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

Υπεύθυνος 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

km

Introduction

gt

2 Ανάλυση της διαδικασίας επιλογής μοντέλου

2.1 Ανάλυση του προβλήματος

Το πρόβλημα το οποίο αντιμετωπίσαμε είχει να κάνει με την δημιουργία ενός μοντέλου μηχανικής μάθησης για την εκτίμηση των βαθμών που θα επιτύχουν οι μαθητές στην εξέταση ενός μαθήματος , 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. Χρησιμοποιήσαμε 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 που δέχεται ο μαθητής 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(μεταβλητή στόχο) του μοντέλου ενώ οι υπόλοιπες μεταβλητές αποτελούν τις πιθανές ανεξάρτητες μεταβλητές.

2.2 Επεξεργασία των δεδομένων

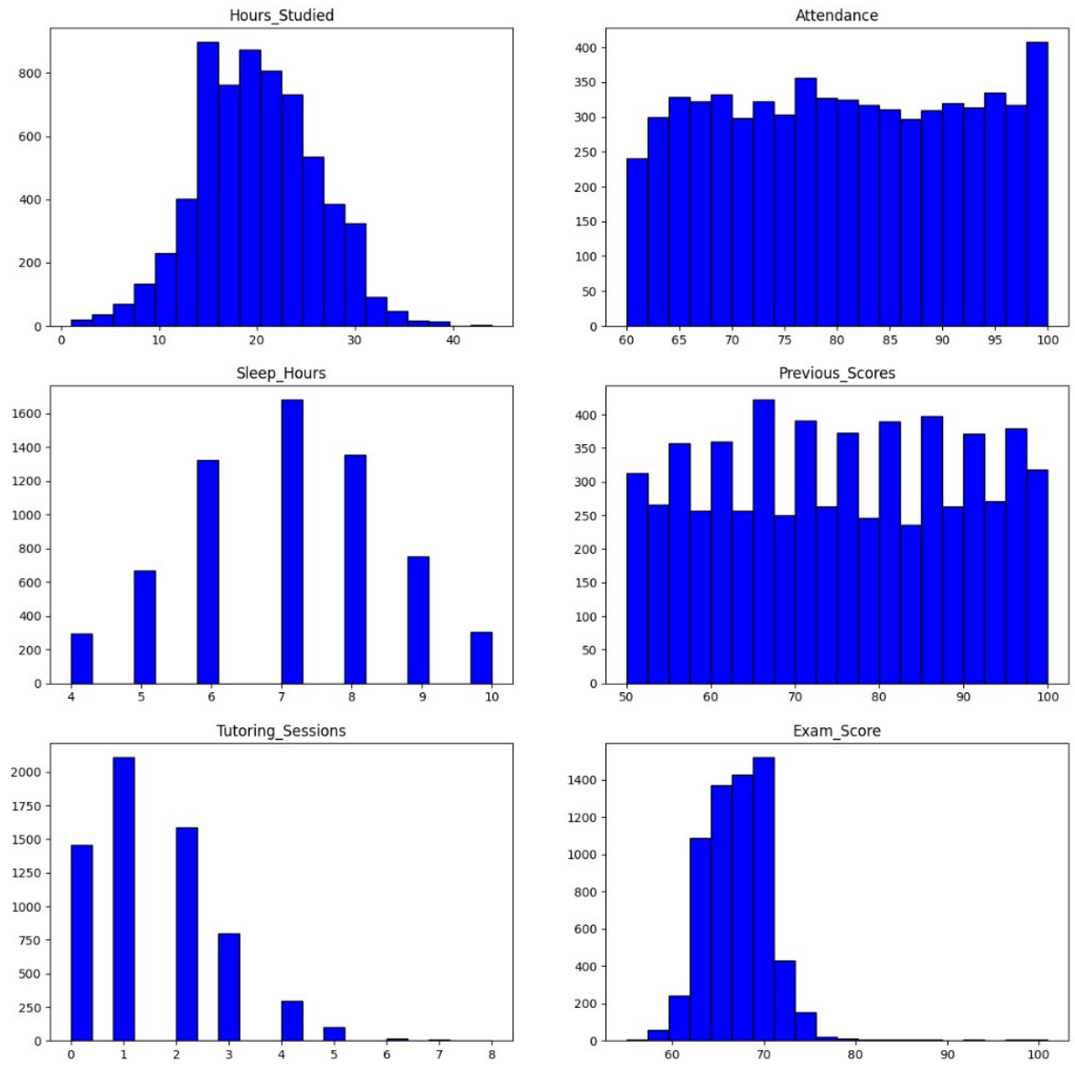
Ρίχνωντας μια ματιά στα δεδομένα παρατηρήσαμε πως υπάρχουν παρατηρήσεις με ελλειπείς τιμές. Στην συγκεκριμένη περίπτωση έχουμε 3 επιλογές. Η πρώτη επιλογή ήταν να διαγράψουμε τις παρατηρήσεις αυτές, η δεύτερη να προβλέψουμε τις τιμές των παρατηρήσεων με κάποια μέθοδο (Παλινδρόμιση για τις ποσοτικές και 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 περιλαμβάνει αριθμητικές τιμές(0-6), αυτές οι τιμές αποτελούν κατηγορίες. Για αυτό και εμείς αλλάξαμε τον τύπο της μεταβλητής σε αντικείμενο. Εύκολα μπορούμε να δούμε πως το dataset μας έχει 14 κατηγορικές μεταβλητές και 6 ποσοτικές.

Παρατηρώντας τα δεδομένα μπορούμε εύκολα να δούμε πως οι ποσοτικές μεταβλητές κυμάινονται σε διαφορετικές κλίμακες. Όταν χρησιμοποιούμε αλγόριθμους μηχανικής μάθησης που βασίζονται σε αποστάσεις (όπως οι K-Neighbors Classifier, K-means, linear regression, polynomial regression κ.λπ.), είναι απαραίτητο να εφαρμόζουμε κάποια μορφή κλιμάκωσης στα δεδομένα μας. Η κλιμάκωση εξασφαλίζει ότι τα χαρακτηριστικά με μεγάλο εύρος τιμών δεν θα κυριαρχούν έναντι εκείνων με μικρό εύρος, αποτρέποντας έτσι την εισαγωγή σφαλμάτων στο μοντέλο. Ιδανικά,θέλουμε όλα τα αριθμητικά χαρακτηριστικά να έχουν την ίδια κλίμακα, έτσι ώστε να συνεισφέρουν ισότιμα στη διαδικασία εκπαίδευσης του μοντέλου. Η επιλογή της κατάλληλης κλιμάκωσης στα δεδομένα εξαρτάται από τα δομή του συνόλου δεδομένων. Για χαρακτηριστικά τα οποία προσεγγίζουν την κανονική κατανομή χρησιμοποιούμε Standard Scaler, για χαρακτηριστικά τα οποία περιέχουν ακραίες τιμές(outliers) χρησιμοποιούμε Robust Scaler και για χαρακτηριστικά τα οποία προσεγγίζουν την ομοιόμορφη κατανομή τον MinMax Scaler.

Μπορούμε να εφαρμόσουμε κλιμάκωση ακόμα και αν οι ποσοτικές τιμές αποτελούν ακέραιες τιμές. Οι μέθοδοι κλιμάκωσης, όπως StandardScaler, MinMaxScaler και RobustScaler, λειτουργούν ανεξάρτητα από τον τύπο δεδομένων, εφόσον οι μεταβλητές είναι αριθμητικές και για κάθε τύπο κλιμάκωσης ισχύουν οι αντίστοιχες προυποθέσεις.

Παρακάτω δίνονται τα ιστογράμματα όλων των ποσοτικών μεταβλητών.



Εύκολα μπορούμε να παρατηρήσουμε πως εφαρμόζουμε Standard Scaler στα χαρακτηριστικά Hours Studied και Sleep Hours αφού όπως μπορούμε να δούμε ακολουθούν κατά προσέγγιση κανονική κατανομή, MinMax Scaler στα χαρακτηριστικά Attendance και Previous Scores αφού προσεγγίζουν την ομοιόμορφη κατανομή και Robust Scaler στο χαρακτηριστικό Tutoring Sessions αφού όπως μπορούμε να δούμε περιέχει ακραίες τιμές. Η μεταβλητή στόχος Exam Scores δεν χρειάζεται να κλιμακωθεί αφού αποτελεί την εξαρτημένη μεταβλητή του μοντέλου μας.

Οι σχέσεις για κάθε τύπο κλιμάκωσης δίνονται παρακάτω:

Μέθοδοι Κλιμάκωσης

StandardScaler

Η μετατροπή με τη μέθοδο StandardScaler ορίζεται ως:

$$X' = \frac{X - \mu}{\sigma}$$

όπου:

- X: Η τιμή της παρατήρησης στο χαρακτηριστικό.
- μ: Ο μέσος όρος του χαρακτηριστικού.
- σ: Η τυπική απόκλιση του χαρακτηριστικού.

Ο Standard Scaler μετασχηματίζει τα δεδομένα έτσι ώστε να έχουν μέση τιμή ίση με 0 και τυπική απόκλιση ίση με 1.

MinMaxScaler

Η μετατροπή με τη μέθοδο MinMaxScaler ορίζεται ως:

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

όπου:

- X: Η τιμή της παρατήρησης στο χαρακτηριστικό.
- X_{min}: Η ελάχιστη τιμή του χαρακτηριστικού.
- X_{max}: Η μέγιστη τιμή του χαρακτηριστικού.

Ο MinMax Scaler κλιμακώνει τα δεδομένα σε μια καθορισμένη περιοχή, συνήθως [0,1].

RobustScaler

Η μετατροπή με τη μέθοδο RobustScaler ορίζεται ως:

$$X' = \frac{X - Q_2}{Q_3 - Q_1}$$

όπου:

- X: Η τιμή της παρατήρησης στο χαρακτηριστικό.
- Q_2: Η διάμεσος (median) του χαρακτηριστικού.
- Q_1: Το πρώτο τεταρτημόριο (25η εκατοστιαία θέση).
- Q_3: Το τρίτο τεταρτημόριο (75η εκατοστιαία θέση).

Ο Robust scaler μετασχηματίζει τα δεδομένα έτσι ώστε να έχουν κέντρο την διάμεσο η οποία γίνεται ίση με 0 και να μειώσει την επίδραση των ακραίων τιμών.

Γνωρίζουμε πως τα μοντέλα μηχανικής μάθησης δεν μπορούν να διαχειριστούν μη αριθμητικά δεδομένα. Επομένως, μέσω των encoders θα τα μετατρέψουμε με κατάλληλο τρόπο σε αριθμητικά χωρίς όμως να αλλάζουμε την κατηγορική τους ιδιότητα. Θα χρησιμοποιήσουμε Ordinal encoder για τα χαρακτηριστικά τα οποία περιέχουν κατηγορίες, οι οποίες έχουν κάποια φυσική σειρά(π.χ. ”Low”,”Medium”,”High”) και OneHot encoder για τα χαρακτηριστικά τα οποία περιέχουν κατηγορίες, οι οποίες δεν έχουν κάποια φυσική σειρά(π.χ. ”Yes”,”No”). Επομένως τα κατηγορικά χαρακτηριστικά ”Parental Involvement”,”Access to Resources”,”Motivation Level”,”Family Income”, ”Teacher Quality”,”Parental Education Level”, ”Distance from Home” και ”Physical Activity” κωδικοποιούνται με ordinal encoding ενώ τα υπόλοιπα κατηγορικά χαρακτηριστικά με Onehot.

Η μέθοδος orindal encoding κωδικοποιεί με αριθμητικές τιμές κάθε κατηγορία με βάση την σειρά ή την θέση τους ξεκινώντας απο το 0. Για παράδειγμα εάν έχουμε τις κατηγορίες ”Low”,”Medium”,”High” θα τους αναθέσει τις τιμές 0,1,2 αντίστοιχα.

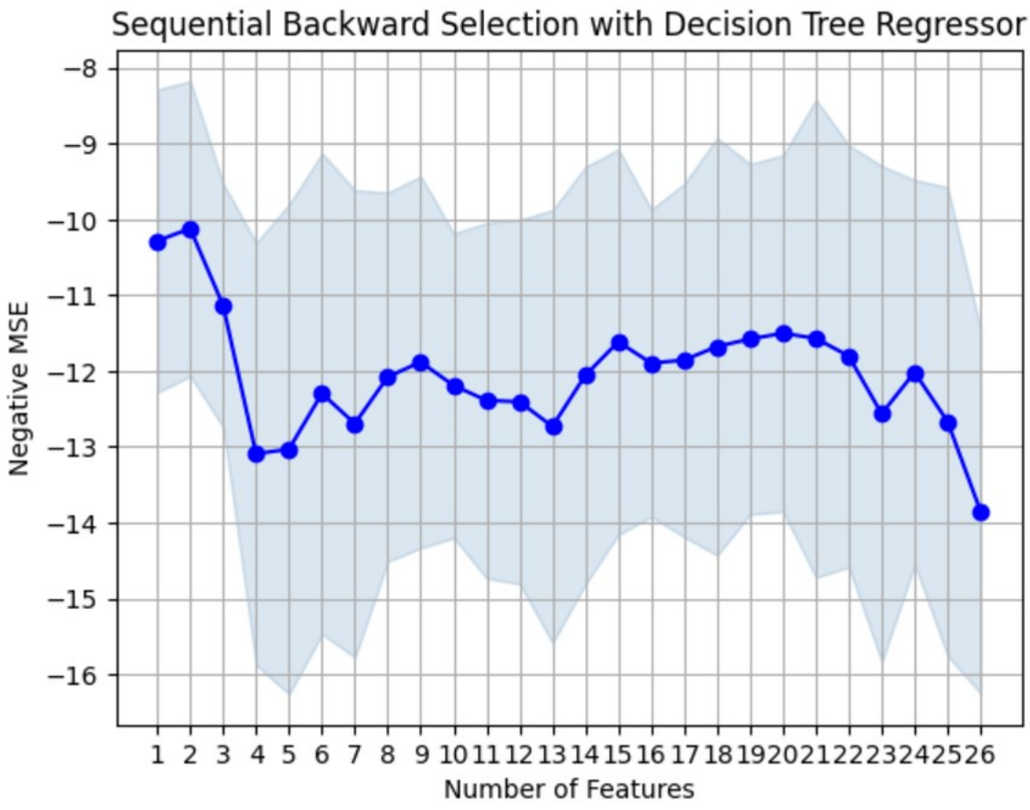
Η μέθοδος onehot encoding διαγράφει την κατηγορική στήλη και δημιουργεί για κάθε κατηγορία μια νέα που για κάθε παρατήρηση περιέχει την τιμή 1 αν το δείγμα ανήκει σε αυτήν την κατηγορία και 0 διαφορετικά.

2.3 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 (training set). Η απόδοση για κάθε αφαίρεση χαρακτηριστικού αξιολογείται με βάση τον μέσο όρο της απόδοσης για κάθε επιλογή fold ώς test set χρησιμοποιώντας 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 παλινδρόμισης Decision Tree και βασίζεται στα χαρακτηριστικά που ελαχιστοποιούν το μέσο τετραγωνικό σφάλμα. Συγκεκριμένα, η γραφική απεικονίζει τη σχέση μεταξύ των χαρακτηριστικών όπως αυτά επιλέχθηκαν μέσω της συγκεκριμένης μεθόδου και του αρνητικού μέσου τετραγωνικού σφάλματος.




```
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.

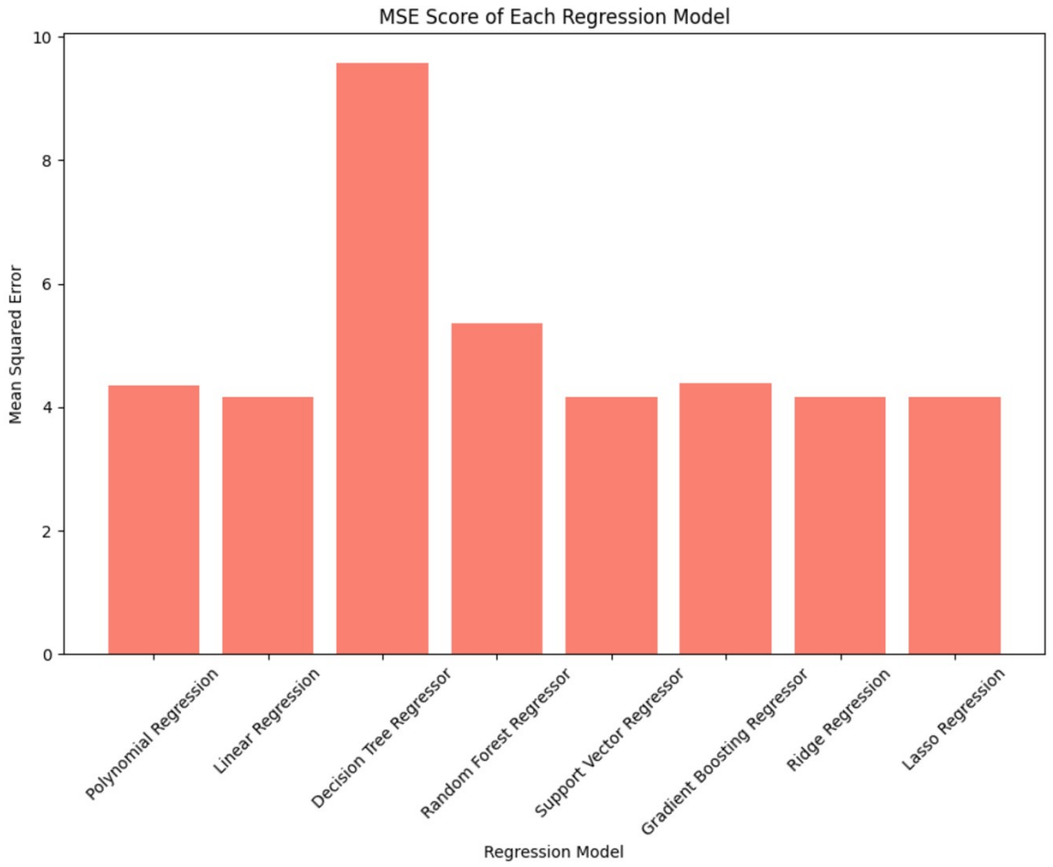
2.4 Επιλογή υπερπαραμέτρων

Για κάθε ένα από τα μοντέλα παλινδρόμησης που ορίσαμε παραπάνω, πραγματοποιήσαμε βελτιστοποίηση υπερπαραμέτρων (hyperparameter tuning) χρησιμοποιώντας τη μέθοδο GridSearchCV. Η βελτιστοποίηση έγινε για κάθε μοντέλο παλινδρόμισης χρησιμοποιώντας 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 πραγματοποιήθηκε 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.

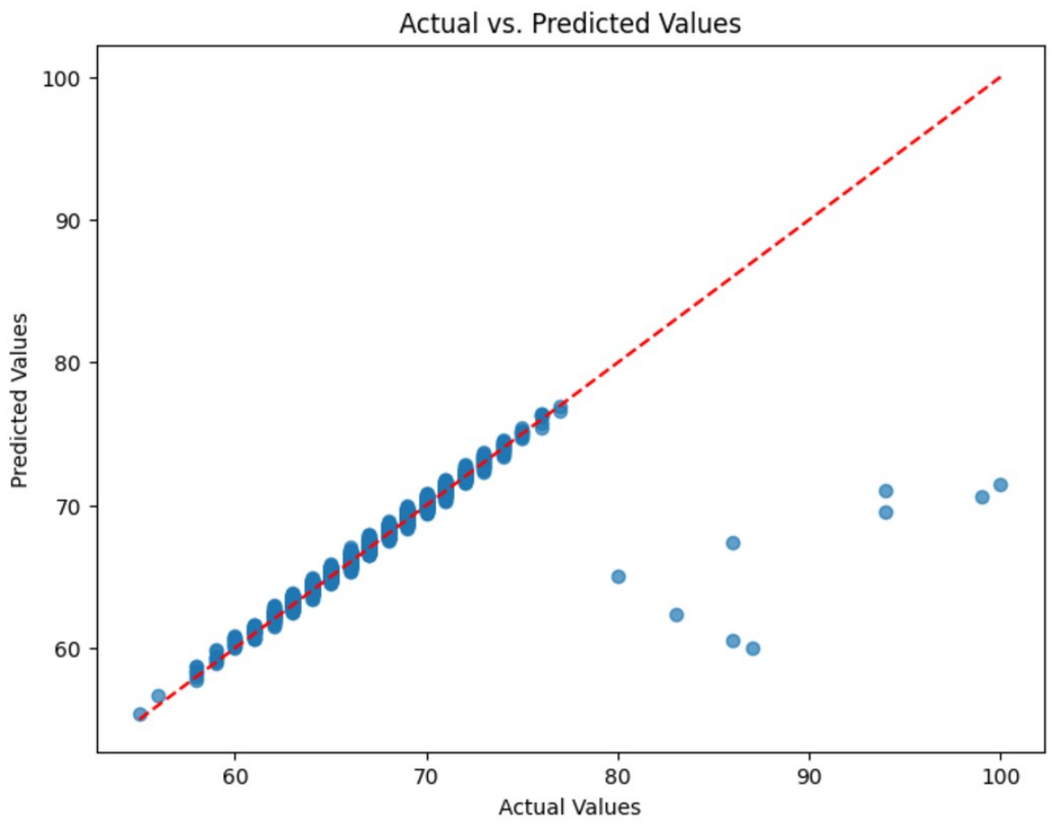
Η μέθοδος GridSearchCV λειτουργεί ώς ακολούθως. Το σύνολο εκπαίδευσης χωρίζεται σε k=cross-validation ίσα μέρη(folds). Στη συνέχεια, για κάθε συνδυασμό υπερπαραμέτρων που ορίζει ο χρήστης το μοντέλο εκπαιδεύεται k φορές, χρησιμοποιώντας κάθε φορά ένα διαφορετικό fold ως σύνολο δοκιμής (test set), ενώ 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.

2.5 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 βελτιστοποιήσαμε. 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 θέσαμε ρυθμό κανονικοποίησης (α) 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 βαθμού του μαθητή(Exam Score). Τέλος, κατασκευάσαμε το γράφημα που φαίνετε παρακάτω όπου παρουσιάζει τις προβλεπόμενες τιμές του συνόλου δοκιμής σε σχέση με τις αντίστοιχες πραγματικές τιμές.



Μπορούμε να παρατηρήσουμε πως το μοντέλο μας με βάση την παραπάνω γραφική παράσταση υπολογίζει σε αρκετά ικανοποιητικό βαθμό τον βαθμό του μαθητή για τιμές μικρότερες απο 80. Για τιμές μεγαλύτερες απο 80 δεν μπορούμε να πούμε το ίδιο. Αυτό οφείλετε στο ότι δεν υπάρχουν αρκετοί μαθητές οι οποίοι πήραν βαθμό μεγαλύτερο απο 80 στο training set (35 σε αριθμό), 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 μέσο τετραγωνικό σφάλμα σε αντίθεση με το να το έκανε γενίκευση(generalize) και για τιμές μεγαλύτερες απο 80.

3 Το μοντέλο παλινδρόμισης Lasso

3.1 Έγκριση ειδικού πλαισίου χωροταξικού σχεδιασμού

Αναλύεται η διαδικασία έγκρισης του πλαισίου.

3.2 Καθορισμός περιοχών αιολικής προτεραιότητας

Αναλύονται οι παράγοντες επιλογής περιοχών.

3.3 Παράμετροι που καθορίζουν τις αποστάσεις

Παρουσιάζονται τα κριτήρια για τις αποστάσεις υποδομών.

4 Σύγκριση με έρευνες

4.1 Επεξήγηση των παραμέτρων

Καταρχάς για να μπορούμε να κάνουμε την σύγκριση μεταξύ συγκεκριμένων ερευνών που υπάρχουν διαθέσιμες στο ευρύκοινόκαιτωναποτελεσμάτωνπουπροέκυψανμέσωτηςδικήςμαςανάλυσης,πρέπει πρώτα να κάνουμε μια ανάλυση της σχέσης μεταξύ του τελικού βαθμού του μαθητή και των διάφορων ανεξάρτητων μεταβλητών.

Μέσω του μοντέλου μηχανικής μάθησης που δημιουργήσαμε πήραμε τα παρακάτω αποτελέσματα:

Feature	**Coefficient**
Attendance	7.945534e+00
Previous_Scores	2.432289e+00
Tutoring_Sessions	4.838539e-01
Hours_Studied	1.769307e+00
Parental_Involvement	1.006635e+00
Access_to_Resources	1.021990e+00
Motivation_Level	5.513339e-01
Family_Income	5.666131e-01
Teacher_Quality	5.520568e-01
Parental_Education_Level	4.946623e-01
Distance_from_Home	-4.593309e-01
Physical_Activity	1.992808e-01
Extracurricular_Activities_No	-5.626076e-01
Extracurricular_Activities_Yes	1.337722e-16
Internet_Access_No	-8.974091e-01
Peer_Influence_Negative	-8.970209e-01
Peer_Influence_Neutral	-3.381165e-01
Peer_Influence_Positive	1.176801e-01
Learning_Disabilities_No	8.672434e-01
Intercept	**57.07844535052128**

Εύκολα μπορούμε να συμπεράνουμε πως:

1. **Parental Involvement** Ο βαθμός εμπλοκής των γονέων στην εκπαίδευση του μαθητή αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: ’Low’, ’Medium’ και ’High’. Για κάθε μεταβολή κατά ένα επίπεδο, από ’Low’ σε ’Medium’ ή από ’Medium’ σε ’High’, ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά μία περίπου μονάδα.

2. **Access to Resources** Η πρόσβαση του μαθητή σε εκπαιδευτικό υλικό αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: ’Low’,’Medium’ και ’High’. Για κάθε μεταβολή κατά ένα επίπεδο, από ’Low’ σε ’Medium’ ή από ’Medium’ σε ’High’, ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 1.02 μονάδες.

3. **Motivation Level** Το κίνητρο του μαθητή συμβάλλει αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: ’Low’, ’Medium’ και ’High’. Για κάθε μεταβολή κατά ένα επίπεδο, από ’Low’ σε ’Medium’ ή από ’Medium’ σε ’High’, ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.55 μονάδες.

4. **FamilyIncome** Το οικογενειακό εισόδημα του μαθητή αναμένετε νασυμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: ’Low’, ’Medium’ και ’High’. Για κάθε μεταβολή κατά ένα επίπεδο, από ’Low’ σε ’Medium’ ή από ’Medium’ σε ’High’, ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.57 μονάδες.

5. **Teacher Quality** Η εκπαιδευτική ικανότητα του καθηγητή του μαθητή αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: 'Low','Medium' και 'High'. Για κάθε μεταβολή κατά ένα επίπεδο, από 'Low' σε 'Medium' ή από 'Medium' σε 'High', ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.55 μονάδες.

6. **Parental Education Level** Το επίπεδο εκπαίδευσης του γονέα του μαθητή αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε 3 επίπεδα: 'High School', 'College' και 'Postgraduate'. Για κάθε μεταβολή κατά ένα επίπεδο, από 'High School' σε 'Col lege' ή από 'College' σε 'Postgraduate', ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.49 μονάδες.

7. **Distance from Home** Η απόσταση του σχολείου απο το σπίτι του μαθητή αναμένετε να συμβάλλει αρνητικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε τρία επίπεδα: 'Near', 'Moderate' και 'Far'. Για κάθε μεταβολή κατά ένα επίπεδο, από 'Near' σε 'Moderate' ή από 'Moderate' σε 'Far', ο βαθμός του μαθητή αναμένετε να μειωθεί κατά περίπου 0.46 μονάδες.

8. **Physical Activity** Το επίπεδο φυσικής δραστηριότητας του μαθητή αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα κατηγοριοποιείται σε 7 επίπεδα: 'Sedentary', 'Light', 'Moderate', 'Active', 'Very Active', 'Highly Active' και 'Athlete'. Για κάθε μεταβολή κατά ένα επίπεδο, από 'Sedentary' σε 'Light', από 'Light' σε 'Moderate' κ.ο.κ. , ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.2 μονάδες.

9. **Peer Influence** Η μορφή της επιρρόης που δέχεται ο μαθητής απο ομήλικους του επιρεάζει και τον βαθμό του. Συγκεκριμένα η επιρροή αυτή χωρίζετε σε 3 κατηγορίες 'Negative', 'Neutral' και 'Positive'. Για μεταβολή αυτού απο 'Negative' σε 'Neutral', ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.56 μονάδες. Για μεταβολή αυτού απο 'Neutral' σε 'Positive', ο βαθμός του μαθητή αναμένετε να αυξηθεί κατά περίπου 0.46 μονάδες.

10. **Extracurricular Activities** Η συμμετοχή του μαθητή σε εξωσχολικές δραστηριότητες αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα εάν ο μαθητής συμμετέχει σε εξωσχολικές δραστηριότητες ο βαθμός του αναμένετε να αυξηθεί κατά περίπου 0.56 μονάδες.

11. **Internet Access** Η πρόσβαση του μαθητή στο διαδύκτιο αναμένετε να συμβάλλει θετικά στον βαθμό του μαθητή. Συγκεκριμένα εάν ο μαθητής έχει πρόσβαση στο διαδίκτυο ο βαθμός του αναμένετε να αυξηθεί κατά περίπου 0.90 μονάδες.

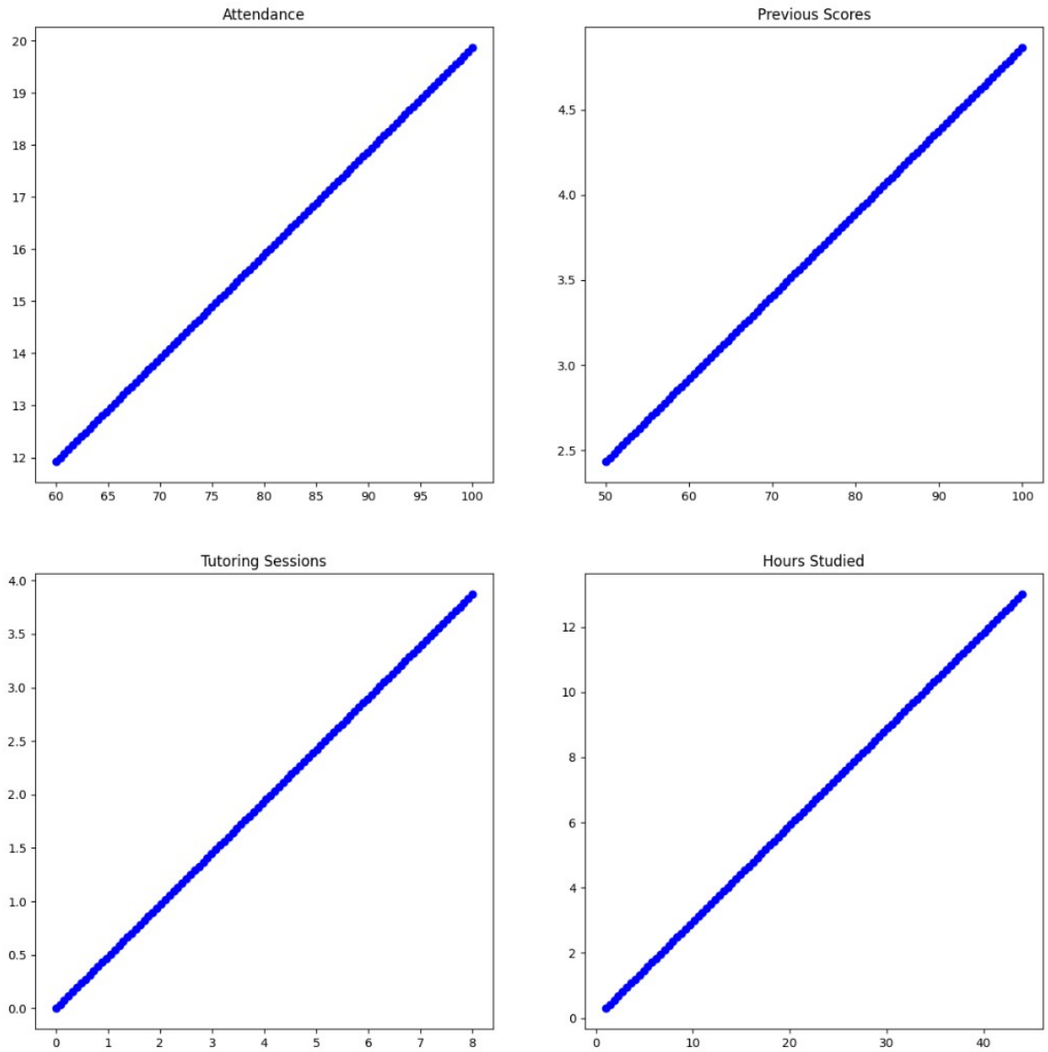
12. **Learning Disabilities** Η ύπαρξη μαθησιακών δυσκολιών στον μαθητή αναμένετε να συμβάλλει αρνητικά στον βαθμό του. Συγκεκριμένα εάν ο μαθητής έχει μαθησιακές δυσκολίες ο βαθμός του αναμένετε να μειωθεί κατά περίπου 0.87 μονάδες.

Η επεξήγηση των μεταβολών για κάθε χαρακτηριστικό το οποίο παίρνει αριθμητικές τιμές θέλει προσοχή. Αυτό οφείλετε στο ότι τα βάρη που υπολογίσαμε δείχνουν την μεταβολή του βαθμού του μαθητή όχι για τις μεταβλητές Attendance, Previous Scores , Tutoring Sessions και Hours Studied αλλά για τις κλιμακωμένες τιμές τους. Για να βρούμε την πραγματική τους σχέση θα κινηθούμε ώς ακολούθως:

- Εάν το χαρακτηριστικό κλιμακώθηκε μέσω Standard Scaler(Συγκεκριμένα η Hours Studied) ,τότε για κάθε μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού θα έχουμε μεταβολή του βαθμού του μαθητή κατά coeff/std(feature) μονάδες, όπου coeff η αντίστοιχη τιμή του coefficient του χαρακτηριστικού όπως φαίνετε στον πίνακα παραπάνω και std(feature) η τυπική του απόκλιση που προκύπτει απο τα δεδομένα.

- Εάν το χαρακτηριστικό κλιμακώθηκε μέσω MinMax Scaler(Συγκεκριμένα οι Attendance και Previous Scores) ,τότε για κάθε μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού θα έχουμε μεταβολή του βαθμού του μαθητή κατά coeff/(max(feature)-min(feature)) μονάδες, όπου coeff η αντίστοιχη τιμή του coefficient του χαρακτηριστικού όπως φαίνετε στον πίνακα παραπάνω, min(feature) η ελάχιστη τιμή του χαρακτηριστικού αυτού στα δεδομένα και max(feature) η μέγιστη τιμή του χαρακτηριστικού στα δεδομένα.
- Εάν το χαρακτηριστικό κλιμακώθηκε μέσω Robust Scaler(Συγκεκριμένα η Tutoring Sessions) ,τότε για κάθε μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού θα έχουμε μεταβολή του βαθμού του μαθητή κατά coeff/IQR(feature) μονάδες, όπου coeff η αντίστοιχη τιμή του coefficient του χαρακτηριστικού όπως φαίνετε στον πίνακα παραπάνω και IQR(feature) το ενδοτεταρτημοριακό του εύρος που προκύπτει απο τα δεδομένα.

Μέσω της ανάλυσης που κάναμε πήραμε τα παρακάτω διαγράμματα που δείχνουν την σχέση μεταξύ των ποσοτικών χαρακτηριστικών και του βαθμού του μαθητή:



Εύκολα μπορούμε να συμπεράνουμε πως:

1. **Attendance** Το ποσοστό της προσέλευση του μαθητή στο μάθημα αναμένετε να συμβάλλει στον βαθμό του μαθητή. Συγκεκριμένα η μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού αναμένετε να μεταβάλλει τον βαθμό του μαθητή κατά περίπου 0.2 μονάδες.
2. **Previous Scores** Ο μέσος όρος των βαθμών του μαθητή απο προηγούμενες εξετάσεις αναμένετε να συμβάλλει στον βαθμό του μαθητή. Συγκεκριμένα η μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού αναμένετε να μεταβάλλει τον βαθμό του μαθητή κατά περίπου 0.05 μονάδες.

3. **Tutoring Sessions** Ο αριθμός των φροντιστηριακών μαθημάτων ανά μήνα του μαθητή αναμένετε να συμβάλλει στον βαθμό του μαθητή. Συγκεκριμένα η μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού αναμένετε να μεταβάλλει τον βαθμό του μαθητή κατά περίπου 0.48 μονάδες.

4. **Hours Studied** Ο αριθμός των ωρών που διάβασε ο μαθητής μέσα στην βδομάδα αναμένετε να συμβάλλει στον βαθμό του μαθητή. Συγκεκριμένα η μεταβολή κατά μια μονάδα του χαρακτηριστικού αυτού αναμένετε να μεταβάλλει τον βαθμό του μαθητή κατά περίπου 0.29 μονάδες.

5 Βιβλιογραφία