ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Πολιτικών Μηχανικών

Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Πρότυπα μηχανικής μάθησης για την πρόβλεψη της τιμής των καυσίμων και η χρησιμότητά τους στις μεταφορές

Διπλωματική Εργασία

Αλέξανδρος Χαραλαμπίδης

Επιβλέπουσα Καθηγήτρια : Ελένη Βλαχογιάννη,

Επίκουρη Καθηγήτρια Σχολής Πολιτικών Μηχανικών ΕΜΠ

Αθήνα, Ιούλιος 2017

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας μου που σηματοδοτεί και την ολοκλήρωση της πενταετούς φοίτησής μου στη Σχολή Πολιτικών Μηχανικών θέλω να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτριά μου, κυρία Ελένη Βλαχογιάννη, Επίκουρη Καθηγήτρια στον Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής της Σχολής Πολιτικών Μηχανικών. Είναι εκείνη που με την πολυετή εμπειρία της στις μεθόδους εκμάθησης μηχανών, με μύησε στα νευρωνικά δίκτυα, τα δένδρα απόφασης και την μέθοδο AdaboostM1. Την ευχαριστώ ολόψυχα για την καθοδήγηση, τις συμβουλές και την αμέριστη συμπαράστασή της που συνέβαλαν στην υπέρβαση δυσκολιών και την ουσιαστική ολοκλήρωση της διπλωματικής μου εργασίας. Θα θυμάμαι πάντα την προθυμία της να με βοηθήσει σε οποιαδήποτε απορία της είχα διατυπώσει.

Θέλω ακόμα να ευχαριστήσω τους ανθρώπους που ήταν δίπλα μου όλον αυτό τον καιρό, την οικογένειά μου και τους φίλους μου που με στήριξαν σε μια ιδιαίτερα απαιτητική και ταυτόχρονα εποικοδομητική περίοδο.

Τίτλος : Πρότυπα μηχανικής μάθησης για την πρόβλεψη της τιμής των καυσίμων και η χρησιμότητά τους στις μεταφορές

Συγγραφέας Διπλωματικής Εργασίας : Αλέξανδρος Χαραλαμπίδης

Επιβλέπουσα Καθηγήτρια : Ελένη Βλαχογιάννη

ΣΥΝΟΨΗ

Οι μεταβολές της λιανικής τιμής των καυσίμων έχουν σημαντικές επιπτώσεις στη ζήτηση των μέσων μαζικής μεταφοράς και στις μεταφορικές εταιρίες που δραστηριοποιούνται στον συγκοινωνιακό κλάδο. Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη προτύπων πρόβλεψης της τιμής της απλής αμόλυβδης σε μεσοπρόθεσμο χρονικό ορίζοντα. Για τον σκοπό αυτό, εφαρμόζονται τρεις μέθοδοι από την οικογένεια εκμάθησης μηχανών, τα νευρωνικά δίκτυα, τα δένδρα απόφασης και τα δάση απόφασης με βάση τη μέθοδο AdaboostM1. Αρχικά, έγινε έρευνα στην βιβλιογραφία για την κατάλληλη προετοιμασία και ομαλοποίηση των δεδομένων εισόδου και διερευνήθηκε η δομή της εκάστοτε μεθόδου εκμάθησης μηχανών. Εν συνεχεία, εκπαιδεύτηκαν και αξιολογήθηκαν οι τρεις μέθοδοι εκμάθησης μηχανών, αλλά έγινε και σύγκριση μεταξύ τους. Τέλος, η βελτιστοποίηση των παραμέτρων της εκάστοτε μεθόδου συνέβαλλε στην μείωση του σφάλματος γενίκευσης και στην επίτευξη ακρίβειας πρόβλεψης της τάξης του 75%.

Λέξεις κλειδιά: μεταβολές καυσίμων, κυκλοφοριακή ζήτηση, μεταφορικές εταιρίες, πρόβλεψη απλής αμόλυβδης, νευρωνικά δίκτυα, δένδρα απόφασης, AdaboostM1, μέθοδοι εκμάθησης μηχανών

Title : Machine learning techniques for the prediction of retail fuel prices and their applications on the transport sector

Thesis Author : Alexandros Charalampidis

Supervising Professor : Eleni Vlahogianni

ABSTRACT

Fuel variations have significant implications on the transport demand for public transport and other transportation related companies. The goal of this diploma thesis is to develop models to predict retail gasoline prices in mid-term time horizon. For the construction of those models we applied three machine learning techniques, artificial neural networks, decision trees and AdaboostM1. Initially, in depth research was conducted to find the most appropriate pre-processing and normalization techniques for the input data and the structure of its machine learning technique. At the next step, all three machine learning techniques were trained and evaluated, as well as compared with each other. Finally, the optimization of the parameters of each machine learning technique significantly contributed to reducing the generalization error and achieve prediction accuracy of approximately 75%.

Key words: Fuel variations, transport demand, transport companies, prediction of retail gasoline prices, artificial neural networks, decision trees, AdaboostM1, machine learning techniques

ΠΕΡΙΛΗΨΗ

Οι μεταβολές της λιανικής τιμής των καυσίμων έχουν σημαντικές επιπτώσεις, τόσο στη ζήτηση των μέσων μαζικής μεταφοράς, όσο και στις μεταφορικές εταιρίες που δραστηριοποιούνται στον συγκοινωνιακό κλάδο. Η αύξηση της λιανικής τιμής των καυσίμων έχει ως συνέπεια οι πολίτες να περιορίζουν τις διαδρομές με τα επιβατικά τους οχήματα, προκειμένου να μετριάσουν το αυξημένο μεταφορικό κόστος. Σε τέτοιες περιπτώσεις οι πολίτες προτιμούν να αντικαθιστούν τα επιβατικά τους οχήματα με κάποιο μέσο μαζικής μεταφοράς, ώστε να εκτελέσουν τις διαδρομές τους. Αυτό έχει ως αποτέλεσμα πολλές φορές να δημιουργούνται προβλήματα εξυπηρέτησης της αυξημένης κυκλοφοριακής ζήτησης από τα μέσα μαζικής μεταφοράς. Ακόμη, η αύξηση της λιανικής τιμής των καυσίμων δημιουργεί πολλά προβλήματα στους οδικούς μεταφορείς. Δεδομένου ότι το κόστος αγοράς καυσίμου αποτελεί σημαντικό ποσοστό του μεταφορικού κόστους οποιαδήποτε μεταβολή της τιμής των καυσίμων έχει άμεση επίπτωση στους οδικούς μεταφορές. Έτσι, δημιουργείται η ανάγκη πρόβλεψης της λιανικής τιμής των καυσίμων, προκειμένου να μπορούν οι πολίτες και οι εταιρίες στον συγκοινωνιακό τομέα να ρυθμίζουν καλύτερα τον χρονισμό αγοράς καυσίμων, ώστε να μετριάσουν τις μεταβολές του μεταφορικού κόστους και οι δεύτεροι να διατηρούν την κερδοφορία και την ανταγωνιστικότητά τους στον κλάδο.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η διερεύνηση της μεσοπρόθεσμης προβλεψιμότητας της λιανικής τιμής των καυσίμων και συγκεκριμένα της απλής αμόλυβδης. Για την επίλυση του προβλήματος της παρούσας διπλωματικής εργασίας κρίθηκε αναγκαία η χρήση της χρονοσειράς της απλής αμόλυβδης, αλλά και του αργού πετρελαίου, καθώς η διύλισή του παράγει την αμόλυβδη. Στην βιβλιογραφική ανασκόπηση εξετάστηκαν διάφορες μέθοδοι γραμμικών και μη γραμμικών μοντέλων πρόβλεψης που εφαρμόζονται στην πρόβλεψη της λιανικής τιμής των καυσίμων. Από την ανάλυση της βιβλιογραφίας κρίθηκε εύλογο να αξιολογηθούν τρεις μέθοδοι, οι οποίες ανήκουν στην ευρύτερη οικογένεια των μεθόδων εκμάθησης μηχανών. Αυτές είναι τα νευρωνικά δίκτυα, τα δένδρα απόφασης και τα δάση με βάση την μέθοδο AdaboostM1, προκειμένου να αναδειχθεί η βέλτιστη μέθοδος για την πρόβλεψη της λιανικής τιμής της απλής αμόλυβδης.

Αρχικά, συλλέχθηκαν τα δεδομένα για την ανάλυση και την εκπαίδευση της εκάστοτε μεθόδου εκμάθησης μηχανών. Προτού αρχίσει η εκπαίδευση έγινε κατάλληλη προετοιμασία και ομαλοποίηση των μεταβλητών εισόδου και εξόδου, προκειμένου να εξασφαλιστούν για κάθε μέθοδο οι κατά το δυνατόν μεγαλύτερες ακρίβειες στο σύνολο δοκιμής. Η προετοιμασία και η ομαλοποίηση των δεδομένων είχε ως σκοπό όλα τα δεδομένα εισόδου να ανήκουν στο πεδίο τιμών (0,1) και η κατανομή τους να προσεγγίζει όσο το δυνατόν την «κανονική». Το πρόβλημα που επιλύει η παρούσα διπλωματική εργασία είναι τύπου 0/1, όπου 1 και 0 είναι η πρόβλεψη για ανοδική και καθοδική κατεύθυνση της τιμής της απλής αμόλυβδης αντίστοιχα.

Εν συνεχεία έγινε η εκπαίδευση των μεθόδων εκμάθησης μηχανών έχοντας πρώτα επιλέξει τις βέλτιστες παραμέτρους από το σύνολο επικύρωσης και υπολογίστηκαν τα στατιστικά μέτρα αξιολόγησης στο σύνολο δοκιμής για κάθε μια από αυτές. Σημειώνεται ότι στην παρούσα διπλωματική εργασία επιτεύχθηκαν ακρίβειες πρόβλεψης της τάξης του 75% για την κατεύθυνση της τιμής της απλής αμόλυβδης στο σύνολο δοκιμής. Η αξιοποίηση ενός από τα μοντέλα πρόβλεψης που κατασκευάστηκαν στην παρούσα διπλωματική εργασία μπορεί να μετριάσει σημαντικά τις μεταβολές του μεταφορικού κόστους και να περιορίσει σε μεγάλο βαθμό τις επιπτώσεις των μεταβολών των καυσίμων στον συγκοινωνιακό τομέα. Τέλος, στο πλαίσιο της παρούσας διπλωματικής εργασίας διατυπώνονται και προβλήματα για περαιτέρω έρευνα.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

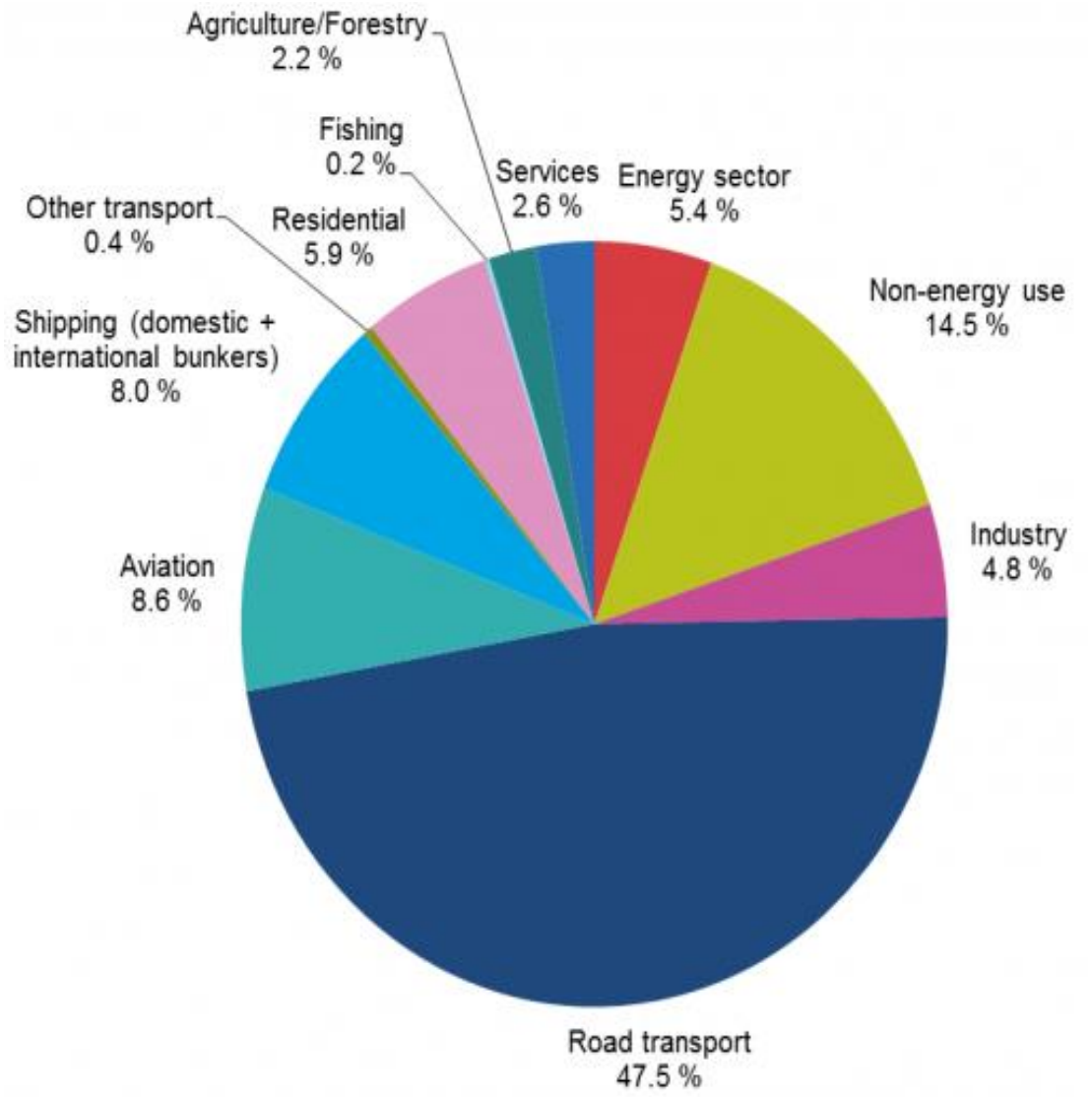
ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1 ΕΝΕΡΓΕΙΑΚΕΣ ΑΝΑΓΚΕΣ ΣΤΟΝ ΣΥΓΚΟΙΝΩΝΙΑΚΟ ΤΟΜΕΑ

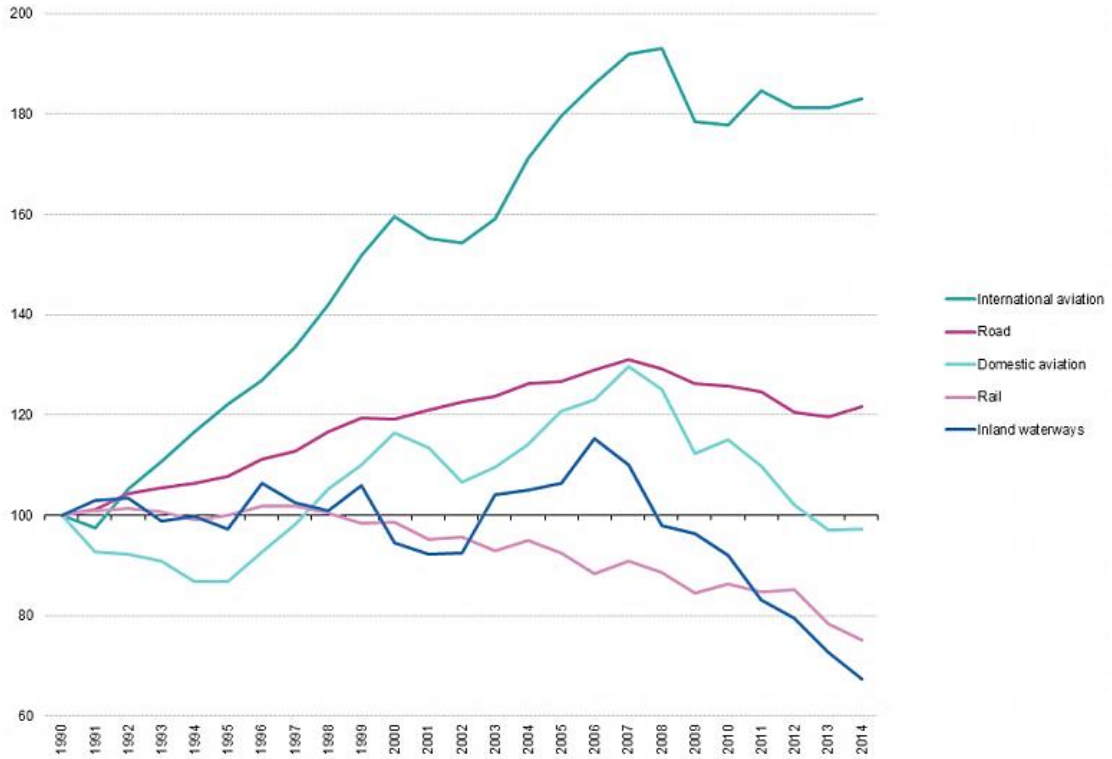
Οι μεταφορές είναι ένας βασικός τομέας δραστηριοτήτων που χαρακτηρίζεται από σημαντικές καταναλώσεις ενέργειας. Κατά την περίοδο 1990 – 2016, η κατανάλωση ενέργειας που οφείλεται στις μεταφορές αυξήθηκε σε μεγαλύτερο βαθμό από ότι στους τομείς της βιομηχανίας και του νοικοκυριού. Συγκεκριμένα, το ποσοστό των μεταφορών επί της συνολικής κατανάλωσης ενέργειας στην Ε.Ε. ανήλθε από το 26% το 1990 στο 33,2% το 2016 (European Commission, 2016). Παρόμοια είναι και η εικόνα στις Ηνωμένες Πολιτείες της Αμερικής , όπου ο τομέας των μεταφορών κατανάλωσε το 28% της συνολικής κατανάλωσης ενέργειας, όπως αναφέρεται από τις Αμερικανικές Αρχές διοίκησης για θέματα ενέργειας (EIA, 2015).

Αξίζει να σημειωθεί ότι ο τομέας των μεταφορών είναι ο κύριος καταναλωτής ορυκτών καυσίμων, καθώς το 97% των ενεργειακών αναγκών του καλύπτονται από προϊόντα που παράγονται από ορυκτά καύσιμα. Αυτό συμβαίνει, διότι τα βιοκαύσιμα που χρησιμοποιούνται στις οδικές μεταφορές και στην Ευρώπη είναι πολύ κοστοβόρα, καθώς απαιτούν μεγάλες εκτάσεις καλλιεργήσιμης γης για να παραχθούν. Δεύτερον, ένας άλλος λόγος είναι ότι τα ηλεκτρικά αυτοκίνητα που θα μπορούσαν να παράσχουν μια εναλλακτική λιγότερο κοστοβόρα, δεν έχουν ακόμη αναπτυχθεί σε σημαντικό βαθμό, καθώς υπάρχουν κάποια τεχνολογικά εμπόδια που πρέπει να επιλυθούν, ώστε να περιοριστεί το κόστος τους. Τρίτον, οι τεχνολογίες υδρογόνου θα εισέλθουν στο κλάδο μακροπρόθεσμα (2030). Κατά συνέπεια, οι οδικές μεταφορές παρουσιάζουν την υψηλότερη ζήτηση για ενέργεια που παράγεται από πετρέλαιο. Συγκεκριμένα, οι οδικές μεταφορές καταναλώνουν το 47,5% της συνολικής κατανάλωσης πετρελαίου, όπως δημοσιεύθηκε από την Ευρωπαϊκή Επιτροπή τον Νοέμβριο του 2016. Στον πίνακα 1.1 παρουσιάζονται αναλυτικά το ποσοστό κάθε κλάδου επί της συνολικής κατανάλωσης πετρελαίου για τα 28 κράτη – μέλη (European Commission, 2016).



Διάγραμμα 1.1: Κατανομή της κατανάλωσης πετρελαίου σε διάφορους τομείς δραστηριοτήτων για τα 28 κράτη-μέλη της Ε.Ε., (Πηγή: Ευρωπαϊκή Επιτροπή Eurostat 2017)

Αξίζει να επισημανθεί ότι υπάρχει μια η δυσανάλογη εξέλιξη των ενεργειακών απαιτήσεων σε κάθε μεταφορικό μέσο ξεχωριστά κατά την περίοδο 1990 - 2014. Συγκεκριμένα, οι διεθνείς αερομεταφορές είχαν την μεγαλύτερη ανάπτυξη σε ότι αφορά την κατανάλωση ενέργειας , 83,2%, ακολουθούμενες από τις οδικές μεταφορές με αύξηση 21,7%. Τα υπόλοιπα μέσα παρουσίασαν μείωση στην κατανάλωση ενέργειας ειδικότερα μετά την έναρξη της οικονομικής κρίσης στις αρχές του 2008. Όλες αυτές οι πληροφορίες απεικονίζονται στο Διάγραμμα 1.2, όπως δημοσιεύθηκε από την Ευρωπαϊκή Επιτροπή τον Ιούλιο του 2016 (European Commission, 2016) και εκφράζει τις μεταβολές στην κατανάλωση ενέργειας υπό την επίδραση των τεχνολογικών εξελίξεων της απόδοσης του καυσίμου στο εκάστοτε μέσο.



Διάγραμμα 1.2: Κατανάλωση ενέργειας ανά μέσο μεταφοράς για τα 28 κράτη-μέλη της Ε.Ε., (Πηγή: Ευρωπαϊκή Επιτροπή Eurostat 2017)

Από όσα αναφέρθηκαν παραπάνω γίνεται κατανοητό ότι το πετρέλαιο καλύπτει περίπου το ένα τρίτο των ενεργειακών αναγκών των 28 κρατών- μελών της Ε.Ε. Λόγω του υψηλού κόστους που έχει συγκρινόμενο με τον άνθρακα και το αέριο, η χρήση του είναι συγκεντρωμένη σε τομείς όπου άλλες μορφές ενέργειας δεν είναι κατάλληλα υποκατάστατα αυτού. Αυτό συμβαίνει κυρίως στις μεταφορές που το 47,5% των ενεργειακών αναγκών καλύπτεται από πετρέλαιο. Το υπόλοιπο ποσοστό του πετρελαίου χρησιμοποιείται στις βιομηχανίες για την λειτουργία των μηχανημάτων, στις αγροτικές εργασίες, αλλά και για την παραγωγή λιπασμάτων και χημικών ουσιών. Έτσι, γίνεται αντιληπτό ότι οποιαδήποτε μεταβολή στην τιμή του πετρελαίου έχει άμεσες επιπτώσεις τόσο στον κλάδο των μεταφορών όσο και στην κοινωνικο - οικονομική ζωή των πολιτών, γενικότερα. Κατά συνέπεια, δημιουργείται η ανάγκη να διερευνηθεί κατά πόσο αυτές οι μεταβολές στις τιμές των καυσίμων μπορούν να προβλεφθούν προς όφελος του κοινωνικού συνόλου γενικότερα.

1.2 ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΙΔΡΟΥΝ ΣΤΗ ΔΙΑΜΟΡΦΩΣΗ ΤΗΣ ΛΙΑΝΙΚΗΣ ΤΙΜΗΣ ΤΩΝ ΚΑΥΣΙΜΩΝ

Στην παρούσα ενότητα θα γίνει αναλυτική περιγραφή των παραγόντων που επιδρούν στην διαμόρφωση της λιανικής τιμής των καυσίμων, καθώς και το ποσοστό που κατέχει κάθε ένας παράγοντας ξεχωριστά.

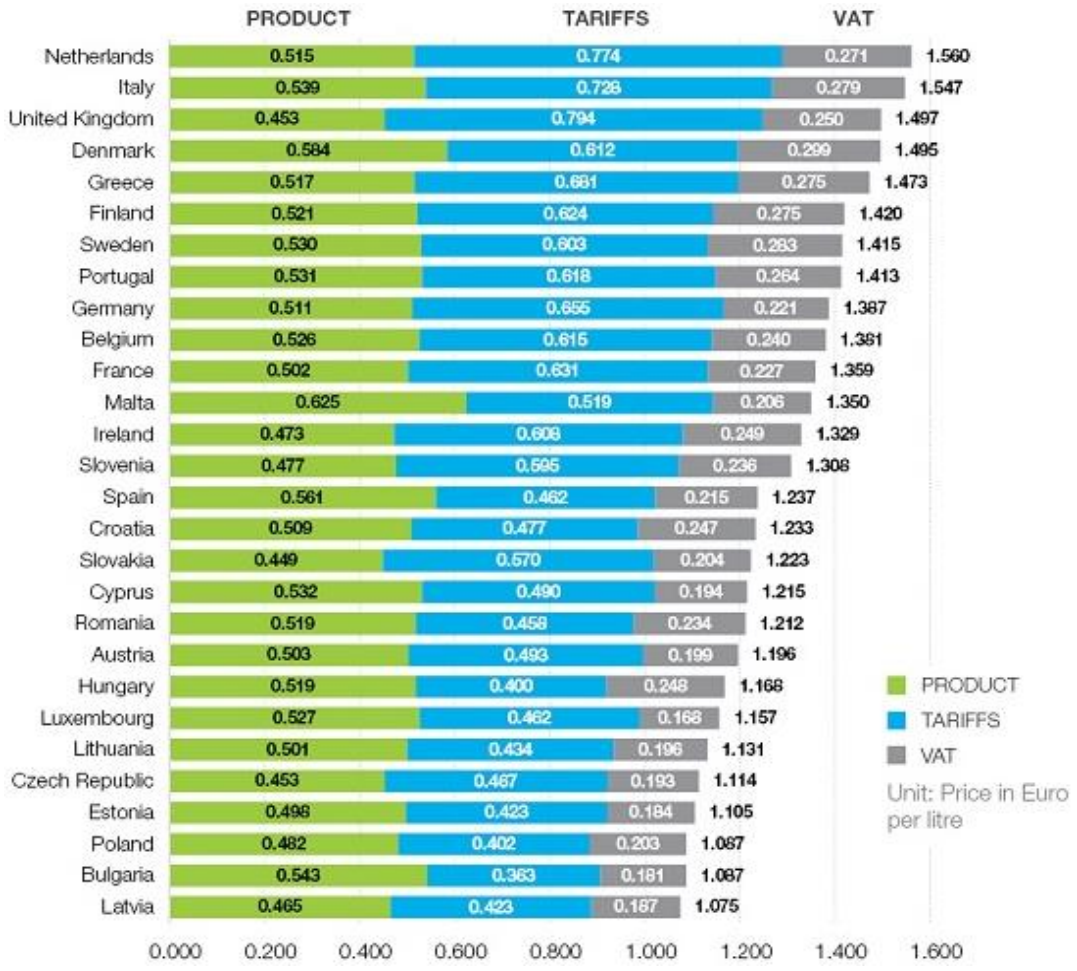
Σημαντικός παράγοντας είναι το αργό πετρέλαιο, καθώς η διύλισή του παράγει προϊόντα, όπως η βενζίνη και το ντίζελ. Εκτιμάται ότι η συνεισφορά του στην λιανική τιμή των καυσίμων στις χώρες της Ε.Ε. είναι περίπου 20 – 22%. Το ποσοστό διαφέρει από χώρα σε χώρα, επειδή οι σταθεροί φόροι και ο φόρος προστιθέμενης αξίας είναι διαφορετικοί. Επισημαίνεται λοιπόν ότι οι αυξομειώσεις στις τιμές του αργού πετρελαίου ενσωματώνονται άμεσα, προκαλώντας σημαντικές μεταβολές στις λιανικές τιμές των καυσίμων. Όσον αφορά τις τιμές του αργού πετρελαίου αυτές καθορίζονται τόσο από την προσφορά, όσο και την ζήτηση. Η παγκόσμια οικονομική ανάπτυξη είναι ο κύριος παράγοντας που επηρεάζει την ζήτηση. Όμως, κάποιες φορές η Ένωση Εξαγωγών Πετρελαιοπαραγωγών Χωρών (OPEC) θέτει ένα πάνω όριο στην ημερήσια παραγωγή βαρελιών αργού πετρελαίου επηρεάζοντας την προσφορά, καθώς ελέγχει τα τρία τέταρτα των αποθεμάτων αργού πετρελαίου παγκοσμίως (EIA, 2016).

Ένας δεύτερος σημαντικός παράγοντας είναι η συναλλαγματική ισοτιμία ευρώ/δολάριο, καθώς επιδρά στον τρόπο τιμολόγησης των προϊόντων από τα διυλιστήρια. Πιο συγκεκριμένα, ως τιμή βάσης για τα προϊόντα διύλισης χρησιμοποιούνται οι διεθνείς τιμές πετρελαιοειδών, όπως αυτές δημοσιεύονται στα χρηματιστηριακά δελτία πετρελαίου Platt's Oil Gram. Οι τιμές αυτές είναι σε δολάριο ανά μετρικό τόνο. Ο ρόλος της συναλλαγματικής ισοτιμίας εισέρχεται για την μετατροπή του δολαρίου σε ευρώ, καθώς το νόμισμα που πραγματοποιούνται οι συναλλαγές στις χώρες της Ε.Ε. είναι το ευρώ. Σημειώνεται ότι ως τιμή βάσης των προϊόντων έχει καθιερωθεί να λαμβάνεται υπόψη η διεθνής τιμή Platt's της Μεσογείου και όχι αυτή της Βόρειας Ευρώπης (Rotterdam). Την πρακτική αυτή ακολουθούν οι χώρες της Μεσογείου όπως η Ελλάδα, με αποτέλεσμα να βρίσκονται μεταξύ των χωρών της Ε.Ε. με τις υψηλότερες τιμές αμόλυβδης (Kathimerini, 2014).

Τρίτος παράγοντας είναι τα περιθώρια κέρδους των εταιριών εμπορίας πετρελαιοειδών και των πρατηριούχων. Οι εταιρίες εμπορίας πετρελαιοειδών προσαυξάνουν τις τιμές ανάλογα με την τιμολογιακή πολιτική που ακολουθούν, τα ποσοστά κέρδους και τις δαπάνες μεταφοράς. Τα ίδια κόστη ισχύουν και για τους πρατηριούχους. Εκτιμάται ότι τα κόστη διανομής πετρελαιοειδών αντιστοιχούν στο 7% της λιανικής τιμής των καυσίμων (European Commission, 2016). Παρόμοιο είναι και το ποσοστό για το κόστος διαφήμισης των πετρελαιοειδών.

Τέταρτος παράγοντας είναι οι φόροι από τους οποίους κάποιοι είναι σταθεροί και κάποιοι εξαρτώνται από την τιμή βάσης του προϊόντος. Αξίζει να σημειωθεί ότι αποτελούν το μεγαλύτερο ποσοστό στην τελική διαμόρφωση της τιμής των καυσίμων και καταβάλλονται στο κράτος από τις εταιρίες εμπορίας πετρελαιοειδών, προτού διανείμουν τα προϊόντα τους στην εσωτερική αγορά. Στην Ευρώπη αυτό το νούμερο είναι ίσο με περίπου 60% της τελικής λιανικής τιμής των καυσίμων. Για την καλύτερη κατανόηση του ρόλου των φόρων στις λιανικές τιμές των καυσίμων θα γίνει μια συνοπτική περιγραφή της κατάστασης που επικρατεί στην Ελλάδα. Ο ειδικός φόρος είναι σταθερός και ισούται με 0,679 ευρώ/λίτρο. Εκτός αυτού υπάρχει και ό φόρος προστιθέμενης αξίας που ισούται με 23% επί της λιανικής τιμής του καυσίμου. Τέλος, υπάρχει μια σειρά επιβαρύνσεων υπέρ του δημοσίου που εκτιμώνται σε 0,009 ευρώ/λίτρο.

Στο Διάγραμμα 1.3 παρατίθεται η ανάλυση της τιμής της απλής αμόλυβδης για τα κράτη – μέλη της Ε.Ε. (European Commission, 2016). Με πράσινο χρώμα υποδηλώνεται η τιμή της απλής αμόλυβδης χωρίς να συμπεριλαμβάνονται οι φόροι (σταθεροί και μη). Το μπλε χρώμα αναφέρεται στους σταθερούς φόρους, ενώ το γκρι χρώμα στο φόρο προστιθέμενης αξίας (Φ.Π.Α.).



Διάγραμμα1.3: Ανάλυση της λιανικής τιμής της απλής αμόλυβδης για τα κράτη – μέλη της Ε.Ε., (Πηγή: Ευρωπαϊκή Επιτροπή FuelsEurope 2017)

1.3 ΣΚΟΠΟΣ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η διερεύνηση της προβλεψιμότητας της απλής αμόλυβδης σε μεσοπρόθεσμο χρονικό ορίζοντα. Θα εφαρμοστούν τρεις μέθοδοι εκμάθησης μηχανών, τα νευρωνικά δίκτυα, δένδρα απόφασης και τα δάση απόφασης με βάση την μέθοδο AdaboostM1 για την ανάπτυξη των μοντέλων πρόβλεψης. Στόχος είναι η ανάπτυξη μοντέλων με μεγάλη ακρίβεια και αξιοπιστία πρόβλεψης, ώστε να μπορούν να αξιοποιηθούν από τους πολίτες και τις εταιρίες οδικών μεταφορών για την εύρεση του βέλτιστου χρονισμού αγοράς καυσίμων για τα οχήματά τους.

Αρχικά, θα γίνει ανάλυση των επιπτώσεων των μεταβολών των καυσίμων σε χώρες εντός και εκτός της Ε.Ε.. Θα παρουσιαστούν οι συνέπειες στην κυκλοφοριακή ζήτηση στα μέσα μαζικής μεταφοράς και στην κυκλοφοριακή ροή αλλά και οι επιπτώσεις στις εταιρίες που δραστηριοποιούνται στον συγκοινωνιακό τομέα. Σκοπός μας είναι μέσω της παράθεσης αυτών των επιπτώσεων να κατανοήσει ο αναγνώστης ότι η πρόβλεψη της τιμής των καυσίμων σε μεσοπρόθεσμο ορίζοντα αποτελεί θέμα μείζονος σημασίας για την πλειοψηφία του κοινωνικού συνόλου.

Εν συνεχεία, θα χρησιμοποιηθούν η υπολογιστική πλατφόρμα της Matlab και οι εφαρμογές της Weka για την εφαρμογή των μεθόδων εκμάθησης μηχανών. Τα δεδομένα εισόδου θα ομαλοποιηθούν κατάλληλα, ώστε να ενισχυθεί θετικά η ακρίβεια πρόβλεψης. Για όλες τις μεθόδους εκμάθησης μηχανών θα γίνει έρευνα ως προς την δομή που θα χρησιμοποιηθεί, καθώς και κατάλληλη επιλογή των παραμέτρων για την ελαχιστοποίηση του παραγόμενου γενικευμένου σφάλματος. Στο τέλος, θα παρουσιαστούν τα συμπεράσματα που προέκυψαν από την εφαρμογή των παραπάνω μεθόδων για την πρόβλεψη της τιμής της απλής αμόλυβδης και θα γίνει σύγκριση μεταξύ τους.

1.4 ΔΙΑΡΘΡΩΣΗ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η διάρθρωση της παρούσας διπλωματικής εργασίας είναι η εξής:

Στο κεφάλαιο 2 πραγματοποιείται η βιβλιογραφική ανασκόπηση της παρούσας διπλωματικής εργασίας. Αρχικά, στο πρώτο μέρος θα αναφερθούν οι επιπτώσεις των μεταβολών των καυσίμων στην κυκλοφοριακή ζήτηση των μέσων μαζικής μεταφοράς και στην κυκλοφοριακή ροή σε χώρες εντός και εκτός της Ε.Ε.. Θα αναφερθούν αναλυτικά τυχόν ιδιότυπες συμπεριφορές στην εκάστοτε χώρα. Στο δεύτερο μέρος της βιβλιογραφικής ανασκόπησης θα παρουσιαστούν γραμμικά και μη γραμμικά μοντέλα που εφαρμόζονται για την πρόβλεψη χρονοσειρών καυσίμων και αργού πετρελαίου. Θα αναφερθούν τα πλεονεκτήματα/μειονεκτήματα της εκάστοτε μεθόδου και διάφορες τεχνικές ομαλοποίησης που χρησιμοποιούνται από τους ερευνητές και ενδέχεται να αξιοποιηθούν και για την έρευνα της παρούσας διπλωματικής εργασίας.

Στο κεφάλαιο 3 παρατίθεται η μεθοδολογία επίλυσης που θα εφαρμοστεί στην παρούσα διπλωματική εργασία και το απαιτούμενο θεωρητικό υπόβαθρο για τις μεθόδους των νευρωνικών δικτύων, δένδρων απόφασης και AdaboostM1. Για κάθε μια μέθοδο θα γίνει αναλυτική περιγραφή της δομής και του τρόπου εκπαίδευσής της καθώς και των μεθόδων αντιμετώπισης προβλημάτων υπερπροσαρμογής. Στο τέλος του κεφάλαιο θα παρουσιαστούν τα στατιστικά μέτρα αξιολόγησης, τα οποία είναι κοινά και για τις τρεις μεθόδους.

Στο κεφάλαιο 4 παρατίθενται οι πηγές συλλογής των δεδομένων και οι τρόποι επεξεργασίας τους στα πλαίσια της παρούσας έρευνας. Θα παρατεθεί η διαδικασία προετοιμασίας και ομαλοποίησης των δεδομένων, προτού χρησιμοποιηθούν από τις μεθόδους εκμάθησης μηχανών.

Στο κεφάλαιο 5 θα γίνει εφαρμογή και των τριών μεθόδων εκμάθησης μηχανών για την πρόβλεψη της κατεύθυνσης της τιμής της απλής αμόλυβδης σε μεσοπρόθεσμο ορίζοντα. Θα παρατεθούν η δομή και ο τρόπος εκπαίδευσης της εκάστοτε μεθόδου. Έπειτα, θα επιλεγούν οι βέλτιστες παράμετροι για κάθε μέθοδο από το σύνολο επικύρωσης με τελικό στόχο την εξασφάλιση της μεγαλύτερης ακρίβειας στο σύνολο δοκιμής. Τα στατιστικά μέτρα θα υπολογιστούν στο σύνολο δοκιμής, ώστε να γίνει η αξιολόγηση των μεθόδων σε πραγματικές συνθήκες. Θα σχολιαστούν τυχόν σημαντικές παρατηρήσεις που προκύπτουν κατά την διαδικασία εκμάθησης.

Στο κεφάλαιο 6 θα παρουσιαστούν τα βασικά συμπεράσματα, όπως αυτά προέκυψαν από την παρούσα διπλωματική εργασία αλλά και θα παρατεθούν ποιοτικά συμπεράσματα αναφορικά με τις μεθόδους εκμάθησης μηχανών που χρησιμοποιήθηκαν. Στο τέλος, του κεφαλαίου θα γίνουν προτάσεις για περαιτέρω έρευνα και διερεύνηση του προβλήματος που επιλύει η παρούσα έρευνα.

Τέλος, στο κεφάλαιο 7 παρατίθεται η πλήρης βιβλιογραφίας που χρησιμοποιήθηκε για την πραγματοποίηση της διπλωματικής εργασίας.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

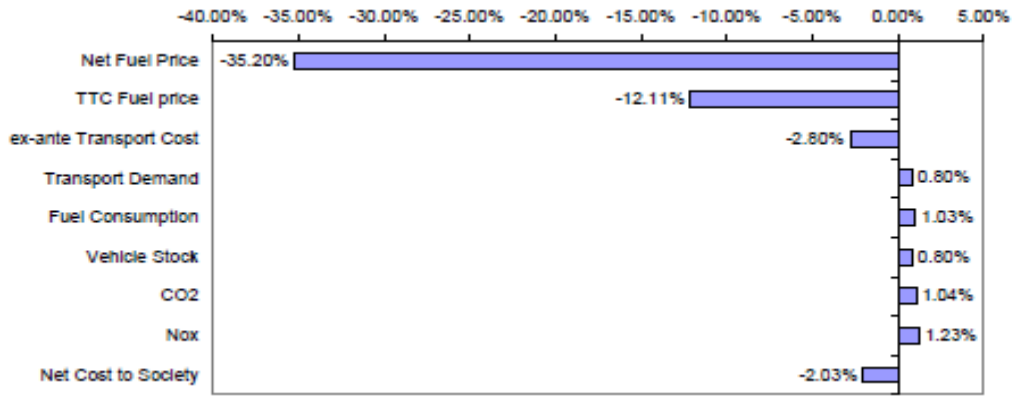
2.1 ΕΠΙΠΤΩΣΕΙΣ ΑΠΟ ΤΙΣ ΜΕΤΑΒΟΛΕΣ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΚΑΥΣΙΜΩΝ ΣΤΟΝ ΣΥΓΚΟΙΝΩΝΙΑΚΟ ΤΟΜΕΑ

2.1.1 ΕΠΙΠΤΩΣΕΙΣ ΣΤΑ ΚΡΑΤΗ – ΜΕΛΗ ΤΗΣ Ε.Ε.

Προσομοίωση σεναρίων στα κράτη μέλη της Ε.Ε.

Ένας τρόπος εύρεσης των επιπτώσεων της μεταβολής της τιμής της βενζίνης στις μεταφορές είναι μέσω προσομοιώσεων υποθετικών μεταβολών. Αυτόν τον τρόπο χρησιμοποίησε η Ευρωπαϊκή Επιτροπή το 2002 με την βοήθεια του μοντέλου TREMOVE υπό την επίβλεψη του γενικού διευθυντή των οικονομικών θεμάτων Jacques Delsalle (European Commission, Auto-Oil II Programme, 2002). Έτσι, στο πρώτο μοντέλο προσομοίωσης που δημιουργήθηκε, θεωρήθηκε πτώση της τιμής του πετρελαίου κατά 47% που σημαίνει μείωση 11,7% και 14,3% για την ενισχυμένη αμόλυβδη και το ντίζελ αντίστοιχα. Υπενθυμίζεται ότι η πτώση των τιμών των καυσίμων είναι μικρότερη από αυτή του πετρελαίου, καθώς στην τελική τιμή στα πρατήρια συμπεριλαμβάνονται και οι φόροι που θέτει το εκάστοτε κράτος στην τελική τιμή του καυσίμου. Συμπεριλαμβάνοντας και τις εννέα χώρες (Ελλάδα, Πορτογαλία, Ιρλανδία, Λουξεμβούργο, Ηνωμένο Βασίλειο, Γαλλία, Ισπανία, Γερμανία, Ιταλία) το αποτέλεσμα ήταν να παρατηρηθεί μείωση στο μέσο μεταφορικό κόστος 2,8% έχοντας σαν δεδομένο ότι οι τιμές των καυσίμων αποτελούν μόνο το 6% του συνολικού μεταφορικού κόστους, παρ' όλο που στην πραγματικότητα η τιμή του καυσίμου αποτελεί το 22,6% του μεταφορικού κόστους. Αξίζει να αναφερθεί ότι η μείωση στο μεταφορικό κόστος από χώρα σε χώρα κυμάνθηκε από 2,1% στο Ηνωμένο Βασίλειο έως 3,7% στην Ιρλανδία. Τέτοιες διαφορές οφείλονται κατά κύριο λόγο στα επίπεδα των φόρων στα καύσιμα στο κάθε

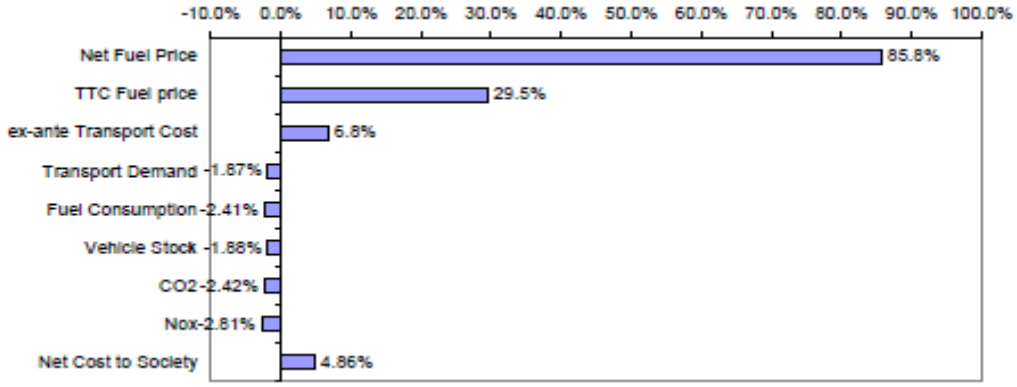
κράτος μέλος της Ε.Ε. Ακόμη, η μείωση στο μεταφορικό κόστος οδήγησε σε μια μικρή αύξηση 0,6% στην κυκλοφοριακή ροή. Σημειώνεται ότι η χώρα με την μεγαλύτερη ευαισθησία ήταν η Ιρλανδία (+0,9%) και λιγότερο ευαίσθητες ήταν η Γαλλία και το Ηνωμένο Βασίλειο (+0,5% και +0,4%). Οι διαφορές οφείλονται στις ελαστικότητες της ζήτησης που έχει κάθε χώρα. Για παράδειγμα, μια χώρα με υψηλή ελαστικότητα ζήτησης στα λεωφορεία, σημαίνει ότι σε μια αλλαγή της τιμής του καυσίμου πολύ εύκολα θα αλλάξουν μέσο κυκλοφορίας οι πολίτες, όπως συνέβη στην Ιρλανδία. Επίσης, η κατανάλωση ενέργειας αυξήθηκε κατά 1,03%, περισσότερο από ότι η κυκλοφοριακή κίνηση. Κατά συνέπεια συμπεραίνεται ότι υπάρχει μια μεταβολή προς οχήματα με μικρότερη εξοικονόμηση καυσίμου, όταν οι τιμές είναι χαμηλά. Όσον αφορά τις εξαγορές καινούριων οχημάτων μια σημαντική πτώση στις τιμές των καυσίμων οδηγεί σε μείωση των αγορών νέων αυτοκινήτων. Εύκολα μπορεί να εξαχθεί το συμπέρασμα ότι η παραπάνω κατάσταση οδηγεί σε αύξηση των εκπομπών διοξειδίου του άνθρακα (CO_2), ωστόσο η αύξηση είναι μόνο 0,5%, διότι υπάρχει μείωση στη χρήση αυτοκινήτων με κινητήρα ντίζελ. Συμπεριλαμβάνοντας όλα τα παραπάνω στα κόστη για την κοινωνία μειώνεται το κοινωνικό όφελος κατά 25%. Παρακάτω στο διάγραμμα 2.1 παρουσιάζονται συνοπτικά τα αποτελέσματα από την μείωση της τιμής των καυσίμων, όπως προέκυψε από το μοντέλο TREMOVE στο Auto-Oil II Programme.



Διάγραμμα 2.1: Παρουσίαση των αποτελεσμάτων από την 1η προσομοίωση, 2002, Ευρωπαϊκή Επιτροπή, Auto – Oil II Programme

Στην δεύτερη προσομοίωση που έγινε στο συγκεκριμένο πρόγραμμα Auto-Oil II Programme προσομοιώθηκε η κατάσταση του φθινοπώρου του 2000, όπου η τιμή του αργού πετρελαίου ήταν 30 δολάρια, δηλαδή σημειώθηκε αύξηση 86% σε σχέση με αυτή που θεωρήθηκε στο βασικό σενάριο. Αυτό είχε ως αποτέλεσμα αύξηση στην τελική τιμή του καυσίμου 29,5%. Κατά συνέπεια, το κόστος μεταφοράς ανήλθε 6,8% με εύρος από 5,1% για το Ηνωμένο Βασίλειο έως 9% για την Ισπανία. Όπως και αναμενόταν, η ζήτηση στις οδικές μεταφορές μειώθηκε κατά 1,4%, με την Ιρλανδία πάλι να είναι η πιο ευαίσθητη χώρα σε μεταβολές (-2%). Επιπρόσθετα, περιορίστηκε ο στόλος των αυτοκινήτων κυρίως εκείνων που κινούνταν με βενζίνη και των βαρέων φορτηγών. Παρατηρήθηκε ότι ο αριθμός των ελαφρών φορτηγών παρέμεινε σταθερός και το ποσοστό των αυτοκινήτων με

εξοικονόμηση ντίζελ καυσίμου αυξήθηκε σημαντικά. Κατά συνέπεια, η κατανάλωση καυσίμου μειώθηκε κατά 2,4% και η κυκλοφοριακή ροή μειώθηκε μόνο κατά 1,4%. Αυτό οφείλεται στην χρήση αυτοκινήτων εξοικονόμησης ενέργειας και στην μείωση της κυκλοφοριακής συμφόρησης. Όπως, και αναμενόταν υπάρχει μείωση των εκπομπών διοξειδίου του άνθρακα (CO_2). Συμμετρικά σε σχέση με το πρώτο σενάριο υπάρχει μείωση του χρόνου διάνυσης διαδρομής, αύξηση των φορολογικών εσόδων και μείωση των ατυχημάτων. Παρακάτω στο διάγραμμα 2.2 παρουσιάζονται συνοπτικά τα αποτελέσματα από την αύξηση της τιμής των καυσίμων, όπως προέκυψαν από το μοντέλο TREMOVE στο Auto-Oil II Programme.



Διάγραμμα 2.2: Παρουσίαση των αποτελεσμάτων από την 2η προσομοίωση, 2002, Ευρωπαϊκή Επιτροπή, Auto – Oil II Programme

Στην τρίτη προσομοίωση αυξήθηκαν οι τιμές των καυσίμων με ίδιο ποσοστό, όπως στην δεύτερη προσομοίωση αλλά τώρα η αύξηση οφειλόταν σε αύξηση της έμμεσης εσωτερικής φορολογίας των καυσίμων. Τα αποτελέσματα έδειξαν ότι οι επιπτώσεις ήταν μικρότερες και οι μεταβολές πιο ήπιες σε σχέση με το πρώτο σενάριο. Σημαντικό είναι το γεγονός ότι μια αύξηση στα έσοδα από εσωτερική φορολογία επιτρέπει την μερική ανακύκλωση των χρημάτων στην κυκλοφοριακή ζήτηση ενισχύοντας έτσι τις οδικές μεταφορές.

Επιπτώσεις κατά την περίοδο 2005 – 2008 στα κράτη μέλη της Ε.Ε.

Με τις επιπτώσεις που είχε η άνοδος της τιμής του πετρελαίου την περίοδο 2004 - 2008 στον τομέα των μεταφορών ασχολήθηκε εκτενώς το τμήμα Policy Department Structural and Cohesion Policies του Ευρωπαϊκού Κοινοβουλίου (European Parliament, 2009). Συγκεκριμένα, αναφέρει ότι οι εταιρίες οδικών μεταφορών εμπορευμάτων δεν μπορούσαν να μετακυλίσουν την αύξηση του κόστους καυσίμου στον τελικό πελάτη. Για παράδειγμα, την περίοδο αυτή στην Ιταλία το κόστος καυσίμου επί του συνολικού μεταφορικού κόστους ανήλθε από το 18,5% στο 21% κατά την παραπάνω περίοδο. Παρόμοιες ήταν και οι τάσεις και στα υπόλοιπα κράτη – μέλη της Ε.Ε. Επίσης, αναφέρεται ότι μια στις τρεις Ευρωπαϊκές εταιρίες οδικών μεταφορών παρατήρησαν αύξηση στα κόστη τους μεγαλύτερο από 20%, με

κύρια αιτία την άνοδο της τιμής του καυσίμου. Στην Ιταλία παρ΄ όλη την αύξηση στις τιμές των καυσίμων δεν σημειώθηκε καμία αύξηση στις τιμές των οδικών μεταφορών πλην ελαχίστων εξαιρέσεων.

Η διαφορά μεταξύ του κόστους καυσίμων και των τιμών των μεταφορών μεγαλώνει στην κεντρική Ευρώπη, με το πλέον χαρακτηριστικό παράδειγμα την Ρωσία που τα κόστη αυξήθηκαν 19,1%, ενώ οι τιμές των δρομολογίων μειώθηκαν κατά 4,5% κατά την περίοδο 2007 – 2008. Δεδομένου ότι οι εταιρίες οδικών μεταφορών δεν μπορούσαν να μετακυλίσουν την αύξηση του μεταφορικού κόστους στους πελάτες τους λόγω ανταγωνισμού, περιορίστηκαν σημαντικά τα κέρδη ολόκληρου του τομέα. Πιο συγκεκριμένα, οι εταιρίες οδικών μεταφορών στην Γερμανία και την Ολλανδία είδαν την κερδοφορία τους να συρρικνώνεται κατά 30% το 2007 συγκριτικά με το 2005. Αναφέρεται ότι λόγω των αυξημένων τιμών των καυσίμων οι οδικοί μεταφορείς άρχισαν να χρησιμοποιούν τεχνικές περιορισμού της κατανάλωσης καυσίμου, όπως ο περιορισμός της μέγιστης ταχύτητας. Ειδικότερα, οι μεγάλες εταιρίες του κλάδου άρχισαν να χρησιμοποιούν δεξαμένες αποθήκευσης καυσίμων για να μπορούν να αντισταθμίσουν τον κίνδυνο αύξησης των τιμών, δημιουργώντας πλεονάσματα καυσίμων ντίζελ, ώστε να παραμένουν ανταγωνιστικοί. Εκείνες οι εταιρίες που επένδυσαν στην αγορά δεξαμενής καυσίμων είδαν ελάχιστες διαφορές στα μεταφορικά τους κόστη την περίοδο 2007 – 2008. Στα πλαίσια συζητήσεων στην Ευρωπαϊκή Επιτροπή είχε προταθεί να επιτραπεί η χρήση μακριών βαρέων οχημάτων (25.5m μήκος και 60tns φορτίο), ωστόσο η πρόταση απορρίφθηκε εξαιτίας των κινδύνων που ενέχουν τέτοιου τύπου φορτηγά.

Όσον αφορά στις σιδηροδρομικές μεταφορές λόγω των τεχνολογικών εξελίξεων και των μειώσεων στα κόστη μεταφοράς την τελευταία δεκαετία οι επιπτώσεις λόγω αύξησης των ενεργειακών κοστών είναι περιορισμένες. Όπως, αναφέρεται στη συγκεκριμένη έρευνα υπήρξαν δυσκολίες στην εξαγωγή συμπερασμάτων για τον κλάδο αλλά κατά κύριο λόγο επηρεάστηκαν περισσότερο οι σιδηρόδρομοι που χρησιμοποιούν καύσιμο ντίζελ για την κίνησή τους. Δεδομένου ότι οι σιδηροδρομικές εταιρίες κατέχουν σημαντικό μερίδιο στην κλάδο των μεταφορών και είναι η μοναδική λύση για κάποιες περιοχές (π.χ. Αλπική ζώνη), μπόρεσαν να μεταφέρουν την αύξηση του κόστους στους πελάτες τους. Έτσι, κατάφεραν να μην συρρικνωθούν τα ποσοστά των κερδών τους. Σημειώνεται επίσης, ότι την περίοδο 2004 – 2008 στην Γερμανία επιβλήθηκαν επιπλέον φόροι στις σιδηροδρομικές μεταφορές που είχαν σαν αποτέλεσμα να αυξηθούν οι τιμές στις σιδηροδρομικές μεταφορές. Γίνεται αντιληπτό ότι οι σιδηροδρομικές μεταφορές έχουν πολύ μεγαλύτερη ευκολία να αναδιαμορφώνουν τις τιμές των δρομολογίων τους σε σύγκριση με τους οδικούς μεταφορείς κυρίως λόγω μικρότερου ανταγωνισμού.

Προτού γίνει ανάλυση των επιπτώσεων των μεταβολών των καυσίμων στις αερομεταφορές αξίζει να αναφερθεί ότι οι εμπορικές μεταφορές αντιπροσωπεύουν το 30% των συνολικών μεταφορών, ενώ το υπόλοιπο 70% αντιστοιχεί στις επιβατικές αερομεταφορές. Στην παραπάνω έρευνα που έγινε για λογαριασμό του Ευρωπαϊκού Κοινοβουλίου οι αεροπορικές εταιρίες παρατήρησαν το κόστος των καυσίμων να διπλασιάζεται από το 2004 στο 2007. Το κύριο πρόβλημα για αυτές τις εταιρίες ήταν η πολύ απότομη αύξηση των τιμών των καυσίμων, καθώς μεγάλο μέρος του ισολογισμού τους είναι προπωλημένο και έτσι οποιαδήποτε αύξηση των τιμών των εισιτηρίων έχει μικρή επιρροή. Ακόμη,

στρατηγικές αντιστάθμισης για τα καύσιμα με συμβόλαια μελλοντικής εκπλήρωσης άργησαν να υιοθετηθούν από τις αεροπορικές εταιρίες. Για την αντιμετώπιση του προβλήματος οι αεροπορικές εταιρίες περιόρισαν τα έξοδα διαφήμισης και συντήρησης κατά 9,2% και 0,8% αντίστοιχα, κατά την παραπάνω περίοδο. Κάποιες αεροπορικές εταιρίες, όπως η EasyJet και η Ryanair ανακοίνωσαν ότι περιόρισαν τα κόστη ανά θέση 13% και 6% αντίστοιχα κατά την περίοδο 2005-2007. Στην παραπάνω μείωση έχει αφαιρεθεί το κόστος του καυσίμου. Ένα άλλο μέτρο που υιοθετήθηκε λόγω των μεταβολών των τιμών του καυσίμου είναι η πώληση των ακριβών και κοστοβόρων αεροσκαφών. Οι αεροπορικές εταιρίες μεταφοράς εμπορευμάτων μετακύλισαν τα αυξημένα κόστη μεταφοράς στους πελάτες τους, με αποτέλεσμα μέχρι και να διπλασιαστούν τα κόστη μεταφοράς σε ορισμένες περιπτώσεις.

Επιπρόσθετα, γίνεται μια εκτίμηση των μελλοντικών επιπτώσεων που θα υπάρξουν από μια πιθανή μελλοντική αύξηση των τιμών των καυσίμων. Δεδομένου ότι ο κλάδος των μεταφορών θεωρείται ανελαστικός οποιαδήποτε μεταβολή θα είναι σημαντική για τον κλάδο. Υπενθυμίζεται ότι βραχυπρόθεσμα οποιεσδήποτε αναπροσαρμογές στις μεταφορές δεν μπορούν να αντισταθμίσουν τις απότομες αυξομειώσεις στις τιμές των καυσίμων. Η αντικατάσταση των παραγώγων του πετρελαίου με άλλες μορφές ενέργειας μπορεί να γίνουν μόνο μακροπρόθεσμα και ακόμη υπάρχουν τεχνολογικές δυσκολίες για αυτήν την μετάβαση. Συγκεκριμένα, επισημαίνεται ότι οι αντιδράσεις του επιβατικού κοινού σε αυξομειώσεις των τιμών των καυσίμων θα είναι άμεσες με κύριο χαρακτηριστικό το εξής: η επιβατική κίνηση θα διασπαστεί σε διάφορα μέσα μεταφοράς με το αυτοκίνητο και τις αέριες μεταφορές να μειώνουν τα ποσοστά τους, ενώ τα μέσα μαζικής μεταφοράς να παρατηρούν μικρή αύξηση στην ζήτηση. Η ζήτηση στην χρήση αυτοκινήτων θα συρρικνωθεί από το 71% στο 67% μέχρι το 2020. Οι αέριες μεταφορές θα σημειώσουν μείωση στο ποσοστό που κατέχουν 20% κατά την περίοδο 2014 – 2020. Αντίθετα, οι σιδηροδρομικές μεταφορές εκτιμάται ότι θα ανέλθουν από το 8% στο 12% μέχρι το 2020. Όλα τα παραπάνω αναφέρονται σε σενάρια ανόδου της τιμής των καυσίμων.

Επιπτώσεις στην Γερμανία

Σε έρευνα που έγινε ειδικά στην Γερμανία για την περίοδο 1996 – 2007 διερευνήθηκαν οι επιπτώσεις των μεταβολών των τιμών των καυσίμων και των εισιτηρίων στην χρήση των δημοσίων μέσων μαζικής μεταφοράς (Frondel and Vance, 2010). Προτού γίνει ανάλυση των συμπερασμάτων της παραπάνω έρευνας αξίζει να αναφερθεί ότι η Γερμανία έχει θεσπίσει μέτρα για την προώθηση των μέσων μαζικής μεταφοράς και τον περιορισμό των εκπομπών διοξειδίου του άνθρακα (CO_2). Παρ' όλα αυτά ενώ έχει πετύχει να αυξήσει τις εκπομπές διοξειδίου του άνθρακα μόνο 1% κατά την περίοδο 1990 – 2005 σε σύγκριση με αύξηση 26% σε όλη την Ε.Ε., η χρήση των μέσων μαζικής μεταφοράς έχει παραμείνει σταθερή ή έχει μειωθεί κατά την παραπάνω περίοδο. Η ανάλυση έγινε με δεδομένα που συλλέχθηκαν από τα νοικοκυριά. Τα αποτελέσματα έδειξαν ότι μια αύξηση της τιμής της απλής αμόλυβδης κατά δύο τρίτα οδηγεί σε αύξηση της χρήσης μέσων μαζικής μεταφοράς κατά 0,7 διαδρομές την εβδομάδα με επίπεδο σημαντικότητας 1%. Ακόμη, επισημαίνεται ότι η επίδραση της τιμής του εισιτηρίου είναι στατιστικά μη σημαντική για όλες τις περιοχές που

έγινε η έρευνα. Έτσι, εξάγεται το συμπέρασμα ότι η επίδραση των μεταβολών της τιμής των καυσίμων είναι μεγαλύτερη στη χρήση των μέσων μαζικής μεταφοράς και αυξάνει μόνο τον ωριαίο κυκλοφοριακό φόρτο στα συγκεκριμένα μέσα μεταφοράς.

Επιπτώσεις στην Ελλάδα

Στην Ελλάδα μετά την έναρξη της οικονομικής κρίσης το 2008 άρχισε μια σφοδρή επιβολή φόρων για την αποκατάσταση του πλεονάσματος. Έτσι, υπήρξε σημαντική αύξηση στην τιμή της βενζίνης και του ντίζελ, καθώς οι φόροι αντιστοιχούσαν πλέον στο 82% και 31% της βενζίνης και του ντίζελ αντίστοιχα. Διερευνήθηκαν οι συνέπειες αυτών των μεταβολών στην κυκλοφοριακή ροή κατά μήκος του αυτοκινητοδρόμου Αθήνα – Τσακώνα (Musso et al., 2013). Παρατηρήθηκε ότι κατά την περίοδο 2006-2011 η κυκλοφοριακή ροή στον αυτοκινητόδρομο και στα δύο ρεύματα μειώθηκε σημαντικά, περίπου 20%, ενώ η κυκλοφοριακή ροή στον παράδρομο παρέμεινε σχεδόν αμετάβλητη. Σημειώνεται ότι η χειρότερη περίοδος ήταν μεταξύ 2010 – 2011 κατά την οποία η κυκλοφοριακή κίνηση μειώθηκε 10%. Επίσης, η έρευνα έδειξε ότι τα ελαφρά οχήματα έχουν μεγαλύτερη ελαστικότητα σε σχέση με τα βαρέα οχήματα, κάτι που υποδηλώνει ότι στην Ελλάδα οι οδηγοί επιβατικών αυτοκινήτων περιορίζουν τις διαδρομές τους σε περίπτωση αύξησης των τιμών των καυσίμων. Επιπλέον, αξίζει να αναφερθεί ότι στην Ελλάδα υπάρχει μεγάλη συσχέτιση μεταξύ του εισοδήματος και της κυκλοφοριακής ροής, κάτι που είναι λογικό καθώς το μεγαλύτερο μέρος της κοινωνίας στην χώρα έχει πολύ χαμηλό εισόδημα. Κατά συνέπεια, είναι λογικό μετά την έναρξη της κρίσης το 2008 και την πτώση των εισοδημάτων σε συνδυασμό με την αύξηση της τιμής των καυσίμων λόγω φορολογίας να μειωθεί σε τέτοιο βαθμό η κυκλοφοριακή κίνηση στον αυτοκινητόδρομο Κορίνθου – Πατρών.

2.1.2 ΕΠΙΠΤΩΣΕΙΣ ΣΤΗΝ ΑΥΣΤΡΑΛΙΑ

Σε έρευνα που έγινε στην Αυστραλία για τις επιπτώσεις στον τομέα των μεταφορών λόγω μεταβολών στις τιμές των καυσίμων αναφέρεται ότι το 72% της κατανάλωσης του πετρελαίου χρησιμοποιείται στις μεταφορές (Rickwood, P., 2010). Ο ίδιος αναφέρει ότι για να κατανοήσουμε την επιρροή των καυσίμων στις μεταφορές χρειάζεται να κατανοηθούν τρία πράγματα: α) πόσο ευμετάβλητη είναι η ζήτηση σε κάθε μέσο μεταφοράς (ελαστικότητα της ζήτησης), β) κατά πόσο μπορούν να χρησιμοποιηθούν άλλες μορφές ενέργειας, γ) κατά πόσο η ζήτηση για ένα μέσο μεταφοράς μπορεί να αντικατασταθεί από ένα άλλο μέσο μεταφοράς. Στον Πίνακα 2.1 παρατίθονται αναλυτικά όλες αυτές οι πληροφορίες ανάλογα με τον σκοπό της μεταφοράς. Για παράδειγμα η αύξηση της τιμής του καυσίμου έχει ως αποτέλεσμα την μείωση της χρήσης ιδιωτικών αυτοκινήτων βραχυπρόθεσμα, ενώ μακροπρόθεσμα αυτό θα ακολουθηθεί από μια τάση των καταναλωτών για οχήματα εξοικονόμησης ενέργειας. Χρησιμοποιώντας τον πίνακα 2.1 για

μία αύξηση στις τιμές των καυσίμων κατά 60% φαίνεται ότι θα υπάρχει μείωσης στα ταξίδια αναψυχής από και προς την Αυστραλία περίπου 50%, ενώ η μείωση στα ταξίδια με επιχειρηματικούς σκοπούς θα είναι 25%. Υπενθυμίζεται ότι τα νούμερα υπολογίζονται με την παραδοχή ότι το κόστος της βενζίνης αποτελεί το 25% του μεταφορικού κόστους.

Σε μακροπρόθεσμο ορίζοντα 3 – 5 ετών αναμένεται μείωση στην κατανάλωση καυσίμου 64%, η οποία αναλύεται σε 30% μείωση των ταξιδιών και σε 50% βελτίωσης της απόδοσης των οχημάτων. Σημειώνεται ότι η τελευταία πληροφορία είναι ευμενής για τόσο σύντομο χρονικό διάστημα, καθώς για τόσο μεγάλη αύξηση στα οχήματα εξοικονόμησης ενέργειας απαιτούνται περίπου 20 χρόνια. Αξιοσημείωτο είναι ότι στις επικρατούσες συνθήκες οι κάτοικοι της Αυστραλίας ακόμη και σε αύξηση της τιμής των καυσίμων δύσκολα θα αντικαταστήσουν την χρήση του αυτοκινήτου με την χρήση δημόσιων μέσων μαζικής μεταφοράς, όπως το λεωφορείο. Για να ξεπεραστεί αυτή η δυσκολία πρέπει η κυβέρνηση να προχωρήσει σε ανανέωση των υπαρχόντων λεωφορείων.

Πίνακας 2.1: Τιμές ελαστικότητας για κάθε μέσο κυκλοφορίας ανάλογα το σκοπό χρήσης, (Πηγή: Rickwood, 2010)

	elasticity (short-run)	elasticity (long-run)	Fuel substitutes	Other substitutes	Likely response to high oil prices
Business air travel	-1.1 (Nairn and Hooper, 1992, p55)	Unclear. Likely somewhat (but not greatly) higher than short-run.	None	Video-conferencing	Demand reduction. Longer-term -- reorganization of business processes for more video-conferencing.
Leisure air travel	-3.2 (Lubulwa, 1986, p208) -2.3 (Nairn and Hooper, 1992, p55)	Unclear. Likely similar to short run	None	Domestic holidays	Demand reduction.
Private passenger vehicles	-0.25 (Goodwin et al., 2004) -0.25 (Graham and Glaister, 2004) -0.15 (Gargett and Hossain, 2008)	-0.64 (Goodwin et al., 2004) -0.77 (Graham and Glaister, 2004) -0.4 (Gargett and Hossain, 2008)	Other liquid fuels Electricity Hydrogen	Public transport	Mixed. Short term some demand reduction, some mode switch. Longer term unclear except for lower vehicle fuel consumption.
Commercial trucking	Depends on good being trucked. -7.42 to 1.72 (Graham and Glaister, 2004) – international mean -1.07, std. dev. 0.84)	Unclear, but higher than short-run.	Other liquid fuels	Rail freight (imperfect)	Reduced demand for a few goods (i.e. price-elastic low value, bulky goods) Some transfer to rail and sea where possible (i.e. non-perishables).

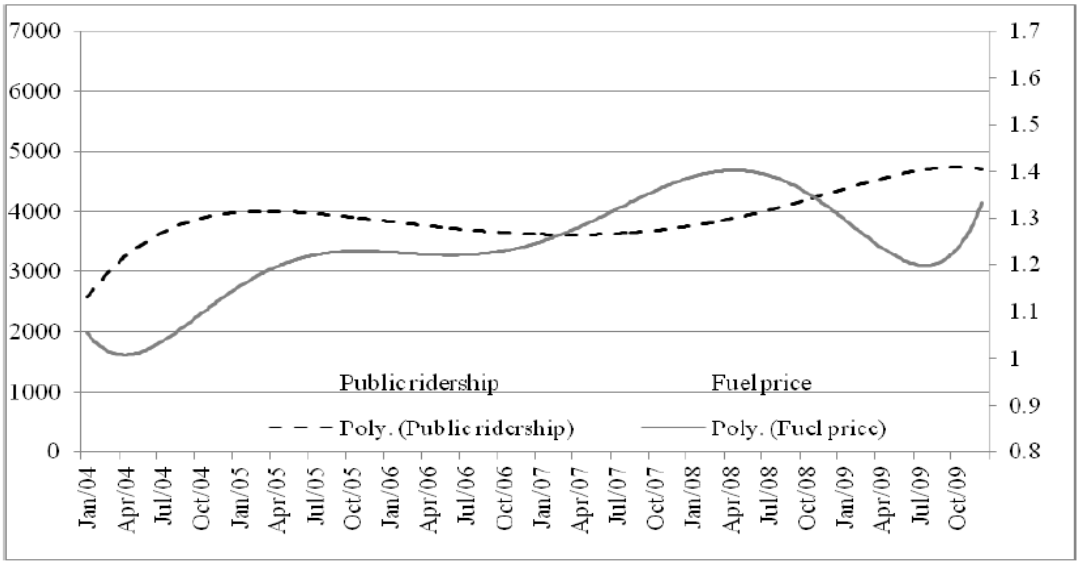
Σχετικά με τις οδικές μεταφορές αναφέρεται ότι κατά κύριο λόγο τα προϊόντα που μεταφέρουν απαιτούν μικρό χρόνο παράδοσης και για αυτό οι οδικές μεταφορές είναι η μοναδική λύση. Εκμεταλλευόμενοι αυτή την κατάσταση οι εταιρίες μεταφέρουν την αύξηση του κόστους στον τελικό πελάτη, αφήνοντας αμετάβλητα έτσι τα κέρδη τους. Φαίνεται λοιπόν ότι οι οδικοί μεταφορείς στην Αυστραλία πλεονεκτούν έναντι των Ευρωπαίων που λόγω του μεγάλου ανταγωνισμού αδυνατούν να αυξήσουν τις τιμές των δρομολογίων. Συμπεραίνεται λοιπόν από όλα τα παραπάνω ότι η κατανάλωση του πετρελαίου είναι συγκεντρωμένη στον τομέα των μεταφορών και βραχυπρόθεσμα δεν υπάρχει τρόπος να αντικατασταθεί από άλλες μορφές ενέργειας. Επίσης, φαίνεται ότι μια αύξηση στις τιμές των καυσίμων επηρεάζει περισσότερο τα νοικοκυριά και τις επιχειρήσεις και έτσι χρήζει άμεσης προτεραιότητας η λήψη προληπτικών μέτρων για την προστασία του κοινωνικού συνόλου.

2.1.3 ΕΠΙΠΤΩΣΕΙΣ ΣΤΗΝ ΔΑΝΙΑ

Στην Δανία σε έρευνα που έγινε κατά την περίοδο 2004 – 2011 στον κλάδο των οδικών μεταφορών διαπιστώθηκε ότι η αύξηση των τιμών των καυσίμων οδηγεί σε μείωση της μέσης απόστασης ταξιδιού (Abate, M., 2014). Όπως αναφέρεται, αυτός είναι ένας τρόπος βελτίωσης της απόδοσης για την εξοικονόμηση καυσίμου. Τα αποτελέσματα προέκυψαν από δεδομένα για βαρέα φορτηγά από το 2004 – 2011. Εκτιμήθηκε ότι για κάθε 1 Κορώνα (δανέζικο νόμισμα ισοδυναμεί σε 0,18 δολάρια) υπάρχει μείωση 0,4% έως 0,7% στην μέση απόσταση ταξιδιού κατά την περίοδο 2004 – 2007. Ωστόσο, αυτά τα αποτελέσματα δεν επιβεβαιώνονται για την περίοδο μετά την έναρξη της οικονομικής κρίσης το 2008. Άλλη μία συνέπεια της αυξημένης τιμής των καυσίμων είναι η αύξηση του αριθμού των φορτωμένων δρομολογίων ανά περίοδο χρήσης. Επισημαίνεται ότι αυτή η επίπτωση είναι σημαντική κατά την περίοδο 2004 – 2007. Στην παραπάνω έρευνα υπάρχουν όμως και περιορισμοί: α) λόγω της οικονομικής κρίσης μετά το 2008 ήταν δύσκολο να συμπεραθούν ισχυρές σχέσεις αιτίου και αιτιατού, β) λόγω έλλειψης δεδομένων ήταν δύσκολο να συνυπολογιστούν αν οι αυξήσεις των τιμών των καυσίμων ήταν αναμενόμενες από τις παραπάνω εταιρίες. Αν κάτι τέτοιο ίσχυε οι εταιρίες οδικών μεταφορών ενδέχεται να είχαν αντισταθμίσει τον κίνδυνο με συμβόλαια μελλοντικής εκπλήρωσης, αφήνοντας

ανεπηρέαστες πολλές από αυτές σε ενδεχόμενες διακυμάνσεις της τιμής του ντίζελ. Επιπρόσθετα στην παραπάνω έρευνα δεν ήταν εφικτό να εντοπιστεί αν η στροφή των οδικών μεταφορέων σε μικρότερες διαδρομές οφειλόταν αποκλειστικά στην αύξηση της τιμής του ντίζελ ή υπήρχε ισχυρός ανταγωνισμός από άλλα μέσα μεταφοράς, όπως πλοίο, αεροπλάνο.

Άλλη μία έρευνα που έγινε στην Δανία αφορούσε τις επιπτώσεις που έχουν οι μεταβολές των τιμών των καυσίμων σε κάθε πολίτη ή νοικοκυριό ξεχωριστά (Yang and Timmermans, 2012). Τα αποτελέσματα έδειξαν ότι υπάρχει μια «εμμονή» οι πολίτες να μην αλλάζουν τις συνήθειες τους, αλλά να αναπροσαρμόζουν τα κόστη τους. Η προσαρμογή των πολιτών στις υψηλές τιμές των καυσίμων περιλαμβάνει ακόμη και αντικατάσταση κάποιων διαδρομών με άλλες δραστηριότητες. Επιπρόσθετα, από την διερεύνηση που έγινε προέκυψε ότι οι επιπτώσεις στα μέσα μαζικής μεταφοράς, λόγω των αυξομειώσεων των τιμών των καυσίμων, διαφέρουν ανάλογα το σκοπό της διαδρομής. Για παράδειγμα, σε διαδρομές που είναι υποχρεωτικές οι επιπτώσεις είναι ελάχιστες σε επίπεδο μήνα. Όμως ,σε βάθος τριμήνου, η προσαρμογή στα μέσα μαζικής μεταφοράς αρχίζει να διαφαίνεται. Αντίθετα, σε περιπτώσεις διαδρομών αναψυχής οι πολίτες αλλάζουν τις συνήθειες τους άμεσα και παρατηρείται αύξηση στη χρήση των μέσων μαζικής μεταφοράς από την συγκεκριμένη κατηγορία πολιτών. Για την καλύτερη κατανόηση των επιπτώσεων των μεταβολών των τιμών των καυσίμων στη ζήτηση των μέσων μαζικής μεταφοράς παρατίθεται το Διάγραμμα 2.3, όπου απεικονίζεται η επιβατική κίνηση για τα μέσα μαζικής μεταφοράς, καθώς και της τιμής της απλής αμόλυβδης για την περίοδο 2003 – 2009.



Διάγραμμα 2.3: Τιμές της επιβατικής κίνησης για τα μέσα μαζικής μεταφοράς και της απλής αμόλυβδης κατά την περίοδο 2004 – 2009 (Πηγή: Abate et al., 2014)

2.1.4 ΕΠΙΠΤΩΣΕΙΣ ΣΤΙΣ ΗΝΩΜΕΝΕΣ ΠΟΛΙΤΕΙΕΣ

Λόγω της αύξησης της τιμής των καυσίμων από το 1999 αποφασίστηκε για 10 Πολιτείες της Αμερικής να διερευνηθούν οι επιπτώσεις στα κυκλοφοριακά μέσα κατά την περίοδο 2002 – 2011 (Iseki and Ali, 2014). Από τα αποτελέσματα της ανάλυσης προέκυψε ότι βραχυπρόθεσμα η επίδραση των αυξημένων τιμών των καυσίμων οδηγεί σε μικρή αύξηση της χρήσης των λεωφορείων, ενώ η ζήτηση του επιβατικού κοινού για την χρήση των σιδηροδρόμων μένει σχεδόν ανεπηρέαστη. Επίσης, από τα αποτελέσματα που εξάγονται φαίνεται ότι η ελαστικότητα της επιβατικής ζήτησης σε αλλαγές της τιμής του εισιτηρίου είναι μεγαλύτερη από ότι αυτή των λεωφορείων. Κατά συνέπεια, συμπεραίνεται ότι για τις 10 πολιτείες της Αμερικής που χρησιμοποιήθηκαν στην παραπάνω έρευνα φαίνεται ότι οι πολίτες δείχνουν να μην επηρεάζονται σημαντικά από τις μεταβολές στις τιμές των καυσίμων. Οι ελαστικότητες της επιβατικής ζήτησης ως προς τις μεταβολές των τιμών των καυσίμων ήταν θετικές σε όλα τα μοντέλα με παρόμοιες τιμές για όλα τα μέσα μεταφοράς. Μόνο στους σιδηροδρόμους οι μεταβολές των καυσίμων αφήνουν ανεπηρέαστη την επιβατική ζήτηση. Το γεγονός αυτό υποδηλώνει ότι οι πολίτες για να αντιμετωπίσουν την αύξηση του μεταφορικού κόστους με τα επιβατικά αυτοκίνητα, χρησιμοποιούν τα λεωφορεία. Έτσι, είναι κατανοητό ότι οι επιβάτες των λεωφορείων έχουν μεγαλύτερη ευαισθησία σε μεταβολές των τιμών των καυσίμων, κάτι που μπορεί να εξηγηθεί από τα χαμηλά εισοδήματα εκείνων που χρησιμοποιούν τα λεωφορεία ως μέσο μετακίνησης. Επισημαίνεται ότι παρ΄όλο που οι ποσοστιαίες μεταβολές στην επιβατική κίνηση είναι

μικρές, αυτές οι μεταβολές απαιτούν μεγάλη αύξηση της προσφοράς των διαφόρων μέσων. Συνεπώς, οι άνθρωποι που ασχολούνται με την διαχείριση των μέσων μαζικής μεταφοράς πρέπει να παίρνουν κατάλληλα μέτρα εκ των προτέρων (π.χ. αύξησης της συχνότητας, χρήση μέσων μεγάλης χωρητικότητας), προκειμένου να αντιμετωπίζουν αποτελεσματικά τις επιπτώσεις από τις μεταβολές των καυσίμων.

Με τις επιπτώσεις των μεταβολών των καυσίμων στην επιβατική κίνηση στο Σικάγο ασχολήθηκαν οι Nowak και Savage (2013). Η έρευνά τους αναφέρεται στην επιβατική κίνηση στα λεωφορεία και τον σιδηρόδρομο. Η συνεισφορά τους στην υπάρχουσα βιβλιογραφία είναι σημαντική, καθώς έδειξαν ότι αναλόγως του εύρους της τιμών καυσίμων οι ελαστικότητες διαφέρουν σημαντικά. Συγκεκριμένα, έδειξαν ότι όταν η τιμή της βενζίνης είναι κάτω από 3 δολάρια/γαλόνι οι ελαστικότητες είναι πολύ μικρές και έχουν εύρος 0,02 έως 0,05. Όταν, η τιμή της βενζίνης είναι από 3 έως 3,99 δολάρια/γαλόνι, τότε η ελαστικότητα της ζήτησης στους σιδηροδρόμους αυξάνεται σε 0,12 έως 0,14. Οι ίδιοι επισημαίνουν ότι κατά την περίοδο Μαΐου – Αυγούστου 2008, όπου η τιμή της βενζίνης ξεπερνούσε τα 4 δολάρια, οι ελαστικότητες της ζήτησης για τα λεωφορεία και τους σιδηροδρόμους εκτιμήθηκαν σε 0.28 – 0.30 και 0.37 αντίστοιχα. Αυτές οι τιμές της ελαστικότητας είναι λίγο μεγαλύτερες από εκείνες κατά την περίοδο της κρίσης του αργού πετρελαίου στην αρχή του 1980. Αξίζει να σημειωθεί ότι δεν είναι σίγουρο αν οι συγκεκριμένες τιμές της ελαστικότητας για τα παραπάνω μέσα μεταφοράς θα μπορούσαν να έχουν διατηρηθεί σε αυτά τα επίπεδα αν δεν είχε μειωθεί τόσο σημαντικά η τιμή του αργού πετρελαίου στα τέλη του 2008. Η παρατήρηση αυτή οφείλονται στο γεγονός ότι αν οι τιμές του αργού πετρελαίου είχαν παραμείνει τόσο ψηλά ενδέχεται οι πολίτες να αναζητούσαν άλλους τρόπους μεταφοράς (π.χ. ομαδικές μετακινήσεις) ή εναλλακτικές διαδρομές.

Την σχέση μεταξύ των μεταβολών των τιμών των καυσίμων και την θέληση των πολιτών να επενδυθούν δημόσια χρήματα για την βελτίωση των μέσων μαζικής μεταφοράς διερεύνησε ο Smart (2014). Στην έρευνα του χρησιμοποίησε μια βάση δεδομένων δημόσιων γνώμεων που είχε δημιουργηθεί από το κράτος για την περίοδο 1984 – 2012. Τα αποτελέσματα της έρευνάς του έδειξαν ότι όσο πιο πολύ αυξάνονται οι μεταβολές των τιμών των καυσίμων τόσο περισσότερο είναι διατεθειμένοι οι Αμερικάνοι πολίτες να ενισχύσουν τα έξοδα που αφορούν τα μέσα μαζικής μεταφοράς. Όπως, προκύπτει οι επιπτώσεις είναι σοβαρές και στατιστικώς σημαντικές. Οι πολίτες θεωρούν ότι τα μέσα μεταφοράς μπορούν να απορροφήσουν τις μεταβολές των τιμών των καυσίμων. Επίσης, η έρευνα επιβεβαιώνει τα αποτελέσματα και προηγούμενων ερευνών που αναφέρουν ότι οι αυξήσεις στις τιμές των καυσίμων έχουν μικρή επίδραση στην αύξηση της επιβατικής κίνησης στα μέσα μαζικής μεταφοράς. Η αιτία σε αυτό, όπως προκύπτει από την συγκεκριμένη έρευνα είναι ότι οι πολίτες επιθυμούν καλύτερες μεταφορικές υπηρεσίες σε σύγκριση με τις υπάρχουσες. Κατά συνέπεια, όταν αυξάνονται οι τιμές των καυσίμων λίγοι είναι εκείνοι που αλλάζουν μέσο μεταφοράς αλλά ταυτόχρονα αυξάνεται ο αριθμός εκείνων που επιθυμούν καλύτερες μεταφορικές υπηρεσίες.

2.2 ΓΡΑΜΜΙΚΑ ΚΑΙ ΜΗ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΚΑΥΣΙΜΩΝ ΚΑΙ ΤΟΥ ΑΡΓΟΥ ΠΕΤΡΕΛΑΙΟΥ

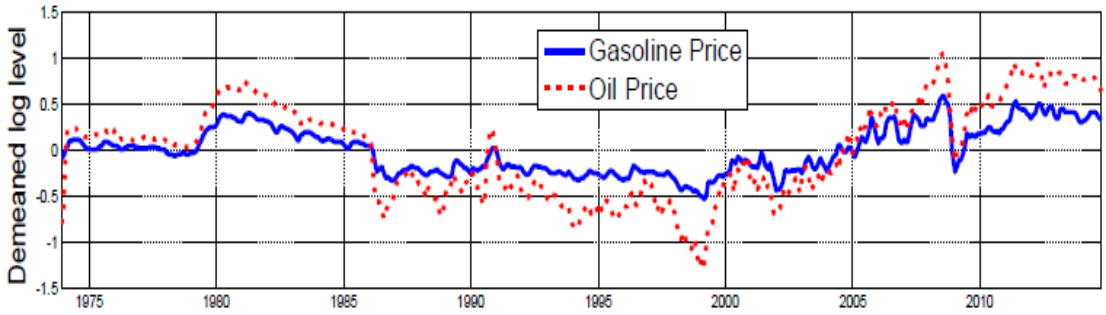
2.2.1 ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΤΗΣ ΤΙΜΗΣ ΤΗΣ ΑΜΟΛΥΒΔΗΣ

Στις Ηνωμένες Πολιτείες έχει διερευνηθεί κατά πόσο ο δείκτης καταναλωτικής εμπιστοσύνης του Michigan (MSC) μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών της αμόλυβδης (Kellogg et al.,2011). Σημειώνεται ότι ο δείκτης MSC έχει μηνιαία χρονοσειρά. Οι ερευνητές χρησιμοποίησαν την μέση τιμή των απαντήσεων του συγκεκριμένου δείκτη για να μπορέσουν να εξάγουν συμπεράσματα σχετικά με την ικανότητα πρόβλεψης της μέσης τιμής της αμόλυβδης. Για την περίοδο 1993 – 2010 φαίνεται ότι η απάντηση του μέσου καταναλωτή ακολουθεί την χρονοσειρά της αμόλυβδης, με εξαίρεση την περίοδο της οικονομικής κρίσης στα τέλη του 2008, όπου υπάρχει σημαντική διαφοροποίηση. Από τα αποτελέσματα το σφάλμα πρόβλεψης χρησιμοποιώντας τον παραπάνω δείκτη είναι περίπου ίσο με το σφάλμα που θα προέκυπτε αν χρησιμοποιούταν το μοντέλο μη αλλαγής της τιμής. Η μόνη διαφορά είναι ότι την περίοδο της οικονομικής κρίσης το σφάλμα του δείκτη MSC είναι πολύ μικρότερο από το σφάλμα του μοντέλου μη αλλαγής. Επίσης, αξίζει να σημειωθεί ότι περίοδοι έντονης μεταβλητότητας της τιμής των καυσίμων συσχετίζονται με περιόδους έντονης μεταβλητότητας στις απαντήσεις που συγκεντρώνονται για την κατασκευή του δείκτη MSC.

Οι Baumeister et al. (2015) διερεύνησαν την εφαρμογή νέων μεθόδων για την πρόβλεψη της λιανικής τιμής της αμόλυβδης. Οι ίδιοι τονίζουν ότι η πρόβλεψη της τιμής της αμόλυβδης είναι μείζων θέμα για τους πολίτες, τους κατασκευαστές αυτοκινήτων, κεντρικούς τραπεζίτες και τους ανθρώπους που σχεδιάζουν την ενεργειακή πολιτική. Έτσι, παρ' ότι το θέμα είναι ύψιστης σημασίας για την κοινωνία αναφέρουν ότι υπάρχει ελάχιστη διαθέσιμη βιβλιογραφία σχετικά με την πρόβλεψη της λιανικής τιμής της αμόλυβδης, καθώς οι ερευνητές θεωρούν ότι με τις πληροφορίες και τα δεδομένα που είναι διαθέσιμα στους πολίτες δεν μπορούν να γίνουν ακριβείς προβλέψεις.

Στην συγκεκριμένη έρευνα οι ερευνητές δημιούργησαν έξι μοντέλα, τα οποία και αξιολόγησαν. Το πρώτο ήταν το ολοκληρωμένο αυτοπαλινδρούμενο μοντέλο κινητού μέσου όρου ARMA(1,1), το οποίο βασίστηκε αποκλειστικά στην χρονοσειρά της λιανικής τιμής των καυσίμων. Το δεύτερο ήταν ένα μοντέλο παλινδρόμησης AR(p) που βασίστηκε στην χρονοσειρά της λιανικής τιμής της βενζίνης. Ακόμα, βρέθηκε ο αριθμός του p(p=1,…….,12) που εξασφαλίζει την μεγαλύτερη ορθότητα των προβλέψεων που είναι για

p=1, δηλαδή το μοντέλο AR(1). Καθώς από το διάγραμμα της λιανικής τιμής της αμόλυβδης φαίνεται ότι η χρονοσειρά δεν έχει κάποια τάση, οι ερευνητές θεώρησαν εύλογο να ελέγξουν το μοντέλο της εκθετικής ομαλοποίησης, δηλαδή: $r_t = ar_t + (1-a)r_{t-1}$, με το a να παίρνει την τιμή 0,8 συνήθως. Το τέταρτο μοντέλο δημιουργήθηκε έχοντας ως ανεξάρτητη μεταβλητή την διαφορά της τιμής μεταξύ αμόλυβδης και αργού πετρελαίου. Το πέμπτο μοντέλο ήταν ένα μοντέλο απλής γραμμικής παλινδρόμησης που βασίστηκε σε μια μεταβλητή, CFNAI, η οποία αποτελείται από 85 δείκτες που αντικατοπτρίζουν την ανάπτυξη ή επιβράδυνση της οικονομικής δραστηριότητας στις Ηνωμένες Πολιτείες. Το έκτο μοντέλο ήταν ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης VAR(p) με δύο ανεξάρτητες μεταβλητές, το αργό πετρέλαιο και την αμόλυβδη. Διερευνήθηκε ο αριθμός p για τον οποίο εξασφαλίζεται η μεγαλύτερη ορθότητα πρόβλεψης και προέκυψε ότι η χρήση μόνο της προηγούμενης παρατήρησης, δηλαδή p=1, επιτυγχάνει την μεγαλύτερη ακρίβεια πρόβλεψης. Για την κατανόηση των μεταβολών των δυο ανεξάρτητων μεταβλητών που χρησιμοποιούνται στο έκτο μοντέλο παρατίθενται στο Διάγραμμα 2.4 οι χρονοσειρές τους για την περίοδο 1973 – 2014 σε επίπεδα log.



Διάγραμμα 2.4: Λιανική τιμή αμόλυβδης και αργού πετρελαίου στις Ηνωμένες Πολιτείες κατά την περίοδο 1973 – 2014 (Πηγή: Baumaeister et al., 2015)

Πραγματοποιήθηκε σύγκριση των παραπάνω μοντέλων και βρέθηκε ότι τα μοντέλα με το μικρότερο MSPE ήταν το AR(1) και το VAR(1). Τα δύο τελευταία μοντέλα χρησιμοποιήθηκαν και για προβλέψεις με μεγαλύτερο χρονικό ορίζοντα έως και 24 μήνες. Όσο αυξανόταν ο χρονικός ορίζοντας οι ερευνητές παρατήρησαν ότι βελτιωνόταν η ακρίβεια των προβλέψεων. Για το χρονικό ορίζοντα των 24 μηνών βρέθηκε ότι το MSPE μειώνεται στο 31% και η ακρίβεια της κατεύθυνσης πρόβλεψης φτάνει στο 73%. Αξίζει να αναφερθεί ότι χρησιμοποιώντας τα δύο τελευταία μοντέλα θα μπορούσε να είχε προβλεφθεί το 39% της πτώσης της τιμής της αμόλυβδης στα τέλη του 2014.

Οι Liao et al. (2016) διερεύνησαν τον λόγο που οι μακροπρόθεσμες προβλέψεις (χρονικό ορίζοντα ενός έτους) του International Energy Agency (IEA) αποτυγχάνουν σε μεγάλο βαθμό. Αρχικά, τονίστηκε ότι οι ορθές προβλέψεις συμβάλλουν στην ενίσχυση της οικονομικής ευημερίας και την αποφυγή του κοινωνικού χάους. Από την έρευνα προκύπτει ότι τα σφάλματα των προβλέψεων οφείλονται στις παραδοχές που γίνονται για το ΑΕΠ και τον πληθυσμό, ενώ επισημαίνεται ότι οι τιμές των καυσίμων απαιτούν ιδιαίτερη προσοχή στην πρόβλεψη τους. Αναφέρεται ότι το ΑΕΠ αποτελεί μόνο το 20% του σφάλματος των προβλέψεων και το υπόλοιπο 80% απαιτεί περαιτέρω διερεύνηση.

2.2.2 ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΤΗΣ ΤΙΜΗΣ ΤΟΥ ΝΤΙΖΕΛ

Οι Nyongesa and Wagala (2015) ασχολήθηκαν με την πρόβλεψη της τιμής του ντίζελ στην Κένυα. Όπως αναφέρουν, η πρόβλεψη της τιμής του ντίζελ είναι θέμα μείζονος σημασίας για μια χώρα, καθώς μία χώρα θα μπορεί να ρυθμίσει κατάλληλα τα αποθέματά της, ώστε να διατηρεί τις χαμηλές τιμές του ντίζελ προς όφελος των πολιτών. Οι συγκεκριμένοι ερευνητές θεώρησαν ότι το κόστος του αργού πετρελαίου, η προσφορά και η ζήτηση του πετρελαίου στην παγκόσμια αγορά και η τιμή του συναλλάγματος δολάριο/σελίνι Κένυας είναι οι ανεξάρτητες μεταβλητές της τιμής του ντίζελ στην Κένυα. Ωστόσο, μόνο η μηνιαία χρονοσειρά της τιμής του ντίζελ θα χρησιμοποιηθεί από το μοντέλο πρόβλεψης. Αξίζει να αναφερθεί ότι υπάρχουν και άλλοι παράγοντες, όπως οι κυβερνητικές πολιτικές, το πολιτικό κλίμα, η γεωγραφική τοποθεσία και ο ανταγωνισμός που επηρεάζουν την τιμή του ντίζελ αλλά δεν μπορούν να χρησιμοποιηθούν άμεσα, καθώς είναι δύσκολο να ποσοτικοποιηθούν.

Για την καλύτερη κατανόηση των παραπάνω μεταβλητών θα γίνει μια εξήγηση για το πως επιδρούν στην διαμόρφωση της τιμής του ντίζελ. Οι κυβερνητικές πολιτικές επιδρούν σημαντικά, καθώς μία αύξηση στο φόρο των καυσίμων έχει άμεση επίπτωση στη τιμή του ντίζελ, αφού αποτελεί σημαντικό ποσοστό της τιμής του καυσίμου. Η τοπογραφική τοποθεσία της χώρας που εισάγει το ντίζελ είναι σημαντικός παράγοντας, καθώς όσο πιο μακριά είναι η συγκεκριμένη χώρα τόσο μεγαλύτερο είναι το κόστος μεταφοράς που τελικώς θα ενσωματωθεί στην τιμή του καυσίμου. Πολλές φορές οι εταιρίες εμπορίας καυσίμων για να προσελκύσουν πιθανούς πελάτες (χώρες) μειώνουν τις τιμές πώλησης με αποτέλεσμα να μειώνεται η τιμή του ντίζελ στα πρατήρια πώλησης καυσίμων.

Στην συγκεκριμένη έρευνα για την πρόβλεψη των χρονοσειρών ντίζελ χρησιμοποιήθηκε το ολοκληρωμένο αυτοπαλινδρούμενο μοντέλο κινητού μέσου όρου ARIMA(p,d,q) και η χρονοσειρά των τιμών του ντίζελ για την περίοδο 2005 – 2012. Από το διάγραμμα της αυτοσυσχέτισης (ACF) προκύπτει ότι η χρονοσειρά δεν είναι στάσιμη και για το λόγο αυτό εφαρμόζεται στα δεδομένα η μετατροπή $y_t = \ln(P_t) - \ln(P_{t-1})$ για την επίτευξη στασιμότητας. Κατασκευάστηκαν τα ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(2,1,2) και φάνηκε ότι το ARIMA(2,1,2) υπερτερεί έναντι των υπολοίπων έχοντας MSE = 0,00208. Η εξίσωση του μοντέλου είναι η παρακάτω:

$$Y_t = 0{,}0038 + 0{,}33804X_{t-1} + 0{,}01455X_{t-2} + 0{,}09682\varepsilon_{t-1} + 0{,}08861\varepsilon_{t-2} + \varepsilon_t \qquad (1)$$

Το συγκεκριμένο μοντέλο χρησιμοποιήθηκε για την πρόβλεψη 5 μηνών και οι προβλεπόμενες τιμές είχαν απόκλιση από τις πραγματικές από $3{,}9\varepsilon_t$ έως $9{,}6\varepsilon_t$. Το συμπέρασμα από την συγκεκριμένη έρευνα είναι ότι οι τιμές ντίζελ στην Κένυα χαρακτηρίζονται από μεγάλη μεταβλητότητα με τυχαίες διακυμάνσεις. Κατά συνέπεια, είναι δύσκολο να κατασκευαστεί ένα μοντέλο, το οποίο να παρέχει στους οδηγούς

αυτοκινήτων έγκυρα σήματα ανοδικών ή καθοδικών τάσεων της τιμής του ντίζελ στην Κένυα.

Ο Bajjalieh (2010) ανέπτυξε μοντέλα πολλαπλής γραμμικής παλινδρόμησης, ολοκληρωμένου αυτοπαλινδρούμενου κινητού μέσου όρου (ARIMA) και συνέκρινε τα αποτελέσματά του με το μοντέλο μη αλλαγής. Αρχικά, τόνισε την σημασία του ντίζελ στον αγροτικό τομέα, καθώς η παραγωγή αγροτικών προϊόντων απαιτεί μεγάλες ποσότητες καυσίμου ντίζελ. Αναφέρθηκε στο ρόλο της πρόβλεψης της τιμής του, ώστε να μπορούν οι αγρότες να αντισταθμίσουν την αύξηση του κόστους σε πιθανή αύξηση της τιμής του ντίζελ. Ακόμη, αναφέρει ότι πριν το 2002 οι τιμές των καυσίμων ήταν αρκετά σταθερές και μπορούσαν να προβλεφθούν, ενώ μετά το 2002 λόγω της ανάπτυξης καινούριων τεχνολογιών ήταν δύσκολη η πρόβλεψη της τιμής του ντίζελ. Επισημαίνει και ο ίδιος ότι υπάρχει ελάχιστη διαθέσιμη βιβλιογραφία σχετικά με την πρόβλεψη παράγωγων προϊόντων του αργού πετρελαίου, όπως είναι το ντίζελ. Η χρονοσειρά που χρησιμοποίησε για την εκμάθηση των συντελεστών του μοντέλου ήταν οι μηνιαίες τιμές του ντίζελ της περιόδου 1994 – 2002 και για σύνολο δοκιμής την περίοδο 2002 – 2008.

Οι προβλέψεις των μοντέλων γίνονται με χρονικό ορίζοντα έναν, δύο ή τριών μηνών και αξιολογούνται αντίστοιχα. Τα δύο πρώτα μοντέλα γραμμικής παλινδρόμησης χρησιμοποιούν ως μεταβλητές εισόδου συμβόλαια μελλοντικής εκπλήρωσης (futures) αργού πετρελαίου και πετρελαίου θέρμανσης αντίστοιχα, αφού για το ντίζελ δεν υπάρχουν συμβόλαια μελλοντικής εκπλήρωσης (futures) και τα παραπάνω προϊόντα είναι παρόμοια. Το τρίτο μοντέλο χρησιμοποιεί την χρονοσειρά των αποθεμάτων του ντίζελ θέλοντας να ενσωματώσει τις οποιεσδήποτε μεταβολές στην τιμή του ντίζελ λόγω μεταβολών στην προσφορά και την ζήτηση. Το μοντέλο της μη μεταβολής θεώρησε σταθερή τιμή του ντίζελ για όλη την περίοδο 1994 – 2008. Για την αξιολόγηση των μοντέλων στο σύνολο δοκιμής για την περίοδο 2002 -2008, εφαρμόστηκε η μέθοδος του κινητού «παραθύρου» των 4-ετών και η κυλιόμενη παλινδρόμηση για να μπορούν να ενσωματωθούν καινούριες συμπεριφορές στα μοντέλα πρόβλεψης. Η πρώτη μέθοδος χρησιμοποιεί σταθερό αριθμό παραδειγμάτων για την εύρεση των συντελεστών, ενώ η δεύτερη αυξάνει τον αριθμό των παραδειγμάτων, όπως περνάει ο χρόνος.

Όσον αφορά τα αποτελέσματα, χρησιμοποιώντας το RMSE ως μέτρο αξιολόγησης για το σύνολο εκπαίδευσης προκύπτει ότι το μοντέλο που χρησιμοποιεί τα συμβόλαια μελλοντικής εκπλήρωσης (futures) του αργού πετρελαίου υπερτερεί σε προβλέψεις με χρονικό ορίζοντα ενός μήνα, ενώ το μοντέλο που χρησιμοποιεί τις τιμές των αποθεμάτων πλεονεκτεί έναντι των υπολοίπων σε προβλέψεις με χρονικό ορίζοντα 2 μηνών. Το μοντέλο που αξιοποιεί μόνο την χρονοσειρά των τιμών του ντίζελ υπερτερεί έναντι των υπολοίπων για προβλέψεις με χρονικό ορίζοντα 3 μηνών. Αναφέρεται ότι σε περιόδους υψηλής μεταβλητότητας οι τιμές των αποθεμάτων του ντίζελ προσφέρουν πολύτιμες πληροφορίες. Σημειώνεται ότι εντός του συνόλου εκπαίδευσης τα μοντέλα προβλέπουν με ακρίβεια 60% - 97% τις μεταβολές της τιμής του ντίζελ, με το ολοκληρωμένο αυτοπαλινδρούμενο μοντέλο κινητού μέσου όρου να έχει την μεγαλύτερη ακρίβεια. Στα σύνολα δοκιμής υπάρχει μικρή ακρίβεια πρόβλεψης με τα μοντέλα να μπορούν να προβλέψουν τις μεταβολές τις τιμής με ακρίβεια 60% - 80%. Αξίζει να σημειωθεί ότι όλα τα παραπάνω μοντέλα δεν μπορούν να προσαρμοστούν γρήγορα σε απότομες μεταβολές.

2.2.3 ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΤΗΣ ΤΙΜΗΣ ΤΟΥ ΑΡΓΟΥ ΠΕΤΡΕΛΑΙΟΥ

Οι Lackes et al. (2009) ασχολήθηκαν με την πρόβλεψη του αργού πετρελαίου σε βραχυπρόθεσμο, μεσοπρόθεσμο και μακροπρόθεσμο χρονικό ορίζοντα. Στην έρευνά τους χρησιμοποίησαν νευρωνικά δίκτυα με ένα κρυμμένο επίπεδο και την χρονοσειρά του αργού πετρελαίου κατά την περίοδο 1999 – 2006. Από τα αποτελέσματα προέκυψε ότι τα νευρωνικά δίκτυα δεν μπορούν να εξασφαλίσουν μεγάλη ακρίβεια σε βραχυπρόθεσμο ορίζοντα (τρεις μέρες), ενώ σε μακροπρόθεσμο ορίζοντα (τρεις μήνες) η ακρίβεια των προβλέψεων βελτιώνεται σημαντικά. Επιπρόσθετα, από την έρευνά τους προέκυψε ότι όσον αφορά τις προβλέψεις σε μακροπρόθεσμο ορίζοντα η χρήση δύο νευρώνων αντί για πέντε στο κρυμμένο επίπεδο βελτιώνει τα αποτελέσματα της πρόβλεψης. Γενικά, αξίζει να σημειωθεί ότι όσο λιγότεροι είναι οι νευρώνες που απαιτούνται στο κρυμμένο επίπεδο τόσο λιγότερο μη γραμμική είναι η χρονοσειρά που χρησιμοποιείται.

Οι Kulkarni και Haidar (2009) διερεύνησαν την προβλεπτικότητα του αργού πετρελαίου σε βραχυπρόθεσμο ορίζοντα με την χρήση συμβολαίων μελλοντικής εκπλήρωσης (futures) αργού πετρελαίου με λήξεις σε έναν, δύο, τρεις και τέσσερις μήνες. Για την προετοιμασία των δεδομένων εισόδου δοκιμάστηκαν δύο τρόποι διαφορετικοί από τις διαφορές λογαριθμικών τιμών. Αυτοί είναι:

α) $y_t = \frac{X_t - X_{t-n}}{X_{t-n}}$ (2α)

β) $y_t = \frac{X_t - 2X_{t-n} + X_{t-2n}}{X_{t-n}}$ (2β)

Για την ομαλοποίηση των δεδομένων χρησιμοποιήθηκε η συνάρτηση:

γ) $y_z = \frac{2*(X_{z,t} - min(X_z))}{(max(X_z) - min(X_z))}$ (2γ)

Έτσι, τα δεδομένα εισόδου έχουν πεδίο τιμών (-1,1). Για την εξαγωγή των προβλέψεων χρησιμοποιήθηκε ένα νευρωνικό δίκτυο με ένα κρυμμένο επίπεδο και ως συνάρτηση ενεργοποίησης λήφθηκε η σιγμοειδής. Για την εύρεση του βέλτιστου αριθμού των προηγούμενων παρατηρήσεων για το δίκτυο ελέγχθηκαν από μία έως 20 παρατηρήσεις και για το εκάστοτε δίκτυο χρησιμοποιήθηκαν από ένας έως δέκα νευρώνες στο κρυμμένο επίπεδο. Από τα αποτελέσματα παρατηρήθηκε ότι η εφαρμογή του κινητού μέσου όρου των τριών ημερών στα παραπάνω δεδομένα και εν συνεχεία η χρησιμοποίηση μιας από τις παραπάνω μεθόδους για την προετοιμασία των δεδομένων εξασφαλίζει την μεγαλύτερη ακρίβεια πρόβλεψης. Για την εύρεση του βέλτιστου αριθμού των προηγούμενων

παρατηρήσεων χρησιμοποιήθηκε μόνο η χρονοσειρά του αργού πετρελαίου και εξήχθη ότι ισούται με 13 παρατηρήσεις. Από τα αποτελέσματα προέκυψε ότι η ακρίβεια πρόβλεψης είναι 78%, 66% και 53% για χρονικό ορίζοντα μίας, δύο και τριών ημερών αντίστοιχα. Επίσης, σημειώνεται ότι τα συμβόλαια μελλοντικής εκπλήρωσης (futures) αργού πετρελαίου με βραχυπρόθεσμη λήξη (ένας ή δύο μήνες) παρέχουν σημαντικές πληροφορίες στο μοντέλο για την πρόβλεψη της κατεύθυνσης της τιμής του αργού πετρελαίου. Αντίθετα, η χρησιμοποίηση των αντίστοιχων συμβολαίων αργού πετρελαίου με λήξεις 3 και 4 μηνών έχουν μικρή συνεισφορά στην αύξηση της ακρίβειας του μοντέλου.

Οι Mahdiani και Khamehchi (2017) ενίσχυσαν την υπάρχουσα βιβλιογραφία της εφαρμογής νευρωνικών δικτύων για την πρόβλεψη του αργού πετρελαίου. Στην έρευνά τους αναφέρουν ότι η χρήση γενετικού αλγορίθμου (GA) είναι απαραίτητη για την κατασκευή του βέλτιστου δικτύου. Ο γενετικός αλγόριθμος που χρησιμοποίησαν βρίσκει τον αριθμό των νευρώνων στο κρυμμένο επίπεδο και τον αριθμό των προηγούμενων παρατηρήσεων του αργού πετρελαίου για τα οποία το σφάλμα πρόβλεψης στο σύνολο δοκιμής ελαχιστοποιείται. Επίσης, αν δύο δίκτυα έχουν περίπου ίση ακρίβεια, ο γενετικός αλγόριθμος επιλέγει εκείνο το δίκτυο με τον μικρότερο αριθμό νευρώνων στο κρυμμένο επίπεδο. Από τα αποτελέσματα συμπεραίνεται ότι το βέλτιστο δίκτυο έχει τρεις προηγούμενες παρατηρήσεις του αργού πετρελαίου ως δεδομένα εισόδου και 14 νευρώνες στο κρυμμένο επίπεδο. Οι συγκεκριμένες τιμές είναι κοινές είτε χρησιμοποιείται ημερήσια είτε μηνιαία χρονοσειρά. Ένα άλλο συμπέρασμα της έρευνας είναι ότι όταν ο αριθμός των νευρώνων στο κρυμμένο επίπεδο γίνει μεγαλύτερος από 14, δεν υπάρχει καμία ουσιώδης βελτίωση στα αποτελέσματα της πρόβλεψης.

Οι Xiong et al. (2013) αξιοποίησαν την χρονοσειρά του αργού πετρελαίου για να αξιολογήσουν μοντέλα πρόβλεψης πολλαπλών σταδίων. Τονίζουν και οι ίδιοι την σημασία της πρόβλεψης της τιμής του αργού πετρελαίου, αφού χρησιμοποιείται εκτενώς από τις κυβερνήσεις, τις επιχειρήσεις αλλά και μεμονωμένα από τους πολίτες. Οι ίδιοι αναφέρουν ότι στην βιβλιογραφία υπάρχουν τρεις κοινές στρατηγικές για την πρόβλεψη πολλαπλών σταδίων και είναι οι παρακάτω: α) επαναλαμβανόμενη στρατηγική, β) η «απευθείας» στρατηγική, γ) η στρατηγική πολλαπλής εισόδου – πολλαπλής εξόδου (MIMO). Όλες αυτές οι στρατηγικές ενσωματώνονται και αξιολογούνται βασισμένες σε νευρωνικά δίκτυα. Η αξιολόγηση βασίζεται στην ακρίβεια του εκάστοτε μοντέλου αλλά και του υπολογιστικού κόστους που απαιτεί. Η χρονοσειρά που χρησιμοποιήθηκε για να αξιολογηθούν τα μοντέλα είναι η εβδομαδιαία χρονοσειρά των τιμών του αργού πετρελαίου κατά την περιόδου 2000 - 2011. Από τα αποτελέσματα εξήχθη το συμπέρασμα ότι η στρατηγική MIMO πλεονεκτεί έναντι των υπολοίπων και αναφέρεται ότι έχει αρκετά μικρότερο υπολογιστικό φόρτο από τις άλλες δύο στρατηγικές.

Οι Abdullah και Zeng (2010) συνείσφεραν σημαντικά στις προσπάθειες των ερευνητών να προβλέψουν τις μελλοντικές τιμές του αργού πετρελαίου. Όπως τονίζεται και από τους ίδιους η χρήση των νευρωνικών δικτύων έχει συμβάλλει σημαντικά στην αύξηση της ορθότητας των προβλέψεων του αργού πετρελαίου, καθώς η χρονοσειρά του παρουσιάζει έντονες μη γραμμικότητες. Οι συγκεκριμένοι ερευνητές εκτός από την χρονοσειρά του αργού πετρελαίου χρησιμοποίησαν και άλλα ποιοτικά δεδομένα και δείκτες που σχετίζονται με το αργό πετρέλαιο και ενισχύουν σημαντικά την ακρίβεια της πρόβλεψης. Οι

δείκτες που θα χρησιμοποιηθούν ως μεταβλητές εισόδου στο νευρωνικό δίκτυο σχετίζονται με την προσφορά αργού πετρελαίου που υπάρχει από τις διάφορες χώρες παραγωγής, την ζήτηση, δηλαδή τις καταναλωτικές απαιτήσεις που υπάρχουν παγκοσμίως, τα αποθέματα, την ανάπτυξη της οικονομίας (δείκτης ανάπτυξης, πληθωρισμός), τις τρεις βασικές ισοτιμίες στερλίνα/δολάριο, γεν/δολάριο, ευρώ/δολάριο και τα στοιχεία που αφορά τον πληθυσμό στις αναπτυσσόμενες και ανεπτυγμένες χώρες. Όλοι αυτοί οι δείκτες εξάγονται μηνιαία και κατά συνέπεια οι προβλέψεις είναι μηνιαίες. Παρακάτω στον πίνακα 2.2 παρατίθενται αναλυτικά όλες οι μεταβλητές που επιδρούν στην διαμόρφωση της τιμής του αργού πετρελαίου.

Πίνακας 2.2: Οικονομικοί δείκτες που επιδρούν στην διαμόρφωση της τιμής του αργού πετρελαίου (Πηγή: Abdullah, Zeng, 2010)

Variables	Factors
S^T	Supply
S_{a1}	Productions of OPEC countries
S_{a2}	Productions of Non-OPEC countries
S_{b1}	Proved reserves of OPEC countries
S_{b2}	Proved reserves of OECD countries
S_{c1}	Number of well drilled
D^T	Demand
D_{a1}	Consumption of OECD countries
D_{a2}	Consumption of China
D_{a3}	Consumption of India
I^T	Inventory
I_{a1}	Ending stocks of OECD countries
I_{a2}	Ending stocks of US
I_{b1}	US petroleum imports from OPEC countries
I_{b2}	US petroleum imports from Non-OPEC countries
I_{c1}	US crude oil imports from OPEC countries
I_{c2}	US crude oil imports from Non-OPEC countries
E^T	Economy
E_{a1}	Foreign Exchange of GBP/USD
E_{a2}	Foreign Exchange of Yen/USD
E_{a3}	Foreign Exchange of Euro/USD
E_{b1}	US Growth Domestic Products (GDP)
E_{c1}	US Inflation rate
E_{d1}	US Consumer Price Index (CPI)
P^T	Population
P_{a1}	Population of developed countries
P_{a2}	Population of less developed countries
WTI	West Texas Intermediate price

Η συγκεκριμένη έρευνα έγινε για την περίοδο 1984 – 2009. Η επιλογή των μεταβλητών εισόδου έγινε με την μέθοδο της δοκιμής – σφάλματος. Δημιουργήθηκαν 8 διαφορετικές ομάδες μεταβλητών εισόδου και η κάθε μια ομάδα ελέγχθηκε με 3, 4 και 5 νευρώνες στο κρυμμένο επίπεδο. Από τα αποτελέσματα φαίνεται ότι η ακρίβεια της πρόβλεψης της

κατεύθυνσης του πετρελαίου με χρονικό ορίζοντα ενός μήνα είναι της τάξης του 86% με 90% αναλόγως την ομάδα δεδομένων που χρησιμοποιείται στο δίκτυο.

Οι Mollasalehi et al. (2011) εφάρμοσαν τα νευρωνικά δίκτυα με ένα κρυμμένο επίπεδο για να προβλέψουν την τιμή του αργού πετρελαίου σε χρονικό ορίζοντα μίας ημέρας και μίας εβδομάδας. Η συνεισφορά τους στην υπάρχουσα βιβλιογραφία έγκειται στο γεγονός ότι έκαναν εκτεταμένη έρευνα των παραγόντων που επιδρούν στην διαμόρφωση της τιμής του αργού πετρελαίου, ώστε να χρησιμοποιηθούν ως μεταβλητές εισόδου στο δίκτυο. Ένας από αυτούς τους παράγοντες είναι η προσφορά και η ζήτηση του αργού πετρελαίου από κάθε χώρα. Για να ληφθεί υπόψιν αυτός ο παράγοντας χρησιμοποιείται ως μεταβλητή εισόδου ένας δείκτης που αφορά τις ενεργειακές ανάγκες κάθε χώρας. Ένας άλλος παράγοντας είναι η τιμή του χρυσού. Στις μέρες μας λόγω χαμηλών επιτοκίων οι πολίτες αποφεύγουν τις προθεσμιακές καταθέσεις και αναζητούν άλλες λύσεις, όπως είναι ο χρυσός που αποτελεί καταφύγιο σε περιόδους κρίσης. Σημαντικός παράγοντας στην διαμόρφωση της τιμής του αργού πετρελαίου είναι οι τιμές των μετοχών στις Ηνωμένες Πολιτείες και αυτές οι μεταβολές ενσωματώνονται σε μια μεταβλητή που ονομάζεται δείκτης «φόβου».

Οι ερευνητές χρησιμοποίησαν ως μεταβλητές εισόδου τις παραπάνω μεταβλητές με τρεις συνεχόμενες ιστορικές παρατηρήσεις για την κάθε μια μεταβλητή προκειμένου να προβλέψουν την κατεύθυνση της τιμής του αργού πετρελαίου την επόμενη ημέρα. Για την πρόβλεψη της κατεύθυνσης της τιμής την επόμενη εβδομάδα οι ερευνητές χρησιμοποίησαν ως μεταβλητές εισόδου τις παρατηρήσεις τριών προηγούμενων ημερών, καθώς και την παρατήρηση της ίδιας ημέρας. Για την εύρεση του βέλτιστου αριθμού νευρώνων στο κρυμμένο επίπεδο οι ερευνητές έλεγξαν τα σφάλματα των δικτύων όταν ο αριθμός των νευρώνων στο κρυμμένο επίπεδο είναι 20, 50, 75 και μετά τις διαμέσους 37 και 62. Από τα αποτελέσματα φαίνεται ότι όσο αυξάνεται ο αριθμός των νευρώνων στο κρυμμένο επίπεδο δεν μειώνεται πάντα το σφάλμα. Για παράδειγμα από 37 έως 50 νευρώνες μειώνεται το σφάλμα αλλά σε όλες τις άλλες περιπτώσεις αυξάνεται. Γενικότερα, υπενθυμίζεται όταν σε μια χρονοσειρά υπάρχουν ομαλές μεταβολές τάσεων, τότε το σφάλμα μειώνεται καθώς αυξάνεται ο αριθμός των νευρώνων στο κρυμμένο επίπεδο. Κατά συνέπεια, φαίνεται ότι η υπάρχουσα χρονοσειρά χαρακτηρίζεται από απότομες μεταβολές τάσεων. Όσον αφορά τις προβλέψεις της κατεύθυνσης του αργού πετρελαίου την επόμενη μέρα η ακρίβεια στο σύνολο δοκιμής ισούται με 80%, ενώ για τις προβλέψεις με χρονικό ορίζοντα μιας εβδομάδας η ακρίβεια είναι ίση με 65%. Έτσι, συμπεραίνεται από τους συγκεκριμένους ερευνητές ότι οι προβλέψεις σε μεγαλύτερο χρονικό ορίζοντα έχουν μεγαλύτερο ρίσκο και είναι λιγότερο αξιόπιστες ανεξαρτήτου της τεχνικής πρόβλεψης που εφαρμόζεται.

2.3 ΣΥΜΠΕΡΑΣΜΑΤΑ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Από την ανάλυση της βιβλιογραφίας γίνεται κατανοητό ότι η αύξηση της τιμής των καυσίμων έχει ως άμεση συνέπεια την αύξηση της ζήτησης στα μέσα μαζικής μεταφοράς και κυρίως τα λεωφορεία, τα οποία οι πολίτες χρησιμοποιούν αντί του επιβατικού οχήματος. Φαίνεται από τις διάφορες έρευνες ότι αν και οι ποσοστιαίες μεταβολές μπορεί να είναι μικρές, ο αριθμός των ανθρώπων είναι τόσο μεγάλος που συχνά απαιτείται αλλαγή της συχνότητας των μέσων μαζικής μεταφοράς και αντικατάσταση κάποιων μέσων με άλλα μεγαλύτερης χωρητικότητας. Πολλές φορές η απουσία κατάλληλου προγραμματισμού οδηγεί σε προβλήματα εξυπηρέτησης της ζήτησης.

Επίσης, από τις διάφορες έρευνες φαίνεται ότι οι πολίτες αναπροσαρμόζουν άμεσα τις συνήθειες τους σε αύξηση ή και σε μείωση της τιμής των καυσίμων, όταν πρόκειται για ταξίδι αναψυχής. Αντίθετα, όσον αφορά τις υποχρεωτικές διαδρομές, οι πολίτες αργούν να αναπροσαρμόσουν τις συνήθειές τους και αυτό μπορεί να πάρει μέχρι και ένα με ενάμισι μήνα. Συνεπώς, είναι κατανοητό ότι η αύξηση του μεταφορικού κόστους επηρεάζει όλους τους πολίτες, οι οποίοι αναζητούν λύσεις για τον μετριασμό του.

Σχετικά με τις οδικές μεταφορές στην Ευρώπη, όλες οι υπάρχουσες έρευνες δείχνουν ότι οποιαδήποτε άνοδος στην τιμή των καυσίμων έστω και μικρή ελαχιστοποιεί τα ποσοστά κέρδους τους, καθώς λόγω του μεγάλου ανταγωνισμού οι εταιρίες δεν μπορούν να μετακυλίσουν αυτή την αύξηση στους πελάτες τους. Συνεπώς, η πρόβλεψη της τιμής των καυσίμων σε μεσοπρόθεσμο ορίζοντα που αποτελεί και αντικείμενο της παρούσας διπλωματικής εργασίας είναι μείζονος σημασίας για τις συγκεκριμένες εταιρίες, ώστε να μπορούν να ρυθμίζουν τα αποθέματα καυσίμων που κατέχουν και να εξασφαλίζουν την ανταγωνιστικότητά τους.

Στο υποκεφάλαιο 2.2 έγινε ανάλυση της βιβλιογραφίας όσον αφορά στις προβλέψεις των τιμών των καυσίμων. Τονίζεται η σημασία των προβλέψεων των τιμών των καυσίμων για την εξασφάλιση της κοινωνικής ευημερίας των πολιτών, καθώς μεγάλο μέρος των επιχειρήσεων και κυρίως ο συγκοινωνιακός τομέας καταναλώνει σημαντικές ποσότητες καυσίμων. Στις υπάρχουσες έρευνες έχουν χρησιμοποιηθεί γραμμικά και μη γραμμικά μοντέλα πρόβλεψης. Ωστόσο, η ακρίβεια των προβλέψεων δεν ξεπερνά το 65% - 70%. Κρίνεται συνεπώς σκόπιμο να διερευνηθεί η χρήση μεθόδων εκμάθησης μηχανών και αν αυτή θα ενισχύσει σημαντικά την ακρίβεια των προβλέψεων.

Επιπρόσθετα, στο υποκεφάλαιο 2.2 γίνεται και ανάλυση της βιβλιογραφίας όσον αφορά στην πρόβλεψη των τιμών του αργού πετρελαίου με μεθόδους εκμάθησης μηχανών. Θεωρήθηκε απαραίτητο να γίνει ανάλυση της συγκεκριμένης βιβλιογραφίας για την εύρεση κατάλληλων μεταβλητών εισόδου, τρόπων ομαλοποίησης και προετοιμασίας των δεδομένων προτού αρχίσει η εκπαίδευση της εκάστοτε μεθόδου εκμάθησης μηχανών, ώστε να επιτευχθεί η κατά το δυνατόν μεγαλύτερη ακρίβεια πρόβλεψης του μοντέλου που θα κατασκευαστεί στην διπλωματική εργασία.

3. ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

3.1 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Στην παρούσα διπλωματική εργασία γίνεται προσπάθεια πρόβλεψης της τιμής της αμόλυβδης σε μεσοπρόθεσμο χρονικό ορίζοντα. Όπως ήδη αναφέρθηκε στο υποκεφάλαιο 2.1 οι μεταβολές των τιμών των καυσίμων έχουν σημαντικές επιπτώσεις τόσο στα μέσα μαζικής μεταφοράς όσο και στον τομέα των συγκοινωνιών γενικότερα. Συνεπώς, η παρούσα διπλωματική εργασία έχει στόχο να συμβάλλει στον μετριασμό αυτών των προβλημάτων, καθώς με την αξιοποίηση της πρόβλεψης για την τιμή της αμόλυβδης οι οδηγοί των επιβατικών οχημάτων αλλά και οι εταιρίες στον συγκοινωνιακό τομέα θα μπορούν να ρυθμίζουν κατάλληλα τα αποθέματα καυσίμων, ώστε να μην διαφοροποιηθεί το μεταφορικό τους κόστος. Επειδή τους οδηγούς και τις εταιρίες που δραστηριοποιούνται στον συγκοινωνιακό τομέα τους ενδιαφέρει να γνωρίζουν αν η τάση της τιμής της αμόλυβδης θα είναι ανοδική ή καθοδική για τον χρονικό ορίζοντα που διερευνάται, το πρόβλημα που πραγματεύεται η παρούσα διπλωματική εργασία θα είναι τύπου 0/1. Όταν η πρόβλεψη είναι 1 θα σημαίνει ανοδική τάση και 0 καθοδική τάση για την τιμή της αμόλυβδης.

3.2 ΕΠΙΛΟΓΗ ΜΕΘΟΔΟΥ ΕΠΙΛΥΣΗΣ

Στο υποκεφάλαιο 2.2 έγινε περιγραφή των μεθόδων που εφαρμόζονται για την πρόβλεψη της τιμής των καυσίμων και του αργού πετρελαίου. Παρατηρήθηκε ότι μετά το 2005 άρχισε να υπάρχει μια δυναμική χρήση των μεθόδων εκμάθησης μηχανών και συγκεκριμένα των νευρωνικών δικτύων για την πρόβλεψη της τιμής του αργού πετρελαίου. Ο λόγος που υπάρχει αυτή η μαζική στροφή των ερευνητών στα νευρωνικά δίκτυα είναι το γεγονός ότι επειδή δημιουργούν ένα μη γραμμικό μοντέλο μπορούν να παρέχουν έγκυρες προβλέψεις για χρονοσειρές που παρουσιάζουν έντονες μη γραμμικότητες, όπως το αργό πετρέλαιο. Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθούν μέθοδοι από την οικογένεια εκμάθησης μηχανών, οι οποίες είναι τα νευρωνικά δίκτυα με ένα κρυμμένο επίπεδο, τα δένδρα απόφασης καθώς και η μέθοδος AdaboostM1, ώστε να επιλυθεί το πρόβλημα της πρόβλεψης της τιμής της απλής αμόλυβδης. Οι παραπάνω μέθοδοι επιλέχθηκαν, επειδή μπορούν να παράξουν έγκυρα αποτελέσματα σε χρονοσειρές που χαρακτηρίζονται από απότομες μεταβολές και μη γραμμικές συμπεριφορές. Θα γίνει ανάλυση των πλεονεκτημάτων και των μειονεκτημάτων της εκάστοτε μεθόδου καθώς και σύγκριση μεταξύ τους.

3.3 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Στο παρόν κεφάλαιο θα γίνει μια περιεκτική, αλλά ταυτόχρονα αναλυτική περιγραφή της λογικής και των θεμελιωδών αρχών των μεθόδων εκμάθησης μηχανών που εφαρμόζονται στην παρούσα διπλωματική εργασία. Ο στόχος είναι ο αναγνώστης να μπορέσει να αντιληφθεί τις μεθοδολογίες που χρησιμοποιούνται χωρίς να χρειάζεται να ανατρέξει σε βιβλία ή στον ιστότοπο. Οι μέθοδοι εκμάθησης μηχανών που θα παρουσιαστούν είναι τα νευρωνικά δίκτυα, δένδρα απόφασης και η μέθοδος AdaboostM1. Η λογική και των τριών μεθόδων βασίζεται σε ένα σύστημα, το οποίο κατηγοριοποιεί μια νέα είσοδο στην σωστή έξοδο, αφού πρώτα έχει εκπαιδευτεί με μεγάλο αριθμό αντιστοιχιών εισόδου – εξόδου. Θα παρουσιαστούν διάφοροι τρόποι διόρθωσης προβλημάτων υπερπροσαρμογής (overfitting). Το κεφάλαιο 3.4 που αφορά τα νευρωνικά δίκτυα βασίζεται κατά κύριο λόγο στις διαδικτυακές σημειώσεις που έχουν γραφτεί από τον καθηγητή του Stanford στο μάθημα Machine Learning Andrew Ng. Το αντίστοιχο μάθημα στο Stanford είναι το CS229: Machine Learning (Stanford, 2017). Στο κεφάλαιο 3.4 θα χρησιμοποιηθούν επίσης και πηγές από το βιβλίο του Michael Nielsen (Nielsen, 2016). Για το κεφάλαιο 3.5 που αφορά τα δένδρα απόφασης η κύρια πηγή αποτελεί το βιβλίο Data Mining and Decision Trees: Theory and Applications που έχει γραφτεί από τους Rokach Lior και Oded Maimon. Το κεφάλαιο 3.6 που αναλύει τα δάση απόφασης με βάση την μέθοδο AdaboostM1 βασίζεται στις δημοσιεύσεις των Yoav Freud και Robert E. Schapire, οι οποίοι ανακάλυψαν την συγκεκριμένη μέθοδο εκμάθησης μηχανών. Οι υπόλοιπες πηγές θα παρατίθενται ξεχωριστά μέσα στο κείμενο.

3.4 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

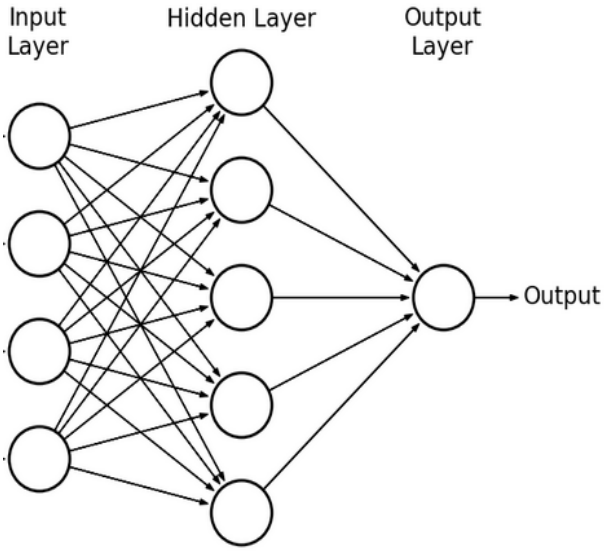
3.4.1 ΣΥΝΤΟΜΗ ΕΙΣΑΓΩΓΗ

Τα νευρωνικά δίκτυα ανήκουν στην οικογένεια της εκμάθησης μηχανών και άρχισαν να χρησιμοποιούνται πολύ την δεκαετία του ’80 και του ’90. Από τα τέλη της δεκαετίας του ’90 η δημοφιλότητα τους άρχισε να συρρικνώνεται σημαντικά. Πρόσφατα έχει αρχίσει πάλι η επαναχρησιμοποίηση τους σε διάφορες επιστήμες, όπως της βιολογίας, των οικονομικών, των μαθηματικών και τις μεταφορές. Ένας λόγος που άρχισαν να εφαρμόζονται πάλι είναι το γεγονός ότι παλαιότερα υπήρχαν τεχνολογικά εμπόδια που δεν επέτρεπαν την χρήση νευρωνικών δικτύων σε μεγάλα σύνολα δεδομένων λόγω του μεγάλου υπολογιστικού κόστους. Ανεξάρτητα από την επιστήμη που τα εφαρμόζει ο στόχος είναι η δημιουργία μιας μηχανής, η οποία να μιμείται τις λειτουργίες του μυαλού. Τα νευρωνικά δίκτυα εκπαιδεύονται σε ένα μεγάλο αντιπροσωπευτικό σύνολο δεδομένων και δίνουν σήματα / ενδείξεις, όταν χρησιμοποιηθούν για ένα άλλο σύνολο δεδομένων. Το μεγάλο πλεονέκτημα τους έναντι της πολλαπλής και λογιστικής παλινδρόμησης είναι το γεγονός ότι λειτουργούν πολύ καλύτερα σε σύνθετες μη γραμμικές εφαρμογές ακόμη και με πολύ μεγάλο αριθμό μεταβλητών εισόδου.

3.4.2 ΔΟΜΗ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Το σημαντικότερο χαρακτηριστικό των νευρωνικών δικτύων είναι ο νευρώνας. Ο νευρώνας αποτελείται από μία ή περισσότερες μεταβλητές εισόδου, τον κορμό και την μεταβλητή εξόδου. Πολλοί νευρώνες μαζί σχηματίζουν αυτό που ονομάζεται νευρωνικό δίκτυο. Ουσιαστικά, ο νευρώνας παίρνει την πληροφορία από τις μεταβλητές εισόδου και την μετατρέπει με μια υπολογιστική συνάρτηση σε μεταβλητή εξόδου. Ένα νευρωνικό δίκτυο αποτελείται από τουλάχιστον δυο επίπεδα, το επίπεδο εισόδου και το επίπεδο εξόδου (Perceptron). Αν υπάρχουν μόνο δύο επίπεδα στο νευρωνικό δίκτυο, τότε αυτό καλείται και λογιστική παλινδρόμηση. Τα κρυμμένα επίπεδα (hidden layers) μπορεί να είναι από 1 έως

25 – 30. Γενικότερα χρησιμοποιούνται νευρωνικά δίκτυα με ένα ή δύο κρυμμένα επίπεδα, όπως στην Εικόνα 3.1.



Εικόνα 3.1: Δομή νευρωνικού δικτύου με ένα κρυμμένο επίπεδο

Στην εικόνα 3.1 φαίνεται ότι κάθε νευρώνας συνδέεται με τις μεταβλητές εισόδου, στο συγκεκριμένο παράδειγμα τέσσερις και αποδίδει μια έξοδο. Αν υπήρχε και άλλο κρυμμένο επίπεδο ο νευρώνας του πρώτου κρυμμένου επιπέδου θα αποτελούσε την μεταβλητή εισόδου για το επόμενο κρυμμένο επίπεδο και αυτή η διαδικασία θα επαναλαμβανόταν μέχρι το επίπεδο εξόδου.

Όπως έχει αναφερθεί κάθε νευρώνας λειτουργεί, όπως ο ανθρώπινος εγκέφαλος. Παίρνει κάποιες μεταβλητές εισόδου και στην συνέχεια θέλει να αποφασίσει για κάτι. Για παράδειγμα, χρησιμοποιώντας την εικόνα 3.1, ο κάθε νευρώνας για να αποφανθεί θετικά ή αρνητικά, τοποθετεί μία βαρύτητα σε κάθε μια από τις τέσσερις μεταβλητές του και εξάγει το αποτέλεσμα. Αν αυτό είναι μεγαλύτερο από κάποιο όριο αποφαίνεται θετικά αλλιώς αποφαίνεται αρνητικά. Συνήθως σε τέτοια διωνυμικά προβλήματα οι μεταβλητές εξόδου είναι 0/1, 1 για θετικά και 0 για αρνητικά. Αυτή η διαδικασία ακολουθείται από κάθε νευρώνα και η μεταβλητή εξόδου είναι ένα σταθμισμένο άθροισμα αυτών που συγκρίνεται με ένα όριο για να αποφανθεί αν είναι 0 ή 1. Σχετικά με την ορολογία που χρησιμοποιείται υπάρχουν τα εξής: κάθε μεταβλητή εισόδου συμβολίζεται με X_i η βαρύτητα σε κάθε μεταβλητή με w_{ji}^{j} (j είναι ο αριθμός του επιπέδου και i ο αριθμός της μεταβλητής), και το όριο που γίνεται η σύγκριση σε κάθε επίπεδο για να αποφανθεί ο νευρώνας ονομάζεται «προκατάληψη» (bias) και είναι κοινό για κάθε επίπεδο ή και για όλο το δίκτυο γενικότερο. Αξίζει να αναφερθεί ότι όσο περισσότερο αυξάνει ο ερευνητής το όριο αυτό τόσο περισσότερο αυξάνεται η ακρίβεια στην πρόβλεψη θετικών αποτελεσμάτων. Όταν άρχησαν να χρησιμοποιούνται τα νευρωνικά δίκτυα από τους ερευνητές έγιναν προσπάθειες να βελτιωθούν τα αποτελέσματα 0/1 αλλά παρ' όλες τις προσπάθειες δεν βρέθηκε κάποια

λύση για αυτό, καθώς η αλλαγή του ορίου βελτίωνε κάποια αποτελέσματα αλλά ταυτόχρονα επηρέαζε κάποια άλλα.

Έτσι, άρχησαν να χρησιμοποιούνται συναρτήσεις που να μπορούν να δώσουν ένα αποτέλεσμα 0/1 ή -1/1 με σωστή συμπεριφορά όσον αφορά τις ακραίες τιμές. Τέτοιες συναρτήσεις είναι: α) σιγμοειδής = $\frac{1}{1+e^{-z}}$ με πεδίο τιμών (0,1), β) υπερβολική εφαπτομένη = $\frac{2}{1+e^{-2z}} - 1$ με πεδίο τιμών (-1,1), ημιτονοειδής = sin(z) με πεδίο τιμών (0,1). Το z είναι ο σταθμισμένος μέσος όρος προσθέτοντας και την «προκατάληψη», δηλαδή z = w*X + b. Η πλέον χρησιμοποιούμενη συνάρτηση είναι η σιγμοειδής σε διάφορα είδη προβλημάτων, ωστόσο για κάθε πρόβλημα καλό είναι να εξάγονται τα αποτελέσματα και με την υπερβολική εφαπτομένη και να κρατάται η βέλτιστη. Χρησιμοποιώντας μία από αυτές τις συναρτήσεις το πρόβλημα αρχίζει να αποκτά μη γραμμικό χαρακτήρα που είναι και ο λόγος που σε πολλά προβλήματα τα νευρωνικά δίκτυα λειτουργούν καλύτερα από ότι η γραμμική παλινδρόμηση. Έτσι, αλλάζει λίγο ο τρόπος που δουλεύει το νευρωνικό δίκτυο, αφού σε κάθε νευρώνα επιτελείται μία ακόμη λειτουργία χρησιμοποιώντας μια από τις παραπάνω συναρτήσεις. Η λειτουργία του νευρωνικού δικτύου περιγράφεται παρακάτω. Κάθε νευρώνας από το πρώτο κρυμμένο επίπεδο δέχεται τις μεταβλητές εισόδου και χρησιμοποιώντας τα βάρη του (είναι διαφορετικά για κάθε νευρώνα) υπολογίζει το σταθμισμένο μέσο z. Ακολούθως χρησιμοποιείται μία από τις παραπάνω συναρτήσεις που ονομάζονται συναρτήσεις ενεργοποίησης (activation function) που συμβολίζεται με a_i^j. Το j υποδηλώνει τον αριθμό του επιπέδου και το i υποδηλώνει τον αριθμό της μεταβλητής. Έτσι, εξάγει ένα αποτέλεσμα που έχει συνεχή τιμή από 0 έως 1. Εν συνεχεία αυτή η μεταβλητή θα αποτελέσει την μεταβλητή εισόδου για το επόμενο κρυμμένο επίπεδο, το οποίο θα επιτελέσει και αυτό την παραπάνω διαδικασία. Αυτή η διαδικασία ακολουθείται μέχρι την μεταβλητή εξόδου. Η παραπάνω διαδικασία ονομάζεται εμπροσθοδρόμηση (forward propagation).

Έτσι, σε κάθε νευρωνικό δίκτυο επιλύεται καθένα από τα μικρά προβλήματα για την επίλυση του τελικού προβλήματος, όπως περιγράφεται παραπάνω. Γίνεται λοιπόν κατανοητό ότι η επίλυση των επιμέρους προβλημάτων επηρεάζει την απάντηση στο τελικό πρόβλημα. Η εφαρμογή των νευρωνικών δικτύων συμβάλλει στην επίλυση τέτοιων πολύπλοκων μη γραμμικών προβλημάτων που ο ανθρώπινος εγκέφαλος αδυνατεί να επιλύσει. Τελικώς τον εκάστοτε ερευνητή τον ενδιαφέρει κατά κύριο λόγο το τελικό αποτέλεσμα δηλαδή τα αποτελέσματα του τελευταίου επιπέδου. Κατά συνέπεια, αντιλαμβάνεται κανείς ότι οι επιμέρους σχέσεις μεταξύ των νευρώνων έχουν μία και μοναδική χρησιμότητα, να συμβάλλουν στην βελτίωση των τελικών αποτελεσμάτων μέσω της συνάρτησης κόστους.

Προτού γίνει επεξήγηση της συνάρτησης κόστους θα γίνει μια σύντομη αλλά σαφής περιγραφή των προβλημάτων κατηγοριοποίησης. Υπενθυμίζεται ότι στην παρούσα διπλωματική εργασία θα επιλυθεί πρόβλημα κατηγοριοποίησης και συγκεκριμένα διωνυμικό τύπου 0/1. Σε ένα πρόβλημα που τα αποτελέσματα είναι 0/1 κάθε παράδειγμα που θα χρησιμοποιείται για την εκμάθηση των νευρώνων θα έχει δύο διαστάσεις στην μεταβλητή εξόδου ($y_i \in R^2$). Επίσης, αν σε ένα πρόβλημα οι μεταβλητές εξόδου είναι

τέσσερις, τότε η μεταβλητή εξόδου σε κάθε παράδειγμα θα έχει τέσσερις διαστάσεις ($y_i \in R^4$), με το ένα να μπαίνει στην σωστή στήλη και οι υπόλοιπες στήλες θα έχουν 0.

Η συνάρτηση κόστους είναι πολύ σημαντική, καθώς η μείωσή της ισοδυναμεί με αύξηση της ακρίβειας του μοντέλου. Η συνάρτηση κόστους εμπεριέχει πάντα τετράγωνα για να μην αλλοιώνονται τα αποτελέσματα σε περίπτωση αρνητικών διαφορών και να βοηθήσει το μοντέλο να συγκλίνει στην καλύτερη λύση. Η χρησιμότητά της είναι για την εύρεση των βέλτιστων βαρών, ώστε να εξασφαλίζεται η μεγαλύτερη ακρίβεια του μοντέλου / μικρότερη τιμή της συνάρτησης κόστους. Αξίζει να αναφερθεί ότι οι συναρτήσεις κόστους χρησιμοποιούνται και σε άλλες μεθόδους γραμμικές ή μη, όπως πολλαπλή γραμμική παλινδρόμηση, λογιστική γραμμική παλινδρόμηση. Ο σκοπός και στις άλλες μεθόδους είναι πάλι ο ίδιος, η εύρεση των βαρών που εξασφαλίζουν την μεγαλύτερη ακρίβεια του μοντέλου / την μικρότερη τιμή της συνάρτησης κόστους. Παρακάτω παρουσιάζονται η τετραγωνική συνάρτηση κόστους (3) και η συνάρτηση κόστους για προβλήματα κατηγοριοποίησης (4). Η δεύτερη συνάρτηση έχοντας και ένα δεύτερο όρο, αυτόν της ομαλοποίησης, θα χρησιμοποιηθεί για την επίλυση του προβλήματος στην παρούσα διπλωματική εργασία.

$$C = \frac{1}{2n}\sum ||y(x) - a^L(x)||^2 \qquad (3)$$

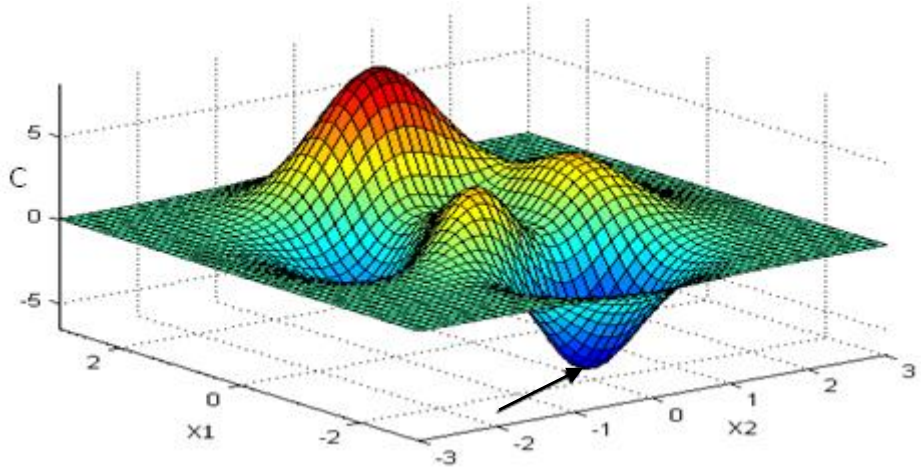
$$C = -\frac{1}{n}\sum_x \sum_j [y_j ln a_j^L + (1 - y_j)\ln(1 - a_j^L)] \qquad (4)$$

Παρατηρείται ότι στις συναρτήσεις κόστους δεν έχει προστεθεί ο όρος της ομαλοποίησης, διότι δεν έχει γίνει ακόμη αναφορά για τον συντελεστή ομαλοποίησης λ και την συνεισφορά του στην εκπαίδευση των νευρώνων. Θα ακολουθήσει αναλυτική περιγραφή στην ενότητα 3.4.4 για τη σημασία της ομαλοποίησης στην εκπαίδευση των νευρώνων.

3.4.3 ΕΚΠΑΙΔΕΥΣΗ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Οπισθοδρόμηση σφάλματος

Παραπάνω περιγράφηκε η διαδικασία της εμπροσθοδρόμησης (forward propagation).Στο παρόν κεφάλαιο θα περιγραφεί η οπισθοδρόμηση σφάλματος (backpropagation) ίσως είναι από τα δυσκολότερη σημεία στην κατανόηση της λειτουργίας των νευρωνικών δικτύων, αλλά είναι πολύ σημαντική διαδικασία, αφού εξασφαλίζει την εκπαίδευση των νευρώνων μέσω της εύρεσης του ολικού ελαχίστου της συνάρτησης του κόστους C. Στην εικόνα 3.2 φαίνεται η συνάρτηση κόστους C συναρτήσει δύο μεταβλητών εισόδου X_1, X_2. Το ολικό ελάχιστο του κόστους C είναι εκεί που δείχνει το βέλος.



Εικόνα 3.2: Συνάρτηση κόστους C με μεταβλητές X_1 και X_2

Η οπισθοδρόμηση σφάλματος (backpropagation) άρχισε να εφαρμόζεται από την δεκαετία του ’70 αλλά η σημασία της αναδείχθηκε το 1986 από μια δημοσίευση που χρησιμοποιεί οπισθοδρόμηση σφάλματος και επιλύει τα προβλήματα πολύ γρηγορότερα από ότι παλαιότεροι αλγόριθμοι (Rumelhart, D., Hinton, G., Williams, R., 1986). Ο στόχος της οπισθοδρόμησης σφάλματος είναι ο υπολογισμός των μερικών παραγώγων $\frac{\partial C}{\partial w}$ και $\frac{\partial C}{\partial b}$ για κάθε βάρος w ή «προκατάληψη» b στο νευρωνικό δίκτυο. Οι μερικές παράγωγοι $\frac{\partial C}{\partial w}$ και $\frac{\partial C}{\partial b}$ δείχνουν πόσο γρήγορα αλλάζει η συνάρτηση του κόστους C, όταν αλλάξουν τα βάρη και η «προκατάληψη» (bias). Για τις πράξεις που θα γίνουν παρακάτω θα χρησιμοποιηθεί η συνάρτηση κόστους (3):

$$C = \frac{1}{2n}\sum ||y(x) - a^L(x)||^2 \qquad (3)$$

Στο σημείο αυτό θα γίνουν δύο παραδοχές για να εφαρμοστεί η οπισθοδρόμηση. Η πρώτη είναι ότι για τον υπολογισμό του κόστους C θα χρησιμοποιηθεί ο μέσος όρος των κόστων όλων των παραδειγμάτων που βρίσκονται στο σύνολο εκπαίδευσης, δηλαδή $C = \frac{1}{n}\sum_x C_x$, όπου C_x το κόστος για κάθε παράδειγμα. Ο λόγος που χρειάζεται να γίνει αυτή η παραδοχή είναι, επειδή πρέπει να υπολογιστούν οι μερικές παράγωγοι για κάθε παράδειγμα μεμονωμένα και εν συνεχεία οι $\frac{\partial C}{\partial w}$ και $\frac{\partial C}{\partial b}$ να είναι ο μέσος όρος των παραπάνω παραγώγων. Η δεύτερη παραδοχή είναι ότι η συνάρτηση του κόστους μπορεί να γραφτεί συναρτήσει των αποτελεσμάτων του νευρωνικού δικτύου C = C(a^L), όπου L είναι το τελευταίο επίπεδο.

Παρακάτω θα αναφερθούν οι τέσσερις βασικές εξισώσεις που συνθέτουν την οπισθοδρόμηση σφάλματος. Αυτές είναι οι μερικές παράγωγοι $\frac{\partial C}{\partial w^l_{jk}}$ και $\frac{\partial C}{\partial b^l_j}$. Για να υπολογιστούν αυτές πρέπει πρώτα να υπολογιστεί το σφάλμα του j νευρώνα στο επίπεδο l, δ^l_j . Ο ρόλος του τελευταίου όρου πέραν από το ότι συμβάλλει στον υπολογισμό των παραπάνω μερικών παραγώγων, προσθέτει και μια μικρή διαφορά Δz^l_j στο βάρος του νευρώνα και κατά συνέπεια ο νευρώνας εξάγει ένα λίγο διαφορετικό αποτέλεσμα σ(z^l_j + Δz^l_j). Αυτή η μεταβολή περνάει και στα υπόλοιπα επίπεδα και τελικά προκαλεί την μεταβολή του κόστους C κατά $\frac{\partial C}{\partial z^l_j}\Delta z^l_j$. Ο ρόλος του δ^l_j είναι να συμβάλλει, ώστε να μειωθεί το κόστος μέσω του Δz^l_j . Αν για παράδειγμα το $\frac{\partial C}{\partial z^l_j}$ έχει μεγάλη τιμή θετική ή αρνητική, τότε το δ^l_j μειώνει το κόστος επιλέγοντας ένα Δz^l_j που έχει το αντίθετο πρόσημο από το $\frac{\partial C}{\partial z^l_j}$. Αν όμως το $\frac{\partial C}{\partial z^l_j}$ έχει πολύ μικρή τιμή το δ^l_j δεν μπορεί να μειώσει και άλλο το κόστος. Κατά συνέπεια, από αυτόν τον όρο μπορεί ο ερευνητής να αντιληφθεί αν είναι κοντά στην βέλτιστη τιμή. Το $\frac{\partial C}{\partial z^l_j}$ είναι ένας τρόπος μέτρησης του σφάλματος στο νευρώνα και είναι ως εξής, $\delta^l_j = \frac{\partial C}{\partial z^l_j}$. Υπενθυμίζεται ότι ο εκθέτης l δείχνει ότι βρίσκεται στο επίπεδο l. Η εξίσωση για το σφάλμα στο τελευταίο επίπεδο (επίπεδο εξόδου) είναι ίση με $\delta^l_j = \frac{\partial C}{\partial a^l_j}$ σ'(z^L_j). Ο πρώτος όρος του γινομένου δείχνει πόσο αλλάζει το κόστος σαν συνάρτηση του a^l_j του επιπέδου l. Για παράδειγμα, αν το κόστος δεν αλλάζει ή αλλάζει ελάχιστα στον συγκεκριμένο νευρώνα, τότε η τιμή του δ^l_j είναι πολύ μικρή. Ο δεύτερος όρος του γινομένου σ'(z^L_j) δείχνει πόσο γρήγορα αλλάζει η συνάρτηση σ, όταν αλλάζει το z^L_j. Επειδή στην οπισθοδρόμηση προτιμώνται τα δ^l σε μορφή πίνακα (εξοικονόμηση υπολογιστικού χρόνου) η μορφή είναι όπως φαίνεται παρακάτω:

$$\delta^L = \nabla_\alpha C \odot \sigma'(z^L) \qquad (5)$$

Ο πρώτος όρος του πολλαπλασιασμού πινάκων κατά στοιχείο είναι ένα διάνυσμα που αποτελείται από τις μερικές παραγώγους $\frac{\partial C}{\partial a^l_j}$.Το $\nabla_a C$ είναι ίσο και με $\nabla_a C = a^L - y$. Για την οπισθοδρόμηση τον ερευνητή τον ενδιαφέρει να δειχθεί πως συνδέεται ο όρος δ^l με

τον όρο δ^{l+1} του επόμενου επιπέδου $l+1$. Υπενθυμίζεται ότι ο τελικός στόχος είναι να μεταφερθεί το σφάλμα προς τα πίσω. Το δ^{l} γράφεται ως εξής:

$$\delta^{l} = ((w^{l+1})^{T}\delta^{l+1}) \odot \sigma'(z^{\mathrm{L}}) \qquad (6)$$

Πλέον συνδυάζοντας τις εξισώσεις (4), (6) μπορεί να υπολογιστεί το σφάλμα σε οποιοδήποτε επίπεδο. Αρχικά, υπολογίζεται το δ^{l}, μετά από την εξίσωση (6) μπορεί να εξαχθεί το δ^{L-1} και ακολουθώντας την ίδια διαδικασία να υπολογιστεί τα σφάλματα όλου του νευρωνικού δικτύου. Η εξίσωση του ρυθμού μεταβολής του κόστους συναρτήσει της «προκατάληψης» (bias) είναι:

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \qquad (7)$$

Όμως, από τις εξισώσεις (4), (6) έχει υπολογιστεί τον όρο δ_j^l. Οπότε η μεταβολή του κόστους ως προς οποιοδήποτε βάρος στο νευρωνικό δίκτυο γράφεται ως εξής:

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l \qquad (8)$$

$$\frac{\partial C}{\partial w} = a_{in}\delta_{out} \qquad (9)$$

Η εξίσωση (9) είναι ίδια με την εξίσωση (8) απλά έχει λιγότερους δείκτες. Από την εξίσωση (9) φαίνεται ότι, όταν το a_{in} είναι μικρό και η μεταβολή του κόστους ως προς το βάρος θα είναι μικρή. Συνεπώς, τα βάρη αργούν να εκπαιδευτούν και κατά συνέπεια και οι νευρώνες αργούν να εκπαιδευτούν.

Σχετικά με τον όρο σ'(z_j^L) αξίζει να υπενθυμιστεί ότι η σιγμοειδής συνάρτηση σ(z_j^L) αποκτά πολύ μικρή κλίση, όταν η τιμή της είναι περίπου 0 ή 1. Σε αυτές τις τιμές η σ'(z_j^L) είναι περίπου μηδέν. Το συμπέρασμα και εδώ είναι ότι ένα βάρος στο τελευταίο επίπεδο θα εκπαιδευτεί αργά αν το αποτέλεσμα του νευρώνα είναι περίπου μηδέν ή περίπου ένα. Σε αυτές τις περιπτώσεις αναφέρεται ότι το αποτέλεσμα του νευρώνα έχει «κορεστεί» και ότι το βάρος έχει σταματήσει να εκπαιδεύεται. Τα ίδια ισχύουν και για τον όρο της «προκατάληψης» (bias). Παρόμοια συμπεράσματα ισχύουν και για τα άλλα επίπεδα, καθώς το δ_j^l μπορεί να γίνει μικρό αν ο νευρώνας είναι κοντά σε «κορεσμό». Κατά συνέπεια, τα βάρη ενός «κορεσμένου» νευρώνα θα εκπαιδευτούν πολύ αργά.

Συνοψίζοντας τα βάρη θα εκπαιδευτούν αργά αν ο νευρώνας εισόδου έχει μικρή τιμή ενεργοποίησης ή ο νευρώνας εξόδου έχει μεγάλη ή μικρή τιμή ενεργοποίησης. Επίσης, σημειώνεται ότι οι παραπάνω εξισώσεις ισχύουν για οποιαδήποτε συνάρτηση ενεργοποίησης, όχι μόνο για την σιγμοειδή που παρουσιάζεται παραπάνω. Η συνάρτηση ενεργοποίησης επιλέγεται, ώστε να φέρει τις ιδιότητες εκπαίδευσης που απαιτεί το κάθε πρόβλημα. Για παράδειγμα, αν απαιτείται να χρησιμοποιηθεί μια συνάρτηση η κλίση της οποίας να είναι πάντα θετική και να μην προσεγγίζει ποτέ το μηδέν, αυτό θα συνέβαλε να μην μειωθεί ο ρυθμός εκπαίδευσης, όταν οι κανονικοί σιγμοειδείς νευρώνες φτάνουν σε συνθήκες «κορεσμού». Παρακάτω φαίνονται συνοπτικά οι τέσσερις βασικές εξισώσεις την οπισθοδρόμησης:

$$\delta^L = \nabla_\alpha \mathrm{C} \odot \sigma'(z^{\mathrm{L}}) \quad (5)$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^{\mathrm{L}}) \quad (6)$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \quad (7)$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \quad (8)$$

Έλεγχος Ορθότητας Υπολογισμού Μερικών Παραγώγων

Την δεκαετία του ’50, όπου δεν είχαν ακόμη ανακαλυφθεί οι αλγόριθμοι της οπισθοδρόμησης σφάλματος για την εκπαίδευση των νευρωνικών δικτύων χρησιμοποιούταν ο Γενικευμένος Κανόνας Δέλτα (Gradient Descent). Χρησιμοποιώντας τον κανόνα της αλυσίδας υπολογίζεται η κλίση του κόστους. Για λόγους απλοποίησης οι μερικές παράγωγοι υπολογίζονται ως εξής:

$$\frac{\partial C}{\partial w_j} = \frac{C(w+\varepsilon e_{j)} - C(w)}{\varepsilon} \quad (10),$$

όπου ε>0 είναι ένας μικρός θετικός αριθμός και e_j ένα διάνυσμα προς την j κατεύθυνση. Δηλαδή υπολογίζεται οι $\frac{\partial C}{\partial w_j}$ από τις διαφορές της συνάρτησης κόστους για δύο ελάχιστα διαφορετικές τιμές w_j και έτσι προκύπτει η εξίσωση (10). Η ίδια λογική εφαρμόζεται και για τον υπολογισμό των μερικών παραγώγων $\frac{\partial C}{\partial b}$. Αυτή η μέθοδος, ενώ δείχνει πολύ απλή στην κατανόηση και στην εφαρμογή είναι πολύ αργή στην εύρεση του ολικού ελαχίστου.

Συνήθως, χρησιμοποιείται για επαλήθευση των μερικών παραγώγων που έχουν υπολογιστεί μέσω της οπισθοδρόμησης σφάλματος. Προτού εφαρμοστεί ο κώδικας της οπισθοδρόμησης σφάλματος σε ένα μεγάλο σύνολο δεδομένων σε μια μικρή ομάδα δεδομένων εφαρμόζεται και η παραπάνω διαδικασία, ώστε να επαληθευτούν ότι οι τιμές των μερικών παραγώγων είναι περίπου ίσες.

Αρχικοποίηση Βαρών (w)

Η αρχικοποίηση των βαρών είναι πολύ σημαντική για την σωστή εκπαίδευση των νευρώνων. Αν κάποιος τοποθετήσει την τιμή μηδέν στα αρχικά βάρη, τότε το νευρωνικό

δίκτυο δεν θα εκπαιδευτεί καθώς, όπως θα οπισθοδρομείται το σφάλμα όλοι οι νευρώνες θα έχουν ίδια τιμή επαναλαμβανόμενα. Έτσι, για να καταργηθεί αυτή η συμμετρία ένας τρόπος είναι να δοθούν στα βάρη w_{ij}^{l} τυχαίες τιμές μεταξύ [-ε, ε], όπου ε = $\frac{\sqrt{6}}{\sqrt{Διάσταση\ σειρών\ w^{l} + Διάσταση\ στηλών\ w^{l}}}$. Αξιοποιώντας το συγκεκριμένο ε w^{l} γράφεται:

w^{l} = 2rand($Διάσταση\ σειρών\ w^{l}, Διάσταση\ στηλών\ w^{l}$+1) – ε (11)

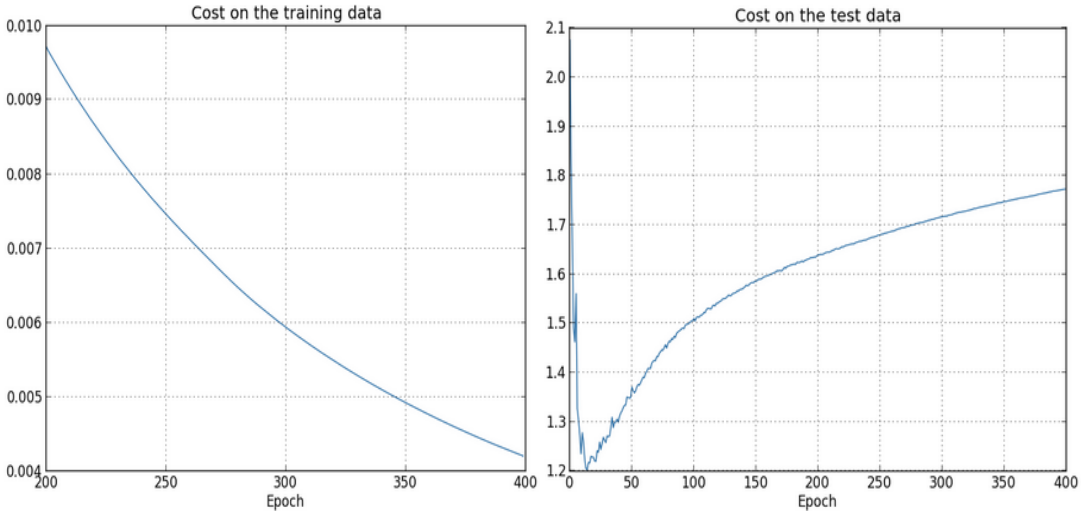
Με την παραπάνω αρχικοποίηση είναι βέβαιο ότι το νευρωνικό δίκτυο θα συγκλίνει αν όχι στο ολικό ελάχιστο πολύ κοντά σε αυτό. Υπενθυμίζεται ότι η συνάρτηση rand() παράγει τυχαίους αριθμούς από το 0 μέχρι το 1 που ακολουθούν ομοιόμορφη κατανομή.

3.4.4 ΑΝΤΙΜΕΤΩΠΙΣΗ ΠΡΟΒΛΗΜΑΤΩΝ ΥΠΕΡΠΡΟΣΑΡΜΟΓΗΣ

Ένας τρόπος που εφαρμόζεται για να μην υπερπροσαρμόζεται το νευρωνικό δίκτυο στο σύνολο δεδομένων για εκπαίδευση είναι η χρήση πολύ μεγάλης βάσης δεδομένων. Ωστόσο, αυτό δεν είναι πάντα εφικτό, είτε επειδή για το υπό επίλυση πρόβλημα δεν υπάρχουν αρκετά δεδομένα, είτε διότι η απόκτηση πολλών δεδομένων απαιτεί μεγάλο χρηματικό κόστος. Κατά συνέπεια, πρέπει να υιοθετούνται άλλοι τρόποι για την αντιμετώπιση της υπερπροσαρμογής (overfitting).

Ένας τέτοιος τρόπος είναι ο διαχωρισμός του συνόλου των δεδομένων σε σύνολο εκπαίδευσης, σύνολο επικύρωσης, σύνολο δοκιμής. Συνήθως, χρησιμοποιείται από 60% έως 70% του συνόλου των δεδομένων για εκπαίδευση και το υπόλοιπο ισομοιράζεται σε σύνολο επικύρωσης και σύνολο δοκιμής. Σε σπάνιες περιπτώσεις μπορεί να χρησιμοποιηθεί το 90% των δεδομένων για εκπαίδευση των νευρώνων, ωστόσο αυτό συμβαίνει όταν η χρονική κλίμακα των χρονοσειρών είναι μεγάλη (π.χ. μήνας, τρίμηνο). Όπως, είναι λογικό όσο αυξάνεται ο αριθμός των «εποχών» τόσο αυξάνεται η ακρίβεια του μοντέλου στο σύνολο εκπαίδευσης και ταυτόχρονα μειώνεται το κόστος της συνάρτησης C. Αν ο αριθμός των «εποχών» αυξηθεί πάρα πολύ ενδέχεται η ακρίβεια του μοντέλου στο σύνολο εκπαίδευσης να αγγίξει το 100%. Αυτό σημαίνει ότι το νευρωνικό δίκτυο έχει εκπαιδευτεί πάρα πολύ καλά στο σύνολο εκπαίδευσης αλλά έχει χάσει την ικανότητα να γενικεύει, με αποτέλεσμα να υπάρχει πρόβλημα υπερπροσαρμογής (overfitting). Για την αποφυγή του προβλήματος της υπερπροσαρμογής (overfitting) πρέπει κανείς να κοιτάει και την συνάρτηση κόστους C αλλά και την ακρίβεια (accuracy) στο σύνολο επικύρωσης. Έτσι, θα παρατηρήσει κανείς ότι μετά από έναν αριθμό εποχών το κόστος C του συνόλου επικύρωσης αρχίζει να αυξάνεται. Ταυτόχρονα και η ακρίβεια (accuracy) στο σύνολο επικύρωσης φτάνει σε κορεσμό και προοδευτικά αρχίζει να μειώνεται. Στην «εποχή», όπου

η συνάρτηση κόστους C έχει ελάχιστο πρέπει να σταματάει η εκπαίδευση των νευρώνων για να μην υπάρξουν προβλήματα υπερπροσαρμογής (overfitting). Κάποιες φορές δεν χρησιμοποιείται σύνολο επικύρωσης και ο βέλτιστος αριθμός «εποχών» βρίσκεται από το σύνολο δοκιμής. Στην εικόνα 3.3 φαίνονται τα διαγράμματα κόστους για τα σύνολα εκπαίδευσης και δοκιμής και παρατηρείται από το δεύτερο διάγραμμα ότι η εκπαίδευση των νευρώνων πρέπει να είχε σταματήσει στις 15 «εποχές».



Εικόνα 3.3: Διαγράμματα Κόστους σε σύνολα εκπαίδευσης και δοκιμής συναρτήσει των «εποχών»

Το σύνολο δοκιμής χρησιμοποιείται για την μέτρησης της ακρίβειας του μοντέλου και είναι αυτό που θεωρείται το πλέον αξιόπιστο. Συνήθως, η ακρίβεια αυτού του συνόλου είναι λίγο μικρότερη από την ακρίβεια των συνόλων εκπαίδευσης και επικύρωσης. Παράλληλα, είναι και ένας τρόπος επιβεβαίωσης ότι το νευρωνικό δίκτυο δεν έχει υπερπροσαρμοστεί και στο σύνολο επικύρωσης.

<u>Συντελεστής ομαλοποίησης λ και υπερπροσαρμογή</u>

Ένας άλλος τρόπος είναι ο περιορισμός του βαθμού στον οποίο υπερπροσαρμόζονται οι νευρώνες στο σύνολο εκπαίδευσης μέσω των παραμέτρων ομαλοποίησης. Από τις πιο ευρέως διαδεδομένες τεχνικές ομαλοποίησης είναι αυτή του συντελεστή αποδόμησης (weight decay) ή L2 ομαλοποίηση. Η L2 ομαλοποίηση εφαρμόζεται τοποθετώντας άλλον ένα όρο στην συνάρτηση κόστους C που ονομάζεται όρος ομαλοποίησης. Έτσι, η συνάρτηση ομαλοποίησης γράφεται:

$$C = -\frac{1}{n}\sum_{xj}[\, y_j ln a_j^L + \big(1 - y_j\big)\ln(1 - a_j^L)] + \frac{\lambda}{2n}\sum_w w^2 \qquad (12)$$

Ο πρώτος όρος είναι γνωστός και έχει αναφερθεί στην ενότητα 3.4.2 και ο δεύτερος όρος προσθέτει τα τετράγωνα των βαρών του δικτύου. Αυτός ο όρος πολλαπλασιάζεται και με τον συντελεστή λ/2n, όπου λ > 0 γνωστό ως συντελεστής ομαλοποίησης και n ο αριθμός

των παραδειγμάτων στο σύνολο εκπαίδευσης. Σημειώνεται ότι ο δεύτερος όρος δεν περιλαμβάνει τις «προκαταλήψεις» (bias). Επισημαίνεται ότι οι «προκαταλήψεις» (bias) δεν εντάσσονται στη συνάρτηση, διότι η ενσωμάτωσή τους δεν αλλάζει τα αποτελέσματα. Δεύτερον, μεγάλες τιμές των «προκαταλήψεων» (bias) δεν επηρεάζουν έναν νευρώνα να ενσωματώσει όλο τον θόρυβο που υπάρχει στο σύνολο εκπαίδευσης, όπως κάνουν τα μεγάλα βάρη. Είναι λογικό ότι και άλλες συναρτήσεις κόστους μπορούν να ομαλοποιηθούν, όπως η τετραγωνική συνάρτηση κόστους που παρατίθεται παρακάτω:

$$C = \frac{1}{2n}\sum_x \left||y - a^L\right|\right|^2 + \frac{\lambda}{2n}\sum_w w^2 \qquad (13)$$

$$C = C_0 + \frac{\lambda}{2n}\sum_w w^2, \quad \text{όπου } C_0 \text{ το αρχικό κόστος} \qquad (14)$$

Η χρήση της παραμέτρου ομαλοποίησης λ έχει ως αποτέλεσμα το νευρωνικό δίκτυο να προτιμά να μαθαίνει μικρά βάρη. Μεγάλα βάρη θα επιτρέπονται εφ' όσον βελτιώνουν σημαντικά τον πρώτο όρο της συνάρτησης κόστους. Συνεπώς, η ομαλοποίηση συνδυάζει την εύρεση μικρών βαρών με την ταυτόχρονη μείωση του αρχικού κόστους. Αν η παράμετρος λ είναι μικρή, τότε είναι σαν υπάρχει μόνο το αρχικό κόστος, ενώ αν είναι μεγάλη τότε το δίκτυο προτιμά μικρά βάρη.

Για την καλύτερη κατανόηση των επιπτώσεων της παραμέτρου ομαλοποίησης στο νευρωνικό δίκτυο, ας υποτεθεί ότι υπάρχει ένα δίκτυο με μικρά βάρη, δηλαδή ένα ομαλοποιημένο δίκτυο. Όταν υπάρχουν μικρά βάρη αυτό σημαίνει ότι η συμπεριφορά του δικτύου δεν αλλάξει πολύ αν αλλάξουν μερικά παραδείγματα. Κατά συνέπεια, το ομαλοποιημένο δίκτυο δεν μπορεί να επηρεαστεί από τον τοπικό θόρυβο των παραδειγμάτων του συνόλου εκπαίδευσης. Σε αντίθεση, το συγκεκριμένο δίκτυο επηρεάζεται από την γενική τάση σε όλο το σύνολο εκπαίδευσης. Αν όμως, το δίκτυο έχει μεγάλα βάρη, η συμπεριφορά του μπορεί να αλλάξει ακόμη και σε μικρές αλλαγές μερικών παραδειγμάτων. Συνεπώς, ένα μη ομαλοποιημένο δίκτυο μπορεί να χρησιμοποιήσει μεγάλα βάρη για να δημιουργήσει ένα πολύπλοκο μοντέλο που εμπεριέχει όλο τον θόρυβο που υπάρχει στα παραδείγματα του συνόλου εκπαίδευσης. Συμπερασματικά, είναι ευκόλως αντιληπτό ότι προτιμώνται τα πρώτα δίκτυα, τα οποία μπορούν να ενσωματώσουν την πληροφορία και να την γενικεύουν, ώστε να μπορούν να αξιοποιηθούν και σε άλλα σύνολα δεδομένων εξασφαλίζοντας εξίσου μεγάλη ακρίβεια (accuracy).

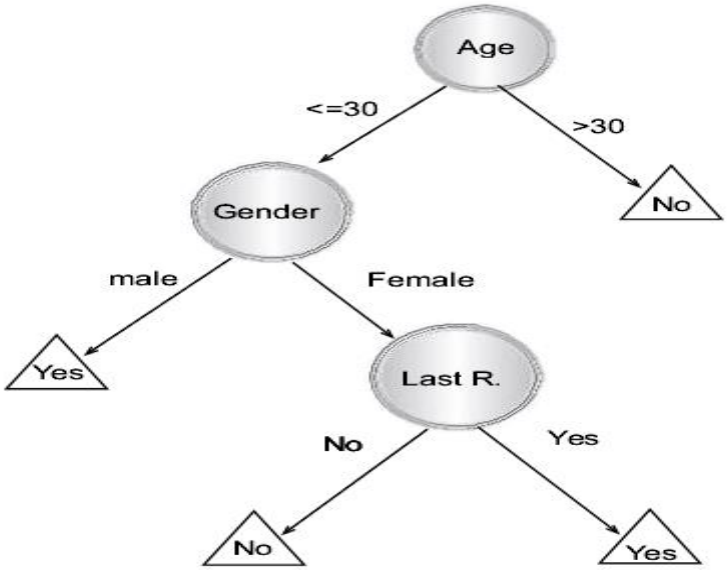
3.5 ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

3.5.1 ΣΥΝΤΟΜΗ ΕΙΣΑΓΩΓΗ

Ένα δένδρο απόφασης είναι ένας ταξινομητής που λειτουργεί σαν ένας αναδρομικός «διαχωριστής» του εκάστοτε εσωτερικού υποχώρου. Κάθε δένδρο απόφασης συνδέεται μέσω των «κλαδιών» του με κόμβους και δημιουργείται έτσι ένα δένδρο. Ο κόμβος που δεν έχει «κλαδιά» να καταλήγουν σε αυτό ονομάζεται «ρίζα» του δένδρου. Οι κόμβοι, στους οποίους καταλήγουν «κλαδιά» χωρίς όμως να φεύγουν «κλαδιά» από αυτά ονομάζονται εξωτερικά «φύλλα». Όλα τα υπόλοιπα ονομάζονται εσωτερικοί κόμβοι.

Σε κάθε δένδρο απόφασης κάθε εσωτερικός κόμβος διαχωρίζει τον υποχώρο σε δύο ή περισσότερους υποχώρους, σύμφωνα με μια διακριτή συνάρτηση για τις μεταβλητές εισόδου. Η πιο απλή και συνήθης μορφή είναι μία συνάρτηση που λαμβάνει υπόψη μόνο μία μεταβλητή και διαχωρίζει τον υποχώρο, σύμφωνα με αυτήν. Σε περίπτωση που η μεταβλητή αυτή είναι συνεχής, τότε η συνθήκη είναι ένα πεδίο τιμών. Κάθε κόμβος έχει ανατεθεί σε μια κατηγορία που αντιπροσωπεύει την πιο κατάλληλη τιμή στόχο ή μπορεί να έχει ένα διάνυσμα πιθανοτήτων που να υποδεικνύει την πιθανότητα μία μεταβλητή να έχει μία συγκεκριμένη τιμή.

Στο διάγραμμα 3.1 απεικονίζεται πώς η ομάδα δεδομένων έχει κατηγοριοποιηθεί από την «ρίζα» του δένδρου προς τα κάτω στα «φύλλα» με βάση τα αποτελέσματα των ελέγχων σε κάθε κόμβο. Το διάγραμμα 3.1 είναι ένα δένδρο απόφασης, στο οποίο υπάρχουν διάφοροι λόγοι για τους οποίους ένας πιθανός πελάτης απαντάει σε μηνύματα ηλεκτρονικής αλληλογραφίας. Οι εσωτερικοί κόμβοι αναπαριστώνται με κύκλο, ενώ τα «φύλλα» με τρίγωνο. Το δένδρο απόφασης αποτελείται από λογικές και αριθμητικές μεταβλητές. Έχοντας αυτό το δένδρο μπορεί ένας αναλυτής να προβλέψει την απάντηση ενός δυνητικού πελάτη αλλά και τα χαρακτηριστικά της συμπεριφοράς μιας ομάδας πελατών. Κάθε κόμβος απεικονίζει τις μεταβλητές που ελέγχει και τα «φύλλα» απεικονίζουν την θετική ή αρνητική απάντηση.



Διάγραμμα 3.1: Απεικόνιση δένδρου απόφασης

Αξίζει να σημειωθεί ότι αυτοί που λαμβάνουν αποφάσεις προτιμούν λιγότερο πολύπλοκα δένδρα, επειδή είναι ευκολότερα στην κατανόησή τους. Η πολυπλοκότητα του δένδρου απόφασης έχει σημαντικές επιπτώσεις στην ακρίβεια του μοντέλου (Breiman, L., 1984). Η πολυπλοκότητα ενός δένδρου ελέγχεται από τρία κριτήρια τερματισμού και την μέθοδο του «κλαδέματος». Η πολυπλοκότητα μετράται από τον αριθμό των κόμβων, τον αριθμό των «φύλλων», το βάθος του δένδρου απόφασης και τον αριθμό των μεταβλητών που χρησιμοποιούνται.

3.5.2 ΚΑΤΑΣΚΕΥΗ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ

Τα δένδρα απόφασης κατασκευάζονται από αλγορίθμους που κατασκευάζουν αυτόματα ένα δένδρο απόφασης για ένα σύνολο δεδομένων. Είναι λογικό ότι ο στόχος είναι η εύρεση του βέλτιστου δένδρου, δηλαδή εκείνου που θα μειώσει σημαντικά το σφάλμα. Παράλληλα, μπορεί να υπάρχουν και άλλοι περιορισμοί, όπως η ελαχιστοποίηση του αριθμού των κόμβων ή η ελαχιστοποίηση του μέσου βάθους. Οι αλγόριθμοι για την εύρεση του βέλτιστου δένδρου χωρίζονται σε δυο κατηγορίες: α) οι από πάνω προς τα κάτω (top down) και οι από κάτω προς τα πάνω (bottom up), με την βιβλιογραφία να προτιμά γενικά την πρώτη κατηγορία. Οι αλγόριθμοι για τα από πάνω προς τα κάτω δένδρα απόφασης είναι οι: α) ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), CART (Breiman 1984). Κάποιοι

αλγόριθμοι είναι συνδυασμοί αυτών , όπως η ταυτόχρονη ανάπτυξη και το «κλάδεμα» του δένδρου απόφασης (C4.5 and CART). Οι αλγόριθμοι αυτοί δουλεύουν, όπως εξηγείται παρακάτω. Σε κάθε επανάληψη ο αλγόριθμος αποφασίζει τον διαχωρισμό των δεδομένων σύμφωνα με το αποτέλεσμα της συνάρτησης χρησιμοποιώντας την μεταβλητή εισόδου. Αφού διαχωριστούν κατάλληλα τα δεδομένα, κάθε κόμβος υποδιαιρείται σε μικρότερα υποσύνολα μέχρι είτε όλα τα δεδομένα του υποχώρου να έχουν διαχωριστεί (ορθή κατηγοριοποίηση όλων των δεδομένων του υποχώρου) είτε να ικανοποιείται κάποιο κριτήριο τερματισμού.

Συγκεκριμένα, ο αλγόριθμος που χρησιμοποιείται από τις εφαρμογές της Weka είναι μια επέκταση του αλγορίθμου ID3 και πιθανά δημιουργεί ένα μικρό δένδρο. Κατασκευάζει το δένδρο από πάνω προς τα κάτω και ακολουθεί τα παρακάτω βήματα. Πρώτα ελέγχει εφ' όσον όλα τα παραδείγματα που θα ταξινομηθούν έχουν κάποιο κοινό χαρακτηριστικό (κατηγορία) και μετά δημιουργεί ένα κόμβο με αυτό το χαρακτηριστικό. Για κάθε χαρακτηριστικό υπολογίζεται η γνώση και το όφελος γνώσης και εντοπίζει εκείνο το χαρακτηριστικό που διαχωρίζει καλύτερα την ομάδα δεδομένων βάσει ενός κριτηρίου επιλογής.

3.5.3 ΜΕΤΡΗΣΗ ΟΦΕΛΟΥΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Εντροπία είναι η διαδικασία που θα χρησιμοποιηθεί για την μέτρηση του οφέλους της πληροφορίας. Ο λόγος που θα χρησιμοποιηθεί η εντροπία είναι η μέτρηση της διαταραχής των δεδομένων και η μονάδα μέτρησής τους είναι σε bit. Η μέτρηση της διαταραχής των δεδομένων ονομάζεται και μέτρηση της αβεβαιότητας μιας τυχαίας μεταβλητής. Για να γίνει κατανοητή αυτή η έννοια θα επεξηγηθεί με ένα παράδειγμα. Έστω ότι υπάρχει ένα κέρμα, κάθε ρίψη του κέρματος έχει εντροπία ίση με ένα bit. Μια σειρά με δύο κέρματα θα έχει εντροπία ίση με δυο bit. Η εντροπία και η δεσμευμένη εντροπία για ένα οποιοδήποτε P υπολογίζεται ως εξής:

$$Entropy(p) = -\sum_{j=1}^{n} \frac{|pj|}{|p|} log \frac{|pj|}{|p|} \qquad (15)$$

$$Entropy(j|p) = -\frac{|pj|}{|p|} log \frac{|pj|}{|p|} \qquad (16)$$

Αν η βάση για τον λογάριθμο ισούται με 2, τότε η μέτρηση της εντροπίας είναι σε bit αλλιώς αν η βάση του λογαρίθμου είναι ίση με 10 η μονάδα μέτρησής της είναι σε did. Το όφελος της πληροφορίας χρησιμοποιείται για την εκτίμηση της σχέσης μεταξύ δεδομένων εισόδου

και δεδομένων εξόδου. Σε κάθε βήμα υπάρχει αλλαγή στην πληροφορία της εντροπίας. Τελικώς , το όφελος της πληροφορίας υπολογίζεται ως εξής:

$$Gain(p,j) = Entropy(p - Entropy(j|p)) \qquad (17)$$

Για την κατασκευή ενός μικρού και αποτελεσματικού δένδρου απόφασης, ο διαχωρισμός των δεδομένων πρέπει να γίνεται με βάση το μεγαλύτερο όφελος. Για να γίνει καλύτερα αντιληπτή η προηγούμενη πρόταση ακολουθεί παράδειγμα. Έστω ότι υπάρχουν εννέα άνδρες και πέντε γυναίκες σε μία τάξη. Θα διαχωριστούν σε δύο διαφορετικές υποομάδες , η μία υποομάδα με έξι άντρες και μια γυναίκα και η άλλη υποομάδα με τρεις άντρες και τέσσερις γυναίκες. Οι πράξεις για τον υπολογισμό της εντροπίας και του οφέλους της πληροφορίας παρατίθενται παρακάτω:

$$Entropy_{bef} = -\frac{5}{14} * \log\left(\frac{5}{14}\right) - \frac{9}{14} * \log\left(\frac{9}{14}\right) \qquad (18\alpha)$$

$$Entropy_{left} = -\frac{3}{7} * \log\left(\frac{3}{7}\right) - \frac{4}{7} * \log\left(\frac{4}{7}\right) \qquad (18\beta)$$

$$Entropy_{right} = -\frac{6}{7} * \log\left(\frac{6}{7}\right) - \frac{1}{7} * \log\left(\frac{1}{7}\right), \qquad (18\gamma)$$

$$Entropy_{aft} = -\frac{7}{14} * Entropy_{left} + \frac{7}{14} * \mathrm{Entropy}_{\mathrm{right}} \qquad (18\delta)$$

$$Information_{Gain} = Entropy_{bef} - \mathrm{Entropy}_{\mathrm{aft}} \qquad (18\varepsilon)$$

3.5.4 ΜΕΘΟΔΟΙ «ΚΛΑΔΕΜΑΤΟΣ»

Ουσιαστικά πρόκειται για την εφαρμογή αυστηρών κριτηρίων τερματισμού, προκειμένου να κατασκευαστούν μικρά και με υποπροσαρμογή (underfitting) δένδρα απόφασης. Ωστόσο, αν παραληφθούν κάποια από αυτά τα κριτήρια ή στα υπάρχοντα κριτήρια τερματισμού δεν εισαχθούν αυστηροί περιορισμοί ενδέχεται να παραχθούν πολύ εκτενή δένδρα απόφασης, τα οποία υπερπροσαρμόζονται (overfitted) στα σύνολα εκπαίδευσης. Κατά συνέπεια, ένα δένδρο απόφασης, το οποίο έχει υπερπροσαρμοστεί στο σύνολο εκπαίδευσης έχει πολύ χαμηλή ακρίβεια στο σύνολο δοκιμής. Οι αλγόριθμοι «κλαδέματος» του δένδρου απόφασης χωρίζονται σε δυο κατηγορίες: α) στους αλγορίθμους που «κλαδεύουν» το δένδρο μετά την κατασκευή του, β) στους αλγορίθμους που «κλαδεύουν» το δένδρο ταυτόχρονα με την κατασκευή του. Οι εφαρμογές της Weka που θα χρησιμοποιηθούν και στην παρούσα διπλωματική εργασία χρησιμοποιούν τους πρώτους αλγορίθμους. Οι μέθοδοι «κλαδέματος» εφαρμόστηκαν για πρώτη φορά το 1984 για

αποφυγή του παραπάνω προβλήματος (Breiman, L., 1984). Σύμφωνα με την μεθοδολογία που προτάθηκε, αρχικά προσδιορίζουν ένα χαλαρό κριτήριο τερματισμού, αφήνοντας έτσι να δημιουργηθεί ένα δένδρο που υπερπροσαρμόζεται στο σύνολο εκπαίδευσης. Έπειτα, από το δένδρο αφαιρούνται κόμβοι και «κλαδιά», τα οποία δεν επηρεάζουν την ακρίβεια (accuracy) του. Από την βιβλιογραφία έχει αποδειχθεί ότι η εφαρμογή μεθόδων «κλαδέματος» μπορεί να βελτιώσει την ακρίβεια (accuracy) του δένδρου απόφασης, ειδικότερα σε δεδομένα που υπάρχει πολύς θόρυβος.

Μία άλλη μέθοδος που άρχισε να εφαρμόζεται παρουσιάστηκε από τους Bratko και Bohanec το 1994. Ο στόχος ήταν η δημιουργία ενός δένδρου απόφασης με επαρκή ακρίβεια που θα ήταν πολύ απλό στην κατασκευή του (Bratko, I., Bohanec, M., 1994). Σε αυτή την μέθοδο το αρχικό δένδρο απόφασης θεωρήθηκε ως απόλυτα ακριβές. Είναι γεγονός ότι η ακρίβεια ενός «κλαδεμένου» δένδρου απόφασης δείχνει πόσο κοντά είναι στο αρχικό δένδρο απόφασης.

Κόστος της πολυπλοκότητας του «κλαδέματος»

Το κόστος της πολυπλοκότητας του κλαδέματος αποτελείται από δυο στάδια (Breiman, L., 1984).Στο πρώτο στάδιο μια ακολουθία από δένδρα T_0, T_1, …., T_k κατασκευάζεται για ένα σύνολο εκπαίδευσης, όπου είναι T_0 είναι το αρχικό δένδρο που δεν έχει «κλαδευτεί» και T_k είναι η ρίζα του δένδρου. Στο δεύτερο στάδιο ένα από αυτά τα δένδρα επιλέγεται ως «κλαδεμένο» με βάση τον γενικό υπολογισμό σφάλματος. Το δένδρο T_{i+1} δημιουργείται αντικαθιστώντας ένα ή περισσότερα υπό – δένδρα από το δένδρο T_i με κατάλληλα «φύλλα». Τα υπό – δένδρα, τα οποία κλαδεύονται είναι αυτά, τα οποία έχουν την μικρότερη αύξηση σφάλματος ανά «φύλλο»:

$$\alpha = \frac{\varepsilon(pruned(T,t),S) - \varepsilon(T,S)}{|leaves(T) - |leaves(pruned(T,t))|} \qquad (19)$$

, όπου $\varepsilon(T,S)$ είναι το σφάλμα του δένδρου T_α ως προς το δείγμα S και τα «φύλλα» στο δένδρο T είναι ο όρος $leaves(T)$. Ο όρος $pruned(T,t)$ υποδηλώνει ότι δένδρο που δημιουργήθηκε αντικαθιστώντας τον κόμβο t στο δένδρο T με ένα κατάλληλο «φύλλο». Στο δεύτερο στάδιο υπολογίζεται το γενικευμένο σφάλμα από κάθε κλαδεμένο δένδρο T_0, T_1, …., T_k. Έπειτα, επιλέγεται το καλύτερο δένδρο. Γενικότερα, όταν υπάρχει ένα μεγάλο σύνολο δεδομένων, η βιβλιογραφία προτείνει τον διαχωρισμό της σε σύνολο εκπαίδευσης και σύνολο δοκιμής. Με αυτόν τον τρόπο το δένδρο κατασκευάζεται από ένα σύνολο εκπαίδευσης και αξιολογείται από το σύνολο δοκιμής.

Μειωμένο Σφάλμα «Κλαδέματος»

Είναι μια απλή διαδικασία «κλαδέματος» δένδρων απόφασης, η οποία προτάθηκε από τον Quinlan το 1987 (Quinlan, J.R., 1987). Διασχίζει από κάτω προς τα πάνω όλους τους κόμβους και ελέγχει αν επηρεάζεται η ακρίβεια του μοντέλου στην περίπτωση που

αντικαταστήσει τον εκάστοτε κόμβο με την κατηγορία που εμφανίζει την μεγαλύτερη συχνότητα. Για την μέτρηση της ακρίβειας ο Quinlan πρότεινε και αυτός να υπολογίζεται η ακρίβεια σε ένα άλλο σύνολο δεδομένων.

Ελάχιστο Σφάλμα «Κλαδέματος»

Το ελάχιστο σφάλμα «κλαδέματος» προτάθηκε από τους Olaru και Wehenkel το 2003 (Olaru, C., Wehenkel, 2003). Διασχίζει το δένδρο από κάτω προς τα πάνω συγκρίνοντας σε κάθε κόμβο την τιμή 1 – πιθανότητα ποσοστού σφάλματος με ή χωρίς «κλάδεμα». Η τιμή 1 – πιθανότητα ποσοστού σφάλματος είναι η διόρθωση στην απλή πιθανότητα εκτίμησης χρησιμοποιώντας συχνότητες. Αν το S_t υποδηλώνει τον αριθμό των παραδειγμάτων που έχουν φτάσει στο «φύλλο» t, τότε το προσδοκώμενο ποσοστό σφάλματος δίνεται από τον ακόλουθο τύπο:

$$\varepsilon'(t) = 1 - \max_{c_i \in dom(y)} \frac{\left|\sigma_{y=c_i} S_t\right| + l * p_{apr}(y = c_i)}{|S_t| + l} \qquad (20)$$

Το ποσοστό σφάλματος ενός εσωτερικού κόμβου είναι ο σταθμισμένος μέσος των ποσοστών σφαλμάτων των «κλαδιών» του. Το βάρος εκτιμάται, σύμφωνα με την αναλογία των δεδομένων που υπάρχουν σε κάθε «κλαδί». Οι υπολογισμοί γίνονται αναδρομικά από κάτω προς τα πάνω. Αν ένας εσωτερικός κόμβος «κλαδευτεί», τότε μετατρέπεται σε «φύλλο» και το ποσοστό σφάλματος υπολογίζεται από την εξίσωση (17). Κατά συνέπεια, υπολογίζεται το ποσοστό σφάλματος πριν και μετά το «κλάδεμα» στον συγκεκριμένο εσωτερικό κόμβο. Αν το «κλάδεμα» αυτού του κόμβου δεν αυξάνει το ποσοστό σφάλματος, τότε το «κλάδεμα» γίνεται αποδεκτό.

Βέλτιστο «Κλάδεμα» Δένδρου Απόφασης

Με το ζήτημα του βέλτιστου «κλαδέματος» δένδρου απόφασης έχουν ασχοληθεί οι Bratko, Bohanec το 1994 και ο Almuallim το 1996. Η πρώτη έρευνα δημιούργησε έναν αλγόριθμο που διασφαλίζει την εύρεση βέλτιστης λύσης, γνωστός και ως OPT. Ο συγκεκριμένος αλγόριθμος βρίσκει το βέλτιστο «κλάδεμα», σύμφωνα με δυναμικό προγραμματισμό με πολυπλοκότητα $\Theta(|leaves(T)|^2)$, όπου Τα είναι το αρχικό δένδρο. Η δεύτερη έρευνα συνέβαλε στην βελτίωση του αλγορίθμου OPT και ονομάστηκε OPT – 2 (Almuallim, H., 1996). Ο συγκεκριμένος αλγόριθμος βασίζεται και αυτός σε δυναμικό προγραμματισμό, όμως τώρα η πολυπλοκότητα είναι και ο χώρος και ο χρόνος, δηλαδή $\Theta(|leaves(T'')| * |internal(T)|)$, όπου Τ” είναι το τελικό δένδρο και Τα το αρχικό δένδρο. Αξίζει να σημειωθεί ότι επειδή το αρχικό δένδρο και ο αριθμός των εσωτερικών κόμβων που έχει είναι μεγαλύτερος από ότι το τελικό δένδρο ισχύει ότι ο OPT – 2 αλγόριθμος απαιτεί μικρότερο υπολογιστικό χρόνο από ότι ο OPT.

Σύγκριση Μεταξύ Μεθόδων «Κλαδέματος»

Έχουν πραγματοποιηθεί διάφορες έρευνες για την σύγκριση των διαφόρων μεθόδων «κλαδέματος» (Quinlan, J.R., 1987, Mingers, J., 1989, Esposito, F., 1997).Τα αποτελέσματα δείχνουν ότι κάποιες μέθοδοι, όπως το κόστος της πολυπλοκότητας του «κλαδέματος» και το μειωμένο σφάλμα «κλαδέματος» τείνουν να μειώνουν πολύ το μέγεθος του δένδρου δημιουργώντας μικρότερα δένδρα απόφασης αλλά με μικρότερη ακρίβεια. Άλλες μέθοδοι, όπως το ελάχιστο σφάλμα «κλαδέματος» έχουν την «προκατάληψη» να υποκλαδεύουν το δέντρο. Συμπερασματικά, αναφέρεται ότι δεν υπάρχει μέθοδος που να υπερέχει των υπολοίπων. Στις εφαρμογές της Weka εφαρμόζονται δύο μέθοδοι για την αποφυγή υπερπροσαρμογής του δένδρου: α) τίθεται περιορισμός για τον ελάχιστο αριθμό στοιχείων που θα υπάρχει σε κάθε «φύλλο», β) εφαρμόζεται η μέθοδος ελαχίστου σφάλματος «κλαδέματος» για την αντικατάσταση εσωτερικών κόμβων με «φύλλα».

3.6 ΔΑΣΗ ΑΠΟΦΑΣΗΣ ΜΕ ΒΑΣΗ ΤΗΝ ΜΕΘΟΔΟ ADABOOSTM1

3.61 ΣΥΝΤΟΜΗ ΕΙΣΑΓΩΓΗ

Η μέθοδος Boosting είναι μια γενική μέθοδος που συμβάλλει στην βελτίωση της ακρίβειας ενός αλγορίθμου εκμάθησης μηχανών. Πιο συγκεκριμένα, η μέθοδος boosting μπορεί να χρησιμοποιηθεί για να περιορίσει το σφάλμα ενός «αδύναμου» αλγορίθμου εκμάθησης μηχανών που παράγει ταξινομητές, οι οποίοι έχουν ακρίβεια λίγο μεγαλύτερη του 50%. Ο τρόπος που δουλεύει η μέθοδος είναι να επαναλαμβάνει πολλές φορές έναν «αδύναμο» αλγόριθμο σε διάφορες κατανομές του συνόλου δεδομένων που χρησιμοποιούνται για εκπαίδευση και εν συνεχεία να συνδυάσει αυτούς τους ταξινομητές σε έναν σύνθετο κατηγοριοποιητή. Ένας τέτοιος αλγόριθμος που είναι αποτελεσματικός και δεν απαιτεί μεγάλη υπολογιστική δύναμη για να δουλέψει είναι ο Adaboost. Αξίζει να σημειωθεί ότι η μέθοδος Adaboost είναι πολύ χρήσιμη σε περιπτώσεις που η ομάδα δεδομένων που

χρησιμοποιείται για εκπαίδευση έχει τις δύο παρακάτω ιδιότητες. Πρώτον, τα δεδομένα έχουν μεγάλο εύρος δυσκολίας εκμάθησης. Σε τέτοιου είδους προβλήματα η μέθοδος Adaboost παράγει κατανομές που συγκεντρώνονται στα δεδομένα που είναι δύσκολο για τον αλγόριθμο εκμάθησης να εκπαιδευτεί. Ο στόχος είναι σε αυτόν τον υποχώρο ο «αδύναμος» αλγόριθμος εκμάθησης να αποκτήσει μεγάλη ακρίβεια. Η δεύτερη ιδιότητα είναι ο αλγόριθμος εκμάθησης που χρησιμοποιείται να είναι ευμετάβλητος σε αλλαγές της υπόθεσης έτσι ώστε σημαντικά διαφορετικές υποθέσεις να δημιουργηθούν για το σύνολο εκπαίδευσης. Η μέθοδος Adaboost παίρνει έναν σταθμισμένο μέσο των υποθέσεων που έχουν δημιουργηθεί από τις διαφορετικές κατανομές δεδομένων. Κατά συνέπεια, περιορίζεται η υπερπροσαρμογή (overfitting) στα δεδομένα εκπαίδευσης. Επίσης, ο συγκεκριμένος αλγόριθμος ενδέχεται να περιορίσει την «προκατάληψη» (bias) του «αδύναμου» αλγορίθμου. Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθούν ως «αδύναμος» ταξινομητής τα δάση απόφασεις, ωστόσο η διαδικασία παρακάτω θα περιγραφεί γενικά για οποιοδήποτε αλγόριθμο εκμάθησης μηχανών, ώστε να γίνει πλήρως κατανοητή η λειτουργία της μεθόδου Adaboost. Δυο αλγόριθμοι έχουν προέλθει από την μέθοδο AdaboostM1, ο πρώτος είναι ο AdaboostM1 και ο δεύτερος είναι ο AdaboostM2. Παρακάτω παρουσιάζεται ο AdaboostM1 που χρησιμοποιείται και από τις εφαρμογές της Weka.

3.6.2 ΑΝΑΛΥΣΗ ΜΕΘΟΔΟΥ ADABOOSTM1

Η ανάλυση της διαδικασίας εκμάθησης θα πραγματοποιηθεί για τον αλγόριθμο AdaboostM1. Ο αλγόριθμος λαμβάνει m δεδομένα εισόδου για εκπαίδευση $S = \langle (x1, y1), \ldots., (xm, ym) \rangle$, όπου yi είναι η κατηγορία στην οποία ανήκει το παράδειγμα xi . Για τις ανάγκες επεξήγησης της παρούσας διαδικασίας θεωρείται ότι το yi μπορεί να πάρει οποιαδήποτε τιμή από ένα σύνολο k τιμών. Ο αλγόριθμος AdaboostM1 χρησιμοποιεί έναν «αδύναμο» αλγόριθμο που θα υποδηλώνεται WeakLearn. Ο αλγόριθμος AdaboostM1 θα καλέσει τον WeakLearn επαναλαμβανόμενα για ένα αριθμό φορών. Την φορά t, ο WeakLearn υπολογίζει έναν ταξινομητή h_t: X→ Y, ο οποίος πρέπει να ταξινομεί λάθος μόνο ένα μικρό ποσοστό των δεδομένων, D_t. Κατά συνέπεια, ο στόχος του WeakLearn είναι να βρει μια υπόθεση h_t που να ελαχιστοποιεί το σφάλμα $\varepsilon_t = Pr_{i\sim Dt}[h_t(x) \neq yi]$. Σημειώνεται ότι το σφάλμα μετράται για την συγκεκριμένη κατανομή D_t που έχει δοθεί στον «αδύναμο» αλγόριθμο. Αυτή η διαδικασία επαναλαμβάνεται για Τα φορές και στον τέλος η μέθοδος Adaboost συνδυάζει τις «αδύναμες» υποθέσεις h_1, ….., h_T σε μια τελική υπόθεση h_{fin}.

Επίσης, πρέπει να εξηγηθεί πως διαμορφώνεται το D_t κάθε φορά και πως υπολογίζεται η υπόθεση h_{fin}. Τα δύο παραπάνω θα επεξηγηθούν για τον αλγόριθμο AdaboostM1 που είναι

αυτός που χρησιμοποιείται για την εξαγωγή των αποτελεσμάτων της παρούσας διπλωματικής εργασίας. Η αρχική κατανομή D_1 είναι ομοιόμορφη σε όλα τα παραδείγματα του S και ισούται με $D_1(i) = 1/m$ για κάθε παράδειγμα i. Για τον υπολογισμό της κατανομής D_{t+1} από την κατανομή D_t και την αδύναμη υπόθεση h_t, θα πολλαπλασιαστεί το βάρος του παραδείγματος i με έναν αριθμό $\beta_t \in \{0, 1\}$ αν η h_t κατηγοριοποιεί σωστά το παράδειγμα x_i αλλιώς το βάρος δεν θα αλλάζει. Όλα τα βάρη ομαλοποιούνται διαιρώντας τα με μια σταθερά ομαλοποίησης Z_t . Κατά συνέπεια, γίνεται κατανοητό ότι τα παραδείγματα, τα οποία έχουν σωστά ταξινομηθεί από τις προηγούμενες «αδύναμες» υποθέσεις αποκτούν μικρότερο βάρος, ενώ τα υπόλοιπα παραδείγματα αποκτούν μεγαλύτερο βάρος. Επισημαίνεται ότι ο αλγόριθμος AdaboostM1 επικεντρώνεται στα παραδείγματα που ο «αδύναμος» αλγόριθμος αδυνατεί να ταξινομήσει σωστά.

Η τελική υπόθεση h_{fin} είναι ένα σταθμισμένο άθροισμα «αδύναμων» υποθέσεων. Για παράδειγμα, για ένα παράδειγμα x εκτός της ομάδας δεδομένων, εξάγει μέσω της υπόθεσης h_{fin} την κατηγορία y για την οποία μεγιστοποιείται το άθροισμα των βαρών των «αδύναμων» υποθέσεων που προβλέπουν αυτή την κατηγορία. Το βάρος σε κάθε υπόθεση h_t δίνεται από το $\log(\frac{1}{\beta_t})$ και κατά συνέπεια το μεγαλύτερο βάρος δίνεται στην υπόθεση με το μικρότερο σφάλμα. Ο αριθμός β_t και το σφάλμα ε_t υπολογίζονται παρακάτω:

$$\varepsilon_t = \sum_{i:ht(xi)\neq yi} D_t(i) \qquad (21)$$

$$\beta t = \frac{\varepsilon t}{1-\varepsilon t} \qquad (22)$$

$$\mathrm{D_{t+1}} = \frac{\mathrm{D_t(i)}}{\mathrm{Z_t}} * \begin{Bmatrix} \beta \mathrm{t}\,, \alpha\nu\ \mathrm{ht(xi)} = \mathrm{yi} \\ 1 \end{Bmatrix} \qquad (23)$$

$$\mathrm{h_{fin}(x)} = \arg\max_{\mathrm{y}\epsilon\mathrm{Y}} \sum_{\mathrm{t:ht(x)=y}} \log\left(\frac{1}{\beta \mathrm{t}}\right) \qquad (24)$$

Ένα σημαντικό χαρακτηριστικό των αλγορίθμων AdaboostM1 παρουσιάζεται στο ακόλουθο θεώρημα. Αυτό το θεώρημα δείχνει ότι αν μια «αδύναμη» υπόθεση έχει συνεχώς σφάλμα λίγο μεγαλύτερο από το 1/2, τότε το σφάλμα της τελικής υπόθεσης h_{fin} μειώνεται εκθετικά στο μηδέν πολύ γρήγορα. Για διωνυμικά προβλήματα αυτό σημαίνει ότι οι «αδύναμες» υποθέσεις πρέπει να είναι λίγο μεγαλύτερες της τυχαιότητας.

Θεώρημα: Έστω ότι ένας αλγόριθμος μάθησης WeakLearn, όταν καλείται από τον αλγόριθμο AdaboostM1 παράγει υποθέσεις με σφάλματα ε_1, ……, ε_T, όπως υπολογίζονται από την εξίσωση (21). Υποθέτοντας ότι $\varepsilon_t \leq 1/2$ και $\gamma_t = 1/2 - \varepsilon_t$. Παρακάτω φαίνονται τα πάνω όρια των σφαλμάτων της τελικής υπόθεσης h_{fin} :

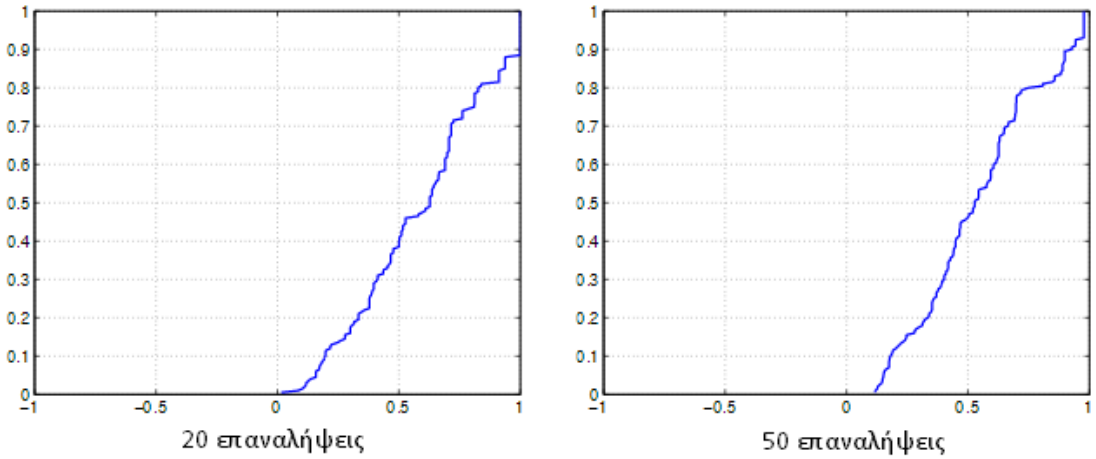
$$\frac{1}{m}\left|\{i : h_{fin}(x_i) \neq y_i\}\right| \leq \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2} \leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) \qquad (25)$$

Το κύριο μειονέκτημα του αλγορίθμου AdaboostM1 είναι ότι δεν μπορεί να χειριστεί αδύναμες υποθέσεις με σφάλμα μεγαλύτερο από 1/2. Το αναμενόμενο σφάλμα της υπόθεσης, η οποία τυχαία υποθέτει την κατηγορία, είναι 1 – 1/k (k είναι ο αριθμός των πιθανών κατηγοριών). Έτσι, για ένα διωνυμικό πρόβλημα (k = 2) η απαίτηση του αλγορίθμου AdaboostM1 είναι η πρόβλεψη να έχει ακρίβεια λίγο μεγαλύτερη του 50%. Όμως, αν το k> 2, η απαίτηση του αλγορίθμου είναι αρκετά μεγαλύτερη από 50% και αυτό είναι ένα σημαντικό εμπόδιο στην χρήση του.

Περιθώρια (Margins)

Ο ρόλος των περιθωρίων για τους αλγορίθμους AdaboostM1 είναι πολύ σημαντικός, καθώς δείχνει την αξιοπιστία του μοντέλου πρόβλεψης. Αν η συγκεκριμένη μέθοδος τρέξει έναν «αδύναμο» κατηγοριοποιητή πολλές φορές είναι βέβαιο ότι μετά από κάποιες φορές η ακρίβεια στο σύνολο εκπαίδευσης θα αυξηθεί σε σημαντικό βαθμό και αντίστοιχα το σφάλμα σχεδόν θα μηδενιστεί. Όμως, για να επιβεβαιωθεί ότι ο αλγόριθμος δεν έχει υπερπροσαρμοστεί στο σύνολο εκπαίδευσης και έχει καταφέρει να γενικεύσει την πληροφορία απαιτείται ένα άλλο μέτρο, τα περιθώρια.

Για την συγκεκριμένη μέθοδο τα περιθώρια θα οριστούν ως η διαφορά μεταξύ των σταθμισμένων ποσοστών (όπως προέκυψαν από τους «αδύναμους» αλγορίθμους) των ορθών προβλέψεων και των σταθμισμένων ποσοστών των μη ορθών προβλέψεων. Τα περιθώρια έχουν πεδίο τιμών (-1,1) και ένα περιθώριο είναι θετικό, όταν η συνδυαστική πρόβλεψη $\mathrm{h_{fin}}$ είναι θετική. Παρατηρείται ότι εκεί που αρχίζει να ελαχιστοποιείται το σφάλμα στο σύνολο εκπαίδευσης η μεταβολή της αθροιστικής κατανομής του περιθωρίου είναι ελάχιστη. Έπειτα, από αρκετές φορές (π.χ. 30 – 40) η αθροιστική κατανομή του περιθωρίου αρχίζει να αυξάνεται. Υψηλές τιμές περιθωρίων στο σύνολο εκπαίδευσης αποτελούν ένδειξη ότι ο αλγόριθμος AdaboostM1 δεν έχει υπερπροσαρμοστεί στα δεδομένα και είναι αξιόπιστος για προβλέψεις και σε άλλα σύνολα δεδομένων. Παρακάτω στην εικόνα 3.4 παρουσιάζονται οι αθροιστικές κατανομές των περιθωρίων σε δύο διαφορετικά σύνολα εκπαίδευσης.



Εικόνα 3.4: Αθροιστικές κατανομές περιθωρίες σε δύο διαφορετικά σύνολα εκπαίδευσης

Εκθετικό Σφάλμα

Ο αλγόριθμος AdaboostM1 αν και δεν είχε κατασκευαστεί προκειμένου να μειώνει κάποια συνάρτηση σφάλματος, παρ΄ όλα αυτά μειώνει την συνάρτηση εκθετικού σφάλματος. Η συνάρτηση εκθετικού σφάλματος είναι η παρακάτω:

$$C = \frac{1}{m} * \sum_{i=1}^{m} \exp(-yi * \mathrm{h_{fin}(x)}) \ , \qquad (26)$$

Από την παραπάνω συνάρτηση προκύπτει ότι η επιλογή των επιμέρους υποθέσεων κάθε φορά μπορεί να μειώσει σημαντικά το σφάλμα. Αυτή η παρατήρηση έγινε για πρώτη φορά από τον Breiman (1999). Ειδικότερα, η ελαχιστοποίηση του εκθετικού σφάλματος συνδέεται άμεσα με την εύρεση της ορθής υπόθεσης $\mathrm{h_{fin}}$ που θα συμφωνεί με το αντίστοιχο yi. Κατά συνέπεια, αυτός είναι και ο τελικός στόχος, η ελαχιστοποίηση των μη ορθών κατηγοριοποιήσεων. Επίσης, η ελαχιστοποίηση των μη ορθών κατηγοριοποιήσεων απαιτεί την βελτιστοποίηση μιας αντικειμενικής συνάρτησης που δεν είναι παντού συνεχής ή να έχει πολλά τοπικά ελάχιστα. Μια τέτοια αντικειμενική συνάρτηση είναι η συνάρτηση εκθετικού σφάλματος.

Έτσι, θα μπορούσε κανείς να εξάγει το συμπέρασμα ότι η ελαχιστοποίηση της συνάρτησης εκθετικού σφάλματος οδηγεί σε σύγκλιση του αλγορίθμου AdaboostM1. Όμως, κάτι τέτοιο δεν ισχύει, αφού η ελαχιστοποίηση του εκθετικού σφάλματος δεν είναι επαρκής συνθήκη για την εξασφάλιση μικρού γενικού σφάλματος ε_t , όπως ορίστηκε από τον τύπο (21). Αντίθετα, είναι πολύ πιθανό να μειωθεί σε σημαντικό βαθμό το εκθετικό σφάλμα έχοντας μειωθεί ελάχιστα το σφάλμα ε_t .

3.7 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΜΑΘΗΣΗΣ ΜΗΧΑΝΩΝ

Τα μέτρα αξιολόγησης των μεθόδων εκμάθησης μηχανών είναι κοινά και για τις τρείς μεθόδους που αναφέρθηκαν στις ενότητες 3.4, 3.5, 3.6. Τα κριτήρια αυτά είναι κάποια στατιστικά μέτρα που σχετίζονται με το είδος του υπό επίλυση προβλήματος. Στην παρούσα διπλωματική εργασία θα επιλυθεί πρόβλημα κατηγοριοποίησης, στο οποίο αντιστοιχούν συγκεκριμένα κριτήρια αξιολόγησης. Επίσης, αξίζει να σημειωθεί ότι τα στατιστικά μέτρα εφαρμόζονται στα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής για να αποφευχθούν ή να εντοπιστούν προβλήματα υπερπροσαρμογής, αλλά και να αξιολογηθεί η ικανότητα πρόβλεψης και αξιοπιστίας του μοντέλου. Ο στόχος είναι η εφαρμογή αυτών των κριτηρίων για την εύρεση πιθανών λαθών του μοντέλου, ώστε τελικά να διορθωθούν και να βελτιωθεί η ακρίβεια του μοντέλου. Παρατίθεται παρακάτω ο πίνακας από τον οποίο θα εξαχθούν αυτά τα στατιστικά μέτρα.

	A1	A2
B1	Γ1	Γ2
B2	Γ3	Γ4

A1 , A2 : Πρόβλεψη θετικών και αρνητικών, αντίστοιχα
B1, B2 : Πραγματικά θετικά και αρνητικά, αντίστοιχα
Γ1, Γ4 : Ορθή πρόβλεψη θετικών και αρνητικών, αντίστοιχα
Γ2, Γ3 : Εσφαλμένη πρόβλεψη αρνητικών και θετικών, αντίστοιχα

Συνολική Ακρίβεια (Accuracy): Δείχνει πόσο ορθό είναι το μοντέλο πρόβλεψης, δηλαδή πόσο κοντά είναι οι προβλέψεις με τις πραγματικές τιμές. Ο τύπος είναι : $\frac{\Gamma 1+\Gamma 4}{\Gamma 1+\Gamma 2+\Gamma 3+\Gamma 4}$

Ακρίβεια (Precision): Δείχνει το ποσοστό θετικών προβλέψεων που έχουν προβλεφθεί ορθά. Ο τύπος είναι : $\frac{\Gamma 1}{\Gamma 1+\Gamma 3}$

Ανάκληση (Recall): Δείχνει το ποσοστό των πραγματικών θετικών τιμών που έχουν προβλεφθεί ορθά. Ο τύπος είναι: $\frac{\Gamma 1}{\Gamma 1+\Gamma 2}$

Ευαισθησία (Sensitivity): Δείχνει το ποσοστό των πραγματικών θετικών τιμών που έχουν προβλεφθεί ορθά, ονομάζεται και Recall. Ο τύπος είναι: $\frac{\Gamma 1}{\Gamma 1+\Gamma 2}$

Εξειδίκευση (Specificity): Δείχνει το ποσοστό των πραγματικών αρνητικών τιμών που έχουν προβλεφθεί ορθά. Ο τύπος είναι: $\frac{\Gamma 4}{\Gamma 3+\Gamma 4}$

Ποσοστό αστοχίας (Miss – Rate): Δείχνει το ποσοστό των μη ορθών προβλέψεων αρνητικών τιμών. Ο τύπος είναι: $\frac{\Gamma 2}{\Gamma 1+\Gamma 2}$

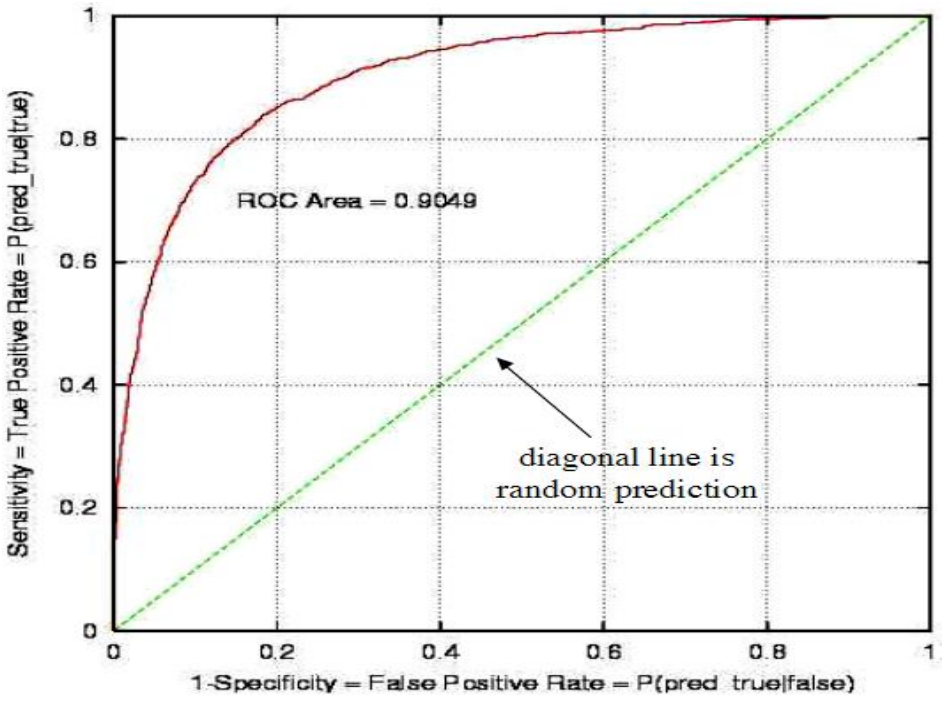
Ποσοστό σφάλματος (Fall – out): Δείχνει το ποσοστό των μη ορθών προβλέψεων θετικών τιμών. Ο τύπος είναι: $\frac{\Gamma 3}{\Gamma 3+\Gamma 4}$

Αρμονικός μέσος (F – measure): Όταν θετικές και αρνητικές τιμές (1/0, αντίστοιχα) δεν είναι περίπου ισοκατανεμημένες, δηλαδή τείνουν να είναι περισσότερο θετικές ή περισσότερο αρνητικές, τότε συνιστάται να χρησιμοποιείται αυτό το μέτρο αντί της συνολικής ακρίβειας (accuracy). Ο τύπος είναι: $F = 2 * \frac{precision*recall}{precision+recall}$

MCC: Χρησιμοποιείται στις μεθόδους εκμάθησης μηχανών ως ένα μέτρο ακρίβειας των προβλημάτων κατηγοριοποίησης. Είναι μια μορφή γραμμικής συσχέτισης μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Οι τιμές που παίρνει είναι από -1 έως 1, όπως και ο συντελεστής γραμμικής συσχέτισης. Η τιμή 1 δείχνει ότι υπάρχει τέλεια πρόβλεψη, ενώ η τιμή 0 δείχνει ότι η πρόβλεψη είναι τυχαία. Ο τύπος είναι: $\frac{\Gamma 1*\Gamma 4-\Gamma 3*\Gamma 2}{\sqrt{(\Gamma 1+\Gamma 3)(\Gamma 1+\Gamma 2)(\Gamma 4+\Gamma 3)(\Gamma 4+\Gamma 2)}}$

Για όλα τα παραπάνω στατιστικά μέτρα εξαιρουμένων των Miss – Rate και Fall – out είναι επιθυμητό να βρίσκονται όσο το δυνατόν κοντά στο 1. Όσο πιο κοντά στο 1 είναι οι παραπάνω τιμές τόσο πιο ακριβείς είναι οι προβλέψεις ενός μοντέλου.

ROC Plot: Η αναπαριστά τα στατιστικά μέτρα Sensitivity και 1 – Specificity σε ένα διάγραμμα. Είναι ένας τρόπος μέτρησης της αξιοπιστίας πρόβλεψης του μοντέλου. Το εμβαδόν κάτω από την καμπύλη ROC ενδιαφέρει τον ερευνητή. Όσο πιο αριστερά είναι η καμπύλη, δηλαδή το εμβαδόν της είναι κοντά στο 1, τόσο μεγαλύτερη αξιοπιστία παρουσιάζουν οι προβλέψεις του μοντέλου. Αντίθετα, αν το εμβαδόν που περικλείει η καμπύλη ROC είναι περίπου 0,5 οι προβλέψεις είναι τυχαίες. Στο διάγραμμα 3.2 παρατίθενται δυο καμπύλες ROC (Cornell, 2003).



Διάγραμμα 3.2 : Παραδείγματα καμπυλών ROC

3.8 ΕΠΙΛΟΓΗ ΛΟΓΙΣΜΙΚΟΥ

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε η γλώσσα προγραμματισμού Matlab, η οποία έχει φιλικό περιβάλλον για τον χρήστη. Η συγκεκριμένη γλώσσα χρησιμοποιείται για μαθηματικές ή στατιστικές εφαρμογές και εφαρμόζεται τόσο από καθηγητές όσο και μαθητές για ερευνητικούς σκοπούς. Επειδή έχει ιδιαίτερα φιλικό περιβάλλον αξιοποιείται από διαφόρους κλάδους και επιστήμες. Επίσης, αξίζει να σημειωθεί ότι πολλές από τις μεθόδους εκμάθησης μηχανών έχουν ήδη κωδικοποιηθεί από ερευνητές και είναι διαθέσιμες για τους χρήστες. Τα πλεονεκτήματά της έναντι άλλων γλωσσών οφείλονται στο γεγονός ότι έχει πολλές γραφικές και στατιστικές συναρτήσεις, οι οποίες δεν είναι διαθέσιμες σε άλλα λογισμικά.

Στο συγκεκριμένο γλωσσικό περιβάλλον κατασκευάστηκε ο αλγόριθμος, ο οποίος εκπαιδεύει ένα νευρωνικό δίκτυο με ένα κρυμμένο επίπεδο με την μέθοδο της οπισθοδρόμησης σφάλματος (backpropagation). Προτού αξιοποιηθεί ο αλγόριθμος για την εκπαίδευση των νευρώνων στο σύνολο εκπαίδευσης κατασκευάστηκε μια συνάρτηση, η οποία επαληθεύει ότι η οπισθοδρόμηση σφάλματος (backpropagation) είναι σωστή. Ο αλγόριθμος έχει δημιουργηθεί, ώστε να λαμβάνει από τον χρήστη ως πληροφορίες τον αριθμό των μεταβλητών εισόδου, τον αριθμό των νευρώνων στο κρυμμένο επίπεδο, τον αριθμό των μεταβλητών εξόδου και τον συντελεστή ομαλοποίησης (λ). Έπειτα από την εκπαίδευση των νευρώνων ο αλγόριθμος εξάγει την ακρίβεια του μοντέλου στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής.

Για την εύρεση της βέλτιστης παραμέτρου ομαλοποίησης (λ) δημιουργήθηκε μια συνάρτηση, η οποία χρησιμοποιεί τον παραπάνω αλγόριθμο για διάφορες τιμές των παραμέτρων ομαλοποίησης (λ) και υπολογίζει τις τιμές των κόστων (C) για το σύνολο εκπαίδευσης και το σύνολο επικύρωσης. Εν συνεχεία αποθηκεύει αυτές τις τιμές σε έναν πίνακα και εξάγει στο χρήστη ένα γράφημα που στον κατακόρυφο άξονα έχει τις τιμές των κόστων (C) και στον οριζόντιο άξονα τις παραμέτρους ομαλοποίησης (λ). Κατά συνέπεια, ο χρήστης παρατηρεί που ελαχιστοποιείται το κόστος (C) του συνόλου επικύρωσης και βρίσκει την βέλτιστη παράμετρο ομαλοποίησης (λ). Όλα τα αποτελέσματα που εξήχθηκαν από τον κώδικα στην Matlab ελέγχθηκαν και από την εφαρμογή MLP του λογισμικού Weka.

Επιπλέον, στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε το λογισμικό Weka, το οποίο περιέχει μεγάλο πλήθος μεθόδων εκμάθησης μηχανών. Πρόκειται για ένα ανοιχτό λογισμικό, το οποίο μπορεί να χρησιμοποιηθεί από οποιονδήποτε ερευνητή. Το συγκεκριμένο λογισμικό κατασκευάστηκε από διδακτορικούς φοιτητές και καθηγητές του Πανεπιστημίου Waikato. Σημειώνεται ότι υπάρχει διαρκής συντήρηση και αναβάθμιση του

λογισμικού με νεότερες εκδόσεις, οι οποίες εντάσσουν καινούριες μεθόδους εκμάθησης μηχανών ή/και αναβαθμίζουν τις υπάρχουσες.

Τα δένδρα απόφασης που δημιουργήθηκαν για την παρούσα διπλωματική εργασία κατασκευάστηκαν από το λογισμικό της Weka εφαρμόζοντας την μέθοδο Decision Trees J48. Το λογισμικό της Weka εκτός από το δένδρο απόφασης παρέχει στο χρήστη διάφορα μέτρα αξιολόγησης, ώστε να μπορεί ο ίδιος να αξιολογήσει την εγκυρότητα του μοντέλου του. Επίσης, δίνεται στον χρήστη η δυνατότητα να χωρίσει τα δεδομένα σε σύνολο εκπαίδευσης και δοκιμής ή να χρησιμοποιήσει την μέθοδο k – fold validation, όταν δεν υπάρχουν αρκετά δεδομένα. Άλλη μια δυνατότητα του λογισμικού είναι, αφού πρώτα κατασκευαστεί το εκάστοτε μοντέλο να χρησιμοποιηθεί μια καινούρια βάση δεδομένων και να αξιολογηθεί και εκεί η εγκυρότητα του μοντέλου.

Η τρίτη μέθοδος που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία ονομάζεται AdaboostM1 και χρησιμοποιεί ως «αδύναμο» κατηγοριοποιητή τα δένδρα απόφασης. Το λογισμικό της Weka χρησιμοποιήθηκε και για αυτή την μέθοδο. Ο χρήστης έχει την δυνατότητα να καθορίσει τον αριθμό των «αδύναμων» κατηγοριοποιητών και αυτός αποτελεί και ένας τρόπος αποφυγής της υπερπροσαρμογής στο σύνολο εκπαίδευσης.

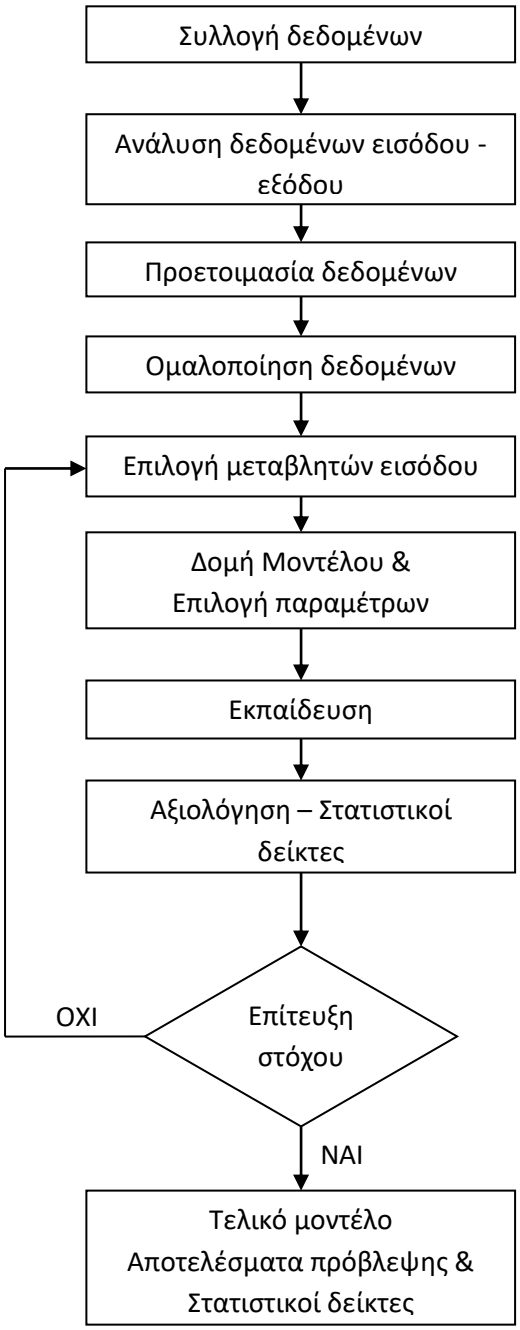
Τέλος, αξίζει να αναφερθεί ότι οι άνθρωποι που ασχολήθηκαν με την κατασκευή του συγκεκριμένου λογισμικού έχουν δώσει ιδιαίτερη έμφαση, ώστε οι αλγόριθμοι να είναι αποτελεσματικοί και να απαιτούν πολύ μικρό υπολογιστικό χρόνο.

3.9 ΔΙΑΓΡΑΜΜΑ ΡΟΗΣ ΠΕΙΡΑΜΑΤΟΣ

Για την εκτέλεση του πειράματος της παρούσας διπλωματικής εργασίας ακολουθείται μια σειρά ενεργειών, οι οποίες περιγράφονται από το Διάγραμμα Ροής (Διάγραμμα 3.9). Αρχικά, συλλέγονται τα δεδομένα της χρονοσειράς της απλής αμόλυβδης και του αργού πετρελαίου σε εβδομαδιαία κλίμακα. Εν συνεχεία γίνεται ανάλυση τόσο των δεδομένων εισόδου όσο και εξόδου. Συγκεκριμένα, δεδομένου ότι το πρόβλημα που επιλύει η παρούσα διπλωματική εργασία είναι τύπου 0/1, μετατρέπεται η μεταβλητή εξόδου από συνεχής σε διακριτή παίρνοντας τιμές 0 ή 1 για καθοδική ή ανοδική κατεύθυνση της τιμής της απλής αμόλυβδης αντίστοιχα. Παράλληλα κατασκευάζονται οι μεταβλητές των κινητών μέσων όρων που βασίζονται στην χρονοσειρά της απλής αμόλυβδης. Έπειτα, γίνεται προετοιμασία και ομαλοποίηση της χρονοσειράς της απλής αμόλυβδης και του αργού πετρελαίου, ώστε τα δεδομένα εισόδου να έχουν πεδίο τιμών (0,1) και η κατανομή τους να προσεγγίζει την «κανονική» για την καλύτερη εκπαίδευση της εκάστοτε μεθόδου

εκμάθησης μηχανών.

Έτσι, έχοντας ολοκληρώσει όλες τις παραπάνω ενέργειες η διαδικασία για κάθε μέθοδο εκμάθησης μηχανών είναι κοινή. Αξιοποιείται η διαδικασία δοκιμής – σφάλματος για την εύρεση των κατάλληλων μεταβλητών εισόδου, της δομής του μοντέλου και των βέλτιστων παραμέτρων από το σύνολο επικύρωσης. Στη συνέχεια, αξιοποιώντας τις βέλτιστες παραμέτρους και την δομή του μοντέλου από την προηγούμενη διαδικασία πραγματοποιείται η τελική εκπαίδευση του μοντέλου και παρατίθονται τα αποτελέσματα πρόβλεψης και οι στατιστικοί δείκτες του τελικού μοντέλου.



Διάγραμμα 3.3: Διάγραμμα Ροής Πειράματος

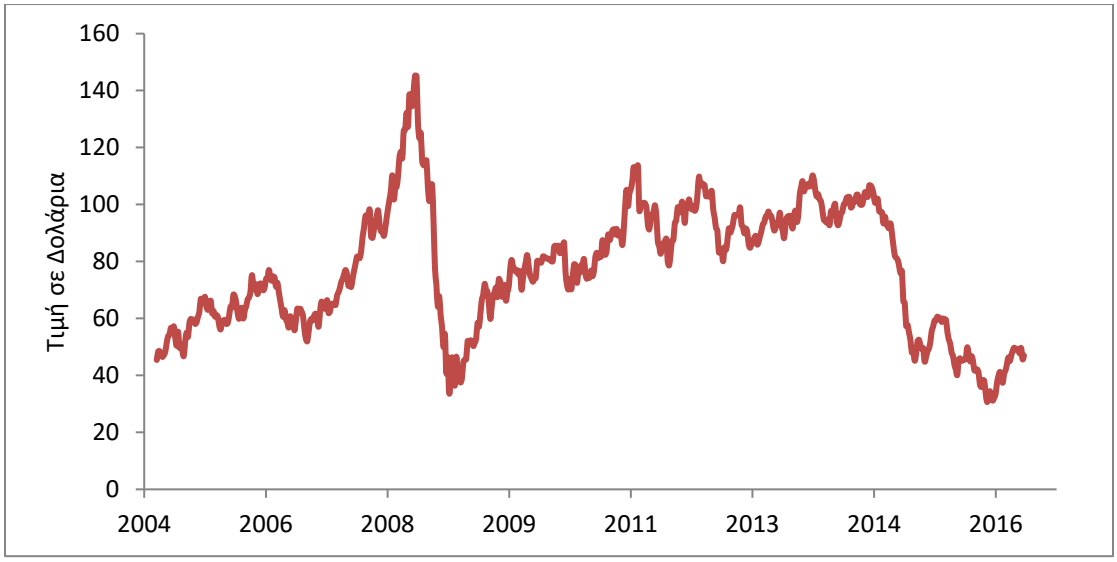
4. ΣΥΛΛΟΓΗ & ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

4.1 ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Για την παρούσα διπλωματική εργασία χρησιμοποιήθηκε η χρονοσειρά της απλής αμόλυβδης (Euro Super 95) και του αργού πετρελαίου για την περίοδο 2005 – 2016. Για την επίλυση του προβλήματος θεωρήθηκε ότι οποιαδήποτε μεταβολή της τάσης της τιμής της αμόλυβδης μπορεί να επεξηγηθεί από την χρονοϊστορία της τιμής της απλής αμόλυβδης, την χρονοϊστορία της τιμής του αργού πετρελαίου και από κάποιες μεταβλητές που σχετίζονται με τους κινητούς μέσους όρους που εφαρμόζονται στην χρονοσειρά της απλής αμόλυβδης και θα επεξηγηθούν αναλυτικά παρακάτω.

Συλλογή Χρονοσειράς Αργού Πετρελαίου

Η συλλογή των δεδομένων του αργού πετρελαίου σε εβδομαδιαία κλίμακα έγινε από μια ηλεκτρονική πλατφόρμα συναλλαγών Metatrader. Οι τιμές που λήφθηκαν είναι σε εβδομαδιαία βάση και αφορούν τιμές κλεισίματος. Ειδικότερα, η χρονοσειρά αυτή έχει δημιουργηθεί από την λήψη της τιμής του αργού πετρελαίου κάθε Παρασκευή και ώρα 12 μ.μ. (GMT: +2.00). Σημειώνεται ότι οι τιμές του αργού πετρελαίου είναι σε δολάριο, αφού οι αγοραπωλησίες αργού πετρελαίου πραγματοποιούνται σε αυτό το νόμισμα. Στο διάγραμμα 4.1 παρατίθεται η χρονοσειρά του αργού πετρελαίου για την περίοδο 2005 – 2016 από το οποίο φαίνεται ότι ο μέσος όρος και η τυπική απόκλιση δεν είναι σταθερά καθ' όλη την περίοδο. Για το λόγο αυτό, όπως θα παρατεθεί στο υποκεφάλαιο 4.2 θα επεξεργαστούν κατάλληλα τα δεδομένα, προκειμένου να αντιμετωπιστεί η μη στασιμότητα των χρονοσειρών, προτού αυτές χρησιμοποιηθούν από τις μεθόδους εκμάθησης μηχανών.



Διάγραμμα 4.1: Χρονοσειρά Αργού Πετρελαίου κατά την περίοδο 2005 - 2016

Μεθοδολογία Συγκέντρωσης Τιμών Απλής Αμόλυβδης

Αρχικά, σημειώνεται ότι η χρονοσειρά της απλής αμόλυβδης αναφέρεται στις λιανικές τιμές της απλής αμόλυβδης, δηλαδή τις τιμές που υπάρχουν στα πρατήρια βενζίνης και αφορούν άμεσα τους καταναλωτές. Για την συγκέντρωση της σταθμισμένης τιμής της απλής αμόλυβδης στις χώρες της Ε.Ε. πρέπει πρώτα να συγκεντρωθούν οι τιμές του κάθε κράτους - μέλους ξεχωριστά. Η μεθοδολογία που ακολουθεί η Ευρωπαϊκή Επιτροπή είναι κοινή για κάθε κράτος – μέλος και παρακάτω θα χρησιμοποιηθεί το παράδειγμα της Ελλάδας, ώστε να επεξηγηθεί η συγκεκριμένη μεθοδολογία. Συγκεκριμένα, για την συγκέντρωση των ημερήσιων τιμών της απλής αμόλυβδης χρησιμοποιούνται οι τιμές από το 70% των πρατηρίων που υπάρχουν στην Ελλάδα και βάσει του όγκου πώλησης του εκάστοτε πρατηρίου καυσίμων εξάγεται ο σταθμισμένος μέσος όρος της ημερήσιας λιανικής τιμής. Έπειτα, η Ευρωπαϊκή Επιτροπή συγκεντρώνει καθημερινά την λιανική τιμή της απλής αμόλυβδης και κάθε Παρασκευή υπολογίζει τον μέσο όρο των τιμών όλης της εβδομάδας. Έτσι, εξάγεται η μία εβδομαδιαία παρατήρηση για την Ελλάδα. Αξίζει να σημειωθεί ότι στην Ελλάδα τα πρατήρια καυσίμων ανήκουν σε μεγάλες εταιρίες εμπορίας καυσίμων και κατά συνέπεια η συγκέντρωση των παραπάνω τιμών δεν απαιτεί πολύ χρόνο. Η διαδικασία αυτή ακολουθείτε και από τις υπόλοιπες χώρες. Για λογαριασμό της Ευρωπαϊκής Επιτροπής καταρτίζεται το παραπάνω δελτίο τιμών από την διεύθυνση της Πετρελαϊκής Πολιτικής του Υπουργείου Περιβάλλοντος, Ενεργειακής και Κλιματικής Αλλαγής κάθε κράτους – μέλους.

Η Ευρωπαϊκή Επιτροπή έχοντας συλλέξει τις εβδομαδιαίες παρατηρήσεις των λιανικών τιμών της απλής αμόλυβδης εξάγει τον σταθμισμένο μέσο όρο των παραπάνω τιμών βάσει του ΑΕΠ κάθε χώρας. Τελικώς, αυτή η χρονοσειρά που δημιουργείται είναι αυτή που χρησιμοποιείται στην παρούσα διπλωματική εργασία και επισημαίνεται ότι δεν περιλαμβάνει σταθερούς φόρους και τον φόρο προστιθέμενης αξίας των κρατών – μελών της Ε.Ε. Από την ιστοσελίδα της Ευρωπαϊκής επιτροπής έγινε λήψη της χρονοσειράς της απλής αμόλυβδης. Παρακάτω στον Πίνακα 4.1 παρατίθενται τα χαρακτηριστικά της

ονομασίας της χρονοσειράς που επιλέχθηκε, όπως αναγράφονται από την Ευρωπαϊκή Επιτροπή.

Πίνακας 4.1: Χρονοσειρές διαφόρων τύπων καυσίμων, όπως αναφέρονται στην ιστοσελίδα της Ευρωπαϊκής Επιτροπής

Consumer prices of petroleum products net of duties and taxes - EU weighted average

Date	Euro-super 95(I)	Gas oil automobileAutomotive gas oilDieselkraftstoff(I)	Gas oil de chauffageHeating gas oilHeizöl(II)	Fuel oil - Schweres Heizöl(III)Soufre	Fuel oil - Schweres Heizöl(III)Soufre > 1%Sulphur > 1%Schwefel > 1%	GPL pour moteurLPG motor fuel
	1000L	1000L	1000L	t	t	1000L

Στο διάγραμμα 4.2 παρατίθεται η χρονοσειρά της λιανικής τιμής της απλής αμόλυβδης κατά την περίοδο 2005 – 2016. Παρατηρώντας τα διαγράμματα 4.1 και 4.2 φαίνεται ότι και οι δυο χρονοσειρές παρουσιάζουν προβλήματα μη στασιμότητας, καθώς δεν έχουν σταθερή μέση τιμή και τυπική απόκλιση κατά την περίοδο της έρευνας. Για το λόγο αυτό, όπως θα παρατεθεί στο υποκεφάλαιο 4.2 θα επεξεργαστούν κατάλληλα τα δεδομένα, προκειμένου να αντιμετωπιστεί η μη στασιμότητα των χρονοσειρών, προτού αυτές χρησιμοποιηθούν από τις μεθόδους εκμάθησης μηχανών.

Διάγραμμα 4.2: Χρονοσειρά Απλής Αμόλυβδης (Euro Super 95) κατά την περίοδο 2005 - 2016

Ανάλυση Δεδομένων Εισόδου και Εξόδου

Όπως, έχει αναφερθεί στο υποκεφάλαιο 3.1 το πρόβλημα που θα επιλυθεί στην παρούσα διπλωματική εργασία είναι η πρόβλεψη της τάσης της απλής αμόλυβδης (ανοδική/καθοδική) με χρονικό ορίζοντα μίας εβδομάδας. Για το λόγο αυτό η χρονοσειρά της απλής αμόλυβδης που θα αποτελεί και την μεταβλητή εξόδου θα πρέπει να μετατραπεί από συνεχής σε διακριτή μορφή με τιμές 0/1, όπου 0 καθοδική τάση και 1 ανοδική τάση. Επομένως, χρησιμοποιώντας την εντολή "if" στο υπολογιστικό φύλλο δημιουργείται μια καινούρια στήλη στην οποία συγκρίνεται η τιμή της απλής αμόλυβδης με την προηγούμενη τιμή και τοποθετείται ο αριθμός 1 στην καινούρια στήλη αν η τελευταία τιμή είναι μεγαλύτερη της προηγούμενης αλλιώς τοποθετείται ο αριθμός 0. Η μετατροπή αυτή φαίνεται και παρακάτω:

$$Out_i = \begin{cases} 1, & αν\ \ Fuel_i > Fuel_{i-1} \\ 0, & διαφορετικά \end{cases} \qquad (27)$$

Αυτή η στήλη αποτελεί την μεταβλητή εξόδου και θα χρησιμοποιηθεί από τα σύνολα εκπαίδευσης και επικύρωσης για την εκπαίδευση της εκάστοτε μεθόδου εκμάθησης μηχανών και από το σύνολο δοκιμής για εξαγωγή των στατιστικών αποτελεσμάτων του αντίστοιχου μοντέλου. Υπενθυμίζεται ότι ο στόχος είναι οι προβλέψεις του κάθε μοντέλου που θα κατασκευαστεί να προσεγγίζουν όσο το δυνατόν καλύτερα τις τιμές της μεταβλητής εξόδου.

Όσον αφορά τις μεταβλητές εισόδου, δεδομένου ότι στην χρονοσειρά της απλής αμόλυβδης υπάρχει πολύς «θόρυβος» θεωρήθηκε εύλογο πέραν των χρονοσειρών της απλής αμόλυβδης και του αργού πετρελαίου να δοκιμαστούν και μεταβλητές εισόδου που σχετίζονται με συγκρίσεις κινητών μέσων όρων (εκθετικών ή απλών), στις οποίες δεν υπάρχει τόσος «θόρυβο». Θεωρείται ότι η χρήση των συγκεκριμένων μεταβλητών θα είναι καθοριστική στην αύξηση της ακρίβειας του μοντέλου πρόβλεψης. Οι συγκεκριμένες μεταβλητές εισόδου θα είναι διακριτές και θα έχουν πεδίο τιμών (0,1), ώστε να μην απαιτούν περαιτέρω ομαλοποίηση. Πριν οριστούν οι συγκεκριμένες μεταβλητές θα οριστούν κάποιοι συμβολισμοί που θα χρησιμοποιηθούν από τις συγκεκριμένες μεταβλητές. Ως $MA(p)$ θα ορίζεται ο κινητός μέσος όρος των τελευταίων p παρατηρήσεων και ως $EMA(p)$ θα ορίζεται ο κινητός εκθετικός μέσος όρος των τελευταίων p παρατηρήσεων. Οι τύποι τους παρουσιάζονται παρακάτω:

$$MA(p)_i = \sum_{j=i-p+1}^{i} Fuel_j \qquad (28)$$

$$EMA(p)_i = MA(p)_i\,, \qquad ισχύει\ για\ i = p \qquad (29)$$

$$EMA(p)_i = \frac{2}{(p+1)} * (Fuel_i - EMA(p)_{i-1}) + EMA(p)_{i-1} \quad , ισχύει\ για\ i > p \qquad (30)$$

Ο τρόπος κατασκευής των μεταβλητών εισόδου που βασίζονται σε συγκρίσεις των κινητών μέσων όρων (εκθετικών ή απλών) παρουσιάζεται παρακάτω:

$$tr1_i = \begin{cases} 1, & \alpha\nu\ MA(2)_i > MA(3)_i \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases} \quad (31)$$

$$tr2_i = \begin{cases} 1, & \alpha\nu\ MA(3)_i > MA(5)_i \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases} \quad (32)$$

$$tr3_i = \begin{cases} 1, & \alpha\nu\ MA(2)_i > MA(3)_i\ \kappa\alpha\iota\ Fuel_i > Fuel_{i-1} \\ 0{,}75, & \alpha\nu\ MA(2)_i > MA(3)_i\ \kappa\alpha\iota\ Fuel_i < Fuel_{i-1} \\ 0{,}25, & \alpha\nu\ MA(2)_i < MA(3)_i\ \kappa\alpha\iota\ Fuel_i > Fuel_{i-1} \\ 0, & \alpha\nu\ MA(2)_i < MA(3)_i\ \kappa\alpha\iota\ Fuel_i < Fuel_{i-1} \end{cases} \quad (33)$$

$$tr4_i = \begin{cases} 1, & \alpha\nu\ MA(3)_i > MA(5)_i\ \kappa\alpha\iota\ Fuel_i > Fuel_{i-1} \\ 0{,}75, & \alpha\nu\ MA(3)_i > MA(5)_i\ \kappa\alpha\iota\ Fuel_i < Fuel_{i-1} \\ 0{,}25, & \alpha\nu\ MA(3)_i < MA(5)_i\ \kappa\alpha\iota\ Fuel_i > Fuel_{i-1} \\ 0, & \alpha\nu\ MA(3)_i < MA(5)_i\ \kappa\alpha\iota\ Fuel_i < Fuel_{i-1} \end{cases} \quad (34)$$

$$tr5_i = \begin{cases} 1, & \alpha\nu\ EMA(2)_i > EMA(3)_i \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases} \quad (35)$$

$$tr6_i = \begin{cases} 1, & \alpha\nu\ EMA(3)_i > EMA(5)_i \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases} \quad (37)$$

$$tr7_i = \begin{cases} 1, & \alpha\nu\ MA(2)_i > MA(2)_{i-1}\ \kappa\alpha\iota\ \ MA(3)_i < MA(5)_i \\ 0{,}75, & \alpha\nu\ MA(2)_i > MA(2)_{i-1}\ \kappa\alpha\iota\ \ MA(3)_i > MA(5)_i \\ 0{,}25, & \alpha\nu\ MA(2)_i < MA(2)_{i-1}\ \kappa\alpha\iota\ \ MA(3)_i < MA(5)_i \\ 0, & \alpha\nu\ MA(2)_i < MA(2)_{i-1}\ \kappa\alpha\iota\ \ MA(3)_i > MA(5)_i \end{cases} \quad (38)$$

Στις μεταβλητές $tr1_i$, $tr2_i$, $tr5_i$, $tr6_i$ η σύγκριση των κινητών μέσων όρων (εκθετικών ή απλών) αποτελεί έναν τρόπο να εντοπιστούν ανοδικές ή καθοδικές τάσεις στην χρονοσειρά που διερευνάται και συνεπώς η εκάστοτε μέθοδος εκμάθησης μηχανών να μπορέσει να εκμεταλλευτεί αυτή την πληροφορία. Χρησιμοποιούνται κινητοί μέσοι όροι μικρής περιόδου, διότι ο στόχος είναι να προσεγγιστεί όσο το δυνατόν καλύτερα η χρονοσειρά της απλής αμόλυβδης. Είναι λογικό ότι όσο μεγαλώνει η περίοδος του κινητού μέσου όρου τόσο πιο ήπιες να γίνονται οι μεταβολές. Για το λόγο αυτό θεωρείται ότι οι μεταβλητές $tr1_i$και $tr5_i$ που συγκρίνουν τους μέσους όρους των τελευταίων δύο και τριών παρατηρήσεων θα προσδώσουν σημαντικές πληροφορίες για την βελτίωση της ακρίβειας των μοντέλων πρόβλεψης.

Οι μεταβλητές $tr3_i$ και $tr4_i$ αποτελούν ένα τρόπο να εντοπιστούν ισχυρά ανοδικές ή καθοδικές τάσεις στην χρονοσειρά. Η χρήση τους εισάγει καινούριες πληροφορίες στην εκάστοτε μέθοδο εκμάθησης μηχανών και αναλόγως τις «συμπεριφορές» της χρονοσειράς, η συνεισφορά τους στην βελτίωση της ακρίβειας πρόβλεψης μπορεί να είναι από σημαντική έως μηδαμινή.

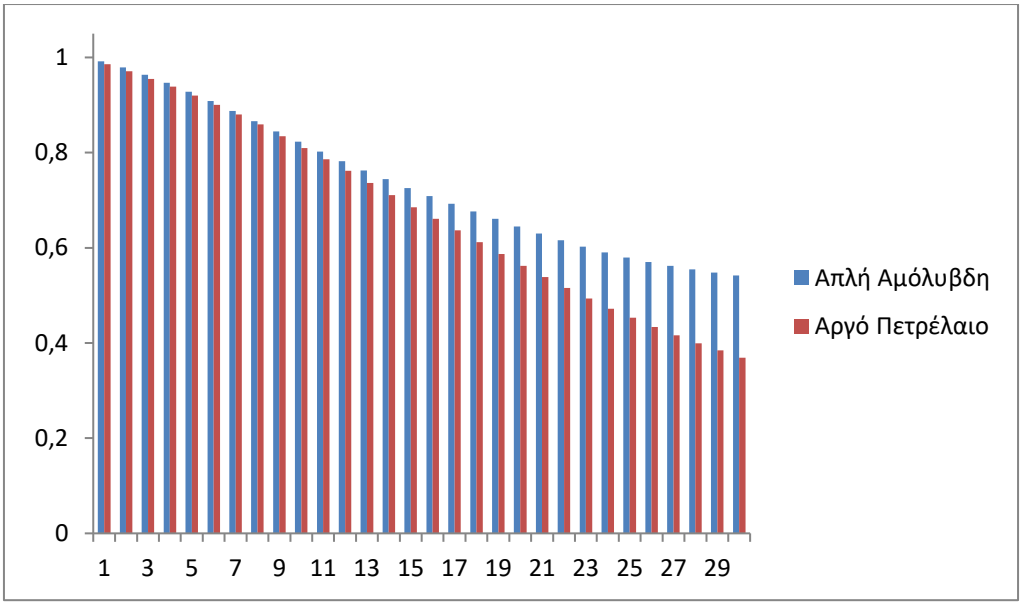
Η μεταβλητή $tr7_i$ θα χρησιμοποιηθεί για να εντοπίσει απότομα ανοδικές τάσεις έπειτα από περιόδους καθοδικών τάσεων και αντίστροφα. Τις περισσότερες φορές θα παίρνει τις τιμές 0,75 και 0,25, καθώς οι συνθήκες που απαιτούνται για να πάρει τις τιμές 0 ή 1 είναι σπάνιες. Ωστόσο, θεωρείται ότι οι πληροφορίες που προσφέρει στο εκάστοτε μοντέλο, όταν πάρει τις τιμές 0 ή 1, θα είναι ιδιαίτερα σημαντικές.

4.2 ΠΡΟΕΤΟΙΜΑΣΙΑ & ΟΜΑΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Προετοιμασία δεδομένων

Όπως αναφέρθηκε στο υποκεφάλαιο 4.1 παρατηρώντας τα διαγράμματα της απλής αμόλυβδης και του αργού πετρελαίου κατά την περίοδο 2005 – 2016 φαίνεται οπτικά ότι οι χρονοσειρές δεν έχουν σταθερό μέσο όρο και τυπική απόκλιση. Συνεπώς, πρόκειται για μη στάσιμες χρονοσειρές. Στο σημείο αυτό αξίζει να επισημανθεί ότι η χρήση στάσιμων χρονοσειρών συμβάλλει θετικά στην βελτίωση της ακρίβειας των προβλέψεων, όταν χρησιμοποιούνται μέθοδοι εκμάθησης μηχανών. Επιπρόσθετα, όταν εφαρμόζονται μοντέλα γραμμικής παλινδρόμησης ή το ολοκληρωμένο αυτοπαλινδρούμενο μοντέλο κινητού μέσου όρου (ARIMA) η χρήση στάσιμων χρονοσειρών είναι μείζονος σημασίας τόσο για την αξιοπιστία των προβλέψεων όσο και την ακρίβεια του μοντέλου.

Για την εξαγωγή του συμπεράσματος ότι οι χρονοσειρές που χρησιμοποιούνται είναι μη στάσιμες θα χρησιμοποιηθεί και η συνάρτηση της αυτοσυσχέτισης (ACF). Ο ερευνητής παρατηρώντας το διάγραμμα της αυτοσυσχέτισης συναρτήσει του αριθμού των υστερήσεων (lags), αν οι αυτοσυσχετίσεις μειώνονται αργά όσο αυξάνεται ο αριθμός των υστερήσεων (lags), αυτό αποτελεί ένδειξη της ύπαρξης μη στάσιμων χρονοσειρών. Η συγκεκριμένη συνάρτηση παρέχει και άλλες πληροφορίες, όπως τον αριθμό των προηγούμενων παρατηρήσεων που πρέπει να δοθούν στο μοντέλο ως μεταβλητές εισόδου. Κατά συνέπεια, αποτελεί έναν τρόπο να αντιληφθεί ο ερευνητής αν οι παρατηρήσεις της χρονοσειράς που χρησιμοποιεί έχουν μεγάλη μνήμη / «εμμονή», δηλαδή αν μια μελλοντική παρατήρηση θα έχει τιμή παρόμοια με τις προηγούμενες παρατηρήσεις. Στο διάγραμμα 4.3 παρουσιάζονται τα διαγράμματα της αυτοσυσχέτισης της απλής αμόλυβδης (μπλε χρώμα) και του αργού πετρελαίου (κόκκινο χρώμα).



Διάγραμμα 4.3: Διαγράμματα Αυτοσυσχέτισης (ACF) Απλής Αμόλυβδης και Αργού Πετρελαίου

Από το διάγραμμα 4.3 φαίνεται ότι και στις δυο χρονοσειρές οι αυτοσυσχετίσεις μειώνονται αργά όσο αυξάνεται ο αριθμός των υστερήσεων (lags). Έτσι, εξάγεται το συμπέρασμα ότι οι χρονοσειρές έχουν μεγάλη μνήμη. Το συγκεκριμένο συμπέρασμα εξάγεται για την πλειονότητα των χρονοσειρών που αφορούν οικονομικούς δείκτες. Γίνεται λοιπόν κατανοητό ότι απαιτείται η εφαρμογή των λογαριθμικών διαφορών στις υπάρχουσες χρονοσειρές για την επίτευξη στασιμότητας, δηλαδή: $r_t = \log\left(\frac{S_t}{S_{t-1}}\right)$,όπου S_t και S_{t-1} είναι οι παρατηρήσεις την χρονική στιγμή t και $t-1$ αντίστοιχα. Η επίτευξη στασιμότητας έχει ως συνέπεια η κάθε παρατήρηση στο σύνολο δεδομένων να είναι ανεξάρτητη από τις άλλες παρατηρήσεις.

Ομαλοποίηση δεδομένων

Ομαλοποίηση είναι η διαδικασία κατά την οποία πραγματοποιείται μετατροπή στις μεταβλητές εισόδου και εξόδου, ώστε να είναι σε κατάλληλη μορφή για να χρησιμοποιηθούν από τα μοντέλα πρόβλεψης. Η πλειοψηφία των ερευνητών υποστηρίζει ότι η ομαλοποίηση των δεδομένων εισόδου έχει θετική συνεισφορά για την βελτίωση της ακρίβειας των μοντέλων πρόβλεψης. Οι Shanker και Hung (1996) στην έρευνά τους υποστήριξαν ότι σε προβλήματα κατηγοριοποίησης που επιλύονται με μεθόδους εκμάθησης μηχανών η ομαλοποίηση προσδίδει μικρότερο σφάλμα πρόβλεψης. Οι ίδιοι αναφέρουν ότι αυτή έχει ακόμη πιο σημαντική επίδραση, όταν χρησιμοποιούνται μικρά σύνολα δεδομένων. Άλλοι ερευνητές υποστήριξαν ότι η μετατροπή αυτή είναι απαραίτητη για την αποφυγή υπολογιστικών προβλημάτων και την διευκόλυνση της διαδικασίας εκμάθησης. Ένα σημαντικό πλεονέκτημα είναι ότι διασφαλίζει να μην υπάρχουν μεταβλητές εισόδου που υπερτερούν έναντι άλλων, επειδή έχουν μεγαλύτερες τιμές. Το γεγονός αυτό περιορίζει σημαντικά το σφάλμα πρόβλεψης.

Ακόμη, σε περιπτώσεις που εφαρμόζονται μέθοδοι εκμάθησης μηχανών που βασίζονται σε συναρτήσεις ενεργοποίησης με περιορισμένο πεδίο τιμών η ομαλοποίηση των μεταβλητών εισόδου είναι απαραίτητη. Για παράδειγμα, στην μέθοδο των νευρωνικών δικτύων, όταν χρησιμοποιείται ως συνάρτηση ενεργοποίησης η σιγμοειδής, τα αποτελέσματα που εξάγει το δίκτυο έχουν πεδίο τιμών (0,1). Κατά συνέπεια, οι μεταβλητές εισόδου και εξόδου δεν πρέπει να έχουν τις μονάδες μέτρησής τους, δηλαδή πρέπει να ομαλοποιούνται. Ο ερευνητής οφείλει να ομαλοποιεί τις μεταβλητές εισόδου και εξόδου, ώστε να έχουν εύρος τιμών (0,1). Αντίστοιχα, αν χρησιμοποιείται ως συνάρτηση ενεργοποίησης η υπερβολική εφαπτομένη, η οποία έχει πεδίο τιμών (-1,1), ο ερευνητής οφείλει να μετατρέψει τις μεταβλητές εισόδου και εξόδου κατάλληλα, ώστε να έχουν και εκείνες το ίδιο πεδίο τιμών.

Υπάρχουν ερευνητές, οι οποίοι θεωρούν ότι η ομαλοποίηση επαρκεί να εφαρμόζεται μόνο στην μεταβλητή εξόδου. Ωστόσο, σύμφωνα με τους Kondratenko και Kuperin (2003) η ομαλοποίηση και των μεταβλητών εισόδου είναι απαραίτητη, καθώς συμβάλλει στην διαδικασία εκμάθησης από την εκάστοτε μέθοδο εκμάθησης μηχανών. Η συνεισφορά της οφείλεται στο γεγονός ότι ομαλοποιεί την κατανομή των δεδομένων εντός του πεδίου τιμών και αυξάνει την εντροπία τους, με αποτέλεσμα η κατανομή να προσεγγίζει τα χαρακτηριστικά κανονικής κατανομής. Στο πρόβλημα που θα επιλυθεί στην παρούσα διπλωματική εργασία, επειδή είναι τύπου 0/1 απαιτείται οι μεταβλητές εισόδου και εξόδου να έχουν πεδίο τιμών (0,1). Έτσι, οι τρόποι ομαλοποίησης είναι οι παρακάτω:

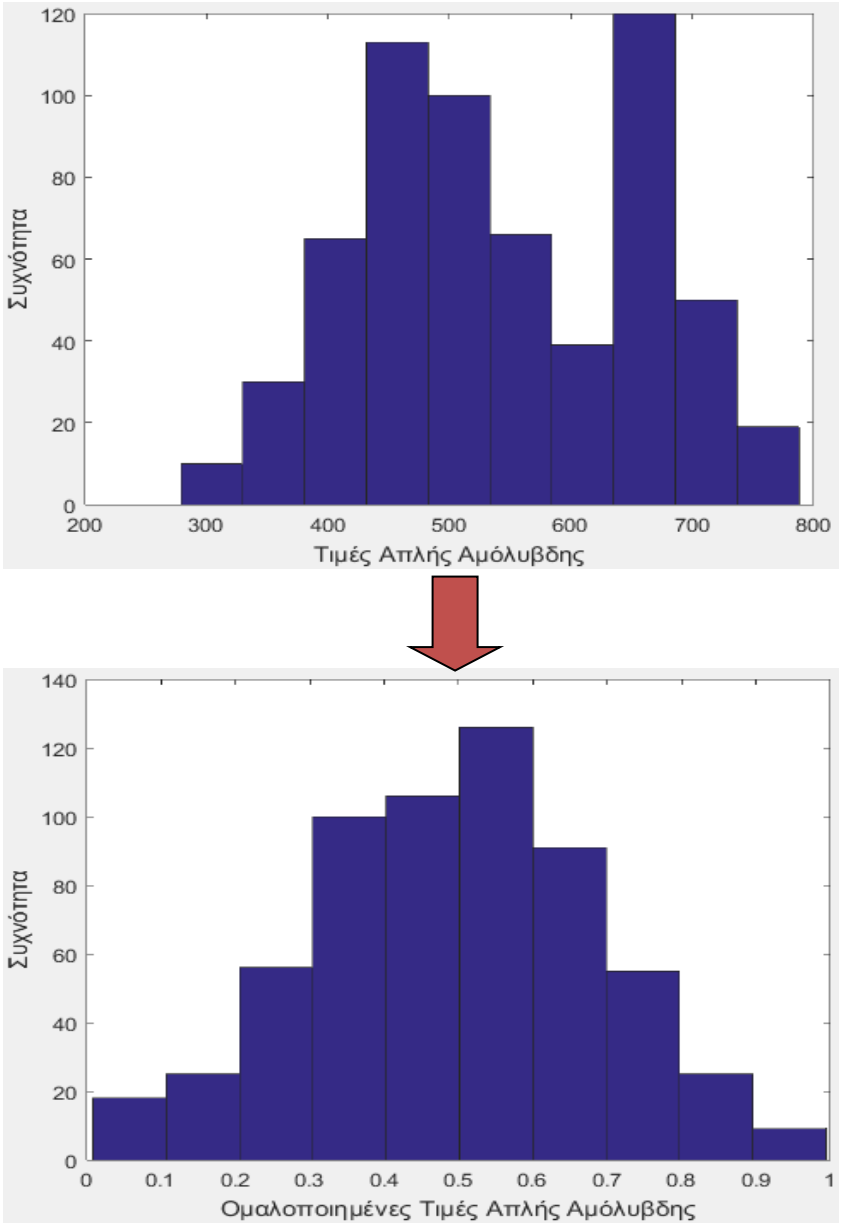
$$X_\iota = \frac{r_i - r_{min}}{r_{max} - r_{min}} \qquad (39)$$

,όπου r_{max} και r_{min} είναι οι ελάχιστες και οι μέγιστες τιμές των r_i στο σύνολο εκπαίδευση

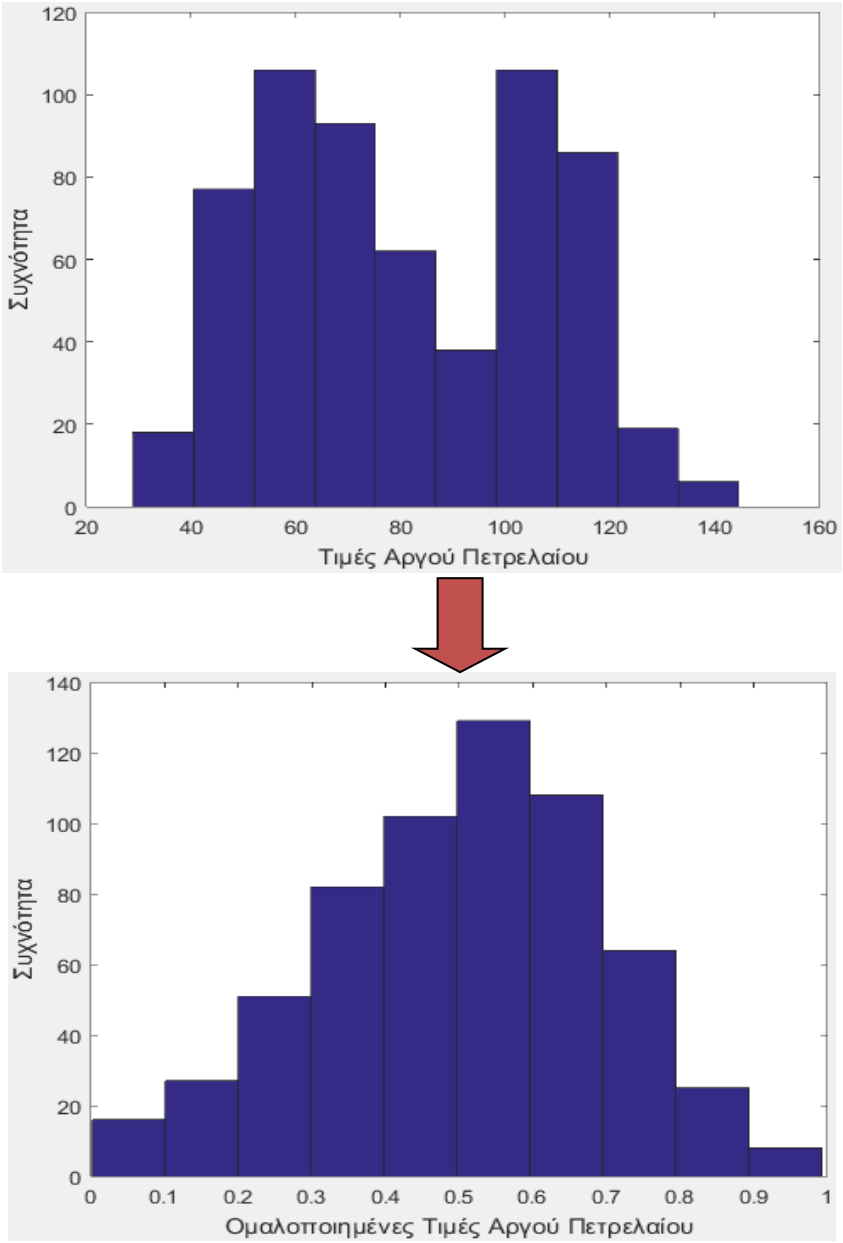
$$X_i = \frac{1}{1 + e^{-(\frac{r_i - r_m}{std})}} \qquad (40)$$

,όπου r_m και std είναι η μέση τιμή και τυπική απόκλιση των r_i του συνόλου εκπαίδευσης

Οι δυο προαναφερθέντες τρόποι ομαλοποίησης μετατρέπουν το σύνολο δεδομένων, ώστε να έχουν πεδίο τιμών (0,1), ωστόσο ο δεύτερος τρόπος επιτυγχάνει καλύτερη κατανομή των δεδομένων εντός του πεδίου τιμών και κατά συνέπεια συμβάλλει περισσότερο στην διαδικασία εκμάθησης. Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθεί ο δεύτερος τρόπος ομαλοποίησης. Με την βοήθεια της γλώσσας προγραμματισμού Matlab κατασκευάστηκαν τα διαγράμματα 4.4 και 4.5, όπου φαίνονται τα ιστογράμματα των δεδομένων πριν και μετά την εφαρμογή της ομαλοποίησης.



Διάγραμμα 4.4: Ιστόγραμμα Χρονοσειράς Απλής Αμόλυβδης πριν και μετά την εφαρμογή της ομαλοποίησης



Διάγραμμα 4.5: Ιστόγραμμα Χρονοσειράς Αργού Πετρελαίου πριν και μετά την εφαρμογή της ομαλοποίησης

Έτσι, ολοκληρώνοντας και αυτές τις ομαλοποιήσεις δημιουργείται η βάση δεδομένων, η οποία θα περιέχει μια στήλη που αντιστοιχεί στις παρατηρήσεις του αργού πετρελαίου της προηγούμενης εβδομάδας ($brent_{t-1}$), 15 στήλες που αντιστοιχούν στις παρατηρήσεις της τιμής της αμόλυβδης τις τελευταίες 15 εβδομάδες ($Fuel_{t-i}$, όπου i=1,….,15), τις επτά μεταβλητές τάσεων trj, (όπου j=1,….,7) και την μεταβλητή εξόδου out_t. Συνεπώς, η βάση δεδομένων αποτελείται από 23 μεταβλητές εισόδου και 1 μεταβλητή εξόδου, οι οποίες θα δοθούν στην εκάστοτε μέθοδο εκμάθησης μηχανών προκειμένου να εξαχθούν τα αποτελέσματα.

5. ΕΚΠΑΙΔΕΥΣΗ & ΑΞΙΟΛΟΓΗΣΗ ΜΕΘΟΔΩΝ ΕΚΜΑΘΗΣΗΣ ΜΗΧΑΝΩΝ

Ο τρόπος εκπαίδευσης της εκάστοτε μεθόδου εκμάθησης μηχανών και τα στατιστικά μέτρα αξιολόγησης έχουν εξηγηθεί αναλυτικά στο κεφάλαιο 3. Στο κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα της έρευνας της παρούσας διπλωματικής εργασίας, η αξιολόγηση τους και ο σχολιασμός τους. Το κεφάλαιο 5 έχει δομηθεί με τέτοιο τρόπο, ώστε στο υποκεφάλαιο 5.1 να γίνει εκπαίδευση και αξιολόγηση ενός νευρωνικού δικτύου, στο υποκεφάλαιο 5.2 να γίνει εκπαίδευση και αξιολόγηση ενός δένδρου απόφασης και στο υποκεφάλαιο 5.3 να γίνει εκπαίδευση και αξιολόγηση της μεθόδου AdaboostM1. Σε κάθε μέθοδο θα παρουσιάζονται τα δεδομένα εισόδου που χρησιμοποιήθηκαν, η βέλτιστη δομή και παράμετροι καθώς και ο σχολιασμός τους. Σημειώνεται ότι στην παρούσα διπλωματική εργασία παρ΄ότι έγιναν πολλές δοκιμές για την επίτευξη αυτών των ακριβειών στην βέλτιστη λύση του προβλήματος, δεν θα παρατεθούν όλες αυτές οι δοκιμές, αλλά μόνο η τελική λύση και η συλλογιστική πορεία που οδήγησε στην εύρεση αυτών των λύσεων. Επίσης, θα γίνονται και σχολιασμοί για το πώς δημιουργήθηκε η εκάστοτε βέλτιστη λύση και ο σχολιασμός των συμπεριφορών κάθε μεθόδου.

5.1 ΕΚΠΑΙΔΕΥΣΗ & ΑΞΙΟΛΟΓΗΣΗ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ

5.1.1 ΔΕΔΟΜΕΝΑ

Όπως αναφέρθηκε η αρχική βάση δεδομένων αποτελείται από 23 μεταβλητές εισόδου και 1 μεταβλητή εξόδου. Κάθε μεταβλητή αποτελείται από 597 παρατηρήσεις και αναφέρεται στην περίοδο 2005 – 2016. Από την συγκεκριμένη βάση δεδομένων κάποιες μεταβλητές δεν θα χρησιμοποιηθούν από το δίκτυο, είτε επειδή η συμβολή τους είναι μηδαμινή και η αξιοποίησή τους απλώς αυξάνει το υπολογιστικό κόστος είτε επειδή μειώνουν την ακρίβεια του μοντέλου. Από το σύνολο των 597 παρατηρήσεων κάθε μεταβλητής αποφασίστηκε να χρησιμοποιηθούν οι 447 ως σύνολο εκπαίδευσης (75%), 59 ως σύνολο επικύρωσης (10%) και 86 ως σύνολο δοκιμής (15%).

5.1.2 ΕΚΠΑΙΔΕΥΣΗ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ & ΕΠΙΛΟΓΗ ΠΑΡΑΜΕΤΡΩΝ

Δομή

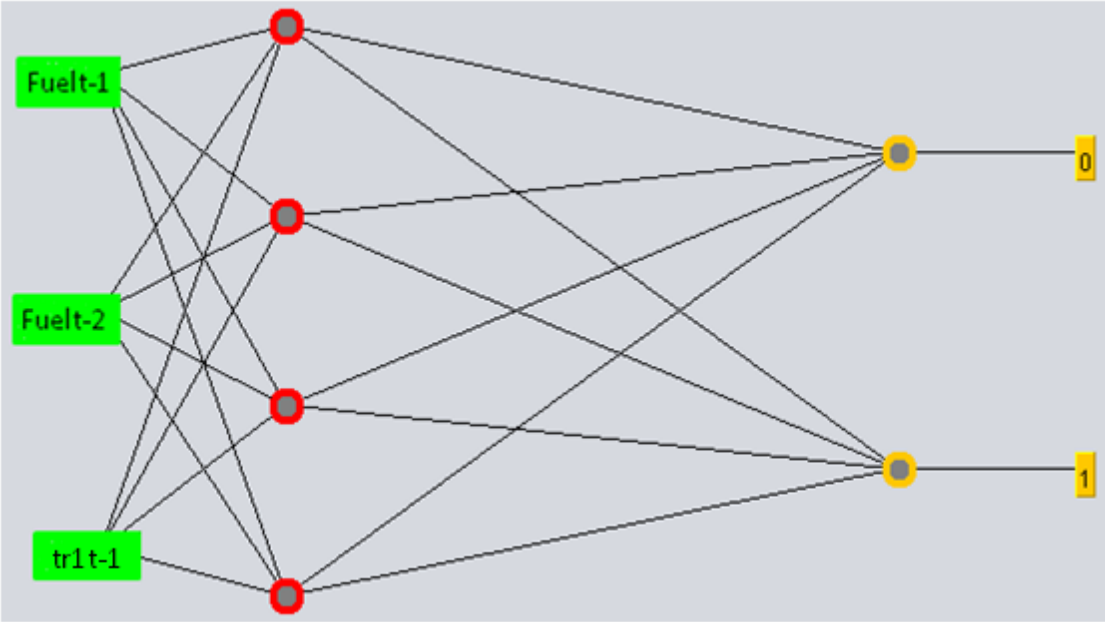
Η δομή του νευρωνικού δικτύου , δηλαδή η κατάλληλη επιλογή των μεταβλητών εισόδου και ο κατάλληλος αριθμός νευρώνων στο κρυμμένο επίπεδο είναι καθοριστικής σημασίας για την ακρίβεια του μοντέλου και την αξιοπιστία πρόβλεψης. Αρχικά, με την μέθοδο δοκιμής – σφάλματος έγιναν προσπάθειες για την εύρεση του βέλτιστου αριθμού των προηγούμενων παρατηρήσεων της απλή αμόλυβδης που πρέπει να χρησιμοποιηθούν από το μοντέλου πρόβλεψης. Υπενθυμίζεται από το διάγραμμα της αυτοσυσχέτισης (ACF) για την χρονοσειρά της απλής αμόλυβδης εξήχθη το συμπέρασμα ότι εμφανίζει μεγάλη μνήμη / οι παρατηρήσεις έχουν «εμμονή». Έτσι, κυρίως για λόγους υπολογιστικού κόστους δόθηκαν στο δίκτυο οι 15 προηγούμενες εβδομαδιαίες παρατηρήσεις της απλής αμόλυβδης. Διερευνήθηκε η ακρίβεια πρόβλεψης του νευρωνικού δικτύου χρησιμοποιώντας από την βάση δεδομένων μόνο αυτές τις 15 μεταβλητές ως μεταβλητές εισόδου. Έχοντας αυτές τις μεταβλητές εισόδου και αλλάζοντας τον αριθμό των νευρώνων από 1 έως 20 με βήμα 1 υπολογίστηκε η ακρίβεια του εκάστοτε μοντέλου. Η καλύτερη ακρίβεια αποθηκεύτηκε μαζί με τον αριθμό των νευρώνων στο κρυμμένο επίπεδο. Έπειτα, αυτή η διαδικασία έγινε χρησιμοποιώντας μια λιγότερη μεταβλητή εισόδου, δηλαδή τις 14 προηγούμενες εβδομαδιαίες παρατηρήσεις ως μεταβλητές εισόδου. Η διαδικασία αυτή ακολουθήθηκε

μέχρι την χρησιμοποίηση ως μεταβλητής εισόδου μόνο την προηγούμενη παρατήρηση, όπου έγιναν δοκιμές με 1 έως 5 νευρώνες στο κρυμμένο επίπεδο. Το αποτέλεσμα ήταν να επιτευχθεί η μεγαλύτερη ακρίβεια πρόβλεψης, όταν χρησιμοποιήθηκαν ως μεταβλητές εισόδου οι δυο προηγούμενες εβδομαδιαίες παρατηρήσεις, δηλαδή $Fuel_{t-1}$, $Fuel_{t-2}$. Η ακρίβεια που επιτεύχθηκε ήταν της τάξεως του 67% -68% στο σύνολο δοκιμής.

Έτσι, κάθε νέα δομή δικτύου που κατασκευαζόταν είχε ως μεταβλητές εισόδου τουλάχιστον τις μεταβλητές $Fuel_{t-1}$, $Fuel_{t-2}$. Η μέθοδος δοκιμής – σφάλματος ακολουθήθηκε και για την εύρεση των απαιτούμενων προηγούμενων παρατηρήσεων του αργού πετρελαίου. Η διερεύνηση έδειξε ότι ακόμη και η συνεισφορά της προηγούμενης παρατήρησης στην βελτίωση της ακρίβειας πρόβλεψης ήταν μηδαμινή. Πιο συγκεκριμένα, φαίνεται από τα αποτελέσματα ότι δίνοντας στο δίκτυο και την προηγούμενη παρατήρηση του αργού πετρελαίου ως μεταβλητή εισόδου βελτιώνεται λίγο η ακρίβεια πρόβλεψης των ανοδικών τάσεων της απλής αμόλυβδης. Αντίθετα, η ακρίβεια πρόβλεψης των καθοδικών τάσεων της απλής αμόλυβδης μειώνεται, με αποτέλεσμα η συνεισφορά της μεταβλητής να είναι μηδαμινή. Αξίζει να σημειωθεί ότι τα αποτελέσματα αυτά επιβεβαιώνουν τα συμπεράσματα της υπάρχουσας βιβλιογραφίας που αναφέρουν τις ασυμμετρίες μεταξύ αργού πετρελαίου και λιανικής τιμής των καυσίμων, δηλαδή το γεγονός ότι σε άνοδο της τιμής του αργού πετρελαίου αυξάνεται και η τιμή της απλής αμόλυβδης άμεσα, ενώ σε μείωση της τιμής του αργού πετρελαίου η τιμή της απλής αμόλυβδης δεν ακολουθεί παρόμοια πορεία. Έτσι, αποφασίστηκε να μην ληφθεί υπόψιν η χρονοσειρά του αργού πετρελαίου για την πρόβλεψη της κατεύθυνσης της τιμής της απλής αμόλυβδης.

Για την επιλογή των μεταβλητών που σχετίζονται με συγκρίσεις κινητών μέσων όρων (απλών ή εκθετικών) που εφαρμόζονται στην χρονοσειρά της απλής αμόλυβδης ελέγχθηκε κάθε μια μεταβλητή ξεχωριστά μαζί με τις μεταβλητές εισόδου $Fuel_{t-1}$, $Fuel_{t-2}$. Για κάθε μια μεταβλητή έγιναν έρευνες χρησιμοποιώντας από 1 έως 6 νευρώνες στο κρυμμένο επίπεδο και αποθηκεύτηκε η μεγαλύτερη ακρίβεια πρόβλεψης στο σύνολο επικύρωσης και ο αριθμός των νευρώνων στο κρυμμένο επίπεδο. Από τα αποτελέσματα προέκυψε ότι οι μεταβλητές $tr2_t$, $tr3_t$, $tr4_t$, $tr5_t$, $tr6_t$ όχι απλώς δεν συνεισφέρουν, αλλά δυσχεραίνουν λίγο την ακρίβεια πρόβλεψης στο σύνολο επικύρωσης. Η μεταβλητή $tr7_t$ είχε μηδαμινή συνεισφορά κάτι που υποδεικνύει ότι η χρονοσειρά της απλής αμόλυβδης δεν εμφανίζει μεγάλες και απότομες μεταβολές έπειτα από περίοδο μεγάλης ανοδικής ή καθοδικής πορείας. Αντίθετα, η συνεισφορά της μεταβλητής $tr1_t$ στην αύξηση της ακρίβειας πρόβλεψης του μοντέλου είναι καθοριστική. Υπενθυμίζεται ότι η μεταβλητή $tr1_t$ συγκρίνει τον κινητό μέσο όρο των δυο εβδομάδων με αυτό των τριών εβδομάδων. Η συνεισφορά της έγκειται στο γεγονός ότι βελτιώνει τις προβλέψεις των καθοδικών τάσεων αλλά ακόμη περισσότερο βελτιώνει τις προβλέψεις των ανοδικών τάσεων.

Ολοκληρώνοντας όλες τις δοκιμές - σφάλματος αποφασίστηκε η χρήση τριών μεταβλητών εισόδου, τις $tr1_t$, $Fuel_{t-1}$, $Fuel_{t-2}$. Παρακάτω στην εικόνα 5.1 φαίνεται η δομή του νευρωνικού δικτύου έχοντας τέσσερις νευρώνες στο κρυμμένο επίπεδο.



Εικόνα 5.1: Δομή Νευρωνικού Δικτύου του υπό Επίλυση Προβλήματος

Εκπαίδευση και Επιλογή Παραμέτρων

Η επιλογή των νευρώνων στο κρυμμένο επίπεδο έγινε με την μέθοδο της δοκιμής - σφάλματος και ο στόχος ήταν να μεγιστοποιείται η ακρίβεια πρόβλεψης στο σύνολο επικύρωσης. Δοκιμάστηκαν από ένας έως δέκα νευρώνες στο κρυμμένο επίπεδο, ωστόσο ο αριθμός των τεσσάρων νευρώνων στο κρυμμένο επίπεδο που παρατίθεται και στην εικόνα 5.1, προέκυψε ο βέλτιστος. Όσο αυξανόταν ο συγκεκριμένος αριθμός βελτιωνόταν η ακρίβεια της πρόβλεψης των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης, ενώ μειωνόταν δυσανάλογα η ακρίβεια της πρόβλεψης των καθοδικών κατευθύνσεων της απλής αμόλυβδης, με αποτέλεσμα την πτώση της σταθμισμένης ακρίβειας του δικτύου. Επιπλέον, από το σύνολο επικύρωσης διαπιστώθηκε αν απαιτείται η χρήση συντελεστή αποδόμησης (decay rate). Ο συντελεστής αποδόμησης (decay rate) είναι ίσος με το πηλίκο της διαίρεσης του ρυθμού εκμάθησης δια τον αριθμό των παρατηρήσεων στο σύνολο εκπαίδευσης. Στη συγκεκριμένη διερεύνηση ο αριθμός αυτός ισούται με 6,7*10^(-4). Ελέγχοντας τα αποτελέσματα του συνόλου επικύρωσης προέκυψε ότι εφαρμόζοντας την παραπάνω τιμή για τον συντελεστή αποδόμησης (decay rate) μειώνεται η ακρίβεια πρόβλεψης του συνόλου επικύρωσης. Κατά συνέπεια, αποφασίστηκε να τεθεί η τιμή του συντελεστή αποδόμησης (decay rate) ίση με 0, όταν θα γίνει η τελική εκπαίδευση των νευρώνων στο σύνολο εκπαίδευσης και ο υπολογισμός των στατιστικών μέτρων στο σύνολο δοκιμής. Στο σημείο αυτό αξίζει να εξηγηθεί ο βαθύτερος λόγος που στο συγκεκριμένο πρόβλημα ο συντελεστής αποδόμησης (decay rate) μειώνει την ακρίβεια του μοντέλου πρόβλεψης. Η χρήση του παραπάνω συντελεστή έχει ως συνέπεια τα βάρη να εκπαιδεύονται αρκετά από τις πρώτες τιμές αλλά μετά ο ρυθμός εκπαίδευσής τους να μειώνεται σημαντικά. Έτσι, στο πρόβλημα που επιλύει η παρούσα διπλωματική εργασία

φαίνεται ότι η χρήση του συντελεστή αποδόμησης δεν αφήνει τα βάρη να εκπαιδευτούν μέχρι το επιθυμητό σημείο και η εκπαίδευσή τους σταματά λίγο νωρίτερα.

Όσον αφορά τον ρυθμό εκμάθησης και την ορμή έγιναν κάποιες δοκιμές με βήματα 0,1 και τελικώς, όπως φαίνεται και στον πίνακα 5.1, αποφασίστηκε να χρησιμοποιηθούν οι τιμές 0,3 και 0,2 αντίστοιχα. Για τις «εποχές» αποφασίστηκε να χρησιμοποιηθούν 500 επαναλήψεις. Όλες οι τιμές των παραμέτρων του τελικού νευρωνικού δικτύου παρατίθενται στον πίνακα 5.1.

Πίνακας 5.1: Δεδομένα & Παράμετροι του υπό Επίλυση Προβλήματος με την μέθοδο των Νευρωνικών Δικτύων

Δεδομένα		**Παράμετροι**	
Αριθμός Μεταβλητών Εισόδου	3	Εποχές	500
Αριθμός Μεταβλητών Εξόδου	2	Ρυθμός Εκμάθησης	0,3
Παρατηρήσεις Συνόλου Εκπαίδευσης	447	Ορμή	0,2
Παρατηρήσεις Συνόλου Επικύρωσης	59	Συντελεστής αποδόμησης	0
Παρατηρήσεις Συνόλου Δοκιμής	86	Κρυμμένοι Νευρώνες	4

5.1.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ

Αξιοποιώντας τις παραμέτρους του πίνακα 5.1 πραγματοποιήθηκε η εκπαίδευση των νευρώνων στο σύνολο εκπαίδευσης και υπολογίστηκαν τα στατιστικά μέτρα αξιολόγησης στο σύνολο δοκιμής, ώστε να εξαχθούν συμπεράσματα για την ακρίβεια και την αξιοπιστία των προβλέψεων του δικτύου.

Πίνακας 5.2: Πίνακας Προβλέψεων / Πραγματικών Μέτρων & Μέτρα Αξιολόγησης του υπό Επίλυση Προβλήματος με την μέθοδο των Νευρωνικών Δικτύων

		Πρόβλεψη	**Πρόβλεψη**
		1	0
Πραγματικά Δεδομένα	1	32	14
Πραγματικά Δεδομένα	0	8	32
<u>Ακρίβεια</u>	0,744		
Ανοδική Τάση (1)		Καθοδική Τάση (0)	
ROC	0,76	ROC	0,76
F-Measure	0,74	F-Measure	0,74
Precision	0,70	Precision	0,80
Recall	0,80	Recall	0,70
Sensitivity	0,80	Sensitivity	0,70
Specificity	0,30	Specificity	0,20

Από τον πίνακα 5.2 φαίνεται ότι η ακρίβεια (accuracy) του μοντέλου πρόβλεψης της κατεύθυνσης της τιμής της απλής αμόλυβδης είναι ίση με 74,4%. Επίσης, ο δείκτης ROC για την κατηγοριοποίηση τόσο της ανοδικής τάσης όσο και της καθοδικής τάσης της αμόλυβδης ισούται με 0,76 που αποτελεί μια πολύ καλή τιμή για την αξιοπιστία του μοντέλου. Υπενθυμίζεται ότι η καμπύλη ROC αναπαριστά τον δείκτη sensitivity συναρτήσει του 1 – specificity και αυτό που ενδιαφέρει τον ερευνητή είναι το εμβαδόν κάτω από την καμπύλη. Συνεπώς, αξιοποιώντας κανείς αυτό το μοντέλο πρόβλεψης μπορεί να παράγει με αρκετά μεγάλη ακρίβεια προβλέψεις για την τάση της τιμής της απλής αμόλυβδης με χρονικό ορίζοντα μιας εβδομάδας.

Επιπρόσθετα, αξίζει να σημειωθεί ότι ο δείκτης True Positive Rate για την πρόβλεψη ανοδικής κατεύθυνσης της τιμής της απλής αμόλυβδης (1) είναι μεγαλύτερος από ότι ο αντίστοιχος δείκτης για την πρόβλεψη της καθοδικής κατεύθυνσης της τιμής (0). Για τον λόγο αυτό η σταθμισμένη ακρίβεια (accuracy) είναι 74,4% αντί για 80% που είναι η ακρίβεια για την πρόβλεψη ανοδικής κατεύθυνσης της τιμής της απλής αμόλυβδης (1). Έτσι, μπορεί να εξαχθεί το συμπέρασμα ότι το μοντέλο πρόβλεψης προβλέπει λίγο καλύτερα ανοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης από ότι καθοδικές. Σε αυτό το σημείο αξίζει να επισημανθεί ότι κατά την διάρκεια των δοκιμών διαφορών δομών δικτύων παρατηρήθηκε το εξής, όσο αυξανόταν ο αριθμός των νευρώνων στο κρυμμένο επίπεδο

τόσο βελτιωνόταν η ακρίβεια της κατηγοριοποίησης των ανοδικών κατευθύνσεων (1) και ταυτόχρονα μειωνόταν η ακρίβεια της κατηγοριοποίησης των καθοδικών κατευθύνσεων (0) σε επίπεδα 55% - 57%. Από την θεωρία του κεφαλαίου 3 υπενθυμίζεται ότι αυξάνοντας τον αριθμό των νευρώνων στο κρυμμένο επίπεδο αυξάνεται η πολυπλοκότητα του δικτύου, δηλαδή αυξάνεται η μη γραμμικότητά του. Έτσι, συμπεραίνεται ότι οι ανοδικές κατευθύνσεις της τιμής της αμόλυβδης εμφανίζουν έντονες μη γραμμικότητες, οι οποίες μπορούν να μοντελοποιηθούν από την πολυπλοκότητα της δομής των νευρωνικών δικτύων. Σε αντίθεση, η αυξημένη πολυπλοκότητα των νευρωνικών δικτύων δεν μπορεί να συμβάλλει στην αύξηση της ακρίβειας πρόβλεψης των καθοδικών κατευθύνσεων της τιμής της αμόλυβδης.

Οι πιθανές αιτίες που η ακρίβεια πρόβλεψης των καθοδικών τάσεων της τιμής της αμόλυβδης είναι μικρότερη από την ακρίβεια πρόβλεψης ανοδικών τάσεων παρουσιάζονται παρακάτω. Αξίζει να υπενθυμιστεί ότι οι νευρώνες εκπαιδεύονται από ένα σύνολο εκπαίδευσης και μαθαίνουν να γενικεύουν με την επιλογή ορθών παραμέτρων από ένα σύνολο επικύρωσης, με αποτέλεσμα αν οι συμπεριφορές των παρατηρήσεων είναι πολύ διαφορετικές στο σύνολο δοκιμής τότε να μην μπορούν να παράγουν πολλές ορθές προβλέψεις. Κάτι τέτοιο μπορεί να συμβαίνει και στην περίπτωση των παρατηρήσεων που έχουν καθοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης. Κατά συνέπεια, συμπεραίνεται ότι οι καθοδικές κατευθύνσεις έχουν αυξημένη στοχαστική συμπεριφορά (τυχαίος περίπατος) που δεν επιτρέπει στο εκάστοτε δίκτυο να παράγει ορθές προβλέψεις με ακρίβεια της τάξεως του 80% - 85% και πάνω.

5.2 ΚΑΤΑΣΚΕΥΗ & ΑΞΙΟΛΟΓΗΣΗ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ

5.2.1 ΔΕΔΟΜΕΝΑ

Όπως αναφέρθηκε στο υποκεφάλαιο 4.2 η βάση δεδομένων αποτελείται από 23 μεταβλητές εισόδου και μια μεταβλητή εξόδου. Κάθε μεταβλητή είτε εισόδου είτε εξόδου αποτελείται από 597 παρατηρήσεις και αναφέρεται στην περίοδο 2005 – 2016. Από την συγκεκριμένη βάση δεδομένων κάποιες μεταβλητές εισόδου δεν θα χρησιμοποιηθούν από το δένδρο απόφασης, είτε επειδή η αφαίρεσή τους κατά την διαδικασία «κλαδέματος» δεν επιδρά αρνητικά στα στατιστικά μέτρα αξιολόγησης είτε επειδή η χρησιμοποίησή τους μειώνει την ακρίβεια πρόβλεψης του δένδρου απόφασης. Από το σύνολο των 597 παρατηρήσεων κάθε μεταβλητής αποφασίστηκε να χρησιμοποιηθούν οι 447 ως σύνολο εκπαίδευσης (75%), 59 ως σύνολο επικύρωσης (10%) και 86 ως σύνολο δοκιμής (15%). Υπενθυμίζεται ότι η συγκεκριμένη αναλογία επιλέχθηκε και κατά την κατασκευή του νευρωνικού δικτύου. Θεωρήθηκε απαραίτητο σε κάθε μέθοδο εκμάθησης μηχανών να τηρηθεί η ίδια αναλογία, προκειμένου να μπορεί να γίνει σύγκριση μεταξύ τους και να εξαχθούν συμπεράσματα σχετικά με τα πλεονεκτήματα και τα μειονεκτήματα της εκάστοτε μεθόδου εκμάθησης μηχανών.

5.2.2 ΚΑΤΑΣΚΕΥΗ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ & ΕΠΙΛΟΓΗ ΠΑΡΑΜΕΤΡΩΝ

Κατασκευή Δένδρου Απόφασης

Μια διαφορά μεταξύ των νευρωνικών δικτύων και των δένδρων απόφασης είναι ότι στην πρώτη μέθοδο ο χρήστης επιλέγει τις μεταβλητές εισόδου, ενώ η διαδικασία στην δεύτερη μέθοδο είναι διαφορετική. Κατά την διάρκεια εκμάθησης, όπως κατασκευάζεται το δένδρο απόφασης ο αλγόριθμος επιλέγει από μόνος του σε κάθε κόμβο την μεταβλητή εκείνη που διαχωρίζει καλύτερα τις αντίστοιχες παρατηρήσεις. Πιο συγκεκριμένα, ξεκινάει από την

«ρίζα» του δένδρου που έχει να διαχωρίσει όλες τις παρατηρήσεις και βρίσκει εκείνη την μεταβλητή που εξασφαλίζει το μικρότερο ποσοστό σφάλματος, καθώς και την τιμή της μεταβλητής που θα διαχωρίζει τα παραδείγματα. Έπειτα εφαρμόζοντας τις μεθόδους «κλαδέματος» ελέγχει από κάτω προς τα πάνω αν η αφαίρεση των εκάστοτε μεταβλητών επηρεάζει την ακρίβεια πρόβλεψης και αφαιρεί όσες είναι περιττές. Αυτό έχει ως αποτέλεσμα να υπάρχει η πιθανότητα κάποιες μεταβλητές εισόδου να μην χρησιμοποιούνται. Έτσι, μια προσπάθεια να αξιοποιηθεί το πλεονέκτημα αυτής της διαδικασίας εκμάθησης, αρχικά δόθηκαν ως μεταβλητές εισόδου και οι 23 μεταβλητές της βάσης δεδομένων. Ωστόσο, κάποιες φορές συμβάλλει θετικά στην διαδικασία εκμάθησης να αφαιρούνται μεταβλητές που δεν χρησιμοποιούνται από το δένδρο απόφασης.

Για την εύρεση των κατάλληλων μεταβλητών εισόδου που πρέπει να δοθούν στον αλγόριθμο, ώστε να παράγει αξιόπιστες και ακριβείς προβλέψεις αλλά και να μην δημιουργείται περιττό υπολογιστικό κόστος χρησιμοποιήθηκε η μέθοδος της δοκιμής - σφάλματος. Για τις εκάστοτε μεταβλητές εισόδου δοκιμάστηκαν διάφοροι συντελεστές εμπιστοσύνης 0,05 έως 0,25 με βήμα 0,05 και ελάχιστος αριθμός στοιχείων σε κάθε κόμβο από 2 έως 20 και υπολογίστηκε η ακρίβεια πρόβλεψης στο σύνολο επικύρωσης. Επίσης, μια σημαντική παρατήρηση σε αυτό το σημείο είναι εκτός από τα αποτελέσματα στο σύνολο επικύρωσης ελέγχθηκαν και το μέγεθος του εκάστοτε δένδρου απόφασης, δηλαδή αν είναι εκτενές (υπερπροσαρμογή) ή περιορισμένο (υποπροσαρμογή), και οι μεταβλητές που αξιοποιεί το εκάστοτε δένδρο απόφασης για να κατηγοριοποιήσει τις παρατηρήσεις. Ο έλεγχος των μεταβλητών που εμφανίζονται στο δένδρο απόφασης είναι μια πολύ σημαντική ενέργεια, καθώς ο ερευνητής οφείλει πρώτα και κύρια να ελέγχει αν υπάρχει συνοχή στο δένδρο απόφασης. Για παράδειγμα, αν δημιουργηθεί ένα δένδρο απόφασης που εμφανίζει ως κατηγοριοποιητή την μεταβολή της απλής αμόλυβδης την προηγούμενη εβδομάδα ($Fuel_{t-1}$) και την μεταβολή της απλής αμόλυβδης πριν από 15 εβδομάδες ($Fuel_{t-15}$), χωρίς να εμφανίζει και μερικές μεταβλητές που αναφέρονται στις μεταβολές της απλής αμόλυβδης τις ενδιάμεσες εβδομάδες ο ερευνητής οφείλει να αποκλείσει την μεταβλητή $Fuel_{t-15}$ από τις μεταβλητές εισόδου. Αντίθετα, αν στο δένδρο απόφασης εμφανιζόταν για παράδειγμα μόνο η μεταβολή του αργού πετρελαίου πριν από 12 ή 13 εβδομάδες ($Brent_{t-12}$, $Brent_{t-13}$), τότε αυτό δεν θα ήταν παράλογο καθώς η ενσωμάτωση της μεταβολής τιμής του αργού πετρελαίου στην απλή αμόλυβδη μπορεί να παίρνει κάποιο χρονικό διάστημα. Γίνεται κατανοητό λοιπόν ότι οι μεταβλητές που εμφανίζονται σε κάθε δένδρο απόφασης πρέπει να ελέγχονται από την κρίση κάθε ερευνητή. Στα πρώτα δένδρα απόφασης που δημιουργήθηκαν υπήρξαν τέτοια προβλήματα, τα οποία οδήγησαν στην αφαίρεση κάποιων μεταβλητών από την βάση δεδομένων.

Αξίζει να αναφερθεί ότι όταν χρησιμοποιήθηκε συντελεστής εμπιστοσύνης (Confidence Factor) ίσος με 0,25 και μικρός αριθμός ελάχιστος στοιχείων (π.χ. 2, 3, 4), στην πλειονότητα των περιπτώσεων το δένδρο απόφασης παρουσίαζε προβλήματα υπερπροσαρμογής. Αντίθετα, όταν χρησιμοποιήθηκε μεγάλος αριθμός ελάχιστος στοιχείων (π.χ. 20) που αποτελεί και αυτός ένας τρόπος «κλαδέματος» το δένδρο απόφασης υποπροσαρμοζόταν. Ακόμη, αφαιρέθηκαν μεταβλητές εισόδου, οι οποίες είτε δεν είχαν «λογική» να χρησιμοποιούνται είτε δεν εμφανίζονταν στο δένδρο απόφασης και απλώς επιβάρυναν την διαδικασία κατασκευής του με επιπλέον υπολογιστικό κόστος. Κατά συνέπεια, αποφασίστηκε να αξιοποιηθούν για το δένδρο απόφασης ως μεταβλητές εισόδου $Brent_{t-1}$,

$Fuel_{t-1}$, $Fuel_{t-2}$, $Fuel_{t-3}$, $tr1_t$, , $tr2_t$, $tr6_t$, $tr7_t$. Στο σημείο αυτό αξίζει να αναφερθεί η παρακάτω παρατήρηση. Εξίσου καλά αποτελέσματα στο σύνολο δοκιμής επέδειξε το δένδρο απόφασης, όταν του δόθηκαν ως μεταβλητές εισόδου οι $Brent_{t-1}$, $Fuel_{t-1}$, $Fuel_{t-2}$, $Fuel_{t-3}$, $tr1_t$, $tr2_t$, $tr5_t$, $tr7_t$ αλλά όταν τοποθετήθηκε μαζί και η μεταβλητή $tr6_t$ υπερτερούσε της $tr5_t$, αφού η $tr5_t$ δεν αξιοποιούταν από το δένδρο απόφασης.

Επιλογή παραμέτρων

Έχοντας επιλέξει τις μεταβλητές εισόδου που θα δοθούν για να κατασκευαστεί το δένδρο απόφασης απαιτείται τώρα η εύρεση των βέλτιστων παραμέτρων των μεθόδων «κλαδέματος» στο σύνολο επικύρωσης. Οι μέθοδοι «κλαδέματος» που εφαρμόζονται από τις εφαρμογές της Weka είναι τρεις: α) η πρώτη είναι η μέθοδος μειωμένου σφάλματος, β) η δεύτερη αφορά τον συντελεστή εμπιστοσύνης (confidence factor) και γ) η τρίτη αφορά τον ελάχιστο αριθμό παρατηρήσεων σε κάθε «φύλλο». Όπως, έχει αναφερθεί στο κεφάλαιο 3 οι μέθοδοι «κλαδέματος» χωρίζονται στους από πάνω προς τα κάτω και τους από κάτω προς τα πάνω. Η μέθοδος μειωμένου σφάλματος αποφασίστηκε να μην χρησιμοποιηθεί για την επίλυση του προβλήματος στην παρούσα διπλωματική εργασία. Ο συντελεστής εμπιστοσύνης (confidence factor) ανήκει στην δεύτερη κατηγορία των μεθόδων «κλαδέματος». Η μέθοδος αυτή ξεκινάει από τα «φύλλα» και τους γειτονικούς κόμβους υπολογίζοντας τα ποσοστιαία σφάλματα και την πιθανότητα σφάλματος εξάγοντας τον συντελεστή εμπιστοσύνης. Αν ένας γειτονικός κόμβος έχει μικρότερο ποσοστιαίο σφάλμα από ότι το «φύλλο» που αντιστοιχεί σε αυτόν, τότε η μέθοδος αυτή τον περικόπτει για να περιορίσει τον συντελεστή εμπιστοσύνης (confidence factor). Γενικότερα, όσο μικρότερος είναι ο συντελεστής εμπιστοσύνης (confidence factor) τόσο εκτενέστερο είναι το «κλάδεμα» που γίνεται στο δένδρο απόφασης και μπορεί να εμφανιστούν προβλήματα υποπροσαρμογής. Για το πρόβλημα που επιλύει η παρούσα διπλωματική εργασία δοκιμάστηκαν συντελεστές εμπιστοσύνης από 0,05 έως 0,25 με βήμα 0,05.

Η άλλη μέθοδος «κλαδέματος» που αφορά τον ελάχιστο αριθμό παρατηρήσεων σε κάθε «φύλλο» είναι επίσης ένας άλλος αποτελεσματικός τρόπος «κλαδέματος». Είναι κατανοητό ότι όσο αυξάνεται ο ελάχιστος αριθμός στοιχείων σε κάθε «φύλλο» τόσο περισσότερο περιορίζεται η έκταση του δένδρου απόφασης και μπορεί να υπάρξουν προβλήματα υποπροσαρμογής και αντίστροφα. Για το πρόβλημα που επιλύει η παρούσα διπλωματική εργασία ο ελάχιστος αριθμός παρατηρήσεων που θα δοκιμαστούν ανά «φύλλο» είναι από 2 μέχρι 20 παρατηρήσεις με βήμα 1.

Από τις δοκιμές που έγιναν για την εύρεση των βέλτιστων παραμέτρων αξίζει να αναφερθούν ορισμένες σημαντικές παρατηρήσεις. Όσον αφορά τον αριθμό των ελάχιστων παρατηρήσεων, όταν χρησιμοποιούνταν 2,3,4 και ορισμένες φορές 5 παρατηρήσεις το δένδρο απόφασης υπερπροσαρμοζόταν στο σύνολο εκπαίδευσης με αποτέλεσμα η ακρίβεια πρόβλεψης να μειώνεται σημαντικά στο σύνολο επικύρωσης. Ωστόσο, όταν τοποθετούνταν τιμές ελάχιστων παρατηρήσεων ανά «φύλλο» μεγαλύτερες από 13 – 14, τότε το αποτέλεσμα ήταν το δένδρο να υποπροσαρμόζεται στο σύνολο εκπαίδευσης και το δένδρο απόφασης να αποτελείται από μία μόνο μεταβλητή την $Fuel_{t-1}$. Σχετικά με τον συντελεστή εμπιστοσύνης (confidence factor) τιμές μικρότερες του 0,2 οδηγούσαν σε

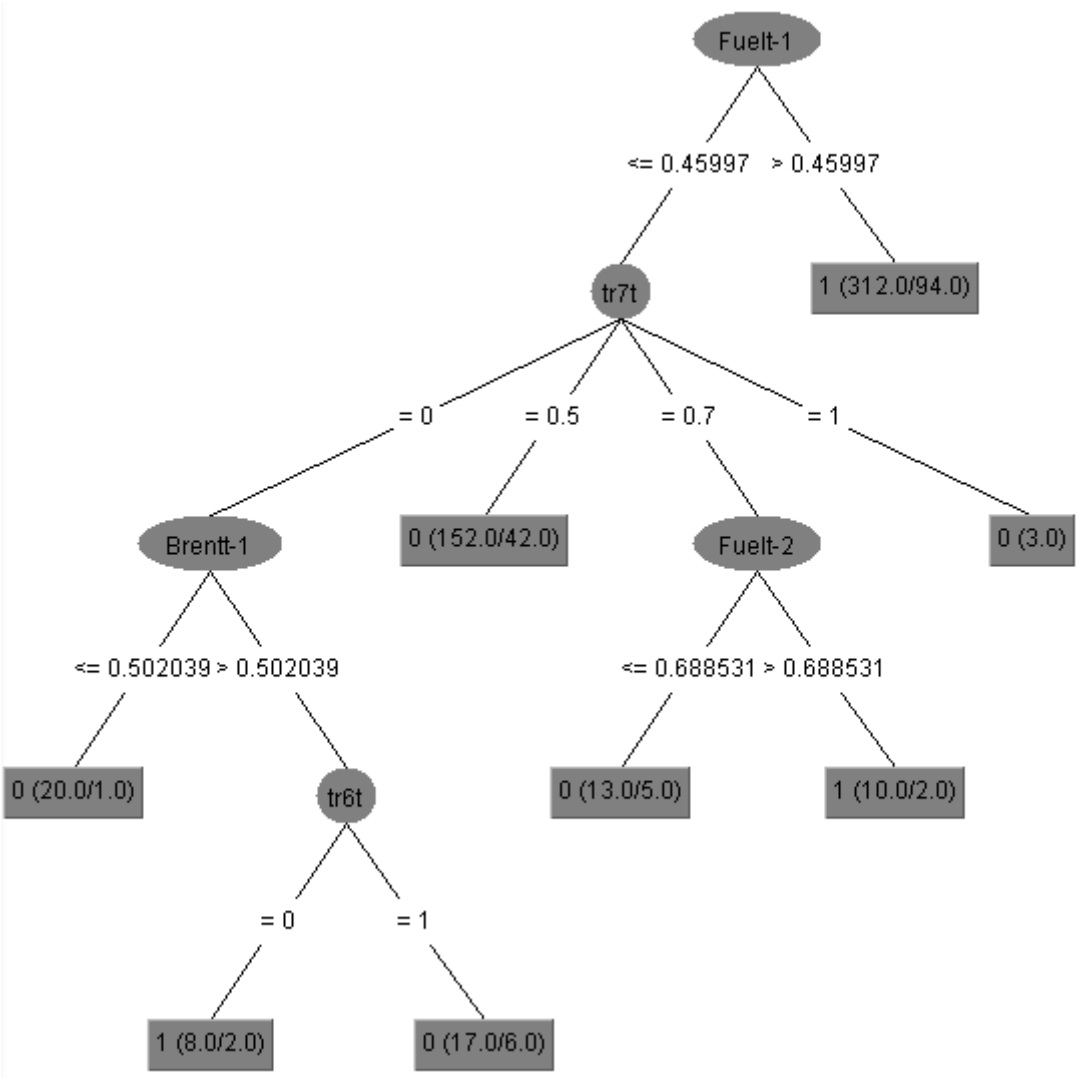
«υπερκλάδεμα» του δένδρου απόφασης, με αποτέλεσμα την δημιουργία ενός πολύ μικρού δένδρου απόφασης με μόνο μια μεταβλητή – κατηγοριοποιητή την $Fuel_{t-1}$, δηλαδή την μεταβολή της τιμής της απλής αμόλυβδης την προηγούμενη εβδομάδα. Οι βέλτιστες τιμές των παραμέτρων, όπως αυτές προέκυψαν από το σύνολο επικύρωσης παρουσιάζονται στον πίνακα 5.3.

Πίνακας 5.3 : Δεδομένα & Παράμετροι του υπό Επίλυση Προβλήματος με την εφαρμογή των δένδρων απόφασης

Δεδομένα		**Παράμετροι**	
Αριθμός Μεταβλητών Εισόδου	8	Subtree Raising	TRUE
Αριθμός Μεταβλητών Εξόδου	1	Συντελεστής Εμπιστοσύνης	0,2
Παρατηρήσεις Συνόλου Εκπαίδευσης	447	Ελάχιστος Αριθμός Παρατηρήσεων ανά "Φύλλο"	8
Παρατηρήσεις Συνόλου Επικύρωσης	59		
Παρατηρήσεις Συνόλου Δοκιμής	86	Reduced Error Pruning	FALSE

5.2.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ

Αξιοποιώντας τις παραμέτρους του πίνακα 5.3 κατασκευάστηκε το δένδρο απόφασης στο σύνολο εκπαίδευσης και υπολογίστηκαν τα στατιστικά μέτρα αξιολόγησης στο σύνολο δοκιμής, ώστε να εξαχθούν συμπεράσματα για την ακρίβεια και την αξιοπιστία των προβλέψεων του δένδρου απόφασης. Παρακάτω στο διάγραμμα 5.1 παρατίθεται η μορφή του δένδρου απόφασης.



Διάγραμμα 5.1: Δένδρο Απόφασης του υπό Επίλυση Προβλήματος

Ερμηνεία Δένδρου Απόφασης

Το δένδρο απόφασης δίνει κάποιες πληροφορίες για την σημαντικότητα των μεταβλητών εισόδου. Από πάνω προς τα κάτω η σημαντικότητα των μεταβλητών είναι φθίνουσα, δηλαδή η σημαντικότερη μεταβλητή εισόδου του δένδρου απόφασης είναι η μεταβολή της τιμής της αμόλυβδης την προηγούμενη εβδομάδα $Fuel_{t-1}$ και η λιγότερη σημαντική μεταβλητή εισόδου είναι αυτή που συγκρίνει τον εκθετικό μέσο όρο των τριών με αυτό των πέντε ημερών, $tr6_t$. Σε κάθε «κλαδί» αναφέρεται η σύγκριση της παρατήρησης με μια συγκεκριμένη τιμή της εκάστοτε μεταβλητής, ώστε να γίνει η κατηγοριοποίηση και τελικώς να καταλήξει η παρατήρηση σε ένα «φύλλο» που να κατηγοριοποιηθεί είτε ως ανοδική κατεύθυνση της τιμής της απλής αμόλυβδης (1) είτε ως καθοδική κατεύθυνση της τιμής της απλής αμόλυβδης (0). Κάθε ορθογώνιο στο αριστερό τμήμα δείχνει πόσες παρατηρήσεις έχουν κατηγοριοποιηθεί ορθά και στο δεξί τμήμα δείχνει πόσες παρατηρήσεις έχουν κατηγοριοποιηθεί λανθασμένα.

Τα αποτελέσματα του συνόλου δοκιμής προέκυψαν, όπως εξηγείται παρακάτω. Έχοντας μία παρατήρηση η κατηγοριοποίησή της ξεκινάει από την «ρίζα» του δένδρου απόφασης. Αν για παράδειγμα, η τιμή της μεταβλητής $Fuel_{t-1}$ της συγκεκριμένης παρατήρησης έχει τιμή μεγαλύτερη από 0,45 η πρόβλεψη είναι ανοδική κατεύθυνση της τιμής της απλής αμόλυβδης και η κατηγοριοποίηση σταματά εκεί. Αντίθετα, αν η τιμή της μεταβλητής $Fuel_{t-1}$ για την συγκεκριμένη παρατήρηση είναι μικρότερη από 0,45 τότε πρέπει να συνεχιστεί η κατηγοριοποίηση παρακάτω. Πρέπει να συγκριθεί και η τιμή της μεταβλητής $tr7_t$ και αυτή η διαδικασία να συνεχιστεί μέχρι να φτάσει σε κάποιο «φύλλο», όπου θα κατηγοριοποιηθεί είτε ως ανοδική (1) είτε ως καθοδική κατεύθυνση (0) της τιμής της απλής αμόλυβδης.

Στον πίνακα 5.4 παρατίθενται τα αποτελέσματα, όπως αυτά προέκυψαν από το σύνολο δοκιμής. Η ακρίβεια (accuracy) του δένδρου απόφασης στο σύνολο δοκιμής ισούται με 70%. Ο δείκτης ROC είναι 0,67 μικρότερος από την τιμή που προέκυψε με την μέθοδο των νευρωνικών δικτύων. Υπενθυμίζεται από τον ορισμό της καμπύλης ROC που αναπαριστά τον δείκτη sensitivity συναρτήσει του 1 – specificity. Επειδή η τιμή του δείκτη specificity είναι 0,348, προκύπτει ότι το εμβαδόν κάτω από την καμπύλη ROC δεν είναι τόσο μεγάλο.

Πίνακας 5.4: Πίνακας Προβλέψεων / Πραγματικών Μέτρων & Μέτρα Αξιολόγησης του υπό Επίλυση Προβλήματος με την κατασκευή Δένδρου Απόφασης

		Πρόβλεψη	**Πρόβλεψη**
		1	0
Πραγματικά Δεδομένα	1	30	16
Πραγματικά Δεδομένα	0	10	30
Ακρίβεια	0,70		
Ανοδική τάση (1)		Καθοδική τάση (0)	
ROC	0,67	ROC	0,67
F-Measure	0,70	F-Measure	0,70
Precision	0,65	Precision	0,75
Recall	0,75	Recall	0,65
Sensitivity	0,75	Sensitivity	0,65
Specificity	0,34	Specificity	0,25

Παρατηρώντας τον πίνακα των προβλέψεων φαίνεται ότι το δένδρο απόφασης μπορεί να κατηγοριοποιήσει καλύτερα τις ανοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης από ότι τις καθοδικές. Αυτό φαίνεται από τον δείκτη που υπολογίζει το ποσοστό των ορθών προβλέψεων, sensitivity, που για τις ανοδικές κατευθύνσεις της τιμής είναι ίσος με 0,75, ενώ για τις καθοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης ισούται με 0,65. Για το λόγο αυτό και η σταθμισμένη ακρίβεια είναι περίπου 0,70. Σημειώνεται ότι το ποσοστό των μη ορθών προβλέψεων των καθοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης είναι 0,25 μικρότερο από ότι των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης. Ένα γενικότερο συμπέρασμα από την κατασκευή του δένδρου απόφασης είναι ότι βοηθά τον ερευνητή να καταλάβει ποιες μεταβλητές εισόδου έχουν μεγάλη σημαντικότητα για την πρόβλεψη της μεταβλητής εξόδου και κατά συνέπεια τον βοηθά να κατανοήσει καλύτερα το υπό επίλυση πρόβλημα. Φαίνεται ότι, ενώ κατηγοριοποιεί με αρκετά μεγάλη ακρίβεια (accuracy) τις ανοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης (1) και η σταθμισμένη ακρίβεια (accuracy) του δένδρου απόφασης είναι καλή, η αξιοπιστία των προβλέψεων είναι μέτρια. Επιπρόσθετα, από τα αποτελέσματα φαίνεται ότι το δένδρο απόφασης δεν μπορεί να περιορίσει σε σημαντικό βαθμό το ποσοστό των αρνητικών προβλέψεων, με αποτέλεσμα η αξιοπιστία των προβλέψεων, όπως αυτή υπολογίζεται από τον δείκτη ROC να είναι 0,67.

Από τα αποτελέσματα φαίνεται ότι τα δένδρα απόφασης παράγουν μοντέλα πρόβλεψης μικρότερης ακρίβειας (accuracy) από ότι τα νευρωνικά δίκτυα, κάτι που επιβεβαιώνεται και από έρευνες στην υπάρχουσα βιβλιογραφία που συγκρίνουν αυτές τις δυο μεθόδους εκμάθησης μηχανών. Αναφορικά με τις πιθανές αιτίες που η κατηγοριοποίηση των καθοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης (0) παράγει αποτελέσματα μικρότερης ακρίβειας (accuracy) οφείλεται στο γεγονός ότι μπορεί να υπάρχουν πολύ διαφορετικές συμπεριφορές των καθοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης μεταξύ του συνόλου εκπαίδευσης και δοκιμής. Κατά συνέπεια, θεωρείται ότι οι καθοδικές κατευθύνσεις της τιμής της απλής αμόλυβδης χαρακτηρίζονται από στοχαστικές συμπεριφορές (τυχαίος περίπατος), οι οποίες δεν επιτρέπουν στην εκάστοτε μέθοδο εκμάθησης μηχανών να παράγουν προβλέψεις πολύ μεγάλης ακρίβειας (accuracy).

5.3 ΕΚΜΑΘΗΣΗ & ΑΞΙΟΛΟΓΗΣΗ ΜΕΘΟΔΟΥADABOOSTM1

5.3.1 ΔΕΔΟΜΕΝΑ

Όπως αναφέρθηκε στο υποκεφάλαιο 4.2 η βάση δεδομένων αποτελείται από 23 μεταβλητές εισόδου και μια μεταβλητή εξόδου, οι οποίες αναφέρονται στην περίοδο 2005 – 2016. Από το σύνολο των 597 παρατηρήσεων κάθε μεταβλητής αποφασίστηκε να χρησιμοποιηθούν οι 447 ως σύνολο εκπαίδευσης (75%), 59 ως σύνολο επικύρωσης (10%) και 86 ως σύνολο δοκιμής (15%). Υπενθυμίζεται ότι η συγκεκριμένη αναλογία επιλέχθηκε και κατά την κατασκευή των δυο προηγούμενων μεθόδων εκμάθησης μηχανών, των νευρωνικών δικτύων και των δένδρων απόφασης. Θεωρήθηκε απαραίτητο σε κάθε μέθοδο εκμάθησης μηχανών να τηρηθεί η ίδια αναλογία, προκειμένου να μπορεί να γίνει σύγκριση μεταξύ τους και να εξαχθούν συμπεράσματα σχετικά με τα πλεονεκτήματα και τα μειονεκτήματα της εκάστοτε μεθόδου εκμάθησης μηχανών.

5.3.2 ΕΚΜΑΘΗΣΗ & ΕΠΙΛΟΓΗ ΠΑΡΑΜΕΤΡΩΝ ΜΕΘΟΔΟΥ ADABOOSTM1

Εκμάθηση Μεθόδου AdaboostM1

Έχοντας δοκιμάσει ήδη από το υποκεφάλαιο 5.2 διαφόρους συνδυασμούς μεταβλητών εισόδου με κατάλληλες τιμές για τον συντελεστή εμπιστοσύνης (confidence factor) και τον ελάχιστο αριθμό στοιχείων ανά «φύλλο» είχαν επιλεγεί να αξιοποιηθούν οι μεταβλητές εισόδου $Brent_{t-1}$, $Fuel_{t-1}$, $Fuel_{t-2}$, $Fuel_{t-3}$, $tr1_t$, , $tr2_t$, $tr6_t$, $tr7_t$. Επειδή ο αλγόριθμος AdaboostM1 χρησιμοποιεί ως «αδύναμο» κατηγοριοποιητή τα δάση απόφασης, αποφασίστηκε να χρησιμοποιηθούν οι παραπάνω μεταβλητές ως μεταβλητές εισόδου για την εκμάθηση του αλγορίθμου και να εφαρμοστεί η μέθοδος της δοκιμής – σφάλματος για την εύρεση των βέλτιστων παραμέτρων του συντελεστή εμπιστοσύνης (confidence factor) και του ελάχιστου αριθμού στοιχείων ανά «φύλλο» στο σύνολο επικύρωσης.

Επιλογή παραμέτρων

Στο σημείο αυτό αξίζει να αναφερθεί πότε ο αλγόριθμος AdaboostM1 μπορεί να παράξει αξιόπιστες προβλέψεις με μεγάλη ακρίβεια. Ο συγκεκριμένος αλγόριθμος επειδή δίνει έμφαση στην κατηγοριοποίηση των μη ορθών παρατηρήσεων τοποθετώντας επιπλέον βάρος σε αυτές, απαιτεί ο «αδύναμος» αλγόριθμος που εφαρμόζεται ως κατηγοριοποιητής να μην υπερπροσαρμόζεται στο σύνολο εκπαίδευσης. Γίνεται κατανοητό ότι αν ο «αδύναμος» κατηγοριοποιητής υπερπροσαρμοστεί στα δεδομένα, τότε οι μη ορθές κατηγοριοποιημένες παρατηρήσεις θα είναι ελάχιστες και ο αλγόριθμος θα σταματήσει πολύ γρήγορα. Έτσι, ο αλγόριθμος AdaboostM1 απαιτεί μη πολύπλοκα δάση απόφασης και γενικότερα απλά δάση απόφασης. Αυτό οδηγεί στο να επικεντρωθεί η έρευνα στην εύρεση κατάλληλων παραμέτρων των μεθόδων «κλαδέματος» που να περιορίζουν σημαντικά την έκταση του δένδρου απόφασης. Από την άλλη μεριά αν ο «αδύναμος» κατηγοριοποιητής έχει υποπροσαρμοστεί στα δεδομένα σε σημαντικό βαθμό, τότε η ακρίβεια πρόβλεψης θα είναι μικρή και ενδέχεται ενισχύοντας τα βάρη των μη ορθών κατηγοριοποιημένων παρατηρήσεων, το επόμενο δένδρο απόφασης που θα κατασκευαστεί να έχει ποσοστό σφάλματος μεγαλύτερο από 0,50. Σε μια τέτοια περίπτωση ο αλγόριθμος AdaboostM1 θα σταματήσει τις επαναλήψεις πολύ γρήγορα και ο ερευνητής δεν θα μπορέσει να οφεληθεί από τα πλεονεκτήματα της συγκεκριμένης μεθόδου εκμάθησης μηχανών.

Συνεπώς, το σύνολο επικύρωσης θα χρησιμοποιηθεί για την εύρεση των βέλτιστων παραμέτρων των μεθόδων «κλαδέματος», καθώς επίσης και τον αριθμό των επαναλήψεων που σταματάει ο αλγόριθμος AdaboostM1, ώστε να μην υπερπροσαρμοστεί στο σύνολο εκπαίδευσης. Επισημαίνεται ότι ο αλγόριθμος σταματά είτε όταν το ποσοστό σφάλματος σε έναν «αδύναμο» κατηγοριοποιητή είναι μεγαλύτερο από 0,5 είτε όταν ο αριθμός των επαναλήψεων γίνει ίσος με τον μέγιστο αριθμό επαναλήψεων που ο ερευνητής έχει θέσει. Οι μέθοδοι «κλαδέματος» που θα εφαρμοστούν είναι: α) ο συντελεστής εμπιστοσύνης (confidence factor), β) ο ελάχιστος αριθμός παρατηρήσεων ανά «φύλλο». Για την πρώτη μέθοδο «κλαδέματος» γίνονται δοκιμές με τιμές από 0,1 έως 0,25 με βήμα 0,05 και για την δεύτερη μέθοδο γίνονται δοκιμές από 4 έως 28 παρατηρήσεις ανά «φύλλο» με βήμα ένα. Για τον μέγιστο αριθμό των επαναλήψεων χρησιμοποιήθηκαν δοκιμάστηκαν οι τιμές από 8 έως 12 με βήμα 1. Υπενθυμίζεται από το υποκεφάλαιο 5.2, όταν τοποθετήθηκε ελάχιστος αριθμός παρατηρήσεων ανά «φύλλο» ίσος με 2, 3, 4 υπήρξαν προβλήματα υπερπροσαρμογής και για τον λόγο αυτόν αποφασίστηκε να ξεκινήσουν οι δοκιμές με την συγκεκριμένη μέθοδο από ελάχιστο αριθμό παρατηρήσεων ίσο με τέσσερα.

Από τις δοκιμές που έγιναν αξίζει να αναφερθούν ορισμένες σημαντικές παρατηρήσεις. Πρώτη σημαντική παρατήρηση αποτελεί το γεγονός ότι για μικρές τιμές των ελάχιστων παρατηρήσεων ανά «φύλλο» η ακρίβεια πρόβλεψης στο σύνολο επικύρωσης δεν ήταν καλή. Επίσης, παρατηρήθηκε ότι κατά την τοποθέτηση ελάχιστου αριθμού παρατηρήσεων ανά «φύλλο» μεγαλύτερο του 26 εμφανίζεται υποπροσαρμογή των δένδρων απόφασης, τα οποία χρησιμοποιούν μόνο μια μεταβλητή – κατηγοριοποιητή, την $Fuel_{t-1}$ και οι επαναλήψεις του αλγορίθμου AdaboostM1 σταματούν πολύ γρήγορα. Ο συντελεστής εμπιστοσύνης (confidence factor) έπαιξε σημαντικό ρόλο στην βελτίωση της ακρίβειας στο

σύνολο επικύρωσης. Ο μέγιστος αριθμός των επαναλήψεων δεν διαφοροποίησε σημαντικά τα στατιστικά μέτρα αξιολόγησης και βρέθηκε ότι ο βέλτιστος αριθμός ισούται με 10. Άλλη μια σημαντική παρατήρηση αποτελεί το γεγονός ότι τοποθετώντας μεγάλο αριθμό ελάχιστων παρατηρήσεων ανά «φύλλο» (μεγαλύτερο από 21) και μεγάλο συντελεστή εμπιστοσύνης (μεγαλύτερο από 0,15) η ακρίβεια πρόβλεψης των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης αυξήθηκε σημαντικά σε επίπεδα 85% - 90%, ενώ η ακρίβεια πρόβλεψη των καθοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης μειώθηκε σε μεγάλο βαθμό. Οι βέλτιστες τιμές των παραμέτρων των μεθόδων «κλαδέματος» καθώς και του μέγιστου αριθμού των επαναλήψεων, όπως αυτές προέκυψαν από το σύνολο επικύρωσης παρατίθονται στον πίνακα 5.5.

Πίνακας 5.5 : Δεδομένα & Παράμετροι του υπό Επίλυση Προβλήματος με την εφαρμογή της μεθόδου AdaboostM1

Δεδομένα		**Παράμετροι**	
Αριθμός Μεταβλητών Εισόδου	8	Subtree Raising	TRUE
Αριθμός Μεταβλητών Εξόδου	1	Συντελεστής Εμπιστοσύνης	0,1
Παρατηρήσεις Συνόλου Εκπαίδευσης	447	Ελάχιστος Αριθμός Παρατηρήσεων ανά "Φύλλο"	24
Παρατηρήσεις Συνόλου Επικύρωσης	59	Μέγιστος Αριθμός Επαναλήψεων	10
Παρατηρήσεις Συνόλου Δοκιμής	86	Reduced Error Pruning	FALSE

5.3.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ ADABOOSTM1

Αξιοποιώντας τις παραμέτρους του πίνακα 5.5 πραγματοποιήθηκε η εκμάθηση του μεθόδου AdaboostM1 με «αδύναμο» κατηγοριοποιητή τα δάση απόφασης στο σύνολο εκπαίδευσης και υπολογίστηκαν τα στατιστικά μέτρα αξιολόγησης στο σύνολο δοκιμής, ώστε να εξαχθούν συμπεράσματα για την ακρίβεια και την αξιοπιστία των προβλέψεων της μεθόδου AdaboostM1.

Πίνακας 5.6: Πίνακας Προβλέψεων / Πραγματικών Μέτρων & Μέτρα Αξιολόγησης του υπό Επίλυση Προβλήματος με την μέθοδο AdaboostM1

		Πρόβλεψη	**Πρόβλεψη**
		1	0
Πραγματικά Δεδομένα	1	27	13
Πραγματικά Δεδομένα	0	13	33
Ακρίβεια	0,70		
Ανοδική Τάση (1)		Καθοδική Τάση (0)	
ROC	0,72	ROC	0,72
F-Measure	0,68	F-Measure	0,72
Precision	0,68	Precision	0,72
Recall	0,68	Recall	0,72
Sensitivity	0,68	Sensitivity	0,72
Specificity	0,28	Specificity	0,33

Τα αποτελέσματα, όπως αυτά προέκυψαν από το σύνολο δοκιμής παρατίθονται στον πίνακα 5.6. Παρατηρώντας το πίνακα 5.6, το σημαντικότερο στατιστικό μέτρο που είναι η ακρίβεια (accuracy) ισούται με 70%, που είναι πολύ καλή τιμή για την πρόβλεψη της κατεύθυνσης της τιμής της απλής αμόλυβδης. Επίσης, ο δείκτης ROC που αντικατοπτρίζει την αξιοπιστία πρόβλεψης είναι ίσος με 0,72 για τις ανοδικές (1) και τις καθοδικές (0)

προβλέψεις της τιμής της απλής αμόλυβδης. Έτσι, μπορεί να εξαχθεί το συμπέρασμα ότι οι προβλέψεις είναι αξιόπιστες. Υπενθυμίζεται ότι ο δείκτης ROC είναι το εμβαδόν της καμπύλης του δείκτη sensitivity συναρτήσει του 1–specificity. Έτσι, αν και το ποσοστό των ορθών προβλέψεων των καθοδικών κατευθύνσεων (0) είναι λίγο μεγαλύτερο από ότι το ποσοστό των προβλέψεων των ανοδικών κατευθύνσεων (1), επειδή το ποσοστό των μη ορθών καθοδικών προβλέψεων είναι μεγαλύτερο, προκύπτει ότι έχουν ίδιο δείκτη ROC.

Επιπρόσθετα, από τα αποτελέσματα προκύπτει ότι ο συνδυασμός πολλών «αδύναμων» κατηγοριοποιητών συμβάλλει στην βελτίωση της ακρίβειας πρόβλεψης των καθοδικών κατευθύνσεων (0) της τιμής της απλής αμόλυβδης. Υπενθυμίζεται ότι στα υποκεφάλαια 5.1, 5.2 είχε παρατηρηθεί ότι οι μέθοδοι των νευρωνικών δικτύων και των δένδρων απόφασης δεν μπορούσαν να κατηγοριοποιήσουν με την ίδια ακρίβεια τις καθοδικές (0) και τις ανοδικές κατευθύνσεις (1) της τιμής της απλής αμόλυβδης. Έτσι, είναι λογικό η μέθοδος AdaboostM1 να παράγει προβλέψεις μεγαλύτερης ακρίβειας για τις καθοδικές κατευθύνσεις (0) από ότι αν χρησιμοποιούταν η εκάστοτε μέθοδος μεμονωμένα, αφού το πλεονέκτημα της μεθόδου είναι ότι επικεντρώνεται στις μη ορθά κατηγοριοποιημένες παρατηρήσεις.

Από τα αποτελέσματα παρατηρείται ότι η σταθμισμένη ακρίβεια είναι παρόμοια με το αν εφαρμοζόταν η κάθε μια μέθοδος εκμάθησης μηχανών ξεχωριστά, ωστόσο το πλεονέκτημα της παρούσας μεθόδου είναι ότι κατηγοριοποιεί με μεγαλύτερη ακρίβεια τις καθοδικές κατευθύνσεις (0) της τιμής της απλής αμόλυβδης από ότι αν εφαρμοζόταν η κάθε μια μέθοδος εκμάθησης μηχανών μεμονωμένα. Αναφορικά με τις πιθανές αιτίες που εμποδίζουν την ακρίβεια πρόβλεψης να φτάσει σε επίπεδα 85% - 90% θεωρείται ότι η κύρια αιτία είναι το γεγονός ότι η χρονοσειρά της απλής αμόλυβδης παρουσιάζει έντονα στοχαστική συμπεριφορά (τυχαίος περίπατος). Οι συμπεριφορές που εμφανίζει η χρονοσειρά στο σύνολο εκπαίδευσης είναι πολύ διαφορετικές από ότι οι συμπεριφορές του συνόλου δοκιμής λόγω της στοχαστικότητας και κατά συνέπεια ακόμη και η μέθοδος AdaboostM1 που συνδυάζει πολλούς «αδύναμους» κατηγοριοποιητές να μην μπορεί να εξασφαλίσει ακρίβεια πρόβλεψης της τάξης του 85% - 90%.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ& ΠΡΟΤΑΣΕΙΣ

6.1 ΓΕΝΙΚΑ

Στο κεφάλαιο 6 θα γίνει μια σύνοψη των βασικότερων σημείων της παρούσας διπλωματικής εργασίας, των συμπερασμάτων που προέκυψαν από την επίλυση του προβλήματος της πρόβλεψης της τιμής της απλής αμόλυβδης και θα παρατεθούν κάποιες προτάσεις για περαιτέρω έρευνα. Η πρόβλεψη της τιμής της αμόλυβδης σε μεσοπρόθεσμο ορίζοντα αποτελεί θέμα μείζονος σημασίας τόσο για τους πολίτες μεμονωμένα όσο και τις εταιρίες μεταφορών στον συγκοινωνιακό τομέα, καθώς οι μεταβολές των τιμών των καυσίμων επιδρούν σημαντικά στο μεταφορικό κόστος. Η ακρίβεια πρόβλεψης που επιτυγχάνεται στην παρούσα διπλωματική εργασία υποδεικνύει ότι η χρήση αυτών των μοντέλων για την πρόβλεψη της τιμής της αμόλυβδης μπορεί να συμβάλλει σημαντικά στον περιορισμό των μεταβολών του μεταφορικού κόστους και να ενισχύσει την ευημερία του κοινωνικού συνόλου.

6.2 ΒΑΣΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι σημαντικές επιπτώσεις των μεταβολών των καυσίμων στον συγκοινωνιακό τομέα δημιούργησαν την ανάγκη κατασκευής μοντέλων πρόβλεψης της τιμής της αμόλυβδης που αποτελεί και τον σκοπό της παρούσας διπλωματικής εργασίας. Η παρούσα διπλωματική εργασία αποσκοπούσε στην κατασκευή μοντέλων με όσο το δυνατόν μεγαλύτερη ακρίβεια πρόβλεψης, ώστε να μπορούν να μετριαστούν αποτελεσματικά οι παραπάνω επιπτώσεις.

Στο κεφάλαιο 2 έγινε αναλυτική περιγραφή των επιπτώσεων από τις μεταβολές των τιμών των καυσίμων σε χώρες εντός και εκτός της Ε.Ε.. Πιο συγκεκριμένα, από όλες τις έρευνες που παρατέθηκαν φαίνεται ότι η αύξηση / μείωση της τιμής των καυσίμων έχει σημαντική επίδραση στην κυκλοφοριακή ζήτηση των μέσων μαζικής μεταφοράς. Έγινε αντιληπτό ότι ακόμη και μικρές αυξομειώσεις της τιμής των καυσίμων οδηγούν σε μεγάλες αυξομειώσεις της ζήτησης σε απόλυτο αριθμό στα μέσα μαζικής μεταφοράς, με αποτέλεσμα να δημιουργούνται αδυναμίες εξυπηρέτησης της ζήτησης σε ορισμένες περιπτώσεις. Επίσης, το λεωφορείο είναι το μέσο μαζικής μεταφοράς που επηρεάζεται περισσότερο από τις αυξομειώσεις των τιμών των καυσίμων, καθώς είναι εκείνο που οι πολίτες προτιμούν να αντικαταστήσουν, όταν αυξάνεται το μεταφορικό κόστος του επιβατικού αυτοκινήτου. Σημαντική παρατήρηση αποτέλεσε μια διερεύνηση που έγινε στις ΗΠΑ και δείχνει ότι ανάλογα του πεδίου τιμών των καυσίμων η ελαστικότητα της ζήτησης για τα μέσα μαζικής μεταφοράς είναι διαφορετική. Για παράδειγμα, για χαμηλές τιμές καυσίμων η ελαστικότητα της ζήτησης είναι μηδαμινή, καθώς οι πολίτες δεν αλλάζουν μέσο μεταφοράς, όταν η τιμή των καυσίμων είναι σε πολύ χαμηλά επίπεδα. Αντίθετα, το 2008 που οι τιμές των καυσίμων ήταν σε ιστορικά υψηλά η ελαστικότητα της ζήτησης ήταν αρκετά μεγαλύτερη από τις τιμές που υπάρχουν στην υπάρχουσα βιβλιογραφία. Αναφορικά με την Ελλάδα φαίνεται ότι η αύξηση της τιμής των καυσίμων λόγω αύξησης του ειδικού φόρου κατανάλωσης είχε ιδιαίτερα σημαντικές επιπτώσεις στην κυκλοφοριακή ροή στους αυτοκινητοδρόμους, καθώς οι πολίτες στην χώρα έχουν μικρά εισοδήματα και δεν μπορούν να απορροφήσουν τις οποιεσδήποτε αυξήσεις.

Ένα ακόμη συμπέρασμα που προκύπτει από τις περισσότερες χώρες που έχουν γίνει έρευνες είναι ότι αναλόγως τον σκοπό της διαδρομής, δηλαδή επαγγελματικό ή αναψυχής οι πολίτες τείνουν να αναδιαμορφώνουν διαφορετικά τις συνήθειες τους. Συγκεκριμένα, σε περιπτώσεις αυξήσεων των τιμών των καυσίμων οι πολίτες περιορίζουν άμεσα τα ταξίδια αναψυχής, ενώ τα επαγγελματικά ταξίδια μειώνονται σε μικρότερο βαθμό και σε μεγαλύτερο χρονικό ορίζοντα. Η σημασία της πρόβλεψης της τιμής της αμόλυβδης είναι πολύ σημαντική για τους οδικούς μεταφορείς, επειδή τυχόν μεταβολές στο μεταφορικό κόστος εκμηδενίζουν την κερδοφορία τους και τους θέτει εκτός ανταγωνισμού. Έτσι, δημιουργείται η ανάγκη πρόβλεψης της τιμής των καυσίμων σε μεσοπρόθεσμο ορίζοντα, ώστε να μπορούν να ρυθμίζουν τα αποθέματά τους, καθώς από όλες τις έρευνες προκύπτει

το συμπέρασμα ότι οποιεσδήποτε αυξήσεις στα μεταφορικά κόστη δεν μπορούν να τις μετακυλίσουν στους πελάτες λόγω του έντονου ανταγωνισμού στον κλάδο.

Στη συνέχεια στο κεφάλαιο 5 για την πρόβλεψη της κατεύθυνσης της τιμής της απλής αμόλυβδης εφαρμόστηκαν τρεις μέθοδοι, τα νευρωνικά δίκτυα, τα δένδρα απόφασης και τα δάση απόφασης με βάση την μέθοδο AdaboostM1. Υπενθυμίζεται ότι επειδή τους πολίτες τους ενδιαφέρει η πληροφορία για το αν η τιμή της απλής αμόλυβδης θα είναι μεγαλύτερη / μικρότερη από ότι την προηγούμενη εβδομάδα, η παρούσα διπλωματική εργασία επιλύει πρόβλημα τύπου 0/1, όπου 0/1 δηλώνουν καθοδική / ανοδική κατεύθυνση της τιμής αντίστοιχα. Η σημαντικότερη παρατήρηση που προέκυψε και από τις τρεις μεθόδους είναι ότι η μεταβολή της αμόλυβδης την προηγούμενη εβδομάδα $Fuel_{t-1}$ αποτελεί την σημαντικότερη μεταβλητή εισόδου και αυτό φαίνεται από το γεγονός ότι στα δένδρα απόφασης ήταν η πρώτη μεταβλητή – κατηγοριοποιητής που χρησιμοποιήθηκε και κάθε «αδύναμος» κατηγοριοποιητής της μεθόδου AdaboostM1 εμπεριείχε την μεταβλητή $Fuel_{t-1}$. Επίσης, αξίζει να αναφερθεί ότι αξιοποιώντας μόνο την μεταβλητή $Fuel_{t-1}$ ως μεταβλητή εισόδου στο νευρωνικό δίκτυο μπορούσαν να επιτευχθούν ακρίβειες της τάξης του 64% - 65% στο σύνολο δοκιμής. Επομένως, γίνεται αντιληπτή η σημασία της συγκεκριμένης μεταβλητής για την επίλυση του προβλήματος της παρούσας διπλωματικής εργασίας.

Έπειτα, σημειώνεται ότι η μέθοδος των νευρωνικών δικτύων μπόρεσε να επιτύχει μεγαλύτερη ακρίβεια πρόβλεψης στο σύνολο δοκιμής από ότι τα δένδρα απόφασης. Επισημαίνεται ότι η μέθοδος των νευρωνικών δικτύων επιτυγχάνει μεγαλύτερη ακρίβεια πρόβλεψης των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης, με αποτέλεσμα να εμφανίζει λίγο μεγαλύτερη σταθμισμένη ακρίβεια πρόβλεψης. Τονίζεται ότι κατά την διάρκεια των δοκιμών του αριθμού των νευρώνων στο κρυμμένο επίπεδο, παρατηρήθηκε ότι όσο αυξάνεται ο συγκεκριμένος αριθμός τόσο βελτιώνεται η ακρίβεια πρόβλεψης των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης, κάτι που υποδηλώνει την έντονη μη γραμμικότητα των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης. Τα επίπεδα της ακρίβειας πρόβλεψης των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης για μεγάλο αριθμό νευρώνων στο κρυμμένο επίπεδο ήταν της τάξης του 85% - 86%, ωστόσο το βέλτιστο μοντέλο πρόβλεψης είχε 4 νευρώνες στο κρυμμένο επίπεδο.

Από την άλλη τα οφέλη της μεθόδου των δένδρων απόφασης έγκεινται στο γεγονός ότι απαιτούν μικρό υπολογιστικό κόστος και είναι απλά στην ερμηνεία τους ακόμη και από ανθρώπους με μικρή ενασχόληση στις μεθόδους εκμάθησης μηχανών. Γενικότερα, από την υπάρχουσα βιβλιογραφία επιβεβαιώνεται ότι η ακρίβεια πρόβλεψης των δένδρων απόφασης είναι λίγο μικρότερη από ότι των υπολοίπων μεθόδων εκμάθησης μηχανών.

Σχετικά με την μέθοδο AdaboostM1 αξίζει να σημειωθεί ότι είναι η μέθοδος που επιτύγχανε την μεγαλύτερη ακρίβεια πρόβλεψης των καθοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης στο σύνολο δοκιμής. Για να εξηγηθεί αυτό το αποτέλεσμα της μεθόδου AdaboostM1 πρέπει να υπενθυμιστεί ότι τα νευρωνικά δίκτυα και τα δένδρα απόφασης εμφάνισαν μικρότερη ακρίβεια πρόβλεψης των καθοδικών κατευθύνσεων από ότι των ανοδικών κατευθύνσεων της τιμής της απλής αμόλυβδης. Το γεγονός αυτό σε συνδυασμό με το χαρακτηριστικό της μεθόδου AdaboostM1 να επικεντρώνεται στις μη ορθά κατηγοριοποιημένες παρατηρήσεις εξηγεί τον λόγο που εξασφάλισε μεγαλύτερη

ακρίβεια πρόβλεψης των καθοδικών κατευθύνσεων από ότι οι άλλες δυο μέθοδοι εκμάθησης μηχανών.

Αναφορικά με τις αιτίες που δεν επέτρεψαν να επιτευχθεί ακρίβεια πρόβλεψης με οποιαδήποτε μέθοδο της τάξης του 90% -95% οφείλεται στο γεγονός ότι η χρονοσειρά της τιμής της απλής αμόλυβδης, όπως και των υπολοίπων καυσίμων έχουν έντονη στοχαστική συμπεριφορά (τυχαίος περίπατος), με αποτέλεσμα να εμφανίζονται αρκετά διαφορετικές συμπεριφορές των παρατηρήσεων στο σύνολο δοκιμής από ότι στο σύνολο εκπαίδευσης. Αυτός είναι και ο κύριος λόγος που οι ερευνητές, όταν διερευνούν την προβλεψιμότητα των λιανικών τιμών των καυσίμων συγκρίνουν τα αποτελέσματά τους με τα αποτελέσματα του μοντέλου μη αλλαγής της τιμής. Ωστόσο, η παρούσα διπλωματική εργασία κατάφερε να πετύχει ακρίβεια πρόβλεψης της κατεύθυνσης της τιμής της απλής αμόλυβδης μέχρι και 75% στο σύνολο δοκιμής ενισχύοντας έτσι σημαντικά τις προσπάθειες της υπάρχουσας βιβλιογραφίας. Συνεπώς, με την χρήση οποιουδήποτε από τα μοντέλα πρόβλεψης που κατασκευάστηκαν στην παρούσα διπλωματική εργασία μπορούν οι πολίτες να βελτιστοποιούν τον χρονισμό αγοράς καυσίμου για τα οχήματά τους και οι οδικοί μεταφορείς να ρυθμίζουν κατάλληλα τα αποθέματα τους, ώστε να παραμένουν ανταγωνιστικοί στον κλάδο και να περιορίσουν σε σημαντικό βαθμό τις διακυμάνσεις του μεταφορικού κόστους.

6.3 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν τρεις μέθοδοι εκμάθησης μηχανών για την διερεύνηση της προβλεψιμότητας της απλής αμόλυβδης και από τα αποτελέσματα φαίνεται ότι η ακρίβεια πρόβλεψης των εκάστοτε μοντέλων είναι πολύ καλή και παρέχουν πληροφορίες, οι οποίες θα μπορούσαν να χρησιμοποιηθούν προς όφελος του κοινωνικού συνόλου. Έτσι, έπειτα από τα αποτελέσματα της παρούσας διπλωματικής εργασίας αξίζει να διερευνηθούν περαιτέρω τα παρακάτω:

Πρώτον, από τα αποτελέσματα εξήχθη το συμπέρασμα ότι οι τρεις μέθοδοι εκμάθησης μηχανών, νευρωνικά δίκτυα, δένδρα απόφασης και AdaboostM1 μπορούν να παράγουν αξιόπιστες και με αρκετά μεγάλη ακρίβεια προβλέψεις της απλής αμόλυβδης σε μεσοπρόθεσμο ορίζοντα. Έτσι, προκύπτει το ερώτημα αν οι παραπάνω μέθοδοι μπορούν να παράξουν προβλέψεις παρόμοιας ακρίβειας σε βραχυπρόθεσμο και μακροπρόθεσμο χρονικό ορίζοντα και να επιβεβαιωθεί ακόμη περισσότερο η υπεροχή αυτών των μεθόδων έναντι των γραμμικών μοντέλων πρόβλεψης.

Δεύτερον, σε άλλη έρευνα θα μπορούσαν να χρησιμοποιηθούν άλλες μέθοδοι εκμάθησης μηχανών για την ίδια χρονοσειρά της απλής αμόλυβδης και να συγκριθούν τα αποτελέσματα με τις μεθόδους που εφαρμόστηκαν στην παρούσα διπλωματική εργασία. Θεωρείται ότι μια τέτοια έρευνα θα είχε σημαντική συνεισφορά για την εύρεση της βέλτιστης μεθόδου εκμάθησης μηχανών για την πρόβλεψη της χρονοσειράς της απλής αμόλυβδης.

Τρίτον, στην παρούσα διπλωματική εργασία διερευνήθηκε η χρονοσειρά της απλής αμόλυβδης. Αν και θεωρείται ότι οι μεταβολές και των υπόλοιπων παράγωγων προϊόντων του αργού πετρελαίου, όπως είναι η ενισχυμένη αμόλυβδη και το ντίζελ θα έχουν παρόμοια χαρακτηριστικά, αυτό μένει να επιβεβαιωθεί από μια μελλοντική έρευνα.

Τέταρτον, μια μελλοντική έρευνα θα μπορούσε να αναφερθεί στις λιανικές τιμές των καυσίμων μόνο σε μια χώρα (π.χ. Ελλάδα) και να διερευνηθεί η προβλεψιμότητα των λιανικών τιμών των καυσίμων. Σε μια τέτοια έρευνα που θα αφορούσε μια χώρα ατομικά θα μπορούσαν να χρησιμοποιηθούν και μεταβλητές εισόδου που σχετίζονται με τις μεταβολές των αποθεμάτων στα διυλιστήρια της υπό διερεύνηση χώρας. Αξιοποιώντας και αυτές τις μεταβλητές εισόδου θεωρείται ότι οι μέθοδοι εκμάθησης μηχανών θα κέρδιζαν σημαντικές πληροφορίες και η ακρίβεια πρόβλεψης θα βελτιωνόταν σημαντικά.

Πέμπτον, αξιοποιώντας ένα από τα παραπάνω μοντέλα πρόβλεψης της τιμής της απλής αμόλυβδης που κατασκευάστηκαν στην παρούσα διπλωματική εργασία θα είχε εξαιρετικό ενδιαφέρον να δει κανείς πως αλλάζουν οι ελαστικότητες της ζήτησης στα μέσα μαζικής μεταφοράς, όταν οι πολίτες μπορούν να εκμεταλλευτούν τις προβλέψεις που παράγουν τα παραπάνω μοντέλα ρυθμίζοντας κατάλληλα τον χρονισμό αγοράς καυσίμων για τα επιβατικά οχήματά τους.

7. ΒΙΒΛΙΟΓΡΑΦΙΑ

Abate, M., (2014). Does fuel price affect trucking industry's network characteristics? Evidence from Denmark. *Centre for Transport Studies,* CTS Working Paper 2014:26.

Abdullah, S. N., Zeng, X., (2010). Machine Learning Approach for Crude Oil Price Prediction with Artificial Neural Networks – Quantitative (ANN-Q) Model. *The 2010 International Joint Conference on Neural Networks (IJCNN), 2010 IEEE World Congress on Computational Intelligence (WCCI),* 2010.

Almuallim, H., (1996). An Efficient Algorithm for Optimal Pruning of Decision Trees, *Artificial Intelligence,* 83(2) : 347 – 362, 1996.

Bajjalieh, J., (2010). Forecasting Diesel Fuel Prices. *Master Thesis for the degree of Master of Science in Agriculture and Consumer Economics in the Graduate College of the University of Illinois at Urbana-Champaign,* 2010.

Baumeister, C., Kilian, L., Lee, T. K., (2015). Inside the Crystal Ball: New Approaches to predicting the Gasoline Price at the Pump. *Journal of Applied Econometrics,* May 2015.

Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M., (2013). Decision Tree Analysis on J48 Algorithm for Data Mining, *International Journal of Advanced Research in Computer Science and Software Engineering,* Volume 3, Issue 6, June 2013.

Bratko, I., Bohanec, M., (1994). Trading accuracy for simplicity in decision trees, *Machine Learning,* 15 : 223-250, 1994.

Breiman, L., (1999). Prediction games and arcing classifiers, Neural Computation II (7), 1493 – 1517, (1999).

Breiman, L., Friedman, J., Olshen, R., Stone, C., (1984). Classification and Regression Trees, *Wadsworth Int. Group,* 1984.

Cornell University, (2003). *Performance Measures for Machine Learning,* Retrieved 03 2017, Cornell University, Computer Science: https://www.cs.cornell.edu/courses/cs578/2003fa/performance_measures.pdf

EIA, (2016). *Gasoline Explained, Factors Affecting Gasoline Prices, 2016,* Retrieved 03 2017, from U.S. Energy Information Administration, Energy Explained: https://www.eia.gov/energyexplained/index.cfm?page=gasoline_factors_affecting_prices

EIA. (2015). *Share of total U.S. energy used for transportation, 2015,* Retrieved 03 2017, from U.S. Energy Information Administration, Monthly Energy Review: https://www.eia.gov/energyexplained/?page=us_energy_transportation

Espisito, F., Malerba, D., Semeraro, G., Kay, J., (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 19, Issue 5, 1997.

European Commission, (2016). *Breakdown of automotive gasoline prices across the EU Member States, 2016,* Retrieved 03 2017, from European Commission, FuelsEurope: https://www.fuelseurope.eu/knowledge/refining-in-europe/economics-of-refining/fuel-price-breakdown

European Commission, (2002). *The effects of fuel price changes on the transport sector and its emissions – simulations with TREMOVE N° 172 - July 2002,* Retrieved 03 2017, European Commission, Economic Papers: http://europa.eu.int/comm/economy_finance

European Commission, (2016). *Consumption of oil EU-28, 2014, percentage update,* Retrieved 03 2017, from European Commission, Eurostat: ,http://ec.europa.eu/eurostat/statistics explained/index.php/ File:Consumption_of_oil_EU-28,_2014,_percentage_update.png

European Commission, (2016). *Energy consumption by transport mode, EU-28, 2014, (1990 = 100, based on tones of oil equivalent,* Retrieved 03 2017, from European Commission, Eurostat: http://ec.europa.eu/eurostat/statistics/explained/index.php/Consumption_of_energy

European Commission, (2016). *Final energy consumption, EU-28, 2014 (% of total, based on tones of oil equivalent)*, Retrieved 03 2017, from European Commission, Eurostat: http://ec.europa.eu/eurostat/statistic/explained/index.php/File:Final_energy_consumption,_EU28,_2014_(%25_of_total,_based_on_tonnes_of_oil_equivalent)_YB16.png

European Parliament, (2009). *The impact of oil prices fluctuations on transport and its related sectors,* Retrieved 03 2017, European Parliament, Policy Department Structural and Cohesion Policies: http://www.europarl.europa.eu/RegData/etudes/etudes/join/2009/419084/IPOL-TRAN_ET(2009)419084_EN.pdf

Freund, Y., Schapire, R., (1996). *Experiments with a New Boosting Algorithm,* Retrieved 03 2017, AT&T Research: http://www.research.att.com/orgs/ssr/people/{yoav, schapire}.

Freund, Y., Schapire, R., (1997). A Decision – Theoretic Generalization of On – Line Learning and an Application to Boosting. *.Journal of Computer and System Sciences 55,* 119 – 139 (1997).

Frondel, M., Vance, C., (2011). Rarely Enjoyed? A Count Data Analysis of Ridership in Germany's Public Transport. *Transport Policy 18 (2): 425 – 433.*

Iseki, H., Ali, R., (2014). Fixed effects panel data analysis of gasoline prices, fare, service supply, and service frequency on transit ridership in ten U.S. urbanized areas. *54th Annual Conference of the Association of Collegiate School of Panning (ACSP) in Philadelphia,* October 2014.

Kellogg, R., Anderson, S., Sallee, J., Curtin, R. T., (2011). Forecasting Gasoline Prices Using Consumer Surveys. *American Economic Review,* 101(3): 110-14, May 2011.

Kondratenko, V.V., Kuperin, Y., (2003). Using Recurrent Neural Networks To Forecasting of Forex. *Disordered Systems and Neural Networks,* 2003.

Kulkarmi, S., Haidar, I., (2009). Forecasting Model for Crude Oil Price Using Artificial Neural Networks and Commodity Futures Prices. *International Journal of Computer Science and Information Security,* Vol. 2, No. 1, June 2009.

Lackes, R., Börgermann, C., Dirkmorfeld, M., (2009). Forecasting the Price Development of Crude Oil with Artificial Neural Networks. *International Work – Conference on Artificial Neural Networks,* IWANN 2009.

Liaggou, C., (2014). *"The reasons why fuels in Europe remain still expensive".* Retrieved 03 2017, from Kathimerini, 2014: http://www.kathimerini.gr/794913/article/oikonomia/ellhnikh-oikonomia/giati-ta-kaysima-sthn-ellada-paramenoyn-akriva

Liao, H., Cai, J. W., Yang, D., W., Wei, Y., M., (2016). Why did the historical energy forecasting succeed or fail? A case study on IEA's projection. *Technological Forecasting and Social Change 107,* April 2016.

Mahdiani, M. R., Khamehchi, E., (2017). A modified neural network model for predicting the crude oil price. *Intellectual Economics,* February 2017.

Mingers, J., (1989). An Empirical Comparison of Selection Measures for Decision – Tree Induction. *Machine Learning,* Volume 3, Issue 4, 319 – 342, 1989.

Minnaar, A., (2015). Implementing the DistBelief Deep Neural Network Training Framework with Akka. *Machine Learning at University College London.*

Mollasalehi, E., Khasian, M., Sadati, H., (2011). Forecasting Petroleum Price Using The Methods of Artificial Neural Networks. *Conference Paper : CANCAM,* June 2011.

Musso, A., Piccioni, C., Tozzi, M., Godard, G., Lapeyre, A., Papandreou, K., (2013). Road transport elasticity: how fuel price changes can affect traffic demand on a toll motorway. *Procedia – Social and Behavioral Sciences 2013,* 87:85 – 102.

Ng, A., (2013). CS229: Machine Learning. *Stanford University.*

Nielsen, M.A. (2016). *Neural Networks and Deep Learning. Determination Press*.

Nowak, W., Savage, I., (2013). The cross elasticity between gasoline prices and transit use: Evidence from Chicago. *Transport Policy29 (2013),* 38 – 45.

Nyongesa, D., Wagala, A., (2016). Non Linear Time Series Modelling Of the Diesel Prices in Kenya. *International Journal of Academic Research in Economics and Management Sciences,* 2016, Vol. 5, No. 4, ISSN: 2226 – 3624.

Olaru, C., Wehenkel, L., (2003). A complete fuzzy decision tree technique. *Fuzzy Sets and Systems 138,* 221 – 254, 2003.

Quinlan, J.R., (1987). Induction of Decision Trees. *International Journal of Man – Machine Studies,* 27, 221-234, 1987.

Rickwood, P., (2010, 12). The impact of rising oil prices on the transport sector. *Australian Planner 2010,* 47 (4-4): 243-252.

Rokach, L., Maimon, O., (2014). Data Mining and Decision Trees: Theory and Applications, 2nd Edition. *Series in Machine Perception and Artificial Intelligence,* Volume 81.

Rumelhart, D., Hinton, G., Williams, R., (1986). Learning Representations by Back Propagating Errors. *Nature,* 323 (6088): 533 – 536, 1986.

Shanker, M., Hung, M.S., (1996). Effect of data standardization on neural network training. *Omega 24(4),* 385-397.

Smart, J., (2014). A volatile relationship: The effect of changing gasoline prices on public support for mass transit. *Transportation Research Part A Policy and Practice,* 61: 178 – 185, March 2014.

Sontag, D., (2012). Introduction to Machine Learning, The AdaboostM1 algorithm. *New York University*.

Xiong, T., Bao, Y., Hu, Z., (2013). Beyond One – Step – Ahead Forecasting: Evaluation of Alternative Multi – Step – Ahead Forecasting Models for Crude Oil Prices. *Energy Economics,* 40, 2013: 405-415.

Yang, D., Timmermans, H., (2012). Effects of fuel price fluctuation on activity – travel behavior by transit and slow modes: Evidence from a pseudo panel data. *European Transport Conference Past Papers Repository.*