Interinstitutional Master of Science "Hydrocarbon Exploration and Exploitation"

Aristotle University of Thessaloniki - National Technical University of Athens - National and Kapodistrian University of Athens - Democritus University of Thrace




Aristotle University of Thessaloniki

Faculty of Science

Department of Geology

ΧΡΙΣΤΟΓΛΟΥ Γ. ΑΝΤΩΝΙΟΣ

Πτυχιούχος Γεωλόγος

"FORECASTING WEST TEXAS INTERMEDIATE (WTI) PRICE SPIKES WITH MACHINE LEARNING"

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων καθηγητής : Περικλής Γκόγκας

Θεσσαλονίκη, Οκτώβριος 2023

ΑΝΤΩΝΙΟΣ Γ. ΧΡΙΣΤΟΓΛΟΥ
Πτυχιούχος Γεωλόγος

FORECASTING WEST TEXAS INTERMEDIATE (WTI) PRICE SPIKES WITH MACHINE LEARNING

Υποβλήθηκε στο Τμήμα Γεωλογίας στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «Έρευνα και Εκμετάλλευση Υδρογονανθράκων»

Τριμελής Εξεταστική Επιτροπή

Καθηγητής Περικλής Γκόγκας, ΔΠΘ, Επιβλέπων
Ομότιμος Καθηγητής Ανδρέας Γεωργακόπουλος, ΑΠΘ Μέλος Τριμελούς Συμβουλευτικής Επιτροπής
Καθηγητής Δημήτριος Δαμίγος, ΕΜΠ, Μέλος Τριμελούς Συμβουλευτικής Επιτροπής

Citation:
Christoglou G. A., 2023. – Forecasting West Texas Intermediate (WTI) price spikes with Machine Learning, Master Thesis, School of Geology, Aristotle University of Thessaloniki.

Ευχαριστίες

Θα ήθελα να αφιερώσω την εργασία αυτή στην οικογένεια μου, για την διαρκή και αμέριστη συμπαράσταση που μου έχει δείξει σε όλη τη διάρκεια των σπουδών μου και συνεχίζει να το κάνει μέχρι και τώρα. Επιπρόσθετα να θα ήθελα να ευχαριστήσω την Χριστίνα για την συνολική της στήριξη όλα αυτά τα χρόνια.

Ακόμα θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Περικλή Γκόγκα και τον Υπ. Δρ. κ. Εμμανουήλ Σοφιανό για την καθοριστική βοήθειά τους και την ουσιαστική καθοδήγησή τους για την διεκπεραίωση αυτής της εργασίας. Οι πολύτιμες συμβουλές τους και η συνεχής επιστημονική υποστήριξή τους, σε συνδυασμό με την βαθιά γνώση του επιστημονικού αντικειμένου τους, συνετέλεσαν καθοριστικά για το άρτιο αποτέλεσμα αυτής της πτυχιακής εργασίας.

Τέλος νιώθω την ανάγκη να εκφράσω τη μεγάλη μου ευγνωμοσύνη και σε όλους τους καθηγητές του μεταπτυχιακού και ιδιαίτερα τον διευθυντή κ. Γεωργακόπουλο, για όλες τις γνώσεις και τις συμβουλές τους που μοιράστηκαν απλόχερα τα δύο αυτά χρόνια.

Πίνακας περιεχομένων

Περίληψη

Η ικανότητα πρόβλεψης της τιμής του αργού πετρελαίου (WTI crude oil) αποτελεί εργαλείο ύψιστης σημασίας για τους επενδυτές, τις κυβερνήσεις και τον εμπορικό – βιομηχανικό κλάδο. Έτσι, στην πτυχιακή αυτή επιχειρείται μια απόπειρα πρόβλεψης των ακραίων τιμών (spikes) του WTI με τη μέθοδο της Μηχανικής Μάθησης. Πιο συγκεκριμένα, χρησιμοποιούνται οι μεθοδολογίες SVM (με τους πυρήνες Linear, Cubic, Quadratic, Fine Gaussian, Medium Gaussian, Coarse Gaussian), Logistic Regression και Bagged Trees για ένα σετ ημερήσιων, εβδομαδιαίων και μηνιαίων δεδομένων. Τα δεδομένα προήλθαν από τα returns των spot prices του WTI Crude oil και καλύπτουν χρονολογικά το διάστημα από το 1986 έως το 2023. Τα όρια προσδιορισμού των spikes ορίστηκαν με τον υπολογισμό του αθροίσματος και της διαφοράς της τυπικής απόκλισης και του μέσου όρου για 30 ημέρες, 30 εβδομάδες και 24 μήνες αντίστοιχα για κάθε ομάδα δεδομένων. Ακολούθως το 90% των δεδομένων αποτέλεσε τα δεδομένα εκπαίδευσης ενώ το 10% αυτών χρησιμοποιήθηκαν ως δεδομένα ελέγχου. Το σύνολο των αποτελεσμάτων, δείχνει, ότι οι ακραίες τιμές του αργού πετρελαίου (WTI) προβλέπονται επιτυχώς σε ποσοστό 8% με το μοντέλο Bagged Trees, 29% με το μοντέλο Cubic SVM και 37% με το μοντέλο Cubic SVM, για τα ημερήσια, εβδομαδιαία και μηνιαία δεδομένα αντίστοιχα. Έτσι, διαφαίνεται ότι οι τιμές του πετρελαίου προκύπτουν από εξαιρετικά περίπλοκες διεργασίες και δεν είναι εύκολο να προβλεφθούν σε ικανοποιητικό βαθμό λαμβάνοντας υπόψη μόνο τις παρελθοντικές τιμές, ακόμα και με την χρήση των εργαλείων Μηχανικής Μάθησης.

Abstract

The ability to predict the price of WTI crude oil is of utmost importance to investors, governments and the commercial – industrial sector. Thus, in this thesis, an attempt is made to predict the extreme values (spikes) of WTI with the Machine Learning approach. More specifically, are used methodologies such as SVM (with the kernels of Linear, Cubic, Quadratic, Fine Gaussian, Medium Gaussian, Coarse Gaussian), Logistic Regression and Bagged Trees. for a set of daily, weekly and monthly data. The data was derived from the WTI Crude Oil spot price returns and covers a period from 1986 to 2023. The limits for the spikes were determined by calculating the sum and difference of the standard deviation and the mean over 30 days, 30 weeks and 24 months respectively for each data group. Subsequently, 90% of the data was used as training data, while 10% was used as test data. The set of results presented shows that the extreme crude oil prices (WTI) are successfully predicted by 8% with the Bagged Trees model, 29% with the Cubic SVM model and 37% with the Cubic SVM model for the daily, weekly and monthly data respectively. It thus, appears that values of WTI crude oil prices, arise from extremely complex processes and are not easy to be predicted satisfactorily by considering only past prices, even with the use of Machine Learning tool.

1. Εισαγωγή

Το West Texas Intermediate (WTI) crude oil συνδέεται άμεσα με την ανάπτυξη και την εξέλιξη της βιομηχανίας του πετρελαίου, τόσο στις Ηνωμένες Πολιτείες της Αμερικής, αλλά και σε ολόκληρο τον πλανήτη. Το πετρέλαιο αυτό είναι γνωστό και ως Texas light sweet oil και αποτελεί μια συγκεκριμένη κατηγορία πετρελαίου που λειτουργεί ως σημείο αναφοράς (benchmark) για την τιμολόγηση του πετρελαίου παγκοσμίως. Έχει διαδραματίσει καθοριστικό ρόλο στην διαμόρφωση της παγκόσμιας αγοράς ενέργειας και ιστορικά υπήρξε καθοριστικός παράγοντας για την οικονομική ανάπτυξη και την γεωπολιτική δυναμική των Ηνωμένων Πολιτειών της Αμερικής.

Οι ρίζες της προέλευσης του WTI φτάνουν πίσω στις αρχές του 20^{ου} αιώνα, όταν η πολιτεία του Τέξας αναδείχτηκε ως μια από τις σημαντικότερες πετρελαιοπαραγωγές περιοχές του πλανήτη. Η ανακάλυψη τεράστιων κοιτασμάτων πετρελαίου στο Περμιακό υπόβαθρο του Τέξας και του νότιο-ανατολικού Νέου Μεξικού σε συνδυασμό με την εξαιρετική του ποιότητα, το καθιέρωσαν ως ένα εξέχον προϊόν αργού πετρελαίου. Η αξία του αναδείχθηκε στην δεκαετία του 1970, όταν οι ΗΠΑ ήρθαν αντιμέτωπες με αλλεπάλληλες γεωπολιτικές εντάσεις και ενεργειακές κρίσεις οι οποίες επηρέασαν άρδην την ενεργειακή αγορά. Η διαχρονικά εξαιρετική ποιότητα του προϊόντος και η αξιοπιστία στην προσφορά του, καθιέρωσαν το WTI Crude Oil ως έναν αξιόπιστο δείκτη παγκόσμιων τιμών πετρελαίου. Καθώς η βιομηχανία του πετρελαίου επεκτάθηκε και η τεχνολογία προχωρούσε έγινε το βασικό σημείο αναφοράς για την τιμολόγηση του αργού πετρελαίου, τόσο στις ΗΠΑ, όσο και διεθνώς. Αναφορικά, άλλοι αντίστοιχοι δείκτες αργού πετρελαίου είναι: το Brend Crude Oil, το Urals Oil (μείγμα ρωσικού πετρελαίου), το Dubai Crude Oil, το Opec Baskett Price Oil και το Tapis Crude oil.

Η αναγκαιότητα και η σημασία του πετρελαίου ως ορυκτού καυσίμου και ως κινητήριας δύναμης του σύγχρονου κόσμου αντικατοπτρίζεται στην ανάγκη, που έχουν οι αγορές, οι επενδυτές και τα κράτη να προβλέψουν την τιμή του πετρελαίου, το οποίο αποτελούσε ανέκαθεν συστηματική προσπάθεια. Αντίθετα, στην ανθρώπινη κοινωνία και την οικονομία, η τιμή του αργού πετρέλαιο είναι άκρως ασταθής και η διατάραξη της αγοράς του πετρελαίου έχει πολύπλευρες επιπτώσεις στην παγκόσμια οικονομία. Χαρακτηριστικό παράδειγμα της αστάθεια της αγοράς του αργού πετρελαίου, είναι η πρόσφατη πανδημία του Covid-19, η οποία οδήγησε σε αρνητικές τιμές την πώληση

του αργού πετρελαίου. Η αστάθεια της αγοράς του αργού πετρελαίου και οι επιπτώσεις της στην παγκόσμια οικονομία έχουν εντείνει την ανησυχία των επενδυτών, των κρατών και κυβερνήσεων αλλά και των εταιριών. Η πρόβλεψη της τιμής του αργού πετρελαίου είναι εξαιρετικά δύσκολη λόγω της περίπλοκης, μη γραμμικής και χαοτικής φύσης του στην οικονομία (Zilin Xu, et al, 2023). Υπάρχουν πολλαπλές μεταβλητές, οι οποίες επηρεάζουν την τιμή του αργού πετρελαίου, όπως η τάση της οικονομίας, οι οικονομικοί κύκλοι, οι διεθνείς σχέσεις και η γεωπολιτική. Κατά συνέπεια, η πρόβλεψη της τιμής του αργού πετρελαίου είναι μια πολύπλοκη, αλλά ταυτόχρονα πολύτιμη προσπάθεια.

Μερικοί από τους λόγους, που καθιστούν αναγκαία την πρόβλεψη του WTI Crude oil είναι:

- Οι επενδυτικές και συναλλαγματικές αποφάσεις: Ακριβείς προβλέψεις της τιμής του αργού πετρελαίου, δίνουν την δυνατότητα στους επενδυτές και στους traders, να λάβουν σημαντικές αποφάσεις σχετικά με το αν θα πουλήσουν, θα κρατήσουν ή θα αγοράσουν μετοχές που σχετίζονται με την αγορά του πετρελαίου, όπως μελλοντικά συμβόλαια (futures contracts), ETFs (Exchange-Trades Funds) και μετοχές πετρελαϊκών εταιριών. Συνεπώς, η πρόβλεψη των τιμών του αργού πετρελαίου βοηθάει του επενδυτές να έχουν πρόσβαση σε πιθανά κέρδη και να περιορίσουν τον κίνδυνο που σχετίζεται με τέτοιας φύσεως επενδύσεις.
- Σταθερότητα στη διεθνή ενεργειακή αγορά: Η πρόβλεψη των τιμών του WTI Crude Oil, δεδομένου ότι αποτελεί ένα benchmark, συμβάλει στην συνολική σταθερότητα και διαφάνεια της παγκόσμιας αγοράς ενέργειας. Εφόσον, είναι ο πιο σημαντικός δείκτης αγοράς, η τιμή του δύναται να επιδράσει σε άλλα προϊόντα ενέργειας, επηρεάζοντας τις τάσεις της αγοράς και τις επενδυτικές αποφάσεις για το ευρύτερο ενεργειακό τοπίο.
- Επίδραση σε καταναλωτές: Η πρόβλεψη της τιμής του WTI μπορεί να έχει άμεση επίδραση και στους απλούς καταναλωτές, καθώς είναι αυτό που καθορίζει σε μεγάλο βαθμό την τιμή της βενζίνης, του πετρελαίου και άλλων παράγωγων προϊόντων του πετρελαίου. Επειδή, η τιμή του πετρελαίου δύναται να αυξάνεται και να μειώνεται, επηρεάζεται το κόστος των μεταφορών, το κόστος της θέρμανσης, ο βιομηχανικός κλάδος και καθορίζεται η καταναλωτική συμπεριφορά των ιδιωτών και των επιχειρήσεων.

- Ανάλυση αγορών ενέργειας: Η πρόβλεψη των τιμών, διευκολύνει την ανάλυση της ενεργειακής αγοράς δίνοντας πολύτιμες πληροφορίες σχετικά με την δυναμική της προσφοράς και της ζήτησης, τις τάσεις παραγωγής του πετρελαίου και τους γεωπολιτικούς παράγοντες, που επηρεάζουνε την τιμή του WTI Crude Oil. Η ιστορική ανάλυση των τιμών επιτρέπει στους αναλυτές να εντοπίζουν μοτίβα, ώστε να λαμβάνονται στρατηγικές αποφάσεις για την αντιμετώπιση επερχόμενων γεωπολιτικών κρίσεων, οικονομικών υφέσεων ή φυσικών καταστροφών, που έχουν αντίκτυπο στο πετρέλαιο.
- Διαχείριση κινδύνου: Η πρόβλεψη και η ανάλυση των τιμών του WTI Crude Oil, είναι ζωτικής σημασίας για τις στρατηγικές διαχείρισης κινδύνου, που εφαρμόζουν οι εταιρίες, που δραστηριοποιούνται με την βιομηχανία του πετρελαίου (upstream sector- εταιρίες παραγωγής, midstream / downstream sector - εταιρίες διύλισης και μεταφοράς). Εντοπίζοντας και κατανοώντας τις πιθανές διακυμάνσεις των τιμών, οι εταιρίες μπορούν να εφαρμόσουν στρατηγικές αντιστάθμισης για να μετριάσουν τους χρηματοοικονομικούς κινδύνους, που προκύπτουν από ξαφνικές μεταβολές των τιμών.

Η σχέση των τιμών του αργού πετρελαίου και από ποιους παράγοντες αυτή καθορίζεται είναι περίπλοκη και έχουν γίνει αρκετές προσπάθειες για να αποτυπωθεί με επιτυχία. Μια από τις πρώτες απόπειρες να περιγράψουν την σχέση ανάμεσα στο πως καθορίζονται οι τιμές του πετρελαίου ήταν αυτή του H. Hotelling, ο οποίος υποστήριξε ότι υπάρχει απευθείας σύνδεση μεταξύ των τιμών του πετρελαίου και της εκάστοτε εφαρμόζουσας νομισματικής πολιτικής (H. Hotelling, 1931). Ο κανόνας αυτός είναι γνωστός ευρέως ως “Hotelling rule”. Ο Hotelling υποστήριξε ότι οι τιμές είναι απόρροια των επιτοκίων που έχουν καθοριστεί. Ωστόσο, η θεωρία αυτή βρίσκεται κάτω από περαιτέρω διερεύνηση (Dimitriadou A., et al, 2018). Περαιτέρω, φαίνεται να υπάρχει μια θετική και ισχυρή σχέση μεταξύ των διακυμάνσεων των οικονομικών κύκλων των επιχειρήσεων και της τιμής του αργού πετρελαίου με την ύπαρξη μιας αμφίδρομης εξάρτησης της οικονομικής κατάστασης και των τιμών στην αγορά (J. D. Hamilton, 1983). Όπως σημειώνεται όμως, η μελέτη αυτή αναφέρεται σε περιόδους ανάπτυξης αφήνοντας ερωτηματικά για τις περιόδους ύφεσης στην οικονομία (Dimitriadou A., et al, 2018).

Στα πλαίσια της μηχανικής μάθησης, τα spikes αντικατοπτρίζουν ξαφνικές απότομες αυξήσεις ή διακυμάνσεις στα εκάστοτε δεδομένα που εξετάζονται. Η

σημαντικότητά τους έγκειται στο γεγονός ότι δίνουν πληροφορίες για το σύνολο των δεδομένων που εκπαιδεύεται.

Στην παρούσα διπλωματική εργασία θα γίνει προσπάθεια πρόβλεψης των ακραίων τιμών που καταγράφονται με την μορφή των spikes και προκύπτουν από τα returns του WTI Crude Oil με τη χρήση μοντέλων Machine Learning. Τα αποτελέσματα, που προκύπτουν από την πρόβλεψη των spikes, μπορεί να διαφέρουν ως προς τη φύση τους και τα χαρακτηριστικά τους, ανάλογα με την βιομηχανία, που εφαρμόζονται ή τις ειδικές μεταβλητές, που λαμβάνονται υπόψη και ως εκ τούτου είναι σημαντικό να αναλυθούν προσεκτικά και να συμπεριληφθούν οι σωστοί παράγοντες και τα κατάλληλα μοντέλα για να μετριαστεί το αντίκτυπο των spikes και να βελτιωθεί η ακρίβεια στην πρόβλεψη τους. Ο στόχος της διπλωματικής αυτής, είναι η δημιουργία ενός αξιόπιστου μοντέλου πρόγνωσης των spikes τα οποία προέκυψαν από τις τιμές των returns του Spot price του WTI crude oil με την εφαρμογή μεθόδων μηχανικής μάθησης (machine learning) μετά από την κατάλληλη επεξεργασία των δεδομένων. Αφού τα δεδομένα επεξεργάστηκαν, στην συνέχεια τροφοδοτήθηκαν σε διάφορα μοντέλα πρόγνωσης μέσω μιας διαδικασίας εκπαίδευσης και ελέγχου (training–testing). Η συχνότητα της περιόδου για τη πρόβλεψη των μοντέλων έγινε για ημερήσια, εβδομαδιαία και μηνιαία δεδομένα, ενώ το εύρος των τιμών καλύπτει το διάστημα από το 1986 έως το 2023.

Τα spikes αναπαριστούν γραφικά τιμές μιας χρονοσειράς, η οποία εμφανίζει απότομα μια έξαρση ή ένα απότομο βύθισμα με τη πάροδο του χρόνου. Με άλλα λόγια η συγκριμένη τιμή της χρονοσειράς αποκλίνει σημαντικά από το σύνολο των υπολοίπων τιμών. Υπολογίζοντας ένα ανώτατο και ένα κατώτατο όριο, οι τιμές που αποκλίνουν από τα όρια αυτά χαρακτηρίζονται ως spikes ενώ οι τιμές εντός των ορίων χαρακτηρίζονται ως non-spikes. Τα spikes στην εργασία αυτή, υπολογίστηκαν για τα returns, δηλαδή την διαφορά των λογαριθμικών τιμών του WTI Crude Oil για δύο διαδοχικές τιμές, ενώ παράλληλα τα δύο όρια, το ανώτατο και το κατώτατο καθορίστηκαν με τον υπολογισμό της τυπικής απόκλισης για κάθε σετ δεδομένων.

Για τα ημερήσια δεδομένα, αρχικά υπολογίστηκε η τυπική απόκλιση για 5 ημέρες (σ5), για 10 ημέρες (σ10), για τριάντα ημέρες (σ30) και ο μέσος όρος για τριάντα ημέρες (SMA30). Ακολούθως το άνω όριο (S+) υπολογίστηκε από το άθροισμα του (σ30) και του (avr30), ενώ το κάτω όριο (S-) από τη διαφορά του (avr30)

και του (σ30). Όσο αναφορά τα εβδομαδιαία δεδομένα, αρχικά υπολογίστηκε ο μέσος όρος για τέσσερις εβδομάδες (SMA4) και ο εκθετικός μέσος όρος για τέσσερις εβδομάδες (EMA4), ο μέσος όρος για 52 εβδομάδες (SMA52), ο εκθετικός μέσος όρος για 52 εβδομάδες (EMA52) και η τυπική απόκλιση για τριάντα εβδομάδες (σ30). Ακολούθως, το άνω όριο (S+) υπολογίστηκε από το άθροισμα του (σ30) και του (avr30), ενώ το κάτω όριο (S-) από τη διαφορά του (avr30) και του (σ30). Για τα μηνιαία δεδομένα, υπολογίστηκε ο μέσος όρος για 6 μήνες (SMA6), ο εκθετικός μέσος για έξι μήνες (EMA6), ο μέσος όρος για 12 μήνες (SMA12), ο εκθετικός μέσος όρος για 12 μήνες (EMA12) και η τυπική απόκλιση για 24 μήνες (σ24). Ακολούθως, το άνω όριο (S+) υπολογίστηκε από το άθροισμα του (σ24) και του (avr24), ενώ το κάτω όριο (S-) από τη διαφορά του (avr24) και του (σ24).

Ακολούθως, αν η τιμή στην οποία γίνεται η πρόβλεψη ξεπεράσει το ανώτατο ή το κατώτατο όριο που έχουν υπολογιστεί, τότε η τιμή αυτή χαρακτηρίζεται ως spike, ενώ στην περίπτωση που η πρόβλεψη είναι εντός των υπολογισμένων ορίων, αυτή χαρακτηρίζεται ως non-spike. Η μέθοδος των χαρακτηρισμών των spikes έχει χρησιμοποιηθεί για διάφορες προβλέψεις σε διάφορες περιπτώσεις, όπως στη πρόβλεψη spikes του S&P500 (Th Papadimitriou et al, 2020), την πρόβλεψη spikes για το Bitcoin (Th Papadimitriou G. Periklis et al, 2022), σε τιμές spikes της αγοράς του ηλεκτρικού ρεύματος (E Stathakis et al, 2017) και σε πολλά άλλα.

1.1 Βιβλιογραφική Ανασκόπηση

Η μηχανική μάθηση (ML - Machine Learning) είναι ένα ανερχόμενο εργαλείο, το οποίο έχει τη δυνατότητα να εκπαιδεύεται σε συγκεκριμένα μοτίβα και να λαμβάνει στη συνέχεια αποφάσεις βασισμένες στα μοτίβα στα οποία εκπαιδεύτηκε. Είναι αδιαμφισβήτητα ένας πολλά υποσχόμενος τομέας και έχει ευρεία εφαρμογή σε τεράστιο φάσμα αντικειμένων, συμπεριλαμβανομένου της ταξινόμησης της ακολουθίας του DNA, την επεξεργασία εικόνων, την αναγνώριση προσώπων, στις προβλέψεις τιμών, αλγορίθμων και πολλών άλλων (Edeh Michael Onyema et al, 2022). Η πρόγνωση τιμών με την βοήθεια του Machine Learning έχει εφαρμοστεί σε διάφορες μελέτες ανά καιρούς και με διαφορετικές τεχνικές.

Οι πιο κλασσικές τεχνικές της πρόβλεψης των μελλοντικών τιμών ενεργειακών προϊόντων μπορούν να κατηγοριοποιηθούν σε τρεις επιμέρους κατηγορίες: 1) Τα οικονομετρικά μοντέλα, 2) τα μοντέλα βάση της τεχνητής νοημοσύνης (Artificial

Intelligent - AI) και 3) τα υβριδικά μοντέλα, που συνδυάζουν τις δύο προηγούμενες μεθόδους (Zilin Xu, et al, 2023). Σε γενικές γραμμές τα οικονομετρικά μοντέλα έχουν χαμηλές επιδόσεις σε μη γραμμικά δεδομένα, ανεξάρτητα από την καλή τους επίδοση σε γραμμικά και δεδομένα χρονοσειράς (Adebayo et al., 2021). Για πιο περίπλοκες χρονοσειρές προτιμώνται τεχνικές, που συμπεριλαμβάνουν μεθόδους τεχνητής νοημοσύνης (AI) αντί των οικονομετρικών μοντέλων, λόγω της ικανότητας τους να αποτυπώνουν καλύτερα τα μη γραμμικά δεδομένα (Zilin Xu, et al, 2023).

Για την πρόβλεψη τιμών του WTI, οι (Dimitriadou A., et al, 2018) εξετάζουν τις σχέσεις των τιμών του πετρελαίου με άλλες οικονομικές μεταβλητές σε μηνιαία βάση. Λαμβάνουν 38 μεταβλητές ανάλογα με την οικονομική θεωρία και τελικά επιλέγουν αυτές που εμφανίζουν την μεγαλύτερη συσχέτιση με την πρόβλεψη των τιμών του WTI. Χρησιμοποιούν μοντέλα Support Vector Machine (SVM) σε συνδυασμό με γραμμικό Kernel και τον μη γραμμικό Radial Basis Function (RBF) και εξετάζουν την απόδοση της πρόβλεψης των μοντέλων αυτών σε σχέση με τη μέθοδο του logistic regression. Οι (Zilin Xu, et al, 2023) χρησιμοποιούν χρονοσειρές και τη μέθοδο logistic regression, για να αξιολογήσουν την πρόβλεψη των τιμών που δίνουν τα μοντέλα για ημερήσιες τιμές του WTI crude oil (Adebayo et al., 2021). Οι (Wang J et al, 2018) έδειξαν, ότι μπορεί να γίνει πρόβλεψη των μηνιαίων τιμών του αργού πετρελαίου για μια συγκεκριμένη περίοδο με τη χρήση extreme learning machine, ενώ η χρήση μεθόδων linear regression στο ίδιο εύρος δείγματος έδωσε χαμηλά αποτελέσματα πρόβλεψης των τιμών. Οι (Zhao Y et al, 2017) συνδύασαν μεθόδους Deep Learning για να προβλέψουν τις μηνιαίες τιμές του WTI crude oil για μια περίοδο από το 1986 έως το 2016 λαμβάνοντας υπόψη ένα εύρος από 193 πιθανές μεταβλητές. Η ιδέα τους, ήταν να επιλέξουν, όχι μόνο τις πιο χρήσιμες μεταβλητές, αλλά να εφαρμόσουν και διαφορετικούς συντελεστές βαρύτητας σε κάθε μεταβλητή με στόχο την βελτίωση του συνόλου της πρόβλεψης αποδεικνύοντας, ότι οι εφαρμογές του Machine Learning υπερισχύουν αυτών των συμβατικών οικονομετρικών μεθοδολογιών. Οι (Yu L et al, 2008) εφάρμοσαν τεχνικές (EMD) empirical mode decomposition ως αρχικό βήμα για την εφαρμογή μοντέλων νευρονικών δικτύων (neural network) για την πρόβλεψη καθημερινών τιμών του πετρελαίου WTI και του Brent. Κατέληξαν στο συμπέρασμα ότι μεταξύ των διαφόρων μοντέλων πρόβλεψης που χρησιμοποιήθηκαν, τα βασισμένα σε EMD μοντέλα των νευρωνικών δικτιών αποδίδουν καλύτερα, ενώ στο σύνολο των δοκιμών, το μοντέλο RMSE (root mean

squared error) είχε τη χαμηλότερη απόδοση και το Dstat (Directional statistc) την υψηλότερη.

Εκτός από την πρόβλεψη των τιμών του WTI crude oil, στην ευρύτερη βιβλιογραφία η μέθοδος του ML εφαρμόζεται και σε πλήθος άλλων τομέων. Οι (Th Papadimitriou et al, 2020) επιχειρούν να προβλέψουν τις ακραίες μεταβολές στις τιμές των returns του S&P500 με την μέθοδο του Support Vector Machine και τη χρήση πυρήνων Kernel. Τα καλύτερα αποτελέσματα προέκυψαν με μοντέλα που χρησιμοποίησαν τον μη γραμμικό RBF Kernel σε συνδυασμό με 2 lags των returns του S&P500 και τα returns του NYSE Composite index. Οι (Enke D et al, 2005) χρησιμοποίησαν Artificial Neural Networks (ANN), για να προβλέψουν τα returns του S&P500 με επιπλέον 31 επεξηγηματικές μεταβλητές. Οι (E Stathakis et al, 2017) πρότειναν την εφαρμογή του SVM με πολλαπλές κλάσεις για την πρόβλεψη των spikes στην αγορά ενέργειας του ηλεκτρικού ρεύματος προβλέποντας τα θετικά και τα αρνητικά spikes των τιμών μαζί με τις κανονικές τιμές του ηλεκτρικού ρεύματος της Γερμανικής αγοράς. Σε γενικές γραμμές το SVM προσφέρει μια ικανοποιητική ακρίβεια για την πρόβλεψη των θετικών και αρνητικών spikes. Τέλος οι (P. Gogas et al, 2021) με τη χρήση πολλών μοντέλων ML όπως SVM, decision trees (DT), random forests (RF) και του μοντέλου elastic-net logistic regression (logit), επιχειρούν να προβλέψουν την κατεύθυνση της τιμής της ανεργίας στην Ευρωπαϊκή Ένωση.

Συνοψίζοντας μερικά από τα οικονομετρικά μοντέλα που εφαρμόζονται είναι το Autoregressive Integrated Moving Average (ARIMA), το Vector Autoregressive Model και το Generalized Auto-regressive Conditional Heteroskedasticity (GARCH). Γενικά τα μοντέλα αυτά, όπως προαναφέρθηκε εμφανίζουν μέτρια απόδοση σε μη γραμμικά δεδομένα, παρά την καλή απόδοση σε γραμμικά (Zilin Xu, et al, 2023). Η πιο διαδεδομένη μέθοδος για τη πρόβλεψη των τιμών των ενεργειακών προϊόντων με τεχνητή νοημοσύνη είναι το Support Vector Machine (SVM) (Zhang D et al, 2021). Ορισμένοι ερευνητές όσο αναφορά την πρόβλεψη τιμών, χρησιμοποιούν το SVM για μη γραμμικά δεδομένα, καθώς είναι πιο ευέλικτο και προσαρμόζεται εύκολα στην εκάστοτε περίσταση. Σε πιο πρόσφατες μελέτες, εφαρμόζονται υβριδικά μοντέλα τα οποία εφαρμόζουν μικτές μεθόδους για το καλύτερο δυνατό αποτέλεσμα. Τα μοντέλα τα οποία χρησιμοποιούν χρονοσειρές, εφαρμόζονται ευρέως για την πρόβλεψη τιμών του πετρελαίου, όμως παράλληλα εμφανίζουν πολλά εγγενή προβλήματα. Το σημαντικότερο είναι, ότι οι τιμές του πετρελαίου δεν εμφανίζουν σταθερότητα με τον

χρόνο, αλλά εμφανίζουν κυκλικότητα και αστάθεια με επακόλουθο να μην είναι εύκολη η αξιόπιστη πρόβλεψη με κριτήριο μόνο τις τιμές του προϊόντος αυτού για μελλοντικές τιμές. (Chang L et al, 2022). Άλλη τεχνική στη πρόβλεψη των μελλοντικών τιμών πετρελαίου είναι η χρήση των νευρωνικών δικτύων (neural networks – NNET).

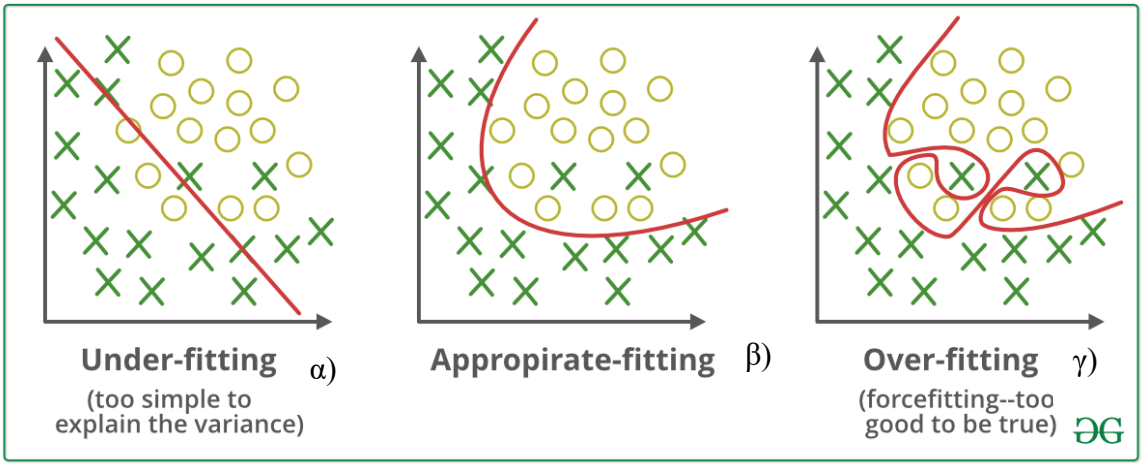
Βάσει όλων των ανωτέρω συνάγεται, ότι η χρήση της μηχανικής μάθησης για τη πρόβλεψη τιμών πετρελαίου, αλλά και η χρήση της μηχανικής μάθησης για γενικότερα χρηματιστηριακά προϊόντα και οικονομικούς δείκτες, αποτελεί έναν τομέα που προσελκύει το ενδιαφέρον της επιστημονικής κοινότητας. Παρόλα αυτά, το πλήθος των μεθόδων και των τεχνικών, που συνηθίζεται να χρησιμοποιούνται, τις περισσότερες φορές δεν προσεγγίζουν αποκλειστικά το ζήτημα της πρόβλεψης για την εμφάνιση των spikes βάση των returns του WTI crude oil αποκλειστικά με τις μεθόδους, που χρησιμοποιήθηκαν στην συγκεκριμένη εργασία. Για αυτό τον λόγο, αξίζει να εξεταστεί αν η χρήση μοντέλων SVM, του Logistic regression και των Bagged trees στις τιμές των returns του WTI Crude oil μπορεί να προβλέψει σωστά την εμφάνιση των spikes.

Μηχανική Μάθηση και Οικονομετρικές μέθοδοι πρόβλεψης

Στις κλασσικές – συμβατικές μεθόδους πρόβλεψης, ιδίως στα μοντέλα πρόβλεψης, που χρησιμοποιούνται, είναι σύνηθες να εμφανίζονται περισσότερα λάθη κατά την πρόβλεψη. Επιπροσθέτως, αρκετές φορές προκύπτουν προβλήματα σχετικά με τις υποθέσεις και τις παραδοχές που απαιτείται να γίνουν, οι οποίες οδηγούν συχνά σε υποκειμενικά συμπεράσματα. Αναμφίβολα, όσο αναφορά τις συμβατικές μεθόδους πρόβλεψης, οι υπολογιστικές απαιτήσεις είναι πολύ μικρότερες σε σχέση με τις υπολογιστικές απαιτήσεις της μηχανικής μάθησης.

Στη μηχανική μάθηση, ένα μοντέλο χαρακτηρίζεται, ότι κάνει Underfitting, όταν το μοντέλο είναι πολύ απλοϊκό για να καταγράψει και να απεικονίσει σωστά την πολυπλοκότητα των δεδομένων. Αντιπροσωπεύει την αδυναμία του μοντέλου να μάθει αποτελεσματικά τα δεδομένα εκπαίδευσης, με αποτέλεσμα τη κακή απόδοση, τόσο στα δεδομένα εκπαίδευσης, όσο και στα δεδομένα ελέγχου. Έτσι, τα μοντέλα αυτά χαρακτηρίζονται ως ανακριβή και εντοπίζονται στις περιπτώσεις, όπου υπερβολικά απλοϊκά μοντέλα επιλέγονται με απλοποιημένες υποθέσεις. Για να αντιμετωπιστεί το φαινόμενο αυτό πρέπει να επιλέξουμε πιο σύνθετα μοντέλα με λιγότερη τακτοποίηση

των δεδομένων. Η αναπαράσταση ενός μοντέλου που παρουσιάζει Underfitting φαίνεται στην εικόνα 1α. (https://www.geeksforgeeks.org/).

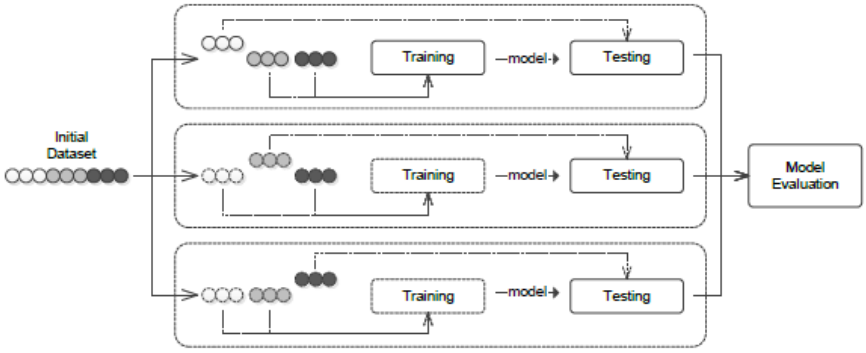


Εικόνα 1. α) Ένα μοντέλο αναφέρεται ότι εμφανίζει Underfitting όταν είναι αρκετά απλό για να περιγράψει και να αποδώσει σωστά τις σχέσεις των δεδομένων, β) μοντέλο το οποίοι αποδίδει σωστά τις σχέσεις που εμφανίζουν οι μεταβλητές του, γ) ένα μοντέλο αναφέρεται ότι εμφανίζει Overfitting όταν εκπαιδεύεται πολύ καλά στα δεδομένα ελέγχου και καταλήγει να περιγράφει αυτά αντί για το φαινόμενο οπότε αδυνατεί να δώσει καλή δυνατότητα γενίκευσης στα άγνωστα δεδομένα. (https://www.geeksforgeeks.org/).

Ένα μοντέλο χαρακτηρίζεται ως Overfitted όταν αυτό εμφανίζεται πολύ καλά προσαρμοσμένο στα δεδομένα εκπαίδευσης αλλά αδυνατεί να περιγράψει την πραγματική φύση των δεδομένων ελέγχου. Η υπερβολική εκπαίδευση στα δεδομένα ελέγχου έχει ως αποτέλεσμα το εκάστοτε μοντέλο να περιγράφει αυτά, και όχι το εκάστοτε φαινόμενο. Το αποτέλεσμα είναι το μοντέλο να μην εμφανίζει καλή δυνατότητα γενίκευσης σε άγνωστα προς αυτό δεδομένα. Όταν το μοντέλο μαθαίνει τις λεπτομέρειες και τον θόρυβο στα δεδομένα εκπαίδευσης, το γεγονός αυτό έχει αρνητική επίδραση στην απόδοσή του κατά την χρήση νέων δεδομένων. Ο θόρυβος και οι τυχαίες μεταβολές προσλαμβάνονται και μαθαίνονται σαν έννοιες από το εκάστοτε μοντέλο. Το φαινόμενο αυτό εμφανίζεται σε μη παραμετρικές και μη γραμμικές μεθόδους της μηχανικής μάθησης, καθώς αυτές οι μέθοδοι διαθέτουν μεγαλύτερη ελευθερία στο να δημιουργούν τα μοντέλα βάση του συνόλου των δεδομένων και έτσι μπορούν να δημιουργήσουν μη ρεαλιστικά μοντέλα. Η αναπαράσταση ενός τέτοιου μοντέλου φαίνεται στην εικόνα 1γ.

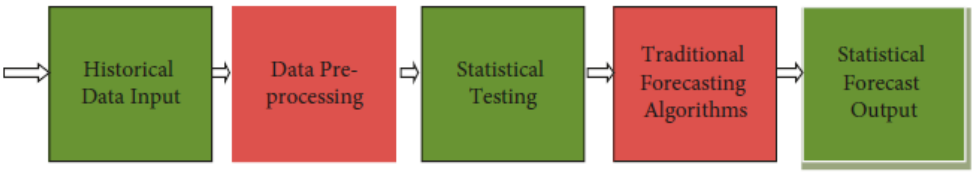
Ένας τρόπος αποφυγής του Overfitting, είναι η χρήση ενός γραμμικού αλγορίθμου για γραμμικά δεδομένα ή η χρήση παραμέτρων όπως το maximal depth στην περίπτωση, που χρησιμοποιούμε δέντρα αποφάσεων (https://www.geeksforgeeks.org/). Επιπρόσθετα, στο στάδιο της εκπαίδευσης,

χρησιμοποιείται η μέθοδος k-fold cross-validation. Κατά την μεθοδολογία αυτή, τα δεδομένα εκπαίδευσης διαχωρίζονται σε k μέρη. Ακολούθως, γίνεται μια αρχική διαμόρφωση (initial configuration) και το μοντέλο εκπαιδεύεται k φορές σε μέρη k-1, κρατώντας κάθε φορά ένα κομμάτι των δεδομένων του για δοκιμή (testing). Μετά την αλλαγή των παραμέτρων του μοντέλου, εκτελείται μια επαναληπτική μέθοδος εκπαίδευσης μέχρι να βρεθεί το σύνολο των παραμέτρων που ελαχιστοποιεί το σφάλμα πρόβλεψης (E Stathakis et al, 2017). Η παραπάνω διαδικασία φαίνεται στην εικόνα 2.



Εικόνα 2. Σχηματική αναπαράσταση της μεθοδολογίας 3-fold cross validation για την αντιμετώπιση του φαινομένου του Overfitting (E. Stathakis et al, 2020).

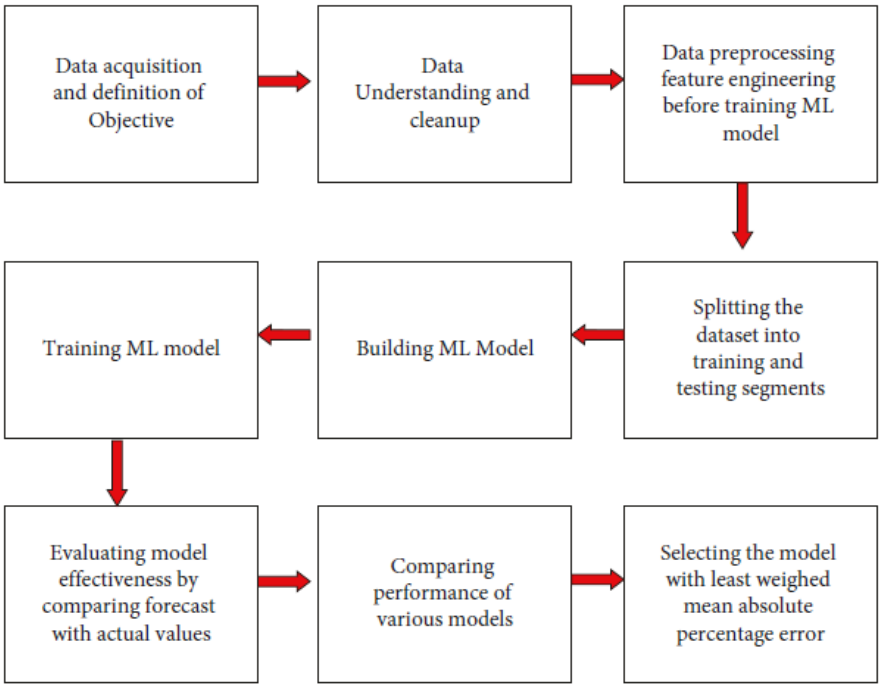
Μερικά ακόμα χαρακτηριστικά που εμφανίζουν οι κλασσικές μέθοδοι πρόβλεψής είναι, ότι λόγου περιορισμένων υπολογιστικών δυνατοτήτων, οι σχέσεις μεταξύ των μεταβλητών πρέπει να είναι συχνά απλουστευμένες, δουλεύουν με περιορισμένο αριθμό ιστορικών δεδομένων και είναι ιδανικές για εφαρμογές με μονές μεταβλητές οι οποίες στοχεύουν στο να αξιολογήσουν ή να συνοψίσουν τα δεδομένα. (Edeh Michael Onyema et al, 2022). Η μεθοδολογία της εφαρμογής μιας τέτοιας κλασσικής μεθόδου αναπαρίσταται στην εικόνα 3.



Εικόνα 3. Η μεθοδολογία της παραδοσιακής μεθόδου πρόβλεψης (Edeh Michael Onyema et al, 2022).

Η χρήση της Μηχανικής Μάθησης, προσφέρει πιο ακριβείς προβλέψεις ακόμα και με πιο περιορισμένη χρήση μεταβλητών, που χρησιμοποιούνται στα ζητήματα που καλείται να λύσει. Επιπρόσθετα είναι πιο αποτελεσματική στο να αποφεύγει να κάνει underfitting και overfitting τα δεδομένα με τη χρήση διαφόρων εργαλείων και μεθόδων, ενώ παράλληλα απαιτεί μεγαλύτερη υπολογιστική δύναμη για την εφαρμογή της.

Επιπλέον, προσανατολίζεται στο αποτέλεσμα που εξάγει και όχι τόσο στη μεταξύ σχέση, που έχουν οι μεταβλητές που χρησιμοποιεί. Η μεθοδολογία των βημάτων, που ακολουθούνται από την αρχή ως το τέλος για τη χρήση ενός μοντέλου στη μηχανική μάθηση αναφέρονται στην Εικόνα 4 (Genpact, 2022). Η χρήση της είναι ιδανική για εφαρμογές, όπου ο στόχος είναι τα εξαχθεί ένα αποτέλεσμα από μια μεγάλη βάση δεδομένων με πολλά διαφορετικά χαρακτηριστικά, ενώ έχει τη δυνατότητα να διαχειρίζεται τεράστιες βάσεις δεδομένων (Edeh Michael Onyema et al, 2022).



Εικόνα 4. Προσέγγιση μεθόδου Μηχανικής Μάθησης όπου αρχικά συλλέγονται τα δεδομένα και ορίζεται ο στόχος της μελέτης. Έπειτα ακολουθεί η κατανόηση και ο καθαρισμό των δεδομένων, η επεξεργασία τους, ο διαχωρισμός τους σε δεδομένα εκπαίδευσης και ελέγχου, η κατασκευή του μοντέλου MLκαι η εκπαίδευση, η αξιολόγηση της αποδοτικότητας του μοντέλου με τα δεδομένα ελέγχου, η σύγκριση του μοντέλου με άλλα μοντέλα και η τελική επιλογή του μοντέλου. (Genpact, 2022)

Η μηχανική μάθηση ως εργαλείο αντιμετωπίζει παράλληλα και ορισμένες προκλήσεις εμφανίζοντας ορισμένες αδυναμίες σε ορισμένες περιπτώσεις. Όπως αναφέρουν ορισμένοι ερευνητές, η μηχανική μάθηση δεν χρησιμοποιεί αποκλειστικά τα δεδομένα εκπαίδευσης για να κάνει τις προβλέψεις της, αλλά μαθαίνει κατά την πορεία της διαδικασίας και καταλαβαίνει τα μοτίβα τα οποία μπορεί να μην υπήρχαν νωρίτερα στα in sample δεδομένα, αλλά αποκτήθηκαν κατά την πορεία της εκπαίδευσης του μοντέλου. Έτσι, δημιουργούνται ζητήματα, όπως η τάση της Μηχανικής Μάθησης σε πιθανές παρερμηνείες. Η μέθοδος, αν και αυτόνομη, είναι άκρως επιρρεπής, τόσο σε «μηχανικά» λάθη, όσο και σε ανθρώπινα. Για παράδειγμα, ένα συνηθισμένο ανθρώπινο λάθος εμφανίζεται, όταν δεν εκπαιδεύεται στη σωστή

αναλογία δεδομένων εκπαίδευσης και ελέγχου και προκύπτουν βεβιασμένα και παραπλανητικά αποτελέσματα, τα οποία μπορεί να οδηγήσουν σε γενικεύσεις. Τα λάθη αυτά σε κάποιες ακραίες περιπτώσεις είναι αρκετά δύσκολο να εντοπιστούν και να διορθωθούν, καθώς απαιτούν περίπλοκους αλγορίθμους και διαδικασίες (Edeh Michael Onyema et al, 2022). Ένα ακόμα μειονέκτημα είναι ο παράγοντας του χρόνου, καθώς όσο περισσότερα είναι τα δεδομένα που χρειάζεται να εκπαιδευτούν, τόσο περισσότερος χρόνος χρειάζεται από τον εκάστοτε αλγόριθμο της μηχανικής μάθησης να λειτουργήσει σωστά.

Παρατηρείται, ότι η μηχανική μάθηση (ML) σε γενικές γραμμές δημιουργεί καλύτερα αποτελέσματα με μεγαλύτερη ακρίβεια σε σχέση με τις κλασσικές οικονομετρικές μεθόδους. Στην εργασία αυτή, επιχειρείται να γίνει η πρόβλεψη εμφάνισης των spikes που προκύπτουν από τις ακραίες τιμές των returns του WTI crude oil για ημερήσιες, εβδομαδιαίες και μηνιαίες τιμές χρησιμοποιώντας μοντέλα μηχανικής μάθησης, με σκοπό να βρεθεί ποιο από αυτά επιτυγχάνει τη καλύτερη δυνατή πρόβλεψη των spikes. Τα μοντέλα που χρησιμοποιήθηκαν για την εκπαίδευση στα πλαίσια της εργασίας αυτής είναι το Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Logistic Regression και Bagged Trees.

2. Μεθοδολογία και Δεδομένα

2.1 Δεδομένα και επεξεργασία

Για την εκπαίδευση και τον έλεγχο των μοντέλων συγκεντρώθηκαν δεδομένα από την ιστοσελίδα eia (U.S. Energy Information Administration). Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι τιμές των Spot Prices για το Cushing, OK WTI FOB (Dollar per Barrel). Χρησιμοποιήθηκαν τρία διαφορετικά σετ δεδομένων, οι ημερήσιες τιμές, οι εβδομαδιαίες τιμές και οι μηνιαίες τιμές. Στον παρακάτω πίνακα 2.1 αναφέρονται οι αρχικές χρονολογίες και οι τελικές χρονολογίες των δεδομένων.

Πίνακας 2.1 Χρονολογίες αρχής και τέλους των δεδομένων.

Σετ δεδομένων	Από	Έως
Ημερήσια	03 Ιανουαρίου 1986	09 Ιανουαρίου 2023
Εβδομαδιαία	03 Ιανουαρίου 1986	26 Μαΐου 2023
Μηνιαία	Ιανουάριος 1986	Μάιος 2023

Για κάθε σετ δεδομένων χρησιμοποιήθηκαν τα returns των ημερήσιων, εβδομαδιαίων και μηνιαίων τιμών του spot price και χρησιμοποιήθηκαν Lags. Τα δεδομένα ακόμα χωρίστηκαν σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου (training–test data) με αναλογία 90% και 10% αντίστοιχα. Τα returns προκύπτουν από τη διαφορά του λογαρίθμου της τρέχουσας τιμής μιας μετοχής με τον λογάριθμο της προηγούμενης τιμής της.

2.1.1 Ημερήσια δεδομένα

Για την επεξεργασία των ημερήσιων δεδομένων χρησιμοποιήθηκαν τα returns (r) των τιμών του WTI spot price, τα οποία προκύπτουν από την διαφορά των λογαριθμικών τιμών δύο διαδοχικών ημερών. Ακόμα, χρησιμοποιήθηκαν 10 Lags των returns (L1 – L10). Στην συνέχεια, στα δεδομένα εφαρμόστηκε ο Κινητός Μέσος Όρος (Simple Moving Average) για 5 μέρες και 10 μέρες (SMA5, SMA10). Η εφαρμογή αυτή είναι ένα απλό εργαλείο τεχνικής ανάλυσης που εξομαλύνει τα δεδομένα των τιμών δημιουργώντας μια ενημερωμένη μέση τιμή για την χρονική περίοδο που έχει οριστεί. Ακόμα εφαρμόστηκε στα δεδομένα ο εκθετικός κινητός μέσος όρος (exponential moving average EMA), ο οποίος είναι ένας τύπος κινητών μέσων όρων, αλλά δίνει μεγαλύτερη βαρύτητα σε πρόσφατες τιμές δεδομένων. Εφαρμόστηκε για 5 μέρες και 10 μέρες αντίστοιχα (EMA5 και EMA10). Στην συνέχεια εφαρμόστηκε η τυπική απόκλιση (standard deviation) για 5 μέρες, 10 μέρες και 30 ημέρες. Η τυπική

απόκλιση δείχνει πόσο απλωμένα είναι τα δεδομένα σε σχέση με το μέσο όρο των δεδομένων. Μια μικρή τυπική απόκλιση υποδηλώνει, ότι τα δεδομένα συγκεντρώνονται στενά γύρω από τη μέση τιμή ενώ μια υψηλή τυπική απόκλιση υποδηλώνει πιο απλωμένα δεδομένα. Τέλος, χρησιμοποιήθηκε ο μέσος όρος (average) για χρονικό όριο των 30 ημερών στις τιμές των returns. Ο υπολογισμός του μέσου όρου και της τυπικής απόκλισης για τις 30 ημέρες χρησιμοποιήθηκε για να οριστούν τα ανώτατα και τα κατώτατα όρια των spikes κατά το στάδιο της μηχανικής μάθησης (S+ και S-). Αφού υπολογίστηκε η τυπική απόκλιση για τα returnstων 30 ημερών, αυτή προστέθηκε με τον μέσο όρο των τιμών των returns για αντίστοιχα 30 ημέρες ώστε να υπολογιστεί το ανώτατο όριο των spikes και όμοια οι δύο τιμές αφαιρέθηκαν για να υπολογιστεί το κάτω όριο των spikes.

Πίνακας 1.2 Επεξεργασία δεδομένων για ημερήσιες τιμές

Επεξεργασία δεδομένων για ημερήσιες τιμές																		
r	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	SMA5	EMA5	SMA10	EMA10	σ(5)	σ(10)	σ(30)	Avrg(30)

2.1.2 Εβδομαδιαία δεδομένα

Για την επεξεργασία των εβδομαδιαίων δεδομένων χρησιμοποιήθηκαν τα returns (r) των τιμών του WTI spot price, τα οποία προκύπτουν από την διαφορά των λογαριθμικών τιμών δύο διαδοχικών εβδομάδων. Ακόμα χρησιμοποιήθηκαν 10 Lags των returns (L1 – L10). Στην συνέχεια στα δεδομένα εφαρμόστηκε ο Κινητός Μέσος Όρος (Simple Moving Average) για 4 εβδομάδες και 52 εβδομάδες (SMA4, SMA52. Ακόμα εφαρμόστηκε στα δεδομένα ο εκθετικός κινητός μέσος όρος (exponential moving average EMA), ο οποίος είναι ένας τύπος κινητών μέσων όρων αλλά δίνει μεγαλύτερη βαρύτητα σε πρόσφατες τιμές δεδομένων. Εφαρμόστηκε για 4 εβδομάδες και 52 εβδομάδες αντίστοιχα (EMA4 και EMA52). Στην συνέχεια εφαρμόστηκε η τυπική απόκλιση (standard deviation) για 30 εβδομάδες. Τέλος, χρησιμοποιήθηκε ο μέσος όρος (average) για χρονικό όριο των 30 εβδομάδων στις τιμές των returns. Ο υπολογισμός του μέσου όρου και της τυπικής απόκλισης για τις 30 βδομάδες χρησιμοποιήθηκε για να οριστούν τα ανώτατα και τα κατώτατα όρια των spikes κατά το στάδιο της μηχανικής μάθησης. Αφού υπολογίστηκε η τυπική απόκλιση για τα returns των 30 εβδομάδων, αυτή προστέθηκε με τον μέσο όρο των τιμών των returns για αντίστοιχα 30 εβδομάδες, ώστε να υπολογιστεί το ανώτατο όριο των spikes και όμοια οι δύο τιμές αφαιρέθηκαν για να υπολογιστεί το κάτω όριο των spikes.

Πίνακας 2.3 Επεξεργασία δεδομένων για εβδομαδιαίες τιμές

Επεξεργασία δεδομένων για εβδομαδιαίες τιμές																
r	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	SMA4	EMA4	SMA52	EMA52	σ(30)	Avrg(30)

2.1.3 Μηνιαία δεδομένα

Για την επεξεργασία των μηνιαίων δεδομένων χρησιμοποιήθηκαν τα returns (r) των τιμών του WTI spot price, τα οποία προκύπτουν από την διαφορά των λογαριθμικών τιμών δύο διαδοχικών μηνών. Ακόμα, χρησιμοποιήθηκαν 12 Lags των returns (L1 – L12). Στην συνέχεια, στα δεδομένα εφαρμόστηκε ο Κινητός Μέσος Όρος (Simple Moving Average) για 6 μήνες και 12 μήνες (SMA6, SMA12). Έπειτα εφαρμόστηκε στα δεδομένα ο εκθετικός κινητός μέσος όρος (exponential moving average EMA), ο οποίος είναι ένας τύπος κινητών μέσων όρων αλλά δίνει μεγαλύτερη βαρύτητα σε πρόσφατες τιμές δεδομένων. Εφαρμόστηκε για 6 μήνες και 12 μήνες αντίστοιχα (EMA6 και EMA12). Στην συνέχεια, εφαρμόστηκε η τυπική απόκλιση (standard deviation) για 24μήνες. Τέλος, χρησιμοποιήθηκε ο μέσος όρος (average) για χρονικό όριο των 24μηνών στις τιμές των returns. Ο υπολογισμός του μέσου όρου και της τυπικής απόκλισης για τους 24 μήνες χρησιμοποιήθηκε για να οριστούν τα ανώτατα και τα κατώτατα όρια των spikes κατά το στάδιο της μηχανικής μάθησης. Αφού υπολογίστηκε η τυπική απόκλιση για τα returns των 24 μηνών, αυτή προστέθηκε με τον μέσο όρο των τιμών των returns για αντίστοιχα 24 μήνες ώστε να υπολογιστεί το ανώτατο όριο των spikes και όμοια οι δύο τιμές αφαιρέθηκαν για να υπολογιστεί το κάτω όριο των spikes.

Πίνακας 2.4Επεξεργασία δεδομένων για μηνιαίες τιμές

Επεξεργασία δεδομένων για μηνιαίες τιμές																		
r	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	SMA6	EMA6	SMA12	EMA12	σ(24)	Avrg(24)

2.2 Μεθοδολογία Μηχανικής Μάθησης

Η μηχανική μάθηση αποτελεί κλάδο της τεχνητής νοημοσύνης, καθώς υποστηρίζει την ιδέα ότι όλα τα συστήματα μπορούν να μάθουν από τα δεδομένα, να εντοπίζουν μοτίβα και να λαμβάνουν δικές τους αποφάσεις με την ελάχιστη ανθρώπινη παρέμβαση. Μπορεί και διερευνά τη μελέτη και τη κατασκευή αλγορίθμων, που μπορούν να εκπαιδευτούν από τα δεδομένα και στη συνέχεια να κάνουν προβλέψεις. Τα μοντέλα αυτά επιτρέπουν στη συνέχεια στους ερευνητές και στους αναλυτές να

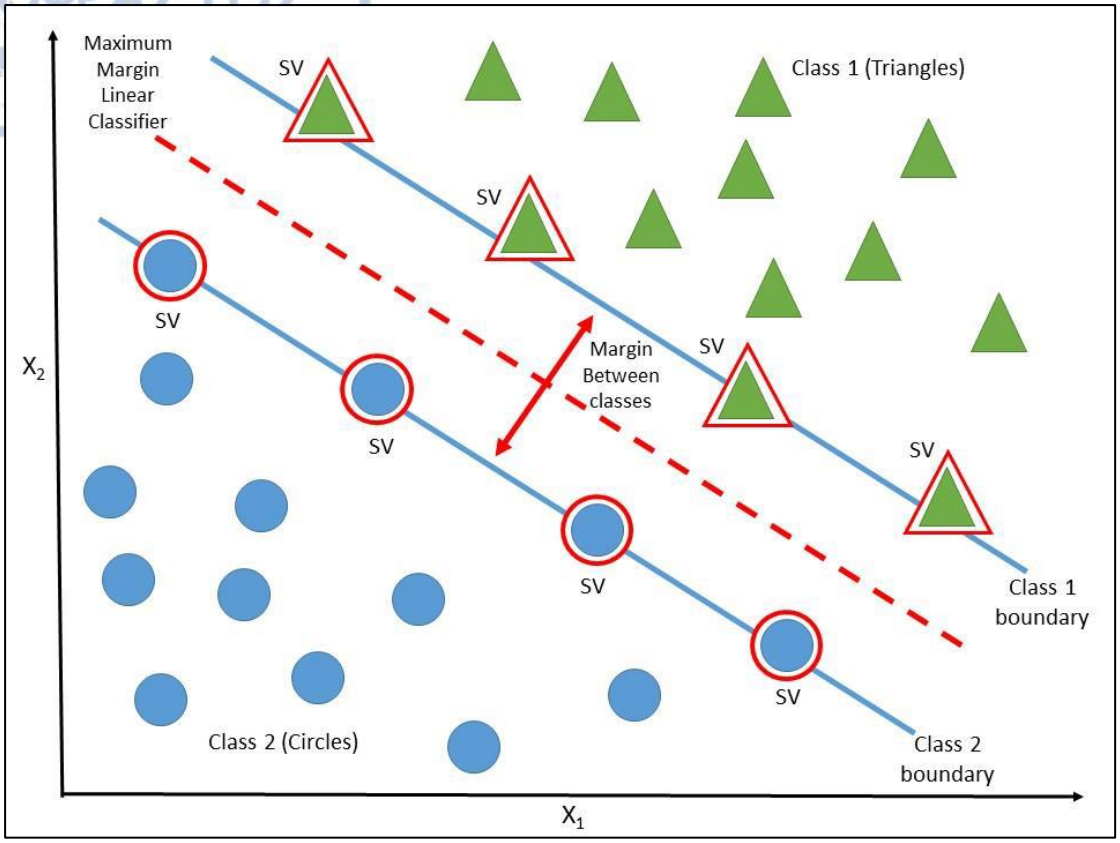
προβούν σε συμπεράσματα και να οδηγηθούν σε αξιόπιστες αποφάσεις και αποτελέσματα, αλλά και να αναδείξουν αλληλοσυσχετίσεις μέσου της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

Η μηχανική μάθηση αποτελεί σήμερα τον πιο αναπτυσσόμενο κλάδο και αμφιταλαντεύεται μεταξύ της επιστήμης των υπολογιστών και της στατιστικής. Η εφαρμογή των τεχνικών της μηχανικής μάθησης μπορεί να εντοπιστεί στην επιστήμη, την τεχνολογία, το εμπόριο και σε άλλους τομείς, δημιουργώντας ένα περιβάλλον, που βασίζεται σε καλύτερες αποδείξεις και τεκμηριωμένες αποφάσεις σε πτυχές της καθημερινής ζωής των ανθρώπων, όπως στο σύστημα υγείας, τον κατασκευαστικό κλάδο, την εκπαίδευση, στα οικονομικά μοντέλα και στις ακολουθούμενες πολιτικές και στο μάρκετινγκ (M.T. Mitchell & M. I. Jordan, 2015).

Στην παρούσα εργασία εφαρμοστήκανε τα μοντέλα Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Logistic Regression και Bagged Trees τα οποία θα αναλυθούν παρακάτω.

2.2.1 Support Vector Machine (SVM)

Το Support Vector Machine (SVM) αποτελεί μια εποπτευόμενη μεθοδολογία μηχανικής εκμάθησης που χρησιμοποιείται σε δεδομένα ταξινόμησης. Η βασική ιδέα γύρω από τη μέθοδο είναι η επιλογή ενός μικρού αριθμού σημειακών δεδομένων από ένα μεγαλύτερο σύνολο δεδομένων, τα οποία ονομάζονται υποστηρικτικά διανύσματα (Support Vectors) και καθορίζουν τα όρια δύο διαφορετικών κλάσεων με ένα γραμμικό όριο το οποίο απέχει το ίδιο από τις δύο κλάσεις. Η μεθοδολογία μπορεί να γενικευτεί για περιπτώσεις, οι οποίες εμπεριέχουν περισσότερες από μία τάξεις. Παρόλα αυτά, στις περισσότερες περιπτώσεις ή σε απλές περιπτώσεις η δυαδική απεικόνιση του μοντέλου είναι επαρκής για να δώσει τα επιθυμητά αποτελέσματα (Dimitriadou A., et al, 2018). Η παραπάνω μέθοδος αναλύεται στην εικόνα 5.



Εικόνα 5. Hyperplane και support vectors. Οι δύο τάξεις αναπαρίστανται με κύκλους και τρίγωνα. Τα support vectors υποδεικνύονται από τους έντονους κοίλους κύκλους και τρίγωνα (με κόκκινου χρώματος), οι margin lines (γραμμές περιθωρίου) αντιπροσωπεύονται από τις συνεχείς γραμμές και το hyperplane (γραμμικός διαχωριστής) αντιπροσωπεύεται από την διακεκομμένη γραμμή. (Dimitriadou A., et al, 2018).

Το SVM μπορεί να επεκταθεί σε προβλήματα μη γραμμικής ταξινόμησης προβάλλοντας τα δεδομένα από τον χώρο, που εισάγονται αρχικώς σε ένα χώρο με περισσότερα χαρακτηριστικά και περισσότερες διαστάσεις, όπου οι δύο κλάσεις διαχωρίζονται γραμμικά. Η προβολή των δεδομένων στον χώρο των περισσότερων διαστάσεων μπορεί να γίνει αναλυτικά μέσω των συναρτήσεων του Kernel (Th Papadimitriou et al, 2020). Τέλος, τα SVM έχουν ισχυρή ικανότητα μάθησης γεγονός, που τα καθιστά το ίδιο αξιόπιστα ως προς τις περιπτώσεις με μικρό σύνολο δειγμάτων, καθώς δίνει μεγάλη επιτυχία πρόγνωσης αλλά και καλά ποσοστιαία αποτελέσματα.

2.2.2 Μαθηματική προσέγγιση

Στη μέθοδο του SVM θεωρείται ένα σετ δεδομένων διανυσμάτων $\mathbf{x}i \in \mathbb{R}2$ (i=1, 2…, n) το οποίο ανήκει σε δύο κλάσεις $yi \in \{-1, +1\}$. Εάν οι δύο αυτές κλάσεις είναι γραμμικά διαχωρισμένες, τότε ορίζουμε ένα όριο διαχωρισμού ως:

$$f(\mathbf{x}i) = \mathbf{w}\boldsymbol{T}\mathbf{x}i - b = 0 \quad (1)$$

Συνυπολογίζοντας ότι:

$$\mathbf{w}\boldsymbol{T}\mathbf{x}i - b > 0 \; for yi \in +1 \text{ και } \mathbf{w}\boldsymbol{T}\mathbf{x}i - b < 0 \; for yi \in -1 \quad (2)$$

Ώστε $yif(xi) > 0$, ∀i. Όπου w είναι το διάνυσμα των παραμέτρων και b είναι το bias (Dimitriadou A., et al, 2018).

Ο βέλτιστος διαχωριστής (hyperplane) αναγνωρίζεται ως το όριο καθορισμού, που ταξινομεί κάθε διάνυσμα δεδομένων στην αντίστοιχη κατηγορία και έχει την μέγιστη απόσταση (πολλές φορές αναφέρεται ως margin) και από τις δύο κατηγορίες. Τα οριακά σημεία δεδομένων (marginal data points) , που καθορίζουν την θέση του hyperplane ονομάζονται Support Vectors (SVs) (Th Papadimitriou et al, 2020).

Η λύση στο πρόβλημα εντοπισμού του βέλτιστου hyperplane μπορεί να δοθεί από την διαδικασία του Lagrange, που χρησιμοποιεί την παρακάτω εξίσωση:

$$\min_{\mathbf{w},b,\xi} \max_{\alpha_n} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{j=1}^{N} a_j[y_j(\mathbf{w}^{\mathrm{T}}\mathbf{x}_j - b) - 1 + \xi_j] - \sum_{k=1}^{N}\mu_k\xi_k \right\} \quad (3)$$

Όπου si μετρά την απόσταση του διανύσματος xi από το hyperplane όταν υπάρξει κάποια λανθασμένη ταξινόμηση και τα α1,α2,...,αn είναι οι μη αρνητικοί πολλαπλασιαστές του Lagrange.

Το hyperplane ορίζεται επομένως ως:

$$\hat{\mathbf{w}} = \sum_{i=1}^{N} a_i \mathrm{y}_i \mathbf{x}_i$$

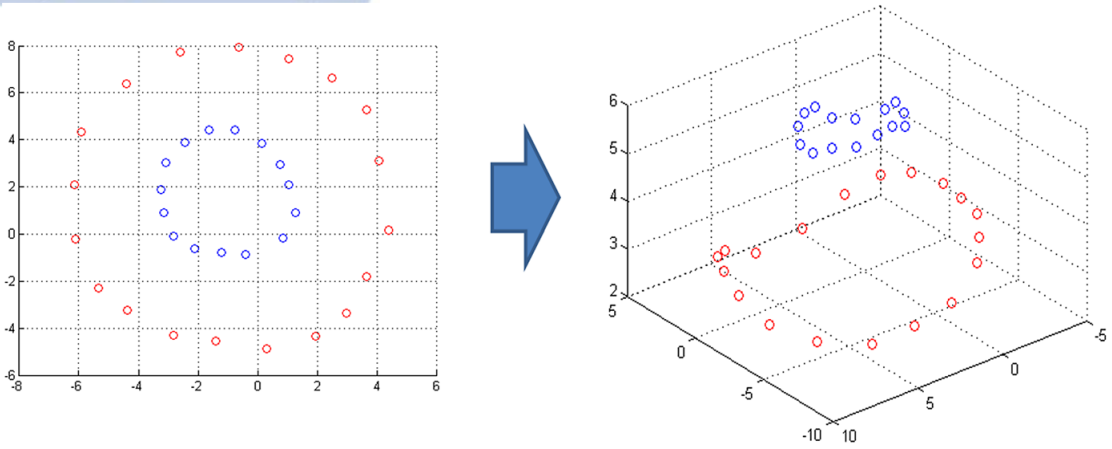
$$\hat{b} = \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_i - \mathrm{y}_i, \; i \in V \quad (4)$$

Όπου V = {i : 0 <yi< C}, είναι το άθροισμα των δεικτών του support vector (Dimitriadou A., et al, 2018).

2.2.3 Kernel Functions

Το SVM χρησιμοποιεί τις μεταβλητές πυρήνα - kernel για να μετατρέψει τα δεδομένα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης που καθιστά ευκολότερη την διαδικασία της διάκρισης μεταξύ των κλάσεων ή τη δημιουργία προβλέψεων. Η διαδικασία προβολής των δεδομένων σε χώρους περισσότερων διαστάσεων συνεχίζεται με διάφορους πυρήνες (kernels) έως ότου τα δεδομένα στο χώρο να είναι γραμμικά διαχωρισμένα ή όσο γίνεται πιο κοντά σε αυτήν την

κατάσταση. Όταν η συνάρτηση του πυρήνα είναι μη γραμμική, το παραγόμενο μοντέλο του SVM είναι επίσης μη γραμμικό (Th Papadimitriou et al, 2020).



Εικόνα 6. Η περίπτωση μη γραμμικού διαχωρισμού στον χώρο εισόδου στα αριστερά. Η προβολή του δισδιάστατου χώρου των δεδομένων σε έναν τρισδιάστατο χώρο χαρακτηριστικών στα δεξιά της εικόνας όπου τα δεδομένα μπορούν να χωριστούν με ένα υπέρ επίπεδο δύο διαστάσεων (Th Papadimitriou et al, 2020).

Η πιο συχνά χρησιμοποιούμενη συνάρτηση πυρήνα είναι η Gaussian SVM ή η radial basis function (RBF). Ο πυρήνας RBF αντιστοιχίζει τα δεδομένα εισόδου σε έναν απεριόριστο χώρο χαρακτηριστικών χρησιμοποιώντας συναρτήσεις Gaussian. Η δημοφιλία της συνάρτησης αυτής οφείλεται στο ότι μπορεί να καταγράψει πολύπλοκες μη γραμμικές σχέσεις στα δεδομένα. Οι μέθοδοι πυρήνα στα SVM είναι μια ισχυρή τεχνική για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης και χρησιμοποιούνται ευρέως στη μηχανική μάθηση, επειδή μπορούν να χειριστούν πολύπλοκες δομές δεδομένων και είναι ανθεκτικές σε θόρυβο και ακραίες τιμές.

Οι μέθοδοι του SVM χρησιμοποιήθηκαν στην παρούσα εργασία περιγράφονται παρακάτω:

- Linear SVM

Η μέθοδος αυτή, είναι ένας τύπος συνάρτησης πυρήνα, που χρησιμοποιείται στη μηχανική εκμάθηση, συμπεριλαμβανομένων των SVM (Support Vector Machines). Είναι η απλούστερη και πιο συχνά χρησιμοποιούμενη συνάρτηση πυρήνα και ορίζει το γινόμενο σημείων μεταξύ των διανυσμάτων εισόδου στον αρχικό χώρο των χαρακτηριστικών.

Ορίζεται ως:

$$K(x, y) = x*y \quad (5)$$

Όπου x και y είναι τα διανύσματα χαρακτηριστικών εισόδου. Το γινόμενο των σημείων των διανυσμάτων εισόδου είναι ένα μέτρο της ομοιότητας ή της απόστασής τους στον αρχικό χώρο χαρακτηριστικών. Όταν χρησιμοποιείται ένας γραμμικός πυρήνας σε ένα SVM, το όριο απόφασης είναι ένα γραμμικό υπερεπίπεδο, που διαχωρίζει τις διαφορετικές κλάσεις στο χώρο χαρακτηριστικών. Αυτό το γραμμικό όριο μπορεί να είναι χρήσιμο, όταν τα δεδομένα είναι ήδη διαχωρίσιμα με ένα γραμμικό όριο απόφασης ή όταν έχουμε να κάνουμε με δεδομένα υψηλών διαστάσεων, όπου η χρήση πιο περίπλοκων συναρτήσεων πυρήνα μπορεί να οδηγήσει σε υπερπροσαρμογή.

- Quadratic SVM

Αποτελεί μια μέθοδο που χρησιμοποιεί τετραγωνικούς πυρήνες (quadratic kernels). Κατά τη φάση της εκπαίδευσης η μνήμη που χρησιμοποιείται για την δυαδική ταξινόμηση είναι μεσαίας τάξης, ενώ για πολλαπλές τάξεις η μνήμη που χρησιμοποιείται είναι μεγάλη. Η ταχύτητα πρόβλεψης για την δυαδική ταξινόμηση είναι γρήγορη και για πολλαπλές κλάσεις είναι πιο αργή. Τέλος η ερμηνεία για τα μοντέλα είναι πιο δύσκολη , ενώ η ευελιξία που έχουν τα μοντέλα είναι πιο μέτρια (Bhoopesh Singh Bhati et al, 2019).Η Quadratic SVM μπορεί να οριστεί ως:

$$K\,(ys, yt)\,(1 + ysyt)^{2} \quad (6)$$

- Cubic SVM

Αποτελεί μέθοδο που χρησιμοποιεί κυβικούς πυρήνες (cubic kernels). Ο τύπος της συνάρτησης αυτής (7), φαίνεται παρακάτω όπου το x και το y είναι τα δεδομένα εισόδου και το c είναι μια σταθερά. Ο στόχος της συνάρτησης αυτής είναι να εντοπιστούν οι βέλτιστες τιμές οι οποίες μεγιστοποιούν το περιθώριο μεταξύ των τάξεων και παράλληλα ελαχιστοποιούνε σφάλματα στην ταξινόμηση.

$$K(x, y) = (x \cdot y + c)^{\wedge}3 \quad (7)$$

- Gaussian SVM (Fine Gaussian – Medium Gaussian – Coarse Gaussian)

Τα Gaussian SVM μοντέλα χρησιμοποιούνται συνήθως για επίλυση απαιτητικών προβλημάτων στη μηχανική μάθηση. Είναι άκρως χρήσιμα και προτιμώνται εξαιτίας της ευέλικτης μη παραμετρικής φύσης τους και της υπολογιστικής απλότητας που εμφανίζουν. Είναι αλλιώς γνωστά και ως radial basis

function (RBF). Χαρακτηρίζονται από μη γραμμική συνάρτηση πυρήνα, που αντιστοιχίζει τα δεδομένα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης χρησιμοποιώντας τις αντίστοιχες συναρτήσεις. Όταν χρησιμοποιείται ένας πυρήνας Gaussian σε ένα μοντέλο SVM, το όριο απόφασης είναι ένα μη γραμμικό υπερεπίπεδο που μπορεί να συλλάβει πολύπλοκες μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου. Το πλάτος της συνάρτησης Gaussian, που ελέγχεται από την παράμετρο γάμμα, καθορίζει τον βαθμό μη γραμμικότητας στο όριο απόφασης. Η εξίσωση μπορεί να χαρακτηριστεί ως η παρακάτω:

$$K(x, y) = \exp(-\text{gamma} * ||x - y||^2) \quad (8)$$

Όπου το x και το y είναι διανύσματα χαρακτηριστικών εισόδου, το γάμμα είναι μια παράμετρος που ελέγχει το πλάτος της gaussian συνάρτησης και το ||x - y||^2 είναι η ευκλείδεια απόσταση στο τετράγωνο μεταξύ των διανυσμάτων εισόδου (javatpoint.com).

Η Gaussian SVM, διαθέτει πολύ καλή απόδοση με ένα σχετικά μικρό σύνολο δεδομένων εκπαίδευσης (training data set) για να κάνει τις προγνώσεις της. Ακόμα είναι ο καλύτερος δυνατός αλγόριθμος που μπορεί να επιλεχθεί όταν οι κλάσεις μπορούν να διαχωριστούν, ενώ το υπερεπίπεδο επηρεάζεται μόνο από τα support vectors και έτσι τα ακραία σημεία έχουν μικρότερο αντίκτυπο. Όμως συχνά εντοπίζονται δυσκολίες λόγω της πολυπλοκότητας στους υπολογισμούς κατά την πρόγνωση, γεγονός που κάνει ανέφικτη την αποτελεσματικότητα του GPR για μεγάλα σύνολα δεδομένων (M. Suetal, 2019).

Στο μοντέλο της Fine Gaussian SVM, κατά τη φάση του training, η ερμηνεία είναι αρκετά δύσκολη, ενώ η ευελιξία του μοντέλου είναι μεγάλη, καθώς μειώνεται με ρύθμιση της κλίμακας kernel (sqrt(P)/4). Η μέθοδος αυτή προσφέρει καλή ευελιξία ως προς την διάκριση μεταξύ των τάξεων. Η ταχύτητα πρόβλεψης για δυαδικά μοντέλα ταξινόμησης είναι γρήγορη και για μοντέλα περισσότερων τάξεων αργή (Bhoopesh Singh Bhati et al, 2019).

Στο μοντέλο της Medium Gaussian SVM, η ταχύτητα πρόβλεψης στη φάση του εκπαίδευσης για τα δυαδικά μοντέλα είναι γρήγορη, ενώ για μοντέλα πολλαπλών κλάσεων είναι αργή. Η ερμηνεία του μοντέλου είναι δύσκολη και η ευελιξία που εμφανίζει μέτρια (Bhoopesh Singh Bhati et al, 2019).

- Logistic Regression

Η μέθοδος της λογιστικής παλινδρόμησης χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης και λειτουργεί προβλέποντας τις κατηγορίες των εξαρτημένων μεταβλητών χρησιμοποιώντας ένα σύνολο δεδομένων ανεξάρτητων μεταβλητών. Μπορεί και ταξινομεί τα δεδομένα χρησιμοποιώντας διακριτά σύνολα δεδομένων. Η μέθοδος είναι ένας τύπος παλινδρόμησης που προβλέπει την πιθανότητα εμφάνισης ενός γεγονότος προσαρμόζοντας δεδομένα σε μια λογιστική συνάρτηση. Ακριβώς, όπως πολλές μορφές ανάλυσης αντίστοιχων παλινδρομήσεων, η μέθοδος της λογιστικής παλινδρόμησης χρησιμοποιεί πολλές μεταβλητές πρόβλεψης, που μπορεί να είναι αριθμητικές ή να ανήκουν σε διάφορες κατηγορίες (Nasteski, 2017).

Η θεωρία του Logistic Regression ορίζεται ως:

$$h_\theta(x)=g(\theta^T x) \quad (9)$$

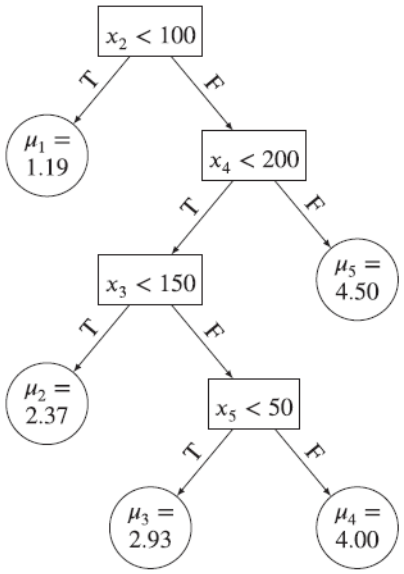
όπου η μεταβλητή του g είναι μια σιγμοειδής συνάρτηση που ορίζεται ως:

$$g(z) = \frac{1}{1+e\hat{}-z} \quad (10)$$

- Bagged Trees

Τα Bagged Trees είναι ένας εξελιγμένος αλγόριθμος της μηχανικής μάθησης ο οποίος χρησιμοποιείται, τόσο για ταξινόμηση, όσο και για παλινδρόμηση. Αποτελούν προγνωστικές προσεγγίσεις μοντελοποίησης, που βρίσκουν εφαρμογή σε τομείς, όπως το datamining, στη στατιστική και στη μηχανική μάθηση. Η δενδροειδείς ταξινόμηση και παλινδρόμηση είναι μη παραμετρικές υπολογιστικές εντατικές μέθοδοι, που έχουν γνωρίσει δημοτικότητα τα τελευταία 12 χρόνια και εφαρμόζονται σε σύνολα δεδομένων, που παρουσιάζουν πολλές παρατηρήσεις και πολλές μεταβλητές, ενώ είναι εξαιρετικά ανθεκτικά ως μοντέλα σε ακραίες τιμές. Τα δένδρα ταξινόμησης και παλινδρόμησης αποτελούνε την βέλτιστη επιλογή για αναλυτές, που θέλουν αρκετά ακριβή αποτελέσματα και σε γρήγορο χρόνο, αλλά μπορεί να μην έχουν την ώρα και την ικανότητα, που απαιτείται για να τα αποκτήσουν χρησιμοποιώντας τις παραδοσιακές μεθόδους. Σε πιο συμβατικές μεθόδους μπορεί να είναι χρήσιμα, όπου υπάρχουν πολλές μεταβλητές, καθώς χρησιμοποιούνται για να προσδιοριστούν οι σημαντικότερες αυτών και οι αλληλεπιδράσεις τους (Clifton D. Sutton, 2005).

Η λειτουργία των regression trees είναι η ακόλουθη. Έστω, ότι έχουν τις μεταβλητές x=(x1, … , x5) και αποτέλεσμα y. Το regression tree αντιπροσωπεύει την υπό όρους προσδοκία του y δεδομένου του x. Η υπόθεση αυτή αναπαρίσταται στην εικόνα 7. Το κάθε μέρος όπου υπάρχει δυαδικός διαχωρισμός-αποτέλεσμα είναι μια απόφαση και ονομάζεται κόμβος. Στον επάνω κόμβο (ρίζα),υπάρχει μια συνθήκη x2 < 100. Αν x2 < 100 είναι αληθής, ακολουθούμε τη διαδρομή προς τα αριστερά. διαφορετικά, ακολουθούμε τη διαδρομή προς τα δεξιά. Υποθέτοντας ότι το x2 < 100 είναι αληθές, βλέπουμε, ότι φτάνουμε σε έναν κόμβο που δεν χωρίζεται. Αυτό ονομάζεται τερματικός κόμβος και η παράμετρος $\mu 1 = 1{,}19$ είναι η εκχωρημένη τιμή του E[y|x] για κάθε x όπου x2 < 100. Ας υποθέσουμε ότι το x2 < 100 δεν είναι αληθές. Στη συνέχεια, ακολουθώντας κατά μήκος της δεξιάς πλευράς, βρίσκουμε έναν άλλον εσωτερικό κόμβο με συνθήκη x4 < 200. Αυτή η συνθήκη θα πρέπει να ελεγχθεί και αν είναι αληθής (false), θα ακολουθηθεί η διαδρομή προς τα αριστερά (right). Αυτή η διαδικασία συνεχίζεται μέχρι να φτάσουμε σε έναν τερματικό κόμβο και η παράμετρος μi σε αυτόν τον τερματικό κόμβο εκχωρείται ως η τιμή του E[y|x], όπου μi είναι ο μέσος όρος του κόμβου i για το regression tree. Έτσι, για παράδειγμα, μια περίπτωση k με xk1 = 30, xk2 = 120, xk3 = 115, xk4 = 191,και xk5 = 56 θα εκχωρηθεί μια τιμή $\mu 2 = 2{,}37$ για E[y|x]. Η τιμή θα ήταν ακριβώς η ίδια για μια άλλη περίπτωση k′ που αντί αυτού είχε συμμεταβλητές xk′1 = 130, xk′2 = 135, xk′3 = 92, xk′4 = 183, xk′5 = 10 (Yaoyuan V. Tan et al, 2019).



Εικόνα 7. Παράδειγμα regression tree (x; T,M) όπου το μi είναι η μέση παράμετρος του i κόμβου (Yaoyuan V. Tan et al, 2019)

Το Bootstrap Aggregation ή αλλιώς bagging αποτελεί μια τεχνική που μπορεί να χρησιμοποιηθεί με πολλές μεθόδους ταξινόμησης και μεθόδους παλινδρόμησης με στόχο τη μείωση της διακύμανσης που σχετίζεται με την πρόβλεψη και ως εκ τούτου την βελτίωση της διαδικασίας πρόβλεψης. Είναι μια σχετικά απλή διαδικασία, όπου πολλά δείγματα “bootstrap” αντλούνται από τα διαθέσιμα δεδομένα, εφαρμόζονται μέθοδοι πρόβλεψης σε πολλά διαφορετικά υποσύνολα δεδομένων στα δεδομένα εκπαίδευσης τα οποία επιλέγονται τυχαία. Στη συνέχεια τα αποτελέσματα σε συνδυασμό με τον υπολογισμό του μέσου όρου για τη παλινδρόμηση με μια απλή ταξινόμηση, λαμβάνεται συνολική πρόβλεψη, όπου η διακύμανση είναι μειωμένη λόγου του μέσου όρου. (CliftonD. Sutton, 2005). Ο μέσος όρος αυτός που προκύπτει από πολλά διαφορετικά δέντρα, είναι ισχυρότερος και αρκετά πιο ακριβής από μόνο ένα δέντρο.

3. Αποτελέσματα

Στο κεφάλαιο αυτό, παρουσιάζονται τα αποτελέσματα και οι αποδόσεις των μοντέλων, που χρησιμοποιήθηκαν για την πρόβλεψη των spikes του WTI Crude oil. Ως δεδομένα, χρησιμοποιήθηκαν οι τιμές των ημερήσιων τιμών, των εβδομαδιαίων και των μηνιαίων τιμών των returns του WTI Crude oil. Για να ελεγχθεί η ικανότητα του κάθε μοντέλου, το σύνολο των δεδομένων χωρίστηκε σε σετ δεδομένων, το 90% αυτών αποτέλεσε το σετ δεδομένων εκπαίδευσης (training data set) και το 10% αυτών αποτέλεσε τα δεδομένα ελέγχου (test data set). Στην συνέχεια ελέγχθηκε η ακρίβεια πρόβλεψης, που έδωσε το κάθε μοντέλο, που χρησιμοποιήθηκε για τα δεδομένα της εκπαίδευσης και τα δεδομένα ελέγχου. Συνολικά για τα ημερήσια και τα εβδομαδιαία δεδομένα χρησιμοποιήθηκαν 10 Lags και για τα μηνιαία δεδομένα χρησιμοποιήθηκαν 12 Lags συνολικά. Παρακάτω, αναλύονται τα σετ των δεδομένων ξεχωριστά για τις ημερήσιες, εβδομαδιαίες και μηνιαίες τιμές των δεδομένων και ορισμένα metrics που χρησιμοποιήθηκαν για την ερμηνεία των μοντέλων.

Σε γενικές γραμμές τα classification metrics είναι αριθμοί που μετράνε την απόδοση ενός μοντέλου μηχανικής μάθησης, όταν σε αυτό ανατίθεται η δουλεία της παρατήρησης ορισμένων τάξεων. Η δυαδική ταξινόμηση (binary classification) είναι μια ιδιαίτερη κατάσταση, όπου το εκάστοτε μοντέλο πρέπει απλώς να καταγράψει αν οι τάσεις είναι αρνητικές ή θετικές συνήθως με τις τιμές 0 και 1.

3.1.1 Classification – Accuracy

Η ακρίβεια είναι μια μέτρηση για την αξιολόγηση των μοντέλων ταξινόμησης (classification models). Εμπειρικά η τιμή αυτή προκύπτει από το κλάσμα των προβλέψεων, που ένα μοντέλο είχε σωστές (neptune.ai). Το accuracy προκύπτει από τον παρακάτω τύπο:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ prediction} \quad (11)$$

Για δυαδικές ταξινομήσεις, όπως τα πλαίσια της εργασίας αυτής το ποσοστό του accuracy μπορεί να υπολογιστεί με βάση τις θετικές και τις αρνητικές προβλέψεις ως παρακάτω:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Οπου:TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Σημαντικό είναι να σημειωθεί, ότι υψηλή ακρίβεια accuracy, δεν μας δίνει την πραγματική εικόνα για ένα σύνολο δεδομένων και δεν πρέπει να εξετάζεται μόνη της. Στις περιπτώσεις όπου τα δεδομένα εκπαίδευσης εμφανίζουν υψηλά ποσοστά accuracy, το γεγονός αυτό μπορεί να οφείλεται σε overfit ή σε imbalanced δεδομένα. Ένα υψηλό ποσοστό ακρίβειας στα δεδομένα εκπαίδευσης δεν εγγυάται απαραίτητα ένα υψηλό ποσοστό πρόβλεψης στα δεδομένα ελέγχου.

Στην συγκεκριμένη εργασία, το ποσοστό του accuracy των ημερησίων δεδομένων για τα δεδομένα εκπαίδευσης κυμάνθηκε από 69.7% έως 71.5% και για τα δεδομένα ελέγχου από 70.87% έως 72.46%, ενώ όλα τα ποσοστά παραθέτονται στον πίνακα 3.2. Το ποσοστό του accuracy των εβδομαδιαίων δεδομένων για τα δεδομένα εκπαίδευσης κυμάνθηκε από 69% έως 69.3% και για τα δεδομένα ελέγχου από 59.78% έως 72.48% ενώ όλα τα ποσοστά παραθέτονται στον πίνακα 3.6. Το ποσοστό του accuracy των μηνιαίων δεδομένων για τα δεδομένα εκπαίδευσης κυμάνθηκε από 60.5% έως 69.2% και για τα δεδομένα ελέγχου από 57.14% έως 78.57%, ενώ όλα τα ποσοστά παραθέτονται στον πίνακα 3.10. Τα ποσοστά υποδεικνύουν ότι τα δεδομένα δεν εμφανίζουν κάποιο overfitting και τα μοντέλα εμφανίζουν σχετικά ικανοποιητικό ποσοστό accuracy το οποίο για όλες τις ομάδες δεδομένων δεν πέφτει κάτω από το 55% και δεν ξεπερνάει το 80%.

3.1.1.2 Confusion Matrix

Το Confusion Matrix είναι ένας συνηθισμένος τρόπος παρουσίασης προβλέψεων true positive (tp), true negative (tn), false positive (fp) και false negative (fn). Οι τιμές αυτές παρουσιάζονται με τη μορφή πίνακα, όπου ο άξονας Y απεικονίζει τις αληθείς κλάσεις, ενώ ο άξονας των X απεικονίζει τις προβλεπόμενες κλάσεις. Το Confusion Matrix προκύπτει, αφού εκπαιδευτούν τα μοντέλα και καθοριστούν τα αποτελέσματα. Τα Confusion Matrix που εξήχθησαν από τα μοντέλα εκπαίδευσης απεικονίζονται στο Παράρτημα Ι α για τα ημερήσια δεδομένα, στο Παράρτημα ΙΙ α για τα εβδομαδιαία δεδομένα και στο Παράρτημα ΙΙΙ α για τα μηνιαία δεδομένα.

3.1.1.3 True Positive / False Negative Rates

Η μεταβλητή True Positive Rates (tr) μετράει το πόσες παρατηρήσεις από το σύνολο των πραγματικών παρατηρήσεων έχουν ταξινομηθεί ως θετικές. Δηλαδή

αντικατοπτρίζει τις τιμές που ήταν πραγματικά θετικές και έχουν προβλεφθεί ως θετικές από το μοντέλο. Ο τύπος είναι:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (13)$$

Συνήθως δεν χρησιμοποιείται μόνο, αλλά σε συνδυασμό με άλλες μετρήσεις, όπως το accuracy.

Όταν δεν προβλέπουμε κάτι πραγματικό, αλλά το μοντέλο προβλέπει ψευδώς αρνητικές τιμές, τότε το metric ονομάζεται False Negative Rate (FNR). Ουσιαστικά αποτελεί ένα κλάσμα χαμένων μεταβλητών, που το μοντέλο προβλέπει και εμφανίζει κατά λάθος. Σχηματικά τα αποτελέσματα απεικονίζονται στο Παράρτημα Ι β για τα ημερήσια δεδομένα, στο Παράρτημα ΙΙ β για τα εβδομαδιαία δεδομένα και στο Παράρτημα ΙΙΙ β για τα μηνιαία δεδομένα.

Ο τύπος που ορίζει την μεταβλητή αυτή είναι ο ακόλουθος:

$$FNR = \frac{FN}{TP+FN} \quad (14)$$

3.1.1.4 Positive Predictive Values – False Discovery Rates

Το metric Positive Predictive Values μετράει το πόσες παρατηρήσεις από τις οποίες προβλέπονται ως θετικές είναι όντως θετικές. Ουσιαστικά μας δίνει το ποσοστό των προβλέψεων, που ταξινομούνται σωστά ως προς αυτές, που ταξινομούνται λάθος. Ο τύπος είναι ο ακόλουθος:

$$PPV = \frac{TP}{TP+FP} \quad (15)$$

Δεν έχει ιδιαίτερο νόημα μόνη της σαν metric, αλλά χρησιμοποιείται σε συνδυασμό με άλλα. Όταν αυξάνονται οι ψευδείς ενδείξεις και θέλουμε να διασφαλίσουμε τις θετικές προβλέψεις, ότι είναι όντως θετικές, τότε θα πρέπει να αναζητήσουμε τρόπους να βελτιστοποιήσουμε την ακρίβεια. Τα αποτελέσματα για το metric απεικονίζονται στο Παράρτημα Ι γ για τα ημερήσια δεδομένα, στο Παράρτημα ΙΙ γ για τα εβδομαδιαία δεδομένα και στο Παράρτημα ΙΙΙ γ για τα μηνιαία δεδομένα.

Το metric False Discovery Rates, ελέγχει το πόσες από τις θετικές προβλέψεις που έγιναν ήταν λανθασμένες. Ο τύπος είναι:

$$FDR = \frac{FP}{FP + TP} \quad (16)$$

3.1.2 Ημερήσια Δεδομένα

Οι τιμές των δεδομένων για τις ημερήσιες τιμές ξεκινάνε από τις 10 Απριλίου του 1986 και σταματάνε στις 06 Ιανουαρίου του 2023 και αναλυτικά οι συνολικές τιμές για την κάθε επιμέρους κατηγορία (δεδομένα εκπαίδευσης – δεδομένα ελέγχου) καταγράφονται στον πίνακα 3.1.

Πίνακας 3.1 Συνολικές τιμές για κάθε κατηγορία δεδομένων για τις ημερήσιες τιμές

Ημερήσια Δεδομένα			
Τύπος δεδομένων	**Από**	**Έως**	**Σύνολο τιμών (ημέρες)**
Σύνολο τιμών	10/04/1986	6/1/2023	9258
Δεδομένα εκπαίδευσης (90%)	10/4/1986	29/4/2019	8332
Δεδομένα ελέγχου (10%)	30/4/2019	6/1/2023	926

Στα ημερήσια δεδομένα των τιμών του WTI crude oil, αρχικά χρησιμοποιήθηκαν μαζί τα 10 Lags, moving average για 5 και 10 μέρες και standard deviation για 30 μέρες. Τα αποτελέσματα κρίθηκαν ανεπαρκή και για αυτόν τον λόγο στην συνέχεια προστέθηκε το standard deviation και για 5 ημέρες αλλά και για 10 μέρες. Τα μοντέλα δίνουν τιμή 0 στην περίπτωση, που το εκάστοτε μοντέλο προβλέπει ότι η τιμή των returns κυμαίνεται εντός των ορίων S+ και S-. Το ανώτατο και κατώτατο όριο S+ και S-ορίστηκαν ως το stdev(30) + average(30) και average(30) - stdev(30). Αντίστοιχα το μοντέλο προβλέπει την τιμή ως 1, όταν η τιμή των returns βρίσκεται εκτός των ορίων, οπότε και δημιουργείται spike. Τα ποσοστά του accuracy για τα δεδομένα εκπαίδευσης αναφέρονται στον πίνακα 3.2 και τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα Ι.

Τα μοντέλα στο σύνολό τους δίνουν ικανοποιητικά αποτελέσματα όσο αναφορά στο accuracy, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Το γεγονός ότι τα ποσοστά κυμαίνονται κοντά μεταξύ τους δείχνει ότι δεν υπάρχει overfit στα δεδομένα. Το καλύτερο accuracy εμφανίζεται στο μοντέλο Medium Gaussian SVM ενώ το χειρότερο εμφανίζεται στο μοντέλο Cubic SVM.

Πίνακας 3.2 Ποσοστά accuracy για τα ημερήσια δεδομένα εκπαίδευσης και για τα ημερήσια δεδομένα ελέγχου.

Μοντέλα πρόβλεψης	**Δεδομένα εκπαίδευσης (%)**	**Δεδομένα ελέγχου (%)**
Linear SVM	71,5	72,16
Quadratic SVM	71,5	72,16
Cubic SVM	69,7	70,87
Fine Gaussian SVM	71,4	72,27
Medium Gaussian SVM	71,5	72,49
Coarse Gaussian SVM	71,5	72,16
Logistic Regression	71,5	72,06
Bagged Trees	70,5	72,06

Τα ποσοστά των αποτελεσμάτων από τα δεδομένα ελέγχου φαίνονται στον πίνακα 3.2. Η πρόβλεψη των τιμών των returns για το αν προβλέπονται εντός ή εκτός των ορίων ώστε να προκύψει spike, έγινε για τη χρονική περίοδο t+1. Τα δεδομένα ελέγχου με τη μορφή binary συγκρίθηκαν με τα αποτελέσματα κατά το στάδιο της εκπαίδευσης των δεδομένων. Το ποσοστό προκύπτει από τον λόγο των τιμών με ένδειξη 1, δηλαδή των τιμών που δίνουν πρόβλεψη spike εκτός των ορίων, που καθορίστηκαν με το συνολικό άθροισμα των τιμών.

Στον πίνακα 3.3 παρουσιάζονται τα True Positive Rates για τα μοντέλα, που χρησιμοποιήθηκαν κατά το στάδιο της εκπαίδευσης για το predicted class με ένδειξη 1 και τα ποσοστά των true positive rates για την κλάση 0. Σημειώνεται ότι τα ποσοστά είναι συμπληρωματικά με το metric False Negative Rate. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα Ι.

Πίνακας 3.3. Ποσοστά των true positive rates για το predicted class 1 και για το predicted class 0 για ημερήσια δεδομένα. Τα True Positives αντιπροσωπεύουν τον αριθμό των περιπτώσεων που είναι πραγματικά θετικές και ταξινομούνται ορθά από το μοντέλο ως θετικές.

Μοντέλα πρόβλεψης	**True Positive Rate 1 (%)**	**True Positive Rate 0 (%)**
Linear SVM	0	100
Quadratic SVM	0	100
Cubic SVM	6	95
Fine Gaussian SVM	1	99
Medium Gaussian SVM	<1	>99
Coarse Gaussian SVM	0	100
Logistic Regression	<1	>99
Bagged Trees	8	96

Στον πίνακα 3.4 παρουσιάζονται τα Positive Predictive Values για την κλάση 0 και τα False Discovery Rates για τη κλάση 1. Οι τιμές των Positive Predictive Values και των False Discovery rates για κάθε κλάση, είτε 0 είτε 1 έχουν άθροισμα 100 % για κάθε μια από αυτές. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα Ι.

Πίνακας 3.4. Ποσοστά των Positive Predictive values για 0 και False Discovery rates για 1 για ημερήσια δεδομένα.Τα Positive Predictive Valus δίνουν την πληροφορία για το πόσες προβλέψεις που έγιναν από το εκάστοτε μοντέλο είναι όντως σωστές. Το False Discovery rate, ορίζει το ποσοστό αναμενόμενων ψευδών προβλέψεων μεταξύ του συνόλου των προβλέψεων που έγιναν.

Μοντέλα πρόβλεψης	**Positive Predictive values 0 (%)**	**False Discovery rate 1 (%)**
Linear SVM	72	
Quadratic SVM	72	
Cubic SVM	72	68
Fine Gaussian SVM	72	55
Medium Gaussian SVM	72	40
Coarse Gaussian SVM	72	
Logistic Regression	72	50
Bagged Trees	72	59

Με βάση τα παραπάνω δεδομένα παρατηρούμε, ότι σε γενικές γραμμές τα ποσοστά πρόβλεψης των τιμών για τις τιμές εκτός των ορίων για την πρόβλεψη των spikes δεν είναι ιδιαίτερα καλά. Τα μοντέλα φαίνεται να μην μπορούν να προβλέψουν ικανοποιητικά την παρουσία spikes για μελλοντικό χρόνο. Τα καλύτερα μοντέλα για την πρόβλεψη των τιμών των spikes για τα ημερήσια δεδομένα είναι το Cubic SVM και το Bagged Trees, τα οποία δίνουν να ποσοστό πρόβλεψης των spikes 6% και 8% αντίστοιχα. Τα υπόλοιπα μοντέλα, όπως το Linear SVM, Quadratic SVM, Fine Gaussian SVM, Medium Gaussian SVM και Logistic regression φαίνεται να αδυνατούν να προβλέψουν σωστά τα spikes με τα ποσοστά να κυμαίνονται στο 0% για τα Linear SVM, Quadratic SVM και Coarse Gaussian SVM. Το ποσοστό πρόβλεψης για το Fine Gaussian SVM είναι 1 και <1% για τα μοντέλα Medium Gaussian SVM και Logistic Regression.

Το αποτέλεσμα αυτό, της χαμηλής πρόβλεψης των spikes, οφείλεται στο γεγονός ότι το target variable είναι imbalanced. Ως εκ τούτου, οι περιπτώσεις όπου τα returns είναι non-spikes είναι πολύ περισσότερες από τις περιπτώσεις όπου

εμφανίζονται ως spikes με απόρροια όλα τα μοντέλα στο σύνολό τους, να προβλέπουν τις τιμές ως επί το πλείστο ως non-spike. Τα non – spikes αποτελούν τα major classes ενώ τα spikes αποτελούν την minority class.

3.1.3 Εβδομαδιαία Δεδομένα

Οι τιμές των δεδομένων για τις εβδομαδιαίες τιμές ξεκινάνε από τις 06 Μαρτίου του 1987 και σταματάνε στις 26 Μαΐου του 2023 και αναλυτικά οι συνολικές τιμές για την κάθε επιμέρους κατηγορία (δεδομένα εκπαίδευσης – δεδομένα ελέγχου) καταγράφονται στον πίνακα 3.5.

Πίνακας 3.5 Συνολικές τιμές για κάθε κατηγορία δεδομένων για τις εβδομαδιαίες τιμές

Εβδομαδιαία Δεδομένα			
Τύπος δεδομένων	**Από**	**Έως**	**Σύνολο τιμών (εβδομάδες)**
Σύνολο τιμών	06/03/1987	26/05/2023	1891
Δεδομένα εκπαίδευσης (90%)	06/03/1987	11/10/2019	1702
Δεδομένα ελέγχου (10%)	18/10/2019	26/05/2023	189

Στα εβδομαδιαία δεδομένα των τιμών του WTI crude oil, χρησιμοποιήθηκαν 10 Lags, moving average για 4 και 52 εβδομάδες και standard deviation για 30 εδβδομάδες.. Τα μοντέλα δίνουν τιμή 0 στην περίπτωση που το εκάστοτε μοντέλο προβλέπει ότι η τιμή των returns κυμαίνεται εντός των ορίων S+ και S-. Το ανώτατο και κατώτατο όριο S+ και S- ορίστηκαν ως το stdev(30) + average(30) και average(30) - stdev(30). Αντίστοιχα το μοντέλο προβλέπει την τιμή ως 1, όταν η τιμή των returns βρίσκεται εκτός των ορίων, οπότε και δημιουργείται spike. Τα ποσοστά του accuracy για τα δεδομένα εκπαίδευσης αναφέρονται στον πίνακα 3.6 και τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙ. Τα ποσοστά των αποτελεσμάτων από τα δεδομένα ελέγχου φαίνονται στον πίνακα 3.6.

Τα μοντέλα στο σύνολό τους δίνουν ικανοποιητικά αποτελέσματα όσο αναφορά στο accuracy, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Το γεγονός ότι τα ποσοστά κυμαίνονται κοντά μεταξύ τους δείχνει ότι δεν υπάρχει overfit στα δεδομένα. Το καλύτερο accuracy εμφανίζεται στα μοντέλα Linear SVM, Fine Gaussian SVM, Coarse Gaussian SVM, Logistic Regression και Bagged Trees, ενώ το χειρότερο εμφανίζεται στο μοντέλο Quadratic SVM.

Πίνακας 3.6. Ποσοστά accuracy για τα εβδομαδιαία δεδομένα εκπαίδευσης και για τα εβδομαδιαία δεδομένα ελέγχου.

Μοντέλα πρόβλεψης	**Δεδομένα εκπαίδευσης (%)**	**Δεδομένα ελέγχου (%)**
Linear SVM	69,3	71,95
Quadratic SVM	69,3	59,78
Cubic SVM	62,2	61,90
Fine Gaussian SVM	69,3	71,95
Medium Gaussian SVM	69	71,95
Coarse Gaussian SVM	69,3	71,95
Logistic Regression	69,2	72,48
Bagged Trees	69,3	71,95

Η πρόβλεψη των τιμών των returns το αν προβλέπονται εντός ή εκτός των ορίων ώστε να προκύψει spike, έγινε για τη χρονική περίοδο t+1. Τα δεδομένα ελέγχου με τη μορφή binary συγκρίθηκαν με τα αποτελέσματα κατά το στάδιο της εκπαίδευσης των δεδομένων.. Το ποσοστό προκύπτει από τον λόγο των τιμών με ένδειξη 1, δηλαδή των τιμών που δίνουν πρόβλεψη spike εκτός των ορίων που καθορίστηκαν με το συνολικό άθροισμα των τιμών.

Στον πίνακα 3.7 παρουσιάζονται τα True Positive Rates για τα μοντέλα που χρησιμοποιήθηκαν κατά το στάδιο της εκπαίδευσης για το predicted classμε ένδειξη 1 και τα ποσοστά των true positive rates για την κλάση 0. Σημειώνεται ότι τα ποσοστά είναι συμπληρωματικά με το metric False Negative Rate. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙ.

Πίνακας 3.7. Ποσοστά των true positive rates για το predicted class 1 και για το predicted class 0 για εβδομαδιαία δεδομένα. Τα True Positives αντιπροσωπεύουν τον αριθμό των περιπτώσεων που είναι πραγματικά θετικές και ταξινομούνται ορθά από το μοντέλο ως θετικές.

Μοντέλα πρόβλεψης	**True Positive Rate 1 (%)**	**True Positive Rate 0 (%)**
Linear SVM	0	100
Quadratic SVM	3	99
Cubic SVM	29	77
Fine Gaussian SVM	<1	>99
Medium Gaussian SVM	<1	99
Coarse Gaussian SVM	0	100
Logistic Regression	<1	>99
Bagged Trees	11	99

Στον πίνακα 3.8 παρουσιάζονται τα Positive Predictive Values για την κλάση 0 και τα False Discovery Rates για τη κλάση 1. Οι τιμές των Positive Predictive Values

και των False Discovery rates για κάθε κλάση, είτε 0 είτε 1 έχουν άθροισμα 100 % για κάθε μια από αυτές. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙ.

Πίνακας 3.8. Ποσοστά των Positive Predictive values για 0 και False Discovery rates για 1 για εβδομαδιαία δεδομένα.Τα Positive Predictive Valus δίνουν την πληροφορία για το πόσες προβλέψεις που έγιναν από το εκάστοτε μοντέλο είναι όντως σωστές. Το False Discovery rate, ορίζει το ποσοστό αναμενόμενων ψευδών προβλέψεων μεταξύ του συνόλου των προβλέψεων που έγιναν.

Μοντέλα πρόβλεψης	**Positive Predictive values 0 (%)**	**False Discovery rate 1 (%)**
Linear SVM	69	
Quadratic SVM	70	52
Cubic SVM	71	64
Fine Gaussian SVM	69	50
Medium Gaussian SVM	69	75
Coarse Gaussian SVM	69	
Logistic Regression	69	60
Bagged Trees	70	60

Με βάση τα παραπάνω δεδομένα παρατηρούμε, ότι σε γενικές γραμμές τα ποσοστά πρόβλεψης των τιμών εκτός των ορίων για την πρόβλεψη των spikes και εδώ δεν αποδίδουν σε ικανοποιητικό βαθμό. Τα μοντέλα φαίνεται να μην μπορούν να προβλέψουν σωστά την παρουσία spikes για μελλοντικό χρόνο. Τα καλύτερα μοντέλα για την πρόβλεψη των τιμών των spikes για τα εβδομαδιαία δεδομένα και εδώ είναι το Cubic SVM και το Bagged Trees τα οποία δίνουν να ποσοστό πρόβλεψης των spikes 29% και 11% αντίστοιχα. Τα υπόλοιπα μοντέλα όπως το Linear SVM, Quadratic SVM, Fine Gaussian SVM, Medium Gaussian SVM και Logistic regression φαίνεται να αδυνατούν να προβλέψουν σωστά τα spikes με τα ποσοστά να κυμαίνονται στο 0% για το Linear SVM, στο 3% για το Quadratic SVM, σε <0,1% για το Fine και Medium Gaussian SVM, 0% για το Coarse Gaussian και <0,1% για το Logistic Regression.

Το αποτέλεσμα αυτό, της χαμηλής πρόβλεψης των spikes, οφείλεται στο γεγονός ότι το target variable και σε αυτό το σετ των δεδομένων είναι imbalanced. Ως εκ τούτου, οι περιπτώσεις όπου τα returns είναι non-spikes είναι πολύ περισσότερες από τις περιπτώσεις όπου εμφανίζονται ως spikes με απόρροια όλα τα μοντέλα στο σύνολό τους, να προβλέπουν τις τιμές ως επί το πλείστο ως non-spike. Ωστόσο είναι χαρακτηριστικό το μεγαλύτερο ποσοστό πρόβλεψης spikes από το μοντέλο Cubic SVM.

3.1.4 Μηνιαία Δεδομένα

Οι τιμές των δεδομένων για τις μηνιαίες τιμές ξεκινάνε από τον Δεκέμβριο του 1988 και σταματάνε στον Μάιο του 2023 και αναλυτικά οι συνολικές τιμές για την κάθε επιμέρους κατηγορία (δεδομένα εκπαίδευσης – δεδομένα ελέγχου) καταγράφονται στον πίνακα 3.9.

Πίνακας 3.9. Συνολικές τιμές για κάθε κατηγορία δεδομένων για τις μηνιαίες τιμές

Μηνιαία Δεδομένα			
Τύπος δεδομένων	**Από**	**Έως**	**Σύνολο τιμών (μήνες)**
Σύνολο τιμών	03/1988	05/2023	415
Δεδομένα εκπαίδευσης (90%)	12/1988	11/2019	372
Δεδομένα ελέγχου (10%)	12/2019	05/2023	42

Στα μηνιαία δεδομένα των τιμών του WTI crude oil, χρησιμοποιήθηκαν 12 Lags, moving average για 6 και 12 μήνες και standard deviation για 24 μήνες. Τα μοντέλα δίνουν τιμή 0 στην περίπτωση, που το εκάστοτε μοντέλο προβλέπει ότι η τιμή των returns κυμαίνεται εντός των ορίων S+ και S-. Το ανώτατο και κατώτατο όριο S+ και S- ορίστηκαν ως το stdev(24) + average(24) και average(24) - stdev(24). Αντίστοιχα το μοντέλο προβλέπει την τιμή ως 1 όταν η τιμή των returns βρίσκεται εκτός των ορίων, οπότε και δημιουργείται spike. Τα ποσοστά του accuracy για τα δεδομένα εκπαίδευσης αναφέρονται στον πίνακα 3.10 και τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙΙ. Τα ποσοστά των αποτελεσμάτων από τα δεδομένα ελέγχου φαίνονται στον πίνακα 3.10.

Τα μοντέλα στο σύνολό τους δίνουν ικανοποιητικά αποτελέσματα όσο αναφορά στο accuracy, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Το γεγονός ότι τα ποσοστά κυμαίνονται κοντά μεταξύ τους δείχνει ότι δεν υπάρχει overfit στα δεδομένα. Το καλύτερο accuracy εμφανίζεται στο μοντέλο Fine Gaussian SVM ενώ το χειρότερο εμφανίζεται στο μοντέλο Bagged Trees.

Πίνακας 3.10. Ποσοστά accuracy για τα μηνιαία δεδομένα εκπαίδευσης και για τα μηνιαία δεδομένα ελέγχου.

Μοντέλα πρόβλεψης	**Δεδομένα εκπαίδευσης (%)**	**Δεδομένα ελέγχου (%)**
Linear SVM	69,1	73,8
Quadratic SVM	62,9	73,8
Cubic SVM	60,5	57,14
Fine Gaussian SVM	69,1	71,42
Medium Gaussian SVM	68,8	73,8
Coarse Gaussian SVM	69,1	73,8
Logistic Regression	68	78,57
Bagged Trees	61,8	73,8

Η πρόβλεψη των τιμών των returns για το αν προβλέπονται εντός ή εκτός των ορίων ώστε να προκύψει spike, έγινε για τη χρονική περίοδο t+1. Τα δεδομένα ελέγχου με τη μορφή binary συγκρίθηκαν με τα αποτελέσματα κατά το στάδιο της εκπαίδευσης των δεδομένων.. Το ποσοστό προκύπτει από τον λόγο των τιμών με ένδειξη 1, δηλαδή των τιμών που δίνουν πρόβλεψη spike εκτός των ορίων που καθορίστηκαν με το συνολικό άθροισμα των τιμών.

Στον πίνακα 3.11 παρουσιάζονται τα True Positive Rates για τα μοντέλα που χρησιμοποιήθηκαν κατά το στάδιο της εκπαίδευσης για το predicted class με ένδειξη 1 και τα ποσοστά των true positive rates για την κλάση 0. Σημειώνεται ότι τα ποσοστά είναι συμπληρωματικά με το metric False Negative Rate. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙΙ.

Πίνακας 3.11. Ποσοστά των true positive rates για το predicted class 1 και για το predicted class 0 για μηνιαία δεδομένα Τα True Positives αντιπροσωπεύουν τον αριθμό των περιπτώσεων που είναι πραγματικά θετικές και ταξινομούνται ορθά από το μοντέλο ως θετικές.

Μοντέλα πρόβλεψης	**True Positive Rate 1 (%)**	**True Positive Rate 0 (%)**
Linear SVM	0	100
Quadratic SVM	33	79
Cubic SVM	37	72
Fine Gaussian SVM	0	100
Medium Gaussian SVM	4	98
Coarse Gaussian SVM	0	100
Logistic Regression	14	90
Bagged Trees	10	91

Στον πίνακα 3.12 παρουσιάζονται τα Positive Predictive Values για την κλάση 0 και τα False Discovery Rates για τη κλάση 1. Οι τιμές των Positive Predictive Values

και των False Discovery rates για κάθε κλάση, είτε 0 είτε 1 έχουν άθροισμα 100 % για κάθε μια από αυτές. Τα διαγράμματα για το κάθε μοντέλο παρουσιάζονται στο Παράρτημα ΙΙΙ.

Πίνακας 3.12. Ποσοστά των Positive Predictive values για 0 και False Discovery rates για 1 για μηνιαία δεδομένα.Τα Positive Predictive Valus δίνουν την πληροφορία για το πόσες προβλέψεις που έγιναν από το εκάστοτε μοντέλο είναι όντως σωστές. Το False Discovery rate, ορίζει το ποσοστό αναμενόμενων ψευδών προβλέψεων μεταξύ του συνόλου των προβλέψεων που έγιναν.

Μοντέλα πρόβλεψης	**Positive Predictive values 0 (%)**	**False Discovery rate 1 (%)**
Linear SVM	69	
Quadratic SVM	72	59
Cubic SVM	72	62
Fine Gaussian SVM	69	
Medium Gaussian SVM	70	44
Coarse Gaussian SVM	69	
Logistic Regression	70	61
Bagged Trees	69	68

Με βάση τα παραπάνω δεδομένα παρατηρούμε ότι σε γενικές γραμμές τα ποσοστά πρόβλεψης των τιμών που βρίσκονται εκτός των ορίων για την πρόβλεψη των spikes και εδώ δεν αποδίδουν σε ικανοποιητικό βαθμό. Τα μοντέλα φαίνεται να μην μπορούν να προβλέψουν ικανοποιητικά την παρουσία των spikes για μελλοντικό χρόνο. Τα καλύτερα μοντέλα για την πρόβλεψη των τιμών των spikes για τα μηνιαία δεδομένα και εδώ είναι το Quadratic SVM και το Cubic SVM, τα οποία δίνουν να ποσοστό πρόβλεψης των spikes 33% και 37% αντίστοιχα. Τα υπόλοιπα μοντέλα, όπως το Linear SVM, Fine Gaussian SVM, Medium Gaussian SVM και Logistic regression και το Bagged Trees φαίνεται να αδυνατούν να προβλέψουν σωστά τα spikes με τα ποσοστά να κυμαίνονται στο 0% για το Linear SVM, σε 0% για το Fine Gaussian SVM, 4% για το Medium Gaussian SVM, 0% για το Coarse Gaussian, 14% για το Logistic Regression και 10% για το Bagged Trees.

Το αποτέλεσμα αυτό, της χαμηλής πρόβλεψης των spikes, οφείλεται στο γεγονός ότι το target variable είναι imbalanced. Ως εκ τούτου, οι περιπτώσεις όπου τα returns είναι non-spikes είναι πολύ περισσότερες από τις περιπτώσεις όπου εμφανίζονται ως spikes με απόρροια όλα τα μοντέλα στο σύνολό τους, να προβλέπουν τις τιμές ως επί το πλείστο ως non-spike.

3.1.5 Σχόλια

Ο βασικός στόχος της εργασίας αυτής ήταν η πρόβλεψη των ακραίων τιμών των returns του WTI Crude oil με την μορφή spikes. Αναλύοντας και παρατηρώντας όλα τα αποτελέσματα στο σύνολό τους, ενδιαφέρον παρουσιάζει το γεγονός, ότι το μοντέλο του Cubic Support Vector Machine φαίνεται να μπορεί να προβλέψει καλύτερα την εμφάνιση spikes σε όλα τα σετ των δεδομένων. Επιπλέον, τα μοντέλα που εκπαιδεύτηκαν στα μηνιαία δεδομένα, δίνουν καλύτερα ποσοστά πρόβλεψης των spikes συγκριτικά με τα ημερήσια σετ και εβδομαδιαία σετ δεδομένων. Το γεγονός αυτό αποδίδεται στον μικρότερο αριθμό in-sample δεδομένων κατά την φάση της εκπαίδευσης Είναι απαραίτητο να σημειωθεί ότι η πρόβλεψη τέτοιων γεγονότων (ακραίες τιμές) με υψηλή ακρίβεια, είναι μια άκρως απαιτητική και πολύπλοκη διαδικασία (Th Papadimitriou et al, 2020). Οι εμφανίσεις των spikes, παρουσιάζουν σημαντικές διακυμάνσεις, που οφείλονται στην περιοδικότητα των τιμών οι οποίες άλλοτε είναι ακραίες και άλλοτε εμφανίζουν μεγαλύτερη ηρεμία. Έτσι, η χρήση της τυπικής απόκλισης χωρίς τα ανώτατα και κατώτατα όρια είναι ανεπαρκής (Th Papadimitriou et al, 2020).

Αναλύοντας τα ποσοστά του accuracy, τα αποτελέσματα που έδωσαν τα μοντέλα στο άθροισμα των σετ των δεδομένων βγήκαν παρόμοια και κοντινά μεταξύ τους για τα in-sample δεδομένα, δηλαδή κατά το training και στο out-of-sample, δηλαδή κατά το testing. Το αποτέλεσμα αυτό, μας οδηγεί στο συμπέρασμα ότι τα δεδομένα δεν εμφανίζουν overfit. Επιπρόσθετα τα χαμηλά ποσοστά που εμφανίζουν τα μοντέλα στο True Positive Rate για το 1 δηλαδή την πρόβλεψη των spikes, τα μοντέλα βλέπουν κάποια συσχέτιση μεταξύ των δεδομένων και των αποτελεσμάτων και τείνουν να προβλέπουν την κλάση με τη μεγαλύτερη συχνότητα. Έτσι εξαιτίας του γεγονότος ότι στο στάδιο της εκπαίδευσης το μεγαλύτερο μέρος των δεδομένων κατατάσσεται ως non-spike (major class), έτσι τα μοντέλα κατατάσσουν το μεγαλύτερο μέρος των δεδομένων ως non-spikes.

Ακόμα, πρέπει να σημειωθεί, ότι η μηχανική μάθηση τείνει να εντοπίζει μοτίβα και συσχετίσεις μεταξύ των δεδομένων, τα οποία είναι δυσδιάκριτα στο ανθρώπινο μάτι εκ φύσεως. Όταν όμως τα δεδομένα προκύπτουν από τυχαία γεγονότα χωρίς να κρύβουν μέσα τους μοτίβα, τότε οι αλγόριθμοι απλουστεύουν την σχέση των δεδομένων και προχωράνε σε εσφαλμένες προβλέψεις με την μορφή απλούστευσης

των αποτελεσμάτων, όπως ακριβώς συνέβη με τις προβλέψεις των spikes του WTI crude oil.

Συνοψίζοντας, φαίνεται η διαδικασία της πρόβλεψης με τα συγκεκριμένα δεδομένα να είναι τυχαία και να μπορεί να γίνει πρόβλεψη των spikes έως ένα συγκεκριμένο ποσοστό κοντά στο 30% μόνο με την χρήση των returns του WTI. Ακόμα, βάση των ποσοστών που προκύπτουν, φαίνεται ότι η τιμή των returns του WTI crude oil από μόνη της δεν αρκεί να προβλέψει ικανοποιητικά τα spikes ενός τόσο περίπλοκου χρηματιστηριακού προϊόντος. Τα μοντέλα, που χρησιμοποιήθηκαν μπορούν παρόλα αυτά σε γενικές γραμμές να αποτυπώσουν καλύτερα τις μη γραμμικές σχέσεις, που παρατηρούνται κατά την επεξεργασία των δεδομένων σε σχέση με άλλα παραδοσιακά στατιστικά και οικονομετρικά μοντέλα (Th Papadimitriou et al, 2020).Η βελτιστοποίηση των παραμέτρων και η προσθήκη μεταβλητών είναι παρόλα αυτά συχνά μια χρονοβόρος και δαπανηρή διαδικασία, που πιθανώς να μας οδηγήσει ωστόσο σε καλύτερα αποτελέσματα.

4. Συμπεράσματα

Είναι κοινώς αποδεκτό, ότι η μηχανική μάθηση κατακτάει όλο και περισσότερο έδαφος σε μια πληθώρα επιστημονικών τομέων και πτυχών της καθημερινότητα των ανθρώπων. Το γεγονός, ότι η υπολογιστική δύναμη έχει αυξηθεί και τα δεδομένα, που συλλέγονται πλέον είναι περισσότερα από ποτέ, δίνει πρόσφορο έδαφος στο machine learning και στο artificial intelligent να καταστούν χρήσιμα εργαλεία στην βιομηχανία, στον επιστημονικό τομέα, σε κυβερνήσεις και πολλούς άλλους τομείς. Ένας τομέας, που βρίσκει εφαρμογή η πρόβλεψη μελλοντικών τιμών με machine learning είναι τα χρηματιστηριακά προϊόντα. Ειδικά όταν αυτό το χρηματιστηριακό προϊόν είναι το WTI crude oil, προϊόν κεντρικό για τον σχεδιασμό και την ανάπτυξη της οικονομικής δραστηριότητας των κρατών και των επιχειρήσεων. Αναμφίβολα, η δυνατότητα της πρόβλεψης της τιμής ενός τέτοιου προϊόντος, επιτρέπει την δυνατότητα στους επενδυτές, τις εταιρίες και τα κράτη να σχεδιάζουν καλύτερα και πιο αποτελεσματικά τις μελλοντικές τους κινήσεις ή ακόμα να πουλάνε και να αγοράζουν μετοχές διαμορφώνοντας την αξία του. Έτσι για τον ίδιο λόγο, είναι δέουσας σημασίας να υπάρχει ένας αξιόπιστος τρόπος πρόβλεψης των ακραίων τιμών των returns του WTI Crude oil υπό τη μορφή των spikes, ώστε η πληροφορία αυτή να αποτελέσει πολύτιμο εργαλείο ανάλυσης και αξιολόγησης της τιμής του WTI crude oil. Αξίζει να σημειωθεί, βέβαια, ότι το αργό πετρέλαιο και στην προκειμένη περίπτωση το WTI crude oil, αποτελεί ένα περίπλοκο χρηματιστηριακό προϊόν, η τιμή του οποίου επηρεάζεται από την παραγωγή του, την ζήτηση του, την εποχικότητα, τις γεωπολιτικές εξελίξεις και πολλά άλλα.

Συνεπώς, εξαιτίας της σπουδαιότητας του αργού πετρελαίου, επιχειρήθηκε η πρόβλεψη των ακραίων τιμών υπό την μορφή spikes των returns του WTI crude oil, και πιο συγκεκριμένα για ομάδες δεδομένων για ημερήσιες, εβδομαδιαίες και μηνιαίες τιμές.. Τα δεδομένα ξεκινάνε από το 1986 και σταματάνε στο 2023, ενώ ο βασικός στόχος ήταν η πρόβλεψη των ακραίων τιμών υπό την μορφή των spikes, που προέκυψαν από τα returns της τιμής του πετρελαίου. Ο ορίζοντας πρόβλεψης ορίστηκε t+1 ενώ υπολογίστηκε ένα ανώτατο όριο S+ και ένα κατώτατο όριο S- για τις τιμές των returns. Το σύνολο των δεδομένων για τα 3 σετ δεδομένων χωρίστηκε σε δύο υποκατηγορίες, τα δεδομένα εκπαίδευσης (90%) με στόχο να εκπαιδεύσουν την πρόβλεψη στα επιλεγμένα μοντέλα και τα δεδομένα ελέγχου (10%) με στόχο τον έλεγχο της ικανότητας τον μοντέλων να προβλέψουν σωστά δεδομένα, που δεν

χρησιμοποιήθηκαν κατά το στάδιο της εκπαίδευσης. Τα μοντέλα, που επιλέχτηκαν για την εκπαίδευση των δεδομένων ήταν μοντέλα Support Vector Machine (SVM), όπως Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM και τα μοντέλα Logistic Regression και Bagged Trees. Τα αποτελέσματα, που έδωσαν τα μοντέλα στο σύνολό του και για τα τρία σετ δεδομένων, οδηγούν στο συμπέρασμα, ότι τα autoregressive μοντέλα δεν μπορούν να χρησιμοποιηθούν παρελθοντικές τιμές για να προβλέψουν σωστά τα μελλοντικά spikes χωρίς την χρήση και άλλων δεικτών. Η πρόβλεψη των ακραίων τιμών του αργού πετρελαίου υπό την μορφή των spikes, δηλαδή των ακραίων διακυμάνσεων της τιμής του, φαίνεται να είναι αδύνατη μόνο με την πληροφορία των παρελθοντικών τιμών του αργού πετρελαίου. Η διαδικασία της πρόβλεψης της εμφάνισης των spikes με μοναδική πληροφορία μόνο της τιμή των returns, φαίνεται να είναι τυχαία, ενώ οι ακραίες διακυμάνσεις με τη μορφή spikes μπορεί να προβλεφθούν σε ποσοστό έως 30%.

Από την έρευνα αυτή, λοιπόν, και με βάση όσα έχουν προαναφερθεί, καταλήγουμε στο συμπέρασμα, ότι για να γίνει επιτυχημένα η πρόβλεψη των ακραίων διακυμάνσεων των τιμών του αργού πετρελαίου, θα πρέπει να ληφθούν υπόψιν και άλλες μεταβλητές, όπως η τιμή του φυσικού αερίου, η τιμή άλλων πετρελαϊκών δεικτών, η τιμή του S&P500 κτλ, με στόχο να μετριαστεί μέχρι κάποιο βαθμό η τυχαιότητα και η μεγάλη επιρρέπεια, που φαίνεται να εμφανίζει η τιμή του αργού πετρελαίου εξαιτίας διάφορων γεγονότων.

Περαιτέρω, με βάση τη βιβλιογραφία, φαίνεται ότι τα μοντέλα autoregressive ARSVM υπερτερούν των μοντέλων random walk (RW), όταν όμως οι προβλέψεις συνδέονται με μεταβλητές, όπως την τιμή του δολαρίου (Dimitriadou A., et al, 2018). Ομοίως, τα αποτελέσματα πρόβλεψης των spikes για την αγορά της ηλεκτρικής ενέργειας έδειξαν ότι το βέλτιστο προγνωστικό μοντέλο το οποίο χρησιμοποίησε την RBF Kernel, απέδωσε σε συνδυασμό με τη χρήση lags του S&P500 και του δείκτη NYSE (E Stathakis et al, 2017). Συμπεραίνουμε, λοιπόν, για αξιόπιστα αποτελέσματα δεν αρκεί η επεξεργασία μόνο των τιμών των returns του πετρελαίου, αλλά αυτή πρέπει να γίνει σε συνδυασμό με άλλους δείκτες που επηρεάζουν άμεσα ή έμμεσα τους πρώτους. Οι (Zilin Xu, et al, 2023) προτείνουν την χρήση αλγορίθμων νευρωνικών δικτύων με τη χρήση συγκεκριμένων βαρών. Χαρακτηριστική είναι η χρήση 14 διαφορετικών μεταβλητών όπως ο χρυσός, το ασήμι, ο S&P500, την ισοτιμία του δολαρίου κτλ.

Επιπροσθέτως, κατά τη χρήση μοντέλων SVM, αποτελεί πρόκληση η δυαδική ταξινόμηση μεγάλων σε αριθμό δεδομένων, όπου χαρακτηρίζονται ως unbalance και τα δεδομένα παρουσιάζουν υψηλή ευαισθησία. Τα μοντέλα αυτά για να εμφανίσουν καλή ακρίβεια, πρέπει να εφαρμοστούν βάρη στα δεδομένα για να μειωθούν οι λάθος προβλέψεις (Tatjana Eitrich et al, 2005).Σε γενικές γραμμές ο μεγάλος όγκος των δεδομένων δημιουργεί προβλήματα σε μοντέλα SVM, σε συνδυασμό με τον θόρυβο που ενδεχομένως εμφανίζουν.

Ενώ η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για τη δημιουργία προγνωστικών μοντέλων για τις χρηματοπιστωτικές αγορές, συμπεριλαμβανομένων των τιμών του αργού πετρελαίου, είναι σημαντικό να κατανοήσουμε τους περιορισμούς και τις αβεβαιότητες που ενυπάρχουν σε τέτοιες προβλέψεις. Τα μοντέλα σε κάθε περίπτωση θα πρέπει να ενημερώνονται και να επικυρώνονται συνεχώς και θα πρέπει να επιλέγονται σε γενικές γραμμές για βραχυπρόθεσμες ή συγκεκριμένες στρατηγικές παρά για μακροπρόθεσμη πρόβλεψη τιμών.

Πρόσφορο πεδίο για μελλοντική έρευνας στο συγκεκριμένο θέμα της πρόβλεψης των τιμών του πετρελαίου, παρουσιάζει η εξέταση των τιμών του WTI crude oil σε συνδυασμό με άλλα benchmark όπως το BRENT, καθώς τα παραπάνω είναι δύο από τους βασικότερους χρηματιστηριακούς δείκτες πετρελαίου. Ακόμη, αξιόλογο ενδιαφέρον παρουσιάζει η σύγκριση των τιμών των spot prices με τα future του WTI, αλλά και η σύνδεση τους με δείκτες, όπως ο χρυσός, το φυσικό αέριο, αλλά και το ίδιο το δολάριο. Τα future διαδραματίζουν σημαντικό ρόλο για τις προσδοκίες των επενδυτών σχετικά με τις μελλοντικές τιμές του αργού πετρελαίου και η σύνδεση με τα returns του spot price αναμένεται να αποδώσει πιο αξιόπιστα αποτελέσματα. Τέλος, προτείνεται και η χρήση άλλων οικονομετρικών μοντέλων, όπως το ARIMA ή άλλα υβριδικά μοντέλα GARCH.

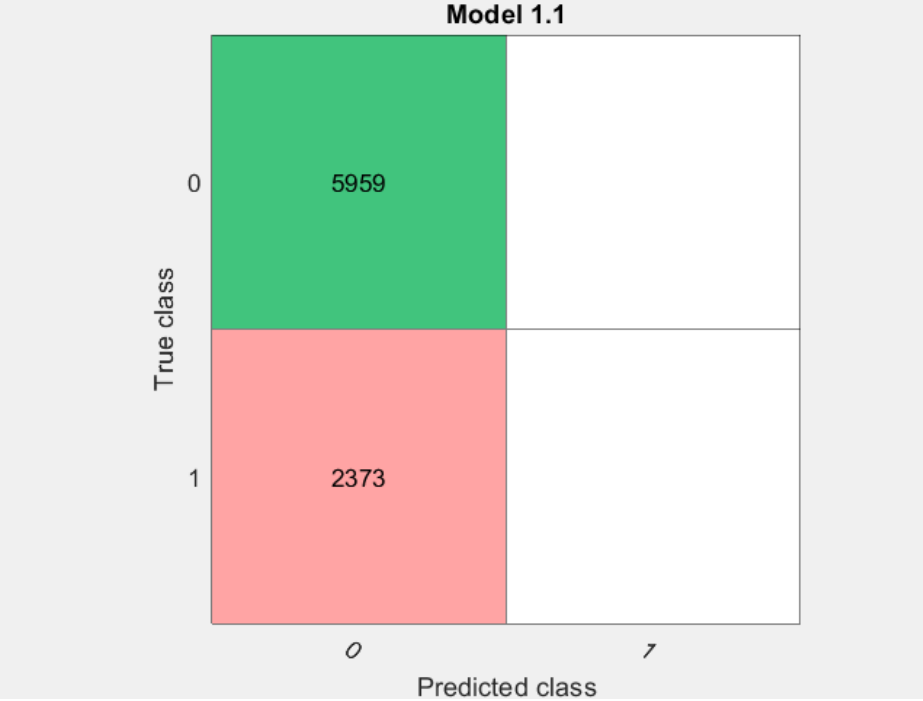
Βιβλιογραφία

- Adebayo Toniwa Sunday, Dervis Kirikkaleli, Ibrahim Adeshola, Dokum Oluwajana, Gbenga Daniel Akinsola, Oseyenbhin Sunday Osemeahon. (2021). Coal consumption and environmental sustainability in South Africa: the role of financial development and globalization.Int. Journal of Renewable Energy Development 10 (3) 2021: 527-536
- Bhoopesh Singh Bhati, C. R. Ral (2019). Analysis of Support Vector Machine-based Intrusion Detection Techniques. Arabian Jurnal for Science and Engineering.
- Chang L et al, Q. C. (2022). Nexus between financial development and renewable energy: empirical evidence from nonlinear autoregression distributed lag.
- Clifton D. Sutton, (2005). Classification and Regression Trees, Bagging, and Boosting.Handbook of Statistics, Vol. 24
- Dimitriadou Athanasia, Periklis Gogas, Theophilos Papadimitriou, Vasilios Plakandaras, 2018. *Oil Market Efficiency under a Machine Learning Perspective.Forecasting 2018, 1, 157–168; doi:10.3390/forecast1010011*
- Efthymios Stathakis, Theophilos Papadimitriou, Periklis Gogas (2017). Forecasting electricity price spikes using suport vector machines.
- Efthymios Stathakis, Theophilos Papadimitriou, Periklis Gogas, (2020). Forecasting Price Spikes in Electricity Markets. *Rview of Economic Analysis forthcoming 12*.
- Edeh Michael Onyema, Khalid K. Almuzaini, Fergus Uchenna Onu, Devvret Verma, Ugdoja Samuel Gregory, Monika Puttaramaiah, Rockson Kwasi Afriyie,(2022). Prospects and Challenges of Using Machine Learning for Academic Forecasting.Computational Intelligence and NeuroscienceVolume 2022, Article ID 5624475, 7 pages.
- Enke D et al,(2005). The use of data mining and neyral networks for forecasting stock market returns. Exprert systems with Application .
- Genpact. (2022). The evolution of forecasting techniques : Traditional versus machine learning methods.
- Gogas Periklis, Theophilos Papadimitriou, Emmanouil Sofianos, (2021). Forecasting unemployment in the euro area with machine learning.Journal of Forecasting. 2021;1–16.
- H. Hotelling, 1931. *The economics of exhaustible resources, The Journal of Political Economy, Volume 39, Number 2* .

- J. D. Hamilton, 1983. *Oil and the Macroeconomy since World War II, Journal of Political eonomy, Vol 91, No 2, University of Chicago Press.*
- javatpoint.com. Major Kernel Functions in Support Vector Machine.
- Moting Su, Zongyi Zhang, Ye Zhu, Donglan Zha, Wenying Wen, (2019). Data Driven Natural Gas Spot Price Prediction Models Using Machine Learning Methods. energies
- M.T. Mitchell, M. I. Jordan. (2015). Machine learning: Trends, perspectives and prospects. Science VOL 349 ISSUE 6245
- Nasteski, Vladimir (2017). An overview of the supervised machine learning methods.
- neptune.ai.com
- Theophilos Papadimitriou, Periklis Gogas, Athanasios Fotios Athanasiou, (2020). Forecasting S&P 500 spikes: an SVM approach. Spiegel, Digital Finance (2020) 2:241–258
- Wang Jue, George Athanasopoulos, Rob J. Hyndman, Shouyang Wang, (2018). Crude oil price forecasting based on internet concern using extreme learning machine.International Journal of Forecasting 34 (2018) 665–677.
- Yaoyuan V. Tan, Jason Roy (2019). Bayesian additive regression trees and the General BART model. Statistics in Medicine. 2019;1–22.https://doi.org/10.1002/sim.8347
- Yu L et al, W. S. (2008). Forcasting crude oil price with an EMD-based neural network ensemble learning paradigm.
- Zhang D et al, (2021). Public spending and green economic growth in BRI region: mediating role of green finance.
- Zhao Y et al, (2017). A deep learning ensemble approach for crude oil price forecasing.
- Zilin Xu,Muhammad Mohsin, Kaleem Ullah, Xiaoyu Ma, 2023. *Using econometric and machine learning models to forecast crude oil prices: Insights from economic history. Resourse Policy, 83 (2023) 103614*

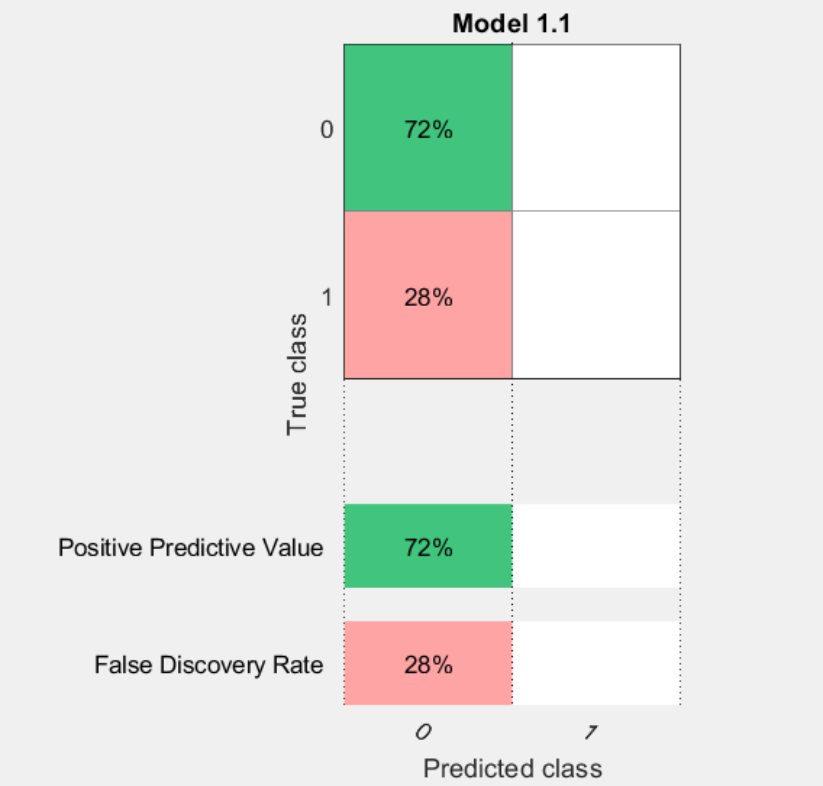
Ημερήσια δεδομένα

Ι.1 Αποτελέσματα με τη μέθοδο Linear SVM



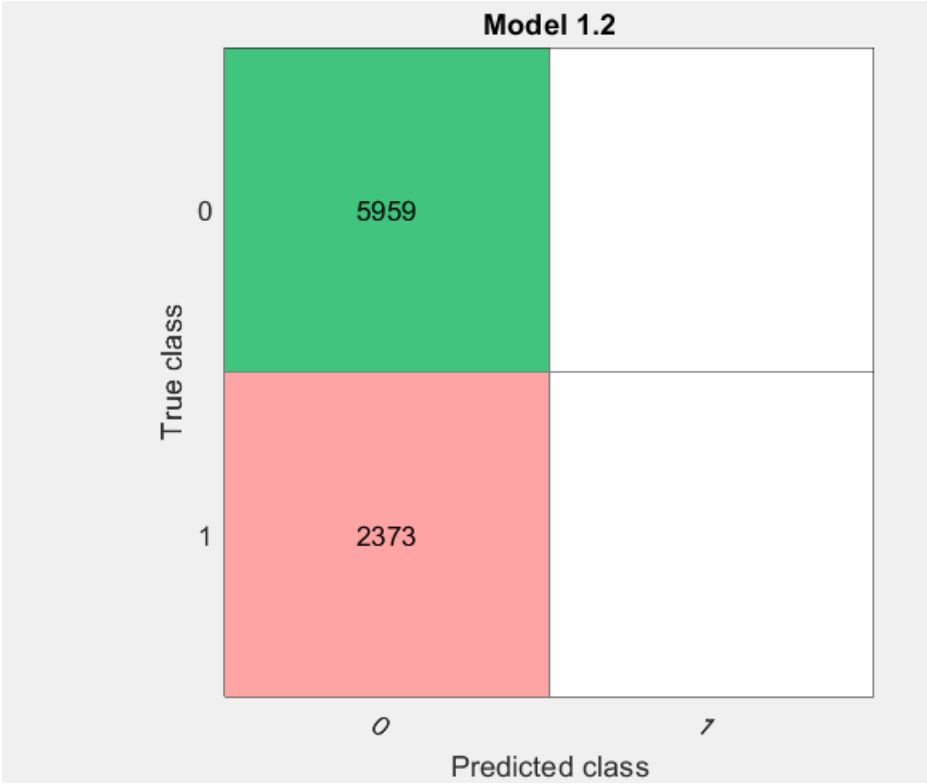
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.



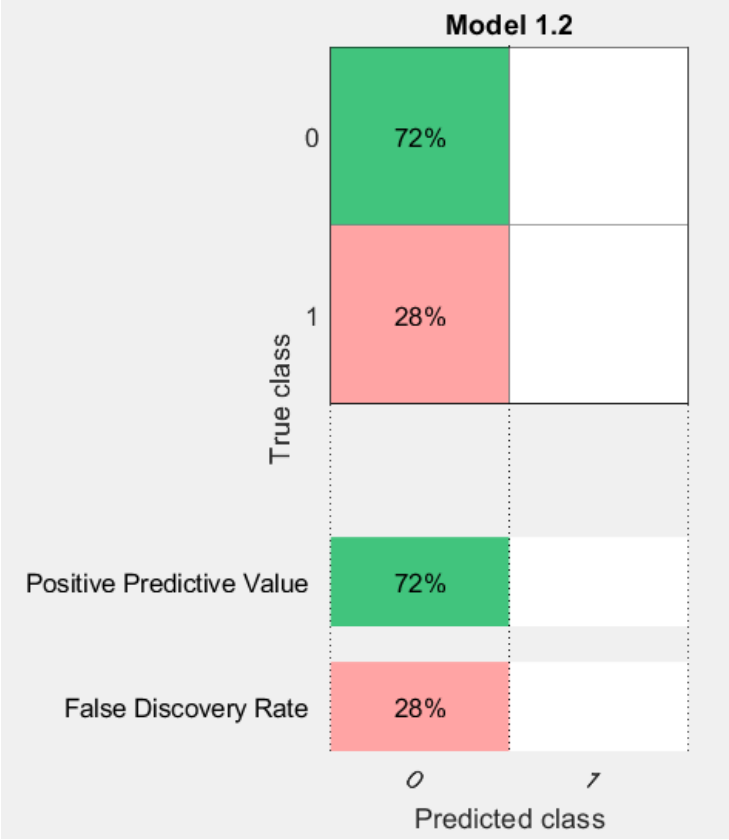
γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

I.2 Αποτελέσματα με τη μέθοδο Quadratic SVM



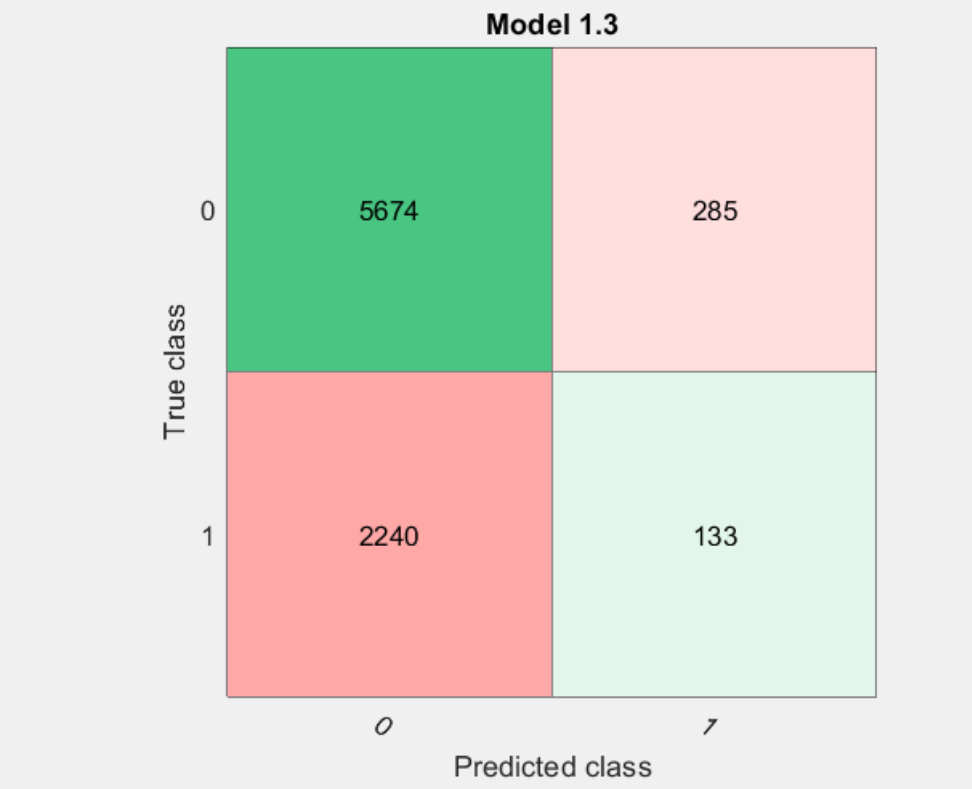
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

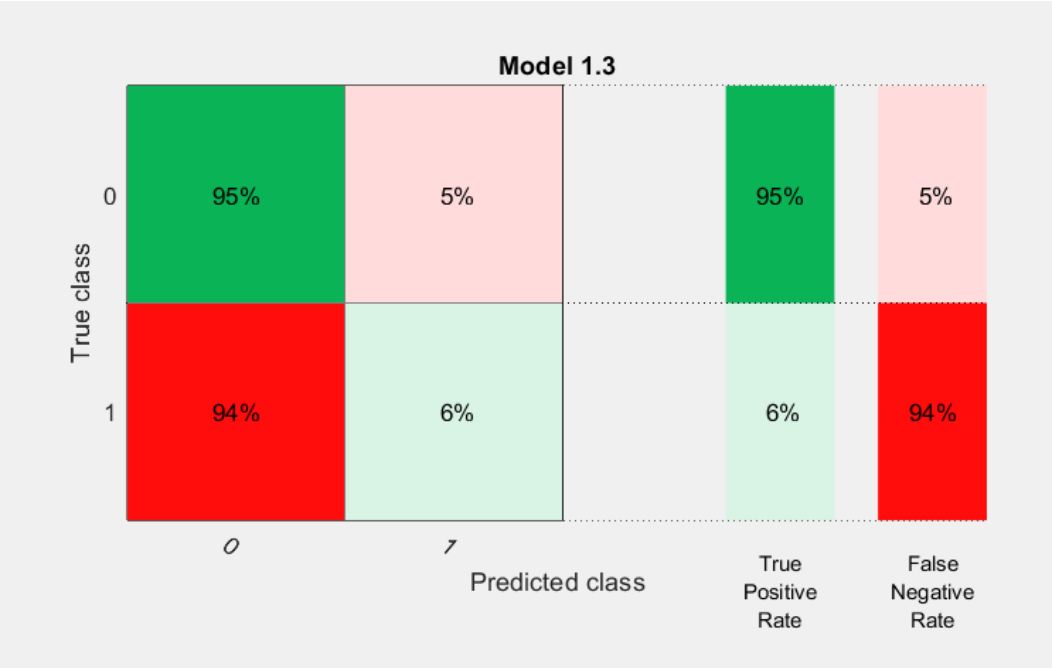


γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

I.3 Αποτελέσματα με τη μέθοδο Cubic SVM



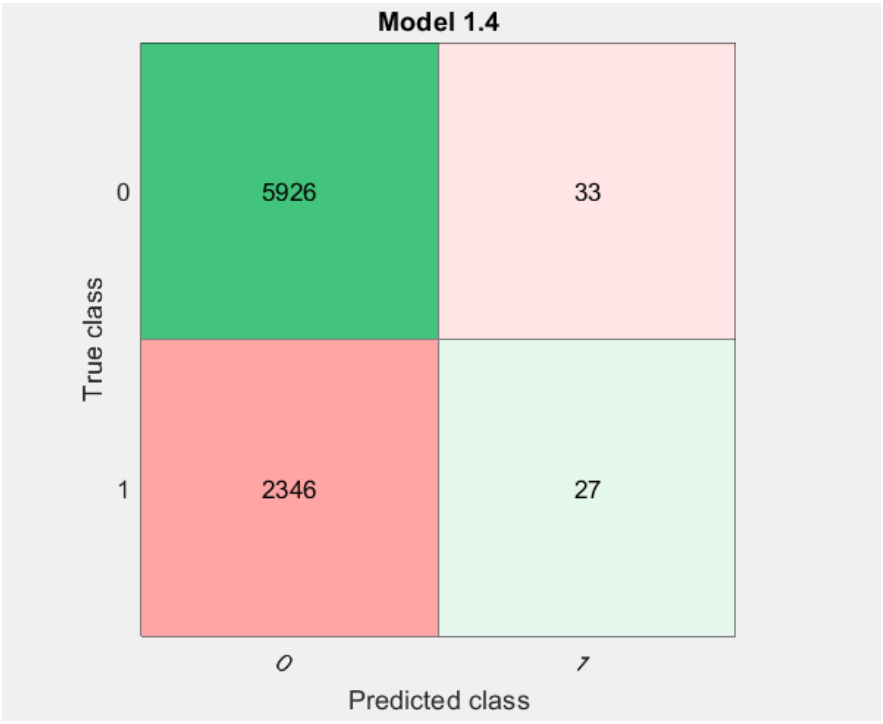
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.



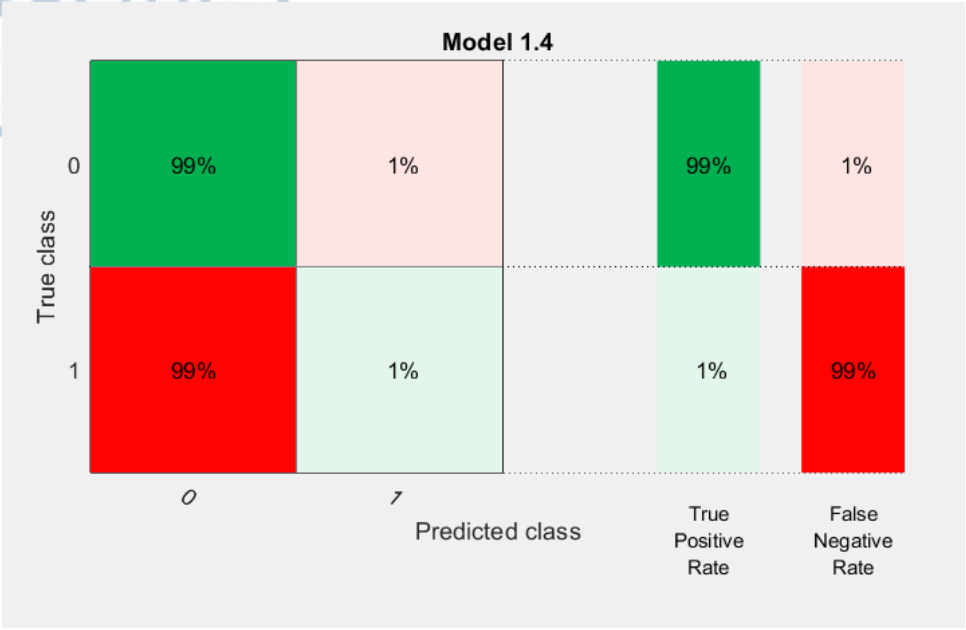
β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

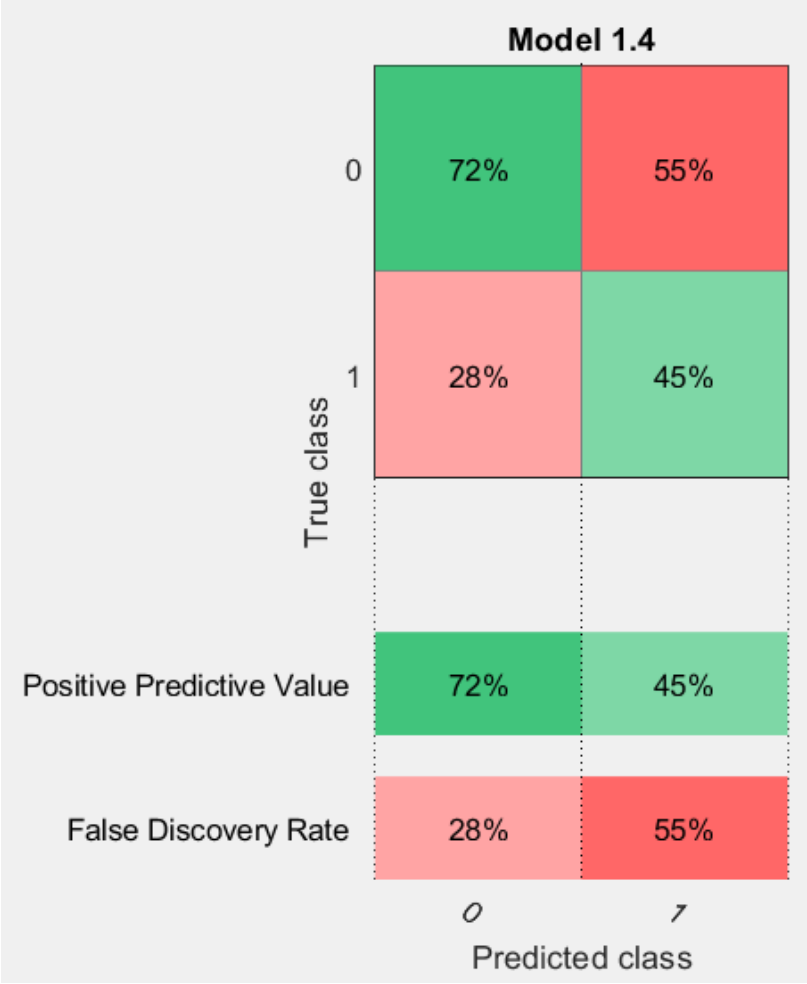
I.4 Αποτελέσματα με τη μέθοδο Fine Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

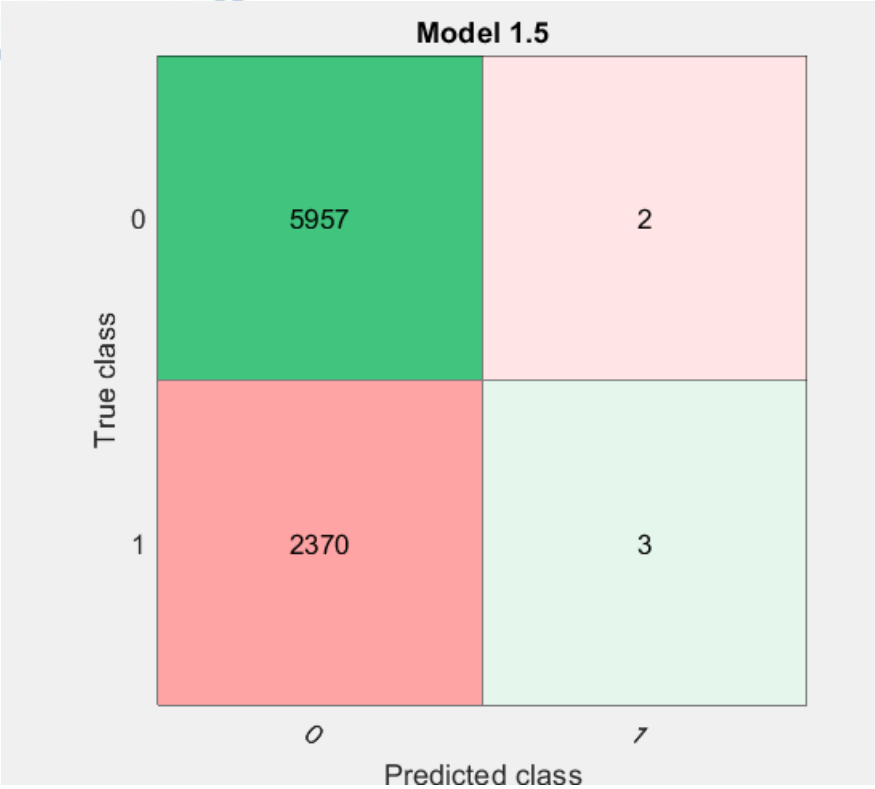


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

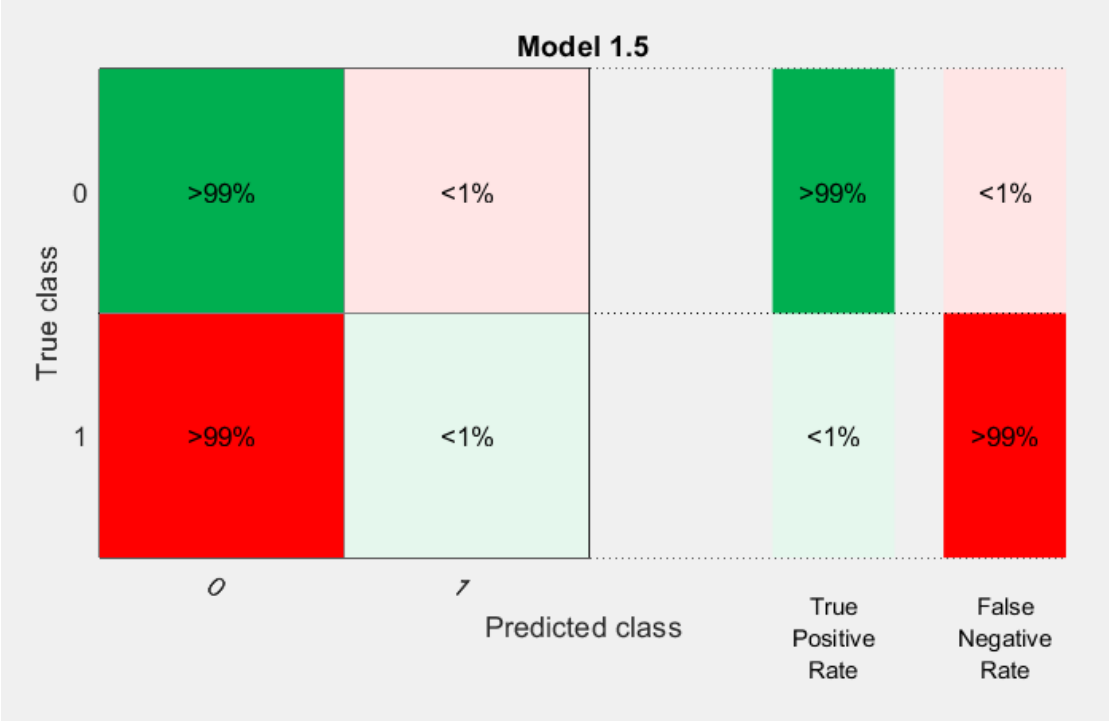


γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

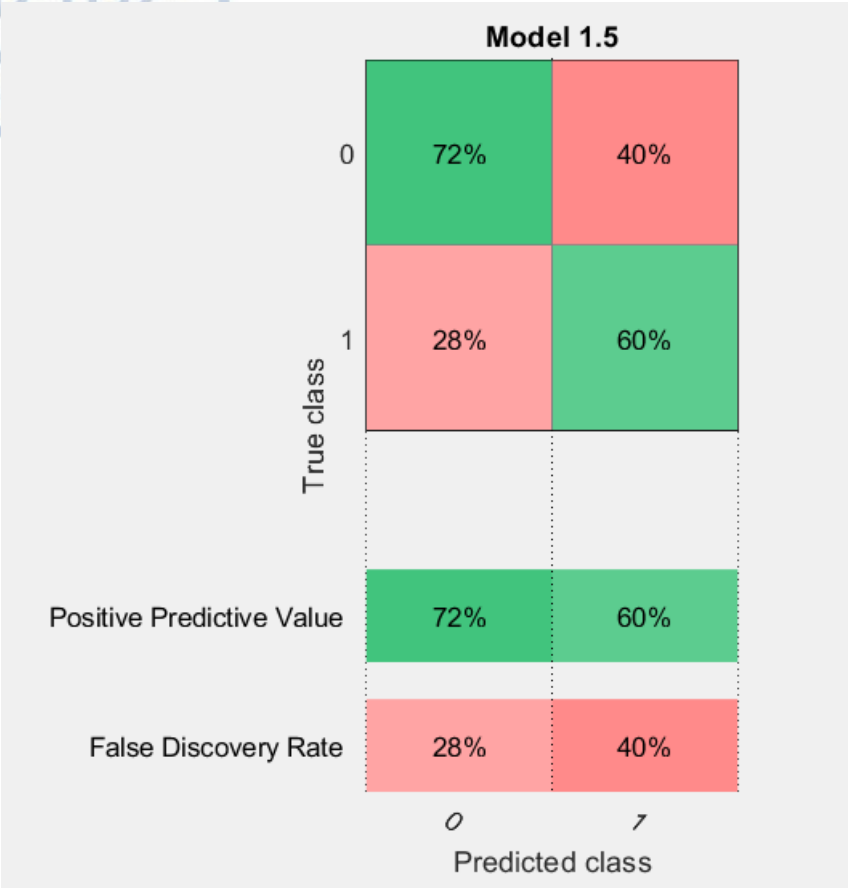
I.5 Αποτελέσματα με τη μέθοδο Medium Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

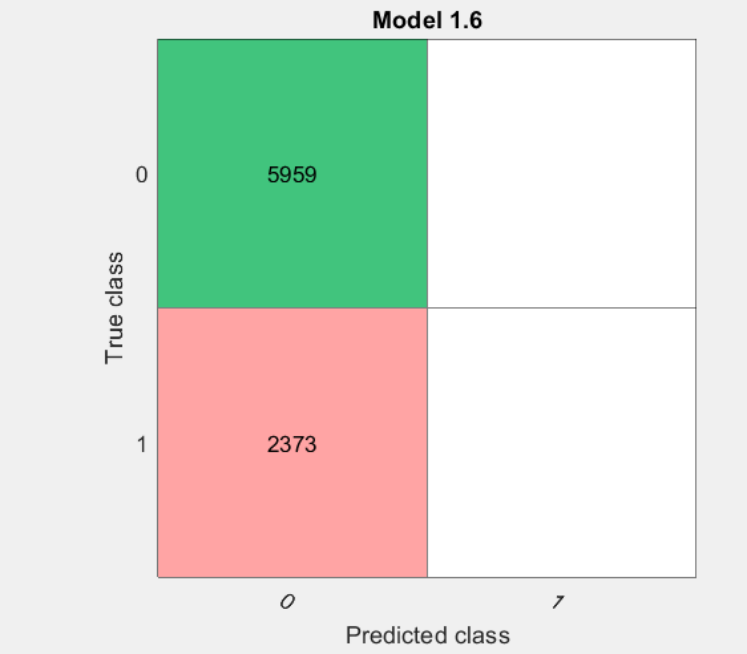


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

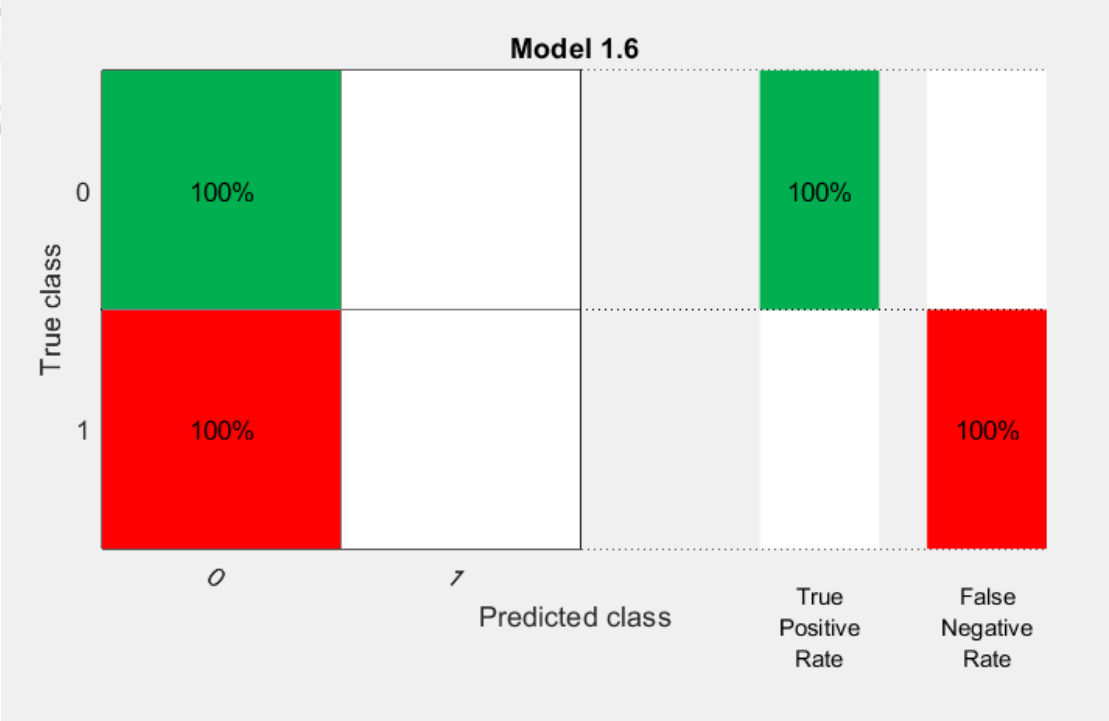


γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

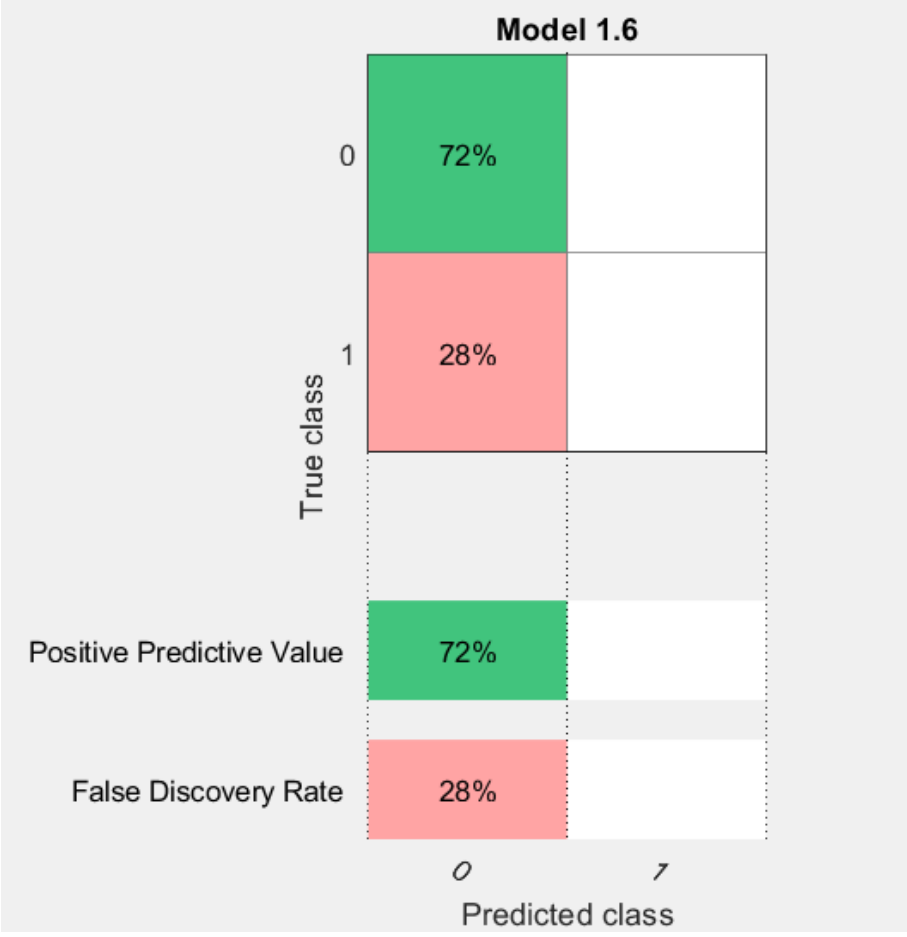
I.6 Αποτελέσματα με τη μέθοδο Coarse Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

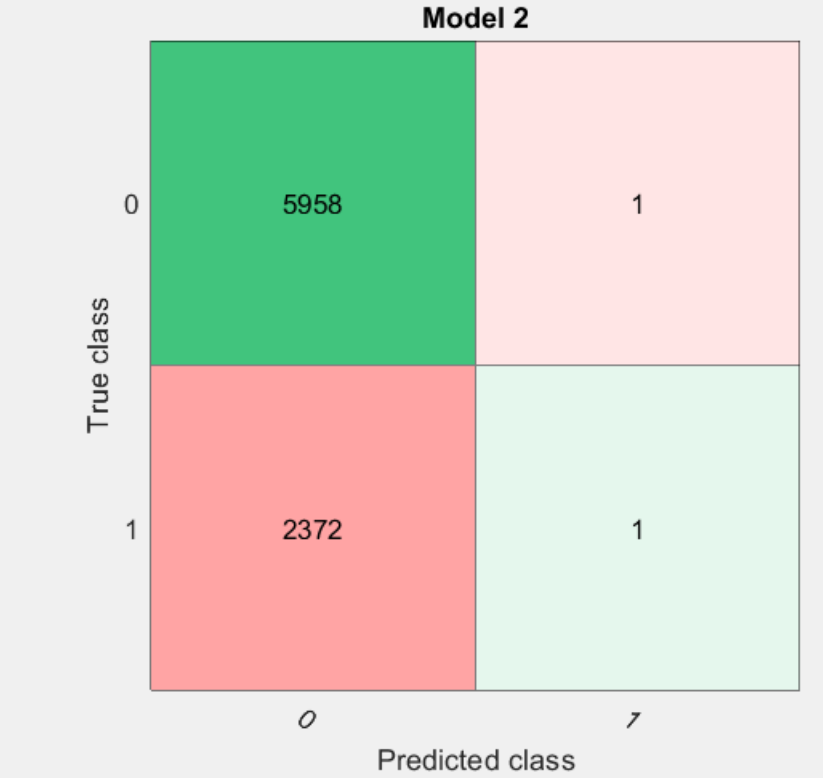


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.



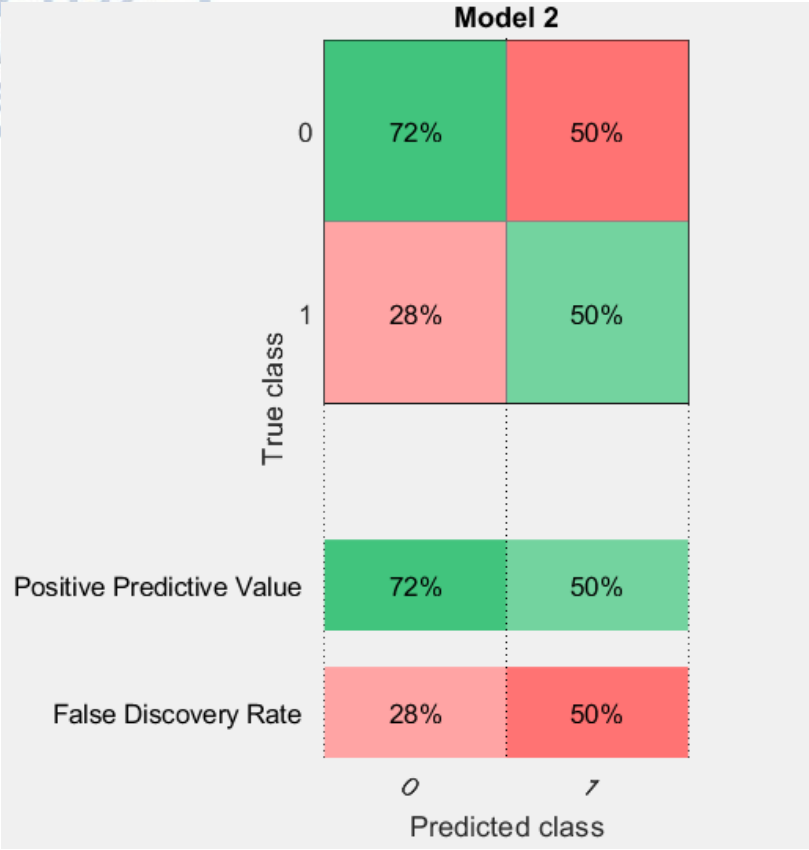
γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

I.7 Αποτελέσματα με τη μέθοδο Logistic Regression



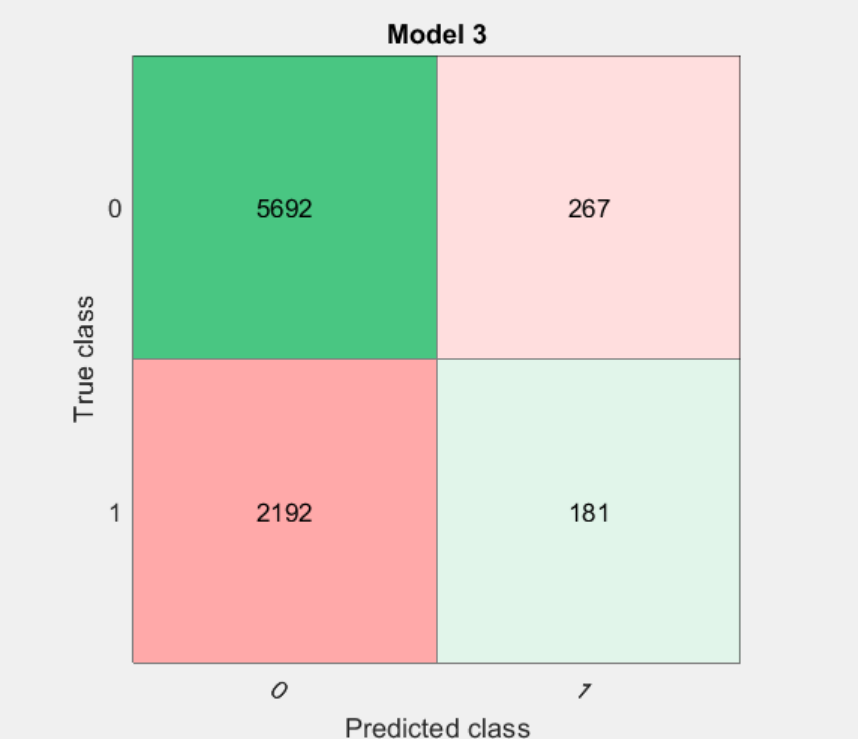
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

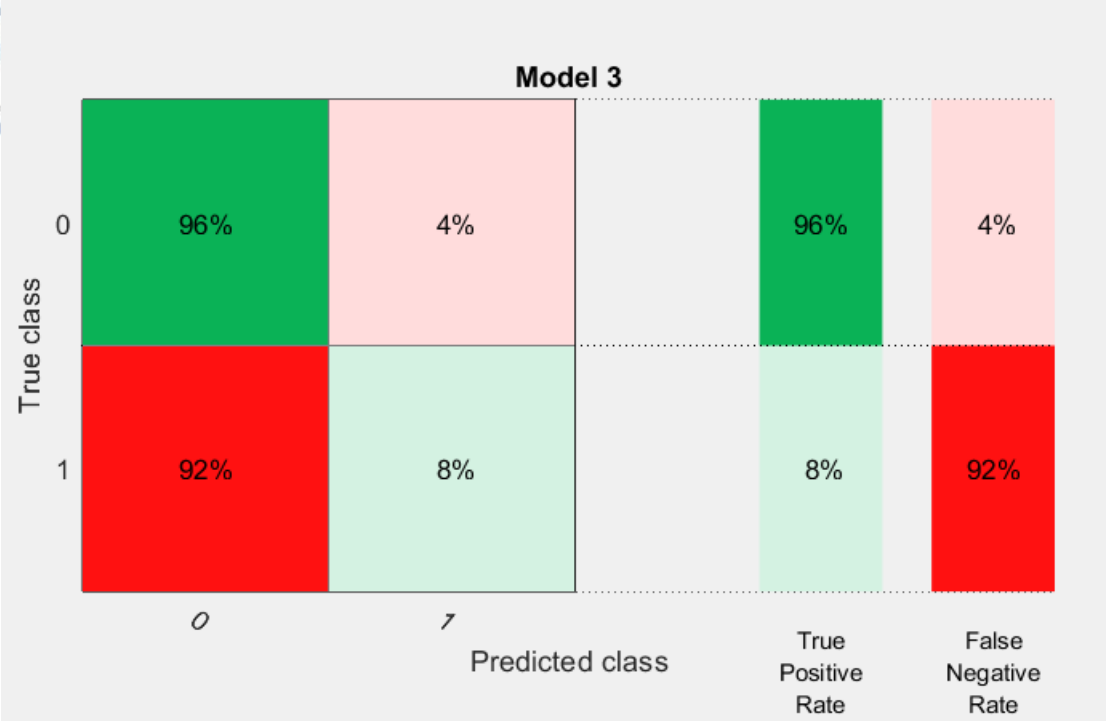


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

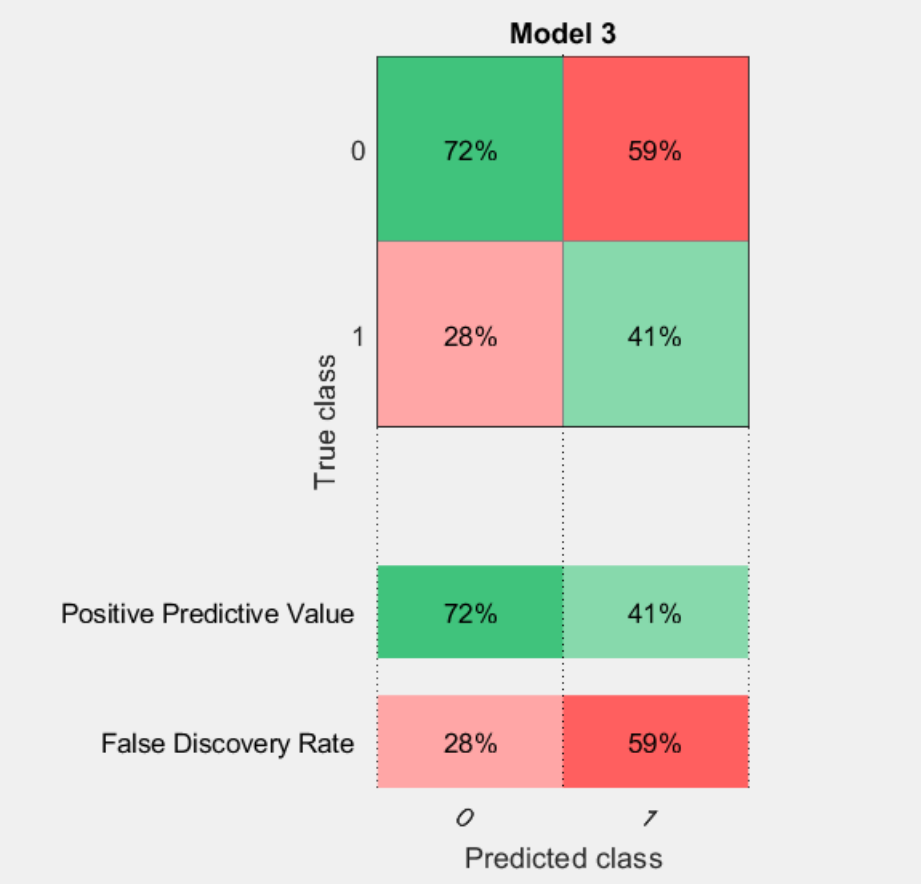
I.8 Αποτελέσματα με τη μέθοδο Bagged Trees



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.



β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

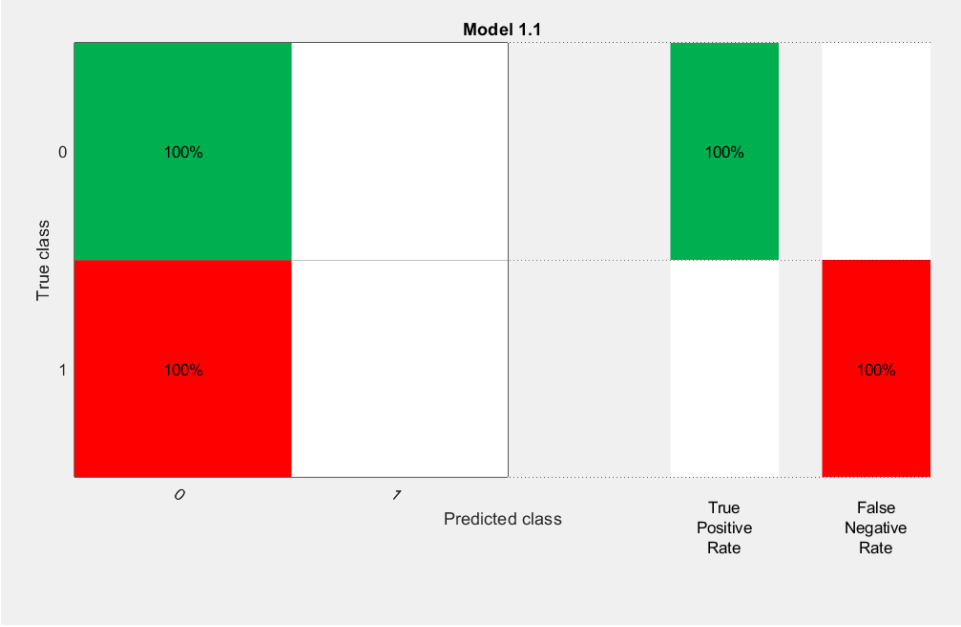


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

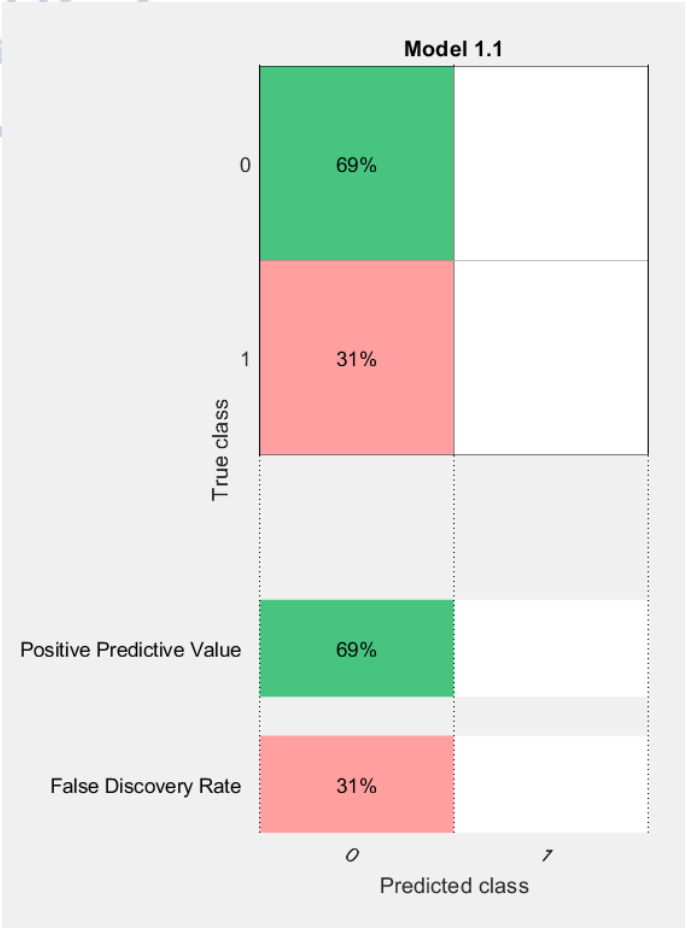
Εβδομαδιαία δεδομένα

ΙΙ.1 Αποτελέσματα με τη μέθοδο Linear SVM

α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

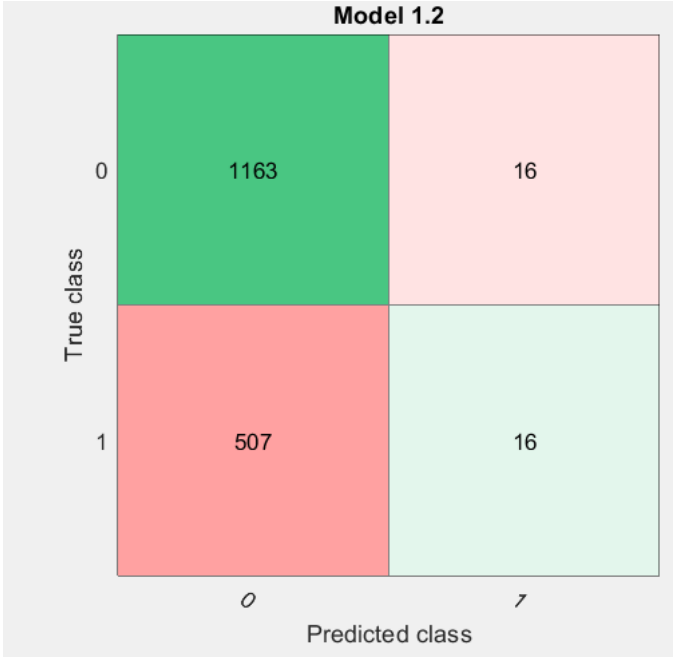


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

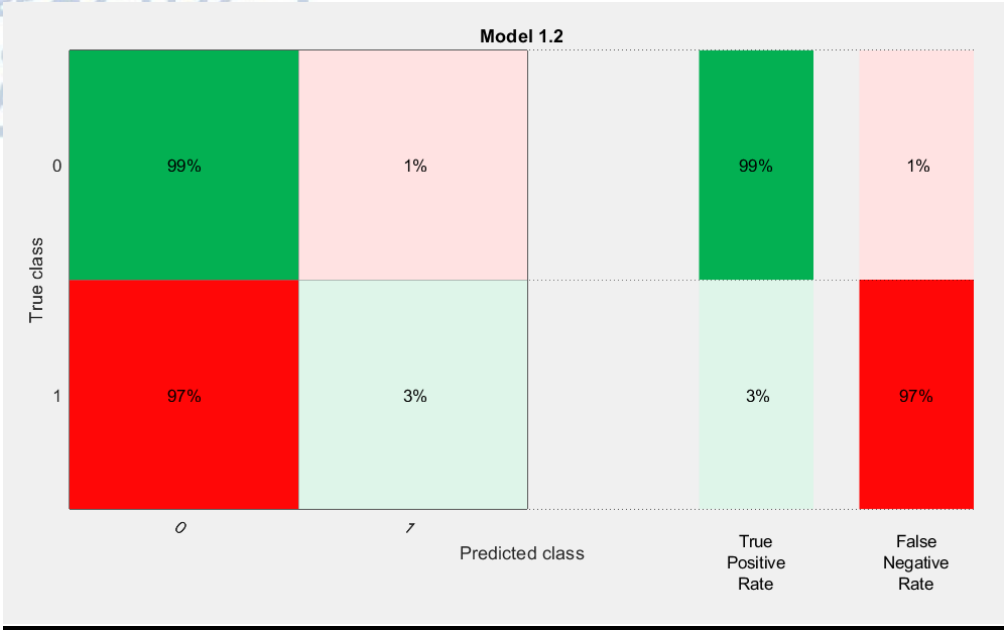


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

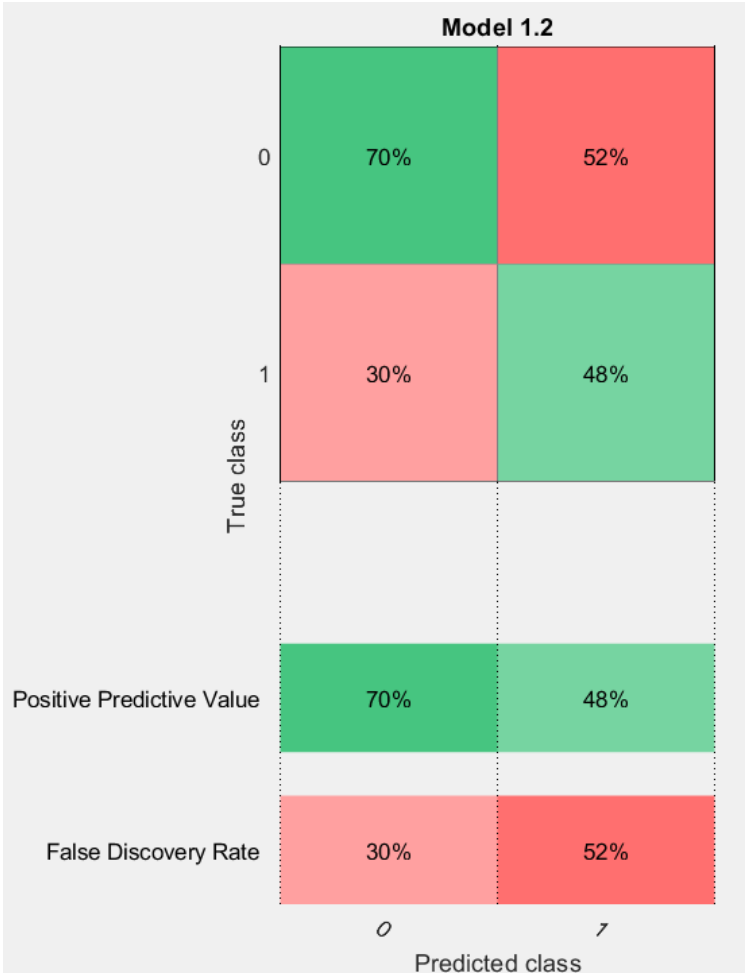
ΙΙ.2 Αποτελέσματα με τη μέθοδο Quadratic SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

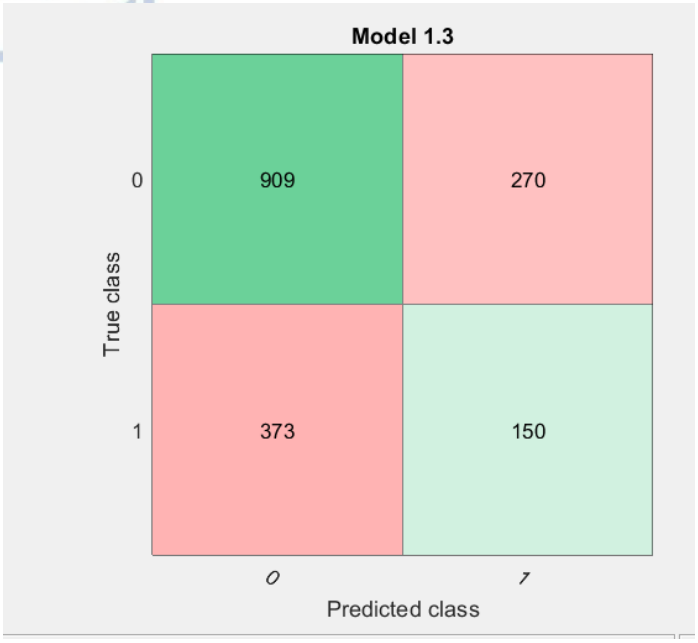


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

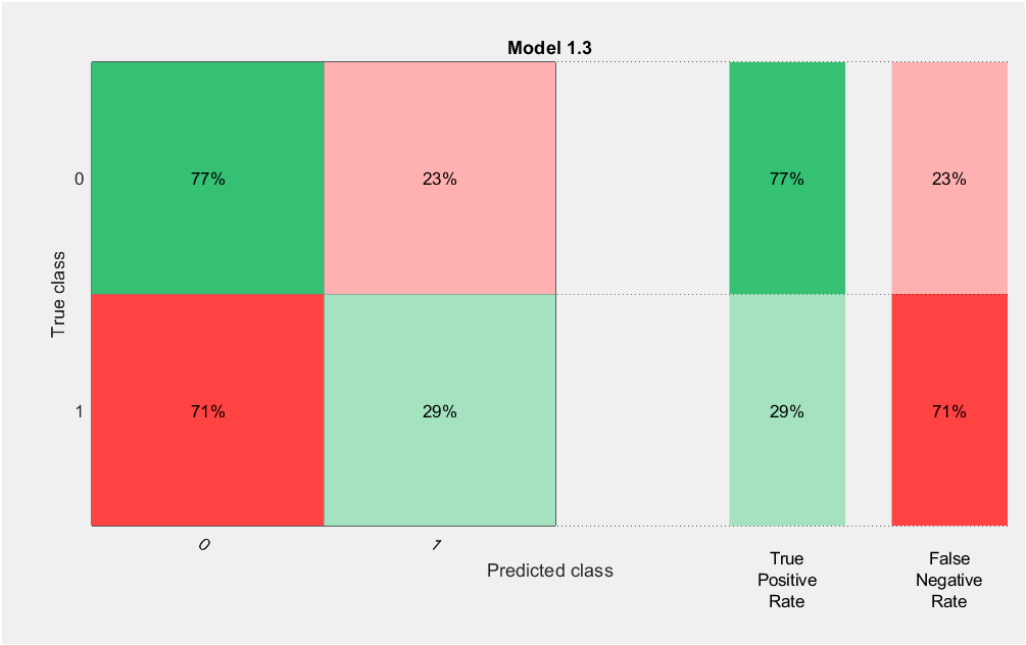


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

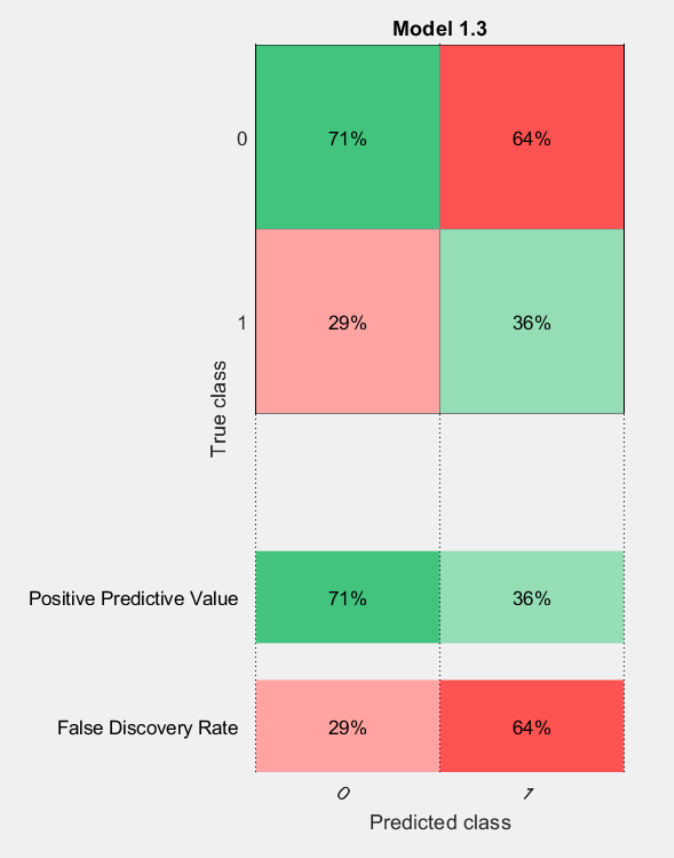
ΙΙ.3 Αποτελέσματα με τη μέθοδο Cubic SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

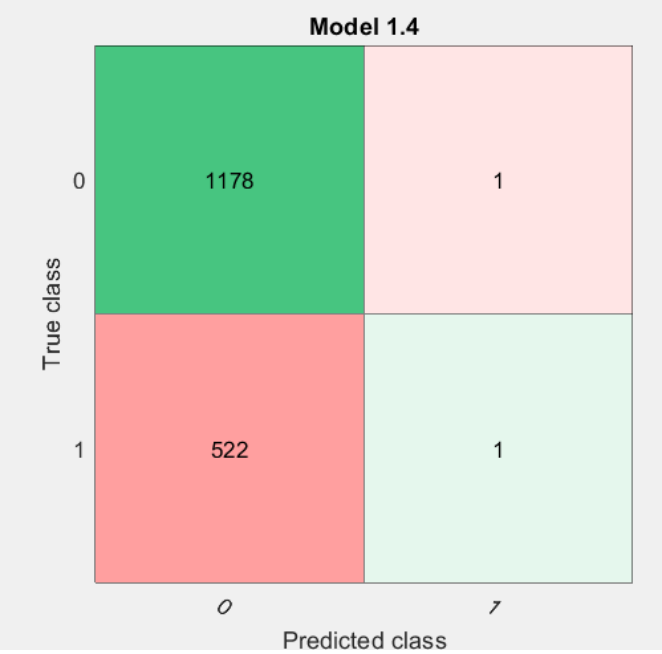


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

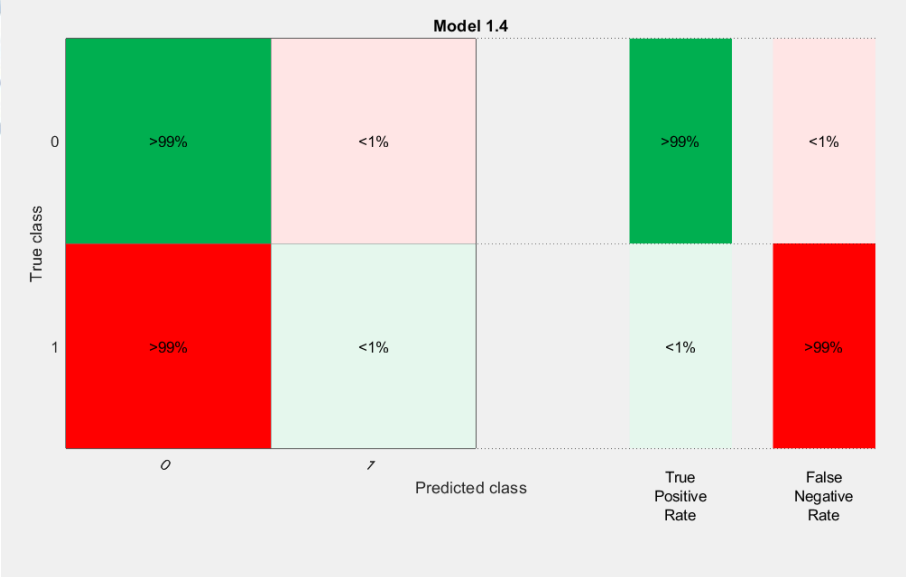


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

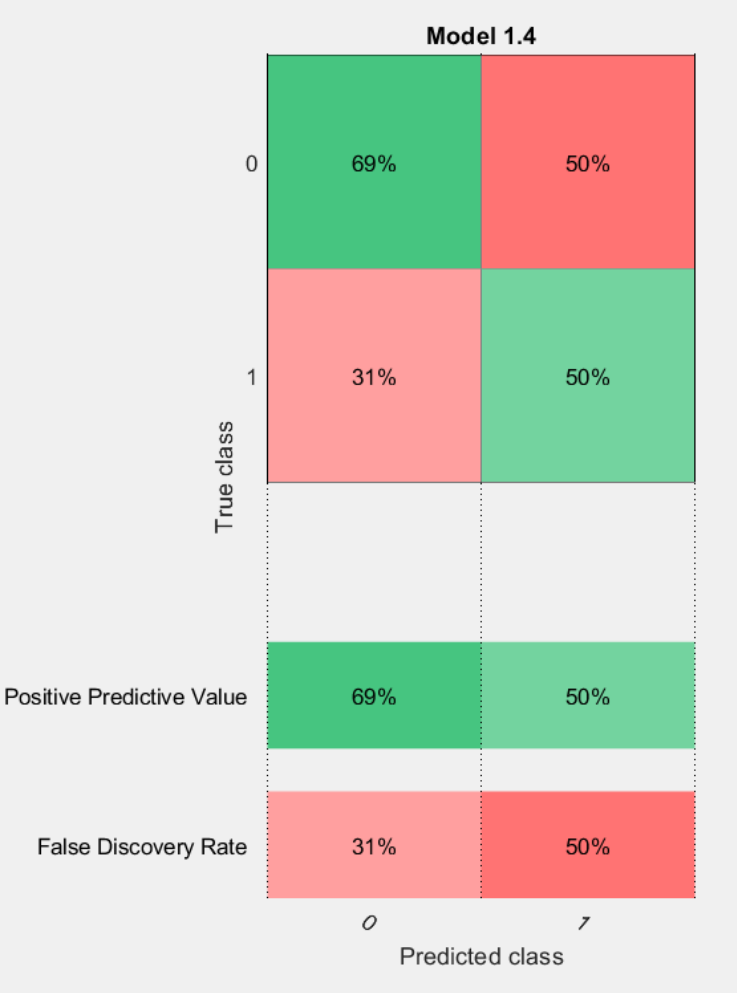
ΙΙ.4 Αποτελέσματα με τη μέθοδο Fine Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

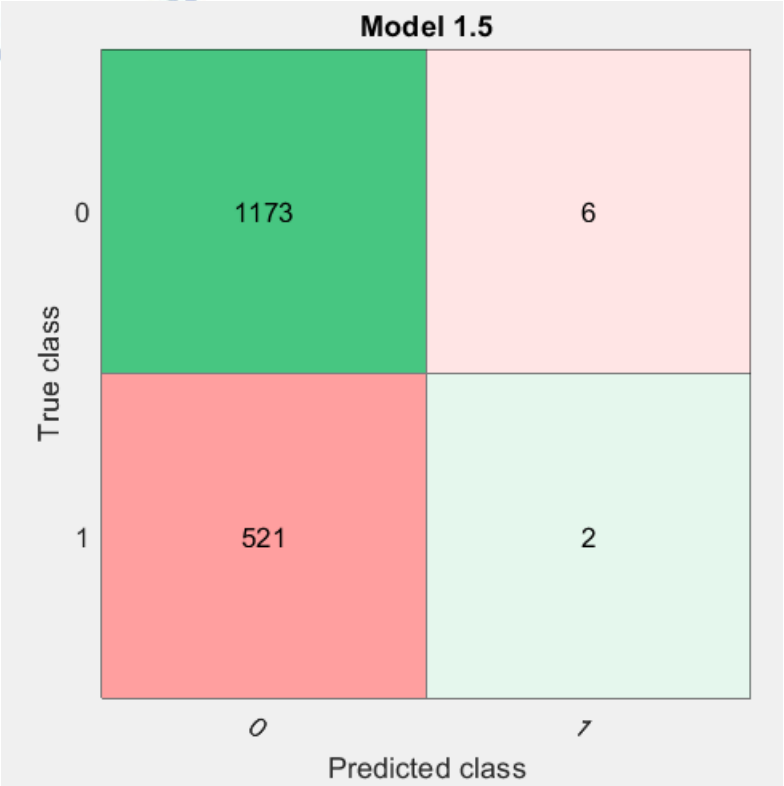


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

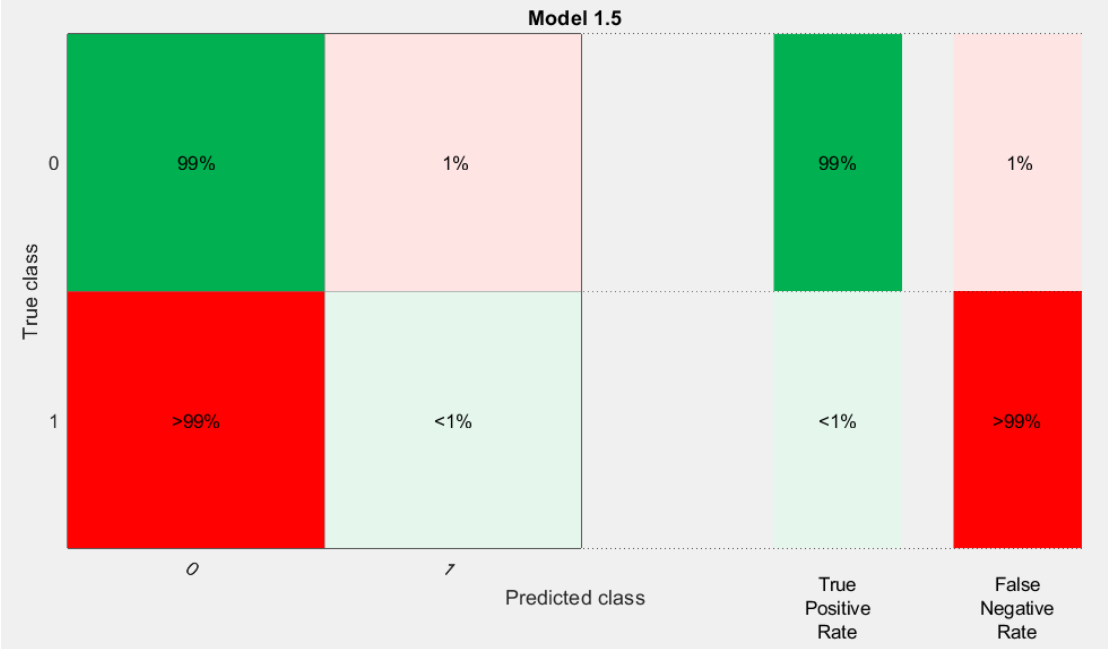


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

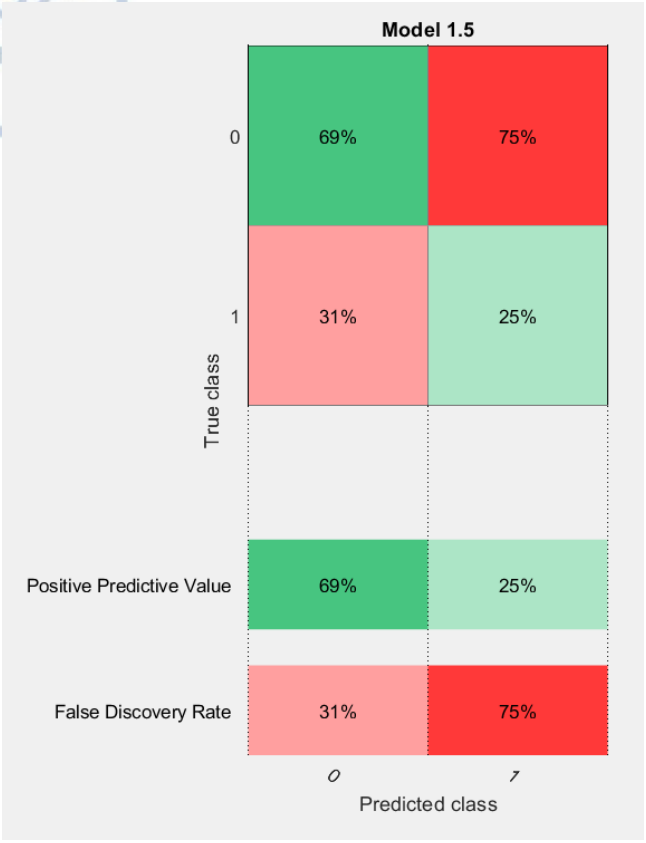
ΙΙ.5 Αποτελέσματα με τη μέθοδο Medium Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

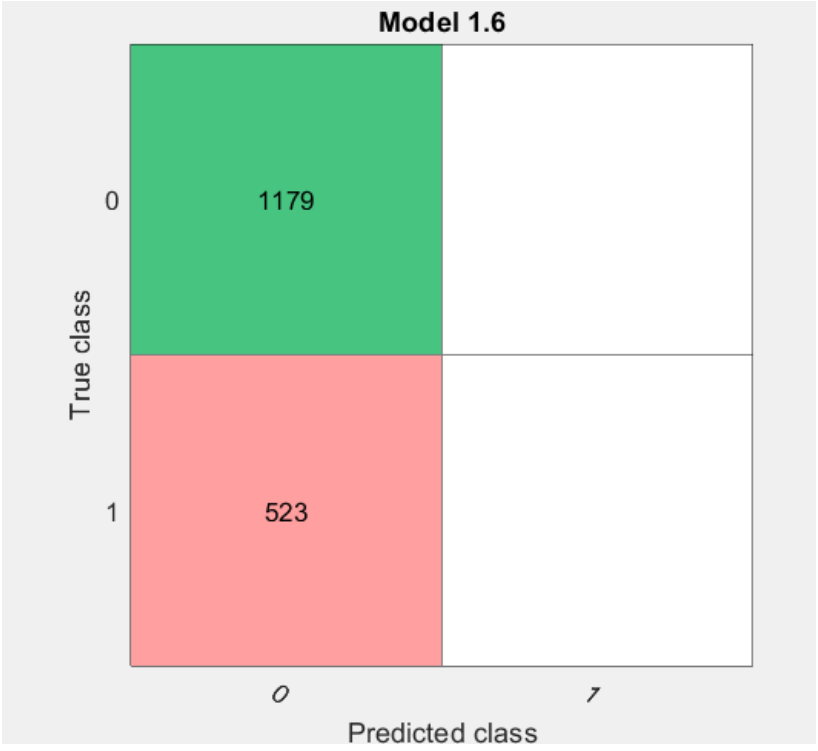


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

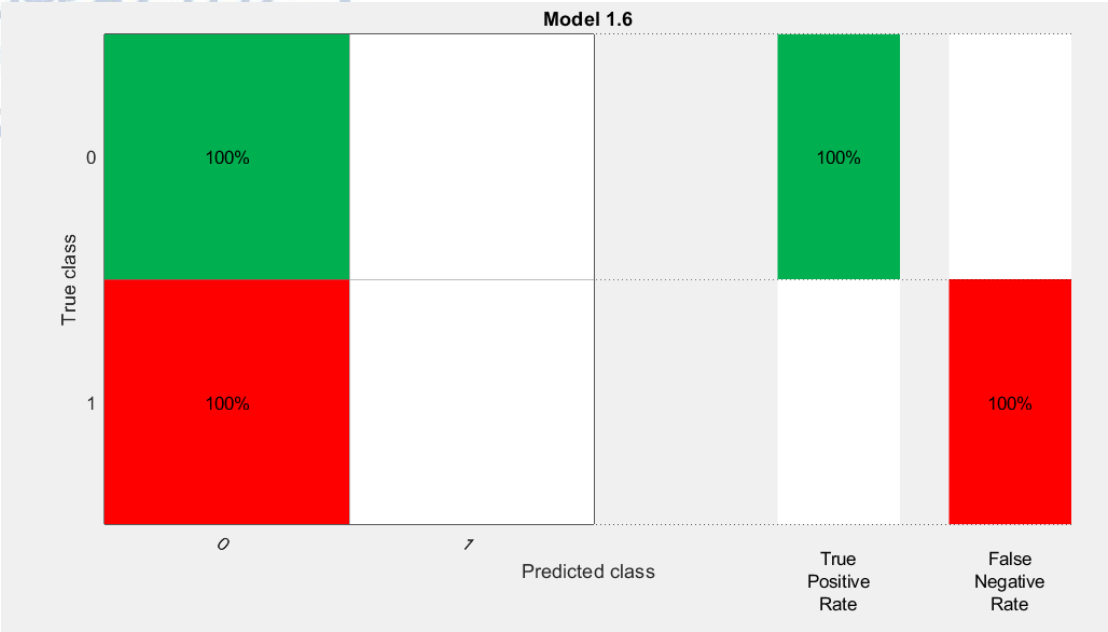


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

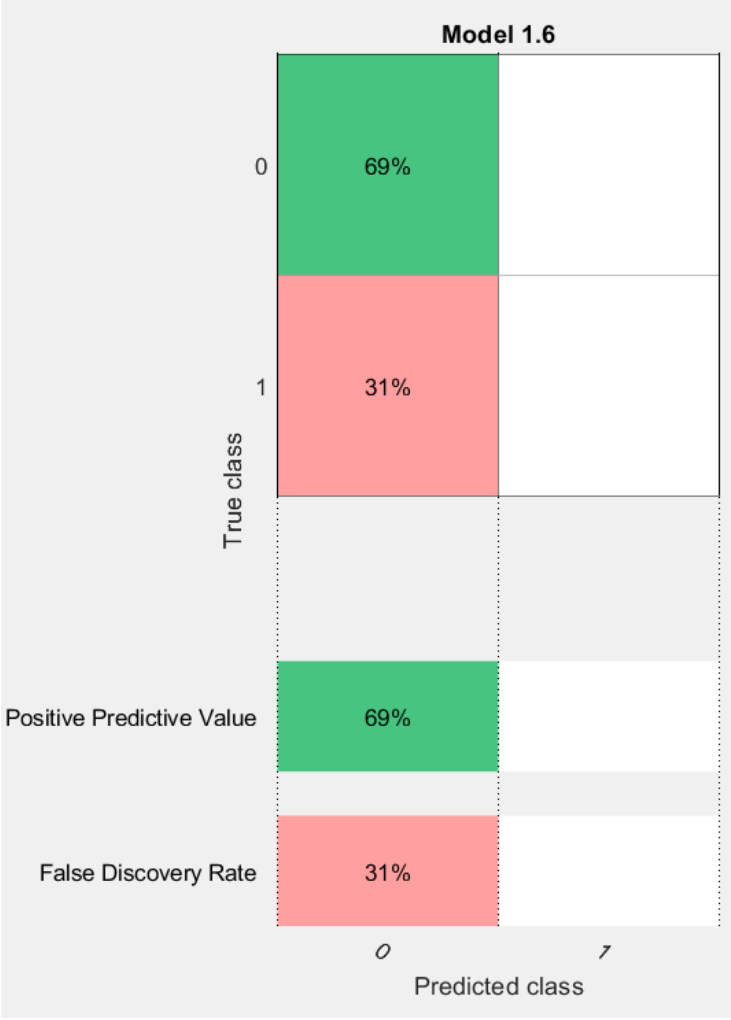
ΙΙ.6 Αποτελέσματα με τη μέθοδο Coarse Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

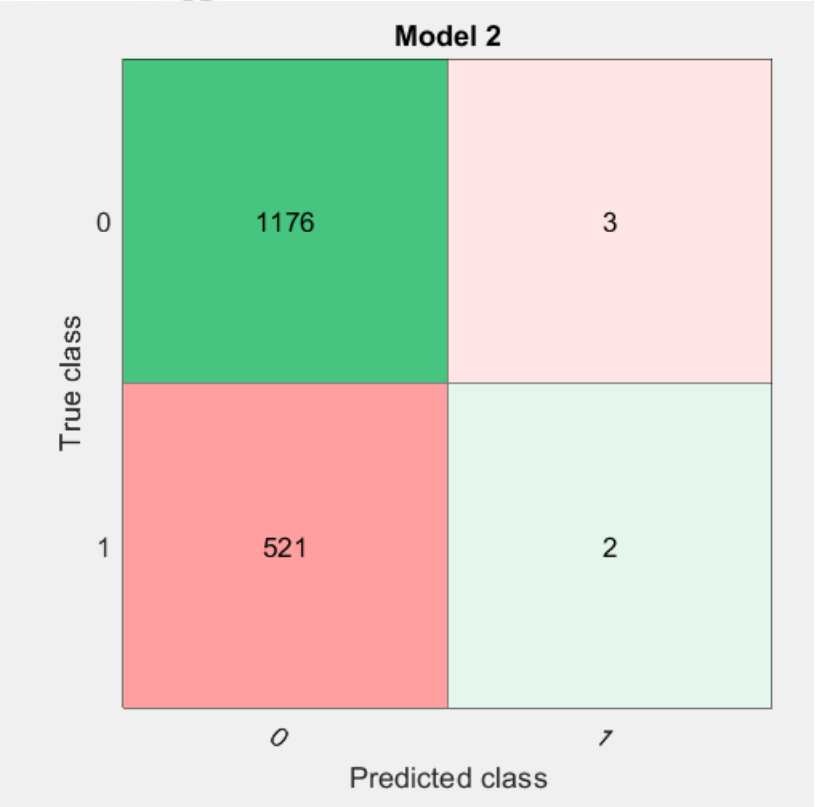


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

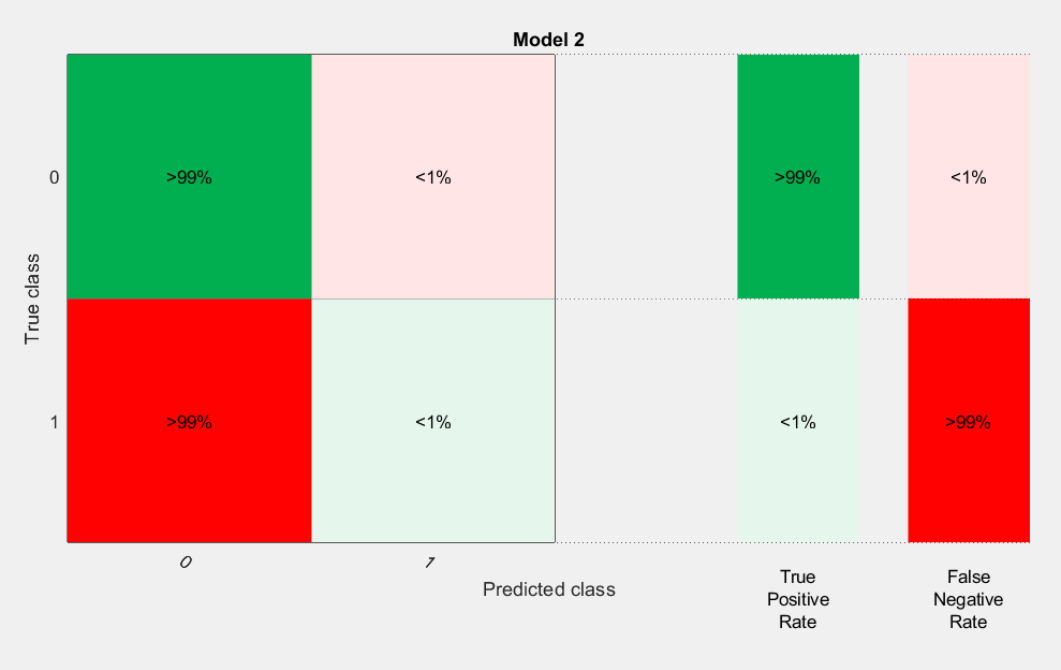


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

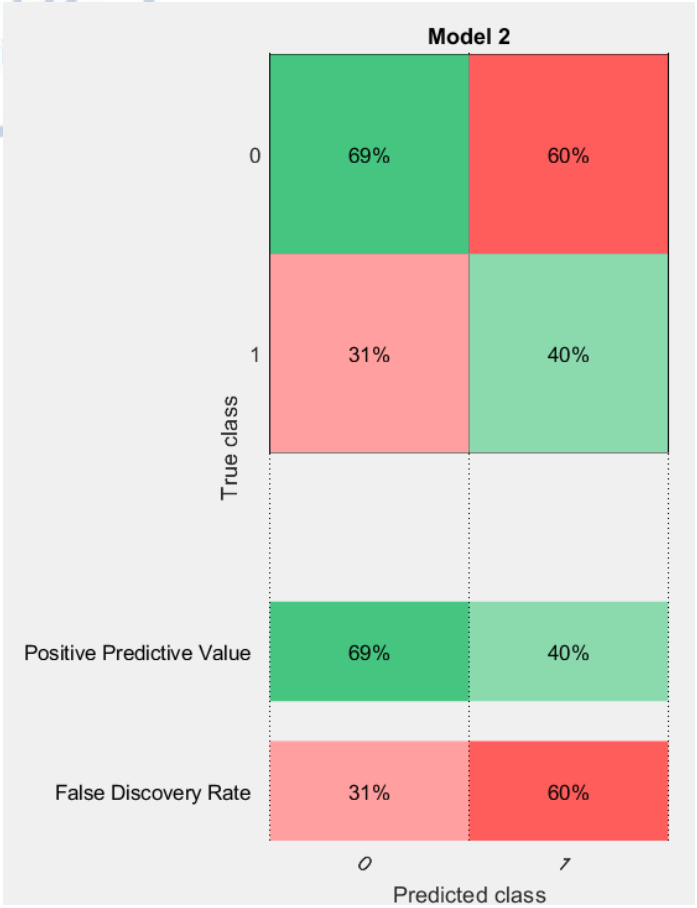
ΙΙ.7 Αποτελέσματα με τη μέθοδο Logistic Regression



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

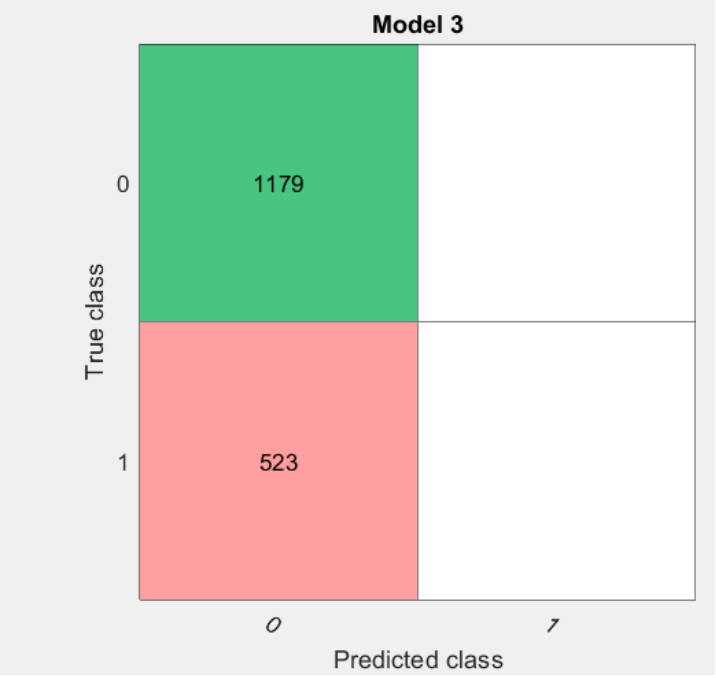


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.



γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

ΙΙ.8 Αποτελέσματα με τη μέθοδο Bagged Trees



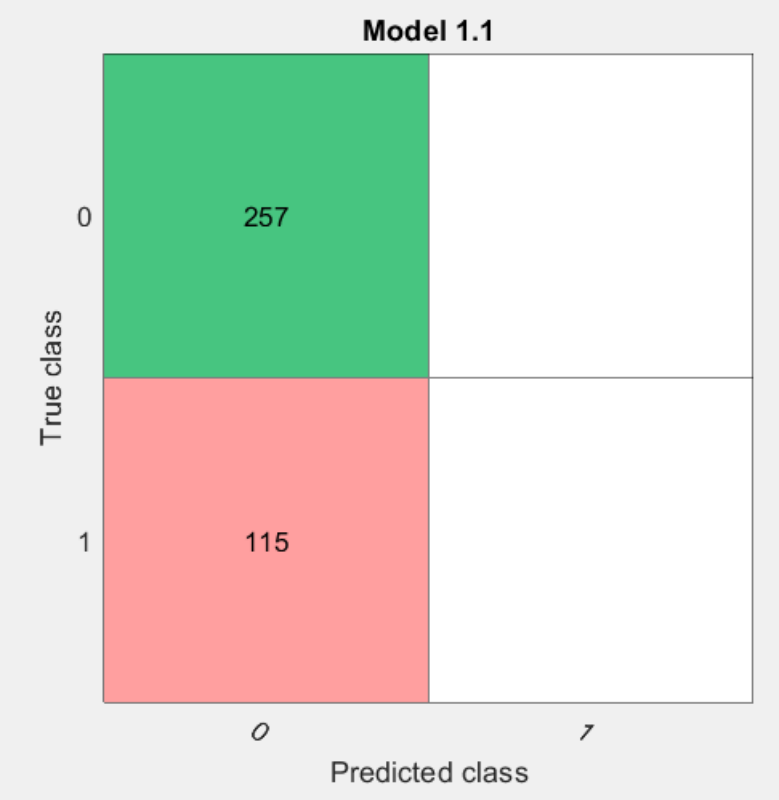
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

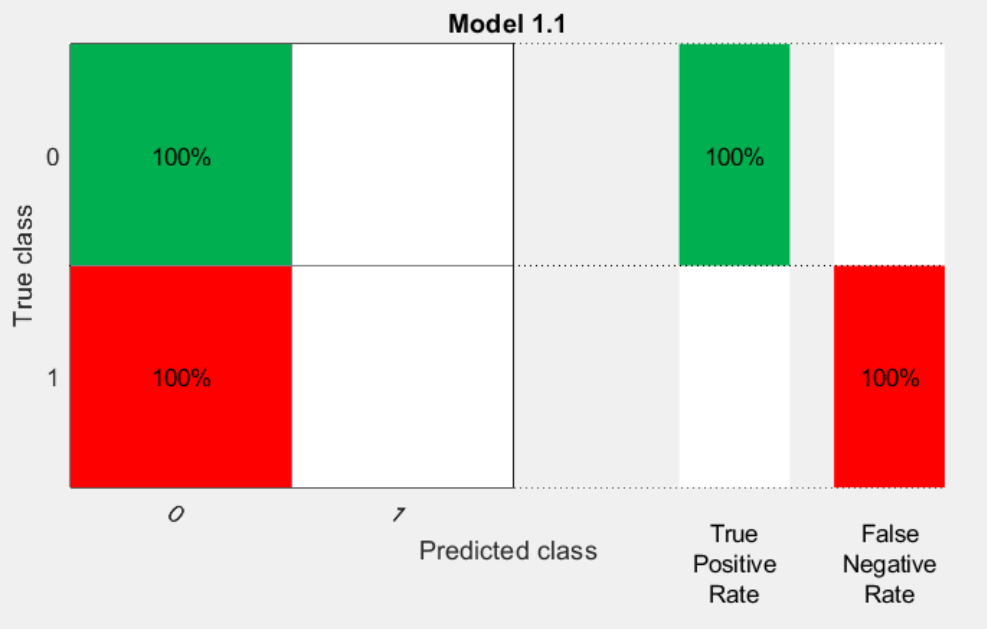
γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

Μηνιαία δεδομένα

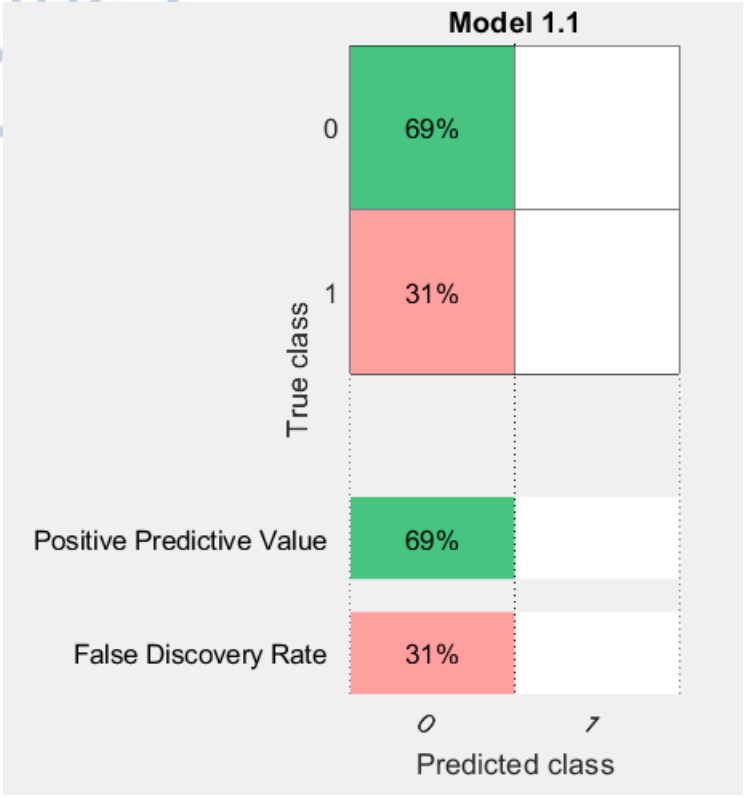
III.1 Αποτελέσματα με τη μέθοδο Linear SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

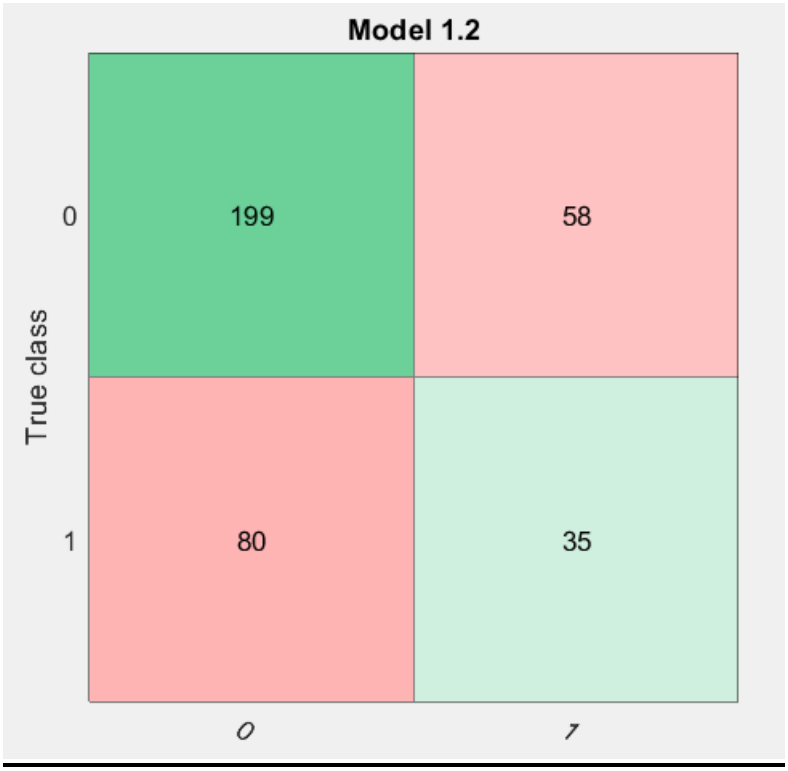


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

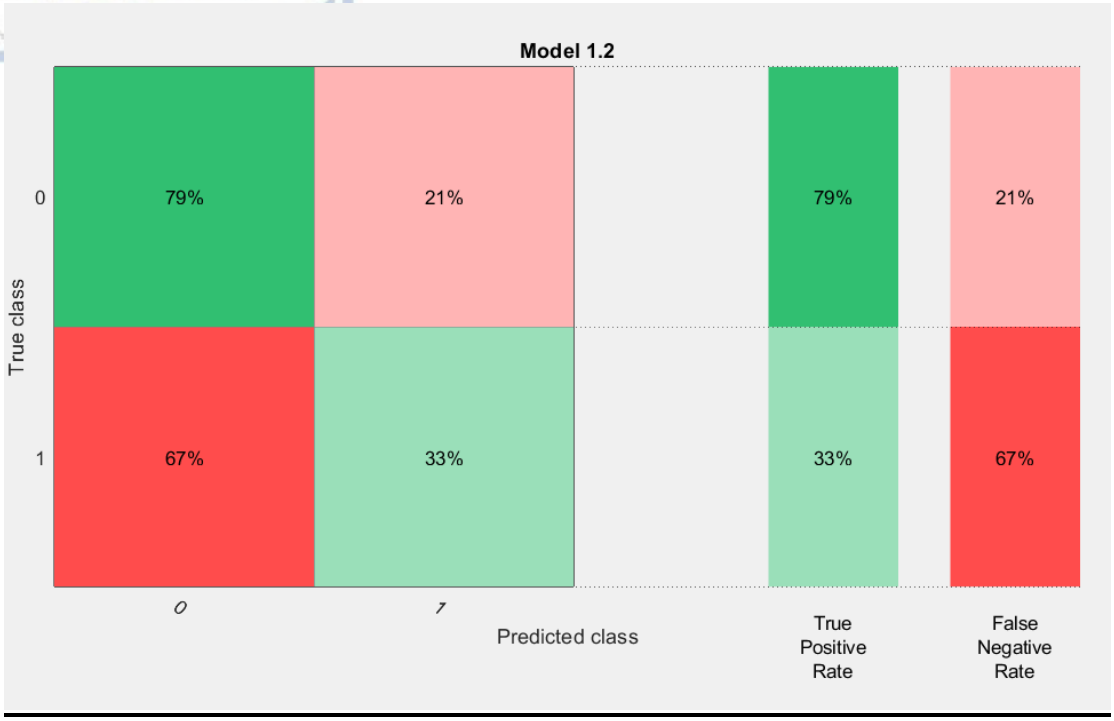


γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

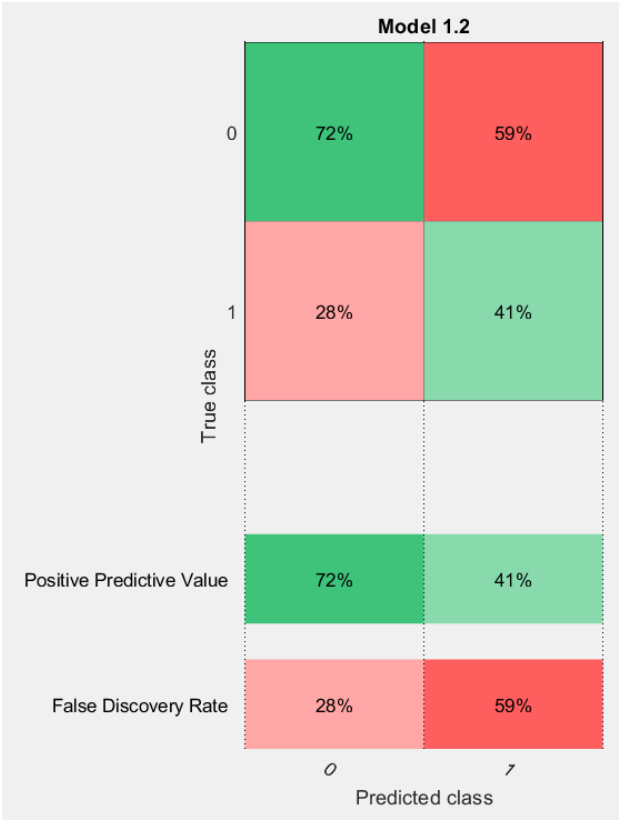
ΙΙΙ.2 Αποτελέσματα με τη μέθοδο Quadratic SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

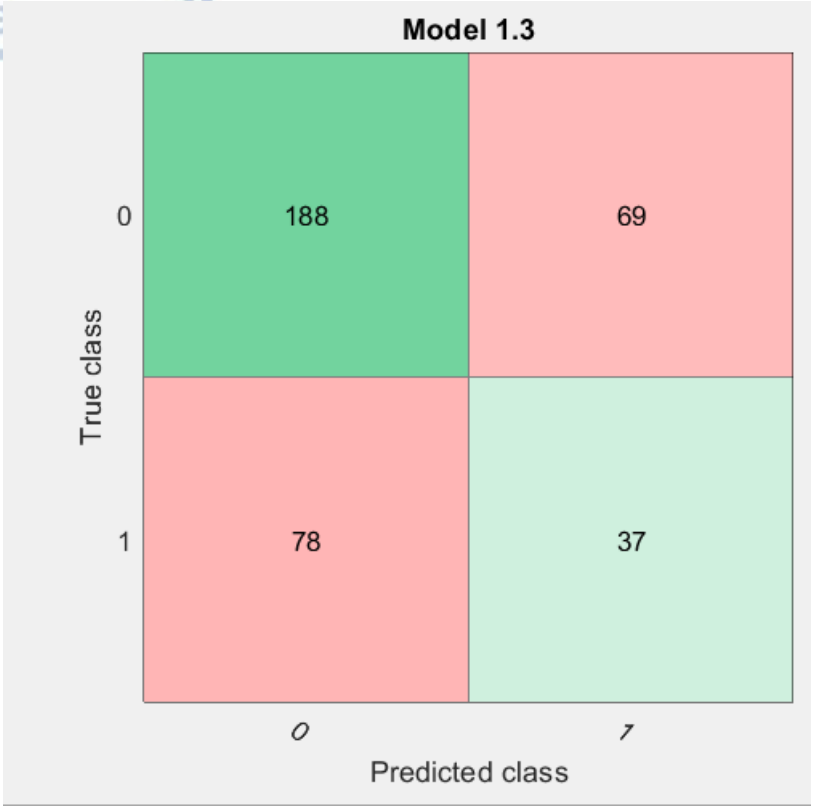


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

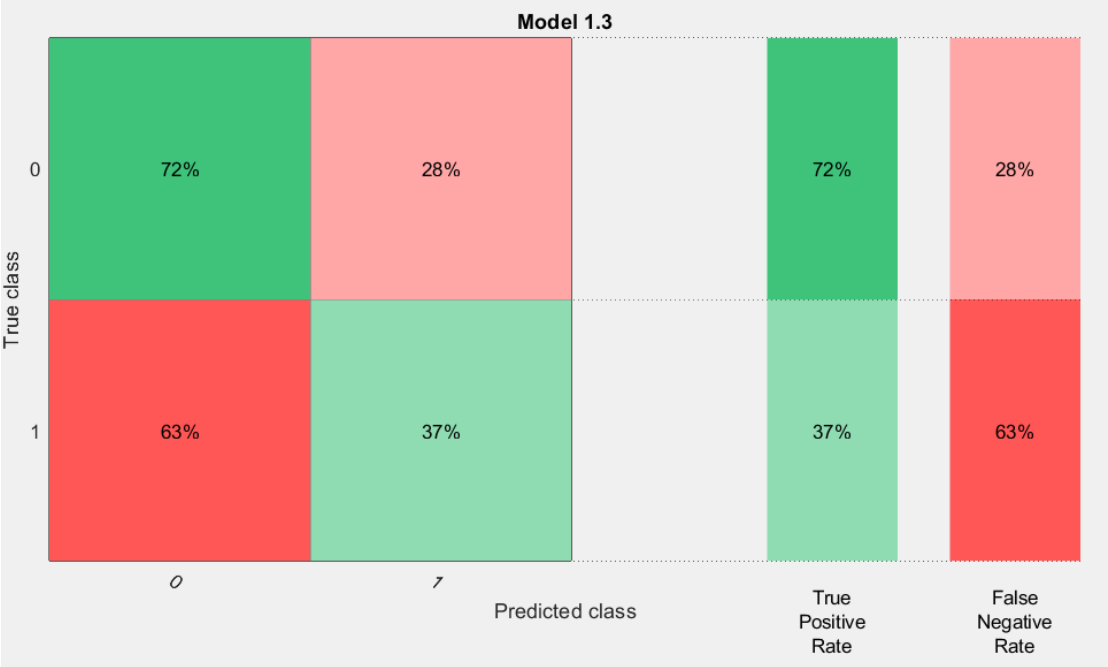


γ) διάγραμμαpositivepredictive value καιfalse discovery rateγιακλάσεις 0 και 1.

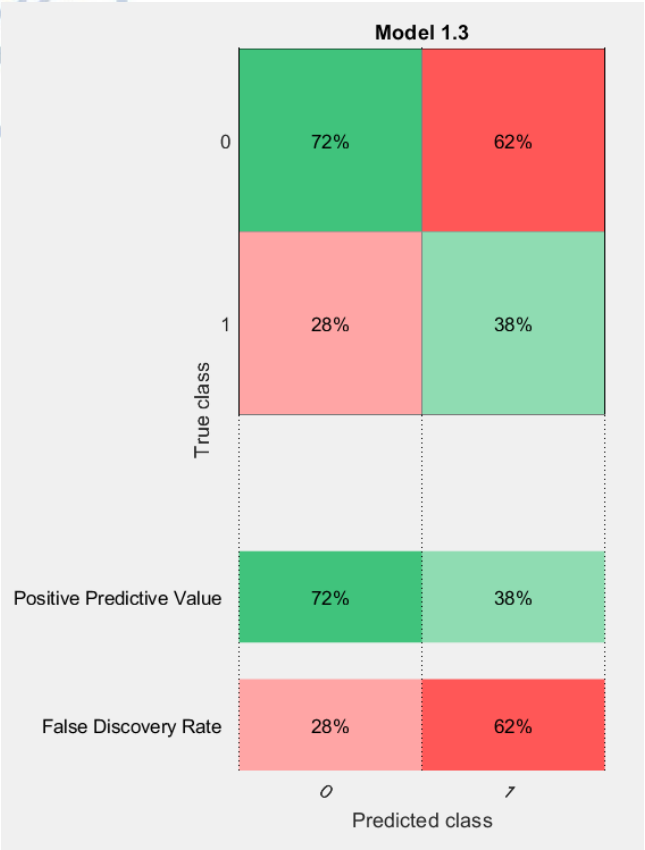
ΙΙΙ.3 Αποτελέσματα με τη μέθοδο Cubic SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

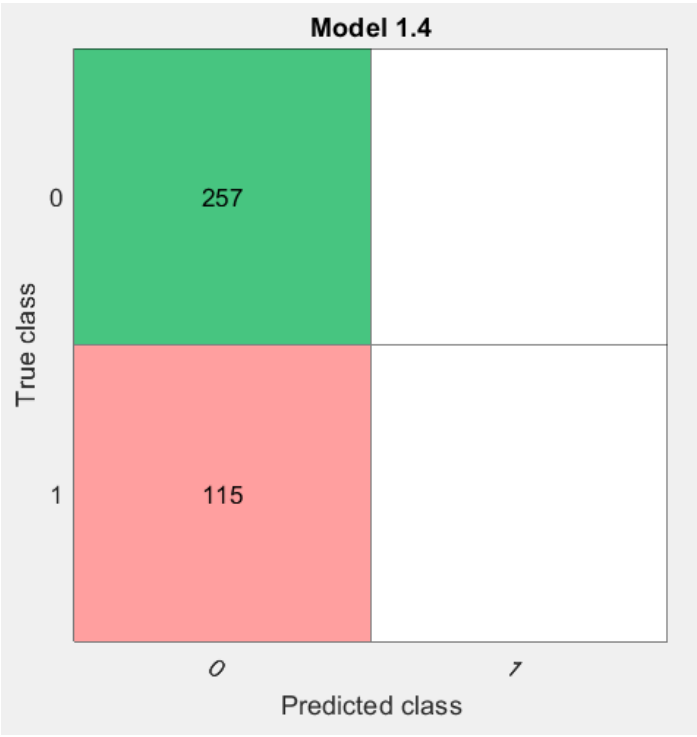


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

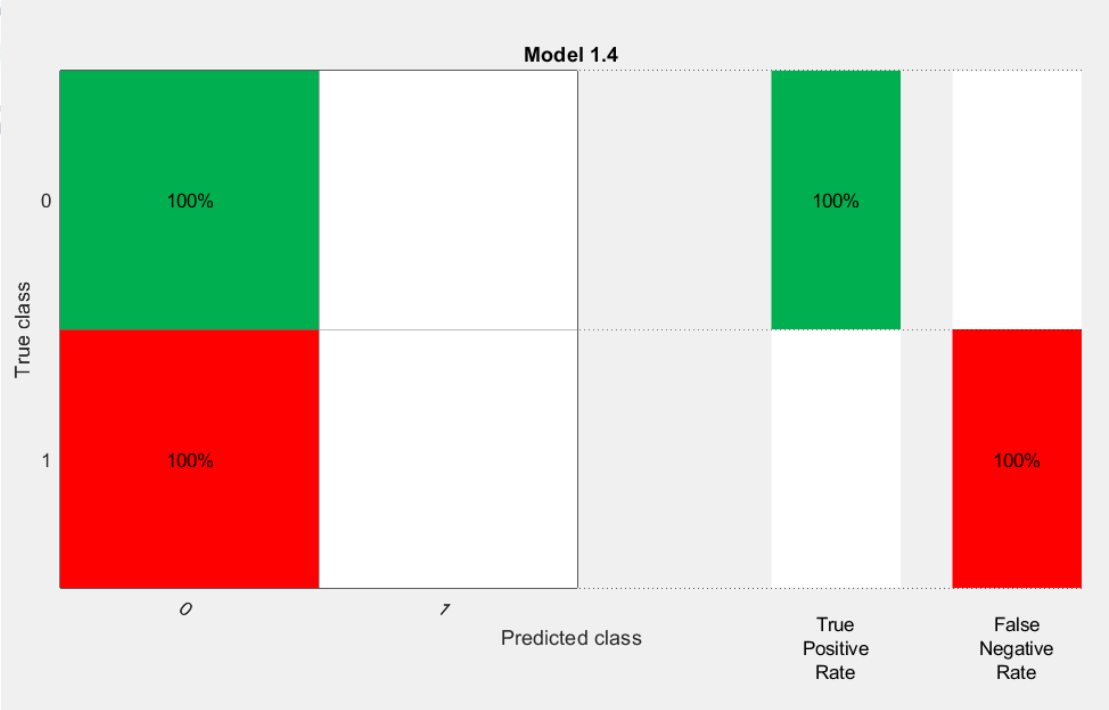


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

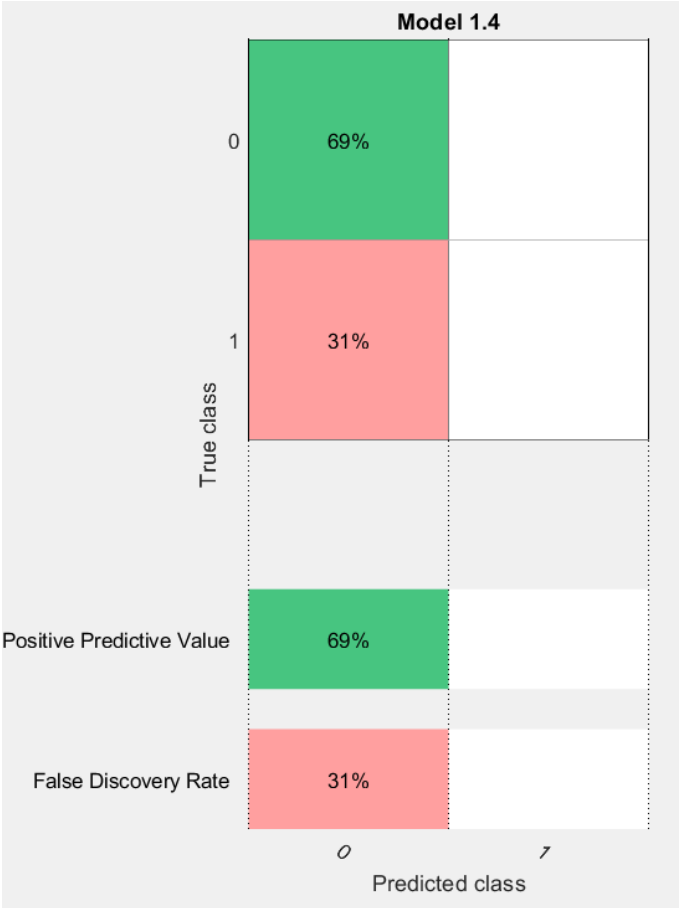
ΙΙΙ.4 Αποτελέσματα με τη μέθοδο Fine Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

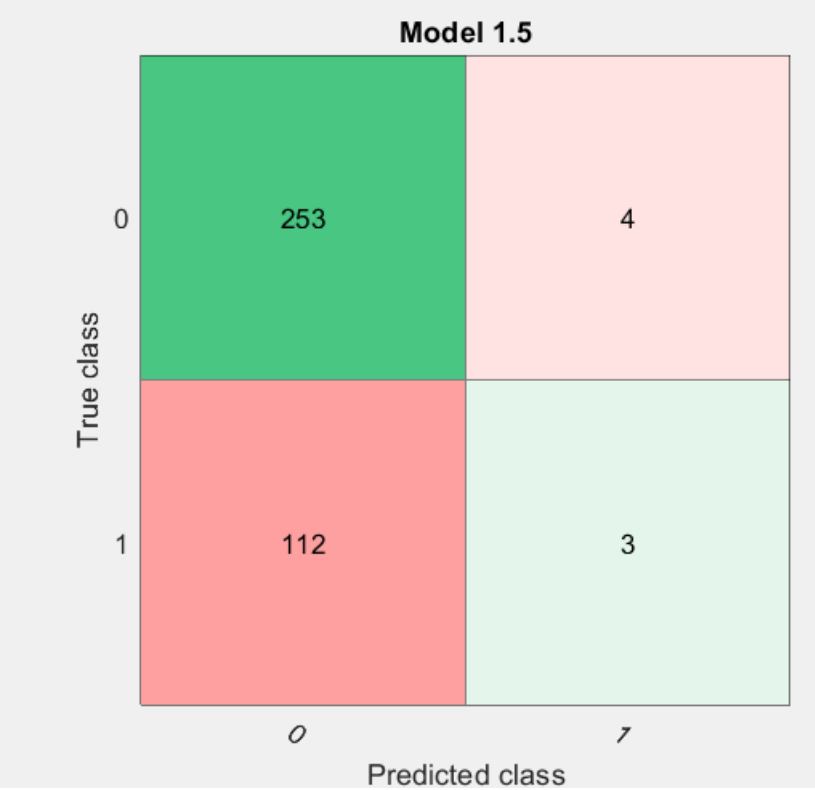


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

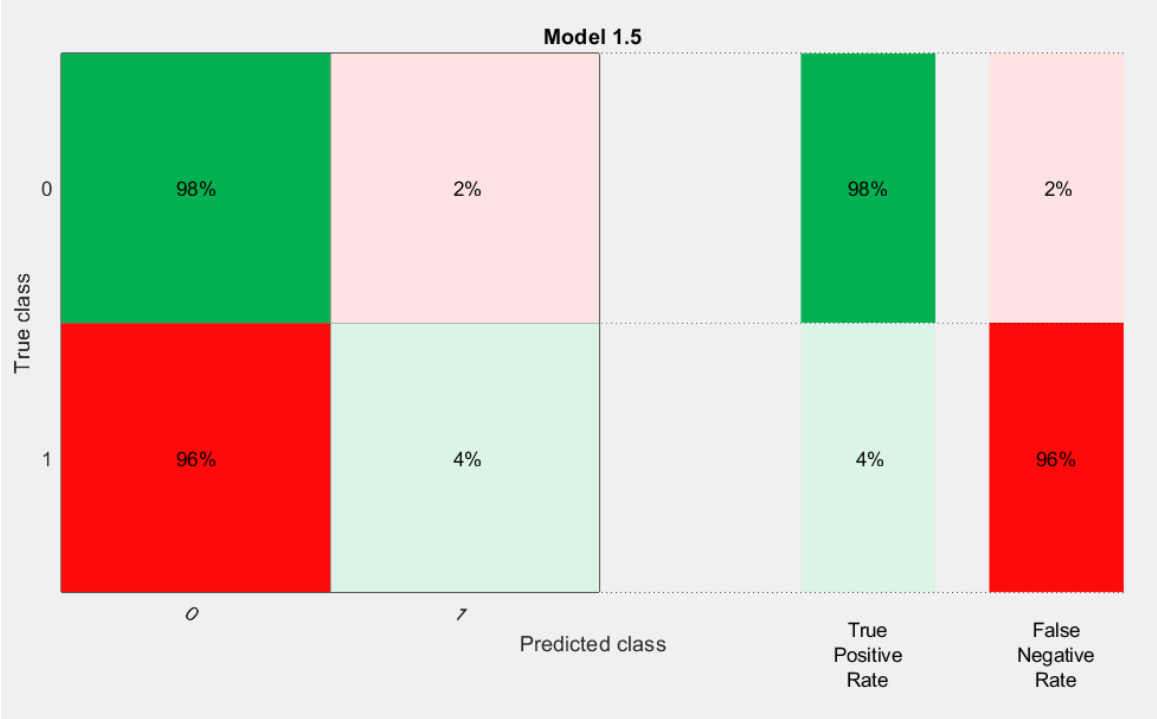


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

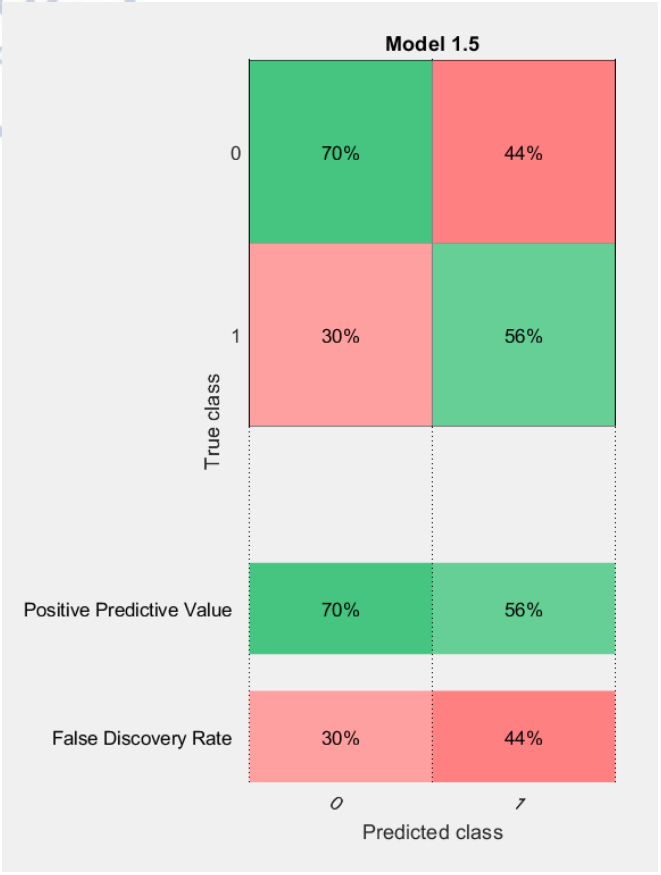
ΙΙΙ.5 Αποτελέσματα με τη μέθοδο Medium Gaussian SVM



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

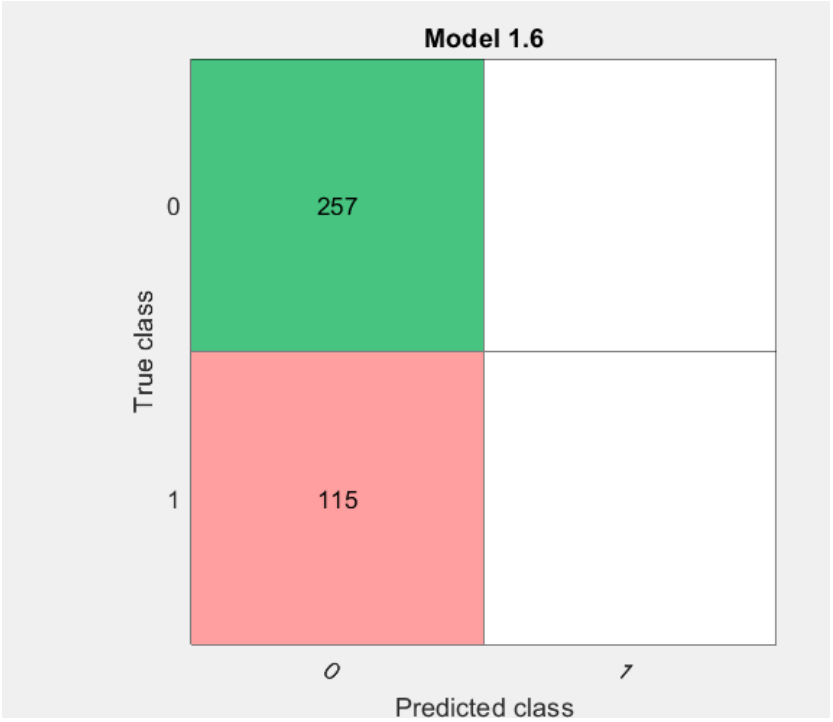


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.



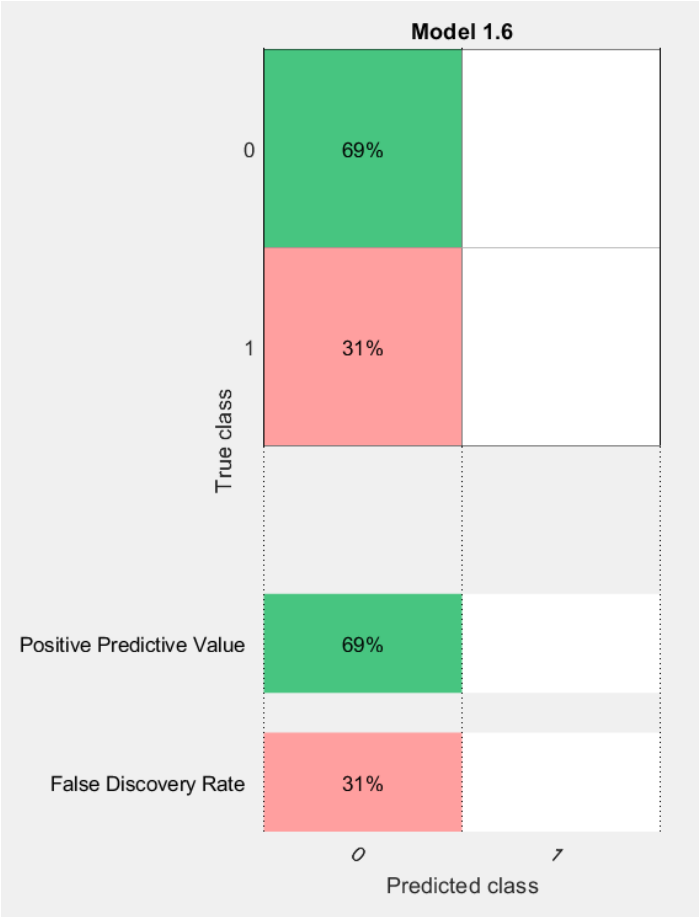
γ) διάγραμμα positive predictive value και false discovery rate για κλάσεις 0 και 1.

III.6 Αποτελέσματα με τη μέθοδο Coarse Gaussian SVM



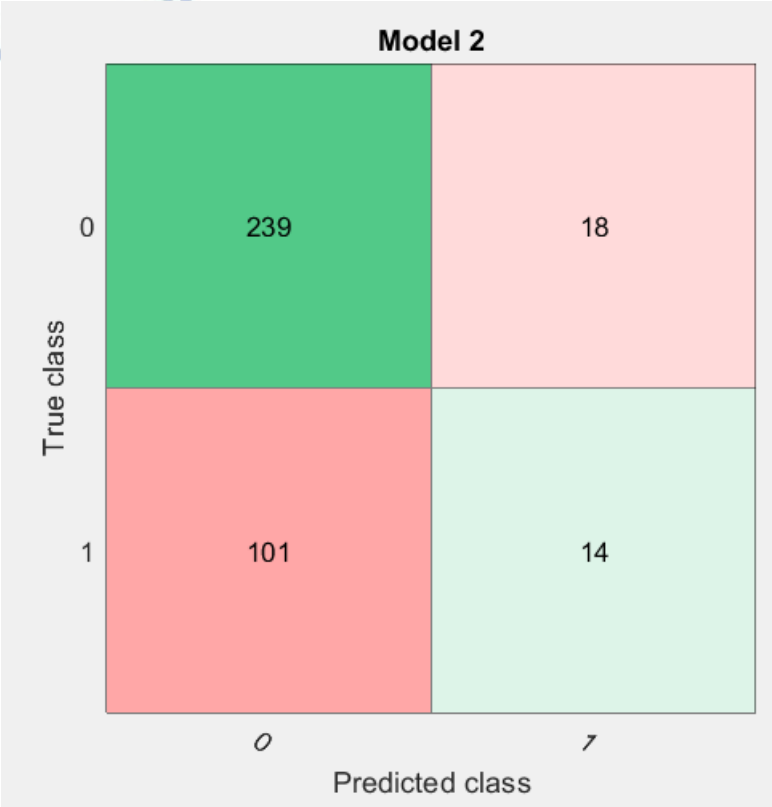
α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

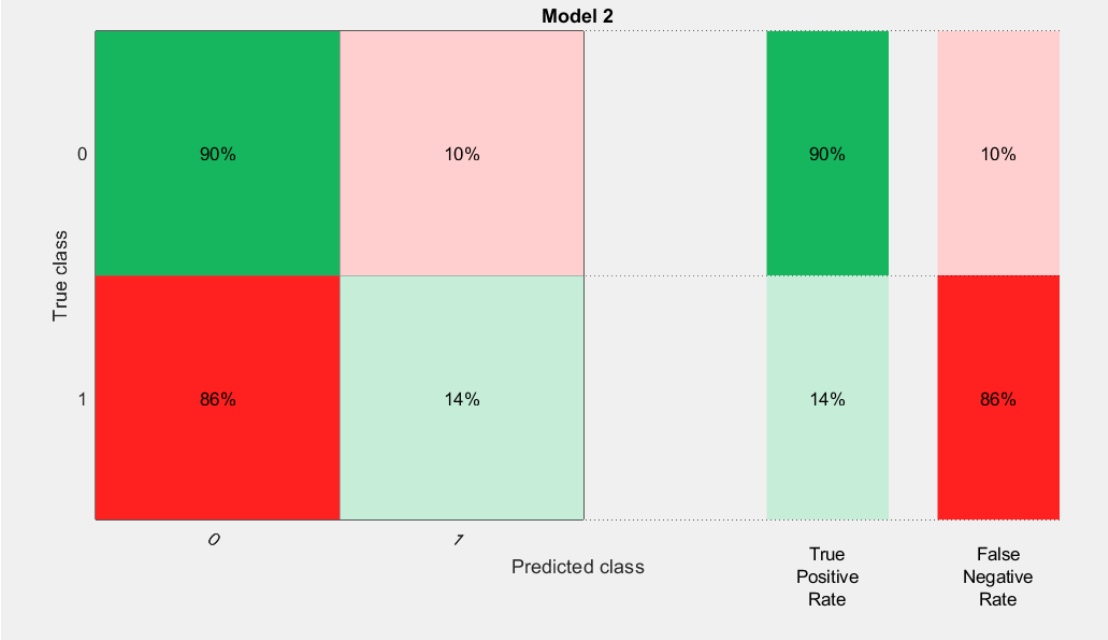


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

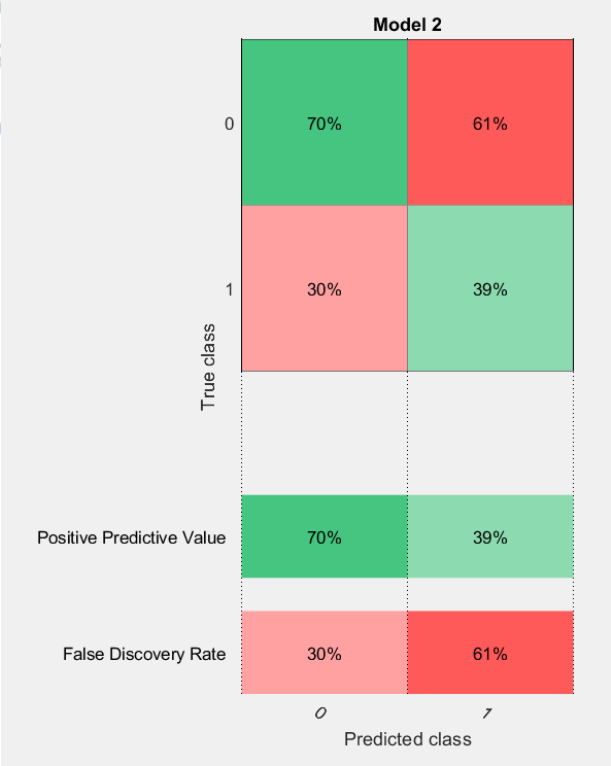
III.7 Αποτελέσματα με τη μέθοδο Logistic Regression



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.

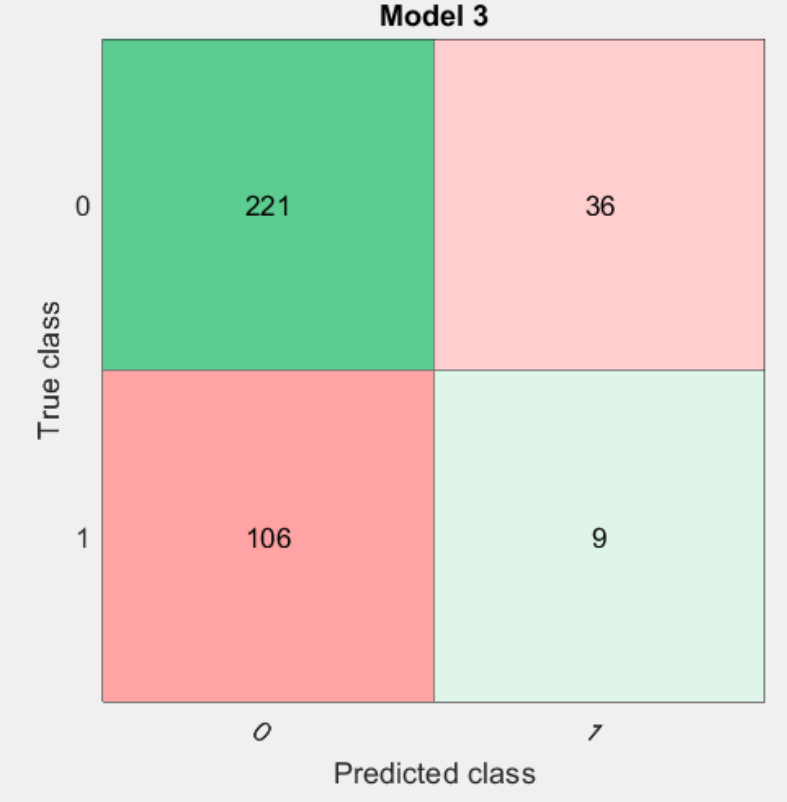


β) διάγραμμα true positive – false negative rate για κλάσεις 0 και 1.

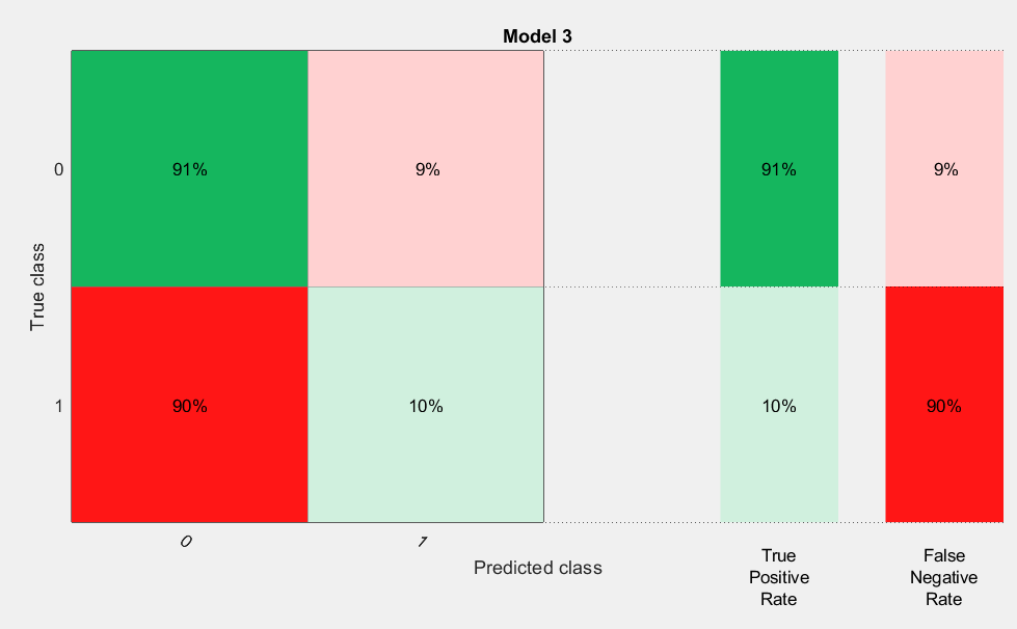


γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.

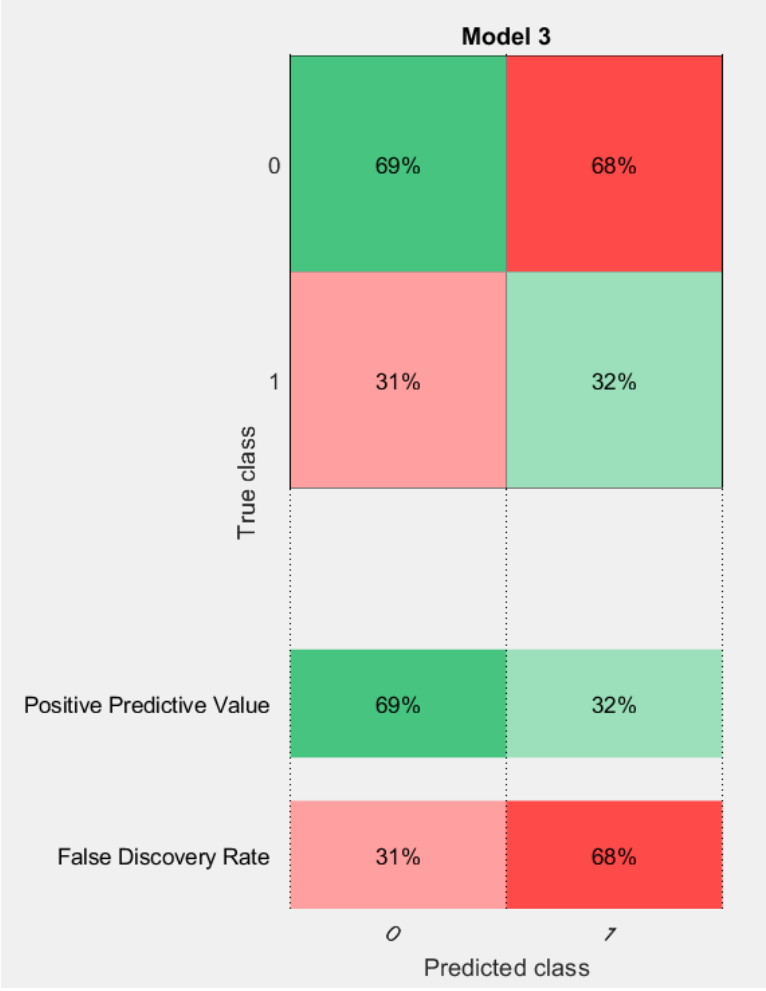
ΙΙΙ.8 Αποτελέσματα με τη μέθοδο Bagged Trees



α) αριθμός παρατηρήσεων για τις κλάσεις 0 και 1.



β) διάγραμματrue positive – false negative rate για κλάσεις 0 και 1.



γ) διάγραμμα positive predictive value καιfalse discovery rate για κλάσεις 0 και 1.