ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη Αναγκών της Τοπικής Αυτοδιοίκησης σε Cloud Storage με Χρήση Μοντέλων Προγνωστικής Αναλυτικής

Σαμιώτης Νικόλαος
Α.Μ. 2008175
Επιβλέπων Καθηγητής : Λουκής Ευριπίδης

Σάμος 2022

Predicting the Needs of the Local Government in Cloud Storage Using Predictive Analytical Models

Samiotis Nikolaos

Supervising Professor: Loukis Evripidis

Samos 2022

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε για την περάτωση του προπτυχιακού προγράμματος σπουδών και απόκτηση διπλώματος του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου.

Την τριμελή Επιτροπή Διδασκόντων του τμήματος αποτελούσαν οι παρακάτω:

- Καθηγητής, Λουκής Ευριπίδης (Επιβλέπων)
- Καθηγητής, Χαραλαμπίδης Ιωάννης
- Μεταδιδακτορική Ερευνήτρια, Κυριακού Νίκη

Στη μητέρα μου Αναστασία και το θείο μου Ιωάννη

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον κύριο Ευριπίδη Λουκή, Καθηγητή Συστημάτων Υποστήριξης Αποφάσεων του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου, διότι με ενέπνευσε να ασχοληθώ με την Προγνωστική Αναλυτική η οποία είναι ουσιαστικά μια από τις πολύ-επίπεδες δυνάμεις που κινούν σήμερα τον κόσμο αλλά και για την βοήθεια και στήριξή του σε ό,τι χρειάστηκα για την διεκπεραίωση αυτής της διπλωματικής εργασίας.

Επίσης θα ήθελα να ευχαριστήσω το ακαδημαϊκό προσωπικό του τμήματος για τη μεγάλη προσπάθειά τους να παρέχουν υψηλό επίπεδο εκπαίδευσης παρά τις αντικειμενικές δυσκολίες που αντιμετωπίζει ένα περιφερειακό πανεπιστημιακό τμήμα αλλά και τον άδικο χαμηλό βαθμό εισαγωγής με ό,τι αυτό συνεπάγεται. Το αναφέρω όχι από ευγένεια αλλά εκ πείρας μιας και έχω φοιτήσει στο τμήμα Βιολογίας του Πανεπιστημίου Αθηνών και στο τμήμα Μοριακής Βιολογίας και Γενετικής του Δημοκριτείου Πανεπιστημίου.

Επιπλέον θα ήθελα να ευχαριστήσω την ξαδέρφη μου Ευστρατία και τον επί πολλά χρόνια φίλο μου Παναγιώτη για την στήριξή τους σε δύσκολη περίοδο της ζωής μου προκειμένου να μπορέσω να τελειώσω τα μαθήματα και να αρχίσω την εκπόνηση αυτής της διπλωματικής.

Περίληψη

Στην σημερινή ψηφιακή εποχή της πληροφορίας όπου σχεδόν τα πάντα γίνονται διαδικτυακά, η τεχνολογία του Cloud Computing κατέχει πρωταγωνιστικό ρόλο συμβάλλοντας αποφασιστικά στη καλύτερη οργάνωση, τα μικρότερα κόστη, τα μεγαλύτερα κέρδη των εταιρειών αλλά και των δημόσιων οργανισμών. Παράλληλα η σχετικά νέα επιστήμη της Προγνωστικής Αναλυτικής μπορεί να επηρεάσει καθοριστικά τη λήψη καλύτερων αποφάσεων για κάθε πτυχή της ανθρώπινης ζωής, αποτελώντας έτσι μία από τις δυνάμεις που κινούν τον κόσμο σήμερα και στον μέλλον. Σε αυτή την εργασία αρχικά αναλύονται θεμελιώδη χαρακτηριστικά του Cloud Computing και εξετάζεται η χρήση της πιο γνωστής υπηρεσίας αυτής της τεχνολογίας, το cloud storage, στην Ελλάδα. Έπειτα γίνεται θεωρητική μελέτη της Προγνωστικής Αναλυτικής με παρουσίαση μερικών ενδιαφερόντων παραδειγμάτων. Στη συνέχεια παρουσιάζονται τέσσερις γνωστοί αλγόριθμοι μηχανικής μάθησης Naive Bayes, ID3, Random Forest, AdaBoost οι οποίοι θα χρησιμοποιηθούν για την ανάπτυξη προγνωστικών μοντέλων στο RapidMiner για τη χρήση του Cloud Storage στους δήμους της Ελλάδας. Στο τελευταίο μέρος της εργασίας γίνεται συγκριτική ανάλυση των αποτελεσμάτων των μοντέλων.

Abstract

In today's digital age of information where almost everything is done online, Cloud Computing technology plays a leading role, contributing decisively to better organization, lower costs, higher profits for companies and public organizations. At the same time the relatively new science of Predictive Analytics can have an influential impact on making better decisions for every aspect of human life, thus becoming one of the forces that drive the world today and in the future. In this thesis we first analyze the basic features of Cloud Computing and examine the use of the best-known service of this technology, cloud storage, in Greece. Afterwards a theoretical study of Predictive Analytics is made with the presentation of some interesting examples. Then four well-known machine learning algorithms Naive Bayes, ID3, Random Forest, AdaBoost are presented which will be used to develop predictive models in RapidMiner about the use of Cloud Storage in municipalities of Greece. In the last part of the thesis a comparative analysis of the models' results is made.

Περιεχόμενα

1. Εισαγωγή

Από το 2000 κυρίως και μετά η εκρηκτική ανάπτυξη της πληροφορικής και του διαδικτύου έχει επηρεάσει σε τέτοιο βαθμό τις πτυχές της ανθρώπινης ζωής ώστε η εποχή μας να χαρακτηρίζεται ως Ψηφιακή Εποχή. Το 2020, σύμφωνα με δεδομένα της Παγκόσμιας Τράπεζας, στο διαδίκτυο είχε πρόσβαση περίπου το 60% του παγκόσμιου πληθυσμού όταν αυτό ήταν 29% μόλις δέκα χρόνια πριν και 7% το 2000. Το Cloud Computing, δηλαδή με απλά λόγια η αξιοποίηση των ανά τον κόσμο υπολογιστικών πόρων, δεν είναι μόνο άλλη μια υπηρεσία πληροφορικής αλλά δημιουργήθηκε για να καλύψει τις τεράστιες τεχνολογικές ανάγκες της εποχής. Είναι χαρακτηριστικό πως το 2022 τα έσοδα από τις υπηρεσίες Cloud Computing έφτασαν το μισό τρις δολάρια. Προσφέροντας εξοικονόμηση χρόνου, χώρου και χρήματος μπορεί να καλύψει ποικιλοτρόπως τις παραγωγικές και οργανωτικές απαιτήσεις των εταιρειών δίνοντάς τους έτσι συγκριτικό πλεονέκτημα έναντι των ανταγωνιστών. Αλλά και στους δημόσιους οργανισμούς και κυβερνήσεις το Cloud Computing δύναται να βελτιώσει την απόδοση, τη διαφάνεια συμβάλλοντας στην παροχή καλύτερων υπηρεσιών στους πολίτες βελτιώνοντας εν τέλει και την ποιότητα της ζωής τους. Εξάλλου ο εν εξελίξει ψηφιακός μετασχηματισμός (Digital Transformation) όλων των πτυχών των κοινωνιών και των οικονομιών θα καταστήσει το Cloud Computing ακόμα πιο σημαντικό.

Η εποχή μας είναι επίσης γνωστή κι ως η Εποχή της Πληροφορίας. Δεν είναι λίγοι οι ειδικοί που υποστηρίζουν πως η πληροφορία πλέον έχει μεγαλύτερη αξία από το πετρέλαιο και το χρυσό. Και πως να μην έχει όταν σχεδόν τα πάντα μπορούν να καταγραφούν και να διακινηθούν ως πληροφορία. Μήπως όμως αυτή η πληροφορία θα μπορούσε και να…προβλεφθεί; Την απάντηση έρχεται να δώσει εκκωφαντικά μια σχετικά νέα αλλά τρομερά ελκυστική επιστήμη η Προγνωστική Αναλυτική. Διότι όχι μόνο η γνώση αλλά και το παρελθόν είναι δύναμη. Μελετώντας τα χαρακτηριστικά παρελθοντικών καταστάσεων μπορούμε να προβλέψουμε το μέλλον. Προβλέποντας το μέλλον μπορούμε να λάβουμε καλύτερες αποφάσεις για γεγονότα πριν συμβούν. Λαμβάνοντας καλύτερες αποφάσεις αποκτούμε καλύτερο έλεγχο, περισσότερα κέρδη, καλύτερες υπηρεσίες, καλύτερη ποιότητα ζωής. Με την προϋπόθεση ότι ένα τέτοιο πανίσχυρο εργαλείο θα χρησιμοποιηθεί υπεύθυνα.

Σε αυτή την εργασία αρχικά θα αναφερθούμε στα χαρακτηριστικά, τα μοντέλα υπηρεσιών και ανάπτυξης του Cloud Computing καθώς και συνοπτική μελέτη στην πιο γνωστή υπηρεσία αυτού το Cloud Storage. Επίσης γίνεται αναφορά της κατάστασης του Cloud Computing στην Ελλάδα και τους παράγοντες υιοθέτησης αυτής της τεχνολογίας. Στη συνέχεια παρουσιάζεται η Προγνωστική Αναλυτική, θεωρητικές πτυχές αυτής και κάποια χαρακτηριστικά παραδείγματα εφαρμογής. Γίνεται ανάλυση τεσσάρων γνωστών αλγόριθμων της μηχανικής μάθησης Naive Bayes, ID3, Random Forest και AdaBoost που χρησιμοποιούνται στη Προγνωστική Αναλυτική. Στην ενότητα της μεθοδολογίας πραγματοποιείται υλοποίηση αυτών των αλγορίθμων για την ανάπτυξη προγνωστικών μοντέλων στο RapidMiner χρησιμοποιώντας δεδομένα ερωτηματολογίου έρευνας για τη χρήση του Cloud Storage στους δήμους της Ελλάδας της εργασίας της Παρασκευής Δημητροπούλου που υπάρχει στο παράρτημα. Τέλος προβάλλονται τα αποτελέσματα των μοντέλων και γίνεται συγκριτική αξιολόγηση αυτών.

2. Cloud Computing και δημόσια τοπική αυτοδιοίκηση

2.1 Εισαγωγή στο Cloud Computing

Σύμφωνα με τον αρχαίο ελληνικό μύθο ο Δίας ερωτεύτηκε την πριγκίπισσα της Αρκαδίας Καλλιστώ, η οποία όμως μεταμορφώθηκε σε αρκούδα από την εξοργισμένη Ήρα. Αργότερα ο γιος της Καλλιστώ Αρκάς τη συνάντησε στο δάσος και ένα αγνοία του ότι η μητέρα του ήταν αρκούδα τράβηξε το τόξο του για να τη σκοτώσει. Ευτυχώς όμως επενέβη ο Δίας, μεταμόρφωσε τον Αρκά σε αρκουδάκι και τους τοποθέτησε στον ουρανό σχηματίζοντας την μικρή και μεγάλη άρκτο.

Όπως οι αστερισμοί δημιουργήθηκαν από ζώα και ήρωες που οι θεοί της μυθολογίας τοποθέτησαν στον ουρανό, κάτι ανάλογο συμβαίνει σήμερα στον κόσμο των υπολογιστών όπου ολοένα και περισσότερα δεδομένα και προγράμματα «τοποθετούνται», εκτελούνται όχι στους επιτραπέζιους ή κινητούς υπολογιστές μας αλλά στο Cloud Computing (Υπολογιστικό Νέφος). Όταν δημιουργούμε ένα υπολογιστικό φύλλο στο Google Drive τόσο το αρχείο αυτό όσο και στοιχεία του προγράμματος διαχειρίζονται από άγνωστους υπολογιστές που μπορούν να βρίσκονται οπουδήποτε στην υφήλιο. Το Cloud Computing έχει αλλάξει δραστικά και συνεχίζει να αλλάζει όχι μόνο στην καθημερινότητα του πολίτη, των εταιρειών, της κυβέρνησης αλλά και την ίδια την επιστήμη των υπολογιστών όπου δε θα ήταν υπερβολή να μιλήσουμε ότι αποτελεί paradigm shift για την επιστήμη αυτή.

Πως θα μπορούσαμε όμως να ορίσουμε το Cloud Computing; Υπάρχουν διάφοροι ορισμοί αλλά ο πιο κοινά αποδεκτός θα λέγαμε ότι είναι αυτός του Εθνικού Ιδρύματος Προτύπων και Τεχνολογιών των ΗΠΑ (NIST) το οποίο ανήκει στο Υπουργείο Εμπορίου:

- Ένα μοντέλο που δίνει τη δυνατότητα της συνεχούς, εύκολης και υψηλών απαιτήσεων πρόσβασης σε μια κοινόχρηστη συλλογή ρυθμιζόμενων υπολογιστικών πόρων (π.χ. δίκτυα, διακομιστές, αποθήκευση, εφαρμογές, υπηρεσίες) οι οποίοι τροφοδοτούνται και απελευθερώνονται με ελάχιστη προσπάθεια διαχείρισης και αλληλεπίδρασης παροχής υπηρεσιών. Αυτό το μοντέλο βασίζεται στην ανά πάσα στιγμή διαθεσιμότητα και αποτελείται από πέντε βασικά χαρακτηριστικά, τρία μοντέλα υπηρεσιών και τέσσερα μοντέλα ανάπτυξης.

Ιστορικά η ιδέα του Cloud Computing θεωρείται ότι ανήκε στον John McCarthy, μια από τις σημαντικότερες και επιδραστικότερες φυσιογνωμίες της θεωρητικής πληροφορικής με μεγάλη συνεισφορά στο τομέα της τεχνητής νοημοσύνης, ο οποίος το 1960 είχε δηλώσει πως «ίσως κάποια μέρα τα υπολογιστικά συστήματα να είναι οργανωμένα και να διατίθενται ως δημόσια αγαθά». Γενικά παλαιότερα με τον όρο cloud εννοούσαν ενοποιημένα δικτυακά κυκλώματα ενώ από τη δεκαετία του 90 και μετά ο όρος αυτός χρησιμοποιήθηκε και για τα εικονικά ιδιωτικά δίκτυα (VPN). Από το 2000 η έκρηξη του Cloud Computing θεωρείται ότι ξεκίνησε από εταιρίες γίγαντες όπως η Amazon οι οποίες πίστευσαν πως η ενοικίαση μη χρησιμοποιούμενων πόρων θα μπορούσε να είναι κερδοφόρα τόσο εξωτερικά όσο και εσωτερικά στη διαχείριση της ίδιας της εταιρείας. Έκτοτε έχουν υπάρξει χιλιάδες μικρές και μεγάλες εταιρείες που προσφέρουν υπηρεσίες Cloud Computing με λογισμικό ανοικτού ή κλειστού τύπου.

2.2 Χαρακτηριστικά του Cloud Computing

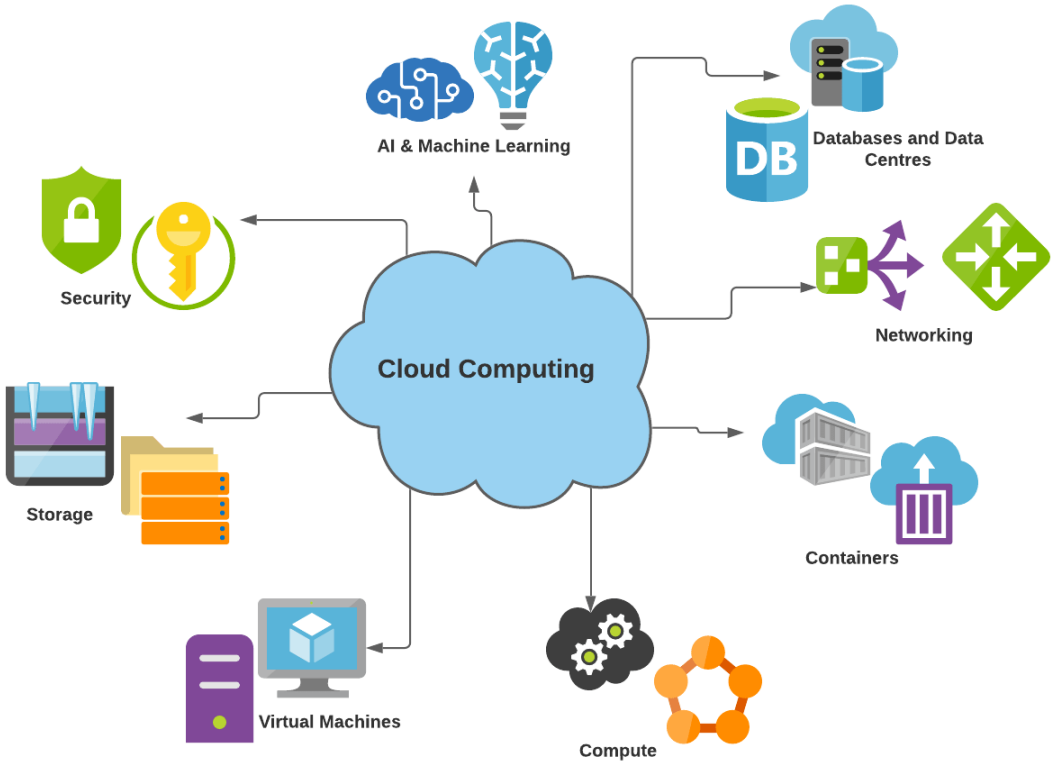
Πριν αναφερθούμε στα χαρακτηριστικά του Cloud Computing θα ήταν χρήσιμο αν όχι απαραίτητο να ορίσουμε συνοπτικά δύο έννοιες οι οποίες είναι πολύ βασικές για την καλύτερη κατανόηση του Cloud Computing: τον τεχνικό όρο Virtualization (Εικονικοποίηση) και τον εννοιολογικό όρο Cloud Services (Υπηρεσίες Νέφους).

- Virtualization. Υπάρχουν δύο κατηγορίες virtualization : application και server. Στο application virtualization υπάρχει μια εφαρμογή εγκατεστημένη σε έναν μόνο υπολογιστή στην οποία μπορούν να έχουν πρόσβαση ταυτόχρονα πολλοί χρήστες από οποιαδήποτε φθηνή υπολογιστική μηχανή, οπουδήποτε κι αν βρίσκονται και όποτε το επιθυμούν μιας και τα δεδομένα αποθηκεύονται στον κεντρικό υπολογιστή. Στο server virtualization γίνεται χρήση φυσικού υπολογιστικού υλικού για την εγκατάσταση άλλων εικονικών υπολογιστικών μηχανών με τα δικά τους λειτουργικά συστήματα και προγράμματα. Όπως γίνεται εύκολα κατανοητό η χρήση virtualization συμβάλλει στη σημαντική μείωση κόστους. Δεν είναι όμως από μόνη της μια υπηρεσία αλλά μια σημαντική τεχνολογία που σε συνδυασμό με άλλα εργαλεία και διαδικασίες να αποτελέσουν μια ολοκληρωμένη υποδομή Cloud Computing.
- Cloud Services. Στην πληροφορική μια υπηρεσία είναι μία συλλογή συστημάτων τεχνολογίας, στοιχείων και πόρων που συνεργάζονται για να παρέχουν αξία στους χρήστες. Δυο σημαντικοί παράμετροι για την αξιολόγηση μια τέτοιας υπηρεσίας είναι το κόστος και το service-level agreement (SLA) δηλαδή η σύμβαση μεταξύ καταναλωτή υπηρεσιών και παρόχου για την ποιότητά της, πόσο γρήγορα θα παραδοθεί και το πεδίο εφαρμογής. Επομένως μια υπηρεσία cloud είναι η υλοποίηση μια επιχειρηματικής διαδικασίας που παρέχει αξία στους καταναλωτές είτε είναι αυτοί εσωτερικοί (π.χ. τμήμα της εταιρείας) είτε εξωτερικοί.

Με βάση τον ορισμό του NIST τα βασικά χαρακτηριστικά του Cloud Computing είναι τα εξής πέντε.

1. Broad Network Access. Πρόσβαση του χρήστη στην υπηρεσία cloud με οποιαδήποτε υπολογιστική συσκευή οπουδήποτε κι αν βρίσκεται. Βέβαια αύξηση της βάσης των χρηστών συνεπάγεται και αύξηση της ανησυχίας για την ασφάλεια αλλά για αυτό το λόγο υπάρχουν και τα διαφορετικά μοντέλα ανάπτυξης που αναφέρονται παρακάτω.
2. On-demand Self-service. Η υπηρεσία cloud θα πρέπει να μπορεί να είναι διαθέσιμη οποτεδήποτε κι αν ζητηθεί από τον πελάτη δηλαδή να συνεχίζει είναι διαθέσιμη όταν δεν την επιθυμεί ο πελάτης αλλά και να μπορεί να είναι χωρίς διακοπές σε όλη τη διάρκεια της χρήσης της.
3. Resource Pooling and Virtualization. Εικονικοποίηση των προγραμματιστικών ή μηχανικών πόρων για τη συγκέντρωση και διάθεση αυτών σύμφωνα με τις ανάγκες των ταυτόχρονων χρηστών.

4. Rapid Elasticity. Οι υποδομές της υπηρεσίας Cloud θα πρέπει να μπορούν να προσφέρουν περισσότερους πόρους αν οι ανάγκες του χρήστη το απαιτούν αλλά και να ελαττώνονται όταν αυτές είναι μειωμένες ώστε αυτοί οι πόροι να μπορούν να χρησιμοποιηθούν από άλλους χρήστες. Για το συγκεκριμένο χαρακτηριστικό αναφέρονται οι όροι οριζόντια κλιμάκωση όταν χρησιμοποιείται ένας μεγάλος αριθμός πόρων για την κάλυψη της ζήτησης και κάθετη κλιμάκωση όταν αναβαθμίζονται οι πόροι προκειμένου να μπορέσουν να καλύψουν τις ανάγκες των πελατών.
5. Measured Service. Παροχή εργαλείων με τα οποία τόσο ο πάροχος όσο και ο πελάτης μπορούν δουν ακριβώς και ανά πάσα στιγμή τα ποιοτικά και ποσοτικά χαρακτηριστικά της χρήσης της cloud υπηρεσίας.

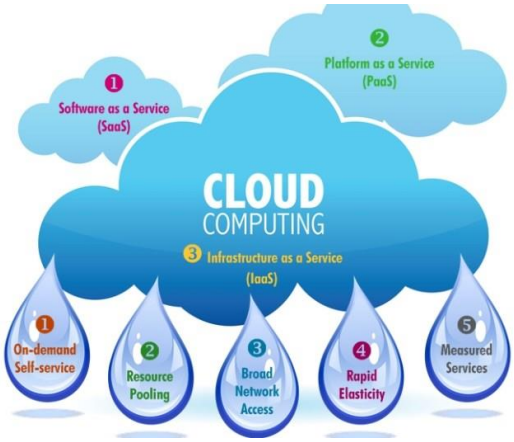


Εικόνα 1. Source: www.linuxhowto.net

2.3 Μοντέλα υπηρεσιών του Cloud Computing

Το Cloud Computing μπορεί να υλοποιηθεί με διάφορα μοντέλα υπηρεσιών ανάλογα με τις ανάγκες των πελατών. Τα πιο γνωστά και ευρέως διαδεδομένα μοντέλα είναι τα εξής τρία: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) και Software as a Service (SaaS).

- Το Infrastructure as a Service (IaaS) καλύπτει ένα ευρύ φάσμα λειτουργιών από μεμονωμένους διακομιστές ή ιδιωτικά δίκτυα, αποθηκευτικούς χώρους, φυσικές ή εικονικές μηχανές, τείχη προστασίας firewall, πακέτα λογισμικού και πολλά άλλα στα οποία ο πελάτης έχει άμεση πρόσβαση, να χρησιμοποιήσει και να παραμετροποιήσει σύμφωνα με τις ανάγκες του. Οι πάροχοι συνήθως χρεώνουν τους πελάτες τους ανάλογα με το πόσους πόρους αυτοί κατανάλωσαν. Amazon CloudFormation και Amazon EC2 αποτελούν γνωστά παραδείγματα τέτοιων παρόχων.
- Στο Platform as a Service (PaaS) προσφέρεται μια ολοκληρωμένη πλατφόρμα λογισμικού όπου ο πελάτης μπορεί να χρησιμοποιήσει προκειμένου να αναπτύξει τα δικά του προγράμματα χωρίς να ενδιαφέρεται για την υλική υποδομή πάνω στην οποία έχει εγκατασταθεί αυτή η πλατφόρμα λογισμικού. Για το λόγο αυτό συνήθως αυτό το μοντέλο προτιμάται κυρίως από εταιρείες ανάπτυξης λογισμικού και προγραμμάτων. Παραδείγματα τέτοιων παρόχων υπηρεσιών περιλαμβάνουν τα Cloud Foundry και OrangeScape. Αυτό το μοντέλο όμως μπορεί να κάνει δύσκολη τη μεταφορά από ένα πάροχο σε ένα άλλο για αυτό και τα τελευταία χρόνια έχει αρχίσει να αναπτύσσεται και το OpenPaaS.
- Το Software as a Service (SaaS) αποτελεί το πιο γνωστό και διαδεδομένο μοντέλο από τα τρία όπου ο πελάτης έχει πρόσβαση μόνο στις εφαρμογές που εξυπηρετούν τις ανάγκες του χωρίς να ενδιαφέρεται ούτε για το λογισμικό ούτε για τη φυσική ή εικονική υποδομή πάνω στις οποίες έχουν εγκατασταθεί αυτές οι εφαρμογές. Η πρόσβαση σε αυτές από τον τελικό χρήστη γίνεται συνήθως με απλά εργαλεία όπως έναν περιηγητή internet και χρεώνονται ανάλογα με το χρόνο χρήσης ή μπορεί να είναι και δωρεάν. Χαρακτηριστικά παραδείγματα τέτοιων μοντέλων είναι το Office365 της Microsoft και το Google Drive. Τα τελευταία χρόνια μάλιστα υπάρχει μια ραγδαία αύξηση του μοντέλου αυτού στον ιδιωτικό τομέα από τα συστήματα ενδοεπιχειρησιακού σχεδιασμού ERP αλλά και στο δημόσιο από τη τοπική αυτοδιοίκηση.



Εικόνα 2. Source: www.csc.com

Εκτός από αυτά τα τρία γνωστά μοντέλα υπηρεσιών υπάρχουν και άλλα όχι τόσο γνωστά αλλά εξίσου σημαντικά. Στο Network as a Service (NaaS) οι πελάτες νοικιάζουν ολοκληρωμένες δικτυακές πλατφόρμες από τους παρόχους cloud απαλλάσσοντάς τους έτσι από τις απαιτήσεις που θα είχε μια τέτοια επένδυση σε κόστος και χρόνο. Με το Business Process as a Service (BPaaS) οι εταιρείες μπορούν να αναθέσουν ολόκληρα τμήματα τους σε άλλες εταιρείες μέσω τεχνολογιών cloud. Ένα άλλο αναδυόμενο μοντέλο cloud είναι το Information as a Service (INaaS) όπου μια εταιρεία παρέχει πληροφορίες (πχ. νομικές, φορολογικές) σε μια άλλη απαραίτητες για τη λειτουργία της. Τέλος υπάρχουν τα μοντέλα Storage as a Service (StaaS) όπου γίνεται παροχή αποθηκευτικού υλικού σε cloud, Hardware as a Service (HaaS) για την παροχή μόνο υπολογιστικού υλικού και Database as a Service (DaaS) για την αξιοποίηση online βάσεων δεδομένων από επιχειρήσεις χρησιμοποιώντας διάφορα web applications.

2.4 Μοντέλα ανάπτυξης του Cloud Computing

Σύμφωνα με τον ορισμό του NIST υπάρχουν τέσσερα μοντέλα ανάπτυξης του cloud τα οποία διακρίνονται κυρίως με βάση το κόστος και την ασφάλεια που προσφέρουν: Public Cloud (Δημόσιο Νέφος), Private Cloud (Ιδιωτικό Νέφος), το Community Cloud (Κοινοτικό Νέφος) και το Hybrid Cloud (Υβριδικό Νέφος).

- Public Cloud. Πρόκειται για το πιο γνωστό και ευρέως χρησιμοποιούμενο μοντέλο ανάπτυξης cloud μέσω του οποίου παρέχονται υπηρεσίες στον οποιοδήποτε τελικό χρήστη που έχει πρόσβαση στο διαδίκτυο. Η υλική υποδομή αυτού μπορεί να βρίσκεται οπουδήποτε στον κόσμο και διαχειρίζεται πλήρως από τον πάροχο. Συνήθως δεν προτιμάται από εταιρείες και μεγάλους οργανισμούς γιατί μπορεί να προκύψουν θέματα ασφαλείας και νομικά προβλήματα.
- Private Cloud. Λόγω του συνήθως υψηλού κόστους για υποδομές και εκπαιδευμένο προσωπικό προτιμάται κυρίως από μεγάλες εταιρείες και κυβερνητικούς οργανισμούς. Οι υπηρεσίες του cloud παρέχονται κυρίως από το εσωτερικό ιδιόκτητο δίκτυο οπότε προσφέρει μεγάλη ασφάλεια αλλά και μεγιστοποίηση της χρήσης των εσωτερικών πόρων.
- Community Cloud. Αποτελεί ένα ενδιάμεσο μοντέλο μεταξύ του public και private cloud. Είναι ιδιαίτερα ελκυστικό για χρήστες, εταιρείες ή οργανισμούς με παρόμοια ενδιαφέροντα, συμφέροντα και προβληματισμούς οι οποίοι δημιουργούν αυτή την οντότητα διαμοιράζοντας τους πόρους τους. Η διαχείριση και εποπτεία του cloud γίνεται από τα ίδια τα μέλη ή κάποιον τρίτο φορέα.
- Hybrid Cloud. Το υβριδικό μοντέλο αποτελεί έναν οποιοδήποτε συνδυασμό από public, private και community clouds τα οποία συνυπάρχουν ως ξεχωριστές οντότητες προσφέροντας το καθένα τα πλεονεκτήματά τους στις διαφορετικές ανάγκες της εταιρείας ή του οργανισμού.

2.5 Cloud Storage

Ο George Carlin, ένας διάσημος Αμερικανός stand up κωμικός και συγγραφέας με κοινωνική κριτική, είχε πει κάποτε πως οι άνθρωποι ουσιαστικά σε όλη τους τη ζωή αυτό που κάνουν είναι να συσσωρεύουν «πράγματα» και να βρίσκουν μέρη για να τα αποθηκεύουν. Αυτά τα «πράγματα» σήμερα είναι κυρίως πληροφορίες και δεδομένα στους υπολογιστές και το διαδίκτυο τα οποία μπορούν πολύ εύκολα πλέον να αποθηκευτούν σε απομακρυσμένους servers.

Πράγματι όταν αναφερόμαστε στο Cloud Storage, στο πιο βασικό του επίπεδο, εννοούμε το σύστημα εκείνο που ένας server με αποθηκευτικό χώρο είναι συνδεδεμένος στο διαδίκτυο στον οποίο ένας χρήστης μπορεί να συνδεθεί και να στείλει αντίγραφα των αρχείων του. Όταν ο χρήστης επιθυμεί να ανακτήσει τα αρχεία, μπορεί ανά πάσα στιγμή μέσω μιας απλής διαδικτυακής εφαρμογής να έχει πρόσβαση σε αυτά ή ακόμα και να τα τροποποιήσει απευθείας στον server ο οποίος εποπτεύεται από κάποιον τρίτο φορέα, συνήθως τον πάροχο. Οι server αυτοί μπορεί να ανήκουν σε μικρά συστήματα ή σε χώρους εκατοντάδων τετραγωνικών μέτρων και να αποτελούν μια server farm.

Οι υπηρεσίες Cloud Storage έχουν εξελιχθεί ώστε πλέον να αποτελούν μόνο ένα μέρος των υπηρεσιών Storage as a Service (StaaS) που όπως είχαμε αναφέρει είναι ένα από τα μοντέλα υπηρεσιών Cloud Computing. Αυτό συμβαίνει γιατί η υψηλή απόδοση αυτών εξαρτάται από παράγοντες όπως είναι η διαθεσιμότητα, η αξιοπιστία, η ακρίβεια στη διατήρηση των δεδομένων και η ασφάλεια. Δηλαδή ο παραδοσιακός τρόπος αποθήκευσης σε απομακρυσμένους διακομιστές που απλά περιλάμβανε μερικά πρωτόκολλα όπως FTP, WebDAV, NFS/CIFS σήμερα δεν αρκεί αλλά χρειάζονται και εξελιγμένα API, τεχνολογίες virtualization, σύγχρονα εργαλεία διαχείρισης αποθηκευτικών χώρων, παροχή υποστηρικτικών υπηρεσιών από τρίτους παρόχους.

Σύμφωνα με την Amazon, το μεγαλύτερο πάροχο υπηρεσιών cloud στο κόσμο, υπάρχουν τρεις τύποι αποθήκευσης δεδομένων σε cloud. Η αποθήκευση αντικειμένων αφορά την ανάπτυξη σύγχρονων εφαρμογών που απαιτούν επεκτασιμότητα, ευελιξία, επεξεργασία metadata των δεδομένων τους, δημιουργία αντιγράφων ασφαλείας και αρχειοθέτηση. Η αποθήκευση αρχείων στηρίζεται σε διακομιστές Network Attached Storage (NAS), χρησιμοποιείται από εφαρμογές που χρειάζονται πρόσβαση σε κοινόχρηστα αρχεία και συστήματα αρχείων, είναι ιδανική για περιβάλλοντα ανάπτυξης εφαρμογών, καταστήματα πολυμέσων και καταλόγους χρηστών. Η αποθήκευση block παρέχεται με εικονικούς διακομιστές και χρησιμοποιείται από βάσεις δεδομένων ή συστήματα ERP που απαιτούν αποκλειστικό χώρο αποθήκευσης χαμηλής δικτυακής καθυστέρησης ώστε να έχουν υψηλή απόδοση.

Οι υπηρεσίες Cloud Storage έχουν πολλά οφέλη. Το πιο σημαντικό θα λέγαμε ότι είναι το κόστος μιας και εξαλείφεται οποιαδήποτε ανάγκη θα υπήρχε σε υποδομές και προσωπικό για τη συντήρηση και υποστήριξη αποθηκευτικών χώρων. Η εξοικονόμηση χρόνου είναι επίσης σημαντική διότι επιτρέπει στο προσωπικό να επικεντρωθεί στον προγραμματισμό και επίλυση άλλων προβλημάτων αντί να χρειάζεται να διαχειρίζεται συστήματα αποθήκευσης. Άλλο ένα πολύτιμο όφελος είναι η διαχείριση δεδομένων στο Cloud Storage με την οποία χρησιμοποιώντας πολιτικές διαχείρισης κύκλου ζωής αυτών μπορούν εύκολα να εισαχθούν νέα δεδομένα ή να γίνει τροποποίηση των υπαρχόντων σε διάφορες βαθμίδες σύμφωνα με τις όποιες ανάγκες.

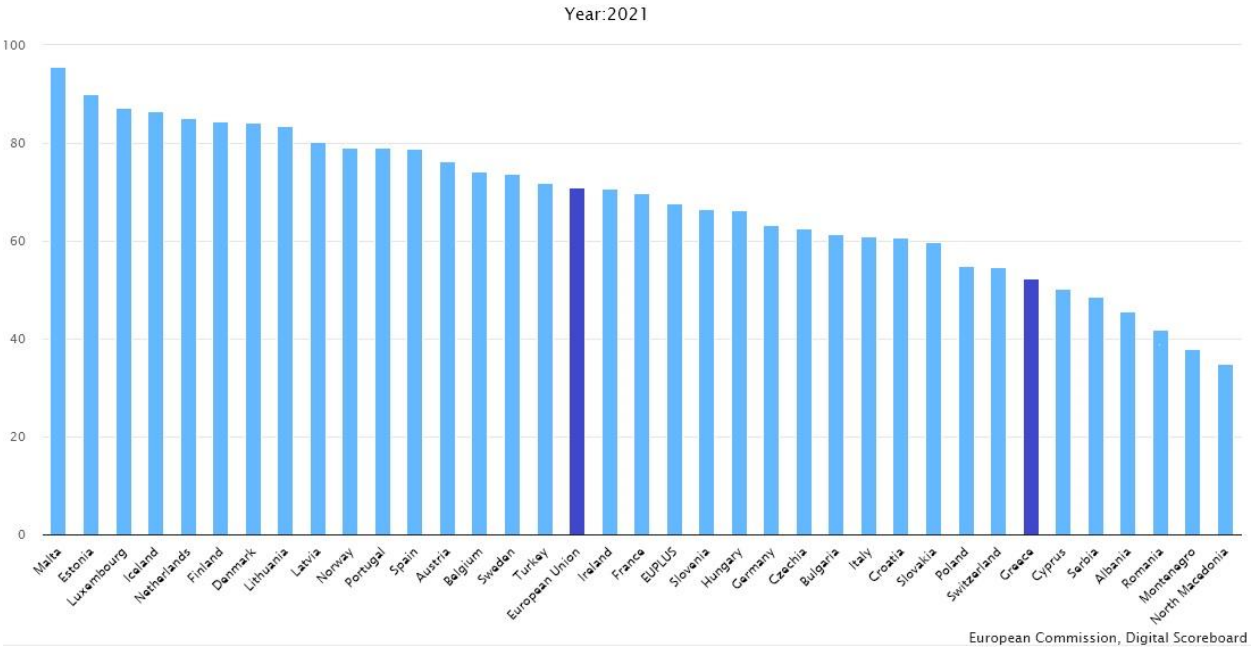
2.6 Ηλεκτρονική Διακυβέρνηση στην Ελλάδα

Όταν στις προηγούμενες υπό-ενότητες αναφερθήκαμε στα χαρακτηριστικά και μοντέλα του Cloud Computing το κάναμε κυρίως υπό το πρίσμα γενικά κάποιων πελατών και εταιρειών. Ποια η σχέση όμως του Cloud Computing με τον κυβερνητικό τομέα σήμερα και ποιοι παράγοντες επηρεάζουν την υιοθέτηση αυτού ειδικά στην Ελλάδα; Για να απαντήσουμε σε αυτό το ερώτημα πρώτα θα ήταν χρήσιμο και σκόπιμο να εξετάσουμε κάποια επίσημα στοιχεία για το 2021 στην ηλεκτρονική διακυβέρνηση.

Σύμφωνα με τα Ηνωμένα Έθνη ο όρος «ηλεκτρονική διακυβέρνηση» συνήθως αναφερόταν στη χρήση τεχνολογιών πληροφοριών και επικοινωνιών για τη βελτίωση της αποτελεσματικότητας των κρατικών υπηρεσιών είτε φυσικά είτε διαδικτυακά. Σήμερα όμως αυτός ο όρος αφορά όχι μόνο τις υπηρεσίες αλλά ένα πολύ μεγαλύτερο φάσμα αλληλεπιδράσεων με πολίτες και επιχειρήσεις χρησιμοποιώντας ακόμα και ευρέως προσβάσιμα κυβερνητικά δεδομένα ώστε να γίνει δυνατή η καινοτομία στη διακυβέρνηση. Η βασική αρχή της ηλεκτρονικής διακυβέρνησης είναι η βελτίωση της εσωτερικής λειτουργίας του δημόσιου τομέα μειώνοντας το οικονομικό κόστος, τους χρόνους συναλλαγών, αποτελεσματικότερη χρήση των πόρων. Έτσι οι κυβερνήσεις, είτε εθνικές είτε τοπικές, θα μπορούν να παρέχουν καλύτερες υπηρεσίες στους πολίτες, να υπάρχει μεγαλύτερη διαφάνεια, λογοδοσία, εμπιστευτικότητα δηλαδή καλύτερη λειτουργία της δημοκρατίας.

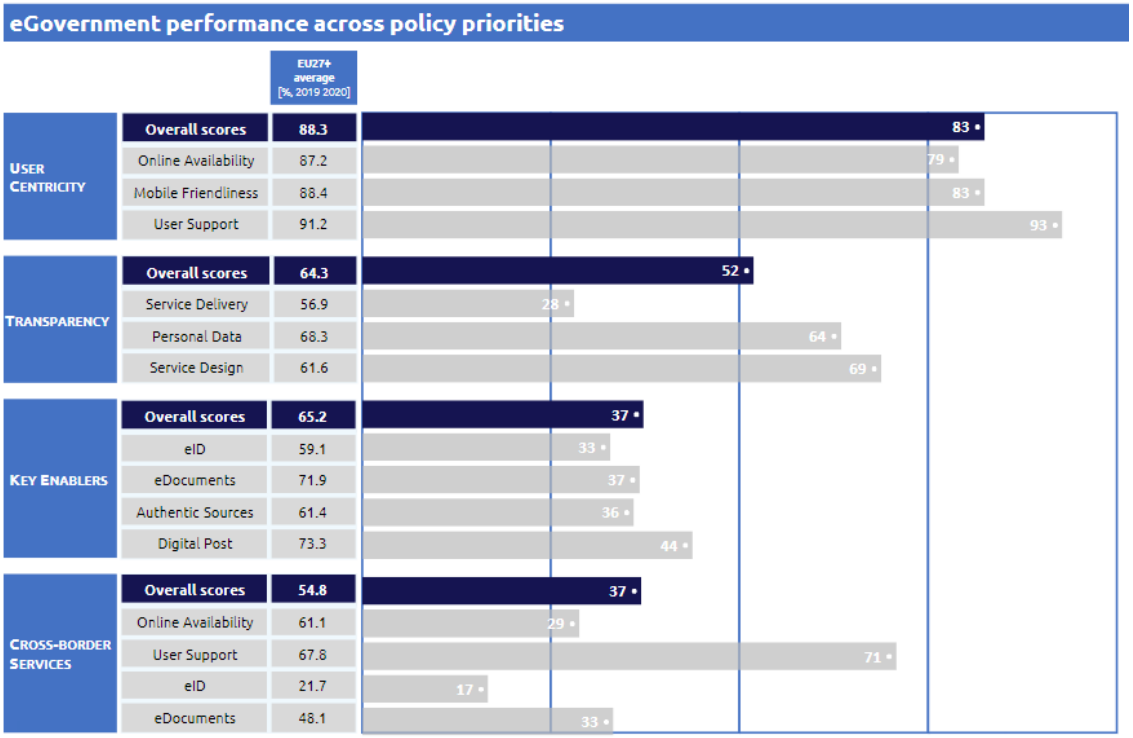
Υπάρχουν τρία είδη υπηρεσιών ηλεκτρονικής διακυβέρνησης. Το πρώτο, που θεωρείται και το πιο σημαντικό, Government to Government (G2G) αφορά την αποτελεσματική διεξαγωγή διαδικτυακών συναλλαγών και ανταλλαγή δεδομένων μεταξύ των κυβερνητικών υπηρεσιών. Το δεύτερο Government to Business (G2B) περιλαμβάνει την αλληλεπίδραση των επιχειρήσεων με το κράτος για την αποδοτικότερη διαχείριση των διάφορων νομικών, οικονομικών, πολιτικών θεμάτων που τις απασχολούν. Το τρίτο Government to Citizen (G2C) αφορά τις δημόσιες υπηρεσίες προς τους πολίτες με στόχο την καλύτερη εξυπηρέτησή τους από το κράτος καθώς και τη συμμετοχή τους σε διαδικασίες διαβούλευσης και λήψης αποφάσεων.

Η Ευρωπαϊκή Επιτροπή στο τέλος κάθε χρόνου διενεργεί και δημοσιοποιεί έρευνα που αφορά την απόδοση των 27 κρατών μελών της Ευρωπαϊκής Ένωσης στην ηλεκτρονική διακυβέρνηση την eGovernment Benchmark. Αν και τα τελευταία χρόνια έχουν βελτιωθεί σε σημαντικό βαθμό οι υπηρεσίες ηλεκτρονικής διακυβέρνησης στην Ελλάδα, παρόλα αυτά το 2021 συνεχίζει να βρίσκεται συνολικά αρκετά κάτω από τον ευρωπαϊκό μέσο όρο αν και κάποιοι επιμέρους δείκτες είναι αρκετά κοντά σε αυτόν. Ενδιαφέρον αποτελεί το γεγονός πως χώρες όπως η Τουρκία, η Λιθουανία, η Λετονία είναι πάνω από τον ευρωπαϊκό μέσο όρο ενώ οι πρώτες θέσεις καταλαμβάνονται από τη Μάλτα, την Εσθονία και το Λουξεμβούργο.



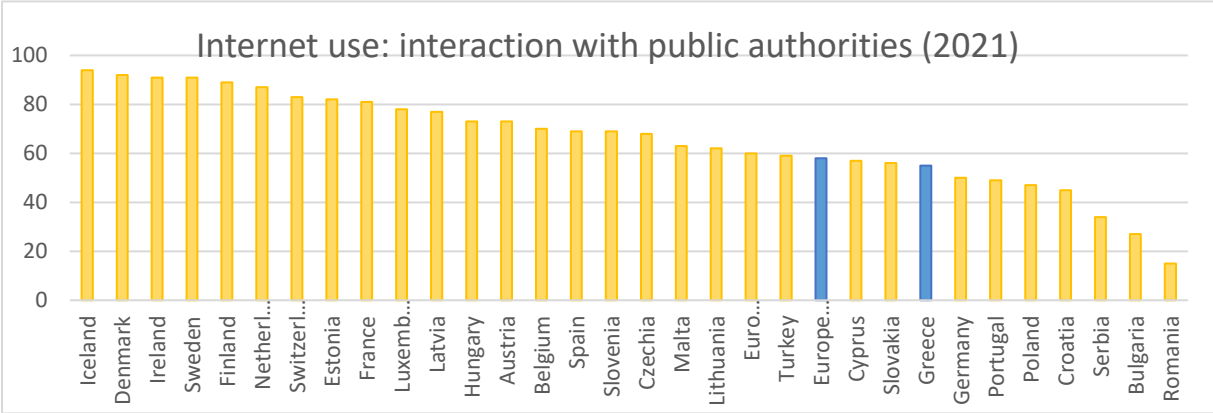
Εικόνα 3. Source: European Commission

Στην παρακάτω εικόνα παρουσιάζονται τα αποτελέσματα της έρευνας για την Ελλάδα το 2021. Παρατηρούμε ότι το User Centricity κυμαίνεται περίπου στον ευρωπαϊκό μέσο όρο, δηλαδή το πόσο φιλικές είναι οι υπηρεσίες ηλεκτρονικής διακυβέρνησης στον πολίτη που περιλαμβάνει την online διαθεσιμότητα, πόσο καλά σχεδιασμένες είναι οι ιστοσελίδες αυτών για κινητές συσκευές και κατά πόσο περιλαμβάνουν βοηθητικές πληροφορίες χρήσης και feedback. Λίγο κάτω από το μέσο όρο βρίσκεται η χώρα όσον αφορά το Transparency που υποδεικνύει σε ποιο βαθμό οι κυβερνήσεις είναι διαφανείς σχετικά με διαφάνεια παροχής υπηρεσιών, διαφάνεια σχεδιασμού των υπηρεσιών, διαφάνεια στη χρήση των προσωπικών δεδομένων για τις υπηρεσίες αυτές. Στις υπόλοιπες δυο κατηγορίες τρίτη και τέταρτη η Ελλάδα βαθμολογείται σημαντικά χαμηλότερα από τον ευρωπαϊκό μέσο όρο. Η τρίτη Key Enablers αναφέρεται στο κατά πόσο χρησιμοποιούνται οι τεχνικές υπηρεσίες της ηλεκτρονικής ταυτοποίησης (eID), των ηλεκτρονικών εγγράφων και της δυνατότητας επικοινωνίας μέσω ηλεκτρονικού ταχυδρομείου. Η τέταρτη κατηγορία Cross-Border Services υποδεικνύει σε ποιο βαθμό οι πολίτες άλλων χωρών της Ευρωπαϊκής Ένωσης μπορούν να χρησιμοποιούν τις υπηρεσίες ηλεκτρονικής διακυβέρνησης της Ελλάδας και περιλαμβάνει την online διαθεσιμότητα αυτών, την υποστήριξη χρηστών και αν μπορούν να χρησιμοποιηθούν εθνικά eID και ηλεκτρονικά έγγραφα από άλλες χώρες.



Εικόνα 4. Source: European Commission

Ενδιαφέρον παρουσιάζουν και τα στοιχεία της Eurostat για το πόσο οι πολίτες χωρών της Ευρωπαϊκής Ένωσης χρησιμοποιούν το διαδίκτυο για αλληλεπίδραση με τις δημόσιες αρχές δηλαδή τις υπηρεσίες ηλεκτρονικής διακυβέρνησης. Η Ελλάδα βρίσκεται κι εδώ κάτω από τον ευρωπαϊκό μέσο όρο ακόμα κι από χώρες όπως η Λιθουανία κι η Ουγγαρία. Δηλαδή εκτός από τη περαιτέρω βελτίωση των υπηρεσιών ηλεκτρονικής διακυβέρνησης είναι απαραίτητο να γίνουν και οι όποιες ενέργειες απαιτούνται από τη πολιτεία προκειμένου οι πολίτες να γνωρίσουν και να εξοικειωθούν με τη χρήση των υπηρεσιών αυτών.



Εικόνα 5. Source: Eurostat

2.7 Παράγοντες υιοθέτησης του Cloud Computing

Σε όλο τον κόσμο οι οργανισμοί του δημόσιου τομέα θεωρούνται λιγότερο ευνοϊκά προσκείμενοι στη χρήση τεχνολογιών πληροφορικής σε σχέση με τον ιδιωτικό κυρίως λόγω της γραφειοκρατίας και της κουλτούρας που υπάρχει σε αυτούς. Για αυτό το λόγο έχουν υπάρξει πολλές μελέτες τόσο σε ακαδημαϊκό όσο και εταιρικό επίπεδο με στόχο τη διερεύνηση των παραγόντων υιοθέτησης των τεχνολογιών πληροφορικής και ειδικά του Cloud Computing στους δημόσιους οργανισμούς. Οι περισσότερες μελέτες βασίζονται στα θεωρητικά πλαίσια του Technology-Organization-Environment (TOE) των Tornatzky and Fleischer, του Diffusion of Innovations (DOI) του Rogers ή συνδυασμό και των δυο. Τα τελευταία χρόνια έχουν αναπτυχθεί και άλλα θεωρητικά πλαίσια όπως το Desires Framework (DF) των Venters and Whitley.

Σύμφωνα με το TOE οι παράγοντες που επηρεάζουν έναν οργανισμό για το αν θα υιοθετήσει και εφαρμόσει καινοτομικές τεχνολογίες μπορούν να χωριστούν σε τρεις κατηγορίες. Στην τεχνολογική περιλαμβάνονται παράγοντες όπως η πολυπλοκότητα, διαθεσιμότητα, συμβατότητα, το κόστος, η ασφάλεια των νέων τεχνολογιών. Η οργανωτική αφορά τα χαρακτηριστικά και τους πόρους του οργανισμού όπως το μέγεθος του, την οργανωτική δομή, το ανθρώπινο δυναμικό. Η περιβαλλοντική κατηγορία περιλαμβάνει το μέγεθος και τη δομή του κλάδου, τους ανταγωνιστές του οργανισμού, το μακροοικονομικό και ρυθμιστικό πλαίσιο. Βέβαια κάποιοι μελετητές έχουν ασκήσει κριτική ότι το TOE δεν έχει εξελιχτεί όσο θα έπρεπε αλλά και το ότι είναι πολύ γενικό.

Το DOI αν και δημοσιεύτηκε για πρώτη φορά το 1962 από τον Rogers είναι το πιο γνωστό θεωρητικό πλαίσιο πάνω στο οποίο έχουν στηριχτεί χιλιάδες μελέτες πολλών επιστημονικών πεδίων. Εξηγεί πως, γιατί και σε ποιο βαθμό γενικότερα οι άνθρωποι, ως μέρος ενός κοινωνικού συστήματος, υιοθετούν μια νέα καινοτόμα ιδέα, συμπεριφορά ή προϊόν. Τα στάδια με τα οποία ένα άτομο ή ένας οργανισμός υιοθετεί μια καινοτομία περιλαμβάνουν την επίγνωση ανάγκης για καινοτομία, την απόφαση υιοθέτησης ή απόρριψης αυτής, τη δοκιμή της και τη συνεχή χρήση της. Υπάρχουν πέντε κύριοι παράγοντες που επηρεάζουν την υιοθέτηση μιας καινοτομίας: ο πρώτος και πιο σημαντικός είναι το συγκριτικό πλεονέκτημα(Relative Advantage) το πόσο καλύτερη θεωρείται η νέα ιδέα ή τεχνολογία σε σχέση με την παλιά, ο δεύτερος αφορά την συμβατότητα(Compatibility) δηλαδή σε ποιο βαθμό μπορεί η νέα τεχνολογία να χρησιμοποιηθεί από τα υπάρχοντα συστήματα, η πολυπλοκότητα(Complexity) πόσο δύσκολη είναι υλοποίηση και χρήση από το υπάρχων προσωπικό, η δοκιμαστική χρήση(Triability) της τεχνολογίας κατά πόσο δηλαδή μπορεί να δοκιμαστεί πριν υιοθετηθεί και η ορατότητα (Observability) των αποτελεσμάτων της χρήσης της καινοτομίας. Το DOI θεωρεί πως όλοι οι άνθρωποι τελικά μπορούν να αποδεχτούν μια νέα ιδέα αλλά διαχωρίζονται σε πέντε διαφορετικές κατηγορίες ανάλογα με τα χαρακτηριστικά τους: οι καινοτόμοι που είναι οι πρώτοι που θέλουν ακόμα και να ρισκάρουν να δοκιμάσουν ή να αναπτύξουν νέες ιδέες, οι πρώιμοι υιοθέτες που συνήθως έχουν ηγετικούς ρόλους και αγκαλιάζουν τις ευκαιρίες αλλαγής, η πρώιμη πλειοψηφία που υιοθετούν τις νέες ιδέες πριν από το μέσο άνθρωπο αν αποδεδειγμένα είναι αποτελεσματικές, η αργοπορημένη πλειοψηφία όπου ανήκουν οι εργαζόμενοι στο δημόσιο τομέα είναι οι δύσπιστοι άνθρωποι που αποδέχονται μια καινοτομία μόνο αν έχει υιοθετηθεί από τους περισσότερους και τέλος οι υστερούντες(laggards) που είναι η πιο δύσπιστη ομάδα οι οποίοι δεσμευμένοι από συντηρητικές παραδόσεις συνήθως αποδέχονται μια νέα ιδέα παρά μόνο μετά από φόβους, πιέσεις, στατιστικά στοιχεία.

Το DF δεν είναι τόσο γενικό όσο τα προηγούμενα δύο θεωρητικά πλαίσια και εστιάζει κυρίως στις επιθυμίες και τις πραγματικές ανάγκες που υπάρχουν σε έναν οργανισμό για να υιοθετηθεί μια καινοτόμα τεχνολογία όπως το Cloud Computing και αυτές χωρίζονται σε δύο πεδία: τεχνολογικό και την υπηρεσία. Το τεχνολογικό αφορά την επιθυμία για ισότητα ώστε ένα χαρακτηριστικό της τεχνολογίας να μην υστερεί ενός άλλου όπως είναι η ασφάλεια, η διαθεσιμότητα, οι μειωμένοι χρόνοι ανταπόκρισης. Την επιθυμία να μπορεί η καινοτομία να χρησιμοποιείται με διαφορετικούς τρόπους αλλά και χωρίς περιττή πολυπλοκότητα. Επίσης σημαντική είναι η επιθυμία για επεκτασιμότητα ώστε η τεχνολογία να μπορεί να ανταποκρίνεται στις διαρκώς μεταβαλλόμενες ανάγκες του οργανισμού. Το πεδίο της υπηρεσίας αφορά τις επιθυμίες η καινοτομία ως υπηρεσία να μπορεί να είναι οικονομικά αποτελεσματική, απλή για να είναι αποδοτική αλλά και να στηρίζει την εφευρετικότητα.

Πρόσφατες μελέτες έχουν δείξει πως η χρήση όχι μόνο ενός θεωρητικού πλαισίου αλλά συνδυασμό αυτών μπορεί να οδηγήσει σε πολύ καλύτερα αποτελέσματα για την έρευνα των παραγόντων που επηρεάζουν την υιοθέτηση νέων τεχνολογιών ειδικά στο δύσκολο τομέα της δημόσιας διακυβέρνησης. Για παράδειγμα τα παραπάνω τρία πλαίσια μπορούν να χρησιμοποιηθούν συμπληρωματικά δίνοντας έτσι μια πολύ καλύτερη, σαφή εικόνα και εξήγηση. Το TOE εστιάζεται κυρίως στους περιβαλλοντικούς και εσωτερικούς παράγοντες που μπορούν να επηρεάσουν την υιοθέτηση μιας καινοτομίας. Το DOI έχει πολύ καλύτερη θεωρητική, εμπειρική βάση και βασίζεται κυρίως στην εξήγηση των χαρακτηριστικών μιας καινοτομίας. Το DF εξηγεί συγκεκριμένα πως μια καινοτόμα τεχνολογία μπορεί να υιοθετηθεί σύμφωνα με τις ανάγκες και επιθυμίες ενός οργανισμού.

Benefit	TOE	DOI	DF	TOE + DOI + DF
Focus on the environment context of any new innovation technology adoption	✓			✓
Better able to explain intra-firm of any new innovation technology adoption	✓			✓
Focus on the innovation characteristics of any new innovation technology adoption		✓	✓	✓
Useful combination framework for studying a variety of IS adoption	✓	✓		✓
Empirical support and theoretical basis	✓	✓	✓	✓
Provides better understanding of the whole IS adoption phenomenon				✓

Εικόνα 6. Source: Ali, O., Shrestha, A., Osmanaj, V. and Muhammed, S. (2021)

2.8 Cloud Storage στην ελληνική τοπική αυτοδιοίκηση

Όπως αναφέρθηκε και προηγουμένως το επίπεδο των υπηρεσιών ηλεκτρονικής διακυβέρνησης και άρα κι η χρήση καινοτόμων τεχνολογιών στον ελληνικό δημόσιο τομέα όπως το Cloud Computing κυμαίνονται κάτω από το μέσο ευρωπαϊκό όρο. Παρόλα αυτά τα τελευταία χρόνια βρίσκεται σε εξέλιξη η υλοποίηση ενός εγχειρήματος της ελληνικής Δημόσιας Διοίκησης που θεωρείται πρωτοποριακό και καινοτόμο το G-Cloud ως μέρος της Εθνικής Ψηφιακής Στρατηγικής. Πρόκειται για μία σύγχρονη δημόσια υποδομή κέντρου δεδομένων υπηρεσιών private cloud και public cloud βασιζόμενο στο μοντέλο IaaS, SaaS και τεχνολογίες virtualization για την υποστήριξη των πληροφοριακών συστημάτων της γενικής και τοπικής κυβέρνησης. Στόχος είναι η μείωση του κόστους με οικονομίες κλίμακας και εξοικονόμηση πόρων, υψηλού επιπέδου ασφάλεια, εγγυημένη διαθεσιμότητα και επεκτασιμότητα, με απώτερο σκοπό τη βελτίωση της παραγωγικότητας του Δημοσίου και άρα και των παρεχόμενων υπηρεσιών προς τους πολίτες. Μέχρι το τέλος του 2021 φιλοξενούσε περίπου 250 πληροφοριακά συστήματα δημόσιων φορέων.

Όσον αφορά τις μελέτες για το Cloud Computing στην τοπική αυτοδιοίκηση είναι γεγονός ότι υπάρχουν σχετικά λίγες σε διεθνές επίπεδο πόσο μάλλον για την ελληνική. Μία από αυτές τις ελάχιστες είναι των Kyriakou Niki, Loukis Euripides and Dimitropoulou Paraskevi “Factors affecting cloud storage adoption by Greek municipalities” της οποίας τα δεδομένα θα χρησιμοποιήσουμε για την εκπαίδευση και αξιολόγηση προγνωστικών αλγοριθμικών μοντέλων σε επόμενη ενότητα. Η έρευνα διενεργήθηκε με τη μορφή ερωτηματολογίου που στάλθηκε στους 332 δήμους της Ελλάδας και τη στατιστική ανάλυση των απαντήσεων από τους 121 δήμους που ανταποκρίθηκαν. Θετικό είναι το γεγονός ότι το 23,1% χρησιμοποιεί ήδη τεχνολογίες Cloud Storage ενώ το 38,8% σχεδιάζει να αποκτήσει στο άμεσο μέλλον. Από τα συμπεράσματα προκύπτει ότι οι δήμοι θεωρούν σημαντικά τα οφέλη της μείωσης του λειτουργικού κόστους, της αύξησης της παραγωγικότητας και τη βελτίωση της συνεργασίας με άλλους δημόσιους φορείς χωρίς όμως αυτά να τα θεωρούν συγκριτικό πλεονέκτημα της τεχνολογίας δηλαδή παράγοντες ικανούς για την υιοθετήσουν, κάτι που αποτελεί έκπληξη δεδομένων των συνθηκών λιτότητας των τελευταίων χρόνων. Αντίθετα τέτοιοι παράγοντες είναι οι δυσκολίες που προκύπτουν από την υιοθέτηση του Cloud Storage όπως η απώλεια δεδομένων ή πρόσβαση σε αυτά από μη εξουσιοδοτημένους και η συμβατότητα με τα υπάρχοντα συστήματα που περιλαμβάνουν τις υλικές υποδομές και τις δεξιότητες του προσωπικού.

Μια άλλη μελέτη είναι αυτή του Νάνου Ιωάννη “Η υιοθέτηση της υπολογιστικής νέφους (cloud computing) στην τοπική αυτοδιοίκηση”. Διενεργήθηκε με τη μορφή ερωτηματολογίου που απάντησαν 211 δήμοι στις απαντήσεις των οποίων έγινε στατιστική ανάλυση. Το 23% χρησιμοποιεί ήδη την τεχνολογία, 57% είναι στα σχέδιά τους για υιοθέτηση ενώ υπάρχει κι ένα 18% που δεν ενδιαφέρεται καθόλου να την αποκτήσει. Οι υπηρεσίες cloud που φαίνεται να χρησιμοποιούνται πιο συχνά είναι φιλοξενία ιστοσελίδων, email, εφαρμογές διαμοιρασμού αρχείων και αποθήκευσης δεδομένων. Από τους δήμους που χρησιμοποιούν υπηρεσίες cloud μόνο το 13% έχουν ψηφιακή στρατηγική και το 16% έχουν ενταχθεί στο G-Cloud. Τα συμπεράσματα αυτής της έρευνας συμφωνούν με αυτά της προηγούμενης προσθέτοντας συμπληρωματικά επιπλέον στοιχεία όπως τους σημαντικούς παράγοντες υιοθέτησης του ρυθμιστικού πλαισίου, το μέγεθος του οργανισμού αλλά και λιγότερο σημαντικούς όπως το γραφειοκρατικό περιβάλλον και την υποστήριξη της ανώτατης διοίκησης. Ο παράγοντας της δοκιμαστικής χρήσης δε φαίνεται να ασκεί καμία επίδραση.

3. Προγνωστική Αναλυτική

3.1 Εισαγωγή στην Προγνωστική Αναλυτική

Σε όλη τη διάρκεια της ιστορίας τους οι άνθρωποι προσπαθούσαν να βρίσκουν τρόπους πρόβλεψης του μέλλοντος κυρίως για να αποφύγουν δυσάρεστες καταστάσεις ή να αποκτήσουν στρατηγικό πλεονέκτημα έναντι άλλων. Οι Μεσοποτάμιοι πίστευαν ότι μπορούσαν να προετοιμαστούν για το μέλλον μελετώντας τους οιωνούς των θεών στα εντόσθια των ζώων και στη παρατήρηση των ουράνιων σωμάτων. Οι αρχαίοι Έλληνες θεωρούσαν κέντρο του τότε γνωστού κόσμου το μαντείο των Δελφών το οποίο επισκέπτονταν τόσο απλοί πολίτες όσο και κορυφαίοι πολιτικοί και στρατιωτικοί ηγέτες προκειμένου να ζητήσουν τις κρυπτικές συμβουλές του για το μέλλον και οι οποίες πολλές φορές έπαιξαν καθοριστικό ρόλο στη λήψη σημαντικών ιστορικών αποφάσεων και γεγονότων. Όμως όλα αυτά καμία σχέση δεν είχαν με επιστήμες. Μία από τις πρώτες επιστημονικές μεθόδους πρόβλεψης μελλοντικών καταστάσεων θεωρείται πως είναι ο περίφημος Μηχανισμός των Αντικυθήρων ο αρχαιότερος μηχανικός, αναλογικός υπολογιστής σχεδιασμένος για να προβλέπει τις κινήσεις των ουράνιων σωμάτων και ειδικά του ηλίου και της Σελήνης με εντυπωσιακή λεπτομέρεια. Θεωρείται δε ένα από τα μεγαλύτερα μυστήρια διότι αν και οι αρχαίοι Έλληνες είχαν το θεωρητικό υπόβαθρο δεν είχαν όμως και την τεχνολογία για την υλοποίηση μιας τέτοιας κατασκευής. Πιο πρόσφατα το 1940 κατά τη διάρκεια του δεύτερου παγκοσμίου πολέμου ο Νόρμπερτ Βίνερ, κορυφαίος μαθηματικός και θεμελιωτής της κυβερνητικής, προσπάθησε να κάνει πιο αποτελεσματικά τα συμμαχικά αντιαεροπορικά όπλα προβλέποντας την μελλοντική θέση των γερμανικών αεροπλάνων σύμφωνα με την μέχρι τότε πορεία τους και τους πιθανούς ελιγμούς των πιλότων.

Στη σημερινή ψηφιακή εποχή πλέον, η οποία εν πολλοίς οφείλεται στην εκρηκτική εξέλιξη της πληροφορικής, υπάρχουν πολύ καλύτερα και πιο αξιόπιστα εργαλεία πρόβλεψης μελλοντικών καταστάσεων συνεπικουρούμενα από μια σχετικά νέα επιστήμη την προγνωστική αναλυτική. Σύμφωνα με τον Eric Siegel, διακεκριμένο επιστήμων της μηχανικής μάθησης και προγνωστικής αναλυτικής, ο ορισμός που θα μπορούσαμε να δώσουμε για αυτήν την επιστήμη είναι ο εξής:

- Προγνωστική Αναλυτική - Τεχνολογία που μαθαίνει από την εμπειρία (δεδομένα) για την πρόβλεψη μελλοντικών συμπεριφορών οντοτήτων με στόχο τη λήψη καλύτερων αποφάσεων.

Ο όρος «οντότητες» μπορεί να περιλαμβάνει ανθρώπους, οργανισμούς, εταιρείες, προϊόντα, μηνύματα σε κοινωνικά δίκτυα, κυριολεκτικά οτιδήποτε μπορεί να προβλεφθεί σε μια εποχή που υπολογίζεται ότι καθημερινά παράγονται τουλάχιστον 2.5 πεντάκις εκατομμύρια bytes δεδομένων. Το αποτέλεσμα είναι να πρόκειται πλέον για μια περιζήτητη επιστήμη σε πολλούς τομείς όπως ο τραπεζικός τομέας, ο υγειονομικός τομέας, στην έρευνα, στην πολιτική, στην εγκληματολογία. Η McKinsey, μία από τις μεγαλύτερες εταιρείες παροχής συμβουλευτικών υπηρεσιών στον κόσμο, αναφέρει πως το 2020 οι προσφερόμενες θέσεις για αναλυτές τριπλασιάστηκαν σε σχέση με το 2013 ενώ αναμένεται ότι μόνο στις Η.Π.Α. θα υπάρξει στο άμεσο μέλλον ανάγκη για κάλυψη εκατοντάδων χιλιάδων κενών θέσεων ατόμων με ικανότητες όχι μόνο να αναλύουν μεγάλες ποσότητες δεδομένων(big data) αλλά και να παίρνουν αποφάσεις με βάση αυτές.

3.2 Κατηγορίες αναλυτικής

Επειδή όπως αναφέραμε η προγνωστική αναλυτική είναι σχετικά νέα επιστήμη κάποια χαρακτηριστικά και έννοιές της συγχέονται με άλλες επιστήμες. Καταρχήν πολλοί άνθρωποι βλέποντας τον όρο «αναλυτική» πιστεύουν ότι πρόκειται για κλάδο της στατιστικής. Η πραγματικότητα είναι όμως ότι η στατιστική αποτελεί ένα μόνο κομμάτι της αναλυτικής και είναι κυρίως ένα νοητικό πλαίσιο, μια υπολογιστικά εξελικτική πρόοδος που εκτός από επιστήμη είναι και τέχνη. Για να κατανοήσουμε καλύτερα τι σημαίνει προγνωστική αναλυτική αναφέρουμε τις τέσσερις κατηγορίες αναλυτικής στις οποίες η μια συμπληρώνει την άλλη:

I. Περιγραφική. Αποτελεί την πιο απλή και ευρέως χρησιμοποιούμενη αναλυτική αλλά και τη βάση πάνω στην οποία στηρίζονται οι υπόλοιπες κατηγορίες. Απαντά στο ερώτημα «Τι συνέβη;» δηλαδή περιγράφει με σαφήνεια καταστάσεις και γεγονότα του παρελθόντος ή του παρόντος. Συνήθως χρησιμοποιούνται οπτικές αναπαραστάσεις των δεδομένων όπως γραφήματα, διαγράμματα και χάρτες.

II. Διαγνωστική. Συμπληρώνει την περιγραφική απαντώντας στο «γιατί» συγκρίνοντας τάσεις και προβάλλοντας τις φανερές ή κρυφές αιτιακές συσχετίσεις μεταξύ των μεταβλητών. Σε ένα απλό παράδειγμα της αύξησης των πωλήσεων ενός προϊόντος μιας επιχείρησης η περιγραφική ανάλυση θα παρουσίαζε τον αριθμό των πωλήσεων ενώ η διαγνωστική θα συμπέραινε ότι η αύξηση αυτή οφείλεται σε κάποια καλή έκπτωση.

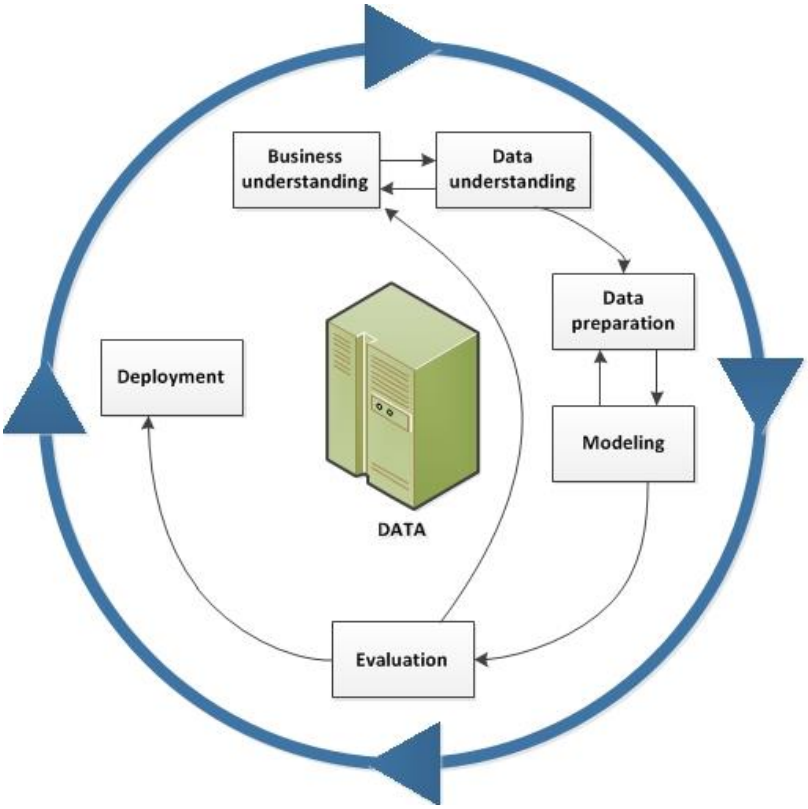
III. Προγνωστική. Πολλές φορές οι επιστήμες της μηχανικής μάθησης (machine learning) και εξόρυξης δεδομένων (data mining) συγχέονται ή ακόμα και εξομοιώνονται με την προγνωστική αναλυτική. Αυτό που συμβαίνει όμως είναι ότι η προγνωστική αναλυτική χρησιμοποιεί τεχνικές εξόρυξης δεδομένων για την ανακάλυψη κρυφών προτύπων δεδομένων μέσω της μηχανικής μάθησης και η γνώση που θα προκύψει από την ανάλυση αυτή να χρησιμοποιηθεί για τεκμηριωμένες προβλέψεις των μελλοντικών τάσεων με στόχο τη διαμόρφωση καλύτερων στρατηγικών.

IV. Καθοδηγητική. Αποτελεί το επόμενο λογικό στάδιο της προγνωστικής αναλυτικής κατά το οποίο τεχνητή νοημοσύνη και big data συνδυάζονται για να ληφθούν υπόψη όλοι οι πιθανοί παράγοντες σε ένα σενάριο και να προταθούν οι κατάλληλες πρακτικές λύσεις για τη λήψη βέλτιστων αποφάσεων.

Εικόνα 7. Source: Eric Siegel (2013)

3.3 CRISP-DM

Η επιστήμη της εξόρυξης δεδομένων αποτελεί βασικό πυλώνα της προγνωστικής αναλυτικής. Πολλές φορές όμως κάθε εταιρεία και οργανισμός ανέπτυσσε και χρησιμοποιούσε τη δικιά της μεθοδολογία εξόρυξης δεδομένων δαπανώντας πολύτιμο χρόνο και κόπο για να συμβεί αυτό. Έτσι το 1999 με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης αναλυτές από διάφορες εταιρείες παρουσίασαν από κοινού μια πρότυπη ελεύθερα διαθέσιμη διαδικασία την Cross-Industry Standard Process for Data Mining (CRISP-DM) σύμφωνα με την οποία κάθε επιχείρηση ή οργανισμός μπορεί να προσαρμόσει την εξόρυξη δεδομένων -άρα και την προγνωστική αναλυτική της- ανάλογα με προβλήματα και τις ανάγκες της. Το 2015 η IBM παρουσίασε μια νέα μεθοδολογία βασισμένη στη CRISP-DM την Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) ως μέρος του λογισμικού αναλυτικής IBM SPSS. Όμως το CRISP-DM εξακολουθεί να είναι η πιο ευρέως χρησιμοποιούμενη πρότυπη διαδικασία εξόρυξης δεδομένων ως μεθοδολογία για την περιγραφή, τις ενέργειες και τις σχέσεις μεταξύ των σταδίων ενός έργου αλλά και ως μοντέλου διαδικασίας για τη συνολική επισκόπηση του κύκλου ζωής του. Αυτό οφείλεται κυρίως γιατί μπορεί να εφαρμοστεί σε οποιοδήποτε βιομηχανικό κλάδο, εργαλείο ή εφαρμογή.



Εικόνα 8. Source: www.ibm.com

Σύμφωνα με το CRISP-DM κάθε διαδικασία αναλυτικής περιλαμβάνει έξι στάδια η σειρά των οποίων δεν είναι αυστηρά καθορισμένη αλλά ευέλικτη ανάλογα με τις ανάγκες και τα προβλήματα που προκύπτουν:

1. Business Understanding. Διατύπωση με σαφήνεια των εταιρικών στόχων και των σκοπών για τους οποίους θα χρησιμοποιηθούν οι διάφορες τεχνικές αναλυτικής. Θα πρέπει όλες οι εμπλεκόμενες πλευρές να ξέρουν μέσω συζητήσεων και ερωτήσεων ακριβώς ποιοι είναι αυτοί οι στόχοι, ενδεχόμενα προβλήματα καθώς και στρατηγικές επίλυσης αυτών, δηλαδή με άλλα λόγια το στάδιο αυτό αποτελεί τον «οδικό χάρτη» (road map) του έργου.
2. Data Understanding. Τα δεδομένα συλλέγονται από διάφορες πηγές και γίνεται διερευνητική ανάλυση αυτών με σκοπό την αξιολόγηση της ποιότητας και της ποσότητάς τους. Αποτελεί ένα ιδιαίτερα σημαντικό στάδιο καθώς εάν γίνει σωστά μπορούν να αποφευχθούν προβλήματα στα επόμενα στάδια όπως για παράδειγμα εάν εντοπιστούν λανθασμένες ή ελλιπείς τιμές σε κάποιες εγγραφές, ασυνέπειες στις στήλες ή μεταξύ σετ δεδομένων, ανεξάρτητες μεταβλητές που δεν προσφέρουν τίποτα για το συγκεκριμένο έργο και πρέπει να αφαιρεθούν.
3. Data Preparation. Θεωρείται το πιο σημαντικό και χρονοβόρο στάδιο της αναλυτικής που μπορεί να καταλάβει ακόμα και το 70% του συνολικού χρόνου διεκπεραίωσης του έργου κάτι το οποίο εξαρτάται από το πόσο καλή δουλειά έγινε στα δυο προηγούμενα στάδια. Γίνεται προετοιμασία και επεξεργασία των δεδομένων για τη χρήση αυτών από τις διάφορες τεχνικές μοντελοποίησης και συνήθως περιλαμβάνει τις εξής διεργασίες: συγχώνευση σετ δεδομένων και εγγραφών, επιλογή δείγματος από το σύνολο των δεδομένων, ομαδοποίηση και ταξινόμηση εγγραφών, τροποποίηση ή δημιουργία νέων μεταβλητών, αφαίρεση ή αντικατάσταση των κενών και προβληματικών τιμών των μεταβλητών, διαχωρισμός των δεδομένων σε σετ εκπαίδευσης και αξιολόγησης (training and testing sets).
4. Modeling. Εφαρμογή των διαθέσιμων και κατάλληλων μοντέλων αναλυτικής στα επεξεργασμένα δεδομένα του προηγούμενου σταδίου με τις απαραίτητες ρυθμίσεις. Είναι σπάνιο αυτό το στάδιο να πραγματοποιηθεί με ένα μόνο μοντέλο και μία εκτέλεση. Συνήθως οι αναλυτές χρησιμοποιούν πολλαπλά αναλυτικά μοντέλα και παραμέτρους ή ακόμα και να επιστρέψουν στο προηγούμενο στάδιο της επεξεργασίας για τη βελτιστοποίηση των αποτελεσμάτων.
5. Evaluation. Αξιολόγηση των ευρημάτων αλλά και των τεχνικών μοντελοποίησης για την αποτελεσματικότητα και ακρίβειά τους. Διερεύνηση αν τα ευρήματα ανταποκρίνονται στους αρχικούς στόχους που είχαν τεθεί κι αν όχι επιστροφή σε προηγούμενο στάδιο ακόμα και στο πρώτο αν χρειαστεί.
6. Deployment. Η επιτυχής δημιουργία ενός ή περισσότερων αναλυτικών μοντέλων δεν αρκεί αλλά θα πρέπει να εφαρμοστεί και σε νέα δεδομένα. Επίσης το στάδιο αυτό περιλαμβάνει τη συνεχή παρακολούθηση υλοποίησης των αποτελεσμάτων και συμπερασμάτων καθώς και τη διενέργεια άλλων απλών εργασιών όπως τη διεξαγωγή έκθεσης και κριτικής του έργου ή πιο περίπλοκων διεργασιών όπως την υλοποίηση του υφιστάμενου αναλυτικού μοντέλου σε άλλο τμήμα της εταιρείας ή του οργανισμού.

3.4 Τεχνικές Προγνωστικής Αναλυτικής

Σε γενικές γραμμές και με απλά λόγια ένα μοντέλο προγνωστικής αναλυτικής λειτουργεί χρησιμοποιώντας αλγοριθμικές τεχνικές εξόρυξης γνώσης, μηχανικής μάθησης, τεχνητής νοημοσύνης και στατιστικής ώστε να καταλήξει σε έναν αριθμό για ένα σενάριο δεδομένων που περιλαμβάνουν ανεξάρτητες μεταβλητές και συνήθως μίας εξαρτημένη μεταβλητή. Όσο μεγαλύτερος είναι αυτός ο αριθμός τόσο πιο πιθανό είναι να συμβεί αυτό το σενάριο δηλαδή η εξαρτημένη μεταβλητή να έχει μια συγκεκριμένη τιμή με βάση τις τιμές των ανεξάρτητων μεταβλητών. Στη συνέχεια αυτός ο αριθμός ή σκορ χρησιμοποιείται από την εταιρεία, τον οργανισμό ή την ομάδα για να ληφθούν οι όποιες αποφάσεις χρειάζονται για την επίτευξη των σκοπών του έργου. Η προγνωστική αναλυτική δηλαδή όπως και όλες οι κατηγορίες αναλυτικής δεν είναι μια αυτοματοποιημένη διαδικασία που συμμετέχουν μόνο αλγόριθμοι αλλά σημαντικό ρόλο παίζει κι ο ανθρώπινος παράγοντας. Να σημειώσουμε εδώ πως γενικά οι ανεξάρτητες μεταβλητές/χαρακτηριστικά των δεδομένων μπορεί να μην είναι ανεξάρτητες και να υπάρχουν περισσότερες από μία εξαρτημένες μεταβλητές/κλάσεις. Τα δεδομένα όμως που θα χρησιμοποιήσουμε σε αυτή την εργασία έχουν ανεξάρτητες μεταβλητές και μία εξαρτημένη μεταβλητή οπότε και η θεωρητική μελέτη θα γίνει με βάση αυτό το σενάριο.

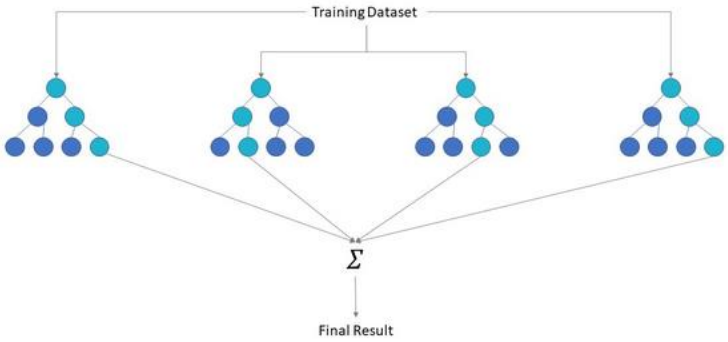
Τα μοντέλα προγνωστικής αναλυτικής μπορούν να χωριστούν σε δυο μεγάλες κατηγορίες: εποπτευόμενης μάθησης (supervised learning), που είναι τα πιο ευρέως γνωστά, τα οποία εκπαιδεύονται χρησιμοποιώντας δεδομένα με γνωστές εξαρχής τις ανεξάρτητες και εξαρτημένες μεταβλητές για την πρόβλεψη νέων εξαρτημένων μεταβλητών και τα μη εποπτευόμενης μάθησης (unsupervised learning) τα οποία μπορεί να μη γνωρίζουν ούτε τις εξαρτημένες μεταβλητές ούτε τις τιμές αυτών και είναι δουλειά των μοντέλων να το κάνουν αυτό. Τα μοντέλα εποπτευόμενης μάθησης χωρίζονται σε δυο κατηγορίες αναλόγως των προβλημάτων που χρησιμοποιούνται για να λύσουν: κατηγοριοποίησης (classification) που αφορά την πρόβλεψη μιας εξαρτημένης μεταβλητής με διακριτές τιμές στα οποία η αποτελεσματικότητα των μοντέλων κρίνεται κυρίως με βάση την ακρίβεια (accuracy) των σωστά ταξινομημένων παραδειγμάτων από όλες τις προβλέψεις που έγιναν και τα παλινδρόμησης (regression) που αφορά την πρόβλεψη μιας εξαρτημένης μεταβλητής με συνεχόμενες τιμές και η αποτελεσματικότητα των οποίων κρίνεται συνήθως με υπολογισμό της ρίζας του μέσου του τετραγώνου του σφάλματος (RMSE). Τα μοντέλα μη εποπτευόμενης μάθησης συχνά χρησιμοποιούνται για τρεις κατηγορίες προβλημάτων: ομαδοποίησης (clustering) για την ομαδοποίηση δεδομένων με βάση τις ομοιότητες ή διαφορές τους, συσχέτισης (association) που χρησιμοποιεί κανόνες για την εύρεση σχέσεων μεταξύ των μεταβλητών και μείωσης των διαστάσεων (dimensionality reduction) για τη μείωση του πολύ μεγάλου αριθμού χαρακτηριστικών των δεδομένων. Βέβαια υπάρχουν και προβλήματα που δεν ανήκουν μόνο στη μία ή άλλη κατηγορία οπότε μπορεί να χρειαστεί ένας συνδυασμός μοντέλων (semi-supervised). Στην εργασία αυτή πάντως θα χρησιμοποιηθούν, όπως θα δούμε στην επόμενη ενότητα, αλγοριθμικά μοντέλα εποπτευόμενης μάθησης και κατηγοριοποίησης μιας και η υπό πρόβλεψη εξαρτημένη μεταβλητή που μας ενδιαφέρει έχει διακριτές τιμές.

Παρακάτω παρουσιάζονται συνοπτικά οι πιο ευρέως διαδομένες αλγοριθμικές τεχνικές επιβλεπόμενης ή μη μηχανικής μάθησης και προγνωστικής αναλυτικής.

Κατηγοριοποίηση Bayes. Αποτελεί μία από τις πιο απλές τεχνικές μοντελοποίησης που βασίζεται στη στατιστική και συγκεκριμένα στο θεώρημα του Bayes το οποίο υπολογίζει την πιθανότητα τιμής της εξαρτημένης μεταβλητής μιας νέας εγγραφής δεδομένου των πιθανοτήτων των τιμών αυτής για τις παλιές εγγραφές. Είναι γρήγορη καθότι χρειάζεται μόνο μια επεξεργασία των δεδομένων εκπαίδευσης και μπορεί να χειριστεί ελλιπή δεδομένα παραλείποντας τις αντίστοιχες πιθανότητες. Όμως η ακρίβεια των αποτελεσμάτων είναι σχετικά μικρή ιδίως όταν πρέπει να ληφθούν υπόψη πολλά χαρακτηριστικά (κατάρα της διαστατικότητας).

Δέντρα Απόφασης. Θεωρείται το πιο γνωστό και ευρέως χρησιμοποιούμενο μοντέλο μηχανικής μάθησης στην προγνωστική αναλυτική λόγω της εύκολης υλοποίησης και σχετικά ακριβών προγνωστικών αποτελεσμάτων. Εφαρμόζεται κυρίως σε προβλήματα κατηγοριοποίησης και λιγότερο σε προβλήματα παλινδρόμησης. Μπορούν εύκολα να ενσωματωθούν σε άλλα μοντέλα δέντρων αποφάσεων και να δίνουν απαντήσεις σε σχεδόν όλες τις περιπτώσεις αλλά τυχόν μικρές αλλαγές ή ελλείψεις στις τιμές των μεταβλητών μπορούν να οδηγήσουν σε μεγάλες αλλαγές στη δομή των δέντρων. Ονομάζονται έτσι γιατί σχεδιαγραμματικά εμφανίζονται ως δέντρα. Ο αρχικός κόμβος (ρίζα) είναι η πιο «αγνή» ανεξάρτητη μεταβλητή της οποίας οι διαφορετικές τιμές αντιστοιχούν σε όσο το δυνατόν λιγότερες διαφορετικές τιμές της εξαρτημένης μεταβλητής. Αυτό όπως και οι κόμβοι που ακολουθούν εξαρτάται από τις μαθηματικές τεχνικές του εκάστοτε αλγόριθμου. Οι κόμβοι-παιδιά συνδέονται με τους γονικούς κόμβους μέσω των κλαδιών δηλαδή των τιμών των γονικών ανεξάρτητων μεταβλητών. Οι τελευταίοι ή εξωτερικοί κόμβοι των υπό-δέντρων λέγονται φύλλα, περιλαμβάνουν τις τιμές της εξαρτημένης μεταβλητής και αποτελούν τις αποφάσεις του μοντέλου.

Μέθοδοι Ensemble. Πρόκειται για μοντέλα που συνδυάζουν λογαριθμικά άλλα παρόμοια μοντέλα επιβλεπόμενης μηχανικής μάθησης συνήθως για προβλήματα κατηγοριοποίησης. Συγκριτικά με άλλες τεχνικές μπορούν να χρησιμοποιηθούν σε σύνολα δεδομένων με πολλές ανεξάρτητες μεταβλητές χωρίς να παρουσιάσουν σημαντικά προβλήματα. Παράλληλα εάν τα επιμέρους μοντέλα είναι ανεξάρτητα και εμφανίζουν ικανοποιητική ποικιλία ως προς τι τιμές των χαρακτηριστικών τους τότε μαθηματικά αποδεικνύεται ότι οι προβλέψεις του τελικού μοντέλου ensemble θα έχουν πολύ καλύτερη ακρίβεια. Αυτό γίνεται είτε με μεθόδους υπολογισμού των μέσων όρων των προβλέψεων των επιμέρους μοντέλων είτε με μεθόδους εκπαίδευσης αδύνατων μοντέλων σε τροποποιημένα δεδομένα.



Εικόνα 9. Source: https://www.ibm.com/cloud/learn/

Μοντέλα Παλινδρόμησης. Σε αυτά τα γνωστά μοντέλα χρησιμοποιούνται στατιστικοί και γραμμικοί ή μη μαθηματικοί μέθοδοι μοντελοποίησης των σχέσεων μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών των οποίων οι τιμές είναι συνεχείς. Το μεγάλο πλεονέκτημα αυτών των μοντέλων είναι ότι μπορούν να δείξουν όχι μόνο τι σχέση έχουν οι ανεξάρτητες μεταβλητές με τις εξαρτημένες αλλά και ποιες από τις πρώτες επιδρούν περισσότερο στις δεύτερες.

Κ Κοντινότεροι Γείτονες (k-NN). Απλή τεχνική μοντελοποίησης η οποία εκτός από το σύνολο των δεδομένων απαιτεί και την επιθυμητή κατηγοριοποίηση των k εγγραφών με βάση την απόσταση των μεταβλητών, δηλαδή την ομοιότητα των τιμών. Όταν προστίθενται νέες εγγραφές χρειάζεται να γίνεται υπολογισμός νέων αποστάσεων και του νέου k οπότε αυτή η τεχνική έχει υψηλό κόστος υπολογισμού ιδιαίτερα όταν πρόκειται για πολλά χαρακτηριστικά. Ο αλγόριθμος δεν μαθαίνει εξαρχής από κάποια δεδομένα εκπαίδευσης αλλά εκτελείται για κάθε νέα εγγραφή για αυτό και χαρακτηρίζεται κι ως lazy learner.

Νευρωνικά Δίκτυα. Προσομοιώνει τη λειτουργία των βιολογικών νευρώνων και αποτελεί δυνητικά ένα από τα πιο περίπλοκα αλλά και αποδοτικά μοντέλα επιβλεπόμενης ή μη μηχανικής μάθησης προβλημάτων κατηγοριοποίησης ή και παλινδρόμησης. Κάθε νευρώνας-κόμβος μπορεί να δέχεται πολλαπλές εισόδους διαφορετικών βαρών και να δίνει μια έξοδο μόνο εάν το ζυγισμένο άθροισμα είναι μεγαλύτερο μια επιθυμητής τιμής. Μπορούν να υπάρχουν πολλαπλά επίπεδα νευρώνων τα οποία είναι δυνατό να επικοινωνούν μεταξύ τους μη γραμμικά. Αν και αυτό αυξάνει την περιπλοκότητα εντούτοις καθιστά αυτά τα μοντέλα πολύ ανθεκτικά σε προβλήματα όπως σφάλματα τιμών και για αυτό η ακρίβεια των αποτελεσμάτων είναι υψηλή. Εκτός από την πολυπλοκότητά τους στα μειονεκτήματά τους συγκαταλέγεται και το γεγονός ότι οι είσοδοι πρέπει να κωδικοποιούνται μεταξύ 0 και 1 για την απλοποίηση και επιτάχυνση των υπολογισμών.

Support Vector Machines. Μοντέλο επιβλεπόμενης μάθησης κατηγοριοποίησης ή παλινδρόμησης που βασίζεται στην εύρεση ενός γραμμικού ή μη ορίου απόφασης (hyperplane) για τη μεγιστοποίηση του διαχωρισμού των δεδομένων σε κατηγορίες σύμφωνα πάντα με τις τιμές της κλάσης. Οπότε για τα νέα δεδομένα η κλάση τους προβλέπεται αναλόγως σε ποια πλευρά του hyperplane ανήκουν. Τα αντικείμενα εκείνα που επηρεάζουν σημαντικά τη διαμόρφωση της hyperplane αποτελούν τα διανύσματα υποστήριξης (support vectors).

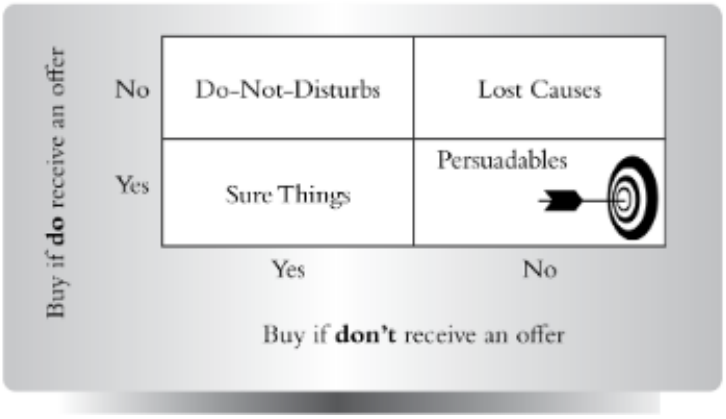
Γενετικοί Αλγόριθμοι. Άλλη μια τεχνική μηχανικής μάθησης που βασίζεται στην προσομοίωση βιολογικής διαδικασίας και συγκεκριμένα της φυσικής επιλογής δηλαδή της δημιουργίας και εξέλιξης αποδοτικών απογόνων που στη περίπτωση της αναλυτικής αφορά τις όσο δυνατό καλύτερες απαντήσεις ή προβλέψεις. Η γενική ιδέα είναι ότι σε ένα πληθυσμό πιθανών λύσεων εφαρμόζονται επαναληπτικά διάφορες τεχνικές όπως της γενετικής επιλογής, διασταύρωσης και μετάλλαξης ελέγχοντας σε κάθε στάδιο την επιθυμητή απόδοση των λύσεων με μια συνάρτηση καταλληλότητας f(x). Μπορούν να συνδυαστούν ιδανικά με τα μοντέλα νευρωνικών δικτύων για την εύρεση των επιθυμητών βαρών των εισόδων σε αυτά.

Time Series Analysis. Αφορά τη στατιστική ανάλυση των προτύπων που προκύπτουν από την αποτύπωση γεγονότων σε τακτά χρονικά διαστήματα και περιλαμβάνει τόσο τη συχνότητα με την οποία αυτά συμβαίνουν όσο και τη χρονική περίοδο. Είναι μια σχετικά απλή προγνωστική μέθοδος συγκριτικά με άλλες και για αυτό χρησιμοποιείται συχνά, ιδιαίτερα για προβλέψεις των χρηματιστηριακών τάσεων και των καιρικών συνθηκών.

3.5 Προγνωστική Αναλυτική στην πράξη

Η προγνωστική αναλυτική χρησιμοποιείται πλέον σε πολλούς τομείς της ζωής για την πρόβλεψη μελλοντικών καταστάσεων όπως για παράδειγμα την πιθανές ανάγκες της τοπικής αυτοδιοίκησης σε cloud storage όπως θα δούμε σε επόμενη ενότητα. Τα τελευταία χρόνια έχει αναπτυχθεί μια εξελιγμένη τεχνική της προγνωστικής αναλυτικής η Uplift Modeling όχι μόνο για την πρόβλεψη καταστάσεων αλλά και το βαθμό που αυτές μπορούν να επηρεάσουν συγκεκριμένες οντότητες που συνήθως είναι οι άνθρωποι. Για το σκοπό αυτό χρησιμοποιούνται δύο ομάδες σετ δεδομένων: μια για τους ανθρώπους εκείνους στους οποίους έχει εφαρμοστεί μια τεχνική επηρεασμού και μια για εκείνους που έχουν μείνει ανεπηρέαστοι. Στόχος είναι να βρεθούν τα άτομα εκείνα τα οποία δεν έχουν ακόμα πάρει κάποια απόφαση για το αν για παράδειγμα θέλουν να αγοράσουν ένα προϊόν και μπορούν να επηρεαστούν αλλά και να αποφευχθεί το κόστος στοχευμένων προωθητικών τακτικών σε εκείνους που έχουν αποφασίσει σίγουρα να το αγοράσουν ή όχι. Το αποτέλεσμα είναι μεγάλες περικοπές στα κόστη και παράλληλη αύξηση των κερδών εφαρμόζοντας απλά Uplift Modeling στην οποία συχνά χρησιμοποιούνται μέθοδοι Ensemble.

Ένα χαρακτηριστικό παράδειγμα Uplift Modeling αποτελεί η περίπτωση της U.S. Bank στα μέσα της δεκαετίας του ’90. Οι τότε διαφημιστικές ενέργειες για χρηματοπιστωτικά προϊόντα μέσω ταχυδρομείου φαινόταν να έχουν καλά αποτελέσματα μιας το ποσοστό των ανθρώπων που ανταποκρίνονταν σε αυτές ήταν ικανοποιητικό. Τούτο όμως είναι ένα λάθος το οποίο συμβαίνει ακόμα και σήμερα διότι αυτό που ενδιαφέρει κυρίως είναι ποιοι επηρεάστηκαν από τις προωθητικές ενέργειες και τελικά αγόρασαν το προϊόν. Μάλιστα σε κάποιους που ήταν ήδη πελάτες φάνηκε πως οι διαφημιστικές ενέργειες είχαν ακριβώς τα αντίθετα αποτελέσματα. Όταν εφαρμόστηκαν τεχνικές προγνωστικής αναλυτικής για τη στόχευση συγκεκριμένων υποψήφιων πελατών τα αποτελέσματα ήταν εντυπωσιακά: 300% αύξηση των εσόδων, 40% μείωση του κόστους της καμπάνιας, πενταπλάσια αύξηση της απόδοσης επένδυσης (ROI) σε σχέση με τις παραδοσιακές μεθόδους ανάλυσης!



Εικόνα 10. Source: Eric Siegel

Uplift Modeling έχει εφαρμοστεί και στην πολιτική. Μία τέτοια περίπτωση ήταν η προεκλογική εκστρατεία του Barack Obama στις Η.Π.Α. το 2012 για την οποία ζητούνταν ειδικοί προγνωστικής αναλυτικής. Πράγματι συστάθηκε ομάδα περισσότερων από 50 άτομα αναλυτών για τη συλλογή ευρέως διαθέσιμων δεδομένων ψηφοφόρων και την κατασκευή Uplift μοντέλων. Στόχος η προσέγγιση των "swing voters" δηλαδή εκείνων που δεν ήξεραν ακόμα τι θα ψηφίσουν αλλά και μπορούσαν να επηρεαστούν θετικά από διαφημιστικές ενέργειες ο προϋπολογισμός των οποίων, μόνο για τα τηλεοπτικά μέσα, ξεπερνούσε τα 400 εκατομμύρια δολάρια.

Ιδιαίτερα σημαντικό ρόλο έχει και θα έχει η προγνωστική αναλυτική στον υγειονομικό τομέα και την έρευνα. Στις μέρες μας η χαρτογράφηση της αλληλουχίας του ανθρώπινου γονιδιώματος κοστίζει 600 δολάρια ενώ όταν αυτό ολοκληρώθηκε για πρώτη φορά μετά από 13 χρόνια το 2003 είχε κοστίσει 2.7 δισεκατομμύρια δολάρια. Αναμένεται λοιπόν πως στο άμεσο μέλλον αυτή η τεράστια βάση δεδομένων που έχει ο κάθε άνθρωπος θα χρησιμοποιείται τακτικά όχι μόνο για διαγνώσεις αλλά και για προβλέψεις ασθενειών όπως είναι οι διάφορες μορφές καρκίνου με στόχο την εξατομικευμένη στρατηγική διαχείριση και πρόληψη αυτών. Έχουν ήδη αναπτυχθεί διάφορα εργαλεία αναλυτικής για την έρευνα της μοριακής βιολογίας και του γονιδιώματος του ανθρώπου όπως είναι το GATK του Broad Institute των MIT και Harvard αλλά και ελληνικές πρωτοβουλίες όπως το Diana Lab του Πανεπιστημίου Θεσσαλίας.

Η προγνωστική αναλυτική βέβαια είναι ένα εργαλείο κι όπως όλα τα εργαλεία μπορεί να χρησιμοποιηθεί για λιγότερο θεμιτούς σκοπούς. Το 2012 αποκαλύφθηκε πως η Hewlett-Packard είχε αναθέσει σε αναλυτές την χρήση προσωπικών δεδομένων των υπαλλήλων της για την κατασκευή προγνωστικών μοντέλων τα οποία υπολόγιζαν έναν αριθμό τον "Flight Risk" που έδειχνε πόσο πιθανό ήταν ένας υπάλληλος να φύγει από την εταιρεία. Τους αριθμούς αυτούς τους γνώριζαν μόνο οι προϊστάμενοι της εταιρείας και να τους χρησιμοποιήσουν όπως εκείνοι νόμιζαν. Την ίδια χρονιά αποκαλύπτεται επίσης πως η αλυσίδα πολυκαταστημάτων Target επεξεργαζόταν στοιχεία των πελατών της για την πρόβλεψη αναγκών τους όπως για παράδειγμα ποιες γυναίκες θα γεννήσουν στο άμεσο μέλλον και χρήση προωθητικών ενεργειών για αντίστοιχα προϊόντα.

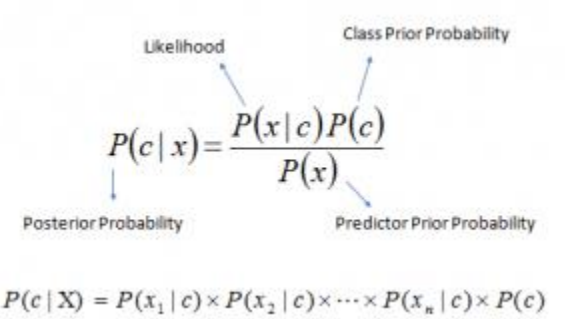
Ακόμα χειρότερα το 2018 πρώην εργαζόμενοι της Cambridge Analytica, εταιρεία παροχής συμβουλών, αποκάλυψαν πως η εταιρεία είχε δημιουργήσει ψυχολογικά προφίλ από 87 εκατομμύρια και πλέον χρήστες του Facebook χρησιμοποιώντας προσωπικά δεδομένα χωρίς τη συγκατάθεσή τους με στόχο την πρόβλεψη και προώθηση σε αυτούς πολιτικών διαφημίσεων με πιο γνωστές περιπτώσεις αυτές της προεκλογικής εκστρατείας του Donald Trump και των υποστηρικτών του Brexit. Τον Οκτώβριο του 2021 το Twitter παραδέχθηκε, μετά από έρευνα ανεξάρτητων αναλυτών στο Ηνωμένο Βασίλειο, τις Η.Π.Α., τον Καναδά, τη Γαλλία, τη Γερμανία, την Ισπανία, την Ιαπωνία, πως ο αλγόριθμος μηχανικής μάθησης παρουσίαζε στους χρήστες του πολύ περισσότερο πολιτικό περιεχόμενο από πηγές προσκείμενες στη συντηρητική ιδεολογία παρά στη προοδευτική ασκώντας όμως έτσι έμμεσα σε κάποιο βαθμό και πολιτική επιρροή. Μέχρι τη στιγμή που γράφονταν αυτές οι γραμμές δεν είχε δοθεί μια πειστική εξήγηση γιατί γινόταν αυτό. Εξάλλου σκοπός της προγνωστικής αναλυτικής και των αλγορίθμων μηχανικής μάθησης είναι να εξάγουν πληροφορία με τη μορφή συμπερασμάτων και προβλέψεων. Το γιατί και το πως θα αξιοποιηθούν αυτές οι προβλέψεις έγκειται στον ανθρώπινο παράγοντα.

4. Προγνωστικά μοντέλα αλγορίθμων

Για την ανάλυση των δεδομένων που προκύπτουν από την έρευνα για τις ανάγκες της τοπικής αυτοδιοίκησης σε cloud storage θα χρησιμοποιήσουμε τέσσερις γνωστούς και ευρείας χρήσης αλγορίθμους τεχνικών προγνωστικής αναλυτικής: τον Naive Bayes διακριτών τιμών της κατηγοριοποίησης Bayes, ID3 των δέντρων απόφασης, Random Forest και AdaBoost των μεθόδων Ensemble.

4.1 Naive Bayes

Ο Naive Bayes είναι ένας πολύ γρήγορος και υπολογιστικά φθηνός αλγόριθμος που χρησιμοποιείται κυρίως σε απλά προβλήματα προβλέψεων διακριτών τιμών με σχετικά καλή ακρίβεια. Σπανιότερα μπορεί να υλοποιηθεί και για προβλήματα συνεχών τιμών χρησιμοποιώντας κανονική κατανομή και συνάρτηση πυκνότητας πιθανότητας. Βασική υπόθεση της λειτουργίας του έγκειται στο ότι όλα τα χαρακτηριστικά ενός αντικειμένου θεωρούνται ανεξάρτητα μεταξύ τους, κάτι βέβαια που στην πραγματικότητα δεν συμβαίνει συχνά και για αυτό το όνομα του αλγορίθμου είναι αφελής (naive). Βασίζεται στις στατιστικές μετρήσεις των πιθανοτήτων των χαρακτηριστικών και της κλάσης καθώς και στο θεώρημα Bayes για να προβλέψει σύμφωνα με τις πιθανότητες των ήδη κατηγοριοποιημένων δεδομένων, τις πιθανότητες των τιμών της κλάσης για ένα νέο αντικείμενο:



Εικόνα 11. Source: https://www.analyticsvidhya.com

όπου P(c|x) η πιθανότητα του ζητούμενου αντικειμένου να έχει κλάση με τιμή c με χαρακτηριστικά x, P(x|c) η πιθανότητα δεδομένου αντικειμένου να έχει χαρακτηριστικά x με κλάσης τιμής c, P(c) η πιθανότητα τιμής c της κλάσης σε όλα τα υπάρχοντα δεδομένα και P(x) η πιθανότητα των χαρακτηριστικών των δεδομένων. Οπότε αν για παράδειγμα μια κλάση έχει δυο τιμές c και y και προκύπτει από τους υπολογισμούς P(c|x) > P(y|x) για ζητούμενο αντικείμενο χαρακτηριστικών x τότε η πρόβλεψη θα είναι c. Εάν κάποιες τιμές χαρακτηριστικών δεδομένου αντικειμένου λείπουν τότε ο αλγόριθμος αγνοεί το αντικείμενο κατά τον υπολογισμό των πιθανοτήτων. Όμως εάν κάποια από τις πιθανότητες είναι μηδενική γιατί ίσως η τιμή της κλάσης ενός αντικειμένου στα δεδομένα αξιολόγησης δεν υπήρχε στα δεδομένα εκπαίδευσης τότε μπορούν να εφαρμοστούν διάφορες τεχνικές όπως η εκτίμηση Laplace.

4.2 ID3

Ο Iterative Dichotomiser 3 (ID3) είναι ένας από τους πιο γνωστούς αλγόριθμους των δέντρων απόφασης. Αναπτύχθηκε από τον ερευνητή μηχανικής μάθησης και τεχνητής νοημοσύνης Ross Quinlan και αποτελεί τον πρόδρομο του αλγόριθμου C4.5 και του εμπορικού C5.0. Ανήκει στη μεγάλη κατηγορία των αλγορίθμων επιβλεπόμενης μηχανικής μάθησης και χρησιμοποιείται κυρίως σε προβλήματα κατηγοριοποίησης δηλαδή σε δεδομένα όπου η κλάση τους έχει διακριτές τιμές. Αποτελεί έναν γρήγορο και αποτελεσματικό αλγόριθμο για τις περισσότερες περιπτώσεις προβλημάτων με προβλέψεις που γίνονται εύκολα κατανοητές.

Ο τρόπος με τον οποίο λειτουργεί είναι σχετικά απλός αλλά και αποτελεσματικός. Εξετάζει κάθε χαρακτηριστικό το οποίο δεν έχει εξετασθεί προηγουμένως και επιλέγει με κάποιο κριτήριο το καλύτερο χαρακτηριστικό για να γίνει ο διαχωρισμός του σετ δεδομένων σε μικρότερα σετ (υπό-δέντρα). Αυτό σταματάει για το συγκεκριμένο σετ όταν η τιμή του εξεταζόμενου χαρακτηριστικού αντιστοιχεί σε μία τιμή της κλάσης ή έχουν εξεταστεί όλα τα χαρακτηριστικά κι αυτή η τιμή της κλάσης είναι η κυριαρχούσα, δηλαδή έχει πολύ μεγάλη πιθανότητα να συμβεί, οπότε και δημιουργείται ένα από τα φύλλα του δέντρου. Η διαδικασία συνεχίζεται αναδρομικά και για τα υπόλοιπα χαρακτηριστικά που δεν έχουν εξεταστεί μέχρι σχηματιστούν όλα τα φύλλα του δέντρου. Τα κύρια μειονεκτήματα αυτού του αλγορίθμου προκύπτουν από το γεγονός ότι κάνει «άπληστη» αναζήτηση (greedy search) διαλέγοντας το καλύτερο χαρακτηριστικό μόνο από τα εξεταζόμενα αλλά και η «υπερπροσαρμογή» (overfitting), κυρίως στα μικρής έκτασης δεδομένα, όπου ο διαχωρισμός των χαρακτηριστικών είναι τόσο λεπτομερής με αποτέλεσμα η πρόβλεψη της κλάσης νέων δεδομένων να μη μπορεί να γίνει με μεγάλη ακρίβεια. Κάποια από αυτά τα θέματα λύνονται είτε με σωστή προ-επεξεργασία των δεδομένων είτε μετά την αλγοριθμική ανάλυση με διάφορες τεχνικές όπως το «κλάδεμα» (pruning) δηλαδή την αφαίρεση των άχρηστων κόμβων του δέντρου που γίνεται από τον αναλυτή ή αυτόματα σε κάποιες υλοποιήσεις του ID3.

Το βασικό κριτήριο με το οποίο γίνεται ο διαχωρισμός των χαρακτηριστικών αφορά την σαφήνεια της πληροφορίας, την ομοιογένεια ή «αγνότητα» ενός κόμβου δηλαδή με άλλα λόγια το πόσες διαφορετικές τιμές ενός χαρακτηριστικού αντιστοιχούν σε διαφορετικές τιμές της κλάσης. Για το σκοπό αυτό χρησιμοποιούνται τα εξής τέσσερα κριτήρια:

- Information Gain. Υπολογίζεται η εντροπία κάθε προς εξέταση χαρακτηριστικού και επιλέγεται για διαχωρισμό εκείνο με τη μικρότερη. Προτιμά χαρακτηριστικά με πολλές τιμές οπότε υπάρχει ο κίνδυνος για πολλούς αλλά ομοιογενείς διαχωρισμούς.
- Gain Ratio. Παραλλαγή του Information Gain ώστε να επιτρέπει το εύρος και την ομοιομορφία των τιμών των χαρακτηριστικών ελαχιστοποιώντας το overfitting.
- Gini Index. Σε παρόμοια λογική με το Information Gain υπολογίζεται το μέτρο της ανομοιογένειας ενός σετ δεδομένων με σκοπό τη μείωση αυτού σε κάθε διαχωρισμό δεδομένων. Δε χρησιμοποιείται συχνά στον ID3 αλλά για άλλους αλγορίθμους δέντρων αποφάσεων.
- Accuracy. Ο διαχωρισμός των χαρακτηριστικών γίνεται ώστε μεγιστοποιηθεί η ακρίβεια ολόκληρου του δέντρου ελέγχοντας το σφάλμα κατηγοριοποίησης σε κάθε κόμβο.

Από αυτές τις τέσσερις τεχνικές ως επί το πλείστων στον ID3 χρησιμοποιείται το Information Gain κάθε χαρακτηριστικού F μετά από τη διάσπαση του σετ S στα υπό-σετ S_i

$$InformationGain(F) = Entropy(S) - Entropy(A)$$

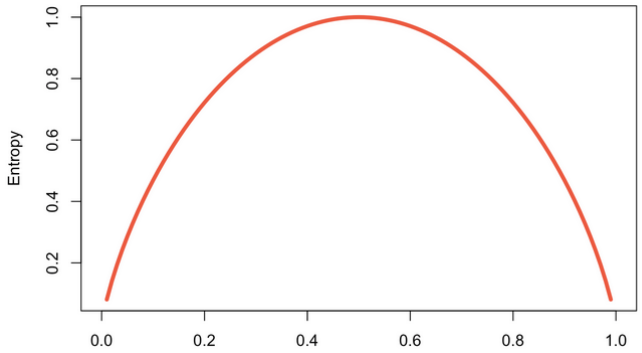
όπου Entropy(S) η εντροπία του αρχικού σετ S και Entropy(A) το άθροισμα του γινομένου της πιθανότητας του χαρακτηριστικού S_i επί την εντροπία του μετά τη διάσπαση

$$\text{Entropy}(A) = \sum_{i=1}^{n} P(S_i) Entropy(S_i)$$

Η εντροπία ως όρος προέρχεται από τη θερμοδυναμική και στην ανάλυση δεδομένων αναφέρεται στο βαθμό αταξίας της πληροφορίας. Όσο πιο μικρή είναι η αταξία και η εντροπία τόσο μεγαλύτερη είναι η ομοιογένεια της πληροφορίας ενός χαρακτηριστικού και άρα η πιθανότητα μιας τιμής ενός χαρακτηριστικού να αντιστοιχεί σε μια τιμή της κλάσης ώστε να επιλεγεί ως κόμβος για τον περαιτέρω διαχωρισμό των άλλων χαρακτηριστικών. Υπολογίζεται από τον τύπο:

$$\text{Entropy}(S) = \sum_{i=1}^{c} -P_i log_2(P_i)$$

Επομένως αν η πιθανότητα P_i ισούται με 1, δηλαδή όλες οι τιμές ενός χαρακτηριστικού αντιστοιχούν σε μία μόνο τιμή της κλάσης, τότε ο λογάριθμος είναι 0 όπως κι εντροπία οπότε μάλλον θα επιλεγεί για το διαχωρισμό ή για τη δημιουργία φύλλου/απόφασης. Αντίθετα η εντροπία ισούται περίπου με 1 όταν μεγιστοποιείται το άθροισμα των λογαρίθμων αν κάθε τιμή i του χαρακτηριστικού S έχει την ίδια πιθανότητα P_i να αντιστοιχεί σε μία μόνο τιμή της κλάσης. Χαρακτηριστικά με μεγάλη εντροπία περιέχουν τη λιγότερη σημαντική πληροφορία και συνήθως δεν επιλέγονται για διαχωρισμό και μπορεί να απορριφθούν πριν ή μετά την ανάλυση.



Εικόνα 12. Source: https://rpubs.com/

4.3 Random Forest

Ο ID3 κι οι άλλοι αλγόριθμοι των δέντρων απόφασης είναι ιδιαίτερα χρήσιμοι για την ανάλυση όλων των περιπτώσεων και των πιθανοτήτων αυτών όπως για παράδειγμα στη λήψη μιας εταιρικής απόφασης. Στους περισσότερους τομείς όμως αυτοί οι αλγόριθμοι έχουν αντικατασταθεί πλέον από τους πολύ πιο ακριβείς μεθόδους Ensemble όπως ο Random Forest και AdaBoost που στηρίζονται στην ιδέα της «σοφίας του πλήθους» ότι δηλαδή η συνισταμένη απόφαση που μπορεί να εξαχθεί από πολλούς ανθρώπους αποδεικνύεται συνήθως καλύτερη από αυτή που θα έπαιρνε μόνο ένας άνθρωπος όσο ειδήμων κι αν είναι αυτός για το συγκεκριμένο πρόβλημα . Είναι εύκολα υλοποιήσιμοι και δεν έχουν τα προβλήματα των δέντρων απόφασης του overfitting ως αποτέλεσμα και της κατάρας της διαστατικότητας όπου η ύπαρξη πολλών στηλών/χαρακτηριστικών οδηγεί μαθηματικά σε μείωση της ακρίβειας των αποτελεσμάτων.

Οι μέθοδοι Ensemble χωρίζονται σε δύο βασικές κατηγορίες ανάλογα με τον τρόπο λειτουργίας: bagging και boosting. Ο όρος bagging (ή bootstrap aggregation) αναφέρεται στην τεχνική με την οποία αναπτύσσονται παράλληλα δέντρα απόφασης σε μέρη του αρχικού σετ δεδομένων με αντικατάσταση, δηλαδή κάθε μέρος μπορεί να περιλαμβάνει πολλές ίδιες εγγραφές από το αρχικό σετ. Αυτή η τυχαιότητα των εγγραφών στα μέρη έχει ως αποτέλεσμα τη δημιουργία αδύναμων δέντρων απόφασης κατηγοριοποιητών (weak learners). Οι προβλέψεις τους όμως είναι και ανεξάρτητες και μπορούν να χρησιμοποιηθούν στο σύνολό τους για την εξαγωγή πρόβλεψης μεγάλης ακρίβειας με διάφορους μαθηματικούς τρόπους.

Ο Random Forest βασίζεται στη τεχνική bagging επιλέγοντας όμως τυχαία μόνο κάποια από τα χαρακτηριστικά για τη δημιουργία των μερών του αρχικού σετ δεδομένων (feature bagging). Στη συνέχεια τα παράλληλα δέντρα απόφασης αναπτύσσονται διαχωρίζοντας τα χαρακτηριστικά αυτά με κριτήρια όπως και ο ID3: Information Gain, Gain Ratio, Gini Index, Accuracy. Επιπλέον μπορεί να χρησιμοποιηθεί και η Least Square όπου το χαρακτηριστικό που επιλέγεται για διαχωρισμό είναι αυτό με το οποίο ελαχιστοποιείται η τετραγωνική απόσταση του μέσου όρου των τιμών με την πραγματική τιμή. Η τελική πρόβλεψη προκύπτει υπολογίζοντας είτε τον μέσο όρο όταν πρόκειται για προβλήματα παλινδρόμησης είτε την πλειοψηφούσα πρόβλεψη για προβλήματα κατηγοριοποίησης. Να αναφερθεί εδώ πως σε αντίθεση με άλλα μοντέλα μηχανικής μάθησης το αρχικό σετ δεδομένων δε χωρίζεται σε σετ εκπαίδευσης και σετ αξιολόγησης επειδή κατά τη δημιουργία των μερών κάποιες εγγραφές δε θα συμπεριληφθούν σε αυτά (out-of-bag sample) οπότε και μπορούν να χρησιμοποιηθούν ως σετ αξιολόγησης της ακρίβειας των αποτελεσμάτων. Μειονέκτημα του Random Forest είναι ότι η υλοποίηση του απαιτεί αρκετή υπολογιστική ισχύ γιατί μπορούν να αναπτυχθούν παράλληλα πολλά δέντρα απόφασης, αν και στη σημερινή εποχή με τους πολυπύρηνους και πολυνηματικούς επεξεργαστές αυτό δεν αποτελεί τόσο μεγάλο πρόβλημα όσο παλαιότερα.

Σε μία παραλλαγή του Random Forest που ονομάζεται Extra Trees, κατά την ανάπτυξη των δέντρων απόφασης επιλέγονται τυχαία τα χαρακτηριστικά με βάση των οποία θα γίνει ο διαχωρισμός κι όχι σύμφωνα με κάποια κριτήρια όπως στον Random Forest και τον ID3. Αυτό έχει ως αποτέλεσμα η ακρίβεια των προβλέψεων να είναι περίπου ίδια αλλά με μικρότερη διακύμανση των τιμών και χρησιμοποιώντας λιγότερη υπολογιστική ισχύ.

4.4 AdaBoost

Μία από τις δυο κατηγορίες αλγορίθμων Ensemble όπως αναφέραμε είναι και η boosting. Αντίθετα με την bagging όπου κατηγοριοποιητές όπως τα δέντρα απόφασης εκπαιδεύονται ταυτόχρονα, στην boosting οι κατηγοριοποιητές εκπαιδεύονται διαδοχικά με τελικό στόχο την εκπαίδευση ενός μοντέλου με όσο το δυνατό πιο μεγάλη ακρίβεια. Θα μπορούσαμε να την παρομοιάσουμε με έναν άνθρωπο που προσπαθεί να βελτιώνεται μαθαίνοντας από τα λάθη του. Αυτό βέβαια μπορεί να οδηγήσει σε φαινόμενα overfitting όπως κι ο ID3 αν και εξακολουθεί να αποτελεί αντικείμενο ακαδημαϊκής έρευνας το κατά πόσο όντως αυτό ισχύει. Επίσης είναι μία τεχνική ευαίσθητη σε τιμές των χαρακτηριστικών που είτε είναι λανθασμένες είτε πολύ απομακρυσμένες από τις υπόλοιπες (outliers).

Ο AdaBoost (Adaptive Boosting) είναι ο πιο γνωστός αλγόριθμος boosting που αναπτύχθηκε από τους Yoav Freund και Robert Schapire. Θεωρείται meta αλγόριθμος διότι μπορεί να εφαρμοστεί σε συνδυασμό με άλλους αλγόριθμους μηχανικής μάθησης για να βελτιώσει τις επιδόσεις τους και μάλιστα εύκολα χωρίς ιδιαίτερες παραμετροποιήσεις (out of the box classifier). Η γενική ιδέα είναι πως ένας αρχικά αδύναμος κατηγοριοποιητής με ποσοστό λάθους μικρότερο του 0.5 μπορεί να προσαρμοστεί διαδοχικά ώστε αυτό το ποσοστό λάθους να ελαχιστοποιηθεί. Αυτό μπορεί να συμβεί αναθέτοντας σε όλες τις αρχικές εγγραφές τον ίδιο αριθμό ή «βάρος» αυξάνοντάς το σε κάθε επανάληψη για τις εγγραφές που κατηγοριοποιούνται λανθασμένα και μειώνοντάς το για τις εγγραφές που κατηγοριοποιούνται σωστά ώστε ο αλγόριθμος να επικεντρώνεται στις πρώτες.

Πιο συγκεκριμένα για εγγραφές m η κάθε μία αρχικά θα έχει βάρος $D_1(i)$:

$$D_1(i) = \frac{1}{m} \in [0,1]$$

Το κάθε βάρος έχει τιμή μεταξύ 0 και 1 κι όλα μαζί αθροιστικά ισούται με 1. Έπειτα υπολογίζουμε την επιρροή α του μοντέλου που προκύπτει:

$$a = \frac{1}{2}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right)$$

όπου ε το ποσοστό λάθους δηλαδή ο αριθμός των λανθασμένων κατηγοριοποιημένων εγγραφών προς τις συνολικές εγγραφές. Όσο πιο λίγα λάθη έχει κάνει το παρών μοντέλο τόσο πιο μεγάλο είναι το α. Αντίθετα όταν το ποσοστό λάθους είναι όσο το ρίξιμο ενός νομίσματος δηλαδή ε = 0.5 τότε α = 0 και όταν ε > 0.5 τότε το α μπορεί να πάρει ακόμα κι αρνητικές τιμές οπότε ο αλγόριθμος ξεκινάει από την αρχή επαναφέροντας τα αρχικά βάρη στις εγγραφές. Εάν ε < 0.5 τότε συνεχίζουμε υπολογίζοντας τα νέα βάρη που προκύπτουν για τις εγγραφές:

$$D_{t+1} = D_t * e^{\pm a}$$

Ο αλγόριθμος τερματίζει όταν επιτευχθεί το επιθυμητό D_t ή παραχθεί ο αριθμός των επαναλήψεων δηλαδή των αδύναμων κατηγοριοποιητών που είχαμε ορίσει εξαρχής.

5. Ανάλυση των αλγοριθμικών μοντέλων με το RapidMiner

5.1 Μεθοδολογία

Ο τίτλος της έρευνας των οποίων τα δεδομένα θα χρησιμοποιήσουμε είναι «Η χρήση του Cloud Storage από τους Δήμους της Ελλάδας» και υλοποιήθηκε στο τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου στο πλαίσιο της ερευνητικής εργασίας "Factors affecting cloud storage adoption by Greek municipalities" των Κυριακού Νίκης, Λουκή Ευριπίδη και Δημητροπούλου Παρασκευής που παρουσιάστηκε στο 13^{ο} διεθνές συνέδριο Theory and Practice of Electronic Governance. Χρησιμοποιήθηκε ερωτηματολόγιο 23 κύριων ερωτημάτων το οποίο στάλθηκε σε 332 δήμους της Ελλάδας σε όλους τους πρώην νομούς της χώρας από τους οποίους απαντήσανε οι 121 (ποσοστό απόκρισης 36,45%). Η μελέτη βασίστηκε στο θεωρητικό πλαίσιο DOI του Rogers για τους παράγοντες υιοθέτησης της τεχνολογίας και οι ερωτήσεις αφορούσαν τη διερευνητική ανάλυση του συγκριτικού πλεονεκτήματος και μειονεκτήματος, της πολυπλοκότητας και της συμβατότητας του cloud storage. Το πρώτο μέρος των ερωτημάτων A1.1 έως A1.5 αφορούν γενικές πληροφορίες για τους δήμους ενώ το δεύτερο μέρος B2.1 έως B2.18 τη χρήση cloud storage. Κάποια ερωτήματα περιείχαν επιμέρους μεταβλητές τις οποίες οι ερωτώμενοι καλούνταν να αξιολογήσουν σε κλίμακα Likert 1-5 όπου 1 όταν διαφωνούσαν κάθετα και 5 όταν συμφωνούσαν απόλυτα. Συνολικά το σετ δεδομένων που προκύπτει περιλαμβάνει 121 γραμμές/εγγραφές και 39 στήλες/χαρακτηριστικά. Ολόκληρο το ερωτηματολόγιο παρουσιάζεται στο παράρτημα στο τέλος της εργασίας.

Τα δεδομένα αυτά στη συνέχεια θα χρησιμοποιηθούν για την εκπαίδευση και αξιολόγηση των αλγοριθμικών μοντέλων Naive Bayes, ID3, Random Forest, AdaBoost στο RapidMiner. Το RapidMiner αποτελεί μια από τις πιο διαδεδομένες και εύχρηστες εφαρμογές προγνωστικής αναλυτικής, μηχανικής μάθησης και εξόρυξης δεδομένων. Το μεγάλο του πλεονέκτημα είναι το απλό, λειτουργικό και παραμετρικό γραφικό περιβάλλον μέσα από το οποίο μπορεί να αναπτυχθεί πλήρως ένα μοντέλο αν και υπάρχει κι η επιλογή για συγγραφή κώδικα αν αυτό είναι επιθυμητό. Εξαιρετική είναι επίσης κι η αυτοματοποίηση της οπτικοποίησης των αποτελεσμάτων και συμπερασμάτων που προκύπτουν από την ανάλυση ενός μοντέλου καθώς και η διαφάνεια των εσωτερικών διεργασιών που γίνονται στα δεδομένα για να συμβεί αυτό. Υποστηρίζεται η διεργασία πληθώρας τύπων αρχείων, βάσεων δεδομένων και υπηρεσιών cloud storage αλλά και custom plug-ins. Ενδιαφέρον και πρωτότυπη είναι επίσης η δυνατότητα παρουσίασης της δημοφιλίας των τιμών των παραμέτρων κατά τη δημιουργία ενός μοντέλου συμβάλλοντας γενικότερα στην ανάπτυξη μιας συνεργατικής κοινότητας αναλυτών.

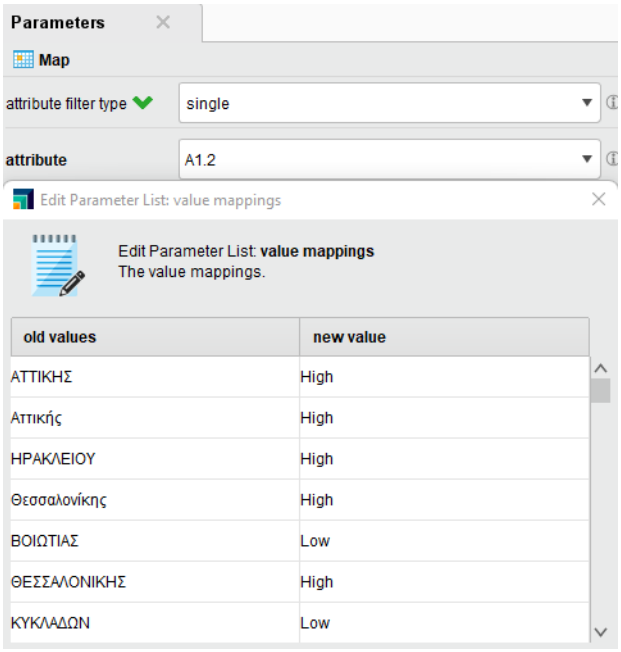
Βέβαια πριν την εισαγωγή των δεδομένων στο RapidMiner θα πρέπει να γίνει όπως έχει αναφερθεί η πολύ σημαντική προετοιμασία αυτών καθώς ο «θόρυβος» μπορεί να επηρεάσει σημαντικά την ακρίβεια των μοντέλων. Αυτό που κάναμε για το συγκεκριμένο σετ δεδομένων ήταν να αφαιρέσουμε 9 στήλες/χαρακτηριστικά (A1.1, A1.4, A1.5, B2.7, B2.8, B2.11, B2.12, B2.13, B2.18) που είτε ήταν σε μεγάλο μέρος τους κενές είτε δε προσέφεραν κάτι για την πρόβλεψη που μας ενδιαφέρει όπως για παράδειγμα το χαρακτηριστικό για την επαγγελματική ιδιότητα του υπαλλήλου που απάντησε στο ερωτηματολόγιο.

Αρχικά το χαρακτηριστικό που επιλέχθηκε ως εξαρτημένη μεταβλητή για την τιμή της οποίας θα γίνεται η πρόβλεψη για νέα δεδομένα ήταν η B2.4 «Πότε σκοπεύετε να κάνετε χρήση του Cloud Storage στον δήμο σας» με τιμές 1) Σκοπεύουμε μέσα στον επόμενο χρόνο 2) Ίσως, μετά τον επόμενο χρόνο 3) Ακόμη δεν έχουμε την κατάλληλη γνώση σχετικά με την χρήση του Cloud Storage 4) Δεν θέλουμε να κάνουμε χρήση αυτής της υπηρεσίας. Όμως τελικά φάνηκε πως η επιλογή αυτή δεν ήταν καλή μιας και η ακρίβεια πρόβλεψης των μοντέλων ήταν χειρότερη ακόμα και του 50% όσο είναι δηλαδή η πιθανότητα να μαντέψει κάποιος το αποτέλεσμα της ρίψης ενός νομίσματος λόγω του μικρού μεγέθους των δεδομένων μας. Για αυτό τελικά επιλέξαμε ως κλάση/εξαρτημένη μεταβλητή B2.3 «Στον δήμο σας κάνετε χρήση Cloud Storage;» η οποία είναι διωνυμική με τιμές 1) Ναι 2) Όχι για την οποία τα αποτελέσματα ήταν πολύ καλύτερα όπως θα δούμε παρακάτω χωρίς να αλλάζει η ουσία της προγνωστικής έρευνας. Η ιδέα είναι ότι ένα τέτοιο προγνωστικό μοντέλο θα μπορούσε να χρησιμοποιηθεί από μια εταιρεία παροχής υπηρεσιών cloud storage ώστε να προβεί σε προωθητικές ενέργειες σε δήμους που δεν έχουν ακόμα cloud storage.

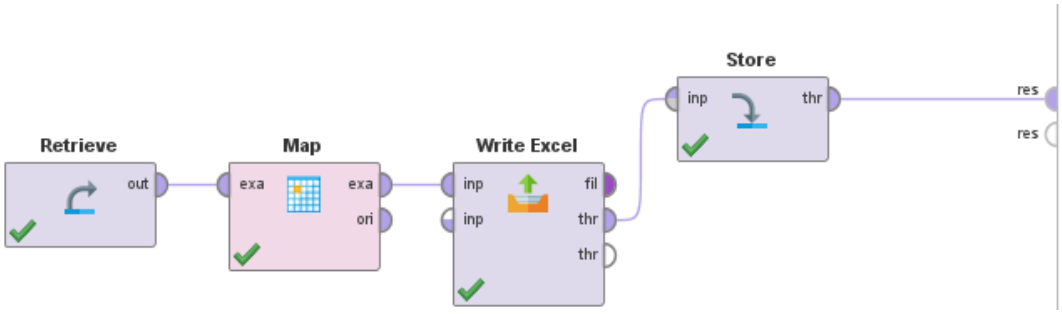
Τα δεδομένα μας είναι καταγεγραμμένα σε αρχείο excel Municipalities.xlsx με 121 γραμμές που είναι οι δήμοι που απάντησαν στο ερωτηματολόγιο και 30 στήλες των ερωτήσεων. Εισάγουμε αυτό το αρχείο στο RapidMiner με το εικονίδιο Import Data προσέχοντας ο τύπος κάθε χαρακτηριστικού να είναι σωστός και ειδικά το B2.3 που είναι η κλάση να είναι binominal και ο τύπος της label. Στο Statistics tab μπορούμε να δούμε τη συνολική εικόνα των χαρακτηριστικών των δεδομένων μας, τον τύπο τους, αν λείπουν τιμές, τα ακρότατα, το μέσο για τις αριθμητικές μεταβλητές, τις πλειάδες για τις ονομαστικές. Αυτά είναι σημαντικά διότι μπορούν να συμβάλουν καθοριστικά στην καλή προετοιμασία των δεδομένων μας. Συγκεκριμένα παρατηρούμε πως τα A1.3 και B2.4 έχουν ελλιπείς τιμές καθώς και ότι το A1.2 έχει 67 διαφορετικές τιμές στο σύνολο των 122 εγγραφών που μπορεί να δημιουργήσει προβλήματα στα μοντέλα μειώνοντας την ακρίβειά τους.

Εικόνα 13

Θα μπορούσαμε απλά να παραλείψουμε την A1.2 στήλη η οποία αναφέρεται στους νομούς της Ελλάδας (κάποιοι από τους οποίους είναι ίδιοι αλλά γραμμένοι διαφορετικά) που χρειάζονται cloud storage. Μία καλύτερη λύση είναι να αλλάξουμε τις τιμές ώστε οι νομοί με μεγάλο πληθυσμό (Αττικής, Θεσσαλονίκης, Ηρακλείου, Αχαΐας) να αντιστοιχούν στη τιμή "High" ενώ οι υπόλοιποι στη τιμή "Low" μειώνοντας έτσι τον αριθμό των τιμών της A1.2 από 67 σε 2. Για να το κάνουμε αυτό θα χρησιμοποιήσουμε τον Map operator. Έπειτα με τους operators Write Excel και Store θα αποθηκεύσουμε το νέο τροποποιημένο αρχείο MunicipalitiesModified.xlsx για τη μετέπειτα χρήση του στα προγνωστικά μοντέλα.



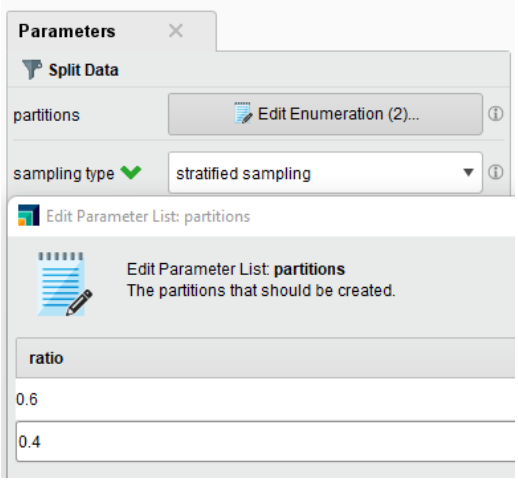
Εικόνα 14



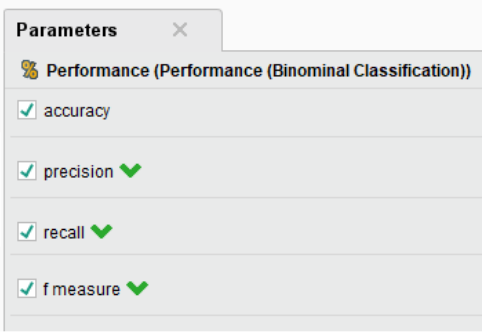
Εικόνα 15

5.2 Εκπαίδευση και Αξιολόγηση - Naive Bayes

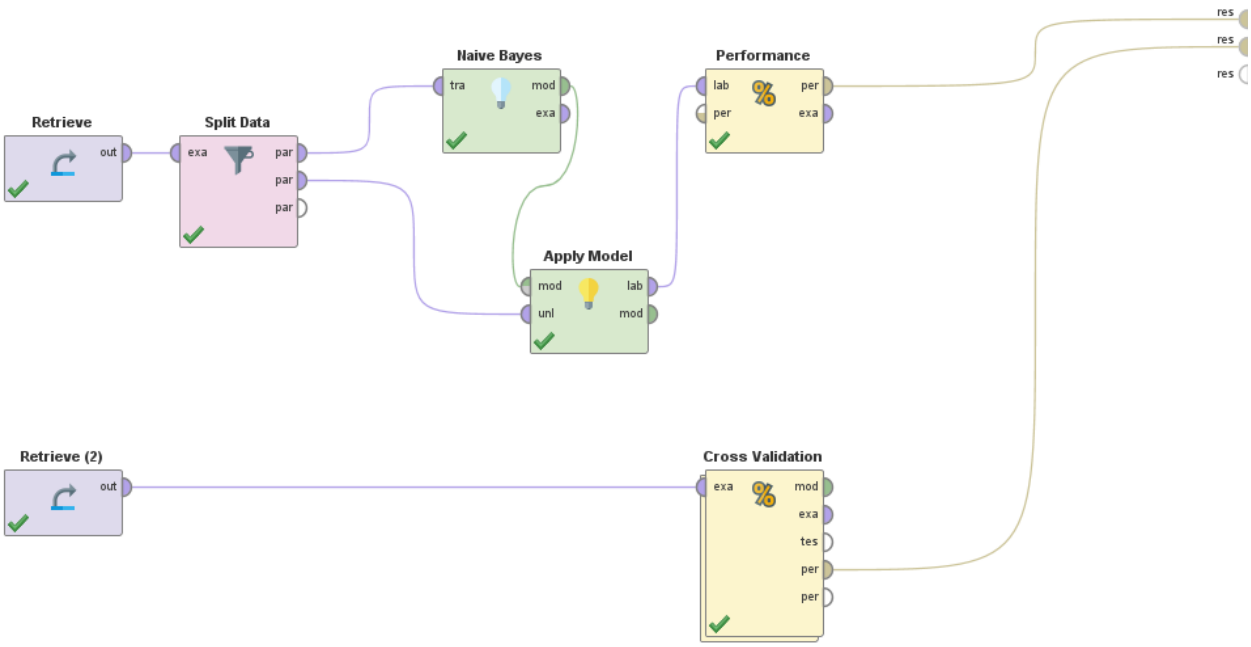
Ξεκινώντας τη δημιουργία μοντέλων προγνωστικής αναλυτικής στο RapidMiner θα υλοποιήσουμε πρώτα τον αλγόριθμο Naive Bayes. Για την εισαγωγή των δεδομένων χρησιμοποιούμε τον operator Retrieve. Με τον operator Split Data διαχωρίζουμε τα δεδομένα σε subsets των training και testing data με αναλογία 60% και 40% αντίστοιχα που είναι καλή για σχετικά μικρά δείγματα. Όσο μεγαλύτερο είναι το μέγεθος των δεδομένων τόσο μικρότερο είναι το subset των testing data. Συνήθως για δείγματα περίπου 1000 αντικειμένων η αναλογία είναι 70/30, για 10000 αντικείμενα 80/20 ενώ για πραγματικά big data μπορεί να γίνει ακόμα και 99/1. Ο διαχωρισμός των subset γίνεται με stratified sampling ώστε να είναι τυχαίος αλλά η κατανομή των κλάσεων να είναι ίδια. Έπειτα τα training data διοχετεύονται στον operator Naive Bayes για τη δημιουργία του αντίστοιχου κατηγοριοποιητή ο οποίος θα εφαρμοστεί στα testing data με τον operator Apply Model ώστε μετά με τον operator Performance (Binominal Classification) να αξιολογηθεί το μοντέλο με βάση το accuracy, precision, recall, f measure που προκύπτουν από τα αποτελέσματα του confusion matrix.



Εικόνα 16

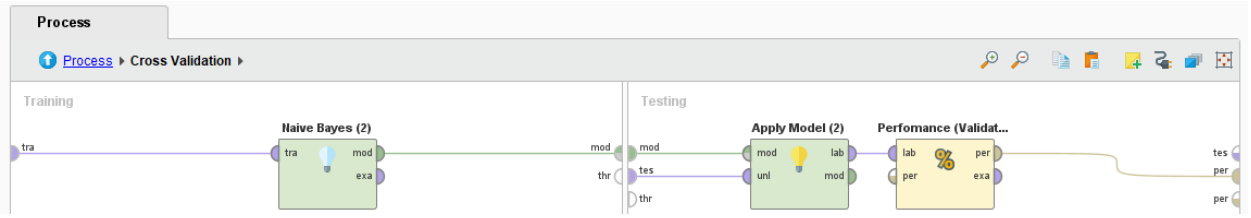


Εικόνα 17

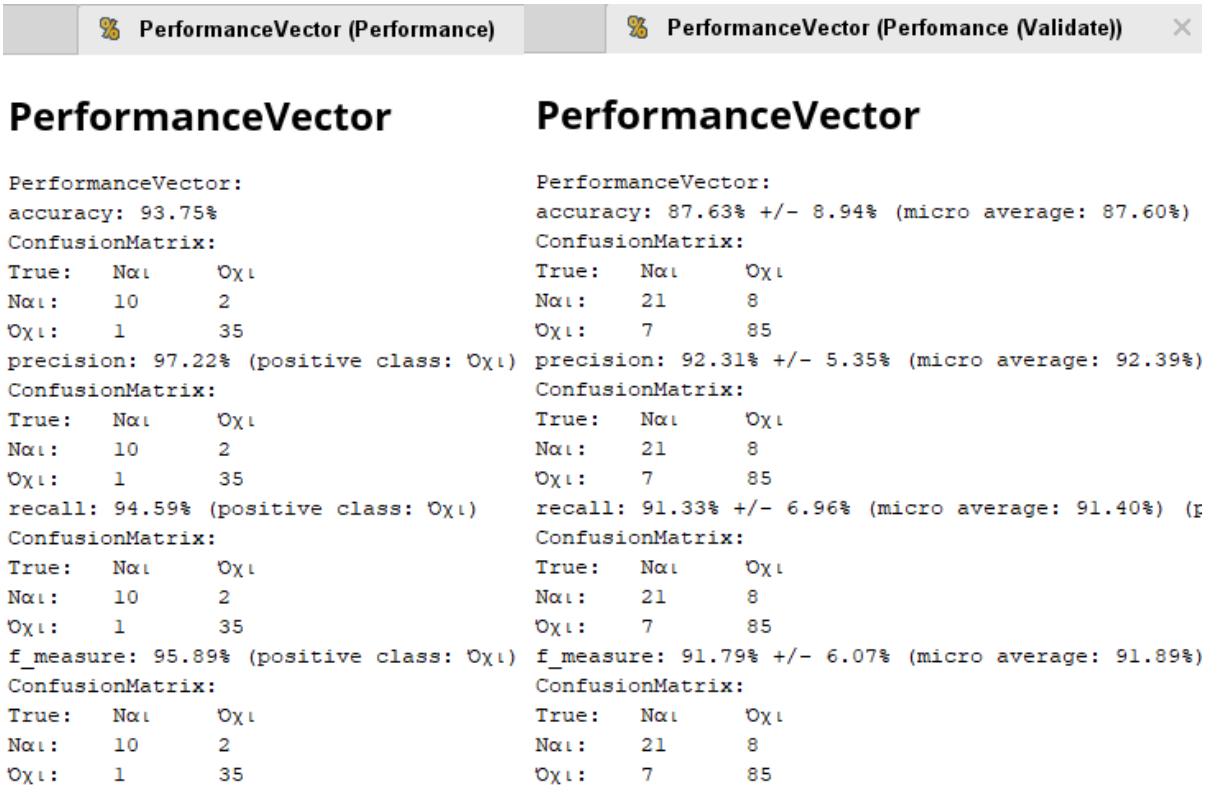


Εικόνα 18

Όπως παρατηρούμε από την παραπάνω εικόνα υλοποιήθηκε μια παράλληλη διεργασία όπου το αρχικό data set διοχετεύεται σε ένα νέο operator τον Cross Validation όπου εμφωλιάζονται οι operators Naive Bayes, Apply Model, Performance (Binominal Classification). Η αλγοριθμική τεχνική αυτού του operator χρησιμοποιείται ευρέως πλέον για την αξιολόγηση μοντέλων. Χωρίζει το αρχικό σετ δεδομένων σε k ίσα σετ (εμείς ορίσαμε k=10 και stratified sampling) από τα οποία το ένα διατηρείται ως testing set και τα υπόλοιπα ως training sets. Αυτό επαναλαμβάνεται k συνολικά φορές με τα υπόλοιπα σετ να χρησιμοποιούνται μία φορά ως δεδομένα δοκιμής. Στο τέλος τα αποτελέσματα συνδυάζονται για να παραχθεί μια ενιαία εκτίμηση της επίδοσης του μοντέλου. Αποφεύγεται έτσι και το φαινόμενο του overfitting (που παρουσιάζεται συνήθως στους αλγορίθμους με δέντρα απόφασης) οπότε και το Cross Validation μπορεί να χρησιμοποιηθεί για τον εντοπισμό αυτού στις κλασσικές μεθόδους ανάπτυξης μοντέλων όταν η ακρίβεια που προκύπτει με Cross Validation είναι πολύ μικρότερη.



Εικόνα 19



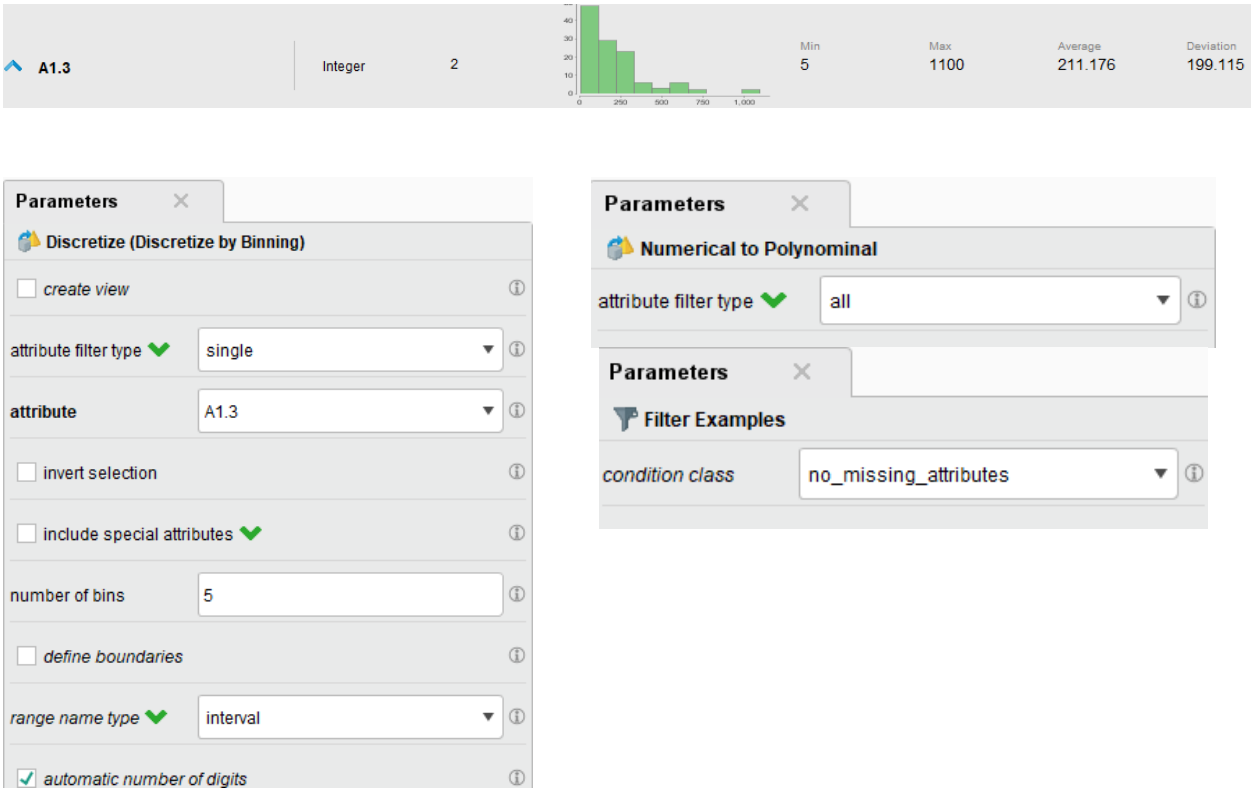
Εικόνα 20

Αριστερά φαίνονται τα αποτελέσματα της κλασσικής μεθόδου αξιολόγησης του μοντέλου με accuracy 93.75% και δεξιά τα αποτελέσματα με Cross Validation με accuracy 87.63% με τυπική απόκλιση +/-8.94%. Αυτό γενικά σημαίνει ότι πρόκειται για ένα πολύ καλό μοντέλο πρόβλεψης αλλά και ότι η αξιολόγηση με Split Data ήταν ακριβής μιας και το accuracy είναι εντός των ορίων της τυπικής απόκλισης του Cross Validation.

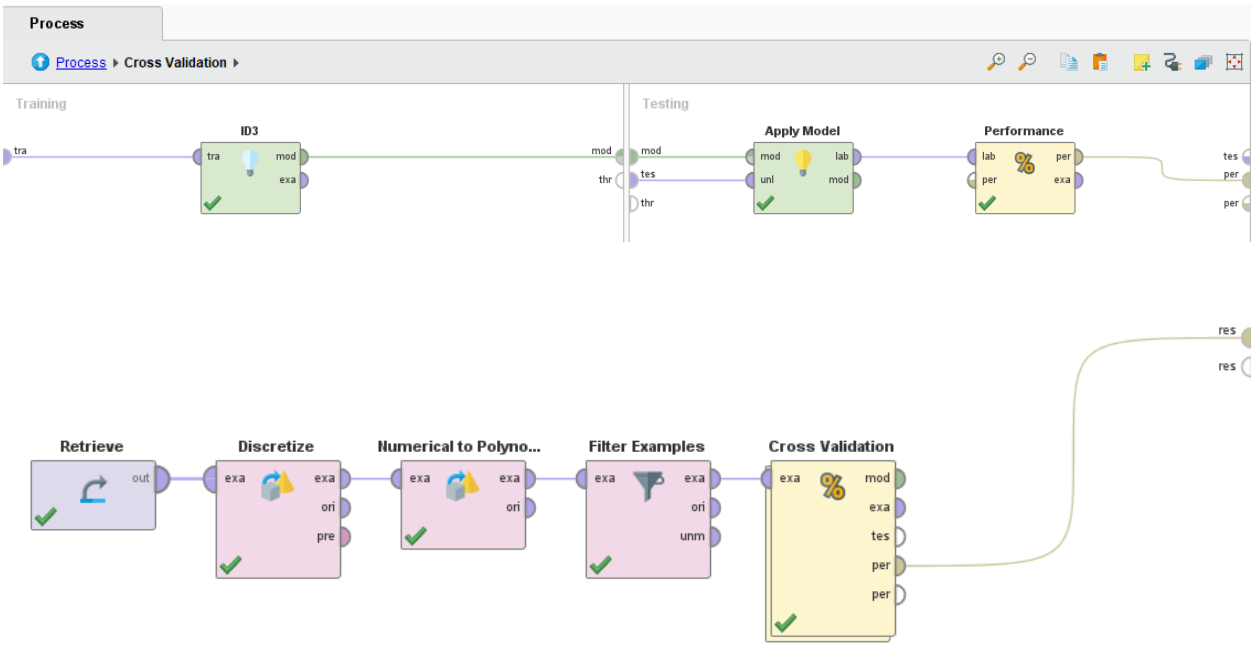
Όσον αφορά τα αποτελέσματα αυτά καθαυτά η παρουσίαση γίνεται με confusion matrix στους οποίους οι γραμμές δείχνουν πόσες προβλέψεις έγιναν σωστά και πόσες λάθος. Οπότε όπως βλέπουμε στο confusion matrix του Cross Validate για τα «Ναι» έγιναν 21 σωστές προβλέψεις (TruePositive) και 8 λάθος (FalsePositive), για τα «Όχι» 7 λάθος προβλέψεις (FalseNegative) και 85 σωστές (TrueNegative). Περαιτέρω ανάλυση αυτών των χαρακτηριστικών και των παραμέτρων αξιολόγησης γίνεται μετέπειτα στην εργασία.

5.3 Εκπαίδευση και Αξιολόγηση - ID3

Τα μοντέλα προγνωστικής αναλυτικής ID3 χαρακτηρίζονται από κάποιους περιορισμούς προκειμένου να έχουν καλή απόδοση όπως το ότι δεν πρέπει να χρησιμοποιούνται χαρακτηριστικά με αριθμητικές τιμές ή/και με ελλιπείς τιμές. Για το σκοπό αυτό επιλέγεται ο Numerical to Polynominal operator ο οποίος θα μετατρέψει όλες τις integer τιμές σε nominal. Όμως υπάρχει ένα χαρακτηριστικό integer το A1.3, που αφορά τον αριθμό των υπαλλήλων των δήμων, το οποίο αν μετατραπεί σε nominal ως έχει θα αποκτήσει πάρα πολλές τιμές με σημαντική επίπτωση στην ακρίβεια του μοντέλου. Οπότε πριν τη χρήση του Numerical to Polynominal θα επιλεγεί ο Discretize operator ο για τον χωρισμό του εύρους των αριθμητικών τιμών του A1.3 σε 5 bins και παράλληλα μετατρέπει τα εύρη αυτά σε nominal. Στη συνέχεια θα χρησιμοποιηθεί ο Filter Examples operator για να απαλειφθούν τα παραδείγματα εκείνα που περιέχουν ελλιπείς τιμές στα χαρακτηριστικά A1.3 και B2.4. Τέλος υλοποιούμε τον Cross Validation operator που περιέχει εμφωλιασμένους τους operators ID3 επιλέγοντας ως criterion το information_gain, Apply Model, Performance (Binary Classification) όπως κάναμε και με τον Naive Bayes για την αξιολόγηση του μοντέλου.



Εικόνα 21



Εικόνα 22

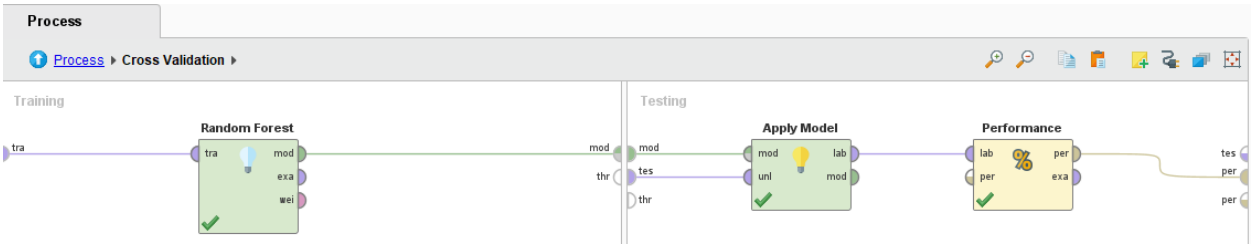
PerformanceVector

```
PerformanceVector:
accuracy: 87.91% +/- 7.64% (micro average: 87.85%)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    9       5
Όχι:    8       85
precision: 92.55% +/- 8.41% (micro average: 91.40%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    9       5
Όχι:    8       85
recall: 94.44% +/- 9.44% (micro average: 94.44%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    9       5
Όχι:    8       85
f_measure: 92.86% +/- 4.63% (micro average: 92.90%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    9       5
Όχι:    8       85
```

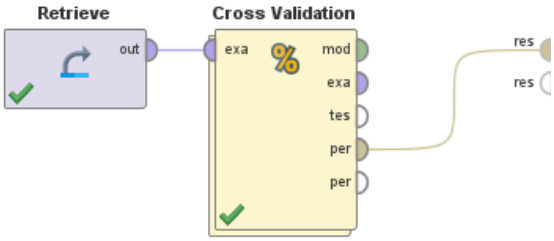
Εικόνα 23

5.4 Εκπαίδευση και Αξιολόγηση - Random Forest

Ο Random Forest αν και χρησιμοποιεί δέντρα απόφασης δεν υπόκειται στους περιορισμούς του ID3 λόγω του τρόπου λειτουργίας του όπως είδαμε στην προηγούμενη ενότητα. Η απόδοσή του δηλαδή δεν επηρεάζεται από τις ελλιπείς τιμές και μπορεί να υλοποιηθεί χωρίς πρόβλημα όταν τα χαρακτηριστικά περιέχουν αριθμητικές τιμές. Οπότε μπορούμε άμεσα να υλοποιήσουμε και να αξιολογήσουμε το μοντέλο με Cross Validation χωρίς ενδιάμεσους operators.



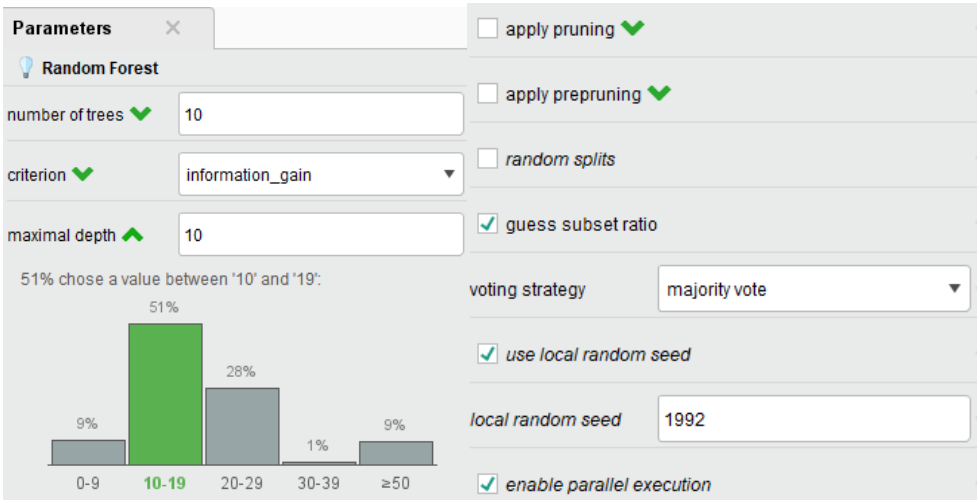
Εικόνα 24



Εικόνα 25

Το ενδιαφέρον εδώ υπάρχει στον ορισμό των ποικίλων παραμέτρων του operator Random Forest. Στο number of trees ορίζουμε τον αριθμό των διαφορετικών δέντρων απόφασης που θα δημιουργηθούν παράλληλα και μάλιστα σε πραγματικό χρόνο χρησιμοποιώντας διαφορετικά threads της CPU επιλέγοντας την τελευταία παράμετρο enable parallel execution. Έπειτα στο πεδίο criterion μπορούμε να ορίσουμε το κριτήριο με το οποίο θα γίνει ο διαχωρισμός στα δέντρα απόφασης όπως και με τον ID3 καθώς και στο maximal depth να περιορίσουμε το μέγιστο βάθος δηλαδή το μέγεθος του κάθε δέντρου. Αξίζει να σημειώσουμε ότι στις περισσότερες παραμέτρους το RapidMiner μας δίνει τη δυνατότητα να δούμε τις δημοφιλέστερες τιμές. Έπειτα υπάρχουν οι επιλογές για pruning πριν ή/και μετά την υλοποίηση του μοντέλου και η επιλογή random splits για τον τυχαίο διαχωρισμό των χαρακτηριστικών με αριθμητικές τιμές υλοποιώντας ουσιαστικά την παραλλαγή του Random Forest τον Extra Trees. Με την επιλογή guess subset ratio χρησιμοποιείται ένας τυχαίος αριθμός int(log(m) + 1) χαρακτηριστικών σε κάθε δέντρο αντί ενός τυχαίου ορισμένου μέρους αυτών. Στην παράμετρο voting strategy ορίζουμε majority vote δηλαδή την επιλογή της κλάσης που προβλέφθηκε από τα περισσότερα δέντρα.

Επιπλέον υπάρχει και η παράμετρος local random seed για τη χρήση ενός αριθμού-φύτρας για τη περαιτέρω βελτιστοποίηση της τυχαιότητας στον Random Forest και των παραμέτρων του.



Εικόνα 26

Οπότε το accuracy του μοντέλου που προκύπτει είναι 88.33%. Χρησιμοποιώντας και τις δυο επιλογές για pruning το accuracy αυξάνεται στο 90.06%. Λίγο μικρότερο accuracy 88.40% φαίνεται ότι προκύπτει επιλέγοντας random splits χωρίς pruning.

PerformanceVector

```
PerformanceVector:
accuracy: 88.33% +/- 7.03% (micro average: 88.43%)
```

PerformanceVector

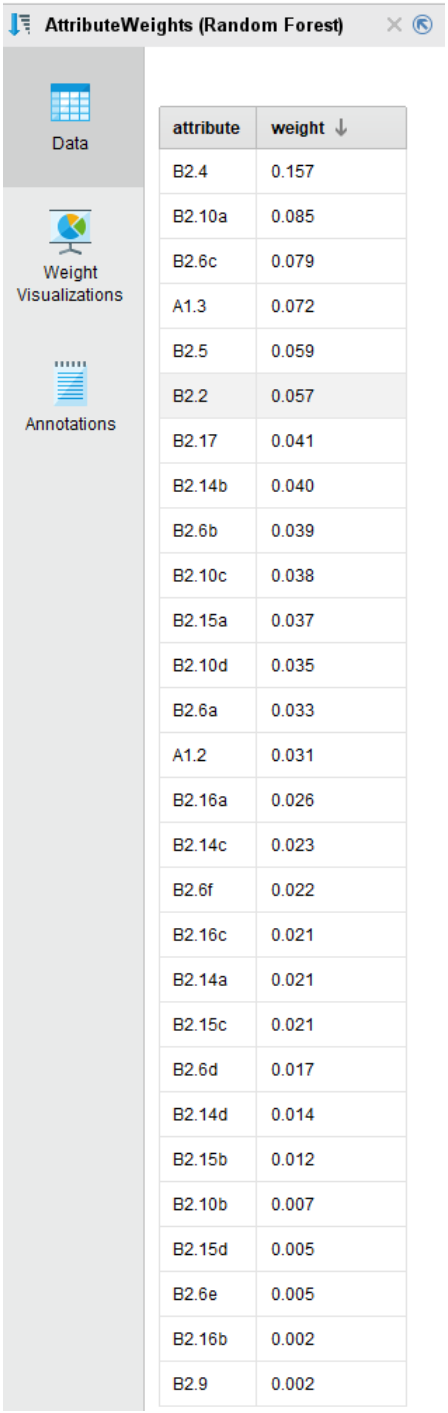
```
PerformanceVector:
accuracy: 88.40% +/- 7.06% (micro average: 88.43%)
```

PerformanceVector

```
PerformanceVector:
accuracy: 90.06% +/- 5.30% (micro average: 90.08%)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    18      2
Όχι:    10      91
precision: 90.85% +/- 7.06% (micro average: 90.10%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    18      2
Όχι:    10      91
recall: 97.89% +/- 4.46% (micro average: 97.85%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    18      2
Όχι:    10      91
f_measure: 93.96% +/- 2.86% (micro average: 93.81%) (positive class: Όχι)
ConfusionMatrix:
True:   Ναι     Όχι
Ναι:    18      2
Όχι:    10      91
```

Εικόνα 27

Αν παρατηρήσουμε τον operator Random Forest θα δούμε πως μία από τις εξόδους ονομάζεται wei. Αυτή μας δίνει τη δυνατότητα να δούμε στα αποτελέσματα τα βάρη των χαρακτηριστικών δηλαδή τη σημαντικότητά τους στη δημιουργία του μοντέλου. Τη δυνατότητα αυτή το RapidMiner μας τη δίνει και μέσω ειδικά διαμορφωμένων operator Weight by X όπου X το κριτήριο (Information Gain, Gini Index, Deviation κτλ) με βάση τον οποίο θα γίνει ο υπολογισμός των βαρών αλλά και βελτιστοποίηση αυτών με τη χρήση διαφόρων τεχνικών μέσω των Optimize Weights operators. Σε κάποιες περιπτώσεις προβλέψεων τα βάρη είναι πολύ χρήσιμα στη λήψη αποφάσεων σε συνδυασμό φυσικά με τα αποτελέσματα των μοντέλων. Επίσης βάρη μπορούν να χρησιμοποιηθούν ακόμα και σε τιμές κλάσεων ή παραδειγμάτων όταν η συχνότητα εμφάνισης της τιμής αυτής είναι πολύ μικρή αλλά πολύ σημαντική όπως για παράδειγμα για την ανίχνευση απάτης στις χρηματοπιστωτικές συναλλαγές ή τον εντοπισμό κάποιου μεταλλαγμένου γονιδίου. Για την περίπτωσή μας, όπως φαίνεται στην παρακάτω εικόνα, το χαρακτηριστικό B2.4 παίζει τον πιο σημαντικό ρόλο για το συγκεκριμένο προγνωστικό μοντέλο ενώ τα χαρακτηριστικά B2.9, B2.16b, B2.6e, B2.15d, B2.10b τον λιγότερο σημαντικό.



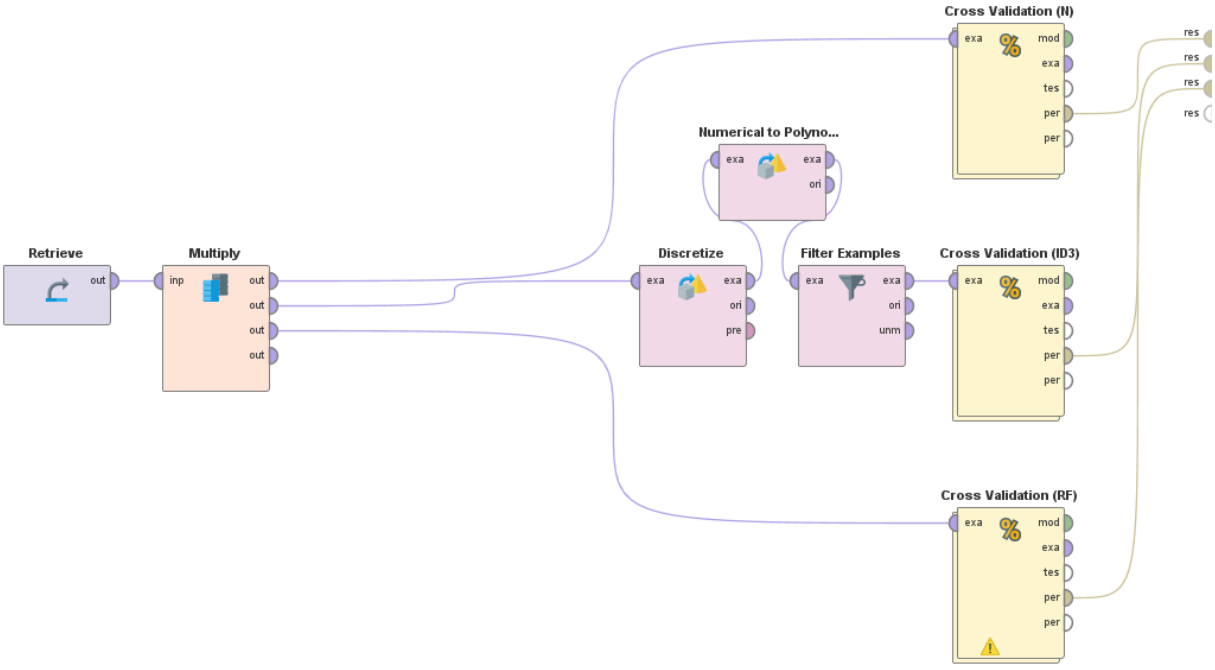
attribute	weight
B2.4	0.157
B2.10a	0.085
B2.6c	0.079
A1.3	0.072
B2.5	0.059
B2.2	0.057
B2.17	0.041
B2.14b	0.040
B2.6b	0.039
B2.10c	0.038
B2.15a	0.037
B2.10d	0.035
B2.6a	0.033
A1.2	0.031
B2.16a	0.026
B2.14c	0.023
B2.6f	0.022
B2.16c	0.021
B2.14a	0.021
B2.15c	0.021
B2.6d	0.017
B2.14d	0.014
B2.15b	0.012
B2.10b	0.007
B2.15d	0.005
B2.6e	0.005
B2.16b	0.002
B2.9	0.002

Εικόνα 28

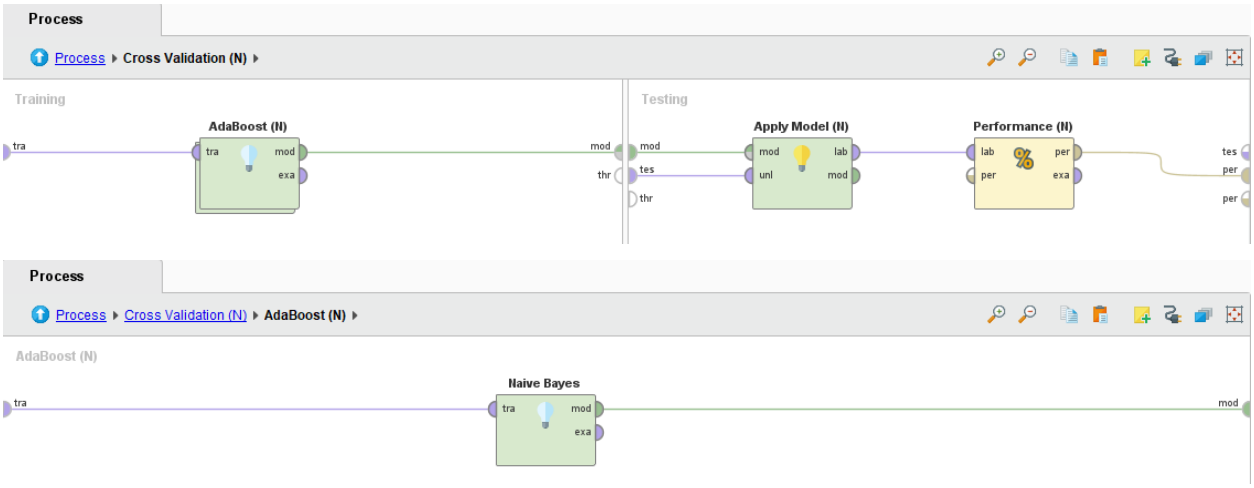
5.5 Εκπαίδευση και Αξιολόγηση - AdaBoost

Ο AdaBoost, όπως αναφέρθηκε και στην προηγούμενη ενότητα, αποτελεί έναν out-of-the-box meta classifier δηλαδή δεν έχει ιδιαίτερες παραμετροποιήσεις και υλοποιείται κυρίως σε συνδυασμό με άλλους κατηγοριοποιητές ως weak learners για την αύξηση των επιδόσεων αυτών. Στην παρούσα εργασία θα εφαρμοστεί στους προηγούμενους τρεις αλγορίθμους μηχανικής μάθησης τον Naive Bayes, ID3, Random Forest. Η πιο ενδιαφέρουσα από τις τρεις περιπτώσεις είναι με αυτή του Random Forest καθώς ουσιαστικά εφαρμόζουμε έναν ensemble boosting αλγόριθμο σε έναν ensemble bagging αλγόριθμο ως weak learner. Το πως ακριβώς συμβαίνει αυτό είναι αρκετά περίπλοκο να εξηγηθεί θεωρητικά και αποτελεί αντικείμενο έρευνας των τελευταίων χρόνων αν και το μοντέλο αυτό έχει υλοποιηθεί σε κάποιες μελέτες όπως την πρόβλεψη επιβίωσης ασθενών από τον καρκίνο του μαστού. Παρόλα αυτά χάρη στην εξέλιξη της τεχνολογίας και των εφαρμογών όπως το RapidMiner ο μάλλον περίεργος συνδυασμός των αλγορίθμων αυτών μπορεί αν γίνει εύκολα.

Η υλοποίηση του AdaBoost με τους άλλους τρεις αλγόριθμους θα γίνει ταυτόχρονα στο RapidMiner. Για το σκοπό αυτό θα χρησιμοποιήσουμε τον operator Multiply ο οποίος μας επιτρέπει να διοχετεύσουμε τα δεδομένα μας σε τρεις Cross Validation operators οι οποίοι περιέχουν τους AdaBoost, Apply Model, Performance (Binary Classification). Ο κάθε AdaBoost operator περιέχει εμφωλιασμένο έναν από τους τρεις άλλους αλγορίθμους. Για την περίπτωση του ID3 θα χρησιμοποιηθούν όπως και προηγουμένως οι operators Numerical to Polynominal, Discretize και Filter Examples. Για τον Random Forest δε θα χρησιμοποιηθούν οι επιλογές pruning και random splits.

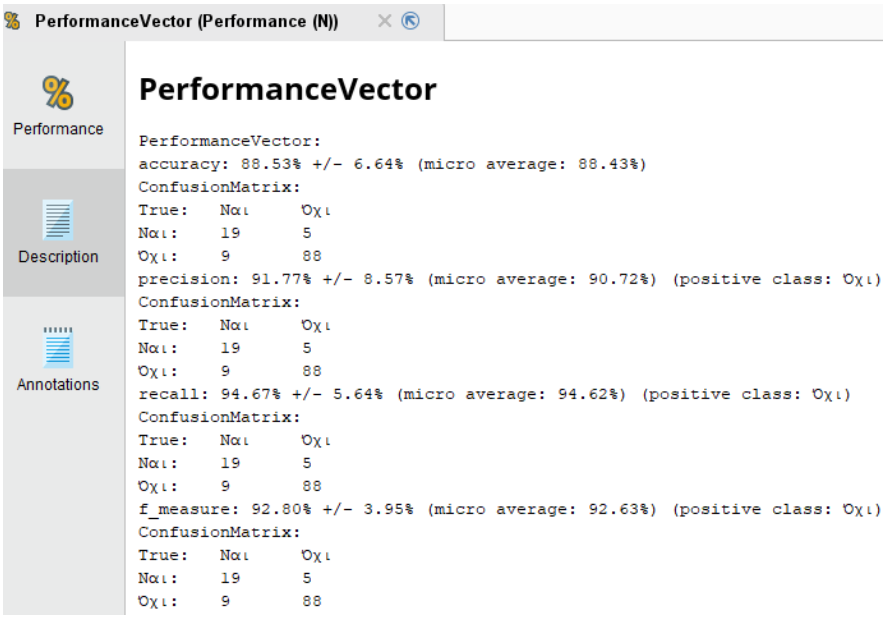


Εικόνα 29

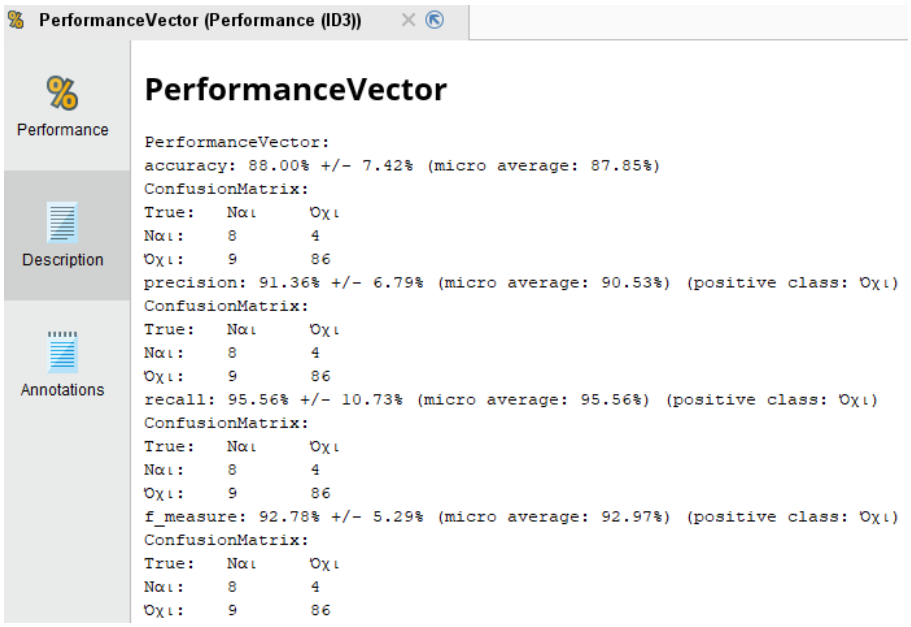


Εικόνα 30

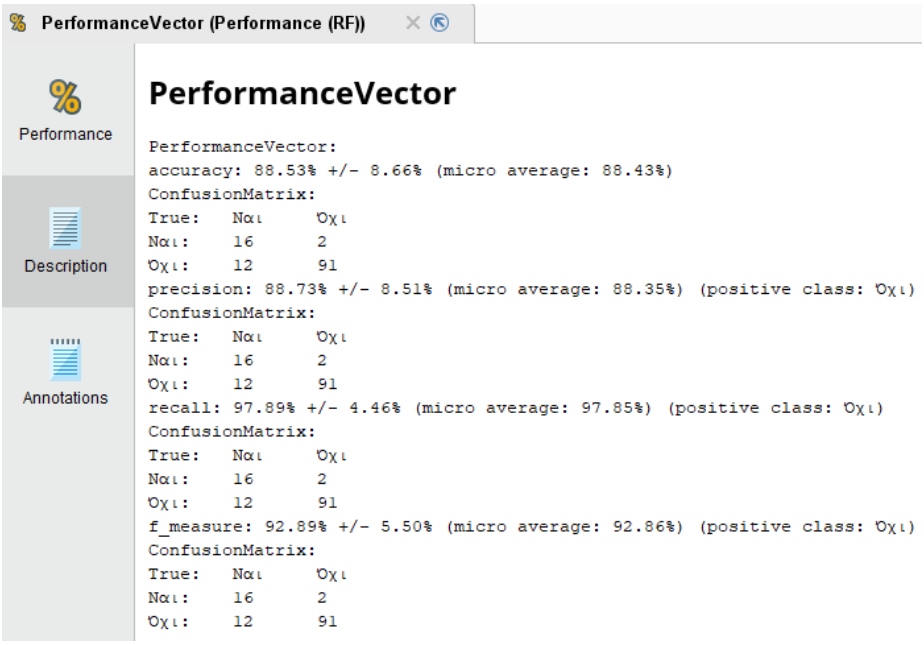
Τα accuracy που προκύπτουν είναι τα εξής: για τον Naive Bayes (N) 88.53%, για τον ID3 (ID3) 88.00%, για τον Random Forest (RF) 88.53%. Τα accuracy των υλοποιήσεων των τριών αυτών αλγορίθμων χωρίς AdaBoost ήταν 87.63%, 87.91%, 90.06% αντίστοιχα. Δηλαδή παρατηρούμε βελτίωση των επιδόσεων 0.90% για Naive Bayes, 0.09% για ID3, μείωση 1.53% για Random Forest. Γεγονός που συμφωνεί με διάφορες έρευνες που έχουν γίνει στον AdaBoost όπου φαίνεται γενικά βελτίωση της επίδοσης των αλγορίθμων.



Εικόνα 31



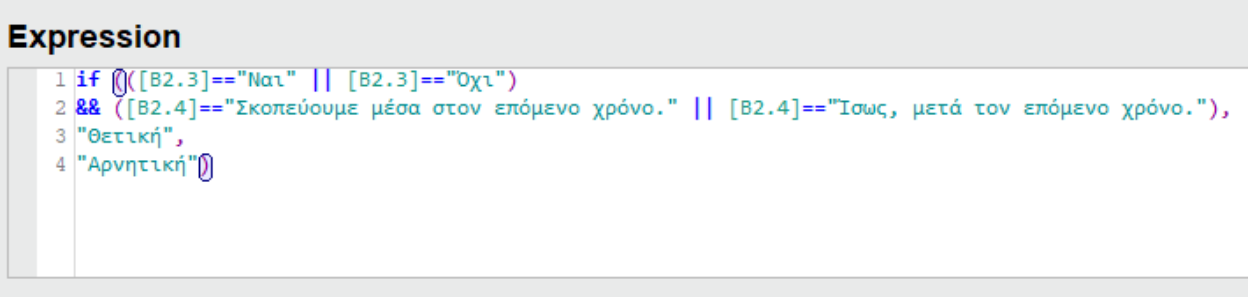
Εικόνα 32



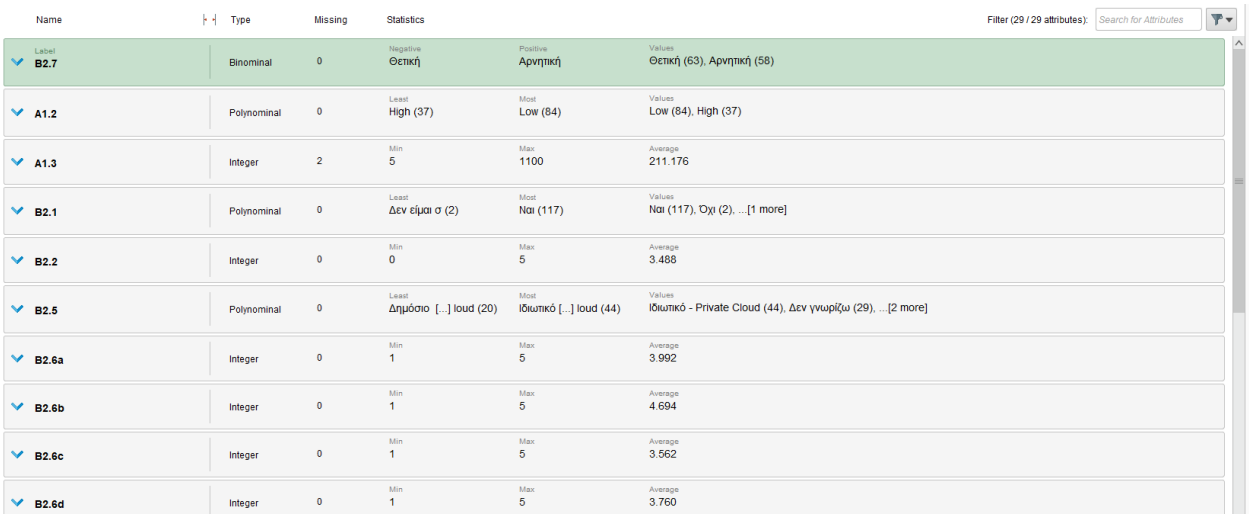
Εικόνα 33

5.6 Πρόσθετη περίπτωση πρόβλεψης

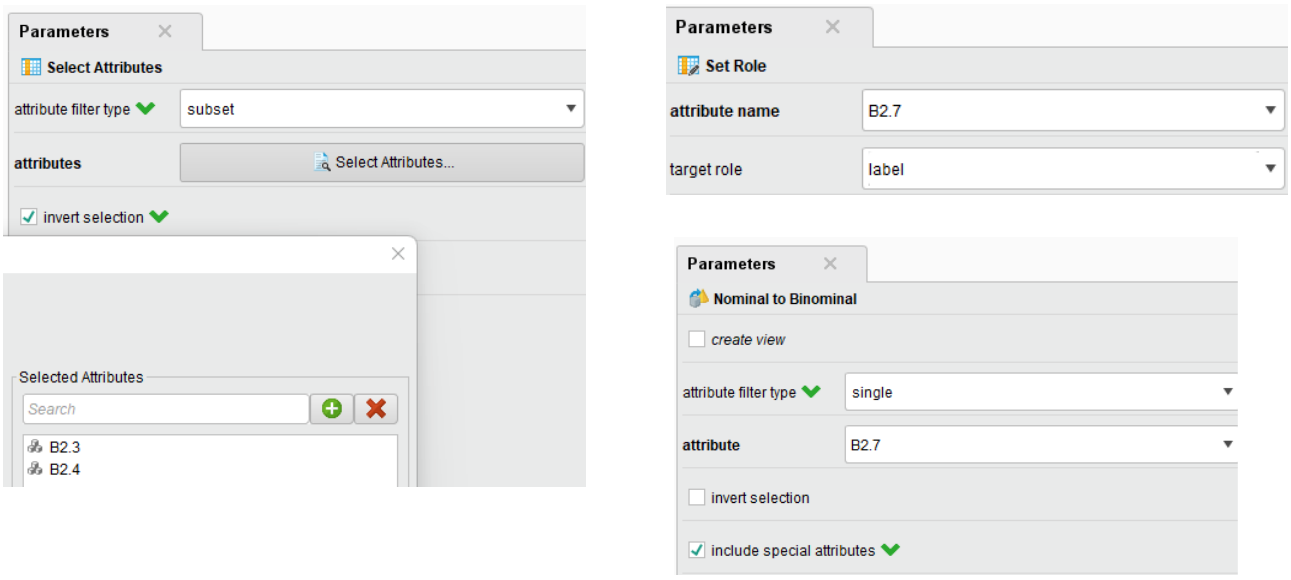
Μια άλλη ενδιαφέρουσα περίπτωση πρόβλεψης είναι να συνδυάσουμε τα δύο χαρακτηριστικά B2.3 «Στον δήμο σας κάνετε χρήση Cloud Storage;» και B2.4 «Πότε σκοπεύετε να κάνετε χρήση του Cloud Storage στον δήμο σας» σε μία νέα διωνυμική κλάση B2.7 με τιμές «Θετική» και «Αρνητική» οι οποίες δηλώνουν την τάση του δήμου για αυτή την τεχνολογία. Για το σκοπό αυτό θα χρησιμοποιήσουμε αρχικά τον operator Generate Attributes για τη δημιουργία της B2.7 η οποία θα παίρνει τιμές σύμφωνα με το script του operator που φαίνεται παρακάτω. Έπειτα με τον Set Role η B2.7 γίνεται η νέα nominal label, με τον Select Attributes επιλέγοντας invert selection αφαιρούμε εξ ολοκλήρου τις στήλες B2.3 και B2.4 από τα δεδομένα μας για να μην επηρεάσουν τα μοντέλα μας λόγω της σχέσης τους με την B2.7, με τον Nominal to Binominal την B2.7 με την επιλογή include special attributes και τον Cross Validation με εμφωλιασμένους τον εκάστοτε αλγόριθμο και τους Apply Model και Performance (Binary Classification) για την ανάπτυξη και αξιολόγηση του μοντέλου. Με τους Multiply, Write Excel και Store operators γίνεται αποθήκευση του αρχείου αυτού ως MunicipalitiesB2.7.xlsx στο repository του RapidMiner. Από τις δοκιμές μας με τους τέσσερις αλγόριθμους της εργασίας το υψηλότερο accuracy 64.29% το είχε ο Random Forest με την επιλογή random splits.



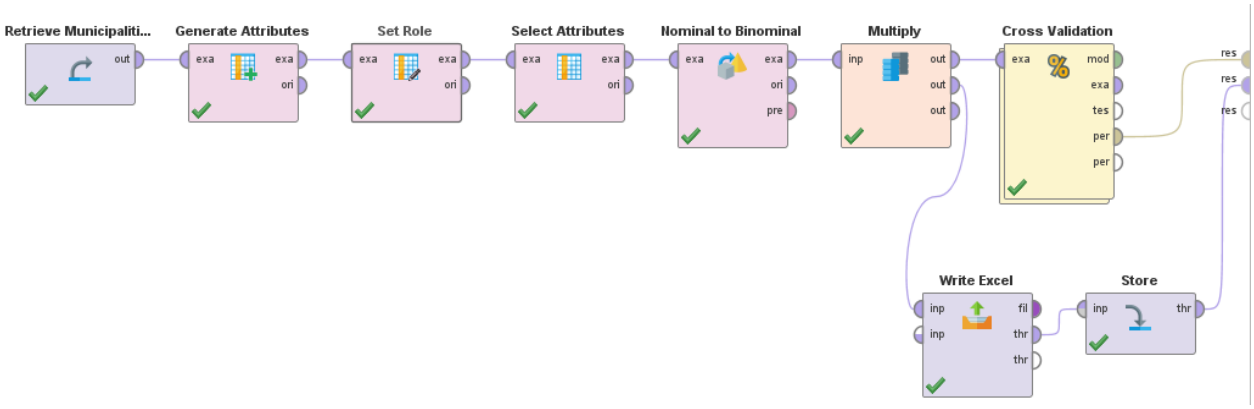
Εικόνα 34



Εικόνα 35



Εικόνα 36



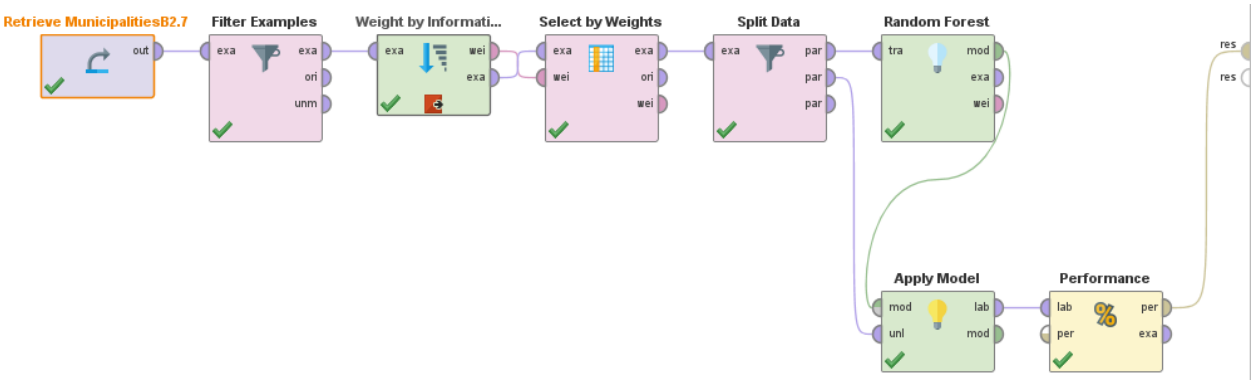
Εικόνα 37

PerformanceVector

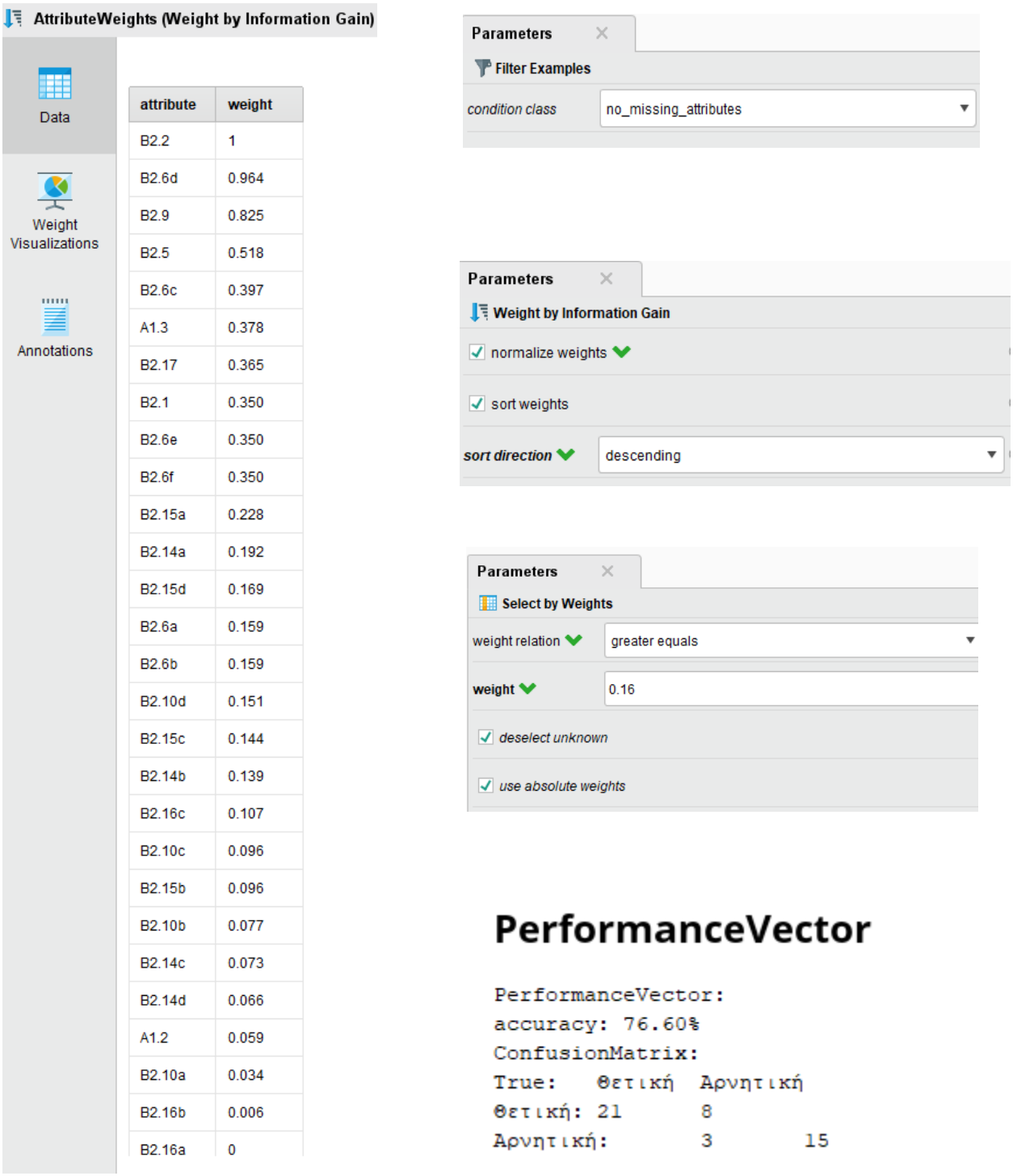
```
PerformanceVector:
accuracy: 64.29% +/- 13.27% (micro average: 64.46%)
ConfusionMatrix:
True:   Θετική  Αρνητική
Θετική: 45      25
Αρνητική:       18      33
```

Εικόνα 38

Για να αυξήσουμε την επίδοση του μοντέλου μπορούμε να υπολογίσουμε και να χρησιμοποιήσουμε τα βάρη των μεταβλητών ώστε να αφαιρέσουμε τις μεταβλητές εκείνες που δεν συνεισφέρουν σημαντικά στην ανάπτυξη του μοντέλου. Η διαχείριση των βαρών μπορεί να γίνει με πολλούς τρόπους στο RapidMiner. Ένας από αυτούς είναι με τη χρήση του operator Weight by Information Gain ο οποίος υπολογίζει τα βάρη σύμφωνα με το κριτήριο information gain, δηλαδή τις εντροπίες, όπως είδαμε και στην υλοποίηση των αλγορίθμων ID3 και Random Forest. Ο operator αυτός όμως δε μπορεί να εφαρμοστεί σε μεταβλητές με ελλιπείς τιμές και για αυτό θα γίνει φιλτράρισμα αυτών με τον Filter Examples operator. Επίσης θα χρησιμοποιηθεί και ο operator Select by Weights με τον οποίο θα γίνει επιλογή των μεταβλητών με βάρος μεγαλύτερο του 0.16 το οποίο ορίστηκε ως βέλτιστο με την παραδοσιακή μέθοδο Split Data κι όχι με Cross Validation μετά από δοκιμές. Με δεξί κλικ στον Weight by Information επιλέγουμε Breakpoint After για διακοπή της διεργασίας μετά τον operator ώστε να δούμε τα βάρη των μεταβλητών. Κατά τα λοιπά εφαρμόζουμε τους operators Split Data, Random Forest, Apply Model, Performance (Binominal Classification) με τις γνωστές παραμετροποιήσεις. Το αποτέλεσμα είναι αύξηση του accuracy του μοντέλου στο 76.60%.



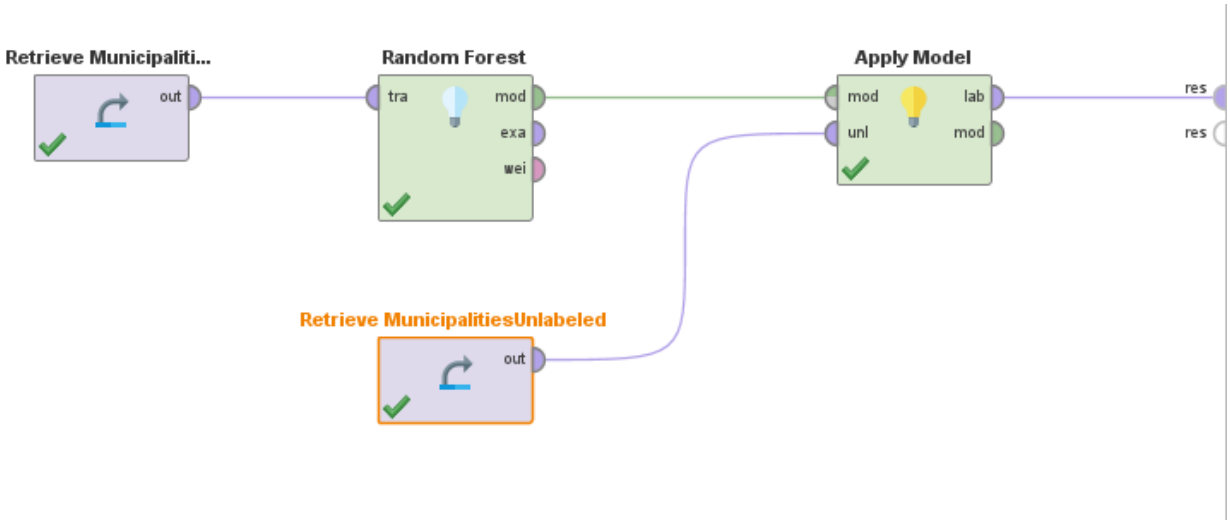
Εικόνα 39


AttributeWeights (Weight by Information Gain)
Data
Weight Visualizations
Annotations
attribute weight
B2.2 1
B2.6d 0.964
B2.9 0.825
B2.5 0.518
B2.6c 0.397
A1.3 0.378
B2.17 0.365
B2.1 0.350
B2.6e 0.350
B2.6f 0.350
B2.15a 0.228
B2.14a 0.192
B2.15d 0.169
B2.6a 0.159
B2.6b 0.159
B2.10d 0.151
B2.15c 0.144
B2.14b 0.139
B2.16c 0.107
B2.10c 0.096
B2.15b 0.096
B2.10b 0.077
B2.14c 0.073
B2.14d 0.066
A1.2 0.059
B2.10a 0.034
B2.16b 0.006
B2.16a 0
Parameters
Filter Examples
condition class
no_missing_attributes
Parameters
Weight by Information Gain
normalize weights
sort weights
sort direction
descending
Parameters
Select by Weights
weight relation
greater equals
weight
0.16
deselect unknown
use absolute weights
PerformanceVector
PerformanceVector:
accuracy: 76.60%
ConfusionMatrix:
True: Θετική Αρνητική
Θετική: 21 8
Αρνητική: 3 15

Εικόνα 40

5.7 Εφαρμογή των μοντέλων

Πώς όμως τα παραπάνω μοντέλα εφαρμόζονται στην πράξη; Μπορούμε να το δούμε με ένα απλό παράδειγμα στο RapidMiner. Έστω ότι έχουμε ένα νέο αρχείο δεδομένων MunicipalitiesUnlabeled.xlsx με 12 νέες εγγραφές (10% του αρχικού σετ) και τον ίδιο αριθμό χαρακτηριστικών αλλά δεν περιέχει το label B2.3. Στόχος μας είναι η δημιουργία και πρόβλεψη τιμών του B2.3 χρησιμοποιώντας ένα από τα παραπάνω μοντέλα προγνωστικής αναλυτικής. Αυτό μπορεί να γίνει εύκολα με δυο retrieve operators για τα δύο αρχεία MunicipalitiesModified.xlsx και MunicipalitiesUnlabeled.xlsx, τον operator Random Forest με τις ίδιες παραμέτρους που χρησιμοποιήσαμε στις προηγούμενες ενότητες και τον operator Apply Model.



Εικόνα 41

Στα αποτελέσματα παρατηρούμε ότι εκτός από τη νέα στήλη prediction(B2.3) και τις τιμές που έχει προβλέψει το μοντέλο, υπάρχουν και άλλες δυο νέες στήλες confidence(Ναι) και confidence(Όχι) με τιμές μεταξύ 0 και 1 οι οποίες δηλώνουν το πόσο «σίγουρες» είναι οι προβλέψεις του μοντέλου για τις συγκεκριμένες τιμές της B2.3. Ουσιαστικά κάθε αλγόριθμος μηχανικής μάθησης διωνυμικών κατηγοριοποιητών κάνει μία πρόβλεψη πιθανότητας για την positive τιμή της κλάσης που στο παράδειγμά μας είναι όπως είχαμε αναφέρει το «Όχι». Εάν αυτή είναι μεγαλύτερη του 0.5 αναθέτει την τιμή «Όχι» και αν είναι μικρότερη την τιμή «Ναι» ορίζοντας και τις αντίστοιχες confidence τιμές. Δηλαδή αν η πιθανότητα για «Όχι» είναι π.χ. 0.3 τότε ο αλγόριθμος προβλέπει «Ναι» με confidence(Ναι)=0.700 και confidence(Όχι)=0.300 όπως φαίνεται στην επόμενη εικόνα. Αυτή η πιθανότητα 0.5 μπορεί να αλλάξει χρησιμοποιώντας τους operators Create Threshold και Apply Threshold κάτι το οποίο μπορεί να είναι πολύ χρήσιμο σε κάποια προβλήματα όταν για παράδειγμα θέλουμε μια πρόβλεψη να γίνεται μόνο όταν η πιθανότητα αυτής είναι πολύ υψηλή. Μπορούμε ακόμα και να βρούμε τα ιδανικά thresholds με διάφορες τεχνικές (καμπύλες ROC, PR κ.α.) η ανάλυση των οποίων, αν και πολύ ενδιαφέρον, είναι εκτός του πλαισίου της εργασίας αυτής.

Row No.	prediction(B2.3)	confidence(Ναι)	confidence(Όχι)	A1.2	A1.3	B2.1	B2.2
1	Όχι	0	1	Low	100	Ναι	3
2	Όχι	0.100	0.900	Low	120	Ναι	2
3	Ναι	0.700	0.300	Low	300	Όχι	3
4	Όχι	0.400	0.600	Low	90	Ναι	5
5	Όχι	0.300	0.700	High	230	Ναι	1
6	Όχι	0.100	0.900	Low	10	Όχι	2
7	Ναι	0.800	0.200	High	80	Ναι	3
8	Όχι	0.200	0.800	Low	140	Ναι	4
9	Όχι	0.200	0.800	Low	70	Ναι	4
10	Όχι	0.200	0.800	High	80	Όχι	5
11	Ναι	0.800	0.200	High	200	Ναι	1
12	Όχι	0.100	0.900	Low	170	Ναι	3

Εικόνα 42

6. Ανάλυση Αποτελεσμάτων - Συμπεράσματα

Όπως είδαμε στην προηγούμενη ενότητα ανάπτυξης των μοντέλων, η αξιολόγησή τους έγινε κυρίως με το accuracy. Αρκετές φορές αυτό όμως δεν είναι αρκετό και χρειάζονται κι άλλες παραμέτρους που για την περίπτωση των διωνυμικών κλάσεων κατηγοριοποίησης οι πιο γνωστές είναι οι precision, recall, f measure. Μπορούν να υπολογιστούν χρησιμοποιώντας στοιχεία του confusion matrix δηλαδή τις σωστές προβλέψεις της positive τιμής (TruePositive – TP), τις λάθος προβλέψεις της positive τιμής (FalePositive – FP), τις σωστές προβλέψεις της negative τιμής (TrueNegative – TN) και λάθος προβλέψεις της negative τιμής (FalseNegative – FN). Το ποια είναι positive και ποια negative καθορίζεται αυτόματα από το RapidMiner και την εισαγωγή του αρχείου των δεδομένων μας αλλά μπορεί να αλλάξει με τον κατάλληλο operator όπως τον Remap Binominals. Στην περίπτωσή μας το RapidMiner όρισε για την B2.3 ως positive την τιμή «Όχι» και ως negative την τιμή «Ναι» κάτι το οποίο είναι θεμιτό καθώς πρωτίστως μας ενδιαφέρουν οι δήμοι που δεν έχουν cloud storage. Οπότε οι παράμετροι αξιολόγησης υπολογίζονται με βάση τις εξής σχέσεις:

$$Accuracy = \frac{Σωστές\ Προβλέψεις}{Αριθμός\ Παραδειγμάτων} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{Σωστές\ Προβλέψεις\ "Όχι"}{Προβλέψεις\ "Όχι"} = \frac{TP}{TP + FP}$$

$$Recall = \frac{Σωστές\ Προβλέψεις\ "Όχι"}{Σωστές\ Προβλέψεις\ "Όχι" + Λάθος\ Προβλέψεις\ "Ναι"} = \frac{TP}{TP + FN}$$

$$F\ Measure = \frac{2(Precision * Recall)}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

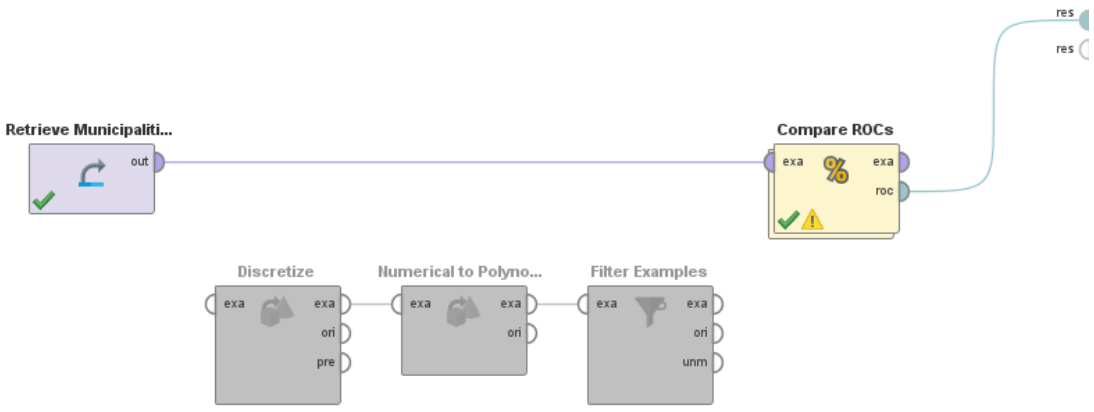
Όπως βλέπουμε από τις σχέσεις Accuracy είναι το ποσοστό των συνολικών σωστών προβλέψεων, Precision (ή Positive Prediction Value – PPV) το ποσοστό των σωστών positive προβλέψεων από όλες τις positive προβλέψεις, Recall (ή Sensitivity ή TP rate ή Hit Rate) το ποσοστό των σωστών positive προβλέψεων σε όλα τα positive παραδείγματα, F Measure (ή F1 Score) ο αρμονικός μέσος των Precision και Recall ιδιαίτερα χρήσιμος για imbalanced classes δηλαδή όταν μια τιμή μιας κλάσης έχει πολύ μεγαλύτερη συχνότητα από άλλες. Κάποιες φορές χρησιμοποιείται και το Specificity (ή TN rate) που είναι το αντίστοιχο του Recall για τα TN δηλαδή το ποσοστό των σωστών negative προβλέψεων σε όλα τα negative παραδείγματα.

Το αν είναι σημαντικό το Precision ή το Recall ή και τα δυο εξαρτάται από την βαρύτητα των FP και FN αντίστοιχα στο εκάστοτε πρόβλημα. Συγκεκριμένα στην περίπτωσή μας FP σημαίνει ότι το μοντέλο πρόβλεψε λανθασμένα ότι ένας δήμος δεν έχει cloud storage οπότε πιθανώς η εταιρεία παροχής υπηρεσιών cloud storage να ξοδέψει χρήματα για προωθητικές ενέργειες σε κάποιους δήμους χωρίς τα αναμενόμενα αποτελέσματα. Αντίστοιχα για το FN σημαίνει ότι το μοντέλο πρόβλεψε λανθασμένα ότι κάποιοι δήμοι έχουν cloud storage χωρίς όμως αυτό να συμβαίνει στην πραγματικότητα με αποτέλεσμα η εταιρεία να χάσει δήμους-πελάτες. Όπως γίνεται εύκολα αντιληπτό στην περίπτωσή μας το FN είναι πολύ σημαντικό, οπότε μας ενδιαφέρει αυτό να είναι χαμηλό κι άρα το Recall υψηλό. Συγκεντρωτικά τα αποτελέσματα των μοντέλων φαίνονται στον παρακάτω πίνακα.

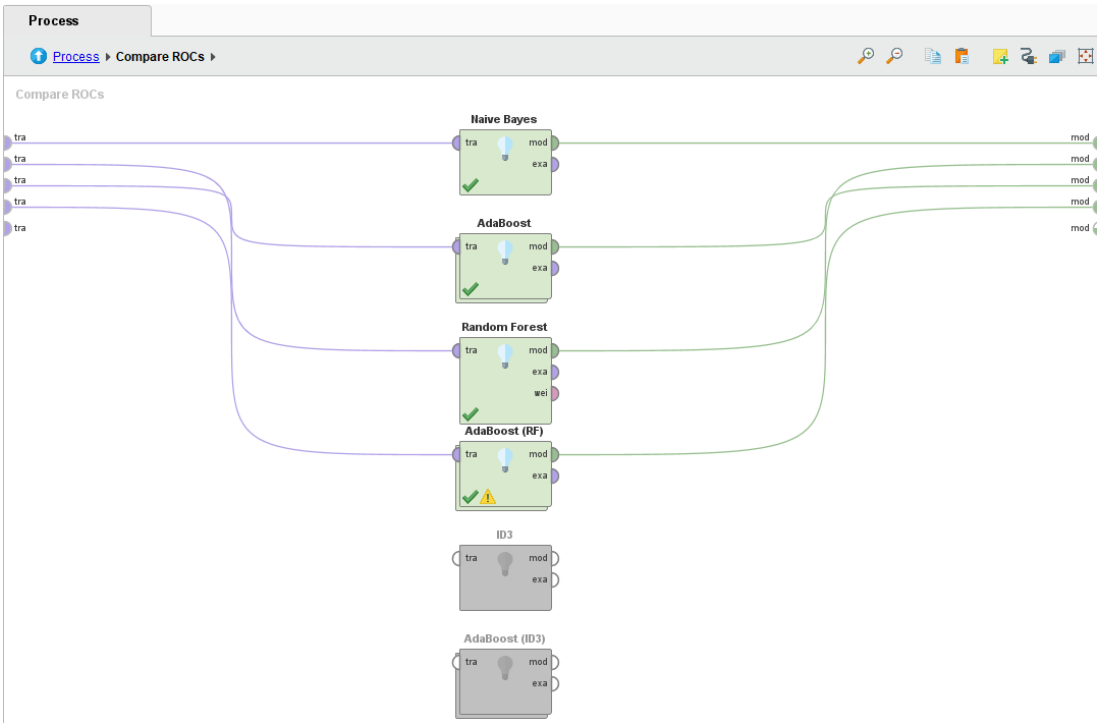
Model	*Accuracy*	*Precision*	*Recall*	*F Measure*
Naive Bayes	87.63%	92.31%	91.33%	91.79%
ID3	87.91%	92.55%	94.44%	92.86%
Random Forest	90.06%	90.85%	97.89%	93.96%
AdaBoost (N)	88.53%	91.77%	94.67%	92.80%
AdaBoost (ID3)	88.00%	91.36%	95.56%	92.78%
AdaBoost (RF)	88.53%	88.73%	97.89%	92.89%

Όπως παρατηρούμε στο Accuracy καλύτερος είναι ο Random Forest και χειρότερος ο Naive Bayes (με Cross Classification), στο Precision καλύτερος ο ID3 και χειρότερος ο AdaBoost(RF), στο Recall καλύτεροι οι Random Forest και AdaBoost(RF) και χειρότερος ο Naive Bayes, στο F Measure οριακά καλύτερος ο Random Forest και χειρότερος ο Naive Bayes. Γενικά φαίνεται πως η αποδόσεις των μοντέλων δεν έχουν μεγάλες διαφορές που είναι ένδειξη ότι έγινε καλή προετοιμασία των δεδομένων πριν την ανάπτυξη των μοντέλων ή ότι απλά δεν υπήρχαν προβλήματα εξ αρχής. Μια επίσης σημαντική παρατήρηση είναι ότι το Recall όλων το μοντέλων είναι σχεδόν ισότιμο δηλαδή με άλλα λόγια οι λανθασμένες προβλέψεις ότι δήμοι έχουν cloud storage ενώ δεν έχουν θα είναι ελάχιστες κι άρα και τα διαφυγόντα κέρδη της εταιρείας.

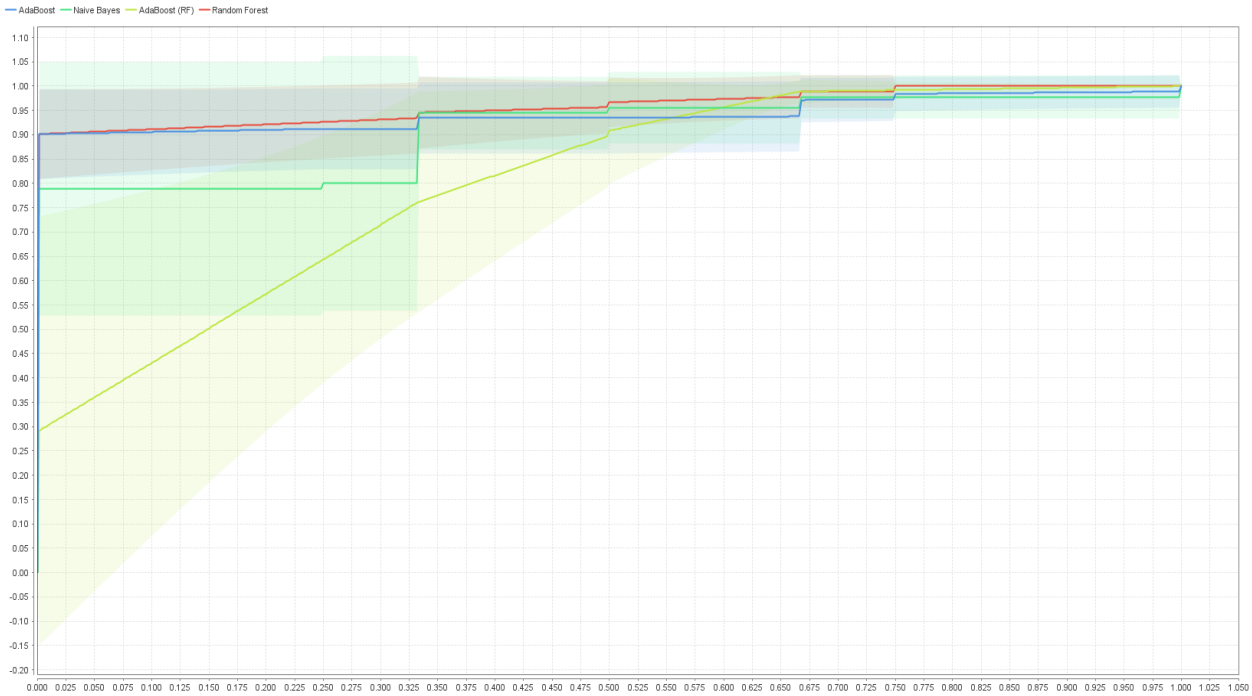
Παρόλα αυτά να αξίζει να δούμε ακόμα ένα κριτήριο της απόδοσης των μοντέλων προγνωστικής ανάλυσης που χρησιμοποιείται συχνά, τις καμπύλες Receiver Operating Characteristic (ROC). Αυτό μπορεί να υλοποιηθεί εύκολα στο RapidMiner με τον operator Compare ROCs ο οποίος περιέχει εμφωλιασμένους τους operators των αλγορίθμων αλλά λειτουργεί κι ως Cross Validation για την αξιολόγηση των μοντέλων. Βέβαια μπορεί να λειτουργήσει κι ως κλασσικός αξιολογητής με split ratio αν θέσουμε -1 στην επιλογή number of folds (το οποίο by default είναι 10 όπως κι ο Cross Validation). Αρχικά θα συγκρίνουμε τις ROC των Naive Bayes, Random Forest, AdaBoost(N), AdaBoost(RF) επειδή ο ID3 operator λειτουργεί μόνο με χαρακτηριστικά nominal και χωρίς ελλιπείς τιμές.



Εικόνα 43



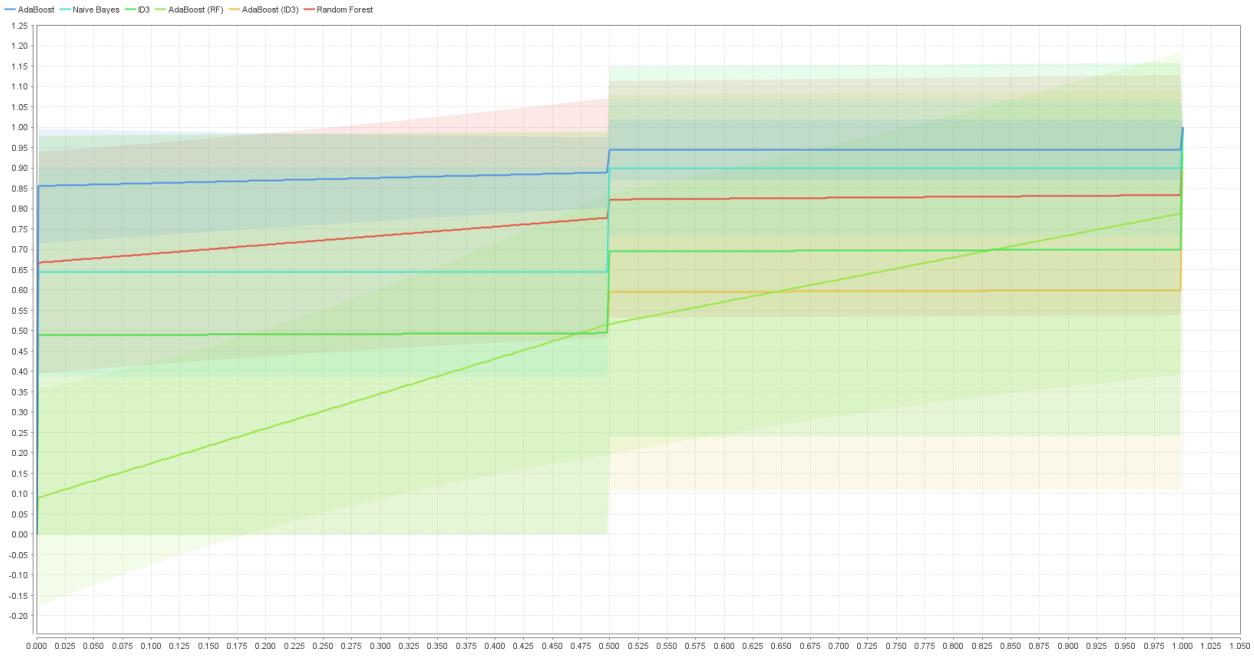
Εικόνα 44



Εικόνα 45

Στις καμπύλες ROC ο άξονας x αναπαριστά το FP Rate (FP/FP+TN) το οποίο στη βιβλιογραφία αναφέρεται πολλές φορές κι ως «Κόστη» κι ο άξονας y αναπαριστά το TP Rate ή Recall γνωστό κι ως «Κέρδη». Η διαγώνιος ευθεία που περνά από την αρχή των αξόνων περιγράφει τους αλγόριθμους που απλά μαντεύουν δηλαδή τις μισές φορές προβλέπουν σωστά τις positive και τις μισές φορές προβλέπουν σωστά τις negative τιμές της εξαρτημένης μεταβλητής. Γενικά καλοί αλγόριθμοι θεωρούνται αυτοί των οποίων οι καμπύλες βρίσκονται στην άνω τριγωνική περιοχή. Για να το πούμε απλά όσο πιο πολύ πλησιάζει η καμπύλη στην άνω αριστερή πλευρά των άνω τριγώνου τόσο καλύτερο είναι το μοντέλο προγνωστικής αναλυτικής.

Το βασικό πλεονέκτημα των ROC καμπυλών έναντι των άλλα κριτηρίων αξιολόγησης είναι ότι δεν εξαρτώνται από τις συχνότητες των τιμών της εξαρτημένης μεταβλητής κάτι το οποίο μπορεί να επηρεάσει την απόδοση κάποιων μοντέλων. Πράγματι στην προηγούμενη εικόνα παρατηρούμε πως οι καμπύλες των μοντέλων που υλοποιήσαμε πλησιάζουν την άνω αριστερή πλευρά πλην ενός: του AdaBoost (RF). Αυτό δηλαδή σημαίνει πως η ανάπτυξη του μοντέλου AdaBoost (RF) επηρεάστηκε σε σημαντικό βαθμό από την ανισοκατανομή των τιμών της κλάσης και οι μελλοντικές προβλέψεις δε θα είναι αντικειμενικές αλλά με υψηλό bias, αν και τα αποτελέσματα των υπολοίπων κριτηρίων αξιολόγησης του ήταν εξαιρετικά. Ακόμα χειρότερα είναι τα αποτελέσματα για τον AdaBoost(RF) όταν στην σύγκριση ROC περιληφθούν οι ID3, AdaBoost (ID3) χρησιμοποιώντας τους operators Discretize, Numerical to Polynominal, Filter Examples. Το συγκεκριμένο ενδιαφέρον φαινόμενο αποτελεί πολύ-επίπεδο πεδίο έρευνας των τελευταίων χρόνων και αποκαλύπτει ότι ακόμα και στα μοντέλα προγνωστικής αναλυτικής κάποια πράγματα ίσως τελικά να μην είναι όπως φαίνονταν αρχικά.



Εικόνα 46

Βιβλιογραφικές Αναφορές

1. Kyriakou, Niki, Loukis Euripides, and Dimitropoulou Paraskevi. "Factors affecting cloud storage adoption by Greek municipalities." In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, pp. 244-253. 2020.
2. Brian Hayes. 2008. Cloud computing. Commun. ACM 51, 7 (July 2008), 9–11.
3. Ruparelia, Nayan B. "Cloud Computing: A Paradigm Shift?" In Cloud Computing. The MIT Press, 2016. MIT Press Scholarship Online, 2016.
4. Bhoyar, R. and Chopde, N., 2013. Cloud computing: Service models, types, database and issues. International Journal of Advanced Research in Computer Science and Software Engineering.
5. Wu, J., Ping, L., Ge, X., Wang, Y. and Fu, J., 2010, June. Cloud storage as the infrastructure of cloud computing. In 2010 International conference on intelligent computing and cognitive informatics.
6. Cloud Computing Services - Amazon Web Services (AWS), https://aws.amazon.com/what-is-cloud-storage/
7. Ali, O., Shrestha, A., Osmanaj, V. and Muhammed, S. (2021), "Cloud computing technology adoption: an evaluation of key factors in local governments"
8. European Commission, Data Visualization Tool - Data & Indicators, https://digital-agenda-data.eu/
9. European Commission, Shaping Europe's digital future, eGovernment benchmark 2021, https://digital-strategy.ec.europa.eu/en/library/egovernment-benchmark-2021
10. Eurostat - DataExplorer, https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_bde15ei&lang=en
11. UN E-Government Knowledgebase, https://publicadministration.un.org/egovkb/en-us/
12. Wayne W. Lamorte, Behavioral Change Models: Diffusion of Innovation Theory, Boston University, 2016
13. G-Cloud Κοινωνία της πληροφορίας Α.Ε. https://www.gsis.gr/en/public-administration/G-Cloud
14. Νάνος, Ιωάννης. "Η υιοθέτηση της υπολογιστικής νέφους (cloud computing) στην τοπική αυτοδιοίκηση." (2019).
15. Siegel, Eric. Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons, 2013.
16. Fitz-Enz, Jac, and I. I. John Mattox. Predictive analytics for human resources. John Wiley & Sons, 2014.
17. Larose, Daniel T. Data mining and predictive analytics. John Wiley & Sons, 2015.
18. Business Insights-Harvard Business School Online, https://online.hbs.edu/blog/post/types-of-data-analysis
19. IBM SPSS Modeler CRISP-DM Guide, https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-introduction-crisp-dm
20. Kumar, Vaibhav, and M. L. Garg. "Predictive analytics: a review of trends and techniques." International Journal of Computer Applications 182, no. 1 (2018): 31-37.
21. Cambridge Analytica Files, https://www.theguardian.com/news/series/cambridge-analytica-files

22. Twitter Admits Algorithm Bias, https://www.theguardian.com/technology/2021/oct/22/twitter-admits-bias-in-algorithm-for-rightwing-politicians-and-news-outlets
23. Dunham, Margaret H. Data mining: Introductory and advanced topics. Pearson Education India, 2006
24. RapidMiner Documentation, https://docs.rapidminer.com/latest/studio/
25. IBM Cloud Education, https://www.ibm.com/cloud/learn/education
26. Loukis, Euripidis, Niki Kyriakou, and Manolis Maragoudakis. "Using Government Data and Machine Learning for Predicting Firms' Vulnerability to Economic Crisis." In International Conference on Electronic Government, pp. 345-358. Springer, Cham, 2020.
27. Schapire, Robert E. "Explaining adaboost." In Empirical inference, pp. 37-52. Springer, Berlin, Heidelberg, 2013.
28. Thongkam, Jaree, Guandong Xu, and Yanchun Zhang. "AdaBoost algorithm with random forests for predicting breast cancer survivability." In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 3062-3069. IEEE, 2008.
29. Oza, Aditya. "Fraud Detection using Machine Learning." TRANSFER 528812, no. 4097 (2018): 532909.
30. Individuals using the Internet (% of population), https://data.worldbank.org/indicator/IT.NET.USER.ZS
31. Public cloud services end-user spending worldwide from 2017 to 2023, https://www.statista.com/statistics/273818/global-revenue-generated-with-cloud-computing-since-2009/
32. Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27, no. 8 (2006): 861-874

Παράρτημα

Το ερωτηματολόγιο περιείχε τα παρακάτω 39 ερωτήματα και υπό-ερωτήματα με πιθανές απαντήσεις (αν υπάρχουν) από τα οποία στην παρούσα εργασία χρησιμοποιήθηκαν τα 29 ως ανεξάρτητες μεταβλητές και 1 ως εξαρτημένη μεταβλητή για την ανάπτυξη των μοντέλων. Η αρίθμηση των ερωτήσεων είναι όπως εμφανίζεται στα ονόματα των στηλών στο Excel.

Η χρήση του Cloud Storage από τους Δήμους της Ελλάδας

Γενικές Πληροφορίες

A1.1 Όνομα Δήμου

A1.2 Νομός

A1.3 Αριθμός υπαλλήλων

A1.4 Ποιος είναι ο ρόλος σας στον δήμο;

A1.5 E-mail για αποστολή αποτελεσμάτων έρευνας

Χρήση Cloud Storage

B2.1 Γνωρίζετε τον όρο Cloud Storage?

- Ναι - Όχι - Δεν είμαι σίγουρος

B2.2 Αν ναι, σε τι βαθμό αξιολογείται την γνώση σας?

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου: 5 = σε πολύ μεγάλο βαθμό, 4 = σε μεγάλο βαθμό, 3 = σε μέτριο βαθμό, 2 = σε μικρό βαθμό και 1 = καθόλου (βάλτε σε κύκλο τον κατάλληλο αριθμό)

1 2 3 4 5

B2.3 Στον δήμο σας κάνετε χρήση Cloud Storage;

- Ναι - Όχι

Β2.4 Αν όχι, πότε σκοπεύετε να κάνετε χρήση του Cloud Storage στον δήμο σας;

- Σκοπεύουμε μέσα στον επόμενο χρόνο.
- Ίσως, μετά τον επόμενο χρόνο.
- Ακόμη δεν έχουμε την κατάλληλη γνώση σχετικά με την χρήση του cloud storage.
- Δεν θέλουμε να κάνουμε χρήση αυτής της υπηρεσίας.

Β2.5 Σύμφωνα με τις ανάγκες σας, τι μοντέλο Cloud Storage χρησιμοποιείτε ή θα χρησιμοποιήσετε στο μέλλον;

- Υβριδικό -Hybrid Cloud
- Ιδιωτικό - Private Cloud
- Δημόσιο - Public Cloud
- Δεν γνωρίζω

Β2.6 Σε τι βαθμό θεωρείτε ότι πρέπει να σας παρέχονται τα παρακάτω από το πάροχο που συνεργάζεστε ή θα συνεργαστείτε;

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου: 5 = σε πολύ μεγάλο βαθμό, 4 = σε μεγάλο βαθμό, 3 = σε μέτριο βαθμό, 2 = σε μικρό βαθμό και 1 = καθόλου (βάλτε σε κύκλο τον κατάλληλο αριθμό)

- Β2.6a Απεριόριστος χώρος 1 2 3 4 5
- Β2.6b Ανάκτηση Αρχείων 1 2 3 4 5
- Β2.6c Υποστήριξη mail -τηλεφώνου 1 2 3 4 5
- Β2.6d Εκπαίδευση Χρηστών 1 2 3 4 5
- Β2.6e Προηγμένα συνεργατικά εργαλεία(π.χ. για προγραμματισμό έργων, οικονομική διαχείριση) 1 3 4 5
- Β2.6f Συνδέσεις που προστατεύονται με κωδικούς 1 2 3 4 5

Β2.7 Υπάρχουν άλλες απαιτήσεις που νομίζετε ότι χρειάζονται για να καλύψουν τον δήμο σας;

Β2.8 Αν χρησιμοποιείτε το Cloud Storage ή σκοπεύετε να το χρησιμοποιήσετε στο μέλλον, τι είδους δεδομένα θα αποθηκεύατε;

Β2.9 Θεωρείτε ότι μπορεί να ωφεληθεί η τοπική αυτοδιοίκηση από την χρήση του Cloud Storage;

- Ναι -Όχι - Δεν γνωρίζω

Β2.10 Σε ποιο βαθμό συμφωνείτε ή διαφωνείτε ότι η χρήση του cloud storage προσφέρει τα παρακάτω οφέλη ;

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου 1= διαφωνώ έντονα, 2 = διαφωνώ, 3 = ουδέτερος, 4 = συμφωνώ, 5 = συμφωνώ έντονα να απαντηθεί ανεξάρτητα της χρήσης ή όχι cloud storage. (βάλτε σε κύκλο τον κατάλληλο αριθμό)

- Β2.10a Μείωση λειτουργικού κόστους 1 2 3 4 5
- Β2.10b Αύξηση Παραγωγικότητας(ευκολία στην πρόσβαση, διαμοιρασμός δεδομένων)
  1 2 3 4 5
- Β2.10c Καλύτερη εξυπηρέτηση δημοτών 1 2 3 4 5
- Β2.10d Καλύτερη συνεργασία με άλλους δήμους ή φορείς
  1 2 3 4 5

Β2.11 Που αλλού θεωρείτε ότι μπορεί να βοηθήσει η χρήση Cloud Storage στην βελτίωση του δήμου;

Β2.12 Πιστεύετε ότι χρειάζεται εκπαίδευση προσωπικού για την χρήση του cloud storage;

- Ναι -Όχι - Δεν γνωρίζω

Β2.13 Αν ναι, αναφέρετε ορισμένους τομείς που χρειάζεται στήριξη. (π.χ. τεχνικά ζητήματα)

Β2.14 Σε ποιο βαθμό συμφωνείτε ή διαφωνείτε ότι η χρήση του cloud storage παρουσιάζει τους παρακάτω κινδύνους;

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου: 1= διαφωνώ έντονα, 2 = διαφωνώ, 3 = ουδέτερος, 4 = συμφωνώ, 5 = συμφωνώ έντονα– να απαντηθεί ανεξάρτητα της χρήσης ή όχι cloud storage

- Β2.14a Πρόσβασης στα δεδομένα από μη εξουσιοδοτημένα/αρμόδια άτομα
  1 2 3 4 5
- Β2.14b Απώλειας/καταστροφής ή τροποποίησης (μη επιθυμητής) δεδομένων
  1 2 3 4 5
- Β2.14c Μη διαθεσιμότητας των δεδομένων κάποιες χρονικές στιγμές
  1 2 3 4 5
- Β2.14d Χαμηλής ταχύτητας πρόσβασης στα δεδομένα
  1 2 3 4 5

Β2.15 Σε ποιο βαθμό συμφωνείτε ή διαφωνείτε ότι η χρήση του Cloud Storage από ένα δήμο για την αποθήκευση δεδομένων του

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου: 1= διαφωνώ έντονα, 2 = διαφωνώ, 3 = ουδέτερος, 4 = συμφωνώ, 5 = συμφωνώ έντονα) – να απαντηθεί ανεξάρτητα της χρήσης ή όχι cloud storage (βάλτε σε κύκλο τον κατάλληλο αριθμό)

- Β2.15a Είναι δύσκολη και πολύπλοκη 1 2 3 4 5
- Β2.15b Απαιτεί σημαντική προσπάθεια 1 2 3 4 5
- Β2.15c Είναι συμβατή με τις διαδικασίες και τον τρόπο λειτουργίας ενός δήμου

  1 2 3 4 5
- Β2.15d Είναι συμβατή με τις ανάγκες και την νοοτροπία ενός δήμου

  1 2 3 4 5

Β2.16 Πως αξιολογείτε την κρισιμότητα των παρακάτω για τον δήμο σας;

Απαντήστε στο ερώτημα με βάση την κλίμακα 1 έως 5, όπου:1=καθόλου κρίσιμο, 2= σε μικρό βαθμό, 3=μέτριο,4= αρκετά κρίσιμο,5= πάρα πολύ κρίσιμο (βάλτε σε κύκλο τον κατάλληλο αριθμό)

- Β2.16a Εμπιστευτικότητα δεδομένων (αποφυγή μη εξουσιοδοτημένης πρόσβασης στα δεδομένα μας) 1 2 3 4 5
- Β2.16b Διασφάλιση συνεχούς πρόσβασης στα δεδομένα μας

  1 2 3 4 5
- Β2.16c Ταχύτητα πρόσβασης στα δεδομένα 1 2 3 4 5

Β2.17 Αποτελούν τα παραπάνω λόγο ώστε να μην κάνατε χρήση του cloud storage;

- Ναι -Όχι

Β2.18 Αναφέρετε άλλον λόγο για τον οποίο θα αποφεύγατε την χρήση του Cloud Storage.

"You don't really understand something unless you can explain it to your grandmother"

Albert Einstein