ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Πρόβλεψη δεικτών OCEAN με μοντέλα μηχανικής μάθησης** **Predicting OCEAN indices with machine learning models**
Ονοματεπώνυμο Φοιτητή	**Μάριος Χαρτσιάς**
Πατρώνυμο	**Θεόδωρος**
Αριθμός Μητρώου	**ΜΠΠΛ/ 21080**
Επιβλέπων	**Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης **Ιούνιος 2024**

Τριμελής Εξεταστική Επιτροπή

Δημήτριος Σωτηρόπουλος Επίκουρος Καθηγητής	Γεώργιος Τσιχριντζης Καθηγητής	Σακκόπουλος Ευάγγελος Αναπληρωτής Καθηγητής

Περίληψη

Η μεταπτυχιακή διατριβή "Predicting OCEAN indices with ML models" επικεντρώνεται στη χρήση τεχνικών μηχανικής μάθησης για την πρόβλεψη χαρακτηριστικών προσωπικότητας με βάση το μοντέλο OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism). Τα δεδομένα, τα οποία συλλέχθηκαν από ένα διαδικτυακό τεστ προσωπικότητας με τη χρήση των Big-Five Factor Markers του IPIP, καλύπτουν το διάστημα από το 2016 έως το 2018. Η έρευνα χρησιμοποιεί διάφορα μοντέλα ML, συμπεριλαμβανομένης της γραμμικής και μη γραμμικής παλινδρόμησης, και έναν γενετικό αλγόριθμο για την πρόβλεψη των βαθμολογιών OCEAN από τις απαντήσεις του τεστ. Η μελέτη επεξεργάζεται σχολαστικά τα δεδομένα, βελτιστοποιεί τις παραμέτρους του μοντέλου και συγκρίνει την αποτελεσματικότητα των διαφόρων προσεγγίσεων, αποδεικνύοντας τις δυνατότητες της ML στην ψυχολογική αξιολόγηση και προσφέροντας πληροφορίες σχετικά με την προβλεπτική δύναμη της μηχανικής μάθησης στην κατανόηση των ανθρώπινων χαρακτηριστικών προσωπικότητας.

Abstract

The master's thesis "Predicting OCEAN indices with ML models" focuses on the use of machine learning techniques to predict personality traits based on the OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism) model. The data, collected from an online personality test using the Big-Five Factor Markers of the IPIP, span from 2016 to 2018. The research utilizes various ML models, including linear and non-linear regression, and a genetic algorithm to predict OCEAN scores from test responses. The study meticulously preprocesses the data, optimizes model parameters, and compares the efficacy of different approaches, demonstrating the potential of ML in psychological assessment and offering insights into the predictive power of machine learning in understanding human personality traits.

Περιεχόμενα

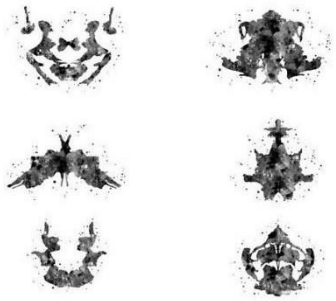
1. Εισαγωγή – Ορισμός του προβλήματος

1.1 Προσωπικότητα

Η προσωπικότητα αντιπροσωπεύει το σύνολο των ψυχολογικών χαρακτηριστικών που διαμορφώνουν τη συμπεριφορά και τις εσωτερικές τάσεις ενός ανθρώπου. Αποτελείται από τον τρόπο σκέψης, τις αξίες, τα συναισθήματα, τις επιδιώξεις και τις προσδοκίες του ατόμου. Η προσωπικότητα επηρεάζει τις διαπροσωπικές σχέσεις, τις επαγγελματικές επιλογές και τη γενικότερη συμπεριφορά ενός ατόμου στο κοινωνικό περιβάλλον. Η αυτογνωσία της προσωπικότητας βοηθά στην αυτοβελτίωση. Κεντρικές θεωρίες περιλαμβάνουν το μοντέλο Big Five, τον δείκτη Myers-Briggs, τον Keirsey Temperament Sorter, το μοντέλο τριών παραγόντων και το ερωτηματολόγιο προσωπικότητας του Eysenck.

1.2 Μοντέλα κατανόησης προσωπικότητας

Γνωστά μοντέλα κατανόησης προσωπικότητας τα οποία προηγήθηκαν του Big Five Personality Trais είναι τα εξής:

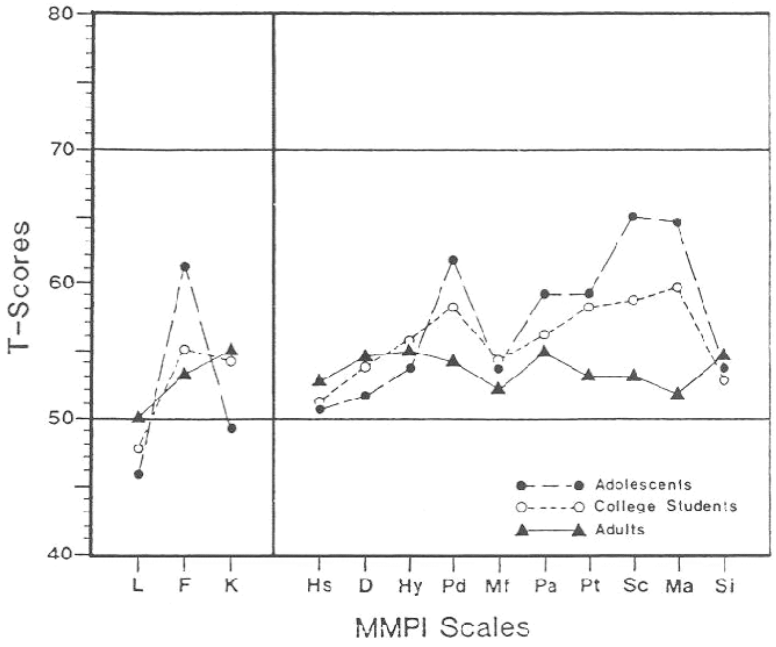


1. Στο Rorschach Inkblot Test, οι συμμετέχοντες παρουσιάζονται με μια σειρά από μελανές εικόνες και καλούνται να περιγράψουν τι βλέπουν. Οι απαντήσεις τους αναλύονται για να εξάγουν συμπεράσματα για την προσωπικότητά τους, βασιζόμενες σε τι επιλέγουν να επικεντρωθούν, τις ερμηνείες τους και την αλληλεπίδρασή τους με τα σχήματα και τις συνδέσεις που κάνουν. Αυτό το τεστ βασίζεται στην υπόθεση ότι οι αντιδράσεις των ατόμων σε ασαφείς και πολυερμηνευόμενες εικόνες αποκαλύπτουν τα υποκείμενα στοιχεία της προσωπικότητάς τους.

Εικόνα 1: Μελανές εικόνες του Rorschach Inkblot Test (1921)

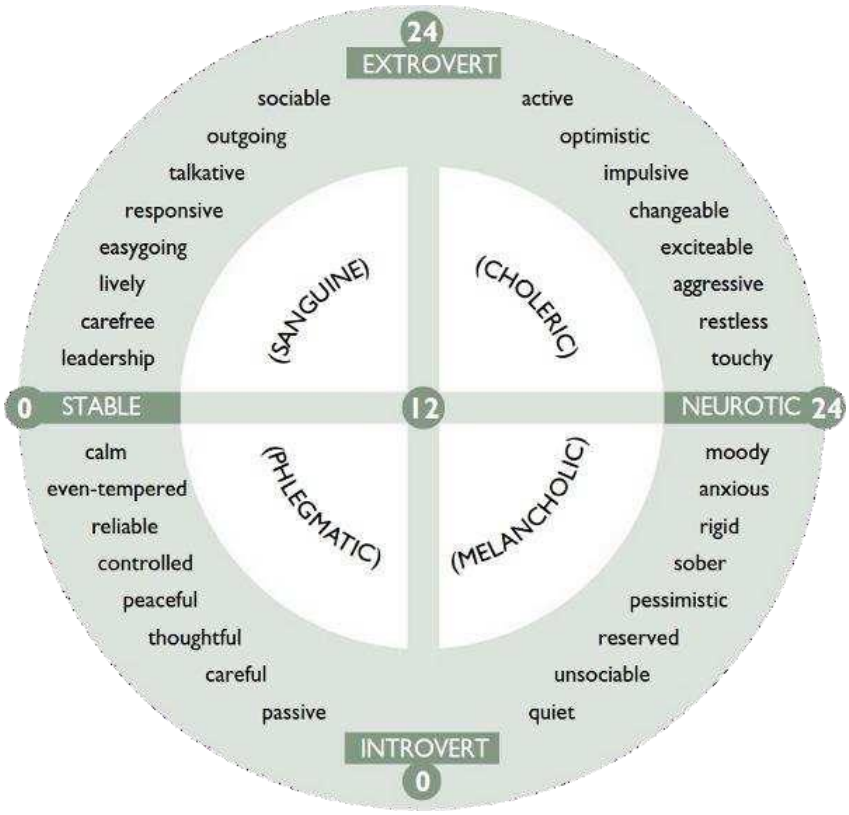
2. Στο Thematic Apperception Test (TAT), δίνονται στους συμμετέχοντες σειρές εικόνων με ασαφείς, πολυσήμαντες καταστάσεις. Οι συμμετέχοντες καλούνται να δημιουργήσουν μια ιστορία για κάθε εικόνα, περιγράφοντας τι συμβαίνει, τις σκέψεις και τα συναισθήματα των χαρακτήρων, καθώς και την έκβαση της σκηνής. Οι αφηγήσεις αυτές αναλύονται στη συνέχεια για να αποκαλυφθούν τα υποκείμενα συναισθήματα, σκέψεις και θέματα της προσωπικότητας του ατόμου.

Εικόνα 2: Thematic Apperception Test (TAT) (1930s): Χρησιμοποιεί εικόνες για να εξάγει αφηγήσεις από τους συμμετέχοντες.



Εικόνα 3: Minnesota Multiphasic Personality Inventory (MMPI) (1943): Εκτενές ερωτηματολόγιο που ανιχνεύει διάφορες ψυχολογικές παραμέτρους.

3. Το Minnesota Multiphasic Personality Inventory (MMPI) είναι ένα ερωτηματολόγιο που αποσκοπεί στην αξιολόγηση της ψυχολογικής κατάστασης και των προσωπικότητας χαρακτηριστικών ενός ατόμου. Αποτελείται από μια σειρά δηλώσεων στις οποίες οι συμμετέχοντες απαντούν συμφωνώντας ή διαφωνώντας. Οι απαντήσεις αυτές στη συνέχεια αναλύονται για να διαπιστωθούν τυχόν ψυχολογικές τάσεις ή διαταραχές, καθώς και για να προσδιοριστούν διάφορες πτυχές της προσωπικότητας.



Εικόνα 4: Eysenck Personality Questionnaire (EPQ) (1960s): Μετρά την εξωστρέφεια-εσωστρέφεια και το νευρωτισμό-σταθερότητα.

4. Το Eysenck Personality Questionnaire (EPQ) αναπτύχθηκε από τον Hans Eysenck και μετρά τρεις κύριες διαστάσεις της προσωπικότητας: την Εξωστρέφεια (E), τη Νευρωτισμό (N), και την Ψυχισμό (P). Οι συμμετέχοντες απαντούν σε ερωτήσεις σε μορφή διλήμματος (ναι ή όχι), και οι απαντήσεις τους αντικατοπτρίζουν την τοποθέτησή τους στις τρεις αυτές κλίμακες. Η εξωστρέφεια σχετίζεται με την κοινωνικότητα και τη δραστηριότητα, ο νευρωτισμός με την συναισθηματική σταθερότητα, και ο ψυχισμός με την εχθρικότητα και την αντικοινωνική συμπεριφορά.

1.2 Big Five Personality traits

Για την αποκρυπτογράφηση της σημασίας του Τεστ Προσωπικότητας Big Five, είναι καίριο να αναλύσουμε τις πέντε κύριες διαστάσεις που διαμορφώνουν το εν λόγω μοντέλο. Οι διαστάσεις αυτές, γνωστές με το ακρωνύμιο OCEAN, καλύπτουν μια εκτενή γκάμα από ψυχολογικά χαρακτηριστικά, προσφέροντας έναν λεπτομερή μηχανισμό για την κατανόηση των προσωπικών ιδιαιτεροτήτων ενός ανθρώπου.

Το ψυχολογικό τεστ των Μεγάλων Πέντε χαρακτηριστικών αναπτύχθηκε μέσω δεκαετιών έρευνας. Η διαδικασία ξεκίνησε με την λεξική υπόθεση, όπου ερευνητές ανακάλυψαν ότι σχεδόν όλα τα χαρακτηριστικά προσωπικότητας μπορούν να περιγραφτούν με λέξεις της καθημερινής γλώσσας. Στη συνέχεια, μέσω εκτενών ερευνών και στατιστικής ανάλυσης, αναδείχθηκαν πέντε κύριες διαστάσεις που κατέστησαν σαφές ότι μπορούν να περιγράψουν με επάρκεια την ποικιλομορφία της ανθρώπινης προσωπικότητας.

Παράγοντες Big5	Υποκλίμακες Big5
(O) Διαθεσιμότητα σε Εμπειρίες vs Συμβατικότητα	1. Φαντασία 2. Καλλιτεχνικό ενδιαφέρον 3. Συναισθηματικότητα 4. Περιπετέτεια 5. Πνευματικοί Ορίζοντες 6. Προοδευτικότητα
(C) Ευσυνειδησία vs Αδιαφορία	1. Αυτεπάρκεια 2. Επιμέλεια 3. Υπευθυνότητα 4. Φιλοδοξία 5. Αυτοπειθαρχία 6. Επιφυλακτικότητα
(E) Εξωστρέφεια vs Εσωστρέφεια	1. Φιλικότητα 2. Κοινωνικότητα 3. Επιβλητικότητα 4. Επιδίωξη ενθουσιασμού 5. Ενεργητικότητα 6. Θετικότητα
(A) Συνεργατικότητα vs Ανταγωνιστικότητα	1. Εμπιστοσύνη 2. Ηθική 3. Αλτρουισμός 4. Συμβιβαστικότητα 5. Σεμνότητα 6. Συμπόνια
(N) Νευρωτισμός vs Συναισθηματική Αστάθεια	1. Άγχος / Αγωνία 2. Θυμός 3. Κατάθλιψη 4. Ανασφάλεια 5. Παρορμητικότητα 6. Τρωτότητα / Ευπάθεια

Εικόνα 5: Πίνακας με τους δείκτες της προσωπικότητας του Big Five Personality Traits

Το ψυχολογικό τεστ αυτό γνωστό και ως OCEAN, είναι ένα ευρέως αναγνωρισμένο στην ψυχολογία που χρησιμοποιείται για την περιγραφή και τη μέτρηση της προσωπικότητας ενός ατόμου. Το OCEAN είναι ένα ακρωνύμιο που σημαίνει Ανοιχτότητα, Ευσυνειδησία, Εξωστρέφεια, Ευπροσάρμοστος και Νευρωτισμός, καθένα από τα οποία αντιπροσωπεύει έναν θεμελιώδη τομέα της ανθρώπινης προσωπικότητας:

Ανοιχτότητα: Αυτό το χαρακτηριστικό διαθέτει χαρακτηριστικά όπως η φαντασία, η διορατικότητα και ένα ευρύ φάσμα ενδιαφερόντων. Τα άτομα με υψηλό βαθμό ανοιχτότητας είναι συχνά περίεργα, δημιουργικά και ανοιχτά σε νέες εμπειρίες. Είναι πιο πιθανό να ασχοληθούν με καλλιτεχνικές και πνευματικές αναζητήσεις και είναι δεκτικά σε νέες ιδέες, αντισυμβατικές αξίες και ποικίλες εμπειρίες.

Ευσυνειδησία: Αυτή η διάσταση αναφέρεται στο επίπεδο οργάνωσης, αξιοπιστίας και πειθαρχίας που επιδεικνύει ένα άτομο. Οι ευσυνείδητοι άνθρωποι τείνουν να είναι επιμελείς, καλά οργανωμένοι και αξιόπιστοι. Συχνά είναι σχολαστικοί στην εργασία τους, σχεδιάζουν εκ των προτέρων και τηρούν τις δεσμεύσεις τους, επιδεικνύοντας έντονη αίσθηση καθήκοντος και ευθύνης.

Εξωστρέφεια: Η εξωστρέφεια χαρακτηρίζεται από κοινωνικότητα, ενθουσιασμό και διεκδικητικότητα. Τα εξωστρεφή άτομα είναι συνήθως ενεργητικά και αναζητούν διέγερση στην παρέα των άλλων. Συχνά γίνονται αντιληπτά ως ομιλητικά, εξωστρεφή και τείνουν περισσότερο να συμμετέχουν σε κοινωνικές δραστηριότητες, δείχνοντας προτίμηση στην εξωτερική διέγερση έναντι των μοναχικών δραστηριοτήτων.

Συμφωνησιμότητα: Αυτό το χαρακτηριστικό αντανακλά την τάση ενός ατόμου προς τον αλτρουισμό, την εμπιστοσύνη, τη συνεργασία και τη συμπόνια. Τα συμφιλιωτικά άτομα είναι γενικά ευγενικά, ενσυναισθητικά και αρμονικά. Είναι πιο πιθανό να εκτιμούν την καλή συνεννόηση με τους άλλους, να δείχνουν ενδιαφέρον για την ευημερία των άλλων και είναι ικανά να συμβιβάζουν τα συμφέροντά τους για χάρη της διατήρησης θετικών σχέσεων.

Νευρωτισμός: Αυτή η διάσταση μετρά την τάση για συναισθηματική αστάθεια, άγχος και κυκλοθυμία. Τα άτομα με υψηλά επίπεδα νευρωτισμού μπορεί να εμφανίζουν μεγαλύτερη τάση για αρνητικά συναισθήματα όπως άγχος, θλίψη ή ευερεθιστότητα. Μπορεί να είναι πιο επιρρεπείς στο να βιώνουν άγχος, ανησυχία και μπορεί να δυσκολεύονται να αντιμετωπίσουν αγχωτικές καταστάσεις.

Το μοντέλο OCEAN χρησιμοποιείται σε διάφορα περιβάλλοντα, από την ακαδημαϊκή έρευνα έως την κλινική πρακτική και τα οργανωτικά περιβάλλοντα, για την κατανόηση και την πρόβλεψη της συμπεριφοράς, των προτιμήσεων και της διαπροσωπικής δυναμικής. Στο πλαίσιο της παρούσας διατριβής, η πρόκληση είναι να προβλεφθούν με ακρίβεια αυτές οι πέντε διαστάσεις χρησιμοποιώντας ένα υποσύνολο δεδομένων που χρησιμοποιούνται συνήθως για την αξιολόγησή τους.

1.3 Υπόδειγμα τεστ του Big Five Personality Trais

1.3.1 Personality Traits and Items

Παρακάτω δίνονται οι ερωτήσεις στις οποίες ο υποψήφιος απαντάει προκειμένου το τελικό αποτέλεσμα να εξαχθεί και να υπολογιστούν οι αντίστοιχοι δείκτες της προσωπικότητας του ατόμου. Κάθε ερώτηση απαντάται με αριθμούς από 1-Διαφωνώ έως 5-Συμφωνώ. Οι ερωτήσεις για τον κάθε δείκτη δίνονται παρακάτω:

- Openness (O)

OPN1: I have a rich vocabulary.

OPN2: I have difficulty understanding abstract ideas.

OPN3: I have a vivid imagination.

OPN4: I am not interested in abstract ideas.

OPN5: I have excellent ideas.

OPN6: I do not have a good imagination.

OPN7: I am quick to understand things.

OPN8: I use difficult words.

OPN9: I spend time reflecting on things.

OPN10: I am full of ideas.

- Conscientiousness (C)

CSN1: I am always prepared.

CSN2: I leave my belongings around.

CSN3: I pay attention to details.

CSN4: I make a mess of things.

CSN5: I get chores done right away.

CSN6: I often forget to put things back in their proper place.

CSN7: I like order.

CSN8: I shirk my duties.

CSN9: I follow a schedule.

CSN10: I am exacting in my work.

- Extroversion (E)

EXT1: I am the life of the party.

EXT2: I don't talk a lot.

EXT3: I feel comfortable around people.

EXT4: I keep in the background.

EXT5: I start conversations.

EXT6: I have little to say.

EXT7: I talk to a lot of different people at parties.

EXT8: I don't like to draw attention to myself.

EXT9: I don't mind being the center of attention.

EXT10: I am quiet around strangers.

- Agreeableness (A)

AGR1: I feel little concern for others.
AGR2: I am interested in people.
AGR3: I insult people.
AGR4: I sympathize with others' feelings.
AGR5: I am not interested in other people's problems.
AGR6: I have a soft heart.
AGR7: I am not really interested in others.
AGR8: I take time out for others.
AGR9: I feel other's emotions.
AGR10: I make people feel at ease.

- Neuroticism (N)

NRT1: I get stressed out easily.
NRT2: I am relaxed most of the time.
NRT3: I worry about things.
NRT4: I seldom feel blue.
NRT5: I am easily disturbed.
NRT6: I get upset easily.
NRT7: I change my mood a lot.
NRT8: I have frequent mood swings.
NRT9: I get irritated easily.
NRT10: I often feel blue.

1.3.2 Φόρμουλα υπολογισμού των δεικτών:

Εφόσον ο υποψήφιος απαντήσει στις παραπάνω ερωτήσεις, το τελικό σκορ για κάθε δείκτη υπολογίζεται με τον εξής τρόπο για τον κάθε δείκτη:

- **O**=8+OPN1-OPN2+OPN3-OPN4+OPN5-OPN6+OPN7+OPN8+OPN9+OPN10
- **C**=14+CSN1-CSN2+CSN3-CSN4+CSN5-CSN6+CSN7-CSN8+CSN9+CSN10
- **E**=20+EXT1-EXT2+EXT3-EXT4+EXT5-EXT6+EXT7-EXT8+EXT9-EXT10
- **A**=14-AGR1+AGR2-AGR3+AGR4-AGR5+AGR6-AGR7+AGR8+AGR9+AGR10
- **N**=38-NRT1+NRT2-NRT3+NRT4-NRT5-NRT6-NRT7-NRT8-NRT9-NRT10

Οι διαφορετικές σταθερές τιμές σε κάθε δείκτη του τύπου των Big Five χαρακτηριστικών προσωπικότητας βοηθούν στην ομαλοποίηση και κλιμάκωση των βαθμολογιών. Αυτές οι σταθερές προσαρμόζονται στη βασική γραμμή κάθε χαρακτηριστικού, λαμβάνοντας υπόψη τον ποικίλο αριθμό ερωτήσεων και το πιθανό εύρος βαθμολογιών για κάθε χαρακτηριστικό. Εξασφαλίζουν ότι η τελική βαθμολογία κάθε χαρακτηριστικού είναι συγκρίσιμη μεταξύ διαφορετικών ατόμων, παρέχοντας ένα τυποποιημένο μέτρο των χαρακτηριστικών προσωπικότητας ανεξάρτητα από τον αριθμό και τον τύπο των ερωτήσεων που χρησιμοποιούνται για την αξιολόγηση του καθενός.

Τα σκορ των ερωτήσεων δεν μπορούν να περιγράψουν εξ ολοκλήρου την προσωπικότητα ενός ατόμου στο βάθος του χρόνου. Η πολυπλοκότητα της προσωπικότητας που στο βάθος χρόνου είναι μεταβλητή και πολυδιάστατη αλλάζει με τον χρόνο. Θα μπορούσαμε να την παρομοιάσουμε ως ένα στιγμιότυπο του χρόνου στο οποίο ο υποψήφιος έχει συγκεκριμένα χαρακτηριστικά που θα τον ωθήσουν με πιθανότητες σε συγκεκριμένες αποφάσεις από τις οποίες ίσως κληθεί να πάρει. Το σίγουρο είναι ότι σε ομάδες ανθρώπων παρατηρούνται μοτίβα συμπεριφορών και επιλογές στην καριέρα. Αυτός είναι ένας λόγος που εταιρείες προσλήψεων χρησιμοποιούν το BFPT προκειμένου να ελέγξουν την καταλληλόλητα ενός υποψηφίου για τις θέσεις εργασίας που παρέχουν.

Επίσης πρέπει να σημειωθεί ότι κάθε οργανισμός χρησιμοποιεί τις δικές του σταθερές βασιζόμενος στα στατιστικά και στην διαφορετική κανονικοποίηση που προτίθεται να εξάγει συμπεράσματα.

1.4 Παραδείγματα εργασιών βάσει αποτελεσμάτων του BFPT

Στην συνέχεια δίνονται παραδείγματα στα οποία είναι πιθανό να βρεθούν άτομα στατιστικά βασισμένα στο μοτίβο των δεικτών που παρουσιάζουν.

Ανοιχτότητα: Τα άτομα με υψηλό βαθμό Ανοιχτότητας είναι ευφάνταστα και περίεργα. Ευδοκιμούν σε περιβάλλοντα που ενθαρρύνουν την καινοτομία και τη δημιουργικότητα, γεγονός που τους καθιστά κατάλληλους για ρόλους που απαιτούν σκέψη έξω από το κουτί, όπως στις τέχνες ή την επιστημονική έρευνα, όπου οι νέες ιδέες εκτιμώνται.

Ευσυνειδησία: Οι ιδιαίτερα ευσυνείδητοι άνθρωποι είναι οργανωμένοι και αξιόπιστοι. Διακρίνονται σε ρόλους που απαιτούν σχολαστικό σχεδιασμό και εκτέλεση, όπως στα οικονομικά ή τη διοίκηση, όπου η ακρίβεια και η τάξη είναι υψίστης σημασίας.

Εξωστρέφεια: Τα εξωστρεφή άτομα ενεργοποιούνται από τις κοινωνικές αλληλεπιδράσεις. Αποδίδουν καλά σε ρόλους που περιλαμβάνουν ομαδική εργασία, επικοινωνία και συμμετοχή στο κοινό, όπως οι πωλήσεις ή η διδασκαλία, όπου η ικανότητά τους να συνδέονται με τους άλλους αποτελεί βασικό πλεονέκτημα.

Συμφωνητικότητα: Όσοι έχουν υψηλή βαθμολογία στη συμφωνητικότητα είναι συνεργάσιμοι και συμπονετικοί. Είναι κατάλληλοι για ρόλους στην υγειονομική περίθαλψη, την εξυπηρέτηση πελατών ή τη συμβουλευτική, όπου η ενσυναίσθηση και οι διαπροσωπικές δεξιότητες είναι ζωτικής σημασίας για την επιτυχία.

Νευρωτισμός: Τα άτομα με υψηλότερα επίπεδα νευρωτισμού μπορεί να διαπρέψουν σε ρόλους που απαιτούν συνεχή επαγρύπνηση και διαχείριση κινδύνων, όπως στη συμμόρφωση με την ασφάλεια ή τη διασφάλιση ποιότητας, όπου η τάση τους να προβλέπουν προβλήματα μπορεί να οδηγήσει σε καλύτερη ετοιμότητα και προσοχή.

1.5 Πρόβλημα-απόπειρα πρόβλεψης των δεικτών με λειψές ερωτήσεις

Στην παρούσα μεταπτυχιακή διατριβή, το πρωταρχικό πρόβλημα που αντιμετωπίζεται είναι η πρόβλεψη των πέντε δεικτών προσωπικότητας OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism) χρησιμοποιώντας ένα μειωμένο σύνολο δεδομένων. Συγκεκριμένα, η πρόκληση έγκειται στην ακριβή πρόβλεψη αυτών των δεικτών με βάση μόνο 40 απαντήσεις από το τεστ προσωπικότητας, σε αντίθεση με το πλήρες σύνολο των 50 απαντήσεων. Αυτή η μείωση στη διαθεσιμότητα των δεδομένων εισάγει πολυπλοκότητα, καθώς κάθε μία από τις 10 απαντήσεις που παραλείπονται θα μπορούσε να περιέχει κρίσιμες πληροφορίες για τα χαρακτηριστικά της προσωπικότητας του ατόμου, επηρεάζοντας ενδεχομένως την ακρίβεια πρόβλεψης των χρησιμοποιούμενων μοντέλων.

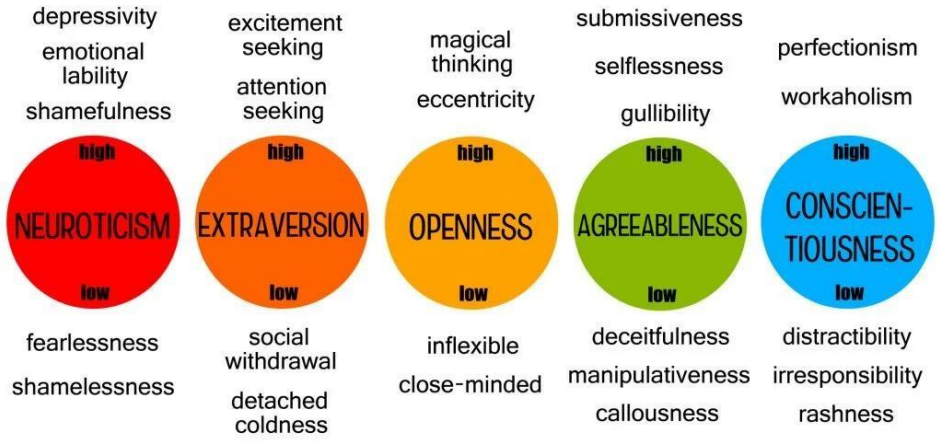
Η έρευνα επιδιώκει να αντιμετωπίσει αυτή την πρόκληση χρησιμοποιώντας προηγμένες τεχνικές μηχανικής μάθησης και έναν γενετικό αλγόριθμο για τη βελτιστοποίηση της απόδοσης του μοντέλου παρά τον περιορισμό των δεδομένων. Ο γενετικός αλγόριθμος διαδραματίζει κρίσιμο ρόλο στην επανάληψη μέσω διαφόρων διαμορφώσεων μοντέλων και υποσυνόλων χαρακτηριστικών, με στόχο τον εντοπισμό του βέλτιστου συνδυασμού που αντισταθμίζει τα δεδομένα που λείπουν, διατηρώντας, ή ακόμη και ενισχύοντας, την ακρίβεια πρόβλεψης των δεικτών OCEAN.

Η αξιοποίηση μοντέλων μηχανικής μάθησης, τόσο γραμμικών όσο και μη γραμμικών, έχει κεντρικό ρόλο σε αυτή την προσπάθεια. Αυτά τα μοντέλα είναι επιφορτισμένα με τη διάκριση μοτίβων και σχέσεων εντός του συνόλου δεδομένων των 40 απαντήσεων που είναι ενδεικτικά των ευρύτερων χαρακτηριστικών προσωπικότητας που συνήθως προκύπτουν από τις πλήρεις 50 απαντήσεις. Η διαδικασία περιλαμβάνει την εκπαίδευση των μοντέλων στα διαθέσιμα δεδομένα, την επικύρωση της απόδοσής τους και τη συνεχή βελτίωσή τους μέσω της διαδικασίας βελτιστοποίησης του γενετικού αλγορίθμου, ώστε να επιτευχθεί η καλύτερη δυνατή ακρίβεια πρόβλεψης.

Το πρόβλημα αυτό είναι σημαντικό όχι μόνο από τεχνικής άποψης, καθώς περιλαμβάνει πολύπλοκη ανάλυση δεδομένων και βελτιστοποίηση μοντέλων, αλλά και από ψυχολογική και πρακτική άποψη. Η επιτυχής πρόβλεψη δεικτών προσωπικότητας με μειωμένα δεδομένα μπορεί να έχει

ουσιαστικές επιπτώσεις σε διάφορες εφαρμογές, από το στοχευμένο μάρκετινγκ έως τις εξατομικευμένες συστάσεις και ακόμη και τις αξιολογήσεις ψυχικής υγείας, όπου οι ολοκληρωμένες αξιολογήσεις προσωπικότητας μπορεί να μην είναι πάντα εφικτές.

Η έρευνα έχει ως στόχο να συμβάλει στον τομέα αυτό, αποδεικνύοντας ότι με τις κατάλληλες αναλυτικές τεχνικές και υπολογιστικές προσεγγίσεις, είναι δυνατόν να αποκτηθούν ουσιαστικές γνώσεις για την ανθρώπινη προσωπικότητα με περιορισμένα δεδομένα. Αυτό θα μπορούσε να ανοίξει το δρόμο για πιο προσαρμόσιμες και αποδοτικές ως προς τους πόρους μεθοδολογίες στην ψυχολογική έρευνα και τις εφαρμογές, αναδεικνύοντας τη δύναμη της μηχανικής μάθησης και των γενετικών αλγορίθμων στην υπέρβαση των περιορισμών δεδομένων και στην εξαγωγή ουσιαστικών μοτίβων από ελλιπή σύνολα δεδομένων.



ΕΙΚΟΝΑ 6: Δείκτες (OCEAN) του ψυχολογικού τεστ Big Five Personality Traits

1.6 Στρατηγική πρόβλεψης των δεικτών με λειψές ερωτήσεις

Στην παρούσα μεταπτυχιακή διατριβή ο τρόπος με τον οποίο θα προσπαθήσουμε να επιλύσουμε το πρόβλημα της πρόβλεψης των δεικτών βασίζεται στην μέθοδο επικύρωσης (validation), όπου πρώτα χρησιμοποιούμε μια μέθοδο-ανίχνευση σε δεδομένα ισχύος τα οποία γνωρίζουμε από πριν και έχουμε στην διάθεση μας όλη την φόρμα υπολογισμού και στην συνέχεια με βάση την ίδια μέθοδο που μας έδωσε ένα έγκυρο αποτέλεσμα θα προσπαθήσουμε με τα λειψά δεδομένα να εξάγουμε τα ίδια συμπεράσματα.

Οι μαθηματικές/ υπολογιστικές μέθοδοι που θα χρησιμοποιήσουμε είναι:

1. Γραμμική παλινδρόμηση (linear regression)

2. Μη γραμμική παλινδρόμηση και συγκεκριμένα πολυωνυμικού τύπου βαθμού 2. (polynomial regression)

3. Γενετικός αλγόριθμος υπολογισμού του διανύσματος από το οποίο θα προκύψουν οι «εύστοχες» ερωτήσεις για τον κάθε δείκτη.

4. Διαδικασία εύρεσης συνάρτησης νευρωνικού δικτύου με χρήση υπερπαραμέτρων για αποφυγή άστοχων συναρτήσεων που οδηγούν σε μεγάλες αποκλίσεις.

Για την ευστοχία των αποτελεσμάτων χρησιμοποιούμε τρείς συναρτήσεις υπολογισμού σφάλματος.

1. Υπολογισμός της ακρίβειας του μοντέλου (συντελεστής προσδιορισμού R-τετράγωνο)

$$r^2 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

2. Υπολογισμός του μέσου απόλυτου σφάλματος (MAE)

$$MAE = \frac{1}{n}\sum|y - \bar{y}|$$

3. Υπολογισμός του μέσου τετραγωνικού σφάλματος (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2$$

1.7 Γλώσσα Προγραμματισμού και τεχνολογίες

Στην παρούσα μεταπτυχιακή διατριβή η γλώσσα προγραμματισμού που χρησιμοποιούμε είναι η Python έκδοσης 3.8 ή νεότερη. Οι βιβλιοθήκες που χρησιμοποιούμε είναι οι εξής:

re # Βιβλιοθήκη για την επεξεργασία κειμένου μέσω τακτικών εκφράσεων
sklearn.metrics # Παρέχει μετρικές για την αξιολόγηση των μοντέλων μηχανικής μάθησης
sklearn.preprocessing # Εργαλεία για την προεπεξεργασία δεδομένων
tqdm # Βιβλιοθήκη για την παραγωγή γραμμών προόδου (progress bars)
math # Παρέχει πρόσβαση σε βασικές μαθηματικές συναρτήσεις
time # Βιβλιοθήκη για τη διαχείριση χρόνου στον κώδικα
datetime # Εργαλεία για τη διαχείριση ημερομηνιών και χρόνου
numpy # Βασική βιβλιοθήκη για επιστημονικούς υπολογισμούς σε Python
numpy.random # Μονάδα για την παραγωγή τυχαίων αριθμών
sklearn # Βιβλιοθήκη μηχανικής μάθησης, παρέχει εργαλεία για την ανάπτυξη προβλεπτικών μοντέλων
pandas # Βιβλιοθήκη για τη διαχείριση και ανάλυση δεδομένων
sklearn.linear_model # Παρέχει αλγόριθμους για γραμμικά μοντέλα
matplotlib.pyplot # Βιβλιοθήκη για τη δημιουργία γραφημάτων
matplotlib.patches # Παρέχει λειτουργίες για τη σχεδίαση γεωμετρικών σχημάτων
sklearn.metrics # Μετρικές για την αξιολόγηση μοντέλων
sklearn.model_selection # Εργαλεία για την επιλογή μοντέλου και διαχωρισμό δεδομένων
pygad # Βιβλιοθήκη για την εφαρμογή γενετικών αλγορίθμων
threading # Επιτρέπει την ταυτόχρονη εκτέλεση κώδικα
concurrent.futures # Εργαλείο για την ταυτόχρονη εκτέλεση κώδικα
os # Παρέχει λειτουργίες διαχείρισης του λειτουργικού συστήματος
tensorflow # Βιβλιοθήκη για νευρωνικά δίκτυα και βαθιά μάθηση
tensorflow.keras.models # Εργαλεία για τη δημιουργία και εκπαίδευση νευρωνικών δικτύων
tensorflow.keras.layers # Εργαλεία για τη δόμηση επιπέδων σε νευρωνικά δίκτυα
logging # Βιβλιοθήκη για την καταγραφή αρχείων καταγραφής
IPython.display # Εργαλεία για την εμφάνιση εξόδου μέσα στο IPython Notebook

Το εργαλείο συγγραφής (IDE) είναι το Jupyter Notebook το οποίο χρησιμεύει στην οπτικοποίηση αποτελεσμάτων, άμεσα μετά από κάθε υπολογισμό.

2. Βιβλιογραφική ανασκόπηση (παρόμοια έγγραφα και περίληψη θεωρίας)

2.1 Ιστορικά επιστημονικά κείμενα υπολογισμού δεικτών OCEAN

Το πρώτο τεστ προσωπικότητας που διεξήχθη ποτέ είναι ευρέως γνωστό ως Woodworth Psychoneurotic Inventory και αναπτύχθηκε κατά τη διάρκεια του Πρώτου Παγκοσμίου Πολέμου. Ο έλεγχος γινόταν για τη διαταραχή μετατραυματικού στρες (PTSD). Ο Αμερικάνικος στρατός το χρησιμοποίησε για την στράτευση ανδρών. Η διαδικασία αυτή ήταν ευρέως γνωστή. Το μοντέλο αυτό ονομάστηκε (PCM) Process communication model και δημιουργήθηκε από τους Taibi Kahler το οποίο και χρησιμοποιήθηκε για αστροναύτες.

Οι βασικές μέθοδοι ανίχνευσης της προσωπικότητας είναι το κείμενο, ο ήχος και τα οπτικά στοιχεία. Όταν χρησιμοποιούνται δεδομένα κειμένου, η προεπεξεργασία των δεδομένων είναι ένα κρίσιμο βήμα που επηρεάζει άμεσα τα αποτελέσματα. Γενικά, τα χαρακτηριστικά κειμένου εξάγονται από τα ακατέργαστα δεδομένα και εισάγονται ως είσοδος σε μοντέλα μηχανικής μάθησης όπως οι μηχανές διανυσμάτων υποστήριξης (SVM), ο ταξινομητής Naïve Bayes κ.λπ. Οι ενσωματώσεις λέξεων αναπαρίστανται ως διανύσματα (Word2Vec, GloVe, κ.λπ.) και αργότερα εφαρμόζονται περισσότερες τεχνικές για την απόδοση αποτελεσμάτων.

Ο Kasula Chaithanya Pramodh και οι συνεργάτες του, το 2016, χρησιμοποίησαν το σύνολο δεδομένων stream-of-consciousness και MyPersonality. Το σύνολο δεδομένων MyPersonality αποτελείται από 250 χρήστες που δημοσίευσαν περίπου 10.000 ενημερώσεις κατάστασης στο Facebook. Για το μοντέλο τους χρησιμοποίησαν το εργαλείο Natural Language Toolkit και κατέγραψαν σκορ F1 0.665, 0.632, 0.625, 0.624 και 0.637 για τα χαρακτηριστικά OPN, CON, EXT, AGR και NEU αντίστοιχα.

Η ερευνητική ομάδα με επικεφαλής τον Carles Venture εφάρμοσε Συγκεχυμένα Νευρωνικά Δίκτυα, ενσωματώνοντάς τα με συστήματα αναγνώρισης μονάδων δράσης και τεχνολογίες αναγνώρισης προσώπου, για να αναλύσει και να κατηγοριοποιήσει τις προσωπικότητες σύμφωνα με το πλαίσιο των Big Five. Η ανάλυσή τους βασίστηκε στο σύνολο δεδομένων "First Impressions", το οποίο περιλαμβάνει περίπου 10.000 βίντεο κλιπ με διάφορα άτομα.

Το 2018, ο Gokul K και η ομάδα του χρησιμοποίησαν έναν ταξινομητή Bayes-Net για τη διάκριση μεταξύ εξωστρεφών και εσωστρεφών προσωπικοτήτων. Χρησιμοποίησαν ένα μοναδικό σύνολο δεδομένων που αποτελούνταν από ηχογραφήσεις που έκαναν οι ίδιοι οι συμμετέχοντες. Η προσέγγισή τους περιλάμβανε τη μοντελοποίηση ακουστικών νεύρων, η οποία συνδυάζει μεθόδους ανίχνευσης φωνητικής δραστηριότητας και επιλογής χαρακτηριστικών για τη μείωση της πολυπλοκότητας των δεδομένων. Αυτή η καινοτόμος μεθοδολογία τους επέτρεψε να επιτύχουν ένα αξιοσημείωτο ποσοστό ακρίβειας 88,3% στις προσπάθειες ταξινόμησης της προσωπικότητάς τους.

Ο Bojan Simoski και οι συνεργάτες του δημιούργησαν μια καινοτόμο προσέγγιση γνωστή ως Μοντέλο Κοινωνικής Μεταδοτικότητας για την κατηγοριοποίηση των προσωπικοτήτων με βάση το πλαίσιο των Big Five. Η έρευνά τους χρησιμοποίησε τις απαντήσεις από μια ομάδα 25 ατόμων που συμπλήρωσαν το τεστ προσωπικότητας Big Five, το οποίο χρησίμευσε ως θεμελιώδες δεδομένο για την ανάλυσή τους.

Το 2018, ο Abir Abyaa και η ερευνητική τους ομάδα ασχολήθηκαν με το σύνολο δεδομένων StudentLife, το οποίο περιλαμβάνει δεδομένα από 48 φοιτητές, για να διερευνήσουν την ταξινόμηση της προσωπικότητας σύμφωνα με το μοντέλο Big Five. Χρησιμοποίησαν μια ποικιλία αλγορίθμων επιβλεπόμενης μάθησης, συμπεριλαμβανομένων των Support Vector Machines, Random Forests, Logistic Regression, C4.5 Decision Tree και του αλγορίθμου k-nearest neighbors, για να αναλύσουν και να ταξινομήσουν τα χαρακτηριστικά της προσωπικότητας των συμμετεχόντων με βάση τα δεδομένα που συλλέχθηκαν.

Το 2019, ο Willy και η ομάδα του υιοθέτησαν τον αλγόριθμο C4.5 Decision Tree για να αναλύσουν και να κατηγοριοποιήσουν τις προσωπικότητες σύμφωνα με το μοντέλο Big Five. Η έρευνά τους χρησιμοποίησε ένα σημαντικό σύνολο δεδομένων που συγκεντρώθηκε μέσω του Twitter API, το οποίο περιλάμβανε περίπου 110 εκατομμύρια καθημερινά tweets. Αυτό το εκτεταμένο σύνολο δεδομένων τους επέτρεψε να επιτύχουν ακρίβεια ταξινόμησης 64,30% στη μελέτη τους.

Το 2019 επίσης, η ερευνητική ομάδα με επικεφαλής τον Tao Hong ανέλαβε μια μελέτη με στόχο τον εντοπισμό συναισθημάτων και προσωπικοτήτων. Χρησιμοποίησαν το σύνολο δεδομένων MDSTC,

το οποίο περιλαμβάνει δεδομένα από την ομιλία, τις εκφράσεις του προσώπου και τη γαλβανική απόκριση του δέρματος, και εφάρμοσαν ένα βαθύ νευρωνικό δίκτυο για την κατηγοριοποίηση των χαρακτηριστικών της προσωπικότητας. Τα ευρήματά τους έδειξαν ενισχυμένη απόδοση σε σύγκριση με τα υπάρχοντα κορυφαία μοντέλα σε αυτόν τον τομέα.

Το Drexel χρησιμοποίησε τον ταξινομητή Gaussian Naïve Bayes για να αναλύσει τα χαρακτηριστικά της προσωπικότητας με βάση το μοντέλο Big Five. Ενσωμάτωσαν τα Word2Vec και FastText word embeddings για να ενισχύσουν την απόδοση του μοντέλου τους και χρησιμοποίησαν μια τεχνική 5πλής διασταυρούμενης επικύρωσης για να διασφαλίσουν την ευρωστία και την αξιοπιστία των ευρημάτων τους.

Το 2020, ο Songcheng Gao και οι συνεργάτες του υιοθέτησαν μια προσέγγιση μάθησης πολλαπλών προβολών και πολλαπλών εργασιών χρησιμοποιώντας το σύνολο δεδομένων StudentLife, το οποίο συγκεντρώθηκε μέσω μιας προσαρμοσμένης εφαρμογής για κινητά. Αυτό το σύνολο δεδομένων περιλάμβανε απαντήσεις από 183 άτομα σε δύο πανεπιστήμια, οι οποίες συλλέχθηκαν μέσω ενός ερωτηματολογίου εντός της εφαρμογής. Αξιολόγησαν την ακρίβεια του μοντέλου τους χρησιμοποιώντας μετρήσεις μέσου απόλυτου σφάλματος και μέσου τετραγωνικού σφάλματος.

Η Marwa S. Salem και η ομάδα της εφάρμοσαν μια σειρά αλγορίθμων, συμπεριλαμβανομένων των Multinomial Naïve Bayes, K-Nearest Neighbor, Support Vector Machine και Decision Trees στο Egyptian Twitter Users Dataset για την ταξινόμηση της προσωπικότητας. Τα ευρήματά τους έδειξαν ότι ο K-Nearest Neighbor υπερείχε των άλλων αλγορίθμων για τα περισσότερα χαρακτηριστικά προσωπικότητας, εκτός από την Ευσυνειδησία, όπου το Δέντρο Αποφάσεων ήταν πιο αποτελεσματικό.

Ο Kamal El-Demerdash και οι συνεργάτες του χρησιμοποίησαν το Universal Language Model Fine-Tuning στο σύνολο δεδομένων stream-of-consciousness για την ανίχνευση χαρακτηριστικών προσωπικότητας με βάση το μοντέλο Big Five, επιτυγχάνοντας ακρίβεια ελαφρώς ανώτερη από τα υπάρχοντα συγκριτικά στοιχεία.

Σε μια μελέτη του 2017, η ομάδα του Navonil Majumder χρησιμοποίησε ένα βαθύ συνελικτικό νευρωνικό δίκτυο για την ταξινόμηση των χαρακτηριστικών προσωπικότητας σύμφωνα με το μοντέλο Big Five, χρησιμοποιώντας τα word2vec embeddings της Google και τα χαρακτηριστικά Mairesse. Η εργασία αυτή βελτιώθηκε περαιτέρω από τον Md. Abdur Rahman και την ομάδα του το 2019, οι οποίοι δοκίμασαν διαφορετικές συναρτήσεις ενεργοποίησης για να βελτιώσουν την απόδοση του μοντέλου, όλα κωδικοποιημένα σε Python 2.7.

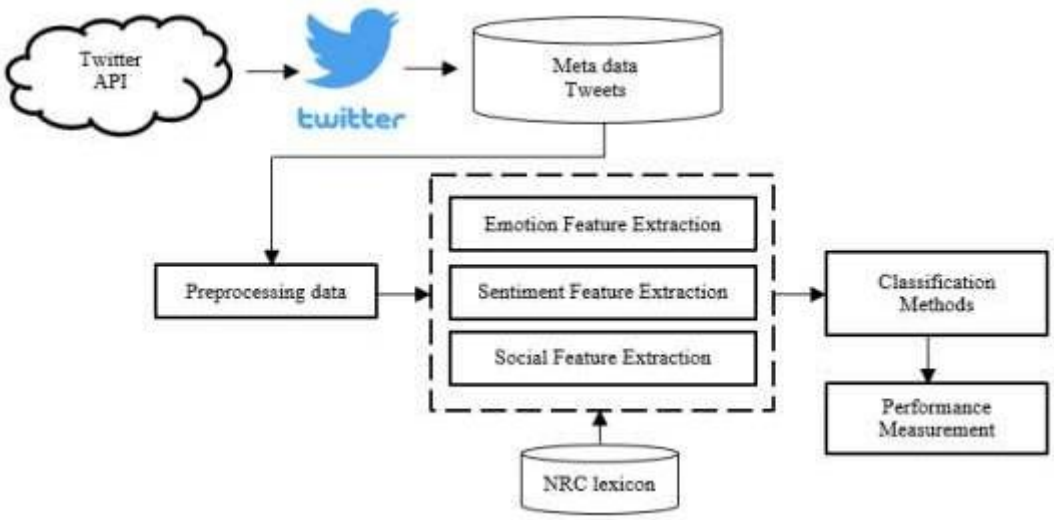
Η ίδια μελέτη του 2017 επαναλήφθηκε σε Python 3.8 από το Πανεπιστήμιο στο Delhi το Σεπ. του 2021 από την Shristi Rauniyar επειδή δεν υποστηριζόταν τότε η Python 2.7. Η ακρίβεια των αποτελεσμάτων της έχει μια συνάφεια με τα δικά μας αποτελέσματα ως προς την ακρίβεια και σχετικότητα των δεικτών, παρόλο που αυτοί πήραν κατευθείαν τους 5 δείκτες και από αυτούς προσπάθησαν με τους 4 να προβλέψουν τον 5ο.

2.2 Επιστημονικά κείμενα υπολογισμού σχετικά με την ίδια ανάλυση.

Υπάρχουν σχετικά λίγες έρευνες σε ότι αφορά τον υπολογισμό των δεικτών αυτών καθαυτών. Οι περισσότερες έρευνες έχουν προσπαθήσει να συσχετίσουν τους δείκτες αυτούς είτε με γραφικούς χαρακτήρες, είτε με πληροφορίες από τα social media, Facebook, Instagram, Twitter κλπ.

Ένα αξιοσημείωτο παράδειγμα είναι η έρευνα του 2021 από τους Warih Maharani και Veronikha Effendy με τίτλο **Big five personality prediction based in Indonesian tweets using machine learning methods** , στο οποίο χρησιμοποιώντας τα tweets 800 χρηστών μέσω τεχνολογίας API και έχοντας αυτοι οι χρήστες συμπληρώσει ένα ερωτηματολόγιο 44 ερωτήσεων , λαμβάνοντας επίσης υπόψιν τους αριθμούς των ακολούθων , τον αριθμό αναφορών, τα αγαπημένα , τα hashtags και άλλα προσωπικά στοιχεία, προσπάθησαν να κάνουν μια ομαδοποίηση και στην συνέχεια να προβλέψουν τους δείκτες της προσωπικότητας. Η έρευνα αυτή βασίστηκε σε Ινδονησιακούς χρήστες του Twitter. Αποδείξανε ότι η προσωπικότητα Big Five μπορεί να προβλεφθεί χρησιμοποιώντας δεδομένα δημόσιων πληροφοριών και των ινδονησιακών tweets που μοιράζονται στο Twitter. Λόγω της φύσης αυτής της μελέτης, η χρήση του Twitter έχει τα μοναδικά της προβλήματα. Η μελλοντική εφαρμογή των εφαρμογών αναγνώρισης της προσωπικότητας ανέφεραν ότι είναι ένα δύσκολο ζήτημα. Υπάρχουν πολυάριθμες ευκαιρίες για τη διαχείριση ταλέντων με τη δυνατότητα εντοπισμού των χαρακτηριστικών προσωπικότητας ενός χρήστη. Συνοπτικά το σχήμα που χρησιμοποιήσαν για να αναδείξουν την μελέτη τους. Χρησιμοποίησαν την λεξικό NRC.

Το λεξικό NRC, επίσης γνωστό ως λεξικό συναισθημάτων NRC ή λεξικό συσχετίσεων λέξης-συναισθήματος NRC, είναι ένας πόρος που αναπτύχθηκε από το Εθνικό Συμβούλιο Έρευνας του Καναδά. Απαριθμεί τις αγγλικές λέξεις και τους συσχετισμούς τους με οκτώ βασικά συναισθήματα (θυμός, φόβος, προσδοκία, εμπιστοσύνη, έκπληξη, θλίψη, χαρά και αηδία) και δύο συναισθήματα (αρνητικό και θετικό). Αυτό το λεξικό χρησιμοποιείται συνήθως στην ανάλυση κειμένου και την ανάλυση συναισθήματος για την κατανόηση του συναισθηματικού τόνου των κειμένων, βοηθώντας σε εργασίες όπως η εξόρυξη γνώμης, η ανάλυση των αναφορών των πελατών και η παρακολούθηση των μέσων κοινωνικής δικτύωσης.



ΕΙΚΟΝΑ 7: Μεθοδολογία αποτελεσμάτων της έρευνας των τους Warih Maharani και Veronikha Effendy

Μια ακόμα σχετική έρευνα που έγινε το 2020 από μια ομάδα ερευνητών στο Μόναχο, με τίτλο **Behavioral Patterns in Smartphone Usage Predict Big Five Personality Traits**. Χρησιμοποιώντας δεδομένα από 743 συμμετέχοντες, η μελέτη αναλύει 15.692 μεταβλητές συμπεριφοράς από αρχεία καταγραφής smartphone για 30 ημέρες. Η έρευνα χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για να συνδέσει αυτές τις συμπεριφορές με τα χαρακτηριστικά της προσωπικότητας, αποκαλύπτοντας συγκεκριμένα μοτίβα χρήσης που αντιστοιχούν σε διαφορετικές διαστάσεις της προσωπικότητας. Η προσέγγιση αυτή προάγει την κατανόηση της ψυχολογίας με την ενσωμάτωση της τεχνολογίας και της μηχανικής μάθησης για την ανάλυση της συμπεριφοράς και της προσωπικότητας.

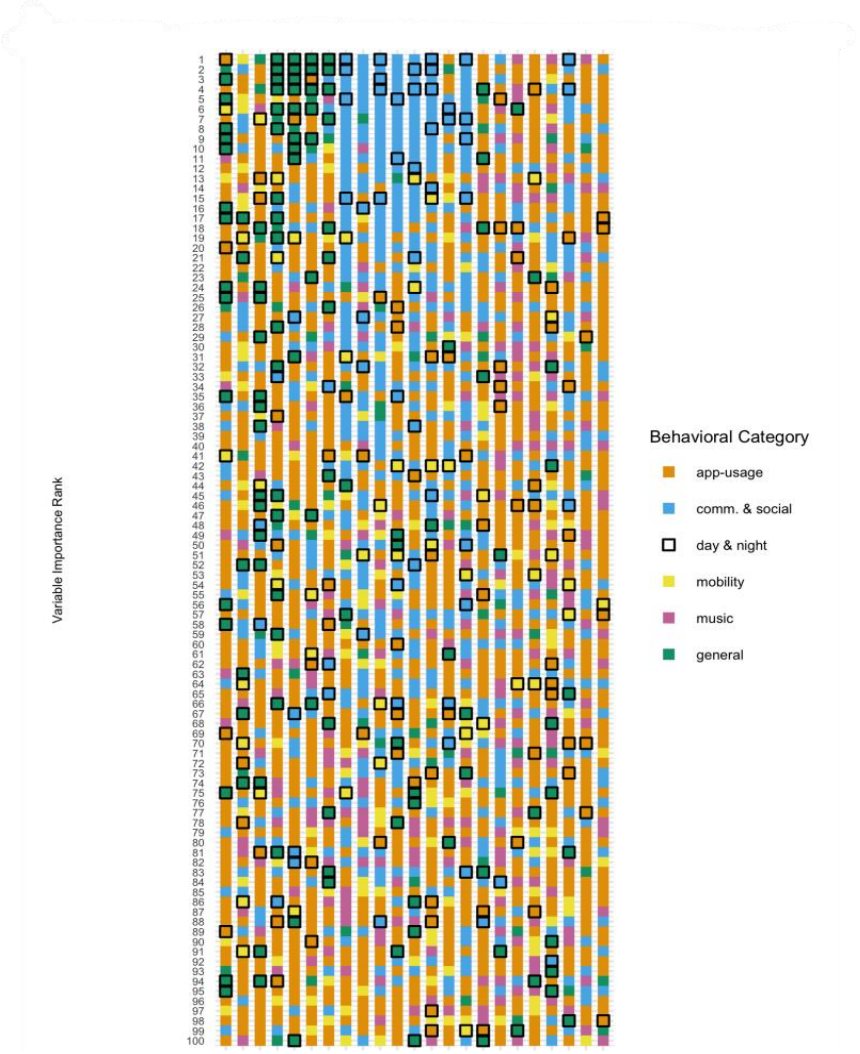
Η ενότητα "Ανάλυση δεδομένων" του εγγράφου εξηγεί πώς οι ερευνητές προ-επεξεργάστηκαν και ανέλυσαν τα δεδομένα χρήσης smartphone από τους συμμετέχοντες για να προβλέψουν τα χαρακτηριστικά προσωπικότητας Big Five. Χρησιμοποίησαν 1821 μεταβλητές που προέκυψαν από τα αρχεία καταγραφής smartphone και εφάρμοσαν τεχνικές μηχανικής μάθησης για να δημιουργήσουν συσχετίσεις με τα αυτοαναφερόμενα χαρακτηριστικά προσωπικότητας. Η ανάλυση περιελάμβανε την αφαίρεση των αντιγράφων, τον εμπλουτισμό των δεδομένων με εξωτερικές πηγές όπως το Google Play Store και το Spotify API και την εφαρμογή διαφόρων βημάτων καθαρισμού δεδομένων για να διασφαλιστεί η ποιότητα και η συνάφεια των δεδομένων που χρησιμοποιήθηκαν στα προγνωστικά τους μοντέλα.

Στην ενότητα "Εξαγωγή μεταβλητών", οι ερευνητές περιγράφουν λεπτομερώς πώς προέκυψαν 15.692 μεταβλητές από τα δεδομένα χρήσης smartphone, εστιάζοντας στην επικοινωνία, τη χρήση εφαρμογών, την κατανάλωση μουσικής, τα πρότυπα δραστηριότητας και την κινητικότητα. Χρησιμοποίησαν διάφορες στατιστικές μεθόδους για να καταγράψουν τις αποχρώσεις των δεδομένων, συμπεριλαμβανομένων των μετρικών της ακανόνιστης κατάστασης, της εντροπίας, της ομοιότητας και της χρονικής συσχέτισης, για να κατανοήσουν καλύτερα τις συμπεριφορές μέσα στα άτομα με την πάροδο του χρόνου. Αυτή η ολοκληρωμένη προσέγγιση είχε ως στόχο να συμπυκνώσει την πολυπλοκότητα των αλληλεπιδράσεων των smartphone και τις πιθανές συνδέσεις τους με τα χαρακτηριστικά της προσωπικότητας.

Οι ερευνητές χρησιμοποίησαν μηχανική μάθηση για να προβλέψουν τα χαρακτηριστικά της προσωπικότητας από δεδομένα smartphone, χρησιμοποιώντας μοντέλα ελαστικού δικτύου και τυχαίου δάσους για την προβλεπτική τους ικανότητα. Εφάρμοσαν μια τεχνική εμφωλευμένης επαναδειγματοληψίας για τη βελτιστοποίηση και την επικύρωση του μοντέλου, εξασφαλίζοντας μια στιβαρή αξιολόγηση με το διαχωρισμό των συνόλων δεδομένων εκπαίδευσης και δοκιμής. Οι μετρικές τους απόδοσης περιλάμβαναν τη συσχέτιση Pearson, το μέσο τετραγωνικό σφάλμα (RMSE) και τον συντελεστή προσδιορισμού (R^2), με στόχο την επικύρωση της αποτελεσματικότητας των μοντέλων στην ακριβή πρόβλεψη των Big Five χαρακτηριστικών προσωπικότητας με βάση τις εξαγόμενες μεταβλητές συμπεριφοράς.

Οι ερευνητές βελτίωσαν την ερμηνευσιμότητα του μοντέλου με τον υπολογισμό των τιμών σπουδαιότητας των μεταβλητών μετατροπής, υποδεικνύοντας τη σημασία κάθε μεταβλητής ή ομάδας μεταβλητών στην προβλεπτική ακρίβεια του μοντέλου. Χρησιμοποίησαν επίσης διαγράμματα συσσωρευμένης τοπικής επίδρασης (ALE) για να απεικονίσουν την επιρροή των μεμονωμένων προβλεπτικών παραγόντων στις προβλέψεις του μοντέλου. Η προσέγγιση αυτή αποσκοπούσε στη διαλεύκανση της εσωτερικής λειτουργίας των πολύπλοκων μοντέλων μηχανικής μάθησης, καθιστώντας την επίδραση συγκεκριμένων μεταβλητών στις προβλέψεις πιο διαφανή και κατανοητή.

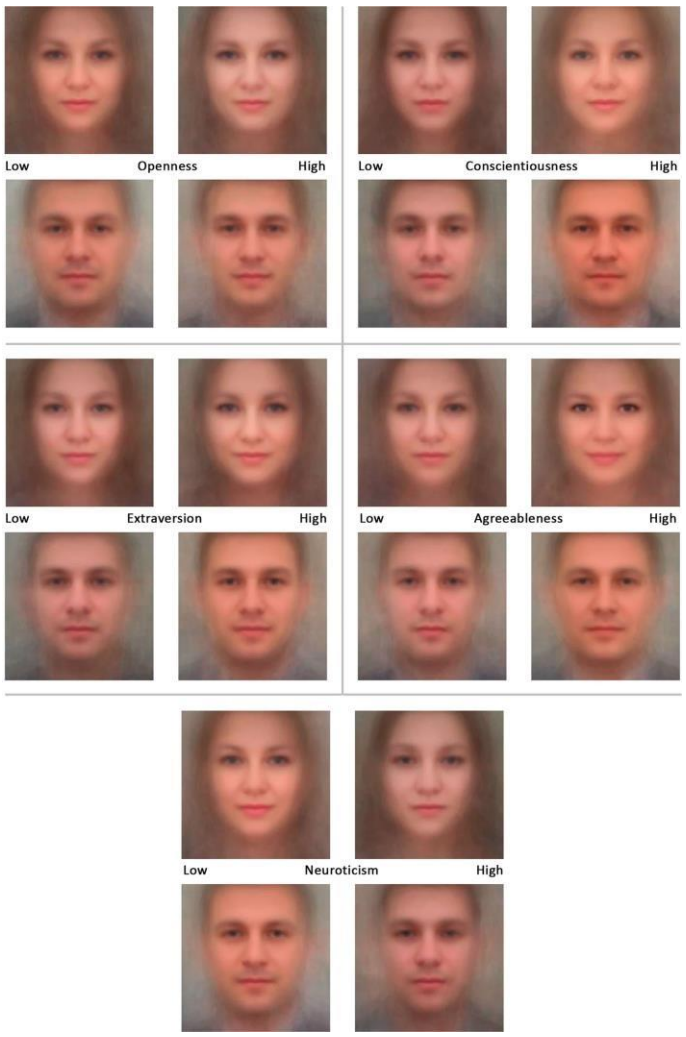
Τελικά διαπίστωσαν ότι η συμπεριφορά των smartphone μπορεί να προβλέψει τα χαρακτηριστικά της προσωπικότητας. Για παράδειγμα, η ανοιχτότητα συνδέθηκε με τις προτιμήσεις μουσικής και εφαρμογών, ενώ η ευσυνειδησία συσχετίστηκε με τις συνήθειες χρήσης του τηλεφώνου. Χρησιμοποίησαν μηχανική μάθηση για να αναλύσουν διάφορες συμπεριφορές, τονίζοντας ότι ενώ οι προβλέψεις είναι δυνατές, δεν φτάνουν την ακρίβεια των λεπτομερών τεστ προσωπικότητας. Η μελέτη αυτή υπογραμμίζει τις δυνατότητες των δεδομένων smartphone στην ψυχολογική έρευνα, τονίζοντας επίσης τις επιπτώσεις στην προστασία της ιδιωτικής ζωής, καθώς η χρήση του τηλεφώνου μπορεί να αποκαλύψει προσωπικά χαρακτηριστικά.



Εικόνα 8: Διάγραμμα συμπεριφορών προτύπων: οι προγνωστικοί παράγοντες με τη μεγαλύτερη σημασία μεταβλητής στο αντίστοιχο μοντέλο βρίσκονται στην κορυφή. Οι σειρές αντιπροσωπεύουν τις τάξεις των προβλεπτών, οι στήλες αντιπροσωπεύουν τα μεμονωμένα μοντέλα, τα τετράγωνα αντιπροσωπεύουν τις μεταβλητές, τα χρώματα αντιπροσωπεύουν τη συμμετοχή σε σημασιολογικές κατηγορίες.

2.3 Επιστημονικά κείμενα υπολογισμού σχετικά με την ίδια ανάλυση και ταυτόχρονα την χρήση εικόνας

Μια ακόμα σχετική έρευνα από το 2020 με τίτλο **Assessing the Big Five personality traits using real-life static facial images** από τους Alexander Kachur και Evgeny Osin. Η μελέτη χρησιμοποιεί μηχανική μάθηση, συγκεκριμένα τεχνητά νευρωνικά δίκτυα (ANN), για να αναλύσει στατικές εικόνες προσώπου και να προβλέψει τα Big Five χαρακτηριστικά της προσωπικότητας. Οι ερευνητές συγκέντρωσαν ένα μεγάλο σύνολο δεδομένων από εθελοντές που παρείχαν φωτογραφίες προσώπου και συμπλήρωσαν ένα τεστ προσωπικότητας Big Five. Η μελέτη είχε ως στόχο να διερευνήσει τη σχέση μεταξύ της μορφολογίας του προσώπου και των χαρακτηριστικών της προσωπικότητας, βασιζόμενη σε προηγούμενες έρευνες που έδειχναν ότι ορισμένα χαρακτηριστικά του προσώπου θα μπορούσαν να συσχετιστούν με τα χαρακτηριστικά της προσωπικότητας.

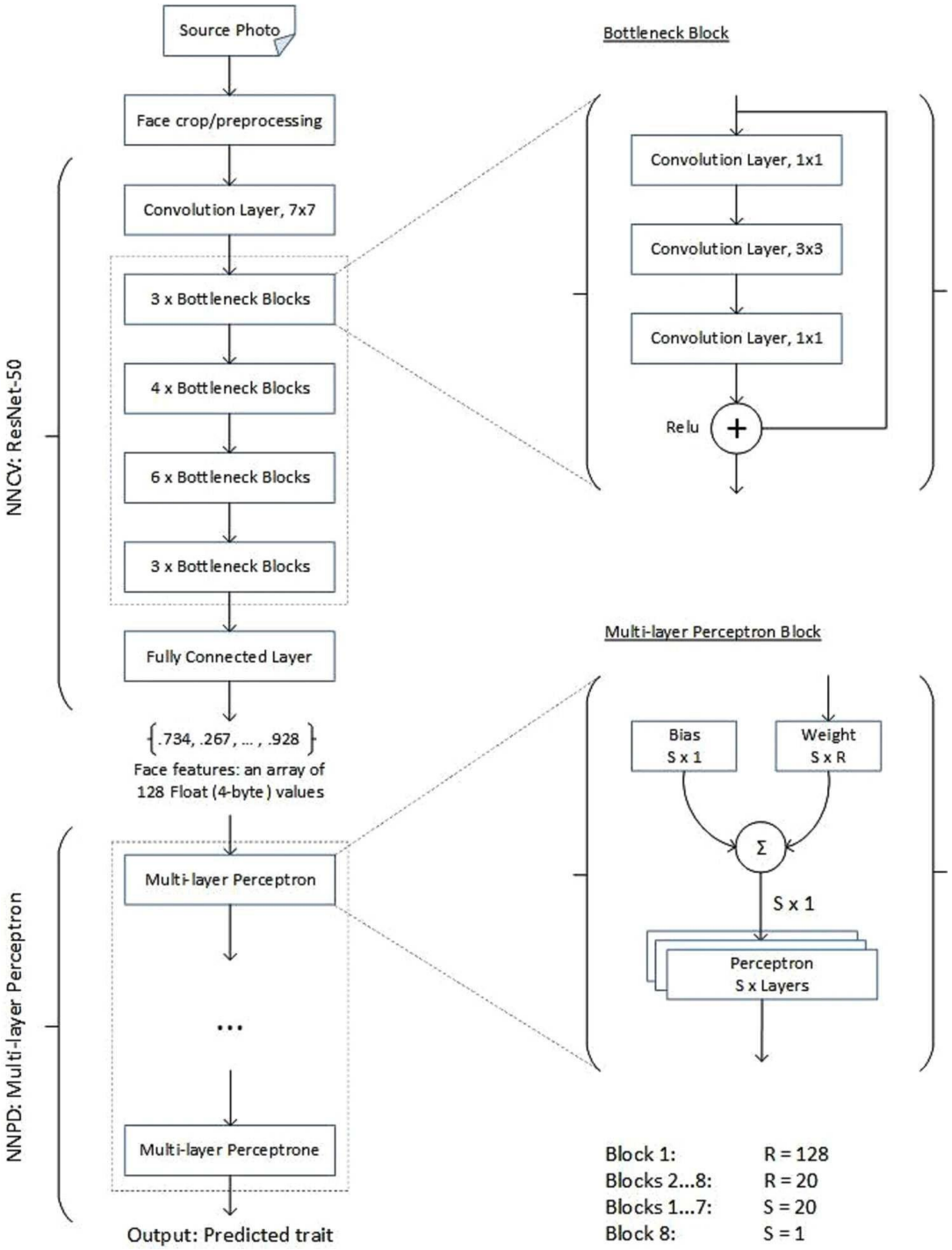


Εικόνα 9: Σύνθετες εικόνες προσώπου που διαμορφώθηκαν σε ομάδες αντίθεσης 100 ατόμων για κάθε χαρακτηριστικό των Big Five.

Οι ερευνητές ανέπτυξαν ένα νευρωνικό δίκτυο δύο επιπέδων: ένα για τον εντοπισμό αμετάβλητων χαρακτηριστικών του προσώπου (NNCV) και ένα άλλο για την πρόβλεψη χαρακτηριστικών της προσωπικότητας (NNPD) με βάση αυτά τα χαρακτηριστικά. Εκπαίδευσαν το NNCV χρησιμοποιώντας μια τεράστια συλλογή μη επισημειωμένων φωτογραφιών της πραγματικής ζωής για να δημιουργήσουν διανύσματα που αντιπροσωπεύουν μοναδικά χαρακτηριστικά του προσώπου. Στη συνέχεια, το NNPD χρησιμοποίησε αυτά τα διανύσματα για να προβλέψει τα Big Five χαρακτηριστικά της προσωπικότητας, παράγοντας αποτελέσματα που ξεπέρασαν προηγούμενες μελέτες σε ακρίβεια πρόβλεψης, ιδιαίτερα για την ευσυνειδησία.

Η εργασία αυτή καταδεικνύει ότι ακόμη και τα ανεπαίσθητα στοιχεία σε στατικές εικόνες προσώπου, που έχουν ληφθεί σε καθημερινές, ανεξέλεγκτες συνθήκες, μπορούν να αποκαλύψουν πτυχές της προσωπικότητας ενός ατόμου. Αξιοσημείωτες είναι οι δυνατότητες της τεχνητής νοημοσύνης και της μηχανικής μάθησης στην ψυχολογική αξιολόγηση και υποδηλώνουν ότι οι εικόνες προσώπου θα μπορούσαν να συμπληρώσουν τα παραδοσιακά μέτρα προσωπικότητας σε διάφορες εφαρμογές. Τα αποτελέσματα της μελέτης δείχνουν ότι τα τεχνητά νευρωνικά δίκτυα μπορούν να προβλέψουν τα πέντε μεγάλα χαρακτηριστικά της προσωπικότητας από στατικές εικόνες προσώπου με διαφορετικό βαθμό ακρίβειας. Οι ισχυρότερες προβλέψεις αφορούσαν την ευσυνειδησία και τα μοντέλα ήταν γενικά πιο ακριβή για τις γυναίκες από ό,τι για τους άνδρες. Οι ερευνητές διαπίστωσαν συνεπείς προβλέψεις χαρακτηριστικών προσωπικότητας σε διαφορετικές εικόνες του ίδιου ατόμου. Τα ευρήματα αυτά υποδηλώνουν ότι τα στοιχεία του προσώπου σε φωτογραφίες μπορούν να προσφέρουν σημαντικές πληροφορίες για τα χαρακτηριστικά της

προσωπικότητας ενός ατόμου, αναδεικνύοντας τις δυνατότητες της μηχανικής μάθησης στην αξιολόγηση της προσωπικότητας.



Εικόνα 10: Αρχιτεκτονική επιπέδων του νευρωνικού δικτύου όρασης υπολογιστών (NNCV) και του νευρωνικού δικτύου διάγνωσης προσωπικότητας (NNPD).

2.4 Επιστημονικά κείμενα σχετικά με γενικότερη θεωρία

Μια ακόμα σχετική έρευνα από τον Αύγουστο του 2022 με τίτλο **How the Big Five personality traits related to aggression from perspectives of the benign and malicious envy** από τους Xinsheng Jiang και Xiaojun Li. Η μελέτη βασίζεται σε προηγούμενες έρευνες που έχουν διαπιστώσει συνδέσεις μεταξύ

των Big Five χαρακτηριστικών προσωπικότητας και της επιθετικότητας. Οι ερευνητές στοχεύουν να εμβαθύνουν στην κατανόηση αυτών των σχέσεων εξετάζοντας τους συναισθηματικούς μηχανισμούς, εστιάζοντας ιδιαίτερα στον φθόνο.

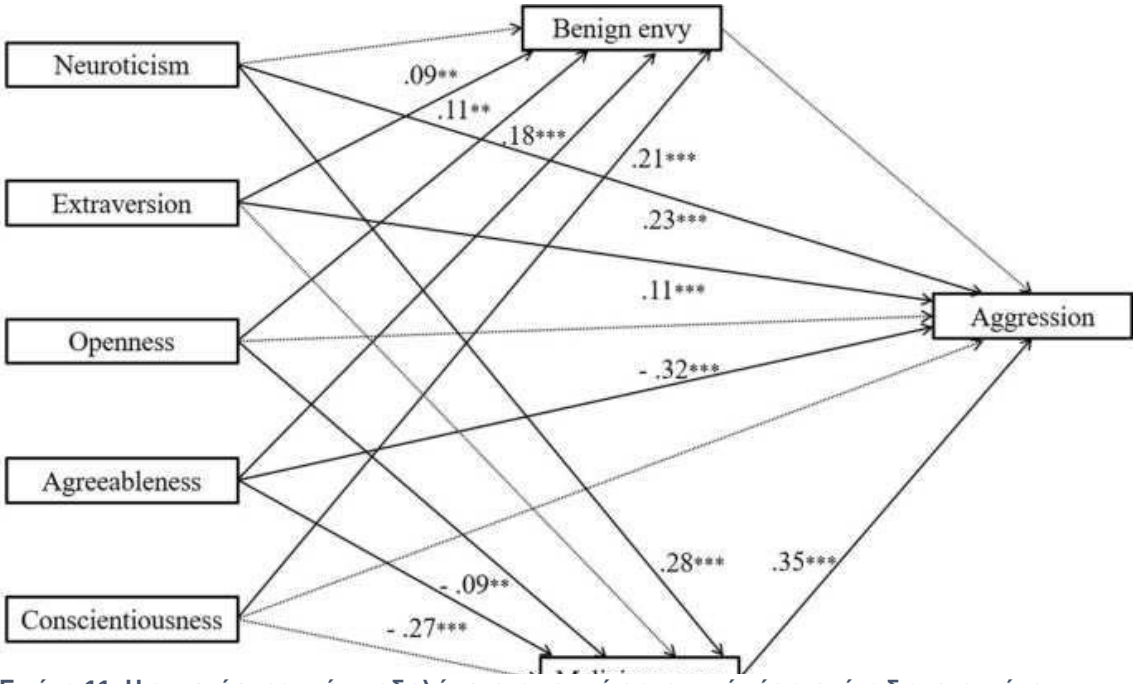
Ο ρόλος του φθόνου: Ο φθόνος αναγνωρίζεται ως ένας κρίσιμος συναισθηματικός παράγοντας που θα μπορούσε να επηρεάσει τη σχέση μεταξύ των χαρακτηριστικών της προσωπικότητας και της επιθετικότητας. Με τη διάκριση μεταξύ καλοήθους φθόνου (ο οποίος μπορεί να σχετίζεται με θετικές πτυχές κινήτρων) και κακόβουλου φθόνου (ο οποίος πιθανότατα συνδέεται με αρνητικά αποτελέσματα όπως η επιθετικότητα), η μελέτη επιδιώκει να προσφέρει αποχρώσεις.

Στόχος: Ο πρωταρχικός στόχος είναι να διερευνηθεί πώς ο καλοήθης και ο κακόβουλος φθόνος μεσολαβούν στη σχέση μεταξύ των Big Five χαρακτηριστικών προσωπικότητας και της επιθετικότητας. Αυτό θα μπορούσε να προσφέρει μια πιο λεπτομερή κατανόηση των συναισθηματικών οδών μέσω των οποίων τα χαρακτηριστικά της προσωπικότητας επηρεάζουν την επιθετική συμπεριφορά.

Μέθοδοι: Οι συμμετέχοντες αξιολογήθηκαν με τη χρήση τριών καθιερωμένων ψυχολογικών εργαλείων για τη μέτρηση των χαρακτηριστικών της προσωπικότητάς τους, του τύπου και του επιπέδου φθόνου που βιώνουν και των τάσεων τους προς την επιθετικότητα.

Δείγμα: Η μελέτη διαθέτει ένα σημαντικό δείγμα 839 συμμετεχόντων, προσφέροντας ένα ισχυρό σύνολο δεδομένων για ανάλυση. Τα δημογραφικά στοιχεία υποδηλώνουν μια διαφορετική ομάδα, κυρίως γυναικών, η οποία θα μπορούσε να προσφέρει μια ευρεία προοπτική για την αλληλεπίδραση μεταξύ της προσωπικότητας, του φθόνου και της επιθετικότητας.

Επιπτώσεις: Με την αποκάλυψη του ρόλου του φθόνου, η μελέτη θα μπορούσε να προσφέρει πολύτιμες πληροφορίες για ψυχολογικές παρεμβάσεις και στρατηγικές για τον μετριασμό της επιθετικότητας, ιδίως σε άτομα με συγκεκριμένα προφίλ προσωπικότητας.



Εικόνα 11: Η συνεχής γραμμή υποδηλώνει στατιστικά σημαντική τάση, ενώ η διακεκομμένη γραμμή υποδηλώνει ότι η τάση δεν είναι σημαντική. Σημείωση N = 839. **p < .01, **p < .001

Στη μελέτη, οι ερευνητές διερεύνησαν τις διαφορές μεταξύ των δύο φύλων στη σχέση μεταξύ των Big Five χαρακτηριστικών προσωπικότητας, του καλοήθους/κακόβουλου φθόνου και της επιθετικότητας. Βρήκαν σημαντικές διαφορές μεταξύ των δύο φύλων στον Νευρωτισμό, τη συμφωνητικότητα, την επιθετικότητα και τον καλοήθη φθόνο, αλλά όχι στην εξωστρέφεια, το άνοιγμα στην εμπειρία, την ευσυνειδησία ή τον κακόβουλο φθόνο.

Η ανάλυση διαδρομής μεταξύ των δύο φύλων αποκάλυψε σημαντικές διαφορές στον τρόπο με τον οποίο τα χαρακτηριστικά της προσωπικότητας και ο φθόνος συνέβαλαν στην επιθετικότητα μεταξύ ανδρών και γυναικών. Παρά τις διαφορές αυτές, ο κακόβουλος φθόνος διαδραμάτισε σταθερό διαμεσολαβητικό ρόλο στη σχέση μεταξύ εξωστρέφειας, ευχάριστης συμπεριφοράς και επιθετικότητας και στα δύο φύλα. Συγκεκριμένα, στους άνδρες, η εξωστρέφεια συσχετίστηκε αρνητικά με την

επιθετικότητα με τη διαμεσολάβηση του κακόβουλου φθόνου, όπως και η ευχάριστη διάθεση. Στις γυναίκες, παρατηρήθηκε παρόμοιο μοτίβο για τη συμφωνητικότητα, αλλά η σχέση μεταξύ εξωστρέφειας και επιθετικότητας διαμεσολαβήθηκε θετικά από τον κακόβουλο φθόνο.

Τα ευρήματα αυτά υποδηλώνουν ότι ενώ τα Big Five χαρακτηριστικά προσωπικότητας επηρεάζουν την επιθετικότητα και διαμεσολαβούνται από τον κακόβουλο φθόνο, τα πρότυπα αυτών των σχέσεων μπορεί να διαφέρουν ανάλογα με το φύλο. Τα αποτελέσματα της μελέτης υπογραμμίζουν την πολυπλοκότητα των αλληλεπιδράσεων μεταξύ της προσωπικότητας, του συναισθήματος και της συμπεριφοράς και υπογραμμίζουν τη σημασία της συνεκτίμησης του φύλου ως παράγοντα στην ψυχολογική έρευνα.

Τέτοιες γνώσεις θα μπορούσαν να είναι πολύτιμες για την ανάπτυξη πιο στοχευμένων παρεμβάσεων ή θεραπευτικών προσεγγίσεων για την αντιμετώπιση της επιθετικότητας, λαμβάνοντας υπόψη τα χαρακτηριστικά της προσωπικότητας ενός ατόμου και τον ρόλο των συναισθημάτων όπως ο φθόνος.

Η παρούσα μελέτη χρησιμοποιεί το Γενικό Μοντέλο Επιθετικότητας (General Aggression Model - GAM) για να διερευνήσει τη σχέση μεταξύ των Big Five χαρακτηριστικών προσωπικότητας και της επιθετικότητας, με ιδιαίτερη έμφαση στο ρόλο του φθόνου, ο οποίος διαφοροποιείται σε καλοήθη και κακόβουλο φθόνο. Τα ευρήματα υπογραμμίζουν τη σημασία του κακόβουλου φθόνου ως διαμεσολαβητικού παράγοντα στη σχέση μεταξύ ορισμένων χαρακτηριστικών της προσωπικότητας (νευρωτισμός και ευχάριστος χαρακτήρας) και της επιθετικότητας, τονίζοντας ότι αυτή η διαμεσολαβητική επίδραση είναι συνεπής σε όλα τα φύλα.
Ουσιαστικά, η έρευνα συμβάλλει στην κατανόηση των ψυχολογικών υποβάθρων της επιθετικότητας με τα εξής:
1. Διαπιστώνοντας μια σύνδεση μεταξύ βασικών χαρακτηριστικών της προσωπικότητας και της επιθετικότητας, ενισχύοντας την κατανόηση του τρόπου με τον οποίο οι ατομικές διαφορές επηρεάζουν την επιθετική συμπεριφορά.
2. Εντοπίζοντας τον κακόβουλο φθόνο ως κρίσιμο ενδιάμεσο παράγοντα, υποδηλώνοντας ότι ο τρόπος με τον οποίο τα άτομα επεξεργάζονται τα συναισθήματα φθόνου μπορεί να επηρεάσει το κατά πόσον είναι επιρρεπή στην επιθετικότητα.
3. Αποδεικνύοντας τη συνέπεια των σχέσεων αυτών ως προς το φύλο, γεγονός που συνεπάγεται την ανθεκτικότητα των ευρημάτων σε διαφορετικούς πληθυσμούς.

Η μελέτη αυτή είναι καίριας σημασίας, καθώς επεκτείνει το GAM, προσφέροντας νέες γνώσεις σχετικά με τους συναισθηματικούς μηχανισμούς που μπορούν να οδηγήσουν σε επιθετική συμπεριφορά με βάση τα χαρακτηριστικά της προσωπικότητας. Η κατανόηση αυτών των συνδέσεων είναι ζωτικής σημασίας για την ανάπτυξη στοχευμένων παρεμβάσεων ή ψυχολογικών στρατηγικών για τον μετριασμό της επιθετικότητας, οδηγώντας ενδεχομένως σε πιο αποτελεσματικές προσεγγίσεις για τη διαχείριση ή την πρόληψη της επιθετικής συμπεριφοράς σε διάφορα περιβάλλοντα.

Τα ακρωνύμια όπως SRMR, RMSEA, GFI, CFI και ECVI είναι στατιστικά μέτρα που χρησιμοποιούνται για την αξιολόγηση της προσαρμογής του μοντέλου που χρησιμοποιείται στην έρευνα, εξασφαλίζοντας την αξιοπιστία και την εγκυρότητα των ευρημάτων που παρουσιάζονται.

3. Περιγραφή συνόλου δεδομένων

3.1 Προέλευση δεδομένων

Τα δεδομένα αυτά συλλέχθηκαν από το 2016 έως το 2018 μέσω ενός διαδραστικού διαδικτυακού τεστ προσωπικότητας. Το τεστ κατασκευάστηκε με τους "Big-Five Factor Markers" από το IPIP. Οι συμμετέχοντες ενημερώθηκαν ότι οι απαντήσεις τους θα καταγραφούν και θα χρησιμοποιηθούν για την έρευνα στην αρχή του τεστ και τους ζητήθηκε να επιβεβαιώσουν τη συγκατάθεσή τους στο τέλος του τεστ.

Το τεστ προσωπικότητας παρουσίαζε τα ακόλουθα στοιχεία σε μία σελίδα, καθένα από τα οποία βαθμολογούνταν σε πενταβάθμια κλίμακα με τη χρήση κουμπιών επιλογής. Η σειρά στη σελίδα ήταν EXT1, AGR1, CSN1, NRT1, OPN1, EXT2 κ.λπ. Η κλίμακα είχε την ένδειξη 1=Διαφωνώ, 3=Ουδέτερη, 5=Συμφωνώ.

3.2 Τι είναι το IPIP

Το IPIP σημαίνει International Personality Item Pool. Πρόκειται για μια δημόσια συλλογή στοιχείων που χρησιμοποιούνται για ψυχολογικές μετρήσεις, ειδικά σχεδιασμένα για την αξιολόγηση διαφόρων πτυχών της προσωπικότητας. Τα στοιχεία αυτά χρησιμοποιούνται συχνά σε ερωτηματολόγια για την αξιολόγηση των χαρακτηριστικών της προσωπικότητας με βάση το μοντέλο των πέντε παραγόντων, γνωστό και ως Big Five, τα οποία περιλαμβάνουν την ανοιχτότητα, την ευσυνειδησία, την εξωστρέφεια, τη συμφωνητικότητα και τον νευρωτισμό.

Το IPIP αποτελεί πλούσιο πόρο για τους ερευνητές και τους επαγγελματίες της ψυχολογίας και των συναφών τομέων, επειδή προσφέρει έναν δωρεάν και πρόσβασιμο τρόπο μέτρησης των χαρακτηριστικών της προσωπικότητας. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για τη δημιουργία αξιολογήσεων προσωπικότητας, τη διεξαγωγή ερευνών και τη διερεύνηση ατομικών διαφορών στην προσωπικότητα.

Τα στοιχεία του IPIP είναι δομημένα με τρόπο ώστε να μπορούν να χρησιμοποιηθούν για τη δημιουργία κλιμάκων για διάφορα χαρακτηριστικά της προσωπικότητας. Κάθε στοιχείο είναι μια δήλωση την οποία οι ερωτώμενοι βαθμολογούν με βάση το πόσο καλά ισχύει γι' αυτούς, συνήθως σε μια κλίμακα Likert. Οι απαντήσεις αναλύονται στη συνέχεια για να παρέχουν ένα προφίλ των χαρακτηριστικών προσωπικότητας ενός ατόμου.

Παρέχοντας ένα τυποποιημένο σύνολο στοιχείων που είναι ανοικτά διαθέσιμα, το IPIP διευκολύνει την έρευνα στην ψυχολογία της προσωπικότητας, επιτρέποντας συγκρίσεις μεταξύ διαφορετικών μελετών και συμβάλλοντας στη γενική κατανόηση της δομής της προσωπικότητας.

3.3 Πληροφορίες δεδομένων

Τα δεδομένα αυτά που συλλέχθηκαν, περιλαμβάνουν επιπλέον πληροφορίες οι οποίες στην εν λόγω εργασία δεν χρησιμοποιήθηκαν.
Πρόσθετες μεταβλητές:

Ο χρόνος που δαπανάται για κάθε ερώτηση καταγράφεται επίσης σε χιλιοστά του δευτερολέπτου, και υποδεικνύεται από μεταβλητές που τελειώνουν σε "_E". Οι ακόλουθες πρόσθετες μεταβλητές περιλαμβάνονται στο σύνολο δεδομένων:

1. dateload: Η χρονοσφραγίδα κατά την έναρξη της έρευνας.
2. screenw: Το πλάτος της οθόνης του χρήστη σε pixels.
3. screenh: Το ύψος της οθόνης του χρήστη σε pixels.
4. introelapse: Ο χρόνος σε δευτερόλεπτα που δαπανάται στην αρχική/εισαγωγική σελίδα.
5. testelapse: Ο χρόνος σε δευτερόλεπτα που δαπανάται στη σελίδα με τις ερωτήσεις της έρευνας.

6. endelapse: Ο χρόνος σε δευτερόλεπτα που δαπανήθηκε στη σελίδα ολοκλήρωσης, όπου ο χρήστης κλήθηκε να δηλώσει αν είχε απαντήσει με ακρίβεια και αν οι απαντήσεις του μπορούσαν να αποθηκευτούν και να χρησιμοποιηθούν για έρευνα.
7. IPC: Ο αριθμός των εγγραφών από τη διεύθυνση IP του χρήστη στο σύνολο δεδομένων. Για μέγιστη καθαρότητα, χρησιμοποιήστε μόνο εγγραφές όπου αυτή η τιμή είναι 1.

χώρα: Η χώρα, η οποία καθορίζεται από τεχνικές πληροφορίες (ΔΕΝ ΕΡΩΤΕΙΤΑΙ ΩΣ ΕΡΩΤΗΣΗ).

8. lat_appx_lots_of_err: Κατά προσέγγιση γεωγραφικό πλάτος του χρήστη (όχι πολύ ακριβές).
9. long_appx_lots_of_err: Κατά προσέγγιση γεωγραφικό μήκος του χρήστη.

3.4 Προεπεξεργασία δεδομένων

Η αρχική μορφή των δεδομένων ήταν ένα αρχείο ονόματι **data-final.csv (416.27 MB).**

EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9
4	1	5	2	5	1	5	2	4
3	5	3	4	3	3	2	5	1
2	3	4	4	3	2	1	3	2
2	2	2	3	4	2	2	4	1
3	3	3	3	5	3	3	5	3
3	3	4	2	4	2	2	3	3
4	3	4	3	3	3	5	3	4
3	1	5	2	5	2	5	2	3
2	2	3	3	4	2	2	2	4
1	5	3	5	2	3	2	4	5
3	3	2	3	3	2	4	3	3
3	1	5	3	5	1	5	5	5
4	1	5	4	5	1	4	1	5
1	5	1	5	1	5	1	5	1
1	5	2	5	1	4	1	2	2

Εικόνα 12: Πίνακας σε μορφή αρχείο excel που δείχνει την αρχική μορφή των δεδομένων. Αποτελεί ένα μέρος των δεδομένων καθώς οι στήλες είναι παραπάνω από 50.

Τα δεδομένα αυτά περιέχουν 1015341 εγγραφές. Από τις οποίες έπρεπε να αφαιρεθούν λειψές απαντήσεις στις οποίες δεν υπήρχαν εγγραφές με όλα τα στοιχεία συμπληρωμένα. Αφαιρέθηκαν λοιπόν 1783 εγγραφές και έτσι τελικά , έμειναν 1013558.

3.5 Καθαρισμός δεδομένων

Σε αυτή την ενότητα, περιγράφουμε λεπτομερώς τη διαδικασία που χρησιμοποιήθηκε για τον καθαρισμό του συνόλου δεδομένων με την αφαίρεση τυχόν γραμμών που περιέχουν κενές στήλες. Το βήμα αυτό είναι ζωτικής σημασίας για να διασφαλιστεί η ακεραιότητα και η ποιότητα των δεδομένων πριν προχωρήσουμε σε περαιτέρω ανάλυση ή μοντελοποίηση.

Υλοποίηση:

Χρησιμοποιήσαμε την Python, μια ευέλικτη γλώσσα προγραμματισμού ευρέως αναγνωρισμένη για τον χειρισμό και την ανάλυση δεδομένων, για να υλοποιήσουμε τη διαδικασία καθαρισμού των δεδομένων. Η κύρια βιβλιοθήκη που χρησιμοποιήθηκε για τον χειρισμό δεδομένων σε πίνακες ήταν το Pandas, ένα εργαλείο ανάλυσης και χειρισμού δεδομένων ανοικτού κώδικα.

Διαδικασία:

Ανάγνωση του αρχείου CSV: Το σύνολο δεδομένων φορτώθηκε αρχικά σε ένα πλαίσιο δεδομένων Pandas DataFrame. Αυτή η δομή διευκολύνει τον χειρισμό και την ανάλυση δεδομένων στην Python.

```
python
import pandas as pd
df = pd.read_csv('path/to/yourfile.csv')
```

Προσδιορισμός κενών στηλών:

Εντοπίσαμε γραμμές με τυχόν κενές στήλες (NaN, None ή ισοδύναμες) χρησιμοποιώντας τη μέθοδο isnull() σε συνδυασμό με τη μέθοδο any(). Η μέθοδος isnull() ανιχνεύει τις ελλείπουσες τιμές. Η μέθοδος any(), όταν χρησιμοποιείται με axis=1, εντοπίζει εάν οποιαδήποτε τιμή σε μια γραμμή είναι True (υποδεικνύοντας μια ελλιπή τιμή).

```
python
rows_with_empty_columns = df.isnull().any(axis=1)
```

Αφαίρεση σειρών με κενές στήλες:

Οι σειρές που εντοπίστηκαν να έχουν μία ή περισσότερες κενές στήλες αφαιρέθηκαν από τον πίνακα, διασφαλίζοντας ότι το σύνολο δεδομένων περιείχε μόνο πλήρεις περιπτώσεις.

```
python
cleaned_df = df[~rows_with_empty_columns]
```

Εξαγωγή των καθαρισμένων δεδομένων:

Μετά τη διαδικασία καθαρισμού, το εκλεπτυσμένο σύνολο δεδομένων εξήχθη σε ένα νέο αρχείο CSV, διατηρώντας τα δεδομένα για τα επόμενα βήματα ανάλυσης.

```
python
cleaned_df.to_csv('path/to/cleanedfile.csv',  index=False)
```

Συμπέρασμα:

Αφαιρώντας τις γραμμές με κενές στήλες, βελτιώσαμε την ποιότητα του συνόλου δεδομένων, ελαχιστοποιώντας πιθανές μεροληψίες ή σφάλματα στα επόμενα στάδια ανάλυσης. Αυτό το βήμα προεπεξεργασίας είναι απαραίτητο στις ροές εργασίας της επιστήμης δεδομένων και της ανάλυσης, διασφαλίζοντας την αξιοπιστία και την εγκυρότητα των ευρημάτων της μελέτης.

3.6 Πρακτική διαχωρισμού δεδομένων

Τα δεδομένα προκειμένου να χρησιμοποιηθούν με την χρήση μηχανικής μάθησης χωρίστηκαν εκ των υστέρων σε όποια διαδικασία το απαίτησε σε 80% δεδομένα εκπαίδευσης και 20% δεδομένα ελέγχου.

Η πρακτική του διαχωρισμού των δεδομένων σε σύνολα εκπαίδευσης και δοκιμών, συχνά σε αναλογία 80/20, αποτελεί θεμελιώδη πτυχή πολλών διαδικασιών μηχανικής μάθησης και στατιστικής μοντελοποίησης. Αυτή η στρατηγική διαχωρισμού έχει ως στόχο να διασφαλίσει ότι το μοντέλο εκπαιδεύεται αποτελεσματικά και αξιολογείται με ακρίβεια. Ακολουθούν οι λόγοι για τους οποίους χρησιμοποιείται συνήθως ο διαχωρισμός 80/20:

Επαρκή δεδομένα εκπαίδευσης: Η ανάθεση του 80% των δεδομένων στην εκπαίδευση διασφαλίζει ότι το μοντέλο έχει πρόσβαση σε ένα αρκετά μεγάλο σύνολο δεδομένων για να μάθει αποτελεσματικά. Όσο περισσότερα δεδομένα βλέπει το μοντέλο κατά την εκπαίδευση, τόσο καλύτερα μπορεί να εντοπίσει μοτίβα, σχέσεις και υποκείμενες δομές. Αυτό το εκτεταμένο σύνολο εκπαίδευσης είναι ζωτικής σημασίας για να μπορεί το μοντέλο να γενικεύει καλά, αντί να απομνημονεύει συγκεκριμένες περιπτώσεις.

Επαρκής δοκιμή: Το υπόλοιπο 20% χρησιμεύει ως σύνολο δοκιμών, το οποίο χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου σε αθέατα δεδομένα. Αυτή η αξιολόγηση είναι κρίσιμη για την εκτίμηση του πόσο καλά γενικεύεται το μοντέλο σε νέες, αθέατες περιπτώσεις, γεγονός που αποτελεί βασικό δείκτη της αποτελεσματικότητάς του σε πραγματικές εφαρμογές. Η κατανομή του 20% θεωρείται συχνά μια καλή ισορροπία μεταξύ της ύπαρξης αρκετών δεδομένων για την αξιόπιστη αξιολόγηση της απόδοσης του μοντέλου και της μη παρακράτησης υπερβολικά πολλών δεδομένων από τη διαδικασία εκπαίδευσης.

Πρόληψη της υπερπροσαρμογής: Η υπερπροσαρμογή συμβαίνει όταν ένα μοντέλο μαθαίνει τις λεπτομέρειες και το θόρυβο στα δεδομένα εκπαίδευσης σε βαθμό που επηρεάζει αρνητικά την απόδοση του μοντέλου σε νέα δεδομένα. Κρατώντας ένα μέρος των δεδομένων για δοκιμή, μπορείτε να παρακολουθείτε και να αξιολογείτε κατά πόσον το μοντέλο προσαρμόζεται υπερβολικά. Εάν ένα μοντέλο αποδίδει εξαιρετικά καλά στα δεδομένα εκπαίδευσης, αλλά ελάχιστα στα δεδομένα δοκιμής, είναι πιθανό να έχει υπερπροσαρμογή.

Επικύρωση μοντέλου και ρύθμιση υπερπαραμέτρων: Εκτός από τον βασικό διαχωρισμό εκπαίδευσης/δοκιμής, τα δεδομένα συχνά διαιρούνται περαιτέρω για τη δημιουργία ενός συνόλου επικύρωσης ή χρησιμοποιούνται σε μια τεχνική που ονομάζεται διασταυρούμενη επικύρωση. Αυτό είναι ιδιαίτερα σημαντικό για τη ρύθμιση των υπερπαραμέτρων και τη λήψη αποφάσεων σχετικά με το ποια μοντέλα θα επιλεγούν χωρίς να αγγίξετε το σύνολο δοκιμής. Ενώ ο διαχωρισμός 80/20 είναι μια γενική κατευθυντήρια γραμμή, όταν απαιτείται επίσης ένα σύνολο επικύρωσης, τα δεδομένα μπορεί να χωριστούν σε τρία μέρη (π.χ. 60% εκπαίδευση, 20% επικύρωση, 20% δοκιμή).

Είναι σημαντικό να σημειωθεί ότι ο διαχωρισμός 80/20 δεν είναι ένας απόλυτος και σταθερός κανόνας. Ο ιδανικός διαχωρισμός μπορεί να διαφέρει ανάλογα με το μέγεθος του συνόλου δεδομένων, την πολυπλοκότητα του μοντέλου και τις ιδιαιτερότητες του προβλήματος που αντιμετωπίζεται. Για πολύ μεγάλα σύνολα δεδομένων, ένα μικρότερο ποσοστό (π.χ. 90/10 ή 95/5) μπορεί να είναι επαρκές για τη δοκιμή, ενώ για μικρότερα σύνολα δεδομένων, ένα μεγαλύτερο σύνολο δοκιμής μπορεί να είναι απαραίτητο για να διασφαλιστεί ότι η αξιολόγηση είναι στατιστικά σημαντική. Το κλειδί είναι να βρεθεί μια ισορροπία που να επιτρέπει την ολοκληρωμένη εκπαίδευση, διατηρώντας παράλληλα αρκετά δεδομένα για τη διεξαγωγή μιας αξιόπιστης αξιολόγησης της απόδοσης του μοντέλου.

Όταν έχουμε να κάνουμε με ένα μεγάλο σύνολο δεδομένων, όπως αυτό με 1013558 εγγραφές, η συμβατική κατανομή 80/20 για την εκπαίδευση και τη δοκιμή μπορεί να μην είναι πάντα η βέλτιστη. Ο λόγος είναι ότι με περισσότερα δεδομένα, το απόλυτο μέγεθος του συνόλου δοκιμής μπορεί να είναι μικρότερο σε ποσοστιαίους όρους, ενώ εξακολουθεί να παρέχει μια ισχυρή αξιολόγηση του μοντέλου. Ακολουθεί ο τρόπος με τον οποίο μπορείτε να σκεφτείτε την καλύτερη στρατηγική διαχωρισμού για ένα σύνολο δεδομένων αυτού του μεγέθους:

Επαρκή δεδομένα εκπαίδευσης: Ακόμα και ένας μικρότερος ποσοστιαίος διαχωρισμός παρέχει σημαντικό όγκο δεδομένων για εκπαίδευση. Για παράδειγμα, ένας διαχωρισμός 90/10 εξακολουθεί να προσφέρει πάνω από 900.000 παραδείγματα για εκπαίδευση, τα οποία είναι πιθανότατα υπεραρκετά για τα περισσότερα μοντέλα για να μάθουν αποτελεσματικά.

Στιβαρό σύνολο δοκιμών: Ένα σύνολο δοκιμών 10% σε αυτό το πλαίσιο θα εξακολουθούσε να περιλαμβάνει πάνω από 100.000 παραδείγματα, τα οποία είναι αρκετά μεγάλα και θα πρέπει να παρέχουν μια στατιστικά σημαντική αξιολόγηση της απόδοσης του μοντέλου.

Διασταυρούμενη επικύρωση: Για ορισμένες εφαρμογές, ειδικά όταν η απόδοση του μοντέλου είναι κρίσιμη, θα μπορούσατε να εξετάσετε το ενδεχόμενο χρήσης διασταυρούμενης επικύρωσης. Αυτή η μέθοδος δεν απαιτεί αρχικά ένα ξεχωριστό σύνολο δοκιμών hold-out, καθώς το μοντέλο εκπαιδεύεται και επικυρώνεται σε διαφορετικές αναδιπλώσεις των δεδομένων. Ωστόσο, εξακολουθεί να είναι καλή πρακτική η ύπαρξη ενός τελικού συνόλου δοκιμών που δεν είδατε ποτέ κατά τη διάρκεια της διαδικασίας διασταυρούμενης επικύρωσης, ώστε να διασφαλίζεται η δυνατότητα γενίκευσης του μοντέλου.

Ανάγκες ειδικού τομέα: Ο βέλτιστος διαχωρισμός μπορεί επίσης να εξαρτάται από τις ειδικές απαιτήσεις του τομέα ή του πεδίου μελέτης σας. Σε ορισμένους τομείς, η σύμβαση μπορεί να διαφέρει ή μπορεί να υπάρχει ένα πρότυπο που είναι σημαντικό να τηρηθεί για λόγους συγκρισιμότητας με άλλες μελέτες ή εφαρμογές.

Υπολογιστικοί πόροι: Παρόλο που διαθέτουμε ένα μεγάλο σύνολο δεδομένων, πρέπει να λάβουμε υπόψη μας τους υπολογιστικούς σας πόρους. Η εκπαίδευση σε εκατοντάδες χιλιάδες ή εκατομμύρια σημεία δεδομένων μπορεί να είναι εντατική σε πόρους και χρονοβόρα. Η εξισορρόπηση του μεγέθους του συνόλου εκπαίδευσής μας με το υπολογιστικό κόστος είναι μια άλλη πρακτική σκέψη.

Συνοψίζοντας, ενώ ο κανόνας 80/20 είναι ένα καλό σημείο εκκίνησης για πολλά σύνολα δεδομένων, για ένα σύνολο δεδομένων τόσο μεγάλο όσο 1013558 σειρές, έχουμε την ευελιξία να διαθέσουμε ένα μικρότερο ποσοστό στο σύνολο δοκιμής μας και να εξακολουθήσουμε να έχουμε αρκετά δεδομένα για μια αξιόπιστη αξιολόγηση. Η ακριβής κατανομή μπορεί να προσαρμοστεί με βάση τις συγκεκριμένες ανάγκες, τους υπολογιστικούς πόρους και τα πρότυπα του τομέα στον οποίο εργάζεστε. Στην δική μας περίπτωση λοιπόν κρίναμε ότι με τα 2/10 των δεδομένων που μας δίνουν μια αξιόπιστη πληροφορία ήταν επαρκές να χρησιμοποιήσουμε διαχωρισμό 80/20.

3.7 Φόρτωση δεδομένων

Τα δεδομένα προκειμένου να χρησιμοποιηθούν φορτώθηκαν με την χρήση της Python και με την χρήση της numpy βιβλιοθήκης τα φτιάξαμε σε μια μεταβλητή πίνακα. Ο κώδικας της Python είναι ο παρακάτω:

```
#METRICS
start_time = time.time()
#max=1013558
answers = getResults(nrows=200000) # Extract specific columns as a
NumPy array

#METRICS
print(answers)
stop_time(start_time)
```

Οι μετρικές του συστήματος αφορούσαν την μέτρηση σε δευτερόλεπτα της διαδικασίας. Χρησιμοποιούνται σε πολλές διεργασίες και θα τις συναντήσουμε αργότερα αρκετές φορές. Η συνάρτηση getResults είναι μια δική μας συνάρτηση η οποία κάνει τα παρακάτω:

```
def getResults(nrows):
    results = pd.read_csv('DataFiles/Q&Adata.csv', nrows=40000)
```

```
    answers = results[answer_columns].to_numpy()
    return answers
```

Η συνάρτηση αυτή έχει προεπιλεγμένο δείγμα 40000 εγραφών προκειμένου να μην επιστρέφει κάποιο λογικό σφάλμα και ταυτόχρονα ένα προσιτό στους περισσότερους υπολογιστές υπολογιστικό πλήθος προκειμένου να μην υπερβαίνει τους υπολογιστικούς πόρους. Το αποτέλεσμα είναι το παρακάτω:

```
[[5 1 4 ... 2 3 2]
 [1 2 4 ... 1 3 1]
 [5 1 2 ... 2 1 3]
        ...
 [4 4 4 ... 2 4 2]
 [3 2 4 ... 4 3 2]
 [5 1 5 ... 4 3 5]]

Execution time: 0.28 seconds
```

Ο πίνακας αυτός είναι διαστάσεων (50,200000) οπου 50 είναι οι απαντήσεις από 1 εως 5 στο αναφερόμενο ερώτημα, κάθε 10 αλλάζει ο δείκτης στο οποίο αναφέρεται και ετσι 50/10 = 5 δείκτες στην σείρά. O, C, E, A, N.

3.8 Υπολογισμός δεικτών O, C, E, A, N

Για να υπολογίσουμε τους δείκτες O , C, E, A και N και να τους ενοποιήσουμε σε έναν numpy πίνακα χρησιμοποιούμε τον παρακάτω κώδικα:

```
#METRICS
start_time = time.time()
progress_bar = tqdm(total=5, bar_format='{bar} {percentage:3.0f}%')
############################

score_O=Result_O(answers, O_array, progress_bar)
score_C=Result_C(answers, C_array, progress_bar)
score_E=Result_E(answers, E_array, progress_bar)
score_A=Result_A(answers, A_array, progress_bar)
score_N=Result_N(answers, N_array, progress_bar)

print("Results_OCEAN: \n")
OCEAN=getOCEAN(score_O,score_C,score_E,score_A,score_N)
print(OCEAN)

#METRICS
progress_bar.close()
stop_time(start_time)
##########################
```

Ουσιαστικά αυτό που κάνουμε είναι να δημιουργήσουμε μοναδιάστατα διανύσματα (1,200000) για το κάθε score του κάθε υποψήφιου και στην συνέχεια να τα ενοποιήσουμε σε ένα τελικό πίνακα. Ετσι λοιπόν οι συναρτήσεις Result_O, Result _C, … Result _N είναι με μικρές διαφορές οι εξής:

```
def Result_O(array, O_array, progress_bar):
    if isinstance(array, np.ndarray) and isinstance(O_array,
np.ndarray):
        n = array.shape[0]   # Get the number of rows in the array
        # Use np.dot for matrix multiplication and then add Oinit
```

```
        output = np.dot(array[:, :10], O_array) + Oinit
        output = output.reshape(-1, 1)   # Reshape to (n, 1)
        progress_bar.update(1)
        return output
    else:
        return None  # Return None if the inputs are not valid numpy
arrays
def Result_C(array, C_array, progress_bar):
    if isinstance(array, np.ndarray) and isinstance(C_array,
np.ndarray):
        n = array.shape[0]  # Get the number of rows in the array
        # Use np.dot for matrix multiplication and then add Oinit
        output = np.dot(array[:, 10:20], C_array) + Cinit
        output = output.reshape(-1, 1) # Reshape to (n, 1)
        progress_bar.update(1)
        return output
    else:
        return None  # Return None if the inputs are not valid numpy
arrays
```

Πρόκειται ουσιαστικά για συναρτήσεις οι οποίες έχουν σαν παραμέτρους τον αρχικό πίνακα διαστάσεων (50,n=πλήθος απαντήσεων) από τον οποίο εξάγουν την δεκάδα απαντήσεων που μας ενδιαφέρει και υπολογίζουν για κάθε γραμμή το αντίστοιχο score. Αυτό προϋποθέτει σαν όρισμα τα διανύσματα υπολογισμού για το κάθε score.

```
O_array = np.array([1,-1,1,-1,1,-1,1,1,1,1]) #(10,1)
C_array = np.array([1,-1,1,-1,1,-1,1,-1,1,1])   #(10,1)
E_array = np.array([1,-1,1,-1,1,-1,1,-1,1,-1])   #(10,1)
A_array = np.array([-1,1,-1,1,-1,1,-1,1,1,1])   #(10,1)
N_array = np.array([-1,1,-1,1,-1,-1,-1,-1,-1,-1])   #(10,1)
```

Ταυτόχρονα με την χρήση numpy.dot

$$numpy.\,dot(a,b) = \sum_{i=1}^{n} a_i\, b_i$$

υπολογίζεται το τελικο διάνυσμα ως αποτέλεσμα αυτής της πράξης. Στην συνέχεια η συνάρτηση getOCEAN κάνει μια ενοποίηση των 5 διανυσμάτων σε ένα ενιαίο που περιέχει για τον κάθε υποψήφιο τα 5 αποτελέσματα από τους δείκτες:

```
def getOCEAN(score_O, score_C, score_E, score_A, score_N):
    global OCEAN
    OCEAN = combine_arrays(score_O, score_C, score_E, score_A,
score_N)
    return np.array(OCEAN)
```

όπου combine_arrays η συνάρτηση ενοποίησης:

```
def combine_arrays(arr1, arr2, arr3, arr4, arr5):
    if all(isinstance(arr, np.ndarray) for arr in [arr1, arr2, arr3,
arr4, arr5]):
        if all(arr.shape[0] == arr1.shape[0] for arr in [arr2, arr3,
arr4, arr5]):
            combined_array = np.concatenate((arr1, arr2, arr3, arr4,
arr5), axis=1)
            return combined_array
        else:
```

```
            return None  # Return None if the arrays have different
number of rows
    else:
        return None  # Return None if the inputs are not valid numpy
arrays
```

Η συνάρτηση combine_arrays στο παρεχόμενο απόσπασμα κώδικα έχει σχεδιαστεί για να συνδέει πέντε πίνακες NumPy κατά μήκος του δεύτερου άξονα (στήλες), εάν πληρούν ορισμένες προϋποθέσεις. Ας αναλύσουμε τη συνάρτηση βήμα προς βήμα:

Έλεγχος παραμέτρων για πίνακες NumPy: Η πρώτη δήλωση if ελέγχει αν όλες οι παράμετροι εισόδου (arr1, arr2, arr3, arr4 και arr5) είναι περιπτώσεις πινάκων NumPy. Αυτό γίνεται με τη χρήση ενός συνδυασμού των all() και isinstance(). Η συνάρτηση συνεχίζει μόνο αν όλες οι είσοδοι είναι πίνακες NumPy- διαφορετικά, επιστρέφει None.

```
if all(isinstance(arr, np.ndarray) for arr in [arr1, arr2, arr3, arr4,
arr5]):
```

Έλεγχος συνέπειας σχήματος: Εάν ικανοποιείται η πρώτη συνθήκη, η συνάρτηση ελέγχει στη συνέχεια εάν όλοι οι πίνακες έχουν τον ίδιο αριθμό γραμμών. Αυτό είναι ζωτικής σημασίας για τη συνένωση κατά μήκος των στηλών. Συγκρίνει τον αριθμό των γραμμών (το μέγεθος της πρώτης διάστασης, προσβάσιμο μέσω της arr.shape[0]) κάθε πίνακα (arr2, arr3, arr4, arr5) με αυτόν του πρώτου πίνακα (arr1). Εάν κάποιος από τους πίνακες έχει διαφορετικό αριθμό γραμμών, η συνάρτηση επιστρέφει None.

```
if all(arr.shape[0] == arr1.shape[0] for arr in [arr2, arr3, arr4,
arr5]):
```

Concatenation: Εάν ικανοποιούνται και οι δύο παραπάνω συνθήκες, η συνάρτηση συνενώνει τους πίνακες κατά μήκος του δεύτερου άξονα (δηλαδή των στηλών). Αυτό γίνεται με τη χρήση της np.concatenate(), με τους πίνακες να παρέχονται ως πλειάδα και με τον προσδιορισμό axis=1 για να δηλωθεί η σύνδεση κατά στήλες.

```
combined_array = np.concatenate((arr1, arr2, arr3, arr4, arr5),
axis=1)
```

Μετά τη συνένωση, επιστρέφεται η combined_array. Αυτός ο πίνακας θα έχει τον ίδιο αριθμό γραμμών με τους πίνακες εισόδου αλλά αριθμό στηλών ίσο με το άθροισμα των στηλών των επιμέρους πινάκων.

Επιστροφή του None: Εάν οποιοσδήποτε από τους ελέγχους αποτύχει (είτε ο έλεγχος τύπου είτε ο έλεγχος συνέπειας σχήματος), η συνάρτηση επιστρέφει το None. Αυτό χρησιμεύει ως ένδειξη ότι η συνένωση δεν μπόρεσε να πραγματοποιηθεί είτε λόγω ασυμβατότητας τύπου είτε λόγω ασυνέπειας σχήματος.

Το αποτέλεσμα ήταν το εξής:

```
Results_OCEAN:

[[35 22 36 29 26]
 [25 27 10 34 25]
 [31 24 15 32 24]
 ...
 [26 34  7 17 17]
 [21 17 17 13 16]
 [34 21 23 31 14]]

Execution time: 0.01 seconds
```

3.9 Δημιουργία Dataframe

Στην προσπάθεια μας να δημιουργήσουμε ένα dataframe των δεδομένων θα έχουμε σε κάθε γραμμή τις 50 απαντήσεις και στην συνέχεια τα 5 score, εκτελέσαμε τον παρακάτω κώδικα:

```
#METRICS
start_time = time.time()
# Define column names for X and y without numbers
DF = pd.DataFrame(answers, columns=[f'{name}{i+1}' for name in
column_names for i in range(10)])
DF['O'] = OCEAN[:, 0]
DF['C'] = OCEAN[:, 1]
DF['E'] = OCEAN[:, 2]
DF['A'] = OCEAN[:, 3]
DF['N'] = OCEAN[:, 4]

#METRICS
stop_time(start_time)
DF
```

Ως αποτέλεσμα πήραμε το παρακάτω:

AGR5	AGR6	AGR7	AGR8	AGR9	AGR10	NRT1	NRT2	NRT3	NRT4	NRT5	NRT6	NRT7	NRT8	NRT9	NRT10	O	C	E	A	N
2	3	2	4	3	4	1	4	4	2	2	2	2	2	3	2	35	22	36	29	26
1	5	3	4	5	3	2	3	4	1	3	1	2	1	3	1	25	27	10	34	25
2	4	1	4	4	3	4	4	4	2	2	2	2	2	1	3	31	24	15	32	24
2	4	2	4	3	4	3	3	3	2	3	2	2	2	4	3	29	15	16	28	21
1	3	1	5	5	3	1	5	5	3	1	1	1	1	3	2	38	38	19	36	31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	3	2	3	4	3	5	3	5	1	4	5	5	5	5	5	26	19	23	26	3
5	1	3	1	1	2	5	2	5	1	3	5	5	5	5	5	40	8	12	6	3
4	2	2	2	2	2	4	3	3	3	5	3	4	2	4	2	26	34	7	17	17
4	2	4	4	2	2	4	2	3	0	3	2	3	4	3	2	21	17	17	13	16
2	5	2	4	4	4	4	4	4	2	3	3	4	4	3	5	34	21	23	31	14

Εικόνα 13: DataFrame δεδομένων με τα τελικά απότελέσματα και τις απαντήσεις κάθε υποψηφίου

4. Σενάρια Πειραματισμού

4.1 Linear Regression

Στην συνέχεια η προσπάθεια μας επικεντρώθηκε μέσω της γραμμικής παλινδρόμησης να πάρουμε το ίδιο αποτελέσματα από τις απαντήσεις με αυτό που υπολογίσαμε προηγουμένως από την θεωρία. Ο κώδικας που εκτελεί αυτή τη διαδικασία είναι ο παρακάτω:

```
X = DF[answer_columns]
Y = DF[score_columns]
Regr = linear_model.LinearRegression()
Regr.fit(Y, X)

# Split the data into training and test sets
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.2, random_state=0)
```

```
print("Y_train:", Y_train.shape)
print("X_train:", X_train.shape)
print("Y_test:", Y_test.shape)
print("X_test:", X_test.shape)

# Create and train the Linear Regression model
Linreg = LinearRegression()
Linreg.fit(X_train, Y_train)

# Make predictions on the test set
Y_pred = Linreg.predict(X_test)

print("Y_pred:", Y_pred.shape, "\n")
#print("Y_pred[0]:\n", Y_pred[0], "\n")

print_metrics(Y_test,Y_pred)
```

Πράγματι τα αποτελέσματα ήταν τα αναμενόμενα.

```
Accuracy of the model: 100.00%
Mean Absolute Error: 0.00
Mean Squared Error: 0.00
```

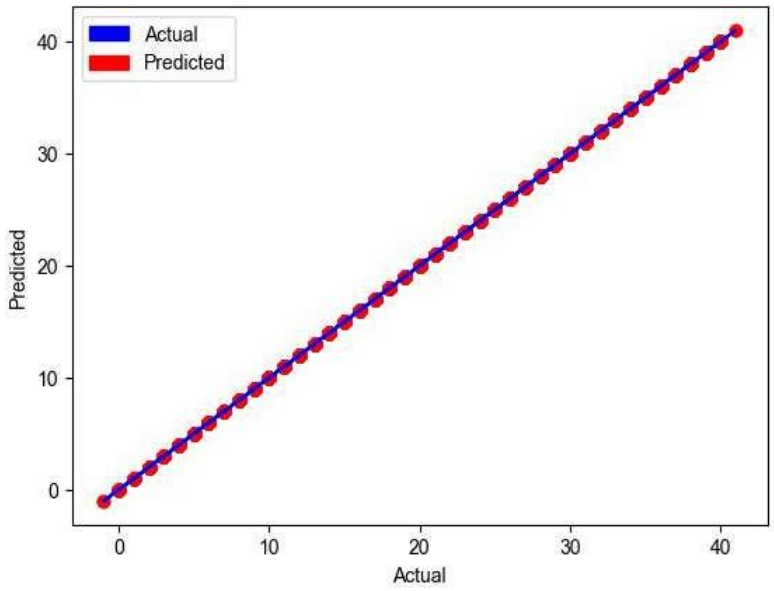
Η ακρίβεια ηταν 100% και με βάση τα αποτελέσματα σχηματίσαμε τις εξισώσεις υπολογισμού, με τον παρακάτω κώδικα:

```
# Get the function used by the model
function_coefficients = Linreg.coef_
function_intercept = Linreg.intercept_

# Print the function used by the model
for i, target_variable in enumerate(score_columns):
    coefficients = function_coefficients[i]

    # Filter out terms with zero coefficients and format non-zero
terms
    non_zero_terms = [f"{coeff:.2f}*{answer_columns[j]}" for j, coeff
in enumerate(coefficients) if coeff != 0]

    # Check if there are non-zero terms to display
    if non_zero_terms:
        # Join the non-zero terms with "+" and format the function
string
        terms_str = " + ".join(non_zero_terms)
        function_str = f"Function for {target_variable}:
{target_variable} = {function_intercept[i]:.2f} + {terms_str}"

        # Display the formatted function string
        print_modify_string_functions(function_str)
```

Χρησιμοποιήσαμε κανονικές εκφράσεις προκειμένου να διώξουμε τους μηδενικούς και τους πολύ μικρούς όρους δύναμης του -3

- `Function for O: O =8.00 + 1.00*OPN1 -1.00*OPN2 + 1.00*OPN3 -1.00 *OPN4 +1.00*OPN5 -1.00*OPN6 + 1.00*OPN7 + 1.00*OPN8 + 1.00*OPN9 + 1.00*OPN10`

- Function for C: C = 14.00+ 1.00*CSN1 -1.00*CSN2 + 1.00*CSN3 -1.00*CSN4 + 1.00*CSN5 -1.00*CSN6 + 1.00*CSN7 -1.00*CSN8 + 1.00*CSN9 + 1.00*CSN10
- Function for E: E = 20.00+ 1.00*EXT1 -1.00*EXT2 + 1.00*EXT3 -1.00*EXT4 + 1.00*EXT5 -1.00*EXT6 + 1.00*EXT7 -1.00*EXT8 + 1.00*EXT9 -1.00*EXT10
- Function for A: A = 14.00-1.00*AGR1 + 1.00*AGR2 -1.00*AGR3 + 1.00*AGR4 -1.00*AGR5 + 1.00*AGR6 -1.00*AGR7 + 1.00*AGR8 + 1.00*AGR9 + 1.00*AGR10
- Function for N: N = 38.00-1.00*NRT1 + 1.00*NRT2 -1.00*NRT3 + 1.00*NRT4 -1.00*NRT5 -1.00*NRT6 -1.00*NRT7 -1.00*NRT8 -1.00*NRT9 -1.00*NRT10



Στην συνέχεια η γραφική παράσταση ακρίβειας δεδομένων υποδεικνύει το ίδιο αποτέλεσμα:

Εικόνα 14: Γράφιμα πρόβλεψης Linear Regression συνάρτησης

4.2 Linear Regression - with lack of data

Το επόμενο βήμα είναι με την ίδια μέθοδο, η πρόβλεψη των δεικτών αυτών, χωρίς την παρουσία όλων των απαντήσεων αυτήν την φορά. Αυτό που θέλουμε να ελέγξουμε είναι αν μπορεί να προβλεφθεί κάποιος δείκτης από τους O,C,E,A,N χωρίς τις 10 απαντήσεις που αφορούν αυτόν τον δείκτη. Για να το καταφέρουμε αυτό θα πάρουμε από το dataFrame κάθε φορά για 5 φορές 40 απαντήσεις άσχετες με τον δείκτη που θέλουμε να υπολογίσουμε.

```
# Define column names for X and y without numbers
df1 = pd.DataFrame(answers[:, 10:], columns=[f'{name}{i+1}' for name
in column_names_without_OPN  for i in range(10)])
df2 = pd.DataFrame(np.concatenate((answers[:, :10],answers[:, 20:]),
axis=1), columns=[f'{name}{i+1}' for name in column_names_without_CSN
for i in range(10)])
```

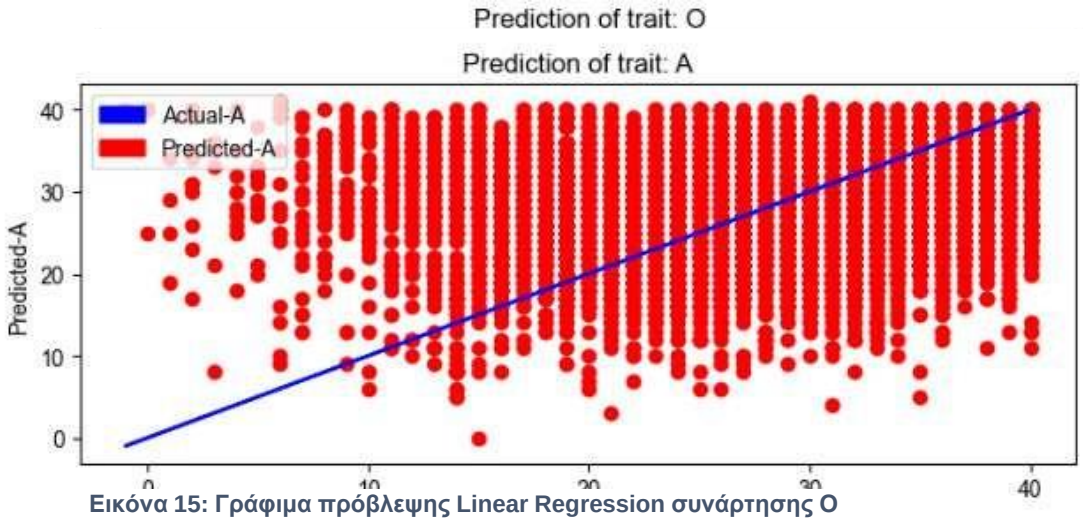
```
df3 = pd.DataFrame(np.concatenate((answers[:, :20],answers[:, 30:]),
axis=1), columns=[f'{name}{i+1}' for name in column_names_without_EXT
for i in range(10)])
df4 = pd.DataFrame(np.concatenate((answers[:, :30],answers[:, 40:]),
axis=1), columns=[f'{name}{i+1}' for name in column_names_without_AGR
for i in range(10)])
df5 = pd.DataFrame(answers[:, :40], columns=[f'{name}{i+1}' for name
in column_names_without_NRT  for i in range(10)])

df1['O'] = OCEAN[:, 0]
df1['C'] = OCEAN[:, 1]
df1['E'] = OCEAN[:, 2]
df1['A'] = OCEAN[:, 3]
df1['N'] = OCEAN[:, 4]

df2['O'] = OCEAN[:, 0]
df2['C'] = OCEAN[:, 1]
df2['E'] = OCEAN[:, 2]
df2['A'] = OCEAN[:, 3]
df2['N'] = OCEAN[:, 4]

df3['O'] = OCEAN[:, 0]
df3['C'] = OCEAN[:, 1]
df3['E'] = OCEAN[:, 2]
df3['A'] = OCEAN[:, 3]
df3['N'] = OCEAN[:, 4]

df4['O'] = OCEAN[:, 0]
df4['C'] = OCEAN[:, 1]
df4['E'] = OCEAN[:, 2]
df4['A'] = OCEAN[:, 3]
df4['N'] = OCEAN[:, 4]

df5['O'] = OCEAN[:, 0]
df5['C'] = OCEAN[:, 1]
df5['E'] = OCEAN[:, 2]
df5['A'] = OCEAN[:, 3]
df5['N'] = OCEAN[:, 4]


x1 = df1[answer_columns_without_OPN]
y1 = df1[score_columns]
regr = linear_model.LinearRegression()
regr.fit(y1, x1)

x2 = df2[answer_columns_without_CSN]
y2 = df2[score_columns]
regr = linear_model.LinearRegression()
regr.fit(y2, x2)

x3 = df3[answer_columns_without_EXT]
y3 = df3[score_columns]
regr = linear_model.LinearRegression()
regr.fit(y3, x3)
```

```
x4 = df4[answer_columns_without_AGR]
y4 = df4[score_columns]
regr = linear_model.LinearRegression()
regr.fit(y4, x4)

x5 = df5[answer_columns_without_NRT]
y5 = df5[score_columns]
regr = linear_model.LinearRegression()
regr.fit(y5, x5)

x1_train,x1_test,y1_train,y1_test=train_test_split(x1,y1,test_size=0.2
,random_state=0)
x2_train,x2_test,y2_train,y2_test=train_test_split(x2,y2,test_size=0.2
,random_state=0)
x3_train,x3_test,y3_train,y3_test=train_test_split(x3,y3,test_size=0.2
,random_state=0)
x4_train,x4_test,y4_train,y4_test=train_test_split(x4,y4,test_size=0.2
,random_state=0)
x5_train,x5_test,y5_train,y5_test=train_test_split(x5,y5,test_size=0.2
,random_state=0)

print("X_train:",x1_train.shape,x2_train.shape,x3_train.shape,x4_train
.shape,x5_train.shape)
print("X_test:",x1_test.shape,x2_test.shape,x3_test.shape,x4_test.shap
e,x5_test.shape)
print("Y_train:",y1_train.shape,y2_train.shape,y3_train.shape,y4_train
.shape,y5_train.shape)
print("Y_test:",y1_test.shape,y2_test.shape,y3_test.shape,y4_test.shap
e,y5_test.shape)
```

Έτσι λοιπόν δημιουργούμε κάθε φορά (5 στο σύνολο) διαφορετικές συναρτήσεις. Τα αποτελέσματα για τον κάθε δείκτη είναι τα παρακάτω:

```
Model without OPN, Metrics:
Accuracy of the model: 84.85%, Mean Absolute Error: 0.89, Mean Squared
Error: 38.32
Model without CSN, Metrics:
Accuracy of the model: 83.87%, Mean Absolute Error: 1.04, Mean Squared
Error: 71.44
Model without EXT, Metrics:
Accuracy of the model: 87.28%, Mean Absolute Error: 1.15, Mean Squared
Error: 103.98
Model without AGR, Metrics:
Accuracy of the model: 86.27%, Mean Absolute Error: 0.95, Mean Squared
Error: 53.28
Model without NRT, Metrics:
Accuracy of the model: 86.58%, Mean Absolute Error: 1.09, Mean Squared
Error: 86.23

Execution time: 0.52 seconds
```

Η επιτυχία κυμαίνεται από το 71%-87% για τον κάθε δείκτη. Αυτό μας δίνει την υποψία για το ότι κάποιες ερωτήσεις από τους άλλους δείκτες είναι πιο πιθανό να έχουν κάποια συσχέτιση περισσότερο έναντι των άλλων. Αν και με τα γραφήματα δεν είναι δυνατό να αναπαρασταθεί ποιοτικά δεδομένου ότι οι εγγραφές είναι πολλές.



Εικόνα 15: Γράφιμα πρόβλεψης Linear Regression συνάρτησης O χωρίς τις απαντήσεις που τον αφορούν

Εικόνα 16: Γράφιμα πρόβλεψης Linear Regression συνάρτησης A χωρίς τις απαντήσεις που τον αφορούν

4.3 Non-Linear Regression - Polynomial function

Στην συνέχεια θα εξετάσουμε αν ένα μοντέλο πολυωνυμικού τύπου είναι ικανό να προβλέψει στον ίδιο βαθμό όπως η γραμμική παλινδρόμηση τους δείκτες. Πρέπει να σημειωθεί ότι το πολυώνυμο θα είναι 2^{ου} βαθμού. Ο κώδικα που το εκτελεί , είναι ο παρακάτω:

```
# Define the degree of the polynomial
degree = 2

# Create polynomial features
Poly = PolynomialFeatures(degree)
X_poly_train = Poly.fit_transform(X_train)
X_poly_test = Poly.transform(X_test)

# Create and train the Polynomial Regression model
Poly_reg = LinearRegression()
Poly_reg.fit(X_poly_train, Y_train)

# Make predictions on the test set
Y_poly_pred = Poly_reg.predict(X_poly_test)

# Calculate and print the evaluation metrics
print_metrics(Y_test, Y_poly_pred)
```

Η λογική του κώδικα είναι η εξής:

Η γραμμή degree = 2 ορίζει τον βαθμό των πολυωνυμικών χαρακτηριστικών που θα δημιουργηθούν. Σε αυτή την περίπτωση, τα πολυωνυμικά χαρακτηριστικά θα είναι βαθμού 2, πράγμα που σημαίνει ότι το μοντέλο θα εξετάσει όχι μόνο τους γραμμικούς όρους (x) αλλά και τους τετραγωνικούς όρους (x^2).

Δημιουργία πολυωνυμικών χαρακτηριστικών:

PolynomialFeatures(degree) δημιουργεί μια περίπτωση του PolynomialFeatures με τον καθορισμένο βαθμό. Αυτή η περίπτωση θα χρησιμοποιηθεί για τη δημιουργία πολυωνυμικών χαρακτηριστικών από τα δεδομένα εισόδου.

X_poly_train = Poly.fit_transform(X_train) μετασχηματίζει τα δεδομένα εκπαίδευσης X_train σε πολυωνυμικά χαρακτηριστικά. fit_transform όχι μόνο προσαρμόζει τον μετασχηματιστή στα δεδομένα αλλά και μετασχηματίζει τα δεδομένα. Τα μετασχηματισμένα δεδομένα X_poly_train περιλαμβάνουν τώρα τα αρχικά χαρακτηριστικά συν τα πολυωνυμικά χαρακτηριστικά μέχρι τον καθορισμένο βαθμό.

X_poly_test = Poly.transform(X_test) μετασχηματίζει τα δεδομένα δοκιμής X_test σε πολυωνυμικά χαρακτηριστικά χρησιμοποιώντας τον ίδιο μετασχηματισμό που εφαρμόστηκε στα δεδομένα εκπαίδευσης. Είναι σημαντικό να χρησιμοποιήσετε εδώ μόνο το transform για να διασφαλίσετε ότι τα δεδομένα δοκιμής μετασχηματίζονται με τον ίδιο τρόπο όπως τα δεδομένα εκπαίδευσης.

Δημιουργία και εκπαίδευση του μοντέλου πολυωνυμικής παλινδρόμησης:

Poly_reg = LinearRegression() δημιουργεί μια περίπτωση ενός μοντέλου γραμμικής παλινδρόμησης.

Poly_reg.fit(X_poly_train, Y_train) εκπαιδεύει το μοντέλο χρησιμοποιώντας τα πολυωνυμικά χαρακτηριστικά των δεδομένων εκπαίδευσης και τις τιμές-στόχους Y_train. Παρόλο που χρησιμοποιούμε ένα μοντέλο γραμμικής παλινδρόμησης, τα πολυωνυμικά χαρακτηριστικά του επιτρέπουν να μοντελοποιήσει μια μη γραμμική σχέση.

Προβλέψεις:

Y_poly_pred = Poly_reg.predict(X_poly_test) χρησιμοποιεί το εκπαιδευμένο μοντέλο για να κάνει προβλέψεις στα δεδομένα δοκιμής. Το μοντέλο προβλέπει τις τιμές-στόχους με βάση τα πολυωνυμικά χαρακτηριστικά των δεδομένων δοκιμής.

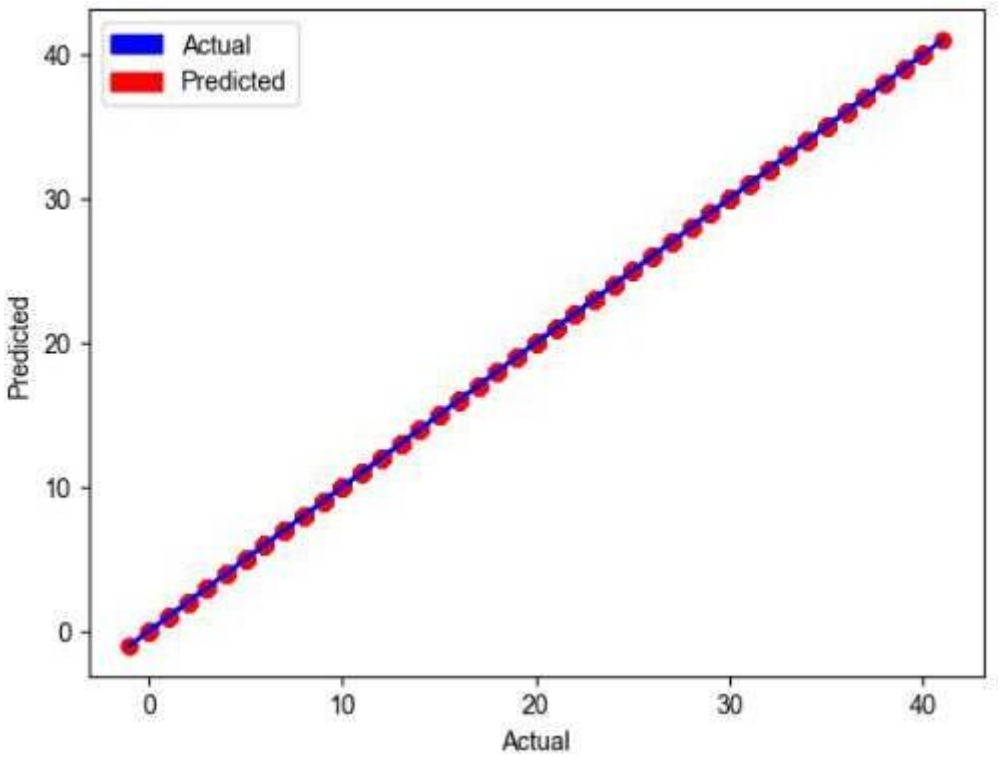
Υπολογισμός και εκτύπωση μετρήσεων αξιολόγησης:

print_metrics(Y_test, Y_poly_pred) είναι κατά πάσα πιθανότητα μια συνάρτηση που υπολογίζει και εκτυπώνει τις μετρικές αξιολόγησης για το μοντέλο. Αυτή η συνάρτηση θα συγκρίνει τις πραγματικές τιμές στόχου Y_test με τις προβλεπόμενες τιμές Y_poly_pred για να καθορίσει πόσο καλά αποδίδει το μοντέλο. Οι συνήθεις μετρικές θα μπορούσαν να περιλαμβάνουν το μέσο απόλυτο σφάλμα (MAE), το μέσο τετραγωνικό σφάλμα (MSE) ή το R-τετράγωνο.

Η ακρίβεια του μοντέλου ήταν 100%

```
Accuracy of the model: 100.00%, Mean Absolute Error: 0.00, Mean Square
d Error: 0.00
```

Με τις συναρτήσεις να είναι πάλι ίδιες όπως στο Linear Regression.



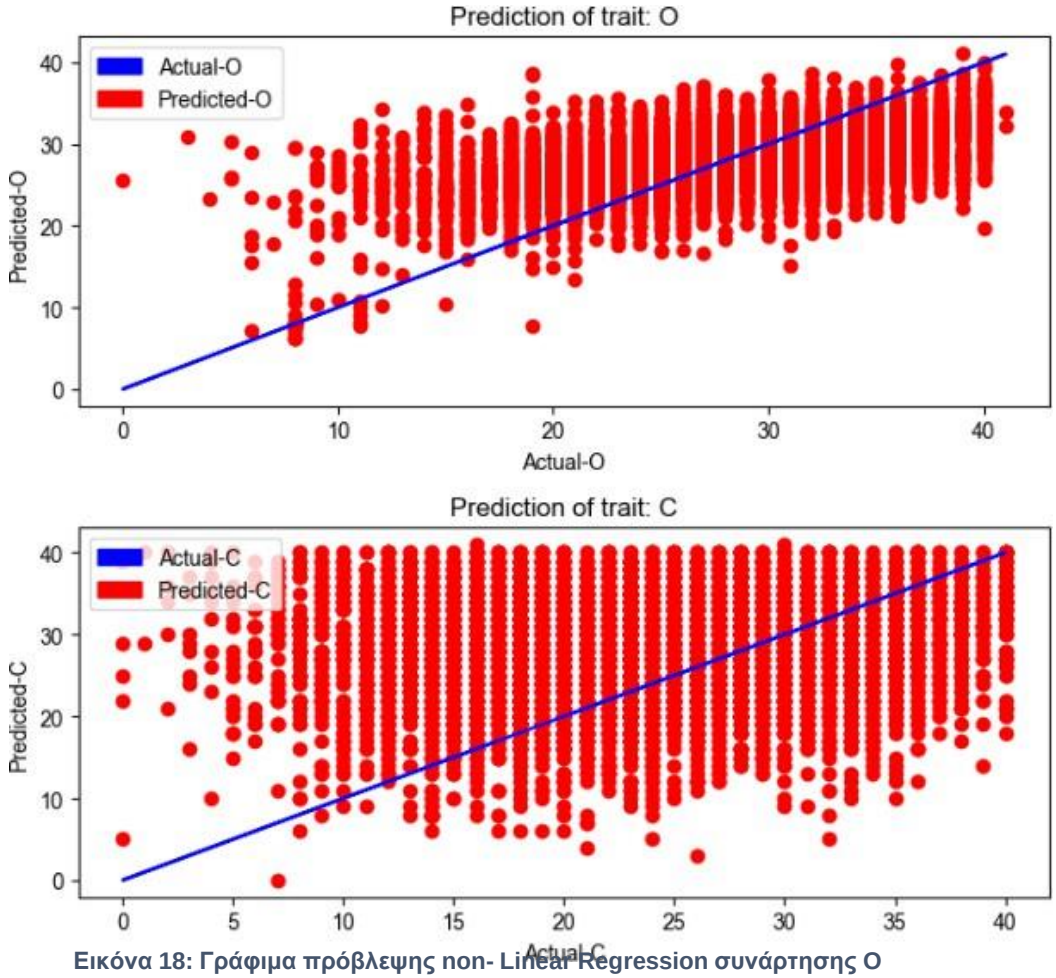
Εικόνα 17: Γράφιμα πρόβλεψης non - Linear Regression συνάρτησης

4.4 Non-Linear Regression - Polynomial function with lack of data

Στην περίπτωση αυτή εκτελούμε ακριβώς την ίδια συνάρτηση χωρίς αυτή τη φορά να έχουμε τις σχετικές με τους δείκτες απαντήσεις. Τα αποτελέσματα είναι τα εξής:

```
Model without OPN, Metrics (Polynomial Regression):
Accuracy of the model: 85.46%, Mean Absolute Error:  0.86, Mean Squared
Error: 35.26
Model without CSN, Metrics (Polynomial Regression):
Accuracy of the model: 83.95%, Mean Absolute Error:  1.03, Mean Squared
Error: 70.70
Model without EXT, Metrics (Polynomial Regression):
Accuracy of the model: 87.41%, Mean Absolute Error:  1.14, Mean Squared
Error: 101.87
Model without AGR, Metrics (Polynomial Regression):
Accuracy of the model: 86.42%, Mean Absolute Error:  0.94, Mean Squared
Error: 52.09
Model without NRT, Metrics (Polynomial Regression):
Accuracy of the model: 86.66%, Mean Absolute Error:  1.08, Mean Squared
Error: 85.24
```

Αρκετά καλύτερα σε σχέση την περίπτωση του Linear Regression, καθώς η αποτελεσματικότ ητα αυξήθηκε από το κατώτατο όριο που ήταν 71% στο 83%. Παρόλα αυτά πάλι είναι δύσκολο να αποτ υπωθεί το σφάλμα σε γραφήματα λόγω των πολλών δεδομένων:



Εικόνα 18: Γράφιμα πρόβλεψης non- Linear Regression συνάρτησης O κκαι C χωρίς τις απαντήσεις που τον αφορούν

4.5 Γενετικός αλγόριθμος. Μια προσαρμοσμένη λογική.

Επειδή ο κώδικας που εκτελεί αυτήν την διαδικασία είναι πολύ μεγάλος και αρκετά πολύπλοκος, στην συνέχεια θα εξηγήσουμε την λειτουργία του. Ο αλγόριθμος προσπαθεί τη βελτιστοποίηση ενός πληθυσμού λύσεων σε σχέση με μια συνάρτηση καταλληλόλητας. Το πλαίσιο σχετίζεται με την εξαγωγή βέλτιστων διανυσμάτων που συσχετίζονται με τα χαρακτηριστικά της προσωπικότητας με βάση τις απαντήσεις. Ας αναλύσουμε τα κύρια συστατικά και τη λειτουργικότητα:

Αρχική ρύθμιση:

Το **OCEAN** αντιπροσωπεύει τον πινακα με τα πέντε χαρακτηριστικά της προσωπικότητας (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism).

Οι συναρτήσεις getOCEAN και getResults ορίζονται για τη ρύθμιση του πίνακα OCEAN και τη φόρτωση των απαντήσεων του ερωτηματολογίου, αντίστοιχα.

Ρύθμιση γενετικού αλγορίθμου:

Ορίζεται μια κλάση Genetic_algorythm_Linear_Fitness με μεθόδους για την εκτέλεση διαφόρων λειτουργιών γενετικού αλγορίθμου, όπως αρχικοποίηση, δημιουργία πληθυσμού, αξιολόγηση καταλληλότητας, διασταύρωση, μετάλλαξη και βελτιστοποίηση.

Ο γενετικός αλγόριθμος χρησιμοποιεί ένα μοντέλο γραμμικής παλινδρόμησης (LinearRegression) για τον υπολογισμό της καταλληλόλητας, η οποία αντιπροσωπεύεται από το μέσο

τετραγωνικό σφάλμα (MSE) μεταξύ των προβλεπόμενων και των πραγματικών βαθμολογιών των χαρακτηριστικών προσωπικότητας.

ΔημιΟυργία πληθυσμού:

Μέθοδοι όπως οι **getSpecificPopulation** και **getPopulation** δημιουργούν αρχικούς πληθυσμούς δυαδικών διανυσμάτων. Αυτά τα διανύσματα θα μπορούσαν να αντιπροσωπεύουν, για παράδειγμα, επιλογές στοιχείων ερωτηματολογίου. Η **loadPopulation** φορτώνει έναν αποθηκευμένο πληθυσμό ή δημιουργεί έναν νέο αν δεν υπάρχει.

ΑξιΟλόγηση καταλληλότητας:

Οι μέθοδοι **Fitness_vector** και **Fitness** αξιολογούν την καταλληλόληα των ατόμων του πληθυσμού. Αυτό γίνεται με τη χρήση των διανυσμάτων για την πρόβλεψη των βαθμολογιών των χαρακτηριστικών προσωπικότητας από τις απαντήσεις του ερωτηματολογίου και τη σύγκριση των προβλέψεων με τις πραγματικές βαθμολογίες χρησιμοποιώντας το MSE.

Διασταύρωση καΙ μετάλλαξη:

Η μέθοδος διασταύρωσης αναμειγνύει γενετικές πληροφορίες μεταξύ των ατόμων του πληθυσμού, δημιουργώντας ενδεχομένως καλύτερες λύσεις. Η μέθοδος μετάλλαξης εισάγει τυχαίες αλλαγές στα άτομα, βοηθώντας τον αλγόριθμο να εξερευνήσει νέες λύσεις και να αποφύγει τοπικά ελάχιστα.

Βρόχος βελτΙστοποίησης:

Η μέθοδος **genetic_algorythm** εφαρμόζει επαναληπτικά τη διασταύρωση και τη μετάλλαξη για να εξελίξει τον πληθυσμό προς καλύτερες λύσεις σε έναν καθορισμένο αριθμό γενεών. Ο αλγόριθμος στοχεύει στην ελαχιστοποίηση της συνάρτησης καταλληλότητας (MSE), βρίσκοντας έτσι τα διανύσματα που προβλέπουν καλύτερα τις βαθμολογίες των χαρακτηριστικών προσωπικότητας από τις απαντήσεις του ερωτηματολογίου.

Αποθήκευση καΙ φόρτωση:

Ο αλγόριθμος διαθέτει λειτουργικότητα για την αποθήκευση και τη φόρτωση πληθυσμών και πινάκων καταλληλότητας σε/από αρχεία CSV, επιτρέποντας την παύση και τη συνέχιση της διαδικασίας βελτιστοποίησης.

ΧρηστΙκές λειτουργίες:

Πρόσθετες βοηθητικές συναρτήσεις, όπως οι **sort_arrays, swap** και **return_X_y**, παρέχουν απαραίτητες λειτουργίες για τη διαχείριση πληθυσμών, την τροποποίηση πινάκων και την προετοιμασία δεδομένων για ανάλυση παλινδρόμησης.

Συνολικά, αυτός ο γενετικός αλγόριθμος είναι προσαρμοσμένος για τη βελτιστοποίηση ενός συνόλου διανυσμάτων που χρησιμοποιούνται για την πρόβλεψη χαρακτηριστικών προσωπικότητας από απαντήσεις σε ερωτηματολόγια, με κριτήριο βελτιστοποίησης την ακρίβεια αυτών των προβλέψεων που μετριέται από το MSE.

Ταυτόχρονα γίνεται χρήση multithreading προκειμένου το αποτέλεσμα του τελικού διανύσματος να επιτελείται όσο το δυνατόν πιο γρήγορα. Καταλήγει τελικά ο αλγόριθμος να επιστρέφει μοναδικά διανύσματα για τον κάθε δείκτη τα οποία μετέχουν με σημαντικό τρόπο στο τελικό αποτέλεσμα. Ο Γενετικός αλγόριθμος εκτελέστηκε 2 φορές με την εξής διαφορά. Την πρώτη φορά χρησιμοποιήθηκε για μέσο τετραγωνικό σφάλμα η συνάρτηση linear regression ενώ την δεύτερη φορά η συνάρτηση non-linear regression. Τα αποτελέσματα είναι τα εξής:

	O	C	E	A	N
MSE-LINEAR	38.23	50.43	75.01	53.16	58.67
MSE NON-LINEAR	37.18	49.31	73.93	47.53	57.16

Όπως και χωρίς την χρήση γενετικού αλγορίθμου, η πολυώνυμική εξίσωση έδωσε καλύτερα αποτελέσματα, έτσι και με την χρήση γενετικού αλγορίθμου βλέπουμε ότι τα αποτελέσματα βελτιστοποιούνται σε μικρότερο βαθμό βέβαια.

4.6 Μηχανική μάθηση. Αναζήτηση καλύτερης δυνατής λύσης.

Επειδή ο κώδικας που εκτελεί αυτήν την διαδικασία είναι πολύ μεγάλος όπως και ο γενετικός αλγόριθμος και αρκετά πολύπλοκος, στην συνέχεια θα εξηγήσουμε την λειτουργεία του. Αρχικά να αναφέρουμε ότι ξεκίνησε η διαδικασία εύρεσης βέλτιστης λύσης με την χρήση του καλύτερου διανύσματος του γενετικού αλγορίθμου όπου το μέσο τετραγωνικό σφάλμα το παίρναμε από την πολυώνυμική εξίσωση.

Αυτός ο αλγορυθμος ορίζει ένα πλαίσιο μηχανικής μάθησης για τη δημιουργία, την εκπαίδευση και τη βελτιστοποίηση μοντέλων νευρωνικών δικτύων για διάφορες εργασίες, αξιοποιώντας την κλάση γενετικού αλγορίθμου που ορίστηκε προηγουμένως. Παρέχει λειτουργίες για την αποθήκευση, τη φόρτωση και την αξιολόγηση μοντέλων, καθώς και για τη δυναμική κατασκευή και εκπαίδευση μοντέλων με διαφορετικές αρχιτεκτονικές και συναρτήσεις ενεργοποίησης. Ας αναλύσουμε τις κύριες λειτουργικότητες:

Λεξικό συναρτήσεων ενεργοποίησης:

Το λεξικό ενεργοποιήσεων αντιστοιχίζει αναγνωριστικά σε διάφορες συναρτήσεις ενεργοποίησης νευρωνικών δικτύων, παρέχοντας μια σύντομη περιγραφή και μια τυπική χρήση για την καθεμία.

```
activations = {
1: ('relu', 'Ranges from 0 to positive infinity. Commonly used for
hidden layers.'),
2: ('tanh', 'Ranges from -1 to 1. Often used for hidden layers.'),
3: ('sigmoid', 'Ranges from 0 to 1. Commonly used in the output layer
for binary classification.'),
4: ('softmax', 'Used in the output layer for multiclass
classification. Converts logits to probabilities.'),
5: ('linear', 'No specific range. Often used in the output layer for
regression tasks.'),
6: ('softplus', 'Ranges from 0 to positive infinity. Smooth
approximation of ReLU.'), #36
7: ('softsign', 'Ranges from -1 to 1. Similar to tanh but with a
simpler shape.'),
8: ('hard_sigmoid', 'Approximates sigmoid but computationally cheaper.
Ranges from 0 to 1.'),
9: ('elu', 'Ranges from negative infinity to positive infinity. A
variant of ReLU.'),
10:('selu', 'Self-normalizing variant of ReLU. Maintains mean and
variance during training.'),
11:('exponential', 'Ranges from 0 to positive infinity. An activation
with exponential growth.')
}
```

Κλάση μοντέλου μηχανικής μάθησης:

Η κλάση Machine_Learning_Model ενσωματώνεται με την κλάση γενετικού αλγορίθμου, παρέχοντας μεθόδους για τον χειρισμό μοντέλων νευρωνικών δικτύων.

Διαχείριση αρχείων μοντέλων:

Μέθοδοι όπως **save_model**, **load_model** και **remove** χειρίζονται την αποθήκευση και ανάκτηση των αρχιτεκτονικών και των βαρών του μοντέλου, διευκολύνοντας τη διατήρηση των εκπαιδευμένων μοντέλων.

Εκπαίδευση και αξιολόγηση μοντέλων:

Η μέθοδος ακρίβειας αξιολογεί την απόδοση ενός μοντέλου χρησιμοποιώντας το μέσο τετραγωνικό σφάλμα (MSE).

Η μέθοδος **generate_model** κατασκευάζει ένα επίπεδο νευρωνικού δικτύου με καθορισμένες παραμέτρους, προσθέτοντάς το σε ένα υπάρχον μοντέλο.

Αναζήτηση και βελτιστοποίηση μοντέλου (ml_search):

Αυτή η μέθοδος διερευνά επαναληπτικά διαφορετικές διαμορφώσεις νευρωνικών δικτύων, προσαρμόζοντας τα στρώματα, τις συναρτήσεις ενεργοποίησης και τον αριθμό των μονάδων για την ελαχιστοποίηση της συνάρτησης απωλειών (MSE σε αυτή την περίπτωση).

Χρησιμοποιεί πρώιμη διακοπή για να αποτρέψει την υπερβολική προσαρμογή και έναν προσαρμοσμένο βρόχο εκπαίδευσης για να αποφασίσει πότε θα σταματήσει την εκπαίδευση με βάση τις τάσεις των απωλειών.

Η διαμόρφωση του μοντέλου με τις καλύτερες επιδόσεις αποθηκεύεται και η διαδικασία μπορεί να προσαρμόζει δυναμικά την πολυπλοκότητα του μοντέλου (αριθμός στρωμάτων και μονάδων) με βάση τις βελτιώσεις των επιδόσεων.

ΔημιΟυργία μοντέλων (generate_ml_models):

Αυτή η μέθοδος πραγματοποιεί βρόχους σε διαφορετικές διαστάσεις (που πιθανώς αντιπροσωπεύουν διαφορετικές εργασίες ή σύνολα δεδομένων) και εφαρμόζει την **ml_search** για να βρει και να αποθηκεύσει βελτιστοποιημένα μοντέλα για κάθε εργασία.

Αξιοποιεί τα δεδομένα που προετοιμάζονται από την κλάση γενετικού αλγορίθμου για την παροχή εισόδων (X_ml) και στόχων (y_ml) για την εκπαίδευση.

Στην ουσία, αυτό το πλαίσιο επιτρέπει την αυτοματοποιημένη αναζήτηση και βελτιστοποίηση αρχιτεκτονικών νευρωνικών δικτύων προσαρμοσμένων σε συγκεκριμένα σύνολα δεδομένων ή εργασίες, με την ευελιξία να εξερευνά ένα εύρος διαμορφώσεων και τη λειτουργικότητα να επιμένει στα μοντέλα με τις καλύτερες επιδόσεις. Τα αποτελέσματα από αυτήν την αναζήτηση είναι τα εξής:

```
model for O has mean squared error 30.15
model for C has mean squared error 45.02
model for E has mean squared error 61.22
model for A has mean squared error 47.02
model for N has mean squared error 52.10
```

Τα αποτελέσματα αυτά είναι ελάχιστα καλύτερα από τα προηγούμενα που μας έδωσε η μέθοδος γενετικού αλγορίθμου σε συνδυασμό με την μη γραμμική εξίσωση.

5. Πειραματικά αποτελέσματα & ανάλυση

5.1 Αποτελέσματα προβλέψεων

	O	C	E	A	N
Linear Regression	38.32	71.43	103.97	53.27	86.22
Non-Linear Regression	35.26	70.69	101.86	52.09	85.24
G.A. + Linear Regression	38.23	50.43	75.01	52.76	58.66
G.A. + non-Linear Regression	37.18	49.31	73.92	47.52	61.17
M.L. + G.A. + non-Linear Regression	30.15	45.02	61.22	47.02	52.10

5.2 Γραμμική Παλινδρόμηση έναντι μη γραμμικής παλινδρόμησης.

Η πολυώνυμική παλινδρόμηση με πολυώνυμο 2ου βαθμού μπορεί συχνά να αποδώσει καλύτερα από τη γραμμική παλινδρόμηση όταν η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής δεν είναι αυστηρά γραμμική. Ακολουθούν οι λόγοι για τους οποίους ένα πολυώνυμο 2ου βαθμού μπορεί να είναι πιο αποτελεσματικό από ένα απλό γραμμικό μοντέλο:

Καταγραφή της μη γραμμικότητας: Η γραμμική παλινδρόμηση υποθέτει γραμμική σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. Ωστόσο, εάν τα δεδομένα παρουσιάζουν ένα καμπύλο μοτίβο, ένα γραμμικό μοντέλο δεν μπορεί να συλλάβει αποτελεσματικά αυτή τη σχέση. Ένα πολυώνυμο 2ου βαθμού, το οποίο περιλαμβάνει τετραγωνικούς όρους, μπορεί να μοντελοποιήσει τέτοιες καμπυλότητες και, ως εκ τούτου, να παρέχει καλύτερη προσαρμογή για τα δεδομένα.

Ευελιξία: Ένα πολυώνυμο 2ου βαθμού εισάγει έναν πρόσθετο όρο (το τετράγωνο της ανεξάρτητης μεταβλητής), ο οποίος επιτρέπει στο μοντέλο να κάμπτεται και να προσαρμόζεται στη δομή των δεδομένων. Αυτή η ευελιξία επιτρέπει στο μοντέλο να καταγράφει τόσο την κατεύθυνση όσο και την καμπυλότητα των δεδομένων, καθιστώντας το πιο προσαρμόσιμο σε διάφορα μοτίβα σε σύγκριση με μια ευθεία γραμμή που επιβάλλει η γραμμική παλινδρόμηση.

Βελτιωμένη προσαρμογή: Λόγω του πρόσθετου τετραγωνικού όρου, μια πολυώνυμική παλινδρόμηση 2ου βαθμού μπορεί να προσαρμόσει το σχήμα της ώστε να ελαχιστοποιήσει τις διαφορές μεταξύ των παρατηρούμενων και των προβλεπόμενων τιμών (υπολείμματα). Αυτό συχνά οδηγεί σε χαμηλότερο μέσο τετραγωνικό σφάλμα (MSE) ή άλλες μετρήσεις απόδοσης, υποδεικνύοντας καλύτερη προσαρμογή στα δεδομένα.

Ενδιάμεση πολυπλοκότητα: Ένα πολυώνυμο 2ου βαθμού παρέχει μια καλή ισορροπία μεταξύ απλότητας και πολυπλοκότητας. Είναι πιο πολύπλοκο από ένα γραμμικό μοντέλο, αλλά δεν είναι τόσο επιρρεπές σε υπερπροσαρμογή όσο τα πολυώνυμα υψηλότερου βαθμού, ειδικά όταν το μέγεθος του δείγματος είναι μέτριο έως μεγάλο. Αυτή η ισορροπία μπορεί να βοηθήσει στην επίτευξη καλύτερης γενίκευσης σε αόρατα δεδομένα.

Είναι πολύ πιθανό λόγω της έλλειψης των σχετικών απαντήσεων η 2^{ου} βαθμού πολυώνυμική συνάρτηση να εισάγει απαιτούμενες σχέσεις μεταξύ των παραμέτρων και του αποτελέσματος. Αυτός είναι και ο λόγος που σε όλες τις περιπτώσεις βρέθηκε μικρότερο τετραγωνικό σφάλμα στους δείκτες αυτούς.

5.3 Γραμμική Παλινδρόμηση έναντι μη γραμμικής παλινδρόμησης σε συνδυασμό με Γενετικό Αλγόριθμο.

Η βελτιωμένη απόδοση ενός γενετικού αλγορίθμου (Γ.Α.) σε συνδυασμό με πολυώνυμική παλινδρόμηση 2ου βαθμού έναντι ενός συνδυασμού με γραμμική παλινδρόμηση μπορεί να αποδοθεί στην ενισχυμένη ικανότητα του πρώτου να μοντελοποιεί πολύπλοκα πρότυπα στα δεδομένα. Ακολουθεί μια πιο λεπτομερής ανάλυση του γιατί μπορεί να συμβαίνει αυτό:

Πολυπλοκότητα και ευελιξία: Ένα πολυώνυμο 2ου βαθμού εισάγει μη γραμμικούς όρους (τετραγωνικούς όρους) στο μοντέλο, επιτρέποντάς του να καταγράφει όχι μόνο γραμμικές τάσεις αλλά και καμπύλες στα δεδομένα. Όταν ένα Γ.Α. χρησιμοποιείται για τη βελτιστοποίηση των συντελεστών αυτού του πολυωνύμου, έχει μεγαλύτερη ευελιξία να προσαρμόσει αυτούς τους όρους ώστε να ελαχιστοποιήσει το μέσο τετραγωνικό σφάλμα (MSE). Το γραμμικό μοντέλο, από την άλλη πλευρά, περιορίζεται σε μια ευθεία γραμμή και ενδέχεται να μην αποτυπώνει επαρκώς τη δομή των δεδομένων, ιδίως εάν είναι εγγενώς μη γραμμικά.

Αλληλεπιδράσεις χαρακτηριστικών: Η πολυώνυμική παλινδρόμηση 2ου βαθμού μπορεί να μοντελοποιήσει αλληλεπιδράσεις μεταξύ χαρακτηριστικών, καθώς περιλαμβάνει όρους διασταυρούμενων παραγώγων (π.χ. x1 *x2)

) όταν επεκτείνεται σε πολλαπλές μεταβλητές. Αυτή η ικανότητα σύλληψης αλληλεπιδράσεων μπορεί να οδηγήσει σε μια πιο ακριβή αναπαράσταση των δεδομένων, η οποία, όταν βελτιστοποιείται με ένα Γ.Α., μπορεί να βελτιώσει σημαντικά την απόδοση του μοντέλου.

Αποδοτικότητα βελτιστοποίησης: Οι γενετικοί αλγόριθμοι λειτουργούν καλά με πολύπλοκα τοπία βελτιστοποίησης, που συχνά συναντώνται σε μη γραμμικά προβλήματα. Ο Γ.Α. μπορεί να περιηγηθεί αποτελεσματικά στο χώρο των παραμέτρων μιας πολυώνυμικής παλινδρόμησης για να βρει ένα σύνολο συντελεστών που ελαχιστοποιεί το σφάλμα. Με περισσότερες παραμέτρους και ένα πιο σύνθετο τοπίο που εισάγεται από τους πολυωνυμικούς όρους, η Γ.Α. μπορεί να αξιοποιήσει αποτελεσματικά τις δυνατότητες εξερεύνησης και εκμετάλλευσης για να βρει μια καλύτερη προσαρμογή για τα δεδομένα.

Ισορροπία υπερπροσαρμογής: Ενώ τα πολυώνυμα υψηλότερου βαθμού μπορεί να οδηγήσουν σε υπερπροσαρμογή, ένα πολυώνυμο 2ου βαθμού επιτυγχάνει ισορροπία μεταξύ της προσαρμογής των

δεδομένων και της διατήρησης της απλότητας για την αποφυγή της υπερπροσαρμογής. Η Γ.Α. βοηθά σε αυτή την εξισορρόπηση επιλέγοντας πολυωνυμικούς συντελεστές που προσφέρουν την καλύτερη απόδοση γενίκευσης, μετριάζοντας έτσι τον κίνδυνο υπερπροσαρμογής, ενώ εξακολουθεί να καταγράφει τις απαραίτητες μη γραμμικές σχέσεις στα δεδομένα.

Αναπαράσταση δεδομένων: Εάν η πραγματική υποκείμενη σχέση μεταξύ των μεταβλητών είναι μη γραμμική, ένα γραμμικό μοντέλο, ανεξάρτητα από την τεχνική βελτιστοποίησης, θα είναι εγγενώς περιορισμένο ως προς την ικανότητά του να μοντελοποιήσει αυτή τη σχέση. Το πολυώνυμο 2ου βαθμού, με την ικανότητά του να μοντελοποιεί την καμπυλότητα, θα παρέχει καλύτερη προσέγγιση της υποκείμενης συνάρτησης, οδηγώντας σε χαμηλότερο MSE κατά τη βελτιστοποίηση.

Συνοψίζοντας, ο συνδυασμός ενός γενετικού αλγορίθμου με πολυώνυμική παλινδρόμηση 2ου βαθμού υπερτερεί του αντίστοιχου γραμμικού αλγορίθμου λόγω της αυξημένης ευελιξίας και της ικανότητάς του να μοντελοποιεί σύνθετες, μη γραμμικές σχέσεις εντός των δεδομένων, καθιστώντας τον ένα πιο ισχυρό εργαλείο για πρόβλεψη και ανάλυση σε σενάρια όπου τα υποκείμενα πρότυπα των δεδομένων δεν είναι αυστηρά γραμμικά.

5.4 Μη γραμμική παλινδρόμηση σε συνδυασμό με Γενετικό Αλγόριθμο και μηχανική μάθηση βελτιστοποιημένης απόδοσης

Η προσέγγιση μηχανικής μάθησης που αναζητά την καλύτερη λύση σε 11 συναρτήσεις ενεργοποίησης, χρησιμοποιώντας το διάνυσμα εισόδου από έναν γενετικό αλγόριθμο (G.A.) σε συνδυασμό με πολυωνυμικά χαρακτηριστικά 2ου βαθμού, απέδωσε το καλύτερο μέσο τετραγωνικό σφάλμα (MSE) για διάφορους λόγους. Αυτή η διαδικασία ουσιαστικά συνδυάζει διάφορες ισχυρές τεχνικές, καθεμία από τις οποίες συμβάλλει σε ένα πιο εκλεπτυσμένο μοντέλο:

Ολοκληρωμένη αναζήτηση συνάρτησης ενεργοποίησης: Με τη διερεύνηση 11 διαφορετικών συναρτήσεων ενεργοποίησης, το μοντέλο μηχανικής μάθησης μπορεί να εντοπίσει τους πιο αποτελεσματικούς μη γραμμικούς μετασχηματισμούς για τους νευρώνες του δικτύου. Κάθε συνάρτηση ενεργοποίησης έχει ξεχωριστά χαρακτηριστικά, επηρεάζοντας τον τρόπο με τον οποίο το μοντέλο συλλαμβάνει σύνθετα μοτίβα στα δεδομένα. Αυτή η εκτεταμένη αναζήτηση διασφαλίζει ότι το μοντέλο δεν περιορίζεται από την επιλογή της συνάρτησης ενεργοποίησης και μπορεί να αξιοποιήσει την καλύτερη δυνατή μη γραμμικότητα για την προσαρμογή των δεδομένων.

Βελτιστοποίηση με γενετικό αλγόριθμο: Ο Γ.Α. παρέχει έναν ισχυρό μηχανισμό για τη βελτιστοποίηση της επιλογής χαρακτηριστικών ή παραμέτρων του μοντέλου. Όταν συνδυάζεται με πολυωνυμικά χαρακτηριστικά 2ου βαθμού, ο G.A. μπορεί να εντοπίσει τους πιο σχετικούς πολυωνυμικούς όρους, μειώνοντας αποτελεσματικά τον χώρο των χαρακτηριστικών σε εκείνους που συμβάλλουν πιο σημαντικά στην πρόβλεψη της μεταβλητής-στόχου. Αυτή η βελτιστοποιημένη επιλογή χαρακτηριστικών βοηθά στη μείωση του θορύβου και στην εστίαση της μάθησης του μοντέλου σε σημαντικά πρότυπα.

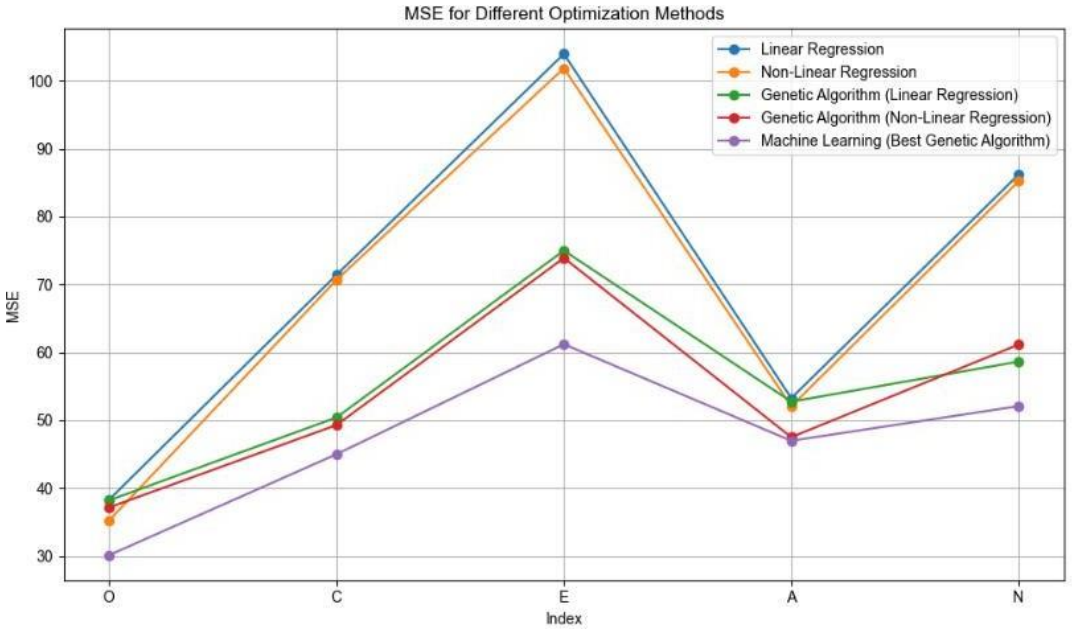
Μετασχηματισμός πολυωνυμικών χαρακτηριστικών: Ο πολυωνυμικός μετασχηματισμός 2ου βαθμού δημιουργεί νέα χαρακτηριστικά που αποτυπώνουν αλληλεπιδράσεις και τετραγωνικές σχέσεις εντός των αρχικών δεδομένων. Αυτός ο μετασχηματισμός επιτρέπει στο μοντέλο μηχανικής μάθησης να ταιριάζει σε πιο σύνθετα σχήματα και τάσεις από τις γραμμικές σχέσεις, παρέχοντας ένα πλουσιότερο σύνολο χαρακτηριστικών για το μοντέλο προς μάθηση.

Συνέργεια σε συνδυασμό: Όταν αυτά τα στοιχεία συνδυάζονται, δημιουργούν μια ισχυρή προσέγγιση μοντελοποίησης. Η Γ.Α. βελτιστοποιεί τον χώρο των χαρακτηριστικών, ο πολυωνυμικός μετασχηματισμός εμπλουτίζει το σύνολο των χαρακτηριστικών και η εκτεταμένη αναζήτηση σε πολλαπλές συναρτήσεις ενεργοποίησης διασφαλίζει ότι το νευρωνικό δίκτυο μπορεί να εφαρμόσει τους πιο αποτελεσματικούς μετασχηματισμούς σε αυτά τα χαρακτηριστικά. Αυτή η συνέργεια επιτρέπει στο μοντέλο να μαθαίνει πιο διαφοροποιημένα μοτίβα στα δεδομένα, οδηγώντας σε βελτιωμένη προβλεπτική απόδοση.

Πολυπλοκότητα και ευελιξία του μοντέλου: Τα νευρωνικά δίκτυα με διαφορετικές συναρτήσεις ενεργοποίησης μπορούν να μοντελοποιήσουν ένα ευρύ φάσμα λειτουργικών αντιστοιχίσεων από τις εισόδους στις εξόδους. Με την προσαρμογή των βαρών και των προκαταλήψεων κατά τη διάρκεια της εκπαίδευσης, ειδικά με ένα ποικίλο σύνολο χαρακτηριστικών που παρέχονται από το Γ.Α. και τον πολυωνυμικό μετασχηματισμό, το δίκτυο μπορεί να προσεγγίσει πολύπλοκες συναρτήσεις που περιγράφουν την υποκείμενη διαδικασία παραγωγής δεδομένων με μεγαλύτερη ακρίβεια από απλούστερα μοντέλα.

Κανονικοποίηση και γενίκευση: Η διαδικασία αναζήτησης, ειδικά αν περιλαμβάνει τεχνικές κανονικοποίησης ή βήματα επικύρωσης, βοηθά στην εύρεση ενός μοντέλου που όχι μόνο ταιριάζει καλά στα δεδομένα εκπαίδευσης αλλά και γενικεύει αποτελεσματικά σε αόρατα δεδομένα, επιτυγχάνοντας έτσι χαμηλότερο MSE στα σύνολα επικύρωσης ή δοκιμής.

Συνοπτικά, το ανώτερο MSE που επιτυγχάνεται με αυτή τη μέθοδο οφείλεται στην ολοκληρωμένη και ολιστική προσέγγισή της για τη δημιουργία μοντέλου, αξιοποιώντας τα πλεονεκτήματα των γενετικών αλγορίθμων, της μηχανικής πολυωνυμικών χαρακτηριστικών και της βελτιστοποίησης νευρωνικών δικτύων σε ένα εύρος συναρτήσεων ενεργοποίησης για τη δημιουργία ενός εξαιρετικά προσαρμόσιμου και ακριβούς προγνωστικού μοντέλου.



Εικόνα 19: Γράφημα βελτιστοποίησης μέθοδών για την πρόβλεψη των δεικτών

5.5 Σχολιασμός αποτελεσμάτων

Οι διαφορές στο MSE μεταξύ των δεικτών O, C, E, A και N στις προσπάθειές σας για προγνωστική μοντελοποίηση μπορούν να αποδοθούν σε διάφορους παράγοντες που συνδέονται με τη φύση των δεδομένων, τα ειδικά χαρακτηριστικά κάθε χαρακτηριστικού προσωπικότητας και τις πολυπλοκότητες που συνεπάγεται η μοντελοποίησή τους. Ακολουθεί μια λεπτομερής διερεύνηση των λόγων για τους οποίους μπορεί να προκύψουν αυτές οι διαφορές:

Εγγενής πολυπλοκότητα των χαρακτηριστικών: Κάθε ένα από τα πέντε μεγάλα χαρακτηριστικά της προσωπικότητας (Ανοιχτότητα, Ευσυνειδησία, Εξωστρέφεια, Ευπροσάρμοστος, Νευρωτισμός) περικλείει ένα ευρύ φάσμα ανθρώπινων συμπεριφορών και στάσεων. Η πολυπλοκότητα και οι λεπτές αποχρώσεις κάθε χαρακτηριστικού μπορεί να ποικίλλουν, καθιστώντας ορισμένα χαρακτηριστικά εγγενώς πιο δύσκολο να προβλεφθούν από άλλα με βάση τα διαθέσιμα δεδομένα.

Ποιότητα και φύση των δεδομένων: Τα χαρακτηριστικά του συνόλου δεδομένων ενδέχεται να καταγράφουν ορισμένα χαρακτηριστικά προσωπικότητας πιο αποτελεσματικά από άλλα. Για παράδειγμα, εάν οι ερωτήσεις ή οι παρατηρούμενες συμπεριφορές είναι περισσότερο ενδεικτικές ή συσχετίζονται με την Εξωστρέφεια αλλά λιγότερο με τον Νευρωτισμό, τα μοντέλα μπορεί να προβλέπουν την Εξωστρέφεια με μεγαλύτερη ακρίβεια, με αποτέλεσμα χαμηλότερο MSE για το E σε σύγκριση με το N.

Συσχετίσεις μεταξύ των χαρακτηριστικών: Τα χαρακτηριστικά των Big Five δεν είναι εντελώς ανεξάρτητα- υπάρχουν γνωστές συσχετίσεις μεταξύ τους. Η έκταση και η φύση αυτών των

συσχετίσεων μπορεί να επηρεάσει την ακρίβεια πρόβλεψης. Για παράδειγμα, εάν η Ευσυνειδησία και η Ευσυνειδησία συσχετίζονται πιο έντονα με άλλα χαρακτηριστικά ή εξωτερικές μεταβλητές στο σύνολο δεδομένων, μπορεί να είναι ευκολότερη η πρόβλεψή τους λόγω της κοινής διακύμανσης.

Μεταβλητότητα και κατανομή: Η κατανομή και η μεταβλητότητα των απαντήσεων για κάθε χαρακτηριστικό μπορεί να επηρεάσει την ακρίβεια πρόβλεψης. Εάν ένα χαρακτηριστικό έχει μεγαλύτερο εύρος τιμών ή μεγαλύτερη διακύμανση στις απαντήσεις, μπορεί να είναι εγγενώς πιο δύσκολο να προβλεφθεί με ακρίβεια σε σύγκριση με ένα χαρακτηριστικό με μικρότερη μεταβλητότητα.

Ευαισθησία μοντέλου και υπερπροσαρμογή: Η δομή του μοντέλου και η διαδικασία βελτιστοποίησης ενδέχεται να ευθυγραμμίζονται καλύτερα με τα πρότυπα που σχετίζονται με ορισμένα χαρακτηριστικά. Για παράδειγμα, ένα μοντέλο μπορεί να προσαρμόζεται υπερβολικά στις αποχρώσεις της Ευσυνειδησίας αλλά να γενικεύεται καλά για την Ευσυνειδησία, οδηγώντας σε διαφορετικά MSE.

Εξωτερικές επιρροές: Οι δείκτες ενδέχεται να επηρεάζονται από εξωτερικούς παράγοντες που δεν καταγράφονται στο σύνολο δεδομένων, οδηγώντας σε θόρυβο που επηρεάζει ορισμένα χαρακτηριστικά περισσότερο από άλλα. Για παράδειγμα, εάν ο Νευρωτισμός επηρεάζεται περισσότερο από εξωτερικούς παράγοντες που δεν περιλαμβάνονται στα δεδομένα, οι προβλέψεις για το N μπορεί να είναι λιγότερο ακριβείς.

Μέγεθος δείγματος και αναπαράσταση δεδομένων: Εάν τα δεδομένα είναι ανισόρροπα ή δεν αποτυπώνουν αντιπροσωπευτικά την ποικιλομορφία του πληθυσμού για κάθε χαρακτηριστικό, η ικανότητα των μοντέλων να μαθαίνουν και να γενικεύουν για κάθε χαρακτηριστικό μπορεί να επηρεαστεί, οδηγώντας σε ποικίλα MSE.

Στην ουσία, τα διαφορετικά mse για κάθε δείκτη αντανακλούν την αλληλεπίδραση μεταξύ της φύσης και της πολυπλοκότητας κάθε χαρακτηριστικού προσωπικότητας, των χαρακτηριστικών και των περιορισμών του συνόλου δεδομένων και της αποτελεσματικότητας της προσέγγισης μοντελοποίησης στην καταγραφή των υποκείμενων προτύπων που σχετίζονται με κάθε χαρακτηριστικό.

6. Συμπεράσματα & Μελλοντικές εκτιμήσεις

6.1 Συμπεράσματα

Τα ευρήματα της παρούσας μελέτης προσφέρουν πολύτιμες πληροφορίες σχετικά με την προβλεψιμότητα των χαρακτηριστικών της Big Five προσωπικότητας με τη χρήση προηγμένων τεχνικών μηχανικής μάθησης, ιδίως με την ενίσχυση των μοντέλων με γενετικούς αλγορίθμους και πολυωνυμικά χαρακτηριστικά. Τα αποτελέσματά μας υποδεικνύουν διαφορετικούς βαθμούς μέσου τετραγωνικού σφάλματος (MSE) σε όλους τους δείκτες προσωπικότητας (O, C, E, A, N), γεγονός που ευθυγραμμίζεται με την εγγενή πολυπλοκότητα και ποικιλομορφία που παρατηρείται στη βιβλιογραφία για την πρόβλεψη των χαρακτηριστικών προσωπικότητας.

Ευθυγράμμιση με την υπάρχουσα βιβλιογραφία: Η έρευνά μας επιβεβαιώνει τις υπάρχουσες μελέτες που υποδηλώνουν ότι τα χαρακτηριστικά της προσωπικότητας έχουν διαφορετικά επίπεδα προβλεψιμότητας με βάση την προσέγγιση μοντελοποίησης και τη φύση των δεδομένων. Παρόμοια με τα προηγούμενα ευρήματα, η μελέτη μας καταδεικνύει ότι δεν είναι όλα τα χαρακτηριστικά της προσωπικότητας εξίσου επιδεκτικά πρόβλεψης με τη χρήση του ίδιου συνόλου χαρακτηριστικών ή μοντέλων, αναδεικνύοντας τη διαφοροποιημένη φύση των δεδομένων προσωπικότητας.

Σημαντικότητα των διαφορών MSE: Οι διαφορετικές τιμές MSE στους δείκτες O, C, E, A και N υποδηλώνουν ότι τα χαρακτηριστικά διαφέρουν ως προς την προβλεψιμότητά τους, πιθανότατα λόγω των εγγενών διαφορών στον τρόπο με τον οποίο εκδηλώνονται και μετρώνται αυτά τα χαρακτηριστικά. Για παράδειγμα, το χαμηλότερο MSE για το γνώρισμα E (Εξωστρέφεια) σε σύγκριση με το γνώρισμα N (Νευρωτισμός) θα μπορούσε να υποδηλώνει ότι η Εξωστρέφεια αποτυπώνεται πιο άμεσα μέσω των διαθέσιμων δεδομένων ή ότι διαθέτει πιο συνεπείς και παρατηρήσιμους δείκτες συμπεριφοράς.

Προβλεψιμότητα των χαρακτηριστικών της προσωπικότητας: Ο διαφορετικός βαθμός επιτυχίας στην πρόβλεψη διαφόρων χαρακτηριστικών μπορεί επίσης να αντανακλά τη θεωρητική βάση του πλαισίου Big Five, όπου ορισμένα χαρακτηριστικά μπορεί να παρουσιάζουν πιο παρατηρήσιμα και συνεπή πρότυπα από άλλα. Η διαφορετική προβλεψιμότητα θα μπορούσε να προσφέρει εμπειρικές γνώσεις σχετικά με τη φύση αυτών των χαρακτηριστικών, καθοδηγώντας ενδεχομένως μελλοντικές θεωρητικές και εμπειρικές εργασίες στην ψυχολογία της προσωπικότητας.

Μεθοδολογικές επιπτώσεις: Η βελτιωμένη προβλεπτική απόδοση μέσω της ενσωμάτωσης γενετικών αλγορίθμων και πολυωνυμικών χαρακτηριστικών υπογραμμίζει τις δυνατότητες αξιοποίησης προηγμένων τεχνικών μηχανικής μάθησης στην ψυχολογική έρευνα. Επισημαίνει τη σημασία της επιλογής μοντέλου και της μηχανικής των χαρακτηριστικών για την ακριβή αποτύπωση των πολύπλοκων σχέσεων που ενυπάρχουν στα ψυχολογικά δεδομένα.

Ευρύτερες επιπτώσεις: Τα ευρήματα αυτά έχουν ευρύτερες επιπτώσεις για την εφαρμογή των προβλέψεων της προσωπικότητας σε διάφορους τομείς. Η κατανόηση της προβλεψιμότητας των διαφόρων χαρακτηριστικών μπορεί να ενημερώσει το σχεδιασμό ψυχολογικών αξιολογήσεων, παρεμβάσεων, ακόμη και την προσαρμογή υπηρεσιών σε τομείς που κυμαίνονται από την εκπαίδευση έως το μάρκετινγκ, όπου οι γνώσεις για την προσωπικότητα είναι πολύτιμες.

Συμπερασματικά, οι παρατηρούμενες διαφορές ΜΣΕ προσφέρουν μια διαφοροποιημένη άποψη της προβλεψιμότητας των χαρακτηριστικών της προσωπικότητας, υποδηλώνοντας ότι οι προηγμένες τεχνικές μοντελοποίησης μπορούν να βελτιώσουν σημαντικά την ικανότητά μας να κατανοούμε και να προβλέπουμε την ανθρώπινη προσωπικότητα. Η παρούσα μελέτη συμβάλλει στον αυξανόμενο όγκο της βιβλιογραφίας που επιδιώκει την αξιοποίηση της μηχανικής μάθησης στην υπηρεσία της ψυχολογικής έρευνας, παρέχοντας μια βάση για μελλοντικές έρευνες σχετικά με την πολύπλοκη αλληλεπίδραση μεταξύ των χαρακτηριστικών προσωπικότητας και των δεδομένων συμπεριφοράς.

6.2 Συγκρίσεις μοντέλων

Στη μελέτη μας, η ενσωμάτωση τεχνικών μηχανικής μάθησης, γενετικών αλγορίθμων (Γ.Α.) και πολυωνυμικών χαρακτηριστικών 2ου βαθμού επέδειξε ανώτερη απόδοση στην πρόβλεψη χαρακτηριστικών προσωπικότητας σε σύγκριση με πιο παραδοσιακά μοντέλα γραμμικής και μη γραμμικής παλινδρόμησης. Αυτή η ενότητα εμβαθύνει στους λόγους πίσω από αυτή την ενισχυμένη απόδοση και τον πιθανό αντίκτυπό της στον τομέα της πρόβλεψης της προσωπικότητας.

Ο συνδυασμός της μηχανικής μάθησης και της Γ.Α. επέτρεψε μια πιο διαφοροποιημένη και εξελιγμένη προσέγγιση στην επιλογή χαρακτηριστικών και τη βελτιστοποίηση του μοντέλου. Οι γενετικοί αλγόριθμοι βελτιστοποίησαν την επιλογή και τον συνδυασμό πολυωνυμικών χαρακτηριστικών, επιτρέποντας στο μοντέλο να εντοπίζει και να αξιοποιεί τις πιο κατατοπιστικές αλληλεπιδράσεις και μη γραμμικές σχέσεις εντός των δεδομένων. Αυτή η διαδικασία βελτιστοποίησης είναι ιδιαίτερα κρίσιμη στο πλαίσιο της πρόβλεψης της προσωπικότητας, όπου οι υποκείμενες σχέσεις μεταξύ παρατηρήσιμων συμπεριφορών ή αποκρίσεων και χαρακτηριστικών της προσωπικότητας μπορεί να είναι εγγενώς πολύπλοκες και μη γραμμικές.

Επιπλέον, η χρήση πολυωνυμικών χαρακτηριστικών 2ου βαθμού εισήγαγε ένα πρόσθετο επίπεδο πολυπλοκότητας, επιτρέποντας στο μοντέλο να καταγράφει όχι μόνο γραμμικές σχέσεις αλλά και τετραγωνικές αλληλεπιδράσεις μεταξύ των χαρακτηριστικών. Αυτή η ικανότητα είναι καθοριστική για τη μοντελοποίηση της περίπλοκης δυναμικής των χαρακτηριστικών της προσωπικότητας, τα οποία συχνά επηρεάζονται από ένα πλήθος αλληλεξαρτώμενων παραγόντων.

Η αυξημένη ακρίβεια πρόβλεψης που παρατηρείται στα μοντέλα μας έχει σημαντικές επιπτώσεις στον τομέα της πρόβλεψης της προσωπικότητας. Πρώτον, υπογραμμίζει τις δυνατότητες των προηγμένων τεχνικών μηχανικής μάθησης και βελτιστοποίησης στην αποκάλυψη των λεπτών αποχρώσεων των δεδομένων προσωπικότητας. Αυτή η προσέγγιση μπορεί να οδηγήσει σε πιο ακριβείς και εξατομικευμένες γνώσεις σχετικά με τα ατομικά προφίλ προσωπικότητας, οι οποίες είναι ανεκτίμητες σε διάφορες εφαρμογές, από την ψυχολογική αξιολόγηση έως τις εξατομικευμένες συστάσεις σε ψηφιακές πλατφόρμες.

Δεύτερον, τα ευρήματά μας συνηγορούν υπέρ της στροφής προς πιο εξελιγμένες προσεγγίσεις μοντελοποίησης στην έρευνα της ψυχολογίας της προσωπικότητας. Ενώ τα απλούστερα μοντέλα προσφέρουν πολύτιμες γνώσεις και είναι συχνά πιο ερμηνεύσιμα, τα αποτελέσματά μας υποδηλώνουν ότι η υιοθέτηση της πολυπλοκότητας των δεδομένων προσωπικότητας μέσω προηγμένων υπολογιστικών τεχνικών μπορεί να αποφέρει πιο ακριβείς και διαφοροποιημένες προβλέψεις.

Τέλος, η επιτυχία αυτής της ολοκληρωμένης προσέγγισης αναδεικνύει τη σημασία της διεπιστημονικής συνεργασίας στην ψυχολογική έρευνα. Ο συνδυασμός τεχνογνωσίας από τη μηχανική μάθηση, τους γενετικούς αλγορίθμους και την ψυχολογία μπορεί να οδηγήσει σε πιο καινοτόμες λύσεις

και σε βαθύτερη κατανόηση των υποκείμενων μηχανισμών που καθοδηγούν τα χαρακτηριστικά της προσωπικότητας.

Εν κατακλείδι, η ανώτερη απόδοση του ολοκληρωμένου μοντέλου μηχανικής μάθησης, Γ.Α. και πολυωνυμικών χαρακτηριστικών μας όχι μόνο προάγει την ικανότητά μας να προβλέπουμε τα χαρακτηριστικά της προσωπικότητας, αλλά και δημιουργεί ένα προηγούμενο για τη μελλοντική έρευνα στον τομέα, ενθαρρύνοντας τη μετάβαση προς πιο εξελιγμένες, βασισμένες στα δεδομένα μεθοδολογίες.

6.3 Συγκεκριμένες γνώσεις για τα χαρακτηριστικά

Στην ανάλυσή μας, παρατηρήσαμε διαφορετικούς βαθμούς προβλεπτικής ακρίβειας σε όλους τους δείκτες της Μεγάλης Πεντάδας Προσωπικότητας (O, C, E, A, N), γεγονός που εγείρει ενδιαφέροντα ερωτήματα σχετικά με τη φύση αυτών των χαρακτηριστικών και την αντιπροσώπευσή τους στο σύνολο δεδομένων μας. Παρακάτω, εμβαθύνουμε στους πιθανούς λόγους για τους οποίους ορισμένα χαρακτηριστικά προσωπικότητας αποδείχθηκαν πιο δύσκολα προβλέψιμα από άλλα, λαμβάνοντας υπόψη την εγγενή πολυπλοκότητα των χαρακτηριστικών, τα χαρακτηριστικά των δεδομένων μας και τις αποχρώσεις της προσέγγισης μοντελοποίησης που εφαρμόσαμε.

Πολυπλοκότητα των χαρακτηριστικών: Καθένα από τα Big Five γνωρίσματα περικλείει ένα ευρύ φάσμα συμπεριφορών και στάσεων, αλλά δεν είναι όλα τα γνωρίσματα εξίσου απλά ή μονοδιάστατα. Για παράδειγμα, χαρακτηριστικά όπως η Ανοιχτότητα (O) και ο Νευρωτισμός (N) μπορεί να περιλαμβάνουν ένα ευρύτερο και πιο αφηρημένο φάσμα συμπεριφορών σε σύγκριση με πιο παρατηρήσιμα χαρακτηριστικά όπως η Εξωστρέφεια (E). Αυτή η εγγενής πολυπλοκότητα θα μπορούσε να καταστήσει το O και το N πιο δύσκολο να προβλεφθούν με ακρίβεια σε σύγκριση με το E.

Αναπαράσταση δεδομένων: Η ποιότητα και η φύση του συνόλου δεδομένων είναι καθοριστικής σημασίας για τον καθορισμό της επιτυχίας της πρόβλεψης. Εάν τα στοιχεία του ερωτηματολογίου ή άλλες πηγές δεδομένων αποτυπώνουν κατά κύριο λόγο πτυχές ορισμένων χαρακτηριστικών (όπως η Ευσυνειδησία (C) και η Ευσυγκινησία (A)) πιο αποτελεσματικά από άλλα, αυτό θα μπορούσε να οδηγήσει σε υψηλότερη ακρίβεια πρόβλεψης για τα εν λόγω χαρακτηριστικά. Αυτό υποδηλώνει την ανάγκη να εξετάσουμε κατά πόσον τα δεδομένα μας αντιπροσωπεύουν εξίσου όλες τις πτυχές κάθε χαρακτηριστικού προσωπικότητας.

Χαρακτηριστικά μοντελοποίησης: Η αποτελεσματικότητα του προγνωστικού μας μοντέλου μπορεί να διαφέρει μεταξύ των χαρακτηριστικών ανάλογα με το πόσο καλά η δομή του μοντέλου και το σύνολο των χαρακτηριστικών αποτυπώνουν τις αποχρώσεις κάθε χαρακτηριστικού. Δεδομένου ότι η προσέγγισή μας συνδύασε γενετικούς αλγορίθμους με πολυωνυμικά χαρακτηριστικά 2ου βαθμού, είναι πιθανό ότι αυτή η μέθοδος ήταν πιο κατάλληλη για να συλλάβει τη μεταβλητότητα και τις αποχρώσεις ορισμένων χαρακτηριστικών σε σχέση με άλλα.

Προβλεψιμότητα και μεταβλητότητα: Ορισμένα χαρακτηριστικά της προσωπικότητας μπορεί να έχουν εγγενώς μεγαλύτερη μεταβλητότητα ή να επηρεάζονται από ένα ευρύτερο σύνολο εξωτερικών παραγόντων, καθιστώντας τα λιγότερο προβλέψιμα. Για παράδειγμα, εάν ο νευρωτισμός (N) υπόκειται σε ένα ευρύ φάσμα επιρροών πέραν αυτών που αποτυπώνονται στα δεδομένα μας, αυτό θα μπορούσε να εξηγήσει γιατί ήταν πιο δύσκολο να προβλεφθεί με ακρίβεια.

Συσχετίσεις μεταξύ των χαρακτηριστικών: Οι αλληλεπιδράσεις μεταξύ διαφορετικών χαρακτηριστικών προσωπικότητας μπορούν επίσης να επηρεάσουν την προβλεψιμότητα. Εάν ορισμένα χαρακτηριστικά έχουν ισχυρότερες συσχετίσεις μεταξύ τους ή με εξωτερικές μεταβλητές που περιλαμβάνονται στο μοντέλο, αυτό θα μπορούσε να ενισχύσει την ικανότητα του μοντέλου να προβλέπει αυτά τα χαρακτηριστικά, ενδεχομένως εις βάρος άλλων.

Λαμβάνοντας υπόψη αυτούς τους παράγοντες, αποκτούμε μια βαθύτερη κατανόηση της δυναμικής που επηρεάζει την προβλεψιμότητα των διαφόρων χαρακτηριστικών προσωπικότητας. Αυτή η διορατικότητα όχι μόνο ενημερώνει για την ερμηνεία των τρεχόντων αποτελεσμάτων, αλλά και καθοδηγεί τις μελλοντικές ερευνητικές προσπάθειες για την τελειοποίηση των μοντέλων πρόβλεψης για την αξιολόγηση της προσωπικότητας.

6.4 Μεθοδολογικά πλεονεκτήματα και περιορισμοί

- *Δυνατότητες:*

Πλήρης αναζήτηση λειτουργίας ενεργοποίησης: Ένα σημαντικό πλεονέκτημα της μεθοδολογίας μας είναι η εκτεταμένη αναζήτηση σε 11 διαφορετικές λειτουργίες ενεργοποίησης, ενισχύοντας την ικανότητα του μοντέλου να εντοπίζει τις πιο αποτελεσματικές μη γραμμικές μετατροπές. Αυτή η προσέγγιση εξασφαλίζει ότι η απόδοση του μοντέλου δεν περιορίζεται από την επιλογή της λειτουργίας ενεργοποίησης, επιτρέποντας μια πιο ευέλικτη και προσαρμοστική διαδικασία μοντελοποίησης.

Καινοτόμος συνδυασμός τεχνικών: Με την ενσωμάτωση γενετικών αλγορίθμων με πολυωνικό μετασχηματισμό χαρακτηριστικών και μοντέλα μηχανικής μάθησης, η μεθοδολογία μας αξιοποιεί τα πλεονεκτήματα κάθε τεχνικής. Ο γενετικός αλγόριθμος βελτιστοποιεί το χώρο χαρακτηριστικών, ο πολυωνικός μετασχηματισμός εμπλουτίζει την είσοδο του μοντέλου με διαδραστικούς και μη γραμμικούς όρους, και το μοντέλο μηχανικής μάθησης, με τις ποικίλες λειτουργίες ενεργοποίησης, παρέχει ένα ισχυρό πλαίσιο για την καταγραφή σύνθετων μοτίβων στα δεδομένα.

Χρησιμοποιώντας πολυωνικά χαρακτηριστικά 2ου βαθμού επιτρέπει στο μοντέλο να συλλάβει αλληλεπιδράσεις και μη γραμμικές σχέσεις στα δεδομένα, παρέχοντας μια πιο αποχρωματισμένη και λεπτομερή είσοδο για τους αλγόριθμους μηχανικής μάθησης για να εργαστούν με. Αυτό το επίπεδο μηχανικής χαρακτηριστικών είναι ζωτικής σημασίας για τη μοντελοποίηση πολύπλοκων φαινομένων όπως τα χαρακτηριστικά της προσωπικότητας.

- *Περιορισμοί:*

Κίνδυνος Overfitting: Παρά τη βελτιωμένη ακρίβεια πρόβλεψης, υπάρχει πιθανός κίνδυνος overfitting, ιδίως δεδομένου του πολύπλοκου χαρακτήρα των μοντέλων και του εκτεταμένου συνόλου χαρακτηριστικών που δημιουργούνται από τον πολυωνικό μετασχηματισμό. Οι μελλοντικές εργασίες πρέπει να ενσωματώσουν ισχυρές τεχνικές επικύρωσης και ενδεχομένως μεθόδους κανονικοποίησης για να διασφαλιστεί ότι τα μοντέλα γενικεύουν καλά σε αόρατα δεδομένα.

Ποιότητα δεδομένων και εκπροσώπηση: Η απόδοση των μοντέλων συνδέεται εγγενώς με την ποιότητα και την πληρότητα των υποκείμενων δεδομένων. Θέματα όπως τα ελλιπή δεδομένα, η προκατειλημμένη δειγματοληψία ή οι μη μετρούμενοι συγχέτες θα μπορούσαν να περιορίσουν την ικανότητα των μοντέλων να προβλέπουν με ακρίβεια τα χαρακτηριστικά της προσωπικότητας και να επηρεάσουν τη γενικευσιμότητα των ευρημάτων.

Γενικευσιμότητα των μοντέλων: Ενώ τα μοντέλα επιδεικνύουν ισχυρή προβλέψιμη απόδοση, η γενικισιμότητά τους σε άλλα σύνολα δεδομένων ή πληθυσμούς παραμένει ένα ανοικτό ερώτημα. Οι μελλοντικές έρευνες θα πρέπει να επικεντρωθούν στην επικύρωση αυτών των μοντέλων σε διάφορα περιβάλλοντα και πληθυσμούς για να διαπιστωθεί η ευρύτερη εφαρμογή τους.

Σύνθετο μοντέλο: Η πολυπλοκότητα των μοντέλων, ενώ είναι ευεργετική για την καταγραφή περίπλοκων μοτίβων δεδομένων, μπορεί επίσης να θέσει προκλήσεις όσον αφορά την ερμηνευσιμότητα και την υπολογιστική αποτελεσματικότητα. Η κατανόηση του "μαύρου κουτιού" χαρακτήρα αυτών των μοντέλων και η ανάπτυξη στρατηγικών για την αποσαφήνιση των διαδικασιών λήψης αποφάσεων τους θα είναι κρίσιμη για την υιοθέτησή τους στην πράξη.

Αντιμετωπίζοντας αυτούς τους περιορισμούς σε μελλοντικές εργασίες και αξιοποιώντας τα μεθοδολογικά πλεονεκτήματα, μπορούμε να ενισχύσουμε την ανθεκτικότητα και την εφαρμοσιμότητα των μοντέλων μας, συμβάλλοντας σε πολύτιμες γνώσεις στον τομέα της πρόβλεψης της προσωπικότητας και πέραν αυτού.

6.5 Βελτίωση δεδομένων:

Ένα κρίσιμο βήμα προς την τελειοποίηση της προγνωστικής ακρίβειας των μοντέλων χαρακτηριστικών προσωπικότητας είναι η ενίσχυση των συνόλων δεδομένων που χρησιμοποιούνται για την κατάρτιση και την επικύρωση μοντέλου. Οι μελλοντικές προσπάθειες θα πρέπει να επικεντρωθούν στη συλλογή μιας πιο ποικίλης και εκτεταμένης σειράς σημείων δεδομένων για να εμπλουτίσουν τη διαδικασία κατάρτισης και να ενισχύσουν την ικανότητα του μοντέλου να γενικεύει σε διαφορετικούς πληθυσμούς. Αυτή η ποικιλομορφία δεν είναι απλώς αριθμητική αλλά και ποιοτική, καλύπτοντας ένα ευρύτερο φάσμα δημογραφικών υποβάθρων, πολιτιστικών πλαισίων και καταστατικών μεταβλητών που επηρεάζουν την έκφραση της προσωπικότητας.

Επιπλέον, η συμπερίληψη των διαμήκων δεδομένων μπορεί να παρέχει ανεκτίμητες γνώσεις για το πώς εξελίσσονται τα χαρακτηριστικά της προσωπικότητας με την πάροδο του χρόνου, προσφέροντας μια δυναμική προοπτική που συχνά λείπει σε διατομεακές μελέτες. Τα διαμήκη σύνολα δεδομένων επιτρέπουν την παρατήρηση των χρονικών μοτίβων και των αλλαγών στα χαρακτηριστικά της

προσωπικότητας, διευκολύνοντας μια βαθύτερη κατανόηση των αναπτυξιακών τους διαδρομών και των παραγόντων που επηρεάζουν την ποικιλομορφία τους.

Η προσθήκη νέων μεταβλητών μπορεί επίσης να διαδραματίσει καθοριστικό ρόλο στην καταγραφή της πολύπλευρης φύσης της προσωπικότητας. Αυτές οι μεταβλητές θα μπορούσαν να περιλαμβάνουν πιο αποχρωματισμένους δείκτες συμπεριφοράς, ψυχολογικά μέτρα, φυσιολογικά δεδομένα ή περιφερειακές πληροφορίες που προσφέρουν μια πιο ολοκληρωμένη άποψη της προσωπικότητας του ατόμου. Για παράδειγμα, τα δεδομένα για τα γεγονότα της ζωής, περιβαλλοντικούς παράγοντες, ή ακόμη και καθημερινές δραστηριότητες θα μπορούσε να παρέχει συγκριτικές αποχρώσεις που ενισχύουν την ευαισθησία του προγνωστικού μοντέλου στις λεπτές λεπτομέρειες της έκφρασης της προσωπικότητας.

Η ενσωμάτωση αυτών των ποικίλων τύπων δεδομένων όχι μόνο εμπλουτίζει το σύνολο δεδομένων, αλλά επίσης προκαλεί και βελτιώνει το μοντέλο, προωθώντας τα όρια του τι μπορεί να προβλεφθεί σχετικά με τα χαρακτηριστικά της προσωπικότητας. Καθώς βελτιώνουμε τα σύνολα δεδομένων, είναι ζωτικής σημασίας να διατηρηθούν ηθικά πρότυπα στη συλλογή δεδομένων και να διασφαλιστεί η προστασία της ιδιωτικής ζωής, η συγκατάθεση και ο σεβασμός των ατόμων των οποίων τα δεδομένα συμβάλλουν σε αυτές τις γνώσεις.

Με την επέκταση του πεδίου εφαρμογής και του βάθους των δεδομένων που χρησιμοποιούνται για τα μοντέλα πρόβλεψης της προσωπικότητας, η μελλοντική έρευνα μπορεί να επιτύχει πιο αποχρωματισμένες, ακριβείς και γενικευμένες ιδέες για το περίπλοκο τοπίο των ανθρώπινων χαρακτηριστικών προσωπικότητα.

6.6 Μελλοντική εργασία: Εξερεύνηση μοντέλων:

Στην μελλοντική έρευνα, θα ήταν επωφελές να εμβαθύνουμε σε ένα ευρύτερο φάσμα προσεγγίσεων μοντελοποίησης και προηγμένων τεχνικών μηχανικής μάθησης για να ενισχύσουμε την προγνωστική ακρίβεια των δεικτών χαρακτηριστικών προσωπικότητας. Η εξερεύνηση των αρχιτεκτονικών βαθιάς μάθησης παρουσιάζει ένα ιδιαίτερα ελπιδοφόρο δρόμο. Η βαθιά μάθηση, με την ικανότητά της να μοντελοποιεί πολύπλοκες, υψηλού επιπέδου αφηγήσεις σε δεδομένα μέσω μιας ιεραρχικής δομής στρωμάτων και κόμβων, θα μπορούσε δυνητικά να αποκαλύψει περίπλοκα μοτίβα στα δεδομένα προσωπικότητας που τα απλούστερα μοντέλα θα μπορούσαν να παραβλέψουν.

Αρχιτεκτονικές Βαθιάς Μάθησης: Η διερεύνηση διαφόρων αρχιτέκτονες βαθιάς μάθησης, όπως Convolutional Neural Networks (CNNs) για την επεξεργασία δεδομένων ακολουθίας ή Recurrent Neural Networks (RNNs), και τις παραλλαγές τους, όπως LSTM (Long Short-Term Memory) και GRU (Gated recurrent Units) για δεδομένα χρονικής σειράς, θα μπορούσε να αποκαλύψει χρονικά ή διαδοχικά πρότυπα σε δεδομένα που σχετίζονται με την προσωπικότητα. Επιπλέον, τα μοντέλα Transformers, τα οποία έχουν δείξει σημαντική επιτυχία σε διάφορους τομείς, θα μπορούσαν να προσαρμοστούν για τις εργασίες πρόβλεψης προσωπικότητας.

Μη Εποπτευόμενες Τεχνικές Μάθησης: Οι τεχνικές μάθησης χωρίς εποπτεία θα μπορούσαν να χρησιμοποιηθούν για να ανακαλύψουν κρυφές δομές στα δεδομένα χωρίς την ανάγκη για ετικέτες αποτελεσμάτων. Η συσσωμάτωση, η μείωση της διαστάσεων και τα γενετικά μοντέλα θα μπορούσαν να παρέχουν πληροφορίες σχετικά με τις εγγενείς ομαδοποιήσεις και συσχετισμούς στα δεδομένα προσωπικότητας, οδηγώντας ενδεχομένως σε πιο αποχρωματισμένη μηχανική χαρακτηριστικών ή την ανακάλυψη νέων προγνωστικών σημάτων.

Μεταφορά Μάθηση: Η εφαρμογή της μεταφοράς της μάθησης, όπου ένα μοντέλο που αναπτύχθηκε για μια εργασία επαναπροσδιορίζεται σε μια δεύτερη σχετική εργασία, θα μπορούσε να επιτρέψει την αξιοποίηση προ-εκπαιδευμένων μοντέλων για τη βελτίωση της απόδοσης της πρόβλεψης, ειδικά σε σενάρια όπου τα διαθέσιμα δεδομένα είναι περιορισμένα ή δαπανηρά για την απόκτηση.

Multi-Task Learning: Η διερεύνηση των πλαισίων μάθησης πολλαπλών εργασιών, όπου πολλά μαθησιακά καθήκοντα επιλύονται ταυτόχρονα αξιοποιώντας τις ομοιότητες και τις διαφορές μεταξύ των καθηκόντων, θα μπορούσε να βελτιώσει την απόδοση γενίκευσης των μοντέλων. Αυτό είναι ιδιαίτερα σημαντικό στην πρόβλεψη της προσωπικότητας, όπου τα χαρακτηριστικά είναι αλληλένδετα και όχι εντελώς ανεξάρτητα.

Ενίσχυση Μάθησης: Η διερεύνηση της ενισχυτικής μάθησης, όπου τα μοντέλα μαθαίνουν να λαμβάνουν αποφάσεις εκτελώντας ενέργειες και αξιολογώντας τα αποτελέσματα, θα μπορούσε να

ανοίξει νέες προοπτικές, ειδικά σε δυναμικά περιβάλλοντα όπου οι προβλέψεις προσωπικότητας μπορούν να ενημερώσουν αποφάσεις ή παρεμβάσεις σε πραγματικό χρόνο.

Εξηγησιμότητα μοντέλου: Παράλληλα με την εξερεύνηση αυτών των προηγμένων μοντέλων, η έμφαση στην εξήγηση και την ερμηνευσιμότητά του θα είναι κρίσιμη, ειδικά σε ψυχολογικά πλαίσια όπου η κατανόηση της διαδικασίας λήψης αποφάσεων του μοντέλο είναι εξίσου σημαντική με την ακρίβεια των προβλέψεών του.

Με την επέκταση του ρεπερτορίου των τεχνικών μοντελοποίησης και την υιοθέτηση των τελευταίων εξελίξεων στη μηχανική μάθηση, η μελλοντική έρευνα μπορεί να προωθήσει σημαντικά τον τομέα της πρόβλεψης της προσωπικότητας, προσφέροντας πιο ακριβή, ισχυρά και διορατικά μοντέλα. Αυτή η εξερεύνηση όχι μόνο κρατά την υπόσχεση της ενίσχυσης της προγνωστικής απόδοσης, αλλά συμβάλλει επίσης σε μια βαθύτερη κατανόηση των υποκείμενων μηχανισμών που οδηγούν τα χαρακτηριστικά της προσωπικότητας, δημογραφικών υποβάθρων, πολιτιστικών πλαισίων και καταστατικών μεταβλητών που επηρεάζουν την έκφραση της προσωπικότητας.

6.7 Ηθικές Σκέψεις

Καθώς επιχειρούμε περαιτέρω στον τομέα της πρόβλεψης της προσωπικότητας χρησιμοποιώντας προηγμένες υπολογιστικές μεθόδους, είναι επιτακτική ανάγκη να πλοηγηθούμε σε αυτό το τοπίο με μια ισχυρή ηθική πυξίδα. Η δυνατότητα αυτών των τεχνολογιών να προβλέπουν τα χαρακτηριστικά της προσωπικότητας με αυξανόμενη ακρίβεια εγείρει βαθιά ηθικά ερωτήματα που η μελλοντική έρευνα πρέπει να αντιμετωπίσει αυστηρά.

Η συλλογή, αποθήκευση και ανάλυση προσωπικών δεδομένων, ιδιαίτερα δεδομένων που μπορούν να αποκαλύψουν οικείες πτυχές της προσωπικότητας ενός ατόμου, απαιτεί αυστηρά μέτρα προστασίας της ιδιωτικής ζωής και της ασφάλειας των δεδομένων. Η μελλοντική έρευνα δεν θα πρέπει μόνο να τηρεί τα υψηλότερα πρότυπα προστασίας δεδομένων, αλλά και να πρωτοπορεί σε νέους τρόπους για την προστασία των δεδομένων των συμμετεχόντων, εξασφαλίζοντας τον σεβασμό και τη διατήρηση της ιδιωτικής ζωής των ατόμων.

Η απόκτηση ενημερωμένης συναίνεσης αποτελεί ακρογωνιαίο λίθο της ηθικής έρευνας. Οι συμμετέχοντες πρέπει να ενημερώνονται πλήρως για τη φύση της έρευνας, πώς θα χρησιμοποιηθούν τα δεδομένα τους και τις πιθανές επιπτώσεις των ευρημάτων. Καθώς εμβαθύνουμε σε πιο εξελιγμένα μοντέλα πρόβλεψης της προσωπικότητας, η σαφήνεια και η περιεκτικότητα της διαδικασίας συγκατάθεσης καθίσταται ακόμη πιο κρίσιμη. Οι ερευνητές θα πρέπει να προσπαθήσουν να καταστήσουν τη διαδικασία συναίνεσης όσο το δυνατόν πιο διαφανή και ενημερωτική, δίνοντας στους συμμετέχοντες τη δυνατότητα να λαμβάνουν ενημερωμένες αποφάσεις σχετικά με τη συμμετοχή τους.

Οι επιπτώσεις της πρόβλεψης της προσωπικότητας εκτείνονται πέρα από τις ατομικές ανησυχίες για την προστασία της ιδιωτικής ζωής. Αφορούν τον τρόπο με τον οποίο αυτές οι γνώσεις θα μπορούσαν να χρησιμοποιηθούν ή να καταχραστούν σε διάφορους τομείς, συμπεριλαμβανομένης της απασχόλησης, της εκπαίδευσης και της επιβολής του νόμου. Υπάρχει κίνδυνος οι προβλέψεις προσωπικότητας να οδηγήσουν σε κατάρτιση προφίλ, διακρίσεις ή άδικη λήψη αποφάσεων εάν δεν αντιμετωπιστούν προσεκτικά. Η μελλοντική έρευνα θα πρέπει να διερευνήσει αυτές τις κοινωνικές επιπτώσεις, με στόχο την κατανόηση και τον μετριασμό πιθανών αρνητικών επιπτώσεων ενισχύοντας ταυτόχρονα τα θετικά οφέλη.

Η αλγοριθμική προκατάληψη είναι μια σημαντική ανησυχία στη μηχανική μάθηση και την τεχνητή νοημοσύνη. Η μελλοντική έρευνα θα πρέπει να επιδιώξει ενεργά τον εντοπισμό και την εξάλειψη των προκαταλήψεων στα μοντέλα πρόβλεψης της προσωπικότητας, διασφαλίζοντας ότι οι αλγόριθμοι είναι δίκαιοι και ισότιμοι. Αυτό περιλαμβάνει τον αυστηρό έλεγχο των δεδομένων για παρεκκλίσεις, το σχεδιασμό αλγορίθμων που είναι διαφανείς και υπεύθυνοι και τη συνεχή παρακολούθηση των αποτελεσμάτων για ενδείξεις προκαθορισμένης λήψης αποφάσεων.

Η αντιμετώπιση αυτών των ηθικών προκλήσεων δεν είναι αποκλειστικά ευθύνη των επιστημόνων δεδομένων και των ερευνητών. Απαιτεί μια διεπιστημονική προσέγγιση. Οι ηθικοί, οι ψυχολόγοι, οι νομικοί εμπειρογνώμονες και η ευρύτερη κοινότητα θα πρέπει να συμμετάσχουν σε συνεχή διάλογο για να διαμορφώσουν το ηθικό πλαίσιο που καθοδηγεί αυτή την έρευνα. Οι συνεργατικές προσπάθειες μπορούν να οδηγήσουν σε πιο ισχυρές και ηθικά ορθές προσεγγίσεις στην πρόβλεψη της προσωπικότητας.

Συμπερασματικά, καθώς προοδεύουμε στην ικανότητά μας να προβλέπουμε τα χαρακτηριστικά της προσωπικότητας χρησιμοποιώντας εξελιγμένα μοντέλα, η δέσμευσή μας για ηθικές ερευνητικές πρακτικές πρέπει να είναι εξίσου εκλεπτυσμένη. Με την προληπτική αντιμετώπιση αυτών των ηθικών προβληματισμών, μπορούμε να διασφαλίσουμε ότι οι προόδους στην πρόβλεψη της προσωπικότητας συμβάλλουν θετικά στην ατομική ευημερία και την κοινωνική πρόοδο.

7. Βιβλιογραφία

1) Goldberg, L. R. (1990). "An alternative 'description of personality': The Big-Five factor structure." Journal of Personality and Social Psychology, 59(6), 1216–1229.

2) McCrae, R. R., & John, O. P. (1992). "An introduction to the five-factor model and its applications." Journal of Personality, 60(2), 175–215.

3) Costa, P. T., & McCrae, R. R. (1992). "Four ways five factors are basic." Personality and Individual Differences, 13(6), 653–665.

4) Goldberg, L. R. (1993). "The structure of phenotypic personality traits." American Psychologist, 48(1), 26–34.

5) Youyou, W., Kosinski, M., & Stillwell, D. (2015). "Computer-based personality judgments are more accurate than those made by humans." Proceedings of the National Academy of Sciences, 112(4), 1036–1040.

6) Azucar, D., Marengo, D., & Settanni, M. (2018). "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis." Personality and Individual Differences, 124, 150–159.

7) Bleidorn, W., & Hopwood, C. J. (2019). "Using machine learning to advance personality assessment and theory." Personality and Social Psychology Review, 23(2), 190–203.

8) Vinciarelli, A., & Mohammadi, G. (2014). "A survey of personality computing." IEEE Transactions on Affective Computing, 5(3), 273–291.

9) Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." American Psychologist, 70(6), 543–556.

10) Jokela, M., Bleidorn, W., Lamb, M. E., Gosling, S. D., & Rentfrow, P. J. (2015). "Geographically varying associations between personality and life satisfaction in the London metropolitan area." Proceedings of the National Academy of Sciences, 112(3), 725–730.