



Ανώτατο Εκπαιδευτικό Ίδρυμα Ιονίου

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία:

**ΣΥΣΤΗΜΑΤΑ ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΗΝ
ΠΡΟΒΛΕΨΗ ΤΡΑΠΕΖΙΚΟΥ MARKETING**

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: κος ΕΞΑΡΧΟΣ ΘΕΜΗΣ

ΦΟΙΤΗΤΗΣ: ΖΟΥΡΤΟΥΜΙΔΗΣ ΜΑΡΙΟΣ

ΜΗΤΡΩΟ: Π2015064

ΚΕΡΚΥΡΑ, ΜΑΪΟΣ 2020

Πρόλογος

Η παρούσα Πτυχιακή εργασία με τίτλο “Συστήματα Υποστήριξης Απόφασης για την πρόβλεψη τραπεζικού marketing” εκπονήθηκε στα πλαίσια της ολοκλήρωσης των προϋποθέσεων για τη λήψη του πτυχίου μου από το Α.Ε.Ι Ιονίου τμήματος Πληροφορικής, με έδρα το νησί της Κέρκυρας. Η ανάληψη της ορίστηκε τον Νοέμβριο του 2019, με υπεύθυνο καθηγητή τον κ. Θέμη Έξαρχο. Η ολοκλήρωσή της πραγματοποιήθηκε εντός των προβλεπόμενων, από το Α.Ε.Ι Ιονίου, χρονικών ορίων, τον Μάιο του 2020.

Σκοπός μου κατά την διάρκεια της συγγραφής, δεν ήταν μόνο η ορθή και όσο το δυνατόν πληρέστερη ανάλυση του θέματος. Έγινε προσπάθεια, έτσι ώστε το περιεχόμενο της εργασίας να είναι κατανοητό και σαφές, γι’ αυτό και η ανάλυση του θέματος έγινε με χρήση πληθώρας διαγραμματικών αναπαραστάσεων, παραδειγμάτων, γραφημάτων και συγκεντρωτικών πινάκων. Ελπίζω το περιεχόμενό του να καλύπτει, όχι μόνο το εξεταζόμενο θέμα, αλλά να ανταποκρίνεται και στις απαιτήσεις των καθηγητών μου.

Για την όσο καλύτερη διατύπωση και συγγραφή της εργασίας χρησιμοποιήθηκαν διάφορα παραρτήματα και πηγές χάρις των οποίων μπορώ αυτή τη στιγμή να σας παρουσιάσω την εργασία-έρευνα μου.

Ευχαριστίες

Αρχικά, ευχαριστώ τον Επιβλέπων καθηγητή μου κ. Θέμη Έξαρχο, γιατί μου έδωσε την ευκαιρία να ασχοληθώ με ένα ιδιαίτερα ενδιαφέρον θέμα, για την αμέριστη και ουσιαστική επιστημονική βοήθεια και καθοδήγηση που μου παρείχε. Τον ευχαριστώ θερμά για τις εξαιρετικά ωφέλιμες κριτικές παρατηρήσεις του στην επεξεργασία του θέματος αυτού.

Επιπλέον, ευχαριστώ όλους τους καθηγητές και τις καθηγήτριες που με δίδαξαν στα μαθήματα της σχολής, αφού μου έδωσαν τα κατάλληλα κίνητρα και τις απαραίτητες γνώσεις για να φτάσω σε αυτό το στάδιο παρουσίας της πτυχιακής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους φίλους μου και τους συμφοιτητές μου για τη συμπαράσταση τους, αφού όλοι βρισκόμαστε στην ίδια κατάσταση αλλά και για τη συνεχή συνεργασία και αλληλοβοήθεια που υπήρχε μεταξύ μας.

Περισσότερο απ' όλους, οφείλω να ευχαριστήσω την οικογένεια μου, διότι χωρίς εκείνους η απόκτηση ενός πτυχίου θα ήταν αδύνατη ή έστω ένα αρκετά δύκολο εγχείρημα. Τους ευχαριστώ ιδιαίτερα που στάθηκαν δίπλα μου, όλα αυτά τα χρόνια των σπουδών μου, και για την υπομονή που υπέδειξαν, μέχρι την επιστροφή μου στην οικογενειακή εστία.

Τέλος, θα ήθελα να απευθύνω τις ευχαριστίες μου στα μέλη της εξεταστικής επιτροπής, οι οποίοι ευγενικά και χωρίς καμία αντίρρηση δέχθηκαν να αξιολογήσουν την παρούσα πτυχιακή εργασία.

Περιεχόμενα

Πρόλογος.....	3
Ευχαριστίες.....	4
<i>Κεφάλαιο 1^ο:</i>	6
1.1 Εισαγωγή.....	6
1.2 Εννοιολογικοί Προσδιορισμοί.....	7
1.3 Τι είναι το Bank Marketing.....	8
<i>Κεφάλαιο 2^ο:</i>	9
2.1 Περίληψη Εργασιών.....	9
2.2.1 Πίνακες Αποτελεσμάτων Εργασιών.....	21
2.2.2 Γραφήματα Εργασιών.....	25
<i>Κεφάλαιο 3^ο:</i>	26
3.1 Περιγραφή Αλγορίθμων.....	26
3.2 Περιγραφή Συνόλου Δεδομένων.....	30
3.3 Περιγραφή Αποτελεσμάτων.....	31
<i>Κεφάλαιο 4^ο:</i>	35
4.1 Επιλογή Βέλτιστου Μοντέλου.....	35
4.2 Κανόνες Μοντέλου με χρήση Αλγορίθμου PART.....	35
<i>Κεφάλαιο 5^ο:</i>	40
Πίνακας Εργασιών.....	40
<i>Κεφάλαιο 6^ο:</i>	43
6.1 Πειράματα για Προτεινόμενο Μοντέλο.....	43
6.2 Συγκρίσεις Εργασιών.....	43
<i>Κεφάλαιο 7^ο:</i>	47
7.1 Συμπεράσματα.....	47
7.2 Αναφορές άρθρων - Βιβλιογραφία.....	49
7.3 Πηγές – Εργαλεία.....	51

Κεφάλαιο 1^ο:

1.1 Εισαγωγή

Το θέμα της πτυχιακής εργασίας είναι τα “**Συστήματα Υποστήριξης Απόφασης για την πρόβλεψη του τραπεζικού marketing**”. Η πτυχιακή εργασία εστιάζει στην εύρεση του βέλτιστου μοντέλου για την αποτελεσματικότητα του τραπεζικού marketing. Η εύρεση του μοντέλου αυτού γίνεται μέσω της μελέτης και επεξεργασίας του συνόλου δεδομένων του UCI bank telemarketing data set (με 45211 στοιχεία και 17 κατηγορίες). Για την επεξεργασία των παραπάνω στοιχείων χρησιμοποιήθηκε το εργαλείο εξόρυξης δεδομένων Weka. Όσον αφορά την αξιολόγηση των δεδομένων αυτή επιτεύχθηκε με διάφορους αλγορίθμους όπως (C4.5, Naive Bayes, Random Forest, Logistic Regression). Η ανάλυση των δεδομένων έγινε σε δύο κατηγορίες: imbalanced δεδομένα και balanced δεδομένα. Η πρώτη απόπειρα ανάλυσης των δεδομένων έγινε στα imbalanced δεδομένα και η δεύτερη στα balanced. Το συγκεκριμένο σύνολο δεδομένων χαρακτηρίζεται από τη φύση του ως imbalanced. Η συγκεκριμένη ανάλυση δεν μας έδωσε και τα καλύτερα αποτελέσματα, γι’ αυτό το λόγο έπρεπε τα δεδομένα αυτά να μετατραπούν σε balanced. Αυτό είχε ως συνέπεια την εξασφάλιση καλύτερων αποτελεσμάτων. Κλείνοντας χρησιμοποιήθηκε και feature selection για την καλύτερη δυνατή επιλογή χαρακτηριστικών.

Πέραν όμως από την επεξεργασία των δεδομένων για την εύρεση του καλύτερου δυνατού μοντέλου στην έρευνα μας ασχοληθήκαμε και με την ανάλυση εργασιών παρόμοιου θέματος προκειμένου να μπορέσουμε να μελετήσουμε και άλλους τρόπους επεξεργασίας δεδομένων σε ίδια ή παρόμοια σύνολα δεδομένων, αλλά και να δούμε και άλλα αποτελέσματα και να τα συγκρίνουμε με τα δικά μας.

Η δομή της εργασίας έχει ως εξής:

- **Στο Πρώτο Κεφάλαιο** γίνεται η εισαγωγή του αναγνώστη στον τρόπο με τον οποίο συντάχθηκε η εργασία αλλά και στο τι θα ακολουθήσει στο κυρίως κομμάτι. Επίσης, γίνεται και μία μικρή αναφορά σε διάφορες έννοιες και ορισμούς που θα συναντήσει κανείς στο κείμενο καθώς και το τι πραγματεύεται ο όρος Bank Marketing.
- **Στο Δεύτερο Κεφάλαιο** γίνεται η ανάλυση-περίληψη των εργασιών-άρθρων που επιλέχθηκαν ως κατάλληλες για το θέμα που εξετάζουμε, στη συνέχεια παραθέτονται οι πίνακες αποτελεσμάτων για κάθε εργασία ξεχωριστά και τέλος τα γραφήματα για τις παραπάνω εργασίες.
- **Στο Τρίτο Κεφάλαιο** γίνεται η περιγραφή του συνόλου δεδομένων, των αλγορίθμων που χρησιμοποιήθηκαν για τη δημιουργία του βέλτιστου μοντέλου μας, καθώς και η αναλυτική παρουσίαση των αποτελεσμάτων μέσω πινάκων.
- **Στο Τέταρτο Κεφάλαιο** γίνεται η επιλογή και ο σχολιασμός του βέλτιστου μοντέλου με βάση τα αποτελέσματα που αντλήθηκαν από τους παραπάνω αλγόριθμους που χρησιμοποιήθηκαν καθώς και η παρουσίαση των κανόνων του προτεινόμενου μοντέλου μέσω της χρήσης του αλγορίθμου PART.
- **Στο Πέμπτο Κεφάλαιο** συναντάμε το πινακάκι των εργασιών, στο οποίο παραθέτονται όλα τα σημαντικά στοιχεία των εργασιών που συγκρίθηκαν με τη δικιά μας, συμπεριλαμβανομένου και της εργασίας μας.
- **Στο Έκτο Κεφάλαιο** συναντάμε την ανάλυση της μεθοδολογίας που εφαρμόστηκε στην εργασία μας αλλά και αναλυτικά τις συγκρίσεις των εργασιών-άρθρων, που επιλέχθηκαν, με την εργασία μας.
- Η ολοκλήρωση της εργασίας έρχεται με το **Έβδομο Κεφάλαιο** στο οποίο περιλαμβάνονται τα συμπεράσματα της συνολικής εργασίας, οι αναφορές (δηλαδή δομή: συγγραφέας / τίτλος / χρονολογία δημοσίευσης / τεύχος / σελίδα που δημοσιεύτηκε) όλων των άρθρων που επιλέχθηκαν, αλλά και η βιβλιογραφία-πηγές-εργαλεία που χρησιμοποιήθηκαν για διεκπεραίωση της εργασίας.

1.2 Εννοιολογικοί Προσδιορισμοί

- ✓ **Feature Selection:** είναι η διαδικασία κατά την οποία το εργαλείο που χρησιμοποιούμε (Weka) μέσω των αλγορίθμων που εφαρμόζονται.
- ✓ **Data Mining:** είναι η εξεύρεση μιας πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.
- ✓ **Machine learning:** ή αλλιώς μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.
- ✓ **Churn Prediction:** είναι μία από τις πιο δημοφιλείς περιπτώσεις χρήσης Big Data στις επιχειρήσεις-τράπεζες. Στην ουσία είναι ο εντοπισμός πελατών που είναι πιθανόν να ακυρώσουν μια συνδρομή σε μια υπηρεσία.
- ✓ **Classifier – Ταξινομητής:** ένας ταξινομητής είναι ένα σύστημα στο οποίο εισάγετε δεδομένα και στη συνέχεια λαμβάνετε εξόδους που σχετίζονται με την ομαδοποίηση (δηλαδή ταξινόμηση) στην οποία ανήκουν αυτές οι εισοδοί.
- ✓ **10 fold cross validation:** είναι μια τεχνική για την αξιολόγηση προβλεπόμενων(predictive) μοντέλων με διαίρεση του αρχικού δείγματος σε ένα εκπαιδευτικό set για την εκπαίδευση του μοντέλου και ένα test set για την αξιολόγησή του.
- ✓ **Instances – Attributes:** instances είναι το σύνολο των στοιχείων του συνόλου δεδομένων και attributes ή features είναι οι κατηγορίες που χωρίζεται το κάθε data set.
- ✓ **Training set – Test set:** το training set ή σεντ εκπαίδευσης χρησιμοποιείται για την κατασκευή του μοντέλου. Συνήθως χρησιμοποιείται ένα σύνολο δεδομένων ή ένα σύνολο δοκιμών (test set) για την αξιολόγηση του πόσο καλά κάνει το μοντέλο με δεδομένα εκτός του εκπαιδευτικού συνόλου (training set).
- ✓ **Τιμές Precision – Recall – F-measure:** Precision (ακρίβεια) ποσοτικοποιεί τον αριθμό των θετικών κλάσεων που ανήκουν στην θετική κλάση. Η Recall ποσοτικοποιεί τον αριθμό των θετικών προβλέψεων κλάσης που έγιναν από όλα τα θετικά παραδείγματα στο σύνολο δεδομένων και F-measure παρέχει ένα μόνο σκορ που εξισορροπεί τόσο τις ανησυχίες του Precision όσο και του Recall σε έναν μόνο αριθμό.
- ✓ **Confusion Matrix:** ή αλλιώς πίνακας σύγχυσης είναι ένας πίνακας που χρησιμοποιείται συχνά για να περιγράψει την απόδοση ενός μοντέλου ταξινόμησης σε ένα σύνολο δεδομένων δοκιμής (test set) για τα οποία είναι γνωστές οι πραγματικές τιμές. Επιτρέπει την οπτικοποίηση της απόδοσης ενός αλγορίθμου. Τα περισσότερα μέτρα απόδοσης υπολογίζονται από τον πίνακα σύγχυσης.

1.3 Τι είναι το Bank Marketing

Ο τελικός στόχος μιας τράπεζας είναι να προσφέρει την πλήρη ικανοποίηση στον πελάτη. Η ικανοποίηση του πελάτη επηρεάζεται από την απόδοση του συνόλου όλων των τομέων της τράπεζας. Όλες οι τεχνικές και οι στρατηγικές του marketing έτσι ώστε τελικά να ωθούν τους ανθρώπους να δουλεύουν με μια συγκεκριμένη τράπεζα. Το marketing είναι μια οργανωτική φιλοσοφία. Αυτή η φιλοσοφία απαιτεί την ικανοποίηση των αναγκών των πελατών ως προϋπόθεση για την ύπαρξη και την επιβίωση της τράπεζας. Το πρώτο και σημαντικότερο βήμα στην εφαρμογή της ιδέας marketing είναι να έχουμε έναν ολοκληρωμένο προσανατολισμό του πελάτη από όλους τους εργαζομένους. Το marketing είναι ένα 'attitude of mind'. Αυτό σημαίνει ότι ο κεντρικός στόχος όλων των δραστηριοτήτων μιας τράπεζας είναι ο πελάτης. Το marketing δεν αποτελεί ξεχωριστή λειτουργία για τις τράπεζες. Η λειτουργία marketing απαιτείται να ενσωματωθεί στην λειτουργία. Το marketing είναι κάτι περισσότερο από απλή διαφήμιση και προώθηση, αποτελεί βασικό μέρος της συνολικής επιχειρηματικής λειτουργίας. Αυτό που απαιτείται για την τράπεζα είναι ο προσανατολισμός της αγοράς και η συνείδηση του πελάτη. Για την ανάπτυξη της φιλοσοφίας και της κουλτούρας marketing, μια τράπεζα μπορεί να απαιτήσει έναν συντονιστή marketing ή έναν ολοκληρωτή στο κεντρικό γραφείο που να αναφέρει απευθείας στον Διευθύνοντα Σύμβουλο για τον αποτελεσματικό συντονισμό των διαφόρων λειτουργιών, όπως έρευνα, εκπαίδευση, δημόσιες σχέσεις, για να εξασφαλιστεί η ικανοποίηση των πελατών. Ο εκτελεστικός διευθυντής είναι το καταλληλότερο πρόσωπο για να κάνει αυτό το έργο του συντονισμού. Κλείνοντας, η συνολική λειτουργία marketing περιλαμβάνει τα εξής:

- **Έρευνα αγοράς**, δηλαδή προσδιορισμό των οικονομικών αναγκών του πελάτη και επιθυμία και πρόβλεψη και έρευνα των μελλοντικών αναγκών της χρηματοπιστωτικής αγοράς και των δραστηριοτήτων των ανταγωνιστών.
- **Ανάπτυξη προϊόντων**, δηλαδή κατάλληλα προϊόντα για την κάλυψη των οικονομικών αναγκών των καταναλωτών.
- **Τιμολόγηση της υπηρεσίας**, δηλαδή διαφημιστικές δραστηριότητες και σύστημα διανομής σύμφωνα με τις κατευθυντήριες γραμμές και τους κανόνες της Τράπεζας Αποθεμάτων και ταυτόχρονα αναζητώντας ευκαιρίες καλύτερης ικανοποίησης των πελατών.
- Τέλος, **Ανάπτυξη της αγοράς** δηλαδή κουλτούρα marketing. Έτσι είναι σημαντικό να αναγνωρίσουμε τις θεμελιώδεις διαφορετικές λειτουργίες που πρέπει να εκτελέσει το τραπεζικό marketing. Δεδομένου ότι οι τράπεζες πρέπει να προσελκύσουν καταθέσεις και κατ' επέκταση χρήστες χρημάτων και άλλων υπηρεσιών, τα προβλήματα marketing είναι πιο περίπλοκα στις τράπεζες απ' ό,τι σε άλλους 'εμπορικούς οίκους'.

Κεφάλαιο 2^ο:

2.1 Περίληψη Εργασιών

Οι συγγραφείς στην εργασία [1] εστίασαν την προσοχή τους σε τέσσερις βασικές τεχνολογίες νέας γενιάς:

- το Internet of things
- Artificial Intelligence (τεχνητή νοημοσύνη)
- Machine Learning (μηχανική μάθηση)
- BlockChain, και οι αντίστοιχοι ρόλοι τους στο marketing

Συγκεκριμένα, η μελέτη εξετάζει σε βάθος καθεμιά από αυτές τις τέσσερις τεχνολογίες, κατανοώντας τα βασικά στοιχεία τους, τους τομείς στους οποίους λειτουργούν και τρέχουσες περιπτώσεις χρήσης που σχετίζονται με αυτές τις τεχνολογίες. Η υιοθέτηση αυτών των νέων τεχνολογιών αναμένεται να έχει αντίκτυπο στα αποτελέσματα για τις επιχειρήσεις και τους πελάτες. Ωστόσο, πρόκειται για δυναμικά εξελισσόμενες τεχνολογίες που δεν έχουν εξερευνηθεί πλήρως και των οποίων το πλήρες δυναμικό δεν έχει ακόμα αποκαλυφθεί. Η μελέτη αυτή επισημαίνει επίσης σημαντικούς τομείς που χρειάζονται managerial focus στην υιοθέτηση αυτών των νέων τεχνολογιών.

Αρχικά, θα μιλήσουμε λίγο για το τι ακριβώς πραγματεύονται αυτές οι τέσσερις τεχνολογίες:

- **Internet of Things (IoT)**, είναι μια παγκόσμια υποδομή για την κοινωνία της πληροφορίας, η οποία επιτρέπει προηγμένες υπηρεσίες μέσω διασύνδεσης αντικειμένων που βασίζονται σε υπάρχουσες και εξελισσόμενες διαλειτουργικές τεχνολογίες πληροφοριών και επικοινωνιών.

Κύρια στοιχεία τους είναι:

- **Artificial Intelligence**, η επιστήμη των training machines για την εκτέλεση ανθρωπίνων εργασιών, από την επεξεργασία μεγάλων ποσοτήτων δεδομένων και η αναγνώριση προτύπων στα δεδομένα μέσω της χρήσης μιας σειράς τεχνολογιών όπως machine learning, η επεξεργασία της φυσικής γλώσσας κλπ. Το AI δίνει τη δυνατότητα στις μηχανές να εκτελούν ανθρώπινα καθήκοντα.
- **Machine Learning**, μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την ανάπτυξη αναλυτικού μοντέλου. Είναι ένας κλάδος της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι τα συστήματα μπορούν να μάθουν από τα δεδομένα, και να λάβουν αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση.
- **Blockchain**, είναι ένα κατανεμημένο ημερολόγιο και μια αμετάβλητη βάση δεδομένων για την ασφαλή μεταφορά δεδομένων. Το όνομα είναι ένας συνδυασμός δύο λέξεων (το block) που περιέχει δοσοληψίες και μια (chain) που αντιπροσωπεύει κρυπτογραφικά συνδεδεμένα blocks.

Η υιοθέτηση των τεχνολογιών νέας γενιάς από τη μεριά των επιχειρήσεων αυξάνει σε μεγάλο βαθμό την ποιότητα του marketing τους. Ενδεικτικά πολλές εταιρείες χρησιμοποιούν στοιχεία AI και μεγάλα δεδομένα για τον εντοπισμό παραπονημένων προϊόντων, τη βελτίωση των πωλήσεων και την ανάπτυξη και ενίσχυση προσωπικών σχέσεων με τους πελάτες. Επιπλέον, αξιοποιώντας τη δύναμη των γνώσεων από τις τεχνικές εξόρυξης δεδομένων, οι επιχειρήσεις μπορούν να προσδιορίσουν τις επικείμενες τάσεις, προκειμένου να δημιουργήσουν νέα προϊόντα που θα καλύπτουν τις ανάγκες των καταναλωτών. Επίσης, μελλοντικά μέσω αυτών των τεχνολογιών αυτών οι επιχειρήσεις θα μπορούν να έρχονται σε επαφή με τους πελάτες τους in real time για να τους ενημερώσουν για τα προϊόντα τους χωρίς την παρέμβαση των media. Ακόμα μία σημαντική επιρροή είναι ότι θα μπορούν οι επιχειρήσεις μελλοντικά να αυξήσουν την παραγωγική τους δραστηριότητα καθώς και την ποιότητα των παραγόμενων προϊόντων μειώνοντας ταυτόχρονα και τα λειτουργικά τους κόστη. Για όλους αυτούς τους λόγους θα υπάρχει σταδιακή αύξηση του marketing των επιχειρήσεων.

Στην εργασία [2] γίνεται αναφορά στην ανάλυση μεγάλων δεδομένων (big data analysis) από τις επιχειρήσεις. Η ανάλυση αυτή επιτυγχάνεται με τεχνικές machine learning. Το ενδιαφέρον της ανάλυσης αυτής έγκειται στον τρόπο με τον οποίο οι επιχειρήσεις αυτές, μέσω της επεξεργασίας των δεδομένων, προβλέπουν ποιοι πελάτες μπορούν να χαθούν σε βάθος χρόνου. Σκοπός της πρόβλεψης αυτής είναι να δώσει στην εταιρεία την ευκαιρία να προσεγγίσει αυτούς τους πελάτες ως μια προσπάθεια να διασφαλίσει τη διατήρησή τους και κατ' επέκταση να χρησιμοποιήσει τις δυνάμεις marketing που διαθέτει. Εκπαιδεύονται classifiers (ταξινομητές) με χρονολογική σειρά που εκτιμούν το απόθεμα του πελάτη και ταξινομούν τους πελάτες ως χαμένους ή

διατηρημένους σε περιόδους αδράνειας. Εξετάζονται διαφορετικά μήκη των χρονοσειρών που χρησιμοποιούνται ως είσοδοι στον αλγόριθμο machine learning, διαφορετικά μήκη για την περίοδο αδράνειας που χαρακτηρίζουν τον πελάτη ως χαμένο και διαφορετικούς ορίζοντες πρόβλεψης, δηλαδή πόσο νωρίτερα εντοπίζεται η απώλεια του πελάτη. Η δομή του άρθρου έχει ως εξής:

- Αρχικά, γίνονται οι υπολογισμοί για την προετοιμασία των δεδομένων εισόδου για τους classifiers.
- Στη συνέχεια, παρουσιάζονται τα πειράματα και τα αποτελέσματα.
- Τέλος, συζητούνται οι συνέπειες των ευρημάτων.

Αφού γίνει η προετοιμασία των δεδομένων, σειρά έχουν τα πειράματα ή αλλιώς τα σενάρια που δημιουργούνται με βάση τα αποτελέσματα. Για το σενάριο όπου η διατήρηση πελατών πρέπει να προβλεφθεί για μικρά ή μεσαία σύνολα δεδομένων, τα ευρήματα υποδεικνύουν ότι ο αλγόριθμος Random Forest είναι η καλύτερη επιλογή (πράγμα που συμβαίνει με τα δεδομένα μας), ενώ για προβλέψεις σχεδόν σε πραγματικό χρόνο για μεγάλα σύνολα δεδομένων η SVM μπορεί να αποδεχθεί επαρκής και πολύ ταχύτερη από τη Random Forest. Τώρα όσον αφορά το τμήμα marketing έχει αρκετές ευκαιρίες να προσεγγίσει τους πελάτες που προβλέπεται να χαθούν με διάφορες προσφορές. Μέσω του αλγορίθμου machine learning μειώνεται σημαντικά ο αριθμός των πελατών που πρέπει να στοχευθούν από το τμήμα marketing, το οποίο πρέπει να έρθει σε επαφή με αυτούς, εξοικονομώντας έτσι χρόνο και χρήματα στην εταιρεία. Ο πίνακας με τα αποτελέσματα που προκύπτουν από την εργασία παραθέτουμε στον [πίνακα 1](#).

Στην επόμενη εργασία [3] περιγράφεται ένα νέο σύστημα προσέγγισης των πελατών με εκστρατείες άμεσου marketing που βασίζονται σε χαρακτηριστικά που επιλέγονται από μια μεγάλη βάση δεδομένων με συναλλαγές από μία από τις μεγαλύτερες τράπεζες λιανικής στην Πολωνία. Κατά την προσέγγιση, τα δεδομένα κάθε πελάτη αντιστοιχίζονται σε πολυδιάστατα χρονοδιαγράμματα. Υπάρχουν συγκεκριμένοι χρονικοί περίοδοι που περιέχουν πληροφορίες που περιγράφουν τη συμπεριφορά του πελάτη σε κάποιο συγκεκριμένο χρονικό παράθυρο. Η αναπράσταση αυτή μας παρέχει ένα training set μέσω του οποίου μπορούμε να ταξινομήσουμε κάθε παρατήρηση σε δύο κατηγορίες, στους πελάτες που ενδιαφέρονται και σε αυτούς που δεν ενδιαφέρονται για αγορά πιστωτικού προϊόντος. Τα set των δεδομένων που μελετούνται στο άρθρο έχουν να κάνουν με:

- Account balance history,
- Money transfers history,
- Credit and debit card transactions history

Οι πελάτες χωρίζονται σε δυο sets:

- Σε αυτούς που κλήθηκαν (A),
- Και σε αυτούς που δεν κλήθηκαν (B)

Με βάση το set (B) λειτουργεί και ο αλγόριθμος, προκειμένου να γίνει και η αξιολόγηση των μετρήσεων του campaign. Λόγω του ότι οι πελάτες κάνουν πολλές συναλλαγές ετησίως, ο όγκος των δεδομένων ήταν αρκετά μεγάλος και έτσι ήταν αρκετά δύσκολη η δημιουργία machine learning models. Για να διευκολύνουν το machine learning χρησιμοποιήθηκαν δύο δομές:

- Apache Spark framework,
- H2O framework

Για τη δημιουργία του συστήματος υψηλού επιπέδου για την πρόβλεψη της αγοράς πιστώσεων ([γράφημα 1](#)) χρησιμοποιήθηκαν 3 machine learning αλγόριθμοι: CART και Random Forest (classification trees) και ο DBN. Τέλος, λόγω του αυξημένου αριθμού features (Dimensionality problem) ήταν δύσκολη η εφαρμογή στατιστικών αλγορίθμων, αλλά και ο υπολογισμός της ακρίβειας (accuracy). Γι' αυτό το λόγο, ήταν απαραίτητη η εφαρμογή feature selection. Αυτή επιτεύχθηκε με την εφαρμογή του Boruta αλγορίθμου.

Στο πρώτο στάδιο, τα δεδομένα μετεγκαθίστανται από την αποθήκη δεδομένων (σχεσιακή βάση δεδομένων) στο cluster Hadoop και στη συνέχεια οι εργασίες του Apache Spark κάνουν εξαγωγή των feature και ταυτόχρονα κατασκευάζουν τόσο το training όσο και το test set. Στο τρίτο στάδιο, εκπαιδεύονται διάφορα μοντέλα (Random Forest, CART με τη χρήση του Apache Spark, του H2O και Deep Belief Networks) και βαθμολογούνται στο τελευταίο στάδιο με τον προσομοιωτή καμπάνιας.

Ο συντάκτης της εργασίας [4] παρουσιάζει την περίπτωση μιας βελτιστοποίησης χαρτοφυλακίου μέσης διακύμανσης (mean variance customer portfolio), στην οποία χρησιμοποιούνται priori γνώσεις σχετικά με τις ενέργειες για τη βελτίωση της απόδοσης για την εύρεση των εκτιμώμενων πολιτικών χρησιμοποιώντας μια διαδικασία reinforcement learning για το MDP (Markov decision process) με διακριτό χώρο δράσης. Το πλεονέκτημα marketing μιας εταιρείας έγκειται στην ικανότητα πρόβλεψης της μελλοντικής αγοραστικής συμπεριφοράς των καταναλωτών και όχι απλά να ανταποκριθεί στις ανάγκες των πελατών. Τα Markov models

χρησιμοποιούνται για την πρόβλεψη της συμπεριφοράς των πελατών. Προκειμένου να αυξηθεί η ταχύτητα εκμάθησης προτείνεται μια νέα επαναλαμβανόμενη αρχιτεκτονική μάθησης ενίσχυσης (novel recurrent reinforcement learning architecture) που προσαρμόζει τις πολιτικές της σύμφωνα με την αρχιτεκτονική προεπεξεργασίας (preprocessing architecture). Η προεπεξεργασία επωφελήθηκε από το γεγονός ότι είχε σταθερά ομοιόμορφη κατανομή κατά την εκμάθηση μιας νέας εργασίας από reinforcement (η ομοιόμορφη κατανομή παράγεται εύκολα). Στη συνέχεια, προτάθηκε μια actor-critic αρχιτεκτονική με στόχο την εκμάθηση του παράγοντα να αλληλεπιδράσει βέλτιστα με την εμπειρία του περιβάλλοντος. Ωστόσο, η διαδικασία RL χρησιμοποιεί μια βελτιστοποιημένη κατανομή των στρατηγικών που υποστηρίζονται από Iterated Quadratic/Lagrange programming method. Τρεις διαφορετικοί κανόνες μάθησης σχετικά με τις μεταβιβάσεις, τις υπηρεσίες κοινής οφείας και το κόστος προτάθηκαν για την εκτίμηση της λειτουργίας των τωρινών πολιτικών. Ο κανόνας μάθησης που σχετίζεται με το κόστος το οποίο ήταν ενδιαφέρον, διότι επιβάλλει περιορισμούς στη διατύπωση του προβλήματος: δεν μπορεί να υποτιμηθεί ή να υπερεκτιμηθεί εάν οι περιορισμοί υποτίθεται ότι πληρούνται. Με άλλα λόγια, στο άρθρο συναντάμε ένα προσομοιωμένο σενάριο μέσω του οποίου η τράπεζα προσφέρει χρυσές πιστωτικές κάρτες στους πελάτες της προκειμένου να μελετήσει μέσω των παραπάνω αλγορίθμων τα αποτελέσματα. Μέσω των δεδομένων που διαθέτει η τράπεζα χωρίζει τους πελάτες σε κατηγορίες:

- εάν έχει λογαριασμό στην τράπεζα και μπορεί να χρησιμοποιήσει την πιστωτική κάρτα,
- ένας πελάτης χαμηλής συχνότητας
- ένας πελάτης με συχνή χρήση
- ένας πελάτης υψηλής χρήσης της κάρτας

Τα marketing campaigns της τράπεζας προσφέρουν στους πελάτες τέσσερα διαφορετικά οφέλη:

- το σταθερό ποσοστό της επιστροφής μετρητών
- τα δώρα
- τις εκπτώσεις
- σημεία προς την αγορά με βάση το ποσό που δαπανά ο πελάτης (εκτός πεδίου αεροπορικών εταιρειών)

Τέλος, η τράπεζα έχει περιορισμένο προϋπολογισμό για τα συγκεκριμένα marketing campaigns.

Οι συγγραφείς στην εργασία [5] προβάλλουν αναλύσεις και εφαρμογές των πιο σημαντικών τεχνικών στην εξόρυξη δεδομένων, multilayer perception neural network (MLPNN), το δέντρο Naive Bayes (TAN) που είναι γνωστό ως Bayesian network, Nominal regression ή Logistic regression και Ross Quinlan νέο μοντέλο δέντρου απόφασης (C5.0). Ο στόχος είναι να εξεταστεί η απόδοση των τεχνικών MLPNN, TAN, LR και C5.0 σε δεδομένα πραγματικού κόσμου της καταγραφής τραπεζικών καταθέσεων. Ο σκοπός είναι να αυξηθεί η αποτελεσματικότητα της καμπάνιας εντοπίζοντας τα κύρια χαρακτηριστικά που επηρεάζουν μια επιτυχία (η κατάθεση που έχει πραγματοποιηθεί από τον πελάτη) βάσει των MLPNN, TAN, LR και C5.0. Τα πειραματικά αποτελέσματα δείχνουν, με υψηλότερη ακρίβεια, την επιτυχία αυτών των μοντέλων στην πρόβλεψη της καλύτερης επαφής της καμπάνιας με τους πελάτες για την καταγραφή κατάθεσης. Οι επιδόσεις υπολογίζονται με τη βοήθεια 3 στατιστικών μέτρων: accuracy, sensitivity και specificity.

Λίγα λόγια για την κάθε τεχνική:

- **Multilayer Perception Neural Network (MLPNN):** είναι γνωστό ως μια ισχυρή λειτουργία προσέγγισης για προβλήματα πρόβλεψης και ταξινόμησης. Η δομή MLPNN οργανώνεται σε επίπεδα εισόδου, εξόδου αλλά και σε κρυφά επίπεδα νευρώνων. Κατά την χρήση του MLPNN πρέπει να αντιμετωπιστούν τρία σημαντικά ζητήματα: η επιλογή των δειγμάτων των δεδομένων για εκπαίδευση στο δίκτυο, την επιλογή κατάλληλου και αποτελεσματικού αλγορίθμου εκπαίδευσης και τον προσδιορισμό του μεγέθους του δικτύου. Τέλος το MLPNN έχει πολλά πλεονεκτήματα όπως: καλή learning ικανότητα, λίγη απαίτηση μνήμης, γρήγορη λειτουργία σε πραγματικό χρόνο, απλή και βολική στη χρήση, κατάλληλη για την ανάλυση σύνθετων προτύπων κ.ά.
- **Tree Augmented Naive bayes (TAN):** είναι ένα εκτεταμένο δέντρο έτσι ώστε ένας κόμβος κλάσης να πηγαίνει απευθείας σε όλους τους κόμβους των attributes, επίσης ένας κόμβος attribute μπορεί να έχει μόνο έναν γονέα από έναν άλλο κόμβο attributes. Ωστόσο, στο TAN δεν υπάρχει περιορισμός στις συνδέσεις μεταξύ κόμβων attributes.
- **Logistic Regression (LR):** είναι κατάλληλο για την αντιμετώπιση ζητημάτων από διαφόρων ειδών data sets. Παρέχει αρκετά και καλά καταναμεμένα δείγματα. Επιπλέον, είναι κατάλληλο για την περιγραφή και τον έλεγχο υποθέσεων σχετικά με τις σχέσεις μεταξύ μιας κατηγοριατικής μεταβλητής outcome και

συνεχών χαρακτηριστικών πρόβλεψης. Το LR είναι ισχυρό, επειδή οι ανεξάρτητες μεταβλητές δεν χρειάζεται να διανέμονται συνήθως ή να έχουν ίση διακύμανση σε κάθε ομάδα.

- **C5.0 Decision Tree:** είναι ένας αλγόριθμος μοντελοποίησης και είναι μια βελτιωμένη έκδοση των αλγορίθμων C4.5 και ID3. Το δέντρο του C5.0 χρησιμοποιεί αλγορίθμους διαχωρισμού που περιλαμβάνει εντροπία με βάση την απόκτηση πληροφοριών. Το μοντέλο λειτουργεί χωρίζοντας το δείγμα με βάση το χαρακτηριστικό που παρέχει το μέγιστο κέρδος πληροφοριών.

Το data set που χρησιμοποιείται στην έρευνα είναι από το UCI repository και περιέχει 45211 instances και 17 attributes. Χρησιμοποιείται supervised learning της ανάλυσης δεδομένων για να φτάσει στην καλύτερη πρόβλεψη για το χαρακτηριστικό Y (output), το οποίο είναι ο στόχος. Το data set χωρίζεται σε δύο μέρη: σε αυτά με ranges (για αριθμητικά χαρακτηριστικά) και σε κλάσεις (για τις κατηγορίες και τα binary categorical attributes).

Όσον αφορά τα αποτελέσματα ο [πίνακας 2](#) δείχνει την αριθμητική απεικόνιση της σημασίας των χαρακτηριστικών σε σχέση με τα μοντέλα MLPNN, TAN, LR και C5.0.

Στον [πίνακα 3](#) το μοντέλο MLPNN έχει ακρίβεια 90.92% για training δείγματα και 90.49% για testing δείγματα, 89.16% και 88.75% για training και testing αντίστοιχα, σύμφωνα με το TAN. Η ακρίβεια ταξινόμησης του μοντέλου LR είναι 90.09% και 90.43% για training και testing δείγματα αντίστοιχα, επίσης το μοντέλο C5.0 έχει 93.23% και 90.09% αντίστοιχα.

Η υψηλότερη τιμή στα training δείγματα εμφανίζεται για το C5.0. Με αυτόν τον τρόπο συνεχίζεται και η ανάλυση για τα υπόλοιπα δεδομένα του πίνακα.

Ο υπολογισμός των accuracy, sensitivity και specificity γίνεται μέσω:

- $\text{Accuracy} = (TP + TN) / N$ (total number of cases)
- $\text{Sensitivity} = TP / (TP + FN)$
- $\text{Specificity} = TN / (TN + FP)$

Η εργασία [6] πραγματεύεται τα μέτρα για την πρόβλεψη κέρδους των Τούρκικων αγορών.

Παρουσιάζονται μοντέλα πρόβλεψης τα οποία βοηθάνε στην πρόβλεψη της απόδοσης ή της κερδοφορίας του τούρκικου IADB.

Στην έρευνα αυτή χρησιμοποιούνται 3 tree-based machine learning μέθοδοι:

- namely DStump
- RTree
- REPTreeare

Έπειτα, 4 ensemble μοντέλα μάθησης δημιουργούνται από την χρησιμοποίηση του Bagging ensemble learning method:

- Bag-DStump
- Bag-RTree
- Bag-MLP
- Bag-REPTree

Ο κύριος σκοπός του Bagging είναι να μειώσει το error που προέρχεται από τη διακύμανση του βασικού classifier, τα οποία είναι ασταθή. Για την κατασκευή του dataset το οποίο χρησιμοποιείται έχει ως output την απόδοση καθαρής θέσης (ROE) και ως inputs 10 βασικοί δείκτες ισολογισμού ως καθοριστικοί παράγοντες του κέρδους στις τράπεζες και διακυμαίνονται από 0 έως 1.

Οι tree-based machine learning μέθοδοι και τα Bagging ensemble μοντέλα που χρησιμοποιούνται στο άρθρο αξιολογούνται με τη χρήση στατιστικών δέντρων (R2, RAE, RRSE). Η αξιολόγηση αυτή διεξάγεται με k-fold cross validation. Η διαδικασία έχει ως εξής:

- Αρχικά, το σύνολο δεδομένων είναι τυχαίο.
- Στη συνέχεια, κατασκευάζονται 10 ξεχωριστά folds.
- Τέλος, κάθε fold χρησιμοποιείται για το test του accuracy των προτεινόμενων μοντέλων και τα υπόλοιπα 9 folds χρησιμοποιούνται για training.

Το μοντέλο είναι ένα σύνολο από βάρη και παραμέτρους που συνδέουν τα input και output στρώματα. Σε αυτό το άρθρο για τη δημιουργία του συνόλου (ensemble) προτείνεται μια instance-base ensemble learning προσέγγιση που χρησιμοποιεί σάκους (bagging). Για αυτή τη προσέγγιση απαραίτητα συστατικά μέρη είναι οι παραπάνω μέθοδοι (DStump, RTree και REPTree). Για να μετρηθεί το κέρδος απαραίτητη είναι η μεταβλητή ROE. Στην παρούσα μελέτη χρησιμοποιείται ένα dataset με 10 attributes για predictive modeling. Η bagging method χρησιμοποιείται για τη δημιουργία ensemble models. Τα bagging ensemble models είναι χρήσιμα

ώστε να εισαχθούν τεχνικές για πρόβλεψη bank profitability. Τέλος, τα ensemble models βοήθησαν στην ακρίβεια (accuracy) έτσι ώστε να αυξηθεί η χρησιμότητα της εφαρμογής για bank profitability.

Στο paper της εργασίας [7] αποδεικνύεται ότι με τη βοήθεια K-means clustering μπορεί να ορισθεί μια ομάδα πελατών που είναι πιο επιρρεπείς στην αγορά στεγαστικού δανείου. Γίνεται μια παρουσίαση της πραγματικής (real life) machine learning χρήσης, χρησιμοποιώντας πραγματικά real life δεδομένα.

Η διαδικασία ξεκινάει με το Data Collection. Προκειμένου να δημιουργηθεί ένα επιτυχημένο prediction model, πρέπει να δημιουργηθεί ένα καλό dataset. Το dataset που δημιουργείται περιέχει 400 derived variables. Για την επικύρωση των αποτελεσμάτων δημιουργείται ακόμα μία μεταβλητή, ονόματι flag. Το flag παίρνει τιμές 0 και 1. Όταν το flag είναι 1 σημαίνει ότι ο πελάτης αγοράζει στεγαστικό δάνειο μέσα στους επόμενους 3 μήνες και όταν είναι 0 το αντίθετο.

Επόμενο βήμα είναι η τελική επιλογή των μεταβλητών. Αν αυτό το βήμα παραληφθεί υπάρχει κίνδυνος οι μεταβλητές που θα περιέχονται να είναι άσχετες με τα clusters που θα δημιουργηθούν. Σύμφωνα με τις δοκιμές που έγιναν κατέληξαν στο $1-R^2$ ratio, όπου το R^2 είναι ένα στατιστικό μέτρο που χρησιμοποιείται για να εκτιμηθεί η καλή συμπεριφορά του μοντέλου. Οι μικρές τιμές αυτού του ratio υποδηλώνουν ότι η σχετική μεταβλητή έχει μια ισχυρή συσχέτιση με το δικό του cluster και μια αδύναμη συσχέτιση με τα άλλα clusters. Με άλλα λόγια προτιμάται η μεταβλητή που παράγει τη μικρότερη τιμή $1-R^2$ σε κάθε cluster. Στη διαδικασία ανάπτυξης του μοντέλου χρησιμοποιήθηκε το $1-R^2$ ratio ως μέθοδος επιλογής. Εδώ εμφανίζεται η συνάρτηση

της σχέσης που περιγράφηκε πιο πάνω: $1-R^2 = \frac{1 - R_{owner_cluster}^2}{1 - R_{next_closest_cluster}^2}$ (αριθμητής: owner_cluster και παρονομαστή:

next_closest_cluster).

Στη συνέχεια έχουμε το Data clusterization. Προκειμένου να εκπαιδευτεί (train) και να δοκιμαστεί (test) το μοντέλο επιτυχώς, δημιουργήθηκαν ξεχωριστά datasets, από το τελικό dataset. Το dataset χωρίστηκε σε δύο datasets, το (A) το οποίο περιέχει το 70% των παρατηρήσεων και το (B) που περιέχει το υπόλοιπο 30% το οποίο και χρησιμοποιήθηκε ως testing dataset.

Για τον καθορισμό του αριθμού των clusters, το οποίο είναι απαραίτητο για την κατασκευή του segmentation model, χρησιμοποιήθηκε το PST2 statistic. Όσο μικρότερη ή μεγαλύτερη είναι η τιμή του καθορίζεται αν θα πρέπει να συνδυαστούν ή όχι, αντίστοιχα, οι 2 clusters μεταξύ τους. Τέλος μέσω της μεθόδου του Ward δημιουργήθηκαν 8 centroids που αντιστοιχούν σε κάθε cluster, προκειμένου να εκτελεστεί ο διαχωρισμός (segmentation) τόσο του testing dataset όσο και του final dataset.

Τα αποτελέσματα που παρουσιάζονται στην έρευνα μπορούν να χρησιμοποιηθούν προκειμένου να αυξηθεί η αποτελεσματικότητα κατά την επικοινωνία με τον πελάτη καθώς και να μειωθούν τα κόστη στο τμήμα marketing.

Σύμφωνα με τον συντάκτη της εργασίας [8] οι μέθοδοι υποστήριξης vector machine που χρησιμοποιούν τον αλγόριθμο Adaboost εξετάζονται για να ταξινομήσουν ενδεχόμενους πελάτες για direct marketing μεγάλου όγκου δεδομένων marketing. Το direct (άμεσο) marketing είναι μια τεχνική marketing που εφαρμόζεται στην τράπεζα προκειμένου να ενημερωθεί ο πελάτης για κάποιο προϊόν ή υπηρεσία μέσω τηλεφώνου, ηλεκτρονικού ταχυδρομείου και άλλων. Λόγω του συνεχούς αυξανόμενου data των πελατών απαιτείται data mining προκειμένου να βγει μια απόφαση. Μία μέθοδος data mining είναι μια classification μέθοδος που είναι μια τεχνική ομαδοποίησης δεδομένων.

Η μέθοδος για την οποία γίνεται αναφορά στην έρευνα είναι η Support Vector Machine (SVM). SVM είναι η μέθοδος που αποτελεί τον καλύτερο διαχωριστή στα δεδομένα που χωρίζουν τις δύο κατηγορίες μεγιστοποιώντας την τιμή περιθωρίου των δύο κλάσεων. Σύμφωνα με το accuracy rate (1), η SVM δείχνει καλή επίδοση στην ταξινόμηση, αλλά είναι ευαίσθητη στις ρυθμίσεις δείγματος και παραμέτρων. Η επίδοση και η ακρίβεια του SVM μπορεί να βελτιωθεί με τη χρησιμοποίηση του Ensemble method. Ensemble method είναι η μέθοδος που συνδυάζει τις προβλέψεις που παράγονται από κάποιον ταξινομητή. Σε αυτή την έρευνα συγκρίνεται η επίδοση του απλού SVM με το SVM χρησιμοποιώντας Adaboost algorithm. Adaboost είναι μία αποτελεσματική μέθοδος που δημιουργεί έναν accurate classifier από την ένωση 2 weak classifiers. Ο adaboost εκπαιδεύει (train) τον βασικό classifier με συνεχείς επαναλήψεις. Εάν ο classifier εκπαιδευτεί όσο πρέπει, τότε ο classifier συνδυάζεται για να σχηματίσει μια τελική απόφαση για το μοντέλο που εμφανίζει την καλύτερη απόδοση. Το αποτέλεσμα της προτεινόμενης μεθόδου αξιολογείται βάσει του βαθμού ακρίβειας (accuracy) και ευαισθησίας (sensitivity). Η αξιολόγηση της μέτρησης λαμβάνεται από το confusion matrix.

Το dataset που χρησιμοποιείται είναι ένα data Bank marketing με 20 attributes και μια target class με 41188 παρατηρήσεις.

Τα αποτελέσματα παρουσιάζονται αναλυτικά στον [πίνακα 4](#).

Κλείνοντας, στο άρθρο αυτό προτείνεται η μέθοδος adaboost SVM για τη βελτίωση των αποτελεσμάτων πρόβλεψης του direct (άμεσου) marketing του τραπεζικού συνόλου δεδομένων σε σύγκριση με το απλό SVM

(ordinary SVM). Ο βαθμός ακρίβειας υπολογίζεται με τον εξής τύπο:
$$\text{Accurate rate} = \frac{TP + TN}{TN + FP + FN + TP}$$

Η εργασία [9] ξεκινάει με την τοποθέτηση ότι το direct marketing είναι χρήσιμο για την προσφορά νέων προϊόντων στους πελάτες. Το direct marketing data analysis των τραπεζών μπορεί να χρησιμοποιηθεί για την επιλογή του τύπου direct marketing των τραπεζών. Η πρόβλεψη μη ισορροπημένων δεδομένων (imbalance data) οδήγησε σε λιγότερη ακρίβεια. Γι' αυτό το λόγο προτάθηκε ο αλγόριθμος SMOTE. Συνεπώς το παρόν έγγραφο προτείνει μια μέθοδο βασισμένη στον αλγόριθμο SMOTE σε συνδυασμό με το Rotation Forest(PCA)-J48 για την πρόβλεψη των εξαιρετικά μη ισορροπημένων δεδομένων.

Αυτό που κάνει ο SMOTE αλγόριθμος είναι να εντοπίζει την κλάση με τα imbalance δεδομένα και να αυξάνει το αριθμό των στοιχείων της κλάσης με τη μειονότητα αυτών, κάνοντας έτσι oversampling ή up-sampling.

Τώρα, όσον αφορά το Rotation Forest (PCA)-J48:

- Το J48 decision tree είναι ένας αλγόριθμος για τη δημιουργία ενός αναδρομικού κλαδεμένου και μη δένδρου απόφασης με τη χρήση της εντροπίας πληροφοριών στο σύνολο του training data. Τα feature data χωρίζονται σε υποτομές και λαμβάνουν πληροφορίες μετρώντας τη διαφορά στην εντροπία που χρησιμοποιείται για τον προσδιορισμό του καλύτερου χαρακτηριστικού.
- Στο Rotation Forest το set των χαρακτηριστικών χωρίζεται σε υποσύνολα και στη συνέχεια εφαρμόζεται η κύρια ανάλυση συνιστώσεων PCA σε κάθε υποσύνολο.

Το data set απαρτίζεται από 45211 instances και 17 attributes.

Σε αυτό το paper, η απόδοση της ταξινόμησης μετριέται από τα: specificity, sensitivity και accuracy.

Κλείνοντας, σύμφωνα με την ensemble μέθοδο που προτάθηκε από την ένωση του SMOTE και Rotation Forest(PCA)-J48 αποδείχθηκε κατάλληλη για την ταξινόμηση των imbalance στοιχείων του direct marketing των τραπεζών.

Όσον αφορά τα αποτελέσματα, σε αυτό το άρθρο η απόδοση του προτεινόμενου μοντέλου αξιολογείται και συγκρίνεται με το δέντρο απόφασης, το Rotation Forest, το MLP neural network, το RBF neural network, το Bayesian network και το Naive Bayes όπως φαίνεται στον πίνακα 1. Τα πειράματα διεξάγονται σε 2 δεδομένα, τα αρχικά και αυτά που έχουν υποστεί επεξεργασία-βελτίωση από τον αλγόριθμο SMOTE. Αυτά τα δύο data sets εκπαιδεύονται(trained) και δοκιμάζονται(tested) σε 14 αλγορίθμους μηχανικής μάθησης. Τα συγκριτικά αποτελέσματα παρουσιάζονται στον [πίνακα 5](#). Από τα αποτελέσματα, ο αλγόριθμος που δίνει την υψηλότερη τιμή ευαισθησίας στα αρχικά δεδομένα είναι ο Rotation Forest-ADTree που μπορεί να επιτύχει 67.73% της ευαισθησίας στο αρχικό σύνολο δεδομένων. Ο αλγόριθμος που παράγει την υψηλότερη τιμή εξειδίκευσης(specificity) είναι ο Decision Tree J48 που μπορεί να επιτύχει το 93.19% specificity στο αρχικό σύνολο δεδομένων. Ωστόσο, μετά την εφαρμογή του αλγορίθμου SMOTE, ο αλγόριθμος που δίνει την υψηλότερη τιμή ευαισθησίας είναι το BayesNet που μπορεί να επιτύχει το 94.48% της ευαισθησίας(sensitivity). Επιπλέον, ο αλγόριθμος που δίνει το υψηλότερο specificity και την υψηλότερη τιμή ακρίβειας(accuracy) είναι το Rotation Forest (PCA)-J48 που μπορεί να επιτύχει 93.17% specificity και 90.81% accuracy.

Η έρευνα που γίνεται στην εργασία [10] στοχεύει στη σύγκριση 2 επιλεγμένων machine learning αλγορίθμων σε σύνολα δεδομένων διαφόρων μεγεθών που αναπτύσσονται σε διαφορετικές πλατφόρμες όπως:

- το Weka,
- το Scikit-Learn,
- και το Apache Spark

Αξιολογούνται με βάση το χρόνο εκπαίδευσης (training time), την ακρίβεια (accuracy) και το Root mean squared error. Αυτή η σύγκριση βοηθάει στο να διαπιστωθεί ποια πλατφόρμα είναι η πλέον κατάλληλη για να εργαστεί κάποιος εφαρμόζοντας παράλληλα υπολογιστικά ακριβούς επιλεγμένους machine learning αλγορίθμους σε ένα συγκεκριμένο μέγεθος δεδομένων.

Αρχικά θα αναφέρουμε λίγα λόγια για κάθε εργαλείο:

- Το Weka είναι ένα machine learning εργαλείο γραμμένο σε Java. Περιέχει μια συλλογή από εργαλεία visualization και αλγορίθμους για ανάλυση δεδομένων και πρόβλεψης μοντέλων, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτή τη λειτουργικότητα.
- Το Scikit-Learn είναι μια open source machine learning δομή για γλώσσα προγραμματισμού Python. Παρέχει έναν απλό τρόπο εκτέλεσης της εξόρυξης και της ανάλυσης δεδομένων.
- Τέλος, ο Pache Spark είναι μια γρήγορη μηχανή για επεξεργασία δεδομένων μεγάλης κλίμακας.

Στην έρευνα χρησιμοποιούνται 2 διαφορετικά data sets. Το Higgs Boson data set το οποίο είναι ένα μεγάλο data set και περιέχει 11x106 instances και 28 features και το Bank Marketing data set το οποίο είναι μικρό και περιέχει 41188 instances και 17 features.

Όσον αφορά την μεθοδολογία, χρησιμοποιείται μία 2 phase προσέγγιση που αξιολογεί τους αλγορίθμους μηχανικής μάθησης σε διαφορετικές πλατφόρμες όταν εφαρμόζεται σε σύνολα διαφορετικών μεγεθών. Κάθε φάση είναι ένα πείραμα που διεξάγεται σε διαφορετικά δεδομένα. Στην πρώτη φάση επιλέγεται το μικρότερο σύνολο δεδομένων (δηλαδή Bank marketing data set) στο οποίο αναπτύσσονται οι επιλεγμένοι αλγόριθμοι machine learning σε διάφορες πλατφόρμες. Ο στόχος ταξινόμησης εδώ είναι να προβλέψει εάν ο πελάτης θα εγγραφεί σε προθεσμιακή κατάθεση. Στη δεύτερη φάση επιλέγεται το μεγαλύτερο σύνολο δεδομένων (δηλαδή Higgs Boson data set) για να ελεγχθεί ο χρόνος εκπαίδευσης και η σχετική ακρίβεια μεταξύ των αλγορίθμων και επίσης να αναλύσουν οι πλατφόρμες που μπορούν να χειριστούν τόσο μεγάλα δεδομένα.

Οι machine learning αλγόριθμοι που επιλέχθηκαν είναι :

- Logistic Regression,
- Linear SVM

Επιλέχθηκαν αυτοί ως ταξινομητές διότι χρησιμοποιούνται ευρέως και εφαρμόζουν σταθερά σενάρια για διαφορετικά σύνολα δεδομένων. Είναι 2 από τους βασικούς αλγορίθμους και ως εκ τούτου εφαρμόζονται σε όλα σχεδόν τα frameworks.

Η σύγκριση των frameworks όπως είδαμε και πιο πάνω γίνεται μέσω:

- CPU Utilization, (χρησιμοποιείται για να διαπιστωθεί η αποτελεσματικότητα του κάθε αλγορίθμου)
- Training Time,
- Accuracy,
- Root mean squared error (RMSE)

Συνοψίζοντας, σκοπός του πειράματος είναι να συγκρίνει την απόδοση των αλγορίθμων μηχανικής μάθησης όταν αναπτύσσεται σε διαφορετικά πλαίσια όσον αφορά τον χρόνο εκπαίδευσης, την ακρίβεια και το RMSE. Αυτή η αξιολόγηση βοηθά να αποφασιστεί ποια πλατφόρμα είναι καλύτερα προσαρμοσμένη για ένα επιλεγμένο τύπο και μέγεθος δεδομένων και ποιοι αλγόριθμοι μηχανικής μάθησης πρέπει να εφαρμοστούν για μια συγκεκριμένη εφαρμογή.

Στο άρθρο [11] χρησιμοποιούνται data mining τεχνικές προκειμένου να προσδιοριστούν τα σημαντικά features του data set για να αυξηθεί η αποτελεσματικότητα του campaign, δηλαδή εάν ο πελάτης υπογράψει την προθεσμιακή κατάθεση. Δύο είναι οι μέθοδοι feature selection με τους οποίους ασχολείται η έρευνα:

- Information Gain,
- Chi-square

Οι μέθοδοι συγκρίνονται χρησιμοποιώντας έναν supervised αλγόριθμο μηχανικής μάθησης, τον Naive Bayes. Για την ανάλυση των δεδομένων χρησιμοποιείται το Weka.

Η μέθοδος Feature Selection είναι σημαντική διότι με την μείωση του αριθμού των features έχουμε καλύτερη αποδοτικότητα στην διαδικασία training και παράλληλα αύξηση της ακρίβειας της κατανομής.

Λίγα πράγματα για την κάθε μέθοδο:

- Chi-square, μετρά την έλλειψη ενός χαρακτηριστικού (feature) και μιας κλάσης. Το Chi-square στατιστικό μπορεί να οριστεί χρησιμοποιώντας το 2x2 πίνακα ενδεχόμενων κινδύνων ([πίνακας 7.3](#)).
- Information Gain, εμετρά το περιεχόμενο της πληροφορίας με την ανάλυση του feature που είναι γνωστό ([πίνακας 7.2](#)).

Ο αλγόριθμος Naive Bayes βασίζεται στον κανόνα του Bayes για τον υπολογισμό των υποθετικών πιθανοτήτων και υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα.

Το data set που χρησιμοποιείται περιέχει 17 features και 41118 instances.

Ο υπολογισμός της Chi-square μεθόδου:

$$x^2 = \sum_{i=1}^I c \sum_{j=1}^J r \frac{(O_{ij} - e_{ij})^2}{e_{ij}}, \text{ όπου } I, j \text{ στήλες και σειρές αντίστοιχα σε ένα ενδεχόμενο πίνακα, } O_{ij} \text{ είναι η}$$

παρατηρηθείσα συχνότητα και e_{ij} η προβλεπόμενη συχνότητα γίνεται μέσω της συνάρτησης.

$$\text{Ο υπολογισμός του IG: } IG(f) = - \sum_{i=1}^m P(c_i) \log P_r(c_i) + P_r(f) \sum_{i=1}^m m P_r(c_i|f) \log P_r(c_i|f) + P_r(\bar{f}) \sum_{i=1}^m m P_r(c_i|\bar{f}) \log P_r(c_i|\bar{f}), \text{ όπου } m \text{ ο αριθμός των κλάσεων και } p_i \text{ πιθανότητα του του feature } f \text{ στην κλάση } c_i.$$

Συγκρίνεται η απόδοση του ταξινομητή με βάση των precision, recall και F-measure values.

Όσον αφορά τα αποτελέσματα, στόχος είναι να αξιολογηθεί η δυνατότητα εφαρμογής μεθόδων feature selection σε σύνολο δεδομένων direct τραπεζικού marketing. Επομένως, πρώτα λαμβάνονται τα βασικά αποτελέσματα χρησιμοποιώντας τον ταξινομητή NB στο data set τραπεζικού marketing για περαιτέρω ανάλυση. Το πείραμα επικυρώνεται με 5-fold cross validation. Ο πίνακας 7.1 παρουσιάζει την απόδοση των βασικών ρυθμίσεων από την άποψη του precision, recall και F-measure του NB ταξινομητή. Το average weighted αποτέλεσμα του F-measure εμφανίζεται με έντονα γράμματα στον πίνακα 7.1.

Πέρα από τα κανονικά αποτελέσματα που προέκυψαν από τα πειράματα (πίνακα 7.1) δοκιμάστηκαν 2 μέθοδοι feature selection ονόματι IG και Chi-square για μείωση των χαρακτηριστικών προκειμένου να βρεθεί η απόδοση του ταξινομητή. Δοκιμάστηκαν 4 μεγέθη χαρακτηριστικών (5, 8, 10, 15) για το σύνολο δεδομένων τραπεζικού marketing. Το πείραμα επικυρώνεται χρησιμοποιώντας 5-fold cross validation που σημαίνει ότι ολόκληρο το set δεδομένων διαιρείται σε 5 set ίσου μεγέθους και ο ταξινομητής εκπαιδεύεται σε 4 train sets και δοκιμάζεται σε ένα test set. Αυτή η διαδικασία επαναλαμβάνεται 5 φορές και στη συνέχεια λαμβάνεται ένας μέσος όρος όλων των folds. Όπως φαίνεται στους πίνακες 2 και 3 το καλύτερο αποτέλεσμα επιτεύχθηκε με τη μέθοδο επιλογής χαρακτηριστικών IG με 5 χαρακτηριστικά:

- (1) Duration
- (2) poutcome
- (3) month
- (4) pdays
- (5) contact

Το F-measure αυξάνεται από 0.887 σε 0.888. Ωστόσο, όπως φαίνεται από τους πίνακες αποτελεσμάτων και οι δυο μέθοδοι IG και Chi-square επιτυγχάνονται με 0.880 σε F-measure για 10 χαρακτηριστικά. Αυτά τα δέκα υψηλότερα χαρακτηριστικά είναι: (1) duration, (2) poutcome, (3) month, (4) pdays, (5) contact, (6) previous, (7) age, (8) job, (9) housing και (10) balance.

Η δουλειά που γίνεται στην εργασία [12] σύμφωνα με το συγγραφέα είναι να αναπτύξει ένα μοντέλο που μπορεί να προβλέψει την καλύτερη δυνατή ακρίβεια του churn. Customer churn prediction είναι στην ουσία η πρόβλεψη που θα κάνει η τράπεζα για το ποιοι πελάτες είναι πιθανόν να ακυρώσουν κάποια εγγραφή τους σε μια υπηρεσία της τράπεζας. Χρησιμοποιήθηκαν υπάρχοντα μοντέλα SVM και kernel function αυτουνού και σε συνεργασία με τεχνητά νευρικά δίκτυα οδηγήθηκαν στην πρόβλεψη. Στην παρούσα εργασία ο ερευνητής έχει δημιουργήσει μια υβριδική προσέγγιση (KSVM + ANN) για να ανακαλύψει churn customers στον ιδιωτικό τομέα των τραπεζών και το αποτέλεσμα συγκρίνεται με 3 state of the art αλγόριθμους SVM, KSVM, ANN.

Classification είναι μία data mining μέθοδος η οποία έχει αποδειχθεί ότι παράγει καλύτερα αποτελέσματα στην πρόβλεψη για το churn των πελατών.

Η έρευνα αυτή πραγματοποιήθηκε με τη βοήθεια του εργαλείου R Studio.

Το input data set περιλαμβάνει προσωπικά στοιχεία πελατών και οικονομικά στοιχεία. Λαμβάνεται από το UCI Repository και αποτελείται από 20 attributes.

Λίγα στοιχεία για την ταξινόμηση χρησιμοποιώντας κάθε αλγόριθμο ξεχωριστά αλλά τον συνδυασμό αυτών:

- **Classification χρησιμοποιώντας KSVM αλγόριθμο:** χρησιμοποιείται για την ομαδοποίηση της εγγραφής 20 χαρακτηριστικών. Τα είδη ταξινόμησης που υποστηρίζει είναι τόσο αριθμητικά όσο και χαρακτήρες. Το Support Vector Machine (SVM) είναι ένας supervised υπολογισμός μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για αμοιβαίες ρυθμίσεις και λογιστικές δοκιμές, ως επί το πλείστον όμως αξιοποιείται ως μέρος των ζητημάτων χαρακτηρισμού.
- **Classification χρησιμοποιώντας ANN αλγόριθμο:** Feed-Forward νευρονικά δίκτυα και Multinomial Log-Linear μοντέλα χρησιμοποιούνται για την ταξινόμηση των δεδομένων. Είναι ένας επαναληπτικός αλγόριθμος. Στο Feed-Forward νευρικό σύστημα οι πληροφορίες μετακινούνται μόνο μπροστά. Είναι

βασισμένα σε προσομοιωμένο Neural Network και η απόφαση λαμβάνει πληροφορίες με μεγαλύτερη ακρίβεια. Υποστηρίζει μόνο αριθμητικές τιμές.

- **Classification χρησιμοποιώντας KSVM και ANN αλγόριθμους:** Το μοντέλο KSVM & ANN έδειξε καλύτερα αποτελέσματα ακρίβειας και είχε ως αποτέλεσμα τη βελτίωση του μοντέλου όταν εξετάστηκαν οι μετρήσεις με τους παραδοσιακούς ταξινομητές. Ο υβριδικός συνδυασμός ήταν ένα αποτέλεσμα της ανάκτησης της output παραγωγής από το KSVM και της εκτέλεσης του μοντέλου με τον ταξινομητή ANN με αποτέλεσμα την καλύτερη πρόβλεψη των τιμών του μοντέλου των churn δεδομένων στη ληφθείσα βάση δεδομένων.

Με άλλα λόγια τα βήματα που ακολουθούνται ([γράφημα 2](#)) για τον υβριδικό αλγόριθμο (KSVM + ANN) για το churn prediction είναι:

- **Βήμα 1:** φόρτωση του αρχικού data set
- **Βήμα 2:** εκτέλεση προεπεξεργασίας δεδομένων
- **Βήμα 3:** ταξινόμησης χρησιμοποιώντας τον KSVM
 - **Βήμα 3.1:** διαχωρισμός του data set σε Training data (Dtr) και Validation (Dval)
 - **Βήμα 3.2:** εκπαίδευση του data χρησιμοποιώντας KSVM
 - **Βήμα 3.3:** παραγωγή του αριθμού των κλάσεων
 - **Βήμα 3.4:** επικύρωση των δεδομένων (Dval) από την κλάση που προέρχονται
- **Βήμα 4:** πρόβλεψη του churn από τα επικυρωμένα δεδομένα
- **Βήμα 5:** εκχώρηση των προβλεπόμενων τιμών ως τιμές της κλάσης στόχου για το νέο μοντέλο
- **Βήμα 6:** ταξινόμηση με χρήση του ANN
- **Βήμα 7:** πρόβλεψη του churn του πελάτη από τα επικυρωμένα δεδομένα

Τέλος, τα αποτελέσματα αποδείχθηκαν εξαιρετικά όταν τα δύο μοντέλα ενσωματώθηκαν μεταξύ τους. Αναλυτικά τα αποτελέσματα που προκύπτουν από την αξιολόγηση των μοντέλων παρουσιάζονται στον [πίνακα 8](#).

Σύμφωνα με το συντάκτη της εργασίας [[13](#)] η συνεισφορά αυτού του άρθρου είναι η εφαρμογή αλγορίθμων μηχανικής μάθησης όπως:

- Deep Artificial Neural Network (DANN),
- Naive Bayes,
- Logistic Regression,
- k-Nearest Neighbors,
- Decision Trees,
- Support Vector Machine (SVM),

για την αναγνώριση της πιο κατάλληλης μεθόδου για να αξιολογήσει εάν ο πελάτης στόχος, της εκστρατείας του bank telemarketing, πρόκειται να κάνει κατάθεση ή όχι. Η αποτελεσματικότητα αυτών των αλγορίθμων έχει αξιολογηθεί με Receiving Operator Characteristic (ROC) και Cumulative Accuracy Profile (CAP) ανάλυση καμπύλης όπως φαίνεται και στο [γράφημα 3](#).

Όσον αφορά το data set είναι βασισμένο σε πραγματικά δεδομένα από το UCI, περιέχει 45211 καταγραφές. Για σκοπούς αξιολόγησης το data set χωρίζεται σε training set (36168 contacts) και test set (9043 contacts).

Είναι ξεκάθαρο από τον [πίνακα 9](#) ότι τα καλύτερα μοντέλα για την πρόβλεψη της αποτελεσματικότητας του bank telemarketing είναι ο Deep Artificial Neural Network και ο Random Forest. Οι αλγόριθμοι Logistic Regression και Naive Bayes δεν είναι τόσο κατάλληλοι όσο οι άλλοι μέθοδοι ταξινόμησης για τέτοιου είδους προβλήματα λόγω της κακής ακρίβειας και των προβλημάτων υπερφόρτωσης. Η καλύτερη δυνατή στρατηγική για να αποφευχθεί η υπερφόρτωση είναι συγκεκριμένα η 10-fold cross validation.

Η εργασία [[14](#)] ασχολείται με την πρόβλεψη churn για τους πελάτες της εμπορικής τράπεζας με βάση το μοντέλο SVM και χρησιμοποιεί τη μέθοδο τυχαίας δειγματοληψίας (random sampling method) για τη βελτίωση του SVM μοντέλου λαμβάνοντας υπόψη τα imbalance χαρακτηριστικά του συνόλου δεδομένων των πελατών.

Η SVM είναι μέθοδος μηχανικής μάθησης. Στοχεύει στην εύρεση ενός υπερπλάνου που μπορεί να ικανοποιήσει τις απαιτήσεις ταξινόμησης και να κάνει trained σημεία πολύ μακριά από την επιφάνεια ταξινόμησης, δηλαδή αναζητώντας ένα πρόσωπο ταξινόμησης για να κάνει το περιθώριο στις 2 πλευρές του μέγιστο.

Οι βέλτιστες συναρτήσεις ταξινόμησης των μοντέλων SVM και Nonlinear SVM διατυπώνονται παρακάτω:

$$f(x) = \text{sng}\left(\sum_{i=1} l a_i y_i () x_i \cdot x\right) + b \quad \text{και} \quad f(x) = \text{sng}\left(\sum_{i=1} l a_i y_i K(x_i, x) + b\right), \text{αντίστοιχα.}$$

Λόγω του χαμηλού ποσοστού των churners των εμπορικών τραπεζών η κατανομή του συνόλου δεδομένων είναι πολύ imbalanced. Για την αλλαγή της κατανομής δεδομένων προκειμένου να μειωθεί η ανισορροπία του συνόλου των δεδομένων χρησιμοποιείται η μέθοδος random sampling. Υπάρχουν 2 τύποι random sampling:

- i) under-sampling, εξαλείφει τα δείγματα της majority κλάσης για να βελτιώσει την ακρίβεια της πρόγνωσης της minority κλάσης, ενώ
- ii) oversampling, δεν αλλάζει τον αριθμό των δειγμάτων της majority κλάσης αλλά αυξάνει τον αριθμό των δειγμάτων της minority κλάσης.

Για την αξιολόγηση της προγνωστικής ικανότητας του μοντέλου χρησιμοποιείται η F-measure, η οποία υπολογίζεται από το confusion matrix της ταξινόμησης και της αρμονικής σχέσης του recall και του precision. Η

συνάρτηση υπολογισμού του F-measure είναι:
$$F = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

Το data set που χρησιμοποιείται είναι δεδομένα πελατών από Κινέζικη εμπορική τράπεζα. Μέσω της προεπεξεργασίας του συνόλου δεδομένων όπως η συμπλήρωση των τιμών που λείπουν και η αφαίρεση των ακραίων τιμών τελικά υπάρχουν 46406 έγκυρες εγγραφές δεδομένων.

Κλείνοντας, αλλάζοντας την κατανομή του δείγματος, η μέθοδος Random sampling έχει υψηλότερο βαθμό αναγνώρισης. Επομένως, χρησιμοποιείται η random sampling μέθοδος για τη βελτίωση της SVM μεθόδου και την επιλογή του F-measure αξιολογείται η προγνωστική ισχύς. Για την σύγκριση του αποτελέσματος πρόβλεψης δημιουργείται επίσης μοντέλο Logistic Regression. Τα αποτελέσματα δείχνουν ότι ο συνδυασμός της μεθόδου random sampling και του μοντέλου SVM μπορεί να βελτιώσει την προγνωστική ισχύ και να βοηθήσει τις εμπορικές τράπεζες να προβλέψουν τα churners ακριβέστερα.

Σύμφωνα με τον συγγραφέα της εργασίας [15] πραγματοποιήθηκε μια μελέτη πρόβλεψης για να διαπιστωθεί εάν η εκστρατεία μιας τράπεζας οδηγεί σε απόκτηση νέων πελατών. Το data set του UCI χρησιμοποιήθηκε δημιουργώντας μοντέλα με τους ίδιους αλγόριθμους ταξινόμησης σε διαφορετικά προγράμματα εξόρυξης δεδομένων. Χρησιμοποιήθηκαν κριτήρια όπως accuracy, precision και f-measure για τη δοκιμή των επιδόσεων των μοντέλων ταξινόμησης. Κατά την δημιουργία των μοντέλων ταξινόμησης τα test και training data sets χωρίστηκαν τυχαία με τη μέθοδο αναμονής για την αξιολόγηση της απόδοσης του συνόλου δεδομένων. Το data set διαιρέθηκε σε training και test data sets με αναλογίες διαχωρισμού 60-40%, 75-25% και 80-20%. Τα προγράμματα εξόρυξης δεδομένων που χρησιμοποιούνται για αυτές τις διαδικασίες είναι τα:

- R,
- Knime,
- RapidMiner,
- Weka

Και οι αλγόριθμοι ταξινόμησης που χρησιμοποιούνται συνήθως σε αυτές τις πλατφόρμες είναι:

- k-nearest neighbor (k-nn)
- Naive Bayes
- C4.5 decision tree

Το data set της έρευνας είναι το Bank marketing data set του UCI το οποίο αποτελείται από 45211 εγγραφές πελατών και 17 attributes.

Στους παρακάτω πίνακες εμφανίζονται τα αποτελέσματα στο κάθε πρόγραμμα ξεχωριστά ανάλογα με το ποσοστό holdout separation.

Στους πίνακες 10.1, 10.2, 10.3 ελήφθησαν τα ίδια αποτελέσματα για την αξιολόγηση της απόδοσης. Και στα 3 κριτήρια, η καλύτερη απόδοση ήταν το δέντρο αποφάσεων C4.5. Επιπλέον, το πρόγραμμα Weka έδωσε καλύτερα αποτελέσματα στο κριτήριο ακρίβειας (accuracy) και το πρόγραμμα R έδωσε καλύτερα αποτελέσματα στα άλλα δύο κριτήρια.

Στην εργασία [16] προτείνεται ένα υβριδικό σύστημα, το οποίο πρώτα ταξινομεί τους πελάτες για να αποφασίσουν αν ενδιαφέρονται για τα προσφερόμενα προϊόντα και μετά τις συγκεντρώνει προκειμένου να προτείνει προϊόντα και κανάλια (channels). Με τον όρο κανάλια θέλει να δείξει τον τρόπο με τον οποίο θα καταφέρει η τράπεζα να φτάσει στον πελάτη. Τα πειράματα με πραγματικά τραπεζικά δεδομένα δείχνουν αρκετά ακριβή αποτελέσματα για την πρόβλεψη του κατάλληλου προϊόντος και καναλιού για τους πελάτες. Επιπλέον, η ανάλυση κόστους-κέρδους προστίθεται επίσης σε αυτό το πρόβλημα.

Λόγω του μεγάλου όγκου των δεδομένων χρησιμοποιούνται τεχνικές data mining, όπως clustering και classification τεχνικές προκειμένου να ομαδοποιηθούν και να μοντελοποιηθούν του πελάτες σύμφωνα με τα χαρακτηριστικά τους.

Το πρόβλημα στην ουσία είναι η πρόβλεψη αν ο πελάτης θα αποδεχθεί την προσφορά κάποιου προϊόντος αλλά και η πρόβλεψη του προϊόντος και του καναλιού επικοινωνίας. Για την λύση αυτού του προβλήματος προτάθηκαν 2 μέθοδοι:

- Partitioning method, δημιουργείται ένα set από clusters σε σχέση με τα προϊόντα, τα κανάλια και το IsSold label και μετά χρησιμοποιώντας αυτούς τους clusters στόχος είναι η πρόβλεψη της απόφασης των πελατών.
- Model based prediction method, χρησιμοποιούνται μοντέλα που δημιουργούνται από τα αρχεία καταγραφής της προσφοράς των προϊόντων στους πελάτες, όπου σε περίπτωση που οι πελάτες αποδεχθούν την προσφορά να γίνει σωστή πρόβλεψη του προϊόντος και του καναλιού.

Ο κύριος σκοπός είναι να βελτιωθεί το return rate για το bank direct marketing. Δεδομένου ότι το return rate της προώθησης εξαρτάται από την ακρίβεια της πρόβλεψης και το κίνητρο για τη partitioning based method είναι η αύξηση της ακρίβειας και η ακρίβεια βελτιώνεται όσο αυξάνεται ο αριθμός των cluster. Παρ' όλα αυτά για κάποια campaigns η τιμή ευαισθησίας μπορεί να είναι πιο σημαντική από την τιμή ακρίβειας, γι' αυτό το λόγο προτείνεται μια δεύτερη μέθοδος που χρησιμοποιεί μια προσέγγιση με βάση το μοντέλο overpartitioning. Έτσι λόγω αυτής της υβριδικής προσέγγισης η τιμή ευαισθησίας αυξάνεται δραματικά ενώ η τιμή ακρίβειας όχι.

Το data set που χρησιμοποιείται στην έρευνα προβάλλεται από μια τούρκικη τράπεζα η οποία εφαρμόζει direct marketing campaigns. Κάθε καταχώριση στον σύνολο δεδομένων εμφανίζει στον πελάτη τα διάφορα χαρακτηριστικά του και αν αγοράζει το προσφερόμενο προϊόν μέσω του καναλιού που χρησιμοποιείται για την επαφή με τον πελάτη. Περιέχει 81915 instances και 13 attributes.

Όσον αφορά τα πειράματα που έγιναν:

Αρκετά είναι τα classification methods που εφαρμόζονται στο model based method με partitioning και τα καλύτερα αποτελέσματα λαμβάνονται με τον Naive Bayes και C4.5 decision tree. Τα αποτελέσματα του C4.5 είναι καλύτερα από του Naive Bayes. Σε σύγκριση με τη partitioning based method η τιμή TN (true negative) καθώς και η ακρίβεια μειώνονται με τον αλγόριθμο ταξινόμησης Naive Bayes. Από την άλλη πλευρά, τα αποτελέσματα του δέντρου απόφασης C4.5 δείχνουν ότι με ποσοστό ακρίβειας 74.47%, επιτυγχάνεται το καλύτερο αποτέλεσμα της πρόβλεψης προϊόντος, καναλιού και ετικέτας (label). Η καλύτερη TP (true positive) τιμή επιτυγχάνεται με τον C4.5. Οι καλύτερες τιμές επιτεύχθηκαν με τον αριθμό των 100 clusters, αλλά δεν σημειώθηκε κάποια άλλη βελτίωση μετά τον αριθμό αυτό.

Η μελέτη που συναντάται στην εργασία [17] προτείνει ένα νέο δέντρο απόφασης για αβέβαια δεδομένα που ονομάζεται uncertain decision tree (UDT), βασισμένο στον uncertain genetic clustering αλγόριθμο (UGCA). Το UDT επεκτείνει το δέντρο αποφάσεων ώστε να χειρίζεται δεδομένα με αβέβαιες πληροφορίες, όπου η αβεβαιότητα πρέπει να λαμβάνεται υπόψη για την επίτευξη αποτελεσμάτων υψηλής ποιότητας. Στο UDT, το UGCA αναζητά αυτόματα τον κατάλληλο αριθμό κλάδων κάθε κόμβου, βάση της τιμής σφάλματος της ταξινόμησης και του χρόνου ταξινόμησης του UDT. Με άλλα λόγια, το UGCA μειώνει τόσο την τιμή σφάλματος ταξινόμησης όσο και το χρόνο υπολογισμού και στη συνέχεια βελτιστοποιεί το προτεινόμενο UDT. Προτού σχεδιαστεί το UDT χρησιμοποιώντας το UGCA, αναπτύσσεται επίσης ακόμα ένας uncertain αλγόριθμος, ο uncertain merging algorithm (UMA) που βοηθάει στην μείωση του αβέβαιου συνόλου δεδομένων, επιτρέποντας έτσι στο UGCA να επεξεργάζεται αποτελεσματικά ένα μεγάλο σύνολο δεδομένων. Είναι σημαντικό ότι τα πειραματικά αποτελέσματα δείχνουν ότι το προτεινόμενο UDT υπερτερεί έναντι των κλασικών uncertain δέντρων απόφασης.

Στα πειράματα το UDT με βάση το UGCA συγκρίνεται με το CFDT με βάση το FCM. Για να ελεγχθεί η απόδοση του UGCA, το UDT συγκρίνεται επίσης με τα δύο δέντρα αποφάσεων το MGP και το GCFDT που βασίζονται στους genetic algorithms. Επιπλέον, το προτεινόμενο UGCA συγκρίνεται με τον αλγόριθμο UK-means για αβέβαια δεδομένα.

Σε αυτό το πείραμα χρησιμοποιήθηκαν 10 data sets, 8 από το UCI repository, ένα σύνολο δεδομένων ομιλίας που προέκυψε από τη βάση δεδομένων ISOLET και ένα σύνολο δεδομένων τραπεζικού marketing από το UCI.

Κλείνοντας στην εργασία [18], με βάση δεδομένα που θα μελετηθούν με τη βοήθεια των μοντέλων μηχανικής μάθησης, θα προβλεφθεί ποιοι πελάτες θα εγγραφούν και τι μπορούν να κάνουν οι τράπεζες για να αυξήσουν το ποσοστό συνδρομής.

Η μεθοδολογία που ακολουθείται στο άρθρο έχει ως εξής:

- **Προγραμματίζοντας σε Python:** ολόκληρη η έρευνα κωδικοποιείται χρησιμοποιώντας Python 3. Τα πακέτα/βιβλιοθήκες που χρησιμοποιούνται είναι: το numpy, για χειρισμό των arrays, το pandas για dataframe λειτουργίες και matplotlib & seaborn για οπτικοποίηση. Οι βιβλιοθήκες sklearn είναι επίσης σημαντικές στην παροχή πακέτων για αλγορίθμους μηχανικής μάθησης.
- **Data cleaning και διερευνητική ανάλυση:** το data set που χρησιμοποιείται προβάλλεται από το UCI και περιέχει πληροφορίες για 41188 πελάτες με 20 διαφορετικά features. Η target variable αποτελείται από Yes (client subscribed) και No (client did not subscribed).
 - Το πρώτο βήμα είναι η φόρτωση του συνόλου δεδομένων σε ένα dataframe για εύκολο χειρισμό και εξερεύνηση χρησιμοποιώντας το πακέτο pandas.
 - Το επόμενο βήμα είναι η εξερεύνηση και ο καθαρισμός των κατηγορηματικών μεταβλητών (όπως πχ 'job type', 'marital status').
 - Πολλά από αυτά τα features περιέχουν missing values. Ο τρόπος για να αντιμετωπιστεί αυτό το πρόβλημα είναι μέσω της συνάρτησης cross-tab που παρέχει η Python.
 - Τέλος, δημιουργείται ένα heat map για να δείξει αν υπάρχει ισχυρή συσχέτιση μεταξύ της target variable και τυχόν ανεξάρτητων variables. Ο χάρτης θερμότητας δημιουργείται χρησιμοποιώντας Spearman correlation, ο οποίος μετρά τον βαθμό στον οποίο ευθυγραμμίζονται οι κατατάξεις κάθε variable.
- **Κατασκευή μοντέλου:** το data set χωρίζεται σε training data και testing data και με τη βοήθεια του training data μπορούν να βρεθούν οι παράμετροι του μοντέλου που χρησιμοποιείται και μετά αυτό να εφαρμοστεί στα test data προκειμένου να εξακριβωθεί η απόδοση του μοντέλου και να σχεδιαστούν συμπεράσματα για την ικανότητα πρόβλεψης που διαθέτει. Αυτό μπορεί να πραγματοποιηθεί με την λειτουργία sklearn.crossvalidation.train_test_split. Οι κατασκευές των μοντέλων γίνεται με τη βοήθεια των εξής μεθόδων: Logistic Regression, Decision Trees, Random Forest, AdaBoost και Gradient Boosting (classifiers).
- **Αξιολόγηση μοντέλων:** για την αξιολόγηση όλων των μοντέλων χρησιμοποιείται το AUC score. Η μέση τιμή του AUC υπολογίζεται για κάθε σύνολο επιλεγμένων παραμέτρων. Το τελικό μοντέλο επιλέγεται με βάση την υψηλότερη μέση τιμή του AUC.
- **Hyper-parameter Tuning:** για όλα τα μοντέλα που εφαρμόστηκαν, οι υπερ-παράμετροι χρησιμοποιήθηκαν για να επιτευχθεί η βέλτιστη απόδοση του ταξινομητή.
 - Για τον Logistic Regression 2 υπερ-παράμετροι χρησιμοποιούνται: penalty type (L1, L2 penalty) και το regularization coefficient
 - Για τα Decision Trees, Random Forest και Gradient Boosted Trees χρησιμοποιούνται 2 υπερ-παράμετροι: minimum samples split (ο μικρότερος αριθμός των δειγμάτων) και minimum samples leaf.
 - Τέλος, για τον AdaBoost classifier μόνο μία υπερ-παράμετρος χρησιμοποιείται: ο αριθμός των estimators.

Στα αποτελέσματα η καλύτερη απόδοση μοντέλου δημιουργήθηκε από τον AdaBoost classifier, διότι είχε την υψηλότερη τιμή AUC.

Κλείνοντας, με βάση αυτά τα μοντέλα δημιουργούνται κάποιες σοβαρές και αρκετά βοηθητικές προτάσεις για το τμήμα marketing των τραπεζών.

2.2.1 Πίνακες Αποτελεσμάτων Εργασιών

Πίνακας 1.

Σύμφωνα με την εργασία [2] στον πίνακα 1, όπως μπορούμε με βάση το Balanced accuracy, ότι ο αλγόριθμος Random Forest είναι πιο γρήγορος και πιο αποτελεσματικός από τους υπολοίπους, με τον SVM να έρχεται αμέσως μετά και ο KNN να είναι όπως φαίνεται εξάλλου η χειρότερη επιλογή από όλες.

Measure		SVM	KNN	RF	ADA
Balanced accuracy A_c	min	0.68	0.28	0.73	0.74
	max	0.8	0.44	0.82	0.8
Specificity S_p	min	0.71	0.16	0.71	0.68
	max	0.84	0.4	0.92	0.82
Sensitivity S_c	min	0.64	0.36	0.72	0.72
	max	0.67	0.62	0.81	0.82
Runtime(in seconds)	min	0.74	0.01	3.63	59.5
	max	1.18	0.02	4.06	60.54

Πίνακας 2.

Σύμφωνα με την εργασία [5] στον παρακάτω πίνακα παρουσιάζεται η σημαντικότητα των στοιχείων ανά κατηγορία σχετικά με τα μοντέλα των TAN, LR, MLPNN και C5.0.

Model	Importance															
Attributes	Age	Job	Marital	Education	Default	Balance	Housing	Loan	Contact	Day	Month	Duration	Campaign	Pdays	Previous	Output
MLPNN	0.02	0.05	0.02	0.02	0.01	0.01	0.02	0.02	0.06	0.05	0.18	0.33	0.02	0.04	0.01	0.15
TAN	0.29	0.19	0.15	0.1	0.02	0.01	0.01	0.07	0	0	0	0.15	0.03	0	0	0
LR	--	0.03	0.01	0.01	--	--	0.06	0.02	0.08	0	0.05	0.5	0.03	--	0.01	0.18
C5.0	0	0.02	0.03	0	--	0.02	0	0.02	0	0	0	0.72	0.02	0.01	0.01	0.15

Πίνακας 3.

Σύμφωνα με την εργασία [5] ο παρακάτω πίνακας περιέχει τα ποσοστά των στατιστικών μέτρων MLPNN, TAN, LR & C5.0 για την εκπαίδευση και τη δοκιμή των υποσυνόλων.

Model	Partition	Accuracy	Sensitivity	Specificity
MLPNN	Training	90.92%	65.66%	93.28%
	Testing	90.49%	62.20%	93.12%
TAN	Training	89.16%	55.87%	91.97%
	Testing	88.75%	52.19%	91.73%
LR	Training	90.09%	64.83%	91.76%
	Testing	90.43%	65.53%	92.16%
C5.0	Training	93.23%	76.75%	94.92%
	Testing	90.09%	59.06%	93.23%

Πίνακας 4.

Σύμφωνα με την εργασία [8] ο παρακάτω πίνακας δείχνει ότι ο αυξανόμενος αριθμός επαναλήψεων δεν εγγυάται ότι η απόδοση που επιτυγχάνεται θα είναι καλύτερη. Σε αυτό το αποτέλεσμα η επανάληψη 30 δίνει την καλύτερη απόδοση.

Method	Iteration	Accuracy	Sensitivity
Ordinary SVM	---	91.67%	83.80%
	10	92.53%	85.93%
	20	94.30%	89.68%
Adaboost SVM	30	95.07%	91.65%
	40	94.81%	91.05%
	50	94.86%	91.05%

Πίνακας 5.

Σύμφωνα με την εργασία [9] ο παρακάτω πίνακας δείχνει αναλυτικά Τα αποτελέσματα συγκρίσεων των Decision Tree, Rotation Forest, MLP RBF Neural Network, BayesNet και Naive Bayes. Με σκούρα γραφή είναι οι υψηλότερες τιμές που καταγράφονται για κάθε μοντέλο.

Model	Sensitivity%		Specificity%		Accuracy%	
	Original Data	SMOTE	Original Data	SMOTE	Original Data	SMOTE
J48	57.11	76.63	93.19	92.75	89.51	89.43
ADTree	60.03	80.1	90.98	88.77	89.21	87.47
LADTree	60.87	75.78	92	90.79	89.68	88.01
RandomTree	47.61	71.67	92.64	91.75	87.51	87.54
REPTree	60.05	79.26	92.78	92.15	89.85	89.68
Rotation Forest (PCA) J48	63.35	81.23	92.21	93.17	90.05	90.81
Rotation Forest (PCA) ADTree	67.73	80.12	89.78	87.81	89.12	86.75
Rotation Forest (PCA) LADTree	65.46	77.52	90.93	87.72	89.63	86.28
Rotation Forest (PCA) Random Tree	62.42	82.7	91.89	91.73	89.81	90.12
Rotation Forest (PCA) REPTree	63.48	81.09	91.9	92.08	89.93	90.03
MultilayerPerception	57.56	73.83	93.12	91.77	89.57	88.13
RBFNetwork	53.21	61.86	90.94	86.09	88.48	82.24
BayesNet	50.61	94.48	92.5	89.63	88.18	90.28
NaiveBayes	42.16	54.09	92.07	89.34	86.22	80.51

Πίνακας 6.

Σύμφωνα με την εργασία [11] στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα για κάθε πείραμα. Στον **πίνακα 7.1** παρουσιάζονται τα βασικά αποτελέσματα που προέκυψαν από το σύνολο δεδομένων, χωρίς να υποστεί κάποια επεξεργασία πιο πριν.

Has the client subscribed a term deposit?							
Baseline	Positive class			Negative class			Average
	P	R	F	P	R	F	F
Baseline	0.44	0.51	0.47	0.94	0.92	0.93	0.87

Στον **πίνακα 7.2** παρουσιάζονται τα αποτελέσματα από την εφαρμογή feature selection και πιο συγκεκριμένα μέσω της μεθόδου IG.

Number of top features	Has the client subscribed a term deposit?						Average
	Positive class			Negative class			
	P	R	F	P	R	F	
5	0.52	0.41	0.46	0.93	0.95	0.94	0.88
8	0.46	0.46	0.46	0.93	0.93	0.93	0.88
10	0.48	0.48	0.48	0.93	0.93	0.93	0.88
15	0.44	0.51	0.47	0.94	0.92	0.93	0.87

Στον **πίνακα 7.3** παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή feature selection μέσω της εφαρμογής της μεθόδου Chi-square.

Number of top features	Has the client subscribed a term deposit?						Average
	Positive class			Negative class			
	P	R	F	P	R	F	
5	0.47	0.42	0.45	0.93	0.94	0.93	0.88
8	0.46	0.46	0.46	0.93	0.93	0.93	0.88
10	0.48	0.48	0.48	0.93	0.93	0.93	0.88
15	0.44	0.51	0.47	0.94	0.92	0.93	0.87

Πίνακας 8.

Σύμφωνα με την εργασία [12] στον παρακάτω πίνακα γίνεται η αξιολόγηση του κάθε αλγορίθμου ξεχωριστά αλλά και του υβριδικού μοντέλου με βάση τα στατιστικά μέτρα που απεικονίζονται παρακάτω.

Models	SVM	KSVM	ANN	KSVM+ANN	Derivations
Measure	Value				
Sensitivity	0.94	0.95	0.98	0.99	$TPR = TP / (TP + FN)$
Specificity	0.42	0.71	0.5	0.5	$SPC = TN / (FP + TN)$
Precision	0.92	0.98	0.98	0.99	$PPV = TP / (TP + FP)$
Negative Predictive					$NPV = TN / (TN + FN)$
Value	0.5	0.5	0.5	0.5	
False Positive Rate	0.58	0.29	0.5	0.5	$FPR = FP / (FP + TN)$
False Discovery					$FDR = FP / (FP + TP)$
Rate	0.08	0.02	0.02	0.01	
False Negative					$FNR = FN / (FN + TP)$
Rate	0.06	0.05	0.02	0.01	
Accuracy	0.88	0.93	0.96	0.98	$ACC = (TP + TN) / (P + N)$
F1 Score	0.93	0.96	0.98	0.99	$F1 = 2TP / (2TP + FP + FN)$
Matthews					$TP * TN - FP * FN /$
Correlation					$\sqrt{((TP + FP) * (TP +$
Coefficient	0.39	0.56	0.47	0.49	$FN) * (TN + FN))}$

Πίνακας 9.

Σύμφωνα με την εργασία [13] στον παρακάτω πίνακα παρουσιάζεται η ανάλυση απόδοσης των αλγορίθμων με βάση το Accuracy, CAP curve, Variance και k-fold cross validation accuracy και μέτρηση σε %.

	Accuracy%	Variance%	% of positive observations at 50% (CAP curve)	K-fold cross validation accuracy %	
				Train set	Train set
KNN	86.23	1.03	78.41	88.56	87.45
RF	90.88	1.72	95.83	95.83	90.23
LR	86.19	5.23	60.19	60.19	86.19
ANN	90.29	1.57	92.41	92.41	91.21
NB	86.87	10.94	88.03	88.03	86.4
SVM	89.67	1.35	93.27	89.35	90.12

Πίνακας 10.

Σύμφωνα με την εργασία [15] στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα που προέκυψαν, για κάθε μοντέλο-αλγόριθμο με βάση τα παρακάτω στατιστικά μέτρα για κάθε πρόγραμμα που χρησιμοποιήθηκε ξεχωριστά.

Στον **πίνακα 10.1** τα αποτελέσματα που προκύπτουν με 60% holdout separation.

Criteria		Accuracy			Precision			F-measure	
Algorithms	NB	k-nn	C4.5	NB	k-nn	C4.5	NB	k-nn	C4.5
R	0.87	0.87	0.9	0.93	0.91	0.93	0.93	0.93	0.95
Knime	0.87	0.86	0.9	0.92	0.87	0.94	0.83	0.9	0.94
RapidMiner	0.86	0.85	0.89	0.92	0.86	0.93	0.88	0.89	0.92
Weka	0.88	0.86	0.9	0.94	0.91	0.94	0.93	0.92	0.94

Στον **πίνακα 10.2** τα αποτελέσματα που προκύπτουν με 75% holdout separation.

Criteria		Accuracy			Precision			F-measure	
Algorithms	NB	k-nn	C4.5	NB	k-nn	C4.5	NB	k-nn	C4.5
R	0.88	0.88	0.91	0.93	0.91	0.94	0.92	0.93	0.95
Knime	0.84	0.87	0.91	0.88	0.88	0.93	0.88	0.87	0.9
RapidMiner	0.87	0.84	0.89	0.88	0.88	0.93	0.86	0.86	0.93
Weka	0.88	0.87	0.9	0.93	0.91	0.94	0.93	0.92	0.95

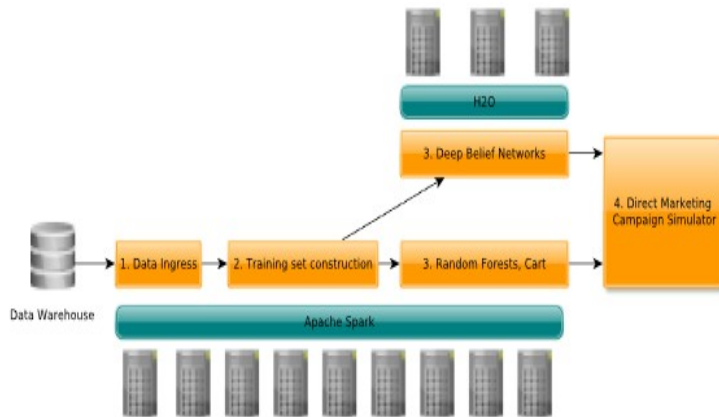
Στον **πίνακα 10.3** τα αποτελέσματα που προκύπτουν με 80% holdout separation.

Criteria		Accuracy			Precision			F-measure	
Algorithms	NB	k-nn	C4.5	NB	k-nn	C4.5	NB	k-nn	C4.5
R	0.88	0.87	0.91	0.93	0.91	0.94	0.93	0.93	0.95
Knime	0.84	0.87	0.91	0.88	0.88	0.93	0.88	0.87	0.9
RapidMine	0.87	0.84	0.88	0.88	0.88	0.93	0.87	0.86	0.93
Weka	0.88	0.86	0.9	0.93	0.91	0.94	0.93	0.92	0.95

2.2.2 Γραφήματα Εργασιών

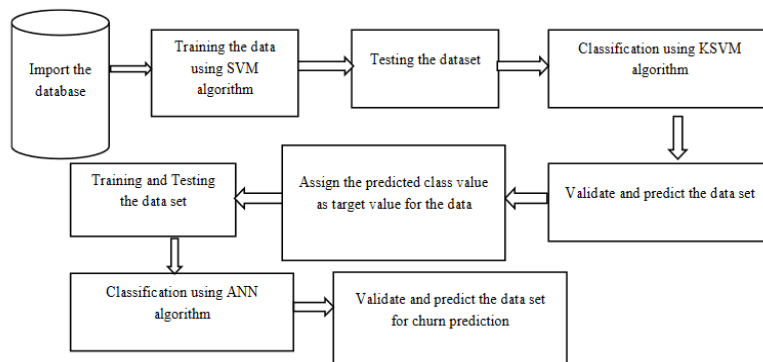
Γράφημα 1.

Σύμφωνα με την εργασία [3] στο παρακάτω σχήμα απεικονίζεται η αρχιτεκτονική συστήματος υψηλού επιπέδου για την πρόβλεψη πιθανότητας αγοράς πίστωσης.



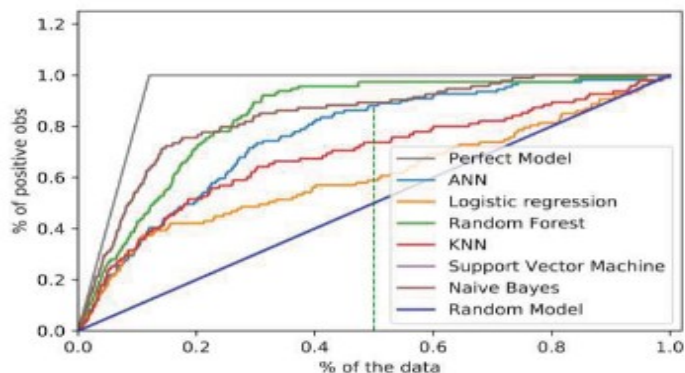
Γράφημα 2.

Σύμφωνα με την εργασία [12] το παρακάτω σχήμα απεικονίζει το προτεινόμενο πλαίσιο για την βήμα προς βήμα αναπαράσταση ροής δεδομένων.



Γράφημα 3.

Σύμφωνα με την εργασία [13] στο παρακάτω γράφημα απεικονίζονται οι CAP καμπύλες για όλους τους αλγορίθμους.



Κεφάλαιο 3^ο:

3.1 Περιγραφή Αλγορίθμων

1) C4.5 αλγόριθμος:

- Χρησιμοποιείται στην εξόρυξη δεδομένων ως decision tree classifier που μπορεί να χρησιμοποιηθεί για την λήψη μιας απόφασης ενός συγκεκριμένου δείγματος δεδομένων.
- Σε κάθε κόμβο του δέντρου, ο C4.5 επιλέγει το χαρακτηριστικό των δεδομένων που χωρίζει αποτελεσματικότερα το σύνολο των δειγμάτων του σε υποσύνολα εμπλουτισμένα στη μία ή στην άλλη κλάση. Το κριτήριο διαχωρισμού είναι το ομαλοποιημένο κέρδος πληροφοριών (normalized information gain). Το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος πληροφοριών είναι αυτό που επιλέγεται για τη λήψη της απόφασης.
- Ο C4.5 αλγόριθμος για την κατασκευή δέντρων απόφασης εφαρμόζεται στο weka ως ταξινομητής που ονομάζεται J48. Οι ταξινομητές, όπως τα φίλτρα είναι οργανωμένοι σε μια ιεραρχία: Το J48 έχει το πλήρες όνομα weka. Classifiers.
- Πλεονεκτήματα του C4.5 έναντι άλλων decision tree συστημάτων:
 - Ο αλγόριθμος χρησιμοποιεί Single Pass Pruning Process προκειμένου να μειώσει την υπερφόρτωση.
 - Μπορεί να λειτουργήσει τόσο με διακριτά(Discrete) όσο και με συνεχή(Continuous) δεδομένα.
 - Μπορεί να χειριστεί το θέμα των ελλিপών δεδομένων πολύ καλά.
- Με λίγα λόγια, το C4.5 εφαρμόζεται αναδρομικά με παρακάτω ακολουθία:
 - Ελέγξτε αν ο αλγόριθμος πληροί τα κριτήρια τερματισμού.
 - Υπολογιστικές πληροφορίες-θεωρητικά κριτήρια για όλα τα χαρακτηριστικά(attributes).
 - Επιλέξτε το καλύτερο attribute σύμφωνα με τα παραπάνω κριτήρια πληροφοριών.
 - Δημιουργείστε έναν κόμβο απόφασης με βάση το καλύτερο χαρακτηριστικό.
 - Διαχωρίστε το σύνολο δεδομένων με βάση τον κόμβο απόφασης που δημιουργήθηκε στο προηγούμενο βήμα.
 - Για όλα τα υποσύνολα στο προηγούμενο βήμα, καλέστε τον αλγόριθμο C4.5 για να λάβετε ένα υπο-δέντρο(αναδρομική κλήση).
 - Συνδέστε το δέντρο που αποκτήθηκε στο βήμα 6 στον κόμβο απόφασης στο βήμα 4. Επιστρέφει δέντρο.

2) NaiveBayes αλγόριθμος:

- Είναι ένας αλγόριθμος ταξινόμησης. Παραδοσιακά υποθέτει ότι οι τιμές εισόδου είναι ονομαστικές(nominal), αν και οι αριθμητικές(numerical) εισοδοί υποστηρίζονται από την υποβολή μιας κατανομής.
- Ο Naive Bayes χρησιμοποιεί μια απλή εφαρμογή του θεωρήματος Bayes όπου η προηγούμενη πιθανότητα για κάθε κλάση υπολογίζεται από τα training δεδομένα και θεωρείται ότι είναι ανεξάρτητη η μία από την άλλη. Αυτή είναι μία μη ρεαλιστική υπόθεση επειδή περιμένουμε οι μεταβλητές να αλληλεπιδρούν και να εξαρτώνται, αν και αυτή η υπόθεση καθιστά τις πιθανότητες γρήγορες και εύκολες στον υπολογισμό. Ακόμα και καί'τω από αυτήν την μη ρεαλιστική υπόθεση, ο Naive Bayes έχει αποδειχθεί ότι είναι ένας πολύ αποτελεσματικός αλγόριθμος ταξινόμησης.
- Ο Naive Bayes υπολογίζει την επόμενη πιθανότητα για κάθε κλάση και κάνει μια πρόβλεψη για την κλάση με την υψηλότερη πιθανότητα. Ως εκ τούτου, υποστηρίζει προβλήματα δυαδικής ταξινόμησης και ταξινόμησης πολλαπλών κατηγοριών.
- Για να καταλάβουμε καλύτερα πως λειτουργεί ο αλγόριθμος Naive Bayes έχουμε ένα παράδειγμα: Παρακάτω υπάρχει ένα σύνολο δεδομένων προπόνησης για τον καιρό και την αντίστοιχη target variable 'Play'. Τώρα πρέπει να ταξινομήσουμε εάν οι παίκτες θα παίξουν ή όχι με βάση τις καιρικές συνθήκες.
- Τα παρακάτω βήματα βοηθάνε στην εκτέλεση:
 - Βήμα 1: Μετατροπή του συνόλου δεδομένων σε πίνακα συχνοτήτων.

- Βήμα 2: Δημιουργήστε έναν πίνακα likelihood εντοπίζοντας τις πιθανότητες όπως η πιθανότητα συννεφιάς = 0,29 και η πιθανότητα playing = 0,64.
- Βήμα 3: Τώρα χρησιμοποιήστε την εξίσωση Naive Bayesian για τον υπολογισμό της posterior πιθανότητας για κάθε κλάση. Η κλάση με την υψηλότερη posterior πιθανότητα είναι το αποτέλεσμα της πρόβλεψης.

Λύση:

Οι παίκτες θα παίξουν αν ο καιρός είναι καλός.

Λύνεται με τη βοήθεια της posterior πιθανότητας. Δηλαδή:

- $P(\text{Yes} | \text{Sunny}) = P(\text{Sunny} | \text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$
- Με βάση τον πίνακα έχουμε, $P(\text{Sunny} | \text{Yes}) = 3/9 = 0.33$, $P(\text{Sunny}) = 5/14 = 0.36$, $P(\text{Yes}) = 9/14 = 0.64$
- Οπότε, $P(\text{Yes} | \text{Sunny}) = 0.33 * 0.64 / 0.36 = 0.60$, η οποία είναι η υψηλότερη πιθανότητα.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table			---	---
Weather	No	Yes	---	---
Overcast		4	4/14	0.29
Rainy	3	2	5/14	0.36
Sunny	2	3	5/14	0.36
All	5	9	---	---
	5/14	9/14	---	---
	0.36	0.64	---	---

Η λύση προέκυψε με βάση:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

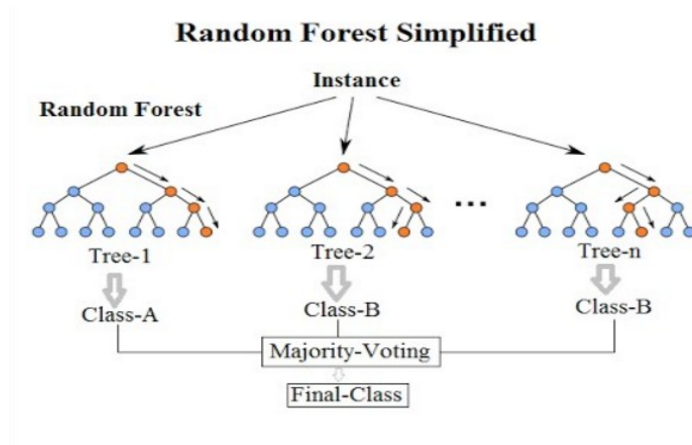
Όπου,

- $P(c|x)$ είναι η posterior πιθανότητα της κλάσης (c , target).
- $P(c)$ είναι η προηγούμενη πιθανότητα της κλάσης.
- $P(x|c)$ είναι το likelihood που είναι η πιθανότητα της δωσμένης κλάσης.
- $P(x)$ είναι η προηγούμενη πιθανότητα πρόβλεψης.

3) Random Forest αλγόριθμος:

- Όπως υποδηλώνει και το όνομα του, αποτελείται από μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο. Κάθε μεμονωμένο δέντρο στο random forest ξετυλίγει μια πρόβλεψη κλάσης και η κλάση με τις περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας.
- Είναι ένας supervised αλγόριθμος μάθησης που χρησιμοποιείται τόσο για την ταξινόμηση όσο και για την παλινδρόμηση (regression).
- Χειρίζεται τα missing values και διατηρεί την ακρίβεια ενός μεγάλου ποσοστού δεδομένων.
- Στην επιστήμη των δεδομένων ο λόγος που το random forest μοντέλο λειτουργεί τόσο καλά είναι: Η χαμηλή συσχέτιση μεταξύ των μοντέλων είναι το κλειδί. Με άλλα λόγια, τα μη συσχετισμένα μοντέλα μπορούν να παράγουν σύνολα προβλέψεων που είναι πιο ακριβείς από οποιαδήποτε από τις μεμονωμένες προβλέψεις. Ο λόγος για αυτό το θαυμάσιο αποτέλεσμα είναι ότι τα δέντρα προστατεύουν το ένα το άλλο από τα ατομικά τους λάθη (αρκεί να μην έχουν πάντα λάθος στην ίδια κατεύθυνση). Ενώ ορισμένα δέντρα μπορεί να είναι λάθος, πολλά άλλα δέντρα θα είναι σωστά, έτσι ως ομάδα τα δέντρα μπορούν να κινηθούν προς τη σωστή κατεύθυνση. Αυτές είναι οι προϋποθέσεις για την καλή απόδοση του random forest.
- **Γράφημα 4.**

Στο παρακάτω γράφημα παρουσιάζεται η οπτικοποίηση της λειτουργίας του Random Forest με βάση όλα τα παραπάνω που περιγράψαμε.



4) Logistic αλγόριθμος:

- Είναι ένας δυαδικός αλγόριθμος ταξινόμησης.
- Υποθέτει ότι οι μεταβλητές εισόδου είναι αριθμητικές και έχουν κατανομή Gaussian (καμπύλη καμπάνας). Αυτό το τελευταίο σημείο δεν είναι απαραίτητο να ισχύει, καθώς η logistic regression μπορεί ακόμα να επιτύχει καλά αποτελέσματα εάν τα δεδομένα σας δεν είναι Gaussian.
- Ο αλγόριθμος μαθαίνει έναν συντελεστή για κάθε τιμή εισόδου, οι οποίοι συνδυάζονται γραμμικά σε μια συνάρτηση παλινδρόμησης και μετατρέπονται χρησιμοποιώντας μια συνάρτηση logistic (σε σχήμα s). Η logistic regression είναι μια γρήγορη και απλή τεχνική, αλλά μπορεί να είναι πολύ αποτελεσματική σε ορισμένα προβλήματα.
- Η logistic regression υποστηρίζει μόνο δυαδικά προβλήματα ταξινόμησης, αν και η εφαρμογή Weka έχει προσαρμοστεί για να υποστηρίζει προβλήματα ταξινόμησης πολλαπλών κλάσεων.
- Logistic Regression model:

Οι συντελεστές (Beta τιμές b) του αλγορίθμου logistic regression πρέπει να εκτιμάται από τα training δεδομένα. Αυτό γίνεται χρησιμοποιώντας εκτίμηση μέγιστης πιθανότητας (maximum-likelihood estimation). Η εκτίμηση μέγιστης πιθανότητας είναι ένας κοινός αλγόριθμος μάθησης που χρησιμοποιείται από μια ποικιλία αλγορίθμων μηχανικής μάθησης. Οι καλύτεροι συντελεστές θα είχαν ως αποτέλεσμα ένα μοντέλο που θα προέβλεπε μια τιμή πολύ κοντά στο 1 (male) για την default κλάση και μια τιμή πολύ κοντά στο 0 (female) για την άλλη κλάση. Η διαίσθηση για τη μέγιστη πιθανότητα του logistic regression είναι ότι μια διαδικασία αναζήτησης αναζητά τιμές για τους συντελεστές (τιμές Beta) που ελαχιστοποιούν το σφάλμα στις πιθανότητες που προβλέπονται από το μοντέλο σε εκείνα τα δεδομένα (π.χ πιθανότητα 1 αν τα δεδομένα είναι η κύρια κλάση).

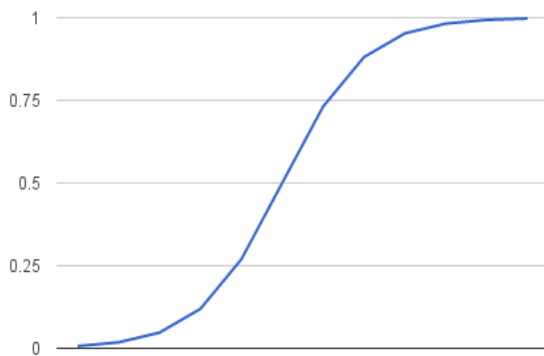
- **Κάνοντας προβλέψεις με Logistic Regression:**

Η πραγματοποίηση προβλέψεων με ένα μοντέλο logistic regression είναι πολύ απλή. Ας το πραγματοποιήσουμε με ένα παράδειγμα:

- Ας υποθέσουμε ότι έχουμε ένα μοντέλο που μπορεί να προβλέψει εάν ένα άτομο είναι άνδρας ή γυναίκα με βάση το ύψος του. Με ύψος 150 εκ. Είναι το άτομο αρσενικό ή θηλυκό.
- Έχουμε μάθει τους συντελεστές $b_0 = -100$ και $b_1 = 0,6$. Χρησιμοποιώντας την παρακάτω εξίσωση μπορούμε να υπολογίσουμε την πιθανότητα να είναι αρσενικό, $P(\text{male} | \text{height}=150)$. Θα χρησιμοποιήσουμε το $\text{EXP}()$ αντί για το e , γιατί αυτό είναι που μπορεί να χρησιμοποιηθεί αν κάνουμε το παράδειγμα στο υπολογιστικό μας φύλλο.
 - $y = e^{(b_0 + b_1 * X)} / (1 + e^{(b_0 + b_1 * X)})$
 - με την αντικατάσταση έχουμε, $y = \exp(-100 + 0.6 * 150) / (1 + \text{EXP}(-100 + 0.6 * X))$
 - αποτέλεσμα: $y = 0.0000453978687$
- Με την πιθανότητα να είναι σχεδόν μηδέν βγαίνει το συμπέρασμα ότι το άτομο είναι αρσενικό.

- **Γράφημα 5.**

Στο παρακάτω γράφημα παρουσιάζεται η γραφική απεικόνιση-συνάρτηση του Logistic Regression machine learning.



3.2 Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στα πειράματά μας προκειμένου να βρεθεί το βέλτιστο μοντέλο έχει να κάνει με το bank telemarketing data set του UCI με 45211 στοιχεία και 17 κατηγορίες. Οι κατηγορίες διακρίνονται τόσο σε nominal όσο και σε numeric. Τα δεδομένα σχετίζονται με εκστρατείες άμεσου marketing (τηλεφωνικές κλήσεις) ενός Πορτογαλικού τραπεζικού ιδρύματος. Σε ολόκληρο το σύνολο δεδομένων σεν παρατηρήθηκε κανένα missing value. Ο στόχος ταξινόμησης (target variable) είναι να προβλέψουμε εάν ο πελάτης θα εγγραφεί σε μια προθεσμιακή κατάθεση (μεταβλητή y). Οι κατηγορίες του data set είναι οι εξής καθώς και οι κλάσεις της κάθε κατηγορίας:

- **Age**, αποτυπώνει την ηλικία του κάθε πελάτη που κλήθηκε. Οι τιμές που εμφανίζονται σε αυτή την κατηγορία είναι αριθμητικές. Η ελάχιστη τιμή που παρατηρήθηκε είναι 18 και η μέγιστη 95.
- **Job**, αφορά την δουλειά του κάθε πελάτη. Το είδος των τιμών-κλάσεων αυτής της κατηγορίας είναι κατηγορηματικές και πιο συγκεκριμένα αυτές που συναντώνται στο data set είναι (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown). Σύμφωνα με το σύνολο δεδομένων ο αριθμός πελατών που αντιστοιχεί σε κάθε κλάση είναι Admin=5171, blue-collar=9732, entrepreneur=1487, housemaid=1240, management=9458, retired=2264, self-employed=1579, services=4154, student, 938, technician=7597, enunemployed=1303, unknown=288.
- **Marital**, είναι η κατηγορία που δείχνει αν ο πελάτης είναι χωρισμένος (divorced), παντρεμένος (married) και ανύπαντρος (single). Επομένως, το είδος και αυτής της κατηγορίας είναι κατηγορηματικό. Οι πελάτες που αντιστοιχούν σε κάθε κλάση είναι divorced=5207, married=27214 και single=12790.
- **Education**, είναι η κατηγορία που δείχνει το επίπεδο εκπαίδευσης του κάθε πελάτη. Οι τιμές-κλάσεις που λαμβάνονται σε αυτή την κατηγορία είναι κατηγορηματικές και πιο συγκεκριμένα αυτές είναι (unknown, secondary, primary, tertiary). Ο αριθμός των πελατών που αντιστοιχούν σε κάθε κλάση που προαναφέρθηκε είναι primary=6851, secondary=23202, tertiary=13301 και unknown=1857.
- **Default**, δείχνει αν ο πελάτης έχει πίστωση από προεπιλογή. Οι κλάσεις που συναντώνται σε αυτή την κατηγορία λαμβάνουν binary τιμές, δηλαδή yes ή no. Με βάση τους πελάτες που έχουν καταγραφεί σε κάθε κλάση από τις δύο yes=815 και no=44396 είναι αντιληπτό ότι οι δύο κλάσεις είναι αρκετά imbalanced.
- **Balance**, οι τιμές που εμφανίζονται σε αυτή την κατηγορία είναι αριθμητικές (numeric) και κυμαίνονται σε αρνητικούς αλλά και σε θετικούς δείκτες. Η ελάχιστη τιμή που έχει παρατηρηθεί είναι -8019 και η μέγιστη 102127.
- **Housing**, είναι η κατηγορία στην οποία γίνεται καταγραφή αν ο πελάτης έχει δάνειο σπιτιού ή όχι. Όπως είναι αντιληπτό οι τιμές που καταγράφονται είναι binary, δηλαδή διακρίνονται σε yes και no. Στην κλάση no παρατηρήθηκαν 20081 πελάτες και στην κλάση yes 25130 πελάτες.
- **Loan**, δείχνει αν κάποιος πελάτης έχει κάποιο προσωπικό δάνειο. Και σε αυτή την κατηγορία οι κλάσεις διακρίνονται σε yes και no, δηλαδή binary τιμές. Αυτό που παρατηρείται από την καταγραφή των πελατών σε κάθε κλάση, δηλαδή no=37967 και yes=7244 είναι ότι και σε αυτή την περίπτωση οι δύο κλάσεις είναι αρκετά imbalanced.
- **Contact**, δείχνει τον τύπο επικοινωνίας και οι τιμές που λαμβάνονται είναι κατηγορηματικές. Αυτές διακρίνονται σε cellullar, telephone και unknown. Οι καταγραφές για κάθε προαναφερθήσα κλάση είναι cellullar=29285, telephone=2906 και unknown=13020.
- **Day**, παρουσιάζει την τελευταία μέρα επαφής μέσα στον μήνα. Οι τιμές που παρουσιάζονται είναι αριθμητικές. Η κάθε μέρα του μήνα παρουσιάζεται με έναν αριθμό, για παράδειγμα η πρώτη μέρα του μήνα παρουσιάζεται με τον αριθμό 1, η δεύτερη με τον αριθμό 2 κλπ. Οι τιμές κυμαίνονται από 1 μέχρι και 31, όσες δηλαδή και οι μέρες ενός μήνα.
- **Month**, παρόμοια με την προηγούμενη κατηγορία, μόνο που αυτή παρουσιάζει τον τελευταίο μήνα που ήρθε σε επαφή ο πελάτης μέσα στο έτος. Οι κλάσεις-τιμές σε αυτή την κατηγορία, όπως είναι εύκολα αντιληπτό είναι κατηγορηματικές. Με άλλα λόγια οι τιμές που λαμβάνονται είναι όλοι οι μήνες του χρόνου και σε συνδυασμό με τις καταγραφές των πελατών έχουμε apr=2932, aug=6247, dec=214, feb=2649, jan=1403, jul=6895, jun=5341, mar=477, may=13766, nov=3970, oct=738 και sep=579. Οι περισσότερες καταγραφές έγιναν τον μήνα Μάιο.

- **Duration**, αυτό το attribute παρουσιάζει την τελευταία επαφή και μετρείται σε δευτερόλεπτα. Επομένως είναι αριθμητική. Η υψηλότερη τιμή που καταγράφηκε είναι 4918 δεύτερα και η χαμηλότερη 0. Σημαντική παρατήρηση είναι ότι αυτή η κατηγορία επηρεάζει σημαντικά το output target (δηλαδή αν το duration=0 τότε y='no'). Ωστόσο, η διάρκεια δεν είναι γνωστή πριν πραγματοποιηθεί μια κλήση. Επίσης, μετά το τέλος της κλήσης το y είναι προφανώς γνωστό. Επομένως, αυτή η εισαγωγή θα πρέπει να συμπεριλαμβάνεται μόνο για σκοπούς αναφοράς.
- **Campaign**, δείχνει τον αριθμό επαφών που πραγματοποιήθηκαν κατά τη διάρκεια αυτής της καμπάνιας και για αυτόν τον πελάτη. Η κατηγορία είναι αριθμητική. Η υψηλότερη τιμή που καταγράφηκε είναι 63 και η χαμηλότερη 1.
- **Pdays**, είναι η κατηγορία που παρουσιάζει τον αριθμό των ημερών που πέρασαν μετά την τελευταία επαφή του πελάτη από μια προηγούμενη καμπάνια. Και αυτό το attribute είναι αριθμητικό. Η χαμηλότερη τιμή που παρατηρείται είναι -1 και η υψηλότερη 871 και ο αριθμός 999 σημαίνει ότι ο πελάτης δεν είχε επικοινωνήσει προηγουμένως.
- **Previous**, δείχνει τον αριθμό επαφών που πραγματοποιήθηκαν πριν από αυτήν την καμπάνια και για αυτόν τον πελάτη. Οι τιμές που καταγράφονται και σε αυτήν την κατηγορία είναι αριθμητικές, με την υψηλότερη καταγραφή να σημειώνεται στην τιμή 275 και την χαμηλότερη στο 0.
- **Poutcome**, αυτή η κατηγορία παρουσιάζει το αποτέλεσμα της προηγούμενης καμπάνιας marketing. Το είδος της κατηγορίας είναι κατηγορηματικό. Οι κλάσεις που σημειώνονται είναι 4. Αυτές είναι η failure, η other, η success και η unknown. Οι τιμές ανά κλάση είναι failure=4901, other=1840, success=1511 και unknown=36959. Αυτό που παρατηρείται είναι ότι η τιμή της κλάσης unknown έχει τεράστια διαφορά από τις υπόλοιπες κάτι που καθιστά την κατηγορία αρκετά imbalanced.
- **Output**, δείχνει αν ο πελάτης έχει συνδρομή σε προθεσμιακή κατάθεση. Η κατηγορία είναι binary και οι τιμές που καταγράφονται είναι το 1 και το 2, όπου το 1 είναι το αντίστοιχο 'no' και το 2 το 'yes'. Την υψηλότερη καταγραφή την έχει η τιμή 1 με 39922 πελάτες και η οι υπόλοιποι 5289 είναι για την 2. Και σε αυτήν την κατηγορία αυτό που παρατηρείται είναι ότι είναι αρκετά imbalanced κάτι το οποίο θα δημιουργήσει αρκετά προβλήματα στη διαχείριση και επεξεργασία του data set.

3.3 Περιγραφή Αποτελεσμάτων

1) Imbalanced Data Set:

Τα παρακάτω 3 στατιστικά μέτρα, δηλαδή τα **Precision**, **Recall** και **F-measure** μας βοήθησαν στο να αξιολογήσουμε τα αποτελέσματα που προέκυψαν για κάθε αλγόριθμο. Το στατιστικό μέτρο του F-measure υπολογίζεται βασιζόμενο στο Precision και στο Recall. Ο υπολογισμός τους γίνεται με βάση τα:

$$\begin{aligned} \rightarrow \text{Precision: } & \frac{TP}{(TP+FP)} \\ \rightarrow \text{Recall: } & \frac{TP}{(TP+FN)} \\ \rightarrow \text{F-measure: } & \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \end{aligned}$$

TP: αριθμός των True Positives

FP: αριθμός των False Positives

FN: αριθμός των False Negatives

Η τιμή Precision (ακρίβεια) ορίζεται ως το κλάσμα των στοιχείων που ταξινομούνται σωστά ως θετικά από όλα τα στοιχεία που ο αλγόριθμος ταξινομείται ως θετικός, ενώ η τιμή Recall (ανάκληση) είναι το κλάσμα των

στοιχείων που ταξινομούνται σωστά ως θετικά από όλα τα θετικά στοιχεία. Στην περίπτωση πολλαπλών κλάσεων, κάθε κλάση έχει μία αντίστοιχη τιμή Precision και Recall, στην οποία ένα “true positive” είναι ένα στοιχείο το οποίο προβλέπεται να είναι στο I και είναι πραγματικά σε αυτό και ένα “true negative” ένα στοιχείο που προβλέπεται να μην είναι στο I και δεν είναι σε αυτό. Έτσι, με αυτόν τον ορισμό του Precision και του Recall, κάθε κλάση μπορεί να έχει τη δική της τιμή F-measure κάνοντας τον ίδιο υπολογισμό με τη δυαδική περίπτωση.

Το weighted F-measure είναι ένας μέσος όρος των τιμών F-measure των κλάσεων σταθμισμένος από την αναλογία του αριθμού των στοιχείων σε κάθε κλάση. Η τιμή του F-measure στην ουσία μας βοηθάει στο να βγάλουμε πιο ασφαλή συμπεράσματα για την αποτελεσματικότητα των αλγορίθμων όταν οι τιμές του Precision και του Recall δεν είναι αρκετά σαφείς για να κρίνουμε κατά πόσο ο αλγόριθμος που εξετάζεται είναι αποδοτικός ή όχι.

- 1o setting:

Το πρώτο setting έγινε χωρίς feature selection. Σε αυτό όπως και στα υπόλοιπα εφαρμόστηκε 10 fold cross validation. Οι τιμές των αποτελεσμάτων των κλάσεων όπως θα δούμε έχουν αρκετά μεγάλη απόκλιση μεταξύ τους, λόγω των imbalance δεδομένων. Τα αποτελέσματα μας για κάθε αλγόριθμο εμφανίζονται στον παρακάτω πίνακα:

Models	Precision	Recall	F-measure	Attributes
J48	0.933	0.96	0.95	-----
	0.61	0.48	0.54	
Weighted avg	0.9	0.9	0.9	
Naive Bayes	0.94	0.93	0.93	-----
	0.49	0.53	0.51	
Weighted avg	0.88	0.88	0.88	
Random Forest	0.93	0.97	0.95	-----
	0.64	0.44	0.52	
Weighted avg	0.9	0.91	0.9	
Logistic	0.92	0.975	0.95	-----
	0.65	0.35	0.45	
Weighted avg	0.89	0.9	0.89	

- 2o setting:

Στο δεύτερο setting εφαρμόζουμε feature selection και πιο συγκεκριμένα Attribute Selection Classifier με CfsSubset Evaluator και μέθοδο Best First. Και εδώ οι τιμές έχουν αρκετά μεγάλη απόκλιση μεταξύ τους. Τα αποτελέσματα καθώς και τα attributes που επιλέχθηκαν ως τα κατάλληλα για κάθε αλγόριθμο ξεχωριστά βρίσκονται παρακάτω:

Models	Precision	Recall	F-measure	Attributes
J48	0.92	0.97	0.95	3, 7, 8, 12, 16 (5)
	0.63	0.36	0.46	
Weighted avg	0.89	0.9	0.89	
Naive Bayes	0.92	0.97	0.94	3, 7, 8, 12, 16 (5)
	0.6	0.36	0.45	
Weighted avg	0.88	0.9	0.89	
Random Forest	0.92	0.95	0.93	3, 7, 8, 12, 16 (5)
	0.48	0.35	0.4	
Weighted avg	0.87	0.88	0.87	

Logistic	0.92	0.977	0.95	3, 7, 8, 12, 16 (5)
	0.64	0.31	0.42	
Weighted avg	0.88	0.9	0.88	

- 3o setting:

Στο τρίτο setting όπως και στο προηγούμενο εφαρμόζουμε feature selection και πιο συγκεκριμένα Attribute Selection Classifier με WrapperSubset Evaluator και μέθοδο Best First. Τα αποτελέσματα και τα attributes που επιλέχθηκαν βρίσκονται παρακάτω:

Models	Precision	Recall	F-measure	Attributes
J48	0.94	0.96	0.95	5, 9, 10, 11, 12, 14,
	0.63	0.52	0.57	16 (7)
Weighted avg	0.9	0.91	0.9	
Naive Bayes	0.93	0.96	0.94	9, 11, 12, 13, 16 (5)
	0.58	0.45	0.51	
Weighted avg	0.89	0.9	0.89	
Random Forest	0.9	0.99	0.94	5, 13, 16 (3)
	0.66	0.18	0.28	
Weighted avg	0.87	0.89	0.87	
Logistic	0.92	0.976	0.95	1, 2, 3, 5, 7, 8, 9,
	0.64	0.33	0.44	11, 12, 13, 14, 15,
Weighted avg	0.88	0.9	0.89	16 (13)

2) Balanced Data Set:

- 1o setting:

Στο πρώτο setting όπως και στα υπόλοιπα αυτό που μπορεί να παρατηρηθεί είναι ότι οι τιμές των αποτελεσμάτων είναι παρόμοιες, κάτι το οποίο συμβαίνει λόγω της μετατροπής των δεδομένων σε balanced. Αυτό συνέβη μέσω του φίλτρου Spread Subsample ο οποίος μας βοήθησε να μετατρέψουμε τον αριθμό των δεδομένων της τελευταίας κλάσης έτσι ώστε να μην υπάρχει αυτή η χαοτική διαφορά μεταξύ των instances των 2 χαρακτηριστικών. Επομένως, παρακάτω εμφανίζονται τα αποτελέσματα για τον κάθε αλγόριθμο ξεχωριστά:

Models	Precision	Recall	F-measure	Attributes
J48	0.89	0.81	0.85	-----
	0.83	0.9	0.86	
Weighted avg	0.86	0.86	0.85	
Naive Bayes	0.81	0.76	0.79	-----
	0.78	0.83	0.8	
Weighted avg	0.8	0.99	0.79	
Random Forest	0.89	0.83	0.86	-----
	0.84	0.9	0.87	
Weighted avg	0.87	0.87	0.87	
Logistic	0.83	0.849	0.84	-----
	0.85	0.82	0.83	
Weighted avg	0.84	0.84	0.84	

- 2o setting:

Στο δεύτερο setting εφαρμόζουμε feature selection και πιο συγκεκριμένα Attribute Selection Classifier με CfsSubset Evaluator και μέθοδο Best First. Επομένως, εκτός από τις τιμές των αποτελεσμάτων, θα παρουσιαστούν το σύνολο αλλά και τα συγκεκριμένα attributes που επιλέχθηκαν:

Models	Precision	Recall	F-measure	Attributes
J48	0.81	0.77	0.79	3, 6, 9, 10, 12,16 (6)
	0.78	0.82	0.8	
Weighted avg	0.79	0.79	0.79	
Naive Bayes	0.71	0.9	0.79	3, 6, 9, 10, 12, 16 (6)
	0.86	0.63	0.73	
Weighted avg	0.78	0.76	0.76	
Random Forest	0.8	0.78	0.79	3, 6, 9, 10, 12,16 (6)
	0.78	0.81	0.79	
Weighted avg	0.79	0.79	0.79	
Logistic	0.77	0.838	0.8	3, 6, 9, 10, 12, 16 (6)
	0.82	0.74	0.78	
Weighted avg	0.79	0.79	0.79	

- 3o setting:

Στο τρίτο και τελευταίο setting εφαρμόσαμε feature selection, με Attribute Selection Classifier, η αξιολόγηση έγινε με τον WrapperSubset Evaluator και μέθοδο Best First. Άρα, παρακάτω διατυπώνονται τα αποτελέσματα και η επιλογή των attributes για κάθε αλγόριθμο ξεχωριστά:

Models	Precision	Recall	F-measure	Attributes
J48	0.89	0.82	0.85	3, 4, 7, 8, 9, 10, 11,
	0.83	0.9	0.86	12, 14 (9)
Weighted avg	0.86	0.86	0.86	
Naive Bayes	0.81	0.83	0.82	1, 3, 4, 7, 11, 12,
	0.82	0.8	0.81	13, 16 (8)
Weighted avg	0.82	0.82	0.81	
Random Forest	0.9	0.84	0.87	1, 3, 4, 6, 7, 9, 10,
	0.85	0.91	0.88	11, 12, 13, 14,15(12)
Weighted avg	0.88	0.87	0.87	
Logistic	0.83	0.850	0.84	2, 4, 7, 8, 9, 11, 12,
	0.85	0.82	0.83	13, 15, 16 (10)
Weighted avg	0.84	0.84	0.84	

Κεφάλαιο 4°:

4.1 Επιλογή Βέλτιστου Μοντέλου

Με βάση των παραπάνω αποτελεσμάτων καλούμαστε να επιλέξουμε ποιο είναι το καλύτερο δυνατό μοντέλο. Αυτό που ξεχωρίζει λόγω των τιμών του Precision, Recall και F-measure είναι από την κατηγορία των balanced δεδομένων και πιο συγκεκριμένα το **Balanced Random Forest** με **Wrapper Evaluator** και μέθοδο **Best First**. Δηλαδή με τιμές:

Models	Precision	Recall	F-measure
Random Forest	0.9	0.84	0.87
	0.85	0.91	0.88
Weighted avg	0.88	0.87	0.87

4.2 Κανόνες Μοντέλου με χρήση Αλγορίθμου PART

Με τη χρήση του παρακάτω πίνακα παρουσιάζονται όλοι οι κανόνες που προέκυψαν από τον αλγόριθμο PART του προτεινόμενου μοντέλου. Συνολικά όλοι οι κανόνες που προέκυψαν είναι 265 και βρίσκονται όλοι μέσα στον πίνακα. Ο πίνακας είναι 5 στηλών και 53 σειρών. Κάθε κελί του πίνακα παρουσιάζει έναν ξεχωριστό κανόνα που προέκυψε. Όσον αφορά τον PART, χτίζει ένα μερικό C4.5 δέντρο απόφασης σε κάθε επανάληψη και κάνει το καλύτερο “φύλλο” σε έναν κανόνα.

V12 <= 206 AND V16 = success AND V12 > 118: 2 (223.0/17.0)	V12 <= 206 AND V12 <= 77 AND V16 = unknown AND V11 = may: 1 (222.0)	V12 <= 206 AND V12 <= 77 AND V16 = unknown AND V11 = jun: 1 (169.0)	V12 <= 206 AND V11 = may AND V7 = yes AND V10 <= 13: 1 (322.0)	V12 <= 206 AND V12 <= 88 AND V16 = unknown AND V11 = jul: 1 (178.0/3.0)
V12 > 433 AND V12 > 639 AND V9 = telephone: 2 (101.0/8.0)	V12 <= 206 AND V11 = may AND V9 = unknown AND V16 = unknown: 1 (278.0/2.0)	V12 > 433 AND V12 > 640 AND V9 = cellular AND V16 = success AND V7 = no: 2 (63.0)	V12 > 433 AND V12 > 640 AND V9 = cellular AND V11 = may AND V16 = unknown: 2 (156.0/9.0)	V12 > 433 AND V12 > 640 AND V9 = cellular AND V11 = feb: 2 (84.0/5.0)
V12 > 408 AND V12 > 640 AND V9 = cellular AND V11 = may AND V7 = yes AND V1 <= 52: 2 (81.0/3.0)	V12 <= 204 AND V12 <= 77 AND V16 = unknown AND V11 = aug: 1 (121.0/1.0)	V12 > 433 AND V12 > 640 AND V9 = cellular AND V11 = aug AND V5 = no: 2 (230.0/18.0)	V12 <= 204 AND V11 = jul AND V14 <= 24 AND V10 > 8: 1 (235.0/5.0)	V12 > 433 AND V12 > 703 AND V11 = apr AND V12 > 836 AND V10 > 6: 2 (74.0)
V12 > 392 AND V12 > 640 AND V11 = jul: 2 (303.0/35.0)	V12 <= 184 AND V12 <= 88 AND V16 = other: 1 (55.0/2.0)	V12 > 255 AND V16 = success AND V7 = no AND V11 = aug: 2 (48.0)	V12 > 392 AND V16 = success AND V11 = apr: 2 (22.0)	V12 > 364 AND V16 = success AND V11 = feb: 2 (24.0/1.0)
V12 > 392 AND V12 > 640 AND V11 = nov AND V16 = unknown: 2 (119.0/15.0)	V12 <= 184 AND V11 = may AND V10 <= 19: 1 (110.0/2.0)	V12 > 392 AND V16 = success AND V11 = oct: 2 (21.0)	V12 > 392 AND V12 > 641 AND V11 = jun: 2 (195.0/28.0)	V12 <= 159 AND V11 = aug AND V14 <= -1 AND V5 = no AND V1 > 30 AND V1 <= 59: 1 (230.0/7.0)
V12 > 392 AND V9 = telephone AND V11 = apr: 2 (13.0)	V12 > 408 AND V9 = telephone AND V16 = success: 2 (17.0)	V12 > 433 AND V9 = cellular AND V11 = sep: 2 (50.0/3.0)	V12 > 440 AND V9 = cellular AND V11 = jun: 2 (44.0/1.0)	V16 = success AND V12 > 162 AND V7 = no AND V11 = sep: 2 (46.0/1.0)
V12 > 408 AND	V12 > 408 AND	V12 > 440 AND	V12 > 440 AND	V12 <= 130 AND

V9 = telephone AND V11 = oct: 2 (8.0)	V9 = cellular AND V11 = dec: 2 (26.0)	V12 > 641 AND V16 = failure AND V14 <= 355: 2 (32.0/1.0)	V10 > 29 AND V11 = apr: 2 (42.0)	V12 <= 62 AND V16 = unknown AND V10 <= 25: 1 (106.0/1.0)
V16 = success AND V12 > 162 AND V11 = jul AND V14 > 79: 2 (49.0/1.0)	V12 > 440 AND V11 = mar AND V7 = no: 2 (22.0)	V16 = success AND V7 = no AND V11 = may: 2 (45.0/1.0)	V12 > 440 AND V12 > 521 AND V11 = oct AND V1 <= 54: 2 (20.0)	V12 > 440 AND V12 > 521 AND V11 = aug AND V9 = cellular AND V16 = unknown AND V5 = no: 2 (61.0/11.0)
V16 = success AND V11 = jun AND V7 = no: 2 (36.0)	V12 > 440 AND V12 > 641 AND V16 = other AND V7 = no: 2 (11.0)	V9 = unknown AND V12 <= 478 AND V11 = may: 1 (389.0/12.0)	V12 <= 206 AND V9 = unknown AND V11 = jun: 1 (230.0/2.0)	V12 <= 130 AND V11 = jan AND V10 > 14: 1 (47.0/1.0)
V16 = success AND V11 = apr AND V12 > 99: 2 (37.0/1.0)	V1 > 60 AND V12 > 188 AND V11 = apr AND V14 <= 180: 2 (28.0)	V16 = success AND V11 = mar: 2 (37.0/3.0)	V16 = success AND V11 = feb AND V12 > 109: 2 (36.0)	V12 > 365 AND V14 > 369 AND V11 = may: 2 (12.0)
V12 > 365 AND V1 <= 29 AND V11 = jun: 2 (29.0/5.0)	V12 > 365 AND V1 <= 28 AND V11 = may AND V10 <= 27 AND V12 > 461 AND V12 > 480 AND V1 > 23 AND V7 = yes AND V12 > 635: 2 (21.0/1.0)	V12 > 365 AND V1 > 59 AND V16 = unknown AND V11 = may: 2 (12.0/1.0)	V12 > 365 AND V11 = mar AND V10 <= 28: 2 (20.0)	V12 <= 130 AND V11 = nov AND V10 > 16: 1 (149.0/2.0)
V16 = success AND V11 = oct: 2 (47.0/4.0)	V1 > 60 AND V11 = nov: 2 (32.0/3.0)	V12 > 434 AND V11 = oct AND V12 <= 511: 2 (11.0)	V12 > 434 AND V11 = aug AND V16 = unknown AND V9 = cellular AND V7 = no AND V5 = no AND V12 > 438: 2 (51.0/14.0)	V12 > 521 AND V11 = aug AND V5 = no: 2 (12.0)
V12 > 521 AND V11 = may AND V10 > 12 AND V9 = cellular AND V7 = yes AND V16 = unknown: 2 (35.0/4.0)	V12 > 521 AND V11 = jul AND V9 = telephone AND V1 > 42: 2 (6.0)	V7 = no AND V9 = unknown AND V11 = jul AND V12 <= 492: 1 (15.0)	V7 = no AND V11 = mar: 2 (133.0/15.0)	V12 > 364 AND V12 > 641 AND V16 = unknown AND V11 = may AND V10 > 12 AND V5 = no AND V10 <= 26 AND V7 = yes: 2 (113.0/18.0)
V7 = no AND V1 <= 29 AND V11 = apr: 2 (69.0/5.0)	V16 = success AND V11 = aug AND V12 > 101: 2 (38.0/1.0)	V7 = no AND V16 = success AND V11 = nov AND V10 <= 18: 2 (24.0)	V7 = no AND V9 = unknown AND V11 = jun AND V12 <= 482: 1 (91.0/6.0)	V7 = no AND V11 = sep AND V10 > 5 AND V10 <= 22: 2 (73.0/2.0)
V7 = no AND V11 = dec AND V12 > 196: 2 (45.0/4.0)	V12 > 364 AND V1 <= 29 AND V11 = may AND V16 = unknown AND V9 = cellular: 2 (39.0/6.0)	V7 = no AND V11 = oct AND V12 <= 93 AND V16 = failure: 1 (5.0)	V7 = no AND V11 = oct AND V10 > 18 AND V9 = cellular AND V10 <= 28 AND V10 <= 24: 2 (31.0/1.0)	V12 > 364 AND V14 > 372 AND V12 <= 640: 2 (31.0/1.0)
V12 > 410 AND V1 <= 28 AND V16 = unknown AND V9 = cellular: 2 (45.0/5.0)	V1 > 60 AND V11 = jan: 2 (20.0/2.0)	V7 = yes AND V12 <= 434 AND V11 = sep AND V12 > 181: 2 (34.0)	V7 = yes AND V12 <= 434 AND V10 <= 21 AND V11 = apr AND V10 > 15 AND V9 = cellular AND V10 <= 18 AND V16 = unknown: 1 (43.0/1.0)	V7 = yes AND V12 <= 434 AND V10 <= 21 AND V11 = apr AND V10 > 15 AND V14 > 323: 1 (31.0)
V7 = yes AND V12 <= 434 AND V10 <= 21 AND V11 = jul AND V16 = unknown AND V9 = cellular AND V10 > 6: 1 (58.0/6.0)	V7 = no AND V11 = apr AND V12 > 123 AND V16 = unknown: 2 (144.0/20.0)	V12 > 405 AND V16 = success: 2 (31.0/7.0)	V12 > 408 AND V10 > 26 AND V12 > 691: 2 (83.0/11.0)	V1 > 60 AND V11 = jul AND V12 > 236: 2 (15.0)
V7 = yes AND V12 <= 405 AND V10 <= 21 AND V11 = nov AND V10 > 16 AND V12 <= 273: 1 (81.0)	V11 = mar AND V16 = unknown AND V12 > 97: 2 (31.0/3.0)	V7 = yes AND V12 <= 405 AND V10 <= 20 AND V16 = success AND V11 = may AND V14 <= 203: 2 (9.0)	V7 = yes AND V12 <= 405 AND V10 <= 20 AND V11 = may AND V10 > 4 AND V12 <= 284: 1 (107.0/5.0)	V1 <= 25 AND V7 = no AND V11 = jul: 2 (14.0/2.0)
V16 = success AND V11 = jan: 2 (16.0/2.0)	V11 = oct AND V10 > 15 AND	V12 <= 124 AND V9 = cellular AND	V11 = oct AND V10 > 15: 2 (122.0/22.0)	V16 = success AND V11 = jun AND

	V7 = yes AND V9 = telephone: 2 (11.0)	V10 <= 7 AND V11 = feb: 1 (63.0/3.0)		V12 > 254: 2 (8.0)
V16 = success AND V11 = nov AND V14 <= 104: 2 (8.0)	V14 > 9 AND V12 > 158 AND V7 = no AND V11 = aug: 2 (54.0/8.0)	V11 = mar AND V16 = failure AND V14 <= 328: 2 (10.0/1.0)	V11 = sep AND V9 = cellular AND V1 > 44: 2 (14.0)	V12 > 405 AND V9 = cellular AND V11 = may AND V10 > 5 AND V5 = no AND V1 <= 56 AND V1 <= 40 AND V16 = failure AND V7 = yes: 2 (25.0/4.0)
V11 = dec AND V7 = yes: 2 (13.0/1.0)	V1 <= 25 AND V11 = aug: 2 (21.0/3.0)	V11 = mar AND V16 = failure: 1 (2.0)	V11 = aug AND V1 > 60 AND V16 = unknown AND V7 = no AND V10 > 11 AND V9 = cellular AND V1 > 61: 2 (16.0/6.0)	V11 = aug AND V14 <= 73 AND V1 > 60 AND V7 = no AND V9 = telephone AND V12 > 187: 2 (7.0)
V11 = aug AND V14 <= 73 AND V9 = cellular AND V12 <= 247 AND V5 = no AND V10 <= 24 AND V7 = no: 1 (127.0/23.0)	V7 = yes AND V12 <= 434 AND V14 > 367 AND V11 = aug AND V12 > 189: 2 (7.0)	V7 = yes AND V12 <= 434 AND V9 = unknown AND V11 = jun AND V10 > 2: 1 (60.0/1.0)	V11 = jun AND V9 = cellular AND V12 > 145 AND V1 <= 58 AND V16 = unknown AND V7 = no: 2 (100.0/12.0)	V12 > 434 AND V9 = cellular AND V5 = yes AND V11 = jul AND V10 <= 20: 2 (6.0/1.0)
V11 = jan AND V10 > 27 AND V10 <= 29 AND V9 = cellular AND V7 = yes AND V1 <= 57: 1 (24.0)	V12 > 434 AND V9 = cellular AND V5 = yes: 2 (14.0/3.0)	V11 = aug AND V16 = unknown AND V9 = cellular AND V12 <= 247 AND V7 = no: 1 (34.0)	V12 > 521 AND V9 = cellular AND V11 = may AND V16 = unknown AND V10 > 6 AND V10 <= 17 AND V7 = yes: 2 (18.0/6.0)	V12 > 521 AND V9 = cellular AND V11 = may AND V16 = other: 2 (11.0/1.0)
V11 = jun AND V16 = other AND V7 = no AND V9 = cellular AND V1 > 35: 2 (9.0)	V1 <= 30 AND V11 = feb AND V10 > 8: 2 (39.0/2.0)	V5 = yes AND V9 = cellular AND V11 = aug: 1 (6.0/1.0)	V12 > 434 AND V9 = cellular AND V14 > 254 AND V11 = may AND V7 = yes AND V16 = failure AND V1 > 45: 1 (6.0)	V12 > 434 AND V9 = cellular AND V11 = feb AND V7 = no AND V16 = unknown AND V10 > 5: 2 (11.0)
V12 > 521 AND V9 = cellular AND V11 = may AND V16 = unknown AND V7 = no AND V10 > 17: 2 (5.0)	V12 > 521 AND V9 = cellular AND V11 = may AND V10 <= 6: 2 (9.0)	V11 = aug AND V16 = other AND V14 > 114: 2 (6.0)	V11 = aug AND V16 = unknown AND V9 = telephone: 1 (18.0/5.0)	V11 = aug AND V9 = cellular AND V16 = failure AND V12 > 101: 2 (24.0/4.0)
V11 = aug AND V9 = cellular AND V1 > 59 AND V1 <= 61 AND V1 <= 60 AND V7 = no: 2 (6.0/1.0)	V11 = aug AND V9 = cellular AND V10 > 7 AND V16 = unknown AND V12 <= 347 AND V1 <= 60: 1 (104.0/17.0)	V5 = no AND V7 = yes AND V12 <= 434 AND V10 <= 20 AND V11 = apr AND V10 > 16 AND V9 = cellular AND V10 > 18: 1 (46.0/6.0)	V5 = yes: 1 (25.0/6.0)	V1 > 65 AND V11 = feb AND V10 > 12: 2 (17.0)
V12 <= 185 AND V9 = telephone AND V12 <= 159 AND V11 = apr: 1 (6.0)	V7 = yes AND V10 <= 7 AND V11 = nov AND V10 > 2: 2 (7.0)	V7 = yes AND V10 <= 7 AND V11 = feb AND V10 > 4: 1 (18.0/1.0)	V11 = jun AND V16 = other AND V7 = no AND V9 = cellular AND V1 <= 33: 2 (6.0/1.0)	V11 = jun AND V16 = failure AND V12 > 203 AND V1 <= 62: 2 (17.0)
V7 = yes AND V9 = telephone AND V11 = jul AND V10 <= 24: 1 (9.0)	V10 > 27 AND V11 = jan AND V10 <= 29 AND V9 = cellular AND V7 = no AND V16 = unknown: 1 (35.0/9.0)	V7 = yes AND V10 <= 7 AND V11 = feb AND V16 = failure AND V10 <= 2: 1 (14.0/1.0)	V7 = yes AND V10 <= 7 AND V1 > 54 AND V1 <= 58: 2 (9.0)	V7 = yes AND V10 <= 7 AND V11 = apr AND V12 <= 582 AND V1 > 33: 1 (20.0)
V7 = no AND V10 <= 16 AND V11 = nov AND V12 > 125: 2 (62.0/4.0)	V11 = nov AND V12 <= 371 AND V10 > 16 AND V9 = cellular AND V10 <= 23: 1 (111.0/15.0)	V9 = telephone AND V11 = nov AND V7 = no: 1 (13.0/1.0)	V10 > 28 AND V11 = jan AND V10 > 29: 1 (23.0)	V11 = jan AND V9 = cellular AND V16 = unknown AND V7 = no AND V10 <= 14: 2 (20.0/1.0)
V11 = nov AND V16 = other AND V7 = yes AND V10 > 16 AND V10 <= 20: 2 (9.0/1.0)	V11 = nov AND V16 = failure AND V12 > 86 AND V7 = no AND V10 > 20: 2 (5.0)	V11 = nov AND V16 = other AND V7 = yes: 1 (6.0/1.0)	V11 = nov AND V16 = unknown AND V10 > 20: 2 (28.0/5.0)	V16 = success AND V11 = may: 2 (6.0)

V11 = feb AND V10 > 9 AND V1 <= 52 AND V9 = cellular AND V7 = yes AND V10 > 10: 2 (28.0)	V12 <= 123 AND V11 = may AND V16 = unknown AND V1 > 29: 1 (15.0)	V7 = no AND V11 = may AND V10 <= 4: 2 (16.0/1.0)	V7 = yes AND V11 = feb AND V14 <= 279 AND V14 <= 246 AND V16 = unknown: 1 (41.0/12.0)	V7 = no AND V11 = may AND V9 = cellular AND V16 = other AND V1 <= 53 AND V14 <= 355: 2 (15.0/1.0)
V11 = dec AND V12 > 169: 1 (4.0)	V11 = dec AND V16 = unknown AND V1 <= 58: 2 (3.0)	V11 = dec AND V12 <= 152: 1 (5.0/1.0)	V11 = feb AND V10 > 9 AND V9 = cellular AND V16 = unknown: 2 (43.0/8.0)	V1 <= 35 AND V11 = nov AND V9 = cellular AND V16 = failure AND V7 = yes: 2 (8.0/1.0)
V11 = sep AND V9 = cellular AND V16 = unknown AND V7 = no AND V12 > 178: 2 (8.0/1.0)	V11 = jun AND V16 = failure AND V10 <= 17 AND V7 = yes: 1 (6.0/2.0)	V11 = jun AND V10 > 18: 1 (30.0/9.0)	V11 = jun AND V10 > 3 AND V16 = unknown AND V7 = yes AND V9 = cellular AND V12 > 128: 2 (10.0)	V11 = jun AND V16 = other AND V9 = cellular AND V14 <= 85: 2 (3.0)
V11 = jun AND V16 = failure: 2 (15.0/6.0)	V14 > 378 AND V11 = jul AND V10 > 6: 2 (12.0)	V11 = jun AND V16 = other AND V9 = cellular: 1 (3.0)	V11 = jun AND V16 = unknown AND V9 = cellular AND V7 = no: 2 (29.0/12.0)	V12 <= 134 AND V11 = jul: 1 (17.0/2.0)
V11 = jun AND V16 = unknown AND V9 = cellular AND V1 > 31: 2 (13.0/4.0)	V7 = yes AND V11 = aug AND V16 = unknown AND V1 <= 60 AND V12 <= 458: 1 (24.0/5.0)	V11 = jun AND V16 = unknown AND V9 = unknown AND V10 > 3: 2 (40.0/13.0)	V7 = yes AND V11 = jun: 1 (23.0/5.0)	V11 = jun: 2 (16.0/4.0)
V11 = jul AND V16 = other: 2 (13.0/2.0)	V11 = jul AND V10 <= 6 AND V14 <= 354: 2 (36.0/5.0)	V11 = jul AND V16 = failure AND V10 > 17: 2 (9.0)	V11 = jul AND V12 <= 341 AND V16 = unknown AND V9 = cellular: 1 (103.0/13.0)	V9 = telephone AND V11 = jul AND V7 = no: 1 (21.0/4.0)
V12 <= 99 AND V11 = aug AND V16 = failure: 1 (8.0/1.0)	V10 > 20 AND V9 = cellular AND V12 > 141 AND V11 = apr AND V7 = yes AND V16 = unknown AND V10 <= 29: 2 (10.0)	V10 > 21 AND V9 = cellular AND V12 > 167 AND V16 = failure: 2 (41.0/5.0)	V16 = failure AND V11 = may AND V1 <= 46 AND V7 = no AND V10 > 12: 1 (9.0)	V7 = yes AND V11 = aug AND V14 <= 86: 2 (10.0)
V7 = yes AND V11 = jan: 2 (10.0/3.0)	V7 = yes AND V11 = oct AND V12 <= 134: 1 (5.0)	V7 = yes AND V11 = feb AND V14 <= 279 AND V14 > 246: 1 (11.0)	V7 = yes AND V11 = feb AND V14 <= 224 AND V14 <= 198 AND V12 <= 256: 1 (7.0/2.0)	V11 = oct AND V10 <= 14 AND V7 = no AND V10 > 4 AND V10 <= 6: 2 (7.0)
V11 = feb AND V10 > 9: 2 (28.0/5.0)	V1 > 35 AND V11 = oct AND V16 = unknown: 1 (19.0/8.0)	V11 = oct AND V1 <= 39 AND V7 = no AND V16 = unknown AND V10 <= 13: 2 (10.0/2.0)	V11 = jan AND V9 = cellular AND V16 = unknown AND V10 <= 26 AND V10 <= 16: 2 (4.0/1.0)	V11 = jan AND V9 = telephone: 1 (4.0)
V11 = may AND V7 = yes AND V16 = failure: 1 (38.0/11.0)	V11 = jul AND V9 = telephone AND V1 <= 46: 2 (3.0/1.0)	V11 = apr AND V9 = telephone AND V14 <= 163: 1 (4.0)	V11 = jul AND V9 = cellular AND V12 > 525 AND V7 = yes AND V1 <= 53: 2 (27.0/5.0)	V7 = yes AND V11 = feb AND V16 = failure AND V14 <= 198: 2 (4.0)
V7 = yes AND V11 = jul AND V9 = cellular AND V14 <= 29 AND V12 <= 501 AND V1 > 30: 1 (24.0/10.0)	V7 = yes AND V11 = jul AND V1 > 31: 1 (10.0)	V11 = jul AND V7 = no AND V16 = unknown AND V12 <= 512: 1 (59.0/24.0)	V11 = jul AND V16 = unknown: 2 (43.0/12.0)	V11 = jul AND V7 = no AND V12 > 232: 2 (5.0/1.0)
V11 = jul AND V7 = no: 1 (4.0)	V11 = sep AND V16 = other AND V1 > 27: 2 (9.0/2.0)	V11 = sep AND V16 = unknown AND V7 = no: 1 (7.0/2.0)	V11 = sep AND V10 <= 7: 1 (7.0)	V11 = sep AND V16 = unknown: 2 (3.0)
V11 = oct AND V16 = failure AND V1 <= 46: 2 (8.0/1.0)	V11 = oct AND V1 > 29: 1 (8.0)	V11 = apr AND V9 = cellular AND V16 = failure AND V1 > 33 AND V10 > 15: 1 (16.0/1.0)	V11 = may AND V16 = unknown AND V7 = no AND V9 = cellular: 2 (69.0/27.0)	V11 = may AND V16 = other: 1 (25.0/8.0)
V11 = may AND V16 = failure: 2 (14.0/4.0)	V11 = may AND V7 = yes AND V9 = telephone AND V12 <= 408: 1 (5.0)	V11 = may AND V9 = cellular AND V10 > 4 AND V10 > 11 AND V1 <= 38: 2 (46.0/17.0)	V11 = may AND V9 = cellular AND V10 > 4: 1 (61.0/11.0)	V11 = apr AND V9 = cellular AND V16 = other AND V10 > 22: 2 (11.0/1.0)
V11 = apr AND V9 = cellular AND	V11 = apr AND V9 = cellular AND	V11 = feb AND V7 = yes AND	V11 = feb AND V7 = yes: 2 (6.0)	V11 = feb AND V16 = other AND

V16 = other AND V7 = yes AND V14 > 166 AND V10 > 14: 2 (7.0/1.0)	V16 = other: 1 (14.0/1.0)	V16 = failure AND V14 <= 245: 1 (4.0)		V10 <= 4 AND V14 <= 195: 2 (5.0)
V11 = feb AND V16 = other: 1 (8.0)	V11 = may AND V9 = unknown AND V7 = yes: 1 (135.0/63.0)	V11 = may AND V9 = unknown AND V12 <= 678: 1 (16.0/4.0)	V11 = may: 2 (27.0/4.0)	V11 = nov AND V16 = failure AND V9 = cellular AND V14 > 93 AND V7 = no AND V14 <= 164: 1 (4.0)
V11 = nov AND V16 = unknown AND V10 <= 16 AND V12 > 128: 2 (10.0)	V11 = nov AND V16 = success: 2 (4.0)	V11 = nov AND V9 = unknown AND V1 > 36: 1 (5.0)	V9 = unknown: 2 (9.0/1.0)	V11 = nov AND V9 = cellular AND V16 = unknown AND V7 = no AND V12 <= 592 AND V12 <= 538 AND V10 > 12: 1 (20.0/7.0)
V11 = jan AND V16 = other AND V12 > 192 AND V10 <= 22: 2 (3.0/1.0)	V11 = aug AND V9 = cellular AND V7 = no: 2 (82.0/39.0)	V11 = nov AND V9 = cellular AND V1 > 52: 2 (7.0)	V11 = aug: 1 (6.0)	V16 = other AND V11 = jan AND V1 <= 40: 1 (2.0)
V16 = other: 2 (8.0)	V11 = nov AND V9 = cellular AND V1 <= 30: 2 (7.0/1.0)	V11 = nov AND V9 = cellular AND V7 = no AND V12 > 586: 1 (7.0)	V11 = nov AND V7 = yes AND V9 = cellular AND V14 <= 108 AND V10 <= 18 AND V10 <= 17 AND V12 > 530: 1 (4.0)	V11 = nov AND V7 = yes AND V9 = cellular AND V14 <= 108 AND V1 <= 46: 1 (16.0/6.0)
V11 = nov AND V9 = cellular AND V14 <= 175: 2 (7.0)	V11 = nov: 1 (6.0)	V11 = apr AND V9 = cellular AND V16 = failure AND V7 = no AND V10 <= 14 AND V14 > 220: 1 (4.0)	V11 = apr AND V9 = cellular AND V16 = failure AND V10 <= 7: 2 (4.0)	V11 = apr AND V9 = cellular AND V16 = failure AND V7 = no AND V1 <= 60: 2 (3.0)
V11 = apr AND V9 = cellular AND V16 = failure: 1 (32.0/14.0)	V11 = apr AND V9 = cellular AND V16 = unknown AND V7 = no AND V1 > 36: 1 (18.0/5.0)	V11 = apr AND V16 = unknown AND V7 = yes AND V12 > 121: 2 (75.0/32.0)	V7 = yes AND V16 = unknown: 1 (14.0/1.0)	V11 = apr AND V16 = unknown: 2 (8.0/1.0)
V7 = no AND V11 = feb AND V16 = failure AND V10 > 2 AND V12 > 232: 2 (7.0)	V7 = no AND V11 = feb AND V16 = failure AND V10 <= 4: 1 (8.0/1.0)	V11 = feb AND V16 = unknown AND V9 = telephone AND V12 <= 421 AND V10 <= 4: 1 (10.0/2.0)	V9 = cellular AND V11 = feb: 1 (105.0/52.0)	V9 = telephone: 2 (9.0/1.0)
V11 = jan AND V14 <= 125 AND V10 <= 26: 1 (3.0/1.0)	V11 = jan AND V14 <= 125: 2 (3.0)	V11 = apr: 1 (2.0)	V7 = no: 1 (7.0/2.0)	: 2 (3.0)

Κεφάλαιο 5°:

Πίνακας Εργασιών

Στον παρακάτω πίνακα αποτειπώνεται η παρουσίαση των σημαντικότερων στοιχείων της κάθε εργασίας που πρόκειται να συγκριθεί. Οι εργασίες που πρόκειται να συγκριθούν και παρουσιάζονται στον πίνακα παρακάτω είναι 8, συμπεριλαμβανομένου και της εργασίας μας. Οι εργασίες που αναλύονται στον παρακάτω πίνακα είναι οι [5], [8], [9], [11], [12], [13], [15] και τέλος η έρευνα μας. Ο πίνακας όπως μπορεί να δει κανείς αποτελείται από 6 στήλες και 8 γραμμές (6x8). Σε κάθε στήλη εμφανίζονται οι κατηγορίες με βάση τις οποίες γίνεται η ανάλυση της κάθε εργασίας, ενώ μέσω των γραμμών χωρίζεται η μία εργασία από την άλλη. Οι στήλες του πίνακα είναι οι εξής:

- **Αναφορά**, στην οποία παρουσιάζεται το άρθρο που μελετάτε.
- **Μεθοδολογία**, στην οποία παρουσιάζονται ο βασικός αλγόριθμος-οι που εφαρμόστηκε ή όποια άλλη μέθοδος χρησιμοποιήθηκε.
- **Σύνολο δεδομένων**, στην οποία παρουσιάζεται το data set που χρησιμοποιήθηκε και πιο συγκεκριμένα ο αριθμός των δειγμάτων και των κατηγοριών (attributes) του κάθε data set.
- **Τρόπος αξιολόγησης**, εδώ γίνεται ο διαχωρισμός των δεδομένων σε αυτά που χρησιμοποιήθηκαν για εκπαίδευση (training) και σε αυτά για δοκιμή (testing).
- **Όλες οι μετρικές**, παρουσιάζονται όλα τα στατιστικά μέτρα, βάση των οποίων έγινε η αξιολόγηση της κάθε εργασίας.
- **Καινοτόμα στοιχεία**, παρουσίαση των καινοτόμων χαρακτηριστικών της κάθε εργασίας.

Αναφορά	Μεθοδολογία	Σύνολο δεδομένων	Τρόπος αξιολόγησης	Όλες οι μετρικές	Καινοτόμα στοιχεία
[5]	_ MLPNN (νευρωνικά δίκτυα) _ TAN (δέντρο Naive Bayes) _ Logistic Regression _ C5.0 (δέντρο απόφασης)	_ 45211 δείγματα _ 17 attributes	Διαχωρισμός του συνόλου δεδομένων σε υποσύνολα/δείγματα: _ το 70% για training _ το 30% για testing	<u>3 στατιστικά μέτρα:</u> _ accuracy _ sensitivity _ specificity τόσο για training όσο και για testing	—
[8]	_ ordinary SVM _ Adaboost SVM	_ 41188 δείγματα _ 20 attributes	_ μέσω του adaboost ο οποίος εκπαιδεύει τον βασικό classifier με συνεχείς επαναλήψεις _ όσο καλύτερα εκπαιδευτεί ο classifier τόσο καλύτερα θα συνδυαστεί για τη δημιουργία του καλύτερου δυνατού μοντέλου	_ accuracy _ sensitivity _ αριθμός επαναλήψεων (iteration)	—
[9]	_ Rotation Forest (PCA) J48 (σε συνδυασμό με	_ 45211 δείγματα (μετά το SMOTE τα δείγματα της	_ 70% για train _ 30% για test	_ accuracy _ sensitivity _ specificity	—

	SMOTE μέθοδο) _ MLP νευρωνικό δίκτυο _ RBF νευρωνικό δίκτυο _ Bayesian Networking _ Naive Bayes	minority class έγιναν 10578 από 5289) _ 17 attributes			
[11]	_ feature selection (Information Gain, Chi-square) _ σύγκριση μέσω του αλγορίθμου Naive Bayes	_ 45211 δείγματα _ 17 attributes	_ 5 fold cross validation (εκ των οποίων τα 4 είναι για train και το 1 για test	_ precision _ recall _ F-measure	_____
[12]	_ προτεινόμενο μοντέλο: KSVM(αλγόριθμος ταξινόμησης) + ANN(νευρωνικό δίκτυο) _ συγκρίνεται με: KSVM, ANN, SVM	_ 41188 δείγματα _ 20 attributes	_ Αρχικά αφού εισαχθεί το data set, κάνει training τα data χρησιμοποιώντας τον SVM αλγόριθμο και στη συνέχεια τα κάνει testing _ για την ταξινόμηση χρησιμοποιεί τον αλγόριθμο KSVM _ διαχωρισμός data set σε training data και validation data	_ accuracy _ sensitivity _ specificity _ precision	_____
[13]	_ KNN-ANN (νευρωνικά δίκτυα) _ Random Forest (δέντρο απόφασης) _ Logistic Regression _ Naive Bayes _ SVM	_ 45211 δείγματα _ 17 attributes	_ 10 fold cross validation _ train set (με 36168 επαφές) + test set (με 9043 επαφές)	_ accuracy για train και test sets _ CAP curve _ variance (%)	_____
[15]	_ k-nn (k-nearest neighbor) _ C4.5 δέντρο απόφασης _ Naive Bayes	_ 45211 δείγματα _ 17 attributes	3 πειράματα με: _ 60-40% train-test set _ 70-30% train-test set _ 80-20% train-test set	_ accuracy _ precision _ F-measure	_____
Έρευνα	_ C4.5 _ Naive Bayes _ Random Forest _ Logistic Regression	_ 45211 δείγματα _ 17 attributes	_ 10 fold cross validation _ feature selection (με Cfs subset Evaluator και wrapper Evaluator με μέθοδο Best First) _ Spread	_ precision _ recall _ F-measure _ αριθμός attributes που επιλέγονται από feature selection	_____

			SubSample φίλτρο για δημιουργία balanced κλάσεων		
--	--	--	--	--	--

Κεφάλαιο 6°:

6.1 Πειράματα για Προτεινόμενο Μοντέλο

Αρχικά, για να κάνουμε μία πρώτη δοκιμή του data set και των αποτελεσμάτων του αρχίσαμε να το τρέχουμε με κάποιους αλγόριθμους. Οι αλγόριθμοι με του οποίους το δοκιμάσαμε ήταν: J48, NaiveBayes, Random Forest, Logistic με 10 fold cross validation.

Στη συνέχεια, εφαρμόσαμε feature selection με CfsSubset Evaluator και μέθοδο Best First. Με αυτόν τον τρόπο θέλαμε να δούμε ποια ήταν τα καλύτερα attributes για τον κάθε αλγόριθμο ξεχωριστά, έτσι ώστε να έχουμε καλύτερα αποτελέσματα.

Ξαναεφαρμόσαμε feature selection με ίδια μέθοδο Best First, αλλά με διαφορετικό Evaluator, τον WrapperEvaluator για να βρούμε τα καλύτερα δυνατά attributes.

Με τις παραπάνω εφαρμογές βλέπαμε ότι τα αποτελέσματα μας δεν ήταν και τόσο ικανοποιητικά και αυτό συνέβαινε διότι το attribute output με τις κλάσεις yes και no ήταν αρκετά imbalanced. Προκειμένου να βελτιωθούν οι τιμές των αποτελεσμάτων μας (Precision, Recall, F-measure) μετατρέψαμε τις κλάσεις από imbalanced σε balanced με τη βοήθεια του φίλτρου Spread SubSample.

Αυτό το φίλτρο κάνει τα instances των δύο κλάσεων ισόποσα, μετατρέποντας την μεγαλύτερη κλάση(σε αριθμό instances) στο μέγεθος της μικρότερης. Αυτό είχε σαν αποτέλεσμα να δημιουργηθούν δύο ίσες κλάσεις με 5289 instances η κάθε μία.

Έπειτα, εφαρμόστηκε η ίδια διαδικασία με πάνω, δηλαδή η εφαρμογή των 5 παραπάνω αλγορίθμων, αρχικά χωρίς feature selection και στη συνέχεια με feature selection με Cfs Subset Evaluator και Wrapper Evaluator αντίστοιχα, αλλά αυτή τη φορά σε balanced δεδομένα.

6.2 Συγκρίσεις Εργασιών

Αρχικά, οι αξιολογήσεις και τα αποτελέσματα και των δύο εργασιών προέκυψαν από το ίδιο σύνολο δεδομένων. Στην εργασία [5] οι συγγραφείς περιγράψανε μια μεθοδολογία για το direct τραπεζικό marketing. Αυτή η μεθοδολογία αφορούσε την αξιολόγηση και σύγκριση της απόδοσης ταξινόμησης τεσσάρων διαφορετικών μοντέλων data mining τεχνικών MLPNN (Multilayer Perception νευρωνικό δίκτυο), TAN (δέντρο Naive Bayes), LR (Logistic Regression), C5.0 (δέντρο απόφασης) στο σύνολο δεδομένων direct marketing. Οι επιδόσεις ταξινόμησης των τεσσάρων μοντέλων χρησιμοποιούν τρία στατιστικά μέτρα: accuracy, sensitivity και specificity. Αφού έγινε μία προεργασία του συνόλου δεδομένων με feature selection προκειμένου να αφαιρεθούν τα λιγότερο σημαντικά attributes, το σύνολο δεδομένων χωρίστηκε σε training και testing samples με αναλογία 70% - 30% αντίστοιχα. Με βάση τα παραπάνω subsets προκύπτουν τα confusion matrices του κάθε αλγορίθμου ξεχωριστά για τις κλάσεις yes και no, μέσω των οποίων υπολογίζονται τα classification accuracy, sensitivity και specificity για κάθε sample του κάθε αλγορίθμου ξεχωριστά. Τα αποτελέσματα που προέκυψαν έδειξαν ότι το MLPNN ήταν το μοντέλο με την μεγαλύτερη ακρίβεια με 90.92% σε training δείγματα και 90.49% σε testing δείγματα. Ταυτόχρονα, η ανάλυση sensitivity έχει το υψηλότερο ποσοστό στο μοντέλο C5.0 με 76.75% σε training δείγματα και για testing δείγματα το μοντέλο Logistic Regression με τιμή 65.53%. Τέλος, το μοντέλο με την υψηλότερη τιμή specificity σε training δείγματα αλλά και testing δείγματα είναι το C5.0 με 94.92% και 93.23% αντίστοιχα.

Συγκρίνοντας τις δύο εργασίες, παρατηρούμε ότι και στις δύο εφαρμόστηκε feature selection για την επιλογή των καλύτερων για κάθε μοντέλο attributes. Στην εργασία [5] η απόδοση ταξινόμησης των μοντέλων έγινε με την βοήθεια του υπολογισμού τριών στατιστικών μέτρων: accuracy, sensitivity, specificity, ενώ στην προτεινόμενη με τη βοήθεια των τιμών Precision, Recall και F-measure. Παρόλο που τα δεδομένα της target class ήταν αρκετά imbalanced στην εργασία [5] δεν υπέστησαν κάποια επεξεργασία σε αντίθεση με την προτεινόμενη. Όσον αφορά τα αποτελέσματα της εργασίας [5], το μοντέλο C5.0 έχει επιτύχει ελαφρώς

καλύτερη απόδοση από τα υπόλοιπα, ενώ από την άλλη μεριά στην προτεινόμενη εργασία το μοντέλο που ξεχώρισε ήταν αυτό του Balanced Random Forest με wrapper evaluator. Κλείνοντας, το κοινό στοιχείο που παρατηρείται και στις δύο εργασίες είναι ότι βέλτιστο μοντέλο, και στις δύο, προκύπτει από αλγόριθμο δέντρου απόφασης.

Αρχικά, το ίδιο σύνολο δεδομένων συναντάται και στις δύο εργασίες. Στην εργασία [9] ο συγγραφέας προτείνει μια μέθοδο βασισμένη στον αλγόριθμο SMOTE σε συνδυασμό με τον Rotation Forest(PCA)-J48 για την πρόβλεψη των εξαιρετικά μη ισορροπημένων δεδομένων. Η απόδοση του παραπάνω μοντέλου αξιολογείται και συγκρίνεται με το δέντρο απόφασης, το Rotation Forest, το MLP νευρωνικό δίκτυο, το RBF νευρωνικό δίκτυο, το Bayesian network και το Naive Bayes. Η απόδοση της ταξινόμησης μετριέται από τα: accuracy, sensitivity και specificity. Αυτά τα δύο σύνολα δεδομένων (δηλαδή το κανονικό και αυτό που έχει υποστεί επεξεργασία) εκπαιδεύονται (trained) και δοκιμάζονται (tested) σε 14 αλγορίθμους μηχανικής μάθησης. Όσον αφορά τα αποτελέσματα, για τα κανονικά δεδομένα, το μοντέλο που δίνει την καλύτερη τιμή sensitivity είναι ο Rotation Forest(PCA) ADTree με 67.73%, την καλύτερη τιμή specificity το μοντέλο J48 με 93.19% και την καλύτερη τιμή accuracy ο Rotation Forest(PCA) J48 με τιμή 90.05%. Ενώ, για τα balanced δεδομένα(SMOTE), το μοντέλο που δίνει την καλύτερη τιμή sensitivity είναι το BayesNet με τιμή 94.48%, για το specificity το Rotation Forest(PCA) J48 με 93.17% και τέλος για το accuracy πάλι το Rotation Forest(PCA) J48 με τιμή αυτή τη φορά 90.81%. Από τα πειράματα αποδείχθηκε ότι η SMOTE μέθοδο λειτουργεί αποτελεσματικά για αυτά τα δεδομένα.

Συγκρίνοντας και τις δύο εργασίες παρατηρούμε ότι λόγω των imbalanced δεδομένων, εφαρμόστηκαν, διαφορετικές στην καθεμία, μέθοδοι μετατροπής imbalanced δεδομένων σε balanced. Το SMOTE για την εργασία [9] και το Spread SubSample για την προτεινόμενη. Παρότι στην εργασία μας εφαρμόσαμε feature selection κάτι τέτοιο δεν παρατηρήθηκε στην εργασία [9]. Τέλος, όσον αφορά τα αποτελέσματα, αποδείχθηκε και στις δύο εργασίες ότι οι καλύτερες τιμές δόθηκαν από τα balanced δεδομένα και πιο συγκεκριμένα το μοντέλο που ξεχώρισε για την εργασία [9] ήταν το Rotation Forest(PCA) J48, ενώ για την εργασία μας ένα εξίσου δέντρο απόφασης του Random Forest είναι το επικρατέστερο.

Ξεκινώντας, και οι δύο έρευνες έγιναν στο ίδιο σύνολο δεδομένων. Στην εργασία [15] οι συγγραφείς περιγράψανε μια μεθοδολογία εξόρυξης δεδομένων. Τα προγράμματα εξόρυξης δεδομένων που χρησιμοποιήθηκαν για τη διαδικασία αυτή ήταν: το R, το Knime, το RapidMiner και το Weka. Οι αλγόριθμοι που εφαρμόστηκαν είναι: k-nearest neighbor (k-nn), Naive Bayes και C4.5 decision tree. Για την αξιολόγηση-δοκιμή των επιδόσεων των μοντέλων ταξινόμησης χρησιμοποιήθηκαν τα κριτήρια Precision, Accuracy και F-measure, τα οποία προκύπτουν από το confusion matrix του κάθε αλγορίθμου ξεχωριστά. Στην έρευνα έγιναν τρία πειράματα, όπου σε κάθε ένα από αυτά το data set διαιρέθηκε σε training και test sets με αναλογίες 60-40%, 75-25% και 80-20% αντίστοιχα. Όσον αφορά τα αποτελέσματα, σε κάθε πείραμα που έγινε με οποιοδήποτε ποσοστό holdout separation (60%, 75%, 80%) στο Weka, το μοντέλο που ξεχώρισε είναι του αλγορίθμου C4.5 με βάση τις τιμές Accuracy = 0.900, Precision = 0.940, F-measure = 0.944.

Συγκρίνοντας τις δύο εργασίες συμπαίρνουμε ότι και στις δύο χρησιμοποιήθηκε, για την ανάλυση των δεδομένων, το εργαλείο του Weka. Επίσης, και στις δύο εργασίες για τη δοκιμή των επιδόσεων των μοντέλων χρησιμοποιήθηκαν οι τιμές των Accuracy, Precision και F-measure. Εφαρμόστηκαν δύο κοινοί αλγόριθμοι ταξινόμησης όπως του Naive Bayes και του δέντρου απόφασης C4.5. Και στις δύο εργασίες έγιναν πειράματα σε κανονικά δείγματα, δηλαδή που δεν είχαν υποστεί κάποια επεξεργασία πιο πριν (imbalanced data). Αυτό που παρατηρείται στην εργασία [15] είναι ότι τα πειράματα και τα αποτελέσματα προέκυψαν από υποσύνολα του data set, κάτι που δεν παρατηρείται στην εργασία μας όπου η μόνη μέθοδος αξιολόγησης ήταν το 10-fold cross validation. Στα αποτελέσματα, ο Naive Bayes (και στις δύο εργασίες) είχε αρκετά υψηλές τιμές παρόλ' αυτά τα μοντέλα που ξεχώρισαν και στις δύο ήταν αλγόριθμοι δέντρων απόφασης και πιο συγκεκριμένα ο C4.5 για την εργασία [15] και ο Random Forest για την προτεινόμενη.

Περνώντας στην επόμενη σύγκριση το σύνολο δεδομένων που χρησιμοποιείται στην εργασία [12] είναι διαφορετικό από αυτό της έρευνας μας με 41188 instances και 20 attributes, παρόλ' αυτά είναι αρκετά παρόμοια διότι η κύρια διαφορά τους είναι κάποια επιπλέον attributes που προστίθενται στο data set της εργασίας [12]. Τα πειράματα της εργασίας [12] έγιναν μέσω του R studio. Το θέμα της έρευνας αφορά το churn prediction των πελατών στις τράπεζες. Σύμφωνα με τον συγγραφέα του άρθρου, για την παραπάνω πρόβλεψη χρησιμοποιήθηκε μια υβριδική προσέγγιση δύο αλγορίθμων του KSVM (αλγόριθμος μηχανικής μάθησης για ταξινόμηση) + ANN (νευρωνικό δίκτυο). Τα αποτελέσματα του παραπάνω υβριδικού μοντέλου συγκρίθηκαν με τους αλγορίθμους SVM, KSVM, ANN ξεχωριστά προκειμένου να αποδειχθεί αν είναι κατάλληλο για το churn prediction σε σχέση

με τα υπόλοιπα μοντέλα. Η αξιολόγηση της επίδοσης του κάθε αλγορίθμου ξεχωριστά αλλά και του προτεινόμενου μοντέλου (KSVM + ANN) υπολογίζεται από τα κριτήρια των sensitivity, specificity, precision και accuracy. Αυτό που παρατηρείται από τα αποτελέσματα είναι ότι το προτεινόμενο υβριδικό μοντέλο βγάζει καλύτερα αποτελέσματα από κάθε άλλο αλγόριθμο ξεχωριστά, με τις τιμές sensitivity = 0.9897, specificity = 0.5000, precision = 0.9897 και accuracy = 0.9798 να το καθιστούν το βέλτιστο μοντέλο για την πρόβλεψη churn σε αυτό το data set.

Η ανάλυση του συνόλου δεδομένων έγινε μέσω του Weka σε αντίθεση με την εργασία [12]. Η κύρια διαφορά των δύο εργασιών έχει να κάνει με τους αλγορίθμους που χρησιμοποιήθηκαν για την επιλογή του βέλτιστου μοντέλου πρόβλεψης, δηλαδή η μία χρησιμοποιεί συνδυασμό ενός αλγορίθμου μηχανικής μάθησης (KSVM) με νευρωνικό δίκτυο (ANN), σε αντίθεση με την άλλη που είναι ένα δέντρο απόφασης (Random Forest). Συγκρίνοντας τα αποτελέσματα του προτεινόμενου μοντέλου της εργασίας [12] με αυτά που προκύπτουν από την προτεινόμενη εργασία παρατηρούμε ότι το μοντέλο της πρώτης είναι ελαφρώς πιο ακριβές από το μοντέλο της δεύτερης λόγω των καλύτερων τιμών accuracy και precision.

Στη συνέχεια, σύμφωνα με τον συγγραφέα του άρθρου [11] χρησιμοποιούνται feature selection τεχνικές για να προσδιοριστούν τα σημαντικά attributes του data set, και πιο συγκεκριμένα οι μέθοδοι feature selection που εφαρμόζονται είναι οι Information Gain και Chi-square. Οι παραπάνω μέθοδοι συγκρίνονται χρησιμοποιώντας τον αλγόριθμο Naive Bayes μέσω του εργαλείου Weka. Η απόδοση του ταξινομητή συγκρίνεται με βάση τις τιμές των precision, recall και F-measure. Τα πειράματα επικυρώνονται με 5-fold cross validation, όπου ο ταξινομητής εκπαιδεύεται (train) σε 4 και δοκιμάζεται (test) σε 1. Δοκιμάστηκαν 4 μεγέθη χαρακτηριστικών στο πείραμα (5, 8, 10, 15). Το καλύτερο δυνατό αποτέλεσμα λήφθηκε από τη μέθοδο Information Gain με τα εξής 5 χαρακτηριστικά: duration, outcome, month, pdays και contact, λόγω της της υψηλότερης τιμής F-measure (0.883) σε σύγκριση με τις υπόλοιπες 3 κατηγορίες χαρακτηριστικών (8, 10, 15).

Λαμβάνοντας υπόψη τη μεθοδολογία της εργασίας [11] και της προτεινόμενης εργασίας παρατηρούμε ότι και στις δύο ακολουθείται μία παρόμοια προσέγγιση. Αρχικά, και στις δύο εφαρμόζεται feature selection παρόλα αυτά με διαφορετικές μεθόδους η κάθε μία ξεχωριστά. Παρόλο που στην έρευνά μας εφαρμόστηκαν 5 διαφορετικοί αλγόριθμοι για τη σύγκριση, ένας από αυτούς είναι κοινός με την [11], ο Naive Bayes. Η επικύρωση των πειραμάτων γίνεται στην μία με 5 fold cross validation και στην άλλη με 5 fold cross validation. Το σύνολο δεδομένων και για τις δύο εργασίες είναι το ίδιο. Σύμφωνα με τα αποτελέσματα που προέκυψαν και από τις δύο εργασίες παρατηρούμε ότι και στις δύο εργασίες όπως ήταν αναμενόμενο, στα κανονικά αποτελέσματα (δηλαδή αυτά που δεν υπέστησαν κάποια επεξεργασία) οι τιμές είναι παρόμοιες. Από την άλλη μεριά, λόγω των imbalance δεδομένων του data set παρατηρείται ότι οι τιμές της εργασίας [11] είναι αρκετά χαμηλές στην μία κλάση (πχ positive class: precision=0.472, recall=0.424, f-measure=0.447 / negative class: precision=0.926, recall=0.938, f-measure=0.932), σε αντίθεση με τις τιμές των balance δεδομένων της έρευνάς μας. Το συμπέρασμα είναι ότι χωρίς την μετατροπή των δεδομένων (από imbalanced σε balanced) αυτό που παρατηρείται στα αποτελέσματα και των δύο εργασιών είναι ότι οι τιμές των κριτηρίων στην μία κλάση είναι αρκετά χαμηλές σε αντίθεση με τις τιμές στην άλλη κλάση.

Σύμφωνα με το συντάκτη του άρθρου της εργασίας [13] για την αποτελεσματικότητα της καμπάνιας του bank telemarketing χρησιμοποιήθηκαν οι εξής αλγόριθμοι μηχανικής μάθησης: οι KNN και ANN (νευρωνικά δίκτυα), ο Random Forest (δέντρο απόφασης), ο Logistic Regression, ο Naive Bayes και ο SVM. Για σκοπούς αξιολόγησης το data set χωρίζεται σε training set (με 36168 επαφές) και σε test set (με 9043 επαφές). Η επικύρωση των πειραμάτων γίνεται με 10 fold cross validation. Από τα αποτελέσματα προκύπτει ότι ο ANN και ο Random Forest είναι πιο αποτελεσματικοί, βάση των τιμών ακρίβειας σε train και test set αντίστοιχα. Πιο συγκεκριμένα, οι τιμές των παραπάνω αλγορίθμων είναι για Random Forest 89.574 / 90233 για train και test set αντίστοιχα και για ANN 90.007 / 91.210 για train και test set αντίστοιχα.

Συγκρίνοντας τις μεθοδολογίες της εργασίας [13] με την προτεινόμενη εργασία παρατηρούμε ότι πολλοί από τους αλγόριθμους που χρησιμοποιούνται στις έρευνες είναι ίδιοι, όπως ο Logistic Regression, ο Naive Bayes και ο Random Forest. Τα πειράματα και στις δύο εργασίες γίνονται στο ίδιο data set. Στην εργασία [13] δεν έχει εφαρμοστεί κάποιο φιλτράρισμα των δεδομένων με αποτέλεσμα τα πειράμα να γίνονται σε imbalance δεδομένα. Η αξιολόγηση στην εργασία [13] γίνεται με βάση τις τιμές accuracy του κάθε αλγορίθμου σε αντίθεση με την εργασία μας. Και στις δύο η επικύρωση των πειραμάτων γίνεται με 10 fold cross validation. Λαμβάνοντας υπόψη τα αποτελέσματα και των δύο εργασιών παρατηρούμε ότι Naive Bayes και Logistic Regression δεν είναι αρκετά ακριβείς και στις δύο σε αντίθεση με τον Random Forest ο οποίος είναι αρκετά ακριβής (διότι

εξειδικεύεται και σε μεγάλα δεδομένα). Παρ' όλα αυτά αυτός που ξεχωρίζει για την εργασία [13] είναι το νευρωνικό δίκτυο ANN, βάση των αποτελεσμάτων.

Κλείνοντας με την τελευταία σύγκριση, σύμφωνα με τον συγγραφέα της εργασίας [8] η μέθοδος μέσω της οποίας θα λειτουργούσε αποτελεσματικά το direct marketing της τράπεζας είναι η SVM. Σύμφωνα με τη μεθοδολογία της εργασίας η επίδοση και η ακρίβεια του SVM βελτιώνεται με τη χρησιμοποίηση του Ensemble method. Σε αυτή την έρευνα συγκρίνεται το απλό SVM με το Adaboost SVM. Ο Adaboost αλγόριθμος εκπαιδεύει τον βασικό classifier με συνεχή επαναλήψεις. Εάν ο classifier εκπαιδευτεί όσο πρέπει, τότε ο classifier συνδυάζεται για να σχηματίσει το μοντέλο που εμφανίζει την καλύτερη απόδοση. Τα αποτελέσματα της προτεινόμενης μεθόδου αξιολογούνται βάσει του βαθμού ακρίβειας (accuracy) και ευαισθησίας (sensitivity). Με βάση τα αποτελέσματα προκύπτει ότι το adaboost SVM είναι το καλύτερο δυνατό μοντέλο με 30 επαναλήψεις και πιο συγκεκριμένα με τιμές ακρίβειας και ευαισθησίας με 95.07% και 91.65% αντίστοιχα.

Συγκρίνοντας τις δύο εργασίες παρατηρείται ότι παρότι τα data set που χρησιμοποιείται στην καθεμία είναι διαφορετικά είναι αρκετά παρόμοια μεταξύ τους ως προς τα attributes που χρησιμοποιούνται. Οι αλγόριθμοι που εφαρμόζονται είναι τελείως διαφορετικοί και γενικότερα η μεθοδολογία των δύο εργασιών διαφέρει αρκετά η μία από την άλλη. Από τα αποτελέσματα των δύο εργασιών παρατηρείται ότι το μοντέλο της εργασίας [8] (Adaboost SVM) σε αντίθεση με το βέλτιστο μοντέλο της έρευνας μας (Random Forest) έχει αρκετά υψηλή τιμή ακρίβειας (95.07% έναντι 90.54%).

Κεφάλαιο 7°:

7.1 Συμπεράσματα

Σκοπός της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός μοντέλου που οδηγήσει στην αποτελεσματικότητα του τραπεζικού marketing, μέσω των συνεχών πειραμάτων πάνω σε ένα σύνολο δεδομένων τραπεζικού marketing.

Μέσω της μεθοδολογίας που ακολουθήθηκε για επιλογή του βέλτιστου μοντέλου συμπαιράναμε ότι τα καλύτερα δυνατά αποτελέσματα προήλθαν από τα balanced δεδομένα και πιο συγκεκριμένα με feature selection και με evaluators. Αυτό που ξεχωρίζει όπως είδαμε είναι το μοντέλο Random Forest με feature selection με wrapper Evaluator και μέθοδο Best First.

Στο πρώτο μέρος της πτυχιακής εργασίας γίνεται μια σύντομη αναφορά, μέσω της εισαγωγής στο θέμα της εργασίας καθώς και στη δομή που θα ακολουθηθεί στο κυρίως κομμάτι της πτυχιακής. Έπειτα γίνεται μια μικρή αναφορά σε κάποιες από τις έννοιες που θα συναντηθούν κατά τη διάρκεια της εργασίας. Τέλος, γίνεται αναφορά στην έννοια του τραπεζικού marketing.

Στη συνέχεια, γίνεται η ανάλυση-περίληψη κάποιων εργασιών που επιλέχθηκαν που έχουν να κάνουν με τραπεζικό marketing αλλά και με αποτελέσματα που προκύπτουν από σύνολα δεδομένων τραπεζικού marketing και η ανάλυση αυτών. Μερικά από αυτά τα σύνολα δεδομένων είναι ίδια και με αυτό που χρησιμοποιήσαμε και στην εργασία μας. Από την ανάλυση των εργασιών παρατηρούμε ότι σε πολλές από αυτές, διάφορους τρόπους με τους οποίους επεξεργάστηκαν και αξιολόγησαν τα δεδομένα, είτε με την εφαρμογή διαφορετικών αλγορίθμων, είτε με διάφορων τεχνικών οι οποίες οδήγησαν σε διαφορετικά συμπεράσματα και αποτελέσματα από τις υπόλοιπες. Χρησιμοποιήθηκαν πολλές data mining και machine learning τεχνικές. Επίσης, παραθέτονται και οι σχετικοί πίνακες αποτελεσμάτων της κάθε εργασίας αλλά και γραφήματα. Το συμπέρασμα είναι ότι και τα δύο είδη τεχνικών που χρησιμοποιήθηκαν έδωσαν αρκετά καλά και οφέλιμα αποτελέσματα, τα οποία συζητήθηκαν και αξιολογήθηκαν αλλά και συγκρίθηκαν με την έρευνα μας.

Το αμέσως επόμενο στάδιο στην πτυχιακή μας είναι η αξιολόγηση του κάθε αλγορίθμου ξεχωριστά που χρησιμοποιήθηκε για την ανάλυση του συνόλου δεδομένων μας αλλά και η λεπτομερής περιγραφή του συνόλου δεδομένων που χρησιμοποιήθηκε για τη δημιουργία του προτεινόμενου μοντέλου μας. Αυτό το στάδιο είναι αρκετά σημαντικό διότι δίνει στον αναγνώστη τη δυνατότητα να κατανοήσει την λειτουργία του κάθε αλγόριθμου και πως αυτός αλληλεπιδρά με τα δεδομένα καθώς και πόσο αποτελεσματικός είναι στην λειτουργία του. Η ανάλυση του κάθε αλγορίθμου έγινε με τη βοήθεια παραδειγμάτων και διαγραμμάτων προκειμένου να γίνει πιο κατανοητή η λειτουργία.

Στο στάδιο περιγραφής των αποτελεσμάτων, έχουν καταγραφεί σε πίνακες όλα τα αποτελέσματα που προέκυψαν από την ανάλυση του data set μέσω των αλγορίθμων που χρησιμοποιήθηκαν. Η ανάλυση έγινε σε balanced και imbalanced δεδομένα, με και χωρίς feature selection. Το μοντέλο που ξεχώρισε ήταν αυτό του του αλγορίθμου Random Forest με feature selection των balanced δεδομένων. Αυτό που παρατηρήθηκε από την ανάλυση είναι ότι τα μοντέλα που δημιουργήθηκαν από τα imbalanced δεδομένα δεν έδιναν καλές τιμές (Precision, Recall, F-measure) σε αντίθεση με τα μοντέλα των balanced δεδομένων. Αφού επιλεγεί το βέλτιστο μοντέλο, παρακάτω γίνεται μια παρουσίαση των κανόνων που δημιουργήθηκαν από τη χρήση του αλγορίθμου PART για το προτεινόμενο μοντέλο.

Στη συνέχεια έχουμε τη δημιουργία του πίνακα των εργασιών οι οποίες συγκρίθηκαν πιο πάνω συμπεριλαμβανομένου και της δικιά μας. Αυτός ο πίνακας αποσκοπεί στο να συμπεριληφθούν όλες οι λεπτομέρειες των σημαντικότερων εργασιών σε όσο το δυνατόν λιγότερα λόγια, έτσι ώστε να είναι ευανάγνωστο για τον αναγνώστη αλλά και παράλληλα να μην χάνει το νόημα και τα σημαντικά στοιχεία της κάθε εργασίας.

Κλείνοντας, η εργασία μας ολοκληρώνεται με την ανάλυση της μεθοδολογίας που εφαρμόστηκε. Στη συνέχεια έχουμε τις συγκρίσεις των εργασιών, οι συγκρίσεις έγιναν μεταξύ παρόμοιων εργασιών με την έρευνα μας. Δηλαδή οι συγκρίσεις έγιναν κυρίως με εργασίες οι οποίες είχαν το ίδιο ή και παρόμοιο σύνολο δεδομένων με τη δικιά μας. Αυτό που παρατηρήθηκε είναι ότι οι εργασίες είχαν αρκετά κοινά στοιχεία μεταξύ τους, για παράδειγμα χρησιμοποιήθηκαν αρκετοί ίδιοι αλγόριθμοι μεταξύ των εργασιών, αρκετές παρόμοιες τεχνικές

καθώς και η αξιολόγηση των μοντέλων έγινε κυρίως με βάση τις τιμές-μέτρα των Precision, Recall και F-measure όπως ακριβώς και στη δικιά μας. Παρόλα αυτά υπήρχαν και αρκετές διαφορές.

7.2 Αναφορές άρθρων - Βιβλιογραφία

- [1] V.Kumar, Divya Ramachandran, Binay Kumar / Influence of new-age technologies on marketing: A research agenda / Journal of Business Research / 14 / Science Direct / Elsevier / 2020
- [2] Satu Elisa Schaeffer, Sara Veronica Rodriguez Sanchez / Forecasting client retention — A machine-learning approach / Journal of Retailing and Consumer Services / 9 / Science Direct / Elsevier / 2020
- [3] Piotr Ładyżynski, Kamil Zbikowski, Piotr Gawrysiak / Direct marketing campaigns in retail banking with the use of deep learning and random forests / Expert Systems With Applications / 8 / Science Direct / Elsevier / 2019
- [4] Emma M. Sánchez, Julio B. Clempner, Alexander S. Poznyak / A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean–variance customer portfolio: The case of bank marketing campaigns / Engineering Applications of Artificial Intelligence / 11 / Science Direct / Elsevier / 2015
- [5] Hany A. Elsalamony / Bank Direct Marketing Analysis of Data Mining Techniques / International Journal of Computer Applications / 11 / Open source online / 2014
- [6] Hamit Erdal, İlhami Karahanoğlu / Bagging ensemble models for bank profitability: An empirical research on Turkish development and investment banks / Applied Soft Computing / 19 / Science Direct / Elsevier / 2016
- [7] Stefan M. Kostić, Miloš Đuričić, Mirjana I. Simić, Miroljub V. Kostić / Data Mining and Modeling Use Case in Banking Industry / 26th Telecommunications forum TELFOR / 4 / IEEE Xplore / 2018
- [8] Armin Lawi, Ali Akbar Velayaty, Zahir Zainuddin / On identifying potential direct marketing consumers using adaptive boosted support vector machine / Computer Applications and Information Processing Technology (CAIPT) / 4 / IEEE Xplore / 2017
- [9] Punitara Ruangthong, Saichon Jaiyen / Bank direct marketing analysis of asymmetric information based on machine learning / 12th International Joint Conference on Computer Science and Software Engineering (JCSSE) / 4 / IEEE Xplore / 2015
- [10] Bhuvan M Shashidhara, Siddharth Jain, Vinay D Rao, Nagamma Patil, Raghavendra G S / Evaluation of Machine Learning Frameworks on Bank Marketing and Higgs Datasets / Second International Conference on Advances in Computing and Communication Engineering / 5 / IEEE Xplore / 2015
- [11] Tuba Parlar, Songul Kakilli Acaravci / Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data / International Journal of Economics and Financial Issues / 5 / EconJournals / 2017
- [12] Putta Hemalatha, Geetha Mary Amalanathan / A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector / International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) / 6 / IEEE Xplore / 2019
- [13] Elzhan Zeinulla, Karena Bekbayeva, Adnan Yazici / Comparative study of the classification models for prediction of bank telemarketing / Application of Information and Communication Technologies (AICT) / 5 / IEEE Xplore / 2018

- [14] Benlan He, Yong Shi, Qian Wan, Xi Zhao / Prediction of Customer Attrition of Commercial Banks based on SVM Model / Procedia Computer Science / 8 / Science Direct / Elsevier / 2014
- [15] Muhammet Sinan Başarslan, İrem Düzdar Argun / Classification Of a bank data set on various data mining platforms / Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) / 4 / IEEE Xplore / 2018
- [16] Merve Mitik, Ozan Korkmaz, Pinar Karagoz, Ismail Hakki Toroslu, Ferhat Yucel / Data Mining Approach for Direct Marketing of Banking Products with Profit/Cost Analysis / The Review of Socionetwork Strategies / 15 / SpringerLink / 2017
- [17] Shiueng-Bien Yang, Tai-Liang Chen / Uncertain decision tree for bank marketing classification / Journal of computational and Applied Mathematics / 24 / Science Direct / Elsevier / 2020
- [18] Chaitra Hegde, Aakash Kaku, Neelang Parghi / Predicting Bank Marketing Campaign Success using Machine Learning / 4 / Open source online / 2014

7.3 Πηγές – Εργαλεία

- ✓ Εργαλείο που χρησιμοποιήθηκε για τη δημιουργία του προτεινόμενου μοντέλου είναι το WEKA: <https://www.cs.waikato.ac.nz/ml/weka/>
- ✓ Οδηγός χρησιμοποίησης και κατανόησης εννοιών του εργαλείου WEKA: http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf
- ✓ Για την εύρεση των άρθρων-εργασιών που χρησιμοποιήθηκαν για την ανάλυση κάναμε χρήση τριών ιστοσελίδων:
 - i. Science Direct, <https://www.sciencedirect.com/>
 - ii. IEEE Xplore, <https://ieeexplore.ieee.org/Xplore/home.jsp>
 - iii. Springer Link, <https://link.springer.com/>
- ✓ Αρκετά βοηθητική σελίδα για την κατανόηση διάφορων εννοιών για το machine learning και για το data mining αλλά και λειτουργίας των αλγορίθμων: <https://machinelearningmastery.com/>
- ✓ Πηγή εύρεσης συνόλου δεδομένων: <https://archive.ics.uci.edu/ml/datasets.php>
- ✓ Διαφάνειες και σημειώσεις μαθήματος “Αποθήκες & Εξόρυξη δεδομένων” Ιονίου πανεπιστημίου τμήματος πληροφορικής: <https://e-class.ionio.gr/courses/DCS172/>

