ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ

ΜΕΡΟΣ Β'

ΓΙΑ ΤΟ ΜΑΘΗΜΑ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Υλοποίηση Γενετικού Αλγορίθμου για Συνεργατικό Φιλτράρισμα (Σύσταση Ταινιών)

ΜΑΝΤΑΣ ΕΛΕΥΘΕΡΙΟΣ

Α.Μ. 1047128

ΠΑΤΡΑ 2020

Github link: https://github.com/LefterisMantas/Computational-Intelligence/tree/master/genetic-algorithm

Εισαγωγή

Το project αυτό υλοποιήθηκε σε γλώσσα προγραμματισμού **Python** με τη βοήθεια των παρακάτω βιβλιοθηκών:

- **DEAP:** Υλοποίηση γενετικού αλγορίθμου.
- **Sci-kit Learn:** Χρήστη μετρικών RMSE και MAE για τον υπολογισμό των σφαλμάτων.
- **Pandas και Numpy:** Διαχείριση αρχείων και δεδομένων.
- **Matplotlib:** Γραφικές παραστάσεις απόδοσης και σφαλμάτων.

Ως χρήστης για την εκτέλεση των πειραμάτων των ερωτημάτων B2-B3-B4 χρησιμοποιήθηκε ο χρήστης 269 του δοσμένου ml-100k data set (https://grouplens.org/datasets/movielens/100k/) καθώς είχε αρκετές αξιολογήσεις.

Για τα ερωτήματα B3-B4, η εκτέλεση του αλγορίθμου έγινε 10 φορές και λήφθηκε υπόψη ο μέσος όρος των αποτελεσμάτων. Αυτό συνέβει διότι οι γενετικοί αλγόριθμοι διέπονται από στοχαστική φύση και τα αποτελέσματά τους δεν είναι τετριμμένα.

Β1. Σχεδιασμός ΓΑ

α) Κωδικοποίηση:

Η κωδικοποίηση για τα άτομα του πληθυσμού μπορεί να γίνει με διάφορους τρόπους. Στη βιβλιογραφία υπάρχει η δυαδική κωδικοποίηση (binary), η δεκαδική (integer), η πραγματική (real number) και άλλες.

Οι τιμές των αξιολογήσεων και αντίστοιχα όλων των γονιδίων στα χρωμοσώματα ανήκουν στο διάνυσμα τιμών [1,5] και είναι ακέραιες τιμές. Εάν επιλέγαμε για παράδειγμα την δυαδική κωδικοποίηση τότε θα είχαμε πλεονάζουσες τιμές. Χρησιμοποιώντας δεκαδική κωδικοποίηση, κωδικοποιούμε το πρόβλημα ακριβώς με τις τιμές που μας αφορούν.

Όσον αφορά τις ελλιπής τιμές, και αυτές θα είναι ακέραιοι στο διάστημα [1,5] ώστε να υπάρχει συμφωνία μεταξύ όλων των δεδομένων.

Επομένως, επιλέγουμε την δεκαδική κωδικοποίηση στο διάστημα [1,5] καθώς ταιριάζει ακριβώς με τη φύση του προβλήματός μας.

β) Πλεονάζουσες τιμές:

Για να αποφευχθούν οι πλεονάζουσες τιμές που θα προέκυπταν στη περίπτωση που διαλέγαμε σαν κωδικοποίηση την δυαδική αναθεωρήθηκε η απάντηση του **ερωτήματος (α).** Ως κωδικοποίηση εφαρμόστηκε η δεκαδική των τιμών με την οποία πετύχαμε τη μη ύπαρξη πλεοναζουσών τιμών.

γ) Αρχικός πληθυσμός:

Για την δημιουργία του αρχικού πληθυσμού, από τον χρήστη επιλογής μας, διατηρήσαμε τις ήδη υπάρχουσες τιμές του αναλλοίωτες και στις υπόλοιπες θέσεις τοποθετήσαμε τυχαίους ακέριους αριθμούς από το 1 έως το 5. Έτσι αποφεύγουμε την ομοιομορφία όλων των ατόμων του πληθυσμού.

i. **Οι τιμές των αξιολογήσεων που οι χρήστες δεν έχουν συμπληρώσει**: : Οι τιμές αυτές συμπληρώθηκαν με τυχαίες τιμές. Επιλέξαμε να μην τις συμπληρώσουμε με το μέσο όρο του διανύσματος των χρηστών προκειμένου να μην δημιουργηθούν 20 (και ύστερα αντίστοιχα 200) ολόιδια διανύσματα, ώστε ύστερα στη διαδικασία του γενετικού αλγορίθμου να έχουμε μεγαλύτερο εύρος πιθανών λύσεων και γονιδίων. Επιπλέον, η λειτουργία του γενετικού αλγορίθμου ευνοείται με περισσότερες τυχαίες τιμές καθώς είναι μια στοχαστική διαδικασία και ο μεγαλύτερος χώρος αναζήτησης δίνει μεγαλύτερη ευελιξία στον αλγόριθμο ως προς την εύρεση της βέλτιστης λύσης.

ii. **Οι τιμές των αξιολογήσεων που οι χρήστες έχουν συμπληρώσει**: Οι τιμές αυτές θα παραμείνουν ως έχουν, δηλαδή αναλλοίωτες, όπως και στο δοσμένο data set (https://grouplens.org/datasets/movielens/100k/) στα νέα εμπλουτισμένα χρωμοσώματα. Προφανώς για να γίνει αυτό θα θεωρήσουμε πως μετά από την εφαρμογή του γενετικού αλγορίθμου αυτές οι τιμές πρέπει να είναι ίδιες και να μην αποκλίνουν. Εάν η λύση, σε κάθε βήμα-γενιά του αλγορίθμου, περιέχει τιμές που αποκλίνουν από τις πραγματικές τότε αυτό το αντιμετωπίζουμε με την **Διαδικασία επιδιόρθωσης** όπως εξηγείται και στο ερώτημα **δ)** διορθώνοντας τις τιμές με τις πραγματικές τιμές των δεδομένων.

δ) Διαδικασία επιδιόρθωσης:

Η διαδικασία της επιδιόρθωσης στη συγκεκριμένη περίπτωση έχει νόημα μόνο για τις αξιολογήσεις εκείνες οι οποίες αποκλίνουν από τις αρχικές, πραγματικές τιμές του data set. Εξετάστηκε θεωρητικά και πρακτικά η εφαρμογή τριών μεθόδων επιδιόρθωσης όπως φαίνεται παρακάτω:

i. **Ελιτισμός (Elitism):** Ο ελιτισμός (ή κλωνοποίηση) είναι μία διαδικασία κατά την οποία ένα ποσοστό των πιο κατάλληλων λύσεων-ατόμων αντιγράφονται και συνεχίζουν στην επόμενη γενιά ως γονείς. Ο ελιτισμός ως διαδικασία επιδιόρθωσης θα διόρθωνε τις λύσεις της εκάστοτε γενιάς με τη βέλτιστη λύση. Αυτή η μέθοδος θα μας προσέφερε γρηγορότερες λύσεις καθώς ο χώρος αναζήτησης θα γινόταν ολοένα και μικρότερος και επομένως η σύγκλιση θα ήταν ταχύτερη. Ταυτόχρονα όμως η διαδικασία αυτή περιορίζει τις λύσεις μας, δημιουργώντας το κίνδυνο εγκλωβισμού σε κάποια τοπικά ακρότατα (βέλτιστα). Εφόσον τοποθετήθηκαν τυχαίες τιμές στις κενές θέσεις των χρωμοσωμάτων, μία βέλτιστη λύση, ιδικά στις πρώτες γενιές του αλγορίθμου, θα ήταν αρκετά «τυχαία» και δε μπορεί να θεωρηθεί κατάλληλη για την επιδιόρθωση των υπόλοιπων λύσεων. Προτιμούμε ο αλγόριθμός μας να εκτελεστεί για περισσότερες γενιές και να διορθώνουμε λάθος τιμές με τις ήδη υπάρχουσες προσπαθώντας να αυξήσουμε όσο το δυνατόν περισσότερο την απόδοση - καταλληλότητα των λύσεων.

ii. **Επιδιόρθωση (Repair procedure):** Κατά τη διαδικασία αυτή οι τιμές που είναι λάθος μετά από την εφαρμογή των γενετικών τελεστών (επιλογή, διασταύρωση, μετάλλαξη), διορθώνονται με τις ήδη γνωστές αρχικές τιμές του data set. Έτσι τα νέα χρωμοσώματα δεν αποκλίνουν ποτέ από τα πραγματικά στις ήδη υπάρχουσες τιμές των αξιολογήσεων των ταινιών. «Βοηθάμε», λοιπόν, τον αλγόριθμο μας να μην αναπαράγει λύσεις-παιδιά που έχουν λάθος τιμές, αλλά να προσπαθήσει να εκτιμήσει τις υπόλοιπες, μη υπάρχουσες αξιολογήσεις.

iii. **Εφαρμογή Ποινής (Penalty method):** Αυτή η μέθοδος αντιμετωπίζει τις τιμές που δεν είναι ορθές (ίδιες με τις αρχικές-πραγματικές) εφαρμόζοντας μία ποινή στη συνάρτηση καταλληλότητας. Θεωρούμε πως η απόκλιση από αυτές τις τιμές δε θα βοηθήσει την «εξέλιξη» του γενετικού αλγορίθμου και δεν δεχόμαστε τέτοιες λύσεις, καθώς απέχουν από τη πραγματικότητα. Προτιμούμε όταν βρίσκονται τέτοιες λύσεις να απορρίπτονται κατευθείαν και να διορθώνονται με τις δοσμένες. Επομένως, δεν χρησιμοποιούμε την μέθοδο αυτή.

Επομένως, η διαδικασία της επιδιόρθωσης έγινε με την **Repair Procedure.**

ε) Εύρεση γειτονιάς χρήστη:

Για την εύρεση της γειτονιάς του χρήστη, αρχικά έπρεπε να ακολουθηθεί μία στρατηγική διαχείρισης των μη συμπληρωμένων αξιολογήσεων. Παρακάτω αναλύονται οι πιθανές τεχνικές.

Η συμπλήρωση των ελλιπών τιμών με 0 (ή 1) δε κρίθηκε σωστή επιλογή, αφού οι ταινίες που δε βαθμολόγησε ένας χρήστης δεν μπορεί να θεωρηθούν πως δεν «αρέσουν» καθόλου στο χρήστη και θα τις αξιολογούσε τόσο χαμηλά, αλλά πως απλώς δεν έχει παρακολουθήσει ακόμα τις ταινίες αυτές. Κάλλιστα θα μπορούσε να τις έχει βαθμολογήσει με οποιαδήποτε τιμή στο εύρος [1,5].

Η τυχαίες τιμές για τη συμπλήρωση των κενών επίσης δε θεωρήθηκε σωστή μέθοδος, διότι θεωρήσαμε πως ο κάθε χρήστης «τείνει» να βαθμολογεί με ένα μέσο όρο και δεν είναι τυχαίος ο τρόπος αυτός. Μία τέτοια τεχνική θα επηρέαζε αρκετά την απόσταση μεταξύ των διανυσμάτων με τυχαίο τρόπο, κάτι που είναι πέρα της λογικής της συσχέτισης (ή μη) μεταξύ των χρηστών.

Η αγνόηση των ελλιπών τιμών θα προκαλούσε σφάλμα, εξίσου, διότι η μετρική Pearson που ζητήθηκε να χρησιμοποιηθεί δε λαμβάνει υπόψη της, τις θέσεις ύπαρξης τιμών και έτσι δύο χρήστες που δεν έχουν καμία κοινή αξιολόγηση σε ταινία θα μπορούσαν να θεωρηθούν κοντινοί χρήστες.

Η μέση τιμή σημαίνει πως ο χρήστης αξιολογεί τις περισσότερες ταινίες με περίπου τον ίδιο τρόπο. Αυτό μας βοηθάει να αναγνωρίσουμε ποιοι χρήστες έχουν παρόμοιο τρόπο βαθμολόγησης κατά προσέγγιση. Αυτή η τεχνική θεωρήθηκε η καταλληλότερη μεταξύ των πιθανών. Για τη συμπλήρωση των τιμών αυτών στο data set χρησιμοποιήθηκε το αρχείο **missing_values.py** που δημιούργησε το αρχείο **data_full.csv**

Μετρική Pearson: Ο μαθηματικός τύπος της μεθόδου αυτής για τον υπολογισμό της απόστασης μεταξύ διανυσμάτων X και Y, φαίνεται παρακάτω:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Η μετρική προσπαθεί να βρει τη γραμμική συσχέτιση μεταξύ των δεδομένων και ανάλογα με τη τιμή της συχέτισής τους ποσοτικοποιεί τη διαφορά-συσχέτιση αυτή. Η τιμή του r κινείται μεταξύ [-1,1], αντίθετα π.χ. με την Ευκλείδεια (0,1). Ουσιαστικά υπολογίζει το ποσοστό μεταξύ της συνδιακύμανσης και της τυπικής απόκλισης δύο αντικειμένων.

Η μετρική αυτή εξαλείφει τις μεγάλες διαφορές στις τιμές των μεταβλητών που οφείλονται π.χ. σε διαφορές στη κλίμακα (scale). Έτσι αποφεύγεται η μεγάλη πόλωση στις βαθμολογίες.

Ο υπολογισμός γίνεται στο αρχείο **neighborhood.py** με τη συνάρτηση **pearsonr** της βιβλιοθήκης **scipy.**

στ) Συνάρτηση καταλληλότητας:

Ως αξιολόγηση για τη συνάρτηση καταλληλότητας (fitness function) χρησιμοποιήθηκε η μετρική Pearson. Πιο συγκεκριμένα υπολογίστηκε η απόσταση του κάθε ατόμου του πληθυσμού με καθένα από τα άτομα της γειτονιάς του χρήστη και βρέθηκε ο μέσος όρος αυτών. Επειδή τα αποτελέσματα της μετρικής βρίσκονται στο πεδίο [-1,1] οι τιμές κανονικοποιήθηκαν (κλιμακώθηκαν) στο διάστημα [0,1].

Επικεντρωθήκαμε στην μεγιστοποίηση αυτής της τιμής, καθώς όσο πιο κοντά στο +1 τείνει η συσχέτιση δύο διανυσμάτων-χρωμοσωμάτων, τόσο πιο πολύ σχετίζονται 2 διανύσματα, κατά τη μετρική Pearson και επομένως συμπεραίνουμε πως κάποιο νέο διάνυσμα είναι κοντά στην επιθυμητή λύση.

ζ) Γενετικοί Τελεστές:

i. **Επιλογή:** Αρχικά, η τεχνική της **ρουλέτας βάση του κόστους** ενδιαφέρεται για την «απόσταση» μεταξύ των ατόμων. Αν η διαφορά αυτή είναι πολύ μεγάλη τότε το αντίστοιχο ποσοστό στη ρουλέτα είναι μεγάλο και το αντίθετο. Έτσι στα πιο κατάλληλα άτομα δίνεται περισσότερη πιθανότητα επιλογής στην επόμενη γενιά. Αυτός ο τρόπος κρίθηκε λιγότερο κατάλληλος καθώς περιορίζει τις λύσεις και στο πρακτικό κομμάτι αποδίδει χειρότερα από την μέθοδο επιλογής βάση τουρνουά.

   Η **ρουλέτα βάση της κατάταξης** έχει ως μετρική την κατάταξη με βάση τη καταλληλότητα των λύσεων, επηρεάζει πολύ το βάρος και έτσι ο αλγόριθμος δίνει περισσότερο «χώρο» στη ρουλέτα στο πρώτο άτομο, λιγότερη στο δεύτερο, ακόμα λιγότερη στο τρίτο κ.λ.π. Δεδομένου του γενικότερου σκοπού μας ο αλγόριθμος να διατηρεί ένα tradeoff στις ιδιότητές του δεν επιλέγουμε αυτή τη τεχνική. Επίσης, η απόδοση και αυτή της τεχνικής είναι χαμηλότερη από αυτής του τουρνουά.

   Τέλος, η τεχνική της επιλογής με **τουρνουά** είναι μια τυχαία διαδικασία κατά την οποία τα αποτελέσματα που παίρνουμε εξαρτώνται πολύ από την αρχική «κλήρωση» των ατόμων στο τουρνουά. Με τα από επαναλαμβανόμενες συγκρίσεις των καταλληλοτήτων, επιλέγεται το πιο κατάλληλο άτομο. Η απόδοση της μεθόδου αυτής είναι αρκετά καλή στο πρόβλημά μας. Τονίζεται πως για την απόφαση αυτή λήφθηκε υπόψη η φύση του προβλήματος και η απόδοση σε αυτό και πως σε κάποια άλλη περίπτωση θα μπορούσαμε κάλλιστα να χρησιμοποιήσουμε κάποια διαφορετική μέθοδο.

   Επομένως σαν μέθοδος επιλογής χρησιμοποιήθηκε η μέθοδος του **τουρνουά**.

ii. **Διασταύρωση:** Όσον αφορά την διασταύρωση υπάρχουν πολλοί τρόποι να επιτευχθεί.

Η διασταύρωση μονού σημείου χωρίζει τα διανύσματα σε 2 μέρη και αυτά μεταξύ τους αντικαθίστανται εκατέρωθεν του σημείου. Δεν είναι κατάλληλη για τα δεδομένα μας, καθώς επιφέρει πολύ μεγάλη και απότομη αλλαγή στα διανύσματα χωρίς αυτός ο διαχωρισμός να επιφέρει θετικά αποτελέσματα. Ένα σημείο διαχωρισμού για το πρόβλημά μας δε κρίνεται αρκετό και κατάλληλο.

Η διασταύρωση πολλαπλού σημείου είχε τη βέλτιστη απόδοση στην εκτέλεση του κώδικα.

Η ομοιόμορφη διασταύρωση, διασταυρώνει τα γονίδια επιλέγοντας τυχαία, «ρίχνοντας ένα νόμισμα» για το αν θα γίνει η διασταύρωση ή όχι. Αυτή η τεχνική παρουσιάζει μεγάλες αυξομειώσεις στην απόδοση καθώς είναι τυχαία ενώ σε αυτό το σημείο του αλγορίθμου επιζητούμε κάτι πιο ελεγχόμενο.

Η διασταύρωση OX και PMX μοιάζουν με την διασταύρωση διπλού σημείου, με διαφορετική προσέγγιση ως προς τις αλλαγές των γονιδίων. Η τεχνική PMX «σέβεται» τις απόλυτες θέσεις, ενώ η τεχνική OX τις σχετικές θέσει των γονιδίων.

Η μέθοδος Διασταύρωσης που επιλέχθηκε τελικά είναι η Διασταύρωση πολλαπλού σημείου.

iii. **Μετάλλαξη:** Η μέθοδος του ελιτισμού εξασφαλίζει πως τα καλύτερα άτομα-λύσεις της εκάστοτε γενιάς κληρονομούνται στην επόμενη γενιά. Ένα μειονέκτημα αυτής της τεχνικής είναι πως αν ο γενετικός αλγόριθμος δεν καταφέρει μέσα σε λίγες γενιές να δημιουργήσει νέα και καλύτερα άτομα, τότε ο αριθμός των αντιγράφων της καλύτερης λύσης θα αυξηθούν πάρα πολύ και θα κυριαρχήσουν. Αυτό δε δίνει πολλές δυνατότητες στον αλγόριθμο και τον περιορίζει σε μικρό εύρος λύσεων. Στο τρέχον πρόβλημα δε κρίθηκε σκόπιμο να χρησιμοποιηθεί ο ελιτισμός.

Β2. Υλοποίηση ΓΑ

Η υλοποίηση του γενετικού αλγορίθμου βρίσκεται στο αρχείο **genetic.py**

Επισυνάπτεται επίσης το αρχείο **parse.py** και **missing_values.py** στα οποία γίνεται η απαραίτητη προεπεξεργασία και δημιουργούνται τα αρχεία **data.csv** και **data_full.csv**

Επιπλέον, στο αρχείο **neighborhood.py** βρίσκουμε τους 10 γείτονες του χρήστη που διαλέξαμε. Τέλος, με το αρχείο **errors_computation.py** υπολογίζουμε το σφάλμα του βέλτιστου ατόμου με βάση το αρχείο **ua.test**

Μαζί με τα παραπάνω αρχεία παραδόθηκε και το **genetic_v1.py,** το οποίο χρησιμοποιήσαμε για να βρούμε τις βέλτιστες παραμέτρους για το γενετικό αλγόριθμο και είναι η πρώτη έκδοση του κώδικα, πριν ολοκληρωθεί και γίνει εκτέλεση 10 φορές και ύστερα για 50 χρήστες για τις ανάγκες των ερωτημάτων Β4.β και Β4.γ.

Β3. Αξιολόγηση και Επίδραση Παραμέτρων

α) Σαν κριτήρια τερματισμού ορίστηκαν τα εξής:

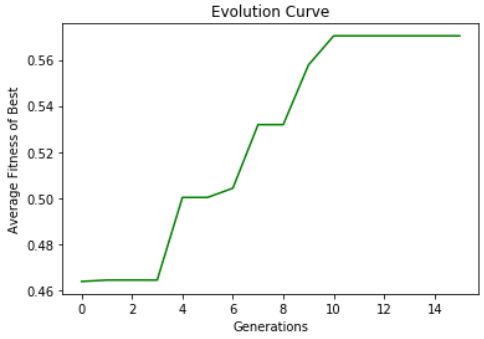
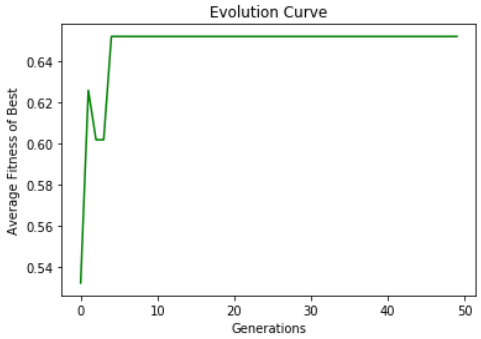
- Η μη βελτίωση του καλύτερου ατόμου του πληθυσμού για 10 γενιές συνεχόμενες πάνω από 1% (ποσοστιαία).
- Η περίπτωση που ο γενετικός αλγόριθμος φτάνει ένα μέγιστο αριθμό γενεών (50). Οι 50 γενιές δεν είναι ιδιαίτερα μεγάλος αριθμός για το σύνολο των γενεών, όμως λόγω της σύντομης σύγκλισης του αλγορίθμου δε κρίθηκε απαραίτητο να συνεχίσει η εκτέλεση για παραπάνω γενιές.

Οι παρακάτω παρατηρήσεις προκύπτουν μετά από 20 εκτελέσεις του γενετικού αλγορίθμου. Οι τιμές είναι οι μέσοι όροι των αποτελεσμάτων. Αυτό συμβαίνει διότι οι γενετικοί αλγόριθμοι είναι στοχαστικοί και τα αποτελέσματα που δίνουν μπορούν να διαφέρουν σε κάθε εκτέλεσή τους αρκετά.

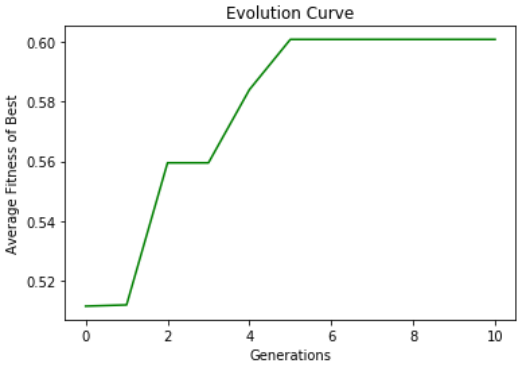
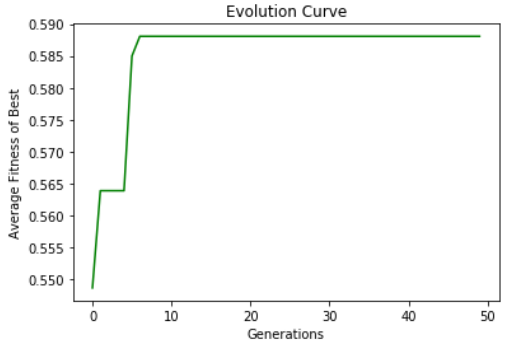
Α/Α	ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00	0.60	11.3
2	20	0.6	0.01	0.63	12.5
3	20	0.6	0.10	0.67	12.1
4	20	0.9	0.01	0.62	12.9
5	20	0.1	0.01	0.57	11
6	200	0.6	0.00	0.76	21.6
7	200	0.6	0.01	0.74	19.4
8	200	0.1	0.01	0.66	12.3
9	200	0.9	0.01	0.77	21.5

β) Οι γραφικές παραστάσεις των παραπάνω αποτελεσμάτων φαίνονται παρακάτω. Οι δύο στήλες αφορούν την εκτέλεση του γενετικού με χρήση κριτηρίων τερματισμού (early stopping) και χωρίς αυτό.

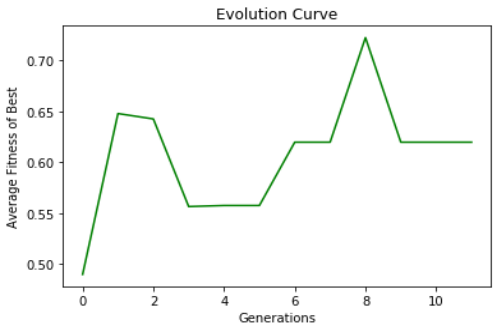
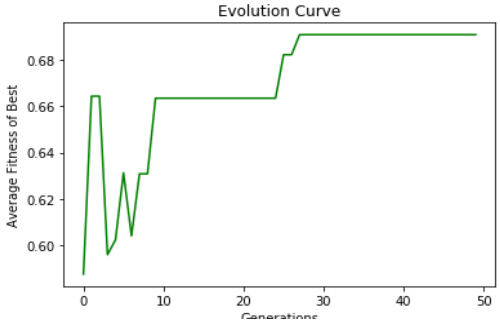
1. **POP_SIZE =20, P_CROS=0.6, P_MUT=0.00**

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

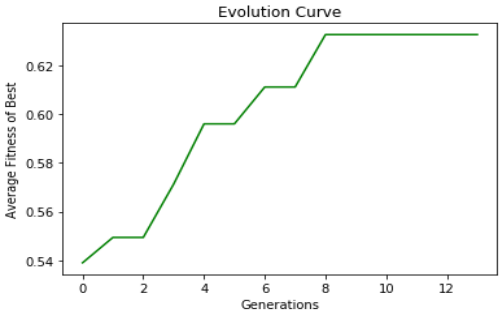
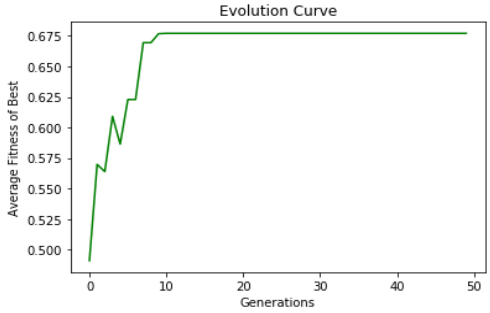
2. **POP_SIZE =20, P_CROS=0.6, P_MUT=0.01**

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

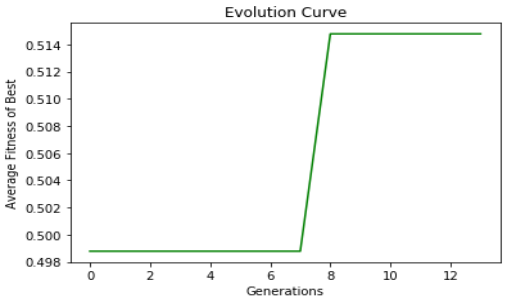
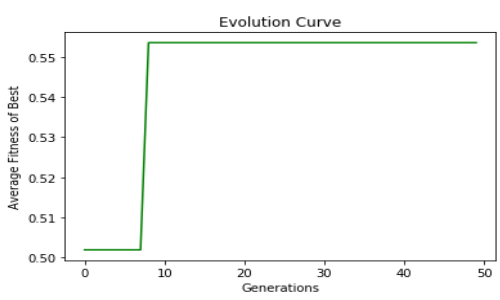
3. POP_SIZE =20, P_CROS=0.6, P_MUT=0.10

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

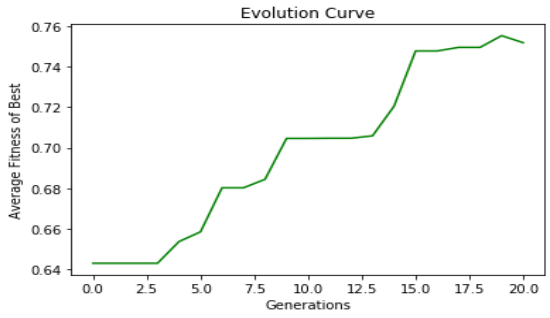
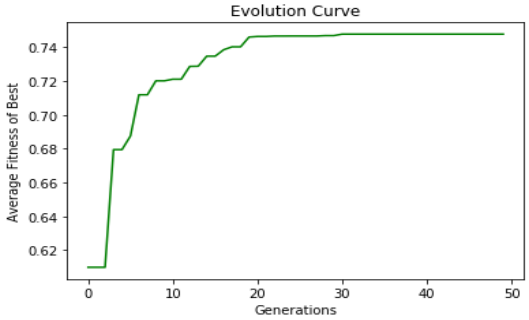
4. POP_SIZE =20, P_CROS=0.9, P_MUT=0.01

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

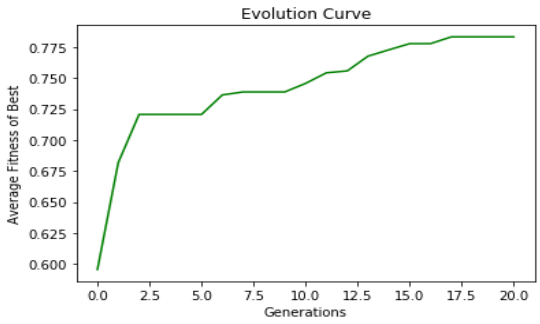
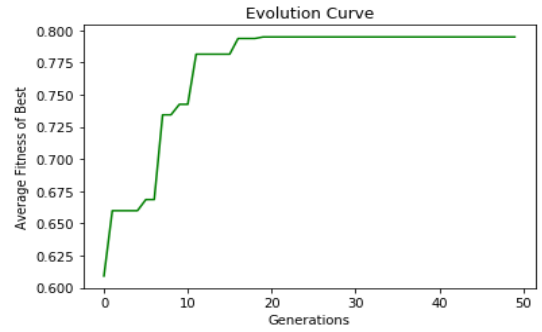
5. POP_SIZE =20, P_CROS=0.1, P_MUT=0.01

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

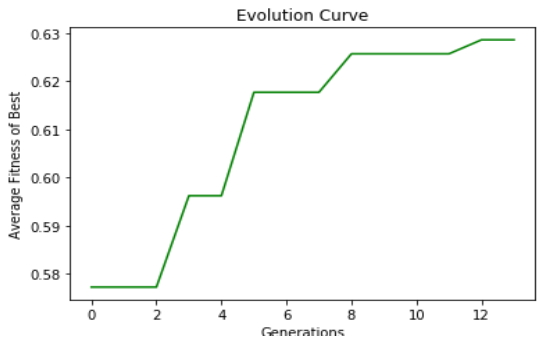
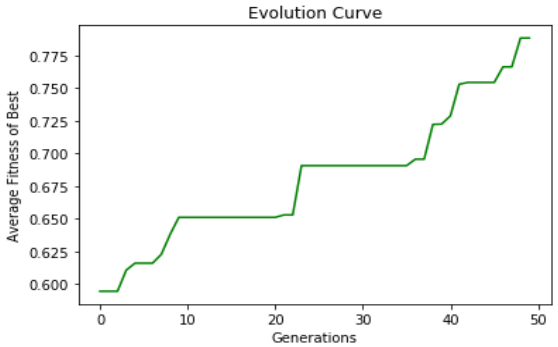
6. POP_SIZE =200, P_CROS=0.6, P_MUT=0.00

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

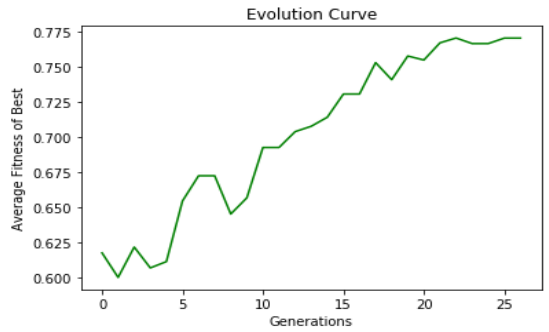
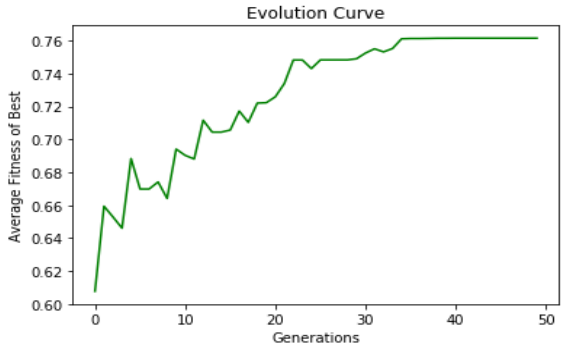
7. POP_SIZE =200, P_CROS=0.6, P_MUT=0.01

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

8. POP_SIZE =200, P_CROS=0.1, P_MUT=0.01

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

9. POP_SIZE =200, P_CROS=0.9, P_MUT=0.01

Με κριτήρια τερματισμού	*Χωρίς κριτήρια τερματισμού*
	

γ) Αρχικά η εκτέλεση του αλγορίθμου για το **πληθυσμό** 200 ατόμων είναι αισθητά πιο αργή από αυτή των 20 ατόμων, κάτι πολύ φυσιολογικό δεδομένου πως το μέγεθος των δεδομένων είναι μίας τάξης μεγέθους μεγαλύτερα. Επίσης, η σύγκλιση του αλγορίθμου είναι φανερά και λογικά πιο αργή. Παρ ’όλα αυτά με μεγαλύτερο αριθμό ατόμων παίρνουμε καλύτερα αποτελέσματα. Ο λόγος είναι πως υπάρχει μεγαλύτερη ποικιλία χρωμοσωμάτων, άρα ο χώρος αναζήτησης μεγεθύνεται, και επομένως μετά από μεγαλύτερο αριθμό γενεών μπορούμε να φτάσουμε σε πιο κατάλληλη λύση, με μεγαλύτερη απόδοση.

Η παράμετρος **πιθανότητα διασταύρωσης** επηρεάζει τον αριθμό των ατόμων που θα διασταυρωθούν σε κάθε γενιά. Όσο μικρότερος είναι αυτός ο αριθμός τόσο λιγότερα τα άτομα που διασταυρώνονται και το αντίστροφο. Όταν ο αριθμός είναι πολύ μικρός (0.1) τα διαθέσιμα γονίδια περιορίζονται σε μικρό αριθμό και εν συνεχεία οι απόγονοι αποτελούνται από λίγα και δίχως ποικιλία γονίδια. Όταν η πιθανότητα είναι μεγάλη (0.9) επιτρέπεται σε άτομα με χαμηλή απόδοση να συνεισφέρουν με τα γονίδια τους στην επόμενη γενιά, κάτι που εμφανώς δε προτιμούμε. Η σύγκλιση με αυτό το τρόπο καθυστερεί. Επομένως, η καλύτερη από τις παραπάνω επιλογές είναι μια μέτρια πιθανότητα διασταύρωσης (0.6).

Η παράμετρος **πιθανότητα μετάλλαξης** ορίζει το αν θα γίνει μια μετάλλαξη σε ένα γονίδιο. Σκοπός αυτής της παραμέτρου είναι να διατηρήσει μία ισορροπία ανάμεσα στη διατήρηση των καλύτερων λύσεων και στην συνολική εξερεύνηση του χώρου των λύσεων. Στις περιπτώσεις που δεν υπάρχει (ισούται με 0.00) τα άτομα τείνουν να γίνονται ίδια από μία γενιά και ύστερα. Όταν αυτή η πιθανότητα αυξάνεται (0.10) περισσότερα γονίδια μεταλλάσσονται και ο χώρος αναζήτησης του αλγορίθμου αυξάνεται. Βγαίνει από το μέχρι ώρας χώρο των μεταβλητών. Αυτά έχουν ως αποτέλεσμα μια πιο αργή σύγκλιση. Από τις 3 τιμές που δίνουμε σαν πιθανότητα μετάλλαξης, η πιο μέτρια (0.01) επιφέρει διατήρηση της ζητούμενης ισορροπίας.

Β4. Αξιολόγηση Συστάσεων

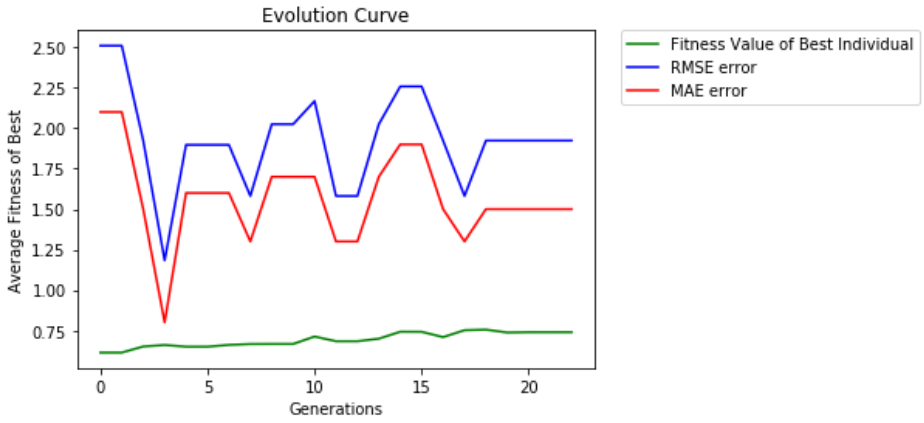
α) Για τις απαιτήσεις αυτού του ερωτήματος χρησιμοποιήθηκε **το αρχείο ua.test** για την αξιολόγηση των λύσεων που παρήγαγε ο γενετικός αλγόριθμος, με χρήση των μετρικών RMSE και MAE, και το αρχείο **ua.dase** για την εκπαίδευση του γενετικού.

β) Οι τιμές των RMSE (Root Mean Squared Error) και MAE (Mean Average Error) μετρικών που παρουσιάζονται παρακάτω είναι ο μέσος όρος 10 εκτελέσεων του αλγορίθμου.

Στις εκτελέσεις χρησιμοποιήθηκε η συνάρτηση επιδιόρθωσης ώστε να ελαχιστοποιηθεί το σφάλμα.

Ως βέλτιστες παράμετροι για το γενετικό αλγόριθμο θεωρήθηκαν οι εξής τιμές των παραμέτρων:

- Μέγεθος Πληθυσμού: 200 άτομα (POP_SIZE)
- Πιθανότητα Διασταύρωσης: 0.6 (P_CROS)
- Πιθανότητα Μετάλλαξης: 0.01 (P_MUT)

RMSE =2.05 ,MAE=1.69


γ) Παρακάτω παρουσιάζονται οι τιμές των μετρικών *RMSE και MAE για 50 χρήστες*

RMSE = 1.99, MAE=1.60

Στην εκκίνηση εκτέλεσης του αρχείου **genetic.py** γίνεται ερώτηση προς το χρήστη για την εκτέλεση του εκτέλεση 10 φορών και υπολογισμού του σφάλματος και παρουσίαση των γραφικών παραστάσεων (Β4.β) και για την εκτέλεση του αλγορίθμου για 50 χρήστες και υπολογισμό του μέσου σφάλματος.