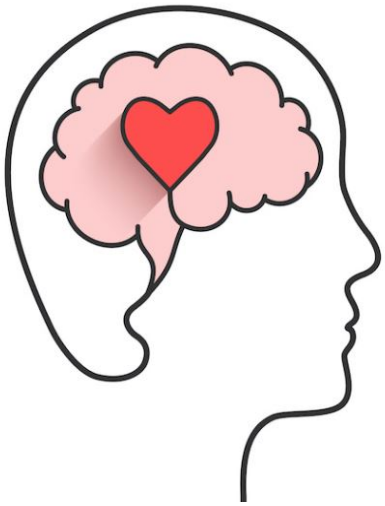
Ανάλυση Δεδομένων Κατάθλιψης εν μέσω πανδημίας, με μεθόδους Μηχανικής και Βαθιάς Μάθησης



Εργασία στο μάθημα "Ανάλυση Βιο-δεδομένων"

Παναγιώτης Κοκκινάκος	03400092
Ιωάννα Μανδηλαρά	03400096
Φίλιππος Σκόβελεφ Ορφανουδάκης	03400107
Άρης Σπύρου	03400113

August 30, 2021

Περίληψη

Η έξαρση της πανδημίας του Covid-19 είχε σημαντικό αντίκτυπο στην ψυχική υγεία των ανθρώπων και ειδικά των φοιτητών, οι οποίοι προσαρμόστηκαν σε νέες συνθήκες στο πανεπιστημιακό τους περιβάλλον. Για τον λόγο αυτό, στα πλαίσια αυτής της εργασίας επεξεργαζόμαστε δεδομένα που προέρχονται από την παγκόσμια μελέτη C19 ISWS, η οποία διεξήχθη σε μορφή ερωτηματολογίου, και διαθέτει ερωτήσεις σχετικά με την ψυχική κατάσταση των φοιτητών.

Πρώτο βήμα είναι ο καθαρισμός των δεδομένων και στη συνέχεια η εφαρμογή στατιστικών ελέγχων στα δεδομένα, έτσι ώστε να εξάγουμε συμπεράσματα διαφορών και εξαρτήσεων των ερωτήσεων συγκριτικά με τα διαφορετικά επίπεδα κατάθλιψης.

Στη συνέχεια με μεθόδους μη επιβλεπόμενης μάθησης, ανιχνεύουμε πιθανά μοτίβα που μπορεί να υπάρχουν σε ανθρώπους που πάσχουν από κατάθλιψη. Τα μοτίβα αυτά μπορεί να σχετίζονται με την καθημερινότητα τους, τις αλλαγές που πραγματοποίησαν στις συνήθειες τους ή με το βαθμό ικανοποίησης που τους προσέφεραν τα μέτρα αντιμετώπισης της κατάστασης εντός του πανεπιστημίου.

Τέλος αντιμετωπίζουμε τα δεδομένα, σαν ένα πρόβλημα επιβλεπόμενης μάθησης και πιο συγκεκριμένα παλινδρόμησης και ταξινόμησης. Πρώτη προσέγγιση είναι η εφαρμογή κλασσικών μεθόδων μηχανικής μάθησης για την πρόβλεψη τις ακριβούς τιμής της κατάθλιψης, και η δεύτερη προσέγγιση είναι η μετατροπή του προβλήματος σε ταξινόμηση 2 ή 3 κλάσεων με τη χρήση τεχνικών βαθιάς μάθησης.

Περιεχόμενα

1 Εισαγωγή

Η συγκεκριμένη εργασία στηρίζεται στην έρευνα COVID-19 International Student Well-being Study (C19 ISWS) [1], μια μεγάλη διαδικτυακή έρευνα η οποία συνέλεξε δεδομένα από 26 διαφορετικές χώρες και 110 Ανώτατα εκπαιδευτικά ιδρύματα.

Στο πλαίσιο της έρευνας, απαντήθηκε από τους φοιτητές που συμμετείχαν ένα ερωτηματολόγιο με ερωτήσεις που σχετίζονται με τις συνθήκες διαβίωσης των συμμετεχόντων, την οικονομική τους κατάσταση και τις αλλαγές που έγιναν στον τρόπο εκπαίδευσης από τα εκπαιδευτικά ιδρύματα για την αντιμετώπιση της πανδημίας, την ψυχολογική ευεξία τους (mental well-being), τους στρεσογόνους παράγοντες, τις αλλαγές στις συνήθειες τους, τις γνώσεις τους σχετικά με την πανδημία και τον τρόπο με τον οποίο ανταποκρίθηκαν στα μέτρα που υλοποιήθηκαν εναντίον του Covid-19. Η έρευνα διενεργήθηκε την άνοιξη του 2020.

Ένα σημαντικό σημείο της έρευνας είναι η δυνατότητα καθορισμού των επιπέδων κατάθλιψης και στρες του εκάστοτε φοιτητή μέσα από σχετικές ερωτήσεις. Είναι ιδιαιτέρως σημαντικό να μπορούν να βρεθούν ομάδες φοιτητών με υψηλά επίπεδα κατάθλιψης και να βρεθεί συσχετισμός της με άλλες απαντήσεις του ερωτηματολογίου, δίνοντας μεγαλύτερη βαρύτητα στις ερωτήσεις που είναι σχετικές με τα μέτρα των πανεπιστημίων και της κυβέρνησης. Με αυτόν τον τρόπο δίνεται η δυνατότητα να καθοριστούν πιθανές αλλαγές ως προς την αντιμετώπιση της πανδημίας με γνώμονα την διατήρηση της ψυχολογικής ευεξίας, ή να μπορούν να βρεθούν χώρες, οι οποίες διαχειρίστηκαν πολύ καλύτερα αυτές τις αλλαγές και να αποτελέσουν παράδειγμα για άλλες χώρες.

Αυτά ζητήματα πραγματεύεται η συγκεκριμένη εργασία με την βοήθεια τεχνικών μηχανικής μάθησης και στατιστικών ελέγχων. Πιο συγκεκριμένα στοχεύουμε στην συλλογή χρήσιμων συμπερασμάτων σε δύο κατηγορίες. Η πρώτη κατηγορία είναι η εξαγωγή μοτίβων, που μπορεί να υπάρχουν στα δεδομένα μας, συγκριτικά με τα επίπεδα κατάθλιψης ενός φοιτητή. Η δεύτερη κατηγορία είναι η υλοποίηση ενός μοντέλου επιβλεπόμενης μάθησης που να μπορεί να προβλέπει τα επίπεδα ή την κατηγορία στην οποία κυμαίνεται η σοβαρότητα της κατάθλιψης του.

Απώτερος σκοπός λοιπόν αυτής της εργασίας αλλά και το ιατρικό αντίκτυπο που έχει, είναι η παροχή ενός πλήρους συστήματος αναγνώρισης ομάδας φοιτητών που πάσχουν από κατάθλιψη, με την παροχή των μοτίβων, που μπορούν να μεταφραστούν σαν συμπτώματα και την εύκολη πρόσβαση τους σε ένα σύστημα διάγνωσης. Επίσης μέσω της αυτής της εργασίας δίνεται η ευκαιρία σε πολλούς φορείς (κρατικούς, πανεπιστημιακούς) να εξετάσουν περαιτέρω τις κινήσεις που οδήγησαν σε διαφορετικά επίπεδα κατάθλιψης.

Η παροχή ενός τέτοιου συστήματος ,όπως συνειδητοποιήσαμε και από τις συνθήκες που επικρατούν στην εποχή της πανδημίας, είναι αναγκαία και λόγω ελάχιστου χρόνου αντιμετώπισης πρέπει να διατίθεται άμεσα. Είναι δεδομένο ότι η ανθρωπότητα θα ξαναπεράσει μια αντίστοιχη κατάσταση και πλέον θα έχει ένα

εργαλείο να παρέχει στους πολίτες της, τις ενδείξεις (μοτίβα) μια υποκείμενης κατάθλιψης αλλά και εύκολης και άμεσης διάγνωσης, μέσω ενός τεστ. Πολύ σημαντικό είναι το γεγονός ότι το τεστ δεν προϋποθέτει την παροχή από τους πολίτες ευαίσθητων πληροφοριών, αλλά απαντήσεις σε ερωτήσεις σχετικά με την καθημερινότητα τους. Θέλουμε να σημειώσουμε πως η παρούσα εργασία δέχεται πολλές επεκτάσεις, με την βασικότερη από όλες , την συλλογή ποιοτικών δεδομένων. Επομένως εμείς προτείνουμε μια μεθοδολογία που μπορεί να εφαρμοστεί και να προσαρμοστεί σε μεγαλύτερης κλίμακα έρευνα και να οδηγήσει σε αντίστοιχα αποτελέσματα.

1.1 Δομή εργασίας

Η συγκεκριμένη εργασία ακολουθεί την δομή που έμμεσα έχει περιγραφεί και παραπάνω.

Αρχικά κάνουμε μια βιβλιογραφική επισκόπηση με τις σχετικές μελέτες που έχουν γίνει και σχετίζονται με το αντικείμενο που πραγματευόμαστε.

Στη συνέχεια παραθέτουμε αναλυτικά τα δεδομένα μας καθώς και την επεξεργασία που δέχθηκαν. Αποτελεί το πιο σημαντικό κομμάτι όλων των επόμενων μεθόδων που πραγματοποιήσαμε.

Έπειτα γίνεται η θεωρητική και πρακτική παρουσίαση των τεχνικών που χρησιμοποιήσαμε και αναλύουμε τις εξής μεθόδους :

- Στατιστικοί έλεγχοι των Δεδομένων
- Κατασκευή μοντέλου παλινδρόμησης για πρόβλεψη τιμής της κατάθλιψης
- Τεχνικές ομαδοποίησης για την εξαγωγή μοτίβων μεταξύ των ομάδων που πάσχουν από κατάθλιψη
- Κατασκευή μοντέλου ταξινόμησης της σοβαρότητας του επιπέδου κατάθλιψης.

2 Βιβλιογραφική Επισκόπηση

Στο σημείο αυτό να αναφέρουμε ότι λόγω του γεγονότος πως η έρευνα C19 ISWS είναι πρόσφατη, δεν υπάρχει μεγάλο πλήθος μελετών οι οποίες κάνουν χρήση των δεδομένων της, και επιπλέον δεν βρέθηκαν δημοσιευμένες μελέτες που να προσεγγίζουν τα δεδομένα με τις μεθόδους που παρουσιάζονται στη συνέχεια.

- Engagement in Health Risk Behaviours before anddduring the COVID-19 Pandemic in German University Students: Results ofa Cross-Sectional Study , [3]

Η συγκεκριμένη μελέτη εξερεύνησε τις αλλαγές στα ΣΡΥ (συμπεριφορές ρίσκου υγείας) κατά την διάρκεια της πανδημίας σε φοιτητές γερμανικών ΑΕΙ, τα χαρακτηριστικά που συνδέονται με αλλαγές στα ΣΡΥ και τα προφίλ των ατόμων που είχαν ΣΡΥ καθώς και τις αλλαγές στα προφίλ αυτά. Τα αντικείμενα που θεωρούνται ΣΡΥ είναι ο αριθμός των τσιγάρων που καπνίζει κανείς, τα ποτά που πίνει, binge drinking, χρήση κάνναβης, έντονη σωματική άσκηση καθώς και μέτριας έντασης σωματική άσκηση. Οι ερευνητές χρησιμοποίησαν περιγραφική ανάλυση, latent transition analysis καθώς και περιγραφική στατιστική και πολυπαραγοντική παλινδρόμηση. Βρέθηκε ότι δεν υπήρξαν σημαντικές διαφορές μεταξύ των συμπεριφορών αυτών πριν και κατά τη διάρκεια της πανδημίας. Η χρήση καπνού επίσης δεν άλλαξε σημαντικά κατά την διάρκεια της πανδημίας. Τα συναισθήματα κατάθλιψης βρέθηκαν να συνδέονται με αύξηση σε συμπεριφορές που επιφέρουν ρίσκο.

Στη συνέχεια υπάρχουν μελέτες όπου κάθε χώρα έχει αναλάβει την οπτικοποίηση των αντίστοιχων δεδομένων μέσω γραφημάτων και γραφικών παραστάσεων, χωρίς όμως την χρήση πολλαπλών στατιστικών ελέγχων και μεθόδων μηχανικής μάθησης. Όλες οι μελέτες ακολουθούν πανομοιότυπες μεταξύ τους αναλύσεις και οπτικοποιήσεις. Συνεπώς τις παραθέτουμε παρακάτω συγκεντρωτικά και θα αναφερθούμε στην αντίστοιχη ελληνική. Μια παρατήρηση είναι ότι ένα πλήθος των συγκεκριμένων ερευνών έχει συνταχθεί στην αντίστοιχη γλώσσα, συνεπώς δεν είναι ευρέως προσβάσιμες.

Paper	Citation
Data from Wageningen University	[5]
STUDENT WELL-BEING DURING THE COVID-19 PANDEMIC IN GREECE	[2]
Student well-being during the first wave of COVID-19 pandemic in Birmingham, UK	[4]

Η ελληνική έρευνα είναι η ”STUDENT WELL-BEING DURING THE COVID-19 PANDEMIC IN GREECE. RESULTS FROM THE C19 ISWS SURVEY” των Stathopoulou Theoni, Mouriki Aliki, Papaliou Olga , η οποία είχε ώς στόχο να πραγματοποιήσει μια στατιστική μελέτη στα συγκεκριμένα δεδομένα της Ελλάδας. Μερικά χρήσιμα συμπεράσματα της έρευνας είναι ότι, παρουσιάζονται υψηλότερα επίπεδα κατάθλιψης σε πρωτοετείς φοιτητές από επιστήμες υγείας και κοινωνικών επιστημών. Επιπλέον, οι φοιτητές των οποίων οι γονείς είχαν χαμηλό εκπαιδευτικό επίπεδο είχαν μεγαλύτερες πιθανότητες να παρουσιάσουν οικονομικά θέματα. Οι φοιτητές, των οποίων οι γονείς είχαν υψηλότερη εκπαίδευση παρουσίασαν πιο συχνό κάπνισμα πριν και μετά την καραντίνα. Τέλος, η συχνότητα των καθημερινών δραστηριοτήτων μέτριας και έντονης σωματικής άσκησης επηρεάστηκε από την

πανδημία του Covid-19. Όλα αυτά τα ενδιαφέροντα συμπεράσματα μας κίνησαν το ενδιαφέρον, με αποτέλεσμα να δούμε το τι συνέβη και σε άλλες χώρες.

3 Δεδομένα

Στο κεφάλαιο αυτό θα δούμε αναλυτικά το σύνολο δεδομένων της εργασία μας. Τα δεδομένα είναι σε μορφή απαντήσεων της μελέτης C19 ISWS. Πριν τη διεξαγωγή της μελέτης αυτής, μόνο λίγες παρόμοιες είχαν διεξαχθεί που εστίαζαν στην ψυχολογική ευεξία κατά τη διάρκεια της πανδημίας Covid-19. Η πλειοψηφία των μελετών αυτών είχαν ως στόχο τους την Κίνα όπου η πανδημία αναγνωρίστηκε πρώτη φορά τον Δεκέμβριο του 2019 αλλά επιπλέον υπάρχουν μελέτες και για άλλες χώρες όπως την Ινδία και το Ηνωμένο Βασίλειο. Αυτές οι μελέτες δείχνουν ότι υπάρχουν αρνητικές ψυχολογικές επιπτώσεις της πανδημίας στον γενικό πληθυσμό και σε συγκεκριμένες ομάδες όπως ιατρικό και νοσηλευτικό προσωπικό. Αυτά τα ευρήματα συμφωνούν με μελέτες που έχουν διεξαχθεί σε προηγούμενες πανδημίες και δείχνουν ότι το αντίκτυπο μίας πανδημίας είναι μεγάλο τόσο στην υγεία όσο και στην κοινωνία. Τα δεδομένα αποτελούνται συνολικά από 123.000 συμπληρωμένα ερωτηματολόγια φοιτητών σε 26 χώρες (Βέλγιο, Γαλλία, Γερμανία, Ολλανδία, Ελβετία, Ηνωμένο Βασίλειο, Τσεχία, Σλοβακία, Ρουμανία, Ουγγαρία, Ρωσία, Σουηδία, Φινλανδία, Νορβηγία, Ισλανδία, Δανία, Ισπανία, Ιταλία, Πορτογαλία, Ελλάδα, Κύπρος, Καναδάς, Ισραήλ, Νότια Αφρική, Τουρκία και ΗΠΑ).

Στη συνέχεια θα δούμε αναλυτικά την δομή του ερωτηματολογίου από το οποίο προέρχονται τα δεδομένα μας.

3.1 Δομή Ερωτηματολογίου

BLOK 0		
StartDate	Start Date	date
EndDate	End Date	date
Duration_in_seconds_	Duration (in seconds)	numeric (count)
Finished	Finished	categorical
RecordedDate	Recorded Date	date
ResponseId	Response ID	String
UserLanguage	User Language	String
BLOK 1		
Q1	What is your gender?	categorical
Q2	Your current age in years	metric
Q3	Are you currently in a steady relationship?	categorical
Q3a	When was the last time you have seen your partner face-to-face?	date
Q4	Were you born in the [country]?	categorical
Q4a	In which country were you born?	Categorical
Q5	Were your parents born in the Netherlands?	categorical
Q5a	In which country?	Categorical
Q5b	In which countries?	Categorical
Q6	What is the highest level of education your mother has completed?	categorical
Q7	What is the highest level of education your father has completed?	Categorical
	From how many of the following people could you easily borrow 500 euros within two days?	
Q8_a	my parents	
Q8_b	my partner	
Q8_c	my siblings	
Q8_d	my grandparents	categorical
Q8_e	friends	
Q8_f	colleagues and/or acquaintances	
BLOK 2		
Q9_1 to Q9_21	Indicate which of the following best describes your field of study?	dummies
Q10	Which program are you currently enrolled in?	categorical
Q11_1 to Q11_	At which higher-education institution are your currently enrolled?	dummies
Q12	Is this your first year in higher education?	categorical
Q13	Please indicate your status in the [country]	categorical
Q14	Did you move back to your previous country of residence since the first implementation of COVID-19 measures	categorical
Q15	How important are your studies compared to other activities?	categorical
Q16_1 to _6	How did you cover the tuition of the current academic year?	dummies
BLOCK 3		
Q17a Q17b	To what extent do you agree with the following statement? 'I had sufficient financial resources to cover my monthly costs'	likert scale
Q18a_1 to _4 Q18b_1 to _4	How many hours did you spend in offline courses, online courses, personal study, and paid jobs on a weekly basis?	numeric (count)
Q19a Q19b	Where did you mainly live (excluding weekends and holidays)?	categorical
Q20_a_1 Q20_b_1	With how many persons did you live together (excluding weekends and holidays)?	numeric (count)
Q21_a Q21_b	On average, how often did you smoke tobacco (cigarettes, cigars, or e-cigarettes)?	categorical
Q21_c_1 Q21_d_1	How many cigarettes did you smoke on average per day?	numeric (count)
Q22a Q22b	On average, how many glasses of alcohol did you drink in one week?	numeric (count)
Q23a Q23b	How often did you drink six or more glasses of alcohol on a single occasion?	categorical
Q24a Q24b	On average, how often did you use cannabis (marijuana, weed, hash,...)?	categorical
Q25a Q25b	On average, how often did you perform vigorous physical activities like lifting heavy things, running, aerobics, or fast cycling for at least 30 minutes?	categorical
Q26a Q26b	On average, how often did you perform moderate physical activities like easy cycling or walking for at least 30 minutes?	categorical
Q27_1 to _10	Do you have any of the following underlying conditions?	dummies
Q28	Have you had symptoms such as coughing, sneezing, or a runny nose during the last month?	categorical
Q28a	Were there occasions that you tried to hide these symptoms from other people	categorical
Q29	Did you have COVID-19, or do you currently have it?	categorical
Q29a	In your opinion, how likely are you to get infected by COVID-19?	metric
Q29b	In your opinion, how likely are you to get re-infected by COVID-19?	metric
Q30a_c	c) How worried are you to get re-infected by COVID-19?	metric
Q30a_d	d) How worried are you that you will get severely ill from a COVID-19 re-infection?	metric
Q30a_c	c) How worried are you to get infected by COVID-19?	
Q30a_d	d) How worried are you that you will get severely ill from a COVID-19 infection?	
Q31a-b	How worried are you that anyone from your personal network ...	
Q31a	... will get infected with COVID-19?	metric
Q31b	... will get severely ill from a COVID-19 infection?	
Q32	How worried are you that doctors and hospitals will not have sufficient medical supplies to handle the COVID-19 outbreak?	metric
Q33	Do you know anyone in your personal network that was or currently is infected with COVID-19?	categorial
Q33a	How severe were the symptoms of that person?	categorial
Q34	To what degree do you adhere to the COVID-19 measures that are currently implemented by the goverment	metric
BLOCK 4		
Q35_1 to _10	During the last week, did you engage in one of the following activities?	dummies
Q36a	Did you have more or less contact with family since the implementation of COVID measures	categorical
Q36b	Did you have more or less contact with friends since the implementation of COVID measures	categorical
Q37	Do you have anyone with whom you can discuss any intimate and personal matters?	categorical
Q38a_n	Please indicate how much of the time during the past week ...	
Q38a	...you felt depressed	
Q38b	...you felt that everything you did was an effort	
Q38c	...your sleep was restless	
Q38d	...you were happy	
Q38e	...you felt lonely	
Q38f	...you enjoyed life	
Q38g	...you felt sad	categorical
Q38h	...you could not get going	
Q38i	...you were bored	
Q38j	...you were frustrated with things in general	
Q38k	...you felt anxious	
Q38l	...you felt calm and peaceful	
Q38m	...you lacked companionship	
Q38n	...felt isolated from others	
BLOCK 5		
Q39a-b	In comparison to the period before the COVID-19 outbreak, did you seek more or less contact with the teaching staff at your university/college:	
Q39a	...to discuss worries about studies	categorical
Q39b	...to discuss psychosocial problems	
Q40	Since the COVID-19 outbreak, did you seek contact with student-counselling services or social services at your university/college?	
Q40a_1 to _5	What was the reason ?	dummies
41a-h	Please indicate to what degree you agree with the following statements:	
Q41_a	My university/college workload has significantly increased since the COVID-19 outbreak.	
Q41_b	I know less about what is expected of me in the different course modules/units since the COVID-19 outbreak.	
Q41_c	I am concerned that I will not be able to successfully complete the academic year due to the COVID-19 outbreak.	
Q41_d	The university/college provides poorer quality of education during the COVID-19 outbreak as before.	Categorical
Q41_e	The change in teaching methods resulting from the COVID-19 outbreak has caused me significant stress.	
Q41_f	The university/college has sufficiently informed me about the changes that were implemented due to the COVID-19 outbreak.	
Q41_g	I am satisfied with the way my university/college has implemented protective measures concerning the COVID-19 outbreak.	
Q41_h	I feel I can talk to a member of the university/college staff (e.g., professor, student counsellor) about my concerns due to the COVID-19 outbreak.	
BLOCK 6		
	Please indicate which of the following statements are true or false:	
Q42_a	The virus survives for days outside the body in open air	
Q42_b	The virus survives for a week outside the body on a plastic surface.	
Q42_c	Most people who get COVID-19 get very ill.	
Q42_d	A possible vaccine will take around 12 to 18 months to produce.	categorical
Q42_e	Smokers who get COVID-19 are more likely to get severely ill than non-smokers.	
Q42_f	You can have the virus without any symptoms.	
Q42_g	On average, children get less ill from the virus than adults.	
Q42_h	Only elderly people die from COVID-19.	
	Please indicate to what degree you agree with the following statement:	
Q43_a	The government provided information concerning the COVID-19 outbreak on time	Categorical
Q43_b	The government provided comprehensive information concerning the COVID-19 outbreak	

Στην πρώτη στήλη βλέπουμε το κωδικό όνομα της ερώτησης το οποίο θα χρησιμοποιούμε για να αναφερθούμε σε αυτήν, στην δεύτερη στήλη βλέπουμε την πλήρη ερώτηση και στην τρίτη και τελευταία στήλη βλέπουμε τον τύπο της μεταβλητής της ερώτησης. Εδώ μπορεί να βρεθεί το αρχείο που περιέχει την πιο πάνω δομή του ερωτηματολογίου με κάποιες επιπλέον επεξηγήσεις σχετικά με τις πιθανές απαντήσεις κάθε ερώτησης [6]. Όσον αφορά τα blocks ο παρακάτω πίνακας περιγράφει συνοπτικά τις θεματικές κατηγορίες των ερωτήσεων κάθε block.

Block 0	Μεταδεδομένα
Block 1	Κοινωνικοδημογραφικές πληροφορίες
Block 2	Πληροφορίες σχετικά με την εκπαίδευση
Block 3	Δραστηριότητες πριν και μετά τον Covid-19 και ανησυχίες
Block 4	Στρεσογόνοι παράγοντες και ψυχολογική ευεξία
Block 5	Ερωτήσεις για τα μέτρα και την αντιμετώπιση των πανεπιστημίων
Block 6	Γνώσεις για τον Covid-19 και πληροφόρηση

Εκτός από το σύνολο δεδομένων της έρευνας C19 ISWS, χρησιμοποιείται το σύνολο δεδομένων Covid-19: Stringency Index [7], το οποίο περιέχει τις τιμές της μετρικής που ορίζει την την αυστηρότητα των μέτρων κατά της πανδημίας κάθε χώρας για κάθε μέρα. Το σύνολο αυτό προέρχεται από το Oxford Coronavirus Government Response Tracker Project. Για τον υπολογισμό της μετρικής αυτής χρησιμοποιούνται εννέα παράγοντες: κλείσιμο σχολείων, κλείσιμο χώρων εργασίας, ακύρωση δημοσίων events, περιορισμοί στις συναθροίσεις, κλείσιμο μέσων μαζικής μεταφοράς, επιβολή παραμονής στο σπίτι, καμπάνιες ενημέρωσης κοινού, περιορισμός στις μετακινήσεις εσωτερικού και έλεγχοι σε ταξίδια εξωτερικού. Το σύνολο δεδομένων ανανεώνεται κάθε μέρα με τις νέες τιμές για την συγκεκριμένη μετρική σε κάθε χώρα.

4 Μεθοδολογία

4.1 Προεπεξεργασία Δεδομένων

Λόγω του μεγάλου αριθμού των μεταβλητών του συνόλου δεδομένων και της ετερογένειάς τους, η προεπεξεργασία και ο καθαρισμός του είναι υψίστης σημασίας. Οι μεταβλητές αυτές σχετίζονται άμεσα με τις απαντήσεις των φοιτητών στο ερωτηματολόγιο. Κατά την προεπεξεργασία, σε κάποιες περιπτώσεις οι απαντήσεις δεν τροποποιούνται, σε άλλες κάποιες απαντήσεις συγχωνεύονται σε μια μεταβλητή, ενώ στις περιπτώσεις στις οποίες οι απαντήσεις των ερωτήσεων είναι στην κλίμακα Likert, κάθε πιθανή απάντηση αντιστοιχίζεται σε έναν αριθμό από το 1 ή το 0 έως το πλήθος των πιθανών επιλογών. Οι αρνητικές απαντήσεις, όπως ”Διαφωνώ πλήρως” και ”Ποτέ” αντιστοιχίζονται με τις μικρότερες τιμές, και όσο αυξάνεται η θετικότητα της απάντησης αυξάνεται και η τιμή. Ο αριθμός των

ερωτήσεων είναι μεγάλος και κάποιες από αυτές δεν έχουν ενδιαφέρον για τη μελέτη μας οπότε τις αφαιρούμε.

Παρακάτω παρουσιάζονται οι μεταβλητές οι οποίες βρίσκονται στο τελικό σύνολο δεδομένων. Για διευκόλυνση χρησιμοποιούνται οι κωδικοί των ερωτήσεων από τον πιο πάνω πίνακα όταν γίνεται αναφορά σε αυτές. Όσες μεταβλητές ταυτίζονται με κάποια συγκεκριμένη ερώτηση, πριν από το όνομά τους έχουν τον κωδικό της ερώτησης αυτής:

- **Q1 - Φύλλο**
- **Q2 - Ηλικία**
- **Q3 - Στάτους Σχέσης**
- **Q6 - Υψηλότερο επίπεδο μόρφωσης μητέρας**
- **Q7 - Υψηλότερο επίπεδο μόρφωσης πατέρα**
- **Αριθμός ατόμων από τα οποία μπορεί να δανειστείς 500 ευρώ** Προκύπτει από το άθροισμα των ερωτήσεων Q8_a έως Q8_f
- **Q9 - Πεδίο σπουδών**
- **Q10 - Πρόγραμμα Σπουδών**
- **Q11 - Κωδικός του εκπαιδευτικού ιδρύματος**
- **Q12 - Πρώτος χρόνος σπουδών** (Ναι, όχι)
- **Q15 - Σημαντικότητα Σπουδών για τον φοιτητή** Likert 3 επιλογών
- **Q17_a - Επαρκείς οικονομικοί πόροι πριν την πανδημία** Likert 5 επιλογών
- **Q17_b - Επαρκείς οικονομικοί πόροι κατά τη διάρκεια της πανδημίας** Likert 5 επιλογών
- **Διαφορά της επάρκειας των οικονομικών πόρων** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Ώρες για προσωπικές δραστηριότητες πριν την πανδημία κάθε εβδομάδα** Προκύπτει από το άθροισμα των ερωτήσεων Q18a_1 έως Q18a_4. Οι δραστηριότητες αυτές είναι Online Μαθήματα, Offline Μαθήματα, Προσωπική Μελέτη, Εργασία.
- **Ώρες για προσωπικές δραστηριότητες κατά τη διάρκεια της πανδημίας κάθε εβδομάδα** Προκύπτει από το άθροισμα των ερωτήσεων Q18b_1 έως Q18b_4. Οι δραστηριότητες αυτές είναι Online Μαθήματα, Offline Μαθήματα, Προσωπική Μελέτη, Εργασία.
- **Διαφορά των ωρών προσωπικών δραστηριοτήτων κάθε εβδομάδα** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.

- **Q20_a_1 - Αριθμός ατόμων με τα οποία συζούσες πριν την πανδημία**
- **Q20_b_1 - Αριθμός ατόμων με τα οποία συζούσες κατά τη διάρκεια της πανδημίας**
- **Διαφορά των αριθμών ατόμων με τα οποία συζούσες** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Q21_a_1 - Πόσο συχνά κάπνιζες πριν την πανδημία**. Likert 6 επιλογών
- **Q21_b_1 - Πόσο συχνά κάπνιζες κατά τη διάρκεια της πανδημίας**. Likert 6 επιλογών
- **Διαφορά της συχνότητας καπνίσματος** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Q21_c_1 - Πόσα τσιγάρα κάπνιζες καθημερινά πριν την πανδημία**
- **Q21_d_1 - Πόσα τσιγάρα κάπνιζες καθημερινά κατά τη διάρκεια της πανδημίας**
- **Διαφορά των τσιγάρων που κάπνιζες καθημερινά** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Q22_a - Πόσα ποτήρια αλκοόλ έπινες εβδομαδιαία πριν την πανδημία**
- **Q22_b - Πόσα ποτήρια αλκοόλ έπινες εβδομαδιαία κατά τη διάρκεια της πανδημίας**
- **Q23_a - Πόσο συχνά έπινες περισσότερα από 6 ποτήρια αλκοόλ πριν την πανδημία** Likert 6 επιλογών
- **Q23_b - Πόσο συχνά έπινες περισσότερα από 6 ποτήρια αλκοόλ κατά τη διάρκεια της πανδημίας** Likert 6 επιλογών
- **Διαφορά της συχνότητας με την οποία έπινες περισσότερα από 6 ποτήρια αλκοόλ** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Q25_a - Πόσο συχνά έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά πριν την πανδημία** Likert 5 επιλογών
- **Q25_b - Πόσο συχνά έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά κατά τη διάρκεια της πανδημίας** Likert 5 επιλογών
- **Διαφορά της συχνότητας με την οποία έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Q26_a - Πόσο συχνά έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά πριν την πανδημία** Likert 5 επιλογών
- **Q26_b - Πόσο συχνά έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά κατά τη διάρκεια της πανδημίας** Likert 5 επιλογών

- **Διαφορά της συχνότητας με την οποία έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά** Προκύπτει αφαιρώντας τις 2 προηγούμενες μεταβλητές.
- **Αριθμός υποκείμενων ασθενειών** Προκύπτει αθροίζοντας τις θετικές απαντήσεις των ερωτήσεων Q27_1 έως Q27_10.
- **Q29 - Έχεις διαγνωστεί με Covid-19** (Ναι, όχι)
- **Q30_a_c - Ανησυχία νόσησης από Covid-19** Τιμές 0-10
- **Q30_a_d - Ανησυχία σοβαρής νόσησης από Covid-19** Τιμές 0-10
- **Q32_a_d - Ανησυχία ότι τα νοσοκομεία δεν έχουν επαρκής πόρους για την αντιμετώπιση της πανδημίας** Τιμές 0-10
- **Q34 - Σε τι βαθμό συμμορφώνεσαι στα μέτρα που υλοποιεί η κυβέρνηση για την πανδημία**
- **Αριθμός δραστηριοτήτων ελεύθερου χρόνου κατά την διάρκεια της τελευταίας εβδομάδας** Προκύπτει αθροίζοντας τις θετικές απαντήσεις των ερωτήσεων Q35_1 έως Q35_10
- **Q36_a - Είχες περισσότερη ή λιγότερη επαφή με την οικογένειά σου από τότε που υλοποιήθηκαν τα μέτρα κατά της πανδημίας** Η επιλογή "λιγότερη" αντιστοιχίζεται στην τιμή 0, "περίπου ίδια" στην τιμή 0 και "περισσότερη" στην τιμή 1.
- **Q36_b - Είχες περισσότερη ή λιγότερη επαφή με τους φίλους σου από τότε που υλοποιήθηκαν τα μέτρα κατά της πανδημίας** Η επιλογή "λιγότερη" αντιστοιχίζεται στην τιμή 0, "περίπου ίδια" στην τιμή 0 και "περισσότερη" στην τιμή 1.
- **Q37 - Ύπαρξη ατόμου για συζήτηση σχετικά με προσωπικά ζητήματα** (Ναι, όχι)

Οι ερωτήσεις Q38_a έως Q38_h αποτελούν την κλίμακα CES-D 8 και χρησιμοποιούνται για την ανίχνευση συμπτώματα κατάθλιψης. Όλες οι ερωτήσεις έχουν 4 επιλογές απαντήσεων σε κλίμακα Likert. Οι ερωτήσεις σχετίζονται με το αν το άτομο νιώθει ή συμπεριφέρεται με κάποιον συγκεκριμένο τρόπο, με τις απαντήσεις να κυμαίνονται από "Ποτέ ή σχεδόν ποτέ" έως "Πάντα ή σχεδόν πάντα". Αντιστοιχίζοντας κάθε απάντηση στο διάστημα 0 έως 3, το τελικό σκορ της κατάθλιψης για κάθε άτομο βρίσκεται αθροίζοντας τις απαντήσεις του. Για τις ερωτήσεις Q38_d και Q38_f πρέπει να γίνει η αντίστροφη αντιστοίχιση αφού σχετίζονται με θετικά συναισθήματα. Είναι προφανές πως το σκορ κατάθλιψης μπορεί να έχει τιμές μεταξύ του 0 και 24.

Επίσης, οι ερωτήσεις Q38_e, Q38_m και Q38_n χρησιμοποιούνται με τον ίδιο τρόπο για τον καθορισμό του σκορ μοναξιάς. Οι 2 τελευταίες εμπεριέχονται στην κλίμακα

RULS-8 η οποία χρησιμοποιείται για ανίχνευση συναισθημάτων μοναξιάς. Το σκορ μοναξιάς έχει τιμές μεταξύ του 0 και 9.

Άρα, συνεχίζοντας με τις μεταβλητές του συνόλου δεδομένων έχουμε:

- **Σκορ κατάθλιψης**
- **Σκορ Μοναξιάς**
- **Q39_a - Αφού ξέσπασε η πανδημία, αναζήτησες περισσότερο ή λιγότερο επικοινωνία με καθηγητές για να συζητήσετε για ανησυχίες σχετικά με τα μαθήματα** Likert 6 επιλογών
- **Q39_b - Αφού ξέσπασε η πανδημία, αναζήτησες περισσότερο ή λιγότερο επικοινωνία με καθηγητές για να συζητήσετε σχετικά με ψυχοκοινωνικά προβλήματα** Likert 6 επιλογών
- **Αριθμός ανησυχιών για τις οποίες μίλησες με ειδικούς από όταν ξέσπασε η πανδημία** Προκύπτει αθροίζοντας τις θετικές απαντήσεις των ερωτήσεων Q40_1 έως Q40_5

Οι ερωτήσεις Q41_a, Q41_b, Q41_c και Q41_e χρησιμοποιούνται για τον υπολογισμό του ακαδημαϊκού στρες και σχετίζονται με την αντιμετώπιση της πανδημίας από το εκπαιδευτικό ίδρυμα στο οποίο σπουδάζει ο φοιτητής και πως τον επηρεάζουν ψυχολογικά. Όλες οι ερωτήσεις έχουν 4 επιλογές απαντήσεων σε κλίμακα Likert, οι οποίες αντιστοιχίζονται σε τιμές 0 έως 4. Το σκορ για το ακαδημαϊκό στρες προκύπτει αθροίζοντας τις απαντήσεις. Οι τιμές του είναι μεταξύ 0 και 16.

Οι ερωτήσεις Q41_d, Q41_f, Q41_g και Q41_h χρησιμοποιούνται για τον υπολογισμό της ακαδημαϊκής ικανοποίησης και όπως και οι προηγούμενες σχετίζονται με την αντιμετώπιση της πανδημίας από το εκπαιδευτικό ίδρυμα στο οποίο σπουδάζει ο φοιτητής και πως τον επηρεάζουν ψυχολογικά. Όλες οι ερωτήσεις έχουν 4 επιλογές απαντήσεων σε κλίμακα Likert, οι οποίες αντιστοιχίζονται σε τιμές 0 έως 4. Το σκορ για την ακαδημαϊκή ικανοποίηση προκύπτει αθροίζοντας τις απαντήσεις. Οι τιμές του είναι μεταξύ 0 και 16.

Συνεχίζοντας με τις μεταβλητές του συνόλου δεδομένων έχουμε:

- **Q41_a έως Q41_h - 8 μεταβλητές που αντιστοιχούν στις προαναφερθείσες ερωτήσεις**
- **Σκορ Στρες**
- **Σκορ Ικανοποίησης**
- **Γνώση για τον Covid-19** Οι ερωτήσεις Q42_a έως Q42_h ελέγχουν αν ο φοιτητής γνωρίζει επαρκώς δεδομένα σχετικά με την πανδημία. Η συγκεκριμένη μεταβλητή προκύπτει αθροίζοντας τις σωστές απαντήσεις κάθε φοιτητή.

- **Q43_a - Η κυβέρνηση παρείχε έγκαιρα πληροφορίες σχετικά με τον Covid-19** Likert 5 επιλογών
- **Q43_b - Η κυβέρνηση παρείχε κατανοητές πληροφορίες σχετικά με τον Covid-19** Likert 5 επιλογών
- **Χώρα στην οποία απαντήθηκε το ερωτηματολόγιο**
- **Μέση Αυστηρότητα Μέτρων** Για κάθε φοιτητή βρίσκεται ο μέσος όρος της αυστηρότητας των μέτρων στη χώρα του από τις 15 Μαρτίου 2020, ημερομηνία κατά την οποία οι περισσότερες χώρες είχαν αρχίσει να εφαρμόζουν μέτρα κατά του Covid-19, έως την ημέρα που ολοκλήρωσε το ερωτηματολόγιο.

Το αρχικό ανεπεξέργαστο σύνολο δεδομένων περιέχει 123,532 εγγραφές. Κρατάμε μόνο εκείνες οι οποίες αντιστοιχούν σε ολοκληρωμένα ερωτηματολόγια και δεν έχουν απουσιάζουσες τιμές στα σκορ κατάθλιψης και στρες. Σε αυτό το σημείο, το σύνολο δεδομένων αποτελείται από 93,087 εγγραφές. Επίσης αφαιρούνται εγγραφές με μη επιτρεπτές τιμές σε κατηγορικές μεταβλητές, και εγγραφές οι οποίες έχουν υπερβολικά υψηλές τιμές στις μεταβλητές που σχετίζονται με τον αριθμό τσιγάρων ανά μέρα και τον αριθμό ατόμων με τα οποία συζεί ο φοιτητής. Κρατάμε μόνο τις εγγραφές οι οποίες έχουν τιμές μικρότερες του 50 σε αυτές τις 2 μεταβλητές. Οι εγγραφές μειώνονται σε 90,579. Σε αυτό το σημείο, οι 2 μεταβλητές που σχετίζονται με τον αριθμό ποτηριών αλκοόλ κάθε εβδομάδα και οι 3 μεταβλητές που σχετίζονται με τον αριθμό ωρών για προσωπικές δραστηριότητες, περιέχουν πολλές απουσιάζουσες τιμές. Γι αυτό, δημιουργούμε 2 ξεχωριστά σύνολα δεδομένων:

- Από το πρώτο αφαιρούμε τις μεταβλητές αυτές, και ύστερα αφαιρούμε τις εγγραφές οι οποίες έχουν απουσιάζουσες τιμές σε κάποια μεταβλητή, οι οποίες είναι λίγες. Το σύνολο δεδομένων αυτό αποτελείται από 85,068 εγγραφές και 76 μεταβλητές.
- Από το δεύτερο αφαιρούμε όλες τις εγγραφές με απουσιάζουσες τιμές σε κάποια μεταβλητή χωρίς να αφαιρέσουμε τις μεταβλητές που αναφέρθηκαν. Λόγω αυτών των μεταβλητών, οι εγγραφές που αφαιρούνται είναι περισσότερες σε σχέση με το πρώτο σύνολο. Αυτό το σύνολο δεδομένων αποτελείται από 60,793 εγγραφές και 81 μεταβλητές.

Στις περισσότερες μεθόδους μηχανικής μάθησης που ακολουθούν, χρησιμοποιείται το πρώτο σύνολο δεδομένων λόγω του ότι είναι μεγαλύτερο. Σε περίπτωση που χρησιμοποιείται το δεύτερο, αυτό θα αναφέρεται ρητά. Επίσης, στις μεθόδους αυτές δεν χρησιμοποιούνται απαραίτητα όλες οι μεταβλητές. Η επιλογή μεταβλητών ως τώρα έγινε με σκοπό να απορριφθούν αυτές που δεν έχουν κανένα ενδιαφέρον και για να είναι πιο διαχειρίσιμο το σύνολο δεδομένων. Παρ' όλα αυτά, σε κάθε μέθοδο επιλέγεται ένα υποσύνολο των 76 μεταβλητών (ή των 81 αν χρησιμοποιείται το δεύτερο σύνολο δεδομένων), είτε επειδή ύστερα από δοκιμές

αποφασίστηκε πως οι υπόλοιπες μεταβλητές δεν συνεισφέρουν σε μεγάλο βαθμό στο να παραχθούν καλύτερα αποτελέσματα, είτε για λόγους εξοικονόμησης υπολογιστικών πόρων όταν έχουμε να κάνουμε με πολύ ακριβά υπολογιστικά μοντέλα. Επίσης, όλες οι μεταβλητές οι οποίες έχουν τιμές σε κλίμακα Likert αντιμετωπίζονται ως συνεχείς μεταβλητές.

4.2 Στατιστικοί Έλεγχοι

Στο παρόν κεφάλαιο, αναλύεται η μεθοδολογία των στατιστικών ελέγχων που χρησιμοποιούνται για την εξαγωγή χρήσιμων συμπερασμάτων, όπως οι αλλαγές σε καθημερινές συνήθειες λόγω της έξαρσης του Covid-19 και οι επιπτώσεις στην ψυχική υγεία των ανθρώπων. Πραγματοποιούνται 2 διαφορετικές προσεγγίσεις στατιστικών ελέγχων, οι οποίες παρουσιάζονται ακολούθως:

4.2.1 Εύρεση εξαρτήσεων μεταξύ της κατάθλιψης και των κατηγορικών μεταβλητών

Λόγω μεγάλου αριθμού κατηγορικών μεταβλητών ή μεταβλητών σε Likert κλίμακα, χρειάζεται να πραγματοποιηθούν ορισμένοι στατιστικοί έλεγχοι, οι οποίοι έχουν στόχο τον προσδιορισμό εξαρτήσεων των συγκεκριμένων μεταβλητών με την μεταβλητή επιπέδων κατάθλιψης. Γι' αυτόν τον σκοπό, χρησιμοποιείται ο έλεγχος One-way Analysis of Variance, ο οποίος καθορίζει αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων 3 ή περισσότερων ομάδων. Πιο συγκεκριμένα, ο έλεγχος υπόθεσης που θέλουμε να μελετήσουμε είναι $H_0 : m_1 = m_2 = ...m_k$, όπου κ = αριθμός των ομάδων έναντι H_1 : *Υπάρχουν τουλάχιστον δύο μέσοι 2 ομάδων, οι οποίοι είναι στατιστικά διαφορετικοί μεταξύ τους*. Ο συγκεκριμένος στατιστικός έλεγχος βασίζεται σε 3 βασικές υποθέσεις για να δώσει αξιόπιστα αποτελέσματα, οι οποίες είναι:

- Η εξαρτημένη μεταβλητή,πρέπει να ακολουθεί την κανονική κατανομή για κάθε ομάδα της κατηγορικής (εξαρτημένης) μεταβλητής
- Ύπαρξη ομοιογένειας των διασπορών
- Ανεξαρτησία μεταξύ των παρατηρήσεων, δηλαδή να μην υπάρχει κάποια σχέση μεταξύ των παρατηρήσεων μεταξύ των διαφορετικών ομάδων

Ο παραπάνω έλεγχος υλοποιείται με την βοήθεια της βιβλιοθήκης *scipy* [8] Στην συγκεκριμένη περίπτωση, ελέγχεται για κάθε κατηγορική μεταβλητή, αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων των επιπέδων κατάθλιψης ανάμεσα στις κατηγορίες κάθε μεταβλητής. Οι κατηγορικές μεταβλητές που χρησιμοποιούνται είναι η χώρα (Q4), ο τομέας σπουδών (Q9), το πρόγραμμα

σπουδών (Q10), η ερωτική κατάσταση (Q3), το φύλο (Q1) και το ανώτατο εκπαιδευτικό ίδρυμα (Q11). Σχετικά με τις μεταβλητές σε Likert Scale, πραγματοποιείται ο μη παραμετρικός έλεγχος X^2,με την βοήθεια της scipy [9], ο οποίος εξετάζει αν υπάρχει συσχέτιση μεταξύ κατηγορικών ή ordinal μεταβλητών, δηλαδή των μεταβλητών σε Likert Scale. Η μηδενική υπόθεση του ελέγχου είναι ότι δεν υπάρχει κάποια συσχέτιση μεταξύ των κατηγορικών μεταβλητών, δηλαδή είναι ανεξάρτητες. Για τον σκοπό της συγκεκριμένης εργασίας, χρησιμοποιούνται οι ordinal μεταβλητές δηλαδή η εκπαίδευση μητέρας/πατέρα του φοιτητή(Q_6/Q_7), σημαντικότητα των σπουδών (Q_15), η ανησυχία σχετικά με τον επαρκή ιατρικό εξοπλισμό (Q_32), η εφαρμογή των μέτρων σχετικά με Covid-19 (Q_34), επαφή με οικογένεια/φίλους (Q36_a/Q36_b), η κάλυψη των μηνιαίων αναγκών (Q17_a,Q17_b), η συχνότητα καπνίσματος και αλκοόλ (Q21_a/Q21_b και Q23_a/Q23_b), η συχνότητα έντονης/μέτριας σωματικής άσκησης (Q25_a/Q25_b και Q26_a/Q26_b), ανησυχία μόλυνσης/σοβαρής μόλυνσης (Q30_a_c/Q30_a_d), συζήτηση με το διδακτικό προσωπικό σχετικά με ανησυχίες για τα μαθήματα/ ψυχο - κοινωνικά προβλήματα (Q39_a/39_b), συζήτηση για προσωπικά ζητήματα(Q_37), οι ερωτήσεις σχετικά με τα μέτρα των πανεπιστήμιων (Q41_a - Q41_h) και οι ερωτήσεις σχετικά με την κυβέρνηση (Q43_a/Q43_b). Χρησιμοποιούνται οι συγκεκριμένες μεταβλητές σε συνδυασμό με την μεταβλητή των επιπέδων κατάθλιψης. Σε περίπτωση σοβαρών ενδείξεων της μηδενικής υπόθεσης, χρησιμοποιείται ο συντελεστής Cramer's V, ο οποίος αποτελεί ένα μέτρο συσχέτισης μεταξύ δύο μεταβλητών σε εύρος τιμών (0,1). Όσο πιο κοντά στο 1 είναι η τιμή του συγκεκριμένου συντελεστή, τόσο μεγαλύτερη εξάρτηση υπάρχει μεταξύ των δύο μεταβλητών.

4.2.2 Επιρροή πανδημίας σε άτομα με χαμηλή και υψηλή κατάθλιψη

Με την βοήθεια των στατιστικών ελέγχων, θέλουμε να ελέγξουμε αν η έξαρση της πανδημίας του Covid-19 παρουσιάζει στατιστικά σημαντικά διαφορές ανάμεσα στους φοιτητές με χαμηλά και υψηλά επίπεδα κατάθλιψης. Για τον σκοπό αυτό, το σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα, εκ των οποίων το ένα αφορά τους φοιτητές με επίπεδα κατάθλιψης μικρότερα του 8 και το άλλο φοιτητές με επίπεδα κατάθλιψης μεγαλύτερα ή ίσα του 8. Στην συνέχεια, πραγματοποιείται ένας αμφίπλευρος στατιστικός έλεγχος t-test [10]. Ο συγκεκριμένος έλεγχος εξετάζει αν υπάρχουν στατιστικά σημαντικά διαφορές μεταξύ των μέσων 2 ομάδων, δηλαδή η μηδενική υπόθεση του ελέγχου είναι $H_0 : m_1 = m_2$. Οι υποθέσεις του συγκεκριμένου ελέγχου είναι οι ακόλουθες:

- Τα δεδομένα ακολουθούν την κανονική κατανομή
- Ύπαρξη ομοιογένειας των διασπορών
- Ανεξαρτησία μεταξύ των παρατηρήσεων, δηλαδή να μην υπάρχει κάποια σχέση μεταξύ των παρατηρήσεων μεταξύ των διαφορετικών ομάδων

Ο συγκεκριμένος έλεγχος χωρίζεται σε δύο είδη στατιστικών ελέγχων, τον έλεγχο t-test με ίσες διασπορές ανάμεσα στις ομάδες και τον Welch's t-test με άνισες διασπορές ανάμεσα στις ομάδες.
Ο στατιστικός έλεγχος εφαρμόζεται στις μεταβλητές που δηλώνουν την αλλαγή με την έξαρση του Covid-19, δηλαδή τις μεταβλητές η συχνότητα καπνίσματος και αλκοόλ (Q21_a/Q21_b και Q23_a/Q23_b), η συχνότητα έντονης/μέτριας σωματικής άσκησης (Q25_a/Q25_b και Q26_a/Q26_b), κάλυψη των μηνιαίων αναγκών (Q17_a,Q17_b), αριθμός των τσιγάρων ανά ημέρα (Q21_c_1/Q21_d_1) και ο αριθμός ατόμων, με τους οποίους ζει κάθε φοιτητής (Q20_a_1/Q20_b_1). Για κάθε μεταβλητή ελέγχονται οι υποθέσεις του ελέγχου και εξετάζεται η διασπορά τους ανάμεσα στις δύο ομάδες για την επιλογή ελέγχου.

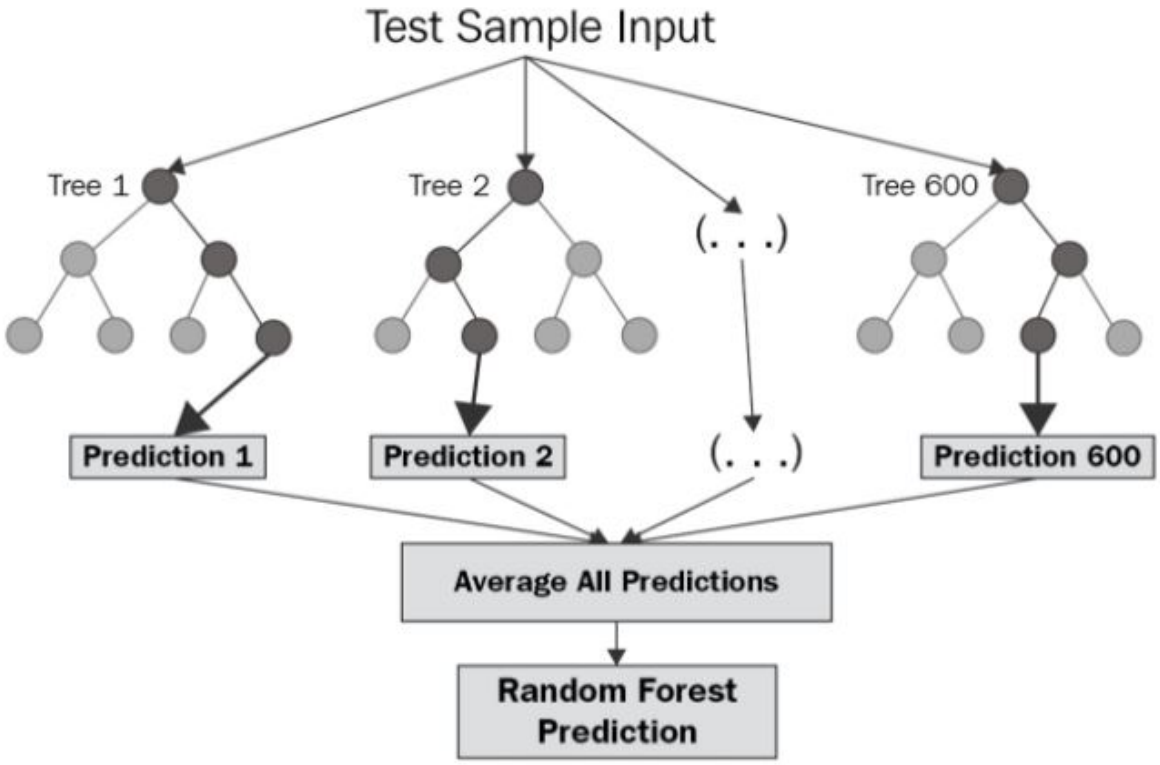
4.3 Παλινδρόμηση

Σε αυτήν την ενότητα εξετάζουμε το κομμάτι της παλινδρόμησης. Ένας από τους πρωταρχικούς στόχους μας είναι να μπορέσουμε να αποφανθούμε για την ψυχολογική κατάσταση ενός φοιτητή όσον αφορά συναισθήματα κατάθλιψης. Αυτό μπορεί να γίνει με δύο τρόπους: Ο ένας είναι να γίνει ένας διαχωρισμός σε ομάδες με βάση το σκορ κατάθλιψης και στην συνέχεια να εφαρμοστούν μέθοδοι ταξινόμησης και ο άλλος και πιο άμεσος τρόπος χωρίς την περαιτέρω εφαρμογή δικής μας μεροληψίας είναι να προβλέψουμε κατευθείαν το σκορ κατάθλιψης. Ο λόγος που θέλουμε να δημιουργήσουμε ένα τέτοιο μοντέλο είναι γιατί αν θεωρητικά αν μπορούσαμε να μοντελοποιήσουμε μία δύσκολη ψυχική κατάσταση και συγκεκριμένα την κατάθλιψη το οποίο είναι μείζον θέμα που απασχολεί πολλά επιστημονικά πεδία και εν γένει δεν έχει βρεθεί ακόμα κάποιο καλό εργαλείο για pretesting. Αυτή η προσέγγιση δοθέντων των δεδομένων που είναι γενικής φύσεως αλλά και την cross-sectional φύση της μελέτης αποδεικνύεται ένας αρκετά δύσκολος στόχος. Τελικά μετά από πολλές προκαταρκτικές δοκιμές καταλήξαμε στην ανάπτυξη ενός μοντέλου με τη χρήση Random Forest. Οι μεταβλητές του μοντέλου μας είναι 96 συνολικά το σύνολο. Στον παρακάτω πίνακα φαίνονται αναλυτικά τα κωδικά ονόματα των ερωτήσεων.

People live together - Q20_a_1
People live together - Q20_b_1
#Cigarettes per day - Q21_c_1
#Cigarettes per day - Q21_d_1
Age - Q2
Education(Mother) - Q6
Education(Father) - Q7
University - Q11
Importance of Study - Q15
Borrow 500 - Q8
Sufficient Medical Supplies C19 - Q32
Measurement_Adherence - Q34
SocialActivities - Q35
FamilyContact - Q36a
FriendsContact - Q36b
On time government info - Q43_a
Comprehensive government info - Q43_b
CoverMonthlyCostsb - Q17_a
CoverMonthlyCostsa - Q17_b
TobaccoOftenb - Q21_a
TobaccoOftena - Q21_b
AlcoholOftenb - Q23_a
AlcoholOftena - Q23_b
VigoPhysiActb - Q25_a
VigoPhysiActa - Q25_b
ModPhysiActb - Q26_a
ModPhysiActa - Q26_b
number_conditions - Q27
InfectionWorries - Q30b_1
SevereInfectionWorries - Q30b_2
Metric_i - Q38_i
Metric_j - Q38_j
Metric_k - Q38_k
Metric_l - Q38_l
DiscussStudies - Q39_a
DiscussStudies - Q39_b
NumberWorries - Q40a
Stress_Score
Satisfaction_Score
Loneliness
COVID_Knowledge - Q42
Gender_female
Gender_male
Gender_x
Status_It is complicated
Status_No, I am single
Status_Yes
FoS_ social sciences, business and law
FoS_Engineering, manufacturing and construction
FoS_Health and welfare
FoS_agriculture
FoS_education
FoS_humanities and arts
FoS_other
FoS_science
FoS_services
Program_Bachelor program
Program_Doctoral program
Program_Master program
Program_Other, specify
HadCovid_no
HadCovid_yes
DiscussPersonal_2
DiscussPersonal_1
FirstYear_No
FirstYear_Yes
Country_USA
Country_UK
Country_Turkey
Country_Switzerland
Country_ Czech
Country_Belgium
Country_Canada
Country_Cyprus
Country_Denmark
Country_Finland
Country_France
Country_Germany
Country_Greece
Country_Hungary
Country_Iceland
Country_Israel
Country_Italy
Country_Netherlands, the
Country_Norway
Country_Portugal
Country_Romania
Country_Russia
Country_Slovakia
Country_South Africa
Country_Spain
Country_Sweden

4.3.1 Random Forest

Σε αυτό το σημείο θα δούμε εποπτικά πώς λειτουργεί ο αλγόριθμος Random Forest. Πρόκειται για έναν αλγόριθμο supervised μάθησης ο οποίος χρησιμοποιεί ensemble learning μέθοδο. Αυτό πρακτικά σημαίνει ότι εκεί που με ένα απλό δέντρο αποφάσεων θα είχαμε 1 έξοδο τώρα έχουμε N δέντρα (συχνά 100-500) με N εξόδους οι οποίες συνδυάζονται για να πάρουμε εν τέλει 1 αποτέλεσμα. Ωστόσο αυτή η τεχνική δεν εφαρμόζεται κατευθείαν στο σετ εκπαίδευσής μας αλλά αυτό που συμβαίνει είναι ότι το κάθε δέντρο αποφάσεων εκπαιδεύεται με ένα bootstrap δείγμα του αρχικού σετ. Αυτή η τεχνική προσδίδει ευρωστία στη μέθοδο, μειώνει το overfitting και εξερευνάται καλύτερα η υποθάλπτουσα στατιστική κατανομή των δεδομένων.



Σχήμα 1: Η δομή της μεθόδου Random Forest.

Ο λόγος που επιλέξαμε Random Forest έγκειται στην απλότητά του. Πρώτον και κυριότερο μπορεί να διαχειριστεί δεδομένα διαφόρων ειδών όπως αυτά που έχουμε σε αντίθεση με γραμμικές μεθόδους που αντιμετωπίζουν προβλήματα. Δεύτερον ο Random Forest είναι από τις λίγες μεθόδους που έχουν built-in επεξηγησιμότητα το οποίο είναι πολύ χρήσιμο στο πεδίο της ιατρικής και συγκεκριμένα της ψυχιατρικής καθώς μας επιτρέπει να βρούμε βαθύτερα πρότυπα μέσα από τη διαδικασία της μοντελοποίησης. Τέλος η μέθοδος είναι πλήρως παραλληλοποιήσιμη το οποίο είναι επίσης σημαντικό.

Εύρεση Βέλτιστου Μοντέλου: Αφού πλέον έχουμε πει για τις μεταβλητές εισόδου και την αρχιτεκτονική του βασικού μοντέλου, θα μιλήσουμε για την εύρεση του βέλτιστου μοντέλου που μας οδήγησε στα αποτελέσματά μας. Για τον σκοπό αυτό κάναμε grid search στις υπερπαραμέτρους τροποποιώντας μόνο τις βασικές οι οποίες είναι υπεύθυνες για τις αποφάσεις του μοντέλου δηλαδή των αριθμό των ταξινομητών (n_estimators) και τα μέγιστα χαρακτηριστικά (max_features). Μετά το grid search καταλήξαμε ότι το καλύτερο μοντέλο είναι αυτό με 300 ταξινομητές και 30% των συνολικών χαρακτηριστικών ανά split.

4.4 Ομαδοποίηση

Ένας από τους βασικούς στόχους αυτής της εργασίας είναι η εφαρμογή τεχνικών ομαδοποίησης με σκοπό την εύρεση ομάδων φοιτητών που διαφοροποιούνται σύμφωνα με τα επίπεδα κατάθλιψης και η συσχέτιση της μεταβλητής αυτής με άλλες. Για τον σκοπό αυτόν χρησιμοποιούνται δύο διαφορετικές προσεγγίσεις, όπως παρουσιάζεται ακολούθως:

- Προσέγγιση 1: Πολλαπλές διαφορετικές ομαδοποιήσεις των φοιτητών σύμφωνα με τα επίπεδα στρες και κατάθλιψης, την πρόθεση τους για συζήτηση σχετικά με ανησυχίες που έχουν και την αλλαγή στις καθημερινές τους συνήθειες. Για τις ομαδοποιήσεις αυτής της προσέγγισης γίνεται χρήση του αλγορίθμου *Kmeans*.
- Προσέγγιση 2: Ομαδοποιήσεις των φοιτητών λαμβάνοντας υπόψιν όσο τον δυνατόν μεγαλύτερο πλήθος μεταβλητών, κυρίως αυτών που σχετίζονται με τα μέτρα των πανεπιστημίων και της κυβέρνησης, και την εύρεση διαφορών μεταξύ των ομάδων. Για τις ομαδοποιήσεις αυτής της προσέγγισης γίνεται χρήση των δικτύων *Self Organizing maps*.

Κάθε προσέγγιση αναλύεται με περισσότερες λεπτομέρειες στην αντίστοιχη ενότητα.

Σε κάθε περίπτωση, οι κατηγορικές μεταβλητές μετατρέπονται σε δυαδική αναπαράσταση One-Hot Encoding και οι μεταβλητές σε κλίμακα Likert κανονικοποιούνται στο διάστημα [0,1] ή στο [-1,1], αναλόγως με το αν περιλαμβάνουν αρνητικές τιμές ή όχι, όπως οι μεταβλητές που δηλώνουν αλλαγές στις συνήθειες των φοιτητών. Η κανονικοποίηση γίνεται με την βοήθεια της βιβλιοθήκης *sklearn* [11].

Όσον αφορά τους συντελεστές αξιολόγησης των ομαδοποιήσεων που πραγματοποιούνται χρησιμοποιούνται:

1. Ο συντελεστής Silhouette, ο οποίος λαμβάνει τιμές στο (0,1) και δίνεται από τον τύπο
$$s = \frac{b - a}{max(a, b)}$$

a: μέση απόσταση μεταξύ ενός δείγματος με όλα τα υπόλοιπα σημεία στην ίδια ομάδα

b: μέση απόσταση μεταξύ ενός δείγματος με όλα τα υπόλοιπα σημεία στην κοντινότερη ομάδα

Όσο η τιμή του είναι πιο κοντά στο 1, τόσο πιο καλά διαχωρισμένες και πυκνές είναι οι ομάδες

2. Ο συντελεστής Calinski-Harabasz Index ή αλλιώς γνωστός ως Variance Ratio Criterion, ο οποίος ορίζεται ως η αναλογία του αθροίσματος της διαχωρισιμότητας μεταξύ των ομάδων και της διαχωρισιμότητας εντός της κάθε ομάδας για όλες τις ομάδες. Σαν διαχωρισιμότητα, ορίζεται το άθροισμα των τετραγωνικών αποστάσεων.

   Όσο πιο μεγάλες τιμές λαμβάνει ό συγκεκριμένος συντελεστής, τόσο καλύτερα υψηλότερη διαχωρισιμότητα έχουν οι ομάδες μεταξύ τους.

Η εύρεση των παραπάνω συντελεστών αξιολόγησης γίνεται με την βοήθειά της *sklearn* [12] [13]

4.4.1 Προσέγγιση 1

Η χρήση του αλγόριθμου *Kmeans* γίνεται με την χρήση της sklearn [14]. Πιο συγκεκριμένα, για την εύρεση του αριθμού των ομάδων με χρήση του *Kmeans* χρησιμοποιείται η μέθοδος Elbow, η οποία είναι μια ευριστική μέθοδος με σκοπό την ελαχιστοποίηση του κριτηρίου Inertia ή το άθροισμα των τετραγώνων μέσα στις ομάδες, που δίνεται από τον τύπο $\sum_{i=0}^{n} min_{m_j \in C}(\| x_i - m_j \|^2)$.

Σύμφωνα με αυτή την προσέγγιση, στόχος είναι να ομαδοποιηθούν οι φοιτητές σε ομάδες που διαχωρίζονται πολύ καλά ενώ παράλληλα χρησιμοποιείται ένα μικρότερο υποσύνολο των μεταβλητών που διαθέτουμε. Πιο συγκεκριμένα, θέλουμε να δημιουργήσουμε ομάδες με ευδιάκριτη και ουσιώδη πληροφορία. Στην συνέχεια, έχοντας διαχωρίσει πολύ καλά τις ομάδες φοιτητών προσπαθούμε να εξάγουμε συμπεράσματα σχετικά με τις χώρες που ανήκουν σε αυτές τις ομάδες. Σύμφωνα με τα αποτελέσματα που προκύπτουν, μπορούμε να λάβουμε αντίστοιχα μέτρα για κάθε χώρα σε περίπτωση επερχόμενης καραντίνας λόγω του Covid-19 ή οι χώρες αντιμετώπισαν πιο βαριά την καραντίνα να βελτιωθούν σύμφωνα με χώρες που την αντιμετώπισαν καλύτερα και να βελτιώσουν την ψυχική υγεία των φοιτητών. Οι ομαδοποιήσεις που πραγματοποιούνται είναι οι παρακάτω:

- Ομαδοποίηση των φοιτητών σύμφωνα με τα επίπεδα στρες και κατάθλιψης

  Για την συγκεκριμένη ομαδοποίηση, χρησιμοποιούνται σαν μεταβλητές τα επίπεδα κατάθλιψης και στρες. Αρχικά εφαρμόζεται η μέθοδος Elbow για την εύρεση του αριθμού των ομάδων και στην συνέχεια εξετάζεται η διαχωρισιμότητα των ομάδων σύμφωνα με τους συντελέστες Silhouette και

Calinski-Harabasz Index. Τέλος, για κάθε χώρα υπολογίζεται η συνεισφορά της σε κάθε ομάδα, υπολογίζοντας το ποσοστό των φοιτητών που ανήκουν σε αυτή την χώρα και σε καθεμιά από τις ομάδες που προκύπτουν. Συγκρίνονται τα αποτελέσματα που προκύπτουν με την αυστηρότητα των μέτρων κάθε χώρας.

- Ομαδοποίηση των φοιτητών σύμφωνα με την πρόθεση τους για συζήτηση

  Για την συγκεκριμένη ομαδοποίηση, χρησιμοποιούνται σαν μεταβλητές τα επίπεδα κατάθλιψης και στρες, συζήτηση με το διδακτικό προσωπικό σχετικά με ανησυχίες για τα μαθήματα/ ψυχο - κοινωνικά προβλήματα(Q39_a/Q39_b) και η συζήτηση των φοιτητών σχετικά με τα προσωπικά τους ζητήματα(Q37). Ακολουθείται η ίδια διαδικασία με την πρώτη ομαδοποίηση και συγκρίνονται τα αποτελέσματα με τα αποτελέσματα της πρώτης ομαδοποίησης.

- Ομαδοποίηση των φοιτητών σύμφωνα με την υγεία και τις συνήθειες τους

  Για την συγκεκριμένη ομαδοποίηση, χρησιμοποιούνται σαν μεταβλητές η συχνότητα καπνίσματος και αλκοόλ (Q21_a/Q21_b και Q23_a/Q23_b), η συχνότητα έντονης/μέτριας σωματικής άσκησης (Q25_a/Q25_b και Q26_a/Q26_b) και συγκρίνονται τα αποτελέσματα με τις 2 παραπάνω ομαδοποιήσεις.

4.4.2 Προσέγγιση 2

Όπως προαναφέρθηκε, για τις ομαδοποιήσεις αυτής της προσεγγίσεις χρησιμοποιούνται τα δίκτυα SOM (Self Organizing Maps), τα οποία είναι νευρωνικά δίκτυα που εκπαιδεύονται χρησιμοποιώντας μη επιβλεπόμενη μάθηση. Χρησιμοποιούνται για την παραγωγή μιας δισδιάστατης απεικόνισης των δεδομένων εισόδου η οποία ονομάζεται χάρτης (map), οπότε πολύ συχνά χρησιμοποιούνται ως μέθοδος μείωσης διαστάσεων, οπτικοποίησης δεδομένων πολλών διαστάσεων και ομαδοποίησης (clustering).

Το δίκτυο SOM διαθέτει 2 επίπεδα, το επίπεδο εισόδου και το επίπεδο εξόδου. Το επίπεδο εισόδου αποτελείται από τόσους νευρώνες όσες είναι οι διαστάσεις των δεδομένων εισόδου. Το επίπεδο εξόδου αποτελείται από όσους νευρώνες επιλέξουμε στους οποίους αντιστοιχίζεται ένα διάνυσμα βαρών ίδιου μήκους με τις διαστάσεις εισόδου. Κάθε νευρώνας του επιπέδου εισόδου συνδέεται με όλους τους νευρώνες του επιπέδου εξόδου. Η διαδικασία της μάθησης αποτελείται από εποχές, και σε κάθε εποχή γίνεται ανανέωση των διανυσμάτων βαρών κάθε νευρώνα. Για κάθε δεδομένο εισόδου, βρίσκεται ο νευρώνας ο οποίος βρίσκεται πιο κοντά σε αυτό εφαρμόζοντας μια συνάρτηση απόστασης μεταξύ των τιμών των μεταβλητών του δεδομένου και των διανυσμάτων βαρών των νευρώνων. Τα βάρη του νευρώνα αυτού, ο οποίος για το συγκεκριμένο δεδομένο ονομάζεται BMU (Best Matching Unit) ανανεώνονται έτσι ώστε να προσεγγίσει ακόμα περισσότερο το δεδομένο. Οι νευρώνες που βρίσκονται κοντά σε αυτόν, δηλαδή στη γειτονιά του, προσεγγίζουν επίσης το δεδομένο αλλά μετατοπίζονται σε μικρότερο βαθμό, ανάλογα με την απόστασή τους από τον BMU. Το μέγεθος της γειτονιάς των νευρώνων φθίνει σε κάθε εποχή. Αφού ολοκληρωθεί

η εκπαίδευση του δικτύου, κάθε δεδομένο εισόδου θα βρίσκεται κοντά στον BMU του, άρα ουσιαστικά θα μπορεί να αναπαρασταθεί από αυτόν. Είναι εμφανές πως υπάρχει περίπτωση κάποιοι νευρώνες να μην είναι BMUs για κανένα δεδομένο, αφού 2 ή περισσότερα δεδομένα μπορεί να έχουν τον ίδιο BMU. Όταν τα δίκτυα SOM χρησιμοποιούνται για ομαδοποίηση, συνήθως ακολουθούνται 2 τακτικές. Στην πρώτη και πιο απλή, το πλήθος των νευρώνων εξόδου του δικτύου επιλέγεται έτσι ώστε να είναι ίσο με το πλήθος των ομάδων που επιθυμείται να προκύψουν. Με αυτό τον τρόπο, κάθε ομάδα αποτελείται από τα δεδομένα που έχουν ως BMU τον εκάστοτε νευρώνα.

Για την δεύτερη τακτική, επιλέγεται σχετικά μεγάλο πλήθος νευρώνων εξόδου και γίνεται η εκπαίδευση του δικτύου. Ύστερα, οι νευρώνες ομαδοποιούνται χρησιμοποιώντας τα τελικά βάρη τους με κάποιον άλλο αλγόριθμο ομαδοποίησης που επιλέγεται (K-Means, DBSCAN κ.λ.π.). Αφού δημιουργηθούν οι ομάδες, κάθε δεδομένο αντιστοιχίζεται σε αυτή την ομάδα στην οποία βρίσκεται ο BMU του. Με αυτό τον τρόπο επιτρέπουμε στο δίκτυο SOM να προσαρμοστεί καλύτερα στα δεδομένα λόγω του μεγαλύτερου αριθμού νευρώνων εξόδου που περιλαμβάνει. Στο συγκεκριμένο πρόβλημα γίνεται χρήση της δεύτερης τακτικής.

Κύρια επιθυμία των ομαδοποιήσεων της παρούσας προσέγγισης είναι η πιθανή παραγωγή ομάδων φοιτητών οι οποίες διαφοροποιούνται μεταξύ τους σε κάποιον βαθμό ως προς το σκορ κατάθλιψης των φοιτητών που ανήκουν σε κάθε μια από αυτές, και η ανάλυση των υπολοίπων μεταβλητών ώστε να εξεταστούν οι πιθανές διαφορές στις τιμές τους μεταξύ των ομάδων. Δίνεται ιδιαίτερη σημασία στις μεταβλητές οι οποίες έχουν να κάνουν με ερωτήσεις σχετικές με την αντιμετώπιση και τα μέτρα λόγω της πανδημίας από το εκπαιδευτικό ίδρυμα του κάθε φοιτητή και πως αυτά τον επηρέασαν, στις μεταβλητές που σχετίζονται με τις διαφορές των συνηθειών του φοιτητή και στις μεταβλητές που σχετίζονται με την πληροφόρηση και τα μέτρα της εκάστοτε κυβέρνησης για την αντιμετώπιση της πανδημίας. Αυτό γίνεται με στόχο να βρεθούν πιθανώς ομάδες φοιτητών με υψηλότερα επίπεδα κατάθλιψης από άλλες και να καθοριστούν αλλαγές που μπορούν να γίνουν ως προς την αντιμετώπιση της πανδημίας από τα εκπαιδευτικά ιδρύματα και την κυβέρνηση, αλλά και να συσχετιστούν χαρακτηριστικά φοιτητών όπως η ηλικία με τα επίπεδα κατάθλιψης για να δοθεί μεγαλύτερη προσοχή σε αυτές τις υποομάδες.

Πραγματοποιούνται 3 διαφορετικές ομαδοποιήσεις, οι οποίες περιγράφονται αναλυτικότερα στη συνέχεια. Οι κοινές μεταβλητές που χρησιμοποιούνται σε όλες τις ομαδοποιήσεις είναι οι εξής:

Q1 - Φύλλο, Q2 - Ηλικία, Q10 - Πρόγραμμα Σπουδών, Σκορ κατάθλιψης, Σκορ Στρες, οι 8 μεταβλητές που αντιστοιχούν στις ερωτήσεις Q41_a έως Q41_h σχετικές με τα μέτρα των εκπαιδευτικών ιδρυμάτων και την αντιμετώπισή τους από τους φοιτητές, Γνώση για τον Covid-19, Q30_a_d - Ανησυχία σοβαρής νόσησης από Covid-19, Q37 - Ύπαρξη ατόμου για συζήτηση σχετικά με προσωπικά ζητήματα, Διαφορά των τσιγάρων που κάπνιζες καθημερινά, Διαφορά της επάρκειας οικονομικών πόρων, Διαφορά της συχνότητας καπνίσματος, Διαφορά της συχνότητας με την οποία έπινες περισσότερα από 6 ποτήρια αλκοόλ, Διαφορά της

συχότητας με την οποία έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά, Διαφορά της συχνότητας με την οποία έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά, Q43_a - Η κυβέρνηση παρείχε έγκαιρα πληροφορίες σχετικά με τον Covid-19, Q43_b - Η κυβέρνηση παρείχε κατανοητές πληροφορίες σχετικά με τον Covid-19.

Εφόσον το σκορ κατάθλιψης είναι η μεταβλητή που έχει την μεγαλύτερη σημασία για την παρούσα έρευνα, στην συγκεκριμένη ενότητα αποφασίζεται να της δοθεί μεγαλύτερο βάρος για την διαδικασία της ομαδοποίησης σε σχέση με τις υπόλοιπες μεταβλητές. Ο τρόπος που γίνεται αυτό είναι κανονικοποιώντας την σε μεγαλύτερο εύρος τιμών από τις υπόλοιπες, αφού όταν όλες κανονικοποιούνται στο ίδιο εύρος σημαίνει πως τους ανατίθεται το ίδιο βάρος, κάτι που έχει αναλυθεί στο "Finding Groups in Data: An Introduction to Cluster Analysis" των Kaufman Leonard, and Peter J. Rousseeuw [15]. Έτσι, το εύρος της μεταβλητής αυτής παραμένει να είναι το αρχικό, ενώ όλες οι άλλες αριθμητικές μεταβλητές κανονκοποιούνται στο διάστημα [0,1].

Το μέγεθος των δικτύων SOM για τις 2 πρώτες περιπτώσεις 5x5, δηλαδή περιέχει 25 νευρώνες εξόδου και χρησιμοποιούνται 10,000 εποχές για την εκπαίδευσή τους. Δεν υπάρχει προκαθορισμένος τρόπος για τον καθορισμό του πλήθους των νευρώνων εξόδου, η επιλογή εξαρτάται πάντα από το εκάστοτε πρόβλημα, οπότε επιλέγεται ο συγκεκριμένος αριθμός ώστε να συμβαδίζει αρκετά με το πλήθος των μεταβλητών εισόδου. Ένας άλλος λόγος είναι οι περιορισμοί υπολογιστικών πόρων, αφού η εκπαίδευση ενός τέτοιου δικτύου για τόσο μεγάλο μέγεθος δεδομένων διαρκεί πολύ, και αύξηση των νευρώνων ή των εποχών έχει σαν αποτέλεσμα μεγάλη αύξηση του χρόνου εκτέλεσης.

Για την ομαδοποίηση των νευρώνων εξόδου χρησιμοποιείται ο αλγόριθμος ιεραρχικής ομαδοποίησης. Ο αριθμός των ομάδων καθορίζεται παίρνοντας υπ' όψιν τις τιμές των συντελεστών Silhouette και Calinski-Harabasz Index.

Παρακάτω περιγράφονται οι 3 διαφορετικές ομαδοποιήσεις:

- Ομαδοποίηση στο πρώτο σύνολο δεδομένων το οποίο έχει τις περισσότερες εγγραφές. Χρησιμοποιείται επίσης η μεταβλητή Μέση αυστηρότητα Μέτρων.
- Ομαδοποίηση στο δεύτερο σύνολο δεδομένων ώστε να μπορεί να συμπεριληφθεί πληροφορία σχετικά και με τις προσωπικές δραστηριότητες εβδομαδιαία. Εκτός των πιο πάνω μεταβλητών χρησιμοποιείται η Μέση αυστηρότητα Μέτρων και η Διαφορά των ωρών προσωπικών δραστηριοτήτων κάθε εβδομάδα.
- Ομαδοποίηση χρησιμοποιώντας μόνο τις εγγραφές 1 χώρας. Σκοπός είναι κυρίως να ελεγχθεί αν η μελέτη 1 συγκεκριμένης χώρας έχει διαφορετικά αποτελέσματα από τις μελέτες που χρησιμοποιούν και τις 24 χώρες, και για να υπάρχει μεγαλύτερη ευελιξία στις δοκιμές σχετικά με τις μεταβλητές που θα χρησιμοποιηθούν και τις υπερπαραμέτρους του δικτύου SOM, αφού ο χρόνος εκπαίδευσης μειώνεται αισθητά λόγω των σημαντικά λιγότερων δεδομένων. Η χώρα που επιλέγεται είναι η Ελλάδα. Χρησιμοποιούνται οι εγγραφές που

σχετίζονται με αυτήν και βρίσκονται στο δεύτερο σύνολο δεδομένων, 443 σε πλήθος. Εκτός των πιο πάνω μεταβλητών χρησιμοποιείται και η Διαφορά των ωρών προσωπικών δραστηριοτήτων κάθε εβδομάδα. Σε αυτή την περίπτωση, το μέγεθος του δικτύου SOM είναι 8x8, δηλαδή αποτελείται από 64 νευρώνες. Η αύξηση των νευρώνων χωρίς υπερβολική αύξηση του χρόνου εκτέλεσης είναι δυνατή λόγω του μικρότερου πλήθους δεδομένων.

4.5 Ταξινόμηση Επιπέδων Κατάθλιψης

Στο κεφάλαιο αυτό θα αναλύσουμε τη μεθοδολογία που ακολουθήθηκε για την δημιουργία ενός μοντέλου που να προβλέπει το επίπεδο σοβαρότητας της κατάθλιψης ενός ατόμου.

Μερικές βασικές λεπτομέρειες αυτού του στόχου είναι οι εξής :

- Να περιέχεται η πληροφορία και η επίδραση της πανδημίας Covid-19
- Να μην υπάρχει ανάγκη για παροχή ευαίσθητων πληροφοριών
- Ιδιαίτερη έμφαση στο Recall της ευπαθούς ομάδας (υψηλά επίπεδα κατάθλιψης) του μοντέλου

4.5.1 Επιλογή Μοντέλων

Έχοντας πλέον κάποιες βασικές αρχές στις οποίες θέλουμε να στηριχθεί το μοντέλο μας, θα συνεχίσουμε στις τεχνικές λεπτομέρειες και θα αναλύσουμε το κριτήριο επιλογής κάθε βήματος.

Πρώτη παρατήρηση μας είναι η πληθώρα (γραμμές) και η πολυπλοκότητα (στήλες) των δεδομένων μας, το οποίο μας παραπέμπει σε τεχνικές Βαθιάς Μάθησης. Επιπλέον, έχοντας ήδη εξερευνήσει την δυναμική διάφορων μοντέλων Μηχανικής Μάθησης (βλ. Παλινδρόμηση), καταλήγουμε στη χρήση Βαθιών Νευρωνικών Δικτύων. Πιο συγκεκριμένα ακολουθάμε 2 προσεγγίσεις :

- Multilayer Perceptron Δίκτυο
- Convolutional Neural Network

Κάθε ένα από αυτά τα δίκτυα αναζητάει με διαφορετικό τρόπο την σχέση εξάρτησης των δεδομένων, και αφού τα δεδομένα μας έχουν υψηλό δείκτη πολυπλοκότητας θα εφαρμόσουμε και τα 2.

4.5.2 Τροποποίηση Δεδομένων

Όπως είναι κατανοητό ο καθορισμός της αρχιτεκτονικής των μοντέλων Βαθιάς Μάθησης εξαρτάται σε μεγάλο βαθμό από τον αριθμό των features. Ένα μοντέλο που έχει λίγα στρώματα και συνεπώς μικρή χωρητικότητα είναι πολύ πιθανό να οδηγηθεί σε underfit αν τα δεδομένα έχουν μεγάλο αριθμό features και αντίστροφα. Επιπλέον πρέπει να υπολογίσουμε ότι οι υπολογιστικές μονάδες που διαθέτουμε έχουν περιορισμένες ικανότητες και συνεπώς ένα πολύ βαθύ μοντέλο θα μας οδηγούσε σε αδυναμία εκπαίδευσης. Συνοψίζοντας τα παραπάνω καλούμαστε να εκτελέσουμε τις κατάλληλες τροποποιήσεις ώστε να διατηρήσουμε στο Dataset μας όσο το δυνατόν περισσότερη πληροφορία με όσο το δυνατόν μικρότερο αριθμό features.

Πρώτο στάδιο είναι η μετατροπή των κατηγορικών μεταβλητών σε 'one-hot' μεταβλητές. Επομένως από τα δεδομένα που περιγράφηκαν στην αντίστοιχη ενότητα Προεπεξεργασία Δεδομένων τροποποιούμε σε 'one-hot' τα εξής:

- Q1 - Φύλλο: 3 στήλες
- Q3 - Στάτους Σχέσης: 3 στήλες
- Q9 - Πεδίο σπουδών: 9 στήλες
- Q10 - Πρόγραμμα Σπουδών: 7 στήλες
- Q12 - Πρώτος χρόνος σπουδών: 3 στήλες
- Q29 - Έχεις διαγνωστεί με Covid-19: 2 στήλες
- Q37 - Ύπαρξη ατόμου για συζήτηση σχετικά με προσωπικά ζητήματα: 2 στήλες

Στη συνέχεια επιλέγουμε να μην χρησιμοποιούμε στο μοντέλο μας όλες τις μεταβλητές που περιγράφουν "Διαφορά", καθώς λόγω πολυσυγγραμικότητας η πληροφορία από την "Αρχική" και "Τελική" τιμή είναι ικανή να περιγράψει στο απόλυτο την πληροφορία της "Διαφοράς". Επομένως δεν διατηρούμε τις εξής μεταβλητές :

- Διαφορά της επάρκειας των οικονομικών πόρων
- Διαφορά των ωρών προσωπικών δραστηριοτήτων κάθε εβδομάδα
- Διαφορά των αριθμών ατόμων με τα οποία συζούσες
- Διαφορά της συχνότητας καπνίσματος
- Διαφορά των τσιγάρων που κάπνιζες καθημερινά

- Διαφορά της συχνότητας με την οποία έπινες περισσότερα από 6 ποτήρια αλκοόλ
- Διαφορά της συχνότητας με την οποία έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά
- Διαφορά της συχνότητας με την οποία έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά

Επόμενη κίνηση είναι η αφαίρεση των ερωτήσεων που σχηματίζουν το σκορ κατάθλιψης και μοναξιάς, δηλαδή οι μεταβλητές Q38_a έως Q38_n. Στη συνέχεια διατηρούμε τις ερωτήσεις που αφορούν το Ακαδημαϊκό Στρες και Ικανοποίηση και αφαιρούμε τις μεταβλητές Σκόρ Στρες και Σκόρ Ικανοποίησης.

Με αυτά τα βήματα έχουμε μια αρχική προσέγγιση ενός λειτουργικού συνόλου δεδομένων. Παρόλα αυτά, προχωράμε σε επιπλέον αφαιρέσεις στηλών οι οποίες θα βασιστούν στα εξής κριτήρια :

- Χαμηλές τιμές στον Correlation Map.
- Χαμηλές τιμές στον στατιστικό έλεγχο X^2
- Ανάδειξη μη στατιστικής σημαντικής διαφοράς με τον στατιστικό έλεγχο ANOVA (One way Analysis of Variance)

Οι τελικές αποφάσεις σχετικά με τους στατιστικούς ελέγχους παρουσιάζονται στην αντίστοιχη ενότητα των αποτελεσμάτων (Ταξινόμηση Επιπέδου Κατάθλιψης). Τέλος να αναφέρουμε ότι το πρόβλημα που θα μας απασχολήσει έχει 2 target groups. Πρώτο είναι όλες οι χώρες που συμμετείχαν στο Dataset και δεύτερο είναι όλες η πρόβλεψη κατάθλιψης για τους φοιτητές του Βελγίου. Η απόφαση αυτή πάρθηκε από το γεγονός ότι το Βέλγιο παρέχει σημαντικό πλήθος φοιτητών της έρευνας που πραγματοποιήθηκε με 20587 φοιτητές, ενώ ακολουθεί η Ιταλία με αρκετή διαφορά και 9090 φοιτητές. Συνεπώς θέλαμε να εξετάσουμε την επίδραση του τοπικότητας των απαντήσεων και αν μπορεί να οδηγήσει σε καλύτερα αποτελέσματα. Επομένως κατασκευάζουμε 2 Datasets , ένα με τα βήματα που περιγράφηκαν παραπάνω, και ένα με τα βήματα που περιγράφηκαν παραπάνω και έξτρα φίλτρο την χώρα Βέλγιο.

4.5.3 Επιλογή κλάσεων ταξινόμησης

Στο επόμενο βήμα, έχοντας τα δεδομένα μας, πρέπει να τα διακριτοποιήσουμε το Σκορ Κατάθλιψης σε κλάσεις για ταξινόμηση. Η πρώτη μας επιλογή είναι ο χωρισμός σε 3 κλάσεις, με το παρακάτω σκορ κατάθλιψης :

- $[0, 7]$ - Κλάση 0: Μη ευπαθής ομάδα, 28239 στον κόσμο/ 6080 στο Βέλγιο

- $[8, 12]$ - Κλάση 1: Μικρή Ανησυχία, 29549 στον κόσμο/ 7256 στο Βέλγιο
- $[13, 24]$ - Κλάση 2: Ευπαθής ομάδα, 27673 στον κόσμο/ 7251 στο Βέλγιο

Κριτήριο της επιλογής μας υπήρξε η προσεγγιστικά ισόποση κατανομή δεδομένων σε αυτές τις κλάσεις. Επομένη μας επιλογή υπήρξε η εξής :

- $[0, 7]$ - Κλάση 0: Μη ευπαθής ομάδα, 28239 στον κόσμο/ 6080 στο Βέλγιο
- $[8, 24]$ - Κλάση 1: Ευπαθής ομάδα, 57222 στον κόσμο/ 14507 στο Βέλγιο

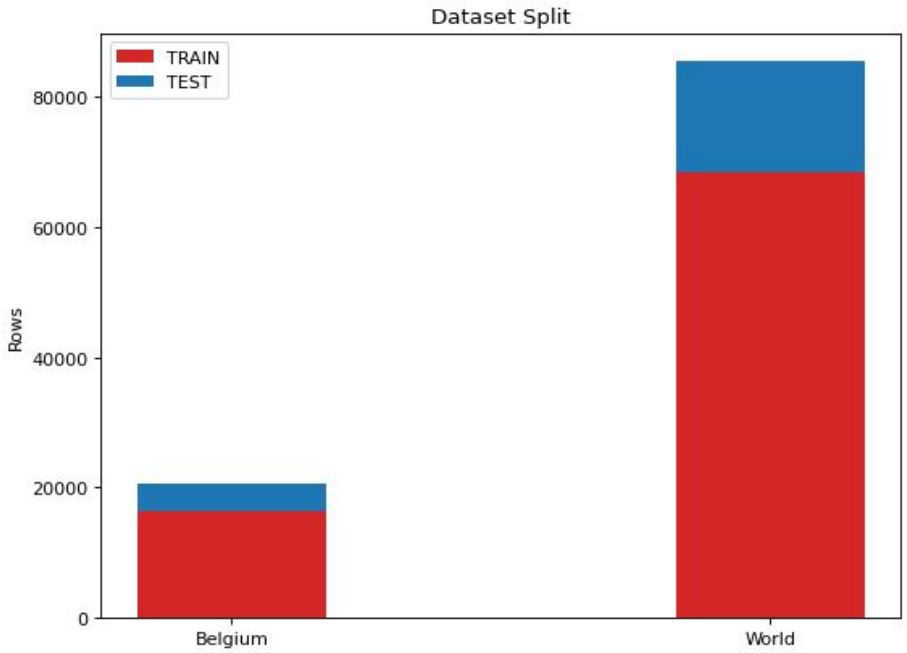
Το κριτήριο επιλογής του 8 ως cut-off value στηρίχθηκε σε αντιστοιχία μελέτης [16] όπου στο CES-D 10 επιλέγεται το 10, καθώς στη βιβλιογραφία δεν υπάρχει ξεκάθαρο cut-off value για την κλίμακα CES-D 8 που χρησιμοποιείται στην συγκεκριμένη έρευνα.

4.5.4 Ανισορροπία Δεδομένων

Παρατηρούμε ότι στη δυαδική ταξινόμηση δεν έχουμε κοντινό αριθμό φοιτητών στις 2 κλάσεις, γεγονός που μπορεί να οδηγήσει σε προβλήματα τόσο στο κομμάτι της εκπαίδευσης των δεδομένων μας όσο και στο κομμάτι της επεξήγησης των αποτελεσμάτων μας. Για να αντιμετωπίσουμε αυτή τη κατάσταση εφαρμόζουμε την τεχνική Ensemble different resampled datasets [17]. Πιο συγκεκριμένα για τα Datasets μας έχουμε :

	Κλάση 0	Κλάση 1	Αναλογία
Belgium_Train1	3891	4165	0.93
Belgium_Train2	3891	4228	0.92
Belgium_Train3	3891	4228	0.92
Belgium_Test	2189	1886	1.16
World_Train1	19767	24319	0.81
World_Train2	19767	24319	0.81
World_Test	8472	8584	0.99

Παράλληλα διατηρούμε την αναλογία 75%-25% στο Train-Test σετ με το αποτέλεσμα να φαίνεται στο διάγραμμα :



Σχήμα 2: Dataset Split

Πραγματοποιήθηκε και επιπλέον split για τη δημιουργία Validation Set απο το Train Set της τάξης του 80%-20%.

4.5.5 Προβλεπτική ικανότητα τελικών μοντέλων

Έχουμε ορίσει 2 διαφορετικά προβλήματα (binary-multiclass classification), διαφορετικά datasets και 2 διαφορετικές αρχιτεκτονικές μοντέλων, συνεπώς θα προκύψουν αρκετά μοντέλα από τα οποία θα πρέπει να επιλέξουμε τα επικρατέστερα για κάθε πρόβλημα και Dataset. Η τελική απόφαση θα παρθεί μέσα από μια συλλογή μοντέλων, μοντέλα που είτε θα προβλέπουν μόνα τους είτε με τεχνική ensemble θα προβλέπουν συλλογικά. Επομένως έχουμε 2 κατηγορίες μοντέλων , τα ανεξάρτητα και τα ομαδοποιημένα. Βασικά κριτήρια αξιολόγησης θα είναι οι μετρικές accuracy, recall, precision, f-score. Τέλος ο τρόπο με τον οποίο θα προκύψει η τελική πρόβλεψη στα ensemble μοντέλα του binary classification θα προκύψει με 4 διαφορετικούς τρόπους:

- Mean Value
- Maximum Value
- Minimum Value
- Dynamic weighting based on certainties

Η τελευταία μέθοδος περιγράφεται με τους εξής τύπους :

$$\bar{y}(x) = \sum_{net=1}^{m} w_{net} \cdot y_{nt}(x)$$

όπου,

$$w_{net} = \frac{C_{net}}{\sum_{i=1}^{m} C_i}, \qquad C_{net} = \begin{array}{l} y_{net}, if y_{net} \geqslant 0,5 \\ 1 - y_{net}, if y_{net} < 0,5 \end{array}.$$

4.5.6 Βελτιστοποίηση αρχιτεκτονικών

Η εύρεση της βέλτιστης αρχιτεκτονικής ενός μοντέλου Βαθιάς μάθησης, είναι ένα πρόβλημα αναζήτησης σε έναν μεγάλο πολυδιάστατο χώρο. Ο τρόπος με τον οποίο εμείς θα πραγματοποιήσουμε αυτή την αναζήτηση είναι η μέθοδος Particle Swarm Optimization ή Βελτιστοποίηση Σμήνους Σωματιδίων.

Η Βελτιστοποίηση Σμήνους Σωματιδίων (ΒΣΣ) [18] είναι ένας αλγόριθμος βελτιστοποίησης ο οποίος προτάθηκε το 1995 από τον James Kennedy και τον Russell C. Eberhart και εμπνεύστηκε από τη συμπεριφορά ζώων να σχηματίζουν ομάδα ή σμήνος και να μοιράζονται πληροφορίες μέσα σε αυτή, με σκοπό την καλύτερη επιβίωση τους. Στόχος της ΒΣΣ είναι να βρει τις κατάλληλες τιμές των δοθέντων μεταβλητών - στην συγκεκριμένη περίπτωση είναι υπερπαράμετροι - για τις οποίες επιτυγχάνεται η βέλτιστη απόδοση, δηλαδή μέγιστη ή ελάχιστη -όταν η μετρική είναι η απώλεια- τιμή μιας συνάρτησης εφαρμογής (fitness) $f(x)$.

Ο αλγόριθμος ΒΣΣ μπορεί να χαρακτηριστεί σαν μια αναζήτηση σε έναν ν-διάστατο χώρο, όπου ν είναι ο αριθμός των υπερπαραμέτρων προς βελτίωση. Για να αναλυθεί το μαθηματικό υπόβαθρο της μεθόδου, παρατίθενται κάποια χαρακτηριστικά της αναζήτησης. Αρχικά η αναζήτηση γίνεται από έναν αριθμό σωματιδίων που ορίζεται εξαρχής. Για κάθε σωματίδιο πρέπει να σημειώνονται:

- οι υπερπαράμετροι που οδήγησαν στην καλύτερη ατομική επίδοση Pbest
- Οι υπερπαράμετροι που οδήγησαν στην καλύτερη ομαδική επίδοση Gbest (πριν από κάθε κίνηση τα σωματίδια επικοινωνούν έτσι ώστε να ενημερώσουν για κάποια πιθανή αλλαγή της Gbest).
- η ταχύτητα (velocity) του, όταν σημειώθηκαν αυτές οι υπερπαράμετροι

Για να κατανοηθεί μαθηματικά ο αλγόριθμος ΒΣΣ, θα πρέπει να οριστούν κάποια μεγέθη. Αρχικά ορίζεται ένα διάνυσμα από τις υπερπαραμέτρους $a_i, b_i, c_i, ...$ που τίθενται προς βελτιστοποίηση

$$\vec{X}^t = [a_i, b_i, c_i, ...]$$

Κάθε σωματίδιο έχει ένα δικό του "προσωπικό" διάνυσμα με τυχαία αρχικοποίηση και ξεχωριστή εξέλιξη στις τιμές των υπερπαραμέτρων, το οποίο δείχνει τη θέση του στον πολυδιάστατο χώρο αναζήτησης. Κάθε σωματίδιο έχει επίσης μια δικιά του ταχύτητα V^t η οποία ορίζεται ως εξής :

$$V^{t+1} = w * V^t + c_1 r_1 * (Pbest - X^t) + c_2 r_2 * (Gbest - X^t) \tag{1}$$

Η νέα θέση κάθε σωματιδίου ανανεώνεται με τον εξής τρόπο :

$$X^{t+1} = X^t + V^{t+1} \tag{2}$$

Ο αλγόριθμος τερματίζει όταν ικανοποιηθεί το κριτήριο τερματισμού, το οποίο ονομάζεται T_{max} και δηλώνει τον μέγιστο αριθμό αναζητήσεων που θα πραγματοποιήσουν. Υπάρχουν και άλλα λιγότερα διαδεδομένα κριτήρια όπως μείωση διαφοράς στη βελτιστοποίηση ή μειωμένη κίνηση στο χώρο.

Algorithm 1: Βελτιστοποίηση Σμήνους Σωματιδίων

for *each particle* **do**
  Initialize X randomly
  Initialize V randomly
  Evaluate the fitness $f(X)$
  Initialize Pbest with a copy of X
end
Initialize Gbest with a copy of the best X
$t = 0$
while $t < T_{max}$ **do**
  for *each particle* **do**
    Update V according to (1)
    Update X according to (2)
    Evaluate the fitness $f(X)$
    $Pbest \leftarrow X$ if $f(Pbest) < f(X)$
    $Gbest \leftarrow X$ if $f(Gbest) < f(X)$
  end
end

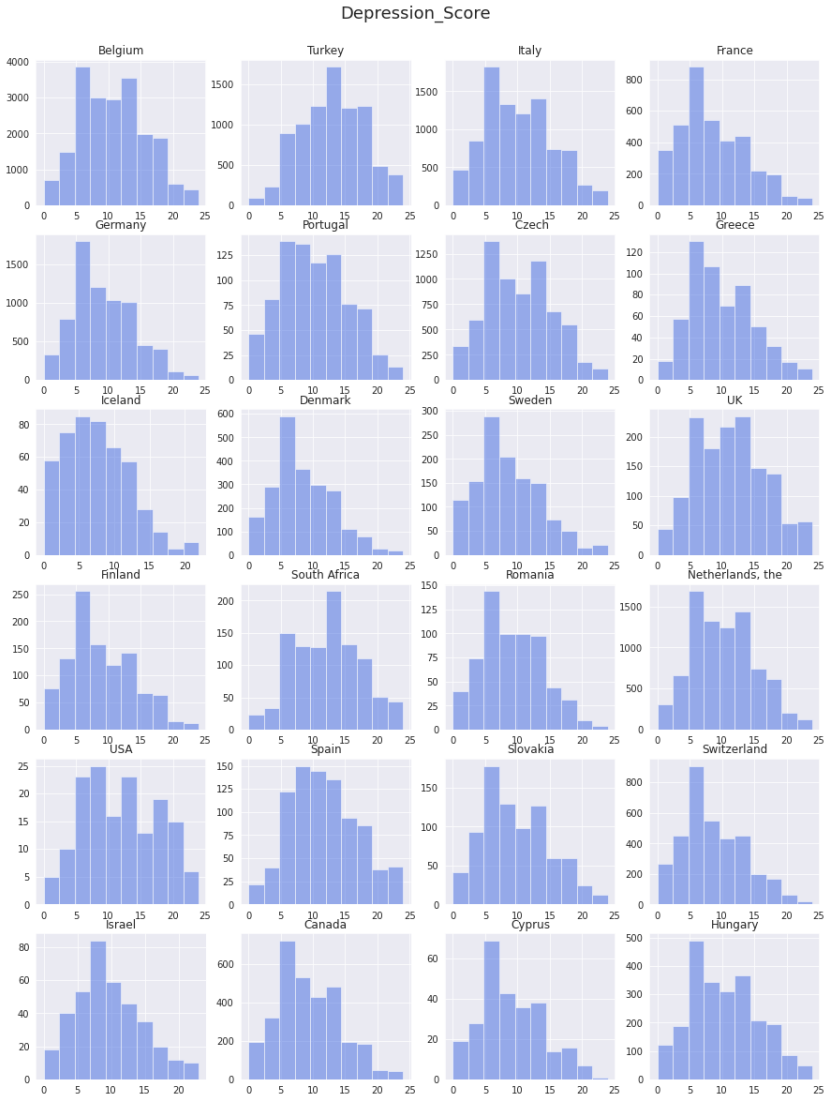
Σύνοψη: Έχουμε αναφέρει θεωρητικά κάθε κομμάτι του pipeline που θα χρησιμοποιήσουμε για την δημιουργία των τελικών μοντέλων μας. Επίσης, έχουμε επισημάνει ορισμένες αλλαγές που πραγματοποιήθηκαν στο Dataset, έτσι ώστε να υπάρχει καλύτερη εποπτεία του προβλήματος .Τέλος στο κεφάλαιο Αποτελέσματα και στην αντίστοιχη ενότητα Ταξινόμηση Επιπέδου Κατάθλιψης αναλύουμε πρακτικά κάθε κομμάτι του pipeline και θα παρουσιάσουμε τα αποτελέσματα κάθε βήματος μέχρι και την αξιολόγηση των τελικών μοντέλων που αποτελούν την υποψήφια εφαρμογή για γρήγορη διάγνωση του επιπέδου κατάθλιψης.

5 Αποτελέσματα

5.1 Στατιστικοί Έλεγχοι

5.1.1 Εύρεση εξαρτήσεων μεταξύ της κατάθλιψης και των κατηγορικών μεταβλητών

Πραγματοποιείται One way Analysis of Variance για τις κατηγορικές μεταβλητές που αναφέρονται στην μεθοδολογία. Για την εφαρμογή του συγκεκριμένου ελέγχου έγινε έλεγχος όλων των υποθέσεων πριν την εφαρμογή του ελέγχου. Ενδεικτικά παρουσιάζεται η υπόθεση της κανονικότητας για την μεταβλητή Χώρα, δηλαδή ένα ιστόγραμμα των επιπέδων κατάθλιψης για κάθε χώρα.



Σχήμα 3: Υπόθεση κανονικότητας

Σύμφωνα με τα αποτελέσματα του ελέγχου, προκύπτει ότι για όλες τις μεταβλητές χώρα, ο τομέας σπουδών (Q9), το πρόγραμμα σπουδών (Q10), η ερωτική κατάσταση (Q3), το φύλο (Q1) και το ανώτατο εκπαιδευτικό ίδρυμα (Q11), υπάρχουν σοβαρές ενδείξεις απόρριψης της μηδενικής υπόθεσης. Συνεπώς υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων επιπέδων κατάθλιψης σε τουλάχιστον δύο κατηγορίες.

Σχετικά με τις μεταβλητές σε likert scale, που αναφέρονται στην μεθοδολογία, ο έλεγχος X^2 δείχνει ότι υπάρχουν ισχυρές ενδείξεις απόρριψης της μηδενικής υπόθεσης, δηλαδή καθεμιά από αυτές τις μεταβλητές είναι εξαρτημένη με το επίπεδο της κατάθλιψης. Στους ακόλουθους πίνακες 1,2 παρουσιάζεται ο συντελεστής Cramer's V, ο οποίος δείχνει τον βαθμό συσχέτισης με την μεταβλητή της κατάθλιψης.

Πίνακας 1: Cramer's V

Μεταβλητή	Cramer's V
AlcoholOftenb - Q23_a	0.029
ModPhysiActb - Q26_a	0.032
Education(Father) - Q7	0.034
Education(Mother) - Q6	0.039
AlcoholOftena - Q23_b	0.039
Measurement_Adherence - Q34	0.044
VigoPhysiActb - Q25_a	0.047
TobaccoOftena - Q21_b	0.051
Importance of Study - Q15	0.052
TobaccoOftenb - Q21_a	0.053
FamilyContact - Q36a	0.070
On time government info - Q43_a	0.072
Comprehensive government info - Q43_b	0.073
Sufficient Medical Supplies C19 - Q32	0.074
DiscussStudies - Q39_a	0.074
SevereInfectionWorries - Q30b_2	0.074
FriendsContact - Q36b	0.084
DiscussStudies - Q39_b	0.085
InfectionWorries - Q30b_1	0.086
Satisf_f - Q41_f	0.089
VigoPhysiActa - Q25_b	0.099
CoverMonthlyCostsb - Q17_a	0.103

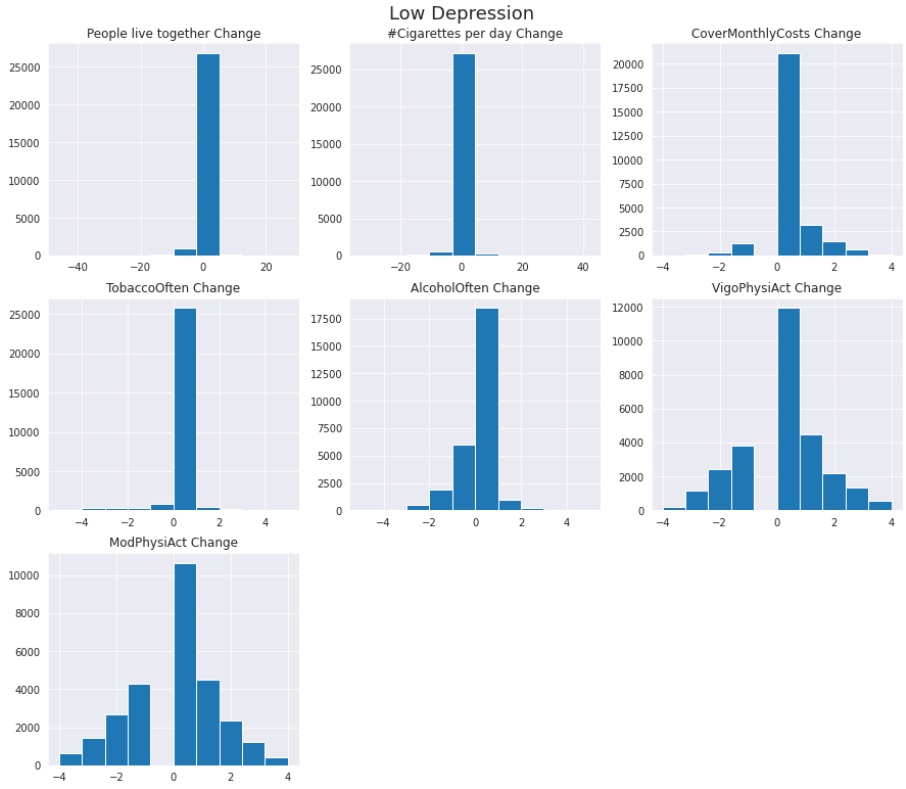
Πίνακας 2: Cramer's V: Top 10

Μεταβλητή	Cramer's V
Satisf_g - Q41_g	0.111
ModPhysiActa - Q26_b	0.116
Satisf_d - Q4_d	0.118
Stress_b - Q41_b	0.121
Stress_a - Q41_a	0.127
CoverMonthlyCostsa - Q17_b	0.140
Satisf_h - Q41_h	0.143
Stress_c - Q41_c	0.200
Stress_e - Q41_e	0.226
DiscussPersonal - Q37	0.246

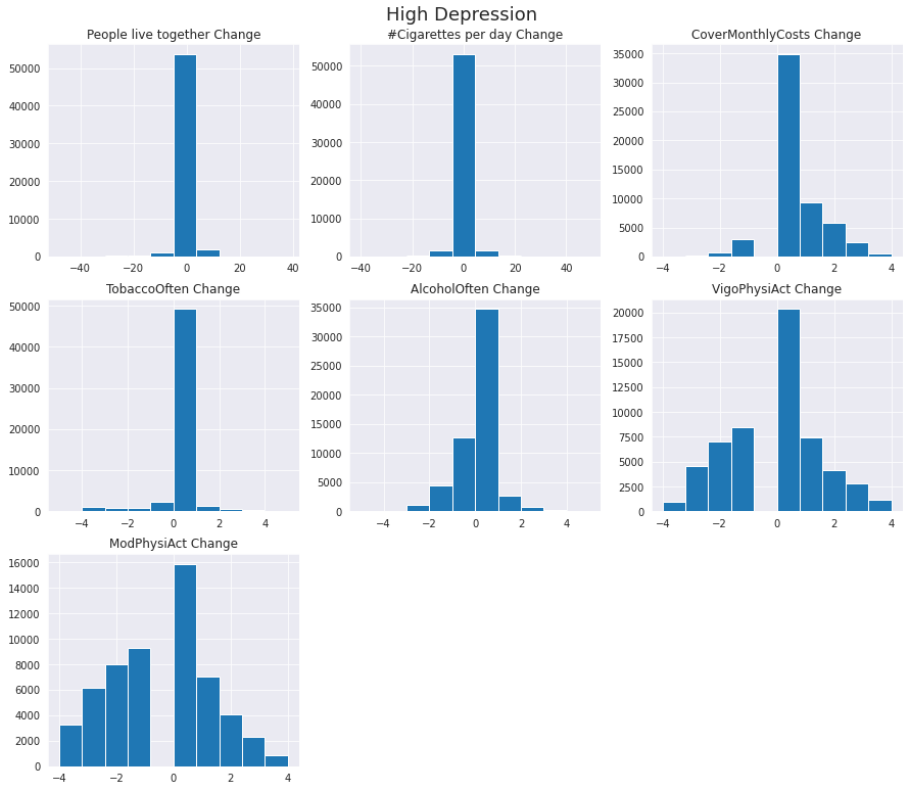
Παρατηρείται ότι οι ερωτήσεις που αφορούν τα μέτρα που έλαβαν τα πανεπιστήμια,οι ερωτήσεις σχετικές με την κάλυψη των μηνιαίων εξόδων και η συζήτηση σχετικά με τα προσωπικά ζητήματα του φοιτητή παρουσιάζουν την μεγαλύτερη συσχέτιση με τα επίπεδα κατάθλιψης.

5.1.2 Επιρροή πανδημίας σε άτομα με χαμηλή και υψηλή κατάθλιψη

Αρχικά διαχωρίζεται το σύνολο δεδομένων σε χωρίζεται σε δύο υποσύνολα, εκ των οποίων το ένα αφορά τους φοιτητές με επίπεδα κατάθλιψης μικρότερα ή ίσα του 8 και το άλλο φοιτητές με επίπεδα κατάθλιψης μεγαλύτερα του 8. Τα άτομα με χαμηλά επίπεδα κατάθλιψης είναι 28145 και τα άτομα με υψηλά επίπεδα κατάθλιψης είναι 56923. Λόγω μεγάλου μέγεθος δειγμάτων και για τις δύο ομάδες και βάσει του κεντρικού οριακού θεωρήματος τα δείγματα από έναν πληθυσμό με πεπερασμένη διασπορά προσεγγίζουν την κανονική κατανομή ανεξάρτητα από την κατανομή του πληθυσμού. Ακολούθως παρουσιάζονται τα ιστογράμματα των μεταβλητών σχετικών με τις αλλαγές με την έξαρση του Covid-19.



Σχήμα 4: Ομάδα με χαμηλή κατάθλιψη



Σχήμα 5: Ομάδα με υψηλή κατάθλιψη

Παρατηρείται ότι τα παραπάνω ιστογράμματα προσεγγίζουν την κανονική κατανομή. Εν συνεχεία, ελέγχονται οι διασπορές των παραπάνω μεταβλητών ανάμεσα στις δύο ομάδες για επιλογή του αντίστοιχου ελέγχου.

Πίνακας 3: Διασπορά μεταβλητών

Μεταβλητή	Υψηλή κατάθλιψη	Χαμηλή κατάθλιψη
People live together Change	7.87	6.04
#Cigarettes per day Change	7.41	2.91
CoverMonthlyCosts Change	1.05	0.66
TobaccoOften Change	0.73	0.42
AlcoholOften Change	0.79	0.62
VigoPhysiAct Change	2.78	2.19
ModPhysiAct Change	3.32	2.45

Οι μεταβλητές People live together Change, #Cigarettes per day Change έχουν μεγαλύτερη απόκλιση στις διασπορές του και γι᾽ αυτό χρησιμοποιείται ο 2ος έλεγχος ενώ όλες οι υπόλοιπες μεταβλητές έχουν περίπου ίσες διασπορές ανάμεσα στις δύο ομάδες και γι᾽ αυτό χρησιμοποιείται ό πρώτος έλεγχος. Λαμβάνουμε τα εξής αποτελέσματα, ότι δηλαδή για τις μεταβλητές AlcoholOften Change και #Cigarettes per day Change δεν υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στις δύο ομάδες ενώ για τις υπόλοιπες μεταβλητές προκύπτει ότι έχουμε σοβαρές ενδείξεις απόρριψης της μηδενικής υπόθεσης, δηλαδή ισότητας των μέσων. Για τα συγκεκριμένα αποτελέσματα, χρησιμοποιείται επίπεδο σημαντικότητας 0.05.

Πίνακας 4: Μέσος όρος μεταβλητών

Μεταβλητή	Υψηλή κατάθλιψη	Χαμηλή κατάθλιψη
People live together Change	-0.04	0.05
#Cigarettes per day Change	-0.07	-0.08
CoverMonthlyCosts Change	0.45	0.22
TobaccoOften Change	-0.11	-0.08
AlcoholOften Change	-0.36	-0.35
VigoPhysiAct Change	-0.20	0.07
ModPhysiAct Change	-0.55	-0.07

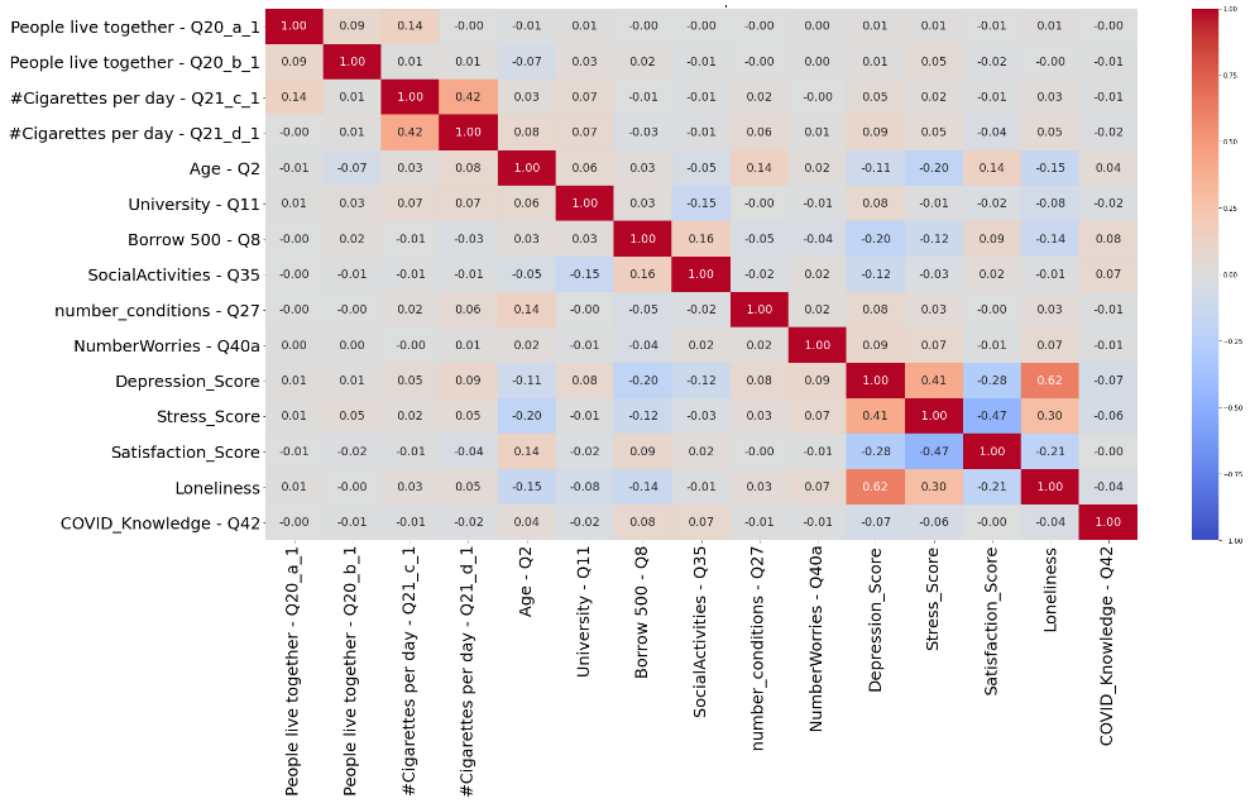
Σύμφωνα με τον πίνακα 4, γίνεται αντιληπτό ότι οι μέσοι όροι των μεταβλητών AlcoholOften Change και #Cigarettes per day Change διαφέρουν κατά 0.01, συνεπώς είναι εύλογο το συμπέρασμα του στατιστικού ελέγχου ότι δεν διαφέρουν ανάμεσα στις δύο ομάδες. Για τις υπόλοιπες μεταβλητές διακρίνονται δύο περιπτώσεις:

- Πιο έντονη αλλαγή ανάμεσα στις δύο ομάδες, δηλαδή αρνητικός μέσος όρος για την μία ομάδα και κατ' επέκταση η συγκεκριμένη συνήθεια επηρεάστηκε αρνητικά από την πανδημία και αντίστοιχα θετικός μέσος όρος όρος για την άλλη ομάδα, δηλαδή η συνήθεια δεν επηρεάστηκε από την πανδημία. Σε αυτή την κατηγορία, οι μεταβλητή που δηλώνει πόσο έντονα οι φοιτητές κάνουν έντονη σωματική άσκηση πριν και μετά την καραντίνα φαίνεται ότι επηρέασε τους φοιτητές με υψηλά επίπεδα κατάθλιψης,δηλαδή η καραντίνα είχε σαν αντίκτυπο να μειώσουν την έντονη σωματική άσκηση. Ενώ οι φοιτητές με χαμηλά επίπεδα κατάθλιψης παρουσίασαν πολύ μικρή αλλαγή.

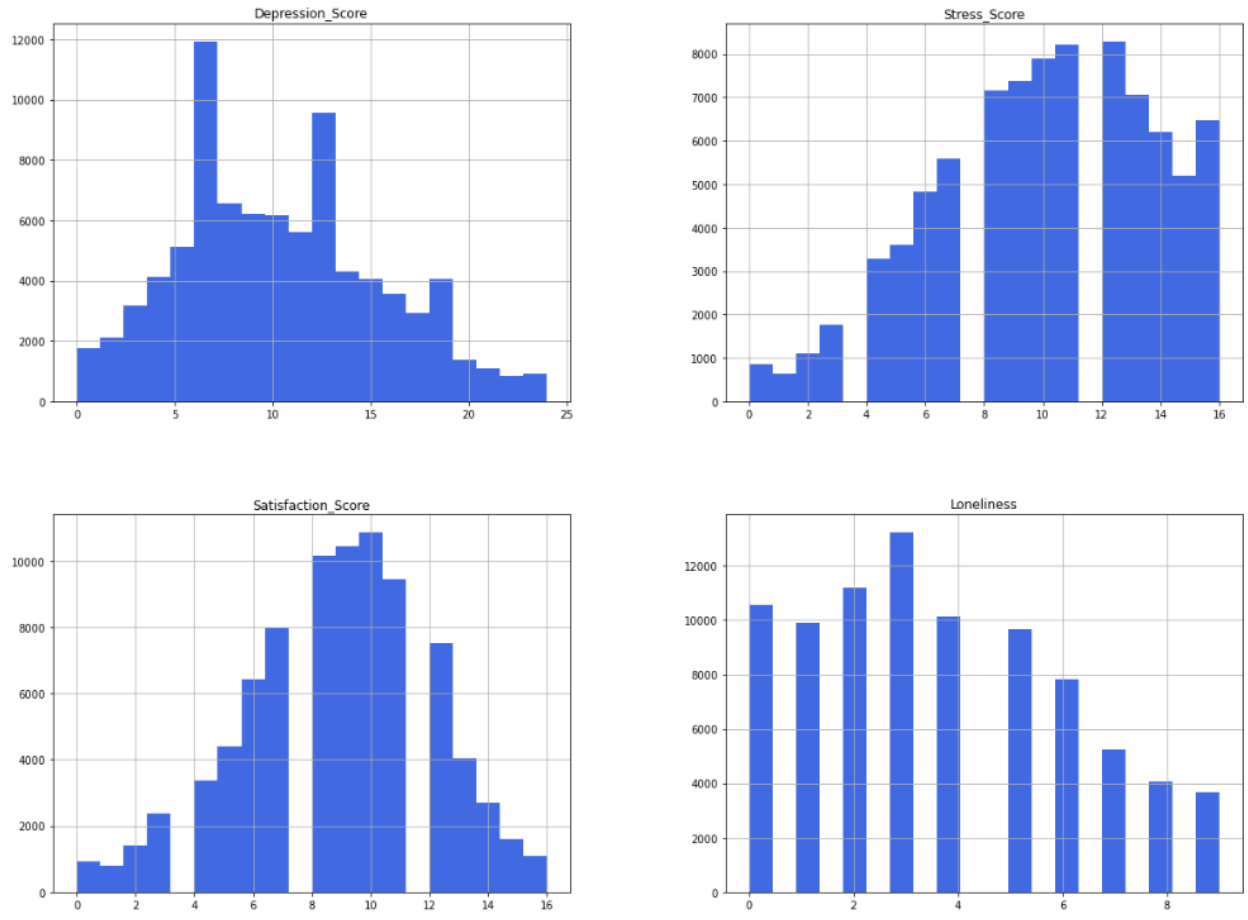
- Ίδιο είδος αλλαγής ανάμεσα στις δύο ομάδες με μεγαλύτερη απόκλιση ανάμεσα στους μέσους όρους των δύο ομάδων. Πιο συγκεκριμένα ο μέσος όρος ανάμεσα στις δύο ομάδες είναι ίδιου προσήμου με μεγάλη διαφοροποίηση. Παρατηρείται ότι η σωματική άσκηση μέτριας έντασης μειώθηκε σε μεγαλύτερο βαθμό σε άτομα με υψηλά επίπεδα κατάθλιψης από ότι σε άτομα με χαμηλά επίπεδα κατάθλιψης. Επιπλέον παρουσιάζει ενδιαφέρον ότι οι φοιτητές με υψηλά επίπεδα κατάθλιψης επιπλέον εμφάνισαν μεγαλύτερη θετική αλλαγή στο ότι έχουν επαρκείς οικονομικούς πόρους από ότι οι φοιτητές με χαμηλότερα επίπεδα κατάθλιψης. Τέλος, για τις ερωτήσεις σχετικά με την αλλαγή ως προς τα άτομα που συζούν μαζί και ως προς την συχνότητα που καπνίζουν πριν και μετά την καραντίνα φαίνεται ότι ήταν αρνητικά ως προς τις δύο ομάδες.

5.2 Παλινδρόμηση

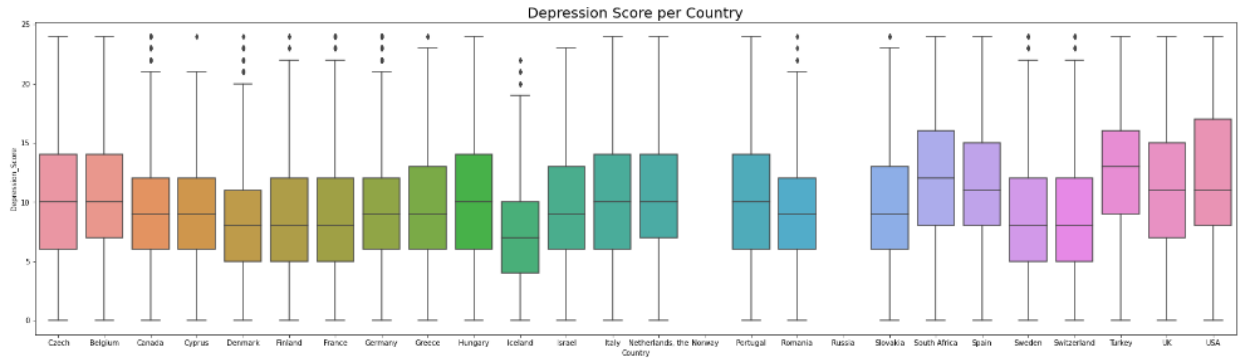
Αρχικά με σκοπό την επιλογή μεταβλητών που είναι και αυτή κομμάτι της επιλογής μοντέλου εξάγουμε έναν πίνακα συσχετίσεων για όλες τις αριθμητικές μεταβλητές για να ανακαλύψουμε τυχόν συσχετίσεις που μπορούν να μας βοηθήσουν κυρίως στην παλινδρόμηση αλλά και γενικότερα στην εξερεύνηση του σετ δεδομένων και ενδεχομένως να αποκτήσουμε κάποιας μορφής επεξηγησιμότητα από την διαδικασία.



Βλέπουμε λοιπόν στον παραπάνω πίνακα την ανά δύο συσχέτιση μεταξύ των αριθμητικών μεταβλητών μας. Όπως φαίνεται οι μεταβλητές μεταξύ τους έχουν από καθόλου έως ελάχιστη συσχέτιση μεταξύ τους παρά μόνο σε μία μικρή ομάδα συναντάμε αξιοσημείωτη συσχέτιση και αυτή η ομάδα είναι οι μεταβλητές depression score, stress score, satisfaction score και loneliness. Το γεγονός αυτό είναι αρκετά λογικό και θα δούμε και παρακάτω στα αποτελέσματα του ταξινομητή μας ότι αυτό επιβεβαιώνεται από τα αποτελέσματα της παλινδρόμησης.
Στη συνέχεια θέλουμε να δούμε πώς είναι κατανεμημένοι αυτοί οι 4 δείκτες στον πληθυσμό των φοιτητών και για αυτό το λόγο εξάγουμε τα ιστογράμματα τους.



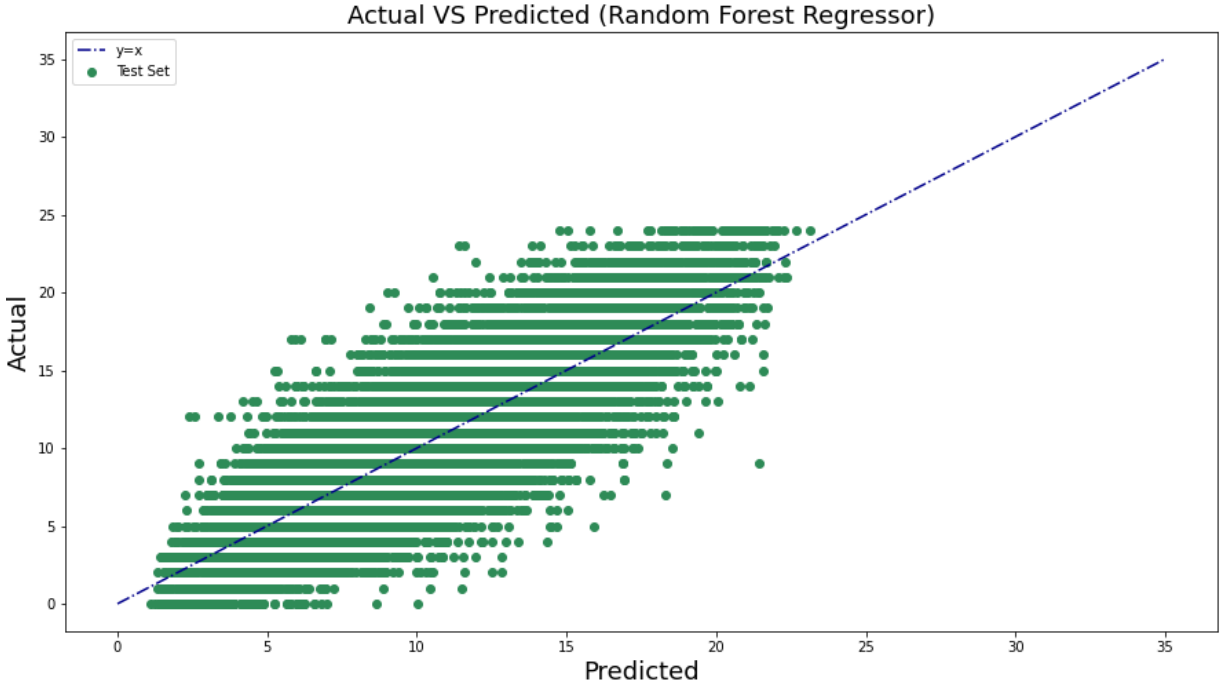
Αυτό το διάγραμμα μας δίνει μία καλύτερη εικόνα για το σετ των δεδομένων και βοηθάει στις αποφάσεις που παίρνουμε. Βλέπουμε για παράδειγμα ότι στο σκορ της κατάθλιψης υπάρχει σημαντική μερίδα του πληθυσμού των φοιτητών με σκορ πάνω από 8 το οποίο στη βιβλιογραφία θεωρείται ένα κρίσιμο κατώφλι. Όσον αφορά το στρες βλέπουμε ότι ο περισσότερος πληθυσμός εμφανίζει μεγάλες τιμές και εκεί. Λόγω της cross-sectional φύσης της μελέτης ωστόσο δεν μπορούμε να εξάγουμε εύρωστα συμπεράσματα για το αν αυτά τα νούμερα είναι από τη φύση τους ψηλά ή αν αυξήθηκαν λόγω της πανδημίας. Ένα προκαταρκτικό συμπέρασμα παρόλα αυτά είναι ότι για οποιοδήποτε λόγο οι φοιτητές εμφανίζουν μεγάλα ποσοστά τόσο κατάθλιψης όσο και στρες. Τώρα ενδέχεται αυτό το γεγονός να είναι λόγω της πανδημίας και των μέτρων περιορισμού της εξάπλωσης, λόγω του ότι το πανεπιστήμιο γενικά είναι στρεσογόνο ή ένας συνδυασμός των δύο.
Στη συνέχεια παρουσιάζουμε ένα boxplot με τα σκορ κατάθλιψης ανά χώρα.



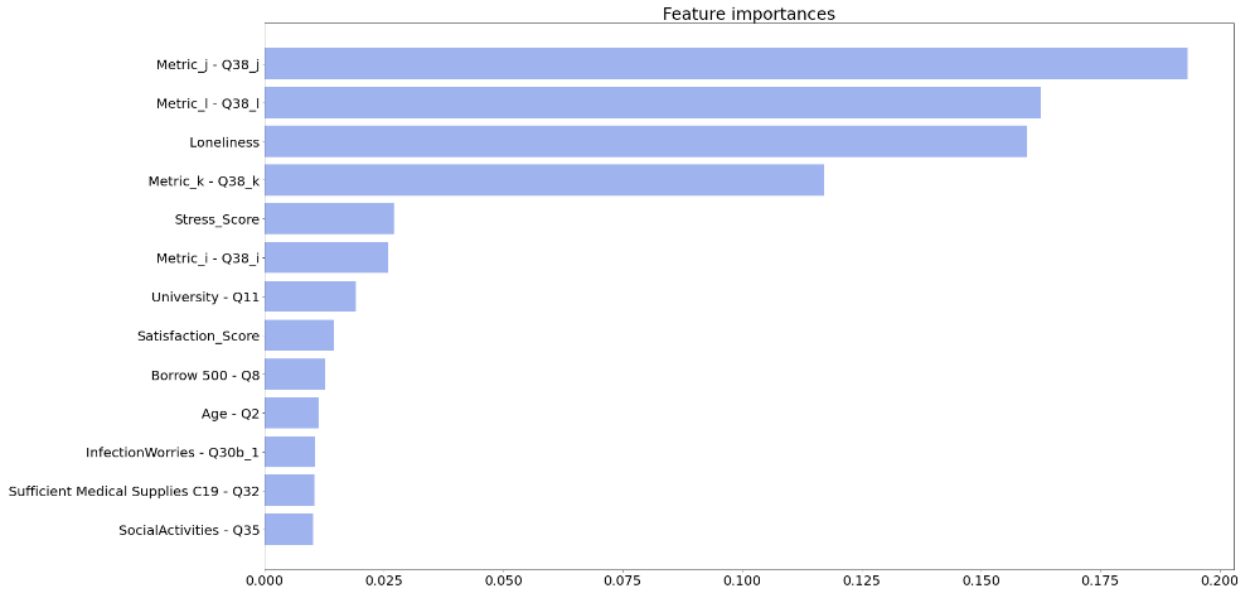
Βλέπουμε ότι κάποιες χώρες όπως η Ισλανδία, η Σουηδία και η Ελβετία έχουν μικρότερα ποσοστά κατάθλιψης σε σχέση με άλλες και χώρες όπως η Νότια Αφρική στην οποία εφαρμόστηκαν αυστηρά μέτρα βλέπουμε μεγαλύτερα ποσοστά.
Έχοντας δει αυτά τώρα θα προχωρήσουμε στα αποτελέσματα του εκτιμητή μας.

	Test Set
MAE	2.14
MSE	7.43
RMSE	2.73
R^2	0.72

Μπορούμε να συμπεράνουμε ότι ο ταξινομητής εν γένει δίνει καλά αποτελέσματα και θα μπορούσε να χρησιμοποιηθεί σε κάποιο ενδεχόμενο lockdown ως δείκτης για την επίπτωση που έχουν τα μέτρα σε νέο πληθυσμό.
Παρακάτω έχουμε ένα διάγραμμα που δείχνει πώς έχουν ταξινομηθεί τα δείγματα στο σετ ελέγχου.



Ουσιαστικά αυτό που βλέπουμε είναι ότι ο ταξινομητής έχει βρει μία γενική τάση αλλά ενδεχομένως στα δεδομένα να λείπει κάτι που επηρεάζει πολύ το αν κανείς εμφανίζει συναισθήματα κατάθλιψης ή όχι. Ενδεχομένως να παίζει ρόλο κάποιου είδους προδιάθεση ή οποία δεν έχει προσμετρηθεί και προφανώς δεν θα μπορούσε να χρησιμοποιηθεί σε ένα τέτοιο μοντέλο pretesting γιατί τότε θα μιλούσαμε για ένα μοντέλο που πάει να κάνει διάγνωση και όχι απλά να φτάσει σε μία προσέγγιση.
Τέλος εξάγουμε τα feature importance για να δούμε ποιες μεταβλητές επηρεάζουν την πρόβλεψή που γίνεται.



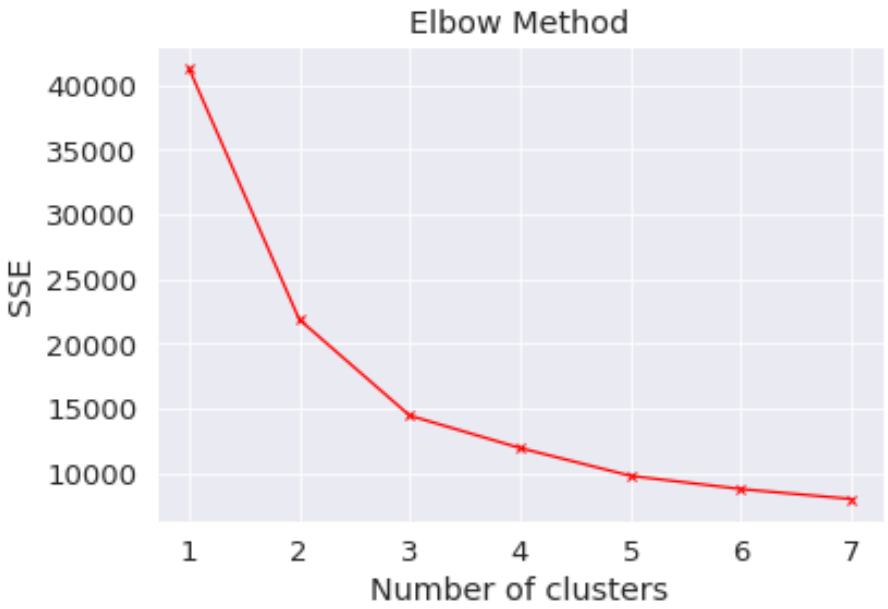
Βλέπουμε ότι μακράν οι πιο σημαντικές μεταβλητές είναι η Q38_i, Q38_l, loneliness, Q38_k. Αυτές αντιστοιχούν στις ερωτήσεις που αναφέρονται σε αισθήματα ηρεμίας και γαλήνης, αισθήματα γενικού εκνευρισμού, τον δείκτη μοναχικότητας και αισθήματα άγχους. Συμπεραίνουμε γενικά ότι υπάρχει σημαντική συσχέτιση μεταξύ άγχους, πλήξης και εκνευρισμού με την κατάθλιψη χωρίς να μπορούμε να υποθέσουμε σχέσης αιτίου-αιτιατού. Λιγότερο σημαντικοί αλλά παρόλα αυτοί αξιόλογοι παράγοντες είναι το στρες (stress score), η πλήξη (Q38_i), το πανεπιστήμιο, ο δείκτης ικανοποίησης, η δυνατότητα να δανειστεί κανείς χρήματα, η ηλικία, η ανησυχία για μόλυνση, η ύπαρξη επαρκών αποθεμάτων στα νοσοκομεία και οι κοινωνικές δραστηριότητες.

5.3 Ομαδοποίηση

5.3.1 Προσέγγιση 1

Ομαδοποίηση των φοιτητών σύμφωνα με τα επίπεδα στρες και κατάθλιψης

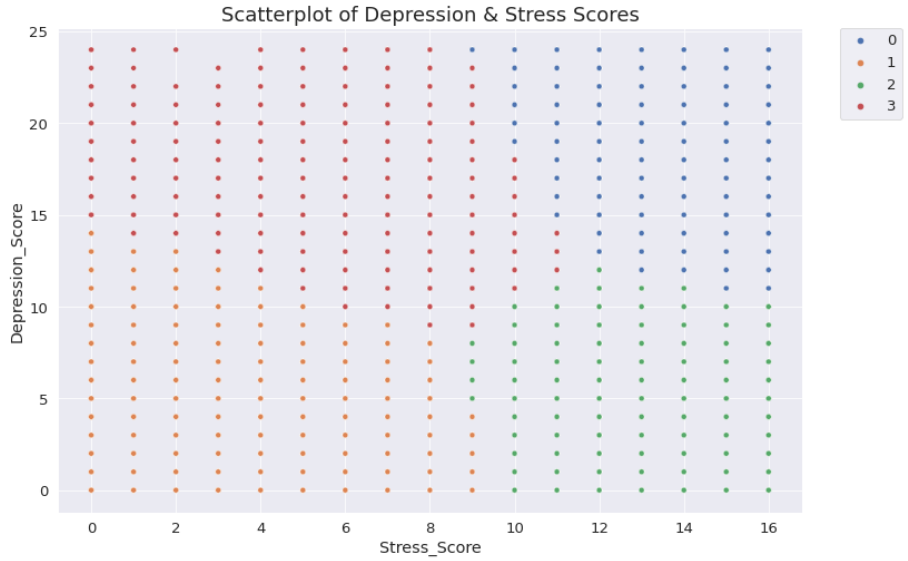
Σύμφωνα με την μέθοδο Elbow, προκύπτει το ακόλουθο διάγραμμα, βάσει του οποίου επιλέγουμε 4 ομάδες αφού η μείωση του κριτηρίου Inertia είναι πολύ μικρή από 3 σε 4 ομάδες.



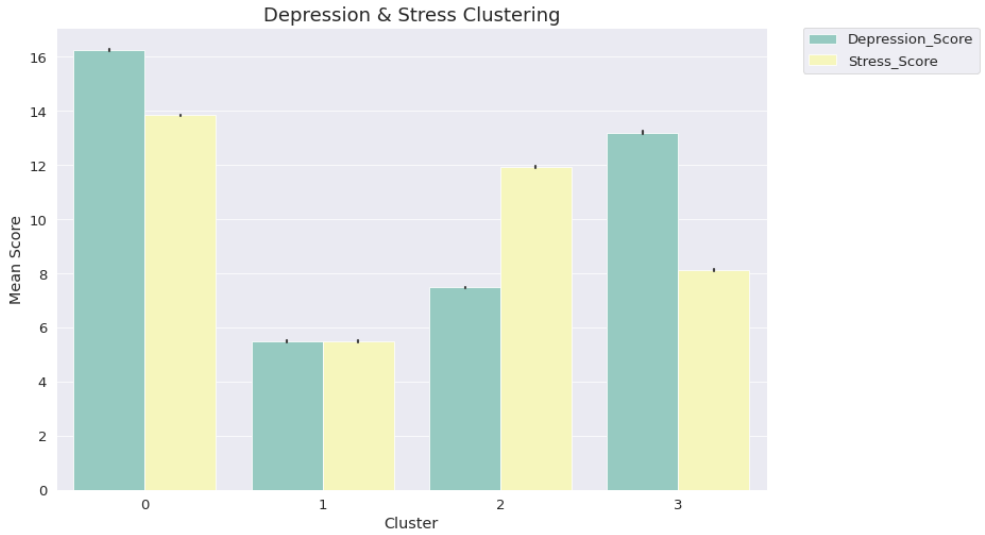
Σχήμα 6: Elbow Μέθοδος πρώτης ομαδοποίησης

Σύμφωνα με τα ακόλουθα διαγράμματα, διακρίνουμε 4 ομάδες από φοιτητές.

- Ομάδα 0: Υψηλή κατάθλιψη/Υψηλό στρες (Μπλε χρώμα)
- Ομάδα 1: Χαμηλή κατάθλιψη/Χαμηλό στρες (Πορτοκαλί χρώμα)
- Ομάδα 2: Χαμηλή κατάθλιψη/Υψηλό στρες (Πράσινο χρώμα)
- Ομάδα 3: Υψηλή κατάθλιψη/ Χαμηλό στρες (Κόκκινο χρώμα)



Σχήμα 7: Επίπεδα στρες με κατάθλιψη



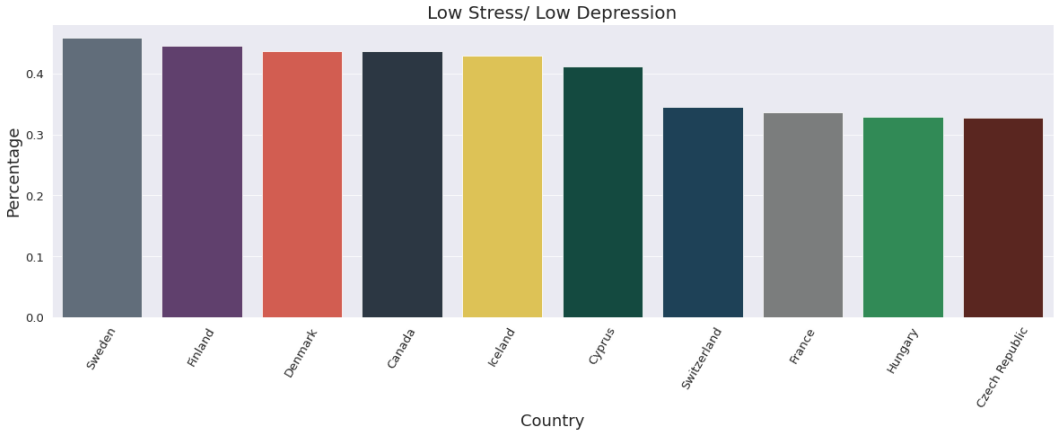
Σχήμα 8: Αποτελέσματα ομαδοποίησης

Για τον έλεγχο διαχωρισιμότητας των ομάδων, στον ακόλουθο πίνακα δίνονται οι συντελεστές Silhouette και Calinski-Harabasz.

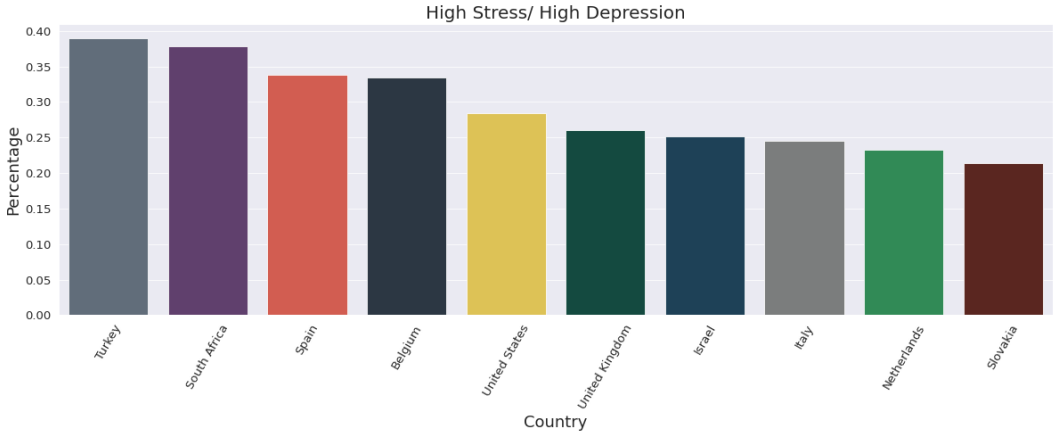
Πίνακας 5: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.37
Calinski-Harabasz Index	71,784

Οι συντελεστές αξιολόγησης που προκύπτουν είναι αρκετοί καλοί και δείχνουν ότι έχει γίνει πολύ καλός διαχωρισμός των ομάδων. Πιο συγκεκριμένα, προκύπτει πολύ μεγάλος συντελεστής Calinski-Harabasz και ο συντελεστής Silhouette είναι 0.37, δηλαδή κοντά στο 1, το οποίο δηλώνει τον ιδιαιτέρως καλό διαχωρισμό. Σύμφωνα με τις ομάδες που δημιουργούνται παρουσιάζονται οι 10 χώρες με το υψηλότερο ποσοστό φοιτητών στις ομάδες 0 και 1.



Σχήμα 9: 10 Κορυφαίες χώρες της ομάδας 1

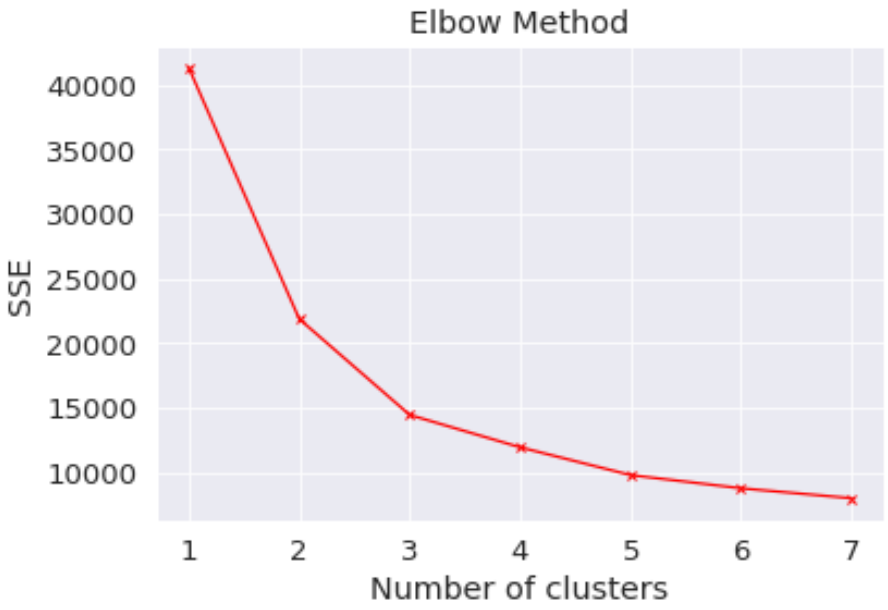


Σχήμα 10: 10 Κορυφαίες χώρες της ομάδας 0

Παρατηρείται ότι στην Ισλανδία,την Σουηδία,την Φιλανδία, την Δανία και τον Καναδά οι φοιτητές παρουσιάζουν χαμηλά επίπεδα κατάθλιψης και στρες. Ενώ στην Τουρκία, την Νότια Αφρική, το Ηνωμένο Βασίλειο, το Ισραήλ, την Ισπανία, το Βέλγιο και τις Ηνωμένες πολιτείες της Αμερικής οι περισσότεροι φοιτητές παρουσιάζουν υψηλά επίπεδα κατάθλιψης και στρες. Συνδυάζοντας αυτά τα αποτελέσματα με τον μέσο δείκτη αυστηρότητας των μέτρων σχετικά με τον Covid-19, οι Σκανδιναβικές Χώρες και ο Καναδάς είχαν τα πιο χαλαρά μετρά σχετικά με τις υπόλοιπες χώρες, με τιμές που κυμαίνονται από 47 έως 70. Ενώ οι χώρες με τα υψηλά επίπεδα στρες και κατάθλιψης βρίσκονται στις χώρες με αυστηρά ή και πολύ αυστηρά μέτρα, με τιμές που κυμαίνονται από 80 και πάνω.

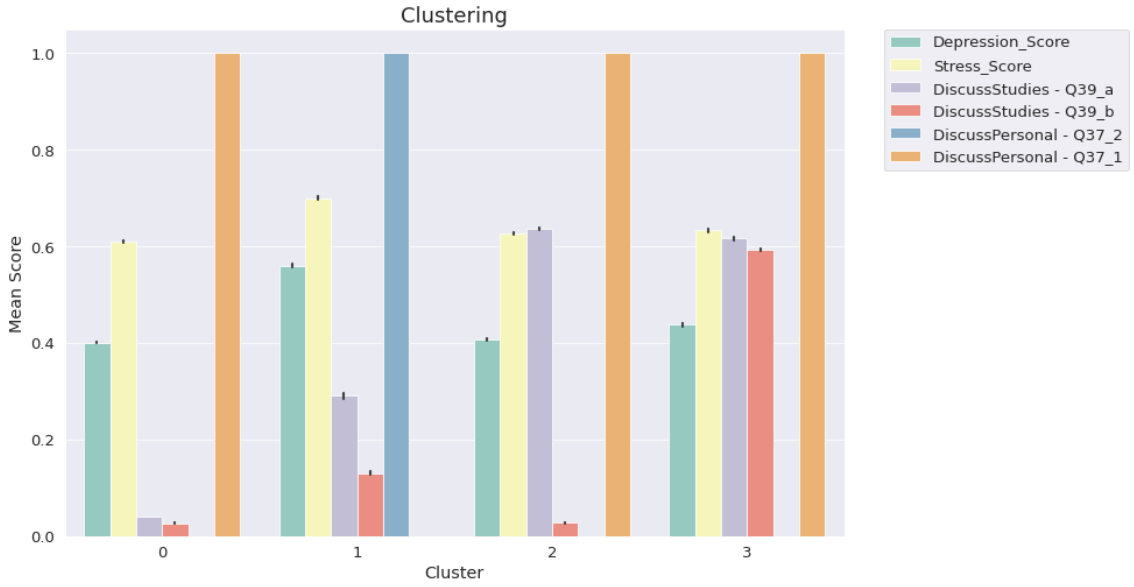
Ομαδοποίηση των φοιτητών σύμφωνα με την πρόθεση τους για συζήτηση

Υλοποιείται η μέθοδος Elbow για την επιλογή των ομάδων και επιλέγονται 4 ομάδες, σύμφωνα με το παρακάτω διάγραμμα.



Σχήμα 11: Elbow Μέθοδος δεύτερης ομαδοποίησης

Στην συνέχεια, στο ακόλουθο διάγραμμα παρουσιάζονται οι μέσοι όροι των κανονικοποιημένων δεδομένων για να διακρίνουμε τις ομάδες που προκύπτουν. Όπως φαίνεται μπορούμε να αποδώσουμε τα παρακάτω ονόματα στις 4 ομάδες που προκύπτουν:



Σχήμα 12: Αποτελέσματα 2ης ομαδοποίησης

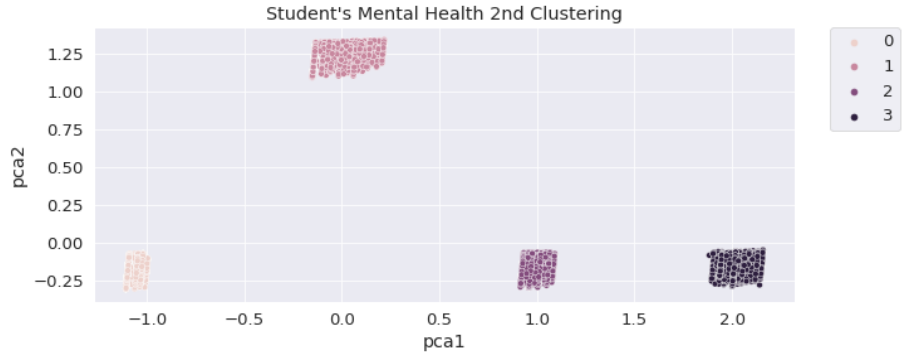
- Ομάδα 0: Πολύ λιγότερη επαφή των φοιτητών με το διδακτικό προσωπικό με την έξαρση της πανδημίας Covid-19 σχετικά με ανησυχίες για τις σπουδές τους και τα ψυχο-κοινωνικά προβλήματα τους και παράλληλα ύπαρξη ατόμου για συζήτηση σχετικά με τα προσωπικά τους ζητήματα.
- Ομάδα 1: Λιγότερη/Πολύ λιγότερη επαφή των φοιτητών με το διδακτικό προσωπικό με την έξαρση της πανδημίας Covid-19 σχετικά με ανησυχίες για τις σπουδές τους/ψυχο-κοινωνικά προβλήματα τους και παράλληλα απουσία ατόμου για συζήτηση σχετικά με τα προσωπικά τους ζητήματα.
- Ομάδα 2: Περισσότερη/Πολύ λιγότερη επαφή των φοιτητών με το διδακτικό προσωπικό με την έξαρση της πανδημίας Covid-19 σχετικά με ανησυχίες για τις σπουδές τους/ψυχο-κοινωνικά προβλήματα τους και παράλληλα ύπαρξη ατόμου για συζήτηση σχετικά με τα προσωπικά τους ζητήματα.
- Ομάδα 3: Περισσότερη/Περισσότερη επαφή των φοιτητών με το διδακτικό προσωπικό με την έξαρση της πανδημίας Covid-19 σχετικά με ανησυχίες για τις σπουδές τους/ψυχο-κοινωνικά προβλήματα τους και παράλληλα ύπαρξη ατόμου για συζήτηση σχετικά με τα προσωπικά τους ζητήματα.

Για τον έλεγχο διαχωρισιμότητας των ομάδων, στον ακόλουθο πίνακα δίνονται οι συντελεστές Silhouette και Calinski-Harabasz.

Πίνακας 6: Συντελεστές αξιολόγησης

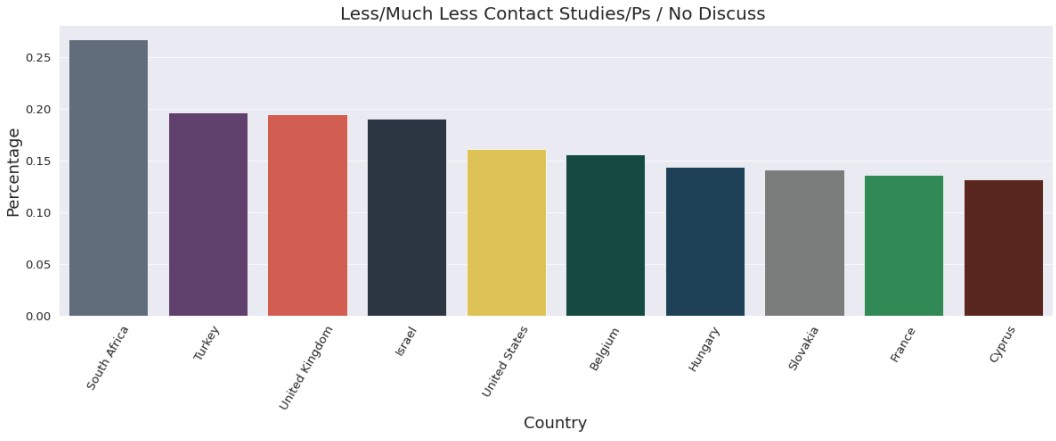
Συντελεστής	Ακρίβεια
Silhouette	0.40
Calinski-Harabasz Index	69,914

Προκύπτει πολύ μεγάλος συντελεστής Calinski-Harabasz και ο συντελεστής Silhouette είναι 0.40, δηλαδή κοντά στο 1, το οποίο δηλώνει τον πολύ καλό διαχωρισμό. Επιπλέον, επιβεβαιώνεται ότι έγινε αρκετά καλός διαχωρισμός των ομάδων, με εφαρμογή της μεθόδου PCA,η οποία χρησιμοποιείται σαν feature extractor. Το ποσοστό πληροφορίας στις 2 διαστάσεις είναι 90%, οπότε μπορούμε να στηρίξουμε τα συμπεράσματα στο ότι διαχωρίζονται πολύ καλά οι ομάδες φοιτητών.



Σχήμα 13: Εφαρμογή PCA 2ης ομαδοποίησης

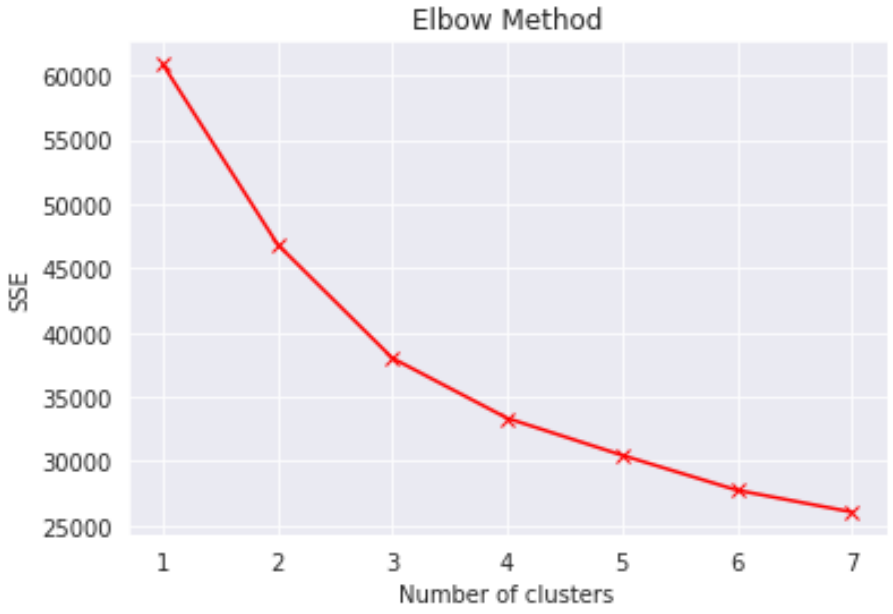
Ιδιαίτερο ενδιαφέρον σε αυτή την ομαδοποίηση παρουσιάζει η ομάδα 1, όπου οι φοιτητές φαίνεται ότι επηρεάστηκαν σε πολύ μεγάλο βαθμό από την πανδημία και παράλληλα δεν είχαν κάποιο άτομο να συζητήσουν τα προβλήματα που τους απασχολούν. Σε αυτή την ομάδα, ανήκουν φοιτητές από χώρες όπως η Νότια Αφρική, η Τουρκία, το Ηνωμένο Βασίλειο,το Ισραήλ, τις Ηνωμένες Πολιτείες της Αμερικής και το Βέλγιο, στις οποίες το μεγαλύτερο πλήθος των φοιτητών έχει υψηλά επίπεδα κατάθλιψης και τους εφαρμόστηκαν αυστηρά μέτρα σχετικά με την πανδημία. Τα αποτελέσματα παρουσιάζονται στο ακόλουθο διάγραμμα.



Σχήμα 14: 10 Κορυφαίες χώρες της ομάδας 1

Ομαδοποίηση των φοιτητών σύμφωνα με την υγεία και τις συνήθειες τους

Η τελευταία ομαδοποίηση της πρώτης προσέγγισης μας δίνει πληροφορίες σχετικά με την υγεία των φοιτητών, δηλαδή αν καπνίζουν ή πίνουν αλκοόλ και με το αν κάνουν σωματική άσκηση μέτριας ή έντονης έντασης. Σύμφωνα με την μέθοδο Elbow, παρατηρούμε ότι το κριτήριο Inertia από 3 έως 4 ομάδες δεν παρουσιάζει έντονες αλλαγές και γι᾽ αυτό επιλέγονται 4 ομάδες.



Σχήμα 15: Elbow Μέθοδος τρίτης ομαδοποίησης

Όπως προηγουμένως, στο ακόλουθο διάγραμμα παρουσιάζονται οι μέσοι όροι των κανονικοποιημένων δεδομένων για να διακρίνουμε τις ομάδες που προκύπτουν.

Σχήμα 16: Αποτελέσματα 3ης ομαδοποίησης

Διακρίνουμε 4 είδη ομάδων:

- Ομάδα 0: Υγιής ομάδα, δηλαδή που δεν καπνίζει ούτε πίνει πολύ με έντονη αρνητική αλλαγή στις καθημερινές δραστηριότητες λόγω της πανδημίας Covid-19
- Ομάδα 1: Υγιής ομάδα με πολύ ήπια αλλαγή στις καθημερινές δραστηριότητες, οι οποίες είναι αυξημένες.
- Ομάδα 2: Υγιής ομάδα και με μέτρια θετική αλλαγή στις καθημερινές δραστηριότητες, οι οποίες είναι σε χαμηλό βαθμό.
- Ομάδα 3: Ομάδα που καπνίζει πολύ και πίνει σχετικά πολύ με αρνητική αλλαγή στις καθημερινές δραστηριότητες, οι οποίες είναι σε μέτριο βαθμό.

Παρουσιάζονται οι συντελεστές αξιολόγησης

Πίνακας 7: Συντελεστές αξιολόγησης

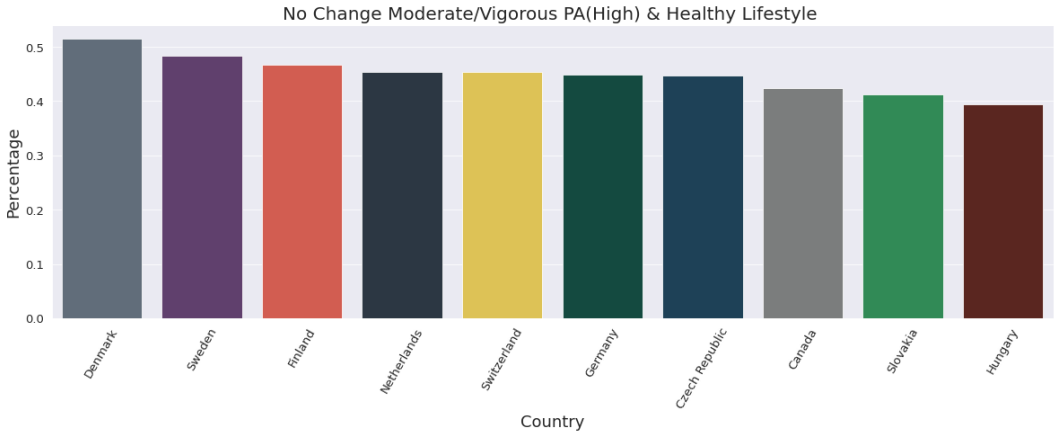
Συντελεστής	Ακρίβεια
Silhouette	0.23
Calinski-Harabasz Index	23,513

Επιβεβαιώνεται ότι έγινε αρκετά καλός διαχωρισμός των ομάδων, με εφαρμογή της μεθόδου PCA. Το ποσοστό πληροφορίας στις 2 διαστάσεις είναι 80%, οπότε

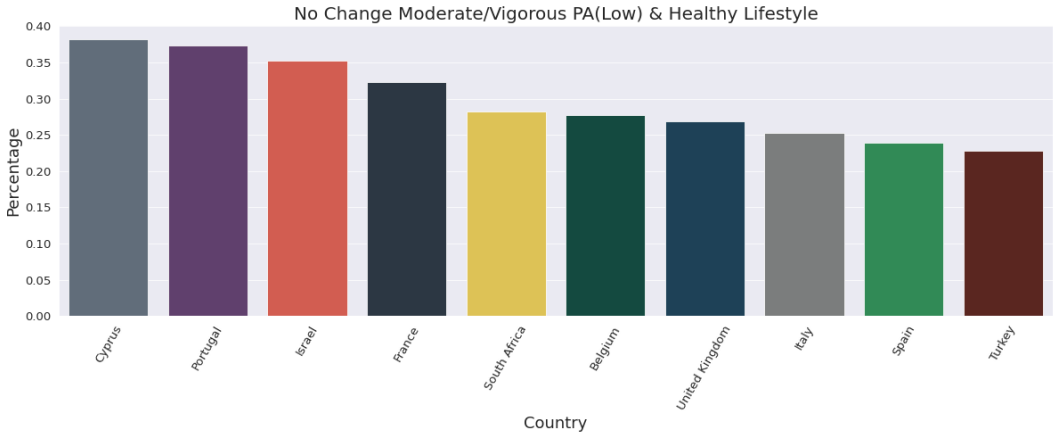
μπορούμε να στηρίξουμε τα συμπεράσματα στο ότι διαχωρίζονται πολύ καλά οι ομάδες φοιτητών.

Σχήμα 17: Εφαρμογή PCA 3ης ομαδοποίησης

Τέλος, όπως παρατηρείται από τα διαγράμματα η Δανία, η Σουηδία, η Φιλανδία έχουν μεγάλο πλήθος φοιτητών της ομάδας 1, η οποία ακολουθεί ένα πιο υγιή τρόπο ζωής με συχνή σωματική άσκηση,η οποία δεν επηρεάστηκε από την πανδημία. Αυτές οι χώρες διαπιστώθηκε ότι έχουν χαμηλά επίπεδα στρες και κατάθλιψης και παράλληλα του εφαρμόστηκαν πιο χαλαρά μέτρα. Σε αντίθεση, χώρες όπως το Ισραήλ, η Νότια Αφρική, το Βέλγιο και το Ηνωμένο Βασίλειο φαίνεται πώς έχουν μεγάλο πλήθος φοιτητών της ομάδας 2. Αυτές οι χώρες είχαν υψηλά επίπεδα κατάθλιψης και στρες και παράλληλα επηρεάστηκαν με την πανδημία ως προς την πρόθεση τους για συζήτηση, σύμφωνα με την 2η ομαδοποίηση και τώρα δείχνουν ότι οι φοιτητές είχαν θετική μέτρια αλλαγή στις καθημερινές δραστηριότητες τους.



Σχήμα 18: 10 Κορυφαίες χώρες της ομάδας 1



Σχήμα 19: 10 Κορυφαίες χώρες της ομάδας 2

Συνεπώς, με τα παραπάνω αποτελέσματα γίνεται αντιληπτό ότι χώρες όπως οι Σκανδιναβικές χώρες αντιμετώπισαν με τέτοιον τρόπο την πανδημία ώστε να υπάρχουν χαμηλότερα επίπεδα κατάθλιψης και στρες και θα μπορούσαν να αποτελέσουν παράδειγμα για άλλες χώρες όπως η Νότια Αφρική. Να τονιστεί σε αυτό το σημείο ότι τα συμπεράσματα που προκύπτουν γίνονται βάσει αυτής της έρευνας και μπορεί να υπεισέρχονται επιπλέον παράγοντες, όπως το βιοτικό επίπεδο για την αύξηση επιπέδων κατάθλιψης. Τέλος, για χώρες όπως η Νότια Αφρική, η Τουρκία ή το Βέλγιο, θα μπορούσε η κυβέρνηση να λάβει αυτά τα αποτελέσματα για να βελτιώσει την ψυχολογία των φοιτητών σε μια επερχόμενη καραντίνα, μιας και ζούμε σε μια εποχή πανδημίας.

5.3.2 Προσέγγιση 2

Ομαδοποίηση στο πρώτο σύνολο δεδομένων

Ο U-matrix που προκύπτει ύστερα από την εκπαίδευση του δικτύου SOM φαίνεται στο σχήμα 20. Αποτελεί μια οπτικοποίηση των αποστάσεων μεταξύ των νευρώνων εξόδου του δικτύου στις 2 διαστάσεις. Οι νευρώνες στις μπλε περιοχές είναι πιο κοντά μεταξύ τους, ενώ όσο το χρώμα αλλάζει προς το κόκκινο, οι νευρώνες βρίσκονται πιο μακριά.

Σχήμα 20: U-matrix 1ου δικτύου SOM

Για την ομαδοποίηση των νευρώνων, χρησιμοποιούνται αρχικά 2 ομάδες, αφού έχουν ως αποτέλεσμα πολύ καλές τιμές στις μετρικές που χρησιμοποιούνται.

Πίνακας 8: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.50
Calinski-Harabasz Index	135,089.72

Στον πίνακα 21 φαίνονται οι μέσες τιμές των αριθμητικών μεταβλητών σε κάθε ομάδα που δημιουργείται, και για κάθε μια κατηγορία των κατηγορικών μεταβλητών, σε κάθε ομάδα φαίνεται το ποσοστό των φοιτητών που ανήκουν σε αυτή την κατηγορία σε σχέση με όλους τους φοιτητές που ανήκουν στην ίδια κατηγορία (για παράδειγμα τι ποσοστό των γυναικών ανήκει σε κάθε ομάδα).

	0	1
Depression_Score	7.378271	16.327652
Stringency Measures Mean	75.621895	75.948212
Stress_Score	9.209247	11.809878
Stress_a - Q41_a	2.35445	2.830195
Stress_b - Q41_b	2.492755	2.884132
Stress_c - Q41_c	2.081523	2.936832
Stress_e - Q41_e	2.293272	3.132187
Satisf_d - Q41_d	2.169163	2.560042
Satisf_f - Q41_f	2.598308	2.324417
Satisf_g - Q41_g	2.70949	2.380461
Satisf_h - Q41_h	2.145307	1.635168
COVID_Knowledge - Q42	6.11087	5.930326
SevereInfectionWorries - Q30b_2	3.541543	4.517918
#Cigarettes per day Change	-0.093519	-0.026608
CoverMonthlyCosts Change	0.295705	0.54405
TobaccoOften Change	-0.094937	-0.121211
AlcoholOften Change	-0.365709	-0.339209
VigoPhysiAct Change	-0.004361	-0.32896
ModPhysiAct Change	-0.214425	-0.770553
On time government info - Q43_a	3.254005	3.000909
Comprehensive government info - Q43_b	3.308302	3.056044
Age - Q2	23.775602	22.779713
DiscussPersonal - Q37_2	43.16%	56.84%
DiscussPersonal - Q37_1	71.26%	28.74%
Program - Q10_Bachelor program	64.63%	35.37%
Program - Q10_Doctoral program	76.17%	23.83%
Program - Q10_Master program	72.08%	27.92%
Program - Q10_Other, specify	75.62%	24.38%
Gender - Q1_female	66.02%	33.98%
Gender - Q1_male	72.54%	27.46%
Gender - Q1_x	47.16%	52.84%
Cluster	0.0	1.0
Size	57554	27514

Σχήμα 21: Στατιστικά των ομάδων

Παρατηρούμε ότι η μέση τιμή του σκορ κατάθλιψης των φοιτητών που κατατάσσονται στην 1η ομάδα είναι 7.3/24 ενώ στην 2η 16.3/24, άρα μπορούμε να πούμε πως η πρώτη ομάδα περιλαμβάνει φοιτητές με μικρότερα επίπεδα κατάθλιψης από ότι η 2η. Η 2η ομάδα έχει επίσης υψηλότερο μέσο όρο για τα επίπεδα στρες, και φαίνεται πως κατά μέσο όρο έχει απαντήσει συμφωνεί περισσότερο με τις ερωτήσεις Q41_a, Q41_b, Q41_c, Q41_e, Q41_d, και λιγότερο με τις ερωτήσεις Q41_g και Q41_h. Αν θυμηθούμε τις ερωτήσεις, οι φοιτητές στη 2η ομάδα πιστεύουν περισσότερο από αυτούς της 1ης ομάδας πως το εκπαιδευτικό τους ίδρυμα αύξησε πολύ τον φόρτο εργασίας κατά την πανδημία, πως η ποιότητα εκπαίδευσης σε αυτό έχει μειωθεί, έχουν ανησυχία για το ότι δεν θα καταφέρουν να ολοκληρώσουν την ακαδημαϊκή χρονιά και δεν τους είναι ξεκάθαρο τι απαιτήσεις έχει το εκπαιδευτικό ίδρυμα από αυτούς. Επίσης, κατά μέσο όρο, οι αλλαγές στις εκπαιδευτικές μεθόδους τους έχουν προκαλέσει περισσότερο στρες, είναι λιγότερο ικανοποιημένοι με τα μέτρα προστασίας του πανεπιστημίου εναντίον της πανδημίας, και δεν αισθάνονται πως μπορούν να μιλήσουν σε κάποιο μέλος του

πανεπιστημίου για τις ανησυχίες τους. Επίσης, φαίνεται η 2η ομάδα έχει ελαφρώς λιγότερη γνώση για τον Covid-19 από την 1η, αν και στις 2 ομάδες οι μέσοι όροι δεν είναι ικανοποιητικοί και οι φοιτητές της 2ης ομάδας ανησυχούν περισσότερο ότι θα νοσήσουν σοβαρά από τον Covid-19. Όσον αφορά τις διαφορές στις δραστηριότητες πριν και κατά τη διάρκεια της πανδημίας, για τους φοιτητές της 2ης ομάδας μειώθηκαν σε μεγαλύτερο βαθμό οι εβδομαδιαίες ήπιες και έντονες δραστηριότητες. Σχετικά με την άποψή τους για τις πληροφορίες της κυβέρνησης, πιστεύουν σε μικρότερο βαθμό πως παρείχε έγκαιρες και κατανοητές πληροφορίες για την πανδημία. Επίσης, ο μέσος όρος ηλικίας στην 2η ομάδα είναι μικρότερος από αυτόν της 1ης. Εστιάζοντας τώρα στις κατηγορικές μεταβλητές, βλέπουμε πως το μεγαλύτερο ποσοστό αυτών που δεν έχουν κάποιον για να συζητήσουν προσωπικά θέματα, βρίσκονται στην 2η ομάδα. Επίσης, το ποσοστό των γυναικών που ανήκει στην 2η ομάδα είναι μεγαλύτερο από αυτό των αντρών, ενώ το μεγαλύτερο ποσοστό των non-binary ατόμων βρίσκεται σε αυτή.

Για να επιβεβαιώσουμε αν για τις παραπάνω μεταβλητές υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στους μέσους τους, εφαρμόζουμε t-test. Επιλέγεται ένας από τους 2 ελέγχους t-test, που αναφέρεται στην ενότητα των στατιστικών ελέγχων, αναλόγως με το αν η διασπορά ανάμεσα στις ομάδες είναι ίση ή όχι. Τα αποτελέσματα φαίνονται στον πίνακα 22.

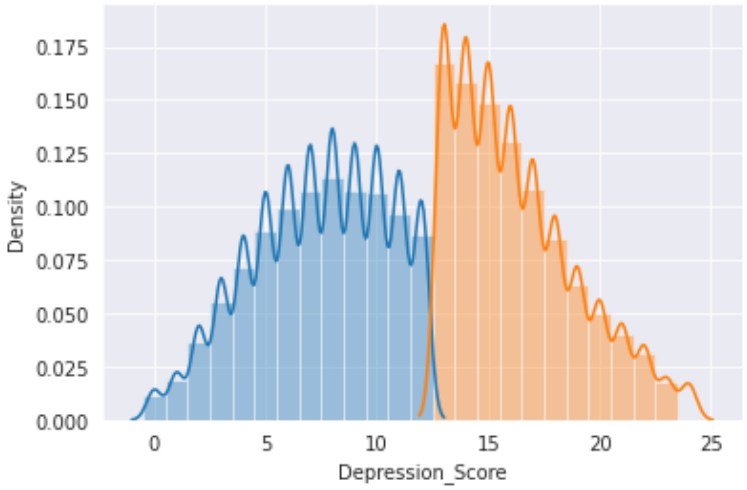
	p-value	Reject Ho?
Depression_Score	0.000000e+00	True
Stress_Score	0.000000e+00	True
Stress_a - Q41_a	0.000000e+00	True
Stress_b - Q41_b	0.000000e+00	True
Stress_c - Q41_c	0.000000e+00	True
Stress_e - Q41_e	0.000000e+00	True
Satisf_d - Q41_d	0.000000e+00	True
Satisf_f - Q41_f	2.894590e-252	True
Satisf_g - Q41_g	0.000000e+00	True
Satisf_h - Q41_h	0.000000e+00	True
COVID_Knowledge - Q42	8.425771e-74	True
SevereInfectionWorries - Q30b_2	0.000000e+00	True
CoverMonthlyCosts Change	6.801377e-270	True
VigoPhysiAct Change	2.427102e-166	True
ModPhysiAct Change	0.000000e+00	True
On time government info - Q43_a	5.747299e-188	True
Comprehensive government info - Q43_b	1.827622e-188	True
Stringency Measures Mean	4.713187e-15	True
#Cigarettes per day Change	1.239886e-03	True
Age - Q2	2.915528e-154	True

Σχήμα 22: T-test για τους μέσους των μεταβλητών

Βλέπουμε ότι πράγματι, για όλες τις παραπάνω αριθμητικές μεταβλητές που αναφέρθηκαν έχουμε σοβαρές ενδείξεις απόρριψης της μηδενικής υπόθεσης, αφού η p-value είναι μικρότερη του 0.05, που επιλέγεται σαν επίπεδο σημαντικότητας.

Μπορούμε να διακρίνουμε δύο ομάδες με στατιστικά σημαντικά διαφορές ανάμεσα στους μέσους όρους κάθε μεταβλητής.

Επίσης στο σχήμα 23 φαίνεται η κατανομή του σκορ κατάθλιψης σε κάθε ομάδα.



Σχήμα 23: Κατανομή του σκορ κατάθλιψης σε κάθε ομάδα

Ο διαχωρισμός γίνεται περίπου στην τιμή 12, η οποία οριοθετεί τις 2 ομάδες αν και υπάρχει αλληλοεπικάλυψη.

Στη συνέχεια, για το ίδιο δίκτυο SOM, η ομαδοποίηση γίνεται χρησιμοποιώντας 3 ομάδες, αρχικά επειδή και αυτός ο διαχωρισμός έχει πολύ καλές τιμές στις μετρικές που χρησιμοποιούνται, αλλά κυρίως επειδή ένας διαχωρισμός σε 3 ομάδες έχει πιθανώς περισσότερο νόημα, ιδίως αν οι ομάδες χωριστούν με τέτοιον τρόπο ώστε να έχουν αυξανόμενα και διαφορετικά επίπεδα κατάθλιψης. Σε αυτή την περίπτωση μπορούμε να επιβεβαιώσουμε προηγούμενες παρατηρήσεις ή και να εντοπίσουμε επιπλέον διαφορές.

Πίνακας 9: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.4
Calinski-Harabasz Index	139,786.21

Στον πίνακα 24 φαίνονται τα στατιστικά των μεταβλητών σε κάθε ομάδα.

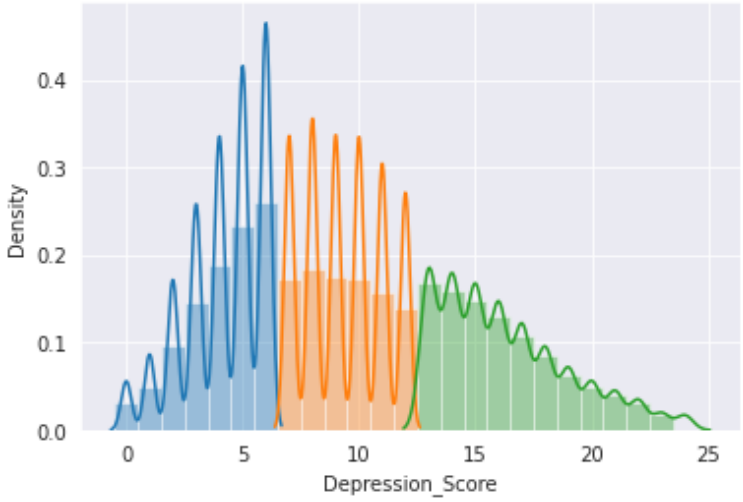
	0	1	2
Depression_Score	4.145879	9.376047	16.327652
Stringency Measures Mean	75.514275	75.6884	75.948212
Stress_Score	8.02943	9.938431	11.809878
Stress_a - Q41_a	2.137236	2.488698	2.830195
Stress_b - Q41_b	2.30786	2.607028	2.884132
Stress_c - Q41_c	1.724345	2.302277	2.936832
Stress_e - Q41_e	1.881186	2.547962	3.132187
Satisf_d - Q41_d	2.001865	2.272561	2.560042
Satisf_f - Q41_f	2.713473	2.52713	2.324417
Satisf_g - Q41_g	2.835835	2.631403	2.380461
Satisf_h - Q41_h	2.306859	2.04546	1.635168
COVID_Knowledge - Q42	6.145606	6.089401	5.930326
SevereInfectionWorries - Q30b_2	3.065184	3.835957	4.517918
#Cigarettes per day Change	-0.074145	-0.105493	-0.026608
CoverMonthlyCosts Change	0.209107	0.349227	0.54405
TobaccoOften Change	-0.075191	-0.107141	-0.121211
AlcoholOften Change	-0.345297	-0.378324	-0.339209
VigoPhysiAct Change	0.08306	-0.058392	-0.32896
ModPhysiAct Change	-0.051401	-0.315181	-0.770553
On time government info - Q43_a	3.32742	3.208631	3.000909
Comprehensive government info - Q43_b	3.389874	3.257886	3.056044
Age - Q2	24.263647	23.473967	22.779713
DiscussPersonal - Q37_2	10.54%	32.62%	56.84%
DiscussPersonal - Q37_1	28.09%	43.16%	28.74%
Program - Q10_Bachelor program	23.37%	41.26%	35.37%
Program - Q10_Doctoral program	32.38%	43.79%	23.83%
Program - Q10_Master program	29.55%	42.53%	27.92%
Program - Q10_Other, specify	32.31%	43.31%	24.38%
Gender - Q1_female	24.01%	42.01%	33.98%
Gender - Q1_male	31.04%	41.50%	27.46%
Gender - Q1_x	13.48%	33.69%	52.84%
Cluster	2.0	1.0	0.0
Size	21984	35570	27514

Σχήμα 24: Στατιστικά των ομάδων

Βλέπουμε πως πλέον η 1η ομάδα χαρακτηρίζεται από χαμηλά επίπεδα κατάθλιψης με μέσο όρο 4.14/24, η 2η από μέσα επίπεδα με μέσο όρο 9.37/24 και η 3η από υψηλότερα με μέσο όρο 16.32/24. Το μέσο σκορ στρες σε κάθε ομάδα αυξάνεται όταν αυξάνεται και η μέση κατάθλιψη. Οι μέσοι όροι των ερωτήσεων Q41_a έως Q41_h διαφοροποιούνται με παρόμοιο τρόπο με αυτόν που περιγράφηκε για τις την ομαδοποίηση σε 2 ομάδες. Η 3η ομάδα πιστεύει σε μεγάλο βαθμό πως τα μέτρα και η αντιμετώπιση των πανεπιστημίων δεν ήταν επαρκή, ενώ αισθάνονται σε πολύ μεγάλο βαθμό κατά μέσο όρο πως οι αλλαγές στις εκπαιδευτικές μεθόδους τους προκαλούν μεγάλο στρες και δεν αισθάνονται πως μπορούν να μιλήσουν με κάποιο μέλος του πανεπιστημίου. Επίσης, παρατηρείται σημαντική αύξηση στο πόσο ανησυχούν οι φοιτητές της 3ης ομάδας ότι θα νοσήσουν σοβαρά από τον Covid-19 σε σχέση με τους φοιτητές της 1ης ομάδας, ενώ για τους φοιτητές αυτούς μειώθηκαν σε μεγαλύτερο βαθμό οι εβδομαδιαίες ήπιες και έντονες δραστηριότητες σε σχέση με τις άλλες 2 ομάδες. Σχετικά με την ενημέρωση της κυβέρνησης, αντίστοιχα με πριν, η ομάδα με τα υψηλότερα επίπεδα κατάθλιψης πιστεύει πως ήταν λιγότερο επαρκής σε σχέση με τις άλλες ομάδες. Ο μέσος όρος ηλικίας στην ομάδα με την χαμηλότερη μέση κατάθλιψη είναι 24, ενώ στην ομάδα με την υψηλότερη 22. Η συντριπτική πλειοψηφία (89%) αυτών που δεν έχουν κάποιο άτομο με το οποίο μπορούν να συζητήσουν να προσωπικά τους θέματα ανήκει στις 2 ομάδες με τα υψηλότερα επίπεδα κατάθλιψης. Σχετικά με τα

προγράμματα σπουδών, φαίνεται πως συνολικά οι διδακτορικοί φοιτητές έχουν χαμηλότερα επίπεδα κατάθλιψης παρατηρώντας την κατανομής τους στις ομάδες, ενώ πολύ μεγάλα επίπεδα έχει ένα μεγάλο ποσοστό των προπτυχιακών φοιτητών. Τώρα είναι πολύ πιο εμφανές πως παραπάνω από τα μισά non-binary άτομα έχουν πολύ υψηλά επίπεδα κατάθλιψης κατά μέσο όρο, ενώ για τα άλλα μισά, στην ομάδα με τα χαμηλά επίπεδα κατάθλιψης ανήκει η μειοψηφία.

Στο σχήμα 25 φαίνεται η κατανομή του σκορ κατάθλιψης σε κάθε ομάδα.

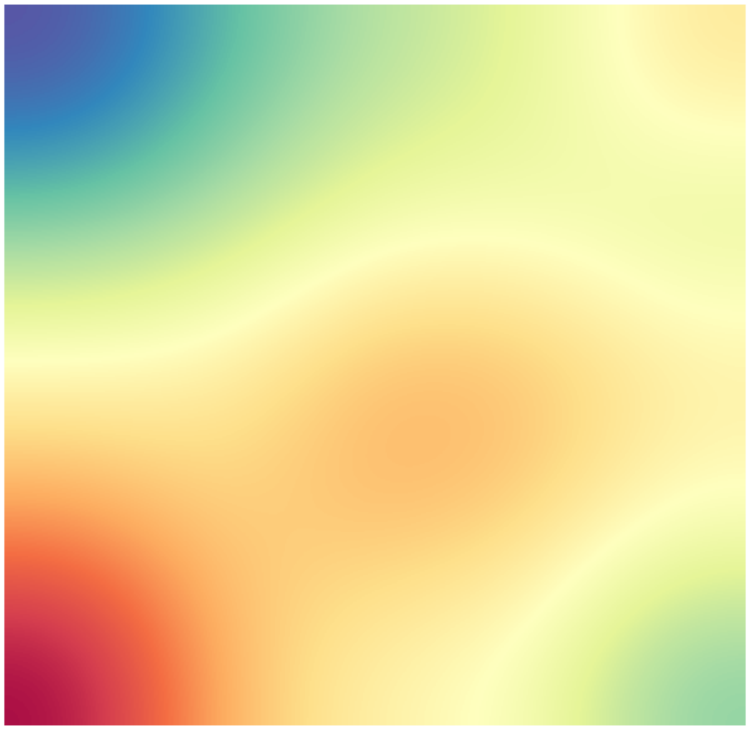


Σχήμα 25: Κατανομή του σκορ κατάθλιψης σε κάθε ομάδα

Πράγματι, επιλέγοντας 3 ομάδες για τον διαχωρισμό, επιβεβαιώθηκαν και ενισχύθηκαν προηγούμενες παρατηρήσεις και οι διαφορές των μέσων όρων των μεταβλητών όπως αυτές διαμορφώνονται στις ομάδες με διαφορετικά επίπεδα κατάθλιψης έγιναν πιο εμφανείς.

Ομαδοποίηση στο δεύτερο σύνολο δεδομένων

Ο U-matrix που προκύπτει ύστερα από την εκπαίδευση του δικτύου SOM φαίνεται στο σχήμα 26.



Σχήμα 26: U-matrix 1ου δικτύου SOM

Για την ομαδοποίηση των νευρώνων του συγκεκριμένου δικτύου SOM, χρησιμοποιούνται 3 ομάδες, αφού οι μετρικές τους έχουν πολύ καλές τιμές και όπως αναφέρθηκε και για το προηγούμενο δίκτυο SOM, οι διαφορές των μέσων όρων των μεταβλητών σε κάθε ομάδα είναι πιο εμφανείς.

Πίνακας 10: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.42
Calinski-Harabasz Index	105,531

Στον πίνακα 27 φαίνονται τα στατιστικά των μεταβλητών σε κάθε ομάδα.

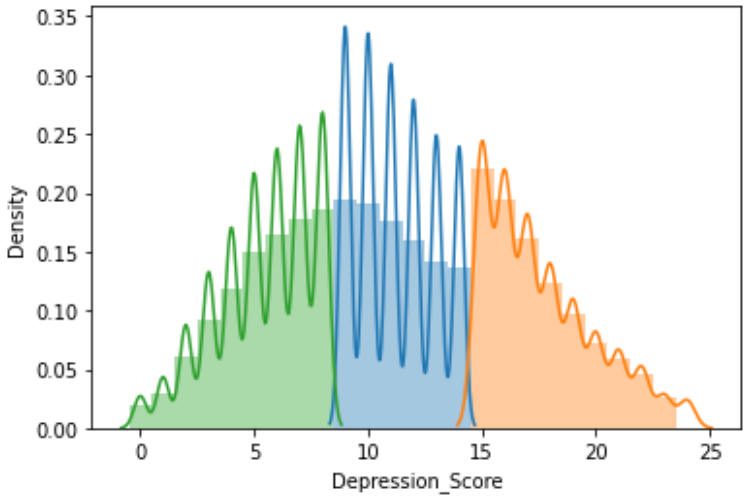
	0	1	2
Depression_Score	11.271924	17.659184	5.380423
Stringency Measures Mean	75.899254	76.088699	75.596214
Stress_Score	10.578368	12.125627	8.530035
Stress_a - Q41_a	2.604358	2.876992	2.225165
Stress_b - Q41_b	2.7096	2.92009	2.401629
Stress_c - Q41_c	2.507426	3.052151	1.868005
Stress_e - Q41_e	2.751149	3.234493	2.055072
Satisf_d - Q41_d	2.358381	2.619229	2.089824
Satisf_f - Q41_f	2.454782	2.243921	2.643848
Satisf_g - Q41_g	2.550433	2.309989	2.777939
Satisf_h - Q41_h	1.920129	1.518593	2.2219
COVID_Knowledge - Q42	6.061925	5.931238	6.180858
Activities Change	-4.467438	-6.374449	-3.811672
SevereInfectionWorries - Q30b_2	3.994652	4.604939	3.257902
#Cigarettes per day Change	-0.110785	-0.023577	-0.091114
CoverMonthlyCosts Change	0.406913	0.582492	0.24363
TobaccoOften Change	-0.121641	-0.141414	-0.08551
AlcoholOften Change	-0.396703	-0.347026	-0.382519
VigoPhysiAct Change	-0.113773	-0.338122	0.083777
ModPhysiAct Change	-0.416947	-0.836364	-0.082204
On time government info - Q43_a	3.131277	2.930041	3.282898
Comprehensive government info - Q43_b	3.176671	2.991096	3.335994
Age - Q2	22.992619	22.452824	23.643766
DiscussPersonal - Q37_2	36.90%	43.95%	19.14%
DiscussPersonal - Q37_1	37.26%	18.74%	44.00%
Program - Q10_Bachelor program	37.88%	24.65%	37.48%
Program - Q10_Doctoral program	35.11%	15.58%	49.31%
Program - Q10_Master program	36.23%	17.89%	45.88%
Program - Q10_Other, specify	35.54%	14.56%	49.90%
Gender - Q1_female	37.86%	23.19%	38.95%
Gender - Q1_male	35.49%	18.28%	46.24%
Gender - Q1_x	37.53%	41.73%	20.74%
Cluster	0.0	2.0	1.0
Size	22624	13365	24804

Σχήμα 27: Στατιστικά των ομάδων

Βλέπουμε πως και πάλι μια ομάδα χαρακτηρίζεται από χαμηλά επίπεδα μέσης κατάθλιψης, μια ομάδα από αυξημένα προς υψηλά και μια από εξαιρετικά υψηλά. Οι παρατηρήσεις που έγιναν για την αντίστοιχη ομαδοποίηση στο προηγούμενο δίκτυο SOM ισχύουν και σε αυτή την περίπτωση, δηλαδή αν εστιάσουμε στις μεταβλητές που έχουν αναφερθεί, όταν κινούμαστε από ομάδα χαμηλότερης κατάθλιψης σε ομάδα υψηλότερης, οι μέσοι όροι των μεταβλητών αυτών αλλάζουν με τον εξής τρόπο: Αν η μεταβλητή εκφράζει κάτι θετικό, όπως για παράδειγμα το ότι οι φοιτητές αισθάνονται πως μπορούν να μιλήσουν σε κάποιο μέλος του πανεπιστημίου για τις ανησυχίες τους, τότε η μέση τιμή μειώνεται. ενώ αν εκφράζει κάτι αρνητικό, όπως το ότι οι αλλαγές στις εκπαιδευτικές μεθόδους του πανεπιστημίου λόγω της πανδημίας προκάλεσαν στους φοιτητές μεγάλο στρες, η μέση τιμή αυξάνεται. Όπως έχει αναφερθεί ήδη, ο λόγος της δημιουργίας του παρόντος δικτύου SOM είναι να συμπεριληφθεί πληροφορία σχετικά με τις προσωπικές δραστηριότητες των φοιτητών χρησιμοποιώντας το μικρότερο σύνολο δεδομένων που περιέχει τις σχετικές μεταβλητές. Εστιάζοντας στην διαφορά των ωρών των προσωπικών δραστηριοτήτων πριν και κατά τη διάρκεια της πανδημίας, αρχικά παρατηρούμε πως σε όλες τις κατηγορίες οι μέσες διαφορές είναι αρνητικές, δηλαδή οι ωρες για τις συγκεκριμένες δραστηριότητες μειώθηκαν για τους περισσότερους φοιτητές ανεξάρτητα από την ομάδα στην οποία

κατηγοριοποιούνται. επίσης, φαίνεται πως όταν κινούμαστε από ομάδα χαμηλότερης κατάθλιψης σε ομάδα υψηλότερης, η μέση διαφορά αυξάνεται προς τις αρνητικές τιμές. Συγκεκριμένα, στην ομάδα των πολύ υψηλών επιπέδων κατάθλιψης, οι ώρες αυτές μειώθηκαν σε μεγαλύτερο βαθμό σε σχέση με τις άλλες ομάδες, με τη μέση διαφορά των ωρών να είναι -6.

Στο σχήμα 28 φαίνεται η κατανομή του σκορ κατάθλιψης σε κάθε ομάδα.

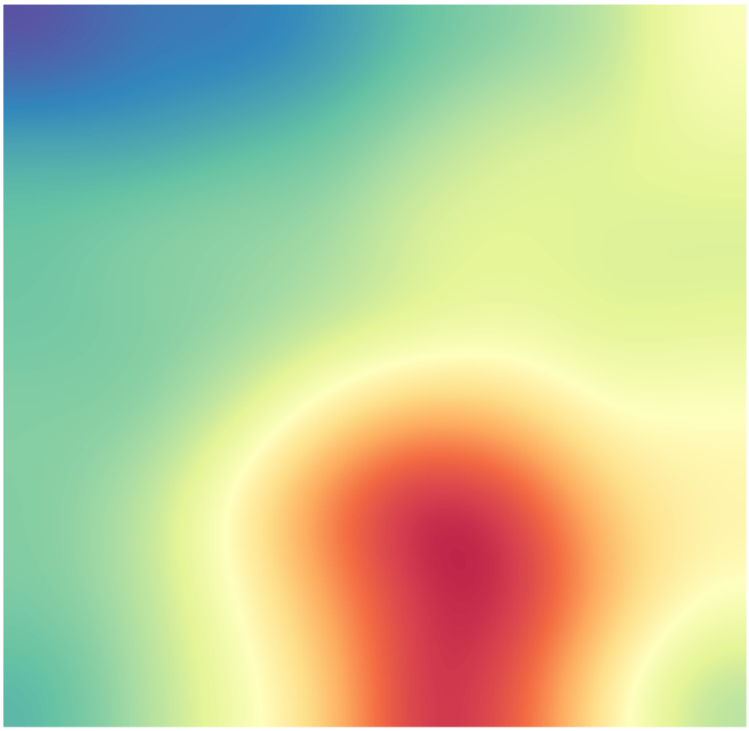


Σχήμα 28: Κατανομή του σκορ κατάθλιψης σε κάθε ομάδα

Έτσι, σε αυτή την ενότητα είδαμε πως διαφοροποιείται ανά τις ομάδες ο μέσος όρος της νέας μεταβλητής που προσθέσαμε, χρησιμοποιώντας αναγκαστικά ένα μικρότερο σύνολο δεδομένων.

Ομαδοποίηση χρησιμοποιώντας τις εγγραφές της Ελλάδας από το δεύτερο σύνολο δεδομένων

Ο U-matrix που προκύπτει ύστερα από την εκπαίδευση του δικτύου SOM φαίνεται στο σχήμα 29.



Σχήμα 29: U-matrix 1ου δικτύου SOM

Για την ομαδοποίηση των νευρώνων του συγκεκριμένου δικτύου SOM, αρχικά χρησιμοποιούνται 2 ομάδες.

Πίνακας 11: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.52
Calinski-Harabasz Index	763.59

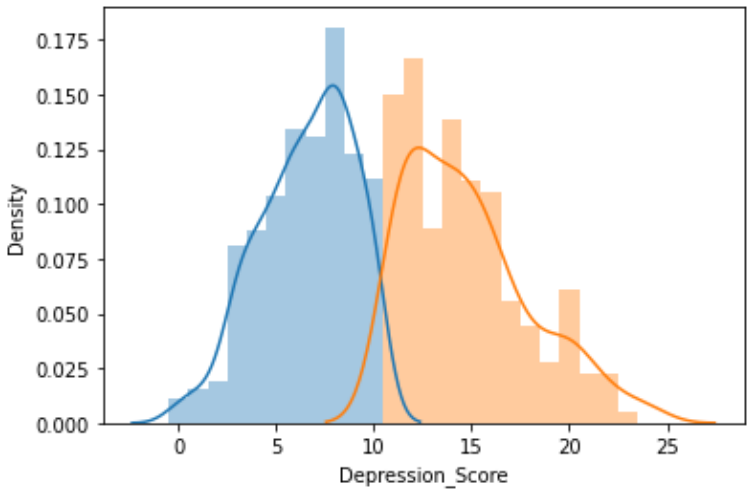
Στον πίνακα 30 φαίνονται τα στατιστικά των μεταβλητών σε κάθε ομάδα.

	0	1
Depression_Score	6.561539	14.836065
Stress_Score	8.026923	10.027323
Stress_a - Q41_a	2.211539	2.573771
Stress_b - Q41_b	1.738461	2.098361
Stress_c - Q41_c	2.292308	2.945355
Stress_e - Q41_e	1.784615	2.409836
Satisf_d - Q41_d	1.688462	2.027322
Satisf_f - Q41_f	2.561538	2.213115
Satisf_g - Q41_g	2.938462	2.699454
Satisf_h - Q41_h	1.973077	1.491803
COVID_Knowledge - Q42	6.092308	6.054645
Activities Change	2.357686	-3.724044
SevereInfectionWorries - Q30b_2	2.476923	2.934426
#Cigarettes per day Change	0.026923	0.36612
CoverMonthlyCosts Change	0.296154	0.497268
TobaccoOften Change	-0.019231	0.153005
AlcoholOften Change	-0.184615	-0.234973
VigoPhysiAct Change	-0.026923	0.010929
ModPhysiAct Change	0.026923	0.005464
On time government info - Q43_a	4.169231	4.032787
Comprehensive government info - Q43_b	3.696154	3.426229
Age - Q2	22.626923	22.606558
DiscussPersonal - Q37_2	35.29%	64.71%
DiscussPersonal - Q37_1	60.64%	39.36%
Program - Q10_Bachelor program	59.11%	40.89%
Program - Q10_Doctoral program	61.54%	38.46%
Program - Q10_Master program	52.38%	47.62%
Program - Q10_Other, specify	75.00%	25.00%
Gender - Q1_female	56.82%	43.18%
Gender - Q1_male	63.91%	36.09%
Gender - Q1_x	0.00%	100.00%
Cluster	1.0	0.0
Size	260	183

Σχήμα 30: Στατιστικά των ομάδων

Βλέπουμε πως όπως ήταν αναμενόμενο, η μια ομάδα χαρακτηρίζεται από χαμηλότερα μέσα επίπεδα κατάθλιψης και η άλλη με υψηλότερα. Οι παρατηρήσεις που έχουν γίνει ήδη για τις μεταβλητές που σχετίζονται με τα μέτρα και την αντιμετώπιση της πανδημίας από τα πανεπιστήμια ισχύουν και στην συγκεκριμένη περίπτωση. Η μεταβλητή για την οποία τα αποτελέσματα που σχετίζονται με αυτήν διαφοροποιούνται, είναι διαφορά των ωρών που αφιερώνουν οι φοιτητές σε προσωπικές δραστηριότητες εβδομαδιαία. Βλέπουμε πως για την ομάδα με την μικρότερη μέση κατάθλιψη, οι δραστηριότητες αυτές αυξήθηκαν κατά τη διάρκεια της πανδημίας κατά μέσο όρο, ενώ για την άλλη ομάδα μειώθηκαν. Αξίζει να αναφερθεί πως σχεδόν οι μισοί φοιτητές (47.62%) που παρακολουθούν κάποιο μεταπτυχιακό πρόγραμμα ανήκουν στην ομάδα με τα μεγαλύτερα επίπεδα κατάθλιψης, ενώ για τους φοιτητές των υπολοίπων προγραμμάτων τα ποσοστά είναι μικρότερα. Επίσης, βλέπουμε πως στην ομάδα υψηλότερης κατάθλιψης ανήκει το 100% των non-binary ατόμων, και ένα πολύ μεγάλο ποσοστό γυναικών (43.1%).

Στο σχήμα 31 φαίνεται η κατανομή του σκορ κατάθλιψης σε κάθε ομάδα.



Σχήμα 31: Κατανομή του σκορ κατάθλιψης σε κάθε ομάδα

Στη συνέχεια, για το ίδιο δίκτυο SOM, η ομαδοποίηση γίνεται χρησιμοποιώντας 3 ομάδες για τους λόγους που έχουν προαναφερθεί.

Πίνακας 12: Συντελεστές αξιολόγησης

Συντελεστής	Ακρίβεια
Silhouette	0.45
Calinski-Harabasz Index	653.87

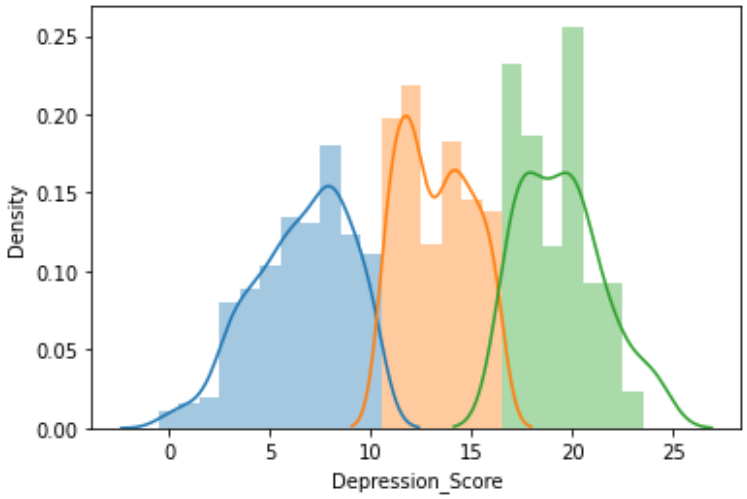
Στον πίνακα 32 φαίνονται τα στατιστικά των μεταβλητών σε κάθε ομάδα.

	0	1	2
Depression_Score	6.561539	13.277372	19.47826
Stress_Score	8.026923	9.59854	11.304348
Stress_a - Q41_a	2.211539	2.547445	2.652174
Stress_b - Q41_b	1.738461	1.948905	2.543478
Stress_c - Q41_c	2.292308	2.875912	3.152174
Stress_e - Q41_e	1.784615	2.226277	2.956522
Satisf_d - Q41_d	1.688462	1.890511	2.434783
Satisf_f - Q41_f	2.561538	2.313869	1.913043
Satisf_g - Q41_g	2.938462	2.781022	2.456522
Satisf_h - Q41_h	1.973077	1.591241	1.195652
COVID_Knowledge - Q42	6.092308	6.10219	5.913043
Activities Change	2.357686	-1.266424	-11.043478
SevereInfectionWorries - Q30b_2	2.476923	2.905109	3.021739
#Cigarettes per day Change	0.026923	0.087591	1.195652
CoverMonthlyCosts Change	0.296154	0.394161	0.804348
TobaccoOften Change	-0.019231	0.116788	0.26087
AlcoholOften Change	-0.184615	-0.255474	-0.173913
VigoPhysiAct Change	-0.026923	0.116788	-0.304348
ModPhysiAct Change	0.026923	0.072993	-0.195652
On time government info - Q43_a	4.169231	4.109489	3.804348
Comprehensive government info - Q43_b	3.696154	3.510949	3.173913
Age - Q2	22.626923	22.562044	22.73913
DiscussPersonal - Q37_2	35.29%	44.12%	20.59%
DiscussPersonal - Q37_1	60.64%	29.83%	9.54%
Program - Q10_Bachelor program	59.11%	30.47%	10.42%
Program - Q10_Doctoral program	61.54%	38.46%	0.00%
Program - Q10_Master program	52.38%	33.33%	14.29%
Program - Q10_Other, specify	75.00%	25.00%	0.00%
Gender - Q1_female	56.82%	32.47%	10.71%
Gender - Q1_male	63.91%	27.82%	8.27%
Gender - Q1_x	0.00%	0.00%	100.00%
Cluster	0.0	1.0	2.0
Size	260	137	46

Σχήμα 32: Στατιστικά των ομάδων

Παρατηρούμε πως σε σχέση με τον προηγούμενο διαχωρισμό σε 2 ομάδες, η ομάδα με την χαμηλότερη κατάθλιψη δεν μεταβλήθηκε καθόλου και δημιουργήθηκε μια ομάδα με μέσα προς υψηλά επίπεδα κατάθλιψης και μια ομάδα με πολύ υψηλά επίπεδα. Συγκρίνοντας και με αποτελέσματα των προηγούμενων SOM δικτύων, στην συγκεκριμένη περίπτωση οι ομάδες υψηλής κατάθλιψης έχουν μεγαλύτερα μέσες τιμές για το συγκεκριμένο σκορ σε σχέση με τις αντίστοιχες ομάδες όταν χρησιμοποιούνται δεδομένα από όλο τον κόσμο. Φαίνεται πως για την ομάδα με τα πολύ υψηλά επίπεδα κατάθλιψης, η αποδοκιμασία για τις τακτικές και τα μέτρα που ακολούθησαν τα πανεπιστήμια είναι μεγάλη. Για την ίδια ομάδα βλέπουμε πως η διαφορά των ωρών στις εβδομαδιαίες προσωπικές δραστηριότητες είναι αισθητά μεγαλύτερη από τις υπόλοιπες, με μέσο όρο -11, δηλαδή οι συγκεκριμένες δραστηριότητες των φοιτητών αυτών μειώθηκαν σημαντικά. Επίσης, αξίζει να αναφερθεί πως κανένας διδακτορικός φοιτητής δεν κατατάσσεται στην ομάδα υψηλής κατάθλιψης, ενώ σε αυτήν ανήκει το 14.29% των μεταπτυχιακών και το 10.42% των προπτυχιακών φοιτητών. Τέλος, το 100% των non-binary ατόμων κατατάσσονται σε αυτή την ομάδα.

Στο σχήμα 33 φαίνεται η κατανομή του σκορ κατάθλιψης σε κάθε ομάδα.



Σχήμα 33: Κατανομή του σκορ κατάθλιψης σε κάθε ομάδα

Έτσι, σε αυτή την ενότητα δείξαμε πως εστιάζοντας σε 1 συγκεκριμένη χώρα, τα αποτελέσματα της ομαδοποίησης είναι γενικά παρόμοια με αυτά των ομαδοποιήσεων που περιλαμβάνουν όλες τις χώρες, αλλά η κατανομή και οι μέσες τιμές κάποιων μεταβλητών μπορεί να διαφοροποιούνται σε μικρό (ή μεγαλύτερο σε κάποιες περιπτώσεις) βαθμό.

5.4 Ταξινόμηση Επιπέδου Κατάθλιψης

Στην ενότητα αυτή παρουσιάζουμε πλέον με πρακτικό τρόπο την πορεία προς την δημιουργία των τελικών μοντέλων που θα προβλέπουν το επίπεδο κατάθλιψης ενός φοιτητή σε καταστάσεις πανδημίας ή αντίστοιχων συνθηκών.

5.4.1 Καθαρισμός Δεδομένων με Στατιστικούς ελέγχους

Όπως αναφέραμε και στην αντίστοιχη ενότητα της μεθοδολογίας (Τροποποίηση Δεδομένων), επιθυμούμε να έχουμε την άκρως απαραίτητη πληροφορία στα δεδομένα μας, και να αφαιρέσουμε μεταβλητές που πιθανόν να μην βοηθήσουν στην προβλεπτική ικανότητα των μοντέλων μας, αλλά αντιθέτως να επιβαρύνουν την εκπαίδευση τους με επιπλέον πολυπλοκότητα. Για αυτό το λόγο από το σύνολο από features που έχουμε καταλήξει στο αντίστοιχο κεφάλαιο θα εφαρμόσουμε κάποια κριτήρια βάση των στατιστικών ελέγχων. Όποιο feature δεν πληροί τα κριτήρια θα αφαιρείται.

Αρχικά σύμφωνα με τα αποτελέσματα του στατιστικού ελέγχου **ANOVA** (Εύρεση εξαρτήσεων μεταξύ της κατάθλιψης και των κατηγορικών μεταβλητών) προκύπτει για όλες τις αναφερθείσες μεταβλητές στατιστικά σημαντικές διαφορές μεταξύ των μέσων

επιπέδων κατάθλιψης, για αυτό το λόγο επιλέγουμε να τις διατηρήσουμε προς το παρόν.

Στη συνέχεια, μελετάμε τον στατιστικό έλεγχο X^2. Όπως έχουμε αναφέρει θέλουμε το μοντέλο μας να έχει την πληροφορία της αλλαγής που έφερε η πανδημία για αυτό θα διατηρήσουμε όλες τις αντίστοιχες μεταβλητές. Δυστυχώς παρατηρούμε ότι οι δύο μεταβλητές με το χαμηλότερο Cramer's Value είναι οι εξής :

- Q23_a - Πόσο συχνά έπινες περισσότερα από 6 ποτήρια αλκοόλ πριν την πανδημία: 0.029
- Q26_a - Πόσο συχνά έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά πριν την πανδημία: 0.032
- Q7 - Υψηλότερο επίπεδο μόρφωσης πατέρα: 0.034

Σε άλλες περιπτώσεις θα αφαιρούσαμε μεταβλητές μέχρι να φτάσουμε σε μεταβλητές που αφορούν την πανδημία, αλλά σε αυτή τη περίπτωση θα αφαιρέσουμε για ζητήματα λιτότητας την **Q7 - Υψηλότερο επίπεδο μόρφωσης πατέρα**.

Στο σημείο αυτό να αναφέρουμε για πληρότητα πως η μεταβλητή **Q12 - Πρώτος χρόνος σπουδών(Ναι, όχι)**, μπορεί να μην αναφέρεται στον πίνακα της αντίστοιχης ενότητας, λόγω του ότι εντοπίστηκε τελικά προβληματική και αφαιρέθηκε, έχει Cramer's Value 0.029 και την αφαιρέσαμε και από αυτό το Dataset.

Ύστερα, επεξεργαζόμαστε τις τιμές συσχέτισης των μεταβλητών με την τιμή της κατάθλιψης και βλέπουμε για ακόμα μια φορά ότι οι 2 μεταβλητές με τη χαμηλότερη συσχέτιση έχουν να κάνουν με την πανδημία και πιο συγκεκριμένα είναι οι εξής :

- Q20_a_1 - Αριθμός ατόμων με τα οποία συζούσες πριν την πανδημία: 0.01
- Q20_b_1 - Αριθμός ατόμων με τα οποία συζούσες κατά τη διάρκεια της πανδημίας: 0.01

Συνεπώς δεν αφαιρούμε καμία, αφού η αμέσως επόμενη μεταβλητή έχει correlation value 5 φορές μεγαλύτερο από αυτές τις 2.

Επόμενο βήμα ήταν η εξέταση των κατηγορικών μεταβλητών που δεν έχουν γίνει ακόμα 'one-hot'. Οι δύο αυτές μεταβλητές είναι :

- Q11 - Κωδικός του εκπαιδευτικού ιδρύματος, διαθέτει 125 διαφορετικές τιμές

- Χώρα στην οποία απαντήθηκε το ερωτηματολόγιο, διαθέτει 24 διαφορετικές τιμές

Θα παρατηρούσαμε μεγάλη αύξηση στη πολυπλοκότητα των παρατηρήσεων κάτι το οποίο θα είχε μεγαλύτερη ζημιά απότι όφελος για την διαδικασία της εκπαίδευσης. Για αυτό το λόγο επιλέγουμε να τις αφαιρέσουμε και να μελετήσουμε την τοπική συμπεριφορά που διαισθητικά θα προσέδισε η μεταβλητή της χώρας, από το Dataset που αφορά αποκλειστικά το Βέλγιο. Όσον αφορά τα πανεπιστήμια, για αποφυγή της "κατάρας της διαστατικότητας" τα αφαιρούμε και μπορεί να αποτελέσει αντικείμενο μελλοντικής επέκτασης η αποκλειστική μελέτη τους.

Τέλος παρουσιάζουμε τις **64** μεταβλητές που περιέχει το Dataset μας.

- **Q1 - Φύλλο**: 3 μεταβλητές
- **Q2 - Ηλικία**
- **Q3 - Στάτους Σχέσης**: 3 μεταβλητές
- **Q7 - Υψηλότερο επίπεδο μόρφωσης πατέρα**
- **Αριθμός ατόμων από τα οποία μπορεί να δανειστείς 500 ευρώ**
- **Q9 - Πεδίο σπουδών** : 9 μεταβλητές
- **Q10 - Πρόγραμμα Σπουδών** : 7 μεταβλητές
- **Q15 - Σημαντικότητα Σπουδών για τον φοιτητή**
- **Q17_a - Επαρκείς οικονομικοί πόροι πριν την πανδημία**
- **Q17_b - Επαρκείς οικονομικοί πόροι κατά τη διάρκεια της πανδημίας**
- **Ώρες για προσωπικές δραστηριότητες πριν την πανδημία κάθε εβδομάδα**
- **Ώρες για προσωπικές δραστηριότητες κατά τη διάρκεια της πανδημίας κάθε εβδομάδα**
- **Q20_a_1 - Αριθμός ατόμων με τα οποία συζούσες πριν την πανδημία**
- **Q20_b_1 - Αριθμός ατόμων με τα οποία συζούσες κατά τη διάρκεια της πανδημίας**
- **Q21_a_1 - Πόσο συχνά κάπνιζες πριν την πανδημία**
- **Q21_b_1 - Πόσο συχνά κάπνιζες κατά τη διάρκεια της πανδημίας**
- **Q21_c_1 - Πόσα τσιγάρα κάπνιζες καθημερινά πριν την πανδημία**
- **Q21_d_1 - Πόσα τσιγάρα κάπνιζες καθημερινά κατά τη διάρκεια της πανδημίας**

- **Q23_a - Πόσο συχνά έπινες περισσότερα από 6 ποτήρια αλκοόλ πριν την πανδημία**
- **Q23_b - Πόσο συχνά έπινες περισσότερα από 6 ποτήρια αλκοόλ κατά τη διάρκεια της πανδημίας**
- **Q25_a - Πόσο συχνά έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά πριν την πανδημία**
- **Q25_b - Πόσο συχνά έκανες έντονες ασκήσεις για τουλάχιστον 30 λεπτά κτά τη διάρκεια της πανδημίας**
- **Q26_a - Πόσο συχνά έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά πριν την πανδημία**
- **Q26_b - Πόσο συχνά έκανες ήπιες ασκήσεις για τουλάχιστον 30 λεπτά κτά τη διάρκεια της πανδημίας**
- **Αριθμός υποκείμενων ασθενειών**
- **Q29 - Έχεις διαγνωστεί με Covid-19**: 2 μεταβλητές
- **Q30_a_c - Ανησυχία νόσησης από Covid-19**
- **Q30_a_d - Ανησυχία σοβαρής νόσησης από Covid-19**
- **Q32_a_d - Ανησυχία ότι τα νοσοκομεία δεν έχουν επαρκής πόρους για την αντιμετώπιση της πανδημίας**
- **Q34 - Σε τι βαθμό συμμορφώνεσαι στα μέτρα που υλοποιεί η κυβέρνηση για την πανδημία**
- **Αριθμός δραστηριοτήτων ελεύθερου χρόνου κατά την διάρκεια της τελευταίας εβδομάδας**
- **Q36_a - Είχες περισσότερη ή λιγότερη επαφή με την οικογένειά σου από τότε που υλοποιήθηκαν τα μέτρα κατά της πανδημίας**
- **Q36_b - Είχες περισσότερη ή λιγότερη επαφή με τους φίλους σου από τότε που υλοποιήθηκαν τα μέτρα κατά της πανδημίας** Η επιλογή "λιγότερη" αντιστοιχίζεται στην τιμή 0, "περίπου ίδια" στην τιμή 0 και "περισσότερη" στην τιμή 1.
- **Q37 - Ύπαρξη ατόμου για συζήτηση σχετικά με προσωπικά ζητήματα** : 2 μεταβλητές
- **Q39_a - Αφού ξέσπασε η πανδημία, αναζήτησες περισσότερο ή λιγότερο επικοινωνία με καθηγητές για να συζητήσετε για ανησυχίες σχετικά με τα μαθήματα**
- **Q39_b - Αφού ξέσπασε η πανδημία, αναζήτησες περισσότερο ή λιγότερο επικοινωνία με καθηγητές για να συζητήσετε σχετικά με ψυχοκοινωνικά προβλήματα**

- **Αριθμός ανησυχιών για τις οποίες μίλησες με ειδικούς από όταν ξέσπασε η πανδημία**
- **Q41_a έως Q41_h - 8 Ακαδημαϊκό Στρες-Ικανοποίηση**
- **Γνώση για τον Covid-19**
- **Q43_a - Η κυβέρνηση παρείχε έγκαιρα πληροφορίες σχετικά με τον Covid-19**
- **Q43_b - Η κυβέρνηση παρείχε κατανοητές πληροφορίες σχετικά με τον Covid-19**
- **Μέση Αυστηρότητα Μέτρων**

Πλέον έχουμε ένα έτοιμο Dataset μεγέθους (85461, 65). Στο οποίο θέτουμε ως target το σκόρ κατάθλιψης.

5.4.2 Εκπαίδευση μοντέλων

Η εκπαίδευση μοντέλων πρέπει να χωριστεί σε 2 υποομάδες, όσα είναι και τα tasks μας (multiclass,binary calssification). Σε αυτές τις 2 υποομάδες υπάρχουν τα μοντέλα που εκπαιδεύονται με δεδομένα που αφορούν όλο τον κόσμο και μοντέλα που εκπαιδεύονται με δεδομένα που προκύπτουν μόνο από το Βέλγιο.

Muliclass Models: Όπως έχουμε αναφέρει οι 3 κλάσεις που δημιουργούνται για τα επίπεδα κατάθλιψης είναι αρκετά ισομοιρασμένες στην έκταση των δεδομένων μας, επομένως δεν χρειάζεται να πραγματοποιήσουμε κάποια μέθοδο για imbalance datas.

Χρησιμοποιούμε το ίδιο Dataset (85461, 65) για όλα τα μοντέλα που αφορούν τον κόσμο (World) και το ίδιο Dataset (20587,65) για όλα τα μοντέλα που αφορούν το Βέλγιο (Belgium). Αν αφαιρέσουμε την target μεταβλητή μας έχουμε 64 features, τα οποία για τις ανάγκες εισόδου των CNN μοντέλων μετατρέπουμε σε διαστάσεις (μήκος εικόνας, πλάτος εικόνας, χρωματικά κανάλια εικόνας) $\rightarrow$ (8,8,1). Το split που πραγματοποιούμε από το αρχικό Dataset είναι :

- Train Set: 64%
- Validation Set: 16%
- Test Set: 20%

Η εκπαίδευση πραγματοποιείται με την χρήση της μεθόδου Particle Swarm Optimization τόσο σε MLP_World,MLP_Belgium αρχιτεκτονική όσο και σε CNN_World,CNN_Belgium. Λεπτομέρειες των παραμέτρων και υπερπαραμέτρων που χρησιμοποιήθηκαν αναλύονται στην επόμενη ενότητα.

Binary Models: Στην ενότητα Ανισορροπία Δεδομένων εξηγούμε αναλυτικά τα τελικά μας Dataset τα οποία είναι 2 για τον κόσμο (World) και 3 για το Βελγιο (Belgium). Το split που χρησιμοποιούμε από το αρχικό Dataset είναι :

- Train Set: 60%
- Validation Set: 15%
- Test Set: 25%

Ομοίως πραγματοποιούμε εκπαίδευση με τη χρήση Particle Swarm Optimization, 2 φορές για MLP_World αρχιτεκτονικές , 2 φορές για CNN_World αρχιτεκτονικές και 3 φορές για MLP_Belgium αρχιτεκτονικές , 2 φορές για CNN_Belgium αρχιτεκτονικές. Λεπτομέρειες των παραμέτρων και υπερπαραμέτρων που χρησιμοποιήθηκαν αναλύονται στην επόμενη ενότητα.

5.4.3 Ορισμός υπερπαράμετρων με τη χρήση Particle Swarm Optimization

Αρχικά οι παράμετροι του Particle Swarm Optimization που ορίστηκαν είναι οι εξής :

- 10 σωματίδια που αναζητούν την βέλτιστη θέση
- συντελεστές επιτάχυνσης c1,c2 = 0.5
- **Τερματική Συνθήκη**: 30 αναζητήσεις-μετακινήσεις των particles ή απόσταση μεταξύ δυο διαδοχικών θέσεων μικρότερη του 1

Στόχος του PSO είναι πάντοτε η ελαχιστοποίηση μια αντικειμενικής συνάρτησης, συνεπώς στη περίπτωση μας θέτουμε αυτή τη συνάρτηση να είναι η

$$f(accuracy) = \frac{1}{1 + accuracy}$$

, έτσι επιτυγχάνουμε την μεγιστοποίηση της ακρίβειας.

Σταθερές τιμές στις υπερπαραμέτρους είναι ο **αριθμός των εποχών** που τέθηκε εμπειρικά ύστερα από μελέτη και πολλαπλές εκτελέσεις του κώδικα:

- Mulitclass MLP_World: 300 epochs
- Mulitclass CNN_World: 20 epochs

- Binary MLP_World (model1,model2): 150 epochs
- Binary CNN_World (model1,model2): 20 epochs

- Mulitclass MLP_Belgium: 300 epochs
- Mulitclass CNN_Belgium: 10 epochs
- Binary MLP_Belgium (model1,model2,model3): 150 epochs
- Binary CNN_Belgium (model1,model2,model3): 20 epochs

Στη συνέχεια ορίζουμε διαστήματα τιμών μέσα στα οποία θα γίνει η αναζήτηση της βέλτιστης αρχιτεκτονικής. Τα διαστήματα που έχουν οριστεί μπορούν να βρεθούν αναλυτικά στον κώδικα, και οι υπερπαράμετροι που τίθονται προς βελτιστοποίηση είναι οι εξής :

- Αριθμός Στρωμάτων δικτύου (Depth)
- Τιμή πιθανότητας Dropout Στρώματος (Drop)
- Batch Size (Batch)
- Αρχική τιμή learning rate (LR)
- Συνάρτηση optimizer (Opt)

Σε κάθε μοντέλο χρησιμοποιούμε LearningRateDecay για να βοηθήσουμε την διαδικασία εκπαίδευσης

Ο αλγόριθμος Particle Swarm Optimization δεν αποθηκεύει την αρχιτεκτονική του βέλτιστου μοντέλου παρά μόνο κάνει αναζήτηση. Για την εύρεση του βέλτιστου μοντέλου θέτουμε σαν threshold αποθήκευσης:

- Multiclass : Validation Accuracy > 0.58
- Binary: Validation Accuracy > 0.75

5.4.4 Συλλογή Μοντέλων

Από την πληθώρα μοντέλων που προκύπτουν, επιλέγουμε με κριτήριο το accuracy που έχουν σημειώσει διάφορα μοντέλα δημιουργώντας μια συλλογή μοντέλων που ύστερα θα χρησιμοποιηθούν μόνα τους ή συλλογικά (ensemble). Παρακάτω παραθέτουμε τα μοντέλα με τις αρχιτεκτονικές που προέκυψαν, θέτουμε **B** ως Binary και **M** ως MultiClass :

MLP								
Model	Depth	Drop1	Drop2	LR	Opt	Batch	Accuracy	Loss
M-World-1	0.6	0.19	0.14	0.001	1	1.6	0.615	0.808
M-World-2	0.9	0.39	0.36	0.001	1	2.6	0.613	0.82
B-World-1	0.1	0.38	0.24	0.001	1	0.86	0.772	0.474
B-World-2	2.9	0.39	0.14	0.001	1	1.2	0.774	0.467
M-Belgium-1	0.01	0.37	0.32	0.001	1	2.8	0.605	0.814
M-Belgium-2	0.9	0.259	0.146	0.001	1	2.6	0.601	0.812
B-Belgium-1	2.7	0.35	0.21	0.001	1	2.9	0.782	0.468
B-Belgium-2	1.8	0.19	0.27	0.001	1	3	0.764	0.508
B-Belgium-3	1.4	0.01	0.36	0.001	1	2.6	0.767	0.475

Και αντίστοιχα για τα CNN μοντέλα.

CNN								
Model	Drop1	Drop2	Drop3	Drop4	Depth	Batch	Accuracy	Loss
M-World-1	0.11	0.18	0.03	0.06	1.89	0.8	0.615	0.804
M-World-2	0.18	0.16	0.04	0.19	0.24	1.82	0.613	0.813
B-World-1	0.06	0.1	0.17	0.2	0.77	0	0.775	0.47
B-World-2	0.07	0.14	0.04	0.01	1.17	0.6	0.773	0.467
M-Belgium-1	0.01	0.368	0.318	0.13	1.19	2.67	0.598	0.81
M-Belgium-2	0.32	0.22	0.29	0.13	1.57	0.88	0.594	0.85
B-Belgium-1	0.17	0.05	0.1	0.05	2.1	1.68	0.774	0.478
B-Belgium-2	0.01	0.1	0.2	0.13	1.25	2.97	0.759	0.495
B-Belgium-3	0.01	0.01	0	0.1	0.9	2.9	0.767	0.475

Οι τιμές που αναγράφονται δεν είναι οι πραγματικές τιμές της αρχιτεκτονικής των μοντέλων, αλλά το κωδικοποιημένο διάνυσμα που αν τοποθετηθεί σαν είσοδο στη συνάρτηση κατασκευής μοντέλου θα παραγάγει την πραγματική αρχιτεκτονική.

5.4.5 Αξιολόγηση Μοντέλων

Έχοντας καταλήξει στα υποψήφια τελικά μοντέλα , θα τα αξιολογήσουμε με το αντίστοιχο Test Set. Τα αποτελέσματα που θα παρουσιαστούν για το Multiclass Classification προκύπτουν από τα εξής μοντέλα :

- Model1: MLP M-World-1
- Model2: MLP M-Belgium-1
- Model3: CNN M-World-1

- Model4: CNN M-Belgium-1
- Model5 :Ensemble (MLP M-World-1,MLP M-World-1,CNN M-World-1,CNN M-World-2) με χρήση του αθροίσματος πιθανοτήτων
- Model6 :Ensemble (MLP M-Belgium-1,MLP M-Belgium-2,CNN M-Belgium-1,CNN M-Belgium-2) με χρήση του αθροίσματος πιθανοτήτων

Και για το Binary Classification από τα εξής μοντέλα:

- Model7,8,9,10: Ensemble (MLP B-Belgium-1,MLP B-Belgium-2,MLP B-Belgium-3) με χρήση της μέσης τιμής πιθανότητας, μέγιστης πιθανότητας, ελάχιστης πιθανότητας και weighted συντελεστών
- Model11,12,13,14: Ensemble (CNN B-Belgium-1,CNN B-Belgium-2,CNN B-Belgium-3) με χρήση της μέσης τιμής πιθανότητας, μέγιστης πιθανότητας, ελάχιστης πιθανότητας και weighted συντελεστών
- Model15,16,17,18: Ensemble (MLP B-World-1,MLP B-World-2) με χρήση της μέσης τιμής πιθανότητας, μέγιστης πιθανότητας, ελάχιστης πιθανότητας και weighted συντελεστών
- Model19,20,21,22: Ensemble (CNN B-World-1,CNN B-World-2) με χρήση της μέσης τιμής πιθανότητας, μέγιστης πιθανότητας, ελάχιστης πιθανότητας και weighted συντελεστών

Για λόγους οικονομίας και ευκολίας του αναγνώστη θα παρουσιάσουμε τα αποτελέσματα που θεωρούμε τα καλύτερα, δηλαδή το μοντέλο που ταιριάζει, σύμφωνα με την άποψη μας, καλύτερα στο προβλημα Multiclass Classification World, Multiclass Classification Belgium, Binary Classification World, Binary Classification Belgium. Για τα πλήρη αποτελέσματα επικοινωνήστε με τους συντάκτες.

	precision	recall	f1-score	support
0.0	0.79	0.78	0.79	2189
1.0	0.75	0.75	0.75	1886
accuracy			0.77	4075
macro avg	0.77	0.77	0.77	4075
weighted avg	0.77	0.77	0.77	4075

Σχήμα 34: Model10-Binary Belgium

	precision	recall	f1-score	support
0.0	0.66	0.65	0.66	1221
1.0	0.50	0.50	0.50	1423
2.0	0.71	0.72	0.72	1474
accuracy			0.62	4118
macro avg	0.62	0.62	0.62	4118
weighted avg	0.62	0.62	0.62	4118

Σχήμα 35: Model6-Multiclass Belgium

	precision	recall	f1-score	support
0.0	0.75	0.78	0.76	8472
1.0	0.77	0.75	0.76	8584
accuracy			0.76	17056
macro avg	0.76	0.76	0.76	17056
weighted avg	0.76	0.76	0.76	17056

Σχήμα 36: Model16-Binary World

	precision	recall	f1-score	support
0.0	0.65	0.72	0.68	5608
1.0	0.49	0.46	0.47	5858
2.0	0.70	0.68	0.69	5627
accuracy			0.62	17093
macro avg	0.61	0.62	0.62	17093
weighted avg	0.61	0.62	0.61	17093

Σχήμα 37: Model1 - Multiclass World

Παραθέτουμε και την τιμή του Area Under the Curve για κάθε ένα από τα παραπάνω μοντέλα:

- Model10: 0.7687564158399245
- Model6: 0.7166757017804245
- Model16: 0.7619975829595347
- Model1: 0.7127730072799813

Να αναφέρουμε πως στην περίπτωση των Model1 και Model6 η τιμή του AUC, έχει υπολογιστεί με την μέθοδο One vs One, δηλαδή βρίσκουμε για κάθε ζεύγος κλάσεων την τιμή και ύστερα χρησιμοποιούμε το average για το τελικό αποτέλεσμα.

Όπως αναφέραμε σε προηγούμενες ενότητες είναι πάρα πολύ σημαντικό η τιμή του Recall της ευπαθούς ομάδας να είναι όσο το δυνατόν πιο ψηλά. Σύμφωνα με αυτό το κριτήριο θα επιλέξουμε το Binary μοντέλο (World είτε Belgium).

5.4.6 Μοντέλο Βελγίου σε Παγκόσμια Δεδομένα

Τελευταίο πείραμα που πραγματοποιήσαμε είναι να εφαρμόσουμε το Test Dataset που αφορά φοιτητές από όλο τον κόσμο στο μοντέλο που έχει εκπαιδευθεί με δεδομένα από το Βέλγιο. Θέλουμε να παρατηρήσουμε αν το τοπικό χαρακτηριστικό μιας μόνο χώρας αλλάζει σημαντικά την προβλεπτική ικανότητα του μοντέλου όταν αυτό θελήσει να γενικεύσει. Ο έλεγχος πραγματοποιήθηκε στο Model10 και τα αποτελέσματα είναι τα εξής :

	precision	recall	f1-score	support
0.0	0.71	0.84	0.77	8472
1.0	0.81	0.67	0.73	8584
accuracy			0.75	17056
macro avg	0.76	0.75	0.75	17056
weighted avg	0.76	0.75	0.75	17056

Σχήμα 38: Belgium Model to World's Dataset

Η τιμή του Area Under the Curve για την παραπάνω περίπτωση είναι : **0.7547370010921345**.

Συνεπώς μπορούμε να θεωρήσουμε πως ένα μοντέλο που έχει προσαρμοστεί στις ανάγκες μια χώρας υπό συνθήκες **μπορεί να μεταφερθεί και σε άλλες χώρες**, χωρίς επιπλέον εκπαίδευση. Αρκετά σημαντικό σε περιπτώσεις όπως οι τωρινές που η πανδημία δεν αφήνει μεγάλα περιθώρια χρόνου

6 Συμπεράσματα

Συνοψίζοντας τα αποτελέσματα μας, είμαστε σε θέση πλέον να συμπεράνουμε ότι οι φοιτητές από διαφορετικές χώρες, διαφορετικά ανώτατα εκπαιδευτικά ιδρύματα, διαφορετικούς τομείς/είδη σπουδών διαφοροποιούνται στατιστικά ως προς τα επίπεδα κατάθλιψης.

Επίσης συμπεραίνουμε ότι τα δεδομένα που παρέχει το ερωτηματολόγιο έχουν την δυνατότητα να εκπαιδεύσουν ως ένα βαθμό διάφορα προβλεπτικά μοντέλα, με μέγιστη τιμή του Recall να φτάνει το 78%. Ένα εργαλείο με αυτή την ικανότητα μπορεί να εφαρμοστεί σε έκτακτες περιπτώσεις, όπως μια πανδημία, άλλα δεν μπορεί να θεωρηθεί απολύτως αξιόπιστο.

Είναι πολύ σημαντικό να επισημάνουμε ότι η προσέγγιση που ακολουθήσαμε δεν έχει εφαρμοστεί σε κάποια άλλη έρευνα. Συνεπώς η παρούσα εργασία, αποτελεί μια καινοτόμα προσέγγιση για την άμεση και επιτακτική ανάγκη διάγνωσης της κατάθλιψης.

Φυσικά μελλοντικές επεκτάσεις επιδέχεται η παρούσα μεθοδολογία, αλλά θα είναι εφικτές με την παροχή ποιοτικότερων δεδομένων. Με αυτή την προϋπόθεση αλλά και την παροχή υπολογιστικών πόρων θα μπορούσαν να εκπαιδευτούν πολύ πιο ισχυρά μοντέλα τόσο ομαδοποίησης όσο και παλινδρόμησης/ταξινόμησης. Επίσης θα μπορούσαν να παρέχονται αντίστοιχα δεδομένα, αλλά με την πληροφορία του επιπέδου κατάθλιψης χωρίς την πανδημία. Με αυτόν τον τρόπο, θα ήταν δυνατό να εκτιμηθεί η ποσοτική αλλαγή στο επίπεδο κατάθλιψης ενός πανεπιστημίου με την εφαρμογή αντίστοιχων μέτρων.

Τέλος παραθέτουμε ορισμένες προσωπικές σκέψεις ύστερα από την εκτενής μελέτη που πραγματοποιήθηκε. Αρχικά για τα εκπαιδευτικά ιδρύματα είναι εμφανής η ανάγκη να αλλάξουν τους τρόπους με τους οποίους λαμβάνουν μέτρα για την αντιμετώπιση τέτοιου είδους κρίσεων όπως η πανδημία του Covid-19 ώστε να διασφαλίζουν την ψυχολογική ευημερία των φοιτητών. Πρέπει σε παρόμοιες περιπτώσεις να φροντίζουν ώστε να μην αυξάνεται υπερβολικά ο φόρτος εργασίας, να μην μειώνεται η ποιότητα της εκπαίδευσης, να παίρνουν ισχυρότερα μέτρα προστασίας και να δώσουν μεγαλύτερη έμφαση σε συμβουλευτικές ομάδες στις οποίες θα μπορούν να απευθυνθούν οι φοιτητές. Ιδιαίτερη σημασία πρέπει να δοθεί στους μεταπτυχιακούς φοιτητές, καθώς και στα non-binary άτομα, μια ήδη περιθωριοποιημένη σε μεγάλο βαθμό κοινωνική ομάδα. Όσο αφορά τις κυβερνήσεις, σε τέτοιες περιπτώσεις θα πρέπει να παρέχουν πληροφορίες πιο έγκαιρα, με μεγαλύτερη σαφήνεια και να οργανώνουν σχετικές εκστρατείες ώστε να ενημερώνεται όσο το δυνατόν περισσότερο το σύνολο του πληθυσμού τις εκάστοτε χώρας κάτι που θα έχει ως αποτέλεσμα την μείωση του φόβου των πολιτών. Για το τελευταίο θα μπορούσαν να συμβάλλουν και τα εκπαιδευτικά ιδρύματα της χώρας. Επίσης, οι κυβερνώντες σε συνδυασμό με τα πανεπιστήμια θα πρέπει να κάνουν προσπάθειες ώστε τα μέτρα να μην διαταράσσουν σε μεγάλο βαθμό τις καθημερινές συνήθειες των φοιτητών, ιδιαιτέρως τις ώρες που χρειάζονται για μελέτη, μαθήματα και εργασία, και να γίνονται συχνές αναφορές στην σημασία της σωματικής άσκησης.

Παρουσιάζεται η συνεισφορά της ομάδας:

Ονοματεπώνυμο	Στατιστικοί έλεγχοι	Παλινδρόμηση	Ομαδοποίηση	Ταξινόμηση
Παναγιώτης Κοκκινάκος	0.40	0.05	0.50	0.05
Ιωάννα Μανδηλαρά	0.50	0.05	0.40	0.05
Φίλιππος Σκόβελεφ Ορφανουδάκης	0.05	0.05	0.05	0.85
Άρης Σπύρου	0.05	0.85	0.05	0.05

Πίνακας 13: Συνεισφορά ατόμων στην εργασία

Βιβλιογραφία

[1] COVID-19 International Student Well-being Study, https://www.uantwerpen.be/en/research-groups/centre-population-family-health/research2/covid-19-internation/

[2] Stathopoulou Theoni, Mouriki Aliki, & Papaliou Olga. (2020, September 19). STUDENT WELL-BEING DURING THE COVID-19 PANDEMIC IN GREECE. RESULTS FROM THE C19 ISWS SURVEY. (Version 1). Zenodo.

[3] Busse, H.; Buck, C.; Stock, C.; Zeeb, H.; Pischke, C.R.; Fialho, P.M.M.; Wendt, C.; Helmer, S.M. Engagement in Health Risk Behaviours before and during the COVID-19 Pandemic in German University Students: Results of a Cross-Sectional Study. Int. J. Environ. Res. Public Health 2021, 18, 1410. https://doi.org/10.3390/ijerph18041410

[4] Rabiee- Khan, Biernat. (2021, March 1). Student well-being during the first wave of COVID-19 pandemic in Birmingham, UK. Zenodo. http://doi.org/10.5281/zenodo.4572408

[5] Super, S., Van Disseldorp, L. (2020, June 24). Covid-19 International Student Well-being Study (C19 ISWS) - Data from Wageningen University Research. Zenodo. http://doi.org/10.5281/zenodo.3906209

[6] Codebook C19 ISWS, https://zenodo.org/record/4074979#.YOsnhFT7SUk

[7] Covid-19: Stringency Index, https://ourworldindata.org/grapher/covid-stringency-index

[8] Scipy One Way Anova, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

[9] Scipy Chi2 Test, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html

[10] Scipy t-test, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

[11] Sklearn MinMax, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[12] Sklearn Silhouette Score,
https://scikit-learn.org/stable/modules/generated
/sklearn.metrics.silhouette_score.html

[13] Sklearn Calinski and Harabasz Score,
https://scikit-learn.org/stable/modules/generated
/sklearn.metrics.calinski_harabasz_score.html

[14] Sklearn Kmeans,
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[15] Finding Groups in Data: An Introduction to Cluster Analysis, Kaufman Leonard, and Peter J. Rousseeuw,
https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

[16] Zhang, Wendy, et al. "Validating a shortened depression scale (10 item CES-D) among HIV-positive people in British Columbia, Canada." PloS one 7.7 (2012): e40793.

[17] Salunkhe, Uma R., and Suresh N. Mali. "Classifier ensemble design for imbalanced data classification: a hybrid approach." Procedia Computer Science 85 (2016): 725-732.

[18] James Kennedy and Russell Eberhart. "Particle swarm optimization".In:Proceedings of ICNN'95-International Conference on NeuralNetworks. Vol. 4. IEEE. 1995, pp. 1942–1948.