ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:
«ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΩΝ ΣΕ ΒΙΒΛΙΟΘΗΚΕΣ, ΑΡΧΕΙΑ, ΜΟΥΣΕΙΑ»
ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΤΙΚΩΝ, ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διαχείριση Δεδομένων στις πλατφόρμες KNIME & WEKA

Παναγιώτα Κωτσάκη (ΑΜ: 186682006)

Επιβλέπων: Ιωάννης Τριανταφύλλου

Στόχος και Σκοπός διπλωματικής

- Σύγκριση λογισμικών εξόρυξης γνώσης μέσω της τεχνικής παλινδρόμησης (regression) σε θεωρητικό αλλά και σε πειραματικό πλαίσιο, με στόχο την εύρεση ενός μοντέλου πρόβλεψης της διάρκειας ψηφιοποίησης αρχειακού υλικού
  - Λογισμικά εξόρυξης γνώσης WEKA (knowledge flow) και KNIME
  - Αλγόριθμοι KNN, SVM, Random Forest, Decision Tree και Linear Regression
  - Το δείγμα προήλθε από την εταιρεία Αρχειοθήκη Α.Ε. (φυλάσσει και διαχειρίζεται αρχειακό υλικό)
  - Το δείγμα αφορά (6.322) εγγραφές (συγκεκριμένα 6.087 φακέλους) προς ψηφιοποίηση
- Σκοπός της σύγκρισης αυτής είναι να παραχθεί ένα μοντέλο πρόβλεψης της διάρκειας ολοκλήρωσης μίας εργασίας ψηφιοποίησης ενός φακέλου με βάση χαρακτηριστικά όπως ο χρήστης, ο αριθμός των σελίδων, το οποίο θα παράσχει χρήσιμες πληροφορίες στην εταιρεία Αρχειοθήκη Α.Ε.

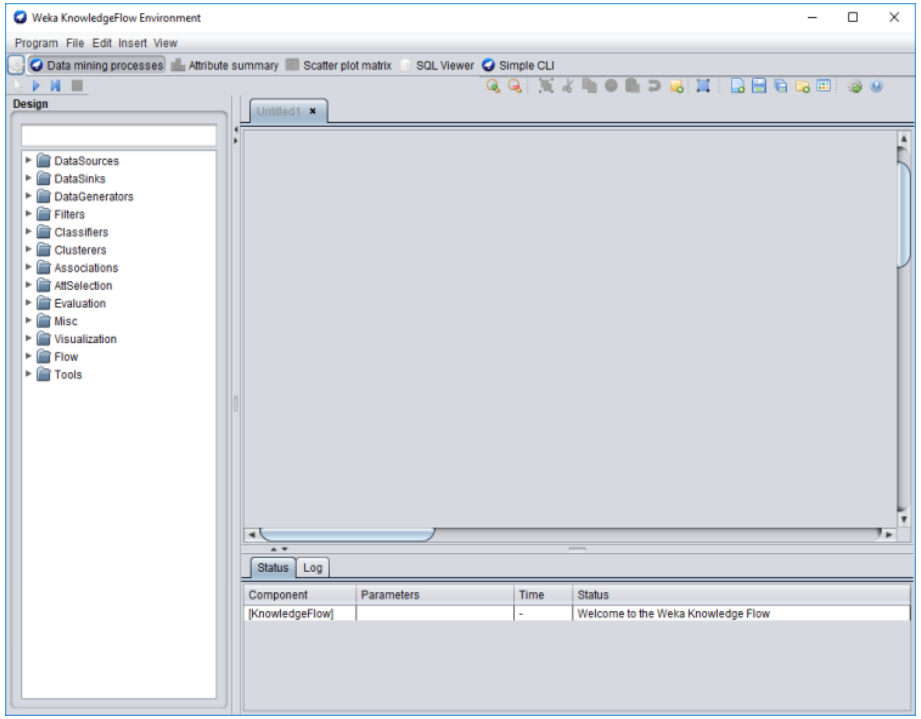
Θεωρητική προσέγγιση της έρευνας

- Εισαγωγικές έννοιες
  - Τι είναι δεδομένα και πληροφορία
  - Ανακάλυψη γνώσης από δεδομένα και εξόρυξη δεδομένων
  - Πρακτικές εφαρμογές
- Εξόρυξη Γνώσης και Μηχανική Μάθηση
  - Οι δύο κεντρικοί τύποι προβλημάτων πρόβλεψης: ταξινόμηση (classification) και παλινδρόμηση (regression)
  - Μοντέλα Πρόβλεψης {Δέντρα αποφάσεων (Decision tress), Νευρωνικά Δίκτυα (Neural networks), Διανυσματικές μηχανές (Support Vector Machines), Random Forest, Αλγόριθμος κ πλησιέστερου γείτονα (KNN), Γραμμική παλινδρόμηση (Linear Regression)}
  - Περιγραφικά Μοντέλα { Συσταδοποίηση (Clustering), Κανόνες συσχέτισης (Association Rules)}
- Προεπεξεργασία Δεδομένων (Σημασία προεπεξεργασίας δεδομένων, Στάδια προεπεξεργασίας, Επιλογή χαρακτηριστικών (attributes)

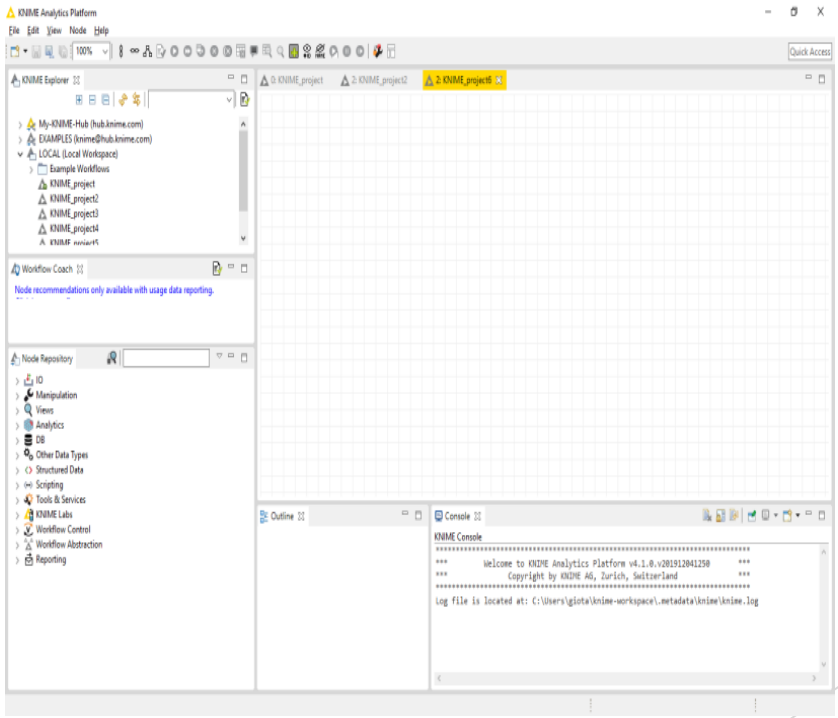
Παρουσίαση και Σύγκριση των εργαλείων γνώσης

- **Περιβάλλον WEKA (**KnowledgeFlow)
  - Waikato Environment for Knowledge Analysis
  - εργαλείο ανοιχτού κώδικα
  - διατίθεται δωρεάν
  - βασίζεται σε γλώσσα java
  - υλοποιημένες μεθόδοι για προεπεξεργασία δεδομένων (data preprocessing), ταξινόμηση (classification) και παλινδρόμηση (regression), συσταδοποίηση (clustering), εύρεση κανόνων συσχέτισης (association rules), επιλογή χαρακτηριστικών (attribute selection) και οπτικοποίηση (visualization)
- **Περιβάλλον KNIME**
  - Konstanz Information Miner
  - πλατφόρμα ανάλυσης δεδομένων ανοιχτού κώδικα
  - ενσωμάτωση νέων αλγορίθμων και μεθόδων επεξεργασίας δεδομένων
  - διαδραστική εκτέλεση μίας ροής δεδομένων και οπτική αναπαράσταση
  - πολλά διαφορετικά λειτουργικά συστήματα (32-bit, 64-bit) Windows, Linux και MAC
  - βασίζεται στην πλατφόρμα ανοιχτού κώδικα Eclipse και την Java

Περιβάλλον WEKA (KnowledgeFlow)



Περιβάλλον KNIME



Συγκεντρωτικός πίνακας χαρακτηριστικών KNIME & WEKA (1/2)

	WEKA Knowledge Flow	KNIME
Website	https://www.cs.waikato.ac.nz/ml/weka/	https://www.knime.com/
Έτος κυκλοφορίας	1992	2006
Τελευταία έκδοση	3.9.4	4.1.1
Γλώσσα προγραμματισμού	JAVA	JAVA
Άδεια χρήσης	General Public *License* (GPL)	General Public *License* (GPL)
Λειτουργικά Συστήματα	Cross platform	Cross platform
Διεπαφή χρήσης	• Μέτρια εμπειρία χρήσης • Όχι αρκετά διαισθητικό περιβάλλον όσον αφορά τη δημιουργία της ροής εργασίας και την ένωση μεταξύ των κόμβων • Καλή οργάνωση των στοιχείων του μενού	• Εύκολη διεπαφή (ευκολία προσθήκης, ένωσης, αντικατάστασης κόμβων) • Διαισθητικό περιβάλλον • Δημιουργία μετακόμβων για μείωση πολυπλοκότητας • Πιο σύγχρονη εμφάνιση
Τύποι υποστηριζόμενων αρχείων	• ARFF,CSV,libSVM,JSON, C45, XRFF, databases, serialized Instances (format .bsi extension) • Δυνατότητα παραγωγής ενός τυχαίου συνόλου δεδομένων και επέκτασης μέσω της προσθήκης επιπλέον πακέτων φόρτωσης αρχείων	• Arff, CSV, ECXEL, PMML, images. (δεν τα υποστηρίζει????) • Μπορεί να διαχειριστεί και συμπιεσμένα αρχεία
Φίλτρα	• Είναι κατηγοριοποιημένα σε supervised - unsupervized και ανά στήλη ανά σειρά • Κανονικοποίηση, διακριτοποίηση, μετασχηματισμός δεδομένων, διαχείριση ελλιπών τιμών, εντοπισμός θορύβου, έκτοπων σημείων κλπ	• Είναι κατηγοριοποιημένα ανά στήλη, ανά σειρά ή ανά πίνακα • Κανονικοποίηση, διακριτοποίηση, μετασχηματισμός δεδομένων, διαχείριση ελλιπών τιμών, εντοπισμός θορύβου, έκτοπων σημείων κλπ
Επιλογή χαρακτηριστικών	• Συσχετίσεις • Με βάση το x squared • Με βάση την εντροπία • Παραγοντική ανάλυση • Μεθόδους που χρησιμοποιούν forward feature selection, backward feature elimination, συνδυασμός μεθόδων • Χρήση αλγορίθμων	• Συσχετίσεις • Forward Feature Selection • Backward Feature Elimination • Γενετικοί αλγόριθμοι • Τυχαία επιλογή

Συγκεντρωτικός πίνακας χαρακτηριστικών KNIME & WEKA (2/2)

Υποστηριζόμενοι αλγόριθμοι κατηγοριοποίησης	Bayes (BayesNet, NaiveBayes), Functions (Logistic Regression, Gaussian Processes, SMO,), Lazy (Ibk, KStar, LWL), Meta (AdaBoostM1, Bagging, Stacking, Vote etc.), Rules (Decision tree, JRip, M5Rules, OneR, PART, ZeroR), Δέντρα (J48, LMT, M5P, Random Forest etc.), διάφοροι ταξινομητές που δεν ταιριάζουν σε καμία άλλη κατηγορία.	Bayes (BayesNet), Neural Network (MLP, PNN), Decision Tree, KNN, Logistic Regression, SVM, Random Forest
Υποστηριζόμενοι αλγόριθμοι Παλινδρόμησης	Linear Regression, k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machines, Multi-Layer Perceptron	Linear Regression, Random Forest
Αλγόριθμοι Συσταδοποίησης	K means, DBSCAN, Fuzzy c-means, Ιεραρχική συσταδοποίηση, EM, CobWeb, Canopy, FarthestFirst	K means, DBSCAN, Fuzzy c-means, Ιεραρχική συσταδοποίηση
Κανόνες συσχέτισης	Apriori, FIlteredAssociator, FPGrowth	Association rule learner (εφαρμόζει τον αλγόριθμο Apriori)
Μέθοδοι επικύρωσης	Cross validation, percentage split	Cross validation, percentage split
Δυνατότητα επέκτασης/ ενσωμάτωσης νέων χαρακτηριστικών	Αφορούν κυρίως επιπρόσθετους αλγορίθμους συσταδοποίησης, κατηγοριοποίησης, επιλογής χαρακτηριστικών ή φίλτρων προεπεξεργασίας	Αφορούν συνήθως ολόκληρες “βιβλιοθήκες”, και την ενσωμάτωση λογισμικών όπως το WEKA, R, κ.ο.κ.
Οπτικοποίηση	Scatter plot, boundary plot, ROC curve	Box plot, ιστόγραμμα, HiLite table, pie chart, scatter matrix

Παρουσίαση Δείγματος (DataSet) (1/2)

- Δείγμα 6.322 εγγραφών προς ψηφιοποίηση
  - Οι 6.322 εγγραφές αντιστοιχούν στα έγγραφα που καταχωρούνται στους φακέλους. Ο τελικός αριθμός των φακέλων είναι 6.087, καθώς κάθε φάκελος μπορεί να περιλαμβάνει παραπάνω από ένα έγγραφα
- ηλεκτρονική μορφή (.xls)
- Παρατηρήσεις από τις εργασίες ψηφιοποίησης του υλικού από 65 διαφορετικούς υπάλληλους σε ένα εύρος εννέα (9) εβδομάδων
- Αποτελείται από δέκα (10) χαρακτηριστικά
  - Creator, Filefolder, Container, process_start, process_end, process_duration, Documentcount, clientcode, contranr, pages

Παρουσίαση Δείγματος (DataSet) (2/2)

- **creator**: ο χρήστης/υπάλληλος - (65) χρήστες με όνομα “users_x”
- filefolder: φάκελοι - (6.087) φάκελοι με όνομα “folder_x” (κάθε φάκελος περιέχει το σύνολο των εγγράφων μια σύμβασης του πελάτη)
- **container**: η κούτα που περιέχει τους φακέλους κάθε έργου - (1.003) κούτες με όνομα “box_x”
- **process_start**: ημερομηνία/ώρα έναρξης επεξεργασίας φακέλου από τον χρήστη
- **process_end**: ημερομηνία/ώρα ολοκλήρωσης επεξεργασίας του φακέλου από τον χρήστη
- **process_duration**: συνολικός χρόνος (σε δευτερόλεπτα) ολοκλήρωσης της επεξεργασίας ενός φακέλου από τον χρήστη
- **documentcount**: ο αριθμός των εγγράφων κάθε φακέλου
- **clientcode**: οι πελάτες του έργου - (5.583) πελάτες με όνομα “ clientcode _x” (Οι πελάτες έχουν πολλούς φακέλους και πολλές διαφορετικές συμβάσεις)
- **contranr**: ο αριθμός των συμβάσεων του πελάτη - (6.217) συμβάσεις με όνομα “ contranr _x” (ο φυσικός φάκελος μπορεί να περιέχει παραπάνω από μια σύμβαση πελάτη)
- **pages**: ο αριθμός των σελίδων του κάθε φακέλου

Σχέδιο Εργασιών

- Μέθοδος ανάλυσης: παλινδρόμηση (regression)
- Στόχος η πρόβλεψη μίας αριθμητικής μεταβλητής και πιο συγκεκριμένα της διάρκειας ολοκλήρωσης μίας διαδικασίας (φακέλου) ψηφιοποίησης (σκαναρίσματος)
- Σύγκριση απόδοσης πέντε (5) γνωστών αλγορίθμων:
  - αλγόριθμος Κ πλησιέστερου γείτονα (KNN)
  - μηχανή διανυσμάτων υποστήριξης (SVM)
  - μέθοδος γραμμικής παλινδρόμησης (Linear Regreassion)
  - αλγόριθμος Random Forest
  - αλγόριθμος Decision Tree
- Αξιολόγηση της απόδοσης των αλγορίθμων με το μέτρο R

Προεπεξεργασία δεδομένων

- Μετατροπή του αρχείου από μορφή «.xls» σε μορφή «.csv»
- Αποτύπωση των χαρακτηριστικών σε μεταβλητές και προσδιορισμός τύπου
  - ονομαστική μεταβλητή (Creator, Filefolder, Container, clientcode, contranr,)
  - αριθμητική μεταβλητή (process_duration, Documentcount, pages)
  - Timestamp (process_start, process_end)
- Αφαίρεση δύο μεταβλητών: χρόνος έναρξης και λήξης (δεδομένου ότι η μεταβλητή process_duration προκύπτει από τη διαφορά τους)
- Μετατροπή ημερομηνίας στις αντίστοιχες εβδομάδες του έτους ((week37 έως week45 (επτά (9) εβδομάδες συνολικά))
- Αφαίρεση ελλιπών τιμών
- Αφαίρεση ακραίων τιμών
- Αφαίρεση των χαρακτηριστικών **crontranr, clientcode, container και filefolder** (τόσες μοναδικές τιμές όσες και ο συνολικός αριθμός των στιγμιότυπων του συνόλου δεδομένων μας)
- Μέθοδος επικύρωσης: η διασταυρωμένη επικύρωση (cross fold validation)

Επιλογή χαρακτηριστικών (attributes)

- Επιλογή χαρακτηριστικών στο WEKA - ευρεία γκάμα τεχνικών επιλογής χαρακτηριστικών
- Δοκιμές με ομάδες χαρακτηριστικών με βάση τη μεταβλητή στόχο - "***process_duration***"
  - μέθοδος συσχετίσεων (correlations) (ομάδα ***document_count***, ***pages***, ***weeks)***
  - μέθοδος αξιολόγησης ενός υποσυνόλου χαρακτηριστικών με βάση τον πλεονασμό και την ατομική ικανότητα πρόβλεψης σε συνδυασμό με μία greedy μέθοδο αναζήτησης (ομάδα ***creator***, ***weeks***, ***document_count)***
  - μέθοδος τεχνικής αξιολόγησης - αλγόριθμος Relief για την βαθμολόγηση των χαρακτηριστικών (ομάδα ***document_count***, ***pages***, ***creator)***
  - **μέθοδος με όλα τα χαρακτηριστικά (*document_count*, *pages*, *creator*, *weeks)***

WEKA RESULTS

Αλγόριθμος	Παράμετροι	Correlation coefficient			
		All attributes	Attributes creator, weeks, docs – *CfsSubsetEva – greedy stepwise - forward*	Attributes docs, pages, weeks – *Correlations*	Attributes docs, pages, creator – *Relief*
KNN	K=3, Euclidean distance	0.826	0.8084	0.7086	0.8057
	K=5, Euclidean distance	0.8264	0.8144	0.7265	0.8156
	K=7, Euclidean distance	0.824	0.8116	0.7382	0.8171
	K=3, Manhattan distance	0.8251	0.8084	0.71	0.8011
	K=5, Manhattan distance	0.828	0.8144	0.7281	0.813
	K=7, Manhattan distance	0.8264	0.8116	0.7372	0.8146
SVM (SMOreg)	Batchsize = 100, c=1, kernel = (polykernel=1), normalized data	0.8207	0.8034	0.7572	0.802
	Batchsize = 100, c=1, kernel = RBF kernel (default parameters), normalized data	0.8254	0.8083	0.7591	0.8023
Linear Regression	M5 method	0.8269	0.8092	0.7589	0.8093
	Greedy method	0.8268	0.8087	0.7589	0.8088
	No attributes	0.827	0.8092	0.7584	0.8093
Random Forest	Depth = 0 (unlimited), iterations = 100	0.845	0.8174	0.7297	0.8302
	Depth = 0 (unlimited), iterations = 300	0.8462	0.8181	0.7305	0.8313
	Depth = 0 (unlimited), iterations = 500	0.8464	0.8184	0.7307	0.8318
Decision Tree	Default parameters	0.8505	0.8319	0.7673	0.8231

KNIME RESULTS

KNIME RESULTS					
Αλγόριθμος	**Παράμετροι**	**Correlation coefficient**			
		All attributes	**Attributes creator, weeks, docs – *CfsSubsetEva – greedy stepwise - forward***	**Attributes docs, pages, weeks – *Correlations***	**Attributes docs, pages, creator – *Relief***
Linear Regression	Default parameters	0.820	0.797	0.759	0.804
Random Forest	Depth = 0 (unlimited), iterations = 100	0.853	0.8348	0.7657	0.8408
	Depth = 0 (unlimited), iterations = 300	0.855	0.8379	0.768	0.8413
	Depth = 0 (unlimited), iterations = 500	0.8555	0.8376	0.7666	0.840
Decision Tree	Default parameters	0.8482	0.8362	0.7645	0.8236

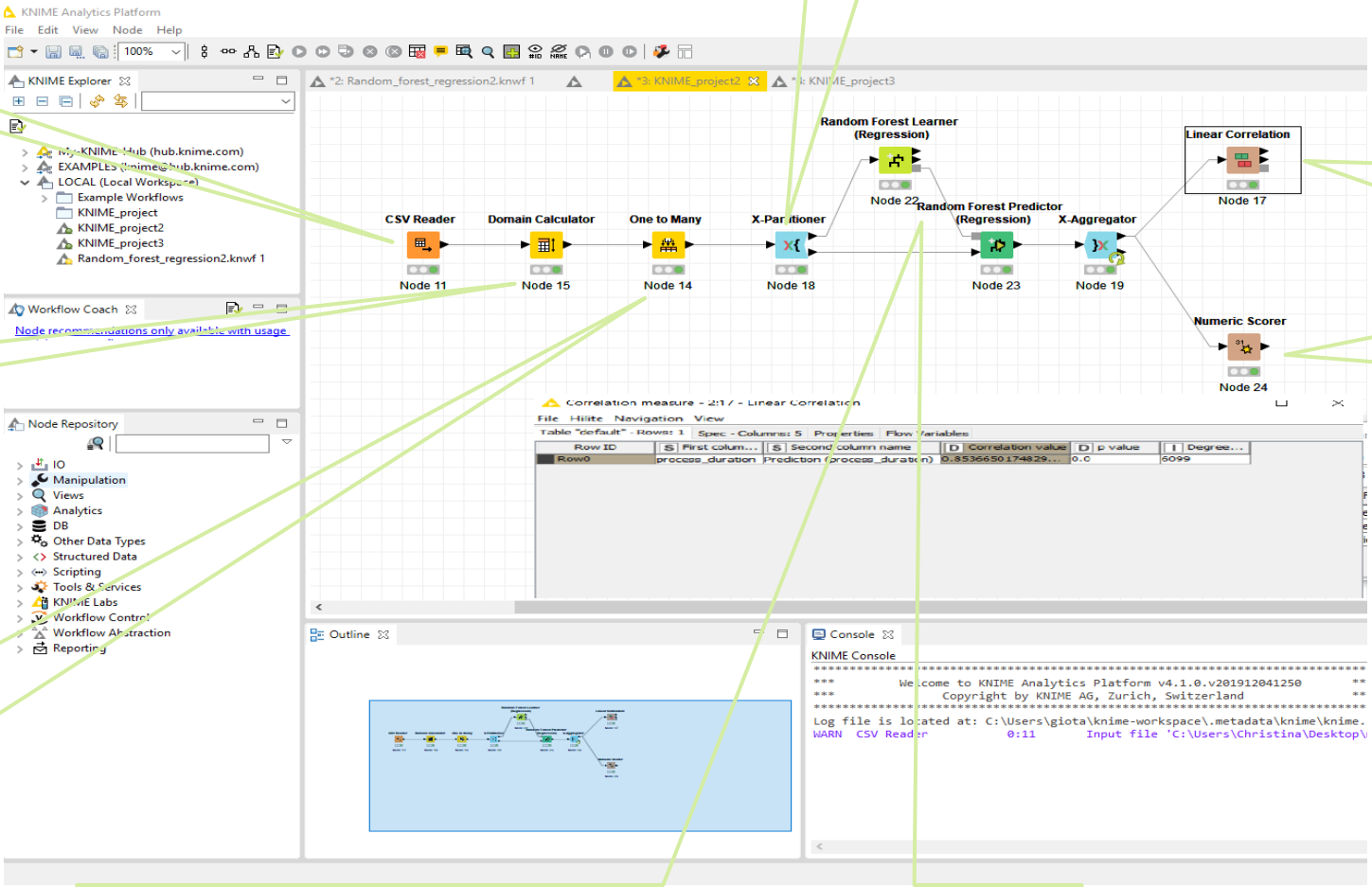
Ροές εργασιών - KNIME

Μέθοδος επικύρωσης 10-fold

Εισαγωγή αρχείου

Καθαρισμός τιμών

Μετατρέπει τη μεταβλητή creator από ονομαστική μεταβλητή σε αριθμητική με τιμές 0,1



Αποτελέσματα με το μέτρο R

Αποτελέσματα με το μέτρο R^2

Κόμβοι για την επιλογή του αλγόριθμου.
Learner δέχεται ως είσοδο το σύνολο εκπαίδευσης.
Predictor δέχεται ως είσοδο το μοντέλο που προέκυψε από την εκπαίδευση και τα δεδομένα για το testing.

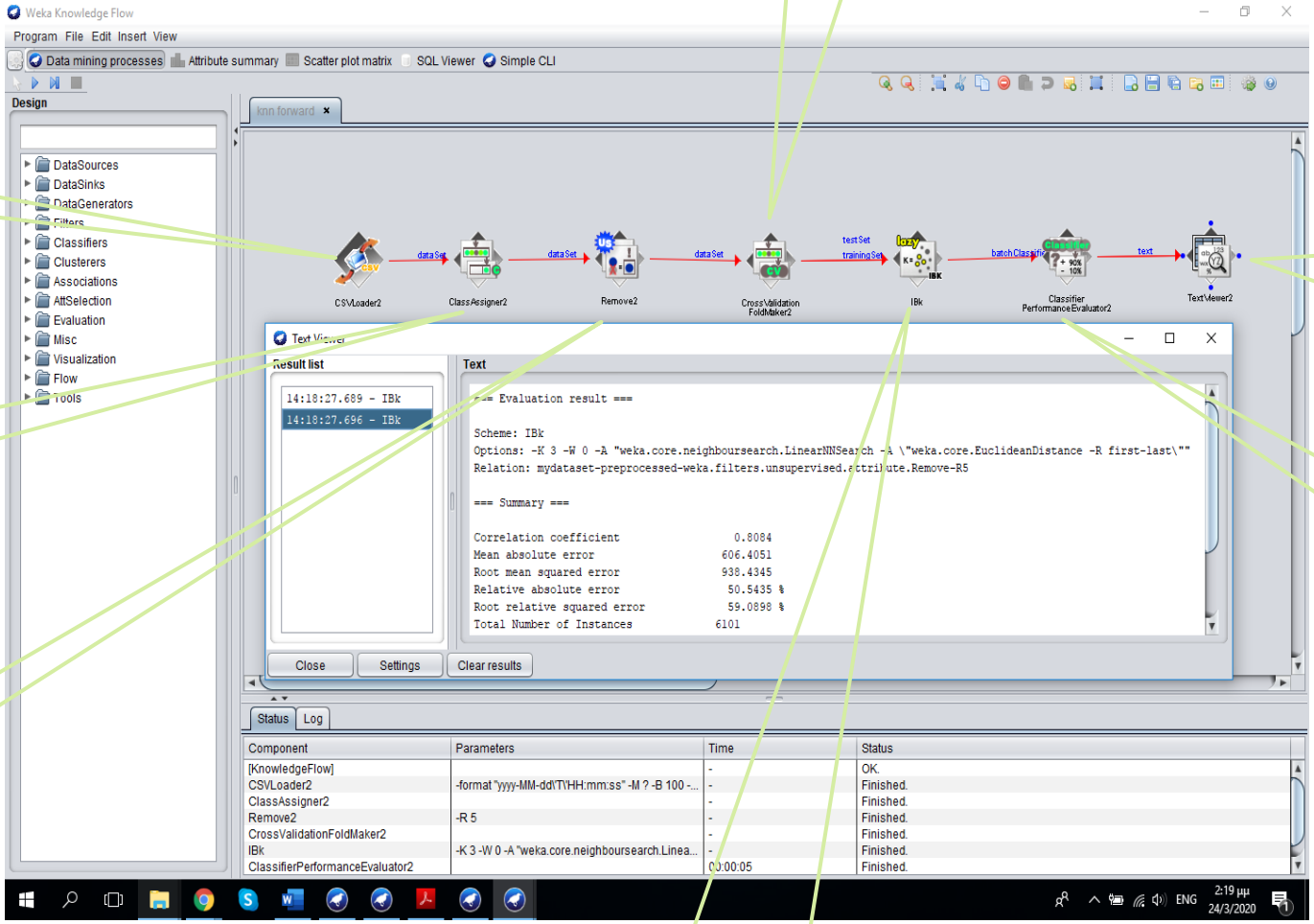
Ροές εργασιών - WEKA

Μέθοδος επικύρωσης 10-fold

Εισαγωγή αρχείου

Μεταβλητή στόχος

Αφαίρεση attributes



Αποτελέσματα με το μέτρο R

Επιλογή μέτρου αξιολόγησης

Κόμβος για την επιλογή του αλγόριθμου.

Αποτελέσματα δοκιμών

- Από τους πέντε (5) αλγορίθμους, μόνο οι τρεις (3) ενσωματωμένοι και στα δύο περιβάλλοντα
  - γραμμική παλινδρόμηση (linear regression), Random Forest, Decision Tree
- Παρόμοια αποτελέσματα και οι τρεις (3) αλγόριθμοι
- Μη ιδιαιτέρως σημαντική η εβδομάδα για την πρόβλεψη της μεταβλητής της διάρκειας
- Ο χρήστης, ο αριθμός των εγγράφων σε συνδυασμό με τον αριθμό των σελίδων, πλησιάζουν στο μοντέλο πρόβλεψης που δημιουργείται βάσει όλων των χαρακτηριστικών
- Απομάκρυνση χαρακτηριστικού - μείωση της τιμή της μεταβλητής στόχου
- Πιο καλή απόδοση (υψηλότερη τιμή R) ο αλγόριθμος Random Forest για το KNIME και ο Decision Tree για το WEKA για το σύνολο των χαρακτηριστικών
- KNIME στην περίπτωση του creator, μετατροπή της μεταβλητής αυτής από ονομαστική σε αριθμητική με τιμές 0,1
- Εξίσου καλά αποτελέσματα η περίπτωση της πρόβλεψης με βάση τον χρήστη, τα έγγραφα και τις σελίδες (μέθοδος relief)

Σημαντικότερα Ευρήματα

WEKA

- KNN
  - τιμές παραμέτρου κ = 3, 5 , 7
  - δύο διαφορετικών μέτρων απόστασης, ευκλείδειας απόστασης (Euclidean distance) και απόστασης Manhattan
  - η καλύτερη απόδοση για κ=7 και την ευκλείδεια απόσταση
- SMOreg (SVM) - παρόμοια αποτελέσματα και για τους δύο διαφορετικούς διαθέσιμους πυρήνες (kernels)
  - Χαμηλότερη απόδοση από τον αλγόριθμο KNN
- Linear Regression παρόμοια αποτελέσματα για κάθε επιλογή της παραμέτρου αναζήτησης
  - χαμηλότερη απόδοση από την τιμή του KNN
- Decision Tree - βέλτιστο για όλα τα χαρακτηριστικά

KNIME

- Decision Tree τα καλύτερα αποτελέσματα
- Εξίσου κοντά ο Random Forest με βέλτιστο αποτέλεσμα τις 500 επαναλήψεις
- Αποδοτικότερος ο αλγόριθμος Random Forest σε σχέση με (Linear regression) και Random Forest για το KNIME
- Βέλτιστο αποτέλεσμα αυτό με τις 500 επαναλήψεις του αλγόριθμου Random Forest.

Συγκριτικά σημαντικότερα αποτελέσματα

Μέθοδος	Παράμετροι	All attributes	
		WEKA	KNIME
Random Forest	Depth = 0 (unlimited), iterations = 500	0.8464	0.8555
Decision Tree	Default parameters	0.8505 (χρησιμοποιήθηκε ο αλγόριθμος M5P)	0.8482 (χρησιμοποιήθηκε ο αλγόριθμος Grandient Boosted Trees)

Συμπεράσματα - Θεωρητική προσέγγιση WEKA

WEKA	
Πλεονεκτήματα	**Μειονεκτήματα**
Περισσότεροι αλγόριθμοι παλινδρόμησης	Δεν επιτρέπει την ενσωμάτωση άλλων εργαλείων
Καλή οργάνωση των στοιχείων του μενού	Δύσκολη επιλογή χαρακτηριστικών που θα λάβουν μέρος στην ανάλυση σε κάθε βήμα, καθώς γίνεται με βάση τον δείκτη όχι την ονομασία της στήλης (στον κόμβο remove όταν γίνεται χειρωνακτικά)
Οι αλγόριθμοι κατηγοριοποίησης KNN και SVM μπορούν να εκτελέσουν ταξινόμηση ή παλινδρόμηση ανιχνεύοντας αυτόματα τον τύπο της μεταβλητής στόχου	Εύχρηστο γραφιστικό περιβάλλον το οποίο χρήζει κάποιας βελτίωσης
Επιτρέπει την εγκατάσταση επεκτάσεων	Δεν είναι προφανές ποιοι κόμβοι πρέπει να χρησιμοποιηθούν και σε ποιο σημείο και με ποιους κόμβους μπορούν να συνδεθούν

Συμπεράσματα - Θεωρητική προσέγγιση KNIME

KNIME	
Πλεονεκτήματα	**Μειονεκτήματα**
Πιο εύχρηστο γραφιστικό περιβάλλον για τον μέσο χρήστη	Λιγότεροι αλγόριθμοι παλινδρόμησης
Εύκολη σύνδεση κόμβων	Περιορίζονται σε εργασίες ταξινόμησης, αφού επιτρέπουν μόνο ονομαστικές μεταβλητές ως μεταβλητές στόχους
Παρέχει περισσότερη πληροφορία μέσω έτοιμων παραδειγμάτων	Η γραμμική παλινδρόμηση, ο Random Forest και το Decision Tree προσφέρουν λιγότερες δυνατότητες παραμετροποίησης
Η τεκμηρίωση του κάθε κόμβου που εισάγεται στο workflow, μας βοηθά να καταλάβουμε ευκολότερα την εργασία που εκτελεί κάθε κόμβος	Υστερεί στις μεθόδους επιλογής χαρακτηριστικών
Προειδοποιήσεις, μηνύματα και χρώματα που εμφανίζονται απευθείας επάνω στον κόμβο που παρουσιάζει το πρόβλημα βοηθά στην κατανόηση και άμεση διόρθωση του προβλήματος	Σημαντικός περιορισμός η μετατροπή της μεταβλητής creator, από ονομαστική σε αριθμητική με τιμές 0,1
Η επιλογή των χαρακτηριστικών που θα λάβουν μέρος στην ανάλυση σε κάθε βήμα είναι πιο εύκολη καθώς γίνεται με βάση την ονομασία της στήλης	
Διαισθητικό περιβάλλον που έχει ως αποτέλεσμα καλύτερη εμπειρία χρήσης	
Υποστηρίζει την ενσωμάτωση εργαλείων όπως το WEKA	
Δίνει τη δυνατότητα στον χρήστη να περιορίσει τις εγγραφές που θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου, ώστε η διαδικασία να εκτελεστεί πιο γρήγορα	

Συγκεντρωτικός πίνακας αποτελεσμάτων - Πειραματική προσέγγιση

- Ο αλγόριθμος Random Forest παρουσίασε συνολικά τα καλύτερα αποτελέσματα
- Με βάση τους αλγορίθμους που υπήρχαν και στα δύο λογισμικά (γραμμικής παλινδρόμησης, Random Forest και Decision Tree) τα αποτελέσματα είναι σχεδόν ταυτόσημα για το WEKA και το KNIME με βάση τη μετρική που χρησιμοποιήθηκε
- Για το μοντέλο πρόβλεψης μείωση των χαρακτηριστικών, μείωση και σε έναν βαθμό η ακρίβεια της πρόβλεψης
- Σημαντικότερα χαρακτηριστικά για την πρόβλεψη της διάρκειας μίας εργασίας: ο χρήστης, ο αριθμός των σελίδων και ο αριθμός των εγγράφων
- Η εβδομάδα μη σημαντικό χαρακτηριστικό για τη διαδικασία της πρόβλεψης
- Στις περιπτώσεις αφαίρεσης των μεταβλητών σελίδες ή αριθμός εγγράφων - μεγάλη συσχέτιση και μείωση αποτελεσμάτων
- Προτεινόμενο μοντέλο αυτό με όλα τα χαρακτηριστικά λόγω καλύτερων αποτελεσμάτων

Περιορισμοί και προτάσεις για μελλοντική έρευνα

- Τα διαθέσιμα δεδομένα αφορούν μόνο λίγες από τις εβδομάδες του χρόνου
  - Δεν υπάρχει πλήρη εικόνα για ολόκληρο το έτος

- Επιβεβαίωση ισχυρισμών στο μέλλον μέσω μεγαλύτερου δείγματος υπαλλήλων, αλλά και σε μεγαλύτερο εύρος χρόνου.

- Μελλοντικές έρευνες: δοκιμές με περισσότερα λογισμικά, αλγορίθμους παλινδρόμησης και μεθόδους διαμέρισης και επικύρωσης
- Έλεγχος επιπλέον παράμετροι: όπως η πρότερη εμπειρία του υπαλλήλου, τα η ηλικία και το φύλο του υπαλλήλου, ο τύπος των εγγράφων που υπάρχουν σε έναν φάκελο
- Εφαρμογή μοντέλων για την βελτίωση διαδικασιών, της ποιότητας των παρεχόμενων υπηρεσιών και επηρεασμός μελλοντικών αποφάσεων

Τέλος Παρουσίασης

Σας ευχαριστώ πολύ !!!