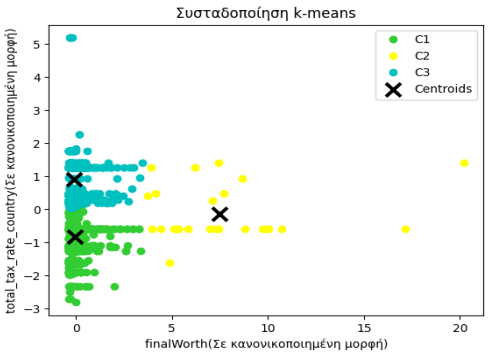
Στη συνέχεια, υλοποιήσαμε τη συνάρτηση **kmeans** που εφαρμόζει k-means συσταδοποίηση στα χαρακτηριστικά του dataset μας που δέχεται ως παράμετρο. Επίσης, ως παράμετρο περνάμε στη συνάρτηση και το k, δηλαδή τον αριθμό των συστάδων, αλλά και μία λίστα που περιέχει τιμές του k και θα εξηγηθεί παρακάτω. Δημιουργήσαμε τις συστάδες με τη βοήθεια της συνάρτησης fit, και χρησιμοποιήσαμε τις εντολές kmeans.labels_ και kmeans.cluster_centers_, ώστε να εμφανίσουμε στο διάγραμμα τις συστάδες με διαφορετικά χρώματα, αλλά με τέτοιο τρόπο ώστε να αντιλαμβανόμαστε και τη θέση των κέντρων τους. Μετά από το διάγραμμα της συσταδοποίησης, ασχοληθήκαμε με το μέτρο αξιολόγησης της συσταδοποίησης SSE(αναλύθηκε παραπάνω). Υπολογίσαμε την τιμή του για k=3, και στη συνέχεια δημιουργήσαμε το διάγραμμα SSE-αριθμός συστάδων. Το ίδιο κάναμε και για το Silhouette Coefficient(για να συγκρίνουμε τους 3 αλγορίθμους με το ίδιο μέτρο αξιολόγησης). Στις περισσότερες πραγματικές εφαρμογές, αν επιλεχθεί ο k-means ως αλγόριθμος ως αλγόριθμος συσταδοποίησης, τότε επιλέγεται ως k εκείνο που αποτελεί την τιμή «γόνατο». Και στα 2 παραδείγματα που υλοποιήσαμε η συγκεκριμένη τιμή είναι η k=7, οπότε αν θέλαμε να πραγματοποιήσουμε την ιδανική k-means συσταδοποίηση θα έπρεπε να ορίσουμε ως 7 τον αριθμό των συστάδων.



Διάγραμμα SSE-k(Παράδειγμα 1)

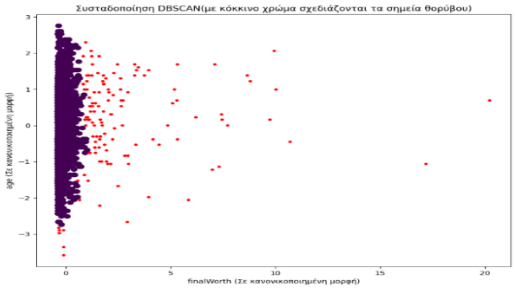


Συσταδοποίηση k-means - Παράδειγμα 1

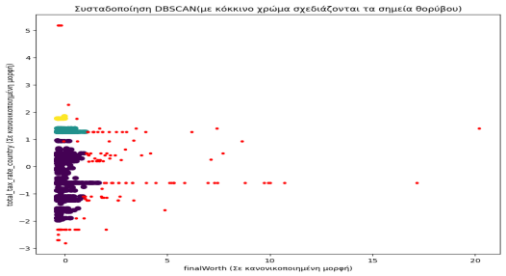


Συσταδοποίηση k-means - Παράδειγμα

Τέλος, δημιουργήσαμε τη συνάρτηση dbscan_clustering που δημιουργεί συστάδες χρησιμοποιώντας τον αλγόριθμο **DBSCAN**. Ως παραμέτρους δέχεται τα χαρακτηριστικά του dataset που συμμετέχουν στην ομαδοποίηση, και τα eps, min_samples που έχουν αναλυθεί παραπάνω, και έχουν ιδιαίτερη σημασία για τη συγκεκριμένη μέθοδο(ο αριθμός των συστάδων δε δίνεται ως όρισμα εκ των προτέρων, αλλά υπολογίζεται μέσω συνδυασμού των 2 παραμέτρων).
Η συγκεκριμένη εντολή υλοποιεί τον αλγόριθμο DBSCAN:
DBSCAN(eps=eps,min_samples=min_samples).fit(chosen_data). Στη συνέχεια, δημιουργούμε το διάγραμμα της ομαδοποίησης και υπολογίζουμε την τιμή του Silhouette Coefficient.



Συσταδοποίηση DBSCAN - Παράδειγμα 1



Συσταδοποίηση DBSCAN - Παράδειγμα 2

Έπειτα, ακολουθεί η συνάρτηση **main()**. Είναι αρκετά σύντομη, περιλαμβάνει 2 παραδείγματα. Σε κάθε παράδειγμα, επιλέγουμε τα χαρακτηριστικά του dataset που θα χρησιμοποιήσουμε, καλούμε τις 3 μεθόδους συσταδοποίησης και συγκρίνουμε τα αποτελέσματα τους διαγραμματικά, ως προς το χρόνο εκτέλεσης και ως προς την ποιότητα της συσταδοποίησης και τυπώνουμε τα αποτελέσματα.
Συνοπτικά, τα χαρακτηριστικά του dataset που χρησιμοποιήσαμε στο πρακτικό μέρος της εργασίας είναι η συνολική του περιουσία(χρησιμοποιείται και στα 2 παραδείγματα), η ηλικία του δισεκατομμυριούχου, και ο συνολικός φορολογικός συντελεστής της χώρας του. Σκοπός της ανάλυσης μας είναι η ομαδοποίηση των δισεκατομμυριούχων σε 3 ομάδες, η ανακάλυψη κρυφής-ενδιαφέρουσας πληροφορίας και η εύρεση ακραίων σημείων.
Ειδικότερα, όσον αφορά το ***1ο παράδειγμα***, χρησιμοποιούμε την περιουσία(finalWorth) και την ηλικία(age), θέλαμε να κατανοήσουμε και να οπτικοποιήσουμε πως κατανέμεται ο πλούτος σε διαφορετικές ηλικιακές ομάδες. Από την ανάλυση που εφαρμόσαμε στα δεδομένα συμπεράναμε πως ο περισσότερος πλούτος που υπάρχει στο dataset βρίσκεται γύρω από το μέσο όρο ηλικιών, αλλά και προς στις μεγαλύτερες ηλικίες.
Από την άλλη πλευρά, στο ***2ο παράδειγμα*** χρησιμοποιήσαμε την περιουσία(finalWorth) και το συνολικό φορολογικό συντελεστή της εκάστοτε χώρας(total_tax_rate_country). Στόχος μας είναι να ανακαλύψουμε εάν ο πλούτος συγκεντρώνεται σε χώρες με μικρότερη φορολογία. Επίσης, το αποτέλεσμα της Ομαδοποίησης είναι πιθανό να φανερώσει πως επηρεάζει η φορολογική πολιτική που εφαρμόζεται τους πλουσιότερους ανθρώπους στον κόσμο. Από τα διαγράμματα συσταδοποίησης παρατηρούμε ότι οι πλουσιότεροι άνθρωποι του dataset(και του κόσμου) κατοικούν σε χώρες που έχουν μικρομεσαίο συντελεστή φορολογίας

- ***Αποτελέσματα Παραδείγματος 1***:
  -Silhouette Coefficient-
  Ιεραρχική Συσταδοποίηση:Silhouette Coefficient για k=3: 0.468
  k-means:Silhouette Coefficient για k=3: 0.496
  DBSCAN:Silhouette Coefficient: **0.629**

  -Χρόνος εκτέλεσης αλγορίθμου-
  Χρόνος εκτέλεσης αλγορίθμου ιεραρχικής συσταδοποίησης για αριθμό συστάδων=3: **0.0062** seconds
  Χρόνος εκτέλεσης αλγορίθμου k-means για k=3: 0.6108 seconds
  Χρόνος εκτέλεσης αλγορίθμου DBSCAN: 0.0675 seconds

- ***Αποτελέσματα Παραδείγματος 2***:
  -Silhouette Coefficient-
  Ιεραρχική Συσταδοποίηση:Silhouette Coefficient για k=3: 0.580
  k-means:Silhouette Coefficient για k=3: **0.599**
  DBSCAN:Silhouette Coefficient: 0.464

  -Χρόνος εκτέλεσης αλγορίθμου-
  Χρόνος εκτέλεσης αλγορίθμου ιεραρχικής συσταδοποίησης για αριθμό συστάδων=3: **0.0098** seconds
  Χρόνος εκτέλεσης αλγορίθμου k-means για k=3: 0.0357 seconds
  Χρόνος εκτέλεσης αλγορίθμου DBSCAN: 0.0708 seconds

Από τα αποτελέσματα της συσταδοποίησης παρατηρούμε ότι και στα 2 παραδείγματα η Ιεραρχική Συσταδοποίηση σημείωσε το μικρότερο(με διαφορά) χρόνο εκτέλεσης. Όσον αφορά την ποιότητα της συσταδοποίησης, με τη βοήθεια του Silhouette Coefficient(σε πραγματικές εφαρμογές πραγματοποιείται συνδυασμός μέτρων αξιολόγησης και όχι μόνο ένα), στο πρώτο παράδειγμα πιο κοντά στο 1 βρίσκεται η μέθοδος DBSCAN, ενώ στο δεύτερο παράδειγμα η μέθοδος k-means.

Επίσης, παρατηρούμε ότι στο 2ο παράδειγμα και οι 3 μέθοδοι δημιουργούν ομαδοποίηση με 3 συστάδες(η DBSCAN 3 + θόρυβο). Από την άλλη πλευρά, στο 1ο παράδειγμα η DBSCAN δημιουργεί μόλις μία συστάδα(οι υπόλοιπες από 3). Ίσως αυτός είναι και ο λόγος που έχει το κοντινότερο στο 1 Silhouette Coefficient. Ο δεύτερος λόγος πιστεύουμε ότι είναι λόγω των δεδομένων θορύβου που δεν περιέχονται στη συσταδοποίηση, σε αντίθεση με τις άλλες μεθόδους.. Και στα 2 παραδείγματα παρατηρούμε ότι οι μέθοδοι k-means και ιεραρχική συσταδοποίηση περιέχουν από μία αρκετά αραιή συστάδα.