1. Ανοίγουμε το αρχείο του glass dataset μας και ορίζουμε τις στείλες βάση των στοιχείων που θέλουμε
2. Κάνουμε κανονικόποιηση των τιμών μάς και βάζουμε τις στείλες των δεδομένων κάνω και αφαίρεση των αουτλιερς καθώς και αφαίρω τον fe που έχει αδείες γραμμές
3. Κάνω την ελμποου ανάλυση που μας ζητάει για να βρώ τα απαραίτητα κλάστερς τα οποια είναι 4 για τα δεδομένα που έχουν υποστεί τον καθαρισμό
4. Το επιβεβαιώνω και απο το σιολουέτ ανάλυσις οτι είναι το K=4 άρα ο ιδανικός αριθμος κλάστερς μάς είναι 4 για το σετ δεδομένων όπως έχω επεξεργαστεί
5. Εφαρμόζω το Kmeans analysis model για 4 κλάστερ όπως έχουμε βρεί και θέτοντας την παράμετρο random_state σε 0 για να διασφαλίστεί ότι τα αποτελέσματα της ομαδοποίησης είναι αναπαραγώγιμα
6. Κάνω σκάττερ πλότ απο το dataframe μας δύο στέιλες και απο κάτω όλες τις στείλές με το ρεφλέλτιβ ιντεξ
7. Χρησιμοποείω το inconcistency index που είναι ένα μέτρο της αστάθειας της ομαδοποίησης κατά τη χρήση ιεραρχικής ομαδοποίησης με τη μέθοδο μονής σύνδεσης. Ορίζεται ως η μέγιστη διαφορά στις αποστάσεις μεταξύ ζευγών παρατηρήσεων στην ίδια συστάδα σε διαφορετικά επίπεδα της ιεραρχίας. Ένας υψηλότερος συντελεστής inconsistency υποδηλώνει ότι η ομαδοποίηση είναι πιο ασταθής και μπορεί να είναι λιγότερο αξιόπιστη.
8. Βρίσκω και για το σετ δεδομένων μου και για τα κέντρα των κλάστερ οτι η απλή σύνδεση είναι καλύτερα αποτελέσματα ώς μέση τιμή του συντελεστή την χαμηλότερη απο όλα
9. Ολές οι μεθόδοι έχουν πολύ μικρό αριθμο inconcistency που συνεπώς σημαινεί οτι όλες παράγουν σωστά κάπως αποτελέσματα με αρκετή ακρίβεια
10. Για να προσδιοριστεί ποιο μοντέλο ομαδοποίησης θα παράγει ένα καλύτερο δενδρόγραμμα, μπορούμε να συγκρίνουμε τους μέσους συντελεστές inconsistency για κάθε μοντέλο και να επιλέξουμε το μοντέλο με τον χαμηλότερο μέσο συντελεστή inconsistency Επομένως, ένα μοντέλο με χαμηλότερο μέσο συντελεστή inconsistency είναι πιθανό να παράγει ένα δενδρόγραμμα
11. Κάνοντας την ανάλυση τον δενδρογραμμάτων βλέπουμε οτι όλα τα δενδρογράμματα παράγουν κάπως καλά δεδομένα όμως παρατηρούμε οτι το δενδρογραμμα απλής single σύνδεσης έχει μικρότερη αποστασή στο κοψιμό y στις τελευταίες κορυφές μας έχουν διακριτό σχήμα και είναι σχετικά συμπαγείς, παρόμοια αποτελέσματα βλέπουμε και στο average που έχει το αμέσως πιο μικρό inconcistency index μετά το single
12. Συμπαιραίνουμε οτι οταν τα δεδομένα καθαριστούν παράγουν καλύτερα σκάττερ πλοτ και έχουμε καλύτερα αποτελέσματα σε όλες τις μετρήσεις και οτι τα κλάστερς μάς είναι 4 και οτι έχουμε καλύτερα αποτελέσματα στα single και average δενδρογράμματα συνεπώς είναι καλύτερες μεθόδοι για τα σέτ δεδομένων μας . αυτο επίσης μπορεί να υποδικνύει ότι τα δεδομένα έχουν ιεραρχική δομή, όπου οι κλάστερς σχηματίζονται από τη συγχώνευση μεμονωμένων δειγμάτων ή μικρών ομάδων δειγμάτων. Αυτό οφείλεται στο γεγονός ότι η μονής σύνδεσης είναι γνωστό ότι σχηματίζει συστάδες που είναι μεγάλες και λεπτές, με τα δείγματα να συγχωνεύονται με διαδοχικό τρόπο.