Η απόδοση των μαθητών λαμβάνοντας υπόψη παράγοντες της καθημερινότητας τους

Ιωάννου Νικόλας - ge20718
Μπέκος Θοδωρής - ge

Δεκέμβριος 2024

Σχολή εφαρμοσμένων μαθηματικών και φυσικών επιστημών

Μάθημα: ΘΕΜΑ

Υπέυθυνος καθηγητής: Δρ. Στεφανέας Πέτρος

Ευχαριστίες:

Θα θέλαμε να ευχαριστήσουμε τον κύριο Πέτρο Στεφανέα για το ενδιαφέρον και την εμπιστοσύνη που μας έδειξε κατά τη διάρκεια πραγματοποίησης της εν λόγω εργασίας. Επίσης, θα θέλαμε να ευχαριστήσουμε τις οικογένειες μας γιατί χωρίς αυτές δεν θα μπορούσαμε να βρισκόμασταν στη θέση που βρισκόμαστε τώρα.

Περιεχόμενα

1 Εισαγωγή **4**

2 Introduction **5**

3 Ανάλυση της διαδικασίας επιλογής μοντέλου **6**

3.1 Ανάλυση του προβλήματος 6

3.2 Επεξεργασία των δεδομένων 7

3.3 Επιλογή χαρακτηριστικών 14

3.4 Επιλογή υπερπαραμέτρων 16

3.5 Επιλογή μοντέλου 17

4 Το μοντέλο παλινδρόμισης Lasso **20**

4.1 Έγκριση ειδικού πλαισίου χωροταξικού σχεδιασμού . . . 20

4.2 Καθορισμός περιοχών αιολικής προτεραιότητας 20

4.3 Παράμετροι που καθορίζουν τις αποστάσεις 20

5 Σύγκριση με έρευνες **21**

6 Βιβλιογραφία **22**

1 Εισαγωγή

km

2 Introduction

gt

3 Ανάλυση της διαδικασίας επιλογής μοντέλου

3.1 Ανάλυση του προβλήματος

Το πρόβλημα το οποίο αντιμετωπίσαμε έχει να κάνει με την κατασκευή μοντέλου μηχανικής μάθησης για την πρόβλεψη των βαθμών των μαθητών μέσω διαφόρων παραγόντων που θα αναλύσουμε παρακάτω. Το πρόβλημα αυτό αποτελεί πρόβλημα παλινδρόμισης αφού ο τελικός βαθμός δεν αποτελεί κατηγορική αλλά ποσοτική μεταβλητή και παίρνει συνεχείς τιμές. Χρησιμοποιήσα το σύνολο δεδομένων Student Performance Factors που βρήκαμε στο Kaggle ,όπου τα χαρακτηριστικά τα οποία το αποτελούν φαίνονται στον πίνακα παρακάτω:

Χαρακτηριστικά	Περιγραφή	Εύρος τιμών
Attendance	Ποσοστό μαθημάτων που παρευρέθηκε ο μαθητής	0-100
Previous Scores	Μέσος όρος βαθμών απο προηγούμενες εξετάσεις	0-100
Sleep Hours	Ώρες ύπνου ανά μέσο όρο κάθε βράδυ	4-10
Hours Studied	Αριθμός ωρών που δίαβασε ο μαθητής μέσα στην βδομάδα	0-50
Tutoring Sessions	Αριθμός φροντιστηριακών μαθημάτων ανά μήνα	0-10
Family income	Οικογενειακό εισόδημα	Low/Medium/ High
Teacher quality	Εκπαιδευτική ικανότητα καθηγητή	Low/Medium/ High
Parental Involvement	Βαθμός εμπλοκής γονέων στην εκπαίδευση του μαθητή	Low/Medium/ High
Access to Resources	Πρόσβαση μαθητή σε εκπαιδευτικό υλικό	Low/Medium/ High

Motivation Level	Κίνητρο μαθητή	Low/Medium/ High
Peer Influence	Επιρροή μαθητή απο ομήλικους του	Negative/ Neutral/ Positive
Distance from home	Απόσταση σχολείου απο το σπίτι	Near/Moderate/ Far
School type	Τύπος σχολείου	Private/Public
Physical activity	Επίπεδο φυσικής δραστηριότητας μαθητή	Sedentary/ Light/Moderate/ Active/ Very Active/ Highly Active/ Athlete
Parental education level	Επίπεδο εκπαίδευσης γονέα	High School/ College/ Postgraduate
Extracurricular activities	Εξωσχολικές δραστηριότητες	Yes/No
Learning disabilities	Μαθησιακές δυσκολίες	Yes/No
Internet Access	Πρόσβαση στο διαδίκτυο	Yes/No
Gender	Φύλο μαθητή	Male/Female
Exam Score	Βαθμός εξέτασης	0-100

Η μεταβλητή Exam Score αποτελεί την εξαρτημένη μεταβλητή(στόχο) του μοντέλου ενώ οι υπόλοιπες μεταβλητές αποτελούν τις πιθανές ανεξάρτητες μεταβλητές.

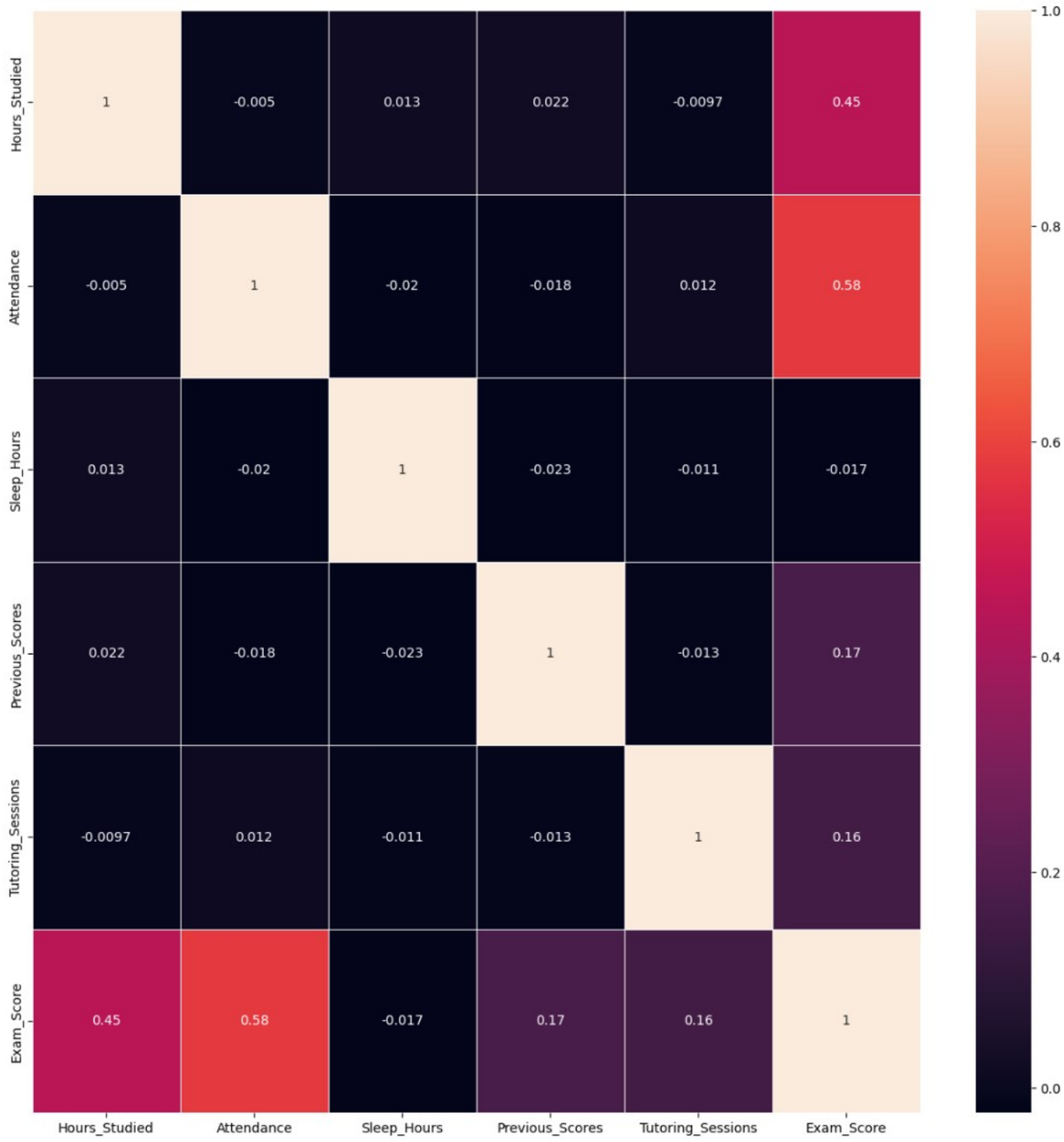
3.2 Επεξεργασία των δεδομένων

Ρίχνωντας μια ματιά στα δεδομένα μπορούμε να παρατηρήσουμε πως υπάρχουν παρατηρήσεις με ελλειπείς τιμές. Σε αυτή την περίπτωση έχουμε

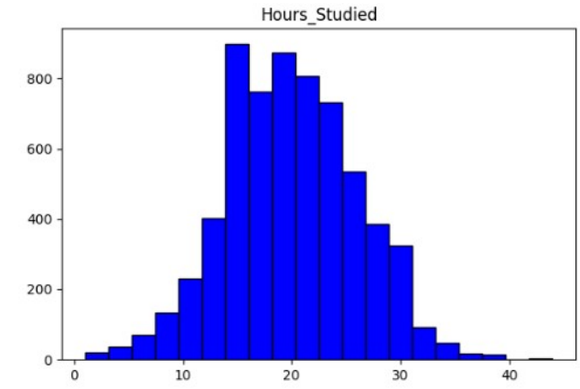
3 επιλογές. Η πρώτη επιλογή είναι να διαγράψουμε τις παρατηρήσεις αυτές, η δεύτερη να προβλέψουμε τις τιμές των παρατηρήσεων με κάποια μέθοδο (Παλινδρόμιση για τις ποσοτικές και ομαδοποίηση για της κατηγορικές μεταβλητές) και η τρίτη να αντικαταστήσουμε τις ελλειπείς τιμές με κάποιο λογικό τρόπο(π.χ. εάν έχουμε ποσοτικές τιμές με την μέση τιμή του χαρακτηρηστικού ή εάν έχουμε κατηγορική με την τιμή, η οποία εμφανίζεται τις περισσότερες φορές). Επιλέξαμε να διαγράψουμε τις παρατηρήσεις οι οποίες έχουν ελλειπείς τιμές μιας και ο αριθμός των παρατηρήσεων είναι ικανοποιητικός και οι ελλειπεις τιμές αποτελούν ένα πολύ μικρό ποσοστό(<10%) των συνολικών παρατηρήσεων και δεν θα επιρεάσουν το αποτέλεσμα.

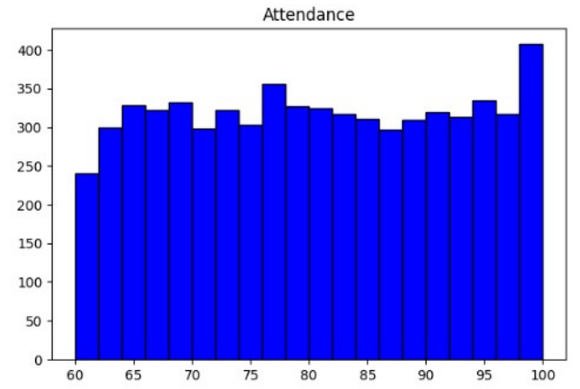
Η μεταβλητή Physical Activity εχει μια ιδιαιτερότητα. Παρόλο που περιέχει αριθμητικές τιμές(0-6), αυτές οι τιμές αποτελούν κατηγορίες. Για αυτό και εμείς αλλάξαμε τον τύπο της μεταβλητής σε αντικείμενο. Μπορούμε να παρατηρήσουμε πως το dataset μας τώρα έχει 14 κατηγορικές μεταβλητές και 6 ποσοτικές.

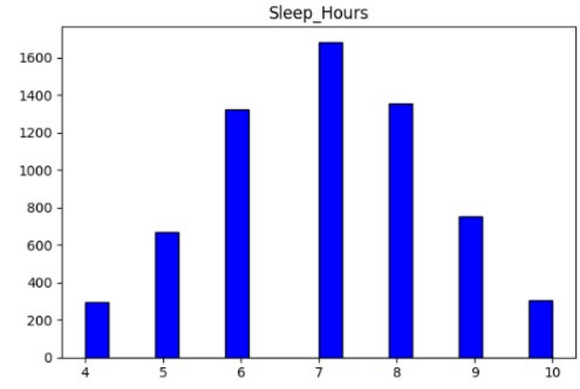
Για να κατανοήσουμε καλύτερα τα δεδομένα μας θα χρησιμοποιήσουμε διάφορες αριθμητικές και γραφικές μεθόδους. Για της ποσοτικές μεταβλητές θα υπολογίσουμε την μέση τιμή, την διασπορά, το εύρος, το ενδοτεταρτημοριακό εύρος, την διάμεσο, το πρώτο τεταρτημόριο(Το 25% των τιμών της μεταβλητής είναι μικρότερα απο αυτή την τιμή), το τρίτο τεταρτημόριο(Το 75% των τιμών της μεταβλητής είναι μικρότερα απο αυτή την τιμή) και την μέγιστη και ελάχιστη τιμή τους. Επίσης, θα κατασκευάσουμε τον πίνακα Pearson των μεταβλητών όπως και τα ιστογράμματα και θηκογράμματα τους τα οποία παρουσιάζονται παρακάτω:

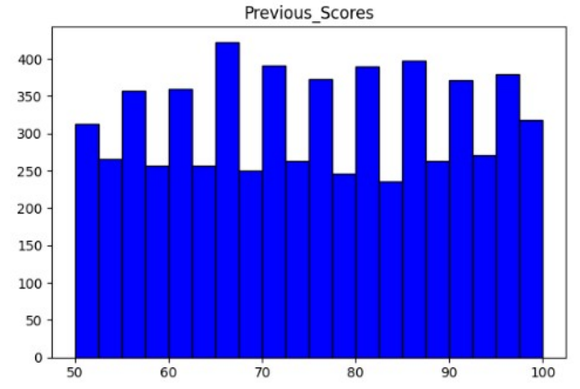


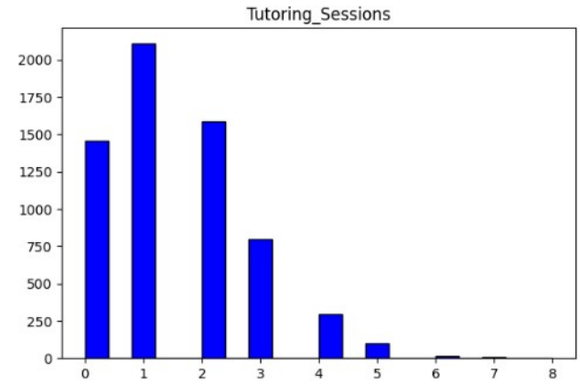
Ο πίνακας Pearson δείχνει την γραμμική συσχέτιση μεταξύ των ποσοτικών μεταβλητών.

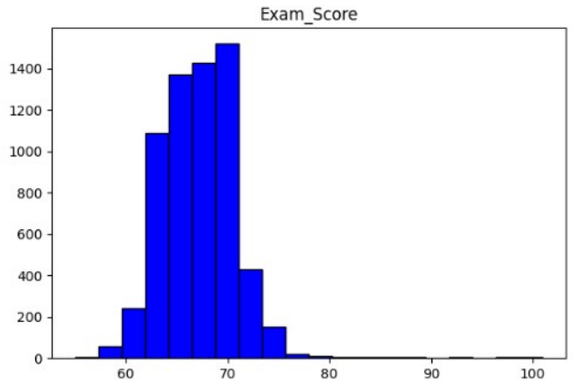

Hours_Studied

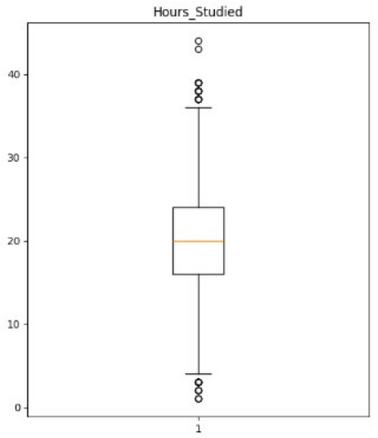

Attendance

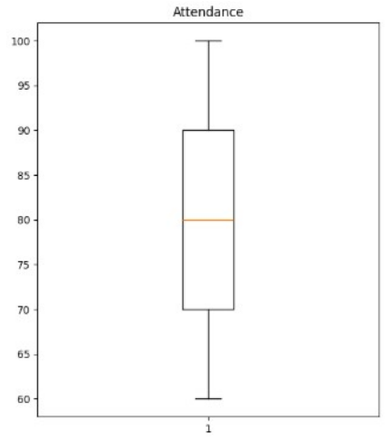

Sleep_Hours

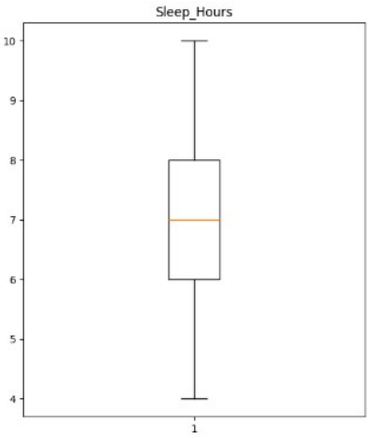

Previous_Scores

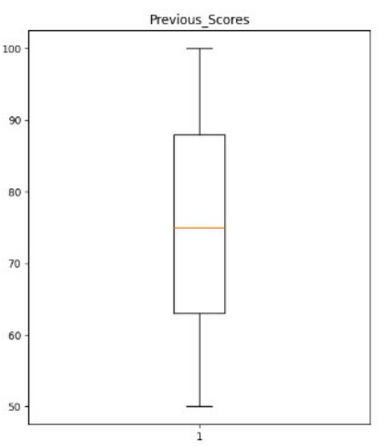

Tutoring_Sessions

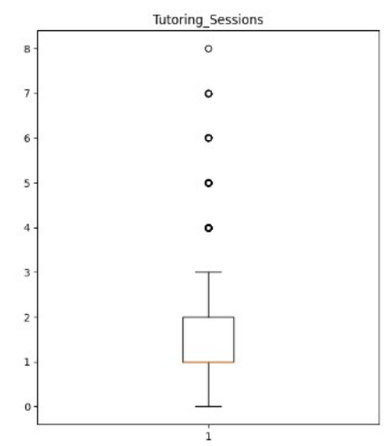

Exam_Score

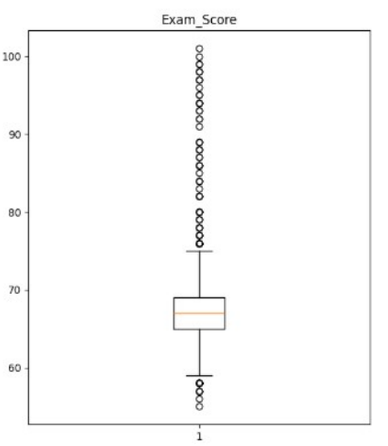

Hours_Studied


Attendance


Sleep_Hours

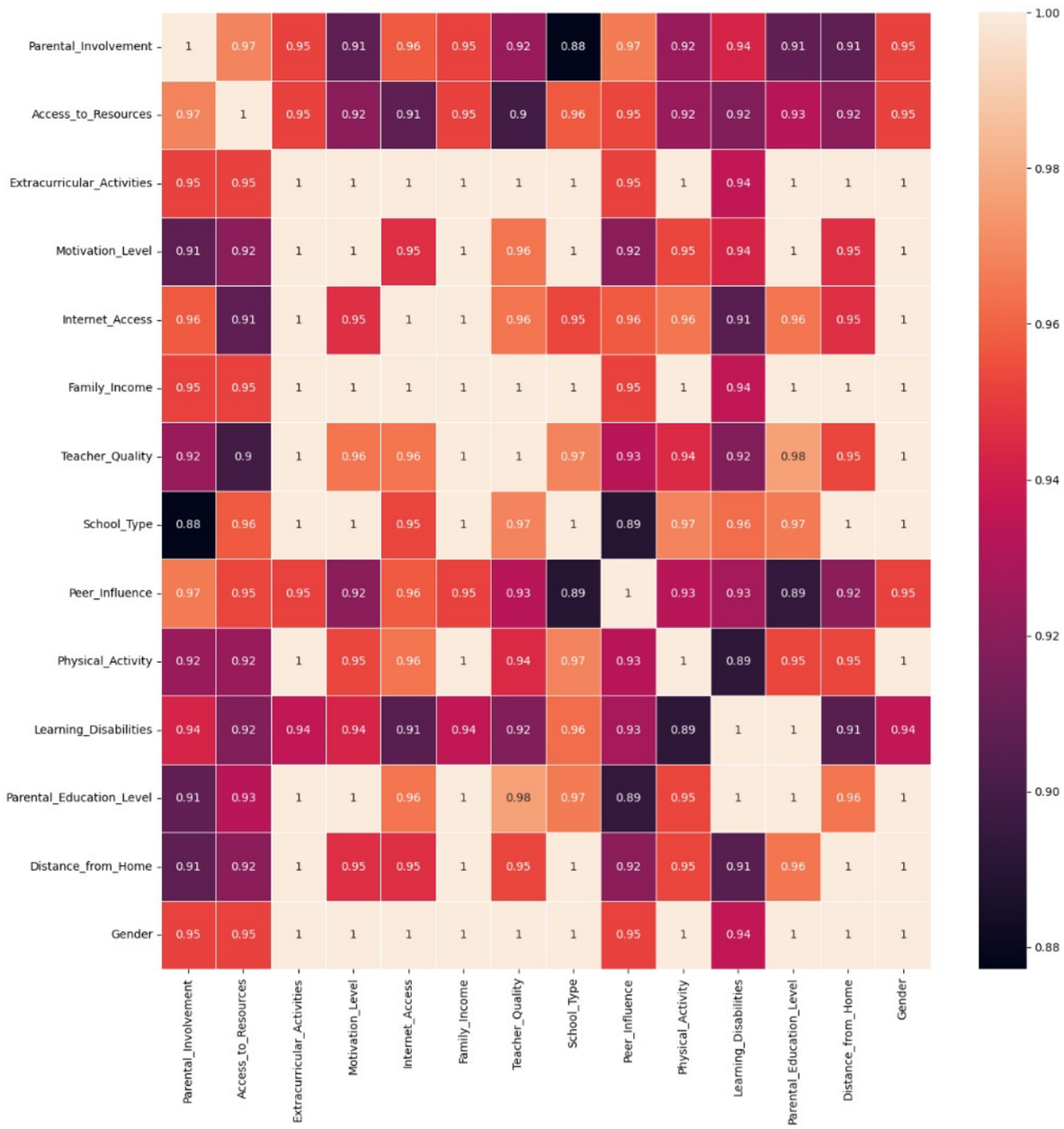

Previous_Scores


Tutoring_Sessions

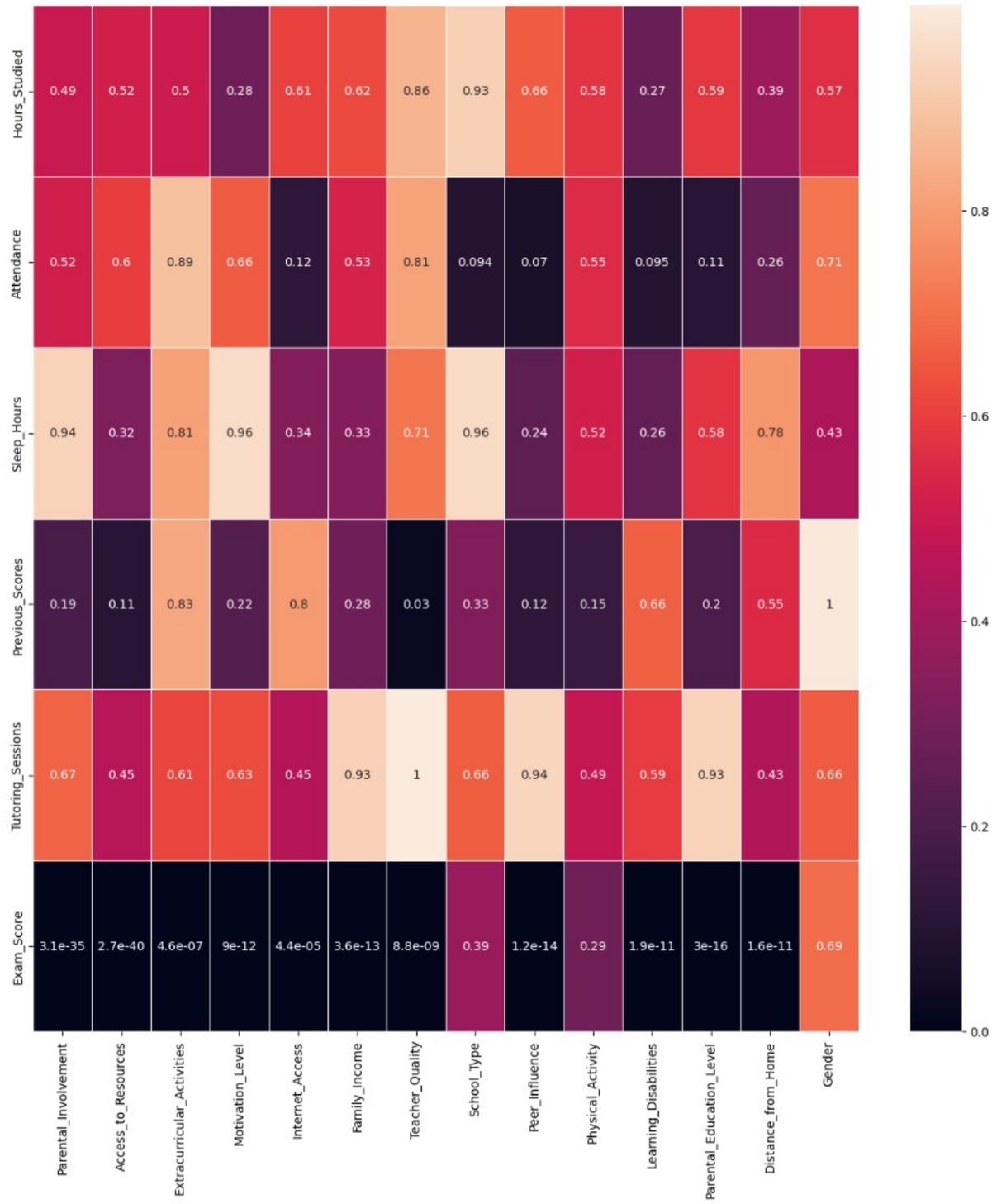

Exam_Score

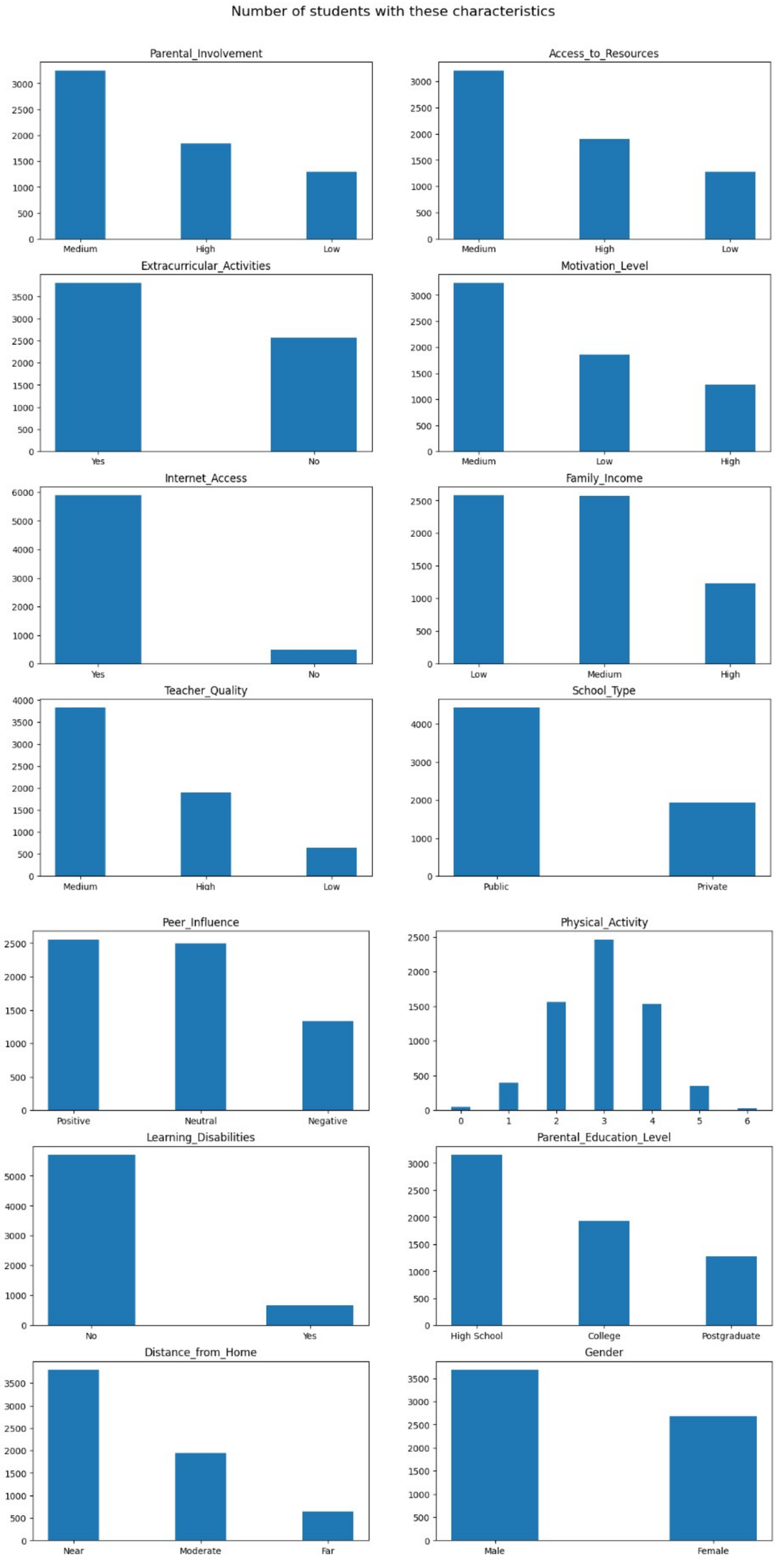
Απο τα παραπάνω διαγράμματα μπορούμε να αντλήσουμε χρήσιμες πληροφορίες οι οποίες θα μας φανούν σημαντικές στην πορεία.

Αντίστοιχα για τις κατηγορικές μεταβλητές υπολογίζουμε τον αριθμό των κατηγοριών, την τιμή η οποία εμφανίζεται τις περισσότερες φορές όπως και το πλήθος της συγκεκριμένης τιμής στην μεταβλητή. Στη συνέχεια κατασκευάζουμε τον πίνακα Cramers' V , τα ραυδογράμματα των κατηγορικών μεταβλητών όπως και ο πίνακας ANOVA μεταξύ των ποσοτικών και κατηγορικών μεταβλητών.



Ο πίνακας Cramer's V χρησιμοποιείται για να μετρήσει τη συσχέτιση μεταξύ κατηγορικών μεταβλητών.


Hours_Studied
Attendance
Sleep_Hours
Previous_Scores
Tutoring_Sessions
Exam_Score
0.49 0.52 0.5 0.28 0.61 0.62 0.86 0.93 0.66 0.58 0.27 0.59 0.39 0.57
0.52 0.6 0.89 0.66 0.12 0.53 0.81 0.094 0.07 0.55 0.095 0.11 0.26 0.71
0.94 0.32 0.81 0.96 0.34 0.33 0.71 0.96 0.24 0.52 0.26 0.58 0.78 0.43
0.19 0.11 0.83 0.22 0.8 0.28 0.03 0.33 0.12 0.15 0.66 0.2 0.55 1
0.67 0.45 0.61 0.63 0.45 0.93 1 0.66 0.94 0.49 0.59 0.93 0.43 0.66
3.1e-35 2.7e-40 4.6e-07 9e-12 4.4e-05 3.6e-13 8.8e-09 0.39 1.2e-14 0.29 1.9e-11 3e-16 1.6e-11 0.69
Parental_Involvement
Access_to_Resources
Extracurricular_Activities
Motivation_Level
Internet_Access
Family_Income
Teacher_Quality
School_Type
Peer_Influence
Physical_Activity
Learning_Disabilities
Parental_Education_Level
Distance_from_Home
Gender
0.8
0.6
0.4
0.2
0.0


Number of students with these characteristics
Parental_Involvement
Medium
High
Low
Access_to_Resources
Medium
High
Low
Extracurricular_Activities
Yes
No
Motivation_Level
Medium
Low
High
Internet_Access
Yes
No
Family_Income
Low
Medium
High
Teacher_Quality
Medium
High
Low
School_Type
Public
Private
Peer_Influence
Positive
Neutral
Negative
Physical_Activity
Learning_Disabilities
No
Yes
Parental_Education_Level
High School
College
Postgraduate
Distance_from_Home
Near
Moderate
Far
Gender
Male
Female

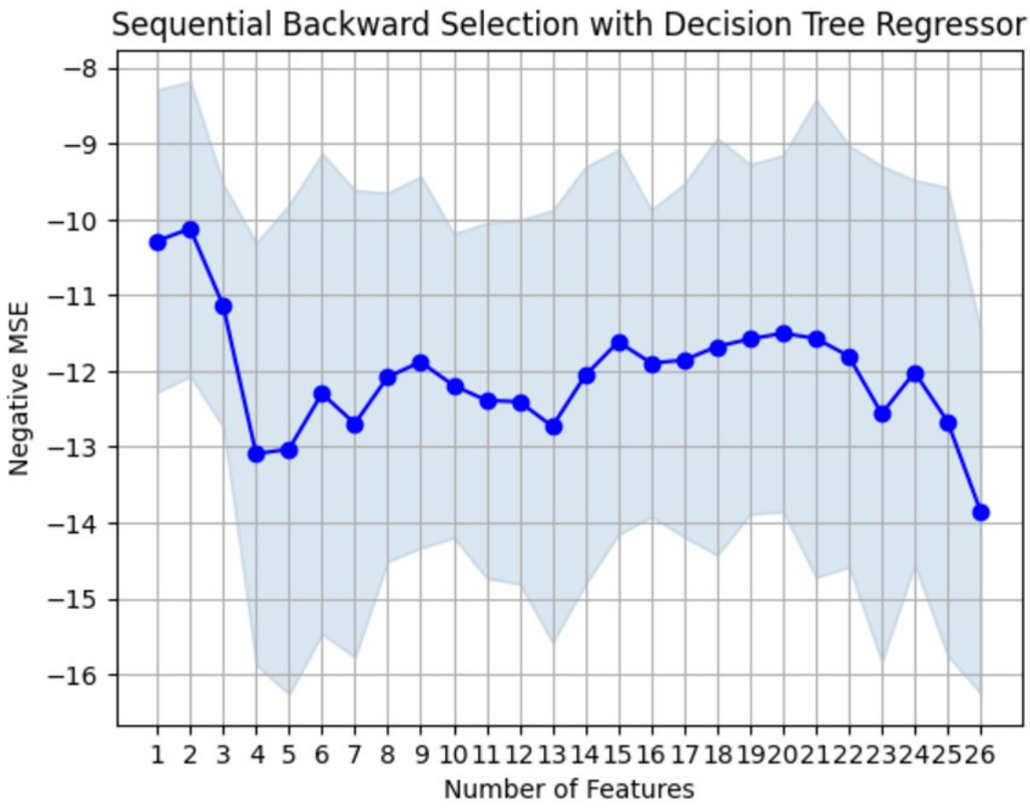
3.3 Επιλογή χαρακτηριστικών

Η επιλογή των χαρακτηριστικών που θα χρησιμοποιήσουμε για την πρόβλεψη του στόχου αποτελέι σημαντικό βήμα. Απο την μια προσπαθούμε να επιλέξουμε όλα τα σημαντικά χαρακτηριστικά έτσι ώστε να αποφύγουμε την υποπροσαρμογή, δηλαδή το μοντέλο μας να μην έχει τις απαραίτητες πληροφορίες για να αποτυπώσει την σχέση μεταξύ των δεδομένων και της μεταβλητής στόχου και απο την άλλη προσπαθούμε να επιλέξουμε όσο το δυνατό λιγότερα μη σημαντικά χαρακτηριστικά έτσι ώστε να αποφύγουμε την υπερπροσαρμογή, δηλαδή το μοντέλο μας να αποδίδει πολύ καλά στα δεδομένα εκπαίδευσης αλλά να αδυνατεί να γενικεύσει σε νέα δεδομένα.Αρχικά θα ορίσουμε 8 στο σύνολο μοντέλα παλινδρόμισης για τα οποία θα εξετάσουμε την απόδοση τους μέσω μιας μετρικής για κάθε συνδυασμό χαρακτηριστικών. Τα μοντέλα παλινδρόμισης τα οποία θα εξετάσουμε είναι τα εξής:

1. Πολυωνιμικό μοντέλο παλινδρόμισης
2. Γραμμική παλινδρόμιση
3. Decision Tree παλινδρόμιση
4. Random Forest παλινδρόμιση
5. Support Vector παλινδρόμιση
6. Gradient Boosting παλινδρόμιση
7. Παλινδρόμιση Ridge
8. Παλινδρόμιση Lasso

Για την πραγματοποίηση της παραπάνω διαδικασίας θα χρησιμοποιήσουμε τον αλγόριθμο Sequential feature selection (SFS). Θα μπορούσαμε να κάνουμε την διαδικασία αυτή και μέσω των πινάκων συσχέτισης Pearson, Cramer's V και ANOVA διαγράφοντας για κάθε ζεύγος μεταβλητών με ισχυρή συσχέτιση την μια απο τις δύο μεταβλητές αλλά με την διαφορά πως δεν θα είχαμε τόσο καλά αποτελέσματα όσο με την μέθοδο SFS. Η μέθοδος SFS με επιλογή για backwards selection ξεκινά με το πλήρες

σύνολο χαρακτηριστικών. Το σύνολο εκπαίδευσης χωρίζεται σε k=cross-validation ίσα μέρη(folds). Σε κάθε επανάληψη, αφαιρείται ένα χαρακτηριστικό από το τρέχον σύνολο, και το μοντέλο εκπαιδεύεται χρησιμοποιώντας τα εναπομείναντα χαρακτηριστικά. Το μοντέλο εκπαιδεύεται k φορές, χρησιμοποιώντας κάθε φορά ένα διαφορετικό fold ως σύνολο δοκιμής (test set), ενώ τα υπόλοιπα k-1 folds χρησιμοποιούνται ως σύνολο εκπαίδευσης(tra set). Η απόδοση για κάθε αφαίρεση χαρακτηριστικού αξιολογείται με βάση τον μέσο όρο της απόδοσης για κάθε επιλογή fold ώς test set χρησιμοποιώντα κάποιο προκαθορισμένο κριτήριο, όπως το μέσο τετραγωνικό σφάλμα (MSE). Το χαρακτηριστικό που οδηγεί στη μικρότερη μείωση της απόδοσης αφαιρείται οριστικά από το σύνολο. Η διαδικασία συνεχίζεται μέχρι να επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών. Για κάθε ένα απο τα μοντέλα παλινδρόμισης που αναφέραμε παραπάνω εντοπίσαμε τα χαρακτηριστικά τα οποία οδηγούν στο μικρότερο μέσο τετραγωνικό σφάλμα. Για την καλύτερη κατανόηση της μεθόδου, παρουσιάζεται παρακάτω η γραφική παράσταση του βέλτιστου αρνητικού μέσου τετραγωνικού σφάλματος. Η γραφική αφορά το μοντέλο παλινδρόμησης Decision Tree και βασίζεται στα χαρακτηριστικά που ελαχιστοποιούν το μέσο τετραγωνικό σφάλμα. Συγκεκριμένα, η γραφική απεικονίζει τη σχέση μεταξύ του αριθμού των χαρακτηριστικών και του μέσου τετραγωνικού σφάλματος, όπως αυτά επιλέχθηκαν μέσω της συγκεκριμένης μεθόδου.




```
Selected features: (0, 5)
MSE: 10.120028860319591
```

Μπορούμε εύκολα να παρατηρήσουμε πως το το μεγαλύτερο αρνητικό μέσο τετραγωνικό σφάλμα επιτυγχάνετε για αριθμό χαρακτηριστικών ίσο με 2. Το αντίστοιχο καλύτερο μέσο τετραγωνικό σφάλμα είναι ίσο με 10.12 και επιτυγχάνετε χρησιμοποιώντας την πρώτη και έκτη στήλη του πίνακα δεδομένων.

3.4 Επιλογή υπερπαραμέτρων

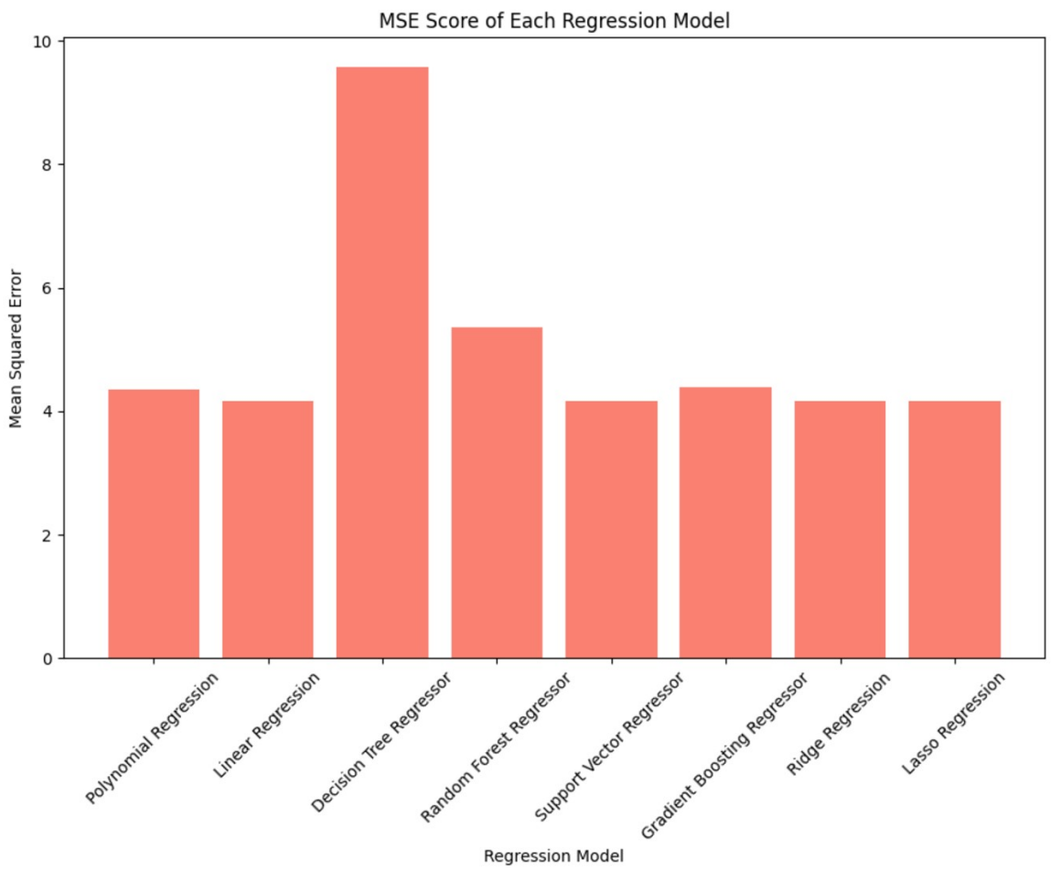
Για κάθε ένα από τα μοντέλα παλινδρόμησης που ορίσαμε παραπάνω, πραγματοποιήσαμε βελτιστοποίηση υπερπαραμέτρων (hyperparameter tuning) χρησιμοποιώντας τη μέθοδο GridSearchCV. Η βελτιστοποίηση έγινε για τα αντίστοιχα χαρακτηριστικά που επιλέχθηκαν ως βέλτιστα μέσω της μεθόδου Sequential Forward Selection (SFS). Στη διαδικασία, θέσαμε cross-validation = 10 και επιλέξαμε ως μετρική απόδοσης το μέσο τετραγωνικό σφάλμα(MSE). Η διαδικασία βελτιστοποίησης πραγματοποιήθ σε δύο στάδια. Αρχικά, εξετάσαμε ένα εύρος παραμέτρων γύρω από τις προεπιλεγμένες τιμές (default parameters), ώστε να αποκτήσουμε μια αρχική εικόνα για τη συμπεριφορά του μοντέλου. Στη συνέχεια, εστιάσαμε στις παραμέτρους που ανέδειξε η πρώτη διαδικασία ως πιο σημαντικές, προσαρμόζοντας το εύρος των τιμών τους για πιο λεπτομερή αναζήτηση.

Με αυτόν τον τρόπο καταφέραμε να βελτιώσουμε περαιτέρω την απόδοση των μοντέλων μας, εξασφαλίζοντας καλύτερα αποτελέσματα.

Η μέθοδος GridSearchCV λειτουργεί ώς ακολούθως. Το σύνολο εκπαίδευσης χωρίζεται σε k=cross-validation ίσα μέρη(folds). Στη συνέχεια, για κάθε συνδυασμό υπερπαραμέτρων που ορίζει ο χρήστης το μοντέλο εκπαιδεύεται k φορές, χρησιμοποιώντας κάθε φορά ένα διαφορετικό fold ως σύνολο δοκιμής (test set), ενώ τα υπόλοιπα k-1 folds χρησιμοποιούνται ως σύνολο εκπαίδευσης(training set). Η απόδοση κάθε συνδυασμού υπερπαραμέτρων αξιολογείται με βάση τον μέσο όρο της απόδοσης για κάθε επιλογή fold ώς test set. Τέλος, επιλέγεται ο συνδυασμός υπερπαραμέτρων που βελτιστοποιεί τη μετρική απόδοσης.

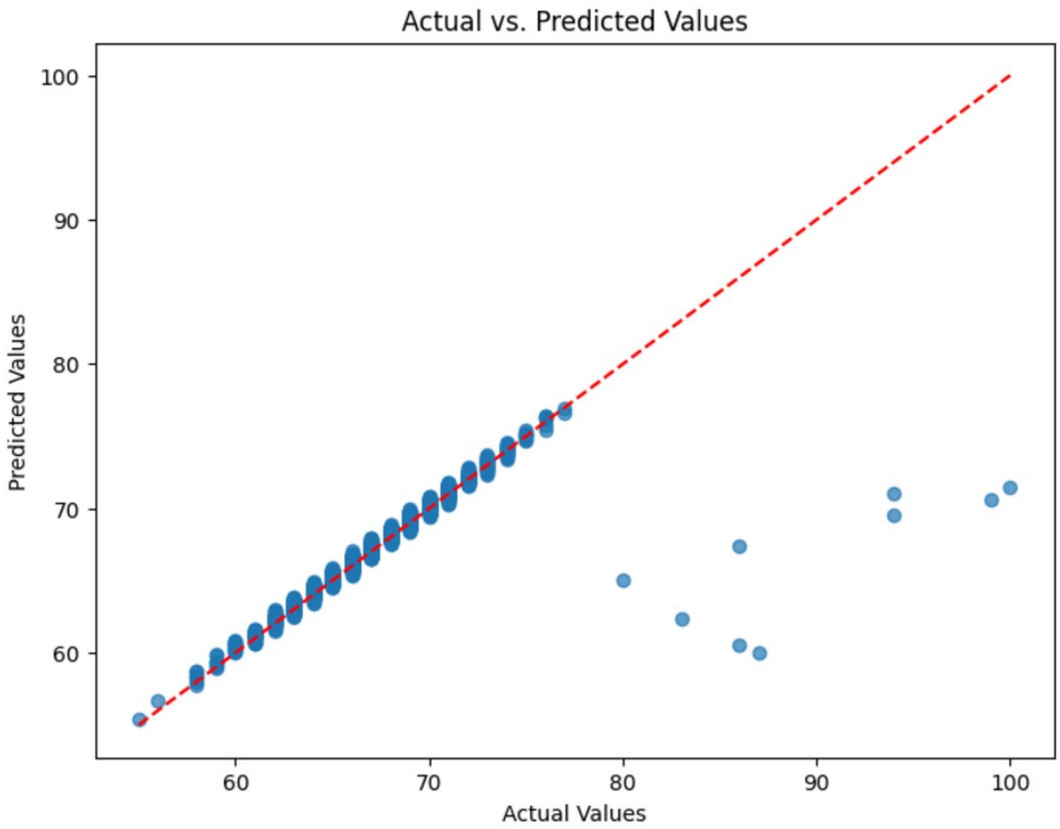
3.5 Επιλογή μοντέλου

Για να βρούμε το τελικό βέλτιστο μοντέλο παλινδρόμισης απο αυτά που εξετάσαμε κατασκευάσαμε το ραυδόγραμμα του μέσου τετραγωνικού σφάλματος για κάθε ένα απο τα μοντέλα παλινδρόμισης που αναλύσαμε. Το ραυδόγραμμα αυτό παρουσιάζεται παρακάτω:



Το μοντέλο το οποίο περιγράφει καλύτερα τα δεδομένα μας και προβλέπει σε καλύτερο βαθμό τον βαθμό του μαθητή με βάση την μετρική απόδοσης

MSE είναι το μοντέλο παλινδρόμισης Lasso με μέσο τετραγωνικό σφάλμα ίσο με 4.155, επομένως έχουμε RMSE=$\sqrt{MSE} = 2.038$ και αυτό σημαίνει πως κατά μέσο όρο η πρόβλεψη μας για κάθε βαθμό εξέτασης διαφέρει κατά 2.038 μονάδες απο την πραγματική τιμή. Με βάση την ανάλυση που κάναμε για τις υπερπαραμέτρους το μοντέλο αυτό έχει παραμέτρους ίσες με ρυθμό κανονικοποίησης(alpha) ίσο με 0.0001, μέγιστο αριθμό επαναλήψεων(max_iter) ίσο με 500 και ανοχή σύγκλισης(tol) ίση με 0.1.Για να εξετάσουμε περαιτέρω την απόδοση του μοντέλου μας υπολογίσαμε τον συντελεστή προσδιορισμού R^2, ο οποίος ισούται με 0.733 επομένως το μοντέλο μας εξηγά το 73.3% της μεταβλητότητας του Exam Score. Τέλος, κατασκευάσαμε το γράφημα που φαίνετε παρακάτω όπου παρουσιάζει τις προβλεπόμενες τιμές του test set σε σχέση με τις αντίστοιχες πραγματικές τιμές.



Μπορούμε να παρατηρήσουμε πως το μοντέλο μας με βάση την γραφική υπολογίζει σε πολύ ικανοποιητικό βαθμό τον βαθμό του μαθητή για τιμές μικρότερες του 80. Για τιμές μεγαλύτερες απο 80 δεν μπορούμε να πούμε το ίδιο. Αυτό μπορεί να οφείλετε στο ότι δεν υπάρχουν αρκετοί μαθητές οι οποίοι πήραν βαθμό μεγαλύτερο απο 80 στο training set, επομένως το μοντέλο μηχανικής μάθησης που δημιουργήσαμε δεν έχει αρκετά δεδομένα για να προβλέψει ικανοποιητικά τον τελικό βαθμό του μαθητή. Επίσης

λόγω της μεγάλης διαφοράς στο πλήθος των δεδομένων στο training set για βαθμούς μικρότερους απο 80 σε σχέση με μεγαλύτερους απο αυτό, μπορεί να οδήγησε το μοντέλο μας σε υπερβολική εστίαση στα δεδομένα μικρότερα απο 80(biased ώς πρός αυτά τα δεδομένα) αφού τελικά θα οδηγούσε σε μικρότερο ολικό μέσο τετραγωνικό σφάλμα σε αντίθεση με το να το έκανε γενίκευση(generalize) και για τιμές μεγαλύτερες απο 80.

4 Το μοντέλο παλινδρόμισης Lasso

4.1 Έγκριση ειδικού πλαισίου χωροταξικού σχεδιασμού

Αναλύεται η διαδικασία έγκρισης του πλαισίου.

4.2 Καθορισμός περιοχών αιολικής προτεραιότητας

Αναλύονται οι παράγοντες επιλογής περιοχών.

4.3 Παράμετροι που καθορίζουν τις αποστάσεις

Παρουσιάζονται τα κριτήρια για τις αποστάσεις υποδομών.

5 Σύγκριση με έρευνες

hb

6 Βιβλιογραφία