ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

ΚΑΤΕΥΘΥΝΣΗ «ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ Τ.Π.Ε. ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ»

ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΚΤΙΜΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΤΩΝ ΜΑΘΗΤΩΝ ΣΤΗ ΔΕΥΤΕΡΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ

Μαροπάκης Αντώνιος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Τασουλής Σωτήριος

Λαμία, 2021

UNIVERSITY OF THESSALY

SCHOOL OF SCIENCE

INFORMATICS AND COMPUTATIONAL BIOMEDICINE

EXPLANATORY ANALYSIS AND PREDICTION OF STUDENT PERFORMANCE IN SECONDARY EDUCATION

Maropakis Antonios

Master thesis

Supervisor

Tasoulis Sotirios

Lamia, 2021

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο [«τίτλος εργασίας»] αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο ΔΗΛΩΝ

28/06/2021

Μαροπάκης Αντώνιος

ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΚΤΙΜΗΣΗ ΑΠΟΔΟΣΗΣ ΜΑΘΗΤΩΝ ΣΤΗ ΔΕΥΤΕΡΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ

Τριμελής Επιτροπή:

Τασουλής Σωτήριος (επιβλέπων),

Πλαγιανάκος Βασίλειος,

Καρανίκας Χαράλαμπος

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον κύριο Τασουλή Σωτήριο για τις πολύτιμες συμβουλές του. Η υποστήριξη σε γνωστικά θέματα, ο τρόπος σκέψης, οι πολλαπλές προσεγγίσεις σε επιμέρους ζητήματα και η γενικότερη καθοδήγησή του καθόλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας ήταν ιδιαιτέρως σημαντική.

Περίληψη

Η αξιολόγηση του μαθητή είναι μέρος της εκπαιδευτικής διαδικασίας. Η επίδοση του μαθητή ή η πιθανή επίδοσή του στο προσεχές μέλλον είναι ένας δείκτης του επιπέδου μάθησης του μαθητή αλλά και της ποιότητας της εκπαίδευσης. Ο σκοπός της εργασίας αυτής είναι διττός. Αρχικά επικεντρωνόμαστε στην διερεύνηση παραγόντων που επηρεάζουν την επίδοση του μαθητή, όπως το αν βρίσκεται σε σχέση (ρομαντική), η κατανάλωση αλκοόλ, ο χρόνος μελέτης, το φύλο, αξιοποιώντας μία ανοιχτή βάση δεδομένων που αφορά μαθητές δευτεροβάθμιας εκπαίδευσης. Στην συνέχεια μελετάμε τεχνικές μηχανικής μάθησης σε μία προσπάθεια να προβλεφθεί η τελική επίδοση των μαθητών καθώς είναι ιδιαίτερα σημαντική η επίτευξη της προγνωστικής ακρίβειας των μοντέλων. Οι πληροφορίες που αντλούνται μπορούν να χρησιμοποιηθούν από την εκπαιδευτική κοινότητα, η οποία λειτουργώντας παρεμβατικά δύναται να βοηθήσει το μαθητή ανάλογα με τις ικανότητες ή τις αδυναμίες του, αλλά και από την πολιτεία με παρεμβάσεις στον ευρύτερο χώρο της παιδείας. Τέλος, μέρος της εργασίας είναι η παρουσίαση του πηγαίου κώδικα με τον οποίο επιτεύχθηκε η ανάλυση σε γλώσσα προγραμματισμού R με σκοπό την άμεση αξιοποίηση και διεύρυνση των ερευνητικών αποτελεσμάτων από τους ενδιαφερόμενους αναγνώστες.

Λέξεις κλειδιά: επίδοση μαθητή, πρόβλεψη επίδοσης, διερευνητική ανάλυση, προγνωστικά μοντέλα, λογιστική παλινδρόμηση, Τυχαία δάση, γλώσσα προγραμματισμού R.

Abstract

Student assessment is part of the educational process. The student's performance or possible performance in the near future is an indicator of the student's level of learning and the quality of education. The purpose of this work is twofold. Initially we focus on investigating factors that affect student performance, such as whether they are in a relationship (romantic), alcohol consumption, study time, gender using an open database of high school students. Then we study machine learning techniques in an attempt to predict the final performance of students as it is particularly important to achieve predictive accuracy of models. The information obtained can be used by the educational community, which by intervening can help the student depending on his abilities or weaknesses, but also by the state with interventions in the wider field of education. Finally, part of the work is the presentation of the source code with which the analysis was achieved in R programming language, in order to immediately utilize and expand the research results by interested readers.

Keys words: performance of students, predictive performance, explanatory analysis, predictive models, logistic regression, Random Forests, program language R.

1 TABLE OF CONTENTS - ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

2 TABLE OF FIGURES – ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ

1 ΕΙΣΑΓΩΓΗ

Κάθε εκπαιδευτικό σύστημα οφείλει να έχει ως στόχο τη γνωστική αλλά και την κοινωνική και συναισθηματική ανάπτυξη του μαθητή. Ο μαθητής αξιολογείται ώστε να έχουμε την εικόνα της μαθησιακής του πορείας. Η επίδοσή του είναι το αποτέλεσμα της αξιολόγησής του μέσα από ένα πλήθος παραγόντων. Μετά τη σχολική προσαρμογή και την ανάπτυξη διαπροσωπικών σχέσεων στο σχολικό περιβάλλον-παράγοντες που επηρεάζουν την επίδοσή του- μπορεί ο μαθητής να επικεντρωθεί στο μαθησιακό του στόχο (Martin-Dowson, 2009). Έρευνα στο Οντάριο από τους Tremblay S. , Ross N. , Berthelot (2001) έδειξε ότι το φύλο αλλά και η οικονομική και κοινωνική θέση του οικογενειακού περιβάλλοντος του μαθητή έχουν επίδραση στην επίδοσή του.

Κατά τον Αθανασίου Λ. (2000 : 42) η επίδοση του μαθητή επηρεάζεται και από εγγενείς και εξωγενείς παράγοντες. Εγγενείς, όπως το ενδιαφέρον του μαθητή για το μάθημα, η μελέτη που κάνει ο μαθητής, η διάθεσή του για εργασία, τα κίνητρά του, ο χρόνος που διαθέτει. Εξωγενείς, όπως ο τρόπος διδασκαλίας από τον εκπαιδευτικό, η ποσότητα της ύλης που διδάσκεται κάθε φορά, τα σχολικά προγράμματα, το αναλυτικό πρόγραμμα, η υλικοτεχνική υποδομή του σχολείου, οι συνθήκες μάθησης στο σπίτι, ο βαθμός επικοινωνίας στο σχολείο με τους εκπαιδευτικούς, η σχέση του μαθητή με τους γονείς του.

Βασικός στόχος της αξιολόγησης του μαθητή είναι η ανατροφοδότηση της εκπαιδευτικής διαδικασίας και ο εντοπισμός των ελλείψεων σε μαθησιακό και ατομικό επίπεδο με απώτερο σκοπό τη βελτίωση της προσφερόμενης εκπαίδευσης και τελικά την πρόοδο όλων των μαθητών. Πιο συγκεκριμένα, με την αξιολόγηση μπορεί να διαπιστωθεί αρχικά σε ποιο βαθμό έχουν επιτευχθεί οι στόχοι της μάθησης αλλά και να σχεδιαστούν τα επόμενα στάδιά της. Με τον τρόπο αυτό διερευνάται και αποτυπώνεται η ατομική και συλλογική πορεία των μαθητών, καθώς και οι ικανότητες, τα ενδιαφέροντα και οι ιδιαιτερότητές τους σε όλα τα στάδια της γνώσης που προσφέρονται από τα προγράμματα σπουδών. Επιπλέον η αξιολόγηση αποσκοπεί στην ποιοτική αναβάθμιση συνολικά της εκπαιδευτικής διαδικασίας, η οποία έχει πολλαπλούς στόχους: την ενίσχυση και ενθάρρυνση των μαθητών, τη δημιουργία κινήτρων μάθησης, τον εντοπισμό

μαθησιακών δυσκολιών και των ελλείψεων των μαθητών με στόχο το σχεδιασμό κατάλληλων παρεμβάσεων για τη βελτίωση της διδακτικής πράξης. Και επιπλέον, την καλλιέργεια ερευνητικού πνεύματος, την ικανότητα επίλυσης προβλημάτων και την απόκτηση γνώσεων και δεξιοτήτων μέσα από διαθεματικές-διεπιστημονικές διαδικασίες. Τέλος, σε προσωπικό επίπεδο στοχεύει στην ανάπτυξη της κριτικής και δημιουργικής ικανότητας, στην ενίσχυση της υπευθυνότητας της αυτοπεποίθησης και αυτοεκτίμησης των μαθητών και συνολικά στη συγκρότηση της προσωπικότητάς τους. (Κωνσταντίνου Χ., 2002:39)

Στην παρούσα εργασία γίνεται διερευνητική ανάλυση δεδομένων από δύο σχολεία δευτεροβάθμιας εκπαίδευσης στην Πορτογαλία (Cortez P., Silva A., 2008). Οπτικοποιούνται οι συσχετίσεις των μεταβλητών μέσω θερμικού χάρτη ή ενός δικτύου ή ραβδογράμματος. Εξετάζεται η συσχέτιση κατανάλωσης αλκοόλ με το φύλο και κατά πόσο η χρήση αλκοόλ επηρεάζει την επίδοση των μαθητών. Μελετάμε τη σχέση χρόνου μελέτης με την επίδοση των μαθητών λαμβάνοντας υπόψη το φύλο. Γίνεται οπτικοποίηση των διαφορών των επιδόσεων αγοριών και κοριτσιών και μέσω ιστογραμμάτων ή διεξάγοντας τεστ υποθέσεων, καταλήγοντας σε συμπεράσματα που αφορούν τη σχέση φύλου και επίδοσης. Επίσης, ιδιαίτερο ενδιαφέρον παρουσιάζει το αν οι προσωπικοί-συναισθηματικοί δεσμοί των μαθητών επηρεάζουν την επίδοσή τους. Στην επίδοση, όμως, των μαθητών έχει επιρροή και το μορφωτικό επίπεδο είτε της μητέρας είτε του πατέρα. Έτσι, και η παρούσα εργασία θα αναδείξει το βαθμό επιρροής της μόρφωσης των γονέων στην πρόοδο των παιδιών.

Η εργασία αυτή δεν επικεντρώνεται μόνο σε μεταβλητές, σε παράγοντες που επηρεάζουν την επίδοση. Σκοπός της είναι και η πρόβλεψη της επίδοσης. Για το λόγο αυτό δοκιμάζονται προγνωστικά μοντέλα, μοντέλα εξόρυξης δεδομένων, όπως λογιστική παλινδρόμηση ή ο αλγόριθμος Random Forest με στόχο την πρόγνωση της προόδου των μαθητών. Στην έρευνα των Ma et al (2000) εφαρμόζοντας αλγόριθμο εξόρυξης δεδομένων στόχος ήταν η επιλογή αδύναμων μαθητών σε σχολείο της Σιγκαπούρης για ενισχυτική διδασκαλία. Η επίδοση και η προβλεπόμενη επίδοση του μαθητή είναι ένα εργαλείο ωφέλιμο για τον εκπαιδευτικό. Μέριμνα του εκπαιδευτικού που αξιολογεί είναι να

εντοπίζει και να ερμηνεύει τις αιτίες των προβλημάτων και στη συνέχεια να παρεμβαίνει κατάλληλα με διδακτικά-παιδαγωγικά μέτρα επαναπροσδιορίζοντας τους μαθησιακούς στόχους. Επιπλέον, οφείλει να ελέγχει την πορεία οργάνωσης της εκπαιδευτικής διαδικασίας για την οποία είναι υπεύθυνος. (Κωνσταντίνου Χ., 2002:40)

Είναι πολύ σημαντικό, λοιπόν, να γνωρίζουμε εκ των προτέρων την πιθανή επίδοση του μαθητή, ώστε να μπορεί να παρεμβαίνει άμεσα το σχολείο, η εκπαιδευτική κοινότητα γενικότερα και η πολιτεία με στόχο πάντα τη βελτίωση της επίδοσης του μαθητή και την καλύτερη ποιότητα της εκπαίδευσης.

1.1 Βιβλιογραφικη επισκοπηση

Στο National Environmental Education and Training Foundation γίνεται αναφορά σε παράγοντες που επηρεάζουν θετικά ή αρνητικά τη σχολική επίδοση των μαθητών (Glenn, Joanne Lozar,2000). Διαχωρίζονται σε κατηγορίες: εξωτερικοί παράγοντες, εσωτερικοί παράγοντες, κοινωνικοί και διοικητικοί παράγοντες, όπως και παράγοντες σχετιζόμενοι με το αναλυτικό πρόγραμμα σπουδών. Υπάρχει συλλογή από μελέτες σχολείων στο Τέξας, τη Βόρεια Καρολίνα, το Ουισκόνσιν, τη Μινεσότα, το Κεντάκι και τη Φλόριντα. Ο στόχος είναι η διαφοροποίηση από τις παραδοσιακές εκπαιδευτικές προσεγγίσεις και η βελτίωση της ακαδημαϊκής απόδοσης σε ολόκληρο το πρόγραμμα σπουδών.

Η απόδοση των μαθητών είναι πολύ σημαντική για τα εκπαιδευτικά ιδρύματα και την πολιτεία για τη λήψη αποφάσεων έχοντας στόχο την ποιότητα της εκπαίδευσης. Οι T.PanduRanga Vital, B.G.Lakshmi , H.Swapna Rekha , M.DhanaLakshmi (2018) ανέλυσαν την απόδοση του μαθητή χρησιμοποιώντας αλγόριθμους στατιστικής και μη επιβλεπόμενης μηχανικής μάθησης. Χρησιμοποιήθηκαν ιεραρχικοί αλγόριθμοι και k-means συσταδοποίησης. Η εκπαίδευση ενός μαθητή εξαρτάται κυρίως από το οικογενειακό υπόβαθρο, το προσωπικό του προφίλ και τις δραστηριότητές του, αλλά και από παράγοντες όπως η κατανάλωση αλκοόλ, η έξοδος με φίλους και οι ρομαντικές σχέσεις. Οι αλγόριθμοι μη επιβλεπόμενης μηχανικής μάθησης δίνουν καλά αποτελέσματα για την πρόβλεψη της απόδοσης (επιτυχία ή αποτυχία) του μαθητή.

Τα περισσότερα τρέχοντα μοντέλα πρόβλεψης είναι δύσκολο για τους εκπαιδευτικούς να τα ερμηνεύσουν. Αυτό δημιουργεί σημαντικά προβλήματα στη χρήση των μοντέλων (για παράδειγμα στην εξατομίκευση της εκπαίδευσης και στην παρέμβαση του εκπαιδευτικού) καθώς και στην αξιολόγηση των μοντέλων αυτών. Οι Xing W., Guo R., Petakovic E., Goggins S. (2015) συνέθεσαν προσεγγίσεις μαθησιακής ανάλυσης, εκπαιδευτική εξόρυξη δεδομένων (EDM) και θεωρία HCI (αλληλεπίδραση ανθρώπου υπολογιστή) για τη διερεύνηση της ανάπτυξης πιο εύχρηστων μοντέλων πρόβλεψης και των αναπαραστάσεων των μοντέλων χρησιμοποιώντας δεδομένα από ένα συνεργατικό περιβάλλον επίλυσης προβλημάτων γεωμετρίας.

Ο Daud A.,et al. (2017) με στόχο την πρόβλεψη επιτυχίας μαθητών χρησιμοποίησε μεθόδους εξόρυξης δεδομένων και ανάλυση δεδομένων μαθητών (LA) τριτοβάθμιας εκπαίδευσης. Χρησιμοποίησε χαρακτηριστικά όπως οικογενειακές δαπάνες και προσωπικά στοιχεία μαθητών, όπως και δεδομένα από διαφορετικά πανεπιστήμια του Πακιστάν (όπως φοιτητές με υποτροφία). Διακριτικά και γενετικά μοντέλα ταξινόμησης εφαρμόζονται για να προβλέψουν εάν ένας μαθητής θα είναι σε θέση να ολοκληρώσει το πτυχίο του ή όχι.

Ο Chokla M. (2013) εστίασε στην έρευνά του στη στατιστική ανάλυση επιδόσεων φοιτητών σε επανασχεδιασμένα αναπτυξιακά μαθηματικά. Τα κολέγια και τα πανεπιστήμια εστιάζουν τις προσπάθειές τους στη βελτίωση της διδασκαλίας στα αναπτυξιακά μαθήματα μαθηματικών με στόχο τη βελτίωση των ποσοστών αποφοίτησης των μαθητών. Ο σκοπός της μελέτης του Chokla είναι να διερευνήσει τις διαφορές στις ακαδημαϊκές βελτιώσεις των μαθητών αναπτυξιακών μαθηματικών προκειμένου να αξιολογηθεί η αποτελεσματικότητα του επανασχεδιασμένου κύκλου μαθημάτων. Πραγματοποιήθηκαν πολλαπλές αναλύσεις. Στην πρώτη φάση της μελέτης δύο μοντέλα γραμμικής παλινδρόμησης αναπτύχθηκαν για την πρόβλεψη της ακαδημαϊκής βελτίωσης των μαθητών σε συγκεκριμένα αναπτυξιακά μαθήματα μαθηματικών. Στη δεύτερη φάση της μελέτης τρία γραμμικά μοντέλα αναλύθηκαν για να προβλέψουν την ακαδημαϊκή απόδοση φοιτητών σε επανασχεδιασμένα μαθήματα. Τρία επιπλέον μοντέλα

παλινδρόμησης αναπτύχθηκαν αλλά και μέθοδος πρόβλεψης εκτός δείγματος για την αξιολόγηση του ποσοστού εσφαλμένης ταξινόμησης μαθητών που χρήζουν προσοχής.

1.2 ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

Δεδομένα μαθητή

Στην Πορτογαλία, η δευτεροβάθμια εκπαίδευση διαρκεί 3 χρόνια, μετά από 9 χρόνια βασικής εκπαίδευσης και ακολουθεί η τριτοβάθμια εκπαίδευση. Οι περισσότεροι μαθητές συμμετέχουν στο δημόσιο και δωρεάν εκπαιδευτικό σύστημα. Εκτός από τα βασικά μαθήματα, Γλώσσα και Μαθηματικά, υπάρχουν και άλλα μαθήματα όπως για παράδειγμα Επιστήμη υπολογιστών, άλλες επιστήμες, Εικαστικές Τέχνες. Όπως και σε πολλές άλλες χώρες (π.χ. Γαλλία ή Βενεζουέλα), έτσι και στην Πορτογαλία χρησιμοποιείται μια κλίμακα βαθμολογίας 20 βαθμών, όπου το 0 είναι η χαμηλότερη βαθμολογία και το 20 η υψηλότερη. Κατά τη διάρκεια του σχολικού έτους οι μαθητές αξιολογούνται σε τρεις περιόδους και η τελευταία αξιολόγηση (G3 του Πίνακα 1) αντιστοιχεί στον τελικό βαθμό.

Αυτή η μελέτη θα εξετάσει τα δεδομένα που συλλέχθηκαν κατά τη διάρκεια του Σχολικού έτους 2005-2006 από δύο δημόσια σχολεία της περιοχής Αλεντέζου της Πορτογαλίας. Αν και υπήρξε μια τάση για αύξηση των επενδύσεων στην νέες τεχνολογίες από την κυβέρνηση, στην πλειοψηφία τους τα πληροφοριακά συστήματα του δημόσιου σχολείου στην Πορτογαλία είναι πολύ φτωχά, βασισμένα κυρίως σε φύλλα χαρτιού (που ήταν η τρέχουσα περίπτωση).

Ως εκ τούτου, η βάση δεδομένων δημιουργήθηκε από δύο πηγές: σχολικές αναφορές, μη μηχανογραφημένες, που περιλαμβάνουν λίγα χαρακτηριστικά (δηλαδή τους βαθμούς τριών περιόδων και το πλήθος των απουσιών) και ερωτηματολόγια, που χρησιμοποιούνται για τη συμπλήρωση των προηγούμενων πληροφοριών. Σχεδιάσαμε τη δεύτερη πηγή με ερωτήσεις κλειστού τύπου (δηλαδή με προκαθορισμένες επιλογές) που σχετίζονται με στοιχεία δημογραφικά (π.χ. εκπαίδευση μητέρας, οικογενειακό εισόδημα), κοινωνικά/ συναισθηματικά (π.χ. κατανάλωση αλκοόλ) (Pritchard and Wilson, 2003) και με το σχολείο (π.χ. αριθμός μαθημάτων που απέτυχε ο μαθητής σε προηγούμενης περίοδο) μεταβλητές που ήταν αναμενόμενο να επηρεάσουν την απόδοση των μαθητών.

Το ερωτηματολόγιο ελέγχθηκε από εκπαιδευτικούς του σχολείου και δοκιμάστηκε σε ένα μικρό σύνολο 15 μαθητών με τα σχόλιά τους πάνω στο ερωτηματολόγιο. Η τελική έκδοση περιείχε 37 ερωτήσεις σε απλό φύλλο Α4 και απαντήθηκε στην τάξη από 788 μαθητές. Επίσης, 111 απαντήσεις απορρίφθηκαν λόγω έλλειψης ταυτοποίησης στοιχείων (απαραίτητο να συνδυαστούν με τις σχολικές αναφορές). Τέλος, τα δεδομένα ενσωματώθηκαν σε δύο σύνολα δεδομένων που σχετίζονται με τα μαθήματα των Μαθηματικών (με 395 παραδείγματα) και της Πορτογαλικής γλώσσας (649 εγγραφές).

Κατά το στάδιο της προεπεξεργασίας, ορισμένα στοιχεία απορρίφθηκαν πιθανώς λόγω προβλημάτων απορρήτου. Για παράδειγμα, μερικοί ερωτηθέντες απάντησαν για το οικογενειακό τους εισόδημα , ενώ σχεδόν το 100% των μαθητών ζουν με τους γονείς τους και έχουν προσωπικό υπολογιστή στο σπίτι.

Τα υπόλοιπα χαρακτηριστικά φαίνονται στον Πίνακα 1, όπου στις τελευταίες τέσσερις σειρές είναι δηλωμένες μεταβλητές με τιμές, πληροφορίες που λαμβάνονται από τις σχολικές αναφορές.

Πίνακας 1: Προεπεξεργασμένες μεταβλητές (σχετικές με μαθητές)

Όνομα	**Περιγραφή**	**(Τιμή)**
sex	φύλο μαθητή	(κορίτσι ή αγόρι)
age	ηλικία μαθητή	(από 15 έως 22)
school	σχολείο μαθητή	(Gabriel Pereira ή Mousinho da Silveira)
address	περιοχή κατοικίας μαθητή	(αστική ή αγροτική)
pstatus	κατάσταση συμβίωσης γονέα	(ζουν μαζί ή χωριστά)
Medu	Εκπαίδευση της μητέρας	(από 0 έως 4) **α**
Mjob	εργασία της μητέρας	(ονομαστικό **b**)
guardian	κηδεμόνας μαθητή	(μητέρα, πατέρας ή άλλος)
famsize	πλήθος ατόμων στην οικογένεια	(≤ 3 ή > 3)
famrel	ποιότητα οικογενειακής σχέσης	(από 1 - πολύ κακή έως 5 - εξαιρετική)

reason	λόγος επιλογής σχολείου	(κοντά στο σπίτι, φήμη σχολείου, προτίμηση μαθημάτων ή άλλα)
traveltime	χρόνος ταξιδιού σπίτι-σχολείο	(1 - <15 λεπτά, 2 - 15 έως 30 λεπτά, 3 – 30 λεπτά έως 1 ώρα ή 4 - > 1 ώρα)
Studytime	Εβδομαδιαίος χρόνος μελέτης	(1 - <2 ώρες, 2 - 2 έως 5 ώρες, 3 – 5 έως 10 ώρες ή 4- > 10 ώρες
failures	αριθμός μαθημάτων αποτυχίας	(n εάν 1 ≤ n <3, αλλιώς 4)
schoolsup	έξτρα εκπαιδευτική σχολική ενίσχυση	(ναι ή όχι)
famsup	οικογενειακή εκπαιδευτική υποστήριξη	(ναι ή όχι)
activities	Εξωσχολικές δραστηριότητες	(ναι ή όχι)
paidclass	έξτρα αμειβόμενα μαθήματα	(ναι ή όχι)
internet	Πρόσβαση στο Διαδίκτυο στο σπίτι	(ναι ή όχι)
nursery	παρακολούθηση νηπιαγωγείου	(ναι ή όχι)
higher	επιθυμία για τριτοβάθμια εκπαίδευση	(ναι ή όχι)
romantic	είναι σε σχέση	(ναι ή όχι)
freetime	ελεύθερος χρόνος μετά το σχολείο	(από 1 - πολύ λίγος έως 5 – πάρα πολύ)
goout	Έξοδος με φίλους	(από 1 - πολύ λίγο έως 5 – πάρα πολύ)
Walc	Κατανάλωση αλκοόλ Σαββατοκύριακου	(από 1 - πολύ μικρή έως 5 – πολύ μεγάλη)
Dalc	Κατανάλωση αλκοόλ εργάσιμης ημέρας	(από 1 – πολύ μικρή έως 5– πολύ μεγάλη)
health	Τρέχουσα κατάσταση υγείας	(από 1 - πολύ κακή έως 5 – πολύ καλή)
absences	Πλήθος απουσιών	(από 0 έως 93)
G1	Βαθμός πρώτης περιόδου	(από 0 έως 20)
G2	Βαθμός δεύτερης περιόδου	(από 0 έως 20)
G3	Τελικός βαθμός	(από 0 έως 20)

όπου

a : 0 - κανένα, 1 - πρωτοβάθμια εκπαίδευση (4η τάξη), 2 - 5η έως 9η τάξη, 3 - δευτεροβάθμια εκπαίδευση ή 4 - τριτοβάθμια εκπαίδευση.

b : δάσκαλος, υγειονομική περίθαλψη, δημόσιες υπηρεσίες (π.χ. διοικητικές ή αστυνομικές), στο σπίτι ή σε άλλο.

2 ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Τυπική μορφή δεδομένων και τύποι διερευνητικής ανάλυσης δεδομένων

Τα δεδομένα από ένα πείραμα συλλέγονται γενικά σε έναν ορθογώνιο πίνακα (π.χ. υπολογιστικό φύλλο ή βάση δεδομένων), συνήθως με μία γραμμή του πίνακα ανά πειραματικό θέμα και μία στήλη για τα δεδομένα εξόδου για κάθε επεξηγηματική μεταβλητή αποτελέσματος. Κάθε στήλη περιέχει τις αριθμητικές τιμές για μία συγκεκριμένη ποσοτική μεταβλητή ή τα επίπεδα για μια κατηγορηματική μεταβλητή. (Μερικά πιο περίπλοκα πειράματα απαιτούν μια πιο περίπλοκη διάταξη δεδομένων). Οι άνθρωποι δεν έχουν την ικανότητα “βλέποντας” μια στήλη αριθμών ή ένα ολόκληρο υπολογιστικό φύλλο να καθορίζουν σημαντικά χαρακτηριστικά των δεδομένων. Θεωρούν ότι οι αριθμοί είναι κουραστικοί ή βαρετοί . Τεχνικές ανάλυσης διερευνητικών δεδομένων έχουν επινοηθεί ως βοήθεια σε αυτήν την κατάσταση. Οι περισσότερες από αυτές τις τεχνικές λειτουργούν εν μέρει αποκρύπτοντας ορισμένες πτυχές των δεδομένων, ενώ δημιουργούν άλλες πιο ξεκάθαρες πτυχές. Η ταξινόμηση των μεθόδων διερευνητικής ανάλυσης δεδομένων γενικά γίνεται με δύο τρόπους: α. κάθε μέθοδος είναι είτε μη γραφική είτε γραφική και β. η ανάλυση είναι είτε μία είτε δύο είτε πολλών μεταβλητών.

Οι μη γραφικές μέθοδοι περιλαμβάνουν γενικά τον υπολογισμό των συνοπτικών στατιστικών, ενώ οι γραφικές μέθοδοι συνοψίζουν προφανώς τα δεδομένα με διαγραμματικό ή εικονογραφικό τρόπο. Οι μέθοδοι μίας μεταβλητής εξετάζουν μία μεταβλητή (στήλη δεδομένων) κάθε φορά, ενώ οι μέθοδοι πολλών μεταβλητών εξετάζουν δύο ή περισσότερες μεταβλητές κάθε φορά για να διερευνήσουν σχέσεις. Συνήθως, η πολυμεταβλητή διερευνητική ανάλυση δεδομένων γίνεται εξετάζοντας δύο μεταβλητές, αλλά περιστασιακά και τρεις ή περισσότερες μεταβλητές. Είναι, σχεδόν πάντα, μία καλή ιδέα να εφαρμοστεί μέθοδος μίας μεταβλητής για καθένα από τα στοιχεία της πολυπαραγοντικής ανάλυσης πριν εφαρμοστεί η ανάλυση πολλών μεταβλητών. Πέρα από τις τέσσερις κατηγορίες ανάλυσης διερευνητικών δεδομένων που αναφέρθηκαν παραπάνω, κάθε μία από τις κατηγορίες επιπλέον διακρίνεται με βάση το ρόλο

(αποτέλεσμα ή επεξηγηματικό) και τον τύπο (κατηγορικές ή ποσοτικές) των μεταβλητών που εξετάζονται.

Αν και υπάρχουν οδηγίες σχετικά με το ποιες τεχνικές ανάλυσης είναι χρήσιμες σε κάθε περίσταση, υπάρχει ένας σημαντικός βαθμός έλλειψης ακρίβειας και δημιουργικότητας στην διερευνητική ανάλυση δεδομένων. Η ικανότητα και η εμπιστοσύνη έρχονται με την πρακτική, την εμπειρία και τη στενή παρατήρηση των άλλων. Επίσης, η ανάλυση δεν χρειάζεται να περιορίζεται σε τεχνικές που έχετε ξαναδεί. Μερικές φορές πρέπει να εφεύρετε έναν νέο τρόπο εξέτασης των δεδομένων σας.

Διερευνητική μη γραφική Ανάλυση Δεδομένων μίας μεταβλητής

Τα δεδομένα που προέρχονται από την πραγματοποίηση μίας συγκεκριμένης μέτρησης σε όλα τα υποκείμενα ενός αντιπροσωπευτικού δείγματος είναι οι παρατηρήσεις μας για ένα μοναδικό χαρακτηριστικό όπως η ηλικία, το φύλο, η ταχύτητα σε μια εργασία ή η αντίδραση σε ένα ερέθισμα. Πρέπει να λογιστούμε ότι αυτά τα μέτρα αντιπροσωπεύουν μια «κατανομή δείγματος» της μεταβλητής, η οποία με τη σειρά της περισσότερο ή λιγότερο αντιπροσωπεύει την «κατανομή πληθυσμού» της μεταβλητής. Ο συνηθισμένος στόχος της διερευνητικής μη γραφικής ανάλυσης δεδομένων μιας μεταβλητής είναι να εκτιμήσουμε καλύτερα την «κατανομή δειγμάτων» και επίσης να βγάλουμε κάποια αρχικά συμπεράσματα σχετικά με την κατανομή του πληθυσμού και κατά πόσο ταιριάζουν με την κατανομή των δειγμάτων. Ακραίες τιμές είναι επίσης μέρος αυτής της ανάλυσης.

Κατηγορικά δεδομένα

Τα χαρακτηριστικά ενδιαφέροντος για μια κατηγορική μεταβλητή είναι απλά το εύρος των τιμών και η συχνότητα (ή σχετική συχνότητα) εμφάνισης για κάθε τιμή. (Για διατάξιμες μεταβλητές είναι μερικές φορές σκόπιμο να τις αντιμετωπίσουμε ως ποσοτικές μεταβλητές). Επομένως, οι μόνες χρήσιμες μη γραφικές τεχνικές μιας μεταβλητής, για κατηγορικές μεταβλητές είναι ο πίνακας των συχνοτήτων, συνήθως μαζί με τον υπολογισμό των σχετικών συχνοτήτων.

Χαρακτηριστικά ποσοτικών δεδομένων

Η διερευνητική ανάλυση δεδομένων μίας μεταβλητής για μια ποσοτική μεταβλητή είναι ένας τρόπος για να κάνετε αρχικές εκτιμήσεις σχετικά με την κατανομή του πληθυσμού όσον αφορά τη μεταβλητή χρησιμοποιώντας τα δεδομένα του παρατηρούμενου δείγματος. Τα χαρακτηριστικά της κατανομής του πληθυσμού μιας ποσοτικής μεταβλητής είναι το κέντρο της κατανομής, η διασπορά, η επικρατούσα τιμή (αριθμός κορυφών), το σχήμα (συμπεριλαμβανομένης της «βαρύτητας των ουρών») και οι ακραίες τιμές.

Τα δεδομένα που παρατηρήθηκαν αντιπροσωπεύουν μόνο ένα δείγμα από έναν άπειρο αριθμό πιθανών δειγμάτων. Τα χαρακτηριστικά από το τυχαία παρατηρούμενο δείγμα μας δεν είναι από μόνα τους ενδιαφέροντα, εκτός από το βαθμό που αντιπροσωπεύουν τον πληθυσμό από τον οποίο προήλθαν. Αυτό που παρατηρούμε στις μετρήσεις του δείγματος για μια συγκεκριμένη μεταβλητή που επιλέγουμε για το συγκεκριμένο πείραμά μας είναι η «κατανομή δείγματος». Χρειαζόμαστε να αναγνωρίσουμε ότι αυτή θα ήταν διαφορετική κάθε φορά που θα μπορούσαμε να επαναλάβουμε το πείραμα, λόγω της επιλογής ενός διαφορετικού τυχαίου δείγματος, μιας διαφορετικής τυχαιοποίησης, σε διαφορετικές τυχαίες πειραματικές συνθήκες.

Επιπλέον, μπορούμε να υπολογίσουμε από τα δεδομένα στατιστικών δειγμάτων, αριθμητικά μέτρα όπως μέση τιμή, διακύμανση, τυπική απόκλιση, αλλά και μέτρα ασυμμετρίας και κύρτωσης. Αυτά και πάλι θα ποικίλουν για κάθε επανάληψη του πειράματος, έτσι δεν αντιπροσωπεύουν καμία βαθιά αλήθεια, αλλά αντιπροσωπεύουν ορισμένες αβέβαιες πληροφορίες σχετικά με την υποκείμενη κατανομή του πληθυσμού και τις παραμέτρους της, οι οποίες είναι αυτές που πραγματικά μας ενδιαφέρουν.

Κεντρική τάση (Θέση κατανομής)

Η κεντρική τάση ή «θέση» μιας κατανομής έχει σχέση με τις κεντρικές τιμές. Τα πιο συνηθισμένα και χρήσιμα μέτρα θέσης είναι η μέση τιμή (αριθμητικός μέσος), η διάμεσος και η επικρατούσα τιμή (κορυφή). Περιστασιακά γίνεται χρήση άλλων μέτρων θέσης όπως ο γεωμετρικός μέσος, ο αρμονικός, ο περικομμένος και ο Winsorized μέσος. Ενώ οι περισσότεροι συγγραφείς χρησιμοποιούν τον όρο «μέσος όρος» ως συνώνυμο για τον

αριθμητικό μέσο, ορισμένοι χρησιμοποιούν τον μέσο όρο με την ευρύτερη έννοια για να συμπεριλάβουν επίσης γεωμετρικό, αρμονικό και άλλους μέσους. Υποθέτοντας ότι έχουμε n τιμές δεδομένων x_1 έως x_n, τότε ο τύπος για τον υπολογισμό του αριθμητικού μέσου του δείγματος είναι:

$$\bar{x}\frac{1}{n}\Sigma_{i=1}^{n}x_i$$

Ο **αριθμητικός μέσος** είναι απλά το άθροισμα όλων των τιμών δεδομένων διαιρούμενο με το πλήθος των τιμών.

Η **διάμεσος** είναι ένα άλλο μέτρο της κεντρικής τάσης. Η διάμεσος του δείγματος είναι η μεσαία τιμή εφόσον όλες οι τιμές έχουν διαταχθεί σε αύξουσα σειρά. Αν το πλήθος των τιμών είναι άρτιος, παίρνουμε τον μέσο όρο των δύο μεσαίων τιμών. Για συμμετρικές κατανομές, η μέση τιμή και η διάμεσος συμπίπτουν. Για μονοκόρυφες ασύμμετρες κατανομές, η μέση τιμή κινείται προς την "ουρά" της κατανομής. Επομένως σε πολλές περιπτώσεις ασύμμετρων κατανομών, η διάμεσος προτιμάται ως μέτρο κεντρικής τάσης.

Ένα σπάνια χρησιμοποιούμενο μέτρο της κεντρικής τάσης είναι η **επικρατούσα τιμή** που είναι η τιμή με τη μεγαλύτερη συχνότητα. Μία κατανομή μπορεί να έχει μία επικρατούσα τιμή- μία κορυφή-(μονοκόρυφη) ή δύο ή περισσότερες κορυφές (δικόρυφες ή πολυκόρυφες). Σε συμμετρικές, μονοκόρυφες κατανομές, η επικρατούσα τιμή ισούται με τη μέση τιμή και τη διάμεσο. Σε μονοκόρυφες, ασύμμετρες κατανομές η επικρατούσα τιμή (είναι στην άλλη πλευρά της διαμέσου από τη μέση τιμή) κινείται αντίθετα προς την ουρά της κατανομής. Σε πολυκόρυφες κατανομές η επικρατούσα τιμή μπορεί να μην αντιπροσωπεύει την κεντρική τάση.

Το πιο συνηθισμένο μέτρο της κεντρικής τάσης είναι η μέση τιμή. Για ασύμμετρη κατανομή ή όταν υπάρχει ανησυχία για τις ακραίες τιμές, μπορεί να προτιμάται η διάμεσος.

Μεταβλητότητα

Διάφορα στατιστικά στοιχεία χρησιμοποιούνται συνήθως ως μέτρα διασποράς μιας κατανομής, συμπεριλαμβανομένης της **διακύμανσης** (διασπορά), της **τυπικής απόκλισης** και του **εύρους** μεταξύ αυτών.

Η διασπορά είναι ένας δείκτης του πόσο μακριά από το κέντρο είναι πιθανό να βρούμε τιμές δεδομένων. Η διακύμανση είναι ένα τυπικό μέτρο διασποράς. Έστω x_1 , x_2 , … , x_n οι παρατηρήσεις ενός δείγματος μεγέθους n. Στη συνέχεια, για οποιαδήποτε δεδομένη τιμή, xi, η αντίστοιχη απόκλιση είναι $(x_i - \bar{x})$, είναι η απόσταση της τιμής δεδομένων από τη μέση τιμή όλων των δεδομένων n. Δεν είναι δύσκολο να αποδειχθεί ότι το άθροισμα όλων των αποκλίσεων ενός δείγματος είναι μηδέν. Η διακύμανση ενός πληθυσμού ορίζεται ως η μέση τετραγωνική απόκλιση. Ο τύπος δείγματος για τη διακύμανση των παρατηρούμενων δεδομένων συμβατικά έχει n − 1 στον παρονομαστή αντί για n για να επιτύχει την ιδιότητα της «αμεροληψίας», πράγμα που σημαίνει ότι όταν υπολογίζεται για πολλά διαφορετικά τυχαία δείγματα από τον ίδιο πληθυσμό, ο μέσος όρος πρέπει να ταιριάζει με τον αντίστοιχη ποσότητα του πληθυσμού. Το πιο συχνά χρησιμοποιούμενο σύμβολο για διακύμανση δείγματος είναι το s^2 και ο τύπος είναι:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Είναι ένα μέτρο μεταβλητότητας, γιατί όσο μεγαλύτερες είναι οι αποκλίσεις από το μέσο όρο, τόσο μεγαλύτερη είναι η διακύμανση. (Στις περισσότερες περιπτώσεις, το τετράγωνο είναι καλύτερο από το να παίρνεις την απόλυτη τιμή γιατί δίνει ιδιαίτερη έμφαση σε εξαιρετικά αποκλίνουσες τιμές). Η διακύμανση είναι μοναδική και σταθερή εκτίμηση δείγματος (μεταβάλλεται από δείγμα σε δείγμα). Λόγω του τετραγώνου, οι διακυμάνσεις είναι πάντα μη αρνητικές και έχουν την κάπως ασυνήθιστη ιδιότητα να έχουμε τετραγωνικές μονάδες σε σύγκριση με τα αρχικά δεδομένα.

Η **τυπική απόκλιση** είναι απλώς η θετική τετραγωνική ρίζα της διακύμανσης. Ως εκ τούτου έχει τις ίδιες μονάδες με τα αρχικά δεδομένα, τα οποία βοηθούν να γίνει πιο ερμηνεύσιμο στατιστικό στοιχείο. Η τυπική απόκλιση δείγματος αντιπροσωπεύεται συνήθως από το σύμβολο s.

$$s = \sqrt{s^2}$$

Σε μία κανονική κατανομή περίπου το 95% των τιμών βρίσκονται εντός δύο τυπικών αποκλίσεων από τη μέση τιμή, δηλαδή στο διάστημα $(\bar{x} - 2s, \bar{x} + 2s)$

Ένα τρίτο μέτρο διασποράς είναι το **ενδοτεταρτημοριακό εύρος**. Για να το ορίσουμε, εμείς πρώτα πρέπει να καθορίσουμε τις έννοιες των τεταρτημορίων. Τα

τεταρτημόρια ενός πληθυσμού ή ενός δείγματος είναι οι τρεις τιμές που χωρίζουν την κατανομή ή τα δεδομένα που παρατηρούνται σε τέταρτα. Έτσι, το ένα τέταρτο των δεδομένων είναι το πρώτο τεταρτημόριο (Q1) το μισό είναι το δεύτερο τεταρτημόριο (Q2) και τα τρία τέταρτα είναι το τρίτο τεταρτημόριο (Q3). Ο οξυδερκής αναγνώστης θα συνειδητοποιήσει ότι οι μισές από τις τιμές πέφτουν πάνω στο Q2, ένα τέταρτο πέφτει πάνω από το Q3, και επίσης ότι το Q2 είναι συνώνυμο της διαμέσου. Μόλις οριστούν τα τεταρτημόρια, είναι εύκολο να οριστεί το ενδοτεταρτημοριακό εύρος ως **IQR = Q_3 - Q_1**. Εξ ορισμού, οι μισές από τις τιμές (και ειδικά το μεσαίο μισό) εμπίπτουν στο διάστημα του οποίου το πλάτος ισούται με το IQR. Εάν τα δεδομένα είναι πιο διασκορπισμένα, τότε το IQR τείνει να αυξάνεται και το αντίστροφο. Το IQR είναι ένα πιο ισχυρό μέτρο μεταβλητότητας από τη διακύμανση ή την τυπική απόκλιση. Οποιοσδήποτε αριθμός τιμών στο πάνω ή κάτω τέταρτο των δεδομένων μπορεί να μετακινηθεί σε οποιαδήποτε απόσταση από τη διάμεσο χωρίς να επηρεαστεί καθόλου το IQR. Περισσότερο πρακτικά, μερικές ακραίες τιμές έχουν ελάχιστη ή καθόλου επίδραση στο IQR.

Σε αντίθεση με το IQR, το εύρος των δεδομένων δεν είναι ισχυρό μέτρο μεταβλητότητας. Το **εύρος** δείγματος είναι η απόσταση της ελάχιστης τιμής από τη μέγιστη : $\mathbf{R=x_{max}-x_{min}}$, εύρος = μέγιστο – ελάχιστο. Εάν συλλέγετε επαναλαμβανόμενα δείγματα από έναν πληθυσμό, το ελάχιστο, το μέγιστο και το εύρος τείνουν να αλλάζουν δραστικά από δείγμα σε δείγμα, ενώ η διακύμανση και η τυπική απόκλιση αλλάζουν λιγότερο, και το IQR λιγότερο από όλα.

Ασυμμετρία (στρέβλωση) και Κύρτωση

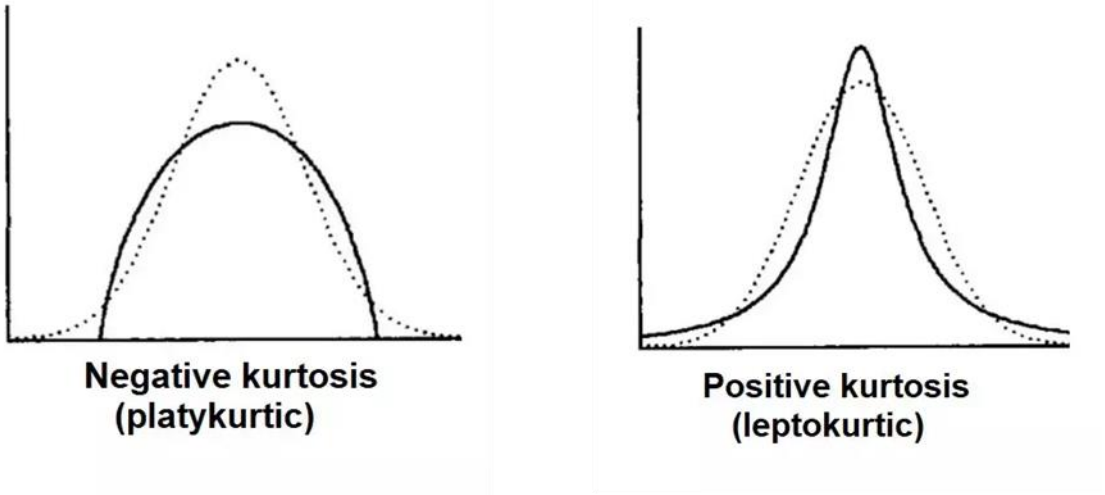
Η στρέβλωση και η κύρτωση μπορούν επιπλέον να περιγράψουν μία κατανομή μίας μεταβλητής. Η στρέβλωση είναι ένα μέτρο ασυμμετρίας και η κύρτωση είναι ένα μέτρο της οξύτητας της κορυφής της κατανομής. Οι εκτιμήσεις δειγμάτων για την στρέβλωση και την κύρτωση λαμβάνονται ως εκτιμήσεις των αντίστοιχων παραμέτρων του πληθυσμού. Εάν η στρέβλωση και η κύρτωση υπολογίζονται μαζί με τα τυπικά σφάλματά τους, μπορούμε να εξάγουμε συμπεράσματα σύμφωνα με τον παρακάτω

πίνακα όπου το e είναι εκτίμηση της ασυμμετρίας και u είναι μια εκτίμηση της κύρτωσης, SE(e) και SE(u) είναι τα αντίστοιχα τυπικά σφάλματα.

ΣΤΡΕΒΛΩΣΗ(E) Η ΚΥΡΤΩΣΗ(U)	**ΣΥΜΠΕΡΑΣΜΑΤΑ**
$-2SE(E) < E < 2SE(E)$	Συμμετρία
$E \leq -2SE(E)$	Αρνητική ασυμμετρία
$E \geq 2SE(E)$	Θετική ασυμμετρία
$-2SE(U) < U < 2SE(U)$	Όχι κύρτωση
$U \leq -2SE(U)$	Αρνητική κύρτωση
$U \geq 2SE(U)$	Θετική κύρτωση

Για μια θετική ασύμμετρη κατανομή, οι τιμές πολύ πάνω από την κορυφή είναι πιο συχνές, από τις τιμές πολύ παρακάτω και το αντίστροφο ισχύει για μια αρνητική ασύμμετρη κατανομή. Όταν ένα δείγμα (ή μία κατανομή) έχει θετική κύρτωση, στη συνέχεια συγκρίνεται με μια κατανομή Gauss (κανονική) με την ίδια διακύμανση ή τυπική απόκλιση και συμπεραίνουμε ότι το σχήμα του ιστογράμματος κορυφώνεται στη μέση τιμή (περισσότερες τιμές κοντά στο μέσο όρο) αλλά με πιο παχιές ουρές. Για μια αρνητική κύρτωση, η κορυφή περιγράφεται μερικές φορές ότι είναι πιο πλατιά (λιγότερο αιχμηρή) από το σχήμα κανονικής κατανομής και οι ουρές είναι λεπτότερες, έτσι ώστε να είναι λιγότερο πιθανό η εμφάνιση ακραίων τιμών.

Εικόνα 1: Κύρτωση κατανομών



2.1 ΓΡΑΦΙΚΗ ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΙΑΣ ΜΕΤΑΒΛΗΤΗΣ (ΜΟΝΟΠΑΡΑΓΟΝΤΙΚΗ)

Εάν εστιάζουμε σε δεδομένα από τις παρατηρήσεις μιας μεταβλητής, σε ένα δείγμα μεγέθους n, στη συνέχεια εκτός από την εξέταση των διαφόρων στατιστικών μέτρων (θέσης, διασποράς) πρέπει επίσης να εξετάσουμε γραφικά την κατανομή του δείγματος. Οι γραφικές και οι μη γραφικές μέθοδοι αλληλοσυμπληρώνονται. Ενώ οι μη γραφικές μέθοδοι είναι ποσοτικές και αντικειμενικές, δεν δίνουν μια πλήρη εικόνα των δεδομένων. Επομένως, απαιτούνται επίσης γραφικές μέθοδοι, οι οποίες είναι πιο ποιοτικές και περιλαμβάνουν έναν βαθμό υποκειμενικής ανάλυσης.

Ιστογράμματα

Η μόνη από αυτές τις τεχνικές που έχει νόημα για κατηγορικά δεδομένα είναι το ιστόγραμμα (βασικά απλώς ένα ραβδόγραμμα του πίνακα των δεδομένων). Το κυκλικό διάγραμμα είναι ισοδύναμο, αλλά δεν χρησιμοποιείται συχνά. Οι έννοιες της κεντρικής τάσης, της διασποράς και της ασυμμετρίας δεν έχουν νόημα για ονομαστικά κατηγορικά δεδομένα. Για διατάξιμα κατηγορικά δεδομένα, μερικές φορές έχει νόημα να τα

αντιμετωπίζουμε ως ποσοτικά για σκοπούς της διερευνητικής ανάλυσης δεδομένων. Εσείς πρέπει να χρησιμοποιήσετε την κρίση σας εδώ.

Το πιο βασικό γράφημα είναι το ιστόγραμμα, το οποίο είναι ένα ραβδόγραμμα στο οποίο κάθε ράβδος (καλείται κλάση) αντιπροσωπεύει τη συχνότητα (μέτρηση) ή την αναλογία (μέτρηση/ συνολικός αριθμός) των περιπτώσεων για ένα εύρος τιμών. Σε σύστημα κάθετων αξόνων, οι κλάσεις στον ένα άξονα και το πλήθος (ή η αναλογία) στον άλλο κάθετο άξονα. Για να δημιουργήσετε ένα ιστόγραμμα, καθορίστε το εύρος των δεδομένων για κάθε κλάση, μετρήστε πόσες περιπτώσεις εμπίπτουν σε κάθε κλάση και σχεδιάστε τις μπάρες αρκετά υψηλές για να υποδείξουν τον αριθμό. Γενικά, τοποθετούνται τιμές που βρίσκονται ακριβώς στο όριο μεταξύ δύο κλάσεων στην προηγούμενη (πιο χαμηλή)κλάση, αλλά αυτός ο κανόνας δεν τηρείται πάντα.

Τα ιστογράμματα είναι ένας από τους καλύτερους τρόπους για να μάθετε γρήγορα πολλά για τα δεδομένα σας, συμπεριλαμβανομένης της κεντρικής τάσης, της διασποράς, της επικρατούσας τιμής, του σχήματος και των ακραίων τιμών.

Φυλλογράφημα (stem and leaf plots)

Ένα απλό υποκατάστατο ενός ιστογράμματος είναι το φυλλογράφημα ή διάγραμμα μίσχου (stem)-φύλλου(leaf). Μερικές φορές είναι πιο εύκολο να γίνει από ένα ιστόγραμμα χωρίς να έχει την τάση για απόκρυψη οποιασδήποτε πληροφορίας. Μας δείχνει όλες τις τιμές δεδομένων και το σχήμα της κατανομής. Αφού διατάξουμε κατά αύξουσα σειρά τις παρατηρήσεις ενός δείγματος μπορούμε να εντοπίζουμε στατιστικά μέτρα ή στοιχεία όπως τη διάμεσο, το πρώτο και το τρίτο τεταρτημόριο, τη μέγιστη και την ελάχιστη τιμή των παρατηρήσεων. Ωστόσο, ένα ιστόγραμμα θεωρείται γενικά καλύτερο για την εκτίμηση μιας κατανομής δείγματος από το φυλλογράφημα.

Θηκόγραμμα (Boxplots)

Μια άλλη πολύ χρήσιμη γραφική τεχνική μίας μεταβλητής είναι το θηκόγραμμα. Μπορεί να έχει κάθετη ή οριζόντια μορφή και αντιπροσωπεύει τα δεδομένα ενός δείγματος. Τα θηκογράμματα είναι πολύ καλά στην παρουσίαση πληροφοριών σχετικά

με την κεντρική τάση, τη συμμετρία και την στρέβλωση, καθώς και τις ακραίες τιμές, αν και μπορεί να αποπροσανατολίσουν (πολυκόρυφες κατανομές).

Από τις καλύτερες χρήσεις είναι η παρουσίαση θηκογραμμάτων δίπλα-δίπλα (πολυμεταβλητή γραφική ανάλυση).Το θηκόγραμμα αποτελείται από ένα ορθογώνιο κουτί που οριοθετείται πάνω και κάτω από τιμές που αντιπροσωπεύουν τα τεταρτημόρια Q_3 και Q_1 αντίστοιχα. Μία οριζόντια γραμμή μέσα στο κουτί αντιστοιχεί στη διάμεσο. Μπορούμε επίσης να δούμε τον πάνω και τον κάτω “φράχτη” και τις ακραίες τιμές. Ο κάθετος άξονας αντιστοιχεί στις μονάδες της ποσοτικής μεταβλητής. Η ερμηνεία των φραχτών και των ακραίων τιμών είναι λίγο πιο περίπλοκη. Οποιαδήποτε τιμή από τα δεδομένα υπερβαίνει το 1,5 IQR (ενδοτεταρτημοριακό εύρος) πέραν των δύο οριζόντιων πλευρών του ορθογωνίου (δηλαδή $> Q_3+1,5IQR$ ή $< Q_1-1,5IQR$) και στις δύο κατευθύνσεις θεωρείται ακραία τιμή και σχεδιάζεται ξεχωριστά. Μερικές φορές αν υπερβαίνει το 3,0 IQR θεωρείται «πολύ ακραία τιμή» και σχεδιάζεται με ένα διαφορετικό σύμβολο. Ο όρος "ακραίες τιμές" δεν ορίζεται καλά στα στατιστικά στοιχεία και ο ορισμός ποικίλλει ανάλογα με το σκοπό και την κατάσταση. Έτσι οι «ακραίες τιμές» που προσδιορίζονται από το θηκόγραμμα, ορίζονται ως οποιοδήποτε σημείο περισσότερο από 1,5 IQR από το Q_3 ή περισσότερο από 1,5 IQR κάτω από το Q_1.

Το θηκόγραμμα είναι μια διερευνητική τεχνική και θα πρέπει να θεωρήσετε τον χαρακτηρισμό ως “θηκόγραμμα ακραίων τιμών” ως μία απλή πρόταση όπου βλέπουμε ότι κάποιες τιμές μπορεί να ασυνήθιστες ή λάθη. Είναι επίσης σημαντικό να συνειδητοποιήσουμε ότι ο αριθμός των ακραίων τιμών εξαρτάται σε μεγάλο βαθμό από το μέγεθος του δείγματος. Στην πραγματικότητα, για δεδομένα κανονικής κατανομής, περιμένουμε 0,70% (ή περίπου 1 σε 150 περιπτώσεις) να είναι ακραίες τιμές, με περίπου το ήμισυ σε κάθε κατεύθυνση. Ένα θηκόγραμμα με πολλές ακραίες τιμές μας οδηγεί στο συμπέρασμα “παχιές ουρές” (θετική κύρτωση), ή πιθανώς πολλά σφάλματα στην εισαγωγή δεδομένων. Επίσης, χαμηλές τιμές στους “φράχτες” δείχνουν αρνητική κύρτωση, τουλάχιστον εάν το μέγεθος του δείγματος είναι μεγάλο. Η συμμετρία εκτιμάται παρατηρώντας εάν η διάμεσος είναι στο κέντρο του κουτιού και αν ο άνω και ο κάτω φράχτης ισαπέχουν από τα Q_3 και Q_1 αντίστοιχα. Σε μια ασύμμετρη κατανομή

περιμένουμε τη διάμεσο να ωθείται προς την κατεύθυνση της μικρότερης απόστασης από τις δύο που αναφέρθηκαν παραπάνω.

Τα θηκογράμματα είναι εξαιρετικές γραφικές παραστάσεις διερευνητικής ανάλυσης δεδομένων επειδή βασίζονται σε ισχυρά στατιστικά μέτρα όπως η διάμεσος και το ενδοτεταρτημοριακό εύρος. Με τα θηκογράμματα είναι εύκολο να συγκρίνουμε κατανομές με υψηλό βαθμό αξιοπιστίας.

Ποσοτικά-Κανονικά Γραφήματα

Αυτή η γραφική διερευνητική ανάλυση δεδομένων είναι η πιο περίπλοκη. Ονομάζεται ποσοτικό-κανονικό γράφημα ή QN ή γενικότερα QQ plot. Χρησιμοποιείται για να δούμε πόσο καλά ακολουθεί ένα συγκεκριμένο δείγμα θεωρητική κατανομή. Αν και μπορεί να χρησιμοποιηθεί για οποιαδήποτε θεωρητική κατανομή, θα περιορίσουμε την προσοχή μας στο να δούμε πόσο καλά ταιριάζει ένα δείγμα δεδομένων μεγέθους n, με μία κανονική κατανομή ίσης μέσης τιμής και διακύμανσης. Μπορούμε να ανιχνεύσουμε αριστερά ή δεξιά ασυμμετρία, θετική ή αρνητική κύρτωση και αν έχουμε δύο κορυφές. Δεν πρέπει να το συγχέουμε με διάγραμμα π.χ. διασποράς δύο μεταβλητών. Πολλές στατιστικές δοκιμές έχουν την υπόθεση ότι το αποτέλεσμα για οποιοδήποτε σταθερό σύνολο των τιμών των επεξηγηματικών μεταβλητών κατανέμεται κανονικά, και αυτός είναι ο λόγος για τον οποίο τα γραφήματα QN είναι χρήσιμα: εάν η υπόθεση παραβιάζεται κατά πολύ, η τιμή p και τα διαστήματα εμπιστοσύνης αυτών των δοκιμών είναι λανθασμένα.

Συμπεραίνοντας τα γραφήματα αυτά επιτρέπουν την ανίχνευση της μη κανονικότητας και τη διάγνωση της ασυμμετρίας και της κύρτωσης.

2.2 ΜΗ ΓΡΑΦΙΚΗ ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

Δείχνει γενικά τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με τη μορφή είτε διασταυρούμενων πινάκων είτε άλλων στατιστικών στοιχείων.

Πίνακας διασταύρωσης

Για κατηγορικά δεδομένα (και ποσοτικά δεδομένα με λίγες μόνο διαφορετικές τιμές) η επέκταση του πίνακα, που ονομάζεται πίνακας διασταύρωσης, είναι πολύ χρήσιμη. Για δύο μεταβλητές, η διασταύρωση γίνεται με έναν πίνακα διπλής εισόδου με τα επίπεδα της πρώτης μεταβλητής στις στήλες του πίνακα και τα επίπεδα της δεύτερης μεταβλητής στις γραμμές του πίνακα. Έχουμε συνδυασμό των τιμών της μίας μεταβλητής με τις τιμές της άλλης και έτσι μπορούμε να μελετήσουμε τη σχέση μεταξύ των μεταβλητών. Ο πίνακας διασταύρωσης μπορεί να επεκταθεί σε τρεις (και μερικές φορές περισσότερες) μεταβλητές ως δημιουργία ξεχωριστών αμφίδρομων πινάκων για δύο μεταβλητές σε κάθε επίπεδο μιας τρίτης μεταβλητής.

Ανάλυση μιας μεταβλητής

Για μία κατηγορική μεταβλητή (συνήθως επεξηγηματική) και μία ποσοτική μεταβλητή (με τιμές εξόδου) είναι σύνηθες να βρίσκουμε μερικά από τα μη γραφικά στατιστικά στοιχεία για την ποσοτική μεταβλητή, ξεχωριστά για κάθε τιμή της κατηγορικής μεταβλητής και στη συνέχεια γίνεται σύγκριση των στατιστικών στοιχείων. Η σύγκριση των μέσων τιμών είναι μια ανεπίσημη εκδοχή ANOVA (ανάλυση διασποράς). Συγκρίνοντας τις διαμέσους έχουμε μια ισχυρή εκδοχή της ανάλυσης. Συγκρίνοντας μέτρα της διασποράς έχουμε ένα καλό τεστ, μια καλή άτυπη δοκιμασία για την υπόθεση ίσων διακυμάνσεων που απαιτούνται για έγκυρη ανάλυση της διακύμανσης. Γίνεται έλεγχος ύπαρξης διαφορών. Η ύπαρξη έστω και μιας διαφοράς δείχνει ότι ο παράγοντας (η κατηγορική μεταβλητή) επηρεάζει σημαντικά την ποσοτική μεταβλητή.

Συσχέτιση και συνδιακύμανση

Για δύο ποσοτικές μεταβλητές και τη σχέση μεταξύ τους, βασικό στατιστικό ενδιαφέρον έχει η συνδιακύμανση δείγματος ή / και η συσχέτιση δείγματος. Η συνδιακύμανση είναι ένα μέτρο για να δούμε για δύο μεταβλητές, πόσο (και σε τι κατεύθυνση) η μία επηρεάζει την άλλη. Θα πρέπει να περιμένουμε μια μεταβλητή να αλλάξει, όταν η άλλη αλλάξει.

Η συνδιακύμανση του δείγματος υπολογίζεται με υπολογισμούς διαφορών κάθε μέτρησης από τον μέσο όρο όλων των μετρήσεων για κάθε μεταβλητή. Στη συνέχεια, κάθε διαφορά για κάθε τιμή της μίας μεταβλητής πολλαπλασιάζεται με την αντίστοιχη διαφορά της άλλης μεταβλητής. Τέλος, αυτές οι τιμές κατά μέσο όρο (στην πραγματικότητα αθροίζονται και διαιρούνται με n-1, για να διατηρηθεί η στατιστική αμεροληψία) δίνουν τη συνδιακύμανση. Θετική συνδιακύμανση υποδηλώνει ότι όταν μία τιμή της μίας μεταβλητής είναι πάνω από τη μέση τιμή σημαίνει ότι η τιμή της άλλης πιθανότατα θα είναι επίσης πάνω από το μέσο όρο και το αντίστροφο. Αρνητική συνδιακύμανση υποδηλώνει ότι όταν μια μεταβλητή είναι πάνω από τη μέση τιμή της, η άλλη μεταβλητή είναι κάτω από τη μέση τιμή της. Και συνδιακύμανση σχεδόν μηδέν υποδηλώνει ότι οι δύο μεταβλητές μεταβάλλονται ανεξάρτητα η μία από την άλλη. Τεχνικά, η ανεξαρτησία συνεπάγεται μηδενική συσχέτιση, αλλά το αντίστροφο δεν ισχύει απαραίτητα. Η συνδιακύμανση δύσκολα ερμηνεύεται επομένως χρησιμοποιούμε συχνά τη συσχέτιση αντ' αυτού. Η συσχέτιση έχει την ιδιότητα να είναι πάντα μεταξύ -1 και +1, με -1 είναι μια «τέλεια» αρνητική γραμμική συσχέτιση, +1 είναι μια τέλεια θετική γραμμική συσχέτιση και το 0 δείχνει ότι οι μεταβλητές είναι γραμμικά ασυσχέτιστες. Ο τύπος που δίνει τη συνδιακύμανση δύο μεταβλητών X και Y είναι:

$\text{Cov(X,Y)}=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})$

ενώ ο τύπος που δίνει τη συσχέτιση δείγματος:

$\text{Cor(X,Y)}=\frac{Cov(X,Y)}{s_x\cdot s_y}$

όπου s_x είναι η τυπική απόκλιση της μεταβλητής X και s_y η τυπική απόκλιση της Y.

Πίνακες συνδιακύμανσης και συσχέτισης

Όταν έχουμε πολλές ποσοτικές μεταβλητές, η πιο συνηθισμένη τεχνική, για μη γραφική διερευνητική ανάλυση δεδομένων, είναι να υπολογίσουμε όλες τις συνδιακυμάνσεις ή / και συσχετίσεις κατά ζεύγη και να τις συγκεντρώσουμε σε ένα πίνακα.

Πολυμεταβλητή γραφική διερευνητική ανάλυση δεδομένων

Υπάρχουν μερικές χρήσιμες τεχνικές για γραφική διερευνητική ανάλυση δεδομένων δύο τυχαίων κατηγορικών μεταβλητών. Συνήθως χρησιμοποιείται ένα ομαδοποιημένο ραβδόγραμμα με κάθε ομάδα να αντιπροσωπεύει ένα επίπεδο μιας από τις μεταβλητές και κάθε ράβδος (ορθογώνιο) μέσα σε μια ομάδα αντιπροσωπεύει τα επίπεδα της άλλης μεταβλητής.

Γραφικές παραστάσεις μιας μεταβλητής μέσω κατηγορικής μεταβλητής

Όταν έχουμε μία κατηγορική και μία ποσοτική (τιμές εξόδου) μεταβλητή, η γραφική διερευνητική ανάλυση δεδομένων χρησιμοποιεί συνήθως ως προϋπόθεση την κατηγορική τυχαία μεταβλητή. Αυτό δείχνει απλώς ότι εστιάζουμε σε όλα τα θέματα με μία συγκεκριμένη τιμή της κατηγορικής τυχαίας μεταβλητής και στη συνέχεια σχεδιάζουμε γραφικές παραστάσεις της ποσοτικής μεταβλητής για αυτά τα θέματα. Το επαναλαμβάνουμε για κάθε τιμή της κατηγορικής μεταβλητής και στη συνέχεια συγκρίνουμε τις γραφικές παραστάσεις.

Το γράφημα που χρησιμοποιείται περισσότερο είναι τα θηκογράμματα που βρίσκονται δίπλα-δίπλα. Τα διπλανά θηκογράματα είναι η καλύτερη τεχνική γραφικής διερευνητικής ανάλυσης δεδομένων για την εξέταση της σχέσης μεταξύ μιας κατηγορικής μεταβλητής και μιας ποσοτικής μεταβλητής, καθώς και την κατανομή της ποσοτικής μεταβλητής σε κάθε μία τιμή της κατηγορικής μεταβλητής.

Διάγραμμα διασποράς

Για δύο ποσοτικές μεταβλητές, η βασική γραφική τεχνική διερευνητική ανάλυση δεδομένων είναι το scatterplot, διάγραμμα διασποράς, το οποίο έχει μία μεταβλητή στον άξονα x, μία στον άξονα y και ένα σημείο για κάθε περίπτωση στο σύνολο δεδομένων σας. Εάν η μία μεταβλητή είναι επεξηγηματική και η άλλη είναι αποτέλεσμα, είναι μία πολύ, πολύ ισχυρή σύμβαση για να βάλουμε το αποτέλεσμα στον άξονα y (κάθετο). Μία ή δύο

επιπρόσθετες κατηγορικές μεταβλητές μπορούν να ενσωματωθούν στο διάγραμμα, κωδικοποιώντας τις πρόσθετες πληροφορίες.

2.3 ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

Κανονική κατανομή

Η κανονική κατανομή αφορά συνεχείς τυχαίες μεταβλητές και είναι η πιο σημαντική κατανομή πιθανότητας με εφαρμογές σε στατιστικές μεθόδους, όπως ο έλεγχος υποθέσεων. Οι τιμές της μεταβλητής που ακολουθεί την κανονική κατανομή είναι συγκεντρωμένες γύρω από τη μέση τιμή και με μικρή συχνότητα όταν είναι πολύ χαμηλές ή υψηλές. Η συνάρτηση πυκνότητας της κανονικής κατανομής είναι:

$f(x)=\frac{1}{\sigma\sqrt{2\pi}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, όπου μ η μέση τιμή και σ η τυπική απόκλιση.

Η κανονική κατανομή είναι μία κωδωνοειδής συμμετρική κατανομή. Συμμετρική γύρω από τη μέση τιμή. Η πιθανότητα μιας μεμονωμένης τιμής είναι μηδενική. Η μέση τιμή, η διάμεσος και η επικρατούσα τιμή συμπίπτουν. Το εμβαδόν της περιοχής που ορίζεται από την καμπύλη της κατανομής και τον οριζόντιο άξονα είναι αντιστοιχεί σε πιθανότητα p=1. Η πιθανότητα μιας μεμονωμένης τιμής της μεταβλητής τείνει στο μηδέν. Έτσι υπολογίζουμε πιθανότητα διαστήματος τιμών μέσω μετατροπής της κατανομής σε τυπική κανονική κατανομή με μ=0 και σ=1 ,και τυχαία μεταβλητή Z με τιμές:

$z_i=\frac{x_i-\mu}{\sigma}$

Υπάρχουν πίνακες που δίνουν πιθανότητες για τις διάφορες τιμές της Z. Τα διαστήματα που ορίζονται από $\pm1,\pm2,\pm3$ τυπικές αποκλίσεις εκατέρωθεν της μέσης τιμής έχουν πιθανότητα περίπου 68% , 95% , 99% αντίστοιχα.

Διάστημα εμπιστοσύνης

Αν έχουμε διαφορετικά δείγματα ίδιου πληθυσμού τότε θα έχουμε διαφορετικές δειγματικές μέσες τιμές, που δε γνωρίζουμε πόσο διαφέρουν από τη μέση τιμή μ. Έτσι, προσδιορίζουμε ένα διάστημα τιμών, το διάστημα εμπιστοσύνης. Η δειγματοληπτική κανονική κατανομή ακολουθεί την κανονική κατανομή με μέση τιμή $\mu_{\bar{X}}$= μ και διακύμανση $\sigma_{\bar{X}}^2$=$\frac{\sigma^2}{n}$

Μετατρέποντας τις δειγματικές μέσες τιμές σε τυπικές τιμές z, έχουμε την τυπική κανονική κατανομή της δειγματοληπτικής κατανομής της μέσης τιμής. Ισχύει ότι:

Z=$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, με $\bar{X}$ οι δειγματικές μέσες τιμές και P ($-1{,}96 \leqslant Z \leqslant 1{,}96$) = 0,95 οπότε έχουμε ότι το διάστημα ($\bar{X}$- 1,96$\frac{\sigma}{\sqrt{n}}$, $\bar{X}$+1,96$\frac{\sigma}{\sqrt{n}}$) ονομάζεται 95% διάστημα εμπιστοσύνης της μέσης τιμής του πληθυσμού. Μπορούμε να συμβολίσουμε με α=1-0,95=0,05 και $z_{\alpha/2}$= 1,96

Αν η τυπική απόκλιση είναι άγνωστη και το δείγμα μικρό τότε έχουμε τη μεταβλητή t που ακολουθεί την κατανομή student-t με n-1 βαθμούς ελευθερίας και όχι την κανονική κατανομή.

t =$\frac{\bar{X}-\mu}{s/\sqrt{n}}$, με s τυπική απόκλιση τυχαίου δείγματος.

Έλεγχοι υποθέσεων

Στους ελέγχους υποθέσεων προκύπτουν συμπεράσματα από τη διαμόρφωση και τον έλεγχο συγκεκριμένων υποθέσεων που αφορούν τις πληθυσμιακές παραμέτρους. Υπάρχουν δύο είδη στατιστικών υποθέσεων. Η μηδενική υπόθεση H_0, η οποία ελέγχεται και απορρίπτεται ή όχι και η εναλλακτική υπόθεση H_1. Συνήθως η μηδενική υπόθεση διατυπώνεται έτσι ώστε να είναι αντίθετη με αυτό που θεωρεί ο ερευνητής αληθινό.

Μονόπλευρος έλεγχος:

Αν H_0: $\mu_1=\mu_2$,, τότε H_1: $\mu_1>\mu_2$ ή $\mu_1<\mu_2$

Αμφίπλευρος έλεγχος:

Αν H_0: $\mu_1=\mu_2$, τότε H_1: $\mu_1 \neq \mu_2$

Σφάλμα τύπου Ι : Απόρριψη της μηδενικής υπόθεσης ενώ είναι αληθινή.

Η πιθανότητα αυτού του σφάλματος συμβολίζεται με α η τιμή του οποίου (συνήθως α=0,01 ή α=0,05) ονομάζεται επίπεδο σημαντικότητας

Σφάλμα τύπου ΙΙ : Μη απόρριψη της H_0 ενώ είναι λανθασμένη.

Η κατανομή που χρησιμοποιείται συχνότερα για τον έλεγχο υποθέσεων είναι η κανονική κατανομή.

Το κριτήρια ελέγχου που χρησιμοποιούνται στον έλεγχο υποθέσεων έχουν τη γενική μορφή:

κριτήριο ελέγχου $=\frac{στατιστική-παράμετρος}{τυπικόσφάλμα}$

Η στατιστική έχει υπολογιστεί από το δείγμα (για παράδειγμα η μέση τιμή). Έτσι όταν έχουμε έλεγχο υποθέσεων για τη μέση τιμή κανονικού πληθυσμού με γνωστή διασπορά το κριτήριο ελέγχου είναι:

$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

Ορίζουμε επίπεδο σημαντικότητας, βρίσκουμε την τιμή του z και στη συνέχεια την κρίσιμη τιμή και μετά τη σύγκριση απορρίπτουμε ή όχι τη μηδενική υπόθεση.

Έλεγχοι υποθέσεων για δύο δείγματα - έλεγχος t (t-test) για σύγκριση αριθμητικών μέσων

Όταν μας ενδιαφέρει να διαπιστώσουμε αν υπάρχει διαφορά στη συμπεριφορά ενός χαρακτηριστικού που μετράται σε δύο ανεξάρτητους πληθυσμούς, υπολογίζουμε τη διαφορά των αριθμητικών μέσων. Σχηματίζουμε τη σχετική δειγματοληπτική κατανομή και με το κατάλληλο κριτήριο ελέγχου συμπεραίνουμε αν οι δύο πληθυσμοί έχουν άνισους αριθμητικούς μέσους. Έστω Χ η μεταβλητή που εξετάζουμε και σε δύο πληθυσμούς έχουμε μέσες τιμές μ_1 και μ_2 αντίστοιχα και γνωστές διακυμάνσεις. Επιλέγουμε τυχαία δείγματα μεγέθους n_1 και n_2 και υπολογίζουμε τους αριθμητικούς μέσους σε αυτά. Δημιουργούμε την δειγματοληπτική κατανομή της διαφοράς των μέσων $\overline{X_1}-\overline{X_2}$ που ακολουθεί την κανονική κατανομή. Όταν όμως τα δείγματα είναι μικρά και δε γνωρίζουμε τις διακυμάνσεις χρησιμοποιείται η κατανομή t. Η μηδενική υπόθεση για τον έλεγχο διαφοράς των αριθμητικών μέσων είναι $H_0 : \mu_1=\mu_2$, ενώ η εναλλακτική είναι $H_1 : \mu_1 \neq \mu_2$. Όταν οι διακυμάνσεις είναι ίσες το κριτήριο ελέγχου είναι:

$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_p^2 \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}$, όπου s_p η σταθμισμένη διασπορά

Οι βαθμοί ελευθερίας της κατανομής t , είναι $df = n_1 + n_2 - 2$.

Ανάλυση διασποράς (ANOVA)

Με την μέθοδο αυτή εξετάζουμε πώς ένας ή περισσότεροι παράγοντες επιδρούν στη μεταβλητότητα ενός συνόλου δεδομένων, πώς επηρεάζεται μία μεταβλητή. Ελέγχεται η διαφορά των μέσων τιμών για περισσότερα από δύο δείγματα. Για τρεις ομάδες ανεξάρτητων παρατηρήσεων με ίση διασπορά οι οποίες ακολουθούν κανονική κατανομή, η μηδενική υπόθεση είναι ότι δεν υπάρχει διαφορά στις μέσες τιμές

$H_0 : \mu_1 = \mu_2 = \mu_3$ και η εναλλακτική υπόθεση

H_1 : Δύο τουλάχιστον μέσες τιμές διαφέρουν μεταξύ τους

Για τον έλεγχο της μηδενικής υπόθεσης υπολογίζεται η αναλογία της διακύμανσης μεταξύ των μέσων τιμών και της συνολικής διακύμανσης των επιμέρους ομάδων.

$F = \frac{διακύμανση μεταξύ των ομάδων}{διακύμανση εντός των ομάδων} = \frac{s_b^2}{s_w^2}$ (κριτήριο F)

Όταν ισχύει η μηδενική υπόθεση τότε F=1. Γίνεται σύγκριση του F με την κρίσιμη τιμή (υπάρχουν πίνακες) για δεδομένο επίπεδο σημαντικότητας και βαθμούς ελευθερίας

$df_1 = k-1$ για τη διακύμανση μεταξύ των ομάδων,

$df_2 = n-k$ για τη διακύμανση εντός των ομάδων, όπου n το σύνολο των παρατηρήσεων και k ο αριθμός των ομάδων. Η αναλογία ακολουθεί την κατανομή της πιθανότητας F, η καμπύλη της οποίας είναι ασύμμετρη.

Έλεγχος X^2

Είναι ένας στατιστικός έλεγχος μη παραμετρικός, μια και δεν αφορά εκτίμηση πληθυσμιακής παραμέτρου. Ο έλεγχος X^2 εξετάζει κατά πόσο δύο ποιοτικές μεταβλητές είναι ανεξάρτητες. Στον έλεγχο αυτό η μηδενική υπόθεση είναι ότι οι μεταβλητές είναι ανεξάρτητες, δεν υπάρχουν δηλαδή διαφορές μεταξύ των κατηγοριών και οι διαφορές που παρατηρούνται στα δεδομένα, οφείλονται σε τυχαίους παράγοντες που συνδέονται με τη δειγματοληψία. Εφαρμόζεται σε κατανομές συχνοτήτων, σε πίνακες

διασταυρώσεων. Οι τιμές των επιμέρους συχνοτήτων προσδιορίζονται δεδομένο ότι ισχύει η μηδενική υπόθεση (αναμενόμενες συχνότητες). Έτσι δημιουργείται ένας νέος πίνακας διασταυρώσεων.

Κριτήριο ελέγχου: $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_{ij} - E_{ij})^2 / E_{ij}$

όπου O_{ij} οι παρατηρούμενες συχνότητες, E_{ij} οι αναμενόμενες συχνότητες, r ο αριθμός γραμμών του πίνακα και c ο αριθμός στηλών του πίνακα. Γίνεται σύγκριση της τιμής με την κρίσιμη τιμή (υπάρχουν πίνακες) για δεδομένο επίπεδο σημαντικότητας και βαθμούς ελευθερίας df=(r-1)(c-1). Αν η τιμή του X^2 είναι μεγαλύτερη από την κρίσιμη τιμή, τότε η μηδενική υπόθεση απορρίπτεται.

3 ΑΛΓΟΡΙΘΜΟΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Στόχος της κατηγοριοποίησης και πρόβλεψης δεδομένων είναι η δημιουργία μοντέλου που να μπορεί να κατηγοριοποιεί τα δεδομένα, να προβλέπει την τιμή εξόδου για τα δεδομένα εισόδου, μέσω εκπαίδευσης αλγορίθμου. Τα δεδομένα εισόδου συνήθως είναι με μορφή διανυσμάτων. Η εκπαίδευση αλγορίθμου επιτυγχάνεται χρησιμοποιώντας τα δεδομένα μας τα οποία μετά τη συλλογή και τη μετέπειτα διαλογή τους επρόκειτο να υποβληθούν σε επεξεργασία.

Έτσι, μετά την εκπαίδευση του αλγορίθμου αναμένουμε την κατάταξη των δεδομένων σε κατηγορίες και προγνωστική ακρίβεια. Σε προβλήματα κατηγοριοποίησης και πρόβλεψης θέλουμε ο αλγόριθμος να προβλέπει με ακρίβεια και να γενικεύει καλά, σε νέα δεδομένα. Αλγόριθμοι, όπως Κ-κοντινότεροι γείτονες, λογιστική παλινδρόμηση, αλγόριθμος Random Forest, τεχνητά νευρωνικά δίκτυα παρουσιάζονται παρακάτω.

Κ-κοντινότεροι γείτονες

Αλγόριθμος μηχανικής μάθησης με δυνατότητα χρήσης σε προβλήματα κατάταξης αλλά και παλινδρόμησης. Μέσω του αλγορίθμου μπορούμε να κατηγοριοποιήσουμε κάθε νέο στοιχείο, με βάση τους Κ-κοντινότερους γείτονές του. Οι κοντινότεροι γείτονες υπολογίζονται συνήθως με βάση την ευκλείδεια απόστασή τους. Έτσι, αν x, x΄σημεία n-διάστατου ευκλείδιου χώρου, τότε η απόστασή τους είναι:

$d(x,x') = \sqrt{\sum_{i=1}^{n} (x_i - x_i)^2}$

Φυσικά υπάρχουν και άλλα μέτρα για τη μέτρηση της απόστασης-εκτός της ευκλείδειας-όπως Manhattan, Chebychev, Hamming, Mahalanobis distance. Σε προβλήματα παλινδρόμησης όπου τίθεται θέμα πρόβλεψης και όχι κατάταξης, το αποτέλεσμα της πρόβλεψης, με τη μέθοδο αυτή, είναι ο μέσος όρος των Κ γειτονικών σημείων. Η επιλογή του Κ είναι σημαντική καθώς ανάλογα με την τιμή του, διαφοροποιείται η ταξινόμηση κάθε νέου στοιχείου, όπως και η ποιότητα πρόβλεψης αν αναφερόμαστε σε προβλήματα παλινδρόμησης.

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση χρησιμοποιείται σε δυαδικά προβλήματα ταξινόμησης. Είναι μία προγνωστική ανάλυση όπου οι τιμές εξόδου δεν είναι συνεχείς αλλά διακριτές (δυαδικές). Ενώ είναι μη γραμμική μέθοδος μέσω της σιγμοειδούς συνάρτησης προσεγγίζεται η γραμμικότητα. Η σιγμοειδής συνάρτηση έχει μορφή S , όπως υποδηλώνει η ονομασία της και έχει τύπο:

$$f(x)=\frac{1}{1+e^{-x}}$$

Μπορούμε από τα δεδομένα μας να αφαιρούμε τις ακραίες τιμές και τους υψηλούς συσχετισμούς και να γίνεται χρήση τεχνικών και αλγορίθμων με στόχο τη βέλτιστη πρόβλεψη.

Τυχαία Δάση

Ένα Τυχαίο Δάσος (Random Forest) είναι ένας αλγόριθμος εκτίμησης που ταιριάζει έναν αριθμό ταξινομημένων Δέντρων Αποφάσεων (Decision Trees) σε διάφορα υποδείγματα του συνόλου δεδομένων και χρησιμοποιεί το μέσο όρο για τη βελτίωση της προγνωστικής ακρίβειας και του ελέγχου της υπερβολικής προσαρμογής (over-fitting). Ένα Δέντρο Απόφασης είναι ένας ταξινομητής που εκφράζεται ως αναδρομική διχοτόμηση του χώρου γεγονότων. Αποτελείται από κόμβους που σχηματίζουν ένα δέντρο με ρίζα, το οποίο σημαίνει ότι είναι ένα δέντρο με κατεύθυνση ένα κόμβο, ο οποίος δεν έχει καμία εισερχόμενη άκρη. Όλοι οι άλλοι κόμβοι έχουν ακριβώς μία εισερχόμενη ακμή. Ένας κόμβος με εξερχόμενη άκρη αναφέρεται ως ένας κόμβος «εσωτερικός» ή «εξέτασης». Όλοι οι άλλοι κόμβοι ονομάζονται "φύλλα" (επίσης γνωστοί ως «τερματικοί κόμβοι» ή «κόμβοι απόφασης»). Τα τυχαία δάση είναι μια παραμετρική μέθοδος. Οι παράμετροι που συνήθως δοκιμάζουμε να ρυθμίσουμε είναι οι εξής:

- **n_estimators:** Αριθμός των δέντρων στο δάσος (default = 100)
- **min_samples_split:** Ελάχιστος αριθμός δειγμάτων που απαιτείται για να διαχωριστεί ένας κόμβος (default = 2)

- **min_samples_leaf:** Ελάχιστος αριθμός δειγμάτων που απαιτείται να υπάρχει σε ένα φύλλο-κόμβο (default = 1)
- **max_depth:** Μέγιστος αριθμός επιπέδων σε κάθε δέντρο αποφάσεων (default = None)

Τεχνητά Νευρωνικά Δίκτυα

Ένα νευρωνικό δίκτυο βασίζεται στον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου: πολλοί νευρώνες που συνδέονται με άλλους νευρώνες περνούν πληροφορίες μέσω των συνδέσεών τους και ενεργοποιούνται όταν το ηλεκτρικό σήμα που εισέρχεται σε αυτούς υπερβαίνει ένα ορισμένο κατώφλι. Ένα τεχνητό νευρωνικό δίκτυο έχει την ίδια δομή, χωρίς βέβαια να έχει φυσική υπόσταση και αποτελείται από τεχνητούς νευρώνες και συνάψεις με πληροφορίες που μεταδίδονται μεταξύ τους. Οι συνάψεις ή οι συνδέσεις, θα σταθμίζονται, δηλαδή θα έχουν κάποιες τιμές/βάρη σύμφωνα με το πόσο έντονα επηρεάζει ο νευρώνας τον προσδιορισμό της εξόδου. Τα βάρη περνούν από μια διαδικασία βελτιστοποίησης που ονομάζεται backpropagation. Για κάθε επανάληψη κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, το backpropagation θα χρησιμοποιηθεί για να επιστρέψει προς τα πίσω, μέσα από κάθε επίπεδο (layer) νευρώνων, του δικτύου και προσαρμόζει τα βάρη ανάλογα με τη συμβολή τους στο σφάλμα του νευρωνικού δικτύου. Τα νευρωνικά δίκτυα είναι ουσιαστικά αυτό-βελτιστοποιημένες συναρτήσεις που χαρτογραφούν τις εισόδους σε σωστές εξόδους. Στη συνέχεια αφότου έχουν εκπαιδευθεί με πολλά δεδομένα, μπορούν ενώ παίρνουν μια νέα είσοδο, να προβλέπουν μια έξοδο με βάση τη συνάρτηση που θα έχει δημιουργηθεί με τα δεδομένα εκπαίδευσης, αλλά και την προσαρμογή των βαρών με το backpropagation.

4 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΑΝΑΛΥΣΗ

Αρχικά καλούμε τις επιπλέον βιβλιοθήκες τις οποίες θα χρησιμοποιήσουμε στην ανάλυση μας. Μέσω των συναρτήσεων της βιβλιοθήκης “ggplot2” (Wickham, 2016) παράγουμε τις περισσότερες οπτικοποιήσεις ενώ η βιβλιοθήκη “dplyr” (Wickham et al., 2020) παρέχει πλήθος συναρτήσεων για την αποδοτική διαχείριση δομών δεδομένων τύπου “data frame”. Η βιβλιοθήκη “corrr” (Kuhn, Jackson, and Cimentada, 2020) παρέχει βοηθητικές συναρτήσεις για τον υπολογισμό και την οπτικοποίηση της συσχέτισης μεταξύ αριθμητικών μεταβλητών.

Ανοίγουμε το αρχείο “student-por.csv”, το οποίο ήδη έχουμε αντιγράψει στο “working directory” που έχουμε επιλέξει να δουλέψουμε.

```
data <- read.csv("student-por.csv")
```

Εξετάζουμε τη δομή του dataset, βλέποντας τις πρώτες γραμμές του

```
head(data)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
```

```
## 4    yes   yes    yes    yes    3     2   2   1   1    5
## 5    yes   yes     no     no    4     3   2   1   2    5
## 6    yes   yes    yes     no    5     4   2   1   2    5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
## 6        6 12 12 13
```

Ας εξετάσουμε αν υπάρχουν κενές τιμές/τιμές που λείπουν (missing data) στο σύνολο δεδομένων μας

```
length(which(is.na(data)))
```

```
## [1] 0
```

```
data <- data[complete.cases(data),]
dim(data)
```

```
## [1] 649  33
```

Αφού συμπεράναμε ότι δεν υπάρχουν τιμές που λείπουν συνεχίζουμε παρατηρώντας τα ονόματα των μεταβλητών που έχουμε στη διάθεση μας.

```
print(names(data))
```

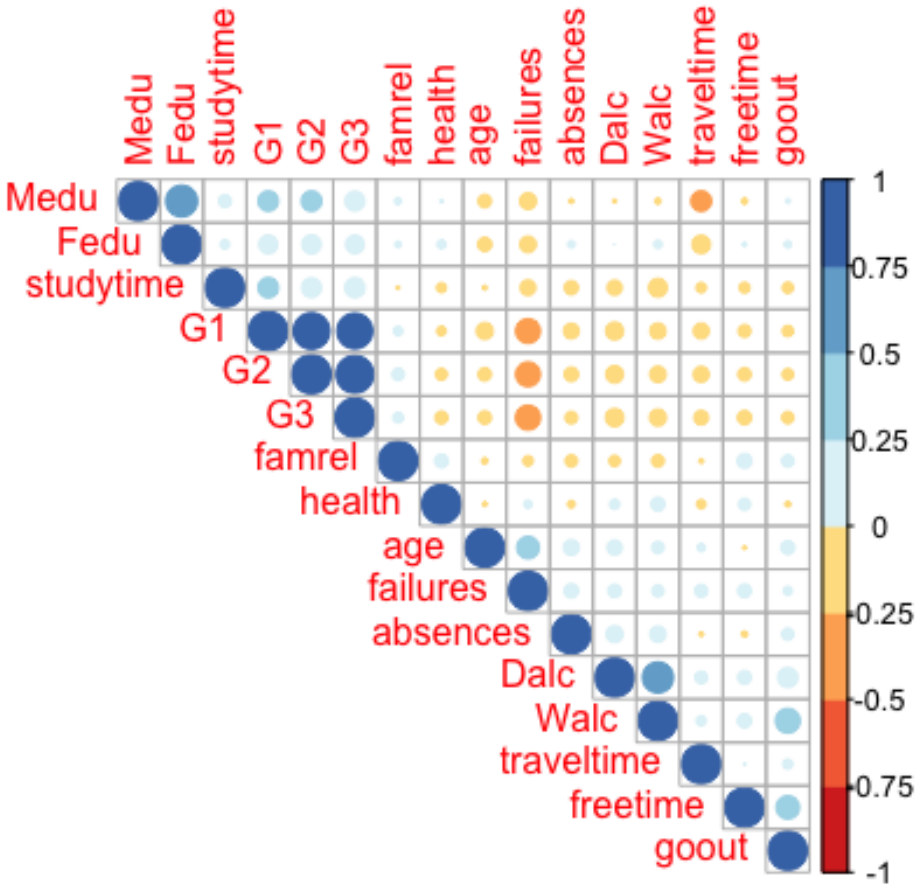
```
##  [1] "school"    "sex"       "age"       "address"   "famsize"
##  [6] "Pstatus"   "Medu"      "Fedu"      "Mjob"      "Fjob"
## [11] "reason"    "guardian"  "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"    "paid"      "activities" "nursery"
## [21] "higher"    "internet"  "romantic"  "famrel"    "freetime"
## [26] "goout"     "Dalc"      "Walc"      "health"    "absences"
## [31] "G1"        "G2"        "G3"
```

Οπτικοποίηση για την διερευνητική ανάλυση δεδομένων

Ένας πολύ καλός τρόπος για να οπτικοποιήσουμε το πώς συσχετίζονται οι (αριθμητικές) μεταβλητές στα δεδομένα μας, είναι η κατασκευή θερμικού χάρτη (heat map), δηλαδή μιας δισδιάστατης απεικόνισης που κάνοντας χρήση μεταβαλλόμενης απόχρωσης αναδεικνύει το πόσο αυξάνεται η κάθε μεταβλητή με την αύξηση οποιασδήποτε άλλης μεταβλητής. Ο θερμικός χάρτης που θα χρησιμοποιήσουμε αφορά τον βαθμό συσχέτισης για κάθε ζευγάρι αριθμητικών μεταβλητών από το σύνολο δεδομένων μας.

```
# συνάρτηση από την βιβλιοθήκη "dplyr"
datanum <- select_if(data, is.numeric)

# συνάρτηση από την βιβλιοθήκη "dplyr"
datanum <- select_if(data, is.numeric)

# συναρτήσεις από την βιβλιοθήκη "corrplot"
library(corrplot)

## corrplot 0.89 loaded

library(RColorBrewer)
M <-cor(datanum)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```

Εικόνα 2 : Διάγραμμα συσχετίσεων για τις αριθμητικές μεταβλητές



Προκειμένου να υπολογίσουμε συσχετίσεις ανάμεσα σε δύο κατηγορικές μεταβλητές θα χρησιμοποιήσουμε το X^2 στατιστικό τεστ, ενώ για τον υπολογισμό συσχέτισης ανάμεσα σε κατηγορική και αριθμητική μεταβλητή θα χρησιμοποιήσουμε ANOVA τεστ. Για την συσχέτιση ανάμεσα σε αριθμητικές μεταβλητές υπολογίζουμε την συσχέτιση Pearson. Στον κώδικα που ακολουθεί καλούμε την συνάρτηση “mixed_assoc” που παρέχεται από τη συλλογή συναρτήσεων (“Holgerbrandl/Datautils: Small Utilities to Make R-Scripting More Fun,” n.d.) η οποία αξιοποιεί επιπλέον συναρτήσεις από τις βιβλιοθήκες “tidyverse” () και “rcompanion” (). Με όρισμα το σύνολο δεδομένων μας σε μορφή “data frame” υπολογίζουμε και κανονικοποιούμε όλες τις συσχετίσεις για κάθε ζευγάρι μεταβλητών. Δείγμα του αποτελέσματος όσο αφορά την συσχέτιση και την μέθοδο που χρησιμοποιήθηκε παίρνουμε στον ακόλουθο πίνακα.

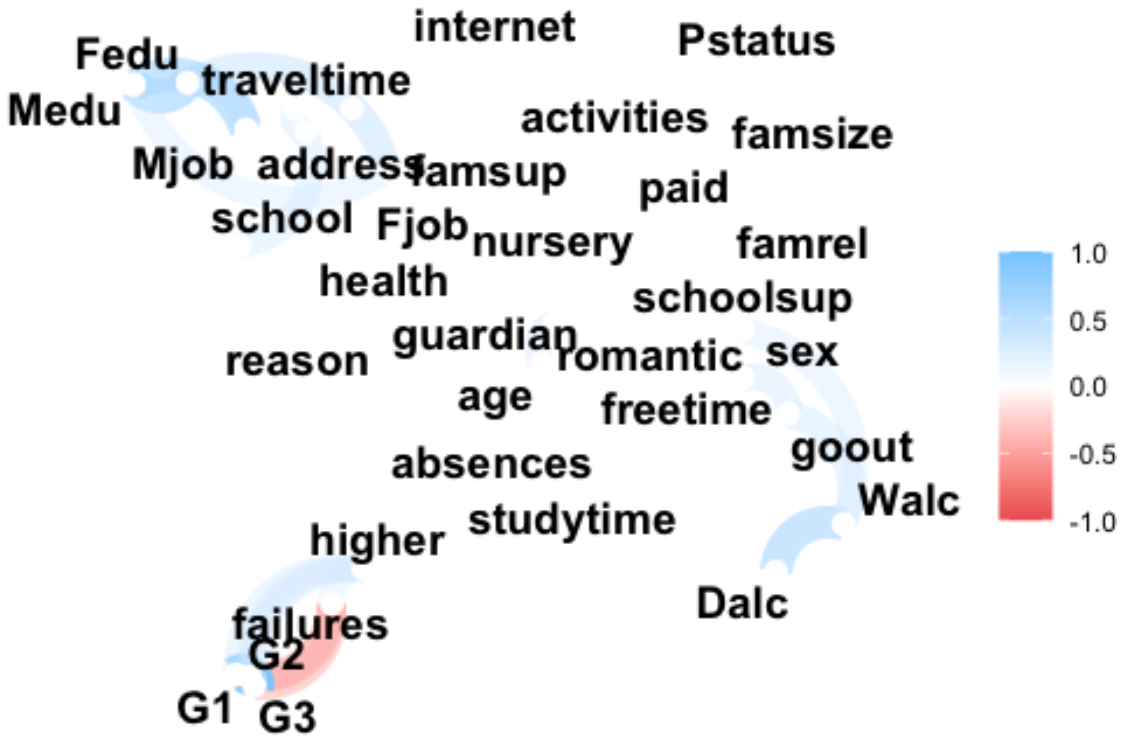
```
source("association.R")
```

```
## Loading required package: tidyverse

res_assoc <- mixed_assoc(data)
head(res_assoc)

##       x      y      assoc     type complete_obs_pairs complete_obs_ratio
## 1  school school 1.00000000 cramersV             649                 1
## 2     sex school 0.07323000 cramersV             649                 1
## 3     age school 0.08716968    anova             649                 1
## 4 address school 0.35260000 cramersV             649                 1
## 5 famsize school 0.00000000 cramersV             649                 1
## 6 Pstatus school 0.00000000 cramersV             649                 1
```

Στη συνέχεια παράγουμε μία ακόμη οπτικοποίηση με την χρήση ενός δικτύου. Κόμβοι στο δίκτυο είναι οι διαθέσιμες μεταβλητές, ενώ η απόχρωση των ακμών που ενώνει κάποιους από αυτούς προκύπτει από την αντίστοιχη τιμή της υπολογισμένης συσχέτισης.

```
data %>%
# select(- name) %>%
mixed_assoc() %>%
select(x, y, assoc) %>%
spread(y, assoc) %>%
column_to_rownames("x") %>%
as.matrix %>%
as_cordf %>%
network_plot()
```

Εικόνα 3 : Δίκτυο συσχετίσεων για όλες τις μεταβλητές



Με μια πρώτη ματιά, παρατηρούμε πως υψηλή συσχέτιση παρουσιάζουν οι βαθμοί μεταξύ τους (G1, G2 και G3). Αρνητική συσχέτιση παρουσιάζει το πλήθος μαθημάτων που αποτυγχάνει ο μαθητής (failures) με τις διαθέσιμες βαθμολογίες (G1, G2, G3). Επίσης παρατηρούμε και κάποιες θετικές συσχετίσεις που θα περιμέναμε να δούμε, όπως για παράδειγμα αυτή μεταξύ της κατανάλωσης αλκοόλ (Walc/ Dalc) και των εξόδων με φίλους (goout). Αξιοσημείωτο είναι το γεγονός ότι παρουσιάζεται χαμηλή συσχέτιση μεταξύ του χρόνου μελέτης και των βαθμών και στα 3 μαθήματα.

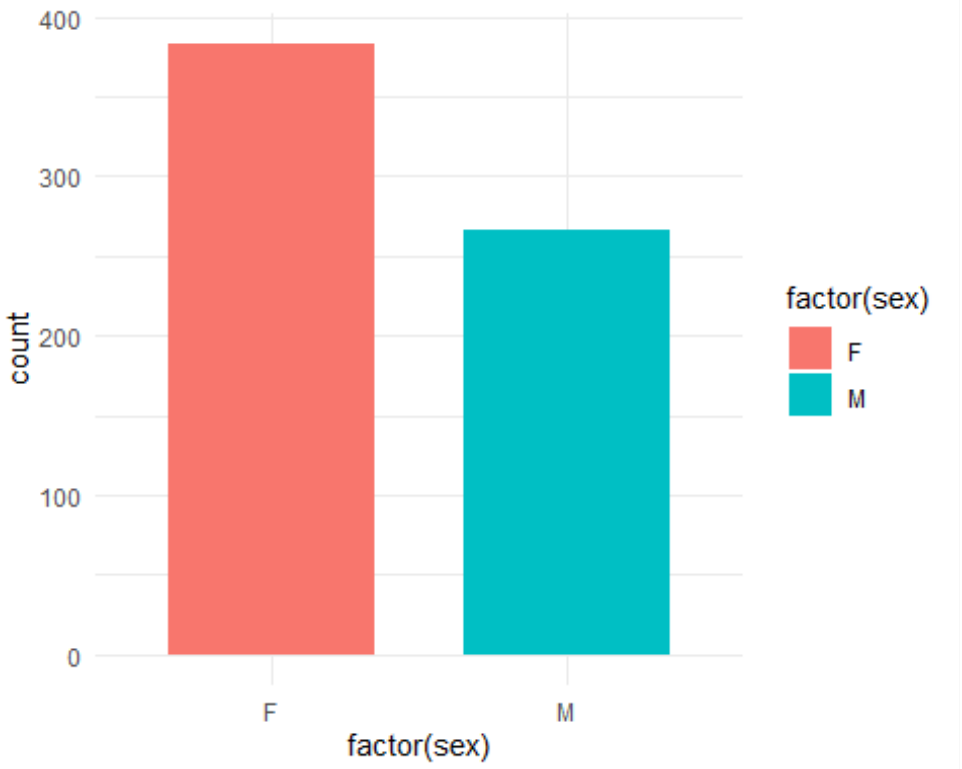
Οπτικοποίηση με Ραβδογράμματα

Ένα άλλο βήμα διερευνητικής ανάλυσης των δεδομένων που μπορούμε να κάνουμε είναι να κατασκευάσουμε ένα ραβδόγραμμα (barplot) ή ένα κυκλικό διάγραμμα (pie-chart), ώστε να δούμε με τι συχνότητα εμφανίζεται η κάθε δυνατή τιμή μιας κατηγορικής μεταβλητής στα δεδομένα μας. Για παράδειγμα, ας συγκρίνουμε τον αριθμό

των κοριτσιών (“F”) και αγοριών (“M”) που εμφανίζονται στη μεταβλητή “sex”, φτιάχνοντας ένα ιστόγραμμα.

```
ggplot(data, aes(x=factor(sex), fill=factor(sex) ))+
geom_bar(stat="count", width=0.7)+
theme_minimal()
```

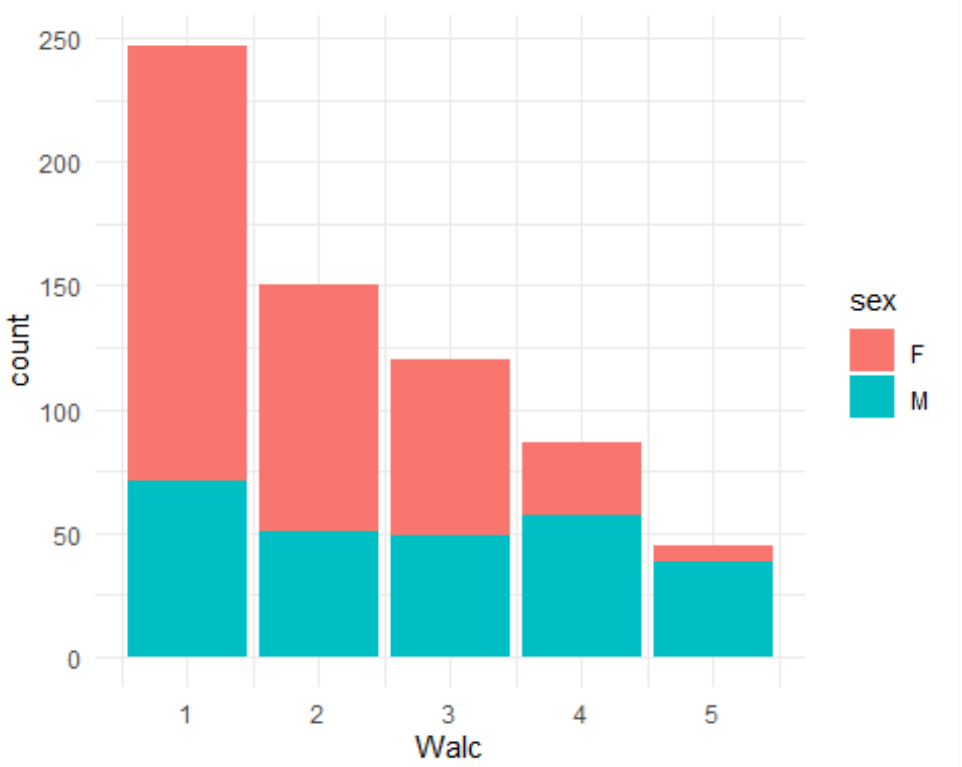
Εικόνα 4: Ιστόγραμμα Φύλων



Ακόμη μεγαλύτερο ενδιαφέρον μπορούμε να βρούμε σε ραβδογράμματα με τα οποία μπορούμε να μελετήσουμε την συσχέτιση μεταξύ δύο κατηγορικών μεταβλητών. Για παράδειγμα, στην επόμενη εικόνα παρατηρούμε την σχέση της κατηγορικής μεταβλητής “Walc” που προσδιορίζει την κατανάλωση αλκοόλ με το φύλο. Παρατηρούμε ότι οι περισσότεροι μαθητές καταναλώνουν την ελάχιστη ποσότητα αλκοόλ και όσο αυξάνεται η ποσότητα αλκοόλ το πλήθος μαθητών μειώνεται. Παρόλα αυτά αν παρατηρήσουμε πιο προσεχτικά την σχέση του φύλου βλέπουμε ότι αυτή η τάση ισχύει κυρίως για τα κορίτσια. Αντίθετα το πλήθος αγοριών παραμένει περίπου ίδιο για τις διαφορετικές ποσότητες αλκοόλ.

```
ggplot(data=data, aes(x=Walc,  fill=sex)) +
geom_bar(stat="count")+
theme_minimal()
```

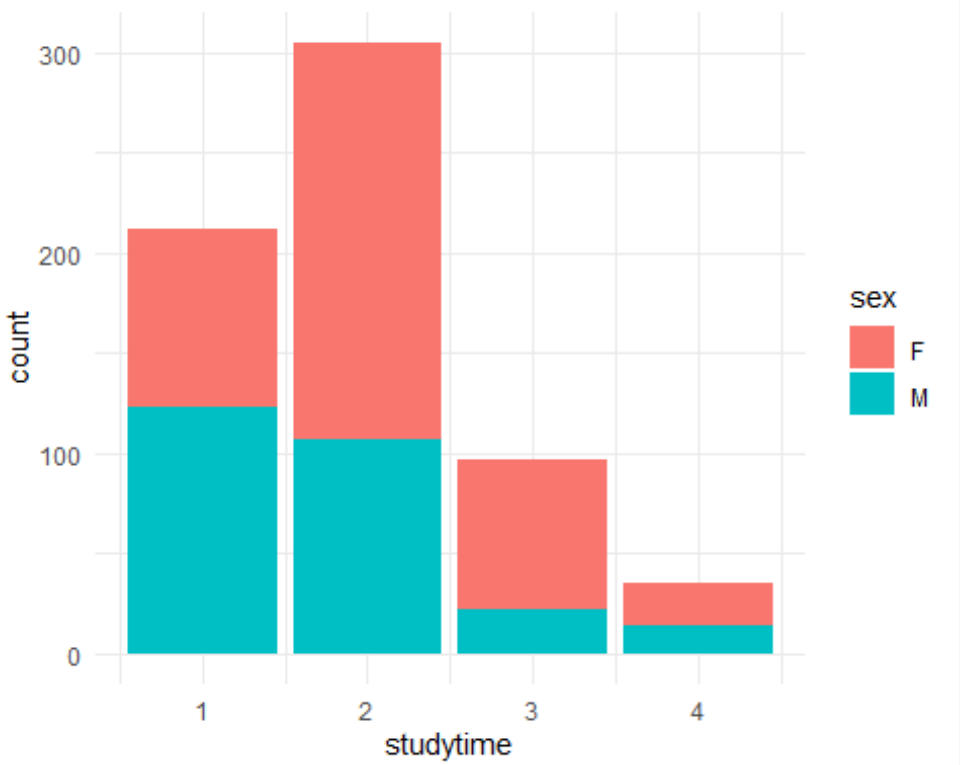
Εικόνα 5: Ραβδόγραμμα μεταβλητής "Walc"



Αντίστοιχη ανάλυση μπορεί να γίνει όσο αναφορά την μεταβλητή “studytime” που προσδιορίζει τις ώρες μελέτης του κάθε μαθητή. Παρατηρούμε ότι οι περισσότεροι μαθητές στο σύνολο τους μελετούν από δύο έως πέντε ώρες την εβδομάδα (τιμή 2 της μεταβλητής studytime), όμως τα περισσότερα αγόρια μελετούν μέχρι δύο ώρες την εβδομάδα (τιμή 1 της μεταβλητής).

```
ggplot(data=data, aes(x=studytime,fill=sex)) +
geom_bar(stat="count")+
theme_minimal()
```

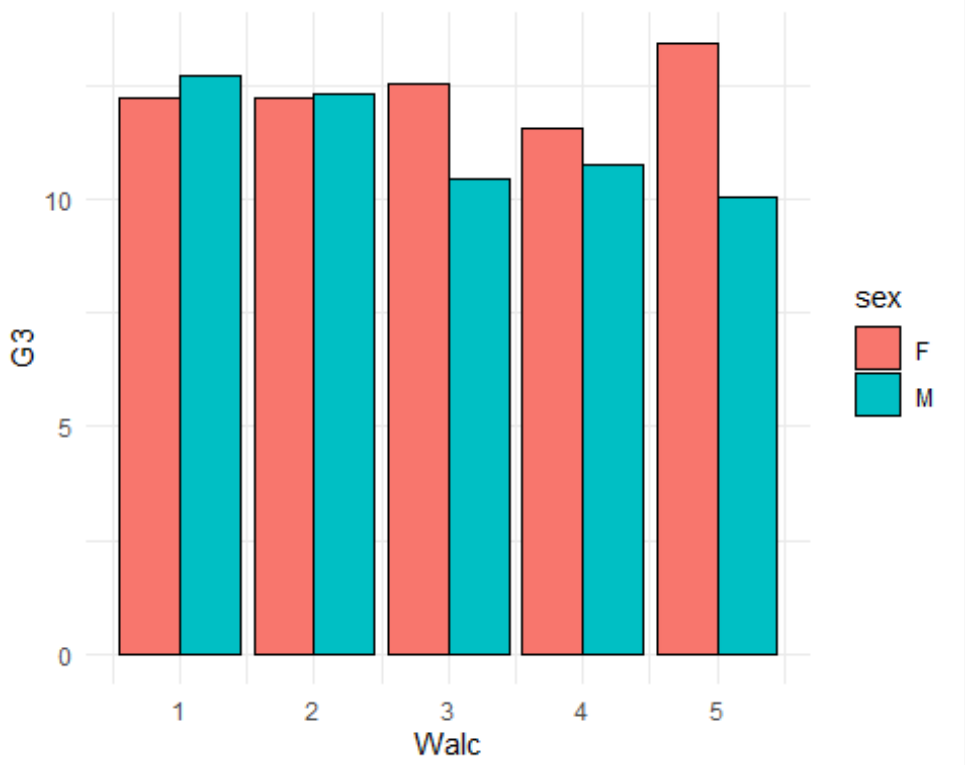
Εικόνα 6: Ραβδόγραμμα μεταβλητής "Studytime"



Στην συνέχεια θέλουμε να μελετήσουμε οπτικά την συσχέτιση δύο κατηγορικών μεταβλητών με μία συνεχή μεταβλητή. Συγκεκριμένα, θέλουμε να δούμε την συσχέτιση της κατανάλωσης αλκοόλ (Walc) με την επίδοση (G3) λαμβάνοντας όμως υπόψη και το φύλο (sex). Για το σκοπό αυτό υπολογίζουμε την μέση επίδοση για κάθε ζευγάρι τιμών των μεταβλητών “Walc” και “sex”. Για τα αγόρια παρατηρούμε ότι όσο η κατανάλωση ποσότητας αλκοόλ αυξάνει, η επίδοση μειώνεται, κάτι που δεν ισχύει για τα κορίτσια. Βέβαια, δεν πρέπει να ξεχνάμε ότι για παράδειγμα το πλήθος κοριτσιών που καταναλώνει αλκοόλ πέντε φορές την εβδομάδα είναι πολύ μικρότερο όπως είδαμε λίγο νωρίτερα.

```
# δημιουργούμε ένα νέο data frame
by_alc <- data %>% group_by(Walc,sex) %>% summarise(G3 = mean(G3))

## `summarise()` regrouping output by 'Walc' (override with `.groups` argument)

ggplot(data=by_alc,aes(x=Walc, y=G3, fill=sex))  +
geom_bar(stat="identity", color="black", position=position_dodge())+
theme_minimal()
```

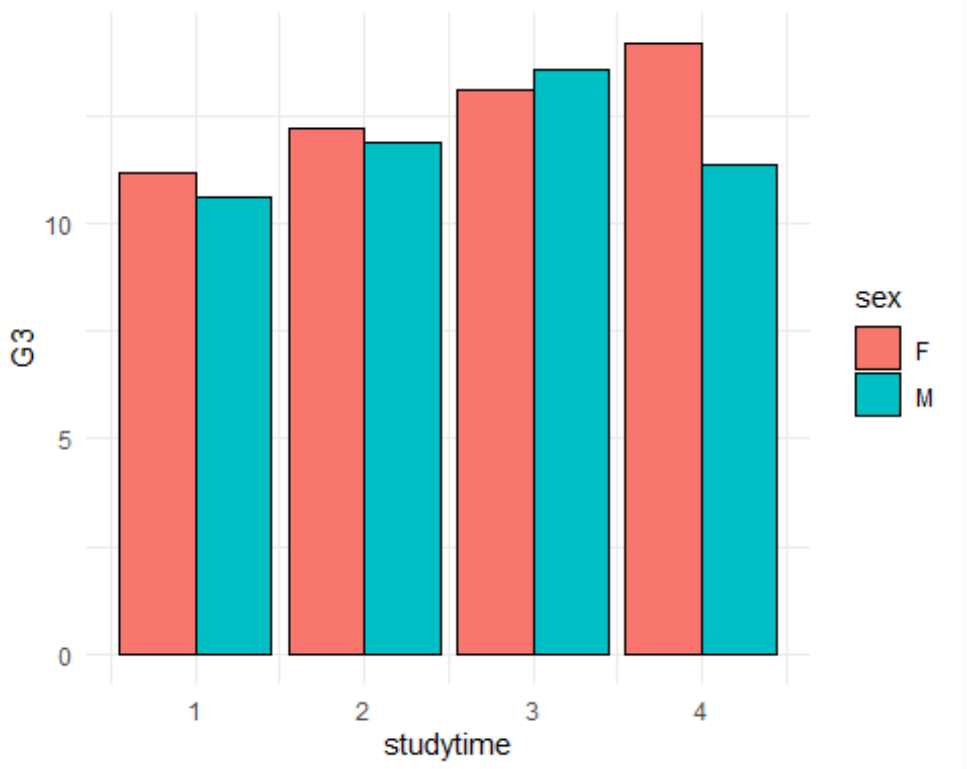
Εικόνα 7: Ιστόγραμμα μέσων τιμών ανά φύλο της μεταβλητής "Walc"



Αντίστοιχα μπορούμε να δούμε την σχέση του χρόνου μελέτης (studytime) με την απόδοση (G3) λαμβάνοντας υπόψη το φύλο. Παρατηρούμε ότι για αγόρια και κορίτσια η απόδοση αυξάνεται όσο αυξάνεται ο χρόνος μελέτης, όμως για τα αγόρια υπάρχει μία πτώση στην απόδοση για εβδομαδιαία μελέτη περισσότερο από 10 ώρες (τιμή 4 της μεταβλητής studytime). Αυτό μπορεί να οφείλεται στο “burnout effect” όπου για κάποιους μαθητές η αύξηση των ωρών μελέτης μπορεί να έχει αρνητικά αποτελέσματα. Είναι επίσης σημαντικό να αναφέρουμε ότι αντίστοιχη συμπεριφορά δεν ισχύει για τα κορίτσια.

```
# δημιουργούμε ένα νέο data frame
by_study <- data %>% group_by(studytime,sex) %>% summarise(G3 = mean(G3))

## `summarise()` regrouping output by 'studytime' (override with `.groups` argument)

ggplot(data=by_study,aes(x=studytime, y=G3, fill=sex))  +
geom_bar(stat="identity", color="black", position=position_dodge())+
theme_minimal()
```

Εικόνα 8: Ιστόγραμμα μέσων τιμών ανά φύλο της μεταβλητής "Studytime"



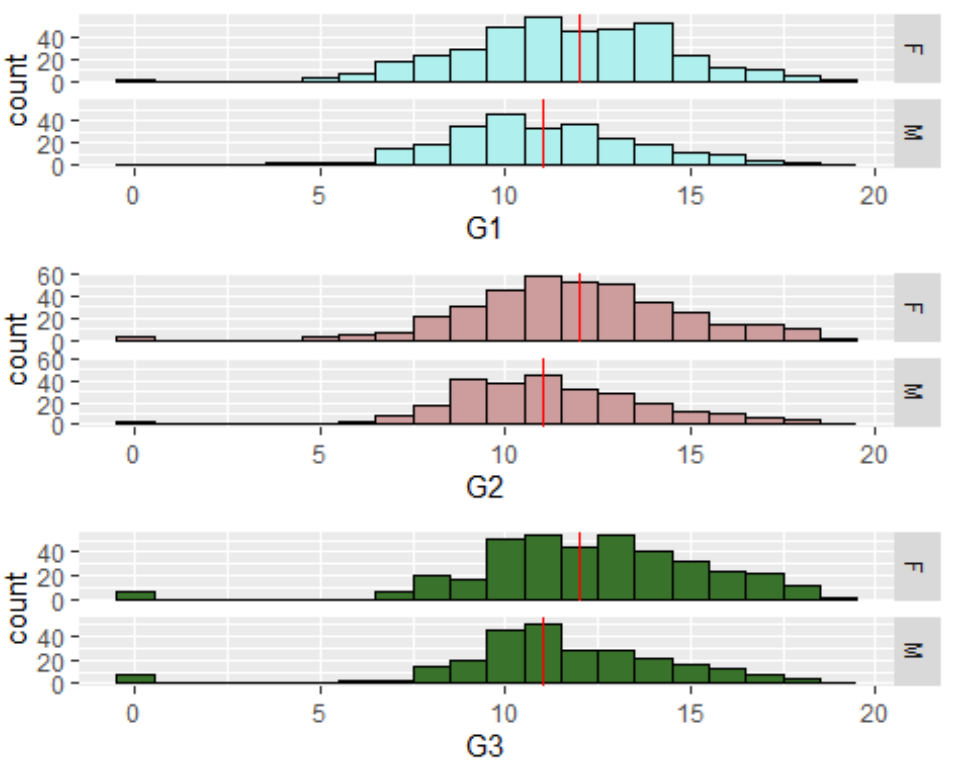
Σχέση φύλου και απόδοσης των μαθητών

Με τη βοήθεια της βιβλιοθήκης “gridExtra” θα οπτικοποιήσουμε σε μία εικόνα τις διαφορές μεταξύ αγοριών και κοριτσιών και για τις τρεις διαθέσιμες βαθμολογίες (G1, G2 και G3). Παρατηρώντας την κατανομή των βαθμών με την αξιοποίηση ιστογραμμάτων βλέπουμε ότι υπάρχουν εξωτερικά σημεία που είναι πιθανόν να επηρεάζουν σημαντικά τη μέση τιμή. Για το λόγο αυτό στα ιστογράμματα συμπεριλαμβάνουμε την διάμεσο (με την κάθετη κόκκινη γραμμή) που μπορεί να είναι περισσότερο αντιπροσωπευτική στην σύγκριση των αποτελεσμάτων. Βλέπουμε ότι και στις τρεις περιπτώσεις τα κορίτσια υπερτερούν όσον αφορά την βαθμολογία.

```
# χρήση της βιβλιοθήκης "gridExtra"
library(gridExtra)

# δημιουργία ενός ακόμα data frame
pp1 <- data %>% select(sex,G1)
p1<-ggplot(data, aes(x=G1)) + geom_histogram(fill="#AFEFED",
colour="black",binwidth=1) +
```

```
facet_grid(sex ~ .)+geom_vline(data=aggregate(pp1[2], pp1[1], median),
    mapping=aes(xintercept=G1), color="red")
# δημιουργία ενός ακόμα data frame
pp2 <- data %>% select(sex,G2)
p2<-ggplot(data, aes(x=G2)) + geom_histogram(fill="#CD9C9C", colour="black",binwidth
= 1) +
facet_grid(sex ~ .)+geom_vline(data=aggregate(pp2[2], pp2[1], median),
    mapping=aes(xintercept=G2), color="red")
# δημιουργία ενός ακόμα data frame
pp3 <- data %>% select(sex,G3)
p3<-ggplot(data, aes(x=G3)) + geom_histogram(fill="#39722A", colour="black",binwidth
= 1) +
facet_grid(sex ~ .)+geom_vline(data=aggregate(pp3[2], pp3[1], median),
    mapping=aes(xintercept=G3), color="red")
# συνδυασμός όλων των παραγόμενων εικόνων σε μία
grid.arrange(p1,p2,p3)
```

Εικόνα 9:Ιστόγραμμα βαθμολογιών ανά φύλο



Συμπληρωματικά μπορούμε να διεξάγουμε και ένα τεστ υποθέσεων όσον αφορά τη σύγκριση της συνολικής επίδοσης των αγοριών με αυτή των κοριτσιών. Για την

ακρίβεια θα χρησιμοποιήσουμε ένα t-test (Student test) για να αποφανθούμε για το αν η μέση συνολική επίδοση των αγοριών είναι ίση με των κοριτσιών ή όχι. Για το σκοπό αυτό υποθέτουμε ότι τα δύο διανύσματα τιμών για τα αγόρια και τα κορίτσια αντίστοιχα περιέχουν τυχαία δείγματα που είναι ανεξάρτητα και προέρχονται από κανονικές κατανομές με άγνωστες αλλά ίσες διασπορές.

οι υποθέσεις είναι οι εξής:

- **H0**: Δεν υπάρχει διαφορά ανάμεσα στις μέσες τιμές
- **H1**: Υπάρχει διαφορά ανάμεσα στις μέσες τιμές

```
# Paired t-test
t.test(data$G3[which(data$sex == "M")],data$G3[which(data$sex == "F")], var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  data$G3[which(data$sex == "M")] and data$G3[which(data$sex == "F")]
## t = -3.3109, df = 647, p-value = 0.0009815
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3497314 -0.3447659
## sample estimates:
## mean of x mean of y
##  11.40602  12.25326
```

Παρατηρούμε ότι το “P-value” είναι μικρότερο από 0.05 άρα απορρίπτουμε την αρχική υπόθεση H0. Με βάση τις υποθέσεις μας η μέση απόδοση των κοριτσιών είναι στατιστικά σημαντικά μεγαλύτερη από αυτή των αγοριών.

Η επίδραση των ρομαντικών σχέσεων

Αρχικά εξετάζουμε τη σχέση μεταξύ της κατανάλωσης αλκοόλ και της ύπαρξης ρομαντικής σχέσης. Επειδή πρόκειται για δύο κατηγορικές μεταβλητές θα εφαρμόσουμε ένα X^2-test αξιοποιώντας την αντίστοιχη συνάρτηση από την βιβλιοθήκη “MASS” (Venables and Ripley, 2002).

```
tb1 <- table(data$Walc,data$romantic)
library(MASS)

chisq.test(tb1)

##
##  Pearson's Chi-squared test
##
## data:  tb1
## X-squared = 1.6351, df = 4, p-value = 0.8025
```

Η τιμή του p-value είναι μεγαλύτερη από 0.05 οπότε συμπεραίνουμε ότι οι δύο κατηγορικές μεταβλητές δεν συσχετίζονται.

Στην συνέχεια μελετάμε την σχέση της ρομαντικής σχέσης με τον τελικό βαθμό (G3) μέσω ενός ακόμη t-test (Student test).

οι υποθέσεις είναι οι εξής:

- **H0**: Δεν υπάρχει διαφορά ανάμεσα στις μέσες τιμές της απόδοσης των μαθητών που είναι σε ρομαντική σχέση και αυτών που δεν είναι.
- **H1**: Υπάρχει διαφορά ανάμεσα στις μέσες τιμές.

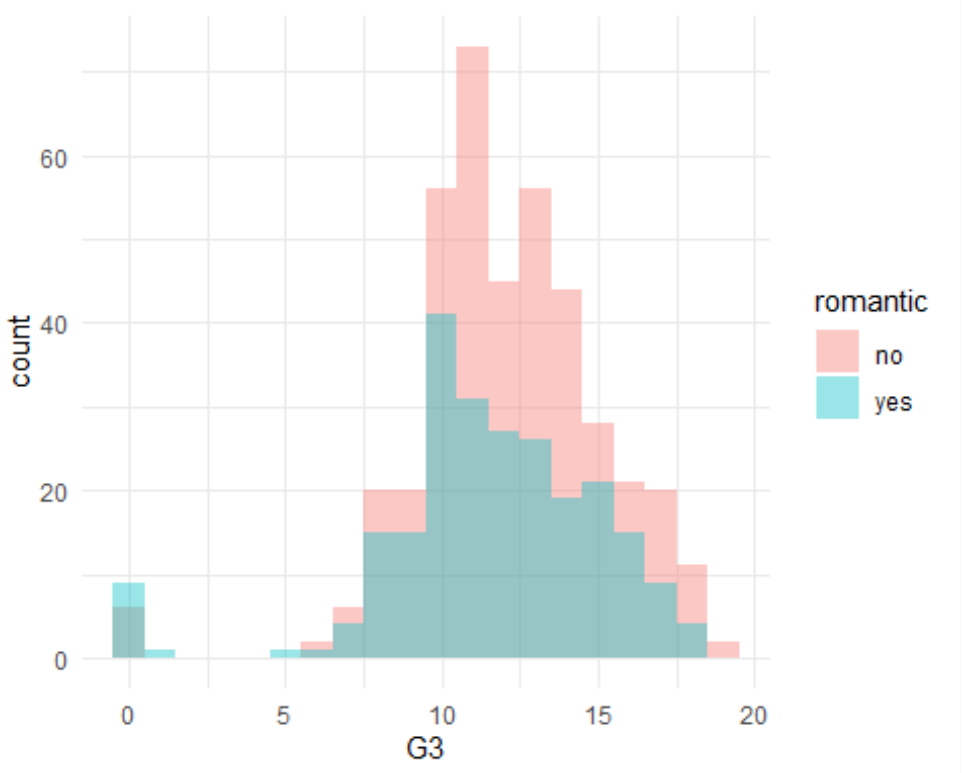
```
# Paired t-test
t.test(data$G3[which(data$romantic == "yes")],data$G3[which(data$romantic == "no")],
var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  data$G3[which(data$romantic == "yes")] and data$G3[which(data$romantic ==
"no")]
## t = -2.3136, df = 647, p-value = 0.021
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.12080832 -0.09170316
## sample estimates:
## mean of x mean of y
##  11.52301  12.12927
```

Σύμφωνα με το αποτέλεσμα του τεστ βλέπουμε ότι το p-value είναι λίγο μικρότερο από 0.05 οπότε μπορούμε να απορρίψουμε την αρχική υπόθεση και να επιβεβαιώσουμε ότι οι μαθητές που δεν είναι σε ρομαντική σχέση έχουν λίγο καλύτερη απόδοση.

Ενδιαφέρον έχει, επίσης, να μελετήσουμε οπτικά την κατανομή για τις δύο διαφορετικές κατηγορίες μέσω ενός ιστογράμματος. Παρατηρούμε ότι σημαντικότερη διαφορά στην απόδοση υπάρχει για τους μαθητές με βαθμό μεγαλύτερο του 10. Οι μαθητές που δεν είναι σε ρομαντική σχέση παίρνουν μεγαλύτερους βαθμούς πιο συχνά.

```
ggplot(data, aes(x=G3, fill=romantic)) +
geom_histogram(position="identity", alpha=0.4,binwidth=1.0)+
theme_minimal()
```

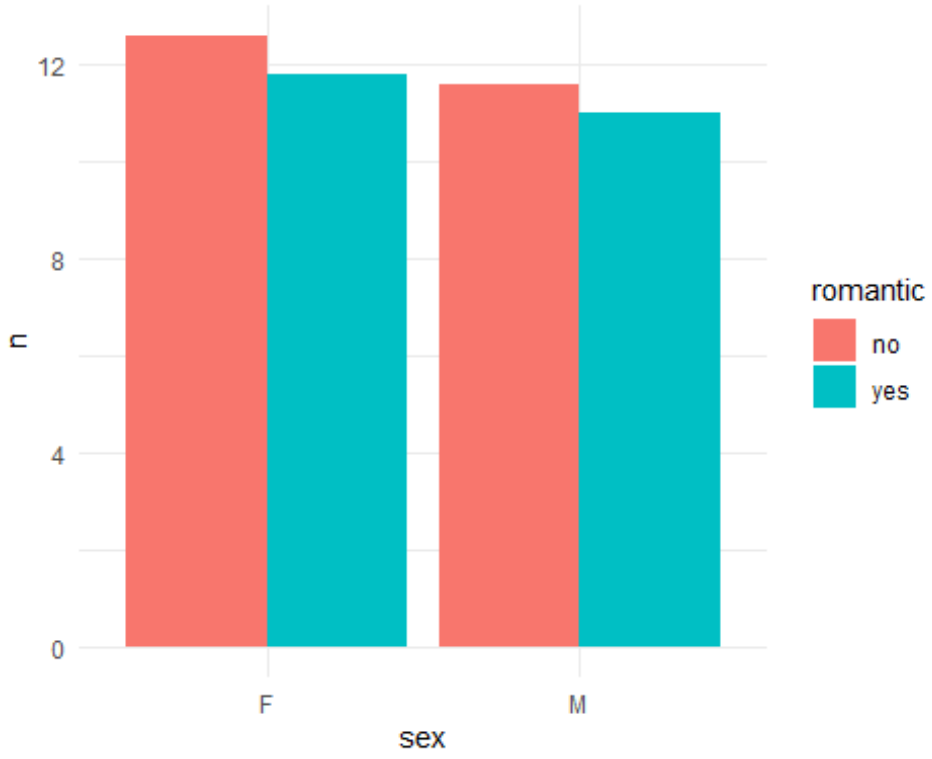
Εικόνα 10: Ιστόγραμμα κατανομής βαθμολογιών ανά τύπου σχέσης



Τέλος, ίσως έχει ενδιαφέρον να δούμε αν οι ρομαντικές σχέσεις επηρεάζουν διαφορετικά τα αγόρια από τα κορίτσια. Παρατηρούμε ότι και για τις δύο κατηγορίες η συμπεριφορά είναι περίπου ίδια.

```
data %>% group_by(sex,romantic) %>% summarise(n=mean(G3,na.rm=T)) %>%
  ggplot(aes(x=sex, y=n, fill=romantic)) +
geom_bar(position="dodge",stat="identity")+
theme_minimal()

## `summarise()` regrouping output by 'sex' (override with `.groups` argument)
```

Εικόνα 11: Ιστόγραμμα βαθμολογιών ανά τύπου σχέσεις για τα διαφορετικά φύλα



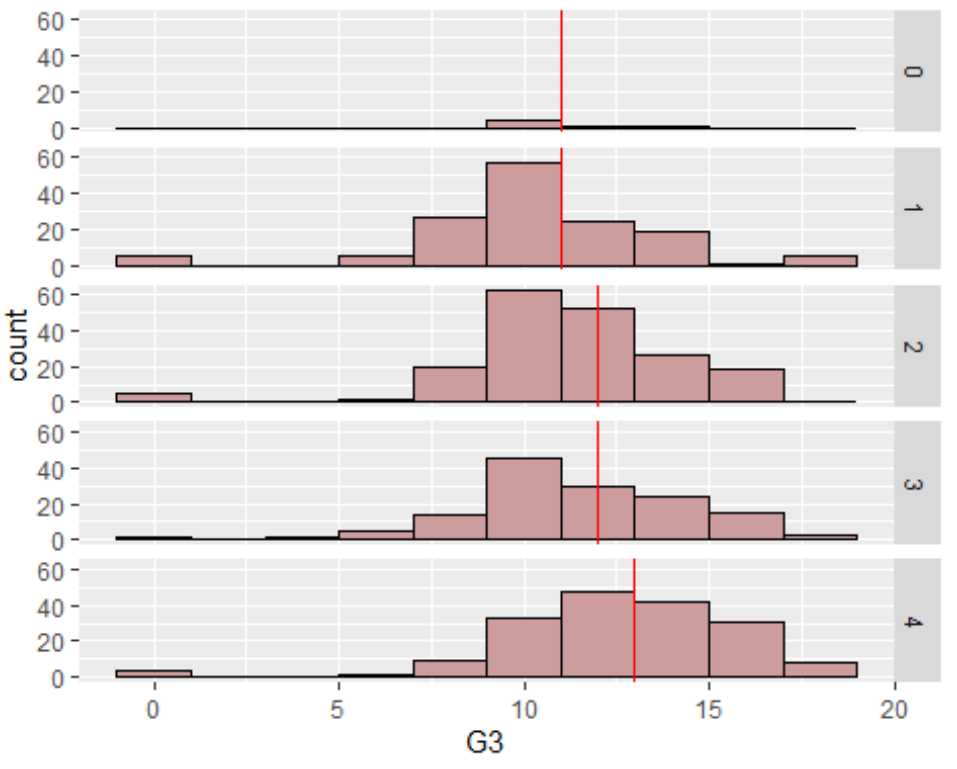
Σχέση μορφωτικού επιπέδου γονέων με την απόδοση των μαθητών

Σε αυτό το σημείο θα μελετήσουμε την σχέση του μορφωτικού επιπέδου της μητέρας με τον τελικό βαθμό των μαθητών. Για το σκοπό αυτό θα χρησιμοποιήσουμε, όπως και νωρίτερα στην ανάλυσή μας, ιστογράμματα για να μελετήσουμε την κατανομή για κάθε τιμή της κατηγορικής μεταβλητής “Medu” και την αντίστοιχη τιμή της διαμέσου. Παρατηρούμε ότι όσο το μορφωτικό επίπεδο της μητέρας αυξάνεται η διάμεσος της συνολικής απόδοσης “G3” επίσης αυξάνεται. Είναι προφανές ότι το μορφωτικό επίπεδο της μητέρας είναι ένας σημαντικός παράγοντας στην απόδοση των μαθητών.

```
ggplot(data, aes(x=G3)) + geom_histogram(fill="#CD9C9C", colour="black",binwidth = 2)
+
```

```
facet_grid(Medu ~ .)+geom_vline(data=aggregate(data[33], data[7], median),
mapping=aes(xintercept=G3), color="red")
```

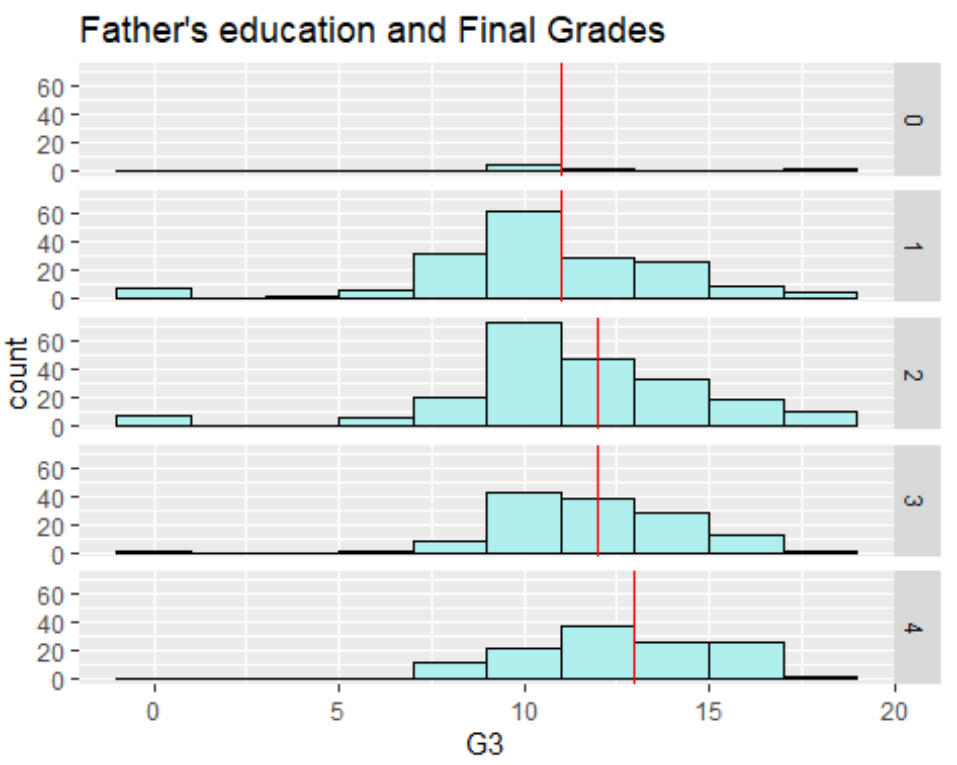
Εικόνα 12: Ιστόγραμμα συσχέτισης βαθμολογίας και μορφωτικού επιπέδου της μητέρας



Στη συνέχεια παράγουμε αντίστοιχη οπτικοποίηση για να παρατηρήσουμε την επίδραση του μορφωτικού επιπέδου του πατέρα και εξάγουμε παρόμοια συμπεράσματα.

```
ggplot(data, aes(x=G3)) + geom_histogram(fill="#AFEFED", colour="black",binwidth = 2)
+
facet_grid(Fedu ~ .)+ggtitle("Father's education and Final
Grades")+geom_vline(data=aggregate(data[33], data[8], median),
mapping=aes(xintercept=G3), color="red")
```

Εικόνα 13: Ιστόγραμμα συσχέτισης βαθμολογίας και μορφωτικού επιπέδου του πατέρα

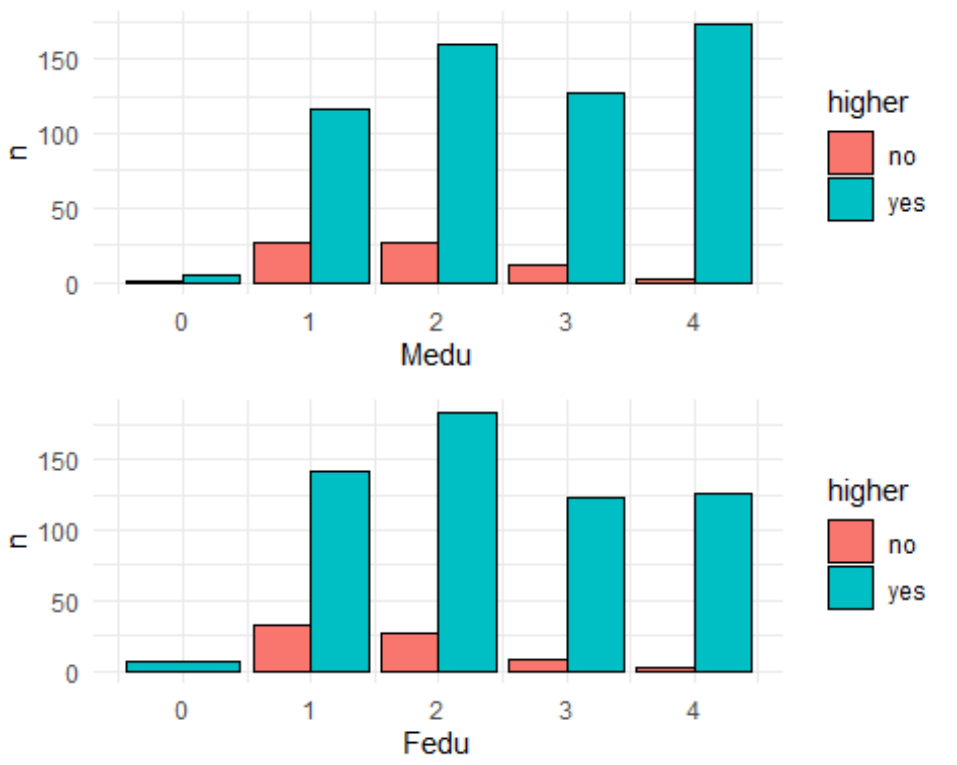


Τέλος, θέλουμε να μελετήσουμε αν το μορφωτικό επίπεδο των γονέων επηρεάζει την απόφαση των μαθητών σχετικά με την συνέχεια στην τριτοβάθμια εκπαίδευση. Για αυτό το λόγο μελετάμε το πλήθος των μαθητών που έχουν σχέδια να συνεχίσουν στην τριτοβάθμια εκπαίδευση για κάθε τιμή των μεταβλητών “Medu” και “Fedu” αντίστοιχα. Και στις δύο περιπτώσεις βλέπουμε αντίστοιχη συμπεριφορά: το ποσοστό των μαθητών που επιλέγει να συνεχίσει στην τριτοβάθμια εκπαίδευση αυξάνεται όσο αυξάνεται το μορφωτικό επίπεδο των γονέων.

```
# δημιουργούμε νέο data frame
by_higher <- data %>% group_by(Medu,higher) %>% summarise(n=n())

## `summarise()` regrouping output by 'Medu' (override with `.groups` argument)

p1 <- ggplot(data=by_higher,aes(x=Medu, y=n, fill=higher))  +
geom_bar(stat="identity", color="black", position=position_dodge())+
theme_minimal()
# δημιουργούμε νέο data frame
by_higher <- data %>% group_by(Fedu,higher) %>% summarise(n=n())
```

```
## `summarise()` regrouping output by 'Fedu' (override with `.groups` argument)

p2 <- ggplot(data=by_higher,aes(x=Fedu, y=n, fill=higher))  +
geom_bar(stat="identity", color="black", position=position_dodge())+
theme_minimal()
grid.arrange(p1,p2)
```

Εικόνα 14: Ιστόγραμμα συσχέτισης επιλογής τριτοβάθμιας εκπαίδευσης και μορφωτικού επιπέδου γονέων



Επιβλεπομένη Μάθηση για την πρόγνωση της τελικής βαθμολογίας

Σε αυτό το σημείο θέλουμε μελετήσουμε αν με βάση τις διαθέσιμες μεταβλητές μπορούμε να προβλέψουμε αν οι μαθητές θα ανήκουν στην κατηγορία υψηλής/χαμηλής τελικής βαθμολογίας. Για το σκοπό αυτό θα μετατρέψουμε την μεταβλητή “G3” σε κατηγορική με δύο κατηγορίες. Αν η τιμή είναι μικρότερη από δεκατρία ο μαθητής θα ανήκει στην πρώτη κατηγορία, ενώ αν είναι μεγαλύτερη στην δεύτερη. Αρχικά κάνουμε τις απαραίτητες προσαρμογές ώστε το σύνολο δεδομένων μας να είναι έτοιμο για την ανάλυση.

```
# δημιουργούμε μία νέα κατηγορική μεταβλητή.
data$final_group <- ifelse(data$G3>=13,1,0)
```

```
data$final_group <- as.factor(data$final_group)
# αφαιρούμε από το σύνολο δεδομένων μας την μεταβλητή G3
data <- data[,c(1:32,34)]
names(data)

##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "final_group"

# μετατρέπουμε σε factor όσες μεταβλητές χρειάζεται
data <- data %>%
mutate(schoolsup=as.factor(schoolsup),famsup=as.factor(famsup),paid=as.factor(paid),
activities=as.factor(activities),nursery=as.factor(nursery),higher=as.factor(higher),intern
et=as.factor(internet),romantic=as.factor(romantic))
```

Στη συνέχεια χωρίζουμε τα δεδομένα μας σε σύνολο εκπαίδευσης (train) και σύνολο επαλήθευσης (test).

```
train <- sample(dim(data)[1],dim(data)[1]*0.9)
test <- c(1:dim(data)[1])[-train]
training_data <- data[train,]
testing_data <- data[test,]
```

Αρχικά επιστρατεύουμε τον αλγόριθμο λογιστικής παλινδρόμησης. Εκπαιδεύουμε το μοντέλο χρησιμοποιώντας τα δεδομένα εκπαίδευσης και στη συνέχεια προβλέπουμε την κατηγορία των δεδομένων επαλήθευσης. Τέλος, αξιολογούμε το αποτέλεσμα χρησιμοποιώντας την μετρική της ακρίβειας. Μετράμε δηλαδή το ποσοστό των δειγμάτων του συνόλου επαλήθευσης των οποίων η πρόβλεψη συμφωνεί με την πραγματική κατηγορία ως προς το συνολικό αριθμό των δεδομένων επαλήθευσης.

```
# εκπαιδεύουμε το μοντέλο
log_model <- glm(final_group~.,data=training_data,family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# summary(log_model)
# προβλέπουμε την κατηγορία των test data
predict_log <- predict(log_model,newdata=testing_data,type="response")
predict_log <- round(predict_log)
# υπολογίζουμε την ακρίβεια του αποτελέσματος
mean(predict_log==testing_data$final_group)
```

```
## [1] 0.9384615
```

Η ακρίβεια είναι 93%, ένα αρκετά καλό ποσοστό λαμβάνοντας όμως υπόψη ότι οι μεταβλητές “G1”, “G2” που επηρεάζουν σημαντικά το αποτέλεσμα λόγω της υψηλής συσχέτισης με την άγνωστη μεταβλητή συμπεριλαμβάνονται στο σύνολο δεδομένων. Μία πιο αναλυτική ματιά στην ερμηνεία του αποτελέσματος μπορεί να δοθεί αξιοποιώντας περισσότερα μέτρα αξιολόγησης της κατηγοριοποίησης. Για το σκοπό αυτό θα χρησιμοποιήσουμε την συνάρτηση “confusionMatrix” της βιβλιοθήκης “caret” (Kuhn, 2020). Στην αναφορά που ακολουθεί βλέπουμε το μητρώο συσχέτισης καθώς και πλήθος μετρικών που αφορούν το αποτέλεσμα της κατηγοριοποίησης.

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
confusionMatrix(as.factor(predict_log),testing_data$final_group)
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction  0  1
##          0 47  1
##          1  3 14
##
##                Accuracy : 0.9385
##                  95% CI : (0.8499, 0.983)
##     No Information Rate : 0.7692
##     P-Value [Acc > NIR] : 0.0002696
##
```

```
##               Kappa : 0.8344
##
##  Mcnemar's Test P-Value : 0.6170751
##
##          Sensitivity : 0.9400
##          Specificity : 0.9333
##       Pos Pred Value : 0.9792
##       Neg Pred Value : 0.8235
##           Prevalence : 0.7692
##       Detection Rate : 0.7231
##   Detection Prevalence : 0.7385
##     Balanced Accuracy : 0.9367
##
##      'Positive' Class : 0
##
```

Παρόλα αυτά μεγαλύτερο ενδιαφέρον έχει η πρόγνωση της κατηγορίας της τελικής επίδοσης παραβλέποντας ακόμη και τις μεταβλητές “G1” και “G2”. Σε αυτή την περίπτωση θα δούμε πραγματικά κατά πόσο είναι εφικτό να εκτιμηθεί η απόδοση των μαθητών λαμβάνοντας υπόψη αποκλειστικά εξωγενείς παράγοντες. Δημιουργούμε λοιπόν ξανά τη μεταβλητή «`data_class`» και ακολουθούμε αντίστοιχη διαδικασία.

```
# αφαιρούμε από το σύνολο δεδομένων μας τις μεταβλητές G1,G2 και  G3.
data_class <- data[,c(1:32,34)]

# εκπαιδεύουμε το μοντέλο
log_model <- glm(final_group~.,data=training_data,family = binomial)
# summary(log_model)
# προβλέπουμε την κατηγορία των test data
predict_log <- predict(log_model,newdata=testing_data,type="response")
predict_log <- round(predict_log)
# υπολογίζουμε την ακρίβεια του αποτελέσματος
mean(predict_log==testing_data$final_group)
```

```
## [1] 0.7846154
```

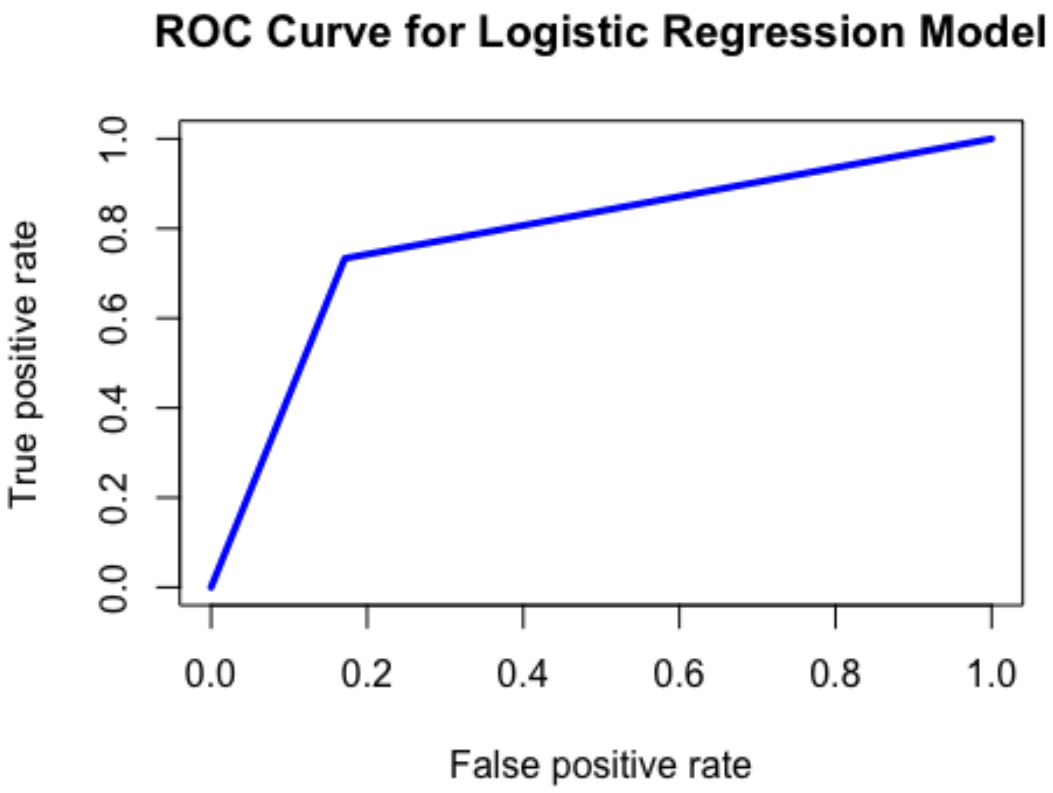
Παρατηρούμε ότι η ακρίβεια μειώθηκε σημαντικά. Παρόλα αυτά διερευνώντας και το μητρώο συσχέτισης βλέπουμε ότι και πάλι η επίδοση μπορεί να προβλεφθεί σε ικανοποιητικό βαθμό.

```
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

confusionMatrix(as.factor(predict_log),testing_data$final_group)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 29  8
##          1  6 22
##
##                Accuracy : 0.7846
##                  95% CI : (0.6651, 0.8769)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 3.386e-05
##
##                   Kappa : 0.5646
##
##  Mcnemar's Test P-Value : 0.7893
##
##             Sensitivity : 0.8286
##             Specificity : 0.7333
##          Pos Pred Value : 0.7838
##          Neg Pred Value : 0.7857
##              Prevalence : 0.5385
##          Detection Rate : 0.4462
##    Detection Prevalence : 0.5692
##       Balanced Accuracy : 0.7810
##
##        'Positive' Class : 0
```

Τέλος μπορούμε να οπτικοποιήσουμε την αποδοτικότητα του μοντέλου μέσω της καμπύλης ROC. Για το σκοπό αυτό θα χρησιμοποιήσουμε την βιβλιοθήκη “ROCR” (Sing et al., 2005).

```
library(ROCR)

## Warning: package 'ROCR' was built under R version 4.0.2

pred <- prediction(predictions = predict_log,labels=testing_data$final_group)
perf <- performance(pred,measure = "tpr",x.measure = "fpr")
plot(perf,main="ROC Curve for Logistic Regression Model",col="blue",lwd=3)
```

Εικόνα 15: Διάγραμμα καμπύλης ROC ως ένδειξη αποδοτικότητας του μοντέλου



Επίσης υπολογίζουμε την επιφάνεια κάτω από την καμπύλη (AUC). Όσο μεγαλύτερη είναι αυτή η τιμή τόσο καλύτερο το αποτέλεσμα της κατηγοριοποίησης.

```
perf.auc <- performance(pred,measure = "auc")
unlist(perf.auc@y.values)

## [1] 0.7809524
```

Η τιμή του AUC είναι περίπου 0.78, μία αρκετά καλή τιμή που δείχνει ότι μία αξιόλογη πρόβλεψη της επίδοσης των μαθητών είναι εφικτή.

Κατηγοριοποίηση με Τυχαία Δάση (Random Forest (rf))

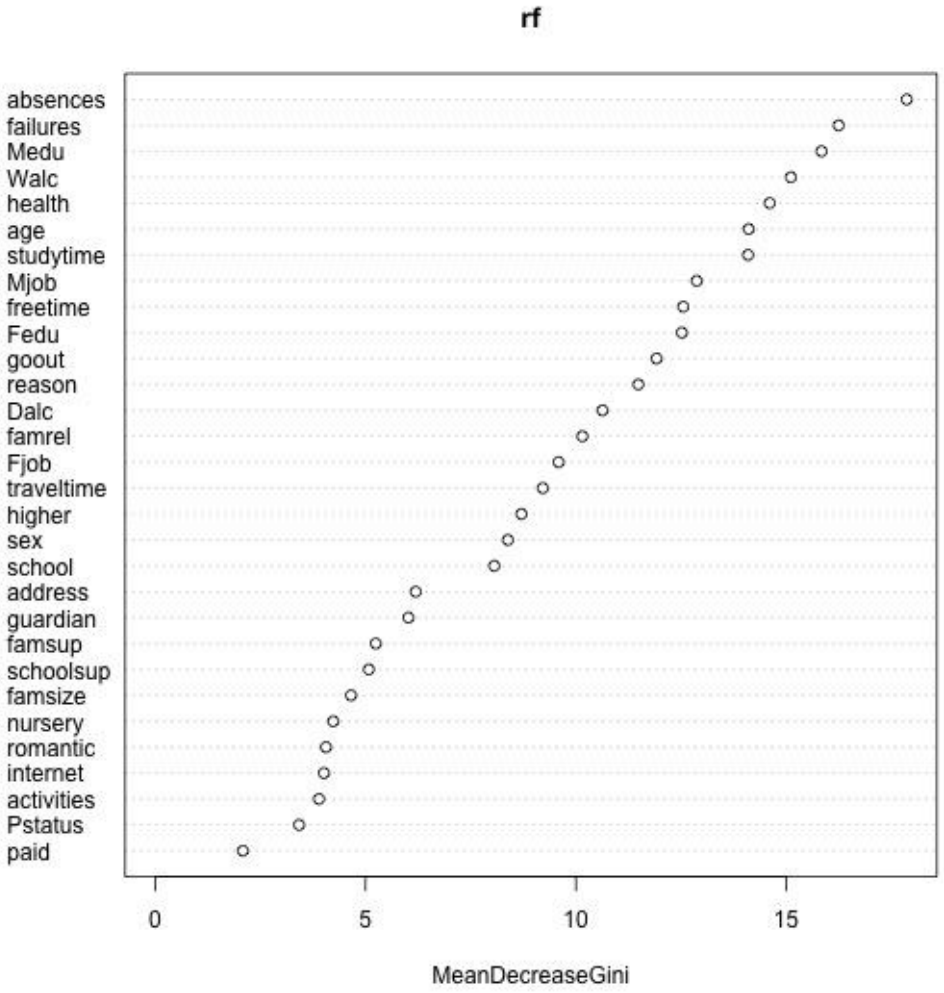
Στο τελευταίο κομμάτι της ανάλυσης θα χρησιμοποιήσουμε τυχαία δάση για την κατηγοριοποίηση που περιγράψαμε στο προηγούμενο βήμα. Χρησιμοποιούμε την βιβλιοθήκη “randomForest” (Liaw and Wiener, 2002). Αρχικά εκπαιδεύουμε το μοντέλο.

```
library("randomForest")

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

mtry <- sqrt(ncol(data))
# επιλέγουμε το πλήθος των δέντρων
ntree <- 100
rf <-
randomForest(final_group~.,data=training_data,mtry=mtry,ntree=ntree)
```

Κάνουμε ανάλυση σημαντικότητας των μεταβλητών στην κατηγοριοποίηση αξιοποιώντας τον αλγόριθμο Τυχαίων δασών. Στην ακόλουθη εικόνα βλέπουμε την σύγκριση όλων των μεταβλητών διατεταγμένων κατά φθίνουσα σειρά σημαντικότητας.

```
varImpPlot(rf)
```

Εικόνα 16: Σημαντικότητα μεταβλητών με το μοντέλο των Τυχαίων Δασών



Βλέπουμε πόσο πολύ επηρεάζει την πρόβλεψη της επίδοσης του μαθητή το πλήθος των απουσιών (absences) και η αποτυχία στα μαθήματα σε προηγούμενες χρονικές περιόδους. Επίσης, σημαντική επιρροή ασκεί το μορφωτικό επίπεδο της μητέρας αλλά και η κατανάλωση αλκοόλ τα Σαββατοκύριακα. Αντίθετα, δεν εμφανίζονται ως σημαντικές μεταβλητές τα έξτρα αμειβόμενα μαθήματα, η κατάσταση συμβίωσης γονέα ή οι εξωσχολικές δραστηριότητες.

Τέλος αξιολογούμε την απόδοση του αλγορίθμου με βάση το σύνολο επαλήθευσης.

```r
# predictions
yhat.rf <- predict(rf,testing_data)
cmRF <- confusionMatrix(yhat.rf,testing_data$final_group)
print(cmRF)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 31  9
##          1  4 21
##
##                Accuracy : 0.8
##                  95% CI : (0.6823, 0.889)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 1.025e-05
##
##                   Kappa : 0.5928
##
##  Mcnemar's Test P-Value : 0.2673
##
##             Sensitivity : 0.8857
##             Specificity : 0.7000
##          Pos Pred Value : 0.7750
##          Neg Pred Value : 0.8400
##              Prevalence : 0.5385
##          Detection Rate : 0.4769
##    Detection Prevalence : 0.6154
##       Balanced Accuracy : 0.7929
##
##        'Positive' Class : 0
##
```

Λαμβάνοντας υπόψη όλες τις μετρικές παρατηρούμε ότι η απόδοση του αλγορίθμου τυχαίων δασών είναι ελάχιστα καλύτερη από αυτή του αλγορίθμου λογιστικής παλινδρόμησης παρότι είναι μια, υπολογιστικά, πιο απαιτητική μέθοδος. Θυμίζετε ότι χρησιμοποιήθηκαν 100 δέντρα για την κατασκευή του μοντέλου.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ-ΠΡΟΤΑΣΕΙΣ

Η παρούσα εργασία έρχεται να αναδείξει παράγοντες που επιδρούν στην επίδοση των μαθητών αλλά και την πρόβλεψη των επιδόσεων μέσω των προγνωστικών μοντέλων που αναφέρθηκαν.

Από τα πραγματικά δεδομένα των μαθητών των δύο σχολείων της Πορτογαλίας βλέπουμε ότι γενικά η απόδοση των κοριτσιών είναι καλύτερη από αυτή των αγοριών. Οι μαθητές που είναι σε ρομαντική σχέση παίρνουν χαμηλότερους βαθμούς. Στην περίπτωση αυτή, και τα αγόρια και τα κορίτσια επηρεάζονται το ίδιο. Οι σχέσεις των εφήβων είναι μία πραγματικότητα που επιδρά στην επίδοση και στη συμπεριφορά τους γενικότερα αφού τα συναισθήματα και η διάθεσή τους μεταβάλλονται με μεγάλη ευκολία. Ίσως σε επίπεδο τοπικής αυτοδιοίκησης να πρέπει να δίνονται διαλέξεις σε γονείς ή κηδεμόνες μαθητών σχετικά με τον τρόπο χειρισμού της δύσκολης ηλικίας στην οποία βρίσκονται τα παιδιά τους με στόχο την καλύτερη επαφή και επικοινωνία μεταξύ τους και την αποφυγή συγκρούσεων.

Το μορφωτικό επίπεδο της μητέρας είναι σημαντικός παράγοντας στην επίδοση των μαθητών, όπως και του πατέρα. Μάλιστα το ποσοστό των μαθητών που επιλέγει να συνεχίσει στην τριτοβάθμια εκπαίδευση αυξάνεται όσο αυξάνεται το μορφωτικό επίπεδο των γονέων.

Επίσης, η κατανάλωση αλκοόλ δε σχετίζεται με την ύπαρξη ρομαντικής σχέσης. Όσο όμως η κατανάλωση αλκοόλ μεγαλώνει, η επίδοση μειώνεται. Αυτό συμβαίνει στα αγόρια που καταναλώνουν μεγαλύτερες ποσότητες αλκοόλ από τα κορίτσια. Μπορεί να απαγορεύεται η κατανάλωση αλκοόλ σε ανήλικους, όμως ο ελεγκτικός μηχανισμός δεν είναι επαρκής.

Ένας ακόμα παράγοντας επίδρασης στην επίδοση είναι ο χρόνος μελέτης. Όσο αυτός αυξάνεται, αυξάνεται και η επίδοση, αλλά στα αγόρια μέχρι το σημείο εξουθένωσης (σύνδρομο burnout). Ένας μαθητής υπό το άγχος των εξετάσεων ή επειδή έχει μεγάλες προσδοκίες θέτοντας μη εφικτούς στόχους οδηγείται σε αδυναμία συγκέντρωσης, ψυχική εξάντληση ακόμα και σε διαταραχή ύπνου. Απαιτείται ο επαναπροσδιορισμός δυσλειτουργικών συμπεριφορών και η ενίσχυση της αυτοεκτίμησης

και αυτοπεποίθησης του μαθητή μέσω συνεδριών με ψυχολόγο. Η πολιτεία οφείλει είτε με μείωση της ύλης σε ορισμένες περιπτώσεις είτε μέσω σωστού σχεδιασμού και δαπανών να δημιουργήσει ένα σχολείο ελκυστικό για το μαθητή. Βέβαια ο μαθητής θα πρέπει να προγραμματίζει και να οργανώνει σωστά το χρόνο του. Ο ελεύθερος χρόνος είναι πολύτιμος για όλους, όπως και ο χρόνος για εξωσχολικές δραστηριότητες, εκτός από το χρόνο μελέτης. Βοήθεια σε αυτό θα μπορούσε να συνεισφέρει η ενίσχυση του ερασιτεχνικού αθλητισμού, η επιδότηση ερασιτεχνικών σωματείων. Τα οφέλη θα είναι η σωματική και η ψυχική υγεία αλλά και ο προγραμματισμός και η αυτοπειθαρχία που προάγει ο αθλητισμός.

Είναι σημαντικό να γνωρίζουμε τι επηρεάζει την επίδοση του μαθητή, είναι σημαντικό όμως να προβλέπουμε και την επίδοσή του. Η έγκαιρη παρέμβαση του εκπαιδευτικού και της πολιτείας θα λειτουργήσει καταλυτικά για το μαθητή. Μαθήματα ενισχυτικής διδασκαλίας για μαθητές που προβλέπεται ότι θα είναι αδύναμοι και όχι μετά που θα αποτύχουν με τα γνωστικά κενά να αυξάνονται. Προώθηση των παιδιών που φαίνεται ότι θα έχουν εξαιρετικές επιδόσεις σε μαθητικούς διαγωνισμούς. Συνεδρίες με ψυχολόγους στα παιδιά είτε εντός είτε εκτός σχολείου για λόγους ενθάρρυνσης. Βελτίωση σχολικών προγραμμάτων, αναπροσαρμογή του αναλυτικού προγράμματος, επαναπροσδιορισμός διδακτικών στόχων. Στόχος πάντα είναι η ανατροφοδότηση της διαδικασία της μάθησης και η ποιότητα της εκπαίδευσης.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΟΓΛΩΣΣΗ:

Αθανασίου Λ., (2000). *Αξιολόγηση της επίδοσης του μαθητή στο σχολείο και του διδακτικού έργου* , Ιωάννινα

Ηλιοπούλου Π., (2015). *Γεωγραφική Ανάλυση* , εκδόσεις Κάλλιπος Chapter 04 (Κατανομές Πιθανότητας- Έλεγχοι Υποθέσεων), URL: https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2063/3/02_chapter_04-Iliopoulou_%ce%91%ce%9d%ce%91%ce%98%ce%95%ce%a9%ce%a1%ce%97%ce%a3%ce%97.pdf

Κωνσταντίνου Χ., (2002). Επιθεώρηση εκπαιδευτικών θεμάτων τ.7:(37-51) *Η αξιολόγηση της επίδοσης του μαθητή σύμφωνα με το Διαθεματικό Ενιαίο Πλαίσιο Προγραμμάτων Σπουδών*

ΞΕΝΟΓΛΩΣΣΗ:

Chockla M., (2013). *Statistical analysis of student performance in redesigned developmental mathematics courses*, URL: https://libres.uncg.edu/ir/wcu/f/Chockla2013.pdf

Cortez P., and Silva A., (2008). *Using data mining to predict secondary school student performance.* EUROSIS.

Daud A., Aljohani N., Abbasi R., Lytras M., Abbas F. and Alowibdi J., (2017). *Predicting student performance using advanced learning analytics.* In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee Google Scholar

Lozar Glenn J., (2000). *Environment-Based Education: Creating High Performance Schools and Students. National Environmental Education and Training Foundation*, Washington, DC. URL: https://files.eric.ed.gov/fulltext/ED451033.pdf

Holgerbrandl/Datautils: *Small Utilities to Make R-Scripting More Fun*. n.d.

Kuhn M., (2020). Caret: *Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Kuhn M., Jackson S., and Cimentada J., (2020). *Corrr: Correlations in R*. https://CRAN.R-project.org/package=corrr

Liaw A., and Wiener M., (2002). *Classification and Regression by randomForest*. R News 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Martin A., and Dowson M., (2009). *Interpersonal Relationships, Motivation, Engagement, and Achievement: Yields for Theory, Current Issues, and Education Practice* Review of Educational Research, Vol.79, No. 1, pp327-365, http://rer.aera.net

Sing T., Sander O., Beerenwinkel N., and Lengauer T., (2005). ROCR: Visualizing Classifier Performance in R. Bioinformatics 21 (20): 7881. http://rocr.bioinf.mpi-sb.mpg.de.

Tremblay S., Ross N., and Berthelot J., (2001). *Factors affecting student performance in Ontario: A multilevel analysis* http://www.geog.mcgill.ca/faculty/grade3ontario.pdf

PanduRanga T., Lakshmi B., Rekha S., and Dhanalakshmi M. (2018). *Student Performance Analysis with Using Statistical and Cluster Studies*
https://link.springer.com/chapter/10.1007/978-981-13-0514-6_71

Venables W., and Ripley B., (2002). *Modern Applied Statistics with S. Fourth*. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Wickham H., (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham H., Francois R., Henry L., and Muller K., (2020). *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xing W., (2015). *Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory*. Comput. Hum. Behav.47, 168–181 CrossRefGoogle Scholar

Ma Y., Liu B., Wong C., Yu P., and Lee S., (2000). *Targeting the right students using data mining. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00)*. Association for Computing Machinery, New York, NY, USA, 457–464. DOI:https://doi.org/10.1145/347090.347184