ePublishing
ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ
NATIONAL DOCUMENTATION CENTRE

Συνέδρια της Ελληνικής Επιστημονικής Ένωσης Τεχνολογιών Πληροφορίας & Επικοινωνιών στην Εκπαίδευση

Τόμ. 1 (2011)

2ο Πανελλήνιο Συνέδριο: «Ένταξη και χρήση των ΤΠΕ στην Εκπαιδευτική Διαδικασία»

Προβλέποντας την επίδοση μαθητών χρησιμοποιώντας τεχνικές εξόρυξης γνώσης: Μια μελέτη περίπτωσης

Σ. Κωτσιαντής, Δ. Παπανικολάου, Π. Πιντέλας

Βιβλιογραφική αναφορά:

Κωτσιαντής Σ., Παπανικολάου Δ., & Πιντέλας Π. (2023). Προβλέποντας την επίδοση μαθητών χρησιμοποιώντας τεχνικές εξόρυξης γνώσης: Μια μελέτη περίπτωσης. *Συνέδρια της Ελληνικής Επιστημονικής Ένωσης Τεχνολογιών Πληροφορίας & Επικοινωνιών στην Εκπαίδευση*, *1*, 0435–0444. ανακτήθηκε από https://eproceedings.epublishing.ekt.gr/index.php/cetpe/article/view/4895

Προβλέποντας την επίδοση μαθητών χρησιμοποιώντας τεχνικές εξόρυξης γνώσης: Μια μελέτη περίπτωσης

Σ. Κωτσιαντής[1], Δ.Παπανικολάου[2], Π. Πιντέλας[3]

[1] Τμήμα Μαθηματικό, Πανεπιστήμιο Πατρών, Εργαστήριο Εκπαιδευτικού Λογισμικού
sotos@math.upatras.gr
[2] Δευτεροβάθμια Εκπαίδευση, 6ο Γυμνάσιο Ρόδου dopap@sch.gr
[3] Τμήμα Μαθηματικό, Πανεπιστήμιο Πατρών, Εργαστήριο Εκπαιδευτικού Λογισμικού
pintelas@upatras.gr

Περίληψη

Σε αυτήν την εργασία μελετάμε με ποιον τρόπο θα μπορέσουμε να εφαρμόσουμε τεχνικές εξόρυξης γνώσης σε εκπαιδευτικά δεδομένα προκειμένου να επιτύχουμε την πρόγνωση της επίδοσης ενός μαθητή. Κύριος στόχος μας είναι ο έγκαιρος εντοπισμός των αδύνατων μαθητών ώστε να μπορέσουμε να τους προσφέρουμε πρόσθετη στήριξη με σκοπό να βελτιώσουν την επίδοση τους και τα μην αποτύχουν στις τελικές εξετάσεις του μαθήματος. Χρησιμοποιήσαμε μερικούς από τους κυριότερους αλγόριθμους κατηγοριοποίησης όπως τα δέντρα απόφασης, τους αλγόριθμους κατασκευής κανόνων, τα τεχνητά νευρωνικά δίκτυα, τις μηχανές διανυσμάτων υποστήριξης, τους κ-κοντινότερους γείτονες και τέλος τους αλγορίθμους στατιστικής ταξινόμησης. Τα δεδομένα μας προέρχονταν από το μάθημα της Γεωγραφίας Α' και Β΄ Γυμνασίου. Αφού δημιουργήσαμε ένα μοντέλο πρόγνωσης αξιολογήσαμε διάφορες περιπτώσεις κατηγοριοποίησης κάνοντας χρήση των παραπάνω αλγορίθμων. Τέλος κατασκευάσαμε ένα εργαλείο το οποίο μπορεί να χρησιμοποιηθεί για πρόγνωση με πολύ μεγάλη ακρίβεια.

Λέξεις κλειδιά: *εξόρυξη γνώσης, πρόγνωση επίδοσης.*

1. Εισαγωγή

Οι τεχνικές εξόρυξης γνώσης είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων ώστε να βρεθούν μη παραχωρηθείσες σχέσεις αυτών των δεδομένων αλλά και να παρουσιαστούν οι σχέσεις αυτές με κατανοητό τρόπο (Hand et al., 2001). Οι τεχνικές εξόρυξης γνώσης χρησιμοποιήθηκαν αρχικά σε διάφορους τομείς της επιστήμης όπως οι λιανικές πωλήσεις, η οικονομική ανάλυση, η βιοπληροφορική κ.λ.π. Τα τελευταία χρόνια όμως υπάρχει ένα όλο και αυξανόμενο ενδιαφέρον για την χρήση των τεχνικών της εξόρυξης γνώσης προκειμένου να ερευνήσουμε επιστημονικά ερωτήματα που αφορούν την εκπαίδευση (Romero & Ventura, 2007). Αυτός ο επιστημονικός τομέας αποκαλείται εξόρυξη γνώσης για εκπαιδευτικούς σκοπούς (Educational Data Mining).

Τα τελευταία χρόνια έχουν πραγματοποιηθεί πολλές έρευνες όσο αφόρα την χρήση τεχνικών μηχανικής μάθησης ή εξόρυξης γνώσης στον τομέα της εκπαίδευσης με

σκοπό την βελτίωση της ποιότητας της εκπαίδευσης και την ενίσχυση της διαδικασίας της μάθησης. Σε αυτή την εργασία μελετήσαμε την χρήση διαφόρων τεχνικών εξόρυξης γνώσης προκειμένου να προβλέψουμε την επίδοση μαθητών δευτεροβάθμιας εκπαίδευσης. Συγκεκριμένα πραγματοποιήσαμε σύγκριση διαφόρων αλγορίθμων όσο αφορά την ακρίβεια τους στην πρόγνωση της επιτυχίας ή αποτυχίας μαθητών δευτεροβάθμιας εκπαίδευσης στο μάθημα της Γεωγραφίας. Τα δεδομένα μας αφορούσαν το μάθημα της Γεωγραφίας της Α' Γυμνάσιου, καθώς και το μάθημα της Γεωγραφίας της Β' Γυμνασίου. Σκοπός μας είναι να διαπιστώσουμε ποιος αλγόριθμος εξόρυξης γνώσης δίνει τα καλύτερα αποτελέσματα όσο αφορά την ακρίβεια πρόγνωσης της επιτυχίας/αποτυχίας των μαθητών στο μάθημα, της κατηγοριοποίησης των μαθητών σε «Κακούς», «Καλούς», «Πολύ Καλούς». Στην συνέχεια κατασκευάσαμε μια εφαρμογή η οποία υλοποιώντας τον αλγόριθμο ο οποίος έδωσε τα καλύτερα αποτελέσματα στα πειράματα μας πραγματοποιεί την κατηγοριοποίηση των μαθητών στις ποιο πάνω κλάσεις.

Χρησιμοποιήσαμε αντιπροσωπευτικούς αλγορίθμους για κάθε μια από τις ποιο γνωστές τεχνικές μηχανικές μάθησης : τα δέντρα απόφασης (Murthy, 1998), τους αλγόριθμους κατασκευής κανόνων (Furnkranz, 1999), τα τεχνητά νευρωνικά δίκτυα (Zhang, 2000), μηχανές διανυσμάτων υποστήριξης (Burges, 1998), κ-κοντινότερων γειτόνων (Aha, 1997) και τέλος τους αλγορίθμους στατιστικής ταξινόμησης (Jensen, 1996).

2. Μεθοδολογία

Όπως αναφέρθηκε και πρωτύτερα κύριος στόχος μας είναι να δοκιμάσουμε διάφορους αλγόριθμους εξόρυξης γνώσης σε μια προσπάθεια να μελετήσουμε τις επιδόσεις μαθητών δευτεροβάθμιας εκπαίδευσης στο μάθημα της Γεωγραφίας Α' και Β' Γυμνασίου. Η μεθοδολογία που θα εξακολουθήσουμε αποτελείται από τα εξής βήματα. :

- Συλλογή των δεδομένων μας
- Προετοιμασία των δεδομένων
- Κατασκευή των μοντέλων κατηγοριοποίησης
- Αξιολόγηση των μοντέλου
- Χρησιμοποίηση του καλύτερου προβλεπτικού μοντέλου για πρόγνωση της επίδοσης νέων μαθητών

2.1 Τα Δεδομένα μας / Προετοιμασία δεδομένων

Στην εργασία αυτή χρησιμοποιήθηκαν δεδομένα από την διδασκαλία του μαθήματος της Γεωγραφίας Α' και Β' Γυμνασίου στο Γυμνάσιο της Μεσαριάς στην Σαντορίνη κατά το χρονικό διάστημα 2003-2006. Συγκεκριμένα το μάθημα της Γεωγραφίας Α' Γυμνασίου αφορά τα έτη 2003-2004, 2004-2005, 2005-2006 και αποτελείται από 192 στιγμιότυπα, ενώ το μάθημα της Γεωγραφίας της Β' Γυμνασίου αφορά τα έτη 2003-2004, 2004-2005 και αποτελείται από 115 στιγμιότυπα άρα στο σύνολο έχουμε 307 εγγραφές στην βάση δεδομένων.

Τα δεδομένα αφορούσαν διάφορα στοιχεία των μαθητών όπως την επίδοση τους και την επιμέλεια τους στο μάθημα. Π.χ. τις βαθμολογίες τους στο διαγώνισμα τριμήνου, τον προφορικό τους βαθμό, αν παρέδωσαν ή όχι τις εργασίες τους στο μάθημα κ.λ.π. Τα δεδομένα αυτά συγκεντρώθηκαν κατά την διάρκεια της διδασκαλίας αυτών των μαθημάτων. Προκειμένου να μπορέσουμε να χρησιμοποιήσουμε τα στοιχεία μας κάναμε μια προετοιμασία στα δεδομένα μας. Στόχος αυτής της φάσης είναι να δεδομένα που έχουμε συλλέξει να τα φέρουμε σε μια τέτοια μορφή η οποία να είναι κατάλληλη ώστε να επιτρέπει την εφαρμογή αλγορίθμων εξόρυξης γνώσης.

Η επιλογή των κατάλληλων χαρακτηριστικών είναι μια δύσκολη διαδικασία. Μια τεχνική επιλογής των κατάλληλων χαρακτηριστικών προτείνουν οι (Ramaswami & Bhaskaran, 2009). Τα χαρακτηριστικά τα οποία τελικά επιλέχτηκαν για την υλοποίηση του μοντέλου μας φαίνονται στον (Πίνακας 1).

Πίνακας 1: Οι μεταβλητές του προτεινόμενου μοντέλου

Χαρακτηριστικό	Περιγραφή	Πεδίο τιμών
DIAG_A_TRIM	Ο Βαθμός του γραπτού διαγωνίσματος Α τριμήνου	Αριθμητική τιμή : (0-20)
1_TEST_A_TRIM	Ο βαθμός στο πρώτο Τεστ του Α τριμήνου	Αριθμητική τιμή : (0-10)
2_TEST_A_TRIM	Ο βαθμός στο δεύτερο Τεστ του Α τριμήνου	Αριθμητική τιμή : (0-10)
ASK_SPIT_A_TRIM	Η βαθμολογία των ασκήσεως για το σπίτι του Α τριμήνου	Αριθμητική τιμή : (0-5)
1_PR_A_TRIM	Βαθμολογία πρώτης προφορικής εξέτασης για το Α τρίμηνο	Αριθμητική τιμή : (0-10)
2_PR_A_TRIM	Βαθμολογία δεύτερης προφορικής εξέτασης για το Α τρίμηνο	Αριθμητική τιμή : (0-10)
BAT_A_TRIM	Η βαθμολογία του μαθητή στο Α τρίμηνο	Αριθμητική τιμή : (1-20)
DIAG_B_TRIM	Ο Βαθμός του γραπτού διαγωνίσματος Β τριμήνου	Αριθμητική τιμή : (0-20)
1_TEST_B_TRIM	Ο βαθμός στο πρώτο Τεστ του Β τριμήνου	Αριθμητική τιμή : (0-10)
ASK_SPIT_B_TRIM	Η βαθμολογία των ασκήσεως για το σπίτι του Β τριμήνου	Αριθμητική τιμή : (0-5)

PR_B_TRIM	Βαθμολογία πρώτης προφορικής εξέτασης για το Β τρίμηνο	Αριθμητική τιμή : (0-10)
BAT_B_TRIM	Η βαθμολογία του μαθητή στο Β τρίμηνο	Αριθμητική τιμή : (1-20)
TEST_C_TRIM	Ο βαθμός στο πρώτο Τεστ του Γ τριμήνου	Αριθμητική τιμή : (0-10)
PR_C_TRIM	Βαθμολογία πρώτης προφορικής εξέτασης για το Γ τρίμηνο	Αριθμητική τιμή : (0-10)
BAT_C_TRIM	Η βαθμολογία του μαθητή στο Γ τρίμηνο	Αριθμητική τιμή : (1-20)
TEL_DIAG	Ο βαθμός του μαθητή στην τελική εξέταση του Ιουνίου	Αριθμητική τιμή : (1-20)
PASS	Πέρασε/Κόπηκε ο μαθητής στο μάθημα	Δυαδική τιμή (Pass/Fail)
3-LEVEL CLASSIFICATION	Χαρακτηρισμός των μαθητών ως Αποτυχόντες, Καλούς, Πολύ Καλούς.	Διακριτές τιμές (“Fail/Good/Very Good”)

Οι δυο τελευταίες γραμμές αφορούν εκείνα τα χαρακτηριστικά βάση των οποίων θα κάνουμε την κατηγοριοποίηση των μαθητών και εκείνα τα οποία θα χρησιμοποιήσουμε για την πρόγνωση.

Για το κάθε μάθημα (Γεωγραφία Α’ Γυμνασίου και Γεωγραφία Β’ Γυμνασίου) θα αναλυθούν τρία διαφορετικά σύνολα δεδομένων τα οποία προέρχονται από την αρχική μας βάση δεδομένων και έχουν προκύψει κατά την φάση της προεπεξεργασίας. Τα τρία αυτά σύνολα είναι τα παρακάτω :

- Α : Περιέχει εκείνα τα χαρακτηριστικά από τον παραπάνω πίνακα που αφορούν την επίδοση του μαθητή στο Α τρίμηνο.
- Β : Περιέχει όλα τα χαρακτηριστικά του συνόλου Α και τα στοιχεία εκείνα που αφορούν την επίδοση ενός μαθητή στο Β τρίμηνο.
- C : Περιέχει όλα τα χαρακτηριστικά του συνόλου Β και τα στοιχεία εκείνα που αφορούν την επίδοση ενός μαθητή στο Γ τρίμηνο (Αυτό δηλαδή είναι το συνολικό σύνολο που βλέπουμε στον παραπάνω πίνακα)

Επίσης θα χρησιμοποιήσουμε δυο προσεγγίσεις όσο αφορά την κατηγοριοποίηση που αφορά την πρόβλεψη της μεταβλητής εξόδου :

- Δυαδική κατηγοριοποίηση (Επιτυχών/Αποτυχών) (*Fail/Pass*)
- Κατηγοριοποίηση 3-επιπεδων *Fail* (0-37), *Good* (38-67), *Very Good* (68-80)

2.2 Εφαρμογή των μοντέλων κατηγοριοποίησης

Προκειμένου να εξετάσουμε την αποδοτικότητα των τεχνικών εξόρυξης γνώσης στο πεδίο της εκπαίδευσης χρησιμοποιήθηκαν οι ποιο διαδεδομένες τεχνικές:

- Δέντρα Απόφασης (Decision Trees)
- Τεχνητά Νευρωνικά Δίκτυα (Neural Networks)
- Αλγόριθμοι Στατιστικής Κατηγοριοποίησης (Naïve Bayes)
- Μάθηση Βασισμένη Σε Στιγμιότυπα (Instance-Based Learning)
- Κανόνες Ταξινόμησης (Rule-based Classification)
- Μηχανές Διανυσμάτων Υποστήριξης (SVM Support Vector Machines)

Για τους σκοπούς αυτής της εργασίας ένας αντιπροσωπευτικός αλγόριθμος από κάθε τεχνική χρησιμοποιήθηκε. Ο αλγόριθμος C4.5 (Quinlan, 1993) ο οποίος είναι ίσως ο πιο διαδεδομένος αλγόριθμος εξόρυξης γνώσης χρησιμοποιήθηκε για τα δέντρα απόφασης. Για να υπολογίσουμε τις τιμές και τα βάρη σε ένα νευρωνικό δίκτυο χρησιμοποιήσαμε τον αλγόριθμο Back Propagation (BP). Ο αφελής ταξινομητής Bayes (Naïve Bayes NB) (Domingos & Pazzani, 1997) χρησιμοποιήθηκε για την αντιπροσώπευση των αλγορίθμων στατιστικής κατηγοριοποίησης. Ο 3-KNN ο οποίος συνδυάζει την πολύ καλή απόδοση με την υψηλή ταχύτητα χρησιμοποιήθηκε από την κατηγορία των αλγορίθμων που παρέχουν μάθηση βασισμένη σε στιγμιότυπα. Από την κατηγορία των αλγορίθμων με κανόνες ταξινόμησης χρησιμοποιήσαμε τον Ripper(Cohen, 1995), ενώ τέλος ο Sequential Minimal Optimazation (SMO) επιλέχτηκε από την κατηγορία των μηχανών διανυσμάτων υποστήριξης (SVM) Support Vector Machines. Για την υλοποίηση των πειραμάτων μας και την υπολογισμό των διαφόρων μετρήσεων ακρίβειας χρησιμοποιήσαμε το εργαλείο Weka το οποίο εχει υλοποιημένους αυτούς τους αλγορίθμους σε ελεύθερα διαθέσιμο κώδικα (Hall et al, 2009).

2.3 Πειραματικές Μετρήσεις και αξιολόγηση του μοντέλου

Σε αυτή την φάση της εργασίας πραγματοποιήθηκαν τα πειράματα και έγιναν διάφορες μετρήσεις. Κατασκευάστηκαν τα προγνωστικά μοντέλα για το μάθημα της Γεωγραφίας Α' Γυμνασίου, και Γεωγραφίας Β' Γυμνασίου για την περίπτωση της κατηγοριοποίησης δυο τιμών *Pass/Fail* για κάθε ένα από τα τρία σύνολα A,B,C που είδαμε παραπάνω. Ενώ στην συνέχεια έγινε η ίδια εργασία για τα ιδία μαθήματα και για τα ίδια σύνολα A,B,C για κατηγοριοποίηση τριών επίπεδων *Fail* / *Good* / *Very Good*.

Σκοπός μας ήταν να διαπιστώσουμε ποιος αλγόριθμος δίνει καλύτερα αποτελέσματα όσο αφορά την ακρίβεια της πρόγνωσης. Η ποιο διαδεδομένη τεχνική προκειμένου να διαπιστώσουμε την ακρίβεια πρόγνωσης είναι η τεχνική της διασταυρωμένης επικύρωσης (Cross-Validation) (Qasem , A. & AL-Radaideh, Q. 2008). Εμείς για να κάνουμε τις μετρήσεις για την ακρίβεια των προγνωστικών μοντέλων χρησιμοποιήσαμε 10-φορές διασταυρωμένη επικύρωση.

Στην περίπτωση των αλγορίθμων κατηγοριοποίησης τα μοντέλα αξιολογούνται

χρησιμοποιώντας το Ποσοστό Σωστών Κατηγοριοποιήσεων που επιτυχαίνει το μοντέλο.

Ένα άλλο χαρακτηριστικό που είναι χρήσιμο να χρησιμοποιήσουμε για την αξιολόγηση των αλγορίθμων που μελετάμε είναι η μήτρα σύγχυσης (confusion matrix) ή αλλιώς πίνακας ενδεχομένων ο οποίος συνοψίζει τα αποτελέσματα μετά τη εξέταση του συνόλου δοκιμής στον αλγόριθμο. Στην διαγώνιό της παρουσιάζονται τα πρότυπα που έχουν ταξινομηθεί σωστά ανά κλάση και στις υπόλοιπες θέσεις τα πρότυπα που έχουν ταξινομηθεί λανθασμένα.

Τόσο η κατασκευή των μοντέλων όσο και οι μετρήσεις τις ακρίβειας έγιναν με χρήση του ελεύθερου λογισμικού Weka. Για κάθε μάθημα και για κάθε αλγόριθμο αξιολογήσαμε τρία διαφορετικά μοντέλα τα Α, Β, C. Το Α όπως έχουμε προαναφέρει περιέχει τους βαθμούς μόνο για το Α τρίμηνο, το Β για το Α και το Β τρίμηνο, ενώ το C για όλα τα τρίμηνα. Εμείς θέλουμε να διαπιστώσουμε ποιο από τα τρία εχει καλύτερη απόδοση αλλά κυρίως ποιος από τους αλγόριθμους δίνει καλύτερη απόδοση για το Α μοντέλο. Ενδεικτικά δίνονται στον (Πίνακας 2) οι υπολογισμένες ακρίβειες πρόγνωσης για το μάθημα της Γεωγραφίας Α' Γυμνασίου για όλους τους αλγορίθμους για όλα τα σύνολα δεδομένων που δίνουμε για την εκπαίδευση του μοντέλου.

Πίνακας 2: Ακρίβεια των αλγορίθμων για κάθε σύνολο Α, Β, C, και Μέσος Όρος

	Γεωγραφία Α' Γυμνασίου (*Pass/Fail*)					
	C4.5 (J48)	**RIPPER (JRip)**	**BP**	**SMO**	**3-NN**	**Naïve Bayes**
A	92,70	92,70	93,23	93,75	95,31◇	92,70
B	93,23	92,70	95,83●	94,27	94,79	94,27
C	96,35 ●	97,91 ●	94,27	94,79 ●	95,83 ●	94,27 ●
Μέσος Όρος	94,09	94,43	94,44	94,27	95,31 ●	93,74

Η δεύτερη κατηγοριοποίηση που κάναμε για την μεταβλητή εξόδου είναι σε 3-επιπεδα (*Fail/Good/Very Good*). Οι μετρήσεις τις αποδοτικότητας πρόγνωσης των αλγόριθμων που δοκιμάσαμε σε αυτή την περίπτωση για τα τρία σύνολα Α,Β,C για το μάθημα της Γεωγραφίας Α' Γυμνασίου φαίνονται παρακάτω (Πίνακα 3). Στους πίνακες 2 και 3 η μαύρη τελεία (●) επισημαίνει την καλύτερη επίδοση για κάθε αλγόριθμο ενώ με (◇) επισημαίνουμε την καλύτερη ακρίβεια συνολικά.

Πίνακας 3: Ακρίβεια των αλγορίθμων για κάθε σύνολο Α, Β, C, και Μέσος Όρος

Γεωγραφία Α' Γυμνασίου κατηγοριοποίηση 3-επιπέδων

	C4.5 (J48)	RIPPER (JRip)	BP	SMO	3-NN	Naïve Bayes
A	94,27	94,79●	95,31●	98,43●◇	96,35	94,79
B	94,79	93,75	95,31	97,91	96,87●	95,83●
C	94,79●	93,75	86,97	97,91	96,35	94,79
Μέσος Όρος	94,61	94,09	92,53	98,08●	96,52	95,13

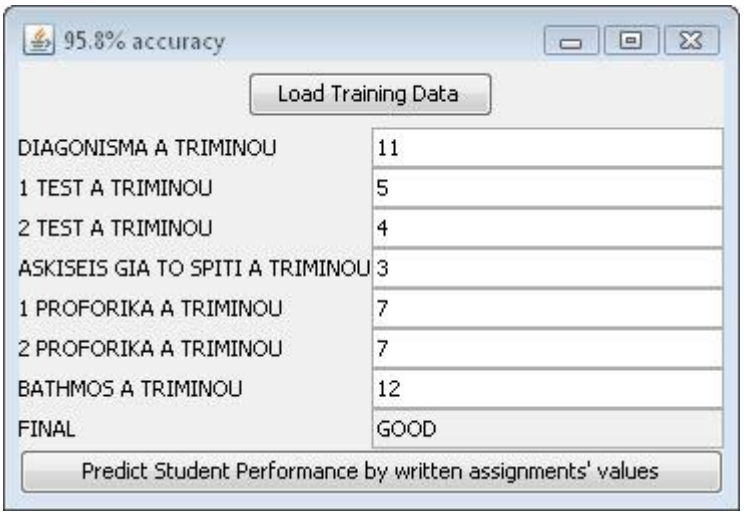
Παρόμοια αποτελέσματα δίνουν και οι μετρήσεις για την Β' Γυμνάσιου Με τον αλγόριθμο 3-NN να μας παρέχει την καλύτερη ακρίβεια πρόγνωσης.

2.4 Κατασκευή Εργαλείου πρόγνωσης

Στην συνέχεια κατασκευάσαμε ένα εργαλείο προκειμένου να κάνουμε πρόγνωση της τελικής επίδοσης ενός μαθητή. Για την κατασκευή αυτού του εργαλείου έγινε χρήση του περιβάλλοντος ανάπτυξης λογισμικού Eclipse το οποίο είναι ένα ελεύθερο λογισμικό που μας επιτρέπει να αναπτύξουμε λογισμικό σε διάφορες γλώσσες προγραμματισμού όπως Java, C, C++, PHP, Python, COBOL, Perl κ.λ.π. Εμείς χρησιμοποιήσαμε το Eclipse για την σχεδίαση του περιβάλλοντος διεπαφής χρήστη και την συγγραφή του κώδικα σε Java .

Το εργαλείο διαβάζει τα δεδομένα εκπαίδευσης καθώς και τα χαρακτηριστικά του μοντέλου μας και τα παρουσιάζει στην οθόνη (στην δική μας περίπτωση οι βαθμοί του Α' τριμήνου). Στο τέλος τοποθετεί την μεταβλητή, την πρόβλεψη της οποίας επιθυμούμε, δηλαδή το αν επέτυχε ή όχι ο μαθητής στην δυαδική κατηγοριοποίηση ή (Fail/Good/Very Good) στην κατηγοριοποίηση 3-επιπεδων. Στην συνέχεια το λογισμικό μας δημιουργεί το μοντέλο πρόγνωσης κάνοντας χρήση του αλγορίθμου 3-NN τον οποίο και αποφασίσαμε να χρησιμοποιήσουμε από την ανάλυση τον μετρήσεων που κάναμε για την ακρίβεια των αλγορίθμων κατηγοριοποίησης. Ο αλγόριθμος εχει ενσωματωθεί στο λογισμικό μας από το Weka χρησιμοποιώντας το Weka API.

Σε πρώτη φάση το εργαλείο μας (Εικόνα 1) ζητά το φορτώσει τα δεδομένα εκπαίδευσης για να κατασκευάσει το μοντέλο, βάση του οποίου θα κάνει την πρόγνωση της επίδοσης των μαθητών. Δηλαδή περιμένει να του φορτώσουμε ένα αρχείο από όπου θα διαβάσει τα χαρακτηριστικά αλλά και τις τιμές των στιγμιότυπων ώστε να προχωρήσει στην δημιουργία του μοντέλου μας.



Εικόνα 1 : Οθόνη του εργαλείου που κάνει την πρόγνωση

3. Συμπεράσματα

Σε αυτήν την εργασία έγινε μια προσπάθεια να προβλέψουμε την επίδοση μαθητών δευτεροβάθμιας εκπαίδευσης στο μάθημα της Γεωγραφίας (Γεωγραφία Α' Γυμνασίου και Γεωγραφία Β' Γυμνασίου), χρησιμοποιώντας τους βαθμούς των μαθητών στο διαγώνισμα του Α τριμήνου καθώς και άλλα στοιχεία που αφορούν την επίδοση τους στο μάθημα. Χρησιμοποιήθηκαν δυο διαφορετικοί στόχοι πρόβλεψης (δυαδική κατηγοριοποίηση *Fail/Pass* και κατηγοριοποίηση 3-επιπεδων *Fail/Good/Very Good*), και έξι διαφορετικοί μέθοδοι εξόρυξης γνώσης. Συγκεκριμένα εξετάσαμε αλγόριθμους από τα δέντρα απόφασης, από τους αλγορίθμους με κανόνες ταξινόμησης, μέθοδοι Bayes, μέθοδος πλησιέστερων γειτόνων, τεχνητά νευρωνικά δίκτυα και μηχανές διανυσμάτων υποστήριξης. Επίσης μελετήσαμε τρία διαφορετικά σύνολα εκπαίδευσης αυτών των αλγορίθμων με δεδομένα από το πρώτο, δεύτερο και τρίτο τρίμηνο αντίστοιχα. Τα λαμβανόμενα αποτελέσματα έδειξαν ότι είναι δυνατόν να επιτύχουμε υψηλή ακρίβεια πρόγνωσης(>90%) ακόμα και από το πρώτο τρίμηνο, βγάζοντας έτσι το συμπέρασμα ότι η βαθμολογίες των μαθητών εξαρτώνται από τις προηγούμενες επιδόσεις τους με ισχυρό τρόπο.

Οι αλγόριθμοι που χρησιμοποιήσαμε ανήκουν στην κατηγορία της κατηγοριοποίησης (classification). Η ακρίβεια πρόγνωσης που μας έδωσαν τα προγνωστικά μοντέλα ήταν μεγάλη. Οι περισσότεροι έδωσαν ακρίβεια πάνω από 90% στην πρόγνωση της κλάσης του μαθητή. Όμως η ποιοτική και ποσοτική σύγκριση που κάναμε για τους έξι αλγορίθμους που μελετήσαμε έδειξε ότι ο ποιο κατάλληλος είναι ο 3-NN. Γι αυτό το λόγο στο εργαλείο λογισμικού που υλοποιήσαμε, επιλέξαμε να χρησιμοποιήσουμε τον 3-NN για την δημιουργία του προγνωστικού μοντέλου. Το εργαλείο αυτό μπορεί να το χρησιμοποιήσει ένας εκπαιδευτικός προκειμένου να κατηγοριοποίηση τους μαθητές μετά το πέρας του πρώτου τριμήνου.

Έτσι με την βοήθεια των τεχνικών μηχανικής μάθησης και την χρήση ενός προγνωστικού λογισμικού εργαλείου σαν αυτό που κατασκευάσαμε στα πλαίσια αυτής της εργασίας οι εκπαιδευτές είναι σε θέση να εντοπίσουν γρήγορα και με μεγάλη ακρίβεια εκείνους τους μαθητές, οι οποίοι δεν θα ανταποκριθούν στις τελικές εξετάσεις στο μάθημα. Ο γρήγορος εντοπισμός θα πρέπει να εχει σαν συνέπεια είτε την παροχή ενισχυτικής διδασκαλίας προς τον μαθητή, είτε την δημιουργία ειδικού υλικού προς αυτόν, προκειμένου να βοηθήσουμε τον μαθητή να αυξήσει τις επιδόσεις του.

Για να μπορεί να είναι χρήσιμο ένα τέτοιο εργαλείο θα πρέπει να απαιτεί από τον χρήση την λιγότερο δυνατή γνώση αυτών των τεχνικών εξόρυξη γνώσης. Για να επιτευχτεί κάτι τέτοιο θα πρέπει να υπάρξει ένας σχεδιασμός και μια υλοποίηση ενός λογισμικού το όποιο αυτόματα θα είναι σε θέση από τα δεδομένα των μαθητών να δημιουργεί τα μοντέλα πρόγνωσης ώστε από τον εκπαιδευτικό να ζητείται απλά η εισαγωγή των στοιχείων που απαιτούνται για την κατηγοριοποίηση του μαθητή.

Αυτό θα μπορούσε να υλοποιηθεί, αφού πλέον σε όλα τα σχολεία της δευτεροβάθμιας εκπαίδευσης από την σχολική χρονιά 2010-2011 θα είναι υποχρεωτική η καταγραφή των βαθμών αλλά και πολλών άλλων στοιχείων των μαθητών(απουσίες, δημογραφικά στοιχεία κ.λ.π) σε μια web εφαρμογή η οποία αποτελεί μια βάση δεδομένων (e-school). Έτσι σε λίγα χρόνια θα υπάρχει ένας πολύ μεγάλος όγκος δεδομένων από τα οποία θα είναι πλέον εύκολο να δημιουργηθούν μοντέλα πρόγνωσης αυτόματα χωρίς να απαιτείται η εισαγωγή στοιχείων από τον εκπαιδευτικό με χρήση φυσικά κάποιου ειδικού εργαλείου. Μια μελλοντική εργασία λοιπόν θα μπορούσε, κάνοντας χρήση της ποιο πάνω εφαρμογής (e-school), να μελετήσει την συμπεριφορά προγνωστικών μοντέλων τα οποία λαμβάνουν υπόψη τους και τέτοια στοιχεία.

Μια μελλοντική βελτίωση ενός τέτοιου εργαλείου πρόγνωσης θα μπορούσε να περιλαμβάνει και διάφορα δημογραφικά στοιχεία των μαθητών (πόσα αδέλφια εχει ο μαθητής, αν μένει με τους δυο γονείς του με έναν ή στην γιαγιά του, αν μιλαει τα ελληνικά καλά, μέτρια ή κακά, αν κάνει φροντιστήρια κ.λ.π.) αλλά και δεδομένα κοινωνικής φύσεως (αν συναναστρέφεται με άλλα παιδιά, η κατανάλωση αλκοόλ στο σπίτι του κ.λ.π.).

Βιβλιογραφία

Aha, D. (1997). Lazy Learning. Dordrecht: Kluwer Academic Publishers

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):1–47

Cohen, W. (1995). Fast Effective Rule Induction. Proceeding of International Conference on Machine Learning 1995, pp. 115-123.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. Vol. 29, pp. 103-130.

Furnkranz, J. (1999). Separate-and-Conquer Rule Learning. Artificial Intelligence Review, Vol. 13, pp. 3-54

Hand, D. Mannila, H., & Smyth, P. (2001). Principles of Data Mining. MIT Press, Cambridge, MA.

Hall, M. Frank, E. Holmes, G. Pfahringer, B. Reutemann, P. Witten, I. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations. Volume 11. Issue 1

Jensen, F. (1996). An Introduction to Bayesian Networks, Springer.

Murthy, S. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery, Vol. 2, pp. 345–389.

Qasem A. & AL-Radaideh, Q. (2008). The Impact Of Classification Evaluation Methods on Rough Set Based Clasifiers. Proceedings of the 2008 International Arab Conference on Information Technology (ACIT2008). Dec 2008. University of Sfax, TunisiaQuinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco

Ramaswami, M. & Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. Journal of computing, Volume 1, Issue 1, December 2009, ISSN: 2151-9617Romero, C. & Ventura, S. (2007). *Educational data Mining: A Survey from 1995 to 2005*. Expert Systems with Applications (33) 135-146.

Zhang, G. (2000). Neural networks for classification: a survey. IEEE Trans Syst Man Cy C 30(4):451–462