ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ
Πανεπιστημιούπολη Σίνδου

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΡΟΜΠΟΤΙΚΗ, STEAM ΚΑΙ ΝΕΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ»

Διπλωματική Εργασία

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΤΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΕΠΙΔΟΣΕΩΝ ΜΑΘΗΤΩΝ ΜΕ ΒΑΣΗ ΔΗΜΟΓΡΑΦΙΚΑ ΚΑΙ ΟΙΚΟΓΕΝΕΙΑΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

της

Χρυσάνθης Μπερετζίκη

Επιβλέπων Καθηγητής
Παναγιώτης Τζιώνας

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος ειδίκευσης Ρομποτική, STEAM και Νέες Τεχνολογίες στην Εκπαίδευση
Θεσσαλονίκη, Ιανουάριος 2023

Υπεύθυνη Δήλωση

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις που προβλέπονται από τον Κανονισμό Σπουδών του Μεταπτυχιακού Προγράμματος Ρομποτική, STEAM και Νέες Τεχνολογίες στην Εκπαίδευση του Διεθνούς Πανεπιστημίου Ελλάδος, δηλώνω υπεύθυνα ότι:

- Η παρούσα Διπλωματική Εργασία αποτελεί έργο αποκλειστικά δικής μου δημιουργίας, έρευνας, μελέτης και συγγραφής.
- Για τη συγγραφή της Διπλωματικής μου Εργασίας δεν χρησιμοποίησα ολόκληρο ή μέρος έργου άλλου δημιουργού ή τις ιδέες και αντιλήψεις άλλου δημιουργού χωρίς να γίνεται σαφής αναφορά στην πηγή προέλευσης(βιβλίο, άρθρο από επιστημονικό περιοδικό, ιστοσελίδα κλπ.).

Θεσσαλονίκη, 24/01/2023

Ο/Η Δηλών/ούσα: Χρυσάνθη Μπερετζίκη

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διατριβής μου θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον καθηγητή του Διεθνούς Πανεπιστημίου Ελλάδος, κ. Τζιώνα Παναγιώτη, για την πολύτιμη -επιστημονική και μη- βοήθεια του αλλά και το κίνητρο που μου έδωσε ώστε να συνεχίσω την παρούσα έρευνα.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου καθώς χωρίς την συνεχή στήριξη και ανιδιοτελή φροντίδα αυτής δεν θα είχα φτάσει στον δρόμο που είμαι σήμερα.

Τέλος, οφείλω να ευχαριστήσω τους ανθρώπους που είναι κοντά μου, οι οποίοι στέκονται άξιοι συνοδοιπόροι της ζωής μου και αποτελούν ανεκτίμητο στήριγμα για εμένα.

Περίληψη

Η αξιοσημείωτη πρόοδος τόσο της μηχανικής μάθησης, όσο και της εξόρυξης δεδομένων τα τελευταία χρόνια έχει συμβάλλει καθοριστικά στον χώρο της εκπαίδευσης μέσω της παραγωγής νέας γνώσης, η οποία συμβάλλει στην λήψη αποφάσεων και στην βελτίωση της εκπαιδευτικής διαδικασίας. Στα πλαίσια της παρούσας έρευνας, θα γίνει η εφαρμογή διαφόρων τεχνικών μηχανικής μάθησης με απώτερο σκοπό την πρόβλεψη των ακαδημαϊκών επιδόσεων μαθητών. Η έρευνα χωρίζεται σε δύο περιόδους, προ και μετά την περίοδο Covid – 19, και αφορά μαθητές και μαθήτριες δευτεροβάθμιας εκπαίδευσης αλλά και φοιτητές πανεπιστημίου. Στο πρώτο μέρος της θα εφαρμοστεί η τεχνική της γραμμικής παλινδρόμησης αλλά και των νευρωνικών δικτύων ώστε να προβλέψουμε τις ακαδημαϊκές επιδόσεις μαθητών και μαθητριών στο μάθημα των Μαθηματικών. Στην συνέχεια, τα μοντέλα θα αξιολογηθούν μέσω διαφορετικών δεικτών επίδοσης ώστε να επιλεγεί αυτό που τα πηγαίνει καλύτερα. Στο δεύτερο κομμάτι, το ενδιαφέρον θα εστιαστεί στην περίοδο Covid-19, κατά την οποία φοιτητές και φοιτήτριες πανεπιστημίου απάντησαν σε διάφορες ερωτήσεις προκειμένου να συλλεγεί το δεύτερο σύνολο δεδομένων που θα χρησιμοποιηθεί. Η πρόβλεψη σε αυτό το κομμάτι θα αφορά τον μέσο όρο βαθμολογίας της τελικής περιόδου και θα επιτευχθεί με την εφαρμογή πέντε διαφορετικών αλγορίθμων μηχανικής μάθησης. Όπως και στο πρώτο πείραμα, έτσι και σε αυτό είναι κρίσιμο να αξιολογήσουμε τους αλγορίθμους ώστε να καταλήξουμε στον βέλτιστο βασιζόμενοι σε σημαντικούς δείκτες όπως η ακρίβεια κ.α. Η εξόρυξη και ανάλυση εκπαιδευτικών δεδομένων μπορούν να θεωρηθούν εξαιρετικά χρήσιμα εργαλεία για την εξέλιξη της εκπαιδευτικής διαδικασίας. Με την κατάλληλη αξιοποίηση τους, καθίσταται εφικτή η αναγνώριση πιθανών παραγόντων που συμβάλλουν στις ακαδημαϊκές επιδόσεις των μαθητών αλλά και η πρόληψη διαφόρων προβλημάτων που ταλανίζουν τον χώρο της εκπαίδευσης.

Λέξεις – Κλειδιά

Εξόρυξη Δεδομένων, Μηχανική μάθηση, Πρόβλεψη ακαδημαϊκών επιδόσεων.

Abstract

The remarkable progress of both machine learning and data mining in recent years has made a decisive contribution to the field of education through the production of new knowledge, which contributes to decision-making and the improvement of the educational process. In the context of this research, various machine learning techniques will be applied with the goal of predicting students' academic performance. The research is divided into two periods, before and after the Covid-19 period, and concerns secondary school students as well as university students. In the first part, the technique of linear regression and neural networks will be applied in order to predict the academic performance of male and female students in the Mathematics course. The models will then be evaluated through different performance indicators to select the one that performs best. In the second part, the interest will be focused on the Covid-19 period, during which university students answered various questions in order to collect the second set of data that will be used. The prediction in this part will be about the average score of the final period and will be achieved by applying five different machine learning algorithms. As in the first experiment, in this one it is crucial to evaluate the algorithms in order to arrive at the optimal one based on important indicators such as accuracy etc. Educational data mining and analysis can be considered extremely useful tools for the development of the educational process. With their proper utilization, it becomes possible to identify possible factors that contribute to the academic performance of students and to prevent various problems that plague the field of education.

Περιεχόμενα

Κατάλογος Εικόνων

1. Εισαγωγή

Την τελευταία δεκαετία η ραγδαία αύξηση των επιδόσεων των υπολογιστικών συστημάτων έφερε την επανάσταση στον τρόπο με τον οποίο αποθηκεύουμε και αξιοποιούμε τα δεδομένα. Εταιρίες, επιχειρήσεις και οργανισμοί έστρεψαν την προσοχή τους στη Ανακάλυψη Γνώσης και συγκεκριμένα στην Εξόρυξη Δεδομένων, αντιλαμβανόμενοι το γεγονός ότι με αυτόν τον τρόπο θα μπορούσαν να παράγουν χρήσιμα αποτελέσματα και προβλέψεις. Σημαντικές αλλαγές σημειώθηκαν και στον χώρο της εκπαίδευσης, όπου, καθημερινά, όλο και περισσότερα εκπαιδευτικά ιδρύματα ενδιαφέρονται να ανακαλύψουν χρήσιμες τεχνικές για την αξιοποίηση των εκπαιδευτικών δεδομένων. Στον τομέα της εκπαίδευσης η έρευνα αναπτύσσεται ολοένα και περισσότερο λόγω του τεράστιου αριθμού πληροφοριών των μαθητών που μπορούν να χρησιμοποιηθούν για την εφεύρεση πολύτιμων προτύπων που σχετίζονται με τη διαδικασία της μάθησης και τη συμπεριφορά των μαθητών. Ωστόσο, λόγω του μεγάλου όγκου δεδομένων μαθητών, η πρόβλεψη της ακαδημαϊκής επιτυχίας ενός μαθητή αποτελεί μια από τις πιο δύσκολες προκλήσεις. Η εξόρυξη δεδομένων σε συνδυασμό με την μηχανική μάθηση προσφέρει μια πληθώρα τεχνικών που μπορούν να χρησιμοποιηθούν ώστε να αξιοποιηθεί με τον κατάλληλο τρόπο ο τεράστιος όγκος δεδομένων που παράγονται από εκπαιδευτικά ιδρύματα. Βασικός στόχος είναι η εξαγωγή συμπερασμάτων, τα οποία θα συντελέσουν στην βελτίωση της εκπαιδευτικής διαδικασίας αλλά και στην αντιμετώπιση διαφόρων εκπαιδευτικών προβλημάτων όπως για παράδειγμα η σχολική διαρροή. Τα συμπεράσματα αυτά θα μπορούσαν να θεωρηθούν εξαιρετικά χρήσιμα και σημαντικά ειδικότερα κατά την περίοδο Covid-19, η οποία αποτέλεσε μια ιδιαίτερη πρόκληση για μαθητές και μαθήτριες.

Στην παρούσα έρευνα, θα χρησιμοποιηθούν διάφορες τεχνικές μηχανικής μάθησης και εξόρυξης δεδομένων με στόχο την πρόβλεψη των ακαδημαϊκών επιδόσεων μαθητών και μαθητριών από διάφορες εκπαιδευτικές βαθμίδες με βάση δημογραφικά και οικογενειακά χαρακτηριστικά. Η έρευνα χωρίζεται σε δύο μέρη, το πρώτο αφορά μαθητές και μαθήτριες της δευτεροβάθμιας εκπαίδευσης, ενώ το δεύτερο φοιτητές και φοιτήτριες Πανεπιστημίου κατά την περίοδο Covid-19. Για τον σκοπό αυτό, χρησιμοποιήθηκε το εργαλείο «Orange», στο οποίο πραγματοποιήθηκε η εισαγωγή, επεξεργασία και οπτικοποίηση των δεδομένων και στην συνέχεια η ανάλυση και αξιολόγηση των επιδόσεων των μοντέλων όπου έγινε η διερεύνηση.

2. Μηχανική μάθηση και Εξόρυξη Δεδομένων

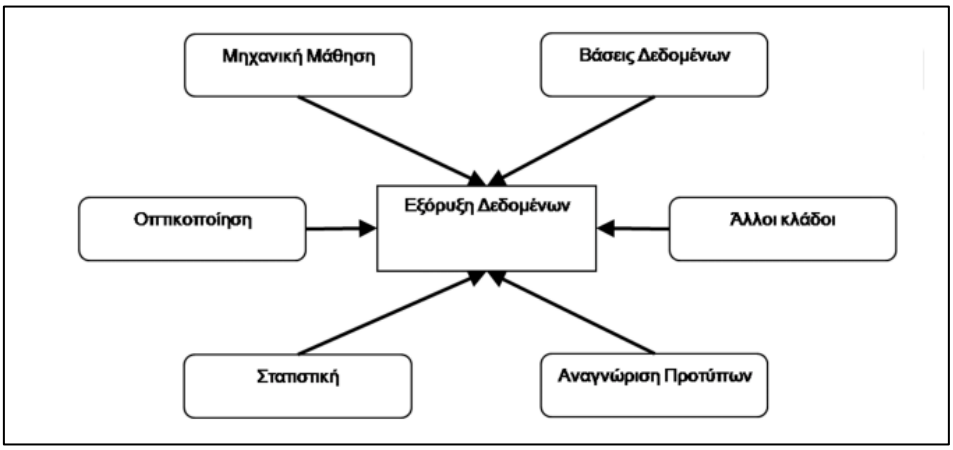
Μηχανική μάθηση (Machine Learning) ονομάζεται το πεδίο της επιστήμης των υπολογιστών που μελετά τη δημιουργία αλγορίθμων οι οποίοι «μαθαίνουν» χωρίς να έχουν προγραμματιστεί με συγκεκριμένους κανόνες. Με άλλα λόγια, οι αλγόριθμοι αυτοί χρησιμοποιούν δεδομένα με σκοπό να ανακαλύψουν μοτίβα και σχέσεις ώστε να κάνουν προβλέψεις ή να πάρουν αποφάσεις. Αναφορές στη μηχανική μάθηση υπάρχουν από τη δεκαετία του 1960, όμως η χρήση των τεχνικών αυτών αυξήθηκε ραγδαία μετά τη δεκαετία του 1990 ως αποτέλεσμα της ανάπτυξης κλάδων της επιστήμης των υπολογιστών. Υπάρχουν τρία διαφορετικά είδη μηχανικής μάθησης, τα οποία παρουσιάζονται παρακάτω:

1) *Επιβλεπόμενη (supervised) μάθηση* ονομάζεται η τεχνική με την οποία ένα πρόγραμμα εκπαιδεύεται για να καταλάβει τη σχέση μεταξύ των δεδομένων

που δίνουμε και ενός επιθυμητού αποτελέσματος. Δηλαδή έχουμε προκαθορισμένη είσοδο (input – δεδομένα) και έξοδο (output – αποτέλεσμα). Αλλιώς, συνηθίζεται να λέμε ότι έχουμε δεδομένα με ετικέτες (labels) που δείχνουν τη σύνδεση με την έξοδο, τις οποίες έχουν βάλει άνθρωποι ή άλλοι κώδικες (επιβλεπόμενη). Σε αυτή τη μορφή μάθησης σκοπός του προγράμματος είναι να καταλάβει τη σχέση μεταξύ εισόδου και εξόδου.

2) Στην περίπτωση της *μη-επιβλεπόμενης (unsupervised) μάθησης* γνωστό είναι μόνο το κομμάτι της εισόδου (input) των δεδομένων και ο υπολογιστής καλείται να αναγνωρίσει τα μοτίβα που μπορεί να υπάρχουν. Μία από τις πιο συνηθισμένες εφαρμογές είναι η ομαδοποίηση (clustering). Σε αυτή την περίπτωση τα δεδομένα κατηγοριοποιούνται σε ομάδες (clusters) που έχουν κοινά στοιχεία/πληροφορίες.
3) Στην *ενισχυτική μάθηση (reinforcement learning)* τα πράγματα είναι κάπως διαφορετικά. Εδώ, κατασκευάζουμε ένα εικονικό «περιβάλλον» που έχει συγκεκριμένους κανόνες και αφήνουμε τον υπολογιστή να αλληλοεπιδράσει με αυτό μέχρι την επίτευξη κάποιου στόχου, όπως για παράδειγμα η μεγιστοποίηση ενός σκορ.

Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, ωστόσο η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων. Τα τελευταία χρόνια, έχει παρατηρηθεί μια εκρηκτική αύξηση του όγκου των δεδομένων, τα οποία ο ανθρώπινος νους είναι δύσκολο να επεξεργαστεί αποτελεσματικά. Η επιστήμη της Στατιστικής προσφέρει λύσεις ανάλυσης δεδομένων, δεν λαμβάνει όμως μέριμνα για το πρόβλημα του πολύ μεγάλου όγκου τους. Επίσης η Μηχανική Μάθηση και η Αναγνώριση Προτύπων διαθέτουν τις δικές τους μεθοδολογίες, όμως και πάλι δεν αντιμετωπίζουν το πρόβλημα του όγκου των δεδομένων. Ο κλάδος των Βάσεων Δεδομένων είναι ο κατ' εξοχήν αρμόδιος για την τήρηση μεγάλου όγκου δεδομένων, όμως η σχεδιαστική φιλοσοφία του είναι προσανατολισμένη στην καταχώρηση, στη διαχείριση και στην ανάκτηση των δεδομένων, όχι όμως και στην ανάλυση τους. Η Εξόρυξη Δεδομένων αποτελεί τέκνο της ανάγκης για επεξεργασία των αποθηκευμένων δεδομένων και εξαγωγή χρήσιμης πληροφορίας. Αντλώντας μεθοδολογίες από όλους τους επιστημονικούς κλάδους που αναφέρθηκαν παραπάνω, καθώς και από άλλους, όπως η Οπτικοποίηση, στοχεύει στην ανακάλυψη πολύτιμης γνώσης, που είναι κρυμμένη σε μεγάλους όγκους δεδομένων.



Εικόνα 1: Η Εξόρυξη Δεδομένων ως αποτέλεσμα συμβολής άλλων κλάδων.

Σύμφωνα με τους Witten and Frank (2000), η Εξόρυξη Δεδομένων (Data Mining) ορίζεται ως η διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, δίνοντας έτσι έμφαση στη διάσταση της Μηχανικής Μάθησης. Σύμφωνα με τους Han and Kamber (2001), η Εξόρυξη Δεδομένων συνίσταται στην ανακάλυψη ή «εξόρυξη» γνώσης από μεγάλους όγκους δεδομένων. Ο ορισμός αυτός τονίζει τη διάσταση του όγκου των δεδομένων. Άλλοι συγγραφείς, όπως οι Maimon and Rokach (2005), χρησιμοποιούν τον όρο Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) για τη συνολική διαδικασία ανακάλυψης προτύπων μέσα από μεγάλα και περίπλοκα σύνολα δεδομένων. Σύμφωνα με το σκεπτικό αυτό, η ανακάλυψη γνώσης από τα δεδομένα συνίσταται σε μια διαδικασία, που ξεκινά από τα πηγαία δεδομένα και καταλήγει στην τελική διατύπωση συμπερασμάτων και στη λήψη αποφάσεων, μέσα από μια αλληλουχία διαδοχικών σταδίων. Η Εξόρυξη Δεδομένων αποτελεί ένα από τα στάδια αυτής της διαδικασίας και συνίσταται στον σκληρό πυρήνα της. Περιλαμβάνει την εφαρμογή αλγορίθμων και την κατασκευή μοντέλων, τα οποία στοχεύουν στην ανακάλυψη και εξαγωγή προτύπων (patterns).

3. Το περιβάλλον Orange

Το Orange είναι ένα πακέτο λογισμικού οπτικού προγραμματισμού το οποίο προσφέρει στους χρήστες μια μεγάλη, ποικιλόμορφη εργαλειοθήκη για οπτικοποίηση δεδομένων, μηχανική μάθηση, εξόρυξη δεδομένων και ανάλυση δεδομένων. Περιλαμβάνει μια πληθώρα από γραφικά στοιχεία, τα οποία ονομάζονται widgets και κυμαίνονται από την απλή οπτικοποίηση δεδομένων, την επιλογή υποσυνόλων και την προεπεξεργασία, έως την εμπειρική αξιολόγηση των αλγορίθμων εκμάθησης και την προγνωστική μοντελοποίηση. Ο οπτικός προγραμματισμός υλοποιείται μέσω μιας διεπαφής, στην οποία δημιουργούνται ροές εργασίας με σύνδεση προκαθορισμένων ή σχεδιασμένων γραφικών στοιχείων από τον χρήστη, ενώ οι προχωρημένοι χρήστες μπορούν να χρησιμοποιήσουν το Orange ως βιβλιοθήκη Python για χειρισμό δεδομένων και τροποποίηση γραφικών στοιχείων. Στα πλαίσια της παρούσας έρευνας, το Orange χρησιμοποιείται για την πρόβλεψη ακαδημαϊκών επιδόσεων σε δύο διαφορετικά σύνολα δεδομένων, ένα για την περίοδο Covid -19 και ένα πριν από αυτήν. Θα πραγματοποιηθεί η εισαγωγή των δεδομένων και στην συνέχεια η οπτικοποίηση και ανάλυση τους ώστε να εκπαιδευτούν με αυτά τα μοντέλα που θα χρησιμοποιηθούν. Έπειτα, τα μοντέλα θα αξιολογηθούν με διάφορους δείκτες, οι οποίοι θα μας οδηγήσουν στο να αποφασιστεί ποιο μοντέλο είναι αυτό που είναι πιο αποτελεσματικό. Το Orange διαθέτει μια πληθώρα γραφικών στοιχείων που θα μας δώσουν χρήσιμα στοιχεία για τα μοντέλα, όπως για παράδειγμα οι ROC καμπύλες και οι πίνακες σύγχυσης.

4. Εφαρμογή τεχνικών μάθησης και νευρωνικών δικτύων για την πρόβλεψη ακαδημαϊκών επιδόσεων μαθητών

4.1 Πληροφορίες συνόλου δεδομένων

Το πρώτο σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα έρευνα αφορά μαθητές και μαθήτριες της δευτεροβάθμιας εκπαίδευσης δύο σχολείων στην Πορτογαλία. Συγκεκριμένα, τα χαρακτηριστικά των δεδομένων περιλαμβάνουν βαθμούς μαθητών, δημογραφικά, κοινωνικά και σχολικά χαρακτηριστικά και

συλλέχθηκαν με τη χρήση σχολικών εκθέσεων και ερωτηματολογίων. Το σύνολο δεδομένων προσεγγίζει τις επιδόσεις των μαθητών στο μάθημα των Μαθηματικών. Παρακάτω δίνεται η πλήρης περιγραφή των χαρακτηριστικών του συνόλου δεδομένων:

1) **school** – το σχολείο που φοιτούν οι μαθητές/τριες ('GP' - Gabriel Pereira ή 'MS' - Mousinho da Silveira)
2) **sex** – το φύλο των μαθητών/τριών ('F' - θηλυκό or 'M' - αρσενικό)
3) **age** – η ηλικία των μαθητών/τριων (numeric: from 15 to 22)
4) **address** – η περιοχή κατοικίας των μαθητών/τριων ('U' – αστικό κέντρο ή 'R' – αγροτική περιοχή)
5) **famsize** – το πλήθος μελών της οικογένειας των μαθητών/τριών ('LE3' – μικρότερο ή ίσο του 3 ή 'GT3' – μεγαλύτερο από 3)
6) **Pstatus** – κατάσταση συμβίωσης γονέων των μαθητών/τριών ('T' – ζουν μαζί ή 'A' – ζουν χωριστά)
7) **Medu** – το εκπαιδευτικό επίπεδο των μητέρων των μαθητών/τριών (0 - καμία, 1 - πρωτοβάθμια εκπαίδευση ($4^{η}$ βαθμίδα), 2 ($5^{η}$ έως $9^{η}$ βαθμίδα), 3 (δευτεροβάθμια εκπαίδευση) or 4 (ανώτατη εκπαίδευση)
8) **Fedu** – το εκπαιδευτικό επίπεδο των πατέρων των μαθητών/τριών (0 - καμία, 1 - πρωτοβάθμια εκπαίδευση (4η βαθμίδα), 2 (5η έως 9η βαθμίδα), 3 (δευτεροβάθμια εκπαίδευση) or 4 (ανώτατη εκπαίδευση)
9) **Mjob** – το επάγγελμα των μητέρων των μαθητών/τριών ('teacher'-δασκάλα, 'health'- επάγγελμα σχετικό με την υγεία, 'services' (πολιτικές υπηρεσίες, για παράδειγμα αστυνομικός, 'at_home'- οικιακά ή 'other'-άλλο)
10) **Fjob** - το επάγγελμα των πατέρων των μαθητών/τριών ('teacher'-δάσκαλος, 'health'- επάγγελμα σχετικό με την υγεία, 'services' (πολιτικές υπηρεσίες, για παράδειγμα αστυνομικός, 'at_home'- οικιακά ή 'other'-άλλο)
11) **reason** – ο λόγος για τον οποίον έχουν επιλέξει το σχολείο τους οι μαθητές/τριες ('home' – βρίσκεται κοντά στο σπίτι, 'reputation' – για την φήμη του σχολείου, 'course' – για το διδακτικό πρόγραμμα του σχολείου ή 'other' – άλλο)
12) **guardian** – ο κηδεμόνας των μαθητών/τριών ('mother'-μητέρα, 'father'-πατέρας ή 'other'-άλλο)
13) **traveltime** – ο χρόνος της διαδρομής από το σπίτι των μαθητών/τριών μέχρι το σχολείο τους (1 – μικρότερο από 15 λεπτά, 2 - 15 έως 30 λεπτά, 3 - 30 λεπτά έως 1 ώρα ή 4 – μεγαλύτερο από 1 ώρα)
14) **studytime** – ο εβδομαδιαίος χρόνος μελέτης των μαθητών/τριών (1 – μικρότερο από 2 ώρες, 2 – από 2 έως 5 ώρες, 3 – από 5 έως 10 ώρες, ή 4 – μεγαλύτερο από 10 ώρες)
15) **failures** – ο αριθμός αποτυχιών της προηγούμενης τάξης των μαθητών/τριών (n εάν ισχύει 1<=n<3, διαφορετικά 4)
16) **schoolsup** - επιπλέον εκπαιδευτική υποστήριξη (yes - ναι ή no - όχι)
17) **famsup** - οικογενειακή εκπαιδευτική υποστήριξη (yes - ναι ή no - όχι)
18) **paid** - επιπλέον αμειβόμενα μαθήματα εντός του σχολικού προγράμματος (yes - ναι ή no - όχι)
19) **activities** – συμμετοχή των μαθητών/τριών σε εξωσχολικές δραστηριότητες (yes - ναι ή no - όχι)

20) **nursery** – ο/η μαθητής/τρια έχει φοιτήσει σε νηπιαγωγείο (yes - ναι ή no - όχι)
21) **higher** – ο/η μαθητής/τρια θέλει να φοιτήσει στην τριτοβάθμια εκπαίδευση (yes - ναι ή no - όχι)
22) **internet** - Πρόσβαση στο διαδίκτυο στο σπίτι των μαθητών/τριών (yes - ναι ή no - όχι)
23) **romantic** – ο/η μαθητής/τρια έχει μια ρομαντική σχέση (yes - ναι ή no - όχι)
24) **famrel** - ποιότητα των οικογενειακών σχέσεων των μαθητών/τριών (από 1 – πολύ κακές μέχρι 5 - εξαιρετικές)
25) **freetime** - ελεύθερος χρόνος μετά το σχολείο (από 1 – πολύ λίγος έως 5 - πάρα πολύς)
26) **goout** – ο μαθητής/τρια αφιερώνει χρόνο σε εξόδους (από 1 – πολύ λίγο έως 5 – πάρα πολύ)
27) **Dalc** – κατανάλωση αλκοόλ κατά τις εργάσιμες ημέρες από τους μαθητές/τριες (από 1 – πολύ λίγο έως 5 – πάρα πολύ)
28) **Walc** – κατανάλωση αλκοόλ τα Σαββατοκύριακα από τους μαθητές/τριες (από 1 – πολύ λίγο έως 5 – πάρα πολύ)
29) **health** - τρέχουσα κατάσταση υγείας (από 1- πολύ κακή έως 5- πολύ κακή)
30) **absences** - αριθμός σχολικών απουσιών (από 0 έως 93)

Επιπλέον στο σύνολο δεδομένων υπάρχουν 3 βαθμοί που σχετίζονται με το μάθημα των Μαθηματικών:

31) **G1** – βαθμός πρώτης περιόδου (από 0 έως 20)
32) **G2** – βαθμός δεύτερης περιόδου (από 0 έως 20)
33) **G3** – τελικός βαθμός (από 0 έως 20)

Ο βαθμός G3, ο τελικός βαθμός του μαθήματος των Μαθηματικών για κάθε μαθητή/τρια είναι ο στόχος, τον οποίον θέλουμε να προβλέψουμε. Είναι σημαντικό να αναφερθεί ότι το χαρακτηριστικό στόχος, G3, έχει ισχυρή συσχέτιση με τα χαρακτηριστικά G2 και G1. Αυτό συμβαίνει επειδή το G3 είναι ο βαθμός του τελευταίου έτους (που εκδίδεται στην 3η περίοδο), ενώ οι G1 και G2 αντιστοιχούν στους βαθμούς 1ης και 2ης περιόδου. Όπως είναι προφανές, είναι πιο δύσκολο να προβλεφθεί το G3 χωρίς την ύπαρξη των G2 και G1, ωστόσο μια τέτοια πρόβλεψη θα ήταν αρκετά χρήσιμη.

4.2 Σκοπός

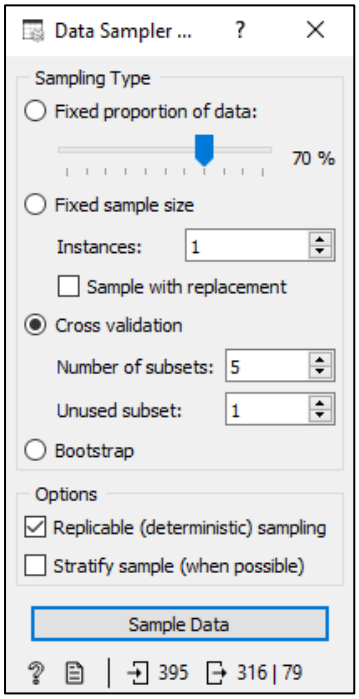
Σκοπός της έρευνας είναι η πρόβλεψη του χαρακτηριστικού στόχου, G3, δηλαδή η πρόβλεψη του τελικού βαθμού του μαθήματος των Μαθηματικών για κάθε μαθητή και μαθήτρια. Η πρόβλεψη αυτή θα επιτευχθεί μέσω τεχνικών μηχανικής μάθησης, όπως θα παρουσιαστεί στα κεφάλαια που ακολουθούν. Συγκεκριμένα, με την βοήθεια του περιβάλλοντος Orange, θα γίνει μια διερεύνηση με χρήση γραμμικής παλινδρόμησης και νευρωνικών δικτύων, ώστε να προβλέψουμε τις ακαδημαϊκές επιδόσεις των μαθητών με την μεγαλύτερη ακρίβεια. Τα μοντέλα που θα χρησιμοποιηθούν για την

πρόβλεψη, θα αξιολογηθούν με διάφορους δείκτες ώστε να βρεθεί το πιο αποτελεσματικό και ακριβές μοντέλο.

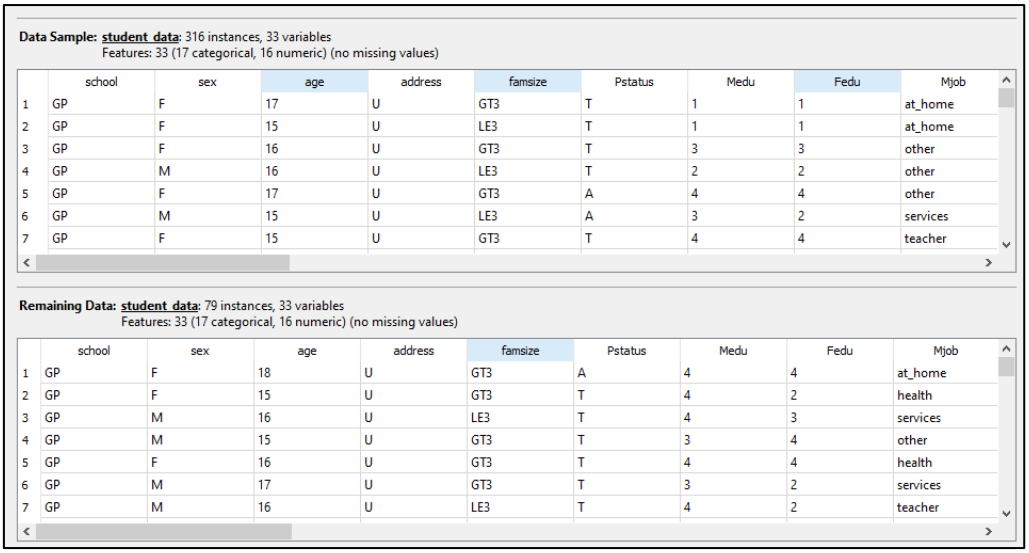
4.3 Διερεύνηση τεχνικών μηχανικής μάθησης για την πρόβλεψη των ακαδημαϊκών επιδόσεων των μαθητών

4.3.1 Διαχωρισμός του συνόλου δεδομένων σε σετ εκπαίδευσης και σετ δοκιμής

Το σύνολο των αλγόριθμων που θα χρησιμοποιηθούν θα αξιολογηθούν με διασταυρωμένη επικύρωση 5 ίσων πτυχών (5-Fold Cross Validation). Η συγκεκριμένη μέθοδος είναι ευρεία γνωστή και είναι μια αποτελεσματική στρατηγική διαχωρισμού των δεδομένων ώστε να χτιστεί ένα περισσότερο γενικευμένο μοντέλο. Ο κύριος σκοπός της εκτέλεσης οποιουδήποτε είδους μηχανικής μάθησης είναι η ανάπτυξη ενός πιο γενικευμένου μοντέλου που μπορεί να έχει καλή απόδοση σε άγνωστα δεδομένα. Θα μπορούσε να δημιουργηθεί ένα τέλειο μοντέλο στα δεδομένα εκπαίδευσης με 100% ακρίβεια ή 0 σφάλμα, αλλά μπορεί να αποτύχει να γενικεύσει για άγνωστα δεδομένα. Το μοντέλο αυτό δεν είναι τόσο αποτελεσματικό, καθώς ταιριάζει υπερβολικά στα δεδομένα προπόνησης. Η μηχανική μάθηση στοχεύει στη γενίκευση που σημαίνει ότι η απόδοση του μοντέλου μπορεί να μετρηθεί μόνο με δεδομένα που δεν έχουν χρησιμοποιηθεί ποτέ κατά τη διαδικασία εκπαίδευσης. Αυτός είναι ο λόγος για τον οποίο συχνά χωρίζουμε τα δεδομένα μας σε ένα σετ εκπαίδευσης και σε ένα σετ δοκιμών. Παρακάτω, παρουσιάζεται ο διαχωρισμός των δεδομένων του πρώτου συνόλου στο περιβάλλον Orange καθώς το πλήθος των τιμών που περιλαμβάνει το σετ εκπαίδευσης και το σετ δοκιμής.



Εικόνα 2: Διαχωρισμός του συνόλου δεδομένων με την μέθοδο Cross Validation



Data Sample: student_data: 316 instances, 33 variables
Features: 33 (17 categorical, 16 numeric) (no missing values)

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob
1	GP	F	17	U	GT3	T	1	1	at_home
2	GP	F	15	U	LE3	T	1	1	at_home
3	GP	F	16	U	GT3	T	3	3	other
4	GP	M	16	U	LE3	T	2	2	other
5	GP	F	17	U	GT3	A	4	4	other
6	GP	M	15	U	LE3	A	3	2	services
7	GP	F	15	U	GT3	T	4	4	teacher

Remaining Data: student_data: 79 instances, 33 variables
Features: 33 (17 categorical, 16 numeric) (no missing values)

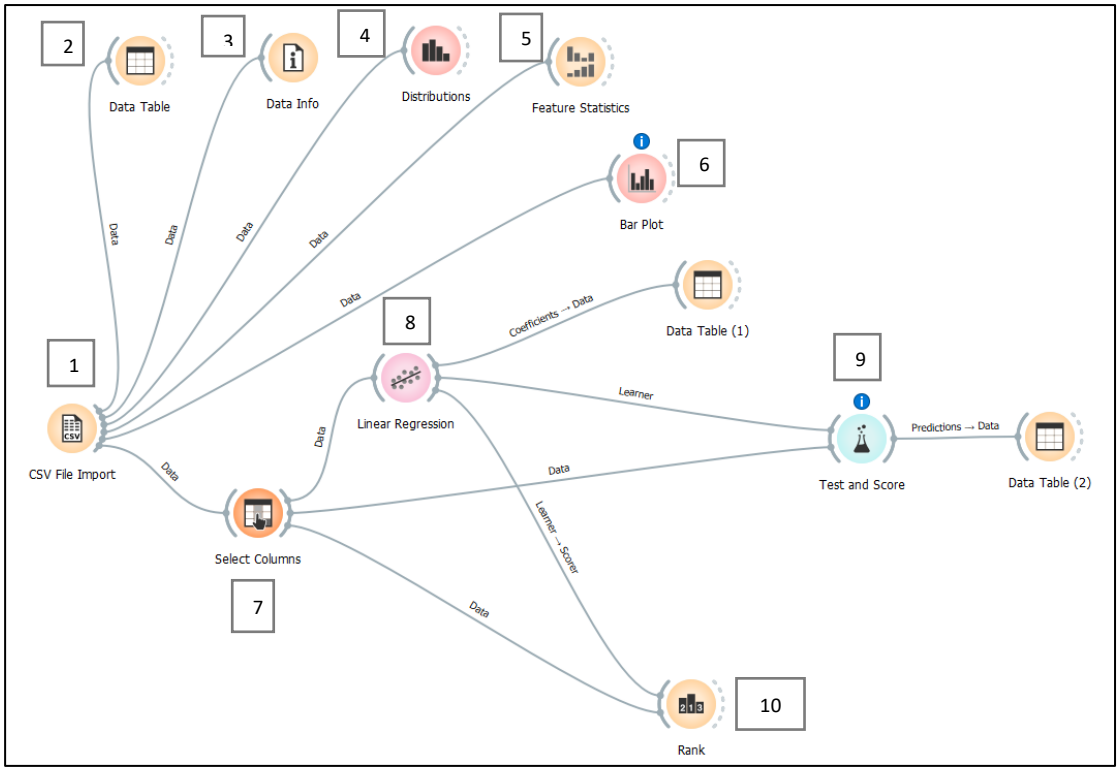
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob
1	GP	F	18	U	GT3	A	4	4	at_home
2	GP	F	15	U	GT3	T	4	2	health
3	GP	M	16	U	LE3	T	4	3	services
4	GP	M	15	U	GT3	T	3	4	other
5	GP	F	16	U	GT3	T	4	4	health
6	GP	M	17	U	GT3	T	3	2	services
7	GP	M	16	U	LE3	T	4	2	teacher

Εικόνα 3: Σετ εκπαίδευσης και σετ δοκιμής

4.3.2 Γραμμική Παλινδρόμηση - Υλοποίηση στο περιβάλλον Orange

Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (Simple Linear Regression), κατά την οποία υπάρχει μόνο μια ανεξάρτητη μεταβλητή X (Independent or Input Variable), και η εξαρτημένη μεταβλητή Y (Dependent or Response Variable), η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X.

Αρχικός στόχος της έρευνας είναι να προβλέψουμε τον χαρακτηριστικό στόχο G3, μέσω της γραμμικής παλινδρόμησης. Ο απώτερος σκοπός είναι να προβλέψουμε τον βαθμό του τελευταίου έτους ($3^{η}$ περίοδο) στα Μαθηματικά για κάθε μαθητή/τρια, έχοντας τους βαθμούς G1, G2 της $1^{ης}$ και της $2^{ης}$ περιόδου, με όσο το δυνατόν μικρότερο σφάλμα. Παρακάτω, φαίνεται η υλοποίηση αυτή στο περιβάλλον Orange και στην συνέχεια αναλύονται τα βήματα που οδήγησαν στην υλοποίηση αυτή.

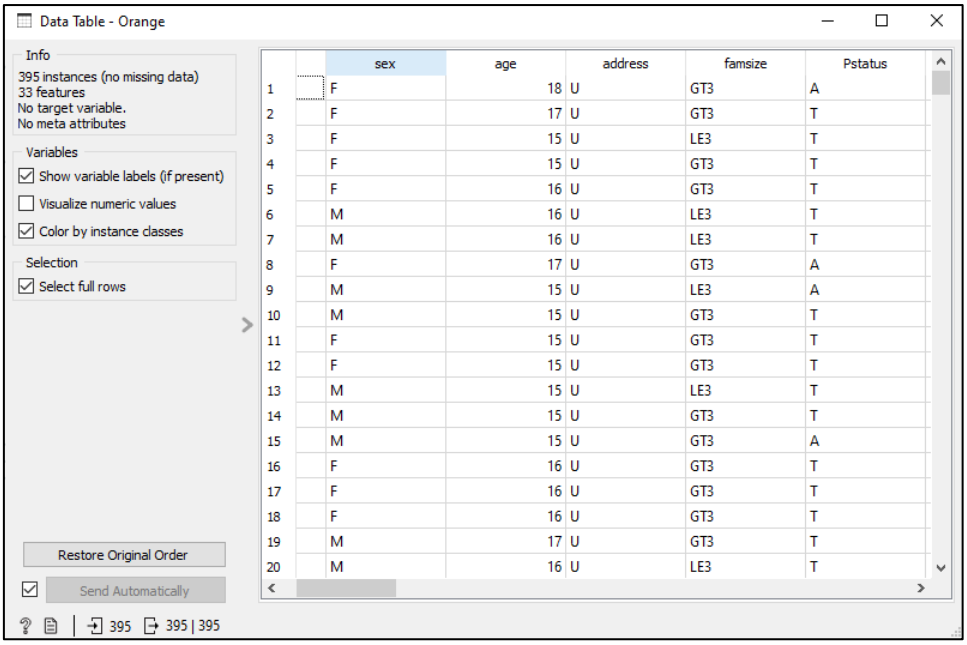


Εικόνα 4: Εφαρμογή Γραμμικής Παλινδρόμησης για την πρόβλεψη της τιμής G3 στο περιβάλλον Orange

4.3.3 Βήματα

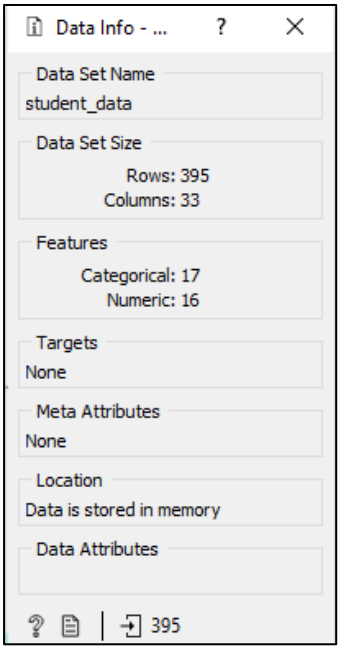
1. Αρχικά γίνεται η φόρτωση του συνόλου δεδομένων στο περιβάλλον του Orange. Το γραφικό στοιχείο «Εισαγωγή αρχείου CSV» (CSV File Import) διαβάζει αρχεία διαχωρισμένα με κόμματα και στέλνει το σύνολο δεδομένων στο κανάλι εξόδου του. Τα διαχωριστικά αρχείων μπορεί να είναι κόμματα, ερωτηματικά, κενά, καρτέλες ή οριοθέτες που ορίζονται με το χέρι. Το ιστορικό των πιο πρόσφατα ανοιγμένων αρχείων διατηρείται στο γραφικό στοιχείο.

2. Το γραφικό στοιχείο «Πίνακας Δεδομένων» (Data Table) λαμβάνει ένα ή περισσότερα σύνολα δεδομένων στην εισαγωγή του και τα παρουσιάζει ως υπολογιστικό φύλλο. Τα στιγμιότυπα δεδομένων μπορούν να ταξινομηθούν κατά τιμές χαρακτηριστικών. Το γραφικό στοιχείο υποστηρίζει επίσης τη μη αυτόματη επιλογή παρουσιών δεδομένων. Στην παρακάτω εικόνα φαίνεται η παρουσίαση των δεδομένων που ανεβάσαμε στο Orange σε μορφή υπολογιστικού φύλλου.



Εικόνα 5: Η παρουσίαση του συνόλων δεδομένων σε μορφή υπολογιστικού φύλου

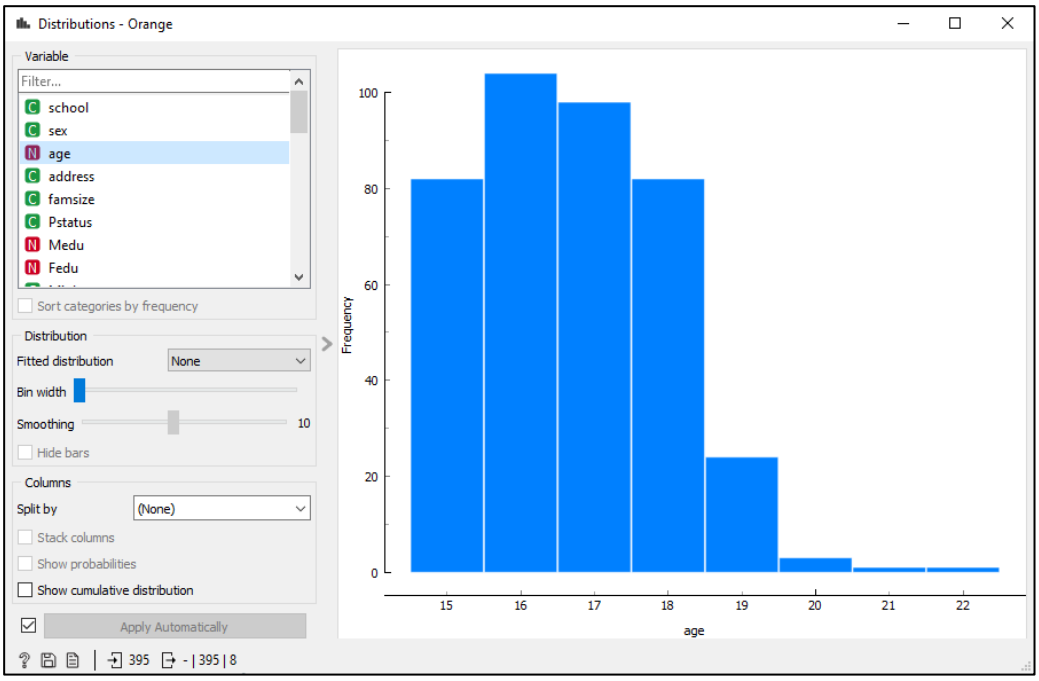
3. Το γραφικό στοιχείο «Data Info» είναι ένα απλό γραφικό στοιχείο που παρουσιάζει πληροφορίες σχετικά με το μέγεθος των δεδομένων, τα χαρακτηριστικά, τους στόχους, τα μετα-χαρακτηριστικά και την τοποθεσία. Οι πληροφορίες του συγκεκριμένου συνόλου δεδομένων φαίνονται στην παρακάτω εικόνα.



Εικόνα 6: Τα χαρακτηριστικά του συνόλου δεδομένων

4. Το γραφικό στοιχείο «Distributions» εμφανίζει την κατανομή τιμών διακριτών ή συνεχών χαρακτηριστικών. Εάν τα δεδομένα περιέχουν μια μεταβλητή κλάσης, οι διανομές μπορούν να εξαρτηθούν από την κλάση. Για το

συγκεκριμένο σύνολο δεδομένων έχουμε τα εξής αποτελέσματα (αφορά την μεταβλητή age):

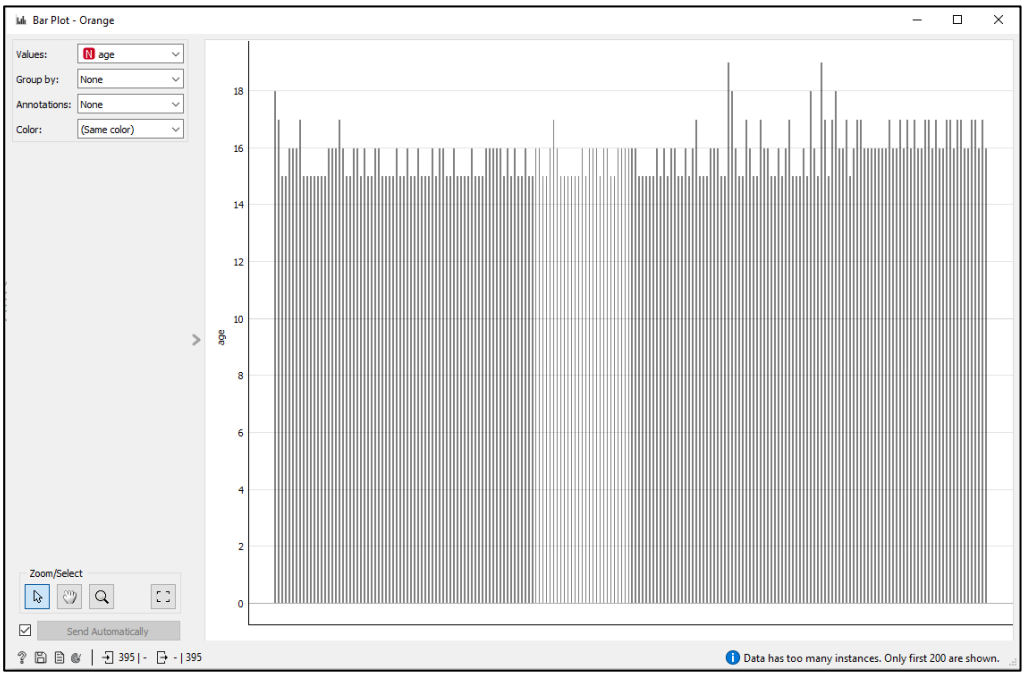


Εικόνα 7: Η κατανομή του χαρακτηριστικού Age

5. Το γραφικό στοιχείο «Feature Statistic» εμφανίζει βασικά στατιστικά στοιχεία για το σύνολο δεδομένων. Συγκεκριμένα, όπως φαίνεται στην παρακάτω εικόνα έχουμε την κατανομή, τον μέσο όρο, την μέγιστη και ελάχιστη παρατήρηση για κάθε μεταβλητή που περιέχεται στο σύνολο δεδομένων.

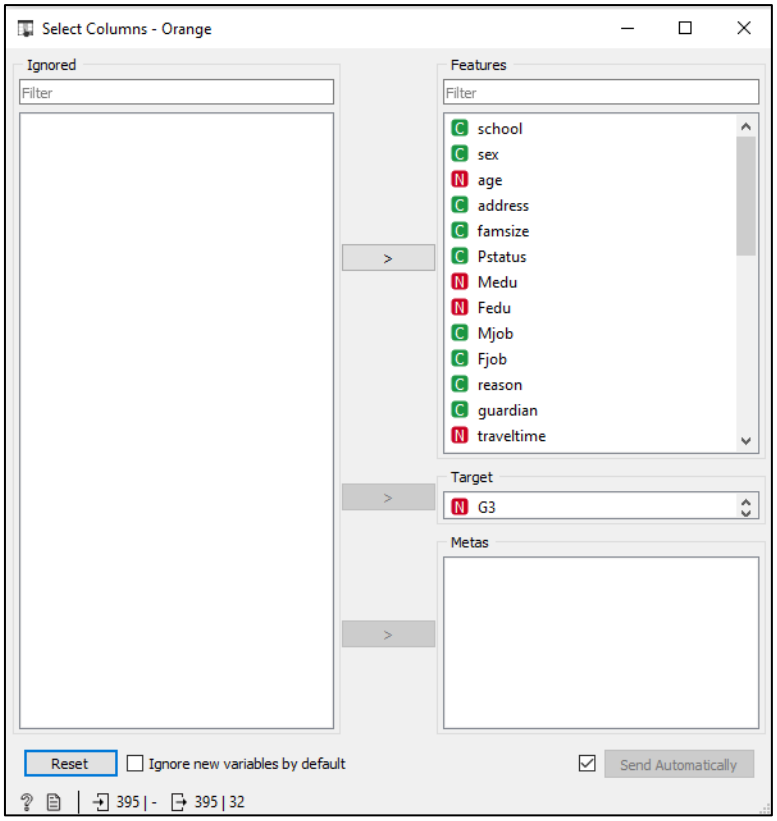
Εικόνα 8: Βασικά στατικά στοιχεία για το σύνολο δεδομένων

6. Το γραφικό στοιχείο «Bar Plot» οπτικοποιεί αριθμητικές μεταβλητές και τις συγκρίνει με μια κατηγορική μεταβλητή. Το γραφικό στοιχείο είναι χρήσιμο για την παρατήρηση ακραίων τιμών, τις κατανομές εντός ομάδων και τη σύγκριση κατηγοριών. Για το συγκεκριμένο σύνολο δεδομένων έχουμε τα εξής αποτελέσματα (αφορά την μεταβλητή age):



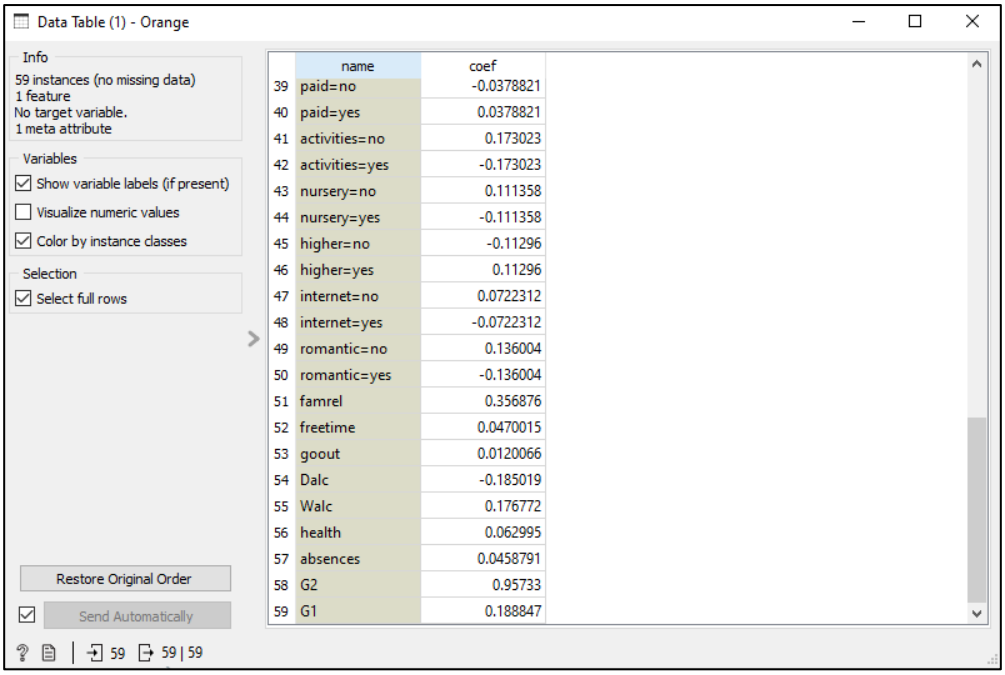
Εικόνα 9: Οπτικοποίηση των τιμών του χαρακτηριστικού Age

7. Το γραφικό στοιχείο «Select Columns» χρησιμοποιείται για τη μη αυτόματη σύνθεση του τομέα δεδομένων. Ο χρήστης μπορεί να αποφασίσει ποια χαρακτηριστικά θα χρησιμοποιηθούν και πώς. Το γραφικό στοιχείο «Select Columns» είναι απαραίτητο για την γραμμική παλινδρόμηση, καθώς εκεί ορίζουμε ότι ο βαθμός G3 είναι το χαρακτηριστικό στόχος.



Εικόνα 10: Καθορισμός του χαρακτηριστικού στόχου G3

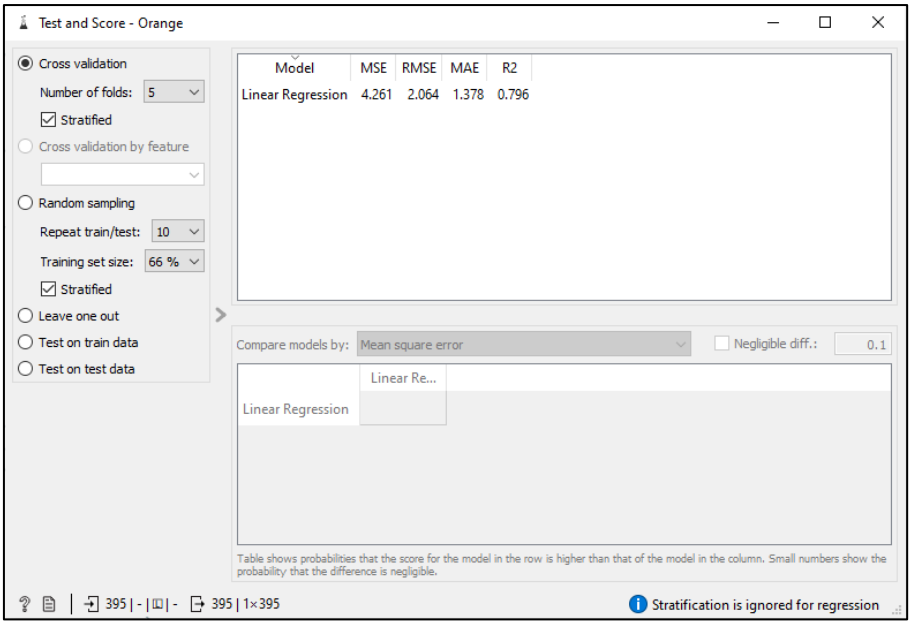
8. Το γραφικό στοιχείο «Γραμμικής παλινδρόμηση» (Linear Regression) κατασκευάζει μια γραμμική συνάρτηση με στόχο την πρόβλεψη μιας συγκεκριμένης μεταβλητής μέσω της μάθησης από τα δεδομένα εισόδου του. Το μοντέλο μπορεί να προσδιορίσει τη σχέση μεταξύ ενός προγνωστικού xi και της μεταβλητής απόκρισης y. Στην παρακάτω φωτογραφία, στην στήλη coef φαίνονται οι συντελεστές γραμμικής παλινδρόμησης. Οι συντελεστές γραμμικής παλινδρόμησης περιγράφουν τη μαθηματική σχέση μεταξύ κάθε ανεξάρτητης μεταβλητής και εξαρτημένης μεταβλητής. Οι τιμές p για τους συντελεστές υποδεικνύουν εάν αυτές οι σχέσεις είναι στατιστικά σημαντικές.



	name	coef
39	paid=no	-0.0378821
40	paid=yes	0.0378821
41	activities=no	0.173023
42	activities=yes	-0.173023
43	nursery=no	0.111358
44	nursery=yes	-0.111358
45	higher=no	-0.11296
46	higher=yes	0.11296
47	internet=no	0.0722312
48	internet=yes	-0.0722312
49	romantic=no	0.136004
50	romantic=yes	-0.136004
51	famrel	0.356876
52	freetime	0.0470015
53	goout	0.0120066
54	Dalc	-0.185019
55	Walc	0.176772
56	health	0.062995
57	absences	0.0458791
58	G2	0.95733
59	G1	0.188847

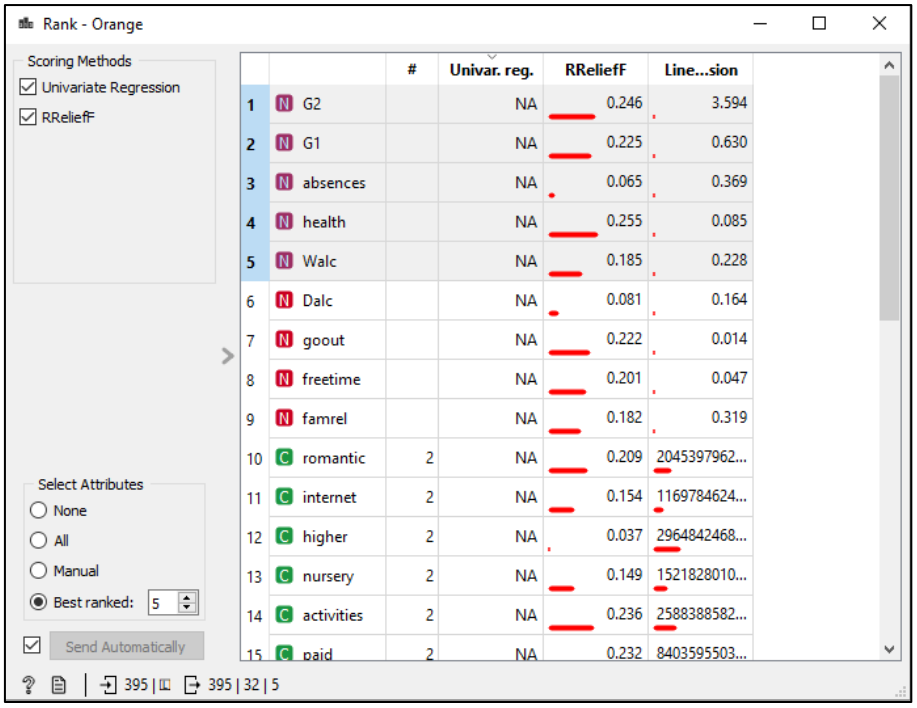
Εικόνα 11: Οι συντελεστές γραμμικής παλινδρόμησης

9. Το γραφικό στοιχείο «Test and Score» δοκιμάζει αλγόριθμους εκμάθησης. Διατίθενται διαφορετικά σχήματα δειγματοληψίας, συμπεριλαμβανομένης της χρήσης ξεχωριστών δεδομένων δοκιμών. Το συγκεκριμένο γραφικό στοιχείο κάνει δύο πράγματα. Πρώτον, δείχνει έναν πίνακα με διαφορετικά μέτρα απόδοσης ταξινομητή, όπως η ακρίβεια ταξινόμησης και η περιοχή κάτω από την καμπύλη. Δεύτερον, εξάγει αποτελέσματα αξιολόγησης, τα οποία μπορούν να χρησιμοποιηθούν από άλλα γραφικά στοιχεία για την ανάλυση της απόδοσης των ταξινομητών, όπως η ανάλυση ROC ή η μήτρα σύγχυσης. Στην παρακάτω φωτογραφία μπορεί κανείς να βρει σημαντικά στατιστικά στοιχεία όπως το MSE, το οποίο μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων—δηλαδή τη μέση τετραγωνική διαφορά μεταξύ των εκτιμώμενων τιμών και της πραγματικής αξίας.



Εικόνα 12: Αποτελέσματα γραφικού στοιχείου "Test and Score" για την γραμμική παλινδρόμηση

10. Το γραφικό στοιχείο «Rank» βαθμολογεί τις μεταβλητές σύμφωνα με τη συσχέτισή τους με τη διακριτή ή αριθμητική μεταβλητή στόχου, με βάση τους ισχύοντες εσωτερικούς βαθμολογητές (όπως κέρδος πληροφοριών, χι-τετράγωνο και γραμμική παλινδρόμηση) και τυχόν συνδεδεμένα εξωτερικά μοντέλα που υποστηρίζουν βαθμολόγηση, όπως γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, τυχαίο δάσος, SGD, κ.λπ. Παρατηρούμε πως για το συγκεκριμένο σύνολο δεδομένων τα 5 σημαντικότερα χαρακτηριστικά, με βάση την συσχέτιση τους με το χαρακτηριστικό στόχο G3, είναι τα G2, G1, absences, health και Walc. Σημειώνεται πως η στήλη «RreliefF» υπολογίζει την σχετική απόσταση μεταξύ των προβλεπόμενων τιμών (κλάσης) των δύο περιπτώσεων.

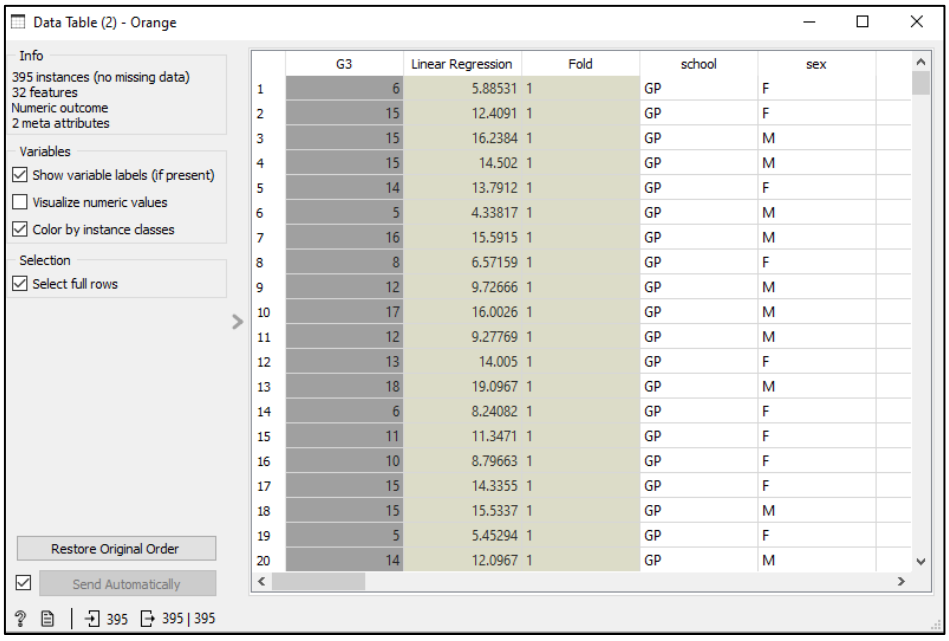


		#	Univar. reg.	RReliefF	Line...sion
1	G2		NA	0.246	3.594
2	G1		NA	0.225	0.630
3	absences		NA	0.065	0.369
4	health		NA	0.255	0.085
5	Walc		NA	0.185	0.228
6	Dalc		NA	0.081	0.164
7	goout		NA	0.222	0.014
8	freetime		NA	0.201	0.047
9	famrel		NA	0.182	0.319
10	romantic	2	NA	0.209	2045397962...
11	internet	2	NA	0.154	1169784624...
12	higher	2	NA	0.037	2964842468...
13	nursery	2	NA	0.149	1521828010...
14	activities	2	NA	0.236	2588388582...
15	paid	2	NA	0.232	8403595503...

Εικόνα 13: Αποτελέσματα του γραφικού στοιχείου "Rank"

4.3.4 Αποτελέσματα Γραμμικής Παλινδρόμησης

Στην παρακάτω φωτογραφία φαίνονται οι προβλεπόμενες τιμές του G3 μετά την εφαρμογή της Γραμμικής Παλινδρόμησης. Συγκεκριμένα στην πρώτη στήλη «G3» έχουμε τις πραγματικές τιμές της βαθμολογίας της 3ης περιόδου στο μάθημα των Μαθηματικών, ενώ στην 2η στήλη «Linear Regression» έχουμε τις προβλεπόμενες τιμές μετά την εφαρμογή της γραμμικής παλινδρόμησης.



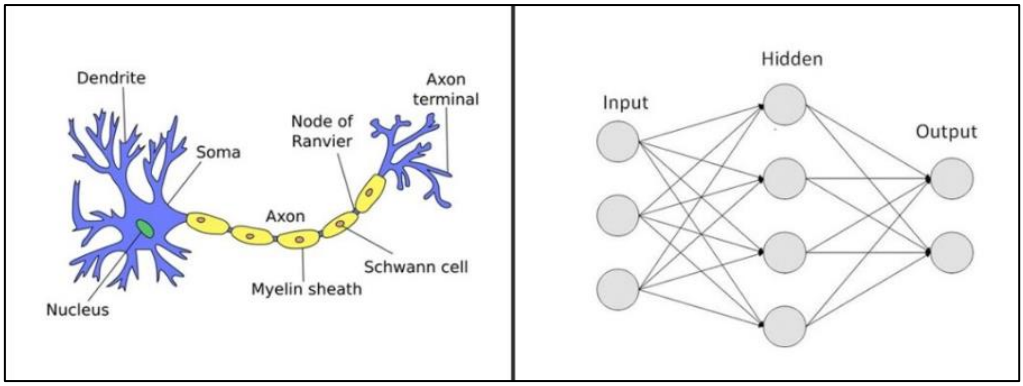
	G3	Linear Regression	Fold	school	sex
1	6	5.88531	1	GP	F
2	15	12.4091	1	GP	F
3	15	16.2384	1	GP	M
4	15	14.502	1	GP	M
5	14	13.7912	1	GP	F
6	5	4.33817	1	GP	M
7	16	15.5915	1	GP	M
8	8	6.57159	1	GP	F
9	12	9.72666	1	GP	M
10	17	16.0026	1	GP	M
11	12	9.27769	1	GP	M
12	13	14.005	1	GP	F
13	18	19.0967	1	GP	M
14	6	8.24082	1	GP	F
15	11	11.3471	1	GP	F
16	10	8.79663	1	GP	F
17	15	14.3355	1	GP	F
18	15	15.5337	1	GP	M
19	5	5.45294	1	GP	F
20	14	12.0967	1	GP	M

Εικόνα 14: Αποτελέσματα Γραμμικής Παλινδρόμησης

Παρατηρούμε πως το μέσο τετραγωνικό σφάλμα (MSE) της γραμμικής παλινδρόμησης είναι 4.261 (εικόνα 10). Το MSE αποτελεί έναν δείκτη, ο οποίος ελέγχει πόσο κοντά είναι οι εκτιμήσεις ή οι προβλέψεις με τις πραγματικές τιμές. Όσο χαμηλότερο είναι το μέσο τετραγωνικό σφάλμα, τόσο πιο κοντά είναι η πρόβλεψη στην πραγματική τιμή. Έτσι, το μέσο τετραγωνικό σφάλμα θεωρείται ως μέτρο αξιολόγησης μοντέλου για μοντέλα παλινδρόμησης και όσο χαμηλότερη είναι η τιμή του υποδηλώνει καλύτερη προσαρμογή. Κάτι αντίστοιχο συμβαίνει και με τον δείκτη RMSE, την ρίζα του μέσου τετραγωνικού σφάλματος. Όσο χαμηλότερο είναι το RMSE, τόσο καλύτερα το μοντέλο είναι σε θέση να «ταιριάξει» με το σύνολο δεδομένων. Επιπλέον, το μέσο απόλυτο σφάλμα, MAE, είναι ένα μέτρο της ακρίβειας του μοντέλου που δίνεται στην ίδια κλίμακα με τον στόχο πρόβλεψης. Με απλά λόγια, το MAE μπορεί να ερμηνευθεί ως το μέσο σφάλμα που έχουν οι προβλέψεις του μοντέλου σε σύγκριση με τους αντίστοιχους πραγματικούς στόχους τους. Στο συγκεκριμένο πείραμα, παρατηρούμε πως RMSE=2.064 και MAE=1.378. Τέλος, ο δείκτης r-squared (R^2) δείχνει πόσο καλά το μοντέλο παλινδρόμησης εξηγεί τα παρατηρούμενα δεδομένα. Στην συγκεκριμένη περίπτωση, έχουμε R^2=0.796, το οποίο μας αποκαλύπτει ότι σχεδόν το 80% της μεταβλητότητας που παρατηρείται στη μεταβλητή στόχο εξηγείται από το μοντέλο παλινδρόμησης.

4.3.5 Νευρωνικά Δίκτυα

Μια από τις πιο διαδεδομένες μεθόδους επιβλεπόμενης μηχανικής μάθησης, είναι αυτή των τεχνητών νευρωνικών δικτύων (Artificial Neural Networks). Ιστορικά, η έρευνα των βιολογικών νευρώνων του ανθρώπινου εγκεφάλου αποτέλεσε έμπνευση για την ανάπτυξη των τεχνητών νευρωνικών δικτύων. Απλουστεύοντας σε μεγάλο βαθμό, μπορούμε να πούμε ότι στην περίπτωση του ανθρώπινου εγκεφάλου, το μυαλό μας μαθαίνει με τα χρόνια ποιους νευρώνες να ενεργοποιεί αναλόγως του ερεθίσματος που λαμβάνει. Στη συνέχεια, αλλάζοντας τον βαθμό ενεργοποίησης των συνάψεων μπορεί και δημιουργεί νέες συνδέσεις και έτσι να μαθαίνει νέες πληροφορίες και διαδικασίες. Οι πρώτες αναφορές για το πως θα μπορούσε να λειτουργήσει ένας τεχνητός νευρώνας, έγιναν από τους Warren McCulloch και Walter Pitts το 1943, ενώ η πρώτη εμφάνισή τεχνητού νευρώνα έγινε το 1957 από τον ψυχολόγο Frank Rosenblatt.



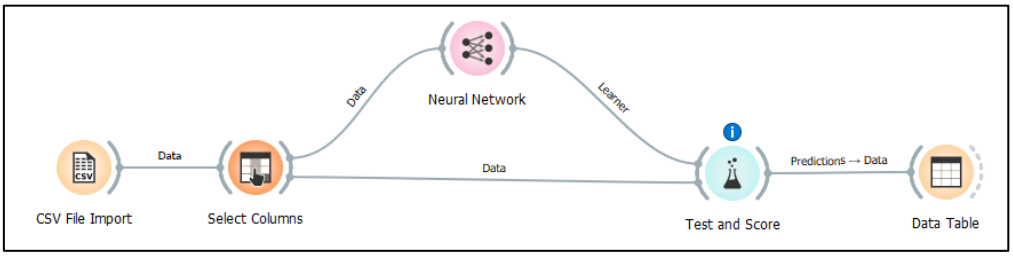
Εικόνα 15: [Αριστερά]: παράδειγμα ενός βιολογικού νευρώνα & [Δεξιά]: παράδειγμα ενός τεχνητού νευρωνικού δικτύου. Κάθε κύκλος αντιπροσωπεύει έναν τεχνητό νευρώνα.

Εν συντομία, ένα νευρωνικό δίκτυο ορίζεται ως ένα υπολογιστικό σύστημα που αποτελείται από έναν αριθμό απλών αλλά πολύ διασυνδεδεμένων στοιχείων ή κόμβων, που ονομάζονται «νευρώνες», οι οποίοι είναι οργανωμένοι σε επίπεδα που

επεξεργάζονται πληροφορίες χρησιμοποιώντας δυναμικές αποκρίσεις κατάστασης σε εξωτερικές εισόδους. Αυτός ο αλγόριθμος είναι εξαιρετικά χρήσιμος, για την εύρεση μοτίβων που είναι πολύ περίπλοκα για να εξαχθούν χειροκίνητα και να διδαχθούν για αναγνώριση στο μηχάνημα. Στο πλαίσιο αυτής της δομής, τα μοτίβα εισάγονται στο νευρωνικό δίκτυο από το επίπεδο εισόδου που έχει έναν νευρώνα για κάθε στοιχείο που υπάρχει στα δεδομένα εισόδου και κοινοποιείται σε ένα ή περισσότερα κρυφά επίπεδα που υπάρχουν στο δίκτυο. ονομάζεται «κρυμμένο» μόνο λόγω του γεγονότος ότι δεν αποτελούν το επίπεδο εισόδου ή εξόδου. Στα κρυφά στρώματα γίνεται στην πραγματικότητα όλη η επεξεργασία, μέσω ενός συστήματος συνδέσεων που χαρακτηρίζονται από βάρη και προκαταλήψεις: η είσοδος λαμβάνεται, ο νευρώνας υπολογίζει ένα σταθμισμένο άθροισμα προσθέτοντας επίσης την προκατάληψη και σύμφωνα με αποτέλεσμα και μια προκαθορισμένη συνάρτηση ενεργοποίησης (το πιο συνηθισμένο είναι το σιγμοειδές, σ, παρόλο που σχεδόν δεν χρησιμοποιείται πια και υπάρχουν καλύτερες όπως το ReLu), αποφασίζει αν θα πρέπει να «πυροδοτηθεί» ή να ενεργοποιηθεί. Στη συνέχεια, ο νευρώνας μεταδίδει τις πληροφορίες σε άλλους συνδεδεμένους νευρώνες σε μια διαδικασία που ονομάζεται «πέρασμα προς τα εμπρός». Στο τέλος αυτής της διαδικασίας, το τελευταίο κρυφό στρώμα συνδέεται με το στρώμα εξόδου που έχει έναν νευρώνα για κάθε πιθανή επιθυμητή έξοδο.

4.3.6 Υλοποίηση στο περιβάλλον Orange

Στην παρακάτω φωτογραφία φαίνεται η ροή των γραφικών στοιχείων που επιλέγουμε ώστε να προβλέψουμε τον βαθμό G3 με την χρήση νευρωνικού δικτύου. Στόχος μας είναι να προβλέψουμε τον βαθμό G3 ο οποίος είναι ο βαθμός του τελευταίου έτους (3η περίοδο) στα Μαθηματικά, έχοντας τους βαθμούς G1, G2 της 1ης και της 2ης περιόδου.



Εικόνα 16: Πρόβλεψη της τιμής G3 με χρήση νευρωνικού δικτύου στο περιβάλλον Orange

4.3.7 Διερεύνηση και Αξιολόγηση επίδοσης για διαφορετικές συναρτήσεις ενεργοποίησης και επιλυτές

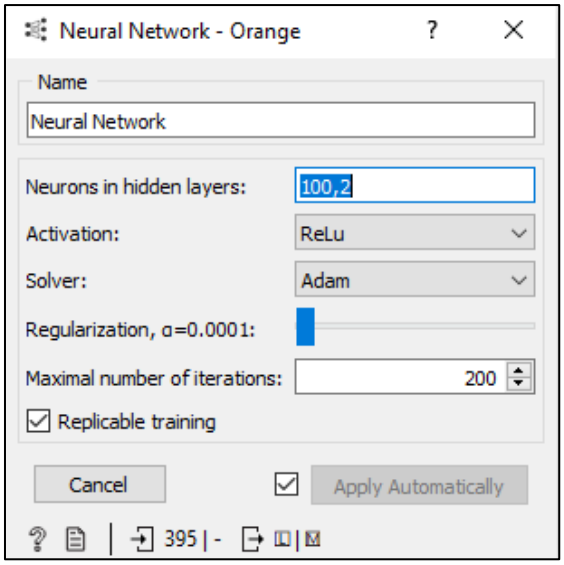
Για τα νευρωνικά δίκτυα είναι απαραίτητο να ορίσουμε την συνάρτηση ενεργοποίησης και τον επιλυτή. Η κλάση επίλυσης αντιπροσωπεύει έναν βελτιστοποιητή που βασίζεται σε στοχαστική κλίση για τη βελτιστοποίηση των παραμέτρων στο γράφημα υπολογισμού. Παρακάτω ακολουθούν κάποιοι από τους σημαντικότερους επιλυτές:

SGD - Η στοχαστική κλίση ενημερώνει τα βάρη με έναν γραμμικό συνδυασμό της αρνητικής κλίσης και της προηγούμενης ενημέρωσης βάρους. Ο ρυθμός μάθησης είναι το βάρος της αρνητικής διαβάθμισης. Η ορμή είναι το βάρος της προηγούμενης ενημέρωσης.
Adam - Ο Adam είναι ένας αλγόριθμος βελτιστοποίησης αντικατάστασης για στοχαστική κλίση κατάβασης για εκπαίδευση μοντέλων βαθιάς μάθησης. Ο Adam

συνδυάζει τις καλύτερες ιδιότητες των αλγορίθμων AdaGrad και RMSProp για να παρέχει έναν αλγόριθμο βελτιστοποίησης που μπορεί να χειριστεί αραιές κλίσεις σε θορυβώδη προβλήματα.
Μια συνάρτηση ενεργοποίησης αποφασίζει εάν ένας νευρώνας πρέπει να ενεργοποιηθεί ή όχι. Αυτό σημαίνει ότι θα αποφασίσει εάν η είσοδος του νευρώνα στο δίκτυο είναι σημαντική ή όχι στη διαδικασία πρόβλεψης χρησιμοποιώντας απλούστερες μαθηματικές πράξεις. Από τις γνωστότερες συναρτήσεις ενεργοποίησης, οι οποίες και θα χρησιμοποιηθούν στην συνέχεια, αναλύονται παρακάτω:

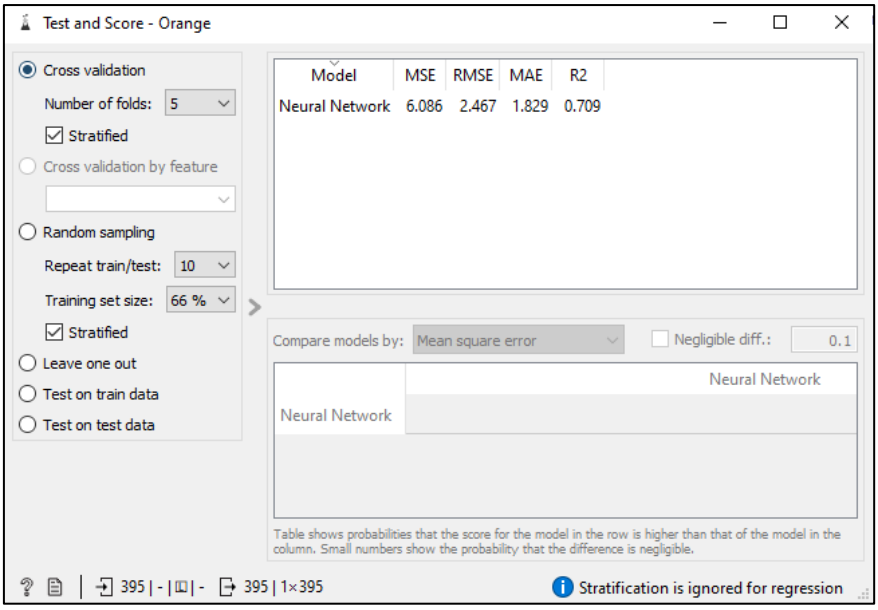
ReLu - Η διορθωμένη συνάρτηση γραμμικής ενεργοποίησης ή ReLU για συντομία είναι μια τμηματικά γραμμική συνάρτηση που θα εξάγει απευθείας την είσοδο εάν είναι θετική, διαφορετικά, θα βγάζει μηδέν. Έχει γίνει η προεπιλεγμένη λειτουργία ενεργοποίησης για πολλούς τύπους νευρωνικών δικτύων, επειδή ένα μοντέλο που το χρησιμοποιεί είναι πιο εύκολο να εκπαιδευτεί και συχνά επιτυγχάνει καλύτερη απόδοση.
Identity - Συνάρτηση Ταυτότητας ή Γραμμικής ενεργοποίησης — Η συνάρτηση ταυτότητας ή γραμμικής ενεργοποίησης είναι η απλούστερη συνάρτηση ενεργοποίησης από όλες. Εφαρμόζει λειτουργία ταυτότητας στα δεδομένα και τα δεδομένα εξόδου είναι ανάλογα με τα δεδομένα εισόδου.
Στην συνέχεια παρουσιάζεται η διερεύνηση και αξιολόγηση επίδοσης για διάφορα ζεύγη συναρτήσεων ενεργοποίησης και επιλυτές.

i. *Συνάρτηση ενεργοποίησης: ReLu, Επιλυτής: Adam*



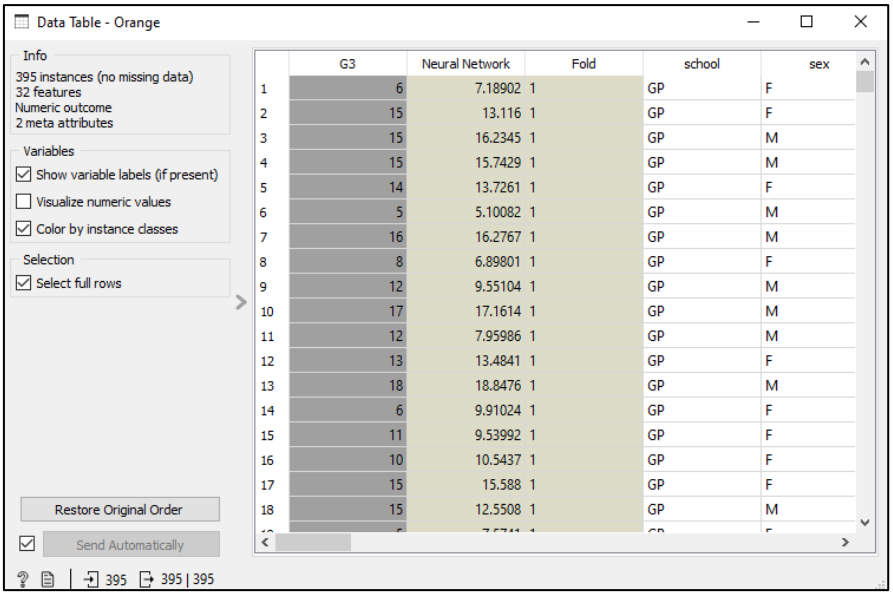
Εικόνα 17: Νευρωνικό δίκτυο με συνάρτηση ενεργοποίησης ReLu και επιλυτή Adam

Με τον συνδυασμό συνάρτησης ενεργοποίησης ReLu και επιλυτή Adam παρατηρούμε πως MSE=6.086.



Εικόνα 18: Αξιολόγηση επίδοσης νευρωνικού δικτύου με συνάρτηση ενεργοποίησης ReLu και επιλυτή Adam

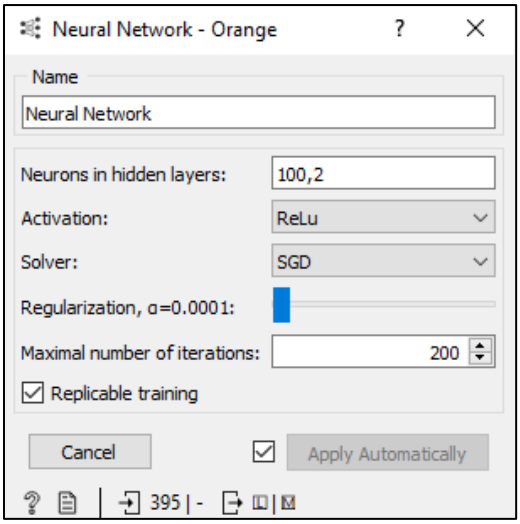
Στην πρώτη στήλη φαίνονται οι πραγματικές τιμές του G3, ενώ στην 2η στήλη φαίνονται οι προβλεπόμενες τιμές του G3 μετά από την εφαρμογή του νευρωνικού δικτύου.



Εικόνα 19: Οι προβλέψεις του νευρωνικού δικτύου για το χαρακτηριστικό G3

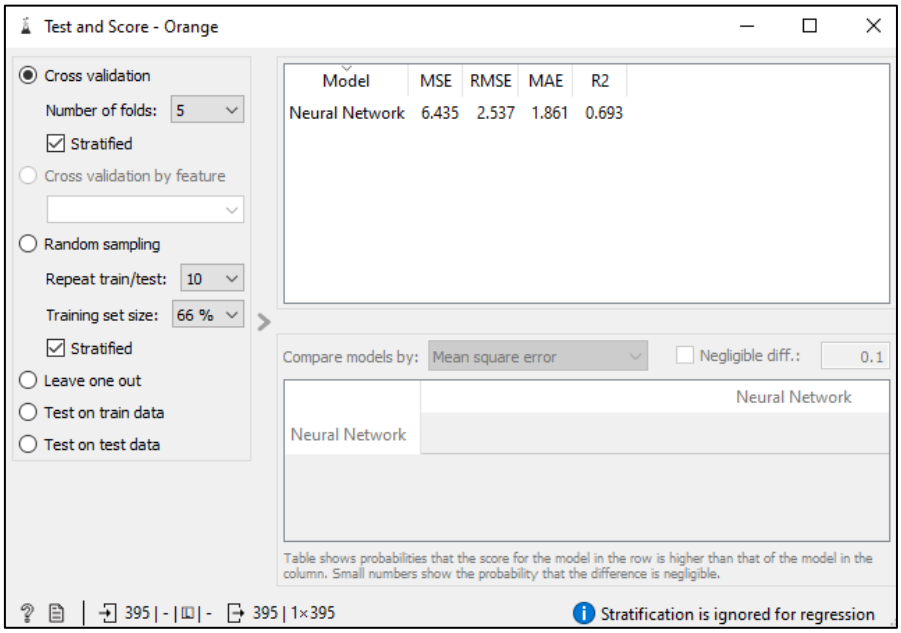
ii. *Συνάρτηση ενεργοποίησης: ReLu, Επιλυτής: SGD*

Με τον συνδυασμό συνάρτησης ενεργοποίησης Relu και επιλυτή SGD παρατηρούμε πως MSE=6.435.

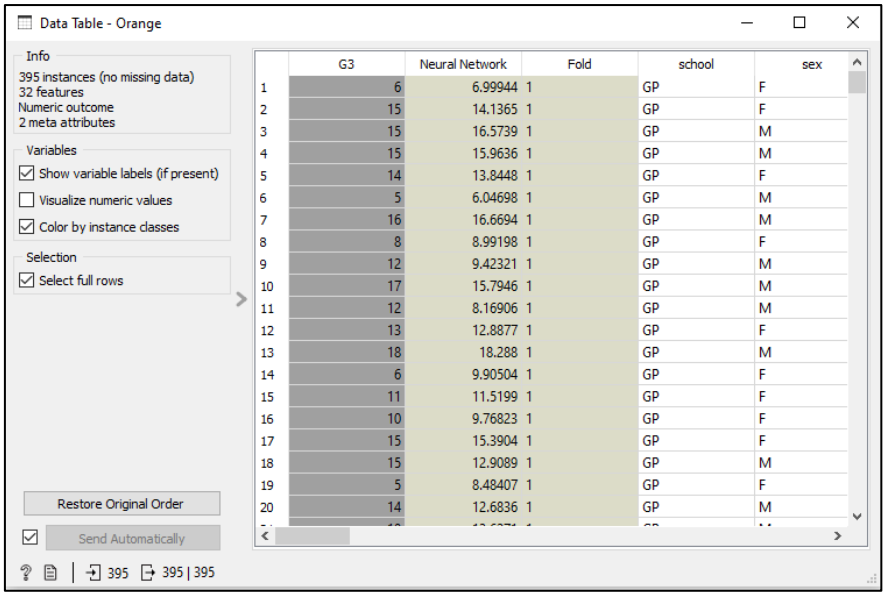


Εικόνα 20: Νευρωνικό δίκτυο με συνάρτηση ενεργοποίησης ReLu και επιλυτή SGD

Στην πρώτη στήλη φαίνονται οι πραγματικές τιμές του G3, ενώ στην 2η στήλη φαίνονται οι προβλεπόμενες τιμές του G3 μετά από την εφαρμογή του νευρωνικού δικτύου.



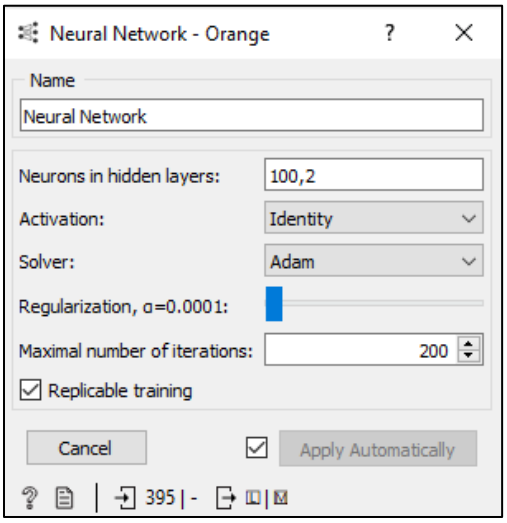
Εικόνα 21: Αξιολόγηση επίδοσης νευρωνικού δικτύου με συνάρτηση ενεργοποίησης ReLu και επιλυτή SGD



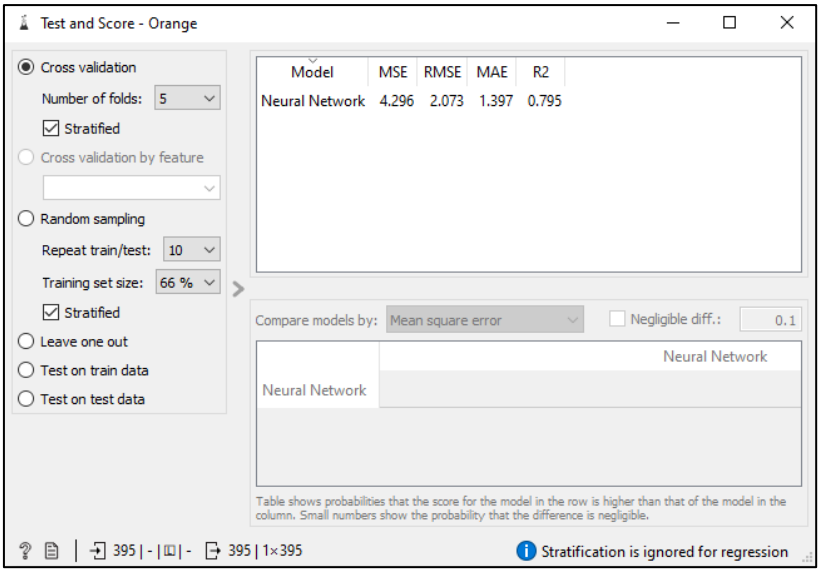
Εικόνα 22: Οι προβλέψεις του νευρωνικού δικτύου για το χαρακτηριστικό G3

iii. Συνάρτηση ενεργοποίησης: Identity, Επιλυτής: Adam

Με τον συνδυασμό συνάρτησης ενεργοποίησης Identity και επιλυτή Adam παρατηρούμε πως MSE=4.296.

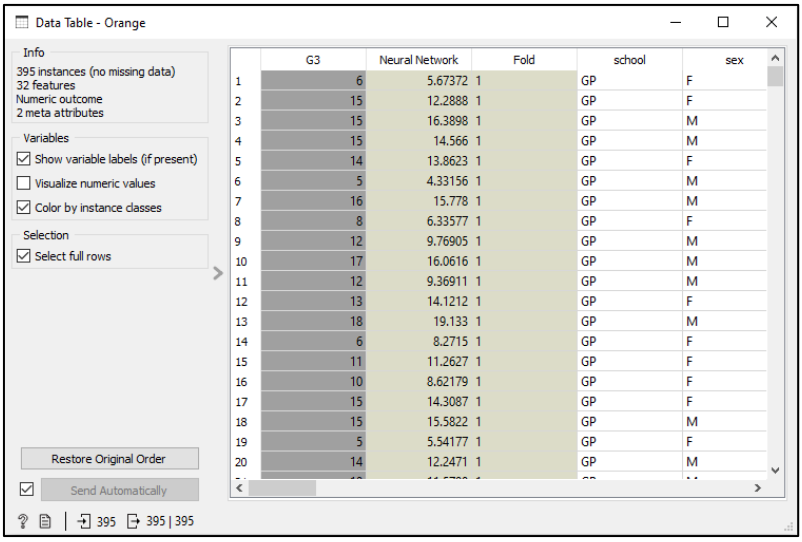


Εικόνα 23: Νευρωνικό δίκτυο με συνάρτηση ενεργοποίησης Identity και επιλυτή Adam



Εικόνα 24: Αξιολόγηση επίδοσης νευρωνικού δικτύου με συνάρτηση ενεργοποίησης Identity και επιλυτή Adam

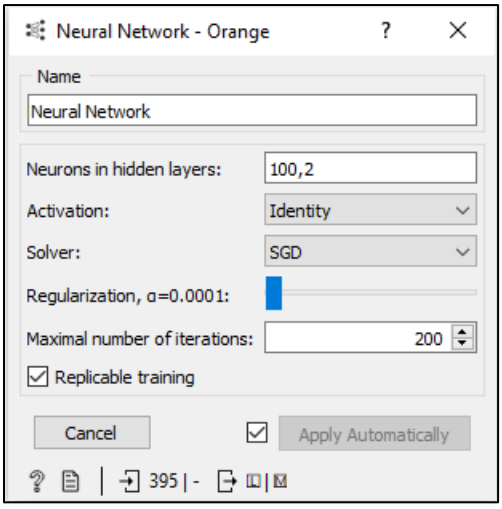
Στην πρώτη στήλη φαίνονται οι πραγματικές τιμές του G3, ενώ στην 2η στήλη φαίνονται οι προβλεπόμενες τιμές του G3 μετά από την εφαρμογή του νευρωνικού δικτύου.



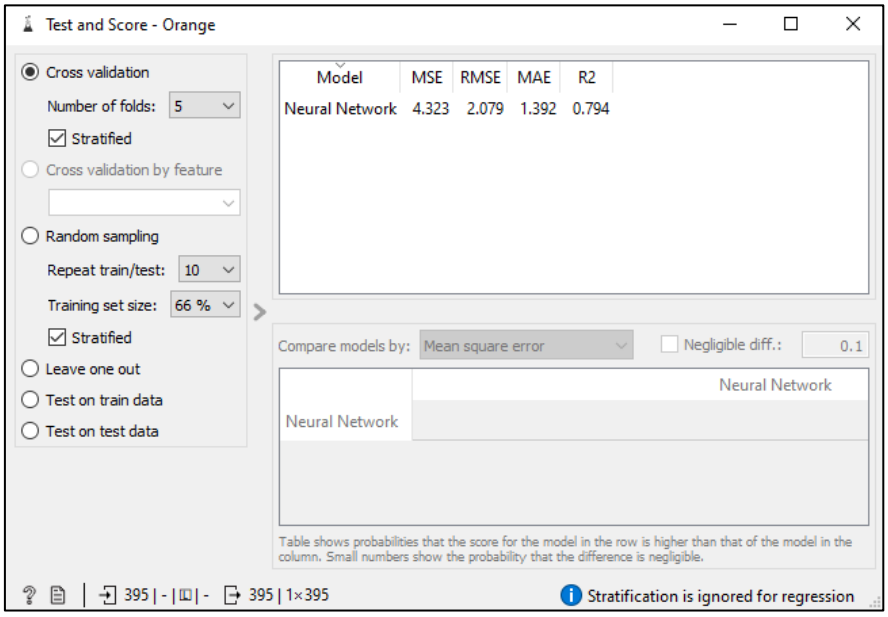
Εικόνα 25: Οι προβλέψεις του νευρωνικού δικτύου για το χαρακτηριστικό G3

iv. Συνάρτηση ενεργοποίησης: Identity, Επιλυτής: SGD

Με τον συνδυασμό συνάρτησης ενεργοποίησης Identity και επιλυτή SGD παρατηρούμε πως MSE=4.323.

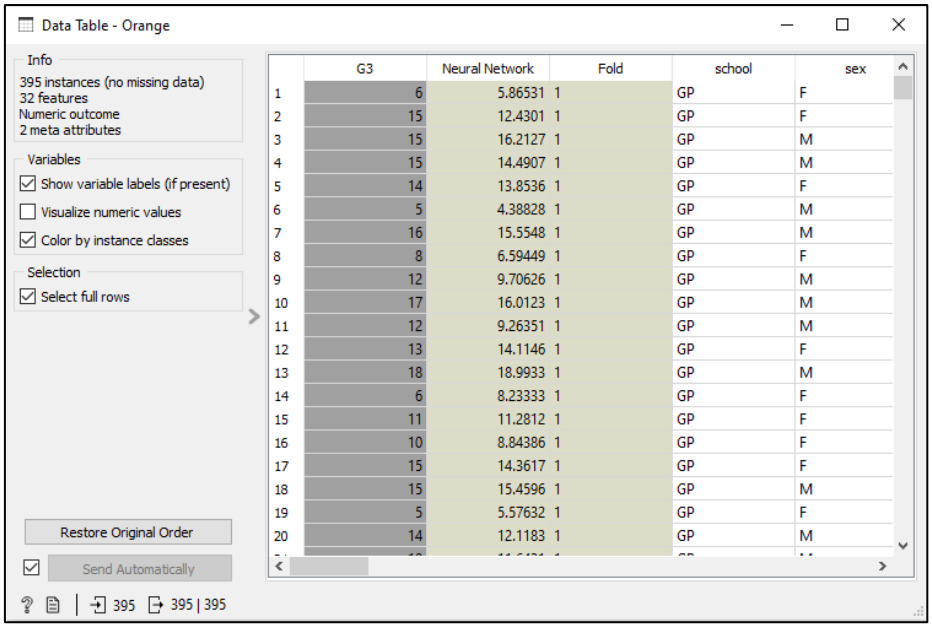


Εικόνα 26: Νευρωνικό δίκτυο με συνάρτηση ενεργοποίησης Identity και επιλυτή SGD



Εικόνα 27: Αξιολόγηση επίδοσης νευρωνικού δικτύου με συνάρτηση ενεργοποίησης Identity και επιλυτή SGD

Στην πρώτη στήλη φαίνονται οι πραγματικές τιμές του G3, ενώ στην 2η στήλη φαίνονται οι προβλεπόμενες τιμές του G3 μετά από την εφαρμογή του νευρωνικού δικτύου.

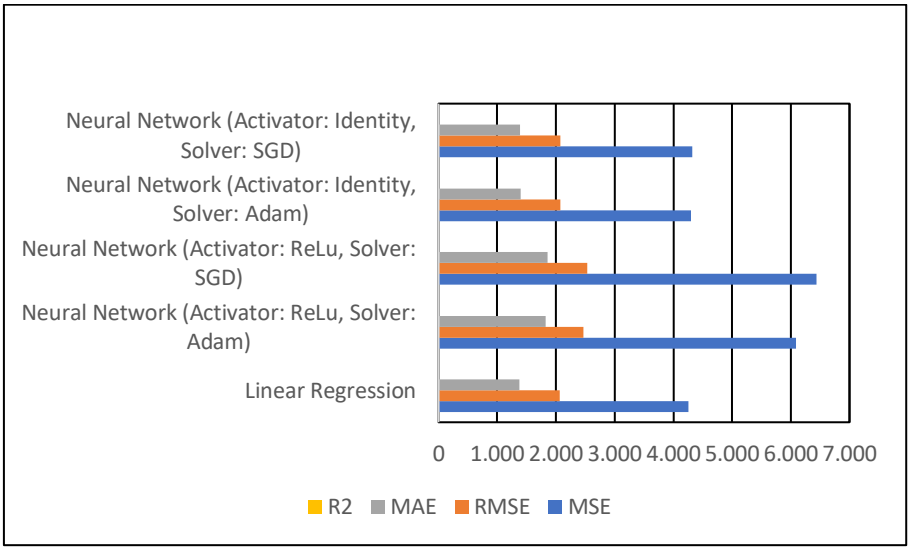


	G3	Neural Network	Fold	school	sex
1	6	5.86531	1	GP	F
2	15	12.4301	1	GP	F
3	15	16.2127	1	GP	M
4	15	14.4907	1	GP	M
5	14	13.8536	1	GP	F
6	5	4.38828	1	GP	M
7	16	15.5548	1	GP	M
8	8	6.59449	1	GP	F
9	12	9.70626	1	GP	M
10	17	16.0123	1	GP	M
11	12	9.26351	1	GP	M
12	13	14.1146	1	GP	F
13	18	18.9933	1	GP	M
14	6	8.23333	1	GP	F
15	11	11.2812	1	GP	F
16	10	8.84386	1	GP	F
17	15	14.3617	1	GP	F
18	15	15.4596	1	GP	M
19	5	5.57632	1	GP	F
20	14	12.1183	1	GP	M

Εικόνα 28: Οι προβλέψεις του νευρωνικού δικτύου για το χαρακτηριστικό G3

4.4 Αποτελέσματα

Έπειτα από τα διαφορετικά πειράματα που παρουσιάστηκαν παραπάνω, διαφοροποιώντας τους επιλυτές και τις συναρτήσεις ενεργοποίησης παρατηρούμε πως το μικρότερο τετραγωνικό σφάλμα προέκυψε μέσω του συνδυασμού συνάρτησης ενεργοποίησης Identity και επιλυτή Adam. Συγκεκριμένα, προκύπτει MSE=4.296, το οποίο είναι λίγο μεγαλύτερο από το τετραγωνικό σφάλμα της γραμμικής παλινδρόμησης (MSE=4.261), που ήταν το πρώτο πείραμα που εκτελέστηκε. Ως επιστέγασμα των όσων αναφέρθηκαν, συμπεραίνουμε πως για το συγκεκριμένο σύνολο δεδομένων, η καταλληλότερη τεχνική να προβλέψουμε τις ακαδημαϊκές επιδόσεις των μαθητών είναι να χρησιμοποιήσουμε το μοντέλο γραμμικής παλινδρόμησης. Παρακάτω παρουσιάζονται συνοπτικά τα αποτελέσματα των επιδόσεων για κάθε μοντέλο:



Εικόνα 29: Γράφημα επιδόσεων ανά μοντέλο

5. Εφαρμογή τεχνικών μάθησης και νευρωνικών δικτύων για την πρόβλεψη ακαδημαϊκών επιδόσεων μαθητών κατά την διάρκεια του Covid-19

5.1 Πληροφορίες συνόλου δεδομένων

Το δεύτερο σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα έρευνα αφορά φοιτητές πανεπιστημίου στην Ιορδανία. Οι φοιτητές και οι φοιτήτριες του πανεπιστημίου απάντησαν σε μια φόρμα ερωτηματολογίου τύπου Likert και τα δεδομένα που παράχθηκαν αποτέλεσαν το κύριο σύνολο των δεδομένων. Αυτό το σύνολο δεδομένων είναι πολύ χρήσιμο καθώς όχι μόνο προκαλεί απαντήσεις από τους μαθητές σχετικά με τη χρήση ψηφιακών εργαλείων για μελέτη, αλλά λαμβάνει επίσης υπόψη τον ψυχολογικό αντίκτυπο που προκαλείται από την υπερβολική χρήση τους, η οποία με τη σειρά της γίνεται κρίσιμος παράγοντας στην ακαδημαϊκή επίδοση ενός μαθητή. Επιπλέον, ως συμπλήρωμα του παραπάνω συνόλου δεδομένων, δημιουργήθηκε ένας σύνδεσμος φόρμας Google και μοιράστηκε σε προπτυχιακούς φοιτητές και φοιτήτριες που φοιτούν σε διάφορα ινδικά κολέγια. Η φόρμα έκανε επίσης παρόμοιες ερωτήσεις στους μαθητές μαζί με τις δημογραφικές τους πληροφορίες όπως ηλικία, επίπεδο/έτος, φύλο και GPA (Grade Point Average: μέσος όρος βαθμολογίας) όπως ακριβώς στο σύνολο δεδομένων της Ιορδανίας. Στη συνέχεια, τα δεδομένα αυτά προστέθηκαν στο κύριο σύνολο και δημιουργήθηκε ένα τελικό ολοκληρωμένο σύνολο δεδομένων. Αξίζει να αναφερθεί πως και τα δύο σύνολα δεδομένων συγκεντρώθηκαν μετά την ανακοίνωση του lockdown, επομένως οι μαθητές είχαν ήδη τη συνήθεια να χρησιμοποιούν ψηφιακά εργαλεία για τη μελέτη. Παρακάτω δίνεται η πλήρης περιγραφή των χαρακτηριστικών του συνόλου δεδομένων:

1) **Gender** – το φύλο των φοιτητών/τριών (Female - θηλυκό, male - αρσενικό)
2) **Level/Year** – το έτος φοίτησης των φοιτητών/τριών (First, First/Freshman, Second, Second/Sophomore, Third, Third/Junior, Fourth, Fourth/Senior, Other)
3) **Age** – η ηλικία των φοιτητών/τριών (18-24, 25-30, 30+)
   Your cumulative average (GPA) - Grade Point Average: μέσος όρος βαθμολογίας (κάτω από 60, κάτω από 60 / Κάτω από 2.0, 60-69, 60-69 / 2-2.9, 60–69 / 2–2.49, 70-79 και 70–79 / 2,5–2,99, 80-89, 80–89 / 3 –3,49, +90, +90 / +3,5)
4) **Before COVID-19: Which of the following digital tools do you usually use?** - Πριν από τον COVID-19: Ποιό από τα παρακάτω ψηφιακά εργαλεία χρησιμοποιείτε συνήθως; (I pad/ Tablet, Laptop, Mobile Phone, Personal Computer, Other)
5) **After COVID-19: Which of the following digital tools do you usually use?** - Μετά τον COVID-19: Ποιό από τα παρακάτω ψηφιακά εργαλεία χρησιμοποιείτε συνήθως; (I pad/ Tablet, Laptop, Mobile Phone, Personal Computer, Other)
6) **Before COVID-19: How much time do you spend using the digital tools in learning?** - Πριν από τον COVID-19: Πόσο χρόνο αφιερώνετε χρησιμοποιώντας τα ψηφιακά εργαλεία στη μάθηση; (1-3, 3-6, 6-9, 9-12, +12)
7) **After COVID-19: How much time do you spend using the digital tools in learning?** - Μετά τον COVID-19: Πόσο χρόνο αφιερώνετε χρησιμοποιώντας τα ψηφιακά εργαλεία στη μάθηση; (1-3, 3-6, 6-9, 9-12, +12)

8) **Before COVID-19: I always use digital tools (mobile, laptop, i-pad) in studying.** - Πριν από τον COVID-19: Χρησιμοποιώ πάντα ψηφιακά εργαλεία (κινητό, φορητός υπολογιστής, i-pad) στη μελέτη. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
9) **After COVID-19: I always use digital tools (mobile, laptop, i-pad) in studying.** - Μετά τον COVID-19: Χρησιμοποιώ πάντα ψηφιακά εργαλεία (κινητό, φορητός υπολογιστής, i-pad) στη μελέτη. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
10) **Before COVID-19: When I use the mobile phone, tablet or laptop in e-learning, I cannot concentrate and I am distracted.** - Πριν από τον COVID-19: Όταν χρησιμοποιώ το κινητό τηλέφωνο, το tablet ή το φορητό υπολογιστή στην ηλεκτρονική μάθηση, δεν μπορώ να συγκεντρωθώ και αποσπώ την προσοχή μου. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
11) **After COVID-19: When I use the mobile phone, tablet or laptop in e-learning, I cannot concentrate and I am distracted.** - Μετά τον COVID-19: Όταν χρησιμοποιώ κινητό τηλέφωνο, tablet ή φορητό υπολογιστή στην ηλεκτρονική μάθηση, δεν μπορώ να συγκεντρωθώ και αποσπώ την προσοχή μου. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
12) **Before COVID-19: I have fixed hours for bedtime and wake-up.** - Πριν από τον COVID-19: Έχω καθορίσει συγκεκριμένες ώρες για την ώρα του ύπνου και του ξυπνήματος. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
13) **After COVID-19: I have fixed hours for bedtime and wake-up.** - Μετά τον COVID-19: Έχω καθορίσει συγκεκριμένες ώρες για ύπνο και ξύπνημα. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
14) **Before COVID-19: Prolonged use of digital tools for learning (mobile, laptop, i-pad) affected my sleeping habits.** - Πριν από τον COVID-19: Η παρατεταμένη χρήση ψηφιακών εργαλείων για μάθηση (κινητό, φορητός υπολογιστής, i-pad) επηρέασε τις συνήθειες ύπνου μου. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
15) **After COVID-19: Prolonged use of digital tools for learning (mobile, laptop, i-pad) affected my sleeping habits.** - Μετά τον COVID-19: Η παρατεταμένη χρήση ψηφιακών εργαλείων για μάθηση (κινητό, φορητός υπολογιστής, i-pad) επηρέασε τις συνήθειες ύπνου μου. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)

16) **Before COVID-19: Continuous exposure to electronic screens in online learning is tiring and exhausting.** - Πριν από τον COVID-19: Η συνεχής έκθεση σε ηλεκτρονικές οθόνες στη διαδικτυακή μάθηση είναι κουραστική και εξαντλητική. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
17) **After COVID-19: Continuous exposure to electronic screens in online learning is tiring and exhausting.** - Μετά τον COVID-19: Η συνεχής έκθεση σε ηλεκτρονικές οθόνες στην ηλεκτρονική μάθηση είναι κουραστική και εξαντλητική. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
18) **The distance learning system, caused by the COVID-19 epidemic, resulted in social distancing.** - Το σύστημα εξ αποστάσεως εκπαίδευσης, που προκλήθηκε από την επιδημία COVID-19, είχε ως αποτέλεσμα την κοινωνική αποστασιοποίηση. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
19) **Prolonged use of digital tools (mobile, laptop, i-pad) causes students' isolation University learning contributes to strengthening the social personality of students.** - Η παρατεταμένη χρήση ψηφιακών εργαλείων (κινητό, φορητός υπολογιστής, i-pad) προκαλεί απομόνωση των μαθητών Η πανεπιστημιακή μάθηση συμβάλλει στην ενίσχυση της κοινωνικής προσωπικότητας των μαθητών. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
20) **Staying home for long periods of time leads to lethargy and laziness.** - Η παραμονή στο σπίτι για μεγάλα χρονικά διαστήματα οδηγεί σε λήθαργο και τεμπελιά. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
21) **Prolonged use of e-learning tools often leads to boredom, nervousness, and tension.** - Η παρατεταμένη χρήση εργαλείων ηλεκτρονικής μάθησης συχνά οδηγεί σε πλήξη, νευρικότητα και ένταση. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
22) **The psychological element is a key factor in the success of the educational process.** - Το ψυχολογικό στοιχείο είναι βασικός παράγοντας για την επιτυχία της εκπαιδευτικής διαδικασίας. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
23) **Some students cannot afford buying all necessary digital tools, which is embarrassing and frustrating.** - Μερικοί μαθητές δεν έχουν την οικονομική δυνατότητα να αγοράσουν όλα τα απαραίτητα ψηφιακά εργαλεία, κάτι που είναι ενοχλητικό και απογοητευτικό. (Strongly disagree – διαφωνώ απόλυτα,

disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)

24) **I don't recommend continuing with the online learning model because it is socially and psychologically unhealthy.** - Δεν συνιστώ να συνεχίσετε με το διαδικτυακό μοντέλο μάθησης επειδή είναι κοινωνικά και ψυχολογικά ανθυγιεινό. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
25) **Measures of lockdown, closures, and quarantine, brought by COVID-19 caused stress, frustration, and depression.** - Τα μέτρα καραντίνας, κλεισίματος και καραντίνας, που ελήφθησαν από τον COVID-19 προκάλεσαν άγχος, απογοήτευση και κατάθλιψη. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
26) **The volume of assignments via e-learning led to confusion, frustration and poor performance.** - Ο όγκος των εργασιών μέσω e-learning οδήγησε σε σύγχυση, απογοήτευση και κακή απόδοση. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
27) **Face-to-face interaction contributes significantly to boosting students' academic achievement.** - Η πρόσωπο με πρόσωπο αλληλεπίδραση συμβάλλει σημαντικά στην ενίσχυση των ακαδημαϊκών επιδόσεων των μαθητών. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)
28) **Taking quizzes and exams online from home was not comfortable and made me nervous.** - Το να κάνω κουίζ και εξετάσεις στο διαδίκτυο από το σπίτι δεν ήταν άνετο και με έκανε νευρικό. (Strongly disagree – διαφωνώ απόλυτα, disagree – διαφωνώ, uncertain – δεν γνωρίζω/δεν απαντώ, agree - συμφωνώ, strongly agree – συμφωνώ απόλυτα)

Παρακάτω δίνεται ο υπερσύνδεσμος που οδηγεί στο σύνολο δεδομένων:
https://1drv.ms/x/s!Ao5Ej5_GIq3Cij_FNMNGYg-sjLmm?e=8nhfDS

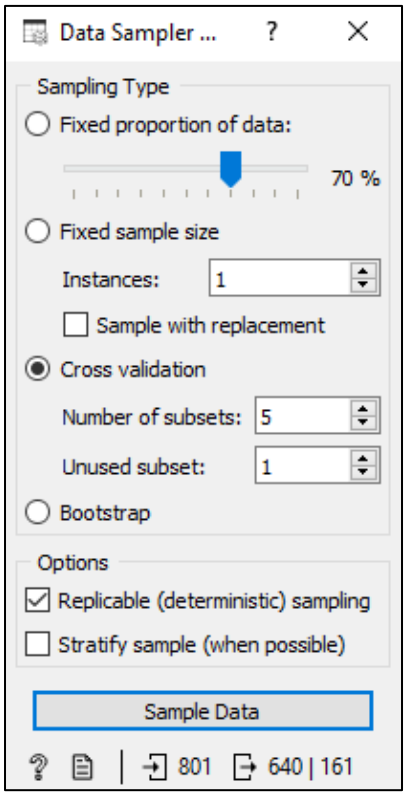
5.3 Σκοπός

Δευτερεύων σκοπός της παρούσας έρευνας είναι η πρόβλεψη του μέσου όρου βαθμολογίας (GPA) των φοιτητών, ο οποίος αποτελεί και τον χαρακτηριστικό στόχο. Η πρόβλεψη αυτή θα υλοποιηθεί με τεχνικές μηχανικής μάθησης με χρήση του περιβάλλοντος Orange. Αναλυτικότερα, θα διερευνηθούν διαφορετικοί αλγόριθμοι και στην συνέχεια θα αξιολογηθούν ώστε να προβλέψουμε τις επιδόσεις των φοιτητών και φοιτητριών με την μεγαλύτερη δυνατή ακρίβεια.

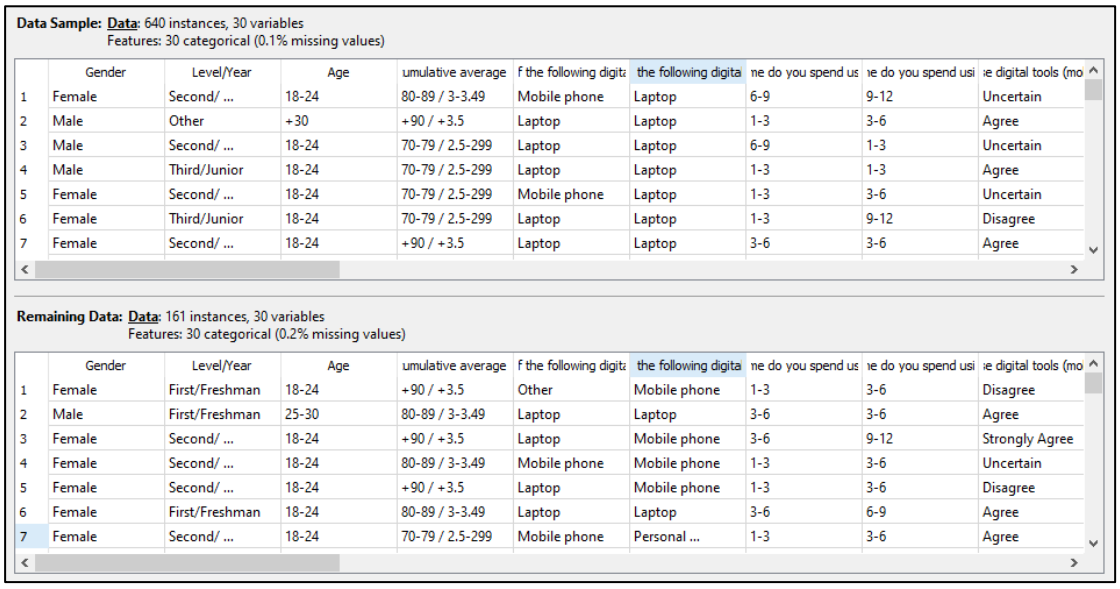
5.3 Διερεύνηση τεχνικών μηχανικής μάθησης για την πρόβλεψη των ακαδημαϊκών επιδόσεων των μαθητών

5.3.1 Διαχωρισμός του συνόλου δεδομένων σε σετ εκπαίδευσης και σετ δοκιμής

Το σύνολο των αλγόριθμων που θα χρησιμοποιηθούν θα αξιολογηθούν με διασταυρωμένη επικύρωση 5 ίσων πτυχών (5-Fold Cross Validation), όπως και στο πρώτο πείραμα. Στις παρακάτω εικόνες φαίνεται ο διαχωρισμός του συνόλου δεδομένων στο περιβάλλον Orange:



Εικόνα 30: Διαχωρισμός του συνόλου δεδομένων με την μέθοδο Cross Validation



Data Sample: Data: 640 instances, 30 variables
Features: 30 categorical (0.1% missing values)

	Gender	Level/Year	Age	umulative average	f the following digita	the following digita	ne do you spend us	ne do you spend usi	e digital tools (mo
1	Female	Second/ ...	18-24	80-89 / 3-3.49	Mobile phone	Laptop	6-9	9-12	Uncertain
2	Male	Other	+30	+90 / +3.5	Laptop	Laptop	1-3	3-6	Agree
3	Male	Second/ ...	18-24	70-79 / 2.5-299	Laptop	Laptop	6-9	1-3	Uncertain
4	Male	Third/Junior	18-24	70-79 / 2.5-299	Laptop	Laptop	1-3	1-3	Agree
5	Female	Second/ ...	18-24	70-79 / 2.5-299	Mobile phone	Laptop	1-3	3-6	Uncertain
6	Female	Third/Junior	18-24	70-79 / 2.5-299	Laptop	Laptop	1-3	9-12	Disagree
7	Female	Second/ ...	18-24	+90 / +3.5	Laptop	Laptop	3-6	3-6	Agree

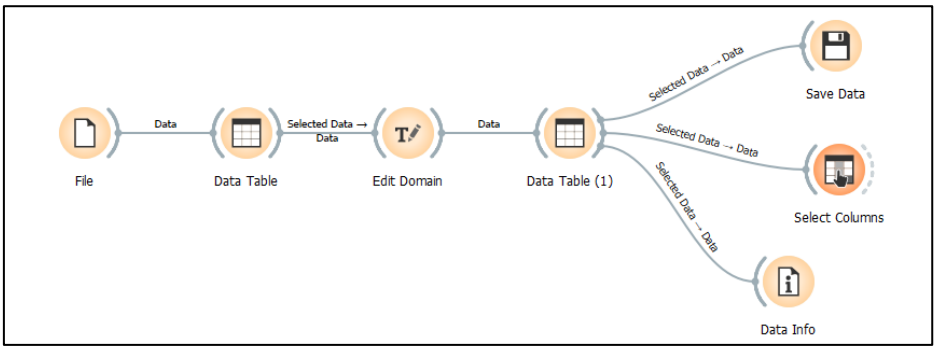
Remaining Data: Data: 161 instances, 30 variables
Features: 30 categorical (0.2% missing values)

	Gender	Level/Year	Age	umulative average	f the following digita	the following digita	ne do you spend us	ne do you spend usi	e digital tools (mo
1	Female	First/Freshman	18-24	+90 / +3.5	Other	Mobile phone	1-3	3-6	Disagree
2	Male	First/Freshman	25-30	80-89 / 3-3.49	Laptop	Laptop	3-6	3-6	Agree
3	Female	Second/ ...	18-24	+90 / +3.5	Laptop	Mobile phone	3-6	9-12	Strongly Agree
4	Female	Second/ ...	18-24	80-89 / 3-3.49	Mobile phone	Mobile phone	1-3	3-6	Uncertain
5	Female	Second/ ...	18-24	+90 / +3.5	Laptop	Mobile phone	1-3	3-6	Disagree
6	Female	First/Freshman	18-24	80-89 / 3-3.49	Laptop	Laptop	3-6	6-9	Agree
7	Female	Second/ ...	18-24	70-79 / 2.5-299	Mobile phone	Personal ...	1-3	3-6	Agree

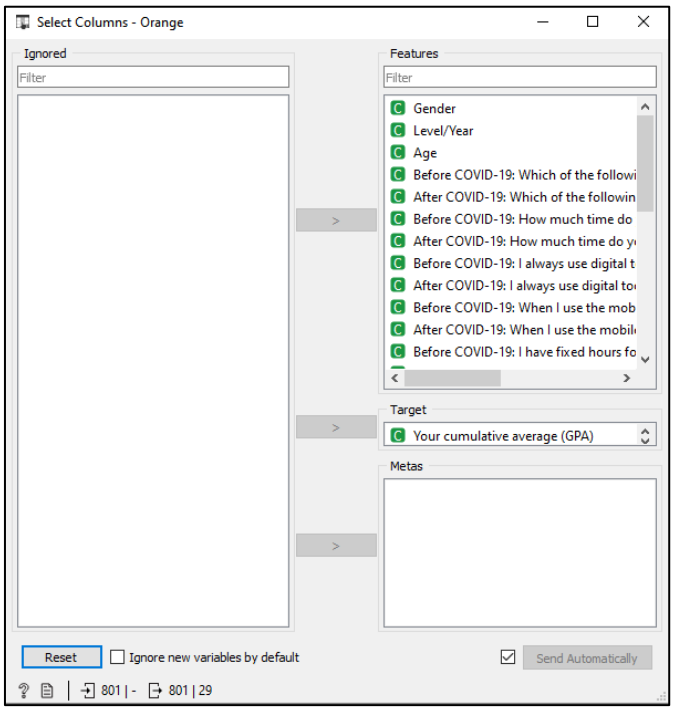
Εικόνα 31: Σετ εκπαίδευσης και σετ δοκιμής

5.3.2 Υλοποίηση στο περιβάλλον Orange

Αρχικά, φορτώνουμε τα δεδομένα στο περιβάλλον Orange και ορίζουμε ως τιμή- στόχο την στήλη GPA. Επιπλέον, θα γίνει επεξεργασία των στηλών των δεδομένων, ώστε κάθε τιμή να είναι αριθμητική ή δυαδική (data preparation).



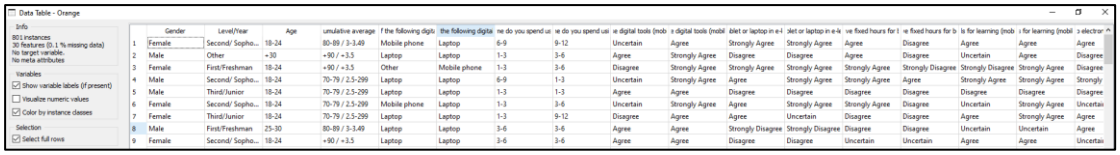
Εικόνα 32: Επεξεργασία των δεδομένων στο περιβάλλον Orange



Εικόνα 33: Καθορισμός του χαρακτηριστικού στόχου GPA

5.3.3 Βήματα επεξεργασίας των δεδομένων

Από τις 30 στήλες χαρακτηριστικών, οι 22 είναι εγγραφές έρευνας με βάση την κλίμακα Likert, η 1 είναι μια δυαδική στήλη (φύλο) και οι υπόλοιπες είναι εγγραφές που βασίζονται σε διαφορετικές κατηγορίες κατηγοριών, συμπεριλαμβανομένης της τιμής στόχου (GPA). Για να κάνουμε τη μηχανή δεδομένων ερμηνεύσιμη, καλύπτουμε τις ακόλουθες τιμές σε αριθμητικές ή δυαδικές τιμές. Στην παρακάτω φωτογραφία αποτυπώνεται όλες οι στήλες του συνόλου δεδομένων χωρίς να γίνει κάποια επεξεργασία:

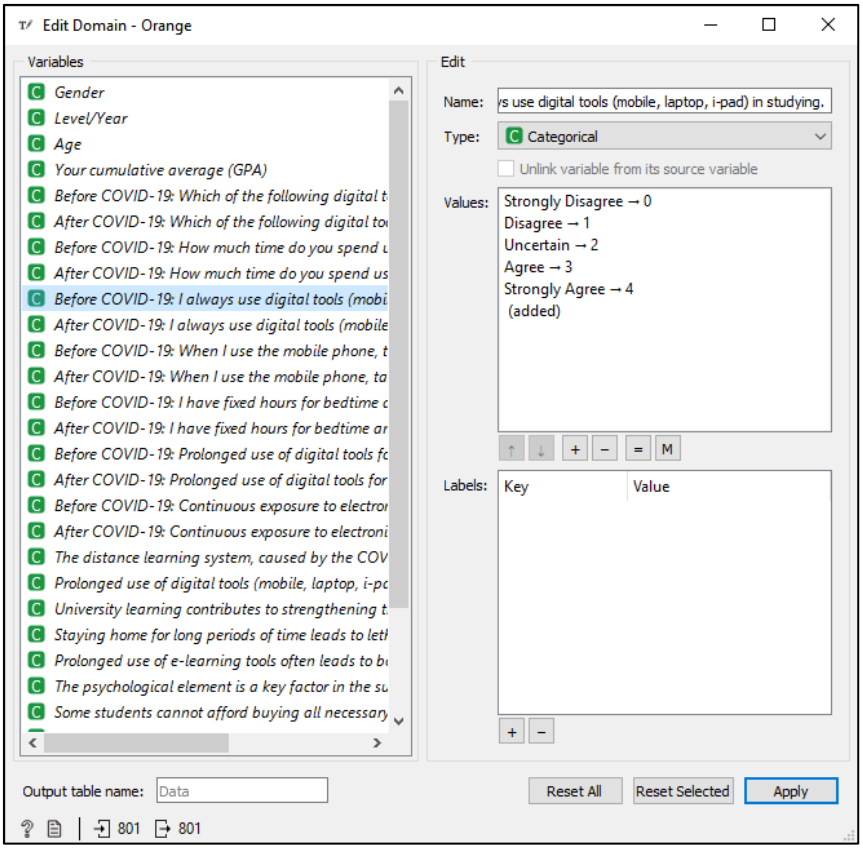


Εικόνα 34: Τα χαρακτηριστικά του συνόλου δεδομένων πριν την επεξεργασία

Τα βήματα που θα ακολουθήσουμε ώστε να μετατρέψουμε όλες τις κατηγορικές τιμές σε αριθμητικές ή δυαδικές αναλύονται παρακάτω:

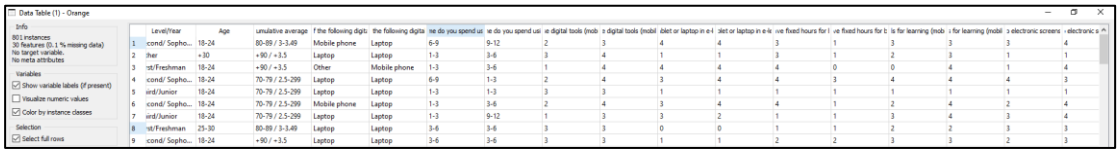
1) Αρχικά θα μετατρέψουμε όλες τις απαντήσεις Likert σε αριθμητικές τιμές σύμφωνα με την εξής κωδικοποίηση:
«Διαφωνώ απολύτως»:0, «Διαφωνώ»:1, «Αβέβαιο»:2, «Συμφωνώ»:3 «Συμφωνώ απόλυτα»:4

   Παρακάτω φαίνεται η υλοποίηση της μετατροπής αυτής για το σύνολο δεδομένων στο περιβάλλον Orange.



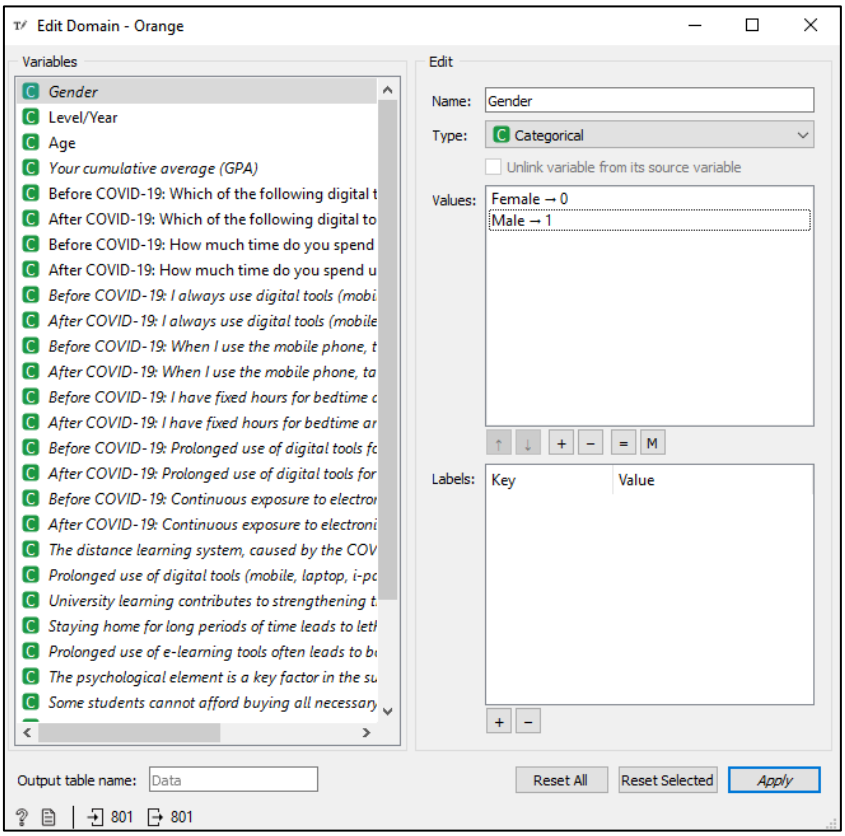
Εικόνα 35: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό «Before COVID-19: I always use digital tools (mobile, laptop, i-pad) in studying».

Μετά την εφαρμογή του πρώτου βήματος, τα δεδομένα έχουν την παρακάτω μορφή:



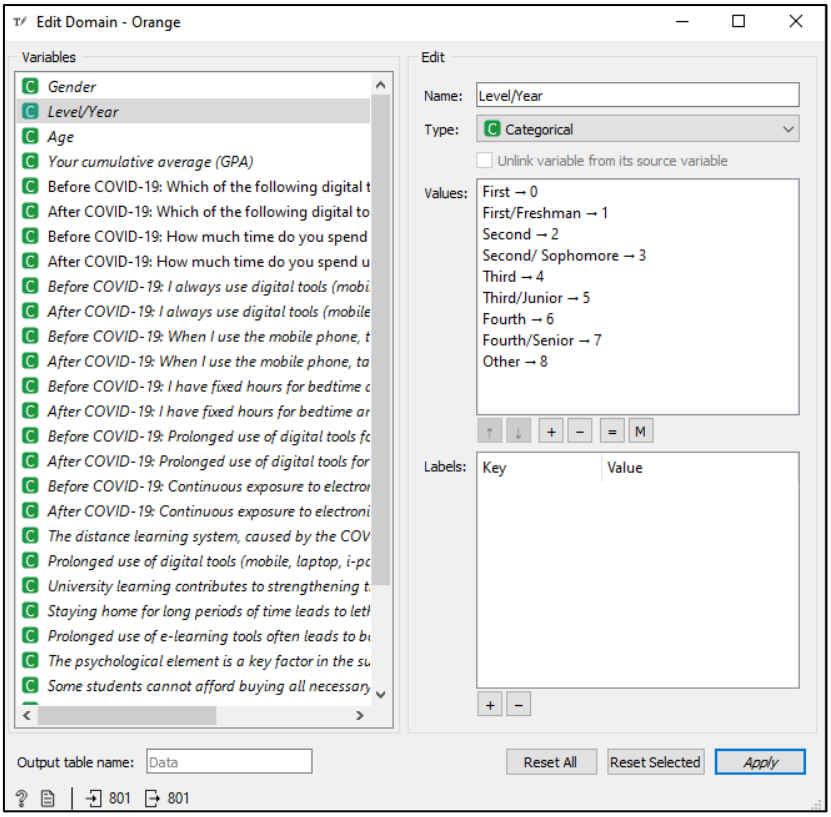
Εικόνα 36: Το σύνολο των δεδομένων μετά την μετατροπή των τιμών

2) Το φύλλο είναι δυαδική στήλη, επομένως για την στήλη «Gender» θα ορίσουμε την κωδικοποίηση: «θηλυκό»: 0 και «αρσενικό»: 1. Η μετατροπή αυτή στο περιβάλλον Orange υλοποιείται με τον τρόπο που φαίνεται στην εικόνα 31.



Εικόνα 37: Μετατροπή των τιμών από κατηγορικές σε δυαδικές για το χαρακτηριστικό Gender

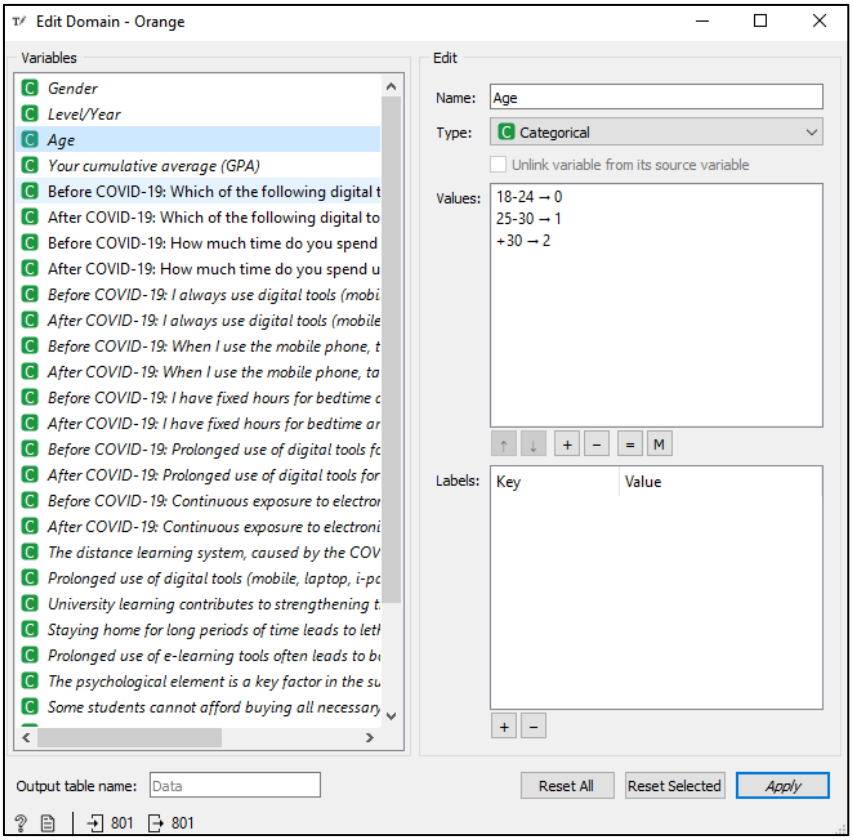
3) Για την στήλη «Level/ Year» η αντιστοιχία της κάθε τιμής θα γίνει με την εξής κωδικοποίηση:
«First»:0, «First/Freshman»:1, «Second»:2, «Second/ Sophomore»:3, «Third»:4, «Third/ Junior»:5, «Fourth»:6, «Fourth/ Senior»:6, «Other»:6



Εικόνα 38: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό Level/ Year

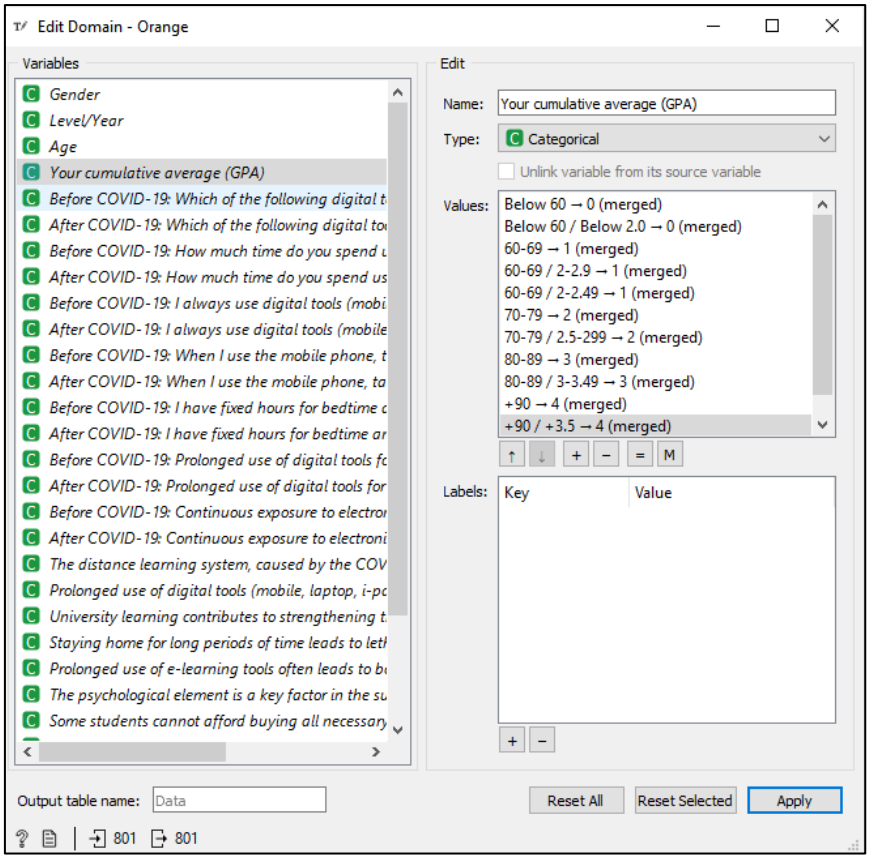
4) Για την στήλη «Age» η αντιστοιχία της κάθε τιμής θα γίνει με την εξής κωδικοποίηση:
«18-24»:0, «25-30»:1, «+30»:2

5)



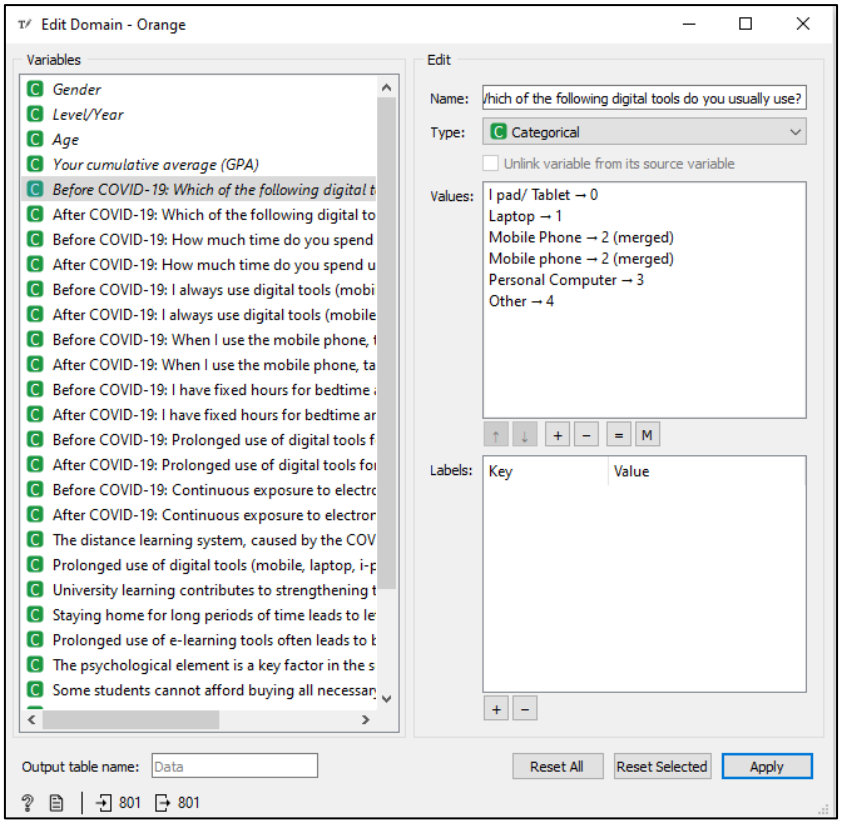
Εικόνα 39: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό Age

6) Για την τιμή στόχο, το GPA, έχουμε αντιστοιχήσει τιμές από το 0 ως το 4 ξεκινώντας από το μικρότερο και πηγαίνοντας προς το μεγαλύτερο. Συγκεκριμένα:
«κάτω από 60» και «κάτω από 60 / Κάτω από 2.0»:0, «60-69», «60-69 / 2-2.9» και «60–69 / 2–2.49»:1, «70-79» και «70–79 / 2,5–299»:2, «80-89» και «80–89 / 3 –3,49»:3, «+90» και «+90 / +3,5»:4

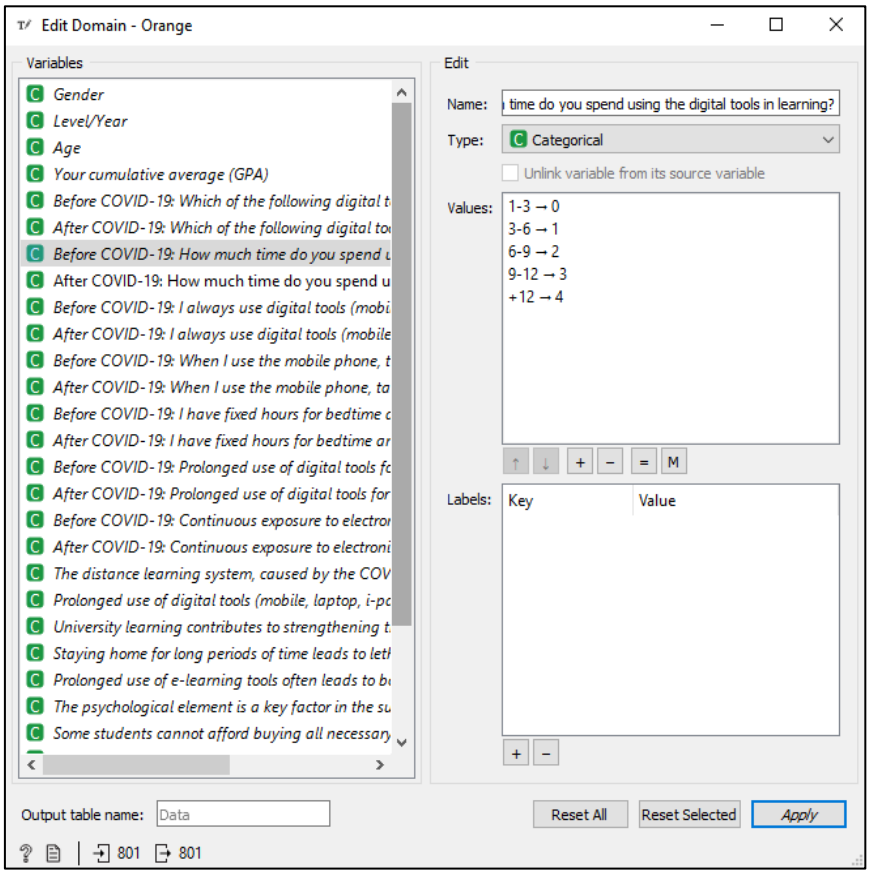


Εικόνα 40: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό GPA

7) Τέλος, για την 5η και 7η στήλη, δηλαδή τα χαρακτηριστικά «Before COVID-19: Which of the following digital tools do you usually use?» και «Before COVID-19: How much time do you spend using the digital tools in learning?» έχουμε την εξής αντιστοιχία:
«I pad/ Tablet»:0, «Laptop»:1, «Mobile Phone»:2, «Personal Computer»:3, «Other»:4 και «1-3»:0, «3-6»:1, «6-9»:2, «9-12»:3, «+12»:4 αντίστοιχα.
Η υλοποίηση φαίνεται στις εικόνες 36 και 37.

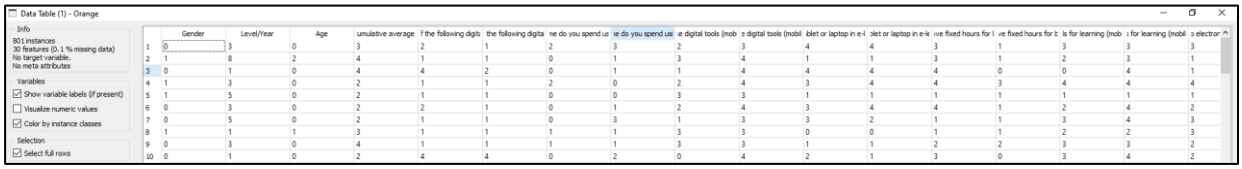


Εικόνα 41: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό «Before COVID-19: Which of the following digital tools do you usually use?»



Εικόνα 42: Μετατροπή των τιμών από κατηγορικές σε αριθμητικές για το χαρακτηριστικό «Before COVID-19: How much time do you spend using the digital tools in learning?»

Αφού ολοκληρώσουμε τα παραπάνω βήματα, τα δεδομένα έχουν έρθει πλέον σε αυτή την μορφή:



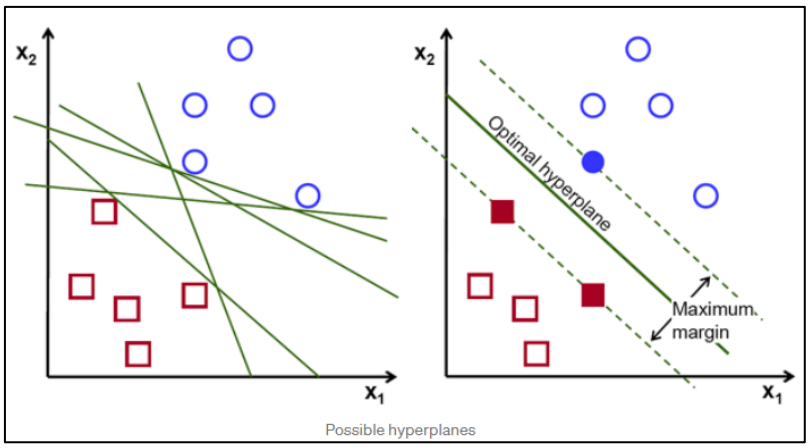
Εικόνα 43: Τελική μορφή του συνόλου δεδομένων μετά την επεξεργασία τους

5.3.4 Διερεύνηση και Αξιολόγηση επίδοσης για διάφορους αλγορίθμους

Στόχος μας είναι να επιλέξουμε ένα μοντέλο το οποίο να κατηγοριοποιεί με όσο το δυνατόν μεγαλύτερη ακρίβεια την πρόβλεψη του μέσου όρου βαθμολογίας (GPA) των φοιτητών σε μια από τις 4 κλάσεις που έχουμε ορίσει. Για να το επιτύχουμε αυτό θα δοκιμάσουμε να εφαρμόσουμε τους παρακάτω αλγορίθμους και στην συνέχεια θα τους αξιολογήσουμε. Οι αλγόριθμοι αυτοί αναφέρονται και παρουσιάζονται παρακάτω:

1) *Support Vector Machine*

Το Support Vector Machine είναι ένας απλός αλγόριθμος που κάθε ειδικός μηχανικής μάθησης πρέπει να έχει στο οπλοστάσιό του. Η μηχανή υποστήριξης διανυσμάτων προτιμάται ιδιαίτερα από πολλούς, καθώς παράγει σημαντική ακρίβεια με λιγότερη υπολογιστική ισχύ. Το Support Vector Machine (SVM) μπορεί να χρησιμοποιηθεί τόσο για εργασίες παλινδρόμησης όσο και για εργασίες ταξινόμησης. Όμως, χρησιμοποιείται ευρέως σε στόχους ταξινόμησης. Ο στόχος του αλγόριθμου της μηχανής διανυσμάτων υποστήριξης είναι να βρει ένα υπερεπίπεδο σε ένα χώρο N-διάστασης (N — ο αριθμός των χαρακτηριστικών) που ταξινομεί ευδιάκριτα τα σημεία δεδομένων.



Εικόνα 44: Αλγόριθμος SVM

Για να διαχωριστούν οι δύο κατηγορίες σημείων δεδομένων, υπάρχουν πολλά πιθανά υπερεπίπεδα που θα μπορούσαν να επιλεγούν. Ο στόχος μας είναι να βρούμε ένα επίπεδο που έχει το μέγιστο περιθώριο, δηλαδή τη μέγιστη απόσταση μεταξύ των σημείων δεδομένων και των δύο κατηγοριών. Η μεγιστοποίηση της απόστασης

περιθωρίου παρέχει κάποια ενίσχυση, έτσι ώστε τα μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη σιγουριά.

2) Decision Tree

Το δέντρο αποφάσεων είναι μια από τις προσεγγίσεις προγνωστικής μοντελοποίησης που χρησιμοποιούνται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. Τα δέντρα αποφάσεων κατασκευάζονται μέσω μιας αλγοριθμικής προσέγγισης που προσδιορίζει τρόπους διαχωρισμού ενός συνόλου δεδομένων με βάση διαφορετικές συνθήκες. Είναι μια από τις πιο ευρέως χρησιμοποιούμενες και πρακτικές μεθόδους για την εποπτευόμενη μάθηση. Τα δέντρα απόφασης είναι μια μη παραμετρική εποπτευόμενη μέθοδος μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Τα μοντέλα δέντρων όπου η μεταβλητή στόχος μπορεί να λάβει ένα διακριτό σύνολο τιμών ονομάζονται δέντρα ταξινόμησης. Τα δέντρα απόφασης όπου η μεταβλητή στόχος μπορεί να λάβει συνεχείς τιμές (συνήθως πραγματικούς αριθμούς) ονομάζονται δέντρα παλινδρόμησης. Το δέντρο ταξινόμησης και παλινδρόμησης (CART) είναι γενικός όρος για αυτό.

3) Random Forest

Το Random Forest, όπως υποδηλώνει το όνομά του, αποτελείται από ένα μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο. Κάθε μεμονωμένο δέντρο στο τυχαίο δάσος βγάζει μια πρόβλεψη τάξης και η τάξη με τις περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας. Η θεμελιώδης ιδέα πίσω από το τυχαίο δάσος είναι μια απλή αλλά ισχυρή ιδέα - η σοφία του πλήθους. Στην επιστήμη των δεδομένων, ο λόγος που το τυχαίο μοντέλο δασών λειτουργεί τόσο καλά είναι: Ένας μεγάλος αριθμός σχετικά μη συσχετισμένων μοντέλων (δέντρων) που λειτουργούν ως επιτροπή θα ξεπεράσει σε απόδοση οποιοδήποτε από τα επιμέρους συστατικά μοντέλα. Η χαμηλή συσχέτιση μεταξύ των μοντέλων είναι το κλειδί. Τα μη συσχετισμένα μοντέλα μπορούν να παράγουν προβλέψεις συνόλου που είναι πιο ακριβείς από οποιαδήποτε από τις μεμονωμένες προβλέψεις. Ο λόγος για αυτό το αποτέλεσμα είναι ότι τα δέντρα προστατεύουν το ένα το άλλο από τα ατομικά τους λάθη (αρκεί να μην κάνουν συνέχεια όλα προς την ίδια κατεύθυνση). Ενώ ορισμένα δέντρα μπορεί να είναι λάθος, πολλά άλλα δέντρα θα είναι σωστά, έτσι ως ομάδα τα δέντρα μπορούν να κινηθούν προς τη σωστή κατεύθυνση. Άρα οι προϋποθέσεις για να έχει καλή απόδοση το τυχαίο δάσος είναι: Πρέπει να υπάρχει κάποιο πραγματικό σήμα στις δυνατότητές μας, έτσι ώστε τα μοντέλα που κατασκευάζονται με αυτές τις δυνατότητες να είναι καλύτερα από την τυχαία εικασία. Οι προβλέψεις (και επομένως τα λάθη) που γίνονται από τα μεμονωμένα δέντρα πρέπει να έχουν χαμηλές συσχετίσεις μεταξύ τους.

4) k-Nearest Neighbors

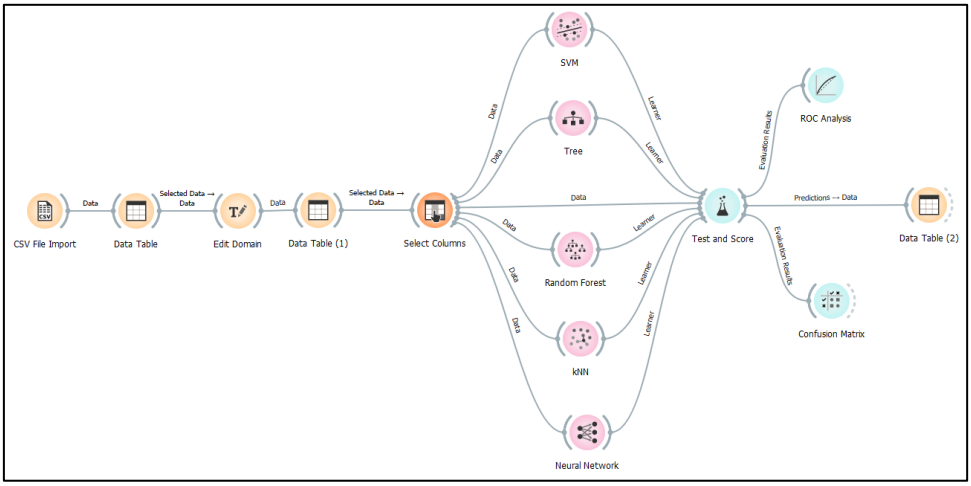
Ο αλγόριθμος k-πλησιέστερων γειτόνων (KNN) είναι μια μέθοδος ταξινόμησης δεδομένων για την εκτίμηση της πιθανότητας ότι ένα σημείο δεδομένων θα γίνει μέλος μιας ή άλλης ομάδας με βάση την ομάδα στην οποία ανήκουν τα σημεία δεδομένων που βρίσκονται πλησιέστερα σε αυτό. Ο αλγόριθμος k-πλησιέστερου γείτονα είναι ένας τύπος εποπτευόμενου αλγόριθμου μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Ωστόσο, χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης. Το KNN είναι ένας «τεμπέλης» μαθησιακός και μη παραμετρικός αλγόριθμος. Ονομάζεται αλγόριθμος «τεμπέλης» μάθησης επειδή δεν

εκτελεί καμία εκπαίδευση όταν παρέχετε τα δεδομένα εκπαίδευσης. Αντίθετα, απλώς αποθηκεύει τα δεδομένα κατά τη διάρκεια του χρόνου εκπαίδευσης και δεν εκτελεί κανέναν υπολογισμό. Δεν δημιουργεί ένα μοντέλο μέχρι να εκτελεστεί ένα ερώτημα στο σύνολο δεδομένων. Αυτό καθιστά το KNN ιδανικό για εξόρυξη δεδομένων. Θεωρείται μη παραμετρική μέθοδος επειδή δεν κάνει υποθέσεις σχετικά με την υποκείμενη κατανομή δεδομένων. Με απλά λόγια, το KNN προσπαθεί να προσδιορίσει σε ποια ομάδα ανήκει ένα σημείο δεδομένων κοιτάζοντας τα σημεία δεδομένων γύρω του.

5) *Artificial Neural Network*

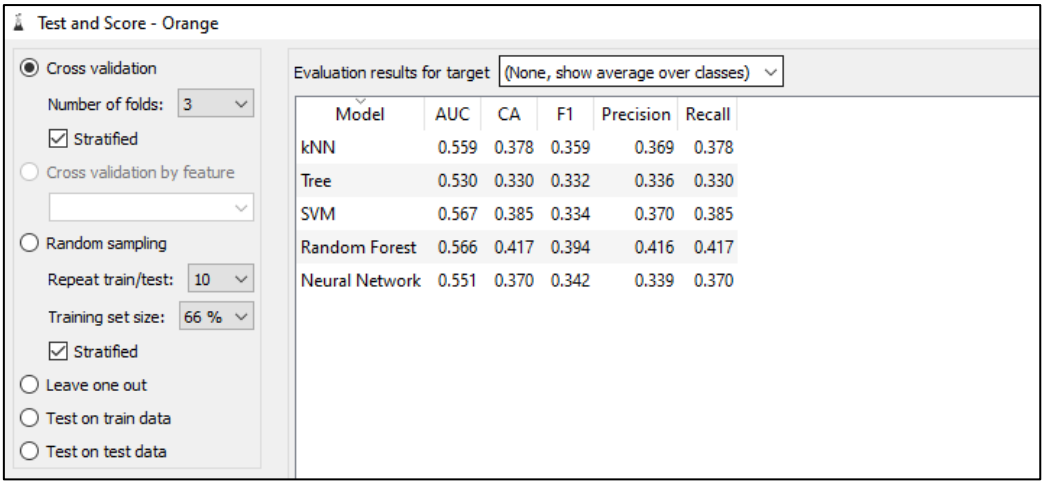
Τα νευρωνικά δίκτυα, γνωστά και ως τεχνητά νευρωνικά δίκτυα (ANN) ή προσομοιωμένα νευρωνικά δίκτυα (SNN), αποτελούν υποσύνολο της μηχανικής μάθησης και βρίσκονται στην καρδιά των αλγορίθμων βαθιάς μάθησης. Το όνομα και η δομή τους είναι εμπνευσμένα από τον ανθρώπινο εγκέφαλο, μιμούμενοι τον τρόπο που οι βιολογικοί νευρώνες δίνουν σήμα ο ένας στον άλλο. Ο τρόπος λειτουργίας των τεχνητών νευρωνικών δικτύων έχει παρουσιαστεί σε προηγούμενο κεφάλαιο αναλυτικά.

Οι αλγόριθμοι που παρουσιάστηκαν παραπάνω θα μας βοηθήσουν στην πρόβλεψη των ακαδημαϊκών επιδόσεων των φοιτητών/φοιτητριών. Συγκεκριμένα, θα εφαρμόσουμε ταυτόχρονα τους αλγορίθμους στο περιβάλλον Orange και στην συνέχεια θα τους αξιολογήσουμε ώστε να συμπεράνουμε ποιος αλγόριθμος τα πηγαίνει καλύτερα. Παρακάτω φαίνεται η εφαρμογή των 5 αυτών αλγορίθμων στο περιβάλλον Orange:



Εικόνα 45: Εφαρμογή των 5 διαφορετικών αλγορίθμων στο περιβάλλον Orange

Το widget «Test and Score» μας δίνει χρήσιμες πληροφορίες για τις επιδόσεις του κάθε μοντέλου αναφέροντας τους δείκτες AUC, CA, F1, Precision και Recall για κάθε μοντέλο ξεχωριστά. Στην παρακάτω φωτογραφία φαίνονται αναλυτικά οι δείκτες για κάθε μοντέλο.



Εικόνα 46: Αξιολόγηση των μοντέλων με χρήση του γραφικού στοιχείου «Test and Score»

Η καμπύλη AUC - ROC είναι μια μέτρηση απόδοσης για τα προβλήματα ταξινόμησης. Το AUC σημαίνει "Περιοχή κάτω από την καμπύλη ROC". Δηλαδή, η AUC μετρά ολόκληρη τη δισδιάστατη περιοχή κάτω από ολόκληρη την καμπύλη ROC από (0,0) έως (1,1). Η AUC παρέχει ένα συνολικό μέτρο απόδοσης σε όλα τα πιθανά όρια ταξινόμησης. Ένας τρόπος ερμηνείας της AUC είναι η πιθανότητα το μοντέλο να κατατάσσει ένα τυχαίο θετικό παράδειγμα υψηλότερα από ένα τυχαίο αρνητικό παράδειγμα. Η τιμή AUC κυμαίνεται από 0 έως 1. Ένα μοντέλο του οποίου οι προβλέψεις είναι 100% λανθασμένες έχει AUC 0,0. Αυτός του οποίου οι προβλέψεις είναι 100% σωστές έχει AUC 1,0. Στην εικόνα 41 παρατηρούμε πως για τον αλγόριθμο SVM έχουμε το μεγαλύτερο AUC=5.667. Αμέσως μετά, έρχεται ο αλγόριθμος Random Forest με AUC=5.666.

Η ακρίβεια (Precision) και η ανάκληση (Recall) είναι μετρήσεις απόδοσης που χρησιμοποιούνται για την αναγνώριση και ταξινόμηση προτύπων στη μηχανική μάθηση. Αυτές οι έννοιες είναι απαραίτητες για τη δημιουργία ενός τέλειου μοντέλου μηχανικής μάθησης που δίνει πιο ακριβή και ακριβή αποτελέσματα. Ορισμένα από τα μοντέλα στη μηχανική μάθηση απαιτούν μεγαλύτερη ακρίβεια και ορισμένα μοντέλα απαιτούν περισσότερη ανάκληση. Επομένως, είναι σημαντικό να γνωρίζουμε την ισορροπία μεταξύ Ακρίβειας και ανάκλησης ή, απλά, αντιστάθμισης ακριβείας-ανάκλησης. Η ακρίβεια ορίζεται ως η αναλογία των σωστά ταξινομημένων θετικών δειγμάτων (True Positive) προς έναν συνολικό αριθμό ταξινομημένων θετικών δειγμάτων (είτε σωστά είτε λανθασμένα). Συγκεκριμένα, ο τύπος της ακρίβειας ορίζεται ως:

Precision = True Positive/True Positive + False Positive ή Precision = TP/TP+FP

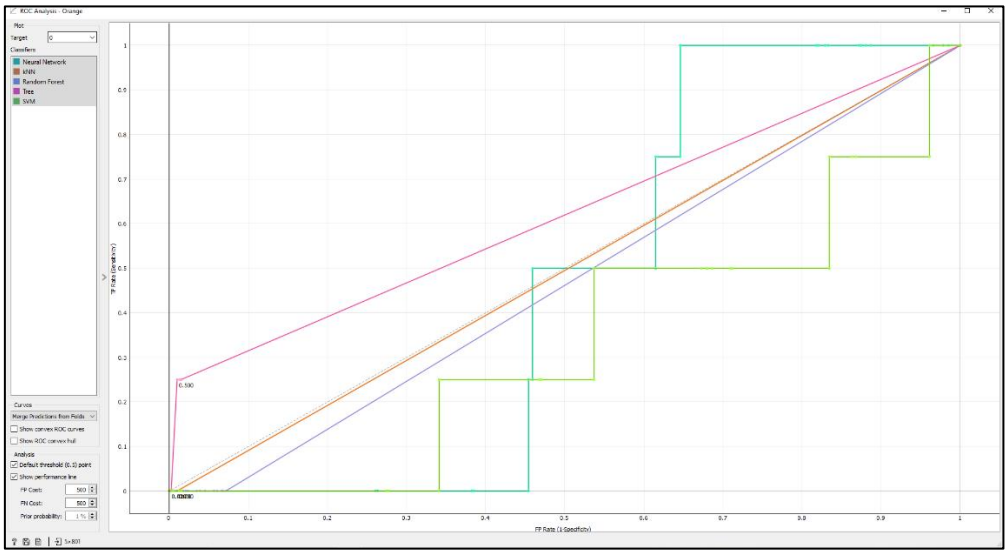
Η ακρίβεια ενός μοντέλου μηχανικής μάθησης θα είναι χαμηλή όταν η τιμή του TP+FP (παρονομαστής) > TP (Αριθμητής), ενώ η ακρίβεια του μοντέλου μηχανικής εκμάθησης θα είναι υψηλή όταν η τιμή του TP (Αριθμητής) > TP+FP (παρονομαστής). Η ανάκληση υπολογίζεται ως η αναλογία μεταξύ των αριθμών των Θετικών δειγμάτων που ταξινομήθηκαν σωστά ως Θετικά προς τον συνολικό αριθμό των Θετικών δειγμάτων. Η ανάκληση μετρά την ικανότητα του μοντέλου να ανιχνεύει θετικά δείγματα. Όσο υψηλότερη είναι η ανάκληση, τόσο περισσότερα θετικά δείγματα ανιχνεύονται. Ο τύπος της ανάκλησης ορίζεται ως:

Recall = True Positive/True Positive + False Negative ή Recall = TP/TP+FN

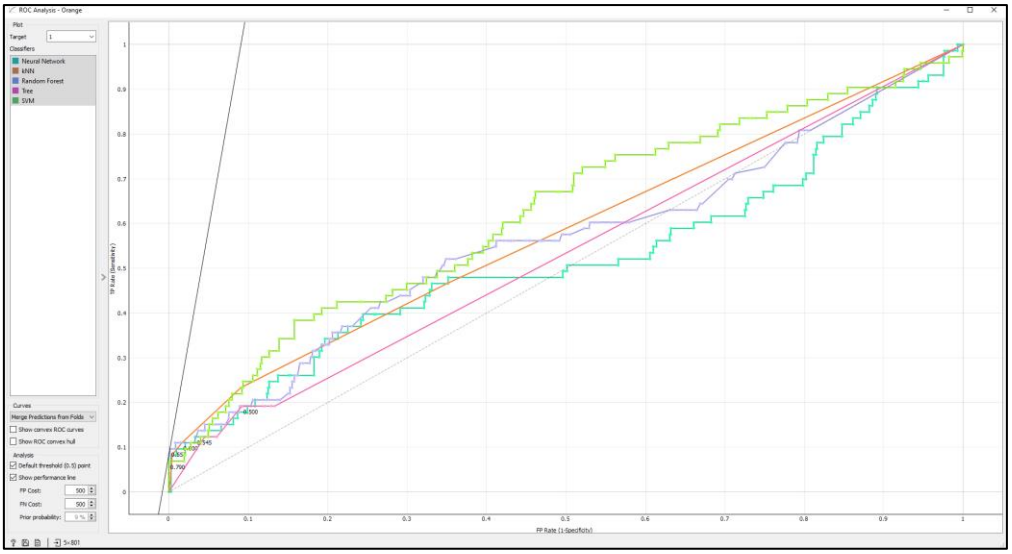
Η ανάκληση ενός μοντέλου μηχανικής μάθησης θα είναι χαμηλή όταν η τιμή του

TP+FN (παρονομαστής) > TP (Αριθμητής), ενώ η ανάκληση του μοντέλου μηχανικής εκμάθησης θα είναι υψηλή όταν η τιμή του TP (Αριθμητής) > TP+FN (παρονομαστής). Σε αντίθεση με το Precision, η Ανάκληση είναι ανεξάρτητη από τον αριθμό των αρνητικών ταξινομήσεων δειγμάτων. Επιπλέον, εάν το μοντέλο ταξινομήσει όλα τα θετικά δείγματα ως θετικά, τότε η Ανάκληση θα είναι 1. Παρατηρούμε στην εικόνα 41 πως ο αλγόριθμος Random Forest έχει την μεγαλύτερη ακρίβεια και ανάκληση, συγκεκριμένα Precision=0.416 και Recall=0.417. Η βαθμολογία F1 (F1 Score) ορίζεται ως η αρμονική μέση ακρίβεια και ανάκληση. Ως σύντομη υπενθύμιση, ο αρμονικός μέσος όρος είναι μια εναλλακτική μέτρηση για τον πιο κοινό αριθμητικό μέσο όρο. Είναι συχνά χρήσιμο κατά τον υπολογισμό ενός μέσου ρυθμού. Στη βαθμολογία F1, υπολογίζουμε το μέσο όρο της ακρίβειας και της ανάκλησης. Είναι και οι δύο ρυθμοί, γεγονός που καθιστά λογική επιλογή τη χρήση του αρμονικού μέσου όρου. Ένα μοντέλο θα λάβει υψηλή βαθμολογία F1 εάν τόσο η Ακρίβεια όσο και η Ανάκληση είναι υψηλή. Όπως είναι προφανές, εφόσον έχουμε την μεγαλύτερη ακρίβεια και ανάκληση για τον αλγόριθμο Random Forest, θα έχουμε και την μεγαλύτερη F1 βαθμολογία (F1=0.394). Τέλος, η ακρίβεια ταξινόμησης (CA) μετρά τον αριθμό των σωστών προβλέψεων που έγιναν διαιρεμένο με τον συνολικό αριθμό των προβλέψεων. Για τον αλγόριθμο Random Forest έχουμε CA=0.417, το μεγαλύτερο από όλα τα άλλα. Το συμπέρασμα που προκύπτει, σύμφωνα με το πείραμα που πραγματοποιήθηκε, είναι πως ο αλγόριθμος Random Forest τα πηγαίνει καλύτερα σε σύγκριση με κάθε άλλο αλγόριθμο για το σύνολο δεδομένων μας.

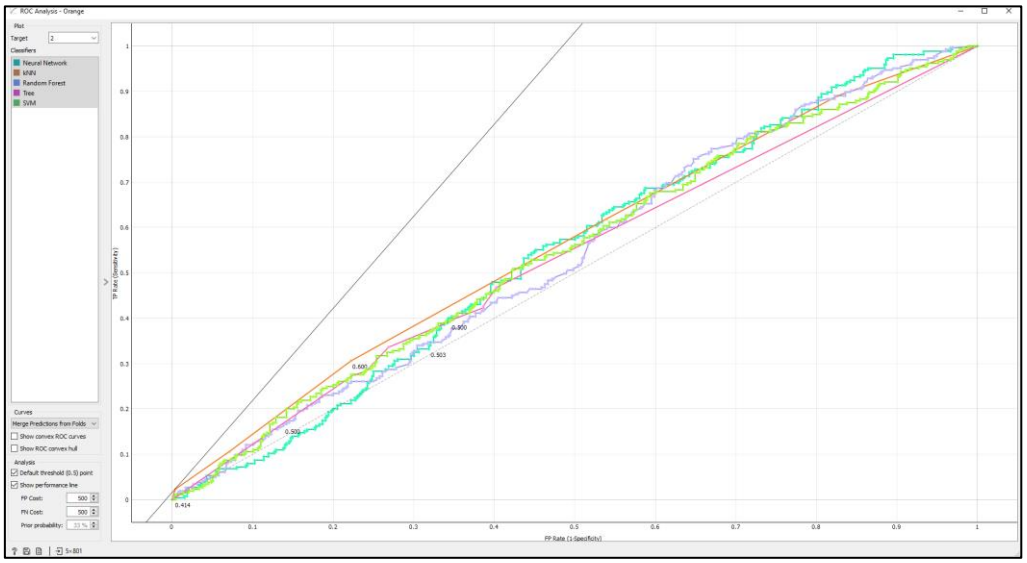
Η καμπύλη ROC και η βαθμολογία ROC AUC είναι σημαντικά εργαλεία για την αξιολόγηση μοντέλων ταξινόμησης. Συνοπτικά, μας δείχνουν τη δυνατότητα διαχωρισμού των κλάσεων με όλα τα πιθανά όρια, ή με άλλα λόγια, πόσο καλά το μοντέλο ταξινομεί κάθε κατηγορία. Στις εικόνες 42 έως 46 παρουσιάζονται οι καμπύλες ROC για κάθε κλάση ξεχωριστά.



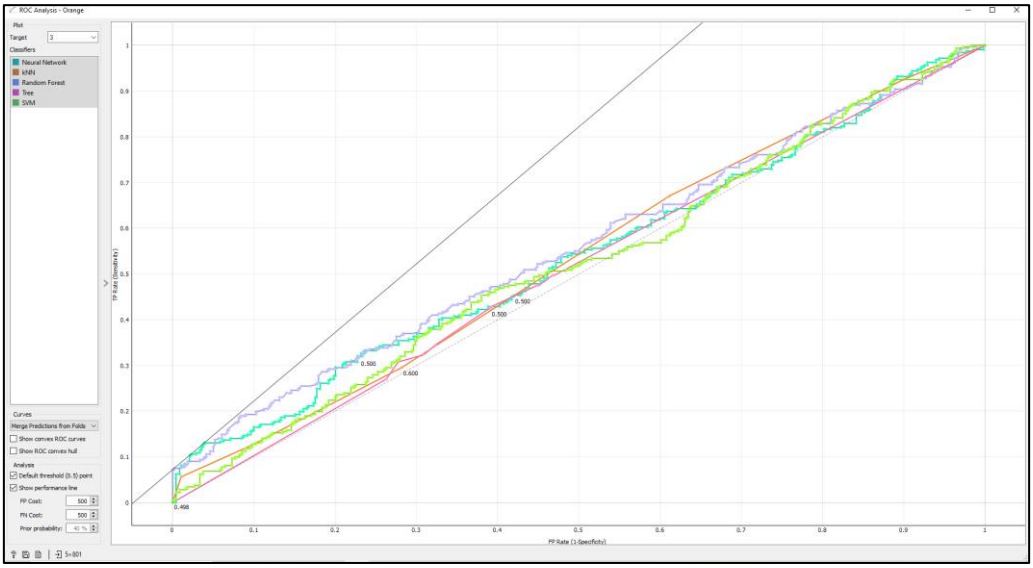
Εικόνα 47: Αποτελέσματα του γραφικού στοιχείου "Roc Analysis" για την κλάση 0 για κάθε μοντέλο



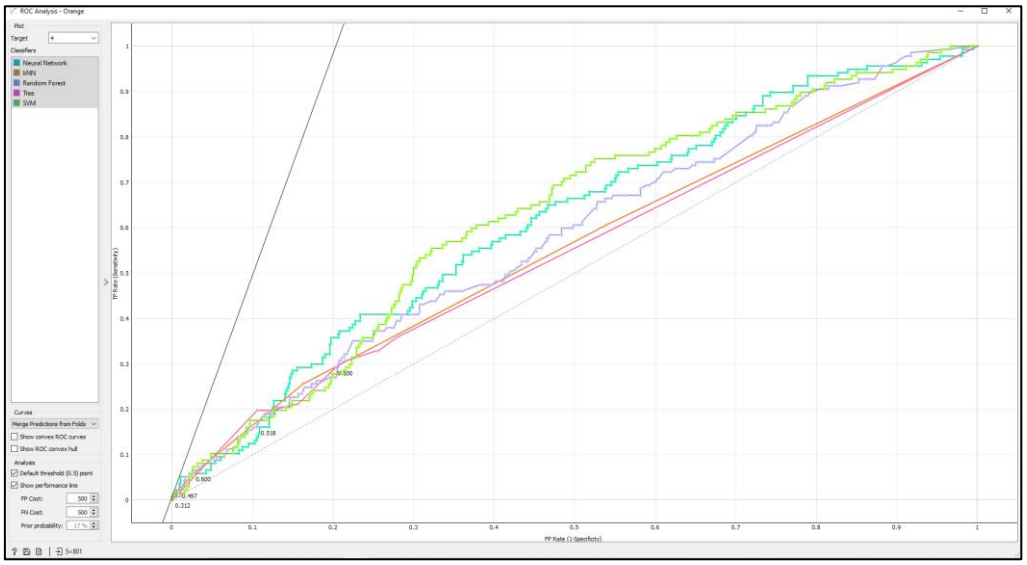
Εικόνα 48: Αποτελέσματα του γραφικού στοιχείου "Roc Analysis" για την κλάση 1 για κάθε μοντέλο



Εικόνα 49: Αποτελέσματα του γραφικού στοιχείου "Roc Analysis" για την κλάση 2 για κάθε μοντέλο

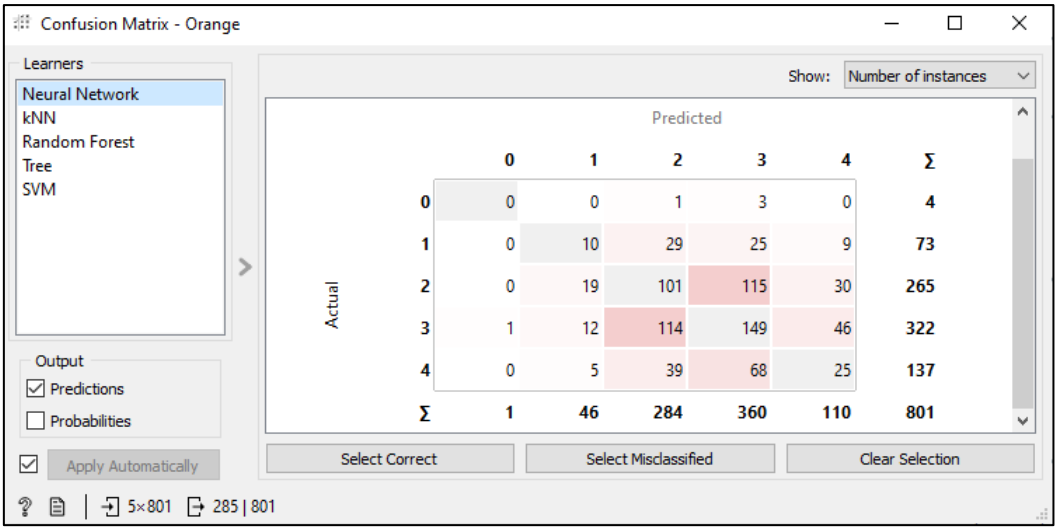


Εικόνα 50: Αποτελέσματα του γραφικού στοιχείου "Roc Analysis" για την κλάση 3 για κάθε μοντέλο

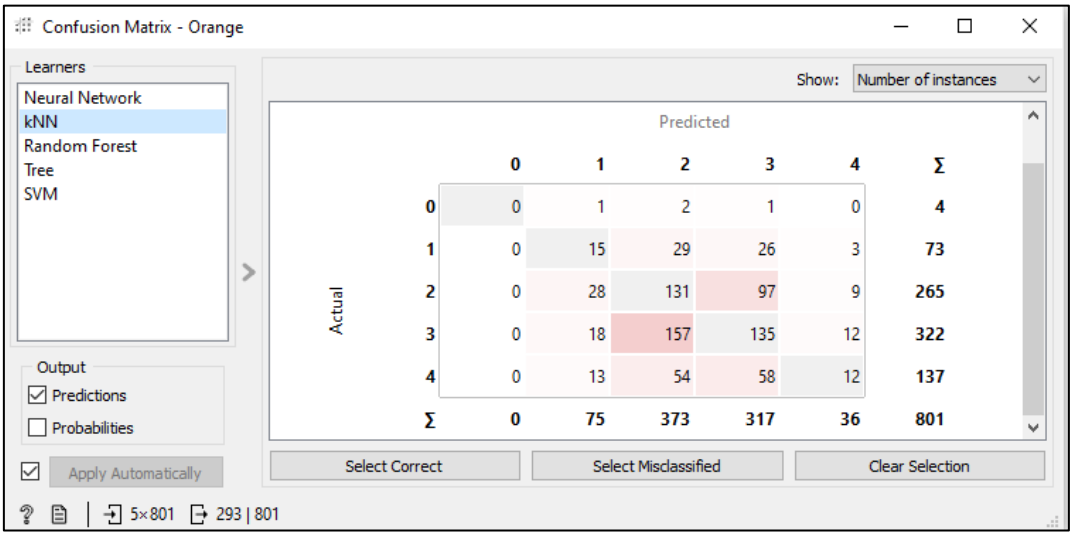


Εικόνα 51: Αποτελέσματα του γραφικού στοιχείου "Roc Analysis" για την κλάση 4 για κάθε μοντέλο

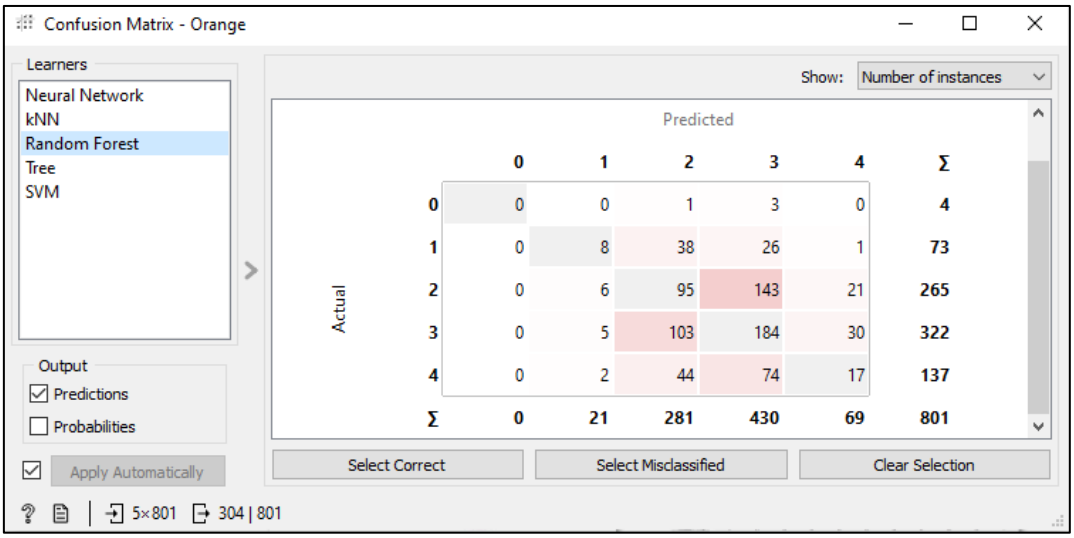
Ο πίνακας σύγχυσης (Confusion Matrix) μας βοηθά να εμφανίσουμε την απόδοση ενός μοντέλου ή πώς ένα μοντέλο έχει κάνει την πρόβλεψή του στη Μηχανική Μάθηση. Ένας πίνακας σύγχυσης είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης σε ένα πρόβλημα ταξινόμησης. Ο αριθμός των σωστών και εσφαλμένων προβλέψεων συνοψίζεται με τιμές μέτρησης και αναλύεται ανά κατηγορία. Αυτό είναι το κλειδί για τη μήτρα σύγχυσης. Ο πίνακας σύγχυσης δείχνει τους τρόπους με τους οποίους το μοντέλο ταξινόμησής σας μπερδεύεται όταν κάνει προβλέψεις. Επιπλέον δίνει πληροφορίες όχι μόνο για τα σφάλματα που γίνονται από τον ταξινομητή, αλλά και για τους τύπους των σφαλμάτων που γίνονται. Στις εικόνες 47-51 παρουσιάζεται ο πίνακας σύγχυσης για κάθε μοντέλο ξεχωριστά.



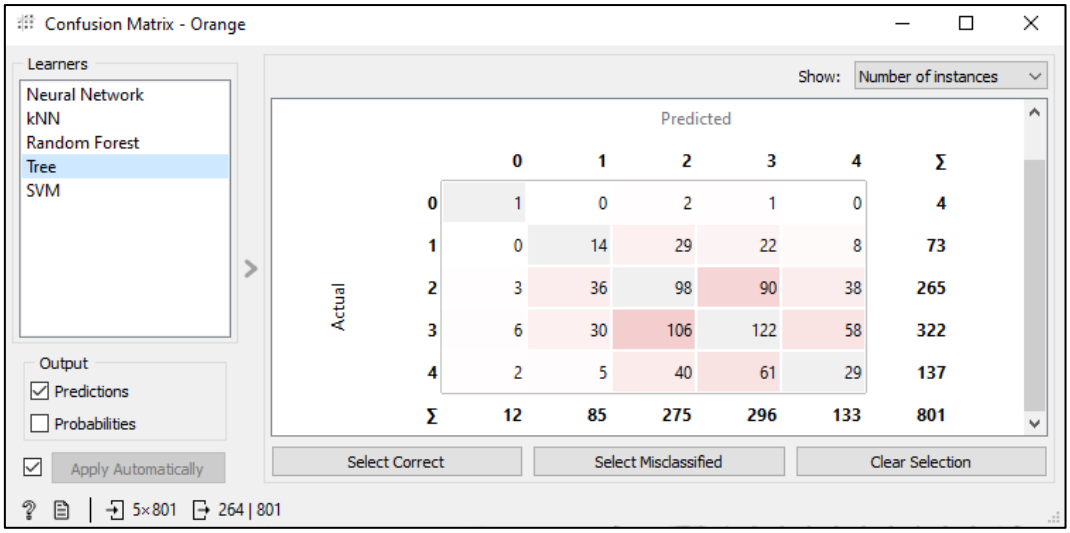
Εικόνα 52: Αποτελέσματα του γραφικού στοιχείου "Confusion Matrix" για το νευρωνικό δίκτυο



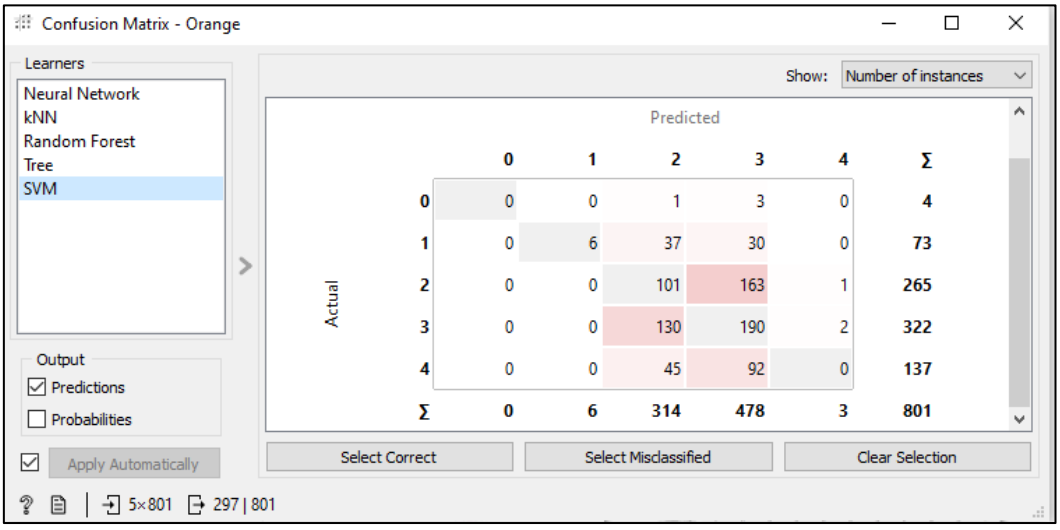
Εικόνα 53: Αποτελέσματα του γραφικού στοιχείου "Confusion Matrix" για τον αλγόριθμο kNN



Εικόνα 54: Αποτελέσματα του γραφικού στοιχείου "Confusion Matrix" για τον αλγόριθμο Random Forest

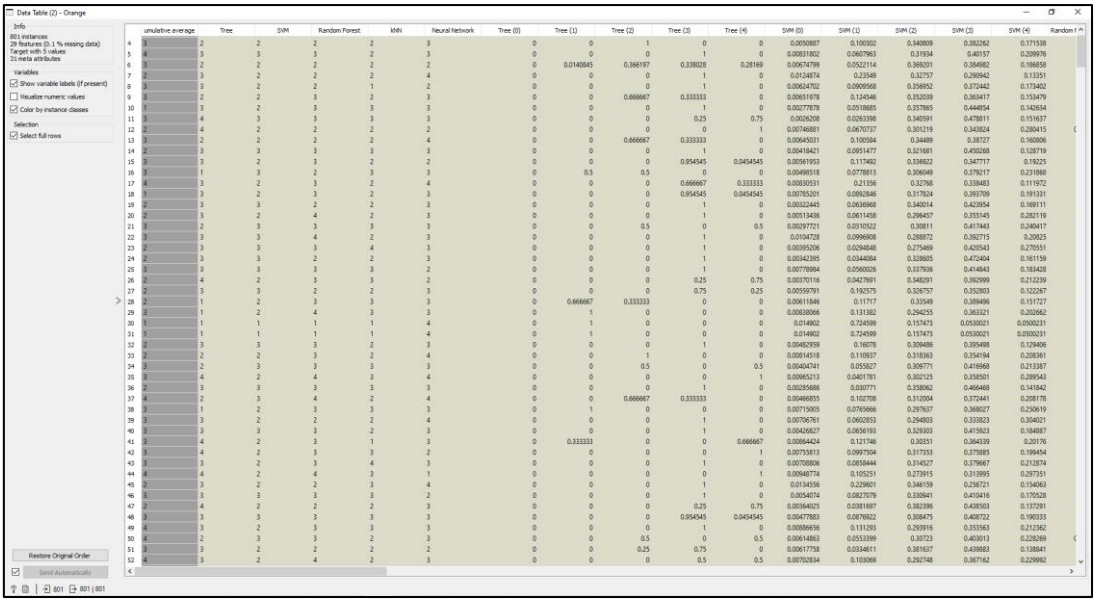


Εικόνα 55: Αποτελέσματα του γραφικού στοιχείου "Confusion Matrix" για τον αλγόριθμο Tree

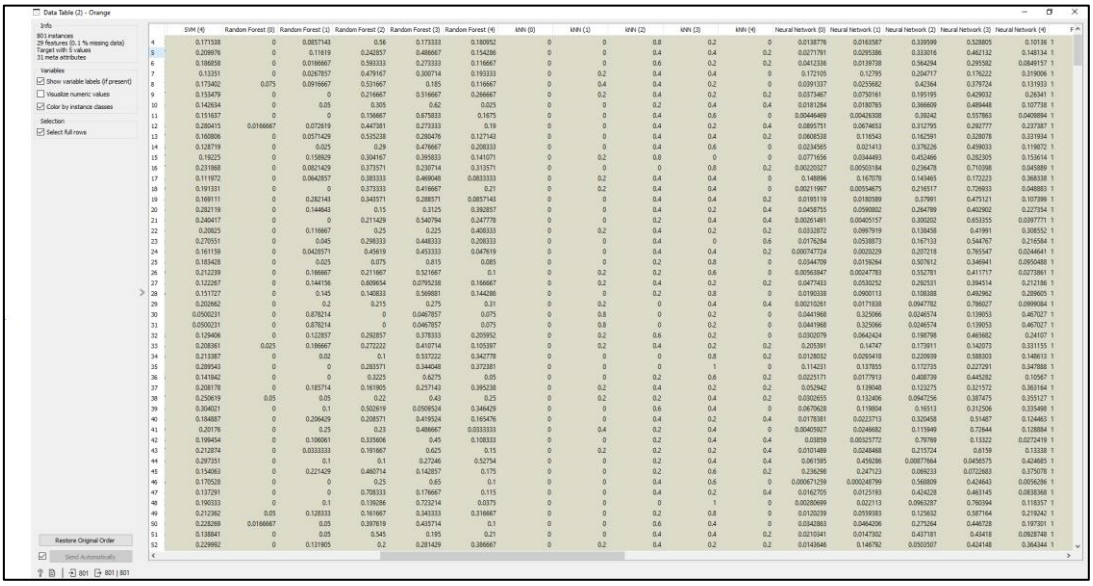


Εικόνα 56: Αποτελέσματα του γραφικού στοιχείου "Confusion Matrix" για τον αλγόριθμο SVM

Στην συνέχεια, ακολουθούν φωτογραφίες από τις πραγματικές και προβλεπόμενες τιμές για κάθε στήλη για κάθε μοντέλο ξεχωριστά.



Εικόνα 57: Οι προβλεπόμενες τιμές ανά μοντέλο (1)



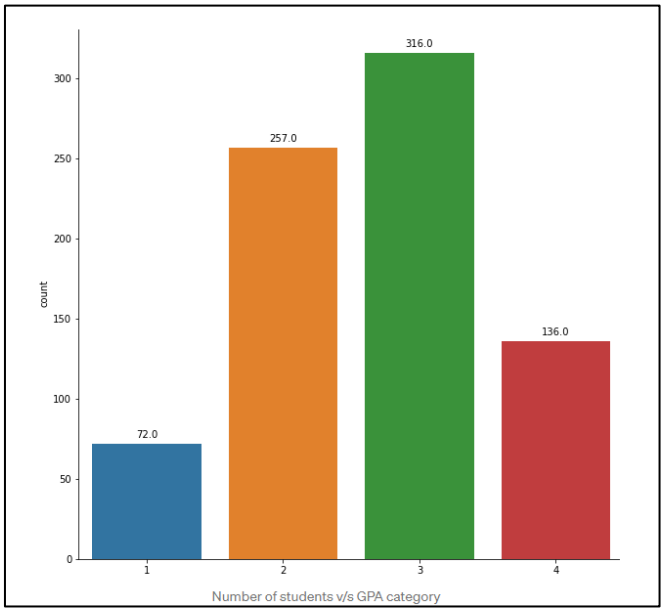
Εικόνα 58: Οι προβλεπόμενες τιμές ανά μοντέλο (2)

5.4 Αποτελέσματα

Έπειτα από τα πειράματα που παρουσιάστηκαν παραπάνω, χρησιμοποιώντας διαφορετικούς αλγορίθμους ταξινόμησης, παρατηρούμε πως καλύτερη επίδοση παρουσίασε ο αλγόριθμος Random Forest. Συγκεκριμένα, οι δείκτες απόδοσης του αλγορίθμου είναι AUC=5.666, Precision=0.416, Recall=0.417, F1=0.394 και CA=0.417. Θα πρέπει να αναφερθεί πως οι δείκτες αυτοί μπορούν να θεωρηθούν χαμηλοί για ένα μοντέλο το οποίο προβλέπει αποτελεσματικά και με ακρίβεια. Το γεγονός αυτό μπορεί να συμβαίνει είτε γιατί μπορεί να υπάρχει κάποια αστοχία στην επιλογή των χαρακτηριστικών του συνόλου δεδομένων, είτε γιατί υπάρχει μια ανισορροπία τάξης στο σύνολο δεδομένων μας.

6. Συμπεράσματα και μελλοντική επέκταση

Η πρόβλεψη των ακαδημαϊκών επιδόσεων των μαθητών, ανεξαρτήτως σχολικής βαθμίδας, αποτελεί μια δύσκολη πρόκληση καθώς είναι αρκετοί οι παράγοντες που επηρεάζουν την ακρίβεια της πρόβλεψης. Ένας από τους βασικότερους παράγοντες που συμβάλλει στις επιδόσεις των μοντέλων είναι το σύνολο δεδομένων και τα χαρακτηριστικά αυτού. Ως μελλοντική επέκταση για την βελτίωση της απόδοσης των αλγορίθμων που χρησιμοποιήθηκαν στα παραπάνω πειράματα, θα μπορούσε να χρησιμοποιηθεί η τεχνική "Synthetic Minority Oversampling Technique" ή αλλιώς "SMOTE" με στόχο την εξισορρόπηση του πλήθους των κλάσεων. Το σύνολο δεδομένων που χρησιμοποιήθηκε στο δεύτερο πείραμα, είχε 4 διαφορετικές κλάσεις ανάλογα με την βαθμολογία (GPA) των μαθητών. Όπως φαίνεται στην παρακάτω εικόνα, 72 μαθητές ανήκουν στην 1η κλάση, 257 στην 2η, 316 στην 3η και 136 στην 4η.



Εικόνα 59: Ο αριθμός των μαθητών ανά κατηγορία GPA

Η επικρατούσα ανισορροπία τάξης στο σύνολο δεδομένων μας θα πρέπει να ληφθεί υπόψιν, καθώς η πρόκληση της εργασίας με μη ισορροπημένα σύνολα δεδομένων είναι ότι οι περισσότερες τεχνικές μηχανικής μάθησης θα αγνοήσουν και στη συνέχεια θα έχουν κακή απόδοση στην κατηγορία μειοψηφίας, αν και συνήθως η απόδοση στην κατηγορία μειοψηφίας είναι η πιο σημαντική. Για να το ξεπεραστεί αυτό το εμπόδιο, χρησιμοποιούμε την τεχνική SMOTE. Η βασική λειτουργία της τεχνικής αυτής είναι ότι θα υπερπληθίσει την τάξη μειοψηφίας συνθέτοντας νέα παραδείγματα επιλέγοντας εκείνα που βρίσκονται κοντά στον χώρο χαρακτηριστικών του. Με άλλα λόγια, θα δημιουργήσει νέες σειρές για την κλάση που υποεκπροσωπείται στο σύνολο δεδομένων μας. Ωστόσο, αυτή η επέκταση δεν είναι σκοπός της παρούσας διπλωματικής έρευνας και δεν θα υλοποιηθεί.

Εν κατακλείδι, σύμφωνα με τα πειράματα που διενεργήθηκαν είναι φανερό ότι υπάρχει ισχυρή σχέση μεταξύ της συμπεριφοράς των μαθητών και των ακαδημαϊκών επιδόσεων τους. Η πρόβλεψη των ακαδημαϊκών επιδόσεων των μαθητών μπορεί να θεωρηθεί χρήσιμη σε πολλά πλαίσια. Αν τα εκπαιδευτικά ιδρύματα είναι σε θέση να προβλέψουν την ακαδημαϊκή επίδοση των μαθητών νωρίτερα από τις τελικές τους

εξετάσεις, τότε θα μπορέσουν να καταβληθούν πρόσθετες προσπάθειες αλλά και η κατάλληλη βοήθεια ώστε να υποστηριχθούν αποτελεσματικά οι μαθητές. Επιπλέον, η αναγνώριση μαθητών που οδεύουν προς την εγκατάλειψη του σχολικού περιβάλλοντος είναι αρκετά σημαντική ώστε να ληφθούν απαραίτητα μέτρα που θα συμβάλλουν στην πρόληψη της σχολικής διαρροής. Είναι λοιπόν προφανές πως η κατανόηση και ανάλυση εκπαιδευτικών δεδομένων των μαθητών που υποδηλώνει την απόδοσή τους στην εκπαίδευση παράγει συγκεκριμένους κανόνες και προβλέψεις που βοηθά τους μαθητές στη μελλοντική τους πρόοδο.

Βιβλιογραφία

Ακολουθούν οι βιβλιογραφικές αναφορές (πηγές) της εργασίας.

Βιβλία:

[1] : Κύρκος Ε. (2015). *Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

Άρθρα:

[1] : Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199).

[2] : Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports* (pp. T2G-7). IEEE.

[3] : Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of medical systems*, *43*(6), 1-15.

[4] : Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial intelligence and machine learning to predict student performance during the COVID-19. *Procedia Computer Science*, *184*, 835-840.

[5] : Alsammak, I. L. H., Mohammed, A. H., & Nasir, I. S. (2022). E-learning and COVID-19: Predicting Student Academic Performance Using Data Mining Algorithms. *Webology*, *19*(1), 3419-3432.

[6] : Nachouki, M., & Abou Naaj, M. (2022). Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm. *International Journal of Distance Education Technologies (IJDET)*, *20*(1), 1-17.

[7] : Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. H. M. (2021). Mining educational data to predict students' performance. *Education and Information Technologies*, *26*(5), 6051-6067.

[8] : Trakunphutthirak, R., & Lee, V. C. (2022). Application of educational data mining approach for student academic performance prediction using progressive temporal data. *Journal of Educational Computing Research*, *60*(3), 742-776.

[9] : Abdelhafez, H. A., & Elmannai, H. (2022). Developing and comparing data mining algorithms that work best for predicting student performance. *International Journal of Information and Communication Technology Education (IJICTE)*, *18*(1), 1-14.

[10] : Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students' performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24*(6), 3577-3589.