2016 US elections classification and cluster analysis

Sofia Nikolakakou

March 2021

Περιεχόμενα

Εισαγωγή

Η συγκεκριμένη εργασία αφορά τις εκλογές των Η.Π.Α. το έτος 2016, όπου ένας από τους υποψήφιους ήταν ο Donald Trump. Έχω δύο σετ δεδομένων τα οποία περιέχουν πληροφορίες για τους υποψήφιους των εκλογών και τις ψήφους που πήραν, καθώς και κοινωνικο-οικονομικά χαρακτηριστικά για τις κομητείες της Αμερικής.

Η ανάλυση χωρίζεται σε δύο μέρη. Το πρώτο μέρος έχει να κάνει με τις ψήφους (τα αποτελέσματα των εκλογών) από τις οποίες θέλω να εξάγω πληροφορίες για το ποσοστό των ψήφων που πήρε ο Donald Trump έτσι ώστε να μπορέσω να το κατηγοριοποιήσω για να προβλέψω σε ποιες κομητείες πήρε πάνω από το 50%, χρησιμοποιώντας τα χαρακτηριστικά των κομητειών. Για αυτή τη διαδικασία εφάρμοσα δύο διαφορετικές μεθόδους και σύγκρινα τα αποτελέσματά τους.

Στη συνέχεια, αφήνοντας στην άκρη τα αποτελέσματα των εκλογών, θέλω με βάση κάποιους δημογραφικούς παράγοντες που αφορούν την κάθε κομητεία, να τις χωρίσω σε clusters. Χρησιμοποιώντας αυτά τα clusters θα ελέγξω στην πορεία αν όντως χωρίστηκαν σωστά οι κομητείες λαμβάνοντας υπόψιν ορισμένα οικονομικής φύσης χαρακτηριστικά.

Classification

Με ενδιαφέρει να κατηγοριοποιήσω αν ο Donald Trump πήρε πάνω από τον 50% των ψήφων σε κάθε κομητεία. Για να το κάνω αυτό χρειάστηκε πρώτα να υπολογίσω τις συνολικές ψήφους όλων των υποψηφίων σε κάθε κομητεία και στη συνέχεια με βάση τις ψήφους που είχε ο Trump να υπολογίσω το ποσοστό που απέκτησε στις κομητείες. Έτσι έφτιαξα μια δίτιμη μεταβλητή η οποία παίρνει την τιμή 0 στην περίπτωση όπου είχε κάτω του 50% των ψήφων, και την τιμή 1 εκεί όπου τελικά απέκτησε το προβάδισμα.

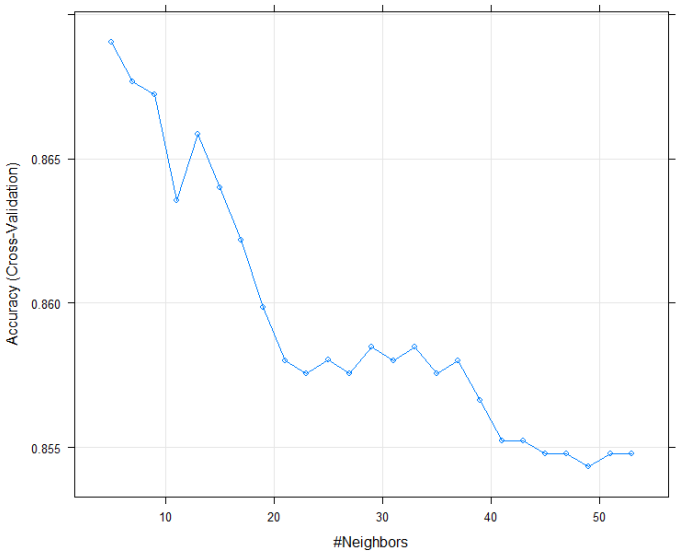
Για το classification χρησιμοποίησα δύο διαφορετικές μεθόδους:

1. K-NN classification
2. Decision Tree

K-NN classification

Πρόκειται για έναν supervised μη-παραμετρικό αλγόριθμο ο οποίος χρησιμοποιεί labeled δεδομένα για να μπορέσει να προβλέψει το αποτέλεσμα των σημείων. Στην ουσία ταξινομεί ένα νέο σημείο στην κλάση ¨στόχο¨ με βάση τα χαρακτηριστικά των γειτονικών σημείων. Εκεί όπου υπάρχει μεγαλύτερη ομοιότητα, εκεί κατατάσσεται το σημείο. Για να γίνει αυτό, χρειάζεται να υπολογιστούν αποστάσεις μεταξύ των σημείων. Επειδή ο αλγόριθμος προβλέπει την κλάση μιας παρατήρησης λαμβάνοντας υπόψιν τις παρατηρήσεις που βρίσκονται πιο κοντά σε αυτή, δηλαδή χρησιμοποιεί αποστάσεις, χρειάστηκε να κεντράρω τα δεδομένα, ώστε να μην υπάρχουν διαφορές μεταξύ της επίδρασης της κάθε μεταβλητής στην απόσταση μεταξύ των παρατηρήσεων και επομένως και στον αλγόριθμο. Επομένως μετασχημάτισα τα δεδομένα, ώστε να έχουν τυπική απόκλιση 1 και μέση τιμή 0, για να βρίσκονται έτσι σε συγκρίσιμη κλίμακα μεταξύ τους.

Για να τρέξω τον αλγόριθμο, χώρισα πρώτα το σετ δεδομένων σε training και test. Αυτό γίνεται γιατί θέλω να φτιάξω το KNN μοντέλο στα training δεδομένα και μετά να αξιολογήσω την απόδοση του στα test δεδομένα. Σημαντικό βήμα στο συγκεκριμένο αλγόριθμο είναι η εύρεση του καλύτερου k, δηλαδή πόσους γείτονες θα χρειαστεί ο αλγόριθμος να χρησιμοποιήσει ώστε να κατατάξει τις παρατηρήσεις. Για να βρω το k έκανα 10-fold cross-validation και κράτησα αυτό που ελαχιστοποιεί το cross-validation σφάλμα. Από αυτή τη διαδικασία βρήκα ότι ο βέλτιστος αριθμός γειτόνων είναι k=5.



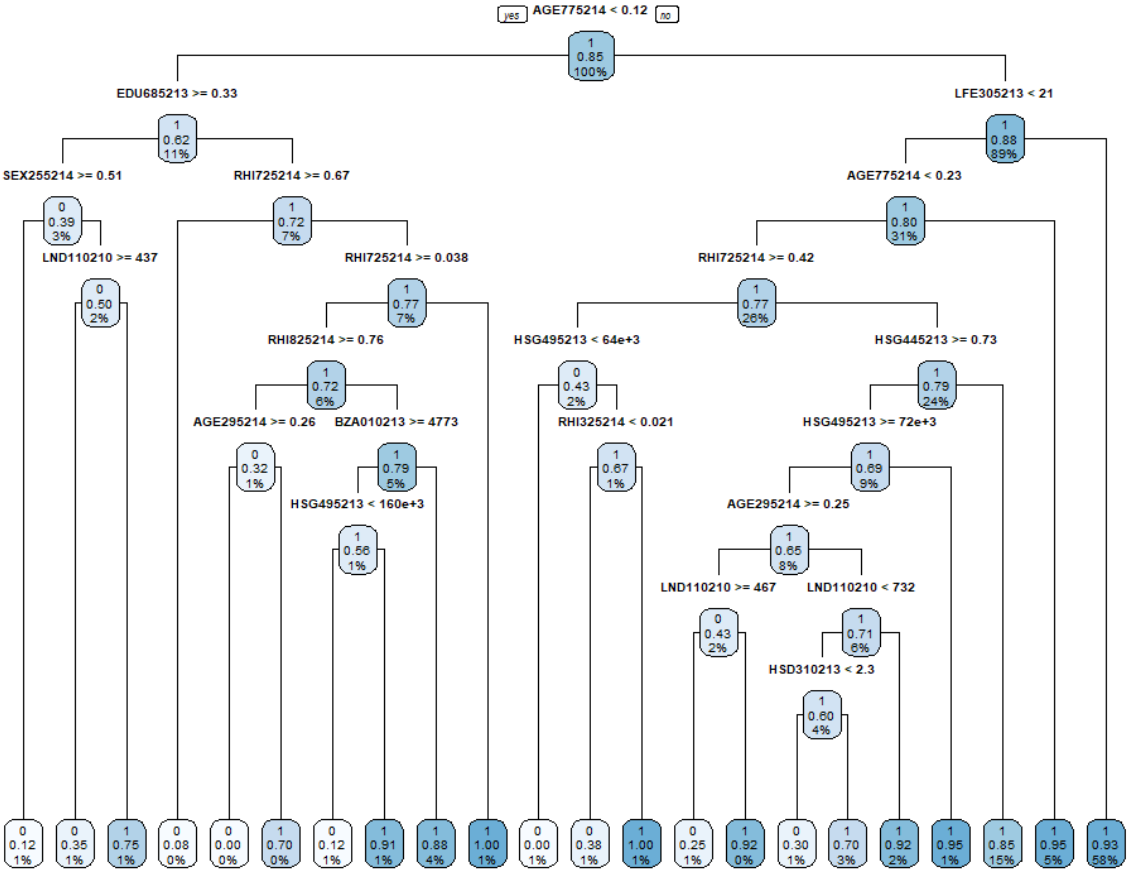
Στην συνέχεια χρησιμοποίησα το μοντέλο για να δω τα αποτελέσματα που θα πάρω από τις προβλέψεις στα test δεδομένα. Ο αλγόριθμος κατάφερε να προβλέψει σωστά τις 442 κομητείες στις οποίες ο Trump πήρε πάνω από το 50% των ψήφων (από τις 541), καθώς και τις 21 στις οποίες είχε ποσοστό κάτω του 50%, ωστόσο η ταξινόμηση δεν έγινε σωστά για τις 78 κομητείες από τις 541. Το accuracy rate, δηλαδή τα true positives και true negatives, ήταν 85.5%, ενώ το misclassification error 1 – 85.5% = 14.5%. Το precision rate ήταν 88.5%, δηλαδή ανάμεσα στις 499 κομητείες, οι 442 ψήφισαν υπέρ του Trump. Σε γενικές γραμμές, η συνολική προβλεπτική ακρίβεια του μοντέλου ήταν αρκετά καλή.

Decision Tree

Πρόκειται για έναν αλγόριθμο που απεικονίζει δυαδικές αποφάσεις, ο οποίες οδηγούν σε κάποια απόφαση σχετικά με την κλάση ενός αντικειμένου. Το δέντρο αποτελείται από την ρίζα η οποία περιέχει πάντα τη μεταβλητή που μου δίνει την περισσότερη πληροφορία σχετικά με την μεταβλητή που με ενδιαφέρει. Στη συνέχεια από τη ρίζα βγαίνουν τα κλαδιά και ο κάθε κόμβος τώρα θα αποτελείται από άλλες μεταβλητές οι οποίες με τη σειρά τους θα οδηγήσουν τελικά στα φύλλα του δέντρου. Το δέντρο σταματάει να μεγαλώνει όταν έχουν εξαντληθεί όλες οι μεταβλητές που είναι διαθέσιμες ή όταν όλες οι παρατηρήσεις έχουν κατηγοριοποιηθεί πλήρως. Χαρακτηριστικό αυτού του αλγορίθμου είναι ότι κάνει από μόνος του επιλογή μεταβλητών, επομένως ακόμα και αν έχω ένα πολυδιάστατο σετ δεδομένων μπορεί να καταλήξω με ένα δέντρο το οποίο αποτελείται από τις μισές ή και λιγότερες μεταβλητές.

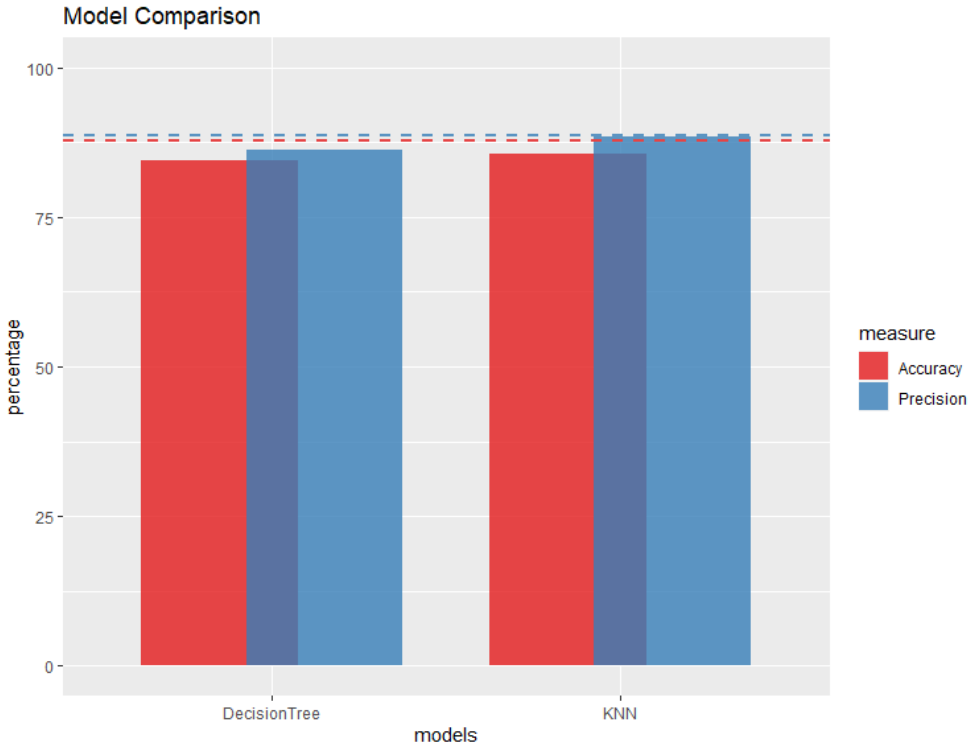
Για να δημιουργηθεί το δέντρο λαμβάνει υπόψιν το Information Gain, το οποίο στην ουσία χρησιμοποιείται για να μετρήσει την ποιότητα με την οποία χωρίζονται κάθε φορά τα κλαδιά του δέντρου. Το Information Gain μετράει τη μείωση της εντροπίας όταν χωρίζουμε ένα σετ δεδομένων σύμφωνα με μια συγκεκριμένη τιμή μια τυχαίας μεταβλητής. Όσο μεγαλύτερο, τόσο μικρότερη η εντροπία. Η εντροπία θα μπορούσε να θεωρηθεί ως το πόσο μη προβλέψιμο είναι ένα σετ δεδομένων., δηλαδή ένα σετ με μια μόνο κλάση θα έχει χαμηλή εντροπία καθώς θα είναι εξαιρετικά προβλέψιμο, ενώ αν αποτελούνταν από μια μίξη κλάσεων τότε θα είχε πολύ μικρότερη προβλεπτικότητα.

Το δέντρο που προέκυψε με βάση τα δεδομένα φαίνεται στο παρακάτω γράφημα.



Στη συνέχεια χρησιμοποίησα τα training test σετ που είχα από πριν για να ελέγξω αν κάνει καλές προβλέψεις ο συγκεκριμένος αλγόριθμος. Κατέληξα να πάρω accuracy rate 84.4%.

Για να δω ποιος από τους δύο αλγορίθμους προβλέπει καλύτερα αν ο Trump πήρε πάνω από το 50% των ψήφων σε κάθε κομητεία, σύγκρινα τα accuracy και precision rates. Από το γράφημα φαίνεται οι δύο προσεγγίσεις να μην έχουν μεγάλη διαφορά, ωστόσο ο KNN αλγόριθμος δίνει λίγο καλύτερα αποτελέσματα.



Clustering

Στο δεύτερο κομμάτι με ενδιαφέρει να κάνω cluster τις κομητείες των Η.Π.Α., δηλαδή να βρω σε ποιες ομάδες ανήκουν με βάση κάποια δημογραφικά χαρακτηριστικά. Στη συνέχεια με βάση κάποιους άλλους οικονομικούς παράγοντες, θα χρησιμοποιήσω να clusters που βρήκα και με βάση αυτά θα ελέγξω αν έγινε σωστά η κατάταξη τους.

Το πρώτο σετ δεδομένων (demographics) με το οποίο έγινε το clustering, περιέχει μεταβλητές που παρέχουν πληροφορίες σχετικά με τον πληθυσμό. Συγκεκριμένα, έχει μεταβλητές σχετικές με το μέγεθος του πληθυσμού στις Η.Π.Α., τις εθνικότητες που απαρτίζουν αυτόν τον πληθυσμό καθώς και το εκπαιδευτικό επίπεδο των ατόμων. Επομένως θέλησα να κάνω cluster τις κομητείες με βάση αυτές τις δημογραφικές ομοιότητες μεταξύ τους.

Ο αλγόριθμος που επέλεξα είναι ο K-Means. Πρόκειται για έναν αλγόριθμο ο οποίος διαχωρίζει να δεδομένα σε K διαφορετικά clusters. Ο τρόπος με τον οποίο χωρίζονται αυτές οι παρατηρήσεις στοχεύει στο να είναι όσο πιο μικρή γίνεται η συνολική διακύμανση μεταξύ των clusters, η οποία ορίζεται ως το άθροισμα των τετραγώνων των Ευκλείδειων διαστάσεων μεταξύ των παρατηρήσεων και των αντίστοιχων centroids. Επομένως αφού χρησιμοποιούνται αποστάσεις χρειάστηκε και πάλι να μετασχηματίσω τα δεδομένα. Για να τρέξει ο αλγόριθμος πρέπει πρώτα να ορίσω τον αριθμό των clusters, ο οποίος βρήκα πως είναι K=9, με βάση το ελάχιστο άθροισμα τετραγώνων μεταξύ των clusters.

Τα clusters που δημιουργήθηκαν και το πως έχουν χωριστεί οι παρατηρήσεις σε αυτά, τα αποθήκευσα σε μία νέα μεταβλητή, η οποία τώρα θα είναι ο ¨στόχος¨ στο δεύτερο σετ δεδομένων με σκοπό να ελέγξω αν έχει γίνει σωστά ο διαχωρισμός. Αυτό το σετ δεδομένων περιέχει μεταβλητές οι οποίες αφορούν πιο οικονομικούς παράγοντες, όπως για παράδειγμα εισόδημα, πόσες επιχειρήσεις κατέχει η κάθε εθνικότητα και λιανικές πωλήσεις καθώς και γεωγραφικούς παράγοντες όπως η έκταση της γης. Επομένως με ενδιαφέρει να δω πως ο διαχωρισμός των clusters, με βάση τα δημογραφικά χαρακτηριστικά, μπορεί να κατηγοριοποιηθεί με βάση κάποιους άλλους παράγοντες.

Για να το κάνω αυτό εφάρμοσα k-Nearest Neighbors ώστε να δω γίνονται σωστά οι προβλέψεις. Και πάλι χώρισα το σετ δεδομένων σε training και test και έκανα 10-fold cross-validation για να βρω τον βέλτιστο αριθμό των k γειτόνων, ο οποίος ήταν 9. Έκανα προβλέψεις χρησιμοποιώντας το test σετ, ο οποίες μου έδωσαν accuracy rate ίσο με 68.7%. Η προβλεπτική ικανότητα επομένως του μοντέλου δεν ήταν και πολύ ικανοποιητική. Συγκεκριμένα, το μοντέλο σε κάποια clusters απέτυχε να κάνει τη σωστή κατηγοριοποίηση, ενώ σε κάποια άλλα κατάφερε να διαχωρίσει σχεδόν τέλεια τις παρατηρήσεις.

Επομένως, ο διαχωρισμός των κομητειών της Αμερικής σε clusters και στη συνέχεια η κατηγοριοποίησή αυτών με βάση οικονομικούς παράγοντες δεν απέφερε τα βέλτιστα αποτελέσματα.