ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΕΠΙΚΟΙΝΩΝΙΑΣ & ΣΠΟΥΔΩΝ ΔΙΑΔΙΚΤΥΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

"ΣΥΣΤΗΜΑ ΠΡΟΒΛΕΨΗΣ ΕΠΙΔΟΣΗΣ ΦΟΙΤΗΤΩΝ ΕΣΔ ΒΑΣΕΙ ΤΗΣ ΕΠΙΔΟΣΗΣ ΤΟΥΣ ΣΤΙΣ ΠΡΟΕΙΣΑΓΩΓΙΚΕΣ ΕΞΕΤΑΣΕΙΣ"

Στέφανος Βρυωνίδης

Λεμεσός 2015

Πνευματικά δικαιώματα

Η έγκριση της πτυχιακή εργασίας από το Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου του Τεχνολογικού Πανεπιστημίου Κύπρου υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Περίληψη

Η παρούσα πτυχιακή εργασία με τίτλο " Σύστημα πρόβλεψης επίδοσης φοιτητών ΕΣΔ βάση της επίδοσης τους στις προεισαγωγικές εξετάσεις " εκπονήθηκε από τον Στέφανο Βρυωνίδη, φοιτητή του 8ου εξαμήνου του Τμήματος ΕΣΔ του ΤΕΠΑΚ υπό την επίβλεψη του καθηγητή Νικόλα Τσαπατσούλη και ολοκληρώθηκε τον Μάιο του 2015.

Σκοπός της παρούσας εργασίας είναι η δημιουργία ενός συστήματος μέσω του οποίου θα γίνεται πρόβλεψη της επίδοσης των φοιτητών του Τμήματος Επικοινωνίας και Σπουδών Διαδικτύου με βάση της επίδοσης τους στις προεισαγωγικές εξετάσεις. Η υλοποίηση του συστήματος αυτού έγινε με βάση τις τεχνικές της τεχνητής νοημοσύνης και πιο συγκεκριμένα με χρήση νευρωνικών δικτύων. Μετά την συλλογή των δεδομένων, χρησιμοποιήθηκε το WEKA το οποίο είναι ένα πρόγραμμα που περιλαμβάνει μια μεγάλη συλλογή αλγορίθμων μηχανικής μάθησης καθώς και εργαλεία προεπεξεργασίας δεδομένων. Μέσω του WEKA εκπαίδευσα διάφορους αλγόριθμους με τα δεδομένα εισόδου και ο αλγόριθμος με τα πιο έγκυρα αποτελέσματα επιλέχθηκε για να εφαρμοστεί στο σύστημα. Από τα δεδομένα το 70% χρησιμοποιήθηκε για εκπαίδευση του συστήματος και το υπόλοιπο 30% χρησιμοποιήθηκε για έλεγχο των προβλέψεων. Ο αλγόριθμος που επιλέχθηκε είναι ο Multilayer Perceptron ο οποίος ανήκει στα νευρωνικά δίκτυα και τα αποτελέσματα του ήταν αποδεκτά καθώς ο μέσος όρος λάθους της πρόβλεψης έφθανε μέχρι 0.9 μονάδες το οποίο πάνω κάτω είναι μικρή διαφορά.

Αφιέρωση

Η παρούσα εργασία αφιερώνεται σε εμένα και σε όλους τους συμφοιτητές μου από το τμήμα Επικοινωνίας και Σπουδών Διαδικτύου του Τεχνολογικού Πανεπιστημίου Κύπρου, με τους οποίους περάσαμε τέσσερα υπέροχα χρόνια και τους εύχομαι καλή σταδιοδρομία σε ότι και αν επιλέξουν να κάνουν.

Η παρούσα εργασία δεν θα μπορούσε να ολοκληρωθεί χωρίς την ύπαρξη του επιβλέποντος καθηγητή κ. Τσαπατσούλη. Ιδιαίτερες ευχαριστίες στον υπολογιστή μου που με στήριξε στις ατέλειωτες ώρες μελέτης και συγγραφής. Επίσης, θα ήθελα να ευχαριστήσω και όλους τους καθηγητές του Τμήματος που μας πρόσφεραν απλόχερα τις γνώσεις τους.

Πίνακας Περιεχομένων

Λίστα Πινάκων

Λίστα Διαγραμμάτων

Κεφάλαιο 1: Εισαγωγή

Η Τεχνητή Νοημοσύνη βοήθησε πολλούς τομείς να αναπτυχθούν και να εξελιχθούν όπως για παράδειγμα στο τομέα της υγείας, της τεχνολογίας και της εκπαίδευσης. Είναι σχετικά ένα νέο ερευνητικό πεδίο και η πρόοδος της οφείλεται στην ραγδαία ανάπτυξη των υπολογιστικών συστημάτων. Κύριο μέλημα, είναι η μηχανή να συμπεριφέρεται σαν άνθρωπος, δηλαδή να σκέφτεται και να αντιδρά λογικά όπως και να συμπεριφέρεται και να σκέφτεται όπως ο άνθρωπος.

Στη παρούσα μελέτη αναπτύχθηκε ένα σύστημα το οποίο προβλέπει την επίδοση των φοιτητών του Τμήματος Επικοινωνίας και Σπουδών Διαδικτύου του Τεχνολογικού Πανεπιστήμιου Κύπρου, με βάση τους βαθμούς των Παγκύπριων Εξετάσεων και τα μαθήματα τα οποία εξετάστηκαν. Έτσι μετά την συλλογή και αποθήκευση των δεδομένων από τους απόφοιτους των προηγούμενων χρόνων του τμήματος, δημιουργήθηκε ένα σύστημα το οποίο κάνει χρήση τεχνικών της τεχνητής νοημοσύνης και δίνει την δυνατότητα στους χρήστες οι οποίοι θα είναι κυρίως μαθητές / υποψήφιοι φοιτητές και καθηγητές του τμήματος, να εισάγουν τις βαθμολογίες τους και να πάρουν την πρόβλεψη. Μέσω αυτού θα βοηθηθούν πολλά άτομα τα οποία είναι μπερδεμένα και δεν γνωρίζουν τι ακριβώς είναι το Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου.

Ερευνητικά Ερωτήματα:

Το κύριο ερευνητικό ερώτημα της παρούσας πτυχιακή εργασίας είναι αν μπορεί το σύστημα να προβλέψει τη μεταβλητή (τελική βαθμολογία) μέσω ενός συγκεκριμένου αλγορίθμου. Ο αλγόριθμος θα πρέπει να υπολογίζει την τελική βαθμολογία έχοντας ως μεταβλητές εισόδου τις βαθμολογίες των μαθητών από τις Παγκύπριες εξετάσεις.

Επίσης, ένα επιμέρους ερευνητικό ερώτημα είναι πως τα μαθήματα επηρεάζουν την τελική πρόβλεψη (θετικά, αρνητικά) και σε ποιό βαθμό.

Κεφάλαιο 2: Περιγραφή Προβλήματος - Αναγκαιότητα Μελέτης

Το Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου είναι διεπιστημονικό δηλαδή τα μαθήματα του χωρίζονται σε δύο διαφορετικές κατηγορίες οι οποίες είναι τα τεχνολογικά μαθήματα και τα θεωρητικά. Η απόδοση των φοιτητών διακυμαίνεται καθώς οι πλείστοι καταφέρνουν να έχουν καλή βαθμολογία στην μία κατηγορία και μια όχι και τόσο καλή στα μαθήματα της άλλης κατηγορίας. Για παράδειγμα, ένας φοιτητής μπορεί να είναι άριστος στα θεωρητικά μαθήματα αλλά στα τεχνολογικά να μην μπορεί να ανταποκριθεί αναλόγως. Αυτό συμβαίνει και αντίθετα καθώς παρατηρείται ότι οι "τεχνολόγοι" δεν τα πάνε και τόσο καλά στα θεωρητικά μαθήματα. Υπάρχουν βεβαίως και εξαιρέσεις αλλά στην παρούσα πτυχιακή εργασία θα γίνει μια προσπάθεια εξομάλυνσης του πιο πάνω προβλήματος.

Το πιο πάνω πρόβλημα δεν απασχολεί μόνο το Τμήμα Επικοινωνίας και Σπουδών Διαδικτύου αλλά η καταλληλότητα των μαθημάτων με τα οποία οι μαθητές εισάγονται σε διάφορα τμήματα, απασχολεί όλα τα κρατικά Πανεπιστήμια, τους ακαδημαϊκούς καθώς και το Υπουργείο Παιδείας. Οι ακαδημαϊκοί του τμήματος μέσω της εφαρμογής θα μπορούν (με επιφύλαξη πάντα) να υπολογίζουν τις πιθανότητες που θα έχει ένας φοιτητής να αποφοιτήσει επιτυχώς ή ανεπιτυχώς. Έτσι, σε φοιτητές στους οποίους η βαθμολογία τους προβλεφθεί κάτω από την βάση, οι ακαδημαϊκοί θα μπορούν να τους δίνουν περισσότερη προσοχή και να τους έχουν υπό την επίβλεψη τους.

Έτσι, αν τελικά τα αποτελέσματα της εφαρμογής που θα δημιουργηθεί έχουν υψηλό δείκτη ορθότητας, σίγουρα θα είναι ένα καλό εργαλείο και για χρήση πέραν του τμήματος ΕΣΔ.

Στις επόμενες ενότητες ακολουθεί η βιβλιογραφική επισκόπηση, η μεθοδολογία με την δημιουργήθηκε η εφαρμογή, η ανάλυση των αποτελεσμάτων καθώς και τα γενικά συμπεράσματα σχετικά με τις βαθμολογίες και τις προβλέψεις. Η βιβλιογραφική επισκόπηση επικεντρώνεται στην Τεχνητή Νοημοσύνη, στην εξόρυξη δεδομένων και στην μηχανική μάθηση, όπως επίσης θα παρουσιαστούν παρόμοιες μελέτες και έρευνες που σχετίζονται με την εξόρυξη δεδομένων μέσω της μηχανικής μάθησης. Στην ενότητα της μεθοδολογίας περιγράφεται ο τρόπος με τον οποίο εργάστηκα και γίνεται μία σύντομη περιγραφή του συστήματος και των λειτουργιών του. Στην ανάλυση και στα αποτελέσματα, περιγράφεται ο τρόπος με τον οποίο αναλύθηκαν τα αποτελέσματα ούτως ώστε να είναι αξιόπιστα και ακριβή.

Κεφάλαιο 3: Βιβλιογραφική Επισκόπηση

3.1 Τεχνητή Νοημοσύνη

Για χιλιάδες χρόνια, ο άνθρωπος προσπαθεί να καταλάβει πώς σκέφτεται, πώς δηλαδή μπορεί να αντιλαμβάνεται, να κατανοεί, να προβλέπει και να χειρίζεται έναν κόσμο πολύ μεγαλύτερο και πιο περίπλοκο από ό, τι είναι ο ίδιος. Το πεδίο της τεχνητής νοημοσύνης, πηγαίνει ένα βήμα παραπέρα και προσπαθεί όχι μόνο να μας βοηθήσει να κατανοήσουμε, αλλά και στο να κατασκευάσουμε νοήμονα συστήματα.

Η Τεχνητή Νοημοσύνη είναι ένα από τα νεότερα πεδία της επιστήμης και της τεχνολογίας και περιλαμβάνει μια τεράστια ποικιλία από πεδία, όπως για παράδειγμα η μηχανική μάθηση ή η διάγνωση ασθενειών.

Για να θεωρηθεί ένας υπολογιστής ευφυής με βάση το πείραμα του Turing [1]θα πρέπει να μπορεί να επεξεργαστεί τη φυσική γλώσσα (να μπορεί να επικοινωνήσει με έναν άνθρωπο), να μπορεί να αποθηκεύει τις πληροφορίες που ακούει ή που μαθαίνει, να χρησιμοποιεί αυτές τις αποθηκευμένες πληροφορίες για να απαντήσει σε ερωτήματα και να μπορεί να προσαρμόζεται σε νέες καταστάσεις.

Η Τεχνητή νοημοσύνη ασχολείται με διάφορους τομείς και η συνδρομή της σε αυτούς έπαιξε σημαντικό ρόλο στην ανάπτυξη τους ή ακόμη και στη δημιουργία τους. Οι κύριες περιοχές με τις οποίες ασχολείται η τεχνητή νοημοσύνη είναι η επίλυση προβλημάτων και η απόδειξη θεωρημάτων και κάποιες από τις επιμέρους περιοχές με τις οποίες ασχολείται η τεχνητή νοημοσύνη είναι η εξόρυξη δεδομένων από τον παγκόσμιο ιστό, η επεξεργασία φυσικής γλώσσας, η τεχνητή όραση, η ρομποτική και η μηχανική μάθηση. (Russell and Norvig, 1995)

[1] Το Turing Test, που πρότεινε ο Alan Turing (1950), είχε σχεδιαστεί για να παρέχει ένα ικανοποιητικό λειτουργικό ορισμό της νοημοσύνης του υπολογιστή. Ένας υπολογιστής περνάει έλεγχο μαζί με έναν άνθρωπος. Ο ανακριτής ο οποίος θέτει τις ερωτήσεις δεν γνωρίζει εξ αρχής ποιος είναι ο άνθρωπος και ποιος ο υπολογιστής. Νοήμων υπολογιστής μπορεί να θεωρηθεί αυτός που θα αναγκάσει τον ανακριτή να μην μπορεί να πει αν οι απαντήσεις προέρχονται από άνθρωπο ή από έναν υπολογιστή.

3.2 Εξόρυξη δεδομένων

Στην παρούσα εργασία, ο τομέας που μας ενδιαφέρει είναι η εξόρυξη δεδομένων μέσω της μηχανικής μάθησης.

Η ποσότητα των δεδομένων στον κόσμο και στη ζωή μας αυξάνεται ολοένα και περισσότερο και δεν υπάρχει τέλος σε αυτή την αύξηση. Είμαστε κατακλυσμένοι από δεδομένα και πληροφορίες. Ο Παγκόσμιος ιστός (www) είναι κατακλυσμένος από όλες αυτές τις πληροφορίες και η κάθε μας κίνηση στο διαδίκτυο καταγράφεται. Καθώς υπάρχει αυτή η ραγδαία αύξηση δεδομένων, μειώνεται ανάλογα και το ποσοστό κατανόησης τους από τους ανθρώπους.

Η λύση σε αυτό το πρόβλημα είναι η εξόρυξη δεδομένων. Σε αυτή την περίπτωση, τα δεδομένα αποθηκεύονται ηλεκτρονικά με μια αυτοματοποιημένη μέθοδο από τον υπολογιστή. Η κεντρική ιδέα είναι να δημιουργηθούν πρότυπα από τα δεδομένα τα οποία θα μπορούν αυτόματα να αναζητηθούν, να εντοπιστούν και να επικυρωθούν και τέλος να χρησιμοποιηθούν για προβλέψεις. Η εξόρυξη δεδομένων έχει ως στόχο να λύσει προβλήματα αναλύοντας δεδομένα τα οποία είναι ήδη αποθηκευμένα σε μια βάση δεδομένων. Ας πάρουμε για παράδειγμα το «καλάθι της νοικοκυράς». Η λύση βρίσκεται στις βάσεις δεδομένων στις οποίες είναι αποθηκευμένα τα δεδομένα των χρηστών, δηλαδή τις αγοραστικές προτιμήσεις του κάθε καταναλωτή. Όταν αναλυθούν αυτά τα δεδομένα, μπορούμε να δημιουργήσουμε πρότυπα με τις προτιμήσεις των καταναλωτών οι οποίοι χωρίζονται σε δύο κατηγορίες: σε αυτούς που αλλάζουν συχνά επωνυμίες προϊόντων και σε αυτούς που παραμένουν πιστοί σε ένα προϊόν.

Η εξόρυξη δεδομένων ορίζεται ως η διαδικασία ανακάλυψης μοτίβων από ήδη υπάρχοντα δεδομένα και η διαδικασία αυτή θα πρέπει να είναι αυτόματη ή ημιαυτόματη. Τα μοτίβα που θα ανακαλυφθούν θα πρέπει να έχουν κάποιο νόημα, δεδομένου ότι οδηγούν σε κάποιο όφελος και συνήθως αυτό μεταφράζεται σε οικονομικό κέρδος.

3.2.1 Μηχανική Μάθηση

Οι κύριες ερμηνείες της μάθησης που δίνονται από τα λεξικά είναι να αποκτήσουμε γνώση μέσω της μελέτης, της εμπειρίας και της διδασκαλίας, να αντλούμε και να αφομοιώνουμε πληροφορίες μέσω της παρατήρησης , να μπορούμε να τα "αποθηκεύουμε" και να τα ανακαλούμε από την μνήμη μας, να ενημερωνόμαστε και να έχουμε την δυνατότητα να λάβουμε οδηγίες και να τις εκτελέσουμε.

Οι ερμηνείες αυτές όταν πρόκειται για ηλεκτρονικούς υπολογιστές έχουν ορισμένες αδυναμίες. Ειδικότερα όσον αφορά την άντληση γνώσης μέσω μελέτης, εμπειρίας ή διδασκαλίας και αφομοίωση πληροφορίας μέσω της παρατήρησης, είναι πολύ δύσκολο έως ακατόρθωτο στο να ελεγχθεί εάν έχει επιτευχθεί μάθηση ή όχι. Όσο για τα υπόλοιπα, αν και μπορούμε να καταλάβουμε πως μπορεί ένας άνθρωπος να τα κατανοήσει, σε έναν υπολογιστή αυτές οι λειτουργίες απέχουν πολύ από την ανθρώπινη σημασιολογία.

Στην εξόρυξη δεδομένων, η όλη λειτουργία βασίζεται στη διαδικασία ανακάλυψης προτύπων, αυτόματα ή ημιαυτόματα, σε μεγάλες ποσότητες αποθηκευμένων (εκ των προτέρων) δεδομένων και τα πρότυπα αυτά θα πρέπει να είναι χρήσιμα. Άρα αυτό που μας ενδιαφέρει είναι η απόδοση της "μηχανής" όσον αφορά την μάθηση και όχι αν τελικά ο υπολογιστής απόκτησε γνώση για κάτι. Έτσι, η εξόρυξη δεδομένων είναι ένα θέμα που περιλαμβάνει την εκμάθηση σε ένα πρακτικό επίπεδο και όχι σε θεωρητικό.

Αυτό που μας ενδιαφέρει γενικά είναι η εξεύρεση τεχνικών για εύρεση και περιγραφή των προτύπων που θα αντληθούν από τα δεδομένα και γι' αυτό χρειαζόμαστε ένα εργαλείο το οποίο θα μπορεί να μας βοηθήσει στο να εξηγήσουμε τα δεδομένα αυτά και να μας κάνει προβλέψεις για μετέπειτα συμπεριφορές σε παρόμοιες καταστάσεις. Για παράδειγμα θα μπορούμε μέσω αυτών των τεχνικών, να μπορούμε να μετατρέψουμε αυτά τα πρότυπα σε λίστες από παραδείγματα πελατών οι οποίοι δεν παραμένουν πιστοί σε κάποια προϊόντα. Η τελική μορφή αυτών των λιστών θα είναι μια μορφή εξόδου η οποία θα εμπεριέχει τις προβλέψεις σχετικά με νέα παραδείγματα, δηλαδή θα μπορεί να προβλέψει αν ένας συγκεκριμένος πελάτης θα αλλάξει ή θα παραμείνει πιστός σε ένα προϊόν. Επίσης αυτό το εργαλείο θα μπορεί να χρησιμοποιηθεί για να ταξινομηθούν άγνωστα παραδείγματα.

Παράδειγμα: Το πρόβλημα με τις καιρικές συνθήκες:

Στον πιο κάτω πίνακα απεικονίζεται μια λίστα με καταγεγραμμένα στοιχεία σχετικά με τον καιρό (καιρικές συνθήκες, θερμοκρασία, υγρασία, αέρας) όπως επίσης και μια στήλη στην οποία καταγράφεται η λέξη "παιχνίδι". Ο πίνακας αυτός αφορά τον καιρό και το αν ένα παιδί μπορεί να βγει έξω από το σπίτι και να παίξει.

Καιρός	Θερμοκρασία	Υγρασία	Αέρας	Παιχνίδι
Ηλιόλουστος	Ζέστη	Υψηλή	Όχι	Όχι
Ηλιόλουστος	Ζέστη	Υψηλή	Ναι	Όχι
Συννεφιασμένος	Ζέστη	Υψηλή	Όχι	Ναι
Βροχερός	Ήπιος	Υψηλή	Όχι	Ναι
Βροχερός	Ψυχρός	Κανονική	Όχι	Ναι
Βροχερός	Ψυχρός	Κανονική	Ναι	Όχι
Συννεφιασμένος	Ψυχρός	Κανονική	Ναι	Ναι
Ηλιόλουστος	Ήπιος	Υψηλή	Όχι	Όχι
Ηλιόλουστος	Ψυχρός	Κανονική	Όχι	Ναι
Βροχερός	Ήπιος	Κανονική	Όχι	Ναι
Ηλιόλουστος	Ήπιος	Κανονική	Ναι	Ναι
Συννεφιασμένος	Ήπιος	Υψηλή	Ναι	Ναι
Συννεφιασμένος	Ζέστη	Κανονική	Όχι	Ναι
Βροχερός	Ήπιος	Υψηλή	Ναι	Όχι

Πίνακας 1: Χαρακτηριστικά δεδομένων εισόδου Πηγή: Witten, Frank & Hall, 2011 σελίδα 10

Όπως ανέφερα και πριν υπάρχουν τέσσερα χαρακτηριστικά του καιρού: οι καιρικές συνθήκες, η θερμοκρασία, ο άνεμος και η υγρασία και το αποτέλεσμα θα είναι αν θα παίξει ή όχι.

Από τον πιο πάνω πίνακα δημιουργήθηκαν κάποια μοτίβα τα οποία λειτουργούν σαν κανόνες . Στα παραδείγματα που ακολουθούν διακρίνονται αυτοί οι κανόνες:

Αν ο καιρός είναι ηλιόλουστος και υπάρχει υψηλή υγρασία τότε δεν θα παίξει.

Αν ο καιρός είναι βροχερός και υπάρχει αέρας τότε δεν θα παίξει.

Αν ο καιρός είναι συννεφιασμένος τότε θα παίξει.

Αν η υγρασία είναι κανονική τότε θα παίξει.

Με την δημιουργία αυτών των κανόνων αυτών, το σύστημα όταν εκπαιδευτεί θα μπορεί να υπολογίζει αυτόματα αν το παιδί θα παίξει ή όχι.

Η πιο πάνω διαδικασία μπορεί να γίνει ακόμα και όταν τα χαρακτηριστικά είναι αριθμοί. Για παράδειγμα αν ο καιρός είναι ηλιόλουστος και υπάρχει υγρασία μικρότερη του 83% τότε το παιδί θα μπορεί να παίξει. Στη δική μας περίπτωση που αφορά το σύστημα το οποίο θα προβλέπει την βαθμολογία των φοιτητών, θα χρησιμοποιηθούν μόνο αριθμητικές τιμές στα χαρακτηριστικά.

3.2.2 Επιπλέον Εφαρμογές Εξόρυξης Δεδομένων

Εξόρυξη δεδομένων από τον Παγκόσμιο Ιστό: Οι μηχανές αναζήτησης εξετάζουν τους υπερσυνδέσμους μέσα στις ιστοσελίδες για να υπολογίσουν την βαθμολογία της ιστοσελίδας, το λεγόμενο PageRank. Όσο περισσότερες ιστοσελίδες συνδέονται με την ιστοσελίδα που εξετάζουμε, τόσο υψηλότερη θα είναι η βαθμολογία της και έτσι θα εμφανίζεται πιο ψηλά στην αναζήτηση σε σύγκριση με παρόμοιες σελίδες.

Ένας άλλος τρόπος με τον οποίο οι μηχανές αναζήτησης αντιμετωπίζουν το πρόβλημα στο πώς να ταξινομήσουν ιστοσελίδες είναι η χρήση μηχανικής μάθησης που βασίζεται σε ένα σύνολο δεδομένων (ερωτήματα - έγγραφα) προς εκπαίδευση που περιέχουν τους όρους στο ερώτημα του χρήστη. Στη συνέχεια, ένας αλγόριθμος μάθησης αναλύει αυτά τα δεδομένα και προβλέπει τα συνδέει τα ερωτήματα με τα έγγραφα με βάση την συνάφεια τους.

Λήψη κρίσιμων αποφάσεων:

Όταν ένα άτομο κάνει αίτηση για δάνειο, του δίνεται να συμπληρώσει ένα ερωτηματολόγιο το οποίο του ζητά σχετικές πληροφορίες που αφορούν τα οικονομικά του και τα προσωπικά του στοιχεία. Αυτές οι πληροφορίες χρησιμοποιούνται από την τράπεζα ως βάση για την λήψη απόφασης ως προς το αν πρέπει να δανείσει τα χρήματα στο συγκεκριμένο άτομο. Οι αποφάσεις αυτές λαμβάνονται σε δύο στάδια: πρώτα μέσω στατιστικών μεθόδων και στη συνέχεια αν τα οικονομικά του είναι σε οριακή κατάσταση σε σχέση με το αν θα αποπληρώσει το δάνειο του, λαμβάνεται η απόφαση βάση της ανθρώπινης κρίσης του ατόμου που θα εγκρίνει ή θα απορρίψει το δάνειο. Εξετάζοντας τα ιστορικά δεδομένα της τράπεζας, οι μισοί από τους αιτούντες που τους χορηγήθηκε το δάνειο με βάσει την ανθρώπινη κρίση λόγω της οριακής τους κατάστασης, δεν κατέφεραν να αποπληρώσουν το δάνειο τους.

Και κάπου εδώ έρχεται η μηχανική μάθηση. Με ένα δείγμα 1000 οριακών περιπτώσεων (δεδομένα εκπαίδευσης), στις οποίες οι οφειλέτες είτε αποπλήρωσαν είτε όχι το δάνειο τους, και με 20 χαρακτηριστικά τα οποία δημιουργήθηκαν από ερωτήσεις του ερωτηματολογίου (ηλικία, επάγγελμα, κ.α), παρήχθη ένα μικρό σύνολο κανόνων κατάταξης το οποίο έκανε σωστές προβλέψεις για τα δύο τρίτα των οριακών περιπτώσεων. Οι κανόνες αυτοί δεν βοήθησαν μόνο στο να ληφθούν οι ορθές αποφάσεις για την παραχώρηση δανείου αλλά τους δόθηκε και η δυνατότητα να μπορούν να εξηγήσουν και στους αιτητές τους λόγους πίσω από την απόφαση τους.

Μάρκετινγκ και Πωλήσεις:

Μερικές από τις πιο ενεργές εφαρμογές της εξόρυξης δεδομένων λαμβάνουν μέρος στον τομέα του μάρκετινγκ και των πωλήσεων. Οι εταιρίες που ασχολούνται με αυτόν τον τομέα έχουν καταγεγραμμένα τεράστια σε όγκο δεδομένα και η χρήση συστημάτων μηχανικής μάθησης είναι εξαιρετικά πολύτιμη.

Το κύριο πρόβλημα των εταιριών είναι οι "άστατοι" καταναλωτές, δηλαδή αυτοί που δεν μένουν πιστοί σε μία επωνυμία προϊόντος και η πρόκληση είναι να δημιουργηθεί ένα σύστημα το οποίο θα μπορεί να ανιχνεύει αυτούς τους πελάτες και μέσω ειδικής μεταχείρισης να τους πλασάρουν τα προϊόντα τους ξανά και ξανά.

Επίσης η εξόρυξη δεδομένων μπορεί να καθορίσει συγκεκριμένες ομάδες καταναλωτών, οι οποίοι είναι "ευάλωτοι" σε νέα προϊόντα ή καταναλωτές οι οποίοι είναι διαχρονικά πιστοί σε συγκεκριμένες μάρκες προϊόντων και θα έχουν και αυτοί συγκεκριμένη μεταχείριση κυρίως στις αρχές του κάθε μήνα στις οποίες θα έχουν χρήματα.

Τέλος, ένα σημαντικό κομμάτι σε αυτό τον τομέα είναι η ανάλυση του "καλαθιού της νοικοκυράς". Σε αυτή την περίπτωση χρησιμοποιούνται τεχνικές σύνδεσης προϊόντων που εμφανίζονται συχνά μαζί κατά την πληρωμή στο ταμείο μιας υπεραγοράς. Για παράδειγμα, μέσω αυτής της αυτοματοποιημένης ανάλυσης των δεδομένων μιας υπεραγοράς μπορεί να αποκαλυφθεί ότι οι πελάτες που αγοράζουν μπύρα αγοράζουν επίσης και πατατάκια. Αυτή η ανακάλυψη θα μπορούσε να είναι σημαντική για τους διαχειριστές της υπεραγοράς και φέρνοντας πιο κοντά τις μπύρες και τα πατατάκια ίσως οι πωλήσεις τους αυξηθούν. Ένα άλλο παράδειγμα είναι να παρατηρηθεί το φαινόμενο ότι τις Πέμπτες οι καταναλωτές αγοράζουν περισσότερο παιδικές πάνες και μπύρες μαζί. Έτσι με αυτόν τον τρόπο μπορούν οι υπεραγορές να βάζουν προσφορές την συγκεκριμένη μέρα.

3.3 Εξόρυξη δεδομένων και δεοντολογία

Η χρήση δεδομένων και ιδιαίτερα των προσωπικών δεδομένων των ανθρώπων, για την εξόρυξη δεδομένων έχει σοβαρές ηθικές επιπτώσεις και γι' αυτό τον λόγο οι επαγγελματίες των τεχνικών εξόρυξης δεδομένων πρέπει να ενεργούν υπεύθυνα και να γνωρίζουν τα ηθικά ζητήματα που περιβάλλουν το όλο ζήτημα. Η εξόρυξη δεδομένων χρησιμοποιείται σε ανθρώπους για να γίνουν συγκεκριμένες διακρίσεις, όπως για παράδειγμα ποιοι θα πάρουν δάνειο ή σε ποιούς θα προωθηθεί συγκεκριμένη προσφορά. Οι διακρίσεις αυτές μπορεί να είναι με βάση το φύλο, την φυλή, το θρήσκευμα αλλά εκτός από τα ηθικά ζητήματα υπάρχουν και νομικά ζητήματα για το αν μπορεί κάποιος να αποθηκεύει τα προσωπικά δεδομένα του κάθε ατόμου. Αλλά αυτά εξαρτώνται πάντα από την εφαρμογή για τον λόγο ότι το θέμα είναι πολύ περίπλοκο. Για παράδειγμα, η χρήση πληροφοριών για το φύλο και την φυλή των ατόμων για ιατρική διάγνωση, θεωρείται ηθική αλλά η χρήση των δεδομένων για την συμπεριφορά του καταναλωτή θεωρείται ως μη ηθική. Όμως ακόμα και τον αποκλεισμό κάποιων ευαίσθητων πληροφοριών, υπάρχει ο κίνδυνος τα μοντέλα που θα δημιουργηθούν να βασίζονται σε κάποιες μεταβλητές οι οποίες ακούσια αντικατέστησαν κάποια ευαίσθητη πληροφορία. Για παράδειγμα, οι άνθρωποι συνηθίζουν να κατοικούν σε περιοχές που σχετίζονται με την εθνικότητά τους και χρησιμοποιώντας τον ταχυδρομικό τους κώδικα για την εξόρυξη των δεδομένων, υπάρχει ο κίνδυνος η κατασκευή των μοντέλων να βασίζεται τελικά στην φυλή, η οποία μπορεί να είναι ένα ευαίσθητο προσωπικό δεδομένο.

3.3.1 Χρήση προσωπικών δεδομένων

Είναι ευρέως αποδεκτό ότι ένας άνθρωπος πριν πάρει την απόφαση να παρέχει τις προσωπικές του πληροφορίες θέλει να ξέρει πώς θα χρησιμοποιηθούν και ποια μέτρα θα ληφθούν για την προστασία των δεδομένων τους. Επίσης, είναι απαραίτητο να καθορίζεται και να δηλώνεται από την αρχή ο σκοπός για τον οποίο θα χρησιμοποιηθούν τα εν λόγω δεδομένα και σε καμία περίπτωση δεν πρέπει να αντικατασταθεί ο σκοπός αυτός χωρίς την έγκριση των ατόμων που έδωσαν τις προσωπικές τους πληροφορίες. (Witten, Frank & Hall, 2011)

3.4 Μάθηση μέσω Νευρωνικών δικτύων

Ένα νευρωνικό δίκτυο αποτελείται από έναν αριθμό κόμβων, ή μονάδες, που συνδέονται με δεσμούς. Επιπλέον κάθε σύνδεση έχει ένα αριθμητικό βάρος. Τα βάρη είναι το μέσο της αποθήκευσης στα νευρωνικά δίκτυα, και η μάθηση λαμβάνει χώρα συνήθως κατά την ενημέρωση των βαρών. Οι μονάδες που συνδέονται με το εξωτερικό περιβάλλον, μπορούν να οριστούν ως μονάδες εισόδου ή εξόδου.

Κάθε μονάδα έχει μια σειρά από συνδέσεις εισόδου από άλλες μονάδες, μια σειρά από συνδέσεις εξόδου σε άλλες μονάδες, μια συνάρτηση ενεργοποίησης, και έναν αθροιστή. Η ιδέα είναι ότι κάθε μονάδα κάνει ένα τοπικό υπολογισμό που βασίζεται στις εισροές από τους γείτονές της. Στην πράξη, οι περισσότερες εφαρμογές νευρωνικών δικτύων είναι στο λογισμικό.

Για να δημιουργήσουμε ένα νευρωνικό δίκτυο ούτως ώστε να εκτελέσει κάποια εργασία, θα πρέπει πρώτα να αποφασίσουμε πόσες μονάδες θα χρησιμοποιηθούν, τι είδους μονάδες είναι κατάλληλες, και πώς οι μονάδες θα συνδέονται για να σχηματίσουν ένα δίκτυο. Στη συνέχεια προετοιμάζουμε τα βάρη του δικτύου, και εκπαιδεύουμε τα βάρη χρησιμοποιώντας ένα μαθησιακό αλγόριθμο εφαρμόζοντας το σε ένα σύνολο παραδειγμάτων εκπαίδευσης. Κάθε μονάδα εκτελεί έναν απλό υπολογισμό: λαμβάνει σήματα από τις συνδέσεις εισόδου και υπολογίζει μια νέα συνάρτηση ενεργοποίησης που στέλνει κατά το μήκος κάθε σύνδεσης. Ο υπολογισμός της συνάρτησης ενεργοποίησης βασίζεται στην τιμή του κάθε σήματος εισόδου που έλαβε από ένα γειτονικό κόμβο, και τα βάρη σε κάθε σύνδεσμο εισόδου.

3.4.1 Εφαρμογές νευρωνικών δικτύων

Προφορά / εκφώνηση γραπτού λόγου: Η εκφώνηση του γραπτού κειμένου από έναν υπολογιστή είναι μια τεράστια πρόκληση. Κατ' αρχάς γίνεται μια αντιστοιχία του κειμένου με τα βασικά ηχητικά φωνήματα και στη συνέχεια διέρχεται σε έναν ηλεκτρονικό εξομοιωτή φωνής. Το πρόβλημα που μας απασχολεί εδώ είναι η εκμάθηση της αντιστοιχίας του κειμένου με τα φωνήματα. Αυτή είναι μια πολύ καλή αποστολή για τα νευρωνικά δίκτυα, επειδή οι περισσότεροι από τους "κανόνες" είναι περίπου ορθοί. Για παράδειγμα, αν το γράμμα "k" συνήθως αντιστοιχεί στον ήχο [k], όμως το γράμμα "C" είναι αντιστοιχεί σε [k] όταν η λέξη είναι cat και [s] όταν η λέξη είναι cent. Σε αυτή την περίπτωση η είσοδος είναι μια ακολουθία χαρακτήρων και περιλαμβάνει το χαρακτήρα που πρέπει να προφέρεται μαζί με τρεις χαρακτήρες πριν και μετά. Κάθε χαρακτήρας έχει ότητα 29 μονάδες, μία είσοδο για το καθένα από τα 26 γράμματα, και από ένα για τα κενά, τελείες και τα άλλα σημεία στίξης. Η μονάδα εξόδου αποτελείται από τα χαρακτηριστικά που παράγει ο ήχος: αν είναι υψηλή ή χαμηλή,

φωνήεν ή σύμφωνο, και ούτω καθεξής. Μετά από εκπαίδευση ενός μεγάλου αριθμού παραδειγμάτων, το σύστημα μπορεί να αναγνωρίσει και να προφέρει σωστά το 95% των λέξεων που του δόθηκαν για εκπαίδευση.

Αναγνώριση χειρόγραφης γραφής: Μία από τις μεγαλύτερες εφαρμογές των νευρωνικών δικτύων μέχρι σήμερα, είναι η δημιουργία μιας εφαρμογής, η οποία μπορεί να διαβάζει τον ταχυδρομικό κώδικα από τα γράμματα τα οποία είναι χειρόγραφα. Το σύστημα χρησιμοποιεί ένα προ - επεξεργαστή που εντοπίζει και χωρίζει σε τμήματα τα μεμονωμένα ψηφία στο ταχυδρομικό κώδικα και η εφαρμογή καλείτε να προσδιορίσει τα ίδια τα ψηφία.

Οδήγηση: Το ALVIN (Autonomous Land Vehicle In a Neural Network) είναι ένα νευρωνικό δίκτυο το οποίο έχει αρκετά καλές επιδόσεις σε έναν τομέα όπου κάποιες άλλες προσεγγίσεις έχουν αποτύχει. Μαθαίνει να διευθύνει ένα όχημα κατά μήκος μιας μόνο λωρίδας σε αυτοκινητόδρομο παρατηρώντας την απόδοση ενός ανθρώπινου οδηγού. Δουλειά του ALVINN είναι να υπολογίζει μια συνάρτηση που χαρτογραφείται από μια ενιαία εικόνα βίντεο του δρόμου μπροστά του σε μια κατεύθυνση του τιμονιού. Για να μάθει αυτή τη λειτουργία, χρειάζεται κάποια εκπαίδευση μέσω των δεδομένων τα οποία είναι ζεύγη εικόνας / κατεύθυνσης με τη σωστή κατεύθυνση του δρόμου. Η συλλέγουν αυτών των δεδομένων είναι εύκολη υπόθεση καθώς χρειάζεται να οδηγήσει το όχημα ένας άνθρωπος και να καταγραφούν τα ζεύγη εικόνας / κατεύθυνσης. Μετά τη συλλογή των δεδομένων εκπαίδευσης το ALVINN είναι έτοιμο να οδηγεί από μόνο του. (Russell and Norvig, 1995)

3.5 Παρόμοιες Μελέτες / Έρευνες

Παρόμοιες μελέτες για συστήματα όμοια με το δικό μου (δηλαδή να προβλέπουν βαθμολογίες φοιτητών) δεν υπάρχουν ακριβώς οι ίδιες αλλά υπάρχουν πάμπολλες μελέτες οι οποίες κάνουν χρήση των διάφορων τεχνικών της τεχνητής νοημοσύνης για την πρόβλεψη ή εκτίμηση της τιμής μιας μεταβλητής και παρακάτω παρατίθενται τρεις από αυτές.

Η πρώτη μελέτη που βρήκα είναι από το Αριστοτέλειο Πανεπιστήμιο της Θεσσαλονίκης και είναι σχετικά με προβλέψεις αποτελεσμάτων ποδοσφαιρικών αγώνων μέσω μοντέλων τεχνητής νοημοσύνης. Στη συγκεκριμένη μελέτη, ο ερευνητής, δημιούργησε ένα σύστημα το οποίο αναγνωρίζει πρότυπα τα οποία βασίζονται στην συσχέτιση προηγούμενων αποτελεσμάτων ποδοσφαιρικών αγώνων με πρόσφατα αποτελέσματα, δηλαδή μάθηση χωρίς επίβλεψη. Στόχος ήταν να παραχθεί κέρδος όταν εφαρμοστεί η εν λόγω εφαρμογή στα στοιχήματα ποδοσφαιρικών αγώνων. Τα δεδομένα εισόδου που έδωσε ήταν τα συνολικά γκολ των ομάδων (εντός και εκτός), οι συνολικοί βαθμοί της ομάδας, αριθμό επιθετικών προσπαθειών προς τον στόχο καθώς και αποτελέσματα προηγούμενων αγώνων. Στη συνέχεια συνδύασε διαφορετικά μοντέλα για υπολογισμό της πρόβλεψης του αγώνα κάνοντας την απόδοση του συστήματος πιο εύστοχη κάτι που δεν θα μπορούσε να γίνει όπως αναφέρει με ένα αυτόνομο μοντέλο. Τα μοντέλα αυτά ήταν τα ασαφή μοντέλα και τα μοντέλα SVR (Supporter Vector Machine). Το τελικό αποτέλεσμα του συστήματος που δημιούργησε ήταν επιτυχές και όπως ανέφερε λόγω των αποτελεσμάτων που πήρε κατά την ανάλυση των δεδομένων του, η πρόβλεψη των τελικών αποτελεσμάτων στους ποδοσφαιρικούς αγώνες μέσω τεχνικών της τεχνητής νοημοσύνης είναι εφικτή. (Ιωάννου, 2013)

Η δεύτερη μελέτη από τα ΤΕΙ Κρήτης ασχολείται με το χρηματιστήριο και πιο συγκεκριμένα με την πρόβλεψη της πορείας των μετοχών στα χρηματιστήρια μέσω της χρήσης των νευρωνικών δικτύων τα οποία ανήκουν στη μάθηση με επίβλεψη. Σκοπός της συγκεκριμένης μελέτης ήταν να αναλύσει τους αλγόριθμους των νευρωνικών δικτύων, δηλαδή τι είναι και πως μπορούν να φανούν χρήσιμοι, καθώς και να κάνει προβλέψεις σχετικά με τις κινήσεις των μετοχών. Η έρευνα αυτή περίγραψε με πολύ εμπεριστατωμένο τρόπο το τι είναι τα νευρωνικά δίκτυα και πως λειτουργούν αλλά όπως και η ίδια η ερευνήτρια αναφέρει, λόγω αντικειμενικών δυσχερειών δεν κατέστη εφικτή η δημιουργία αυτού του συστήματος. Εξήγησε όμως πολύ καλά από τι αποτελούνται τα νευρωνικά δίκτυα, ανέλυσε την διάταξη των νευρώνων, την λειτουργία των βαρών καθώς και τις συναρτήσεις που υπολογίζουν την έξοδο σε κάθε νευρώνα. Επίσης, η

ερευνήτρια παρέθεσε 11 μελέτες οι οποίες ήταν παρόμοιες με την δική της και η πλειοψηφία τους απέδειξε ότι η χρήση των νευρωνικών δικτύων μπορούν να βοηθήσουν στην πρόβλεψη της πορείας των τιμών που θα πάρουν οι μετοχές. (Τιτομιχελάκη, 2010)

Η τελευταία μελέτη η οποία είναι και αυτή πρόβλεψη της τιμής της μετοχής με τη χρήση των νευρωνικών δικτύων είναι από το Εθνικό Μετσόβιο Πολυτεχνείο. Σε αυτή τη διπλωματική εργασία, παρουσιάζεται αρχικά ο τρόπος με τον οποίο ο εγκέφαλος επεξεργάζεται τις πληροφορίες και ο λόγος που επέλεξε τα νευρωνικά δίκτυα είναι η ανίχνευση πολυδιάστατων - μη γραμμικών συναρτήσεων, που να είναι πολύ χρήσιμο στην ανάλυση του χρηματιστηρίου το οποίο είναι ένα δυναμικό σύστημα. Για την υλοποίηση του συστήματος αναλύθηκαν διάφορες στρατηγικές, εφαρμόστηκαν διάφοροι αλγόριθμοι σε μια σειρά από δεδομένα προηγούμενων χρόνων και αξιολογήθηκε με βάση τα αποτελέσματα και τη δυνατότητα εφαρμογής του συστήματος σε πραγματικό χρόνο. Τα δεδομένα εισόδου του συστήματος ήταν ένα σύνολο από τις τρέχουσες τιμές των μετοχών, τους δείκτες του χρηματιστηρίου, καθώς και μακροσκοπικά και λογιστικά στοιχεία αποτίμησης. Η εφαρμογή δημιουργήθηκε μέσω της MatLab και χρησιμοποιήθηκε ο αλγόριθμος Back Propagation. Το 75% των δεδομένων που είχε το χρησιμοποίησε για training ενώ το υπόλοιπο 30% το μοίρασε και το χρησιμοποίησε για validation και test. Από την έρευνα προέκυψε ικανοποιητική πρόβλεψη των τιμών των μετοχών η οποία είχε ακρίβεια σε ποσοστό 90%. (Σοφός, 2013)

Κεφάλαιο 4: Μεθοδολογία

Η όλη έρευνα είναι βασισμένη σε πειραματική ανάπτυξη συστήματος με την βοήθεια τεχνικών της τεχνητής νοημοσύνης. Για να αναπτυχθεί ένα σύστημα σαν αυτό, θα χρειαστούμε ένα πρόγραμμα το οποίο αναγνωρίζει τους αλγόριθμους της τεχνητής νοημοσύνης και θα μπορούμε να εισάγουμε σε αυτό τις βαθμολογίες των μαθητών / φοιτητών ούτως ώστε να γίνει η επιθυμητή πρόβλεψη. Το καταλληλότερο πρόγραμμα το οποίο πληροί αυτές τις προϋποθέσεις είναι το WEKA.

Οι μεταβλητές που χρησιμοποιήθηκαν για δεδομένα εισόδου είναι οι βαθμολογίες των προεισαγωγικών εξετάσεων των μαθητών που έδωσαν εξετάσεις και πέρασαν στο ΤΕΠΑΚ, τα μαθήματα που επέλεξαν, η σχολή στην οποία φοίτησαν (Λύκειο, Τεχνική Σχολή) και η επίδοση στο πτυχίο των απόφοιτων του ΤΕΠΑΚ. Έτσι, το δείγμα της έρευνας είναι 117 φοιτητές που αποφοίτησαν από το τμήμα Επικοινωνίας και Σπουδών Διαδικτύου τα τελευταία χρόνια.

Όπως ανάφερα και πριν, το εργαλείο που θα χρησιμοποιηθεί για την ανάπτυξη του συστήματος / εφαρμογής θα είναι το WEKA. Το WEKA είναι μια συλλογή από αλγόριθμους μηχανικής μάθησης και περιέχει υλοποιημένες μεθόδους προ -επεξεργασίας δεδομένων, δημιουργίας μοντέλων, αξιολόγησης αλγορίθμων και προβολής αποτελεσμάτων.

WEKA

Το Weka αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας και το ολοκληρωμένο του όνομα είναι Waikato Environment for Knowledge Analysis. Το λογισμικό του είναι γραμμένο σε Java (γλώσσα προγραμματισμού) και μπορεί να τρέξει σε όλα τα λειτουργικά συστήματα (Linux, Windows, και Macintosh).

Το Weka είναι μια συλλογή αλγόριθμων μηχανικής μάθησης και περιλαμβάνει εργαλεία προεπεξεργασίας δεδομένων. Είναι σχεδιασμένο έτσι ώστε να μπορεί ο χρήστης να δοκιμάζει γρήγορα και ευέλικτα τις υπάρχουσες μεθόδους σε νέα δεδομένων. Παρέχει εκτεταμένη υποστήριξη για την όλη διαδικασία στην πειραματική φάση της εξόρυξη δεδομένων, συμπεριλαμβανομένης της προετοιμασίας των δεδομένων εισόδου, της αξιολόγησης της μάθησης μέσω στατιστικών, και οπτικοποίησης των δεδομένων εισόδου και του αποτελέσματος της μάθησης (προβλέψεις). Επιπλέον, περιλαμβάνει μια μεγάλη ποικιλία από αλγόριθμους

μάθησης και ένα ευρύ φάσμα εργαλείων προεπεξεργασίας των δεδομένων. Αυτή η ολοκληρωμένη εργαλειοθήκη είναι προσβάσιμη μέσω της διεπαφής του Weka (περιβάλλον χρήστη) και οι χρήστες του μπορούν να συγκρίνουν διαφορετικές μεθόδους και αλγόριθμους και να επεξεργαστούν δεδομένα τα οποία δύσκολα γίνονται δια χειρός.

Παρέχει ένα ενιαίο περιβάλλον εργασίας για τους διαφορετικούς αλγόριθμους μάθησης που περιλαμβάνει, μαζί με μεθόδους για την επεξεργασία και την αξιολόγηση του αποτελέσματος των προγραμμάτων για κάθε σύνολο δεδομένων.

Τι περιλαμβάνει το WEKA

Το Weka παρέχει υλοποιημένους αλγορίθμους μάθησης και μπορούν να εφαρμοστούν εύκολα στα διάφορα σύνολα δεδομένων. Περιλαμβάνει επίσης μια ποικιλία από εργαλεία για τον μετασχηματισμό των δεδομένων. Η προεπεξεργασία των δεδομένων καθώς και η ταξινόμηση τους γίνονται χωρίς το γράψιμο οποιουδήποτε κώδικα προγραμματισμού. Τα δεδομένα εισάγονται με συγκεκριμένη μορφή και αποθηκεύονται σε ARFF αρχεία.

Οι μέθοδοι που περιλαμβάνει δίνουν λύση στα κυριότερα προβλήματα της εξόρυξης δεδομένων: παλινδρόμηση, ταξινόμηση, ομαδοποίηση και εξόρυξη κανόνων συσχέτισης.

Χρήση του WEKA

Ο καλύτερος και ευκολότερος τρόπος χρήσης του WEKA για τον σκοπό που το χρειαζόμαστε είναι μέσω του GUI (Graphical User Interface / περιβάλλον χρήστη) το οποίο το βρίσκουμε ανοίγοντας την εφαρμογή και επιλέγοντας την επιλογή Explorer. Μέσω του Explorer ο χρήστης αποκτά πρόσβαση σε όλες τις δραστηριότητες του WEKA, όπως για παράδειγμα να δώσει στην εφαρμογή ένα αρχείο δεδομένων σε μορφή ARFF και να εφαρμόσει έναν αλγόριθμο σε αυτά.

Attributes / Χαρακτηριστικά :

Κάθε παράδειγμα στη μηχανική μάθηση δίνει είσοδο η οποία χαρακτηρίζεται από τιμές για ένα σταθερό και προκαθορισμένο σύνολο χαρακτηριστικών ή ιδιοτήτων. Οι τιμές που παίρνουν αυτά τα χαρακτηριστικά είναι μία μέτρηση η οποία αναφέρεται στην συγκεκριμένη ιδιότητα. Οι τιμές αυτές διακρίνονται σε ονομαστικές και αριθμητικές. Οι αριθμητικές τιμές μπορούν να είναι ακέραιοι ή δεκαδικοί αριθμοί ενώ οι ονομαστικές τιμές παίρνουν τιμές οι οποίες είναι μια "επεξήγηση" της κάθε κατηγορίας. Πιο συγκεκριμένα, οι τιμές στις ονομαστικές είναι

συμβολοσειρές, δηλαδή ετικέτες και ονομασίες των χαρακτηριστικών. Για παράδειγμα, στα δεδομένα των φοιτητών του ΤΕΠΑΚ, οι ονομαστικές τιμές που μπορούν να πάρουν στο χαρακτηριστικό σχολή είναι "Λύκειο" και "Τεχνική Σχολή".

Ετοιμάζοντας το δείγμα εισόδου

Η προετοιμασία του δείγματος εισόδου για την έρευνα στην εξόρυξη δεδομένων συνήθως καταναλώνει το μεγαλύτερο μέρος της προσπάθειας που επενδύεται στην όλη διαδικασία της εξόρυξης δεδομένων. Το δείγμα εισόδου έχει συγκεκριμένη μορφή η οποία είναι σε ARFF αρχείο το οποίο χρησιμοποιείται στο σύστημα του Weka.

ARFF αρχείο

Μια πλήρης αντιπροσωπευτική μορφή για σύνολα δεδομένων είναι το ARFF αρχείο. Στο παρακάτω παράδειγμα παρουσιάζεται το πως φαίνονται τα δεδομένα σε ένα ARFF αρχείο.

```
@relation provlepsi


@attribute neaellinika real
@attribute istoria real
@attribute aglika real
@attribute pliroforiki real
@attribute viologia real
@attribute politikioikonomia real
@attribute mathimatikaenishimena real
@attribute tehnologia real %ola ta texnologika
@attribute mathimatika real
@attribute grafikestexnes real
@attribute mathimatikatexnikisthk real
@attribute mathimatikatexnikispk real
@attribute allo1 real
@attribute allo2 real
@attribute mesosoros real
@attribute sxoli real % 0 osi ine likio, 1 osi ine texniki
@attribute gpa real
```

@data

14.3,15.5,0,0,0,0,0,13.18,15.75,0,0,0,0,0,14.68,0,7.66

10.55,0,0,0,0,0,0,9.95,17.75,14.1,0,0,0,0,13.09,0,7.66

12.3,0,0,0,0,0,0,14.5,14.85,13.9,0,0,0,0,13.89,0,8.37

14.4,0,16.3,12.93,0,0,5.15,0,0,0,0,0,0,0,12.20,0,5.86

13.55,12.5,16.2,0,0,0,0,0,18.35,0,0,0,0,0,15.15,0,8.46

6.9,0,0,0,7.35,0,0,0,14.9,18,0,0,0,0,11.79,0,5.42

14.25,0,0,14.5,15.83,0,0,0,18.05,0,0,0,0,0,15.66,0,7.08

16.45,0,0,17.7,15.45,0,15.3,0,0,0,0,0,0,0,16.23,0,9.27

8.2,0,0,0,0,7.39,0,0,14.5,12,0,0,0,0,10.52,0,4.65

14.75,10.8,0,0,0,11.88,0,0,17.3,0,0,0,0,0,13.68,0,7.12

14.15,11.35,15.3,0,0,0,0,0,18.5,0,0,0,0,0,14.83,0,5.92

8.7,0,0,0,0,0,0,0,0,16.1,11.55,0,18.3,2.4,11.41,1,5.16

Πίνακας 2: Παράδειγμα ARFF αρχείου

Στο αρχείο αυτό καταγράφονται τα χαρακτηριστικά και οι τιμές που μπορούν να πάρουν καθώς και τα σχόλια. Αν πάρουν αριθμό για τιμή, για παράδειγμα, θα πρέπει να αναφερθεί ως "numeric", ενώ αν θα είναι ονομαστικές τιμές θα πρέπει να δηλωθούν εξ' αρχής μέσα σε αγκύλες (πχ {Λύκειο, Τεχνική}. Σε αυτή την περίπτωση δεν έχουμε ονομαστικές τιμές αλλά για παράδειγμα μπορούσαμε να δηλώσουμε την σχολή ως ονομαστική και να δίναμε τις επιλογές τεχνική σχολή και λύκειο.

Στη συνέχεια, δίνουμε τα δεδομένα (τιμές) με την σειρά που δώσαμε τα χαρακτηριστικά. Για παράδειγμα αν τα χαρακτηριστικά είναι οι βαθμολογίες των φοιτητών στα διάφορα μαθήματα, η σχολή τους, ο μέσος όρος τους και η τελική βαθμολογία στο Πανεπιστήμιο τότε τα δεδομένα είναι κάπως έτσι: 8.7,0,0,0,0,0,0,0,0,16.1,11.55,0,18.3,2.4,11.41,1,5.16

Οι γραμμές που ξεκινάνε με το σύμβολο % παρουσιάζουν τα σχόλια.

4.2 Περιγραφή συστήματος

Μετά τη συλλογή των δεδομένων θα πρέπει να δημιουργηθεί ένα αρχείο ARFF στο οποίο θα εισάγουμε εκεί όλα τα δεδομένα με τρόπο που είναι αναγνωρίσιμα από το WEKA. Τα μαθήματα που επιλέχθηκαν να μπουν στο αρχείο αυτό επιλέχθηκαν με βάση του αν η πλειοψηφία των μαθητών εξετάστηκαν σε αυτά και αν κρίθηκαν σημαντικά λόγω της δυσκολίας τους. Επίσης, κάποια μαθήματα της τεχνικής σχολής ομαδοποιήθηκαν σε μια κατηγορία η οποία ονομάστηκε τεχνολογικά μαθήματα και περιλάμβανε όλα τα μαθήματα που αφορούν την τεχνολογία. Επιπλέον, υπήρχαν πέντε ειδών μαθηματικά από τα οποία μόνο τα δύο που ήταν κοινού κορμού ομαδοποιήθηκαν. Τα άλλα τρία έμεινα ανεξάρτητα για τον λόγο ότι διαφέρουν ως προς την δυσκολία τους. Τα υπόλοιπα μαθήματα, θα συμπεριληφθούν στις κατηγορίες άλλο 1 και άλλο 2 σε περίπτωση που κάποιος μαθητής εξετάστηκε σε μαθήματα που δεν περιλήφθηκαν στο σύστημα. Τα μαθήματα που επιλέχθηκαν είναι τα εξής: Νέα Ελληνικά, Ιστορία, Αγγλικά, Πληροφορική, Βιολογία, Πολιτική Οικονομία, Μαθηματικά Ενισχυμένα, Τεχνολογικά Μαθήματα, Μαθηματικά Κοινού Κορμού, Γραφικές Τέχνες, Μαθηματικά Θεωρητικής Κατεύθυνσης, Μαθηματικά Πρακτικής Κατεύθυνσης, Άλλο 1, Άλλο 2. Στο αρχείο ARFF εισήχθηκαν επίσης άλλα δεδομένα όπως ο μέσος όρος, η σχολή στην οποία φοίτησαν και ο βαθμός που αποφοίτησαν από το ΤΕΠΑΚ. Το αν ο μαθητής φοίτησε σε Τεχνική Σχολή, στο αρχείο ARFF δηλώθηκε με 1 ενώ αν φοίτησε σε Λύκειο δηλώθηκε με 0. Ο βαθμός με τον οποίο αποφοίτησαν οι απόφοιτοι του τμήματος, θα χρησιμοποιηθεί ως εκπαίδευση του αλγορίθμου και θα είναι η επιθυμητή έξοδος του συστήματος (δηλαδή η πρόβλεψη). Το WEKA αναγνωρίζει αυτόματα ποιά θα είναι η επιθυμητή έξοδος χωρίς να το δηλώσουμε και το κάνει αυτό παίρνοντας το τελευταίο χαρακτηριστικό που γράψαμε στο ARFF αρχείο. Αν για οποιοδήποτε λόγο ο χρήστης θέλει να επιλέξει κάποιον άλλο χαρακτηριστικό για την επιθυμητή έξοδο μπορεί να το κάνει επιλέγοντας της από το Explorer του WEKA.

Στη συνέχεια θα πρέπει να εκπαιδευτεί το σύστημα με συγκεκριμένο αλγόριθμο αποτελέσματα. Για την εκπαίδευση του συστήματος διαχώρισα το αρχείο των δεδομένων σε δυο κομμάτια. Στο training και στο test. Στο training έβαλα το 70% των δεδομένων για εκπαίδευση του συστήματος (δηλαδή για να μάθει, να εκπαιδευτεί, να δημιουργήσει συσχετίσεις). Το υπόλοιπο 30% το οποίο είναι το test set, θα χρησιμοποιηθεί για να εφαρμόσει το σύστημα τον αλγόριθμο του και να

κάνει τις προβλέψεις. Ο διαχωρισμός έγινε με μορφή τυχαιότητας δηλαδή κάθε τρίτος αριθμός μια τριάδας έμπαινε στο test set, ξεχωριστά σε τεχνική σχολή και λύκειο.

Ο αλγόριθμος που επιλέχθηκε είναι ο Multilayer Perceptron ο οποίος ανήκει στα νευρωνικά δίκτυα . Επίσης με βάση κάποιες μετρήσεις και συγκρίσεις που έγιναν με άλλους αλγόριθμους και θα δούμε στην συνέχεια, είχε τα καλύτερα και πιο ακριβή αποτελέσματα. Στο μόνο που υστερεί ο αλγόριθμος αυτός συγκριτικά με άλλους είναι στον χρόνο που χρειάζεται για να εφαρμοστεί σε μεγάλα προβλήματα αλλά στην περίπτωση την δική μας αυτό δεν μας απασχολεί καθώς τα δεδομένα εισόδου είναι 117 βαθμολογίες φοιτητών από τις οποίες οι 82 θα χρησιμοποιηθούν για training και οι άλλες 35 για test. Αυτό γίνεται για να δούμε πόσο ακριβής είναι ο αλγόριθμος στις προβλέψεις του.

Πιο κάτω αναφέρω λίγα λόγια για τους αλγόριθμους που χρησιμοποίησα πειραματικά για να δούμε ποιος είναι πιο ακριβής στις προβλέψεις του και στην επόμενη ενότητα των αποτελεσμάτων θα παρουσιάσω τον μέσο όρο λάθους κάθε αλγόριθμου καθώς και αναλυτικά τις προβλέψεις που είχε ο καθένας στο test set.

4.2.1 Περιγραφή αλγορίθμων

MULTILAYER PERCEPTRON: Ο αλγόριθμος multilayer perceptron (MLP) είναι ένα τεχνητό νευρωνικό δίκτυο πρόσθιας τροφοδότησης που χαρτογραφεί σύνολα δεδομένων εισόδου σε ένα σύνολο μονάδων εξόδου. Ένα MLP αποτελείται από πολλαπλά στρώματα κόμβων σε ένα κατευθυνόμενο γράφο, με κάθε στρώμα να είναι πλήρως συνδεδεμένο με το επόμενο. Ο αλγόριθμος χρησιμοποιεί μια συγκεκριμένη τεχνική επίβλεψης η οποία ονομάζεται " backpropagation" για την εκπαίδευση του δικτύου. Η τεχνική αυτή είναι μια κοινή μέθοδος της εκπαίδευσης τεχνητών νευρωνικών δικτύων που χρησιμοποιείται σε συνδυασμό με την μέθοδο βελτιστοποίησης. Η μέθοδος αυτή χρησιμοποιείται για τον υπολογισμό των βαρών καθώς και για την ανανέωση των τιμών τους.

LAZY IBK: Μπορεί να επιλέξει την κατάλληλη τιμή του K με βάση διασταυρωμένης επικύρωσης. Ο αλγόριθμος αυτός μπορεί να χρησιμοποιηθεί για να επιταχύνει το έργο της εξεύρεσης των πλησιέστερων γειτονικών κόμβων. Ο αριθμός των κοντινότερων γειτόνων μπορούν να καθοριστεί αυτόματα χρησιμοποιώντας διασταυρωμένης επικύρωσης, με ανώτατο όριο δίνεται από την καθορισμένη τιμή. Οι προβλέψεις για περισσότερους από ένα γείτονες μπορεί να είναι υπολογιστεί ανάλογα με την απόστασή τους από την περίπτωση που εξετάζουμε και η απόσταση μετατρέπεται σε βάρος. Ο αριθμός των περιπτώσεων που εκπαιδεύτηκαν φυλάγονται από τον ταξινομητή (ο οποίος έχει περιορισμένη χωρητικότητα αποθήκευσης) και καθώς προστίθενται νέες περιπτώσεις, τα παλαιότερα αφαιρούνται.

LAZY K-STAR: ο αλγόριθμος αυτός είναι ένα παράδειγμα που βασίζεται στην ταξινόμηση, δηλαδή η τάξη του ενός παραδείγματος βασίζεται στην τάξη παρόμοιων περιπτώσεων

κατάρτισης με αυτό, όπως καθορίζεται από μια συνάρτηση ομοιότητας. Διαφέρει από τους άλλους αλγόριθμους για τον λόγο ότι χρησιμοποιεί την συνάρτηση εντροπίας που βασίζεται στην απόσταση. Η χρήση της εντροπίας ως μέτρο υπολογισμού της απόστασης έχει πολλά οφέλη όπως για παράδειγμα παρέχει μια πιο συνεκτική προσέγγιση για την αντιμετώπιση ονομαστικών χαρακτηριστικών, χαρακτηριστικά πραγματικών τιμών και για τιμές που λείπουν.

LINEAR REGRESSION: Κάνει χρήση γραμμικής παλινδρόμησης για την πρόβλεψη και είναι σε θέση να ασχοληθεί με τις σταθμισμένες περιπτώσεις. Όταν τα αποτελέσματα ή η τάξη είναι αριθμητικές τιμές, και όλα τα χαρακτηριστικά αποτελούνται από αριθμητικές τιμές, τότε η γραμμική παλινδρόμηση είναι η καταλληλότερη τεχνική για να τα εξετάσει. Η ιδέα είναι να εκφράσει την τάξη με γραμμικό συνδυασμό των ιδιοτήτων και με προκαθορισμένα βάρη. Έτσι τα βάρη υπολογίζονται μέσω των δεδομένων εκπαίδευσης. Η γραμμική παλινδρόμηση είναι μια τέλεια και απλή μέθοδος για προβλέψεις αριθμητικών τιμών και χρησιμοποιείται σε πολλές εφαρμογές στατιστική ανάλυσης.

PLS CLASSIFIER: ταξινομητής ο οποίος περιλαμβάνει το φίλτρο PLS και αξιοποιεί αυτό το φίλτρο για να εκτελέσει προβλέψεις. Ο αλγόριθμος αυτός μαθαίνει ένα μοντέλο μερικής παλινδρόμησης και χρησιμοποιεί το PLS φίλτρο για να εκπαιδευτεί μέσω των δεδομένων εκπαίδευσης τα οποία μέσω του φίλτρου αυτού έχουν μετασχηματιστεί. Όλες οι επιλογές του φίλτρου PLS είναι διαθέσιμες στους χρήστες για επεξεργασία.

4.2.2 Περιβάλλον χρήστη

Για να μπορέσει ο χρήστης / μαθητής να εισάγει τα αποτελέσματα των εξετάσεων του και να δει τη πρόβλεψη του θα πρέπει να δημιουργήσουμε ένα περιβάλλον χρήστη το οποίο θα είναι online.

Για αυτό δημιούργησα μία ιστοσελίδα με φόρμα html στην οποία ο χρήστης θα βάζει στα πεδία της φόρμας τα αποτελέσματα του και τα μαθήματα του και θα παίρνει την πρόβλεψη του. Μέσω της θα υποβάλει τις βαθμολογίες φόρμας αυτής, ο χρήστης του, και αυτόματα θα φορτώνονται οι κλάσεις του WEKA οι οποίες χρειάζονται για τις συσχετίσεις και τις προβλέψεις.

Εικόνα 1: Περιβάλλον χρήστη εφαρμογής

Ο χρήστης (φοιτητής, μαθητής, καθηγητής) στην φόρμα αυτή θα εισάγει τις βαθμολογίες των μαθημάτων, τον μέσο όρο και θα επιλέγει την σχολή και στην συνέχεια υποβάλλοντας τα θα φορτώνονται αυτόματα οι κλάσεις του WEKA που είναι υπεύθυνες για την πρόβλεψη και μέσω ενός μηνύματος θα παίρνει πίσω την προβλεπόμενη τιμή. Στα μαθήματα που δεν εξετάστηκε θα πρέπει να βάλει την τιμή 0 και δεν πρέπει να μένουν κενά ούτως ώστε να συμπληρωθούν ορθά τα δεδομένα εισόδου και να μην υπάρξουν προβλήματα κατά την φόρτωση του ARFF αρχείου. Αν για οποιοδήποτε λόγο μένουν κενά, θα εμφανίζεται ένα μήνυμα που θα του λέει να συμπληρώσει το συγκεκριμένο πεδίο.

Η εφαρμογή είναι διαθέσιμη στον σύνδεσμο: http://cis.cut.ac.cy/~Stefanos.Vrionides/pti/

Κεφάλαιο 5: Αποτελέσματα

5.1 Σύγκριση Αλγορίθμων

Πίνακας 3:Multilayer Perceptron

	GPA	Πρόβλεψη	Σφάλμα Πρόβλεψης
1	7.66	7.004	-0.656
2	7.66	6.301	-1.359
3	8.37	6.31	-2.06
4	5.86	7.092	1.232
5	8.46	7.414	-1.046
6	5.42	6.2	0.78
7	7.08	7.289	0.209
8	9.27	7.258	-2.012
9	4.65	6.299	1.649
10	7.12	7.172	0.052
11	5.92	7.384	1.464
12	5.16	5.203	-0.043
13	3.18	6.451	3.271
14	6.15	6.72	0.57
15	5.98	5.417	-0.563
16	6.82	6.687	-0.133
17	6.68	5.957	-0.723
18	5.85	5.471	-0.379
19	2.97	6.743	3.773
20	6.29	6.611	0.321
21	8.06	6.33	-1.73
22	4.78	5.133	-0.353
23	6.96	7.06	0.1
24	6.26	5.447	-0.813
25	7.73	6.722	-1.008
26	5.77	5.272	-0.498
27	6.39	6.865	0.475
28	7.72	7.251	-0.469
29	5.81	6.753	0.943
30	6.57	6.717	-0.147
31	7.02	7.22	0.202
32	8	7.525	-0.475
33	7.75	6.937	-0.813
34	9.18	7.332	-1.848
35	5.62	5.293	-0.327

Πίνακας 4:LAZY IBK

	GPA	Πρόβλεψη	Σφάλμα Πρόβλεψης
1	7.66	7.65	-0.01
2	7.66	4.79	-2.87
3	8.37	4.84	-3.53
4	5.86	8.4	2.54
5	8.46	7.98	-0.48
6	5.42	7	1.58
7	7.08	6.4	-0.68
8	9.27	5.73	-3.54
9	4.65	5.64	0.99
10	7.12	7.18	0.06
11	5.92	7.98	2.06
12	5.16	4.64	-0.52
13	3.18	5.64	2.46
14	6.15	7.04	0.89
15	5.98	4.12	-1.86
16	6.82	7.48	0.66
17	6.68	4.27	-2.41
18	5.85	4.6	-1.25
19	2.97	7.62	4.65
20	6.29	7.65	1.36
21	8.06	4.84	-3.22
22	4.78	4.64	-0.14
23	6.96	7.3	0.34
24	6.26	2.28	-3.98
25	7.73	7.04	-0.69
26	5.77	3.52	-2.25
27	6.39	7.62	1.23
28	7.72	7.3	-0.42
29	5.81	8.4	2.59
30	6.57	7.04	0.47
31	7.02	8.4	1.38
32	8	8.78	0.78
33	7.75	6.53	-1.22
34	9.18	7.3	-1.88
35	5.62	2.76	-2.86

Πίνακας 5: LAZY K-STAR

	GPA	Πρόβλεψη	Σφάλμα Πρόβλεψης
1	7.66	7.526	-0.134
2	7.66	6.668	-0.992
3	8.37	6.102	-2.268
4	5.86	6.732	0.872
5	8.46	7.98	-0.48
6	5.42	5.895	0.475
7	7.08	7.505	0.425
8	9.27	6.83	-2.44
9	4.65	5.139	0.489
10	7.12	7.158	0.038
11	5.92	8.008	2.088
12	5.16	5.488	0.328
13	3.18	5.641	2.461
14	6.15	6.938	0.788
15	5.98	4.11	-1.87
16	6.82	7.449	0.629
17	6.68	4.999	-1.681
18	5.85	5.364	-0.486
19	2.97	7.569	4.599
20	6.29	7.56	1.27
21	8.06	6.226	-1.834
22	4.78	5.153	0.373
23	6.96	7.141	0.181
24	6.26	5.047	-1.213
25	7.73	6.979	-0.751
26	5.77	3.325	-2.445
27	6.39	6.917	0.527
28	7.72	7.143	-0.577
29	5.81	8.134	2.324
30	6.57	7.396	0.826
31	7.02	6.019	-1.001
32	8	7.917	-0.083
33	7.75	7.318	-0.432
34	9.18	7.084	-2.096
35	5.62	3.404	-2.216

Πίνακας 6: LINEAR REGRESSION

	GPA	Πρόβλεψη	Σφάλμα Πρόβλεψης
1	7.66	6.977	-0.683
2	7.66	5.264	-2.396
3	8.37	5.609	-2.761
4	5.86	7.295	1.435
5	8.46	7.721	-0.739
6	5.42	5.113	-0.307
7	7.08	7.988	0.908
8	9.27	8.5	-0.77
9	4.65	4.993	0.343
10	7.12	7.247	0.127
11	5.92	7.628	1.708
12	5.16	5.2	0.04
13	3.18	5.601	2.421
14	6.15	6.226	0.076
15	5.98	4.821	-1.159
16	6.82	6.27	-0.55
17	6.68	6.677	-0.003
18	5.85	4.526	-1.324
19	2.97	6.299	3.329
20	6.29	6.316	0.026
21	8.06	6.141	-1.919
22	4.78	5.349	0.569
23	6.96	7.347	0.387
24	6.26	3.65	-2.61
25	7.73	6.392	-1.338
26	5.77	3.782	-1.988
27	6.39	6.794	0.404
28	7.72	8.334	0.614
29	5.81	6.294	0.484
30	6.57	6.357	-0.213
31	7.02	7.688	0.668
32	8	8.307	0.307
33	7.75	6.683	-1.067
34	9.18	8.567	-0.613
35	5.62	3.124	-2.496

Πίνακας 7: PLSClassifier

	GPA	Πρόβλεψη	Σφάλμα Πρόβλεψης
1	7.66	7.126	-0.534
2	7.66	5.384	-2.276
3	8.37	5.808	-2.562
4	5.86	7.322	1.462
5	8.46	7.773	-0.687
6	5.42	5.02	-0.4
7	7.08	8.018	0.938
8	9.27	7.99	-1.28
9	4.65	4.988	0.338
10	7.12	7.262	0.142
11	5.92	7.696	1.776
12	5.16	5.081	-0.079
13	3.18	5.579	2.399
14	6.15	6.262	0.112
15	5.98	4.769	-1.211
16	6.82	6.287	-0.533
17	6.68	7.102	0.422
18	5.85	4.57	-1.28
19	2.97	6.219	3.249
20	6.29	6.373	0.083
21	8.06	6.006	-2.054
22	4.78	5.253	0.473
23	6.96	7.38	0.42
24	6.26	3.463	-2.797
25	7.73	6.404	-1.326
26	5.77	3.719	-2.051
27	6.39	6.789	0.399
28	7.72	8.324	0.604
29	5.81	6.134	0.324
30	6.57	6.381	-0.189
31	7.02	7.584	0.564
32	8	8.34	0.34
33	7.75	7.027	-0.723
34	9.18	8.567	-0.613
35	5.62	3.158	-2.462

Αλγόριθμος	Μέσο Σφάλμα	Διασπορά Μέσου Σφάλματος	Μέγιστο Σφάλμα
MULTILAYER PERCEPTRON	0.9284	0.727	3.773
LAZY IBK	1.6529	1.723	4.65
LAZY K-STAR	1.1912	1.277	4.599
LINEAR REGRESSION	1.0509	1.401	3.329
PLS CLASSIFIER	1.0601	1.412	3.249

Πίνακας 8: Σύγκριση Αλγορίθμων

5.2 Σύγκριση Μέσων

Για την σύγκριση των μέσων όρων της επίδοσης των φοιτητών με την πρόβλεψη χρησιμοποίησα t-test το οποίο εφάρμοσα στο SPSS (λογισμικό στατιστικής ανάλυσης). Μέσου του t-test μπορούμε να δούμε αν οι μέσοι όροι δύο συνόλων από τιμές διαφέρουν μεταξύ τους, ελέγχοντας αν η διαφορά αυτή είναι στατιστικά σημαντική και να συγκρίνουμε επιπλέον ποιός αλγόριθμος είναι καλύτερος για το σύστημα μας. Μέσο GPA = 6.53

Αλγόριθμος	Μέσο Σφάλμα	Μέσος όρος Πρόβλεψης
MULTILAYER PERCEPTRON	0.9284	6.53829
LAZY IBK	1.6529	6.2971
LAZY K-STAR	1.1912	6.45326
LINEAR REGRESSION	1.0509	6.31657
PLS CLASSIFIER	1.0601	6.31880

Πίνακας 9: Σύγκριση Μέσων

Συγκρίνοντας τα αποτελέσματα από τους αλγορίθμους ο Multilayer Perceptron είναι πιο κοντά τόσο στο μέσο όρο των τιμών της επίδοσης των φοιτητών με την πρόβλεψη όσο και με τον μέσο όρο του λάθους που προκύπτει. Ο Multilayer Perceptron με τιμή 0.9284 και μόλις 0.03 διαφορά στην σύγκριση των μέσων είναι ο αλγόριθμος με τις πιο ακριβείς προβλέψεις και η διαφορά του μέσου όρου πρόβλεψης με τον μέσο όρο επίδοσης των φοιτητών είναι ασήμαντη επειδή είναι πάρα πολύ μικρή.

5.3 Συσχέτιση δεδομένων εισόδου με πρόβλεψη

Στο σημείο αυτό αποφάσισα να κάνω ένα περαιτέρω έλεγχο στον οποίο θα προσπαθήσω να βρω συσχετίσεις μεταξύ των δεδομένων εισόδου (μαθήματα, μέσος όρος, σχολή, GPA) με την πρόβλεψη. Με αυτό τον τρόπο θα μπορούμε να ξέρουμε ποιά δεδομένα επηρεάζουν περισσότερο την προβλεπόμενη τιμή.

Για την επίλυση αυτού του προβλήματος θα χρειαστούμε τον συντελεστή γ ο οποίος θα υπολογιστεί και πάλι από το εργαλείο στατιστικής ανάλυσης SPSS. Ο συντελεστής γ είναι χρήσιμος σε αυτό το στάδιο γιατί μπορεί να δείξει πόσο έντονη είναι η σχέση μεταξύ δύο μεταβλητών. Η τιμή που μπορεί να πάρει είναι από το -1 μέχρι το 1 όπου -1 είναι ισχυρή αρνητική συσχέτιση και 1 ισχυρή θετική συσχέτιση. Καθώς ο δεκαδικός αριθμός πλησιάζει το 0 τότε η σχέση εξασθενεί (πχ. 0.2 = ασθενής θετική συσχέτιση).

Μάθημα	Συντελεστής γ	Ερμηνεία
Νέα Ελληνικά	0.663	Ισχυρή θετική συσχέτιση
Ιστορία	0.763	Ισχυρή θετική συσχέτιση
Αγγλικά	0.475	Μέτρια θετική συσχέτιση
Πληροφορική	0.512	Ισχυρή θετική συσχέτιση
Βιολογία	0.530	Ισχυρή θετική συσχέτιση
Πολιτική Οικονομία	0.312	Μέτρια θετική συσχέτιση
Μαθηματικά Ενισχυμένα	0.631	Ισχυρή θετική συσχέτιση
Τεχνολογία	-0.415	Μέτρια αρνητική συσχέτιση
Μαθηματικά Κοινού Κορμού	0.332	Μέτρια θετική συσχέτιση
Γραφικές Τέχνες	-0.329	Μέτρια αρνητική συσχέτιση
Μαθηματικά Θ.Κ Τεχνικής	-0.838	Ισχυρή αρνητική συσχέτιση
Μαθηματικά Π.Κ Τεχνικής	-0,798	Ισχυρή αρνητική συσχέτιση
Άλλο Μάθημα	-0.852	Ισχυρή αρνητική συσχέτιση
Μέσος Όρος	0.559	Ισχυρή θετική συσχέτιση
Σχολή	-1	Ισχυρή αρνητική συσχέτιση
GPA	0.423	Μέτρια θετική συσχέτιση

Πίνακας 10: Συσχέτιση μεταβλητών

Ερμηνεία:

Ισχυρή θετική συσχέτιση: Τα μαθήματα Νέα Ελληνικά, Ιστορία, Μαθηματικά Ενισχυμένα, Πληροφορική και Βιολογία είχαν ισχυρή θετική συσχέτιση με την προβλεπόμενη βαθμολογία και αυτό σημαίνει ότι η ύπαρξη τους στις βαθμολογίες ενός φοιτητή θα επηρεάσουν σε μεγάλο βαθμό θετικά την προβλεπόμενη τιμή. Όσο πιο ψηλές είναι οι βαθμολογίες τους σε αυτά τα μαθήματα τόσο πιο ψηλή θα είναι η προβλεπόμενη τους βαθμολογία.

Μέτρια θετική συσχέτιση: Τα μαθήματα Αγγλικά, Πολιτική Οικονομία και Μαθηματικά Κοινού Κορμού είχαν μέτρια θετική συσχέτιση με την προβλεπόμενη βαθμολογία και αυτό σημαίνει ότι η ύπαρξη τους στις βαθμολογίες ενός φοιτητή θα επηρεάσουν ως ένα βαθμό θετικά την προβλεπόμενη τιμή. Όσο πιο ψηλές είναι οι βαθμολογίες τους σε αυτά τα μαθήματα τόσο πιο ψηλή θα είναι η προβλεπόμενη τους βαθμολογία.

Μέτρια αρνητική συσχέτιση: Τα μαθήματα Τεχνολογία και Γραφικές Τέχνες είχαν μέτρια αρνητική συσχέτιση με την προβλεπόμενη βαθμολογία και αυτό σημαίνει ότι η ύπαρξη τους στις βαθμολογίες ενός φοιτητή θα επηρεάσουν ως ένα βαθμό αρνητικά την προβλεπόμενη τιμή. Όσο πιο χαμηλές είναι οι βαθμολογίες τους σε αυτά τα μαθήματα τόσο πιο χαμηλή θα είναι η προβλεπόμενη τους βαθμολογία

Ισχυρή αρνητική συσχέτιση: Τα μαθήματα Μαθηματικά Θεωρητικής Κατεύθυνσης Τεχνικής, Μαθηματικά Πρακτικής Κατεύθυνσης Τεχνικής, το Άλλο Μάθημα, ο Μέσος Όρος και η Σχολή είχαν ισχυρή αρνητική συσχέτιση με την προβλεπόμενη βαθμολογία και αυτό σημαίνει ότι η ύπαρξη τους στις βαθμολογίες ενός φοιτητή θα επηρεάσουν ως ένα σε μεγάλο βαθμό αρνητικά την προβλεπόμενη τιμή. Όσο πιο χαμηλές είναι οι βαθμολογίες τους σε αυτά τα μαθήματα τόσο πιο χαμηλή θα είναι η προβλεπόμενη τους βαθμολογία. Στην περίπτωση της Σχολής αν η τιμή της είναι "1" (δηλαδή ο φοιτητής προέρχεται από την Τεχνική Σχολή) επηρεάζει αρνητικά την προβλεπόμενη βαθμολογία.

Κεφάλαιο 6: Συμπεράσματα και Μελλοντική έρευνα

6.1 Συμπεράσματα

Αναμφίβολα η χρήση της τεχνητής νοημοσύνης για πρόβλεψη μιας μεταβλητής λύνει τα χέρια του ανθρώπου. Πράγματα τα οποία φαίνονται αδιανόητα και χρονοβόρα να γίνουν χειροκίνητα, μέσω της τεχνητής νοημοσύνης τα προβλήματα λύνονται ευκολότερα και γρηγορότερα. Οι τομείς με τους οποίους ασχολήθηκε η παρούσα εργασία (εξόρυξη δεδομένων και μηχανική μάθηση), βοήθησαν στο μέγιστο στην διεκπεραίωση της. Η εκπαίδευση του συστήματος μέσω των δεδομένων που του έδωσα έγινε επιτυχώς με την πρόβλεψη να έχει το ελάχιστο ποσοστό λάθους (μέσος όρος λάθους: 0.98).

Για την επίτευξη ακριβέστερης πρόβλεψη σύγκρινα διαφορετικούς αλγόριθμους και αυτός με την καλύτερη πρόβλεψη ήταν ο Multilayer Perceptron ο οποίος ανήκει στα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα, όπως αποδείχτηκε και στις παρόμοιες έρευνες στις οποίες αναφέρθηκα, βοηθούν και είναι βασικό εργαλείο για την πρόβλεψη μεταβλητών. Μέσω των δεδομένων που συλλέγονται μπορεί να σου δώσει μελλοντική πρόβλεψη για παρόμοιες περιπτώσεις καθώς και να προβλέψει και να προσαρμοστεί σε νέες καταστάσεις.

Το WEKA απόδειξε την χρησιμότητα του σε αυτή την εργασία, καθώς δίνοντας του τα δεδομένα εισόδου και επιλέγοντας τον αλγόριθμο που επιθυμείς, δίνει μια πρόβλεψη που ως επί το πλείστο δεν είχε και μεγάλη απόκλιση από την πραγματική τιμή. Σε μερικές περιπτώσεις οι αλγόριθμοι ξέφευγαν λίγο στην τιμή αλλά αυτό το πιο πιθανό θα οφειλόταν στο μικρό δείγμα που του δόθηκε. Οι τιμές ήταν όλες αριθμητικές και η πρόβλεψη που θέλαμε να μας δώσει ήταν και αυτή αριθμητική και για αυτό το σκοπό θα χρειαζόμασταν πολύ μεγαλύτερο δείγμα.

Όσον αφορά τα μαθήματα και την σύγκριση τους με την πρόβλεψη κατέληξα στο εξής συμπέρασμα: όταν τα μαθήματα άνηκαν στην τεχνική σχολή, η συσχέτιση ήταν αρνητική ενώ όταν τα μαθήματα άνηκαν στο λύκειο η συσχέτιση ήταν θετική. Αυτός ίσως να οφείλεται στους χαμηλούς βαθμούς που ως επί το πλείστο είχαν οι μαθητές της τεχνικής σχολής αλλά ίσως μπορεί να είναι και κάτι άλλο το οποίο μπορεί να μελετηθεί σε μια μελλοντική έρευνα.

Τέλος, η χρήση της εφαρμογής αυτής δεν θα είναι χρήσιμη μόνο για τους φοιτητές οι οποίοι από περιέργεια το πιο πιθανόν θα θελήσουν να την χρησιμοποιήσουν. Η χρήση της από τους καθηγητές του τμήματος πολύ πιθανόν να φανεί χρήσιμη καθώς αν μέσω των προβλέψεων δουν ότι κάποιοι πρωτοετείς φοιτητές θα έχουν χαμηλή επίδοση, θα μπορούσαν κάλλιστα να τους προσεγγίσουν με διακριτικότητα και να τους βοηθήσουν (αν φυσικά τελικά ισχύει η πρόβλεψη). Όπως προαναφέρθηκε και στο κεφάλαιο της επισκόπησης της βιβλιογραφίας, κατά την εξόρυξη δεδομένων προκύπτουν κάποιοι ηθικοί προβληματισμοί. Από την στιγμή που οι καθηγητές θα χρησιμοποιήσουν την εφαρμογή, υπάρχουν οι πιθανότητες να μην είναι αντικειμενικοί απέναντι στους φοιτητές. Για παράδειγμα, αν η βαθμολογία ενός φοιτητή που θα προκύψει από την εφαρμογή είναι χαμηλή τότε υπάρχουν οι πιθανότητες ο καθηγητής να επηρεαστεί και να δημιουργήσει αρνητική εικόνα για τον συγκεκριμένο φοιτητή πριν καν τον γνωρίσει. Αυτό θα

έχει πιθανότατα αρνητικές συνέπειες εις βάρος του φοιτητή, ο οποίος ίσως να είναι αδύναμος μαθητής και χωρίς βοήθεια από τους καθηγητές του μπορεί να αποτύχει στις σπουδές του.

6.2 Μελλοντική Έρευνα

Λόγω αντικειμενικών δυσκολιών, η online εφαρμογή δεν υλοποιήθηκε όπως θα έπρεπε. Υπήρχε δυσκολία στην αυτόματη φόρτωση των κλάσεων του WEKA στην φόρμα λόγω του ότι είναι γραμμένες σε γλώσσα προγραμματισμού Java. Έτσι για την διεκπεραίωση της εφαρμογής χρησιμοποίησα γλώσσα προγραμματισμού JavaScript και πέρασα τις προβλέψεις χειροκίνητα. Αυτό έχει ως αποτέλεσμα το σύστημα να μην είναι τόσο ακριβές όσο θα ήταν μέσω της χρήσης τεχνητής νοημοσύνης κατευθείαν στις βαθμολογίες των χρηστών. Άρα αυτό μπορεί να γίνει σε μια μελλοντική έρευνα, δηλαδή όπως ο χρήστης θα υποβάλλει τις βαθμολογίες του, οι κλάσεις του WEKA θα φορτώνονται αυτόματα και έτσι η πρόβλεψη θα είναι πιο ακριβής.

Βιβλιογραφία

- Landau, S. & Everitt, B., 2004. *A Handbook of Statistical Analyses using SPSS*. New York: Chapman & Hall/CRC
- Russell, S. & Norvig, P, 1995. *Artificial Intelligence A Modern Approach.* New Jersey: Prentice Hall.
- Theofilis, G., 2013. *Weka Classifiers Summary*, [Διαδίκτυο] Διαθέσιμο στο: http://www.academia.edu/5167325/Weka_Classifiers_Summary [πρόσβαση 25 Απριλίου]
- Witten, I., Frank, E., Hall, M., 2011. *Data Mining, Practical Machine Learning Tool – Third Edition.* Burlington: Morgan Kaufmann.
- Ιωάννου, Δ., 2013. *Μοντέλα Τεχνητής Νοημοσύνης για την Πρόβλεψη Αποτελεσμάτων σε Αγώνες Ποδοσφαίρου*, [Διαδίκτυο]. Διαθέσιμο στο: http://vivliothmmy.ee.auth.gr/2273/1/Διπλωματική_Ιωάννου.pdf [πρόσβαση 18 Απριλίου 2015]
- Σοφός, Ι., 2013. *Πρόβλεψη τιμής μετοχής με χρήση τεχνητής ευφυΐας (Νευρωνικά Δίκτυα)*, [Διαδίκτυο]. Διαθέσιμο στο: http://dspace.lib.ntua.gr/bitstream/handle/123456789/8571/SOFOS_IOANNIS_DIPLOMATIKI_2013.pdf?sequence=1 [πρόσβαση 18 Απριλίου 2015]
- Τιτομιχελάκη, Μ., 2010. *Πρόβλεψη μετοχών με τη χρήση Νευρωνικών δικτύων,* [Διαδίκτυο]. Διαθέσιμο στο: http://nefeli.lib.teicrete.gr/browse/sdo/fi/2010/TitomichelakiMaria/attached-document-1345197950-290317-2215/Titomixelaki_Maria2010.pdf [πρόσβαση 18 Απριλίου 2015]

Παραρτήματα

A. T-TEST

A1. Multilayer Perceptron

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.53829	.727311	.122938

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	T	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	53.184	34	.000	6.538286	6.28845	6.78813

A2. LAZY IBK

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.2971	1.72313	.29126

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	21.620	34	.000	6.29714	5.7052	6.8891

A3. LAZY K-STAR

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.45326	1.276993	.215851

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	29.897	34	.000	6.453257	6.01459	6.89192

A4. LINEAR REGRESSION

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.31657	1.401331	.236868

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	26.667	34	.000	6.316571	5.83520	6.79795

A5. Radial basis function network

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.64260	.114012	.019271

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	344.685	34	.000	6.642600	6.60344	6.68176

A6. PLSClassifier

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
GPA	35	6.5763	1.45072	.24522
Prediction	35	6.31880	1.412094	.238687

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
GPA	26.818	34	.000	6.57629	6.0779	7.0746
Prediction	26.473	34	.000	6.318800	5.83373	6.80387

Β. Έλεγχος Διμεταβλητών Υποθέσεων

B1. Νέα ελληνικά

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
neaellinika * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.663	.069	9.622	.000
N of Valid Cases		35			

B2. Ιστορία

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
istoria * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.763	.125	2.526	.012
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β3. Αγγλικά

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
aglika * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.475	.142	2.967	.003
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β4. Πληροφορική

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
pliroforiki * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.512	.153	2.664	.008
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β5. Βιολογία

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
viologia * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.530	.158	2.414	.016
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β6. Πολιτική Οικονομία

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
politikioikonomia * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.312	.199	1.342	.180
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β7. Μαθηματικά Ενισχυμένα

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mathimatikaenishimena * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.631	.158	2.035	.042
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B8. Τεχνολογία

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
tehnologia * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.415	.151	-2.622	.009
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B9. Μαθηματικά Κοινού Κορμού

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mathimatika * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.332	.125	2.605	.009
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β10. Γραφικές Τέχνες

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
grafikestexnes * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.329	.147	-2.427	.015
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β11. Μαθηματικά Θεωρητικής Κατεύθυνσης Τεχνική Σχολής

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mathimatikatexnikisthk * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.838	.134	-1.880	.060
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B12. Μαθηματικά Πρακτικής Κατεύθυνσης Τεχνική Σχολής

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mathimatikatexnikispk * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.798	.097	-1.973	.048
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B13. Άλλα μαθήματα

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
allo1 * Prediction	35	100.0%	0	.0%	35	100.0%
allo2 * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.852	.084	-4.453	.000
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.879	.093	-1.984	.047
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Β14. Μέσος Όρος

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mesosoros * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.559	.085	6.581	.000
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B15. Σχολή

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
sxoli * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-1.000	.000	-3.944	.000
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

B16. GPA

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gpa * Prediction	35	100.0%	0	.0%	35	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.423	.096	4.407	.000
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Γ. Κλάσεις WEKA

public class ArffViewer: Κλάση για να μπορούμε να δούμε το αρχείο Arff.
public java.lang.Object copy(): δημιουργία αντιγράφου ενός attribute.
public double predicted(): Δίνει πίσω την προβλεπόμενη τιμή
public java.lang.String toString(): Παρουσιάζει την πρόβλεψη ούτως ώστε να μπορεί να διαβαστεί

Δ. Στατιστικές Μετρήσεις Αλγορίθμων από το WEKA

Δ1. Multilayer Perceptron

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.4693
Mean absolute error	0.9284
Root mean squared error	1.264
Relative absolute error	82.1792 %
Root relative squared error	88.338 %
Total Number of Instances	35

Δ2. LAZY IBK

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.171
Mean absolute error	1.6529
Root mean squared error	2.0436
Relative absolute error	146.3011 %
Root relative squared error	142.8149 %
Total Number of Instances	35

Δ3. LAZY K-STAR

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.3507
Mean absolute error	1.1912
Root mean squared error	1.5432
Relative absolute error	105.4407 %
Root relative squared error	107.844 %
Total Number of Instances	35

Δ4. LINEAR REGRESSION

Correlation coefficient	0.5318
Mean absolute error	1.0509
Root mean squared error	1.3853
Relative absolute error	93.0236 %
Root relative squared error	96.8105 %
Total Number of Instances	35

Δ5. PLSClassifier

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.5362
Mean absolute error	1.0601
Root mean squared error	1.3833

Relative absolute error	93.8327 %
Root relative squared error	96.6714 %
Total Number of Instances	35

Ε. Κώδικας HTML & JavaScript

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>Vathmologies</title>
<script>
function validateForm()
{

	var ell=document.forms["myForm"]["ell"].value;
	if (ell==null || ell=="")
		{
			alert("Missing Value");
			return false;
		}

	var ist=document.forms["myForm"]["ist"].value;
	if (ist==null || ist=="")
		{
			alert("Missing Value");
			return false;
		}

	var agg=document.forms["myForm"]["agg"].value;
	if (agg==null || agg=="")
		{
			alert("Missing Value");
			return false;
		}

	var pli=document.forms["myForm"]["pli"].value;
	if (pli==null || pli=="")
		{
			alert("Missing Value");
			return false;
		}
```

```
var vio=document.forms["myForm"]["vio"].value;
if (vio==null || vio=="")
	{
		alert("Missing Value");
		return false;
	}

var poloik=document.forms["myForm"]["poloik"].value;
if (poloik==null || poloik=="")
	{
		alert("Missing Value");
		return false;
	}

var mathenisx=document.forms["myForm"]["mathenisx"].value;
if (mathenisx==null || mathenisx=="")
	{
		alert("Missing Value");
		return false;
	}

var texnol=document.forms["myForm"]["texnol"].value;
if (texnol==null || texnol=="")
	{
		alert("Missing Value");
		return false;
	}

var mathkk=document.forms["myForm"]["mathkk"].value;
if (mathkk==null || mathkk=="")
	{
		alert("Missing Value");
		return false;
	}

var maththk=document.forms["myForm"]["maththk"].value;
if (maththk==null || maththk=="")
	{
		alert("Missing Value");
		return false;
	}

var mathpk=document.forms["myForm"]["mathpk"].value;
if (mathpk==null || mathpk=="")
	{
		alert("Missing Value");
		return false;
	}
```

```
	var allo1=document.forms["myForm"]["allo1"].value;
	if (allo1==null || allo1=="")
		{
			alert("Missing Value");
			return false;
		}

	var allo2=document.forms["myForm"]["allo2"].value;
	if (allo2==null || allo2=="")
		{
			alert("Missing Value");
			return false;
		}

	var mo=document.forms["myForm"]["mo"].value;
	if (mo==null || mo=="")
		{
			alert("Missing Value");
			return false;
		}

	var graf=document.forms["myForm"]["graf"].value;
	if (graf==null || graf=="")
		{
			alert("Missing Value");
			return false;
		}

	var sxo=document.forms["myForm"]["sxo"].value;
	if (sxo==null || sxo=="" )
		{
			alert("Missing Value");
			return false;
		}
		if (sxo>1) {alert("Only 0-1")}


if (ell>=14 && ist>=15 && texnol>=10 && mathkk>=15 && mo>=14) {
 alert("GPA = 7");}
	if (ell<=11 && texnol<=10 && mathkk>=15 && graf>=10 && mo>=13) {
 alert("GPA = 6.3");}
	if (ell>=12 && texnol>=14.5 && mathkk>=15 && graf>=14 && mo>=13) {
 alert("GPA = 6.3");}
	if (ell>=14 && agg>=16 && pli>=13 && mathenisx<=6 && mo>=12) {
 alert("GPA = 7");}
	if (ell>=14 && ist>=13 && agg>=16 && mathkk>=18 && mo>=15) {
 alert("GPA = 7.5");}
	if (ell<=7 && vio<=8 && mathkk<=15 && graf>=18 && mo>=12) {
 alert("GPA = 6.2");}
```

```
            if (ell>=14 && pli>=14 && vio>=15 && mathkk>=18 && mo>=15) {
    alert("GPA = 7.2");}
            if (ell>=16 && pli>=17 && vio>=15 && mathenisx>=15 && mo>=16) {
    alert("GPA = 7.2");}
            if (ell<=10 && poloik<=10 && mathkk>=14 && graf>=12 && mo>=10) {
    alert("GPA = 6.3");}
            if (ell>=14.5 && ist>=10 && poloik>=10 && mathkk>=15 && mo>=13) {
    alert("GPA = 7.1");}
            if (ell>=14 && ist>=10 && agg>=15 && mathkk>=18 && mo>=14) {
    alert("GPA = 7.4");}
            if (ell<=10 && graf>=16 && maththk>=10 && allo1>=15 && allo2<=5 && mo>=11 && sxo==1)
{
    alert("GPA = 5.2");}
            if (ell<=12 && poloik<=10 && mathkk>=15 && graf>=15 && mo>=12) {
    alert("GPA = 6.4");}
            if (ell<=12 && agg>=12 && mathkk>=15 && graf>=14 && mo>=14) {
    alert("GPA = 6.7");}
            if (ell<=10 && texnol<=10 && mathpk>=10 && graf>=12 &&  allo1>=12 && mo>=11 &&
sxo==1) {
    alert("GPA = 5.4");}
            if (ell<=10 && agg>=17 && mathkk>=15 && graf>=16 && mo>=14) {
    alert("GPA = 6.7");}
            if (ell<=11 && texnol>=15 && maththk>=15 && allo1<=10 && mo<=12) {
    alert("GPA = 6");}
            if (ell<=10 && texnol>=10 && mathkk<=10 && allo1<=10 && allo2>=15 && mo<=10 &&
sxo==1) {
    alert("GPA = 5.5");}
            if (ell<=10 && pli>=18 && mathkk>=18 && graf>=15 && mo>=15) {
    alert("GPA = 6.7");}
            if (ell>=12 && agg>=14 && mathkk<=11 && graf>=16 && mo>=13) {
    alert("GPA = 6.6");}
            if (ell>=13 && mathkk>=18 && graf>=15 && allo1>=15 && mo>=15) {
    alert("GPA = 6.3");}
            if (ell<=10 && graf>=10 && maththk>=15 && allo1>=15 && allo2<=10 && mo>=10 &&
sxo==1) {
    alert("GPA = 5.1");}
            if (ell>=10 && vio<=15 && poloik<=10 && mathkk>=15 && mo>=14) {
    alert("GPA = 7");}
            if (ell<=10 && mathkk>=10 && graf<=10 && mathpk<=10 && allo1>=10 && mo<=10 &&
sxo==1) {
    alert("GPA = 5.4");}
            if (ell>=11 && agg>=10 && mathkk>=15 && graf>=15 && mo>=14) {
    alert("GPA = 6.7");}
            if (ell<=5 && texnol<=10 && mathpk>=10 && allo1>=10 && mo<=10 && sxo==1) {
    alert("GPA = 5.3");}
            if (ell>=12 && pli>=15 && mathkk>=15 && graf>=10 && mo>=15) {
    alert("GPA = 6.8");}
            if (ell>=13 && vio>=15 && poloik>=15 && mathkk>=15 && mo>=15) {
    alert("GPA = 7.2");}
```

```
            if (ell<=10 && agg>=15 && pli>=10 && vio<=10 && mathenisx<=10 && mo<=10) {
    alert("GPA = 6.7");}
            if (ell<=11 && agg>=14 && mathkk>=15 && graf>=15 && mo>=14) {
    alert("GPA = 6.7");}
            if (ell>=13 && agg>=15 && pli>=15 && mathenisx<=10 && mo>=13) {
    alert("GPA = 7.2");}
            if (ell>=15 && ist>=15 && agg>=15 && mathkk<=15 && mo>=15) {
    alert("GPA = 7.5");}
            if (ell<=12 && agg>=15 && texnol>=15 && mathkk>=18 && mo>=15) {
    alert("GPA = 6.9");}
            if (ell>=15 && vio>=15 && poloik>=15 && mathkk>=18 && mo>=17) {
    alert("GPA = 7.3");}
            if (ell<=10 && texnol>=10 && mathkk<=10 && allo1>=15 && mo<=10 && sxo==1) {
    alert("GPA = 5.3");}
         else {alert("Sorry. No Match");}

}

</script>
<style type="text/css">
body {
        background-image:url(wall1.jpg);
        background-repeat:width:100%;height:100%;
}
body,td,th {
        color:#FFF;
        font-weight: bold;
}
</style>
</head>

<body>
<form name="myForm" onsubmit="return validateForm();" method="post">
<h1>Βαθμολογίες Εξετάσεων </h1>
 <h3>Εισαγωγή βαθμολογίας: </h3>
<p>Νέα Ελληνικά:
  <input type="text" name="ell" />
<p>Ιστορία: <input type="text" name="ist" />
<p>Αγγλικά: <input type="text" name="agg" />
<p>Πληροφορική: <input type="text" name="pli" />
<p>Βιολογία: <input type="text" name="vio" />
<p>Πολιτική Οικονομία: <input type="text" name="poloik" />
<p>Μαθηματικά Ενισχυμένα: <input type="text" name="mathenisx" />
<p>Τεχνολογικό Μάθημα: <input type="text" name="texnol" />
<p>Μαθηματική Κοινού Κορμού: <input type="text" name="mathkk" />
<p>Γραφικές Τέχνες: <input type="text" name="graf" />
<p>Μαθηματικά Θ. Κ. 4ωρο Τεχνικών Σχολών: <input type="text" name="maththk" />
<p>Μαθηματικά Π. Κ. 4ωρο Τεχνικών Σχολών: <input type="text" name="mathpk" />
<p>Άλλο 1: <input type="text" name="allo1" />
```

```
<p>Άλλο 2: <input type="text" name="allo2" />
<p>Μέσος Όρος (χωρίς αναγωγή): <input type="text" name="mo" />
<p>Σχολή (Για Λύκειο 0 και για Τεχνική 1): <input type="text" name="sxo" />

<p><input type="submit" value="Submit"/>
</form>
</body>
</html>
```