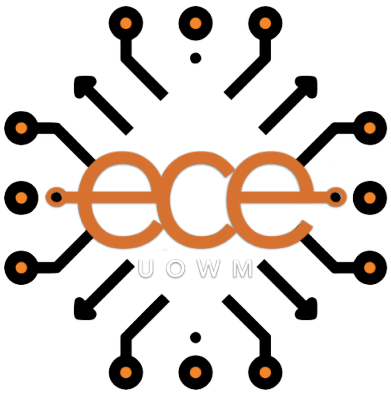


Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Σύγκριση μεθόδων εντοπισμού ακραίων τιμών

Δημήτριος Χατζηαναγνώστου (ΑΜ: 1095)

Επιβλέπων Καθηγητής: Νικόλαος Πλόσκας

Εργαστήριο Ευφυών Συστημάτων & Βελτιστοποίησης

18 Οκτωβρίου 2022

Ευχαριστίες

Με το πέρας της παρούσης διπλωματικής εργασίας θα ήθελα να ευχαριστήσω θερμά τον Επίκουρο Καθηγητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών κ. Νικόλαο Πλόσκα για τη συνεχή καθοδήγηση, τη συνεργασία και τις υποδείξεις του με σκοπό την εκπόνηση της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου που στάθηκε καθόλη τη διάρκεια των σπουδών δίπλα μου, καθώς και τους φίλους μου για τις συμβουλές και τη βοήθεια που μου παρείχαν.

Περίληψη

Τα τελευταία χρόνια λόγω της προόδου της τεχνολογίας, όλο και περισσότερα άτομα ασχολούνται με τη μελέτη δεδομένων και την εξαγωγή αποτελεσμάτων από πειράματα είτε σε επαγγελματικό είτε σε ερασιτεχνικό επίπεδο. Κάποια στιγμή οι ως επί το πλείστον ερευνητές συναντούν το πρόβλημα, αν οι παρατηρήσεις τους είναι σωστές και πρέπει να λάβουν μέρος στην ανάλυση των αποτελεσμάτων. Συνεπώς, θα πρέπει να αποφασίσουν αν οι παρατηρήσεις αυτές θα πρέπει να απαλειφθούν ή αν με κάποιο τρόπο να ομαλοποιηθούν ώστε να μειωθεί η επίδρασή τους στην απόρροια των μετρήσεων. Στόχος της διπλωματικής εργασίας είναι η σύγκριση των μεθόδων εντοπισμού αυτών των ακραίων παρατηρήσεων. Για την επίτευξη αυτού του στόχου στο πειραματικό κομμάτι της εργασίας, διάφορα σύνολα δεδομένων επεξεργάζονται με τη χρήση εργαλείων και βιβλιοθηκών της γλώσσας προγραμματισμού Python, καθώς επίσης και του πακέτου αλγορίθμων μηχανικής μάθησης scikit-learn. Οι αλγόριθμοι εντοπισμού ακραίων τιμών που χρησιμοποιήθηκαν είναι οι DBscan, Elliptical Envelope, Gaussian Mixture Model, Local Outlier Factor, Isolation Forest Mahalanobis distance και OneClassSVM. Με αυτόν τον τρόπο πραγματοποιούνται οι προβλέψεις ως προς την ακρίβεια που παρουσιάζει η κάθε μέθοδος. Με το πέρας των πειραμάτων της διπλωματικής εργασίας, καταλήγουμε στο συμπέρασμα ότι στις περισσότερες περιπτώσεις, η απαλοιφή των ακραίων τιμών βοήθησε στη βελτίωση των ποσοστών ακριβείας.

Λέξεις κλειδιά: Ακραίες τιμές, μηχανική μάθηση, ανάλυση δεδομένων, κατηγοριοποίηση, παλινδρόμηση

Abstract

In recent years, due to the advancement of technology, more and more people are engaged in studying data and extracting results from experiments either at a professional or amateur level. At some point most researchers encounter the problem of whether their observations are correct and must take part in analyzing the results. Therefore, they will have to decide whether these observations should be deleted or if they should be normalized in some way to reduce their influence on the outcome of the measurements. The aim of the thesis is to compare the methods of identifying these extreme observations. To achieve this goal in the experimental part of the work, various data sets are processed using Python programming language tools and libraries, as well as the scikit-learn machine learning algorithmic package. The algorithms used to detect extreme values are DBscan, Elliptical Envelope, Gaussian Mixture Model, Local Outlier Factor, Isolation Forest, Mahalanobis distance and OneClassSVM. In this way, the predictions are made in terms of the accuracy presented by each method. By the end of the thesis experiments, we conclude that in most cases, the removal of outliers helped to improve the accuracy rates.

Keywords: Outliers, machine learning, data analysis, classification, regression

Δήλωση Πνευματικών Δικαιωμάτων

Δήλωση Πνευματικών Δικαιωμάτων Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο ''Σύγκριση μεθόδων εντοπισμού ακραίων τιμών'' καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Πλόσκα Νικόλαου αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Υπογραφή Φοιτητή

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κατάλογος απεικονίσεων

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός του προβλήματος

Ως ακραία τιμή ορίζεται μια παρατήρηση, η οποία απέχει κατά ένα μεγάλο βαθμό από την πλειοψηφία των τιμών ενός σετ δεδομένων. Ένα παράδειγμα που βοηθάει στην κατανόηση του ορισμού των ακραίων τιμών, είναι αυτό των βαθμών που παίρνουν οι μαθητές στο σχολείο. Ας υποθέσουμε ότι έχουμε δέκα μαθητές, εκ των οποίων οι εννέα έχουν βαθμό που κυμαίνεται από το 85 μέχρι το 90, με άριστα το 100 και ο δέκατος έχει βαθμό 20. Όπως είναι φανερό ο μέσος όρος είναι χαμηλότερος εφόσον συμπεριληφθεί και ο δέκατος μαθητής στη μέτρηση. Σε μια άλλη περίπτωση, υποθέτουμε ότι σχεδιάζουμε τα καθίσματα των αεροπλάνων μίας εταιρείας. Το σετ δεδομένων, σχετικά με το ύψος των επιβατών, που λαμβάνουμε υπόψη προέρχεται από μία χώρα με σχετικά μεγάλο μέσο όρο ύψους των ανθρώπων, όπως αυτής της Ολλανδίας. Επομένως οι θέσεις του αεροπλάνου θα είναι λιγότερες και αυτό θα έχει ως αποτέλεσμα τις μικρότερες οικονομικές απολαβές της εταιρείας. Συνεπώς είναι πολύ σημαντικό να ανιχνεύονται οι ακραίες παρατηρήσεις και να λαμβάνονται οι κατάλληλες ενέργειες με σκοπό τη βελτίωση της απόδοσης του εκάστοτε αλγορίθμου ή πειράματος.

1.2 Κίνητρα και Στόχοι Υλοποίησης

Η εποχή που ζούμε έχει χαρακτηριστεί από την καθημερινή εξέλιξη της τεχνολογίας που συνεχώς βελτιώνει και διευκολύνει τον τρόπο ζωής μας. Προκειμένου να λάβουν χώρα αυτές οι τεχνολογίες είναι απαραίτητη η ύπαρξη αρκετών δεδομένων και μετρήσεων με σκοπό την τελειοποίησή τους. Πολλές φορές όμως οι μετρήσεις

αυτές αποτελούνται από σφάλματα που προκύπτουν είτε από ανθρώπινο λάθος είτε από αποτυχία του συστήματος. Βασικό κίνητρο της εργασίας είναι η εκμηδένιση του σφάλματος όσο είναι αυτό εφικτό, ώστε να εφαρμοστούν τα νέα τεχνολογικά συστήματα χωρίς την ύπαρξη ρίσκου για λάθη, ξεκινώντας από την καθημερινή ζωή του ανθρώπου έως και τους τομείς επιστημονικού ενδιαφέροντος. Στόχος της διπλωματικής εργασίας είναι η μελέτη των μεθόδων ανίχνευσης ακραίων τιμών με σκοπό τη σύγκρισή τους ως προς την ακριβή πρόβλεψη.

1.3 Διάρθρωση κειμένου

Η διπλωματική εργασία απαρτίζεται από πέντε κεφάλαια. Στο δεύτερο κεφάλαιο εισάγονται οι έννοιες της μηχανικής μάθησης και των ακραίων τιμών. Επίσης, αναφέρονται οι κατηγορίες στις οποίες χωρίζονται οι έννοιες αντίστοιχα, καθώς και αναλύονται οι γενικές κατηγορίες των μεθόδων ανίχνευσης ακραίων τιμών. Στο τρίτο κεφάλαιο γίνεται μια αναλυτική ανοφορά σε ορισμένες μεθόδους εύρεσης ακραίων τιμών, οι οποίες απαρτίζονται από παραδείγματα και κομμάτια ψευδοκώδικα για κάθε περίπτωση. Έπειτα στο τέταρτο κεφάλαιο παρουσιάζονται τα αποτελέσματα από τις δοκιμές των αλγορίθμων δείχνοντας τις ακραίες τιμές που ανιχνεύονται σε κάθε σετ δεδομένων που χρησιμοποιήθηκε και επιπλέον τα ποσοστά επιτυχίας πρόβλεψης των αλγορίθμων. Τέλος, το πέμπτο κεφάλαιο έχει ως στόχο την παρουσίαση των αποτελεσμάτων της εργασίας και την αναφορά επιπλέον ενεργειών ως προς τη βελτίωσή της.

Κεφάλαιο 2

Θεωρητικό Μέρος

2.1 Μηχανική μάθηση

Ανάλογα με τη φύση του προβλήματος υπάρχουν πολλές τεχνικές μηχανικής μάθησης, οι οποίες χωρίζονται στις εξής κατηγορίες:

1. Μάθηση με επίβλεψη (supervised learning) ή αλλιώς μάθηση με παραδείγματα
2. Μάθηση χωρίς επίβλεψη (unsupervised learning) ή αλλιώς μάθηση από παρατήρηση
3. Ενισχυτική μάθηση (reinforcement learning)

2.1.1 Μάθηση με επίβλεψη

Ο αλγόριθμος λαμβάνει κάποια δεδομένα (input) καθώς και τα προσδοκώμενα αποτελέσματα (output) [11]. Με αυτό τον τρόπο τα δεδομένα που δέχεται ο αλγόριθμος λειτουργούν ως «δάσκαλος». Σκοπός είναι να του μάθουν ένα γενικευμένο κανόνα ώστε να κάνει τις αντιστοιχίες μεταξύ δεδομένων και αποτελεσμάτων. Κάποιοι από τους πιο γνωστούς και δημοφιλείς ως προς τη χρήση τους αλγορίθμους είναι οι εξής: nearest neighbor, naive Bayes, decision trees, support vector machines (SVM) και logistic regression. Επιπροσθέτως, για την καλύτερη κατανόηση της μάθησης με επίβλεψη, παρατίθεται το παρακάτω παράδειγμα που αφορά την πρόβλεψη των τιμών της αξίας σπιτιών. Για αρχή πρέπει να ληφθούν κάποια δεδομένα για τα σπίτια, όπως είναι τα τετραγωνικά μέτρα, ο αριθμός δωματίων, η ύπαρξη ή όχι κήπου, η ύπαρξη ή όχι ιδιωτικού χώρου στάθμευσης και άλλων παρόμοιων δεδομένων. Έπειτα γνωρίζοντας τις τιμές αυτών των σπιτιών γίνονται οι κατάλληλες

αντιστοιχίες των ετικετών. Στη συνέχεια, αξιοποιώντας τα δεδομένα χιλιάδων σπιτιών καθώς και των τιμών τους, φτάνουμε στο σημείο εκπαίδευσης του μοντέλου μηχανικής μάθησης. Με αυτό το τρόπο μπορεί να γίνει η πρόβλεψη της τιμής ενός σπιτιού βασιζόμενη στα δεδομένα που δόθηκαν στον αλγόριθμο.

2.1.2 Μάθηση χωρίς επίβλεψη

Αντίθετα με την προηγούμενη περίπτωση της επιβλεπόμενης μάθησης, ο αλγόριθμος εδώ πρέπει να βρει συσχετίσεις σχετικά με τα εισερχόμενα δεδομένα [11]. Η πιο συνηθισμένη μέθοδος είναι η ομαδοποίηση (clustering), της οποίας στόχος είναι η δημιουργία ομάδων. Ο διαχωρισμός αυτός πραγματοποιείται με τέτοιο τρόπο ώστε κάθε δεδομένο να εμπίμπτει σε μία ομάδα αποτελούμενη και με άλλα σημεία δεδομένων με τα οποία έχει κοινά στοιχεία. Ένα παράδειγμα χρήσης της ομαδοποίησης είναι η κατηγοριοποίηση των τμημάτων πελατών στα δεδομένα του μάρκετινγκ. Ανάλογα δηλαδή με τα χαρακτηριστικά του πελάτη (φύλο, ηλικία, τοποθεσία, εκπαίδευση, εισόδημα κ.λπ.) ο αλγόριθμος δημιουργεί κατηγορίες και εντάσσει τον κάθε πελάτη σε αυτές. Με αυτό τον τρόπο οι ομάδες μάρκετινγκ μπορούν να προσεγγίσουν περισσότερους πελάτες με ξεχωριστούς τρόπους.

2.1.3 Ενισχυτική μάθηση

Σε αυτή την κατηγορία ο αλγόριθμος αλληλεπιδρά σε ένα εικονικό περιβάλλον που έχουμε δημιουργήσει [12]. Η αλληλεπίδραση γίνεται ανάλογα με τους κανόνες που έχουμε ορίσει σε αυτό το περιβάλλον και σκοπός είναι ο αλγόριθμος να πετύχει τον στόχο που του τίθεται χωρίς να υπάρχει άμεση επίβλεψη από ένα «δάσκαλο». Κατά τη διαδικασία της συνεχόμενης εκμάθησης ο αλγόριθμος μέσα από τις εμπειρίες του εξερευνεί όλο το φάσμα των πιθανών καταστάσεων. Με αυτόν τον τρόπο ο πράκτορας λαμβάνει τις καλύτερες αποφάσεις για τη μέγιστη απόδοσή του, μέσα από ένα σύστημα ανταμοιβής – τιμωρίας. Όταν το βήμα της απόφασης της καλύτερης ενέργειας ανάλογα της κατάστασης του περιβάλλοντος επαναλαμβάνεται, το πρόβλημα ονομάζεται Διαδικασία Απόφασης Markov (Markov Decision Process). Καταλυτικό παράδειγμα αυτό του οποίου ο υπολογιστής κέρδισε το 2017 τον παγκόσμιο πρωταθλητή Ke Jie στο επιτραπέζιο Go. Ένα άλλο παράδειγμα βιντεοπαιχνιδιού όπως αυτό στο Dota 2 ο πράκτορας είχε ποσοστό επιτυχίας 99%,

ενώ επιπρόσθετα κέρδισε τους παγκόσμιους πρωταθλητές.

2.2 Κατηγορίες προβλημάτων μηχανικής μάθησης

Ένα είδος προβλημάτων μηχανικής μάθησης, προκύπτει από το επιθυμητό αποτέλεσμα του αλγορίθμου και οι κατηγορίες είναι οι εξής:

2.2.1 Κατηγοριοποίηση

Κατά την κατηγοριοποίηση (classification) οι εισαγόμενες τιμές χωρίζονται σε δύο ή περισσότερες κλάσεις. Ο αλγόριθμος πρέπει να κατασκευάσει ένα μοντέλο σύμφωνα με το οποίο θα αντιστοιχίζονται αυτές οι τιμές σε μία ή περισσότερες κλάσεις. Γνωστό παράδειγμα αυτό του διαχωρισμού των spam email από την ηλεκτρονική αλληλογραφία.

2.2.2 Παλινδρόμηση

Το μοντέλο της παλινδρόμησης (regression) ανήκει στην κατηγορία της μάθησης με επίβλεψη. Στα μοντέλα αυτά η μεταβλητή εξόδου είναι μία πραγματική τιμή. Στόχος του μοντέλου είναι η εύρεση μίας συνάρτησης κατά την οποία γίνεται η αντιστοίχιση μεταξύ των δεδομένων εισόδου (ανεξάρτητη μεταβλητή) και μιας συνεχόμενης μεταβλητής εξόδου (εξαρτημένη μεταβλητή).

2.2.3 Συσταδοποίηση

Η συσταδοποίηση (clustering) είναι μια διαδικασία παρόμοια με αυτή της κατηγοριοποίησης, όπου πρέπει να χωριστούν οι εισακτέες τιμές σε ομάδες. Η σημαντική διαφορά όμως είναι ότι οι ομάδες δεν είναι γνωστές εξαρχής. Αυτό καθιστά τη συσταδοποίηση μία εργασία της μάθησης χωρίς επίβλεψη.

Εν κατακλείδι, η μηχανική μάθηση χρησιμοποιείται σε πολλούς τομείς, όπως αυτοί της βιομηχανίας, των χρηματοπιστωτικών υπηρεσιών, της υγειονομικής περίθαλψης, των κυβερνητικών υπηρεσιών και πολλών άλλων, με σκοπό τη βελτίωση και διευκόλυνση του τρόπου ζωής μας. Ο συνδυασμός του αυξημένου όγκου και των ποικίλων κατηγοριών δεδομένων, σε συνάρτηση με τη φθηνότερη και ισχυρότερη υπολογιστική επεξεργασία, καθώς και την οικονομικά προσιτή

αποθήκευση των δεδομένων, καθιστά τη μηχανική μάθηση πιο αναγνωρίσιμη από ποτέ. Χάρη σε αυτήν έχουμε πλέον τη δυνατότητα να δημιουργούμε τάχιστα και αυτόματα μοντέλα, τα οποία είναι ικανά να αναλύσουν περισσότερα και πιο σύνθετα δεδομένα, αλλά και να προσφέρουν πιο γρήγορα και ακριβή αποτελέσματα. Τέλος, με αυτά τα μοντέλα, οι οργανισμοί μπορούν να προβλέψουν και να αποφύγουν άγνωστους κινδύνους ή να ανακαλύψουν κερδοφόρες ευκαιρίες.

2.3 Ακραίες τιμές

Στη στατιστική, ως ακραία τιμή (outlier) ορίζεται μία παρατήρηση η οποία βρίσκεται σε μη φυσιολογική απόσταση από άλλες τιμές σε μία τυχαία δειγματοληψία ενός πληθυσμού. Οι ακραίες τιμές μπορεί να οφείλονται σε σφάλματα κατά τη μέτρηση, στην ελλιπή συλλογή δεδομένων ή στην εμφάνιση δεδομένων που απλώς δε λαμβάνονται υπόψη κατά τη συλλογή τους. Ένα σύνηθες πρόβλημα είναι η σύγχυση των ακραίων τιμών με τους «θορύβους» σε μία δειγματοληψία [1]. Παρόλη την ομοιότητα που παρουσιάζουν μεταξύ τους, στην ουσία διαφέρουν αρκετά.

Όταν αναφερόμαστε στους «θορύβους» της δειγματοληψίας, μιλάμε για δεδομένα που δε μας προσφέρουν καμία χρησιμότητα και δεν έχουν σκοπιμότητα. Η ύπαρξη «θορυβωδών» δεδομένων οδηγεί σε πληθώρα προβλημάτων, καθώς οι μηχανές δεν μπορούν να τα κατανοήσουν σωστά ή να τα διασταυρώσουν. Ορισμένες αιτίες για την ύπαρξη «θορύβων» είναι: ο λανθασμένος τύπος των δεδομένων, οι εσφαλμένες τιμές δεδομένων και οι ελλιπείς τιμές.

Αντίθετα, οι ακραίες τιμές μας παρέχουν χρήσιμες και ενδιαφέρουσες πληροφορίες. Η παρουσία ακραίων τιμών στα δεδομένα αποδίδεται σε: σφάλματα μέτρησης, σφάλμα δειγματοληψίας, λανθασμένη αναφορά και υπερβολικές αλλά αληθείς τιμές.

Είναι σημαντικό πρωτού ληφθεί οποιαδήποτε απόφαση οι ακραίες τιμές να μελετώνται με προσοχή ώστε να καθοριστεί το αν θα πρέπει να διατηρηθούν ή να αφαιρεθούν από το σύνολο δεδομένων [1]. Οι ακραίες τιμές που εμποδίζουν την ανάλυση των δεδομένων είναι προτιμότερο να διαγράφονται, ενώ εκείνες που προσφέρουν σημαντικές πληροφορίες να διατηρούνται.

Αφού αναφερθήκαμε στον ορισμό και στη σημαντικότητα των ακραίων τιμών, πριν προχωρήσουμε στους τρόπους και τις μεθόδους εντοπισμού τους, είναι σημα-

ντικό να αναλύσουμε τους τύπους τους [13].

1. **Ολικές ακραίες τιμές (Global outliers)**: ορίζονται οι ακραίες τιμές των οποίων η απόσταση από το υπόλοιπο σύνολο δεδομένων είναι πολύ μεγάλη.

2. **Υπό συνθήκες ακραίες τιμές (Contextual/Conditional outliers)**: σε αυτήν την περίπτωση μιλάμε για δεδομένα, όπου η τιμή τους διαφέρει σημαντικά από τα υπόλοιπα δεδομένα του ίδιου πλαισίου. Αυτό σημαίνει ότι η ίδια τιμή αν ανήκει σε διαφορετικό πλαίσιο, ίσως να μη θεωρηθεί ακραία. Τέτοιου είδους ακραίες τιμές είναι σύνηθες φαινόμενο σε δεδομένα χρονοσειρών, διότι τα δεδομένα αυτά είναι εγγραφές συγκεκριμένης ποσότητας με την πάροδο του χρόνου και το πλαίσιο στο οποίο ενεργούν είναι σχεδόν πάντα χρονικό.

3. **Συλλογικές ακραίες τιμές (Collective outliers)**: τα δεδομένα μίας υποομάδας του συνόλου δεδομένων θεωρούνται ως συλλογικές ακραίες τιμές, όταν αυτά ως ομάδα αποκλίνουν από το σύνολο. Παρόλα αυτά ως μεμονωμένες τιμές δεν μπορούν να χαρακτηριστούν από μόνες τους ούτε συλλογικές ακραίες τιμές ούτε υπό συνθήκες ακραίες τιμές.

Για να γίνουν κατανοητοί οι τύποι των ακραίων τιμών μπορούμε να χρησιμοποιήσουμε το εξής παράδειγμα. Ένα αεροπλάνο το οποίο προσγειώνεται σε ένα αυτοκινητόδρομο είναι μια ολική ακραία τιμή, διότι κάτι τέτοιο είναι εξαιρετικά απίθανο να συμβεί. Εάν ο αυτοκινητόδρομος ήταν γεμάτος αμάξια αυτό θα θεωρούταν υπό συνθήκες ακραία τιμή αν η χρονική στιγμή που συνέβαινε ήταν σπάνια, δηλαδή αν γινόταν στις δύο μετά τα μεσάνυχτα, όπου συνήθως δεν έχει κίνηση στους δρόμους προτού ξημερώσει και πάει ο κόσμος στις δουλειές του. Αν τώρα όλα τα αυτοκίνητα κινούνταν στην αριστερή λωρίδα του δρόμου την ίδια στιγμή, θα είχαμε συλλογικές ακραίες τιμές επειδή αν και δεν είναι παράλογο να συμβεί κάτι τέτοιο, είναι ασυνήθιστο να κινούνται όλα τα αμάξια την ίδια ακριβώς χρονική στιγμή στην ίδια κατεύθυνση.

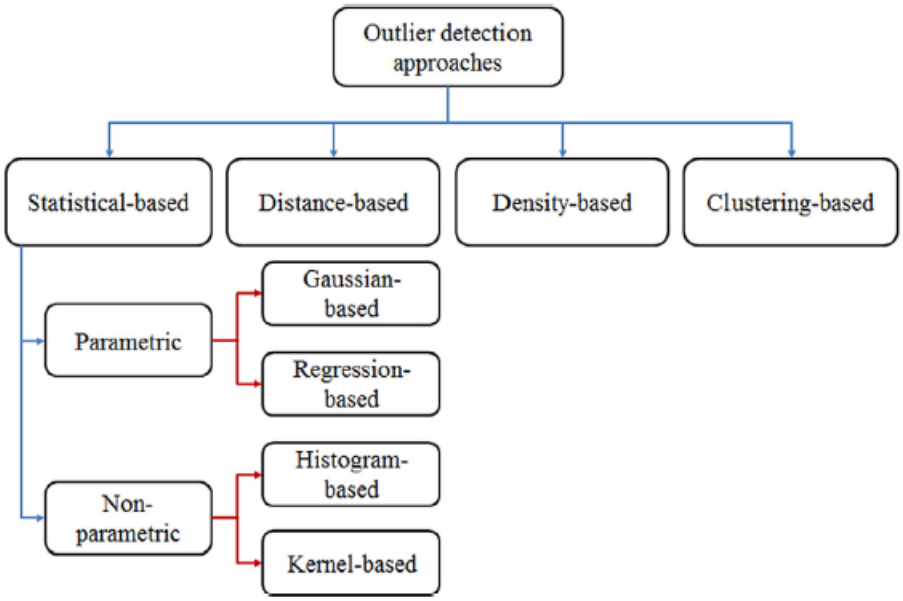
2.4 Ανίχνευση ακραίων τιμών

Η ανίχνευση ακραίων τιμών (Outlier detection) είναι μία διαδικασία που συνήθως χρησιμοποιείται στη φάση της προεπεξεργασίας της ανάλυσης δεδομένων, και

σκοπός της είναι προφανώς η εύρεση των ακραίων τιμών σε ένα σύνολο δεδομένων [14]. Η ανίχνευση των τιμών αυτών είναι πολύ σημαντική. Η ανάλυση δεδομένων μπορεί να επηρεαστεί σε μεγάλο βαθμό και να οδηγήσει σε λάθος συμπεράσματα αν δεν απομακρυνθούν οι ακραίες τιμές. Επίσης, η καλή ανάλυση και κατανόηση των ακραίων τιμών μπορεί να αποφέρει σημαντικές πληροφορίες στον αναλυτή.

Οι μέθοδοι ανίχνευσης ακραίων τιμών έχουν πλέον προσαρμοστεί σε πολλούς κλάδους, όπως αυτοί της ασφάλειας, των επιχειρήσεων και της βιομηχανίας. Στον τομέα της ασφάλειας για παράδειγμα με τις μεθόδους ανίχνευσης, μπορούμε να εντοπίσουμε ύποπτες συναλλαγές των πιστωτικών καρτών. Στον τομέα της υγείας με την ανίχνευση ακραίων τιμών μπορούμε να βοηθήσουμε τους γιατρούς στην ανεύρεση ανωμαλιών, όπως αυτή της πρώιμης ανάπτυξης καρκινικών όγκων. Ανάλογα με τα διάφορα κριτήρια, μπορούμε να κατηγοριοποιήσουμε τις μεθόδους εύρεσης ακραίων τιμών. Εμείς θα τις ομαδοποιήσουμε με βάση τις διαφορετικές τεχνικές που χρησιμοποιούνται. Με αυτόν τον τρόπο ταξινομούνται σε τέσσερις βασικές μεθόδους (Σχήμα 2.1) [1]: βάσει στατιστικών (Statistical-based), απόστασης (Distance-based), ομαδοποίησης (Clustering-based) και πυκνότητας (Density-based).

Σχήμα 2.1: Μέθοδοι προσέγγισης ακραίων τιμών[1]



2.4.1 Στατιστική προσέγγιση

Οι στατιστικολόγοι ήδη από τον 19ο αιώνα είχαν την ανάγκη της ανίχνευσης των ακραίων τιμών. Η παρατήρηση ανωμαλιών ή ασυμφωνιών στα δεδομένα προκατέλαβαν τη στατιστική ανάλυση, κάτι το οποίο οδήγησε στην απομάκρυνσή τους, μέσω διάφορων στατιστικών τεχνικών. Η στατιστική προσέγγιση υποθέτει την ύπαρξη ενός μοντέλου κατανομής ή πιθανοτήτων για το δοθέν σύνολο δεδομένων. Στη συνέχεια, αναγνωρίζει τις ακραίες τιμές σε σχέση με το μοντέλο, με τη χρήση ενός τεστ ασυμφωνίας [15]. Η ανάλυση της στατιστικής προσέγγισης βασίζεται σε πέντε φάσεις:

1. **Συλλογή δεδομένων**: Συλλέγονται τα δεδομένα του εκάστοτε προβλήματος.

2. **Υπολογισμός μέσης τιμής / εξίσωσης γραμμικής παλινδρόμησης**: Υπολογίζεται η μέση τιμή προκειμένου να βρεθεί η κεντρική γραμμή για την τεχνική του διαγράμματος. Το ίδιο συμβαίνει και με τον υπολογισμό της εξίσωσης γραμμικής παλινδρόμησης.

3. **Υπολογισμός των τιμών ανώτατου και κατώτατου ορίου**: Για την τεχνική γραμμικής παλινδρόμησης τα άνω και κάτω όρια ορίζονται από το 95 τις εκατό της εξίσωσης γραμμικής παλινδρόμησης.

4. **Δοκιμή δεδομένων**: Στη φάση αυτή χρησιμοποιώντας τα πραγματικά δεδομένα, τη γραμμή γραμμικής παλινδρόμησης και τα άνω και κάτω όρια, μπορούμε να σχεδιάσουμε την εξίσωση της γραμμικής παλινδρόμησης. Με αυτόν τον τρόπο ανιχνεύουμε τις ακραίες τιμές, όπου βρίσκονται εκτός του άνω και κάτω ορίου της εξίσωσης.

5. **Ανάλυση και σύγκριση των αποτελεσμάτων**: Τέλος, τα αποτελέσματα της δοκιμής δεδομένων χρησιμοποιούνται για τη σύγκριση και ανάλυση των τεχνικών αυτών. Σκοπός αυτών των διαδικασιών είναι η «προπόνηση» των τεχνικών αυτών για τη βελτίωση της απόδοσής τους, σε ότι αφορά την εύρεση των ακραίων τιμών με βάση τη στατιστική προσέγγιση.

Συνεπώς, στη στατιστική προσέγγιση, ο τρόπος εύρεσης των ακραίων τιμών επιτυγχάνεται με την εκπαίδευση ενός μοντέλου για το συγκεκριμένο σύνολο δεδομένων που του δίνεται. Έτσι, οι κανονικές τιμές εμφανίζονται σε περιοχές υψηλής

πιθανότητας για το μοντέλο, ενώ οι ακραίες τιμές σε περιοχές χαμηλής πιθανότητας. Γενικά υπάρχουν αρκετοί τρόποι εκπαίδευσης αυτών των μοντέλων, αλλά όπως φαίνεται και στο Σχήμα 2.1, η στατιστική προσέγγιση χωρίζεται σε δύο μεγάλες κατηγορίες, τις παραμετρικές μεθόδους (parametric methods) και τις μη-παραμετρικές μεθόδους (non-parametric methods) [16].

Παραμετρικοί μέθοδοι

Οι μέθοδοι χωρίζονται ανάλογα των γνώσεων που έχουμε πάνω στο πληθυσμό που μελετάμε. Οι παραμετρικές μέθοδοι είναι κατά κύριο λόγο οι πρώτες μέθοδοι που μελετώνται σε ένα μάθημα στατιστικής [16]. Η γενική ιδέα είναι η ύπαρξη ενός συνόλου δεδομένων που προσδιορίζει το μοντέλο πιθανοτήτων. Στις παραμετρικές μεθόδους συνήθως έχουμε τη γνώση ότι ο πληθυσμός είναι σχεδόν κανονικός ή μπορούμε να τον προσδιορίσουμε με τη βοήθεια της κανονικής κατανομής αφού χρησιμοποιήσουμε το θεώρημα του κεντρικού ορίου. Σε μία κανονική κατανομή υπάρχουν δύο παράμετροι: ο μέσος όρος και η τυπική απόκλιση. Συνεπώς, για να χαρακτηριστεί μια μέθοδος ως παραμετρική, πρέπει να ληφθούν υπόψη οι υποθέσεις που γίνονται για τον πληθυσμό. Το πλεονέκτημα των παραμετρικών μεθόδων σε σύγκριση με των μη-παραμετρικών είναι ότι έχουν μεγαλύτερη στατιστική ισχύ.

1. Μοντέλο Gauss

   Το μοντέλο Gauss (Gaussian based model) είναι ένα στοχαστικό μοντέλο για την αναπαράσταση κανονικά κατανεμημένων υποομάδων σε ένα συνολικό πληθυσμό [17]. Γενικά αυτό το μοντέλο δεν προϋποθέτει τη γνώση σχετικά σε ποια υποομάδα ανήκει ένα σημείο δεδομένων. Με αυτόν τον τρόπο το μοντέλο «μαθαίνει» τις υποομάδες αυτόματα. Λαμβάνοντας υπόψη ότι η ανάθεση υποομάδων δεν είναι γνωστή, το μοντέλο αυτό αποτελεί μία μορφή μη επιβλεπόμενης μάθησης.

2. Μοντέλο Παλινδρόμησης

   Το μοντέλο παλινδρόμησης (regression model) [18] είναι ένα σύνολο στατιστικών διαδικασιών με σκοπό την παρατήρηση των σχέσεων μεταξύ μιας εξαρτημένης μεταβλητής (γνωστή και ως μεταβλητή αποτέλεσμα) και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η ανάλυση των ακραίων τιμών με το

μοντέλο της παλινδρόμησης διακρίνεται σε δύο προσεγγίσεις. Η πρώτη προσέγγιση ονομάζεται αντίστροφη αναζήτηση (reverse search), και στο πλαίσιο της κατασκευάζεται ένα μοντέλο παλινδρόμησης το οποίο εμπεριέχει όλα τα δεδομένα. Έπειτα αυτά με το μεγαλύτερο σφάλμα διαγράφονται από το μοντέλο. Στη δεύτερη προσέγγιση έχουμε το μοντέλο της άμεσης αναζήτησης (direct search), όπου σε αυτή την περίπτωση το μοντέλο που κατασκευάζεται περιέχει ένα μέρος των δεδομένων. Στη συνέχεια το μοντέλο ανακατασκευάζεται αφού προστεθούν κάποια νέα αντικείμενα. Έπειτα ακολουθείται παρόμοια διαδικασία προσθήκης νέων δεδομένων, μόνο που αυτή τη φορά τα αντικείμενα είναι πιο κατάλληλα έχοντας μικρότερες αποκλίσεις. Βασικό μειονέκτημα του μοντέλου αυτού είναι η εξάρτησή του στην υπόθεση κατανομής σφαλμάτων.

Μη Παραμετρικές μέθοδοι

Οι μη παραμετρικές μέθοδοι, σε αντίθεση με τις παραμετρικές, είναι στατιστικές τεχνικές κατά τις οποίες δε χρειάζεται η υπόθεση παραμέτρων για τον πληθυσμό που εξετάζουμε [16]. Δεν έχουν καμία εξάρτηση από το πληθυσμό, καθώς ούτε το σύνολο δεδομένων αλλά ούτε και η κατανομή του είναι πλέον σταθερά. Για αυτόν τον λόγο οι μη παραμετρικές μέθοδοι είναι γνωστές και ως μέθοδοι χωρίς κατανομή. Παρακάτω αναφέρονται κάποιοι λόγοι που καθιστούν τις μη παραμετρικές μεθόδους δημοφιλείς. Ο βασικός λόγος είναι ο μικρότερος βαθμός περιορισμού σε σύγκριση με αυτόν των παραμετρικών μεθόδων. Επίσης, δεν απαιτείται μεγάλος αριθμός υποθέσεων για τον πληθυσμό τον οποίο αναλύουμε. Τέλος, σε αρκετές περιπτώσεις οι μη παραμετρικές μέθοδοι εφαρμόζονται και κατανοούνται πιο εύκολα.

1. Ιστόγραμμα

   Το ιστόγραμμα (Histogram based) είναι μια κατά προσέγγιση αναπαράσταση της κατανομής των αριθμητικών δεδομένων [16]. Συνήθως δείχνουν το πόσες φορές εμφανίζεται ένας συγκεκριμένος τύπος μεταβλητής, σε ένα συγκεκριμένο εύρος. Το πρώτο βήμα για την ανίχνευση των ακραίων τιμών, είναι να κατασκευαστεί το ιστόγραμμα ανάλογα με τα εισαγόμενα δεδομένα. Έπειτα μία απλοποιημένη εξήγηση της εύρεσης των ακραίων τιμών είναι ο έλεγχος του εκάστοτε δεδομένου όσον αφορά την ένταξή του στις ράβδους του ιστογράμματος που δημιουργήθηκαν. Συνεπώς, αν το δεδομένο ανήκει σε κάποια

ράβδο θεωρείται κανονικό, ενώ στην αντίθετη περίπτωση ακραία τιμή. Το πρόβλημα που αντιμετωπίζεται με την επιλογή των μοντέλων του ιστογράμματος είναι το χαμηλό ποσοστό ακριβείας και επιτυχίας των αποτελεσμάτων. Αυτό συμβαίνει διότι αν το μέγεθος των ράβδων δεν επιλεχθεί σωστά, μπορεί να μας οδηγήσει σε λανθασμένα αποτελέσματα και συμπεράσματα.

2. Μέθοδος Kernel

   Η μέθοδος Kernel (Kernel based) [19] μεταφέρει τα δεδομένα από τον χώρο εισόδου σε ένα χώρο υψηλότερων διαστάσεων (χώρος χαρακτηριστικών). Έπειτα η μέθοδος αναζητά συναρτήσεις γραμμικής απόφασης στον χώρο χαρακτηριστικών, οι οποίες γίνονται μη γραμμικές συναρτήσεις στον χώρο εισόδου. Για να επιτευχθεί κάτι τέτοιο, πρέπει η μέθοδος Kernel να αντικαταστήσει το εσωτερικό γινόμενο των παρατηρήσεων με μια λειτουργία Kernel.

Εν κατακλείδι χρειαζόμαστε και τις δύο μεθόδους για την επίλυση των προβλημάτων που μας απασχολούν. Όπως αναφέρθηκε και πιο πάνω οι μη παραμετρικές μέθοδοι είναι πιο κατανοητές και εύκολα διαχειρίσιμες από τις παραμετρικές. Ενώ οι παραμετρικές πολλές φορές είναι πιο αποδοτικές και αποτελεσματικές από τις αντίστοιχες μη παραμετρικές. Παρόλα αυτά η διαφορά απόδοσης των δύο μεθόδων δεν είναι τόσο μεγάλη, κάτι το οποίο μας οδηγεί στην εξέταση του εκάστοτε προβλήματος για την απόφαση και επιλογή της καταλληλότερης μεθόδου.

2.4.2 Προσέγγιση με βάση την απόσταση

Όσον αφορά την εύρεση ακραίων τιμών με βάση την απόσταση [20], αυτό το οποίο διερευνείται είναι η γειτονιά ενός αντικειμένου, η οποία και καθορίζεται από μία δεδομένη ακτίνα της εκάστοτε προσέγγισης. Στην περίπτωση κατά την οποία το αντικείμενο δεν περιβάλλεται από αρκετά άλλα σημεία που ανήκουν στην ίδια γειτονιά, τότε θεωρείται ακραία τιμή βασιζόμενη στην απόσταση. Κάποια εργαλεία εύρεσης της απόστασης μεταξύ των αντικειμένων είναι η ευκλείδια απόσταση, η απόσταση συνημιτόνου, η απόσταση Manhattan κ.λπ. [21]. Οι μέθοδοι αυτοί ασχολούνται με τα πολυδιάστατα δεδομένα, ένα πλεονέκτημα σε αντίθεση με τις προηγούμενες μεθόδους της στατιστικής προσέγγισης. Το κύριο ελάττωμα της χρήσης μεθόδων με βάση την απόσταση είναι η εξάρτησή τους από μία μόνο τιμή της

παραμέτρου που μεταβάλλεται ανάλογα τις συνθήκες. Κάτι τέτοιο μπορεί να καταλήξει σε ψευδείς ενδείξεις στην περίπτωση όπου το σύνολο δεδομένων εμπεριέχει ανομοιογενή πυκνότητα των περιοχών.

2.4.3 Προσέγγιση με βάση την πυκνότητα

Στην περίπτωση προσέγγισης με βάση την πυκνότητα αυτό το οποίο διερευνάται είναι η πυκνότητα ενός αντικειμένου καθώς και των γειτόνων του [21]. Η σύγκριση μεταξύ των πυκνοτήτων του αντικειμένου και των γειτόνων του είναι αυτή που θα καθορίζει αν μία τιμή είναι ακραία ή κανονική. Αν η πυκνότητα του αντικειμένου είναι παρόμοια με αυτή των γειτόνων του τότε αναφερόμαστε σε μία κανονική τιμή, ενώ στην αντίθετη περίπτωση, όπου οι μεταξύ τους πυκνότητες διαφέρουν, το αντικείμενο θεωρείται ως ακραία τιμή. Για να προσδιοριστεί ένα δεδομένο ως ακραία τιμή λαμβάνεται υπόψη η ακραία βαθμολογία που το χαρακτηρίζει. Το εργαλείο που χρησιμοποιείται πιο συχνά σε αυτού του είδους τους αλγορίθμους είναι η απόσταση του κοντινότερου γείτονα k (distance of the kth nearest neighbor) [22]. Ωστόσο είναι ικανοί να ανιχνεύσουν ορισμένους τύπους ”θορύβων”, όταν υπάρχουν ομάδες δεδομένων διαφορετικών πυκνοτήτων. Ορισμένοι αλγόριθμοι της προσέγγισης με βάση την πυκνότητα είναι ο DBscan, ο LOF κ.λπ. [23].

2.4.4 Προσέγγιση με βάση τη συσταδοποίηση

Στη μέθοδο αυτή χρησιμοποιούνται τα μεγέθη των συστάδων που προκύπτουν, προκειμένου να αναγνωριστούν οι ακραίες ομάδες παρατηρήσεων [24]. Η προσέγγιση που είναι βασισμένη στις συστάδες είναι πιο ακριβής σε σχεση με αυτή με βάση την απόσταση. Στο πρώτο στάδιο ελέγχονται οι αριθμοί των δεδομένων που ανήκουν στις συστάδες. Αν ο αριθμός των δεδομένων της συστάδας είναι μικρότερος από τον μέσο όρο, τότε θεωρείται ως ακραίο ολόκληρο το σύνολο δεδομένων της. Σε δεύτερο στάδιο υπολογίζεται η απόλυτη απόσταση ανάμεσα σε ένα σημείο και τη μέση τιμή. Κάθε σύμπλεγμα έχει μία τιμή threshold και στο ενδεχόμενο που η απόλυτη απόσταση είναι μεγαλύτερη από την τιμή αυτή τότε το αντικείμενο αντιμετωπίζεται ως ακραία τιμή. Σε αυτή την ομάδα αλγορίθμων υπάγεται και ο OFP που πρότειναν οι Jiang et al. [25], βασίζεται στον αλγόριθμο k-means και θεωρεί ως ακραίες τιμές τις μικρές συστάδες. Επιπλέον οι Yu et al. [26] εισήγαγαν τη μέθοδο

FindOut με τη βοήθεια του αλγορίθμου WaveCluster [27]. Αυτές οι δύο μέθοδοι χρησιμοποιούνται για αριθμητικά δεδομένα σε αντίθεση με τον αλγόριθμο FindCBLOF που πρότειναν οι He et al. [28], ο οποίος επεξεργάζεται μόνο κατηγορίες χαρακτηριστικών δεδομένων.

Κεφάλαιο 3

Μέθοδοι ανίχνευσης ακραίων τιμών

Στο κεφάλαιο αυτό θα αναλυθούν κάποιες βασικές και γνωστές μέθοδοι ανίχνευσης ακραίων τιμών. Επίσης, θα ακολουθήσει παρουσίαση και ανάλυση παραδειγμάτων και θα ενταχθούν κομμάτια κώδικα για την κάθε μέθοδο αντίστοιχα, με σκοπό την καλύτερη κατανόησή τους.

3.1 Interquartile Range

Η μέθοδος του ενδοτεταρτημοριακού εύρους (IQR) [2] είναι πολύ χρήσιμη για την εύρεση ακραίων τιμών και ανήκει στην κατηγορία της στατιστικής προσέγγισης. Οι ακραίες τιμές είναι παρατηρήσεις που δεν ταιριάζουν με το μοτίβο του συνόλου δεδομένων. Για να αποσαφηνιστεί η μέθοδος αυτή θα πρέπει να αναλυθούν κάποιες έννοιες, με τις οποίες μπορεί να χαρακτηριστεί οποιοδήποτε σύνολο δεδομένων:

1. Ελάχιστη ή χαμηλότερη τιμή του συνόλου δεδομένων.
2. Πρώτο τεταρτημόριο Q1, το οποίο είναι το πρώτο τέταρτο της λίστας του συνόλου δεδομένων.
3. Η διάμεσος του συνόλου, που δείχνει το μέσο όλης της λίστας των δεδομένων.
4. Τρίτο τεταρτημόριο Q3, το οποίο είναι τα τρία τέταρτα της λίστας του συνόλου δεδομένων
5. Μέγιστη ή μεγαλύτερη τιμή του συνόλου δεδομένων.

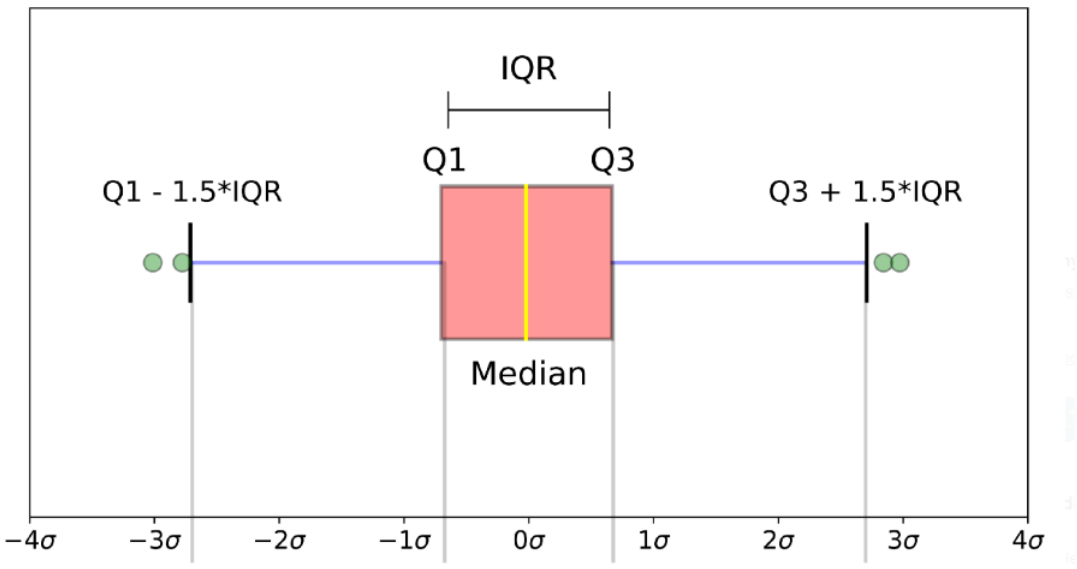
Αυτοί οι πέντε αριθμοί μπορούν να δώσουν πολλά στοιχεία για τα δεδομένα. Το ενδοτεταρτημοριακό εύρος υπολογίζεται από την αφαίρεση του πρώτου από

το τρίτο τεταρτημόριο ($IQR = Q3–Q1$). Το εύρος αυτό μας δείχνει τον τρόπο κατανομής των δεδομένων σε σχέση με τη δίαμεσο. Ο τρόπος τώρα με τον οποίο χρησιμοποιούμε τον ενδοτεταρτημοριακό εύρος για την εύρεση των ακραίων τιμών είναι ο εξής:

1. Υπολογίζουμε το IQR των δεδομένων.
2. Πολλαπλασιάζουμε το IQR επί 1.5.
3. Προσθέτουμε το $IQR \times 1.5$ στο τρίτο τεταρτημόριο, οποιοσδήποτε αριθμός μεγαλύτερος αυτού θεωρείται ακραία τιμή.
4. Αφαιρούμε το $IQR \times 1.5$ στο πρώτο τεταρτημόριο, οποιοσδήποτε αριθμός μικρότερος αυτού θεωρείται ακραία τιμή.

Πρέπει να ληφθεί υπόψη ότι η μέθοδος του ενδοτεταρτημοριακού εύρους είναι μία εμπειρική μέθοδος και τα πιθανά αποτελέσματα της θα πρέπει να εξετάζονται στο πλαίσιο όλου του συνόλου δεδομένων.

Σχήμα 3.1: Παράδειγμα Θηκογράμματος IQR[2]

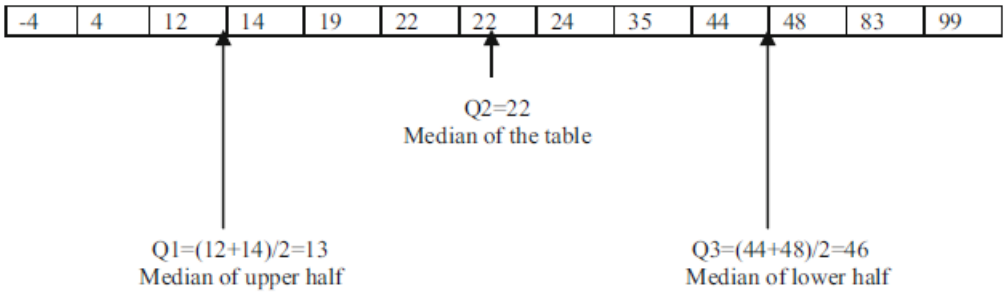


Οι Vinutha et al. [2], παρουσιάζουν το εξής παράδειγμα για την κατανόηση του αλγορίθμου IQR. Έστω ότι έχουμε τις εξής τιμές:

[–4, 4, 12, 14, 19, 22, 22, 24, 35, 44, 48, 83, 99].

Σύμφωνα με το Σχήμα 3.1 [2] καθώς και την ανάλυση της μεθόδου που προηγήθηκε πρέπει να βρούμε τις παραμέτρους κάνοντας τις κατάλληλες πράξεις. Αφού ταξινομηθεί ο πίνακας έχουμε το εξής αποτέλεσμα (Σχήμα 3.2) [2]:

Σχήμα 3.2: Αποτελέσματα IQR[2]



Γνωρίζοντας πλέον τις παραμέτρους και κάνοντας τις κατάλληλες αριθμητικές πράξεις βρίσκουμε το IQR και τα όρια του.

$IQR = Q3 - Q1 = 46 - 13 = 33$

Κάτω όριο $= Q1 - IQR \times 1.5 = 13 - 33 \times 1.5 = -36.5$

Άνω όριο $= Q3 + IQR \times 1.5 = 46 + 33 \times 1.5 = 95.5$

Συνεπώς η τιμή 99 που είναι εκτός των επιτρεπτών ορίων, θεωρείται ακραία τιμή. Στην Απεικόνιση 3.1 [2] παρουσιάζεται μέρος του κώδικα του αλγορίθμου IQR.

Απεικόνιση 3.1: Κώδικας IQR.

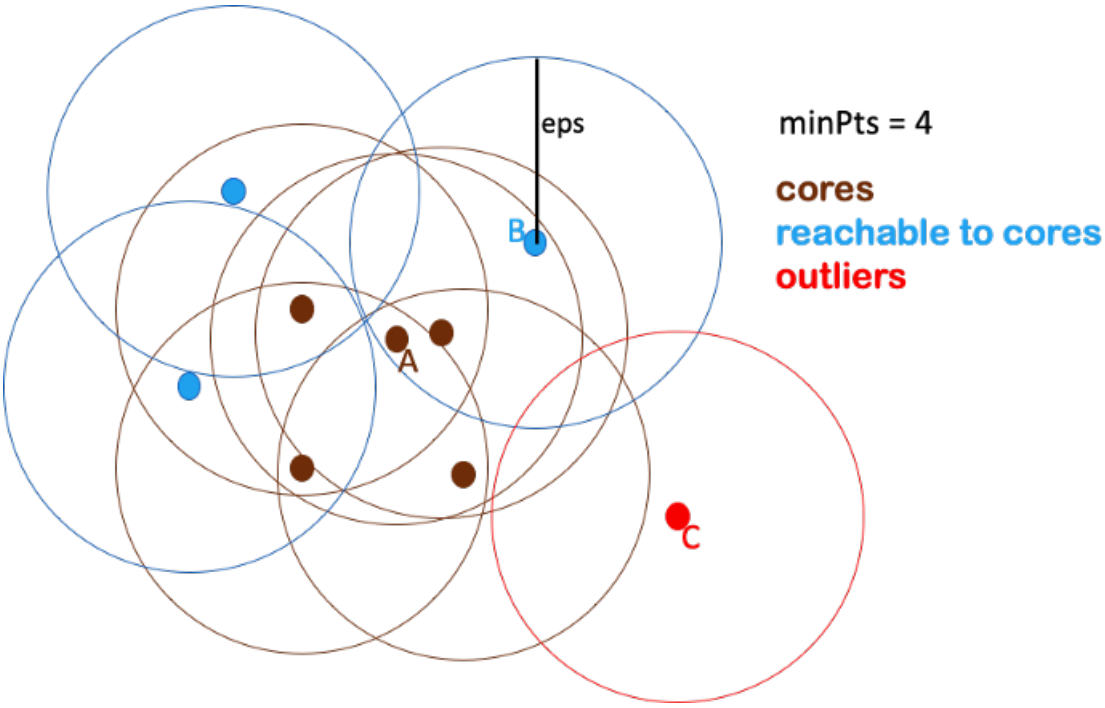
```
data = [-4, 4, 12, 14, 19, 22, 22, 24, 35, 44, 48, 83, 99]
df = data
sort_data = np.sort(data)
quarter1 = np.percentile(data, 25)
quarter2 = np.percentile(data, 50)
quarter3 = np.percentile(data, 75)
IQR = quarter3 - quarter1
low_lim = quarter1 - 1.5 * IQR
up_lim = quarter3 + 1.5 * IQR
outlier =[]
for x in data:
        if ((x> up_lim) or (x<low_lim)):
                outlier.append(x)
```

3.2 Density-Based Spatial Clustering of Applications with Noise

Ο DBscan [29] ανήκει στην κατηγορία των μη επιτηρούμενων αλγορίθμων και επίσης βασίζεται στην πυκνότητα (Density-based). Η λειτουργία του αλγορίθμου ξεκινάει με τη λήψη πολυδιάστατων δεδομένων ως εισόδους, τα οποία ανάλογα με τις παραμέτρους του εκάστοτε μοντέλου χαρακτηρίζονται σε ομάδες. Έπειτα λαμβάνοντας υπόψη τα κριτήρια αυτά καθώς και τις ομάδες που δημιουργήθηκαν, ο αλγόριθμος διαχωρίζει τις ακραίες τιμές του συνόλου δεδομένων. Πιο αναλυτικά κατηγοριοποιεί «πυκνά ομαδοποιημένα» σημεία δεδομένων σε ένα ενιαίο σύμπλεγμα. Εντοπίζει ομάδες σε μεγάλα χωρικά σύνολα δεδομένων αξιολογώντας την πυκνότητα στο συγκεκριμένο σημείο δεδομένων. Ο αλγόριθμος DBscan [1] απαιτεί την ύπαρξη μόνο δύο παραμέτρων. Πρώτη παράμετρος το epsilon, όπου είναι η ακτίνα του κύκλου που δημιουργείται γύρω από κάθε σημείο δεδομένων και βοηθάει στον έλεγχο της πυκνότητας της περιοχής. Δεύτερη παράμετρος τα minpoints, τα οποία συμβολίζουν τον ελάχιστο αριθμό σημείων δεδομένων και είναι απαραίτητο να βρίσκονται μέσα στο κύκλο, ώστε το σημείο δεδομένων να χαρακτηριστεί ως σημείο πυρήνα. Επίσης, απαραίτητη προϋπόθεση είναι ο αριθμός του minpoints να είναι

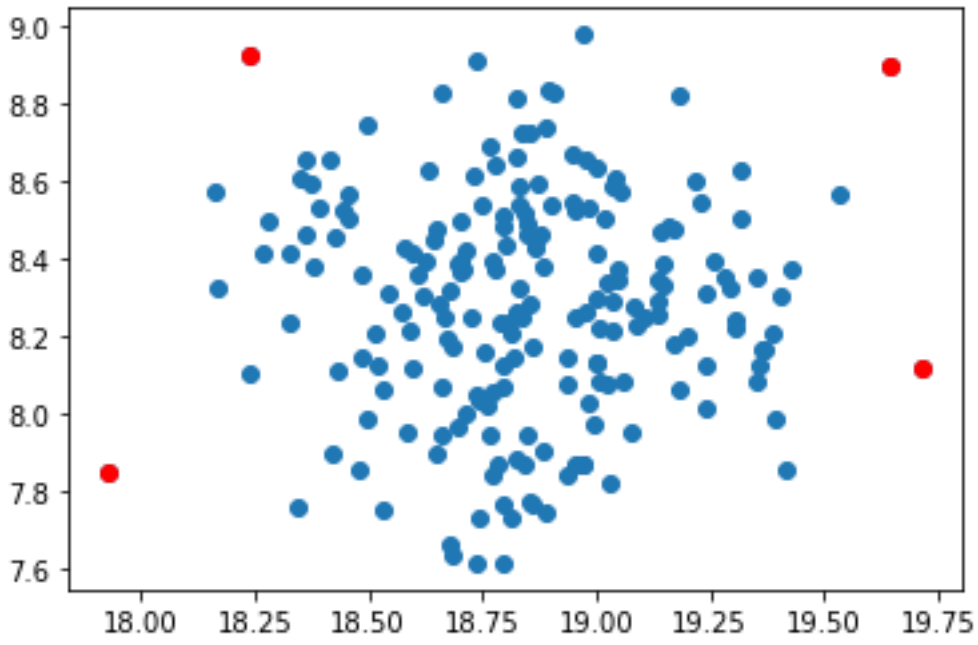
τουλάχιστον μία μονάδα πάνω από τις διαστάσεις των δεδομένων, συνεπώς τουλάχιστον τρεις. Τα σημεία δεδομένων χαρακτηρίζονται σε τρεις καταστάσεις, ανάλογα τα minpoints. Το σημείο πυρήνα (Core point) που όπως προαναφέρθηκε πρέπει να περιέχει τον ελάχιστο αριθμό minpoints. Το σημείο συνόρων (Border point) κατά το οποίο ο αριθμός των σημείων είναι μικρότερος από το ζητούμενο minpoints. Τέλος, τα ακραία σημεία (noise) στα οποία δεν υπάρχουν άλλα σημεία δεδομένων εντός της ακτίνας epsilon.

Σχήμα 3.3: Παράδειγμα αλγορίθμου DBscan[1]



Όπως βλέπουμε και στο Σχήμα 3.3 για κάθε σημείο δεδομένων δημιουργείται ένας κύκλος με ακτίνα epsilon [29]. Σύμφωνα με τους κανόνες που προηγήθηκαν, τα δεδομένα χωρίζονται σε σημεία πυρήνα (καφέ χρώμα), σημεία συνόρων (γαλάζιο χρώμα) και ακραίες τιμές (κόκκινο χρώμα). Επιπλέον, στο Σχήμα 3.4 αναφέρεται ένα παράδειγμα του αλγορίθμου με epsilon = 0,28 και minpoints = 20, στο οποίο με κόκκινο χρώμα έχουν χαρακτηριστεί οι ακραίες τιμές, ενώ με μπλε οι κανονικές. Επιπροσθέτως στην Απεικόνιση 3.2 παρουσιάζεται κομμάτι του κώδικα που χρησιμοποιήθηκε για την παρουσίαση του Σχήματος 3.4.

Σχήμα 3.4: Παράδειγμα γραφήματος DBscan[1]



Απεικόνιση 3.2: Κώδικας DBscan

```
# Creates a 200 random sample dataset
random.seed(7)
x, _ = make_blobs(n_samples=200, centers=1, cluster_std=.3,
   center_box=(20, 5))
dbscan = DBSCAN(eps = 0.28, min_samples = 20)
# Fits the model and gets the prediction data
pred = dbscan.fit_predict(x)
# Puts the negative outputs as outliers
index = where(pred == -1)
outliers = x[index]
```

3.3 Z-score

Η μέθοδος z-score [30] επισημαίνεται ως μέτρο της απόκλισης μίας παρατήρησης από το ενδεχόμενο αποτέλεσμα που είναι η μέση τιμή. Το Z ορίζεται ως ο αριθμός των τυπικών αποκλίσεων από τη μέση τιμή και υπολογίζεται από τον παρακάτω τύπο: $Z = \frac{X-\mu}{\sigma}$.

Όπου:

X: η εκάστοτε εξεταζόμενη τιμή

μ: η μέση τιμή

σ: ο αριθμός τυπικών αποκλίσεων

Σύμφωνα με τον τύπο γίνεται προφανές ότι το μέτρο της απόκλισης που υπολογίζεται από το Z-score παρέχει ένα μηχανισμό, που προσδιορίζει το μέγεθος της απόκλισης του εξεταζόμενου δεδομένου από άλλες παρατηρήσεις του συνόλου δεδομένων. Στην περίπτωση που η τιμή του Z-score είναι υπερβολικά μεγάλη, τότε η παρατήρηση που εξετάζεται θεωρείται ακραία τιμή. Όταν κάθε παρατήρηση του συνόλου δεδομένων λάβει την αντίστοιχη z-score τιμή της, τότε η κατανομή της μεταβλητής ονομάζεται τυπική κανονική κατανομή και έχει μέσο όρο το μηδέν και τυπική απόκλιση ένα. Η μέθοδος Z-score απαιτεί από το χρήστη την επιλογή μίας τιμής για το διαχωρισμό των ακραίων τιμών από τις κανονικές. Ο ευρέως χρησιμοποιούμενος αριθμός τυπικών αποκλίσεων του κάτω ορίου είναι το -3, ενώ του πάνω ορίου το +3. Αυτά τα όρια αποκοπής επιλέγονται διότι το 99.7% των τιμών ενός συνόλου βρίσκονται ανάμεσα στις τιμές -3 και +3 σε μία τυπική κανονική κατανομή.

Στο Σχήμα 3.6 απεικονίζεται το παράδειγμα που παρουσιάζουν οι Kaliyaperumal et al. [3]. Στην πρώτη περίπτωση του πίνακα έχοντας τη μέση τιμή και την τυπική απόκλιση, προκύπτει η τιμή Z-score κατά την οποία μόνο η τιμή της παρατήρησης 50 είναι ακραία τιμή. Αντίστοιχα στη δεύτερη περίπτωση, αφού έχει αφαιρεθεί η τιμή 50 που αντιμετωπίστηκε ως ακραία, έχοντας τις νέες τιμές για τη μέση τιμή και την τυπική απόκλιση, λαμβάνεται ξανά η τιμή Z-score για κάθε δείγμα. Γίνεται φανερό ότι οι παρατηρήσεις 48 και 49 είναι ακραίες τιμές.

Απεικόνιση 3.3: Κώδικας Z-score.

```
# random data points to calculate z-score
data = [5, 5, 5, -99, 5, 5, 5, 5, 5, 5, 88, 5, 5, 5]
# calculate mean
mean = np.mean(data)
# calculate standard deviation
sd = np.std(data)
# determine a threhold
threshold = 2
# create empty list to store outliers
outliers = []
# detect outlier
for i in data:
    z = (i-mean)/sd # calculate z-score
    if abs(z) > threshold:  # identify outliers
        outliers.append(i) # add to the empty list
# print outliers
print("The detected outliers are: ", outliers)
```

Σχήμα 3.5: Αποτελέσματα Z-score[3]

```
print("The detected outliers are: ", outliers)
The detected outliers are:  [-99, 88]
```

Σχήμα 3.6: Παράδειγμα τιμών Z-score[3]

Case – 1 ($\bar{x} = 126.34, sd = 67.55$)						Case – 2 ($\bar{x} = 118.20, sd = 35.78$)					
Obs No.	x_i	**Z-Scores**	**Obs. No.**	x_i	**Z-Scores**	**Obs No.**	x_i	**Z-Scores**	**Obs. No.**	x_i	**Z-Scores**
1	70	-0.83	26	110	-0.24	1	70	-1.35	26	110	-0.23
2	75	-0.76	27	113	-0.20	2	75	-1.21	27	113	-0.15
3	81	-0.67	28	114	-0.18	3	81	-1.04	28	114	-0.12
4	84	-0.63	29	117	-0.14	4	84	-0.96	29	117	-0.03
5	84	-0.63	30	117	-0.14	5	84	-0.96	30	117	-0.03
6	84	-0.63	31	119	-0.11	6	84	-0.96	31	119	0.02
7	85	-0.61	32	121	-0.08	7	85	-0.93	32	121	0.08
8	85	-0.61	33	121	-0.08	8	85	-0.93	33	121	0.08
9	86	-0.60	34	127	0.01	9	86	-0.90	34	127	0.25
10	92	-0.51	35	130	0.05	10	92	-0.73	35	130	0.33
11	93	-0.49	36	131	0.07	11	93	-0.70	36	131	0.36
12	95	-0.46	37	132	0.08	12	95	-0.65	37	132	0.39
13	95	-0.46	38	134	0.11	13	95	-0.65	38	134	0.44
14	96	-0.45	39	135	0.13	14	96	-0.62	39	135	0.47
15	96	-0.45	40	136	0.14	15	96	-0.62	40	136	0.50
16	96	-0.45	41	139	0.19	16	96	-0.62	41	139	0.58
17	98	-0.42	42	153	0.39	17	98	-0.56	42	153	0.97
18	99	-0.40	43	155	0.42	18	99	-0.54	43	155	1.03
19	101	-0.38	44	166	0.59	19	101	-0.48	44	166	1.34
20	101	-0.38	45	169	0.63	20	101	-0.48	45	169	1.42
21	105	-0.32	46	172	0.68	21	105	-0.37	46	172	1.50
22	106	-0.30	47	175	0.72	22	106	-0.34	47	175	1.59
23	108	-0.27	48	236	1.62	23	108	-0.29	**48**	**236**	**3.29**
24	109	-0.26	49	236	1.62	24	109	-0.26	**49**	**236**	**3.29**
25	110	-0.24	**50**	**525**	**5.90**	25	110	-0.23	50	-	-

3.4 Isolation Forest

Ο αλγόριθμος Isolation Forest [4] (Σχήμα 3.7) είναι βασισμένος στα δέντρα απόφασης και ανήκει στην κατηγορία των μοντέλων χωρίς επίβλεψη. Η φιλοσοφία του αλγορίθμου είναι ότι οι ανωμαλίες που ανιχνεύονται στα σημεία δεδομένων είναι «λίγες» και «διαφορετικές». Το Isolation Forest αποτελείται από ένα σύνολο δυαδικών δέντρων που ονομάζονται Isolation Trees. Η ακολουθία του αλγορίθμου είναι η εξής:

1. Ένα τυχαίο υποσύνολο δεδομένων, του συνόλου δεδομένων που εισάγεται, επιλέγεται και μεταβιβάζεται σε ένα δυαδικό δέντρο.

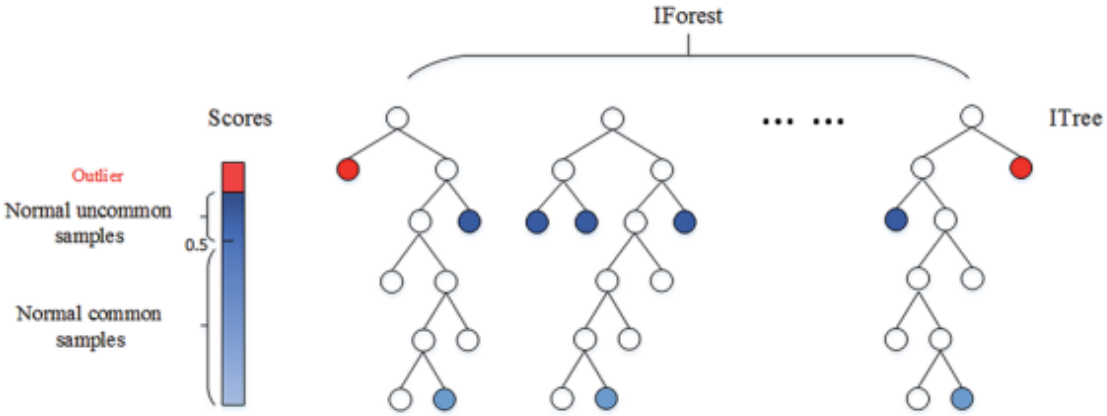
2. Για να ξεκινήσει η διακλάδωση του δέντρου πρέπει πρώτα να επιλεχθεί ένα τυχαίο σημείο από το σύνολο δεδομένων, καθώς αυτή έχει τυχαίο όριο το οποίο

πρέπει να βρίσκεται ανάμεσα στη μέγιστη και ελάχιστη τιμή του επιλεγμένου σημείου.

3. Στην περίπτωση που η τιμή του σημείου δεδομένων είναι χαμηλότερη απο το επιλεγμένο όριο τότε κατευθείνεται στον αριστερό κλάδο του δέντρου. Σε αντίθετη περίπτωση κατευθείνεται στο δεξιό κλάδο.

4. Η διαδικασία από το 2ο βήμα και μετά επαναλαμβάνεται έως ότου κάθε σημείο του συνόλου δεδομένων έχει απομονωθεί ή εάν στην περίπτωση που ορίζεται ο αλγόριθμος φτάσει στο μέγιστο βάθος.

Σχήμα 3.7: Μέθοδος Isolation Forest[4]



Όλα τα παραπάνω βήματα επαναλαμβάνονται για τη δημιουργία τυχαίων δυαδικών δέντρων. Με αυτό το τρόπο δημιουργείται ένα σύνολο από Isolation Trees και επιτυγχάνεται η ολοκλήρωση του μοντέλου εκπαίδευσης. Κατά τη διάρκεια βαθμολόγησης των δέντρων ένα σημείο δεδομένων διασχίζει τα εκπαιδευμένα δέντρα, με σκοπό να αναθέσει μία «βαθμολογία ανωμαλίας» (anomaly score) σε κάθε σημείο δεδομένων, βασιζόμενο στο βάθος των δέντρων που απαιτείται για να φτάσει στο σημείο αυτό. Ο τύπος για τη «βαθμολογία ανωμαλίας» (s) είναι ο εξής [4]:

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}}$$

$$c(n) = 2H(n-1) - (2(n-1)/n)$$

Όπου:

$h(x)$: το μήκος του μονοπατιού του δείγματος x μετρούμενο από το ριζικό κόμβο μέχρι το τερματισμό στον εξωτερικό κόμβο

$H(i)$: είναι ένας αρμονικός αριθμός που δίνεται από τη σχέση

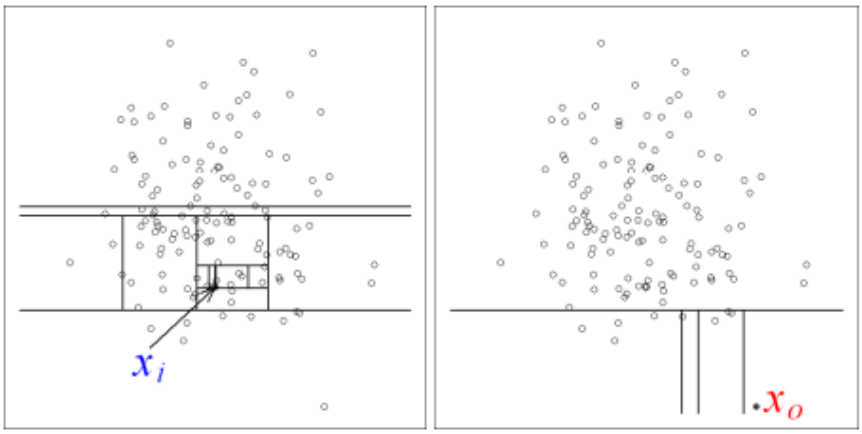
$ln(i) + 0,5772156649$ (Σταθερά Euler)

$c(n)$: είναι ο μέσος όρος του h(x) για τον αριθμό εξωτερικών κόμβων n

Συνεπώς έχουμε τις ακόλουθες αξιολογήσεις:

1. Αν η βαθμολογία τείνει στη μονάδα, τότε θεωρείται ακραία τιμή.
2. Αν η βαθμολογία είναι πολύ μικρότερη του μισού, τότε είναι κανονική τιμή.
3. Αλλίως αν η βαθμολογία είναι ίση με μισό, τότε το συνολικό δείγμα δεν έχει κάποια διακριτή ακραία τιμή.

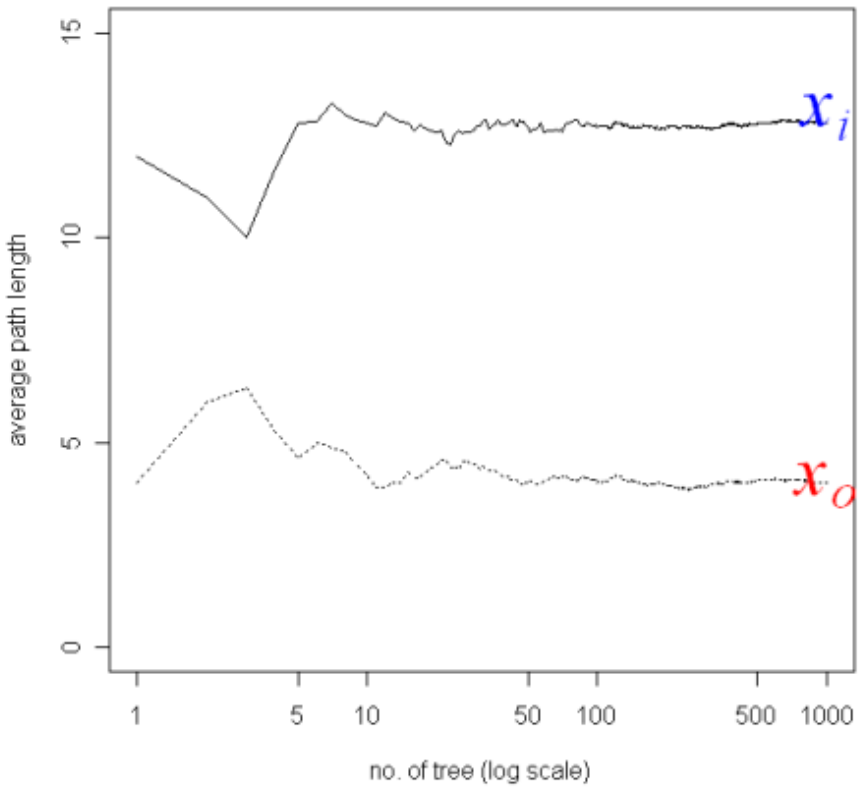
Σχήμα 3.8: Παράδειγμα Isolation Trees[4]



(a) Isolating x_i (b) Isolating x_o

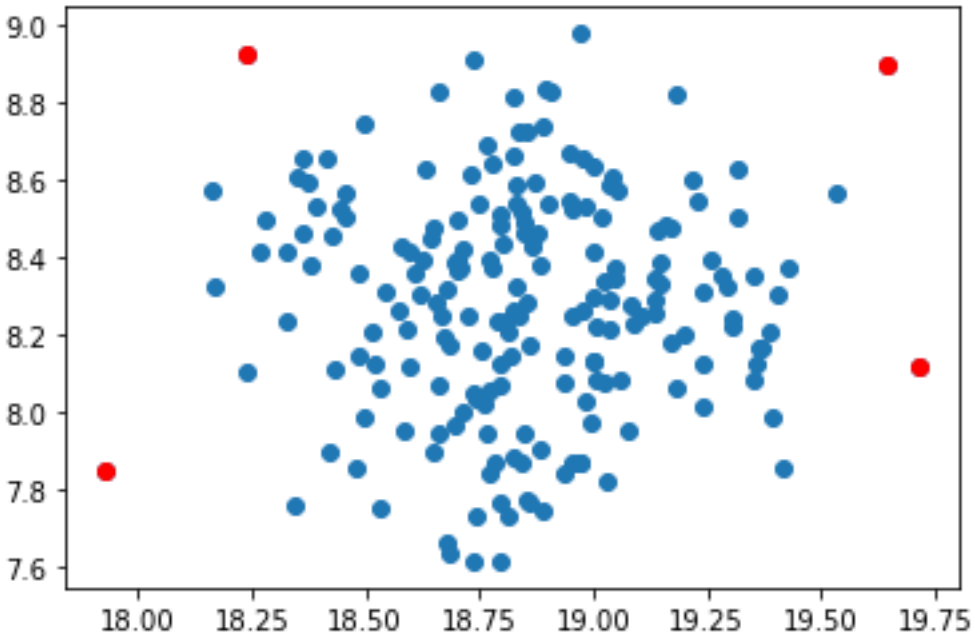
Στο παράδειγμα που χρησιμοποιούν οι Liu et al. [5] όπως φαίνεται και στο Σχήμα 3.8, γίνεται ξεκάθαρη η διαφορά των ακραίων τιμών από τις κανονικές τιμές μέσω του Isolation Forest. Στη συνέχεια στο Σχήμα 3.9 [5] φαίνεται η σύγκριση των αριθμών κόμβων μίας ακραίας και μίας κανονικής τιμής, που όπως προαναφέρθηκε η ακραία τιμή έχει μικρότερο μονοπάτι για να φθάσει στον εξωτερικό κόμβο.

Σχήμα 3.9: Γράφημα μονοπατιών των δέντρων[5]



(c) Average path lengths converge

Σχήμα 3.10: Παράδειγμα γραφήματος Isolation Forest[5]



Αντίστοιχα για τον αλγόριθμο Isolation Forest το παράδειγμα, που χρησιμοποιείται για τη σύγκριση των μεθόδων, φαίνεται στο Σχήμα 3.10, όπου πάλι με κόκκινο χρώμα είναι οι ακραίες τιμές και με μπλε οι κανονικές. Επιπλέον στην Απεικόνιση 3.4 παρουσιάζεται ένα κομμάτι του κώδικα του αλγορίθμου Isolation Forest.

Απεικόνιση 3.4: Κώδικας Isolation Forest.

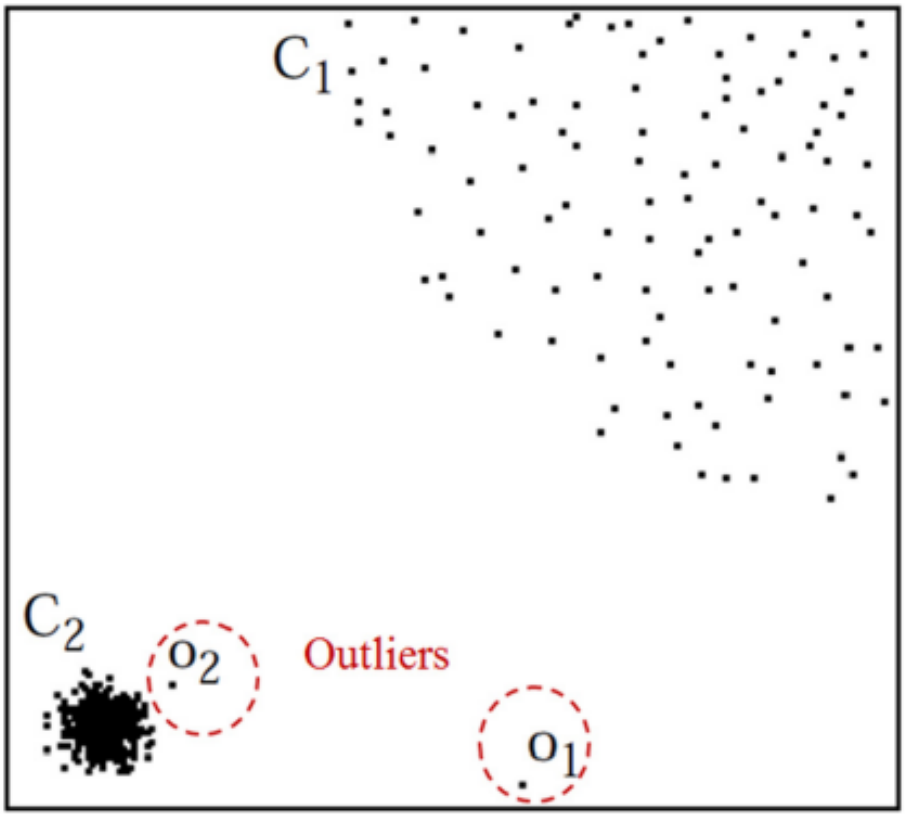
```
# Creates 200 samples with with given variables
random.seed(7)
x, _ = make_blobs(n_samples=200, centers=1, cluster_std=.3,
   center_box=(20, 5))
# Defines the model, creating 100 trees in the forest and
   contamination is the parameter that defines the outliers
iforest = IsolationForest(n_estimators=100)
print(iforest)
# Fits the model in dataset and finds the score for each data
iforest.fit(x)
scores = iforest.score_samples(x)
# Creates the theshhold value and puts the 2% score of the
   samples as anomalies/outliers
thresh = quantile(scores, .02)
print(thresh)
# Finds the outliers using the threshold score
index_out = where(scores <= thresh)
outliers = x[index_out]
```

3.5 Local Outlier Factor

Ο αλγόριθμος LOF [6] ανήκει στην κατηγορία της μάθησης χωρίς επίβλεψη και στις μεθόδους με προσέγγιση την πυκνότητα. Η λειτουργία αυτού του αλγορίθμου είναι η σύγκριση των πυκνοτήτων ενός αντικειμένου του συνόλου δεδομένων με την πυκνότητα των γειτόνων του. Λαμβάνοντας υπόψη ότι οι ακραίες τιμές βρίσκονται σε περιοχές χαμηλής πυκνότητας, ο λόγος θα είναι μεγαλύτερος για ανώμαλα σημεία δεδομένων. Ένας κανόνας που μπορεί να ακολουθηθεί είναι ότι ένα κανονικό σημείο έχει τιμή LOF ανάμεσα στο ένα και ενάμισι , αντίθετα με τα ακραία σημεία που έχουν πολύ υψηλότερη τιμή LOF. Για την εύρεση της τιμής LOF και την κατανόηση του τύπου πρέπει να αναλυθούν οι εξής ορισμοί [31]:

1. dp,q: είναι η απόσταση μεταξύ δύο σημείων p και q.
2. k-distance: ταξινομεί τις αποστάσεις από το σημείο p έως τα άλλα σημεία δεδομένων καθώς και την απόσταση του σημείου p στο k σημείο δεδομένων
3. k nearest neighbors: είναι το σημείο που έχει οριστεί από το σημείο p και έχει απόσταση μικρότερη από $k - distp$ και ορίζεται ως Nkp
4. Απόσταση προσβασιμότητας: $reach - dist_kp, r = max(k - distr, dp, r)$
5. Πυκνότητα τοπικής προσβασιμότητας: είναι η αντίστροφη άλγεβρα (όταν το γινόμενο δύο παραμέτρων ισούται με τη μονάδα) μεταξύ του μέσου όρου της απόστασης προσβασιμότητας του σημείου p και των k πλησιέστερων γειτόνων του και ορίζονται ως:
   $lrd_kp = [\frac{\sum_o(\in N_kp)*reach-dist_kp,r}{|N_kp|}]^{-1}$
6. Τοπικός ακραίος παράγοντας (LOF): είναι ο μέσος όρος των αναλογιών της πυκνότητας τοπικής προσβασιμότητας του p και αυτή των k-nearest neighbors του p και ορίζεται ως:
   $LOF_kp = \frac{\sum_o(\in N_k(p))\frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$

Σχήμα 3.11: Ανάλυση LOF[6]



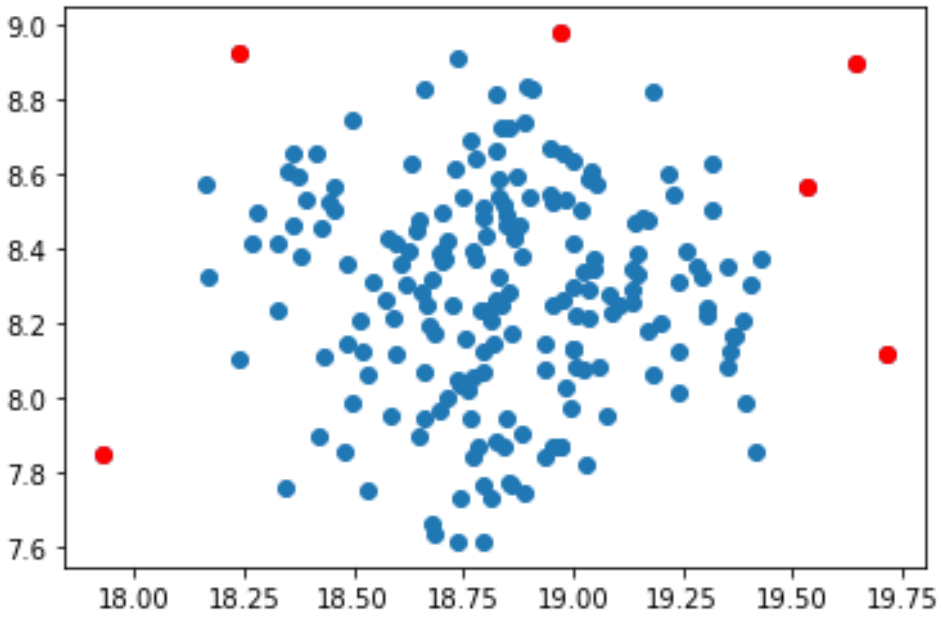
Όπως γίνεται φανερό και στο Σχήμα 3.11 που παρουσιάζουν οι Alshawabkeh et al. [7], το σύμπλεγμα C1 έχει πολλά περισσότερα σημεία δεδομένων από ότι το C2, αντίθετα όμως η πυκνότητα της συστάδας C2 είναι αισθητά μεγαλύτερη από την πυκνότητα της C1. Στο σύμπλεγμα C1 οι αποστάσεις μεταξύ ενός στοιχείου q και του πλησιέστερου γείτονά του, λόγω της πυκνότητας της συστάδας είναι φανερά μεγαλύτερες, σε σύγκριση με την απόσταση του στοιχείου o2 και του πλησιέστερου γείτονά του από το σύμπλεγμα C2. Επομένως το στοιχείο o2 δε θεωρείται ακραία τιμή.

Συνεπώς η χρήση μιας απλής προσέγγισης, όπως αυτή του πλησιέστερου γείτονα, δεν αρκεί για την επιτυχημένη ανίχνευση ακραίων τιμών σε τέτοιου είδους περιπτώσεις. Το στοιχείο o1 όμως, με τη χρήση του εργαλείου του πλησιέστερου γείτονα χαρακτηρίζεται ως ακραία τιμή. Ωστόσο ο αλγόριθμος LOF είναι ικανός να συλλάβει και τα δύο δεδομένα (o1, o2) ως ακραίες τιμές, λόγω του γεγονότος ότι δίνει βάση στην πυκνότητα γύρω από αυτά τα σημεία.

Επιπλέον στο Σχήμα 3.12 εμφανίζονται με κόκκινο χρωματισμό οι ακραίες τιμές,

ενώ αντίστοιχα με μπλε οι κανονικές.

Σχήμα 3.12: Παράδειγμα γραφήματος LOF[7]

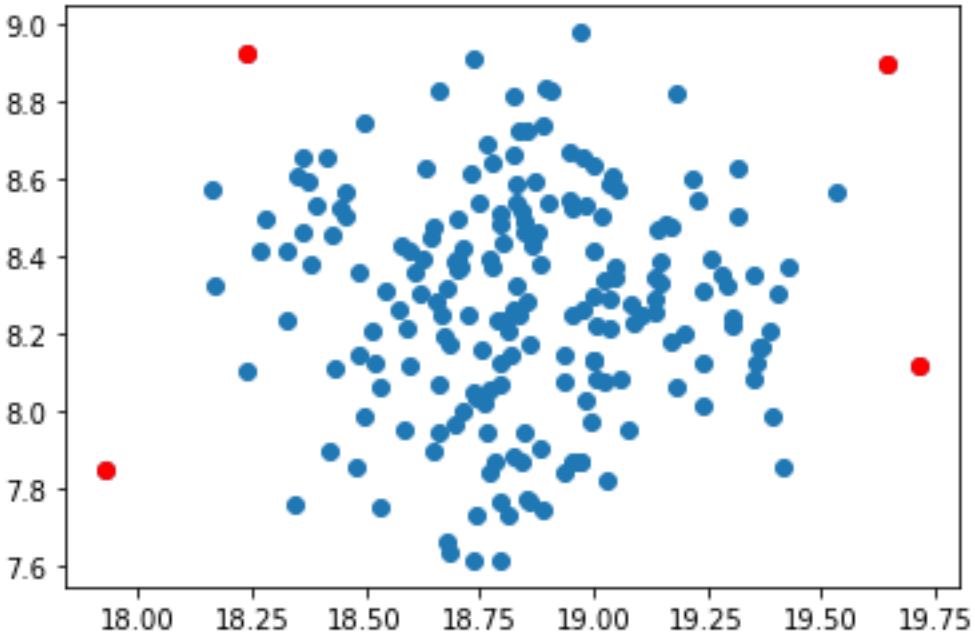


3.6 Elliptical Envelope

Η μέθοδος Elliptical Envelope [8] ανήκει στην κατηγορία της μηχανικής μάθησης χωρίς επίβλεψη. Η μέθοδος αυτή μοντελοποιεί τα δεδομένα ως μία κατανομή Gauss υψηλής διάστασης, με πιθανή συμμεταβλητότητα μεταξύ των χαρακτηριστικών δεδομένων. Έχει ως στόχο τη δημιουργία μίας έλλειψης με τέτοια δομή ώστε η πλειοψηφία των στιγμιοτύπων των παρατηρήσεων να ταιριάζουν σε αυτήν. Τα στιγμιότυπα δεδομένων που βρίσκονται μακριά από την έλλειψη θεωρούνται ως ακραίες τιμές. Η μέθοδος Elliptic Envelope χρησιμοποιεί τον προσδιοριστή ελάχιστης συνδιακύμανσης FAST (FAST-MCD) για την εκτίμηση του σχήματος καθώς και του μεγέθους της έλλειψης. Ο αλγόριθμος FAST-MCD επιλέγει μη τέμνοντα υποσύ-

νολα δεδομένων, με σκοπό τον υπολογισμό του μέσου όρου μ καθώς και του πίνακα συμμεταβλητότητας C, κάθε χαρακτηριστικού δεδομένου που ανήκει στο εκάστοτε υποσύνολο.

Σχήμα 3.13: Παράδειγμα γραφήματος Elliptical Envelope[8]



Όπως φαίνεται και στο Σχήμα 3.13, του οποίου τα αποτελέσματα δίνονται από τον κώδικα της Απεικόνισης 3.5, με κόκκινο χρωματισμό χαρακτηρίζονται οι ακραίες τιμές και με μπλε οι κανονικές.

Απεικόνιση 3.5: Κώδικας Elliptical Envelope

```
# Creates a 200 random sample dataset
random.seed(7)
x, _ = make_blobs(n_samples=200, centers=1, cluster_std=.3,
   center_box=(20, 5))
elenv = EllipticEnvelope()
print(elenv)
elenv.fit(x)
scores = elenv.score_samples(x)
thresh = quantile(scores, .02)
print(thresh)
anom_index = where(scores <= thresh)
outliers = x[anom_index]
```

3.7 Mahalanobis Distance

Με ένα μαθηματικό τρόπο σκέψης για να βρεθούν οι ακραίες τιμές μιας δειγματοληψίας, θα πρέπει να διευκρινιστούν το σχήμα και η δομή του συνόλου δεδομένων. Ας υποθέσουμε ότι υπάρχει ένα σύννεφο από σημεία δεδομένων στον R^2 χώρο με ελλειπτική μορφή, τότε ορισμένα σημεία είναι πιο κοντά στο κέντρο από άλλα. Ωστόσο δε μπορούμε να συμπεράνουμε ότι τα πιο απομακρυσμένα σημεία δεν αποτελούν μέρος του δείγματος σε σύγκριση με τα κοντινά σημεία. Συνεπώς συνίσταται η χρήση μιας απόστασης που λαμβάνει υπόψη το σχήμα του συνόλου των παρατηρήσεων, μια τέτοια απόσταση είναι η Mahalanobis [32] που συμβολίζεται ως εξής:

$$d = \sqrt{(x-\mu)^T \Sigma^{-1} (x - /mu)}$$

Όπου x είναι ένα διάνυσμα μεταβλητών $x = (x_1, x_2, x_3, \cdot, x_k)$, το $= (_1, _2, \cdot, _k)$ ένα διάνυσμα διαστάσεων k και ο που είναι ένας συμμετρικός πίνακας $k \times k$. Από ένα σύνολο δεδομένων σε ένα δείγμα $[X_1, X_2, \cdot, X_n]$, μπορούν να ληφθούν ο

μέσος όρος του δείγματος $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, καθώς και το δείγμα πίνακα διακύμανσης $S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$. Επομένως, με τη χρήση του αλγορίθμου Mahalanobis Distance μία τιμή X_i θεωρείται ακραία όταν τηρεί το εξής κριτήριο: $\sqrt{(X_i - \bar{X})^T S^{-1}(X_i - \bar{X})} > c_k$, για ένα ορισμένο συντελεστή c_k.

Στο Σχήμα 3.14 που παρουσιάζουν οι Maesschalck et al. [9] απεικονίζεται το σύνολο δεδομένων που χρησιμοποιούνται για το παράδειγμα του αλγορίθμου Mahalanobis Distance. Στη συνέχεια στο Σχήμα 3.15 [9] παρουσιάζονται οι τιμές MD για κάθε παρατήρηση του συνόλου δεδομένων για τις τιμές x_1 και x_2. Παρατηρώντας τη στήλη MD καταλαβαίνουμε ότι οι παρατηρήσεις 6, 8 και 9 έχουν μεγαλύτερη τιμή από τις υπόλοιπες. Τέλος στο Σχήμα 3.16 [9] βλέπουμε την έλλειψη που δημιουργείται σύμφωνα με το σύνολο δεδομένων. Η παρατήρηση 6 είναι απομακρυσμένη από τις υπόλοιπες και εκτός των ορίων της έλλειψης.

Απεικόνιση 3.6: Κώδικας Mahalanobis Distance

```
# calculateMahalanobis Function to calculate
# the Mahalanobis distance
def calculateMahalanobis(y=None, data=None, cov=None):
    y_mu = y - np.mean(data)
    if not cov:
        cov = np.cov(data.values.T)
    inv_covmat = np.linalg.inv(cov)
    left = np.dot(y_mu, inv_covmat)
    mahal = np.dot(left, y_mu.T)
    return mahal.diagonal()
data = {'Price':[100000, 800000, 650000, 700000, 860000,
730000, 400000, 870000, 780000, 400000],
'Distance':[16000, 60000, 300000, 10000, 252000, 350000,
260000, 510000, 2000, 5000],
'Emission':[300, 400, 1230, 300, 400, 104, 632, 221, 142, 267],
'Performance':[60, 88, 90, 87, 83, 81, 72, 91, 90, 93],
'Mileage':[76, 89, 89, 57, 79, 84, 78, 99, 97, 99]
}
# Creating dataset
df = pd.DataFrame(data,columns=['Price', 'Distance',
                                'Emission','Performance',
                                'Mileage'])
# Creating a new column in the dataframe that holds
# the Mahalanobis distance for each row
df['Mahalanobis'] = calculateMahalanobis(y=df, data=df[[
  'Price', 'Distance', 'Emission','Performance', 'Mileage']])
# calculate p-value for each mahalanobis distance
df['p'] = 1 - chi2.cdf(df['Mahalanobis'], 3)
# display first five rows of dataframe
print(df)
```

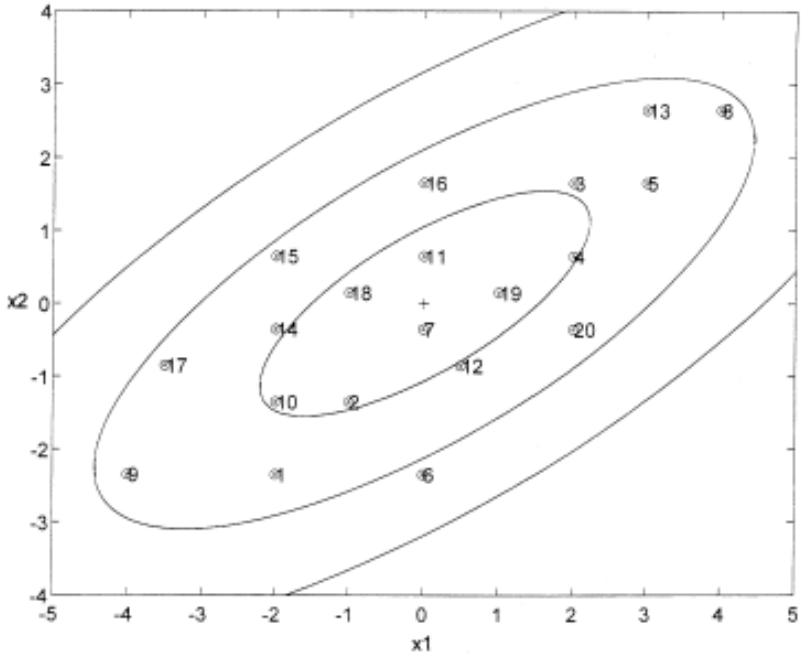
Σχήμα 3.14: Σύνολο Δεδομένων[9]

Object number (i)	x_1	x_2	x_3	x_4	x_1 centered	x_2 centered	x_3 centered	x_4 centered
1	4.00	3.00	1.00	2.00	−2.000	−2.350	−2.125	−1.245
2	5.00	4.00	2.00	3.50	−1.000	−1.350	−1.125	0.255
3	8.00	7.00	3.00	4.00	2.000	1.650	−0.125	0.755
4	8.00	6.00	5.00	4.00	2.000	0.650	1.875	0.755
5	9.00	7.00	2.00	3.00	3.000	1.650	−1.125	−0.245
6	6.00	3.00	5.00	3.00	0.000	−2.350	1.875	−0.245
7	6.00	5.00	3.00	2.50	0.000	−0.350	−0.125	−0.745
8	10.00	8.00	2.00	3.00	4.000	2.650	−1.125	−0.245
9	2.00	3.00	1.50	3.40	−4.000	−2.350	−1.625	0.155
10	4.00	4.00	3.00	3.00	−2.000	−1.350	−0.125	−0.245
11	6.00	6.00	6.00	4.00	0.000	0.650	2.875	0.755
12	6.50	4.50	0.00	2.00	0.500	−0.850	−3.125	−1.245
13	9.00	8.00	5.00	5.00	3.000	2.650	1.875	1.755
14	4.00	5.00	1.00	1.00	−2.000	−0.350	−2.125	−2.245
15	4.00	6.00	3.00	5.00	−2.000	0.650	−0.125	1.755
16	6.00	7.00	2.00	4.00	0.000	1.650	−1.125	0.755
17	2.50	4.50	6.00	4.00	−3.500	−0.850	2.875	0.755
18	5.00	5.50	8.00	3.00	−1.000	0.150	4.875	−0.245
19	7.00	5.50	1.00	2.50	1.000	0.150	−2.125	−0.745
20	8.00	5.00	3.00	3.00	2.000	−0.350	−0.125	−0.245
$\bar{x}$	6.00	5.350	3.125	3.245				

Σχήμα 3.15: Τιμές Mahalanobis Distance[9]

Object number (i)	ED	MD
1	3.0859	1.5464
2	1.6800	0.9122
3	2.5928	1.0814
4	2.1030	0.9652
5	3.4238	1.3576
6	2.3500	2.2133
7	0.3500	0.3296
8	4.7982	1.8947
9	4.6392	1.8278
10	2.4130	0.9549
11	0.6500	0.6122
12	0.9862	1.0640
13	4.0028	1.7186
14	2.0304	1.0983
15	2.1030	1.8097
16	1.6500	1.5540
17	3.6017	1.8037
18	1.0112	0.7664
19	1.0112	0.5629
20	2.0304	1.5710

Σχήμα 3.16: Γράφημα Mahalanobis Distance



3.8 One Class Support Vector Machine

Η μέθοδος One Class SVM [33] ανήκει στην κατηγορία ημι-επιτηρούμενων αλγορίθμων, αλλά θα μπορούσε να ενταχθεί και στους μη-επιτηρούμενους αλγορίθμους. Έχει ως στόχο την εκμάθηση ενός ορίου απόφασης με σκοπό το βέλτιστο διαχωρισμό των σημείων και του σημείου αρχής. Ο One Class SVM χρησιμοποιεί μία άρρητη συνάρτηση μετασχηματισμού φ(*), που ορίζεται από τον πυρήνα με στόχο την προβολή των δεδομένων σε έναν υψηλότερο χώρο διαστάσεων. Έπειτα με την εκμάθηση του ορίου απόφασης ο αλγόριθμος διαχωρίζει την πλειονότητα των δεδομένων από το σημείο προέλευσης. Το υπολειπόμενο κλάσμα των παρατηρήσεων που βρίσκονται στην άλλη πλευρά του ορίου απόφασης χαρακτηρίζονται ως ακραίες τιμές. Ο πυρήνας Gauss εγγυάται την ύπαρξη του ορίου απόφασης. Στην περίπτωση που όλες οι εγγραφές του πυρήνα δεν είναι αρνητικές, μπορούμε να καταλήξουμε στο συμπέρασμα ότι τα δεδομένα του χώρου βρίσκονται στο ίδιο τεταρτημόριο. Συνεπώς, ο πυρήνας Gauss καθιστάται κατάλληλος για την αντιμετώπιση οποιουδήποτε συνόλου δεδομένων. Έτσι η συνάρτηση g(*), ορίζεται ως εξής:

$$g(x) = w^T(x) - \varrho$$

όπου w είναι το κάθετο διάνυσμα στο όριο απόφασης και ο όρος μεροληψίας.

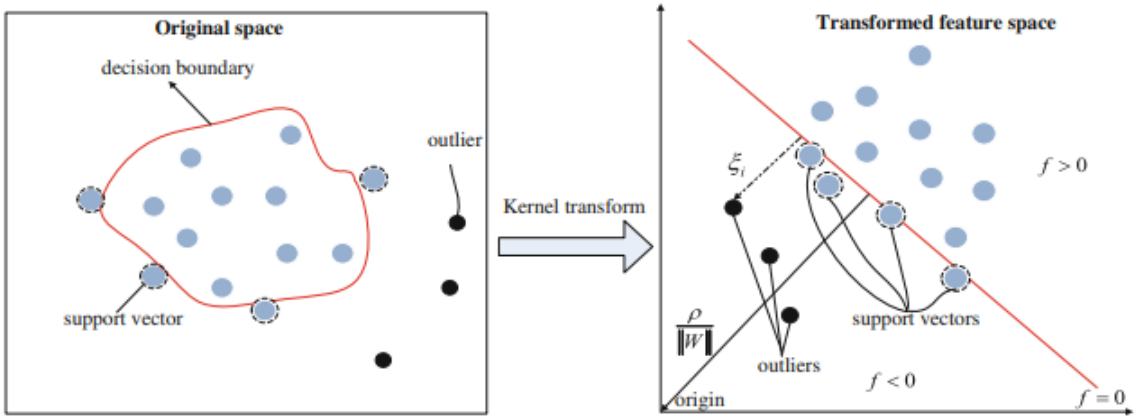
Η συνάρτηση του ορίου απόφασης που χρησιμοποιεί ο One Class SVM είναι η:

$$f(x) = sgn(g(x))$$

Η συνάρτηση επιστρέφει θετική τιμή για τα κανονικά δεδομένα και αρνητική για τα ακραία.

Στο Σχήμα 3.17 παρουσιάζεται ένα παράδειγμα από τους Yang et al. [10]. Στο αριστερό μέρος της εικόνας φαίνεται η ταξινόμηση του συνόλου δεδομένων στον αρχικό χώρο, ενώ στο δεξί ο χώρος μετά το μετασχηματισμό Kernel. Το όριο που περικλείει τα δεδομένα με μπλε χρώμα δημιουργείται από τον ταξινομητή One Class SVM. Συνεπώς τα δεδομένα με μπλε χρωματισμό είναι οι κανονικές τιμές, αυτά με το μαύρο οι ακραίες τιμές και τέλος τα στοιχεία με μπλε χρώμα και μαύρο περίβλημα είναι τα διανύσματα στήριξης που βρίσκονται πάνω στο όριο που δημιουργείται.

Σχήμα 3.17: Παράδειγμα One Class SVM[10]



Απεικόνιση 3.7: Κώδικας One Class SVM

```
# Creates a 200 random sample dataset
X, Y= make_blobs(n_features=2, centers=1, n_samples=100,
   random_state=42)
print("Dataset Size : ", X.shape)
nu = 0.05  # theory says it should be an upper bound of the
   fraction of outliers
ocsvm = OneClassSVM(kernel='rbf', gamma=0.05, nu=nu)
ocsvm.fit(X)
preds = ocsvm.predict(X)
# Defines the outliers and the valid data
X_outliers = X[preds == -1]
X_valid = X[preds != -1]
print("Original Samples : ",X.shape[0])
print("Number of Outliers : ", X_outliers.shape[0])
print("Number of Normal Samples : ", X_valid.shape[0])
print("Outliers :\n", X_outliers)
print("Valid data :\n", X_valid[0:20])
```

Κεφάλαιο 4

Πειραματικό Μέρος

4.1 Στόχος των πειραμάτων

Ο στόχος των πειραμάτων της διπλωματικής εργασίας, είναι η σύγκριση μεθόδων εύρεσης ακραίων τιμών σε μια δειγματοληψία. Η σύγκριση των μεθόδων αυτών μας βοηθάει στο να καταλήξουμε στο συμπέρασμα, αν τελικά είναι σωστό να αφαιρεθούν οι ακραίες τιμές από ένα σετ δεδομένων. Οι αλγόριθμοι που εκτελούνται στο πείραμα αναλύθηκαν στο κεφάλαιο 3 και συγκεκριμένα είναι οι εξής: DBscan, Elliptical Envelope, Gaussian Mixture Model, Isolation Forest, Local Outlier Factor, Mahalanobis distance, OneClassSVM. Τα πειράματα εκτελούνται στο προγραμματιστικό περιβάλλον Spyder μέσω του λογισμικού Anaconda. Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι Pandas, NumPy, Matplotlib και scikit-learn. Η σύγκριση πραγματοποιείται με τους αλγόριθμους μηχανικής μάθησης Random Forest και Support Vector Machine.

4.2 Αναμενόμενα Αποτελέσματα

Ως στόχος της διπλωματικής εργασίας είναι η σύγκριση μεθόδων εύρεσης ακραίων τιμών. Αυτή η σύγκριση πραγματοποιείται σε τρία στάδια. Το πρώτο στάδιο είναι η σύγκριση των ποσοστών επιτυχίας του κάθε αλγορίθμου, με είσοδο τα αρχικά σετ δεδομένα χωρίς καμία παρέμβαση σε αυτά. Στο δεύτερο στάδιο γίνεται η σύγκριση των αλγορίθμων ως προς τα ποσοστά που εμφανίζουν, με είσοδο τα σετ δεδομένων με την προσθήκη 5% και 10% ακραίων τιμών σε αυτά. Τρίτο και τελευταίο στάδιο είναι αυτό κατά το οποίο έχουν ανιχνευθεί και αφαιρεθεί όλες οι ακραίες τιμές του σετ δεδομένων για κάθε αλγόριθμο, με στόχο τη σύγκριση των αποτελεσμάτων τους.

Τα αναμενόμενα αποτελέσματα σύμφωνα με τη θεωρία είναι ότι στο πρώτο στάδιο θα πρέπει να παρουσιαστούν τα μεγαλύτερα ποσοστά επιτυχίας της πρόβλεψης. Έπειτα θα ακολουθούν τα ποσοστά του τρίτου σταδίου των συγκρίσεων και τέλος τα χαμηλότερα ποσοστά θα πρέπει να είναι του δεύτερου σταδίου με τις ακραίες τιμές.

4.3 Συλλογή Δεδομένων

Η συλλογή των σετ δεδομένων έγινε μέσω της ηλεκτρονικής πλατφόρμας UCI Machine Learning Repository. Τα σετ δεδομένων που επιλέχθηκαν διαμερίζονται σε δύο περιπτώσεις, της κατηγοριοποίησης (Classification) και της παλινδρόμησης (Regression). Συνολικά χρησιμοποιήθηκαν δέκα σετ δεδομένων, πέντε για την κατηγορία Classification και αντίστοιχα πέντε για την Regression. Στα σχήματα 4.1 και 4.2 παρουσιάζονται τα dataset που χρησιμοποιήθηκαν. Σε όλα τα σετ δεδομένων οι ακραίες τιμές προστέθηκαν εκ των υστέρων, βασιζόμενες στον εξής κανόνα. Εάν το εύρος των τιμών ενός σετ δεδομένων κυμαίνεται από x έως y, τότε τα outliers θα παίρνουν τιμές στο εύρος του μισού της ελάχιστης τιμής και του διπλάσιου της μέγιστης $[x/2, 2 \times y]$. Επιπλέον στην περίπτωση όπου τα σετ δεδομένων είχαν ελλιπή τιμές στις μετρήσεις, αυτές αφαιρέθηκαν με σκοπό την ακριβέστερη μέτρηση και αξιολόγηση των αλγορίθμων.

Πίνακας 4.1: Σετ δεδομένων κατηγοριοποίησης [https://archive.ics.uci.edu/ml/datasets/]

Breast Cancer Wisconsin	:	Breast+Cancer+Wisconsin+
User Knowledge Modeling	:	User+Knowledge+Modeling
Breast Cancer Coimbra	:	Breast+Cancer+Coimbra
Glass	:	Glass+Identification
HCV	:	HCV+data

Πίνακας 4.2: Σετ δεδομένων παλινδρόμησης [https://archive.ics.uci.edu/ml/datasets/]

Air Quality	:	air+quality
Employees Productivity	:	Productivity+Prediction+of+Garment+Employees
Forest Fires	:	forest+fires
Grisoni	:	QSAR+Bioconcentration+classes+dataset
Qsar	:	QSAR+aquatic+toxicity

4.4 Εργαλεία και μετρικές κώδικα

Προκειμένου να μπορέσουμε να αξιολογήσουμε τους αλγορίθμους πρέπει να χρησιμοποιηθούν κάποια εργαλεία που μας προσφέρει η βιβλιοθήκη scikit-learn. Αυτά τα εργαλεία αφορούν τον διαχωρισμό των στοιχείων σε σύνολα εκπαίδευσης και δοκιμών, καθώς και την εφαρμογή αλγορίθμων μηχανικής μάθησης και συγκεκριμένα των Random Forest και Support Vector Machine. Στην κατηγορία Classification η μετρική που χρησημοποιήθηκε είναι από τη βιβλιοθήκη scikit-learn και ονομάζεται accuracy score. Αντίστοιχα στην κατηγορία Regression και πάλι από την ίδια βιβλιοθήκη η μετρική που εφαρμόστηκε είναι η R squared (r2 score). Στις απεικονίσεις 4.1 και 4.2 παρουσιάζονται τα κομμάτια κώδικα όσων ειπώθηκαν παραπάνω. Στον κώδικα παρατηρούμε την εφαρμογή τόσο των εργαλείων μηχανικής μάθησης Random Forest και Support Vector Machine στην κατηγορία κατηγοριοποίησης, όσο και στην κατηγορία παλινδρόμησης.

Απεικόνιση 4.1: Αξιολόγηση Classification

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn import metrics
## Load the csv file to dataframe ##
df = pd.read_csv (r'...Path of csv file...')
x = df.values['columns of variables']
y = df.values['column of the target']
## Defining the models for SVM and RF ##
clc = SVC(kernel="linear")
clf=RandomForestClassifier(n_estimators=100)
## Split the data to 80% train and 20% test ##
x_train, x_test, y_train, y_test = train_test_split(x, y,
   test_size=0.2)
    clf.fit(x_train,y_train.ravel())
    clc.fit(x_train,y_train.ravel())
    yrf_pred=clf.predict(x_test)
    ysvm_pred=clc.predict(x_test)
## Calculate and print the accuracy of algorithms through
the accuracy score metric ##
print("RF accuracy:", metrics.accuracy_score(y_test,yrf_pred))
print("SVM accuracy", metrics.accuracy_score(y_test,ysvm_pred))
```

Απεικόνιση 4.2: Αξιολόγηση Regression

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.metrics import r2_score
## Load the csv file to dataframe ##
df = pd.read_csv (r'...Path of csv file...')
x = df.values['columns of variables']
y = df.values['column of the target']
## Defining the models for SVM and RF ##
clc = SVC(kernel="linear")
clf=RandomForestClassifier(n_estimators=100)
## Split the data to 80% train and 20% test ##
x_train, x_test, y_train, y_test = train_test_split(x, y,
   test_size=0.2)
    clf.fit(x_train,y_train.ravel())
    clc.fit(x_train,y_train.ravel())
    yrf_pred=clf.predict(x_test)
    ysvm_pred=clc.predict(x_test)
## Calculate and print the accuracy of algorithms through the r
   2 score metric ##
print("RF accuracy:", r2_score(y_test, yrf_pred))
print("SVM accuracy", r2_score(y_test, ysvm_pred))
```

4.5 Αφαίρεση ακραίων τιμών

Ένα πρόβλημα που αντιμετωπίστηκε ήταν η εύρεση του συνόλου των ακραίων τιμών της δειγματοληψίας. Στο σετ δεδομένων που είναι της μορφής csv αρχείου, υπάρχουν αρκετά columns για κάθε αντικείμενο. Συνεπώς πρέπει να ληφθούν υπόψη οι ακραίες τιμές από όλα τα columns. Για αυτόν τον λόγο ο εκάστοτε αλγόριθμος τρέχει σε μία επανάληψη (for) για όλες τις στήλες που θέλουμε να βρούμε τις

ακραίες τιμές. Με αυτό τον τρόπο σε κάθε επανάληψη βρίσκουμε τις ακραίες τιμές εκείνης της συνθήκης και τις προσθέτουμε σε ένα πίνακα. Έπειτα απαλοίφουμε τυχόν τιμές που βρέθηκαν για περισσότερες από μία φορά ως ακραίες και έχουμε την τελική λίστα με τις τιμές που αφαιρέθηκαν. Στην Απεικόνιση 4.3 αναγράφεται ο κώδικας που μας οδηγεί στη λύση του παραπάνω προβλήματος. Σκοπός της απαλοιφής των ακραίων τιμών του κάθε column είναι η βελτιστοποίηση της πρόβλεψης του αλγορίθμου μηχανικής μάθησης.

Απεικόνιση 4.3: Αφαίρεση ακραίων τιμών

```
list = ["...Includes all the column names from the csv file..."
   ]
for i in list:
    ##Puts all the objects that their values are greater
    than the thresh to the out array##
    indeX = where(scores <= thresh)
    outliers = X[indeX]
    out = csv[indeX][:, 0]
    outs.append(out)
##Creates a list from the out array and removes
any duplicate value##
outs_li = [item for sublist in outs for item in sublist]
outs_li.sort()
sums_out = []
[sums_out.append(x) for x in outs_li if x not in sums_out]
```

4.6 Σετ δεδομένων Κατηγοριοποίησης

4.6.1 Breast Cancer Coimbra

Το σετ δεδομένων breast cancer coimbra αποτελείται από εννέα στήλες οι οποίες είναι οι εξής: BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resitin , MCP-1 και η τελευταία Label που προσδιορίζει αν ο ασθενής είναι υγιής ή όχι. Στα Σχήματα 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 και 4.7 παρατηρούμε τις ακραίες τιμές (10%) που ανιχνεύει ο εκάστοτε αλγόριθμος που χρησιμοποιείται, για την κάθε ανεξάρτητη μεταβλητή. Ως σταθερή μεταβλητή έχει οριστεί το BMI. Όπως αναφέρθηκε στην ενότητα 4.4 για την απαλοιφή των ακραίων τιμών λαμβάνεται υπόψη το σύνολο αυτών, έπειτα απ' όλες τις επαναλήψεις εύρεσης τους για κάθε αλγόριθμο. Συνεπώς στους πίνακες 4.5 και 4.6 παρατηρούμε τα ποσοστά από τις προβλέψεις των αλγόριθμων μηχανικής μάθησης Random Forest και Support Vector Machine με 10% ακραίες τιμές αντίστοιχα. Όπως είναι φανερό τα αποτελέσματα δεν είναι ιδανικά και απέχουν από τα αναμενόμενα. Κάτι τέτοιο οφείλεται στις αδυναμίες του κάθε αλγορίθμου καθώς και στην ιδιορρυθμία του εκάστοτε σετ δεδομένων. Συγκεκριμένα παρατηρούμε ότι τα αποτελέσματα με τον αλγόριθμο μηχανικής μάθησης SVM απέχουν αρκετά από αυτά του Random Forest και πως ο αλγόριθμος OneClassSVM δίνει τα χειρότερα αποτελέσματα. Κάτι παρόμοιο συμβαίνει και στους Πίνακες 4.3 και 4.4, με τη διαφορά ότι λόγω της ύπαρξης λιγότερων ακραίων τιμών στο σετ δεδομένων τα ποσοστά που συμπεριλαμβάνουν τις ακραίες τιμές 5% είναι στις περισσότερες περιπτώσεις μεγαλύτερα από αυτά των 10%. Επιπλέον η ακρίβεια που παρουσιάζουν τα σετ δεδομένων με 5% ακραίες τιμές, στην περίπτωση που ανιχνευθούν και αφαιρεθούν είναι μικρότερη σε σύγκριση με τα ποσοστά που παρουσιάζουν τα σετ με 10%.

Πίνακας 4.3: Ποσοστά του Breast Cancer Coimbra για 5% με Random Forest

Coimbra data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	68.75%	70.80%	67.73%
Elliptical Envelope	70%	74%	66.36%
GMM	66.67%	70.80%	70.95%
Isolation Forest	67.08%	66%	73.64%
LOF	69.58%	71.60%	69.05%
Mahalanobis distance	70.42%	75.60%	75.71%
One Class SVM	76.67%	72.40%	65.62%

Πίνακας 4.4: Ποσοστά του Breast Cancer Coimbra για 5% με Support Vector Machine

Coimbra data set with SVM	Original data	With 5% outliers	Without outliers
DBscan	70%	74.80%	71.82%
Elliptical Envelope	70%	78%	67.73%
GMM	68.33%	76%	74.76%
Isolation Forest	71.25%	71.20%	70.91%
LOF	69.58%	72.80%	70.95%
Mahalanobis distance	72.92%	75.60%	75.24%
One Class SVM	73.33%	76.40%	68.13%

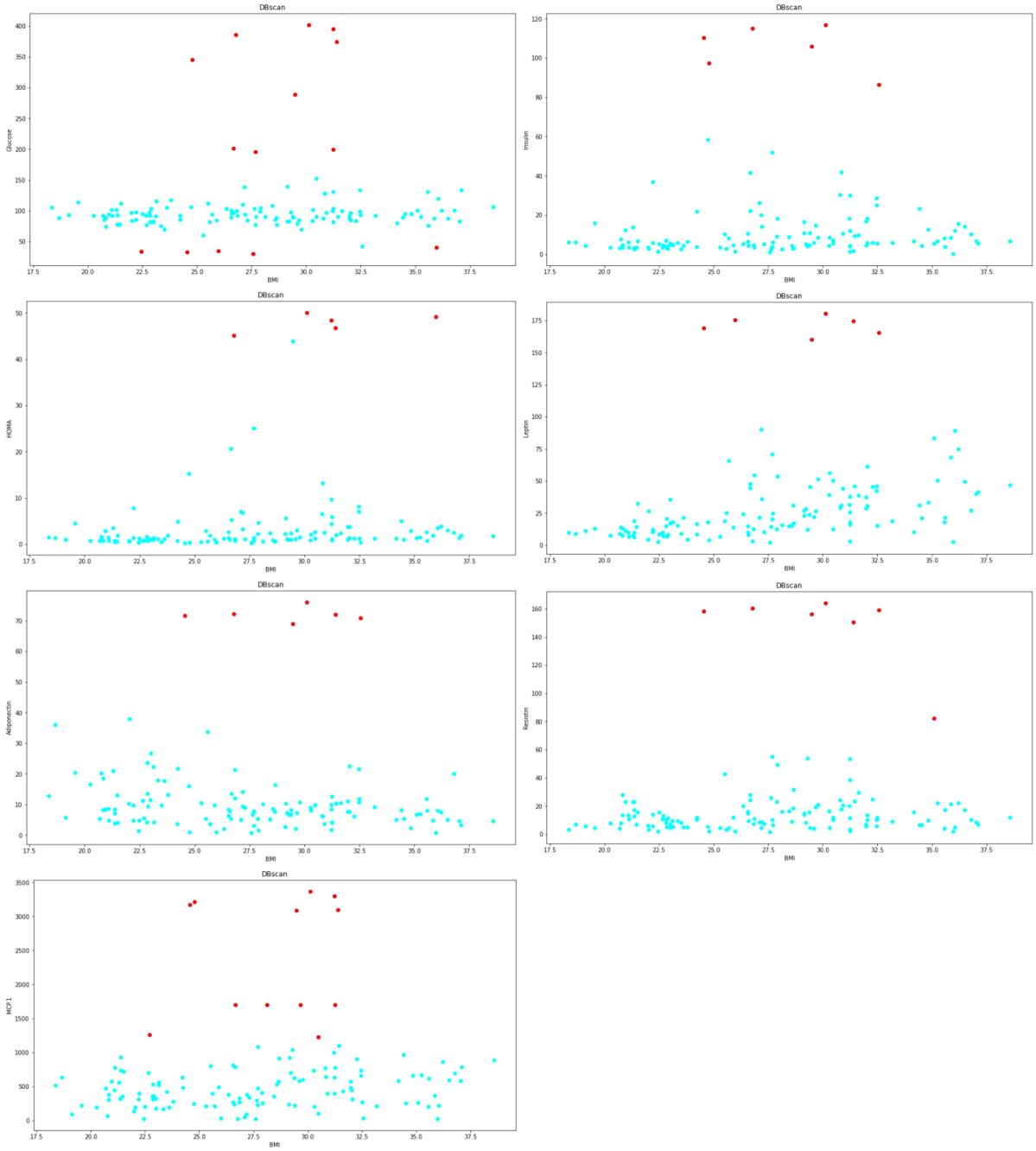
Πίνακας 4.5: Ποσοστά του Breast Cancer Coimbra για 10% με Random Forest

Coimbra data set with RF	Original data	With 10% outliers	Without outliers
DBscan	71.25%	67.69%	73.48%
Elliptical Envelope	75%	71.15%	70.42%
GMM	72.50%	61.92%	72.08%
Isolation Forest	70.83%	71.54%	69.58%
LOF	76.67%	65.77%	69.58%
Mahalanobis distance	78.75%	72.69%	66.67%
One Class SVM	72.92%	72.69%	67.50%

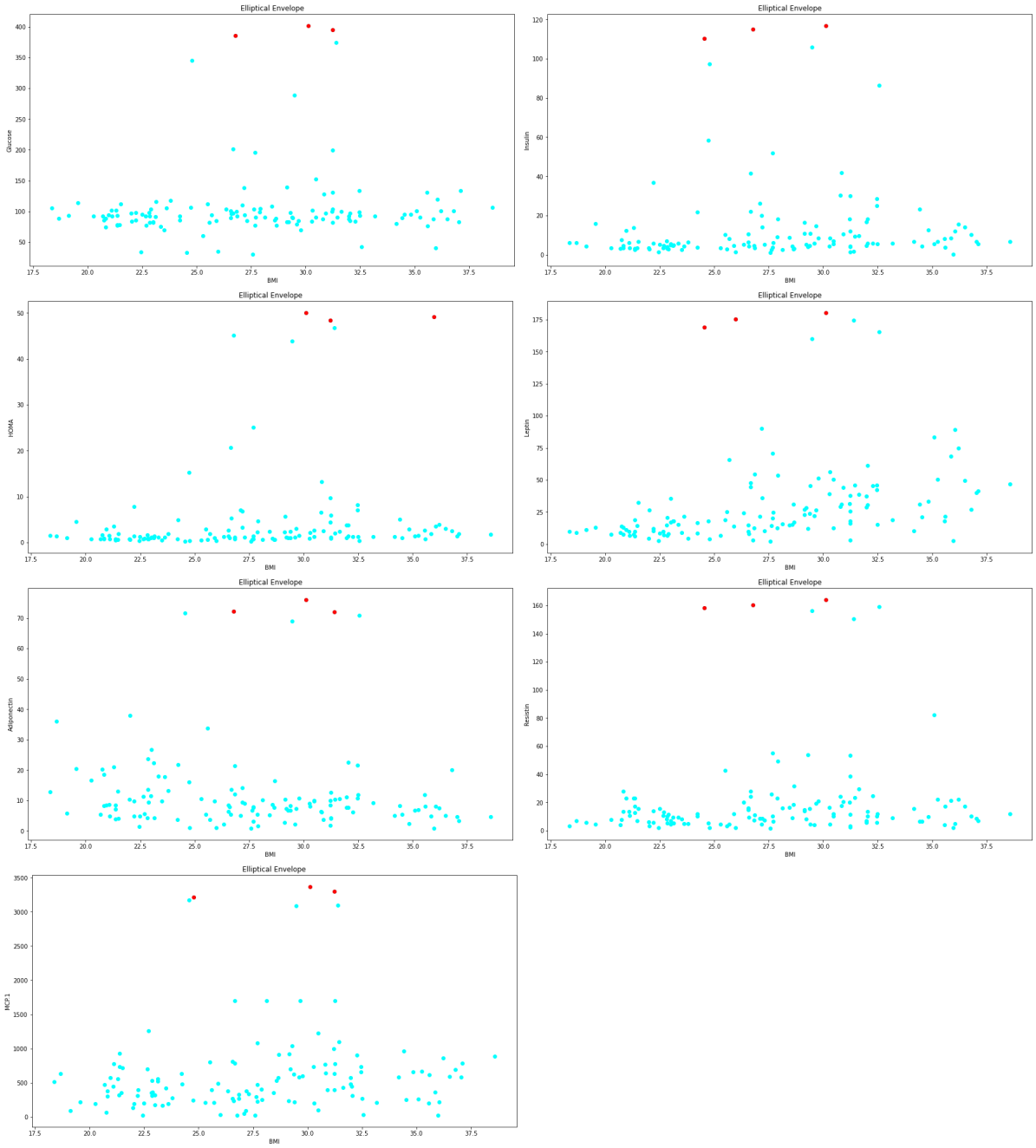
Πίνακας 4.6: Ποσοστά του Breast Cancer Coimbra για 10% με Support Vector Machine

Coimbra data set with SVM	Original data	With 10% outliers	Without outliers
DBscan	69.58%	71.15%	75.22%
Elliptical Envelope	71.67%	75%	70.42%
GMM	70.83%	70%	70.42%
Isolation Forest	72.50%	77.69%	68.33%
LOF	76.25%	69.23%	72.92%
Mahalanobis distance	75.42%	71.92%	70.42%
One Class SVM	72.92%	76.92%	56.87%

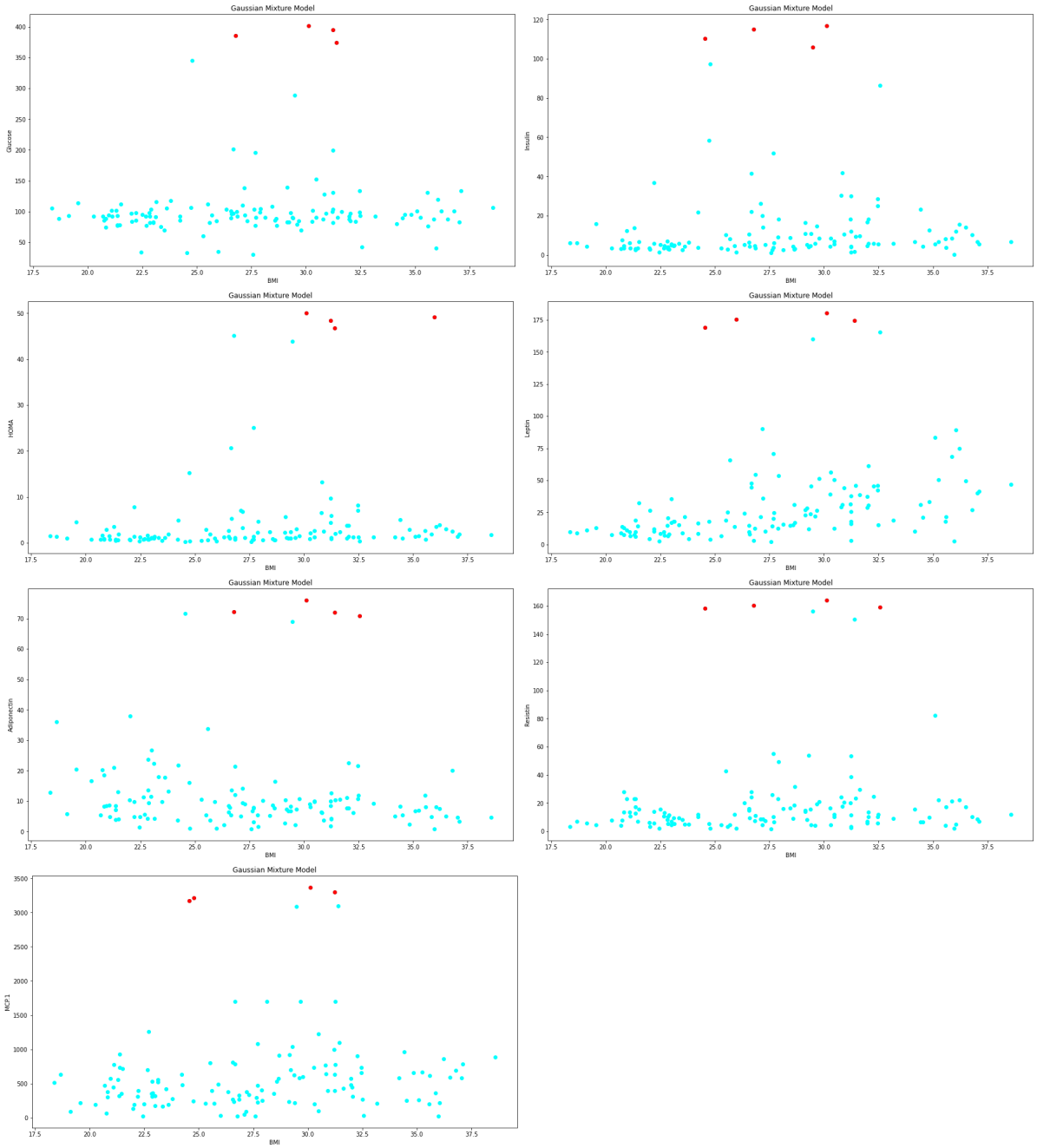
Σχήμα 4.1: Γραφήματα ακραίων τιμών DBscan στο Breast Cancer Coimbra



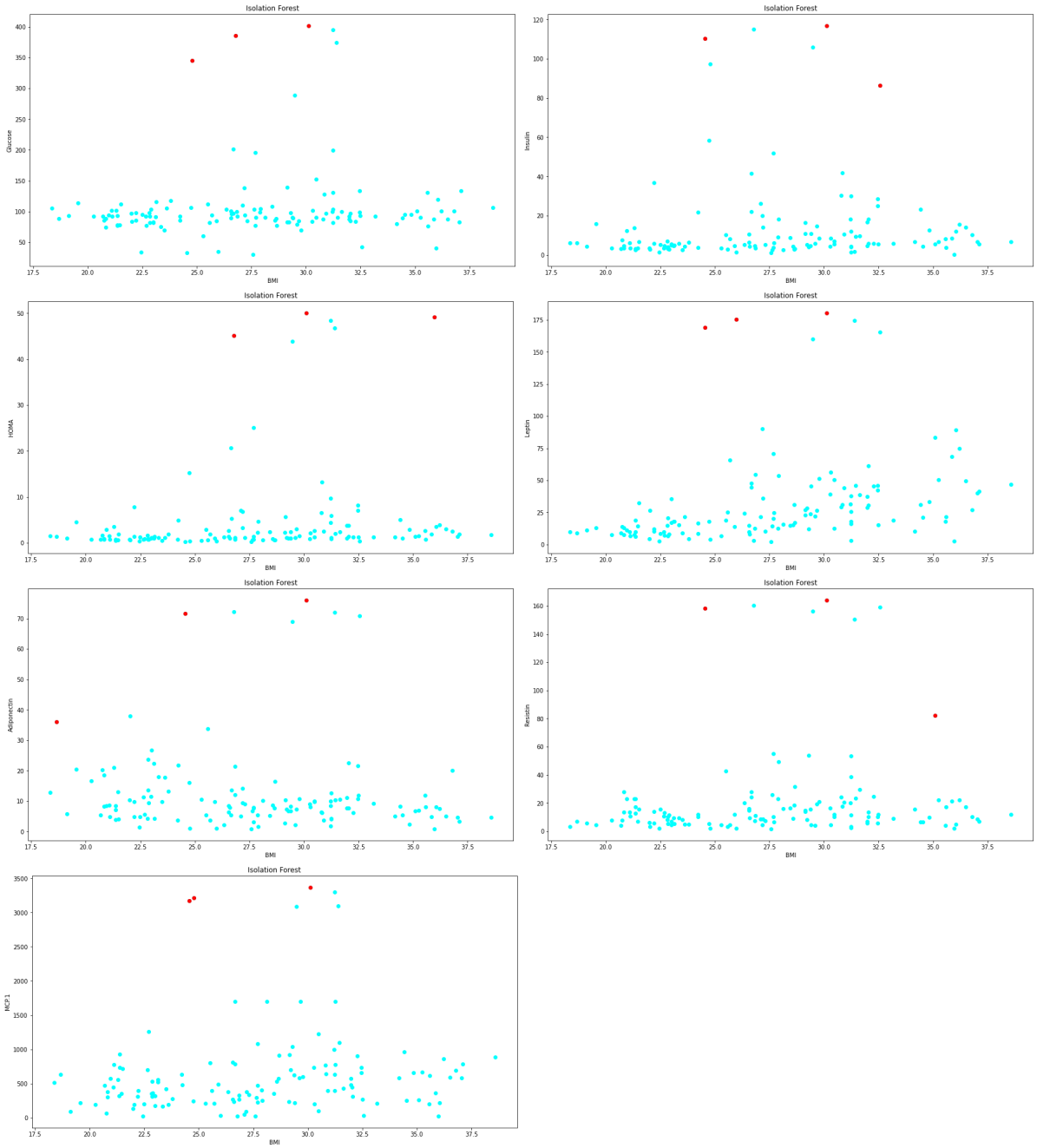
Σχήμα 4.2: Γραφήματα ακραίων τιμών Elliptical Envelope στο Breast Cancer Coimbra



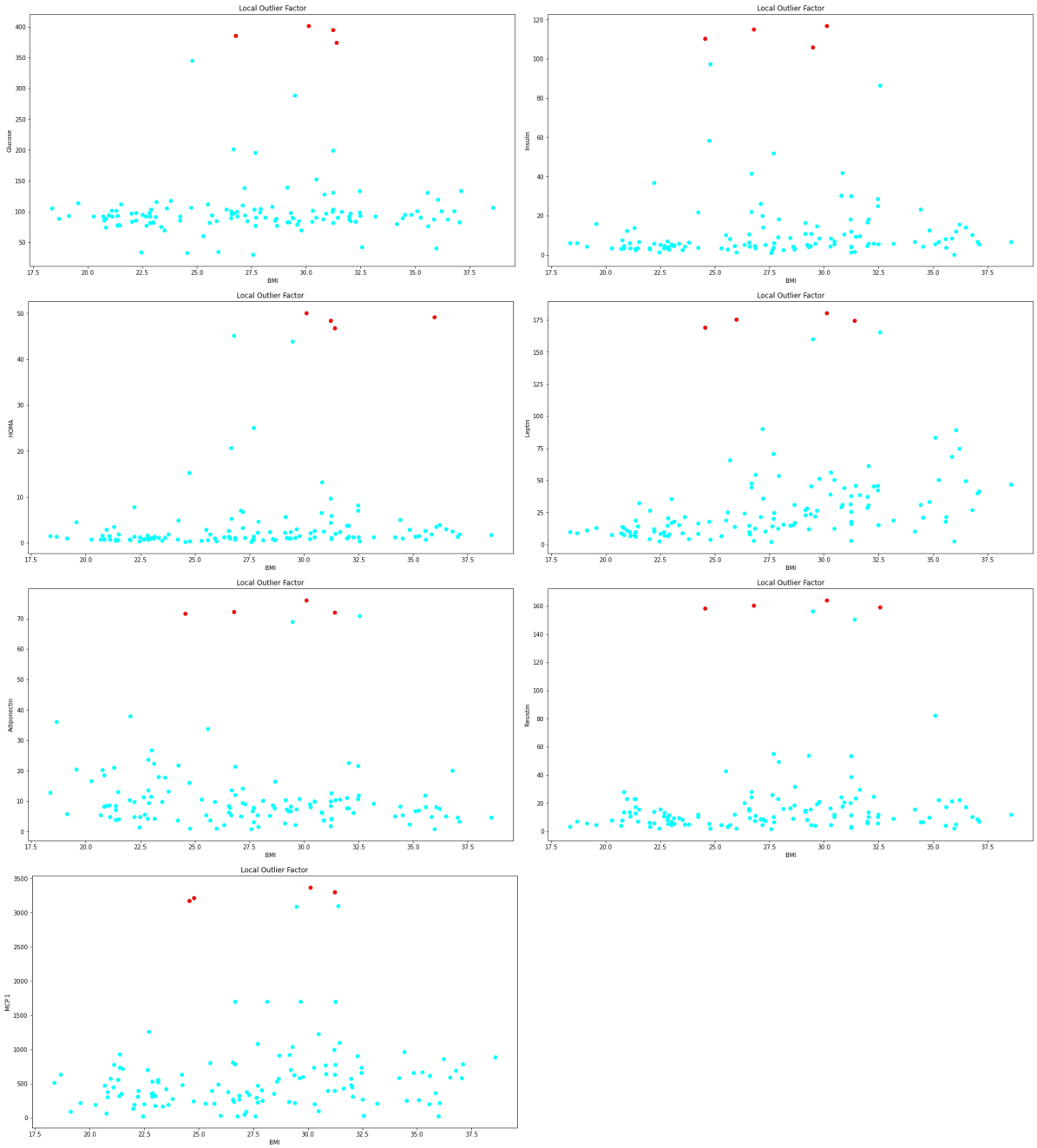
Σχήμα 4.3: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Breast Cancer Coimbra



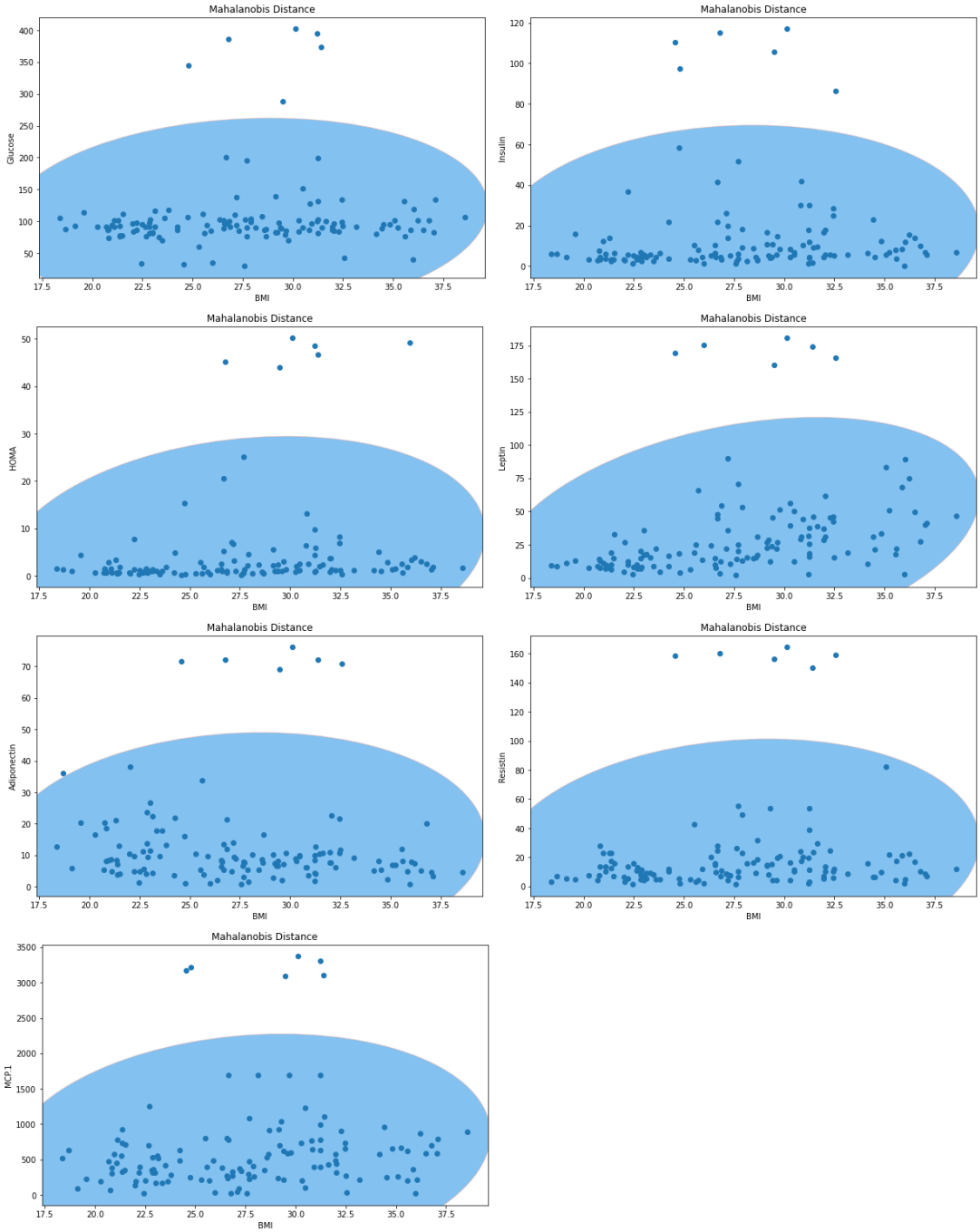
Σχήμα 4.4: Γραφήματα ακραίων τιμών Isolation Forest στο Breast Cancer Coimbra



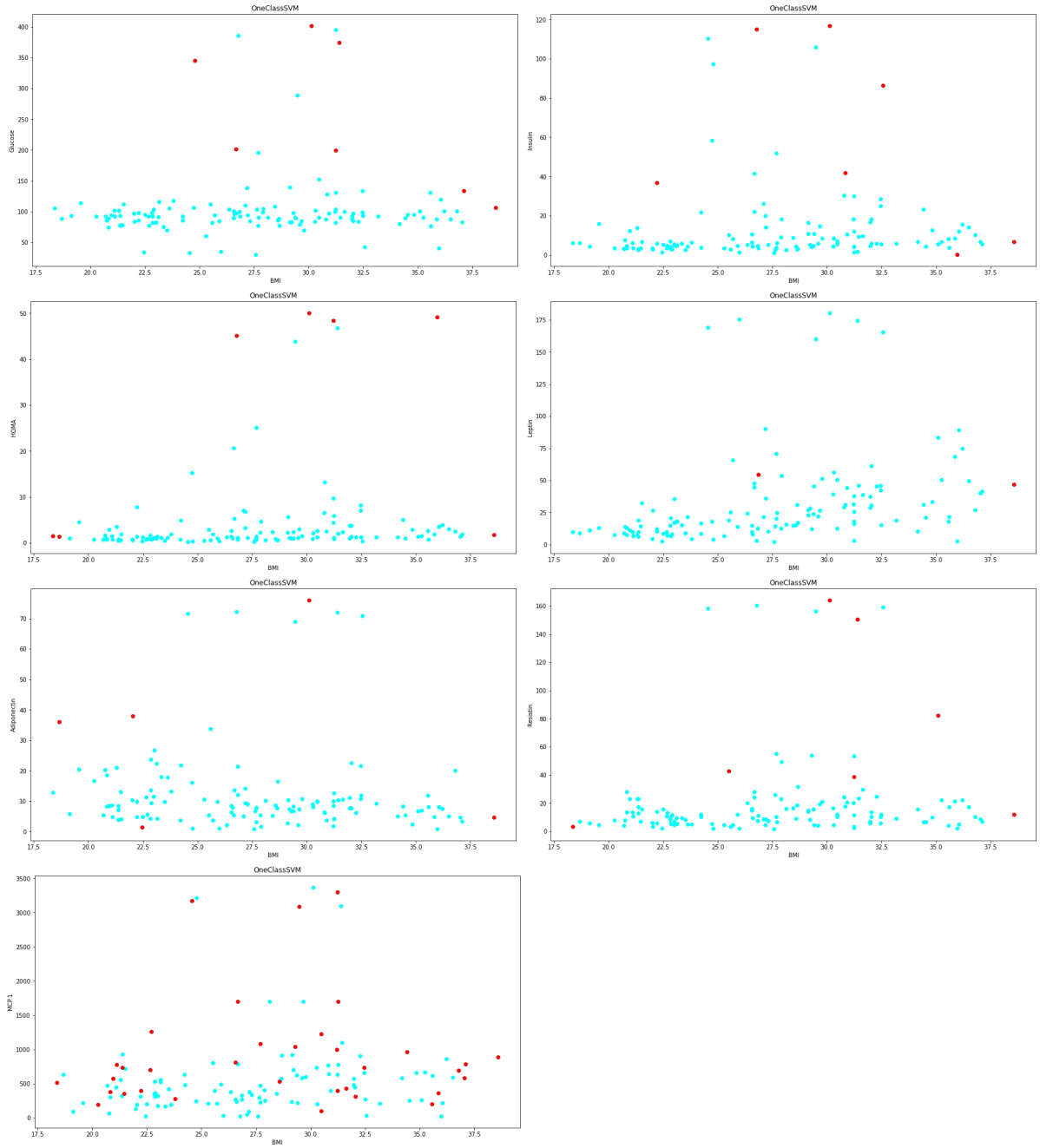
Σχήμα 4.5: Γραφήματα ακραίων τιμών Local Outlier Factor στο Breast Cancer Coimbra



Σχήμα 4.6: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Breast Cancer Coimbra



Σχήμα 4.7: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Breast Cancer Coimbra



4.6.2 Breast Cancer Wisconsin

Το σετ δεδομένων Breast cancer Wisconsin αποτελείται από τις εξής δέκα στήλες: ID, Clump thichness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatic, normal nucleoni, mitoses και τελευταία στήλη class που προσδιορίζει την κατάσταση στην οποία είναι ο ασθενής. Στα Σχήματα 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 και 4.14 παρουσιάζονται οι ακραίες τιμές (10%) που ανιχνεύονται από τους αλγορίθμους. Ως σταθερή μεταβλητή έχει οριστεί το ID. Στους Πίνακες 4.9 και 4.10 παρουσιάζονται αντίστοιχα τα αποτελέσματα για τους αλγορίθμους μηχανικής μάθησης Random Forest και Support Vector Machine με 10% ακραίες τιμές. Παρατηρούμε ότι τα ποσοστά αυτά ταιριάζουν με τα επιθυμητά σχεδόν σε όλες τις περιπτώσεις. Για άλλη μια φορά ο Random Forest έχει μεγαλύτερα ποσοστά επιτυχίας της πρόβλεψης από αυτά του SVM και επίσης ο αλγόριθμος OneClassSVM μας δίνει αποτελέσματα που απέχουν από τα επιθυμητά. Στους Πίνακες 4.7 και 4.8 παρατηρούμε ότι τα ποσοστά είναι επίσης τα επιθυμητά με πολύ μικρές μεταξύ τους διακυμάνσεις. Επιπλέον οι διαφορές μεταξύ των ποσοστών 5% και 10% είναι ότι με την ύπαρξη ακραίων τιμών στο σετ δεδομένων η πρώτη περίπτωση παρουσιάζει μεγαλύτερα ποσοστά επιτυχίας της πρόβλεψης. Ενώ στην περίπτωση που αφαιρεθούν οι ακραίες τιμές τα ποσοστά με 10% είναι μεγαλύτερα από αυτά των 5%.

Πίνακας 4.7: Ποσοστά του Breast Cancer Wisconsin για 5% με Random Forest

Wisconsin data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	97.27%	96.70%	96.79%
Elliptical Envelope	96.73%	94.70%	95.39%
GMM	96.73%	95.57%	96.40%
Isolation Forest	96.27%	94.87%	96.23%
LOF	96.18%	95.74%	95.13%
Mahalanobis distance	96.73%	95.22%	96.02%
One Class SVM	97%	94.26%	92.50%

Πίνακας 4.8: Ποσοστά του Breast Cancer Wisconsin για 5% με Random Forest

Wisconsin data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	96.45%	95.57%	96.42%
Elliptical Envelope	96.27%	94.61%	95%
GMM	96.45%	94.61%	96.50%
Isolation Forest	96.18%	93.83%	95.94%
LOF	96.09%	95.13%	96.51%
Mahalanobis distance	96.09%	95.48%	95.51%
One Class SVM	96.64%	92.96%	91.36%

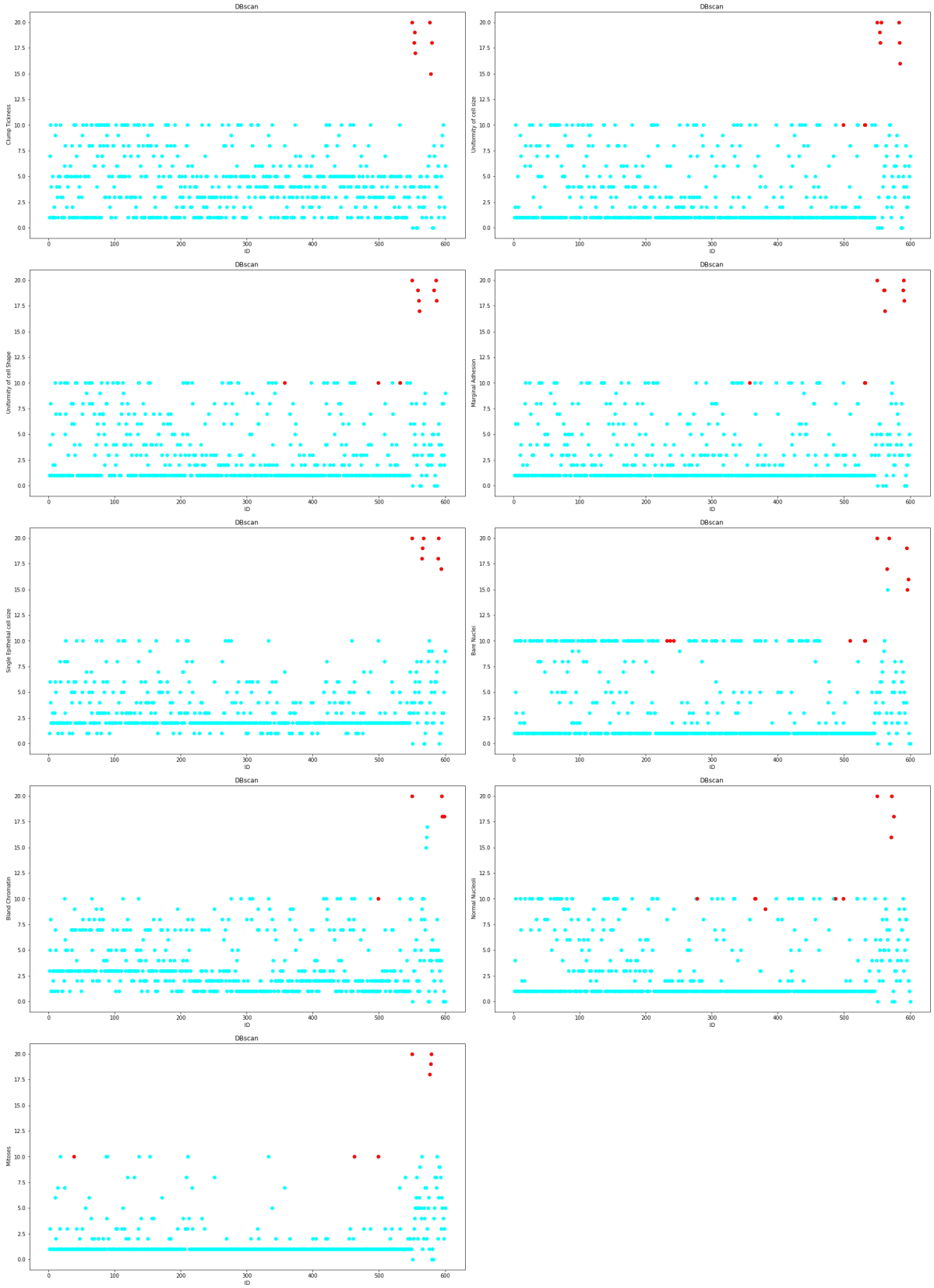
Πίνακας 4.9: Ποσοστά του Breast Cancer Wisconsin για 10% με Random Forest

Wisconsin data set with RF	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	96.64%	89.50%	94.72%
Elliptical Envelope	97%	89.67%	94.86%
GMM	97.36%	90.83%	94.13%
Isolation Forest	97%	90.67%	95.50%
LOF	96.73%	90.83%	94.32%
Mahalanobis distance	97.64%	91%	94.80%
One Class SVM	98.09%	91.50%	81.40%

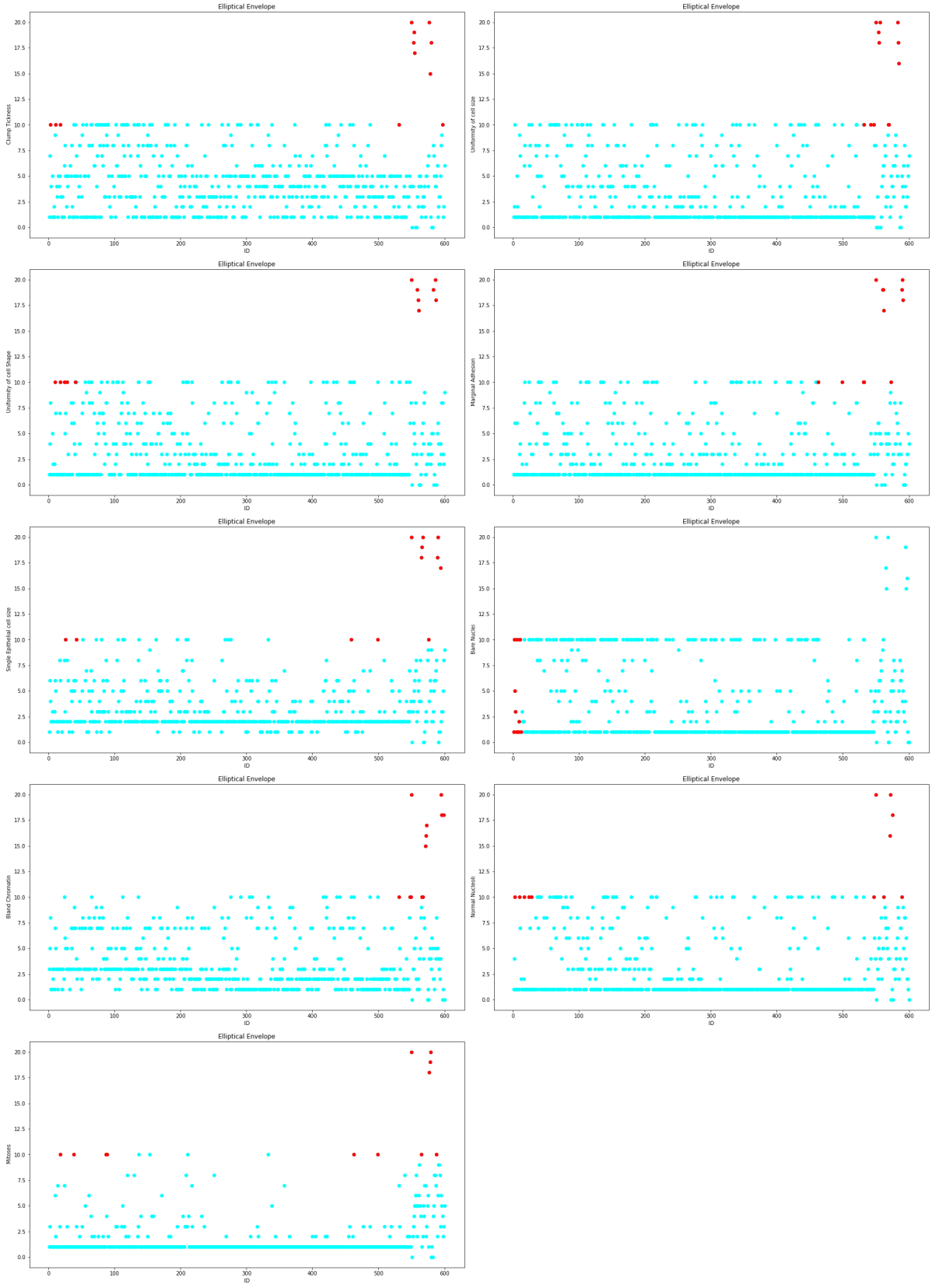
Πίνακας 4.10: Ποσοστά του Breast Cancer Wisconsin για 10% με Support Vector Machine

Wisconsin data set with SVM	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	96.18%	87.92%	94.62%
Elliptical Envelope	96.45%	88.67%	93.27%
GMM	97%	89.58%	94.33%
Isolation Forest	96.64%	88.83%	95.41%
LOF	96.82%	88.92%	93.60%
Mahalanobis distance	97.18%	89.33%	94.61%
One Class SVM	97.45%	89.58%	80.40%

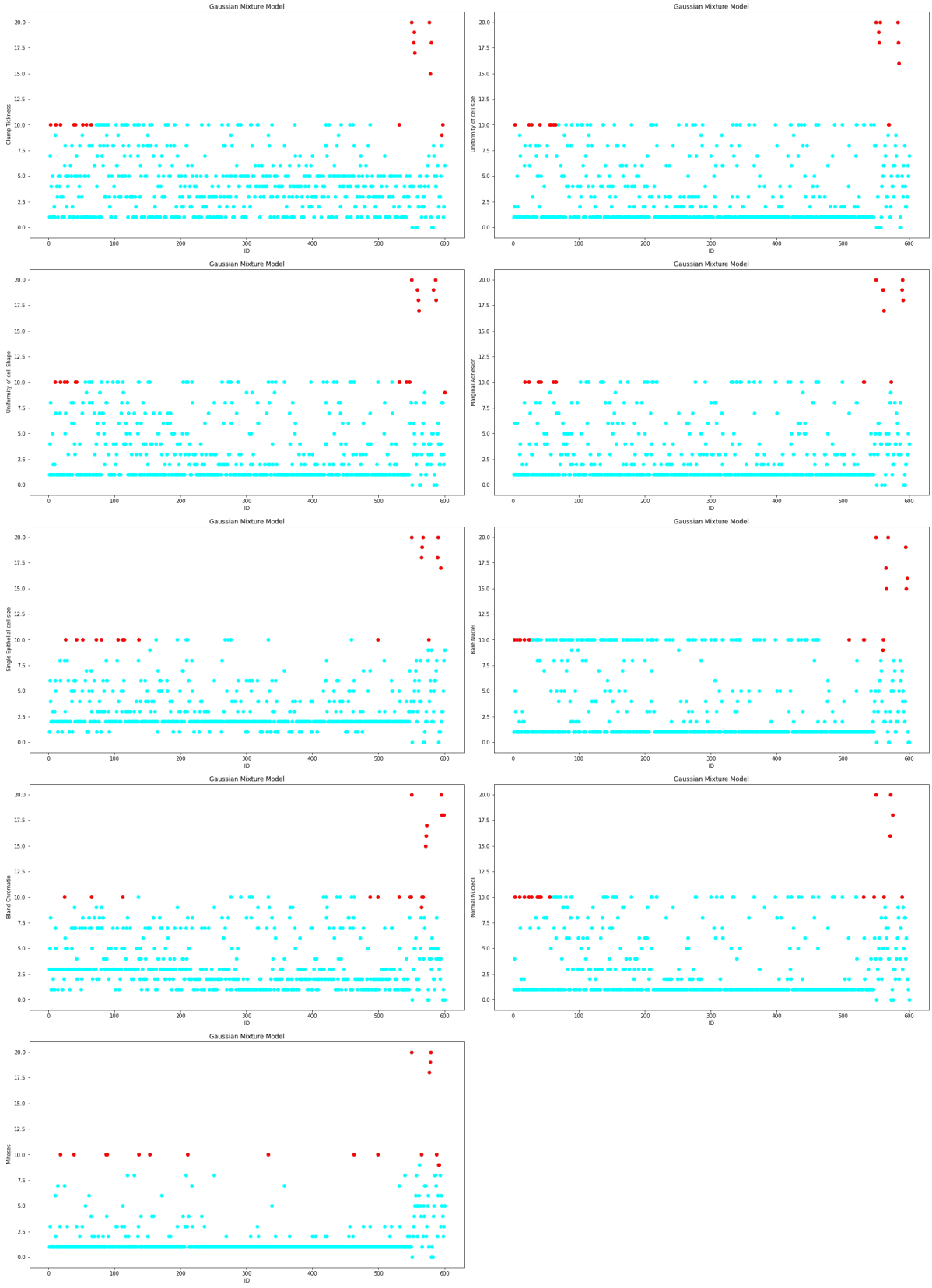
Σχήμα 4.8: Γραφήματα ακραίων τιμών DBscan στο Breast Cancer Wisconsin



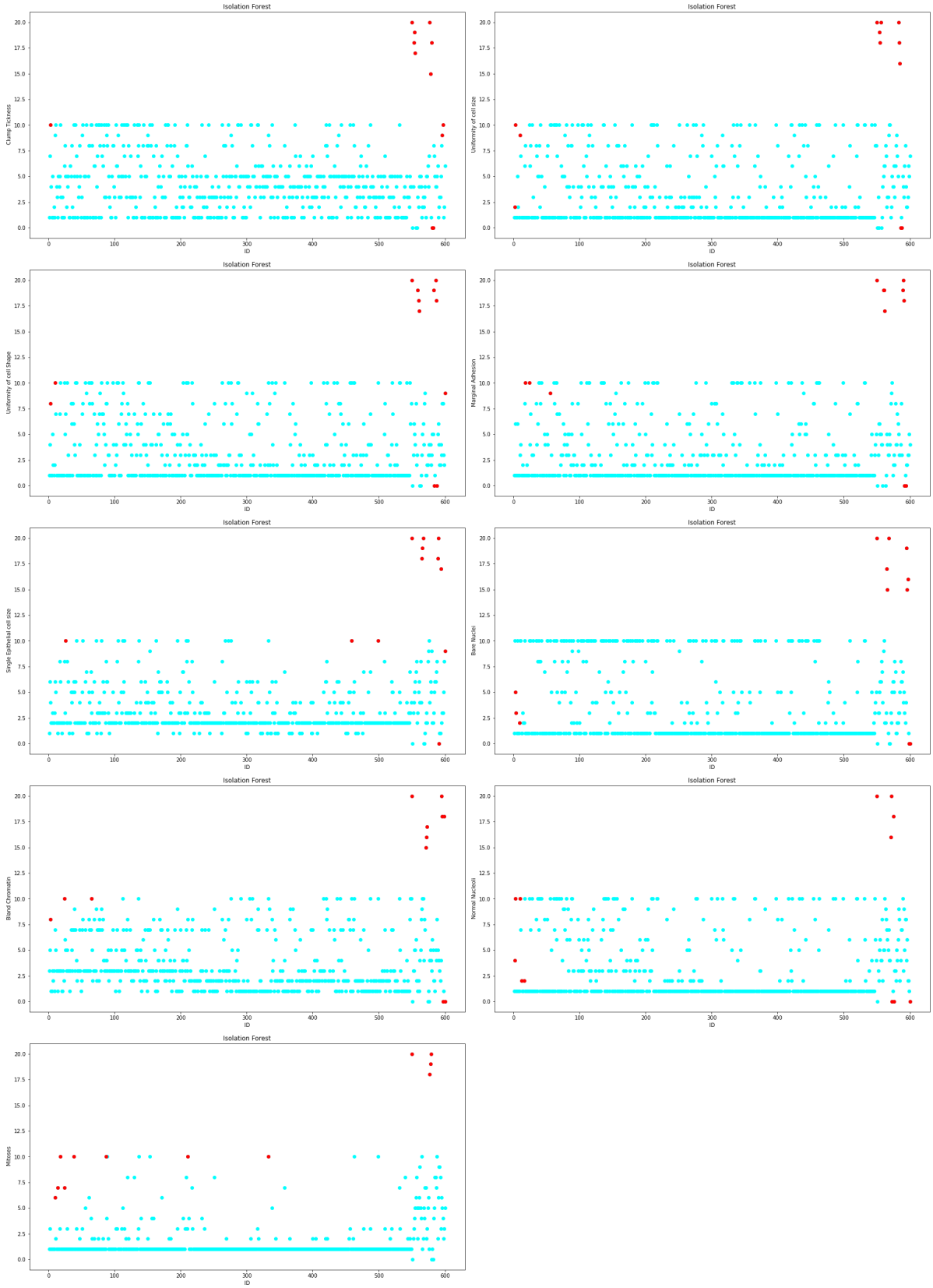
Σχήμα 4.9: Γραφήματα ακραίων τιμών Elliptical Envelope στο Breast Cancer Wisconsin



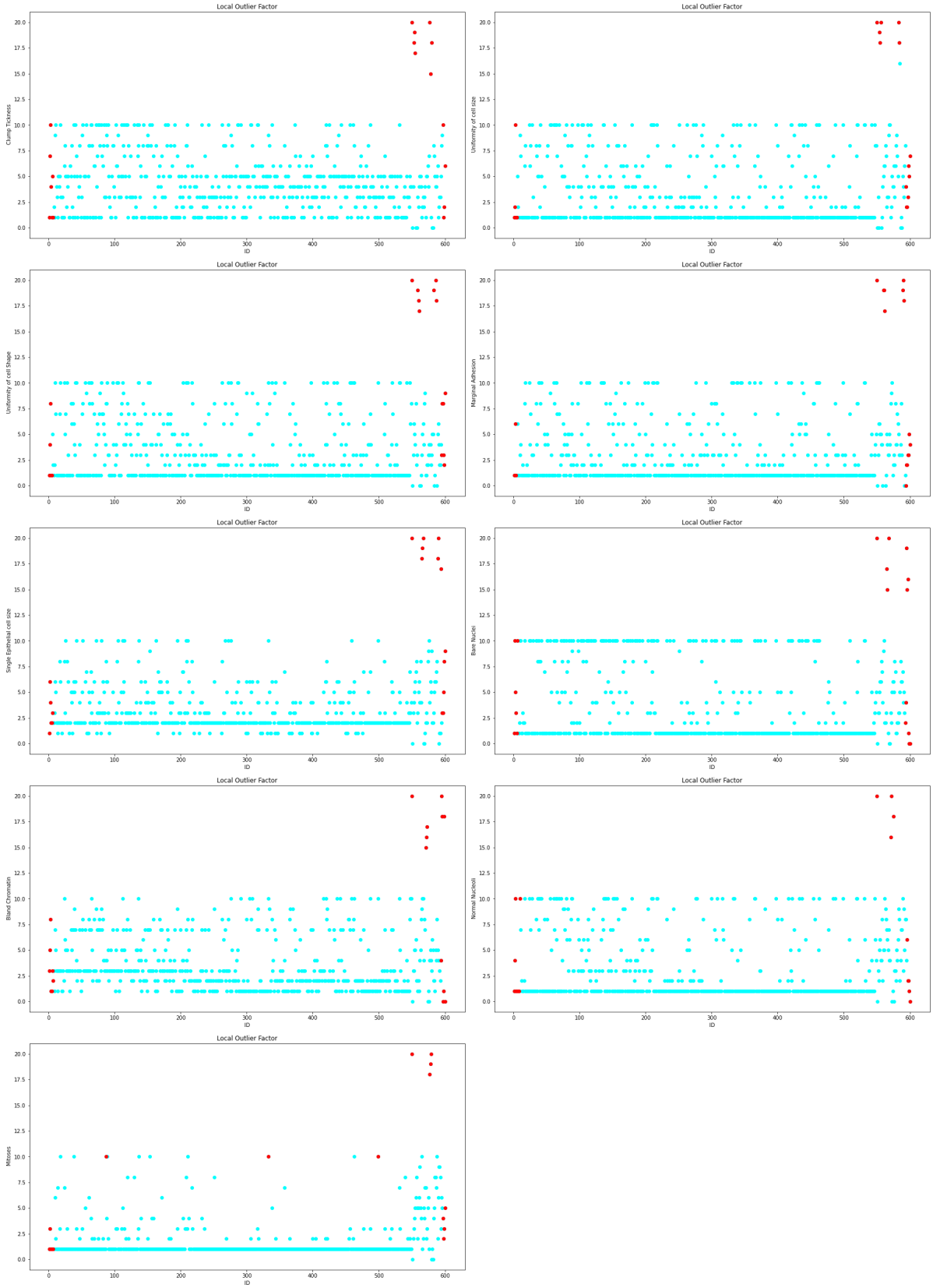
Σχήμα 4.10: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Breast Cancer Wisconsin



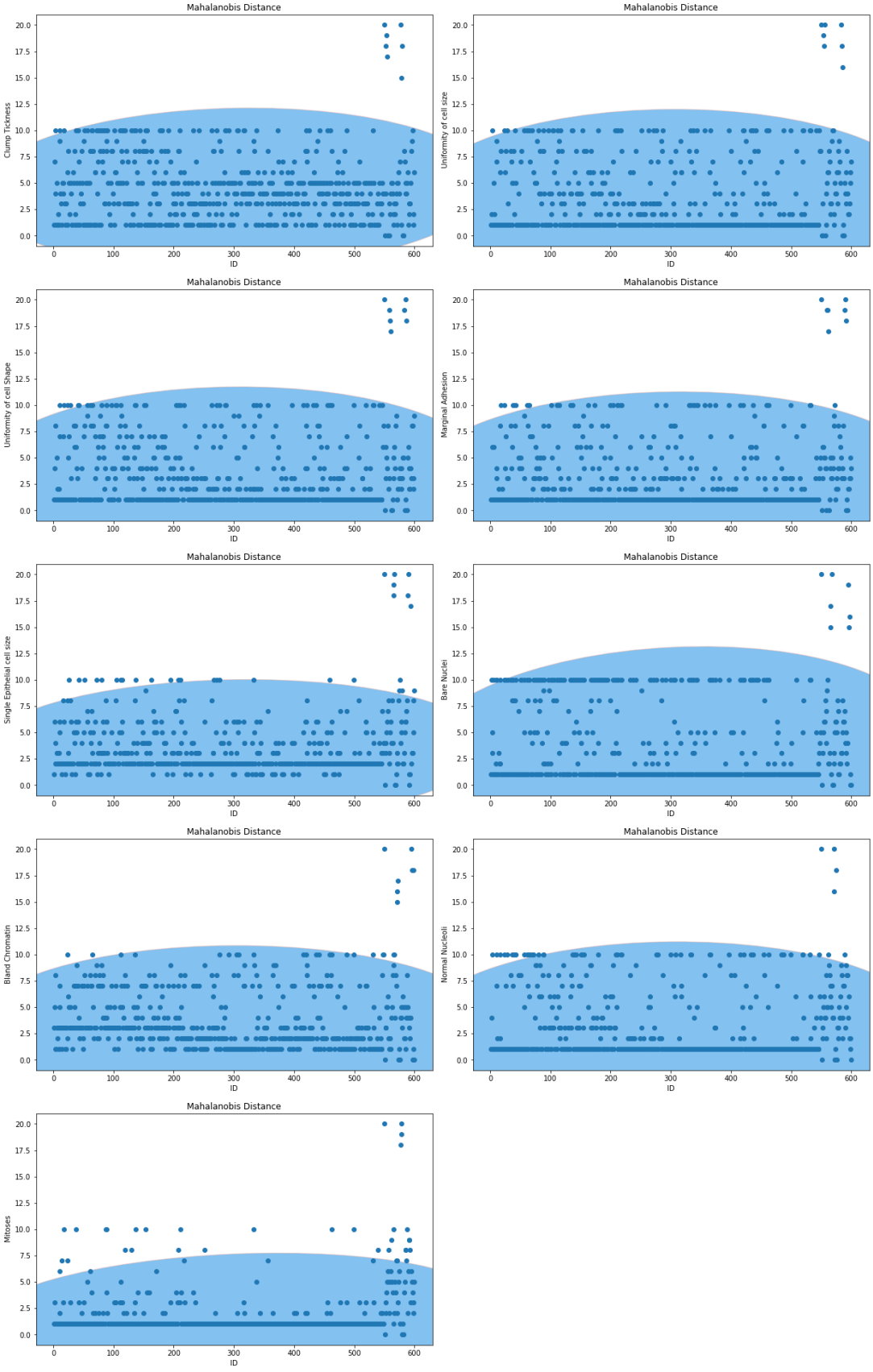
Σχήμα 4.11: Γραφήματα ακραίων τιμών Isolation Forest στο Breast Cancer Wisconsin



Σχήμα 4.12: Γραφήματα ακραίων τιμών Local Outlier Factor στο Breast Cancer Wisconsin



Σχήμα 4.13: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Breast Cancer Wisconsin



Σχήμα 4.14: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Breast Cancer Wisconsin

4.6.3 Glass Identification

Το σετ δεδομένων glass identification αποτελείται από έντεκα στήλες τις: ID, refractive index, Sodium, Magnesium, Aluminum, Silicon, Potassium, Calcium, Barium, Iron και Class όπου καθορίζει το είδος του γυαλιού. Στα Σχήματα 4.15, 4.16, 4.17, 4.18, 4.19, 4.20 και 4.21 παρουσιάζονται οι ακραίες τιμές (10%) που ανιχνεύονται από τον κάθε αλγόριθμο. Επιπλέον στους Πίνακες 4.13 και 4.14 παραθέτονται τα ποσοστά των προβλέψεων των αλγορίθμων μηχανικής μάθησης Random Forest και SVM αντίστοιχα με 10% ακραίες τιμές. Τα αποτελέσματα έχουν μία μικρή απόκλιση από τα αναμενόμενα και αυτό οφείλεται στην ιδιορρυθμία του σετ δεδομένων. Παρ' όλ' αυτά στις περισσότερες περιπτώσεις βλέπουμε ότι με την αφαίρεση των ακραίων τιμών τα ποσοστά πρόβλεψης βελτιώνονται. Επιπλέον τα ποσοστά του Random Forest είναι υψηλότερα από αυτά του SVM. Στην περίπτωση που έχουμε ακραίες τιμές της τάξεως του 5%, όπως φαίνεται και στους Πίνακες 4.11 και 4.12 τα αποτελέσματα με μικρές αποκλίσεις τείνουν στα επιθυμητά. Για άλλη μια φορά τα ποσοστά με τις ακραίες τιμές στα σετ δεδομένων τους, είναι μικρότερα στην περίπτωση με 10% σε σύγκριση με το 5%, ενώ όταν από τα σετ δεδομένων απαλοίφονται οι ακραίες τιμές η περίπτωση του 10% παρουσιάζει μεγαλύτερα ποσοστά από αυτά των 5%.

Πίνακας 4.11: Ποσοστά του Glass Identification για 5% με Random Forest

Glass data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	79.07%	76.22%	78.50%
Elliptical Envelope	79.77%	72.22%	77.80%
GMM	76.74%	72.67%	75.84%
Isolation Forest	75.35%	71.56%	78.54%
LOF	78.37%	74.22%	76.98%
Mahalanobis distance	77.67%	80.22%	78.81%
One Class SVM	76.05%	78%	81.11%

Πίνακας 4.12: Ποσοστά του Glass Identification για 5% με Support Vector Machine

Glass data set with SVM	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	61.63%	62.89%	59%
Elliptical Envelope	62.56%	56%	58.29%
GMM	63.49%	60.44%	63.90%
Isolation Forest	61.63%	59.78%	59.51%
LOF	63.49%	57.11%	59.53%
Mahalanobis distance	63.02%	62.67%	60.71%
One Class SVM	62.09%	65.78%	74.07%

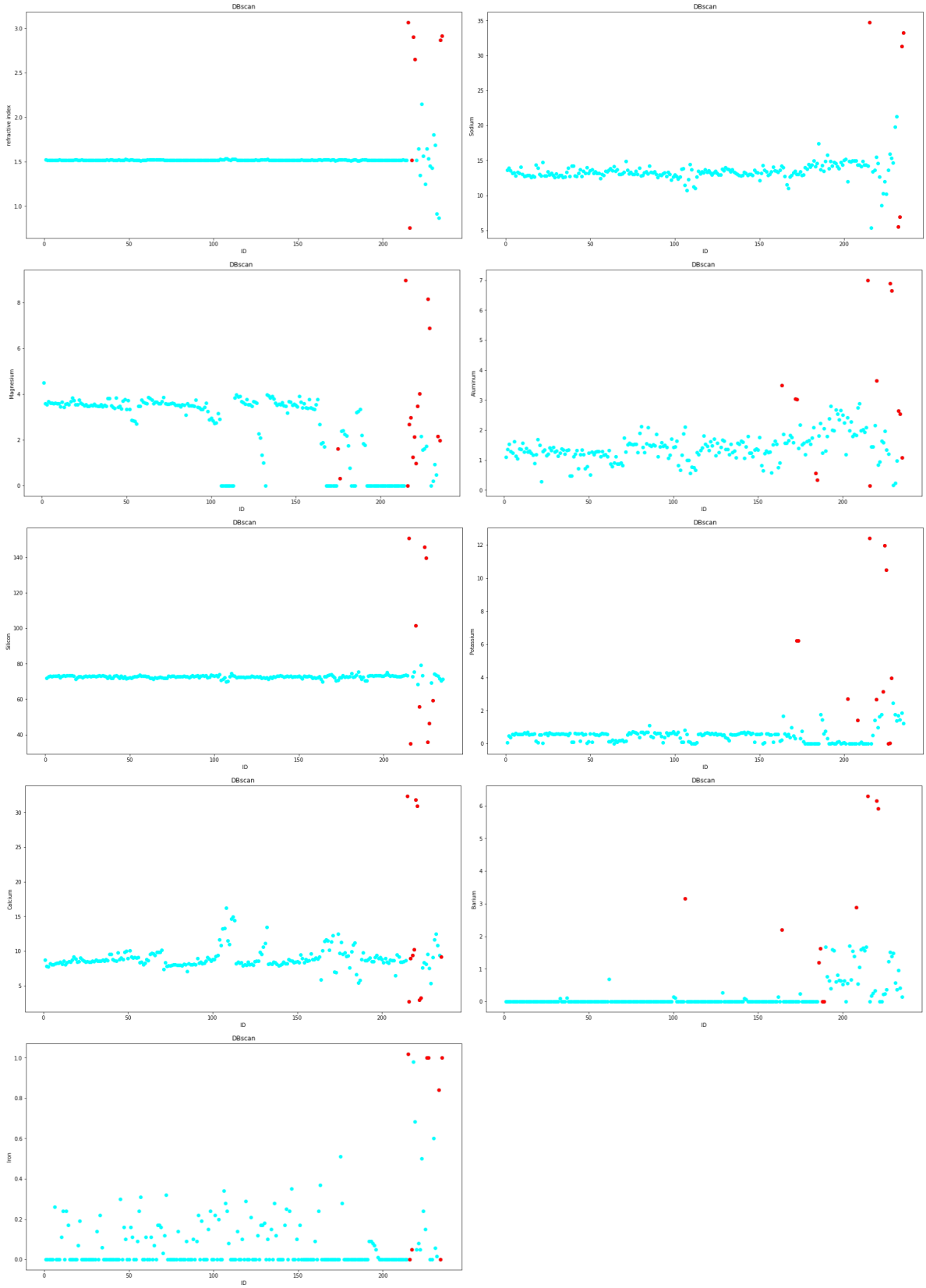
Πίνακας 4.13: Ποσοστά του Glass Identification για 10% με Random Forest

Glass data set with RF	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	77.67%	71.91%	78.38%
Elliptical Envelope	77.44%	77.47%	76.98%
GMM	76.74%	73.19%	78.81%
Isolation Forest	78.60%	73.62%	78.84%
LOF	77.67%	74.89%	78.86%
Mahalanobis distance	74.44%	75.11%	78.33%
One Class SVM	78.14%	71.06%	76.55%

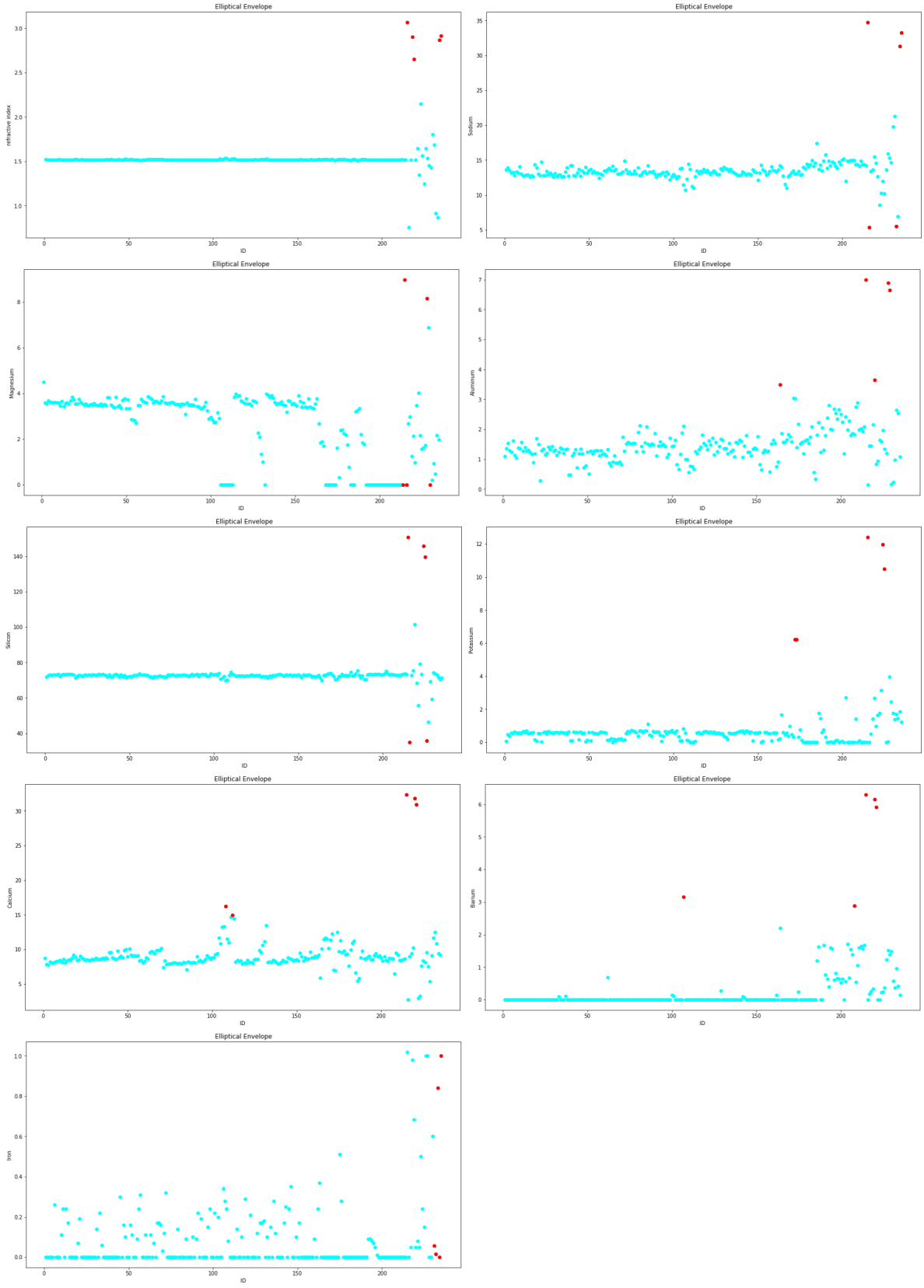
Πίνακας 4.14: Ποσοστά του Glass Identification για 10% με Support Vector Machine

Glass data set with SVM	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	61.63%	59.57%	63.24%
Elliptical Envelope	59.77%	61.70%	59.30%
GMM	55.81%	60.85%	60.95%
Isolation Forest	61.86%	60.64%	62.33%
LOF	63.49%	61.28%	63.18%
Mahalanobis distance	59.53%	60%	64.52%
One Class SVM	62.56%	58.51%	61.38%

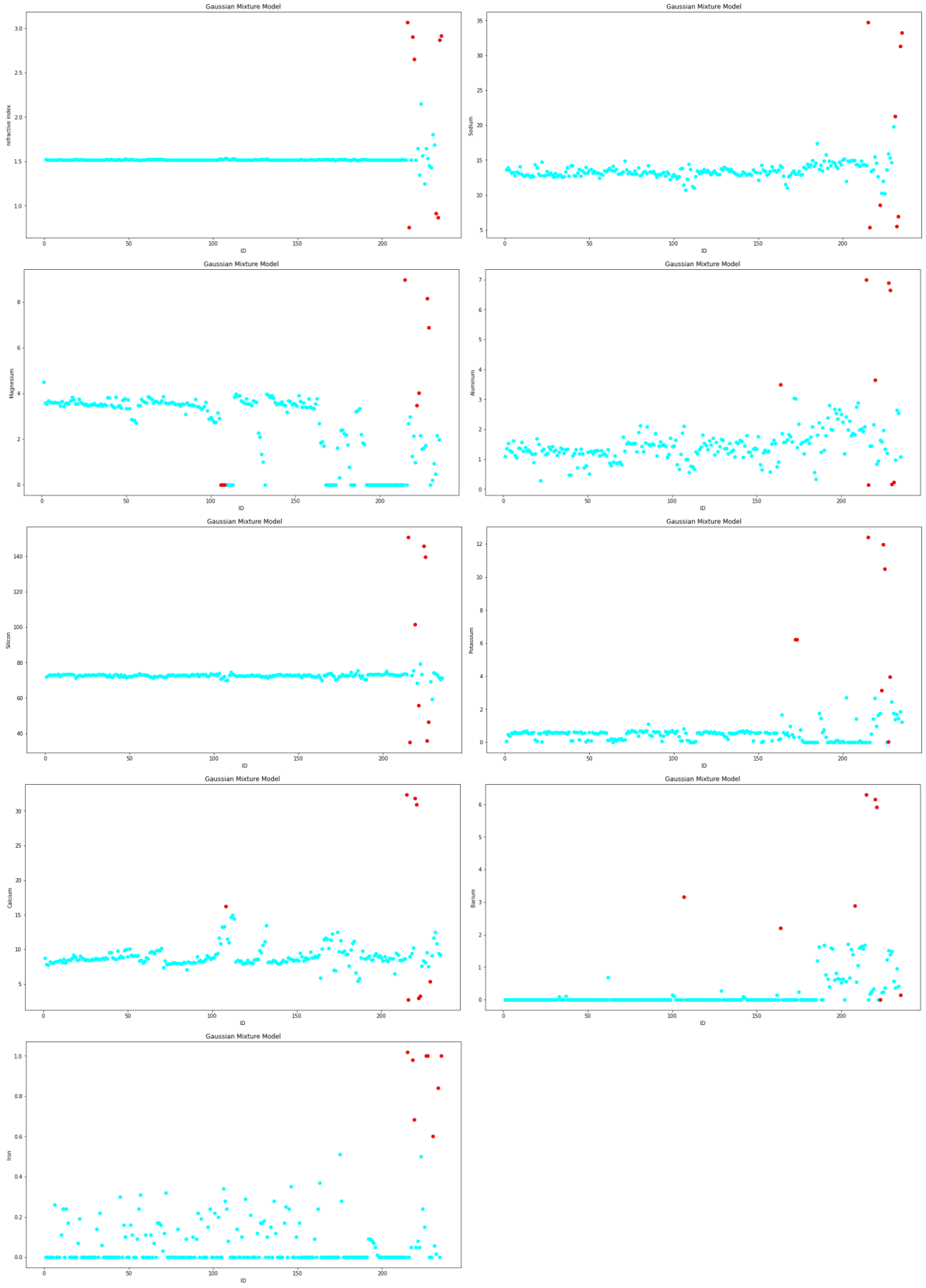
Σχήμα 4.15: Γραφήματα ακραίων τιμών DBscan στο Glass Identification



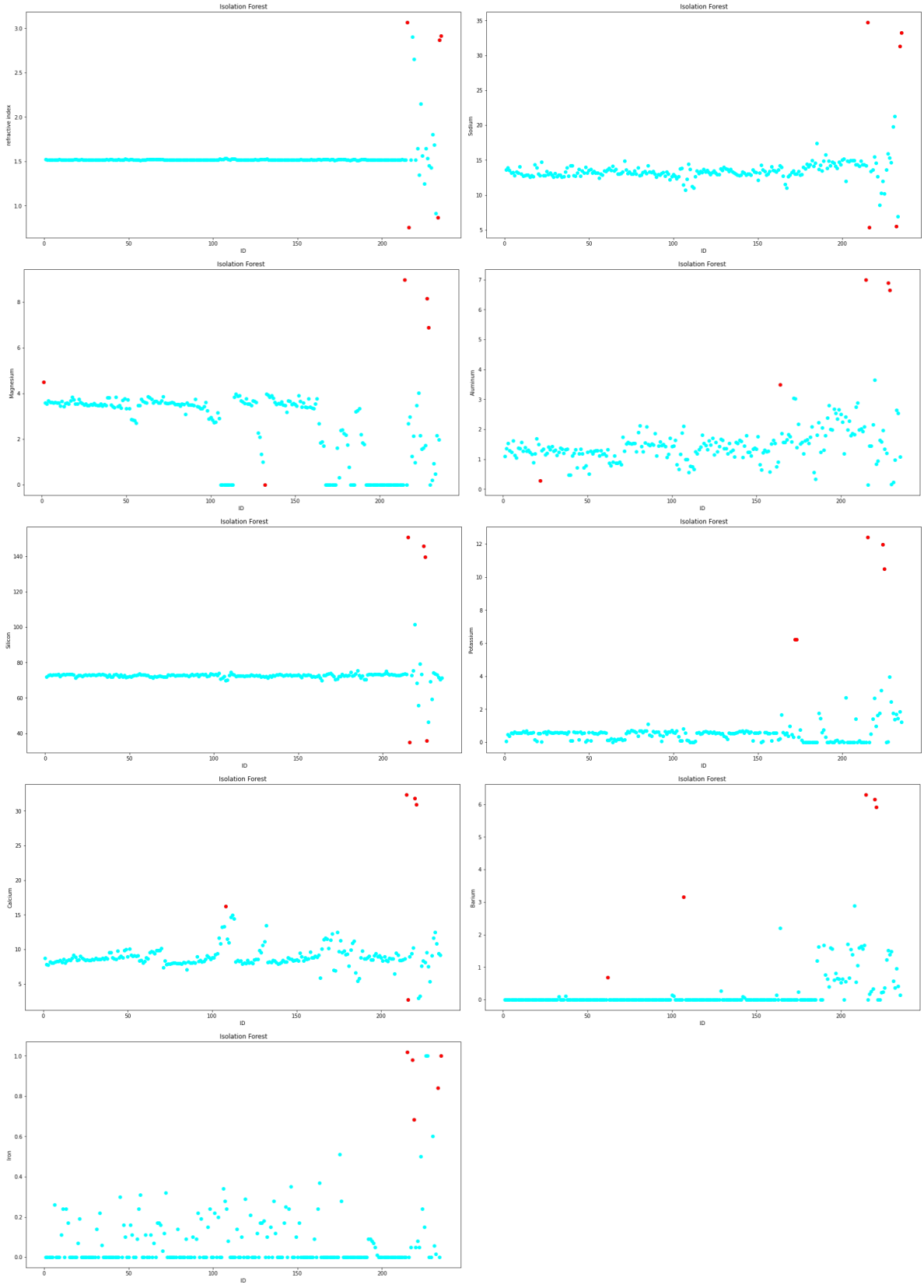
Σχήμα 4.16: Γραφήματα ακραίων τιμών Elliptical Envelope στο Glass Identification



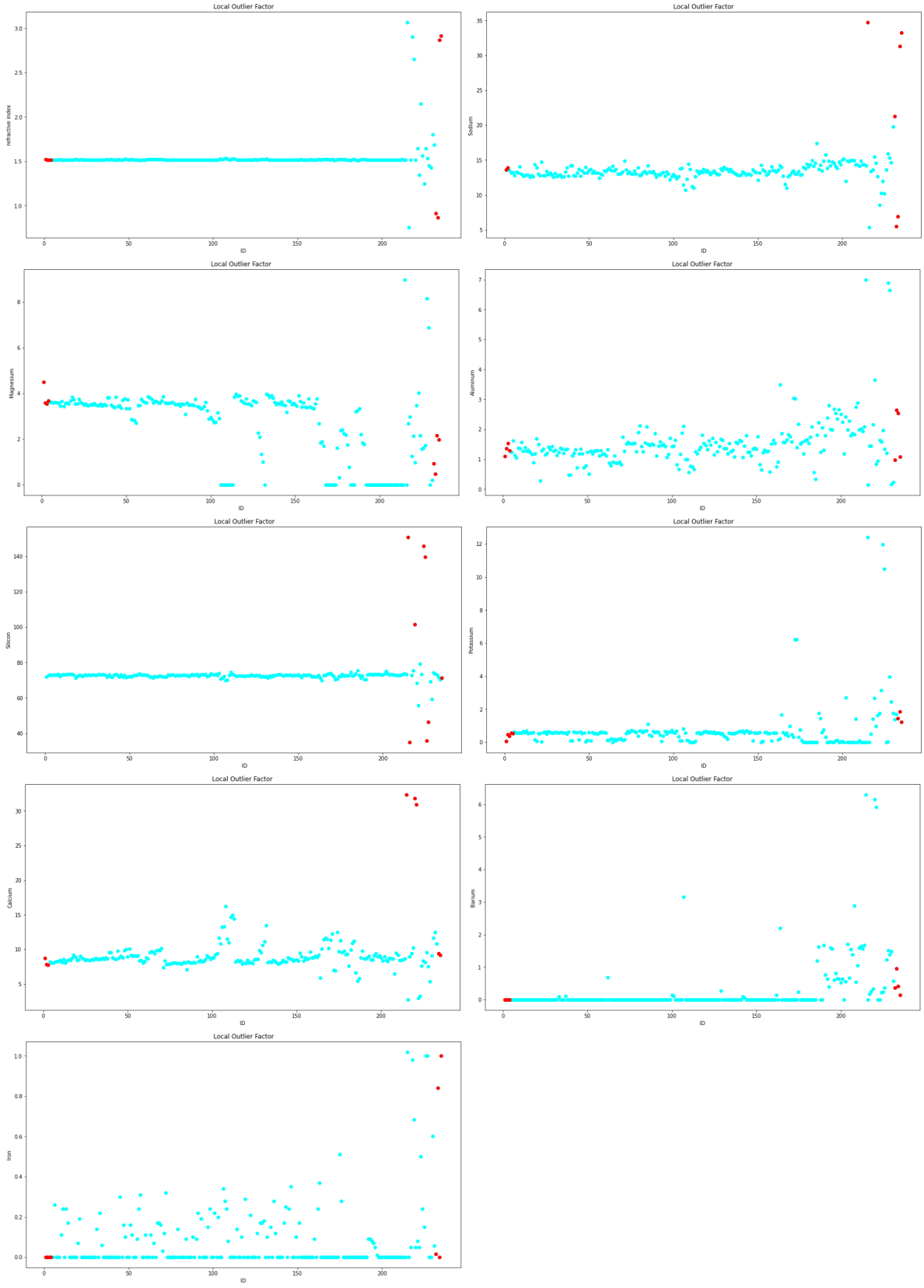
Σχήμα 4.17: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Glass Identification



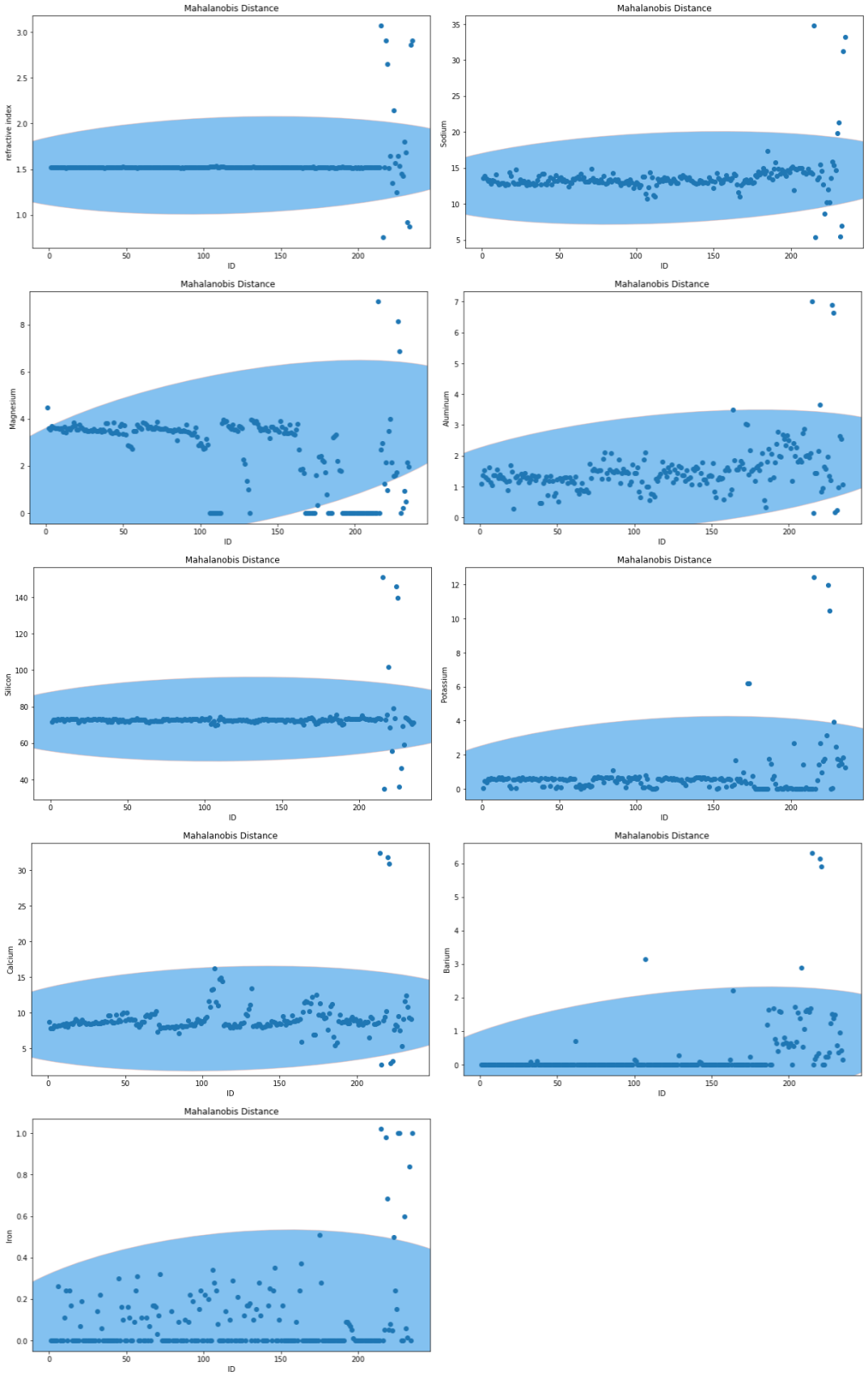
Σχήμα 4.18: Γραφήματα ακραίων τιμών Isolation Forest στο Glass Identification



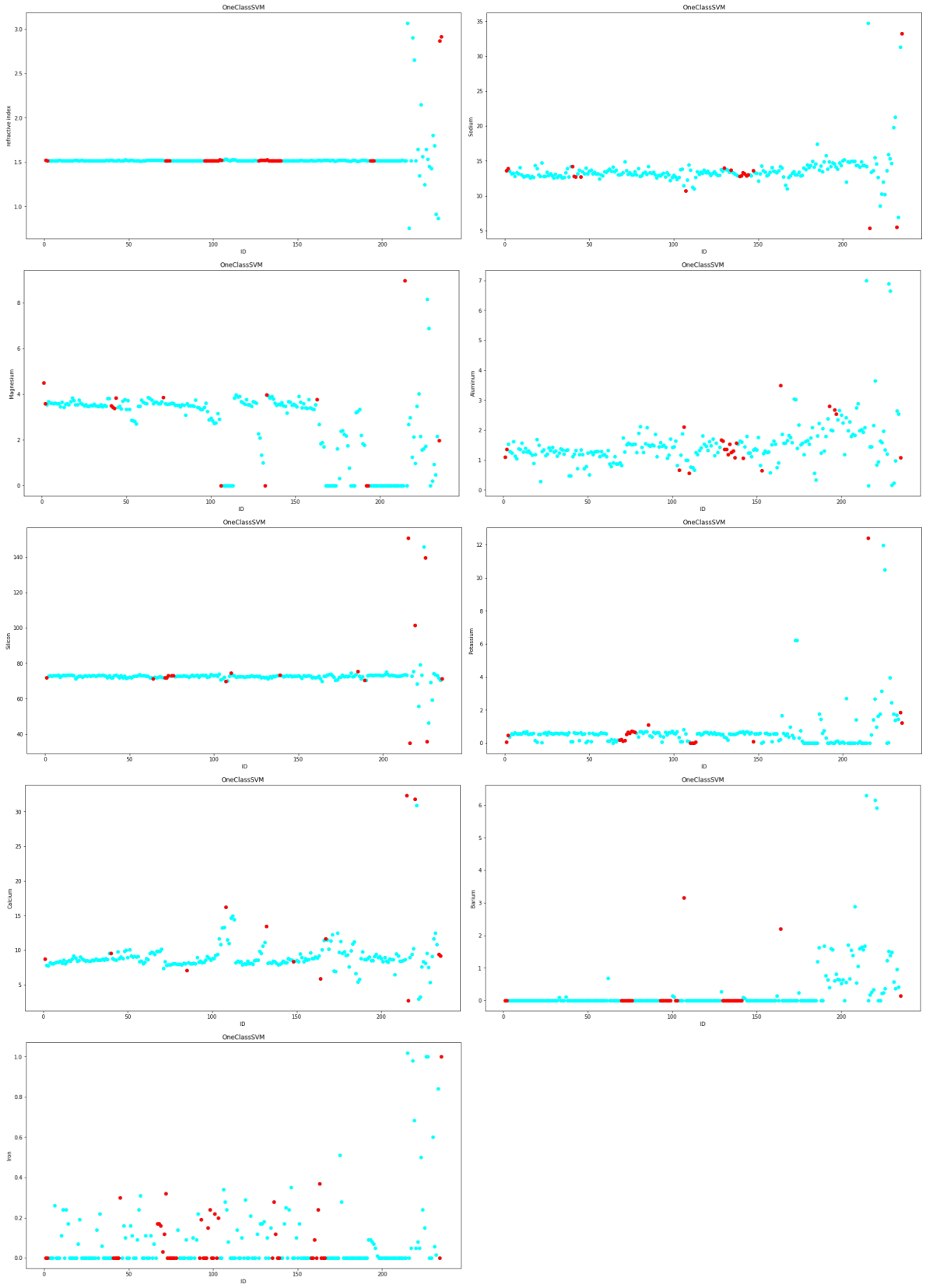
Σχήμα 4.19: Γραφήματα ακραίων τιμών Local Outlier Factor στο Glass Identification



Σχήμα 4.20: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Glass Identification



Σχήμα 4.21: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Glass Identification



4.6.4 HCV

Το σετ δεδομένων HCV αφορά το αν ένα άτομο είναι κατάλληλο για αιμοδοσία. Εμπεριέχει τις εξής στήλες: ID, Age, Sex, ALB, ALP, ALT, AST, BIB, CHE, CHOL, CREA, GGT PROT και Class όπου ανάλογα τα δεδομένα του κάθε ατόμου το καθιστά κατάλληλο ή όχι για αιμοδοσία. Ως ανεξάρτητη μεταβλητή έχει οριστεί το ID. Στα Σχήματα 4.22, 4.23, 4.24, 4.25, 4.26, 4.27 και 4.28 παρουσιάζεται η εύρεση των ακραίων παρατηρήσεων (10%) από το κάθε αλγόριθμο. Επιπλέον στους Πίνακες 4.17 και 4.18 παραθέτονται τα ποσοστά των αλγορίθμων Random Forest και SVM αντίστοιχα με ακραίες τιμές 10%. Παρατηρούμε ότι σε όλες τις περιπτώσεις, εκτός αυτών του Isolation Forest με Random Forest και του Local Outlier Factor με SVM, η ύπαρξη ακραίων τιμών συμβάλλει αρνητικά στην πρόβλεψη του στόχου. Ενώ τα ποσοστά των αλγορίθμων χωρίς τις ακραίες τιμές είναι καλύτερα από αυτά των αλγορίθμων με δεδομένα το αρχικό σετ. Συνεπώς τα αποτελέσματα δεν είναι ιδανικά αλλά επιτυγχάνεται ο στόχος. Επιπλέον στους Πίνακες 4.15 και 4.16 παρουσιάζονται τα αποτελέσματα με ακραίες τιμές της τάξεως του 5%. Είναι φανερό ότι τα ποσοστά 5% και 10% είναι παρόμοια χωρίς να έχουν ιδιαίτερες διαφορές, συνεπώς τείνουν στο επιθυμητό αποτελέσμα. Επιπλέον για τα ποσοστά 5% παρατηρούμε ότι ο αλγόριθμος DBscan τόσο με το Random Forest, όσο και με το SVM παρουσίαζει μεγαλύτερα ποσοστά με την απαλοιφή των ακραίων τιμών απο το σετ δεδομένων από ότι το αρχικό σετ. Παρόμοια συμπεριφορά έχει ο OneClasSVM στην περίπτωση του SVM.

Πίνακας 4.15: Ποσοστά του HCV για 5% με Random Forest

HCV data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	79.44%	79.84%	100%
Elliptical Envelope	92.74%	90.62%	95.14%
GMM	92.98%	91.33%	95.45%
Isolation Forest	93.15%	91.33%	94.26%
LOF	92.58%	91.09%	94.82%
Mahalanobis distance	92.18%	91.87%	96%
One Class SVM	93.39%	91.95%	95%

Πίνακας 4.16: Ποσοστά του HCV για 5% με Support Vector Machine

HCV data set with SVM	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	78.06%	75.39%	100%
Elliptical Envelope	91.77%	88.44%	95.77%
GMM	92.42%	89.14%	95.63%
Isolation Forest	91.94%	89.53%	93.91%
LOF	92.34%	94.30%	94.55%
Mahalanobis distance	92.50%	89.06%	95.91%
One Class SVM	92.90%	89.84%	95.87%

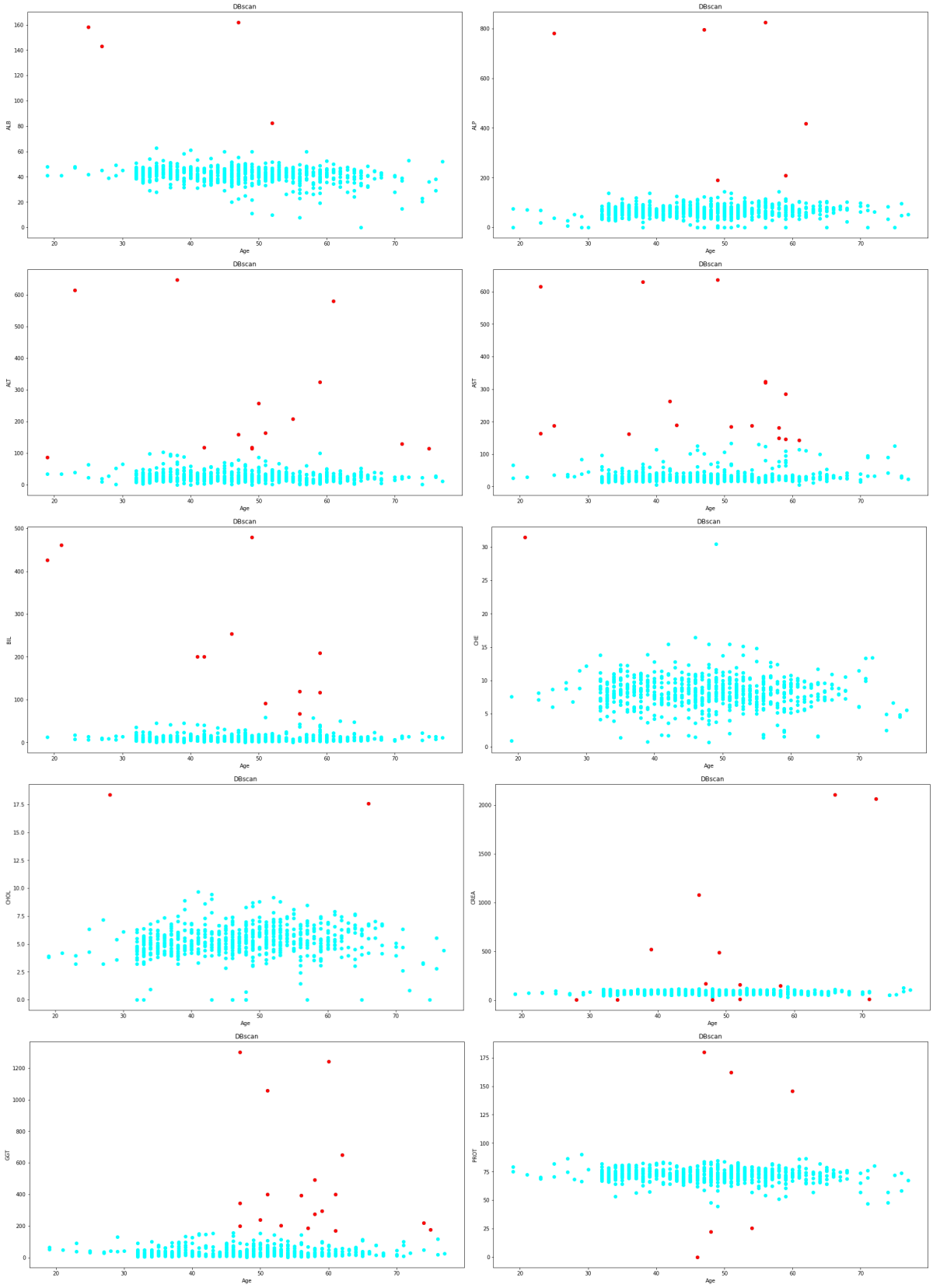
Πίνακας 4.17: Ποσοστά του HCV για 10% με Random Forest

HCV data set with RF	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	81.94%	80.70%	100%
Elliptical Envelope	90.97%	91.48%	95.09%
GMM	92.74%	91.02%	95.98%
Isolation Forest	94.03%	89.92%	95.39%
LOF	92.98%	91.33%	95.36%
Mahalanobis distance	93.63%	91.41%	96%
One Class SVM	92.90%	91.56%	95.67%

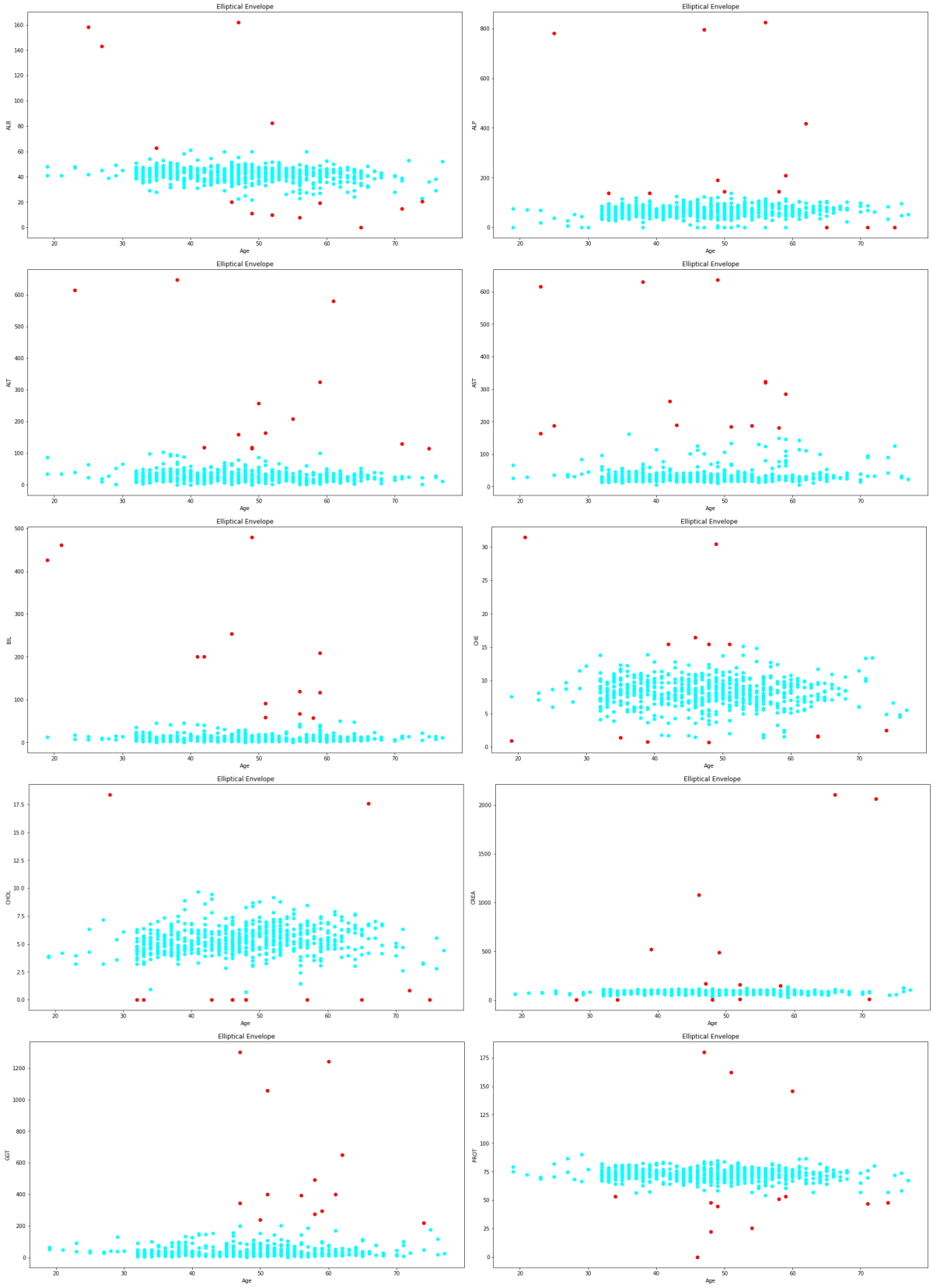
Πίνακας 4.18: Ποσοστά του HCV για 10% με Support Vector Machine

HCV data set with SVM	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	79.27%	77.03%	100%
Elliptical Envelope	91.05%	82.52%	95.45%
GMM	91.69%	88.98%	96.16%
Isolation Forest	93.47%	88.52%	94.87%
LOF	93.55%	94.84%	96.36%
Mahalanobis distance	93.47%	88.83%	95.27%
One Class SVM	91.45%	89.84%	94.33%

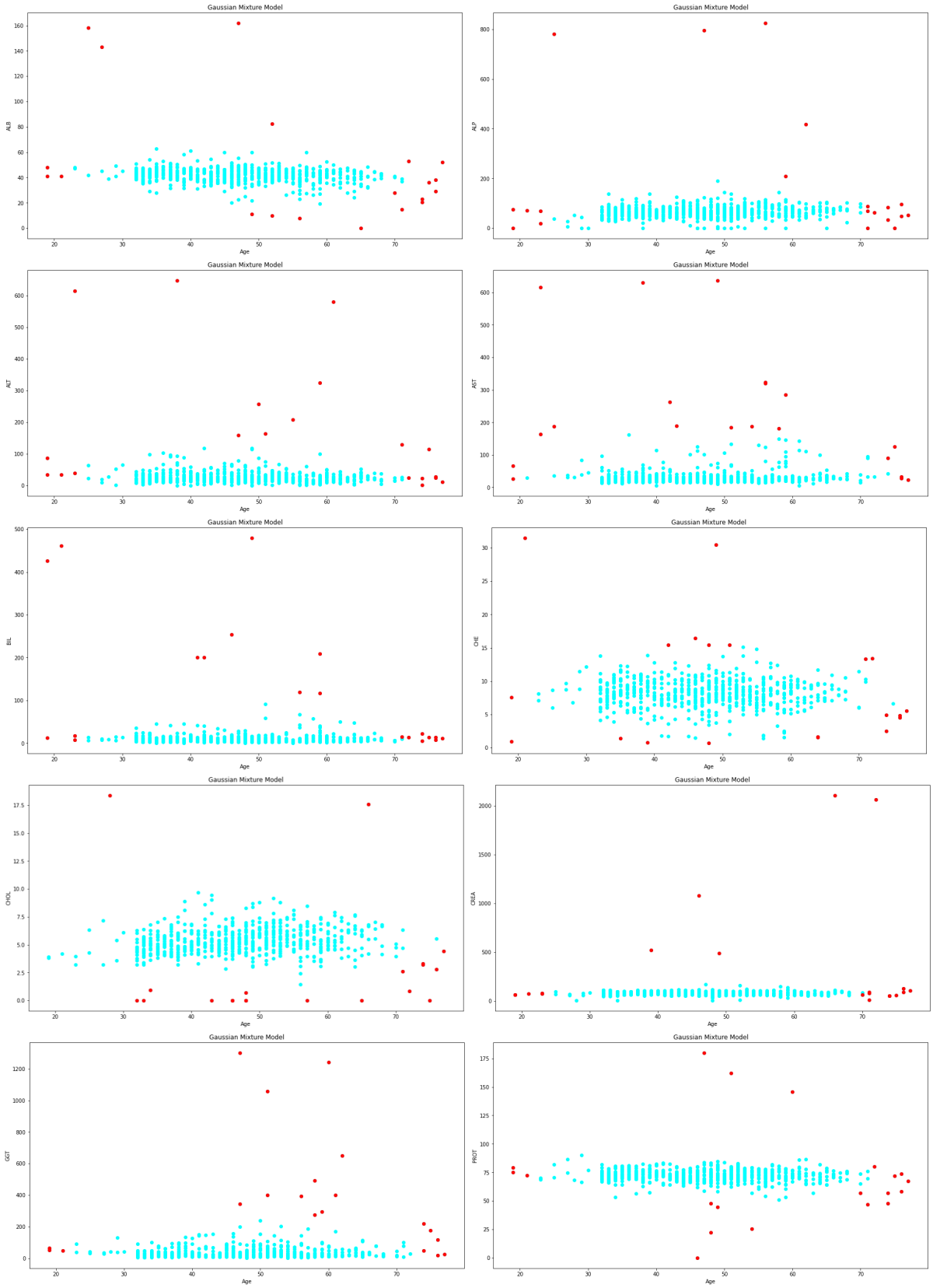
Σχήμα 4.22: Γραφήματα ακραίων τιμών DBscan στο HCV



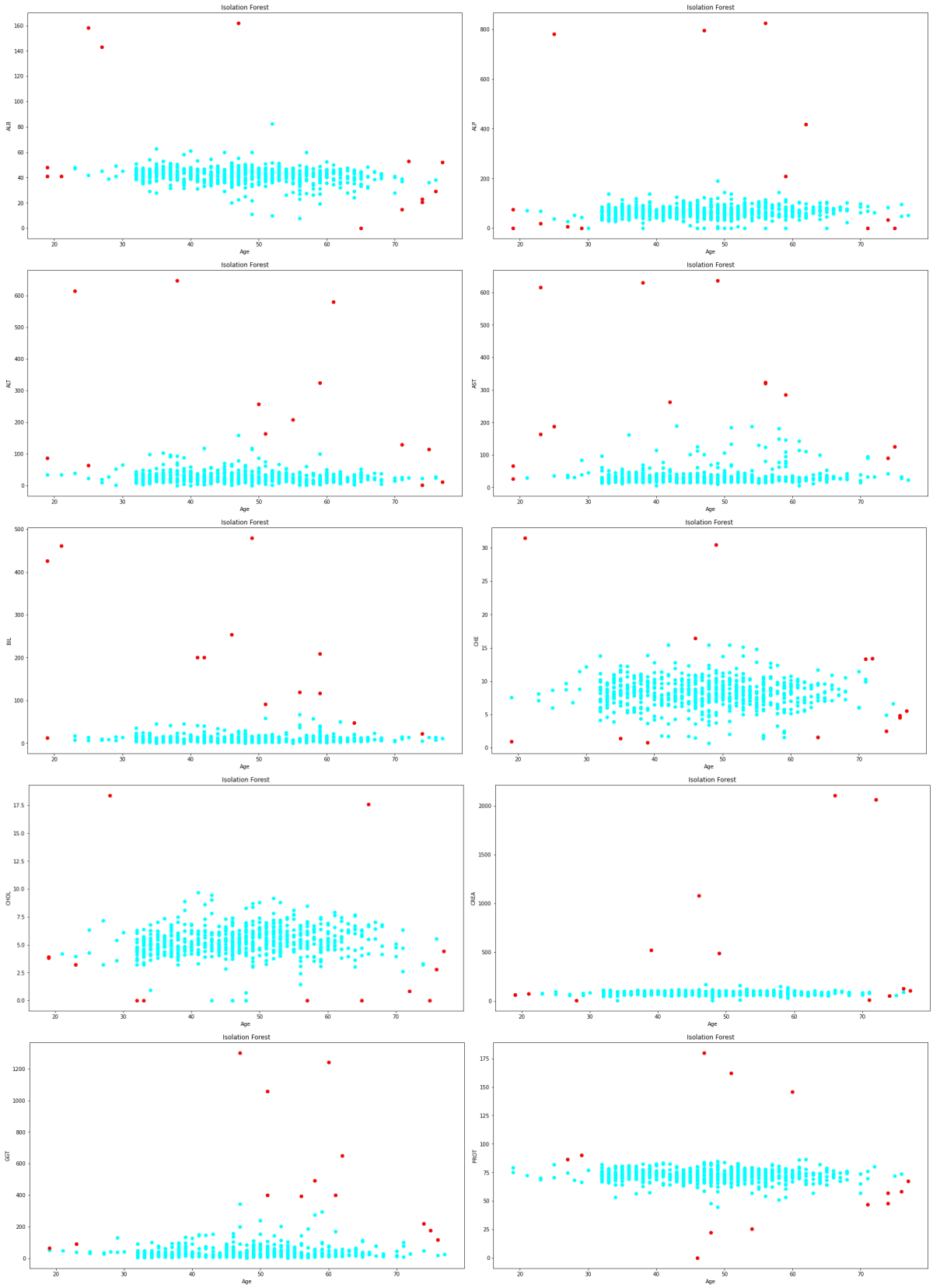
Σχήμα 4.23: Γραφήματα ακραίων τιμών Elliptical Envelope στο HCV



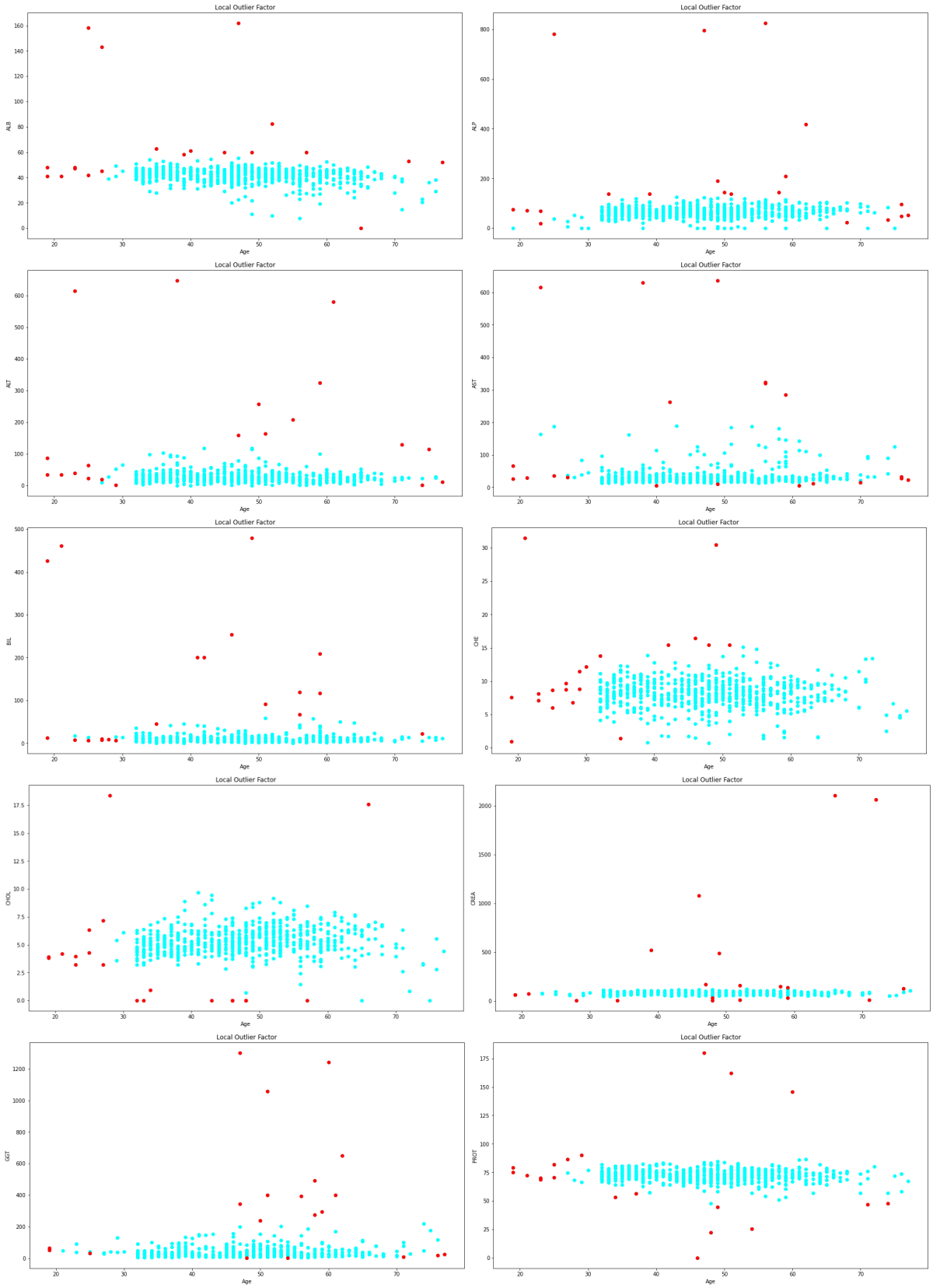
Σχήμα 4.24: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο HCV



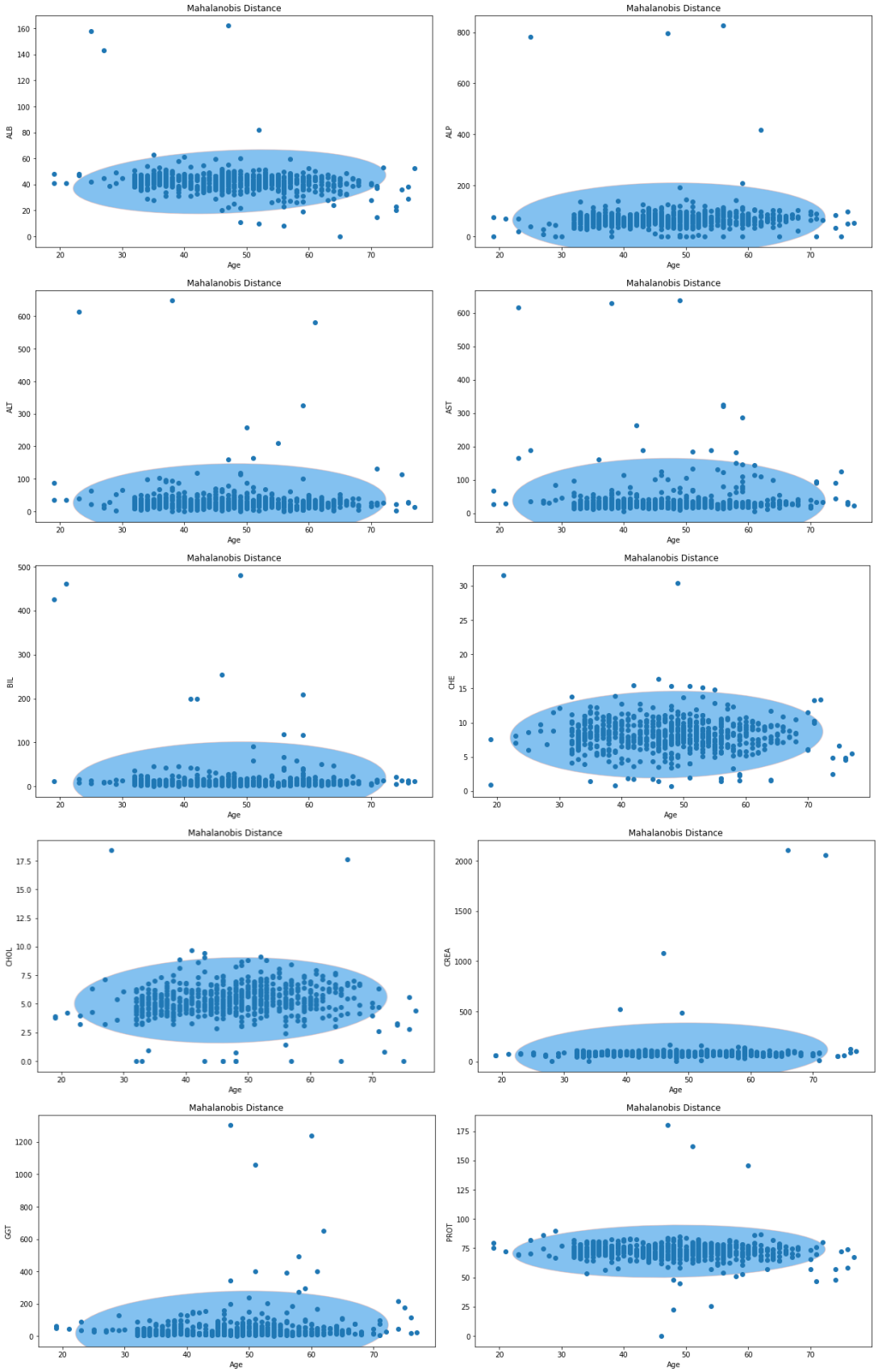
Σχήμα 4.25: Γραφήματα ακραίων τιμών Isolation Forest στο HCV



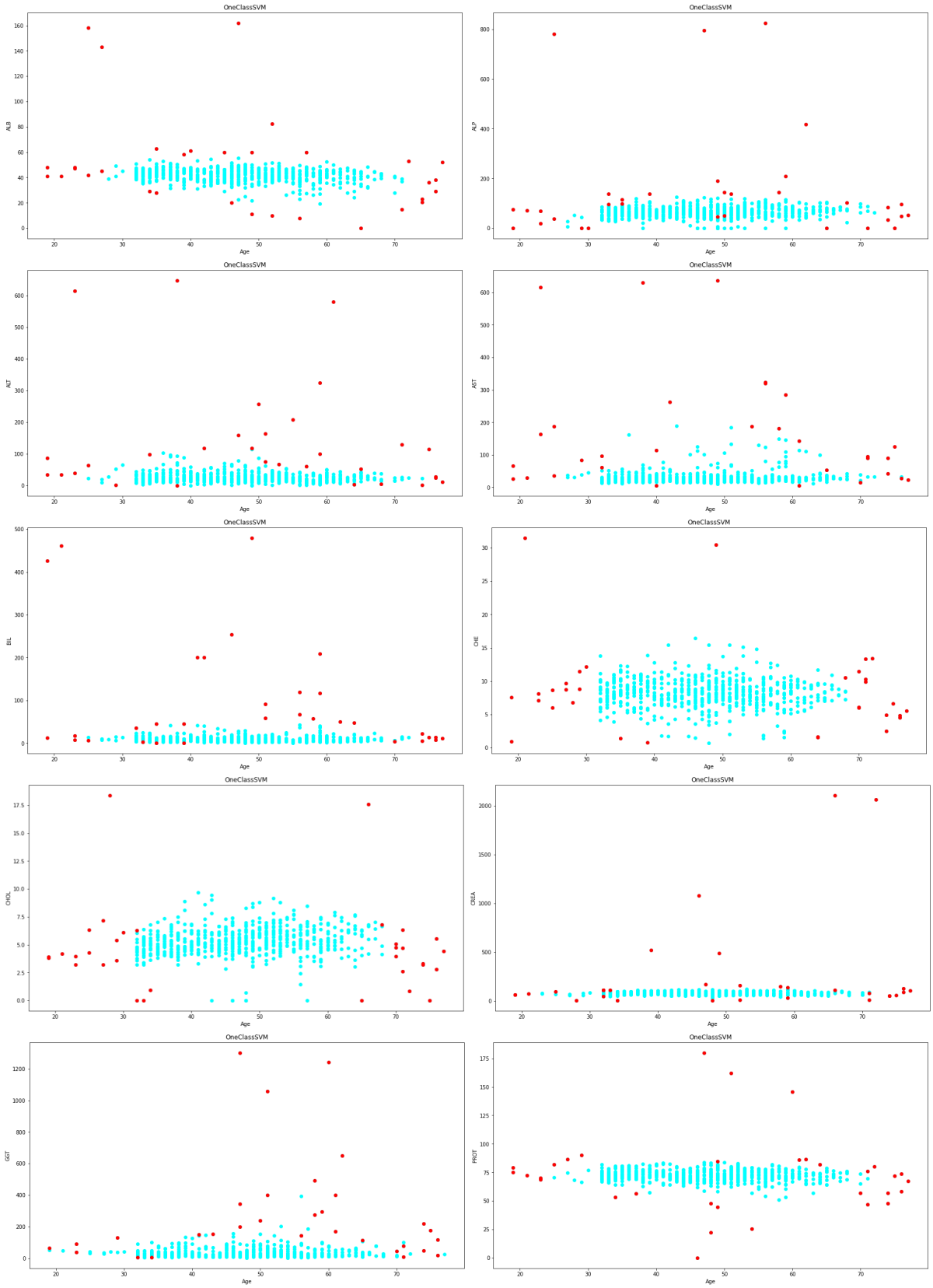
Σχήμα 4.26: Γραφήματα ακραίων τιμών Local Outlier Factor στο HCV



Σχήμα 4.27: Γραφήματα ακραίων τιμών Mahalanobis Distance στο HCV



Σχήμα 4.28: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο HCV



4.6.5 User Knowledge Modeling

Το σετ δεδομένων user knowledge modeling αφορά την κατηγορία που ανήκει το εκάστοτε ανάλογα με τις γνώσεις και τις σπουδές του. Οι μεταβλητές που data set είναι οι: ID, STG, SCG, STR, LPR, PEG και UNS όπου είναι οι κατηγορίες των ατόμων ανάλογα τη βαθμολογία που συλλέγουν από τις σπουδές και γνώσεις τους. Επίσης ως σταθερή μεταβλητή έχει οριστεί το ID. Στα Σχήματα 4.29, 4.30, 4.31, 4.32, 4.33, 4.34 και 4.35 παρουσιάζονται οι ακραίες τιμές (10%) που ανιχνεύονται από τον κάθε αλγόριθμο. Στους Πίνακες 4.21 και 4.22 παραθέτονται τα ποσοστά των αποτελεσμάτων από τις προβλέψεις των αλγορίθμων μηχανικής μάθησης με 10% ακραίες τιμές. Ο αλγόριθμος Mahalanobis distance και με Random Forest και με SVM παρουσιάζει μεγαλύτερα ποσοστά πρόβλεψης χωρίς ακραίες τιμές, από ότι του αρχικού σετ δεδομένων. Επιπλέον και στις δύο περιπτώσεις του Random Forest και του SVM ο αλγόριθμος OneClassSVM εμφανίζει χειρότερα αποτελέσματα αφού αφαιρεθούν οι ακραίες τιμές, παρά όταν υπάρχουν αυτές στο σετ δεδομένων. Αυτό οφείλεται στον τρόπο λειτουργίας του αλγόριθμου καθώς όπως φαίνεται στο Σχήμα 4.35 βρίσκει τις περισσότερες ακραίες τιμές, όπου μέσα σε αυτές ανήκουν και κανονικές τιμές του σετ δεδομένων. Στις υπόλοιπες περιπτώσεις τα αποτελέσματα είναι ιδανικά. Στους Πίνακες 4.19 και 4.20 παρουσιάζονται τα αποτελέσματα για την τάξη των 5% ακραίων τιμών. Παρατηρούμε ότι σχεδόν σε όλους τους αλγορίθμους και τις περιπτώσεις, τα ποσοστά που λάβαμε για το 5% των ακραίων τιμών είναι μεγαλύτερα από αυτά του 10%. Επίσης ο OneClassSVM και στις δύο περιπτώσεις του Random Forest και του SVM παρουσιάζει μικρότερα ποσοστά αφού αφαιρεθούν οι ακραίες τιμές από το σετ δεδομένων παρά από όταν προσθεθούν αυτές στο σετ. Αντίστοιχα ο Mahalanobis distance και στις δύο περιπτώσεις Random Forest και SVM έχει μεγαλύτερο ποσοστό ακριβείας όταν αφαιρεθούν οι ακραίες τιμές, σε σύγκριση με το αρχικό σετ δεδομένων.

Πίνακας 4.19: Ποσοστά του User knowledge modeling για 5% με Random Forest

UKM data set with RF	Original data	With 5% outliers	Without outliers
DBscan	92.50%	86.30%	92.64%
Elliptical Envelope	93.85%	88.89%	92.04%
GMM	91.54%	90.93%	91.22%
Isolation Forest	94.23%	90.56%	92.35%
LOF	92.69%	86.68%	90.57%
Mahalanobis distance	94.04%	90%	99.81%
One Class SVM	92.50%	90%	82.35%

Πίνακας 4.20: Ποσοστά του User knowledge modeling για 5% με Support Vector Machine

UKM data set with SVM	Original data	With 5% outliers	Without outliers
DBscan	84.42%	78.15%	83.77%
Elliptical Envelope	83.65%	81.30%	85.10%
GMM	84.23%	82.41%	84.49%
Isolation Forest	86.35%	83.15%	85.63%
LOF	85.77%	79.81%	83.58%
Mahalanobis distance	88.08%	80.74%	100%
One Class SVM	87.12%	82.78%	77.65%

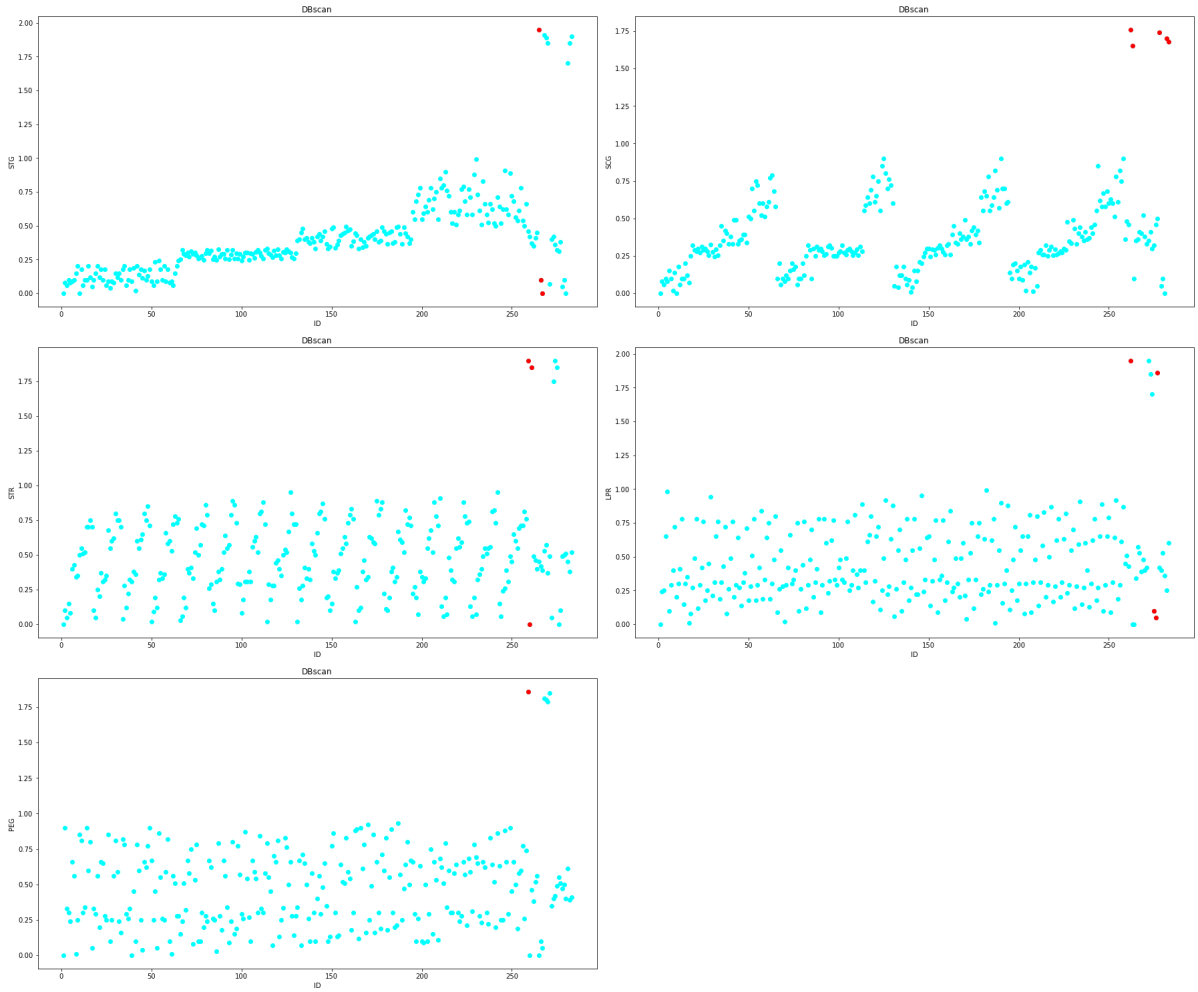
Πίνακας 4.21: Ποσοστά του User knowledge modeling για 10% με Random Forest

UKM data set with RF	Original data	With 10% outliers	Without outliers
DBscan	90.96%	86.84%	90.37%
Elliptical Envelope	91.15%	88.60%	90.19%
GMM	92.88%	88%	91%
Isolation Forest	91.92%	88.77%	90.19%
LOF	93.27%	88.42%	88.73%
Mahalanobis distance	92.69%	85.44%	99.42%
One Class SVM	93.65%	85.09%	82.97%

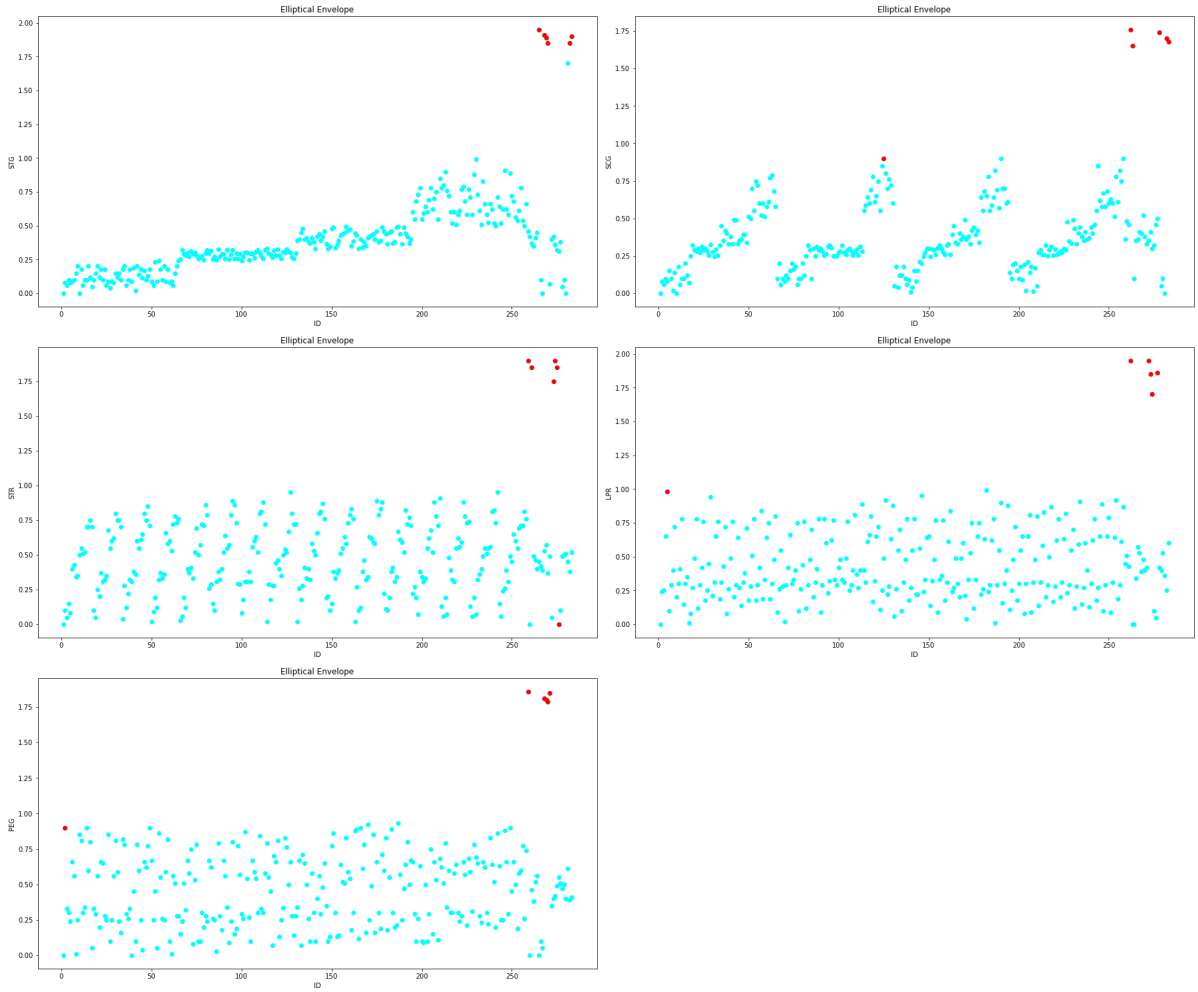
Πίνακας 4.22: Ποσοστά του User knowledge modeling για 10% με Support Vector Machine

UKM data set with SVM	Original data	With 10% outliers	Without outliers
DBscan	84.81%	79.47%	82.59%
Elliptical Envelope	83.46%	79.82%	82.83%
GMM	84%	82%	84%
Isolation Forest	84.81%	82.11%	84.34%
LOF	85.96%	81.05%	82%
Mahalanobis distance	81.54%	79.12%	100%
One Class SVM	86.73%	76.14%	72.16%

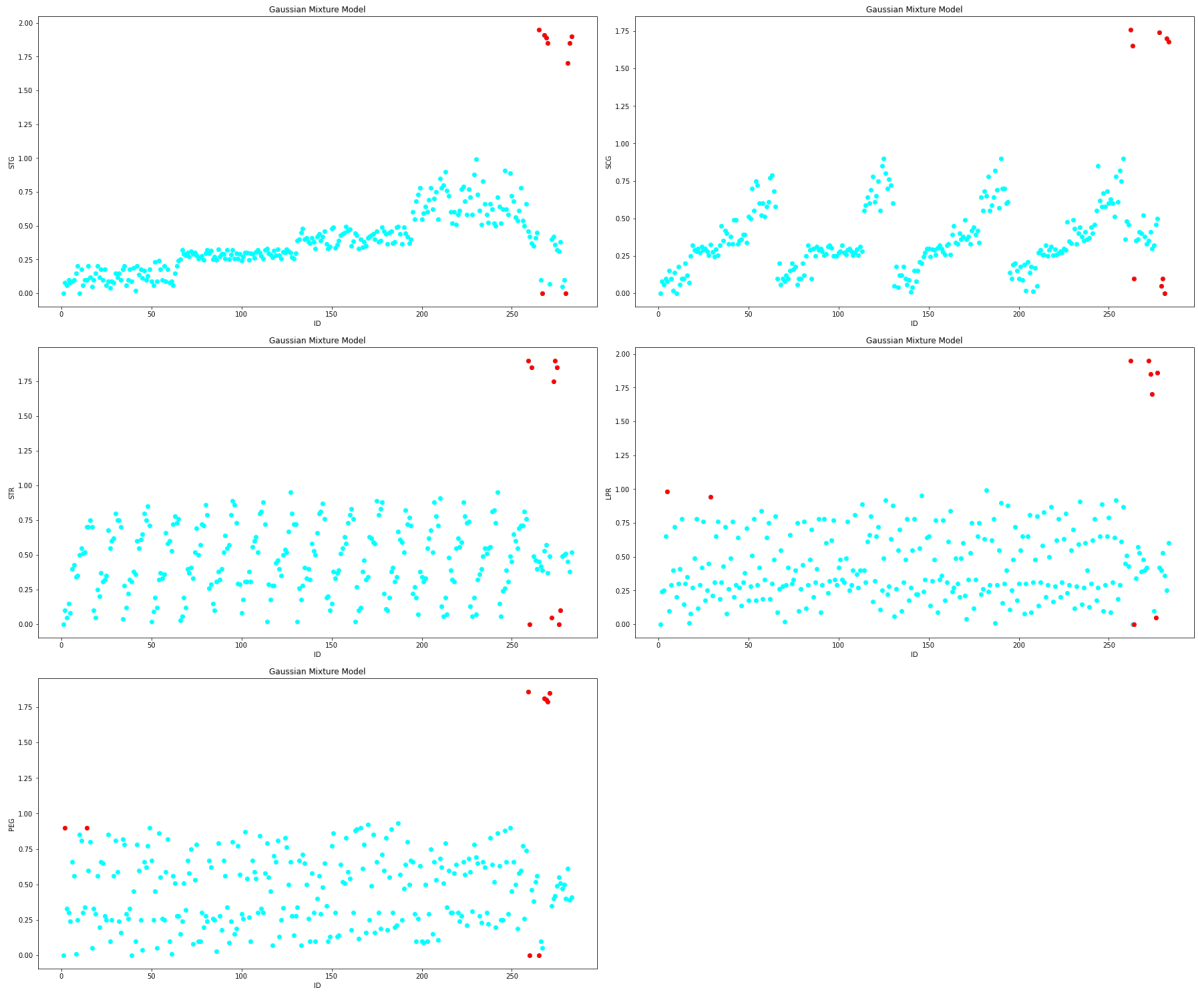
Σχήμα 4.29: Γραφήματα ακραίων τιμών DBscan στο User Knowledge Modeling



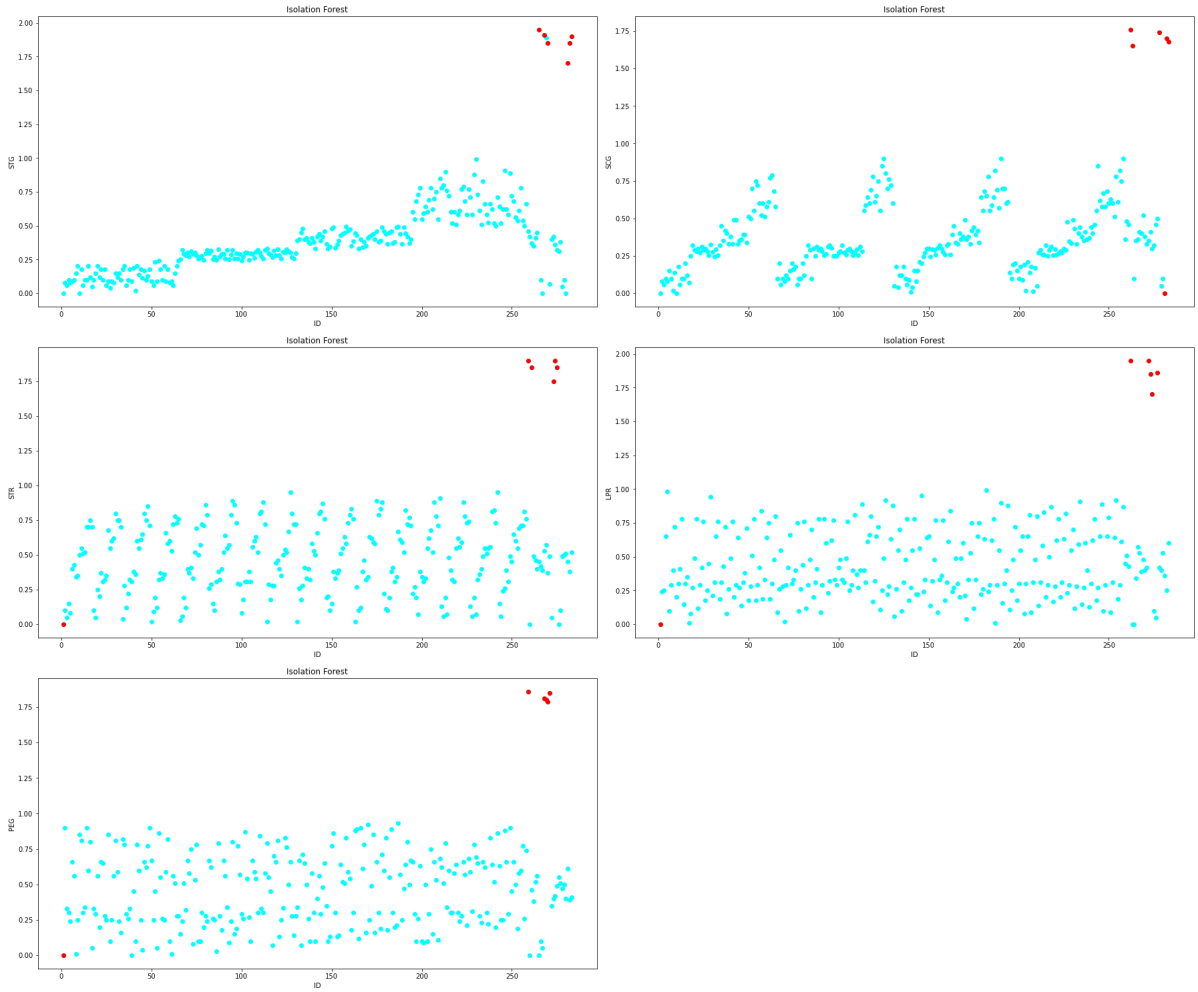
Σχήμα 4.30: Γραφήματα ακραίων τιμών Elliptical Envelope στο User Knowledge Modeling



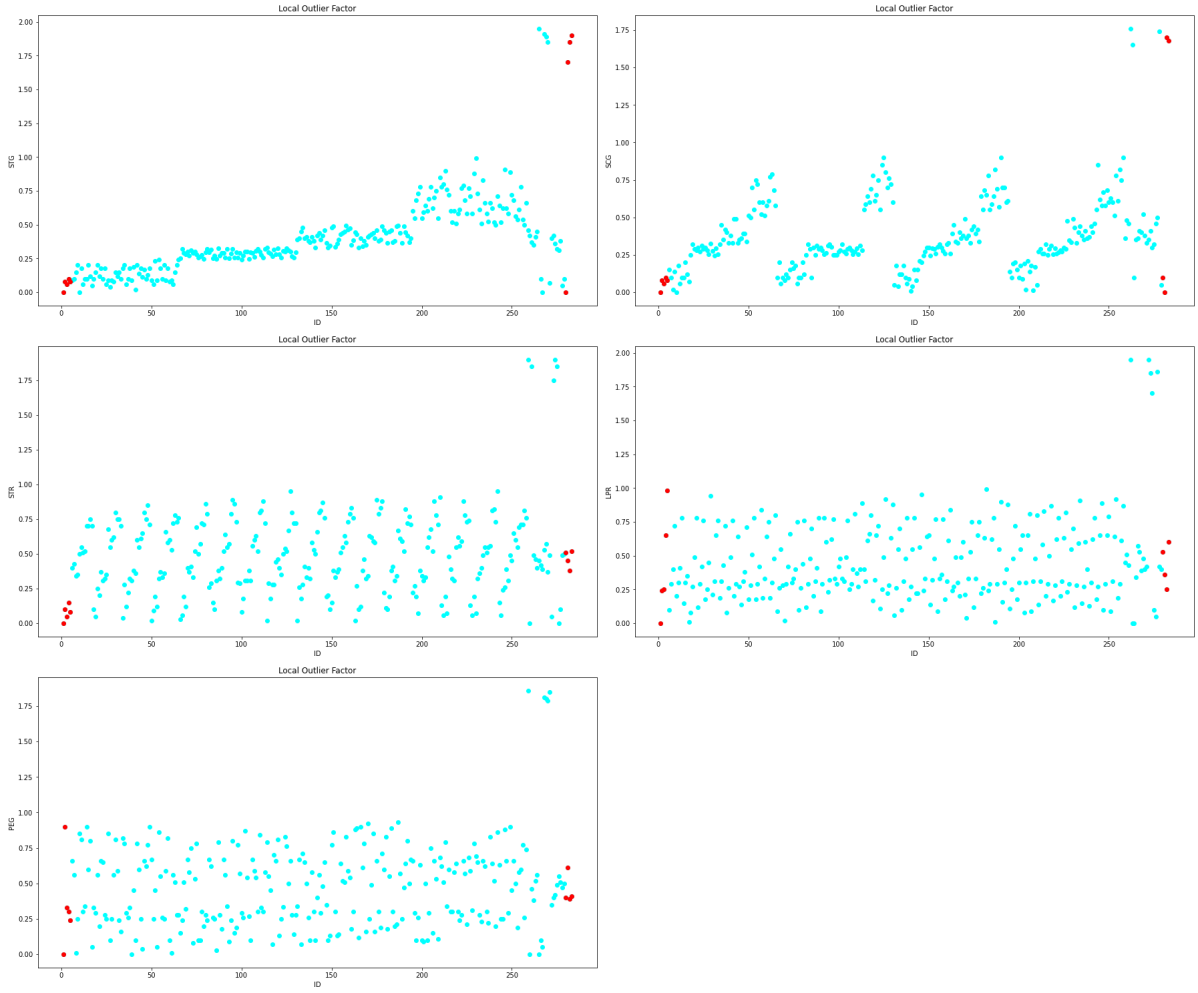
Σχήμα 4.31: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο User Knowledge Modeling



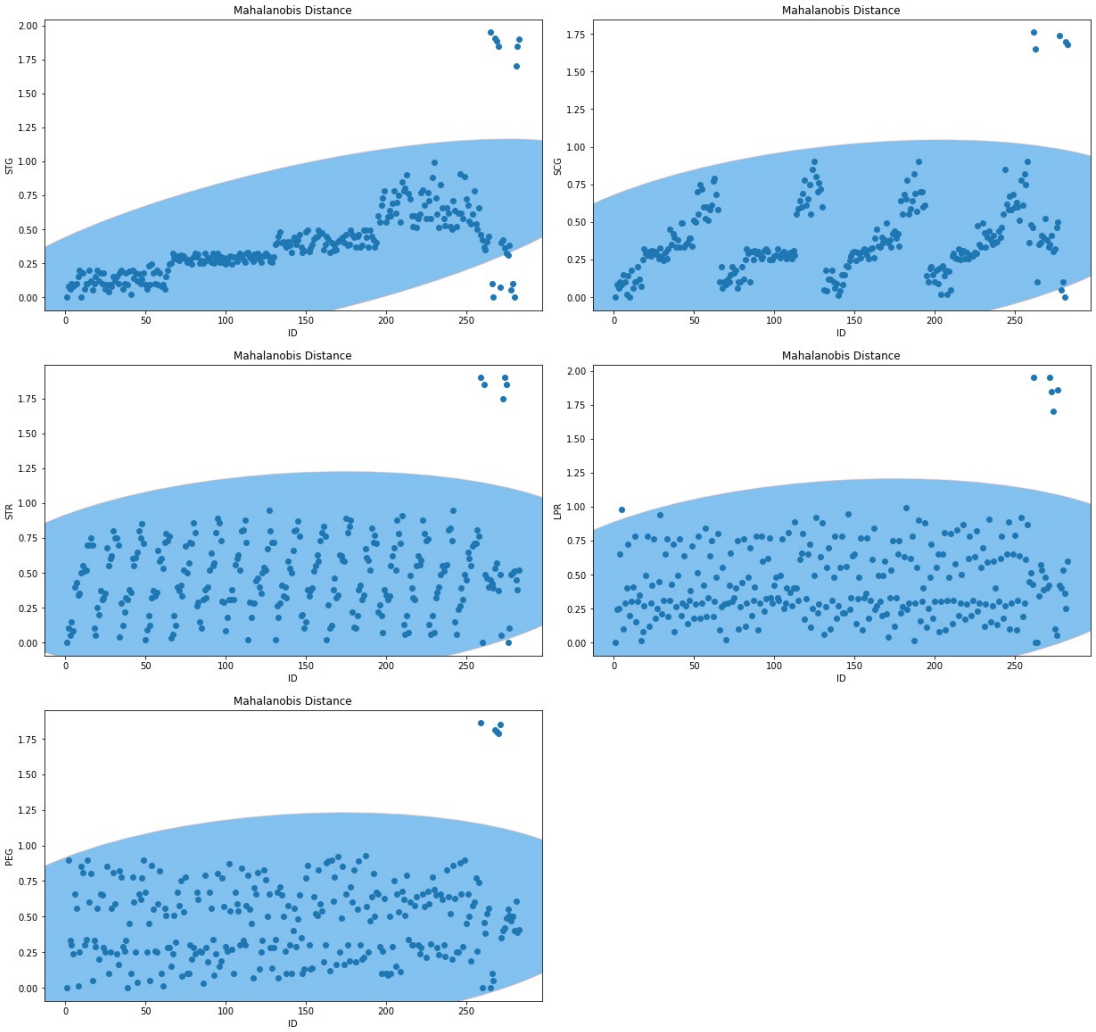
Σχήμα 4.32: Γραφήματα ακραίων τιμών Isolation Forest στο User Knowledge Modeling



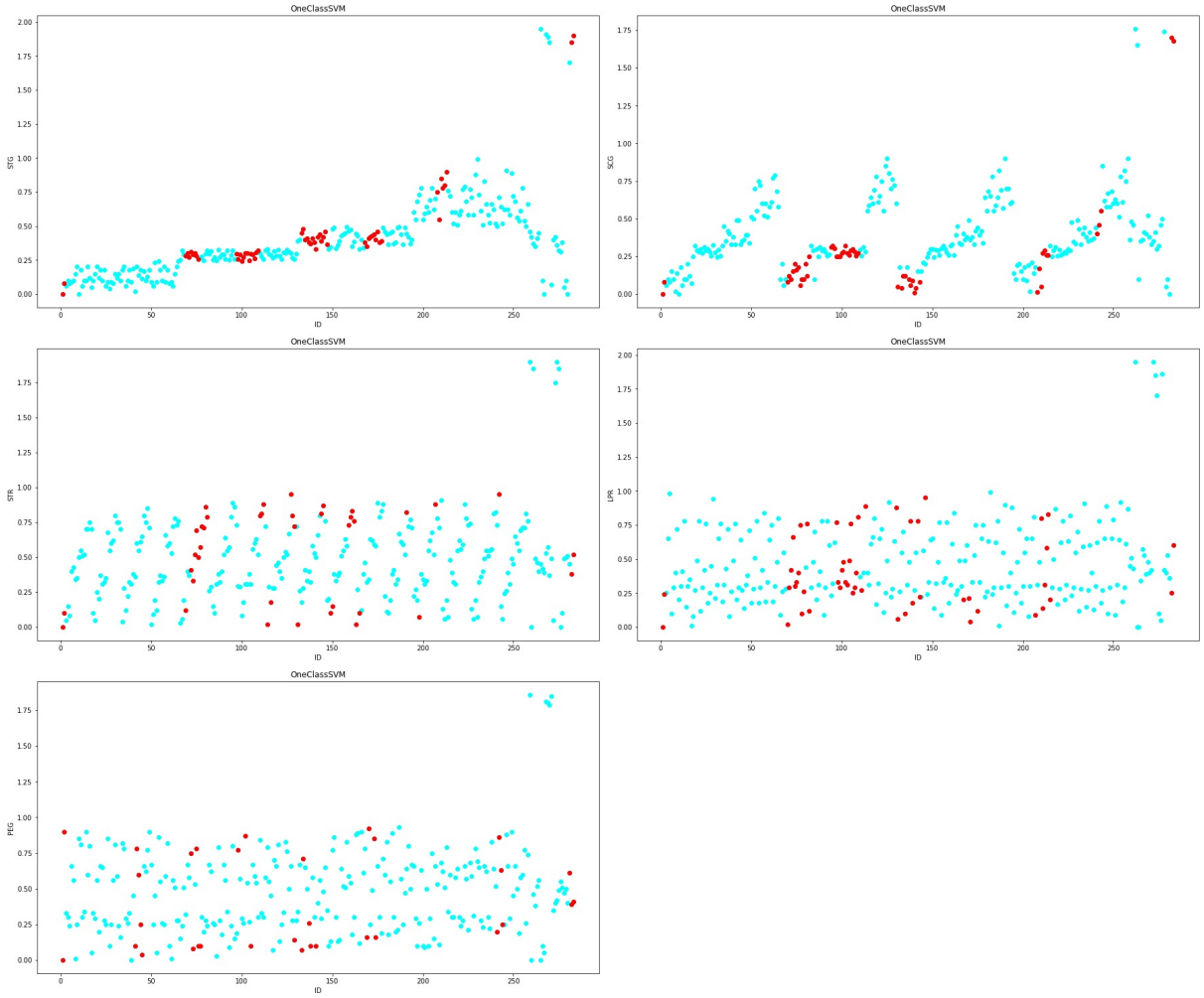
Σχήμα 4.33: Γραφήματα ακραίων τιμών Local Outlier Factor στο User Knowledge Modeling



Σχήμα 4.34: Γραφήματα ακραίων τιμών Mahalanobis Distance στο User Knowledge Modeling



Σχήμα 4.35: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο User Knowledge Modeling



4.7 Σετ δεδομένων Παλινδρόμησης

4.7.1 Air Quality

Το σετ δεδομένων air quality αφορά την υγρασία που υπάρχει στον αέρα, σύμφωνα με τα δοθέντα δεδομένα. Αποτελείται από τις εξής μεταβλητές: ID, Month, CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3), T, RH και AH όπου είναι η στήλη στόχος που αφορά τα ποσά υγρασίας στον αέρα. Στα Σχήματα 4.36, 4.37, 4.38, 4.39, 4.40, 4.41 και 4.42 παρουσιάζονται οι αλγόριθμοι με τις ανιχνευμένες ακραίες τιμές (10%). Ως σταθερή μεταβλητή έχει οριστεί το ID. Στους Πίνακες 4.25 και 4.26 παραθέτονται τα αποτελέσματα από τις προβλέψεις των αλγορίθμων μηχανικής μάθησης Random Forest και SVM αντίστοιχα με 10% ακραίες τιμές. Στα ποσοστά του αλγορίθμου Random Forest τα αποτελέσματα είναι ιδανικά. Δηλαδή πετυχαίνουμε το μεγαλύτερο ποσοστό πρόβλεψης με τα αρχικά δεδομένα, το μικρότερο ποσοστό με τις ακραίες τιμές και ένα ενδιάμεσο ποσοστό αφού αφαιρεθούν οι ακραίες τιμές. Στην κατηγορία του SVM τα ποσοστά είναι σαφώς χαμηλότερα, αλλά σε όλες τις περιπτώσεις πέρα του GMM και Mahalanobis distance τα αποτελέσματα είναι ιδανικά. Τα αρνητικά ποσοστά που εμφανίζονται στη στήλη με τις ακραίες τιμές είναι αποδεκτά αλλά προκύπτουν λόγω των ”κακών” τιμών των δεδομένων. Στην περίπτωση που έχουμε 5% ακραίες τιμές στο σετ δεδομένων όπως φαίνεται και στους Πίνακες 4.23 και 4.24 τα αποτελέσματα είναι ιδανικά για την περίπτωση του Random Forest. Παρατηρώντας όλα τα ποσοστά είναι φανερό ότι κατά το πλείστον τα ποσοστά με 5% των ακραίες τιμών είναι μεγαλύτερα από αυτά με 10%. Επιπλέον ο SVM στην περίπτωση του 5% παρουσίαζει μεγαλύτερα ποσοστά ακριβείας αφού αφαιρεθούν οι ακραίες τιμές από το σετ δεδομένων από ότι το αρχικό σετ, κάτι το οποίο απέχει από τα επιθυμητά αποτελέσματα.

Πίνακας 4.23: Ποσοστά του Air Quality για 5% με Random Forest

Air Quality data set with RF	Original data	With 5% outliers	Without outliers
DBscan	80.72%	16.29%	65.01%
Elliptical Envelope	80.72%	38.06%	73.02%
GMM	80.72%	45.71%	76.04%
Isolation Forest	80.72%	29.43%	73.60%
LOF	80.72%	49.69%	74.54%
Mahalanobis distance	80.72%	37.75%	71.04%
One Class SVM	80.72%	45.77%	58.51%

Πίνακας 4.24: Ποσοστά του Air Quality για 5% με Support Vector Machine

Air Quality data set with SVM	Original data	With 5% outliers	Without outliers
DBscan	58.86%	20.79%	61.09%
Elliptical Envelope	58.86%	-2.99%	45.29%
GMM	58.86%	7.99%	62.95%
Isolation Forest	58.86%	24.80%	62.49%
LOF	58.86%	12.66%	61.82%
Mahalanobis distance	58.86%	10.56%	62.07%
One Class SVM	58.86%	5.70%	56.77%

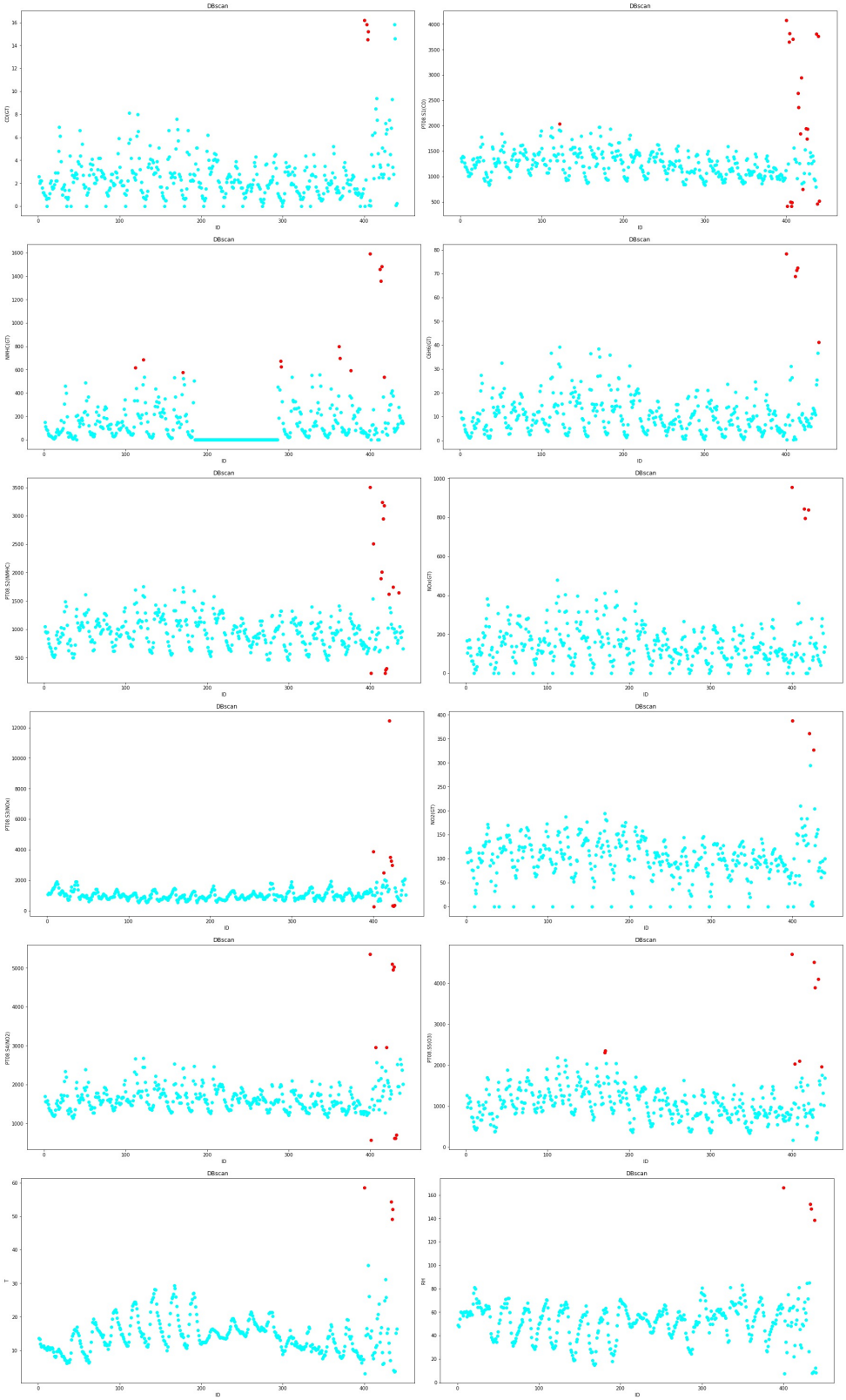
Πίνακας 4.25: Ποσοστά του Air Quality για 10% με Random Forest

Air Quality data set with RF	Original data	With 10% outliers	Without outliers
DBscan	80.72%	26.65%	51.76%
Elliptical Envelope	80.72%	31.60%	59.74%
GMM	80.72%	38.51%	71.76%
Isolation Forest	80.72%	31.24%	70.83%
LOF	80.72%	13.72%	66.40%
Mahalanobis distance	80.72%	41.13%	74.83%
One Class SVM	80.72%	23.97%	30.03%

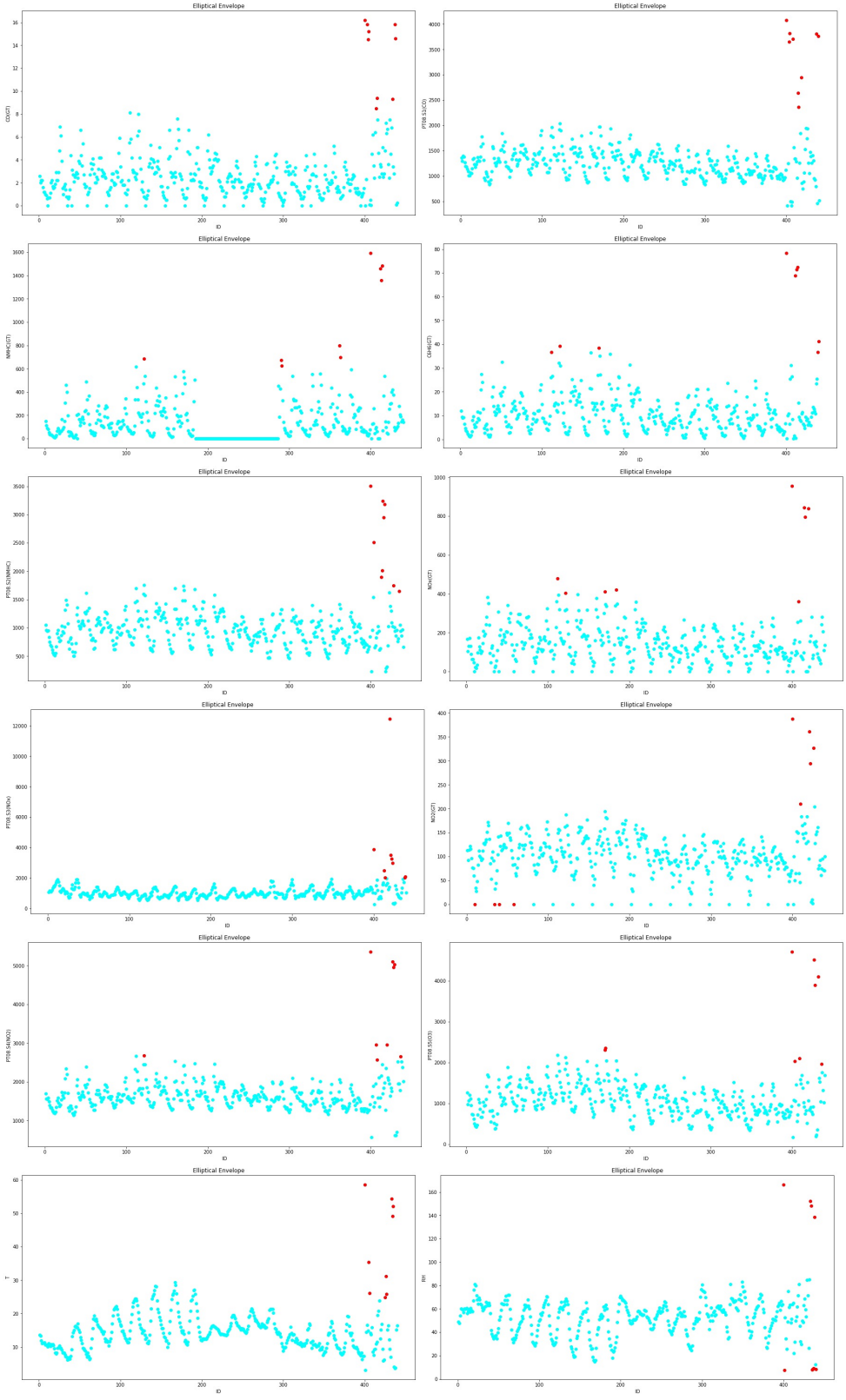
Πίνακας 4.26: Ποσοστά του Air Quality για 10% με Support Vector Machine

Air Quality data set with SVM	Original data	With 10% outliers	Without outliers
DBscan	58.86%	2.08%	51.76%
Elliptical Envelope	58.86%	-3.13%	37.12%
GMM	58.86%	2.11%	63.93%
Isolation Forest	58.86%	5.81%	45.39%
LOF	58.86%	-4.50%	59.73%
Mahalanobis distance	58.86%	6.07%	64.39%
One Class SVM	58.86%	1.88%	32.38%

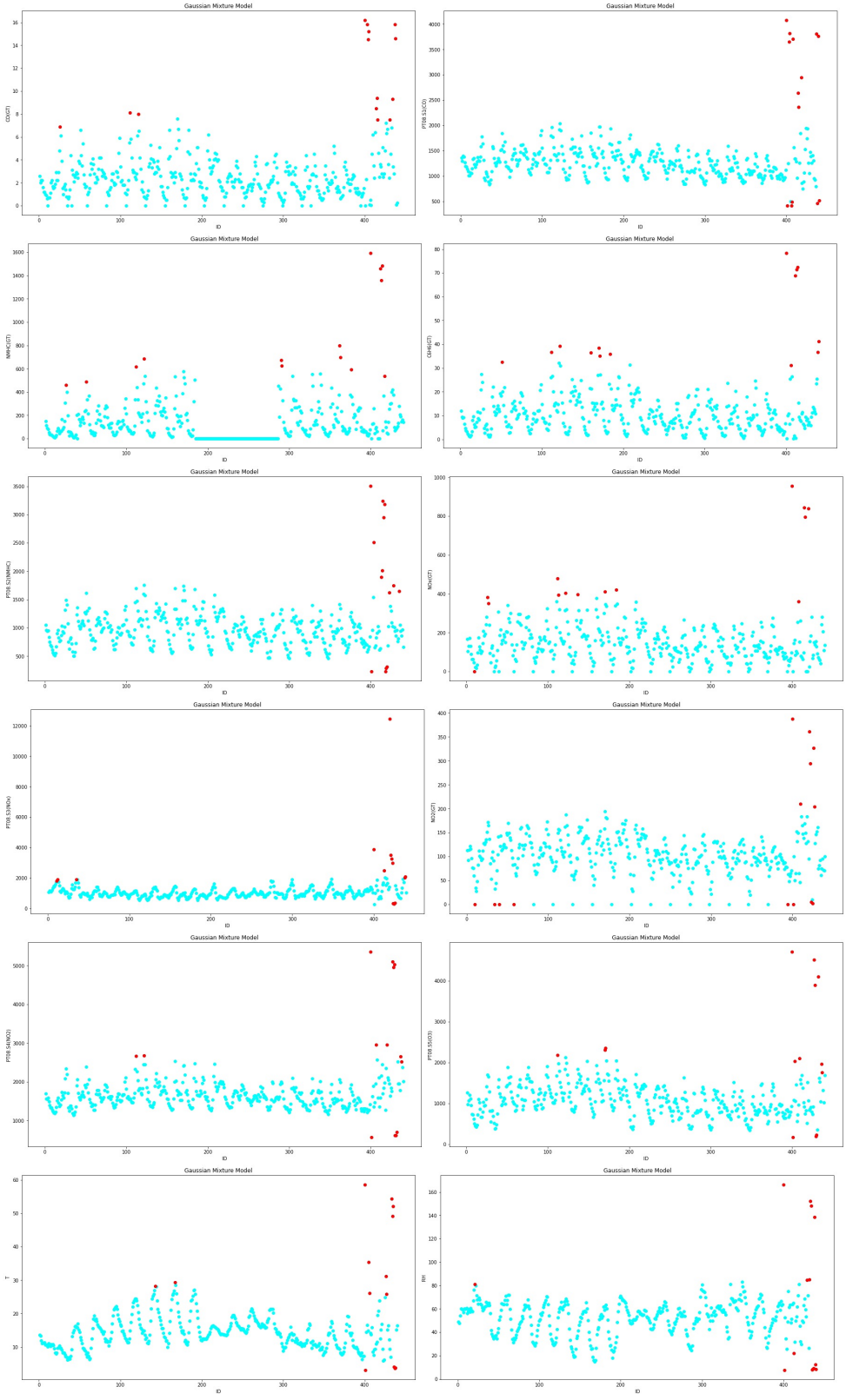
Σχήμα 4.36: Γραφήματα ακραίων τιμών DBscan στο Air Quality



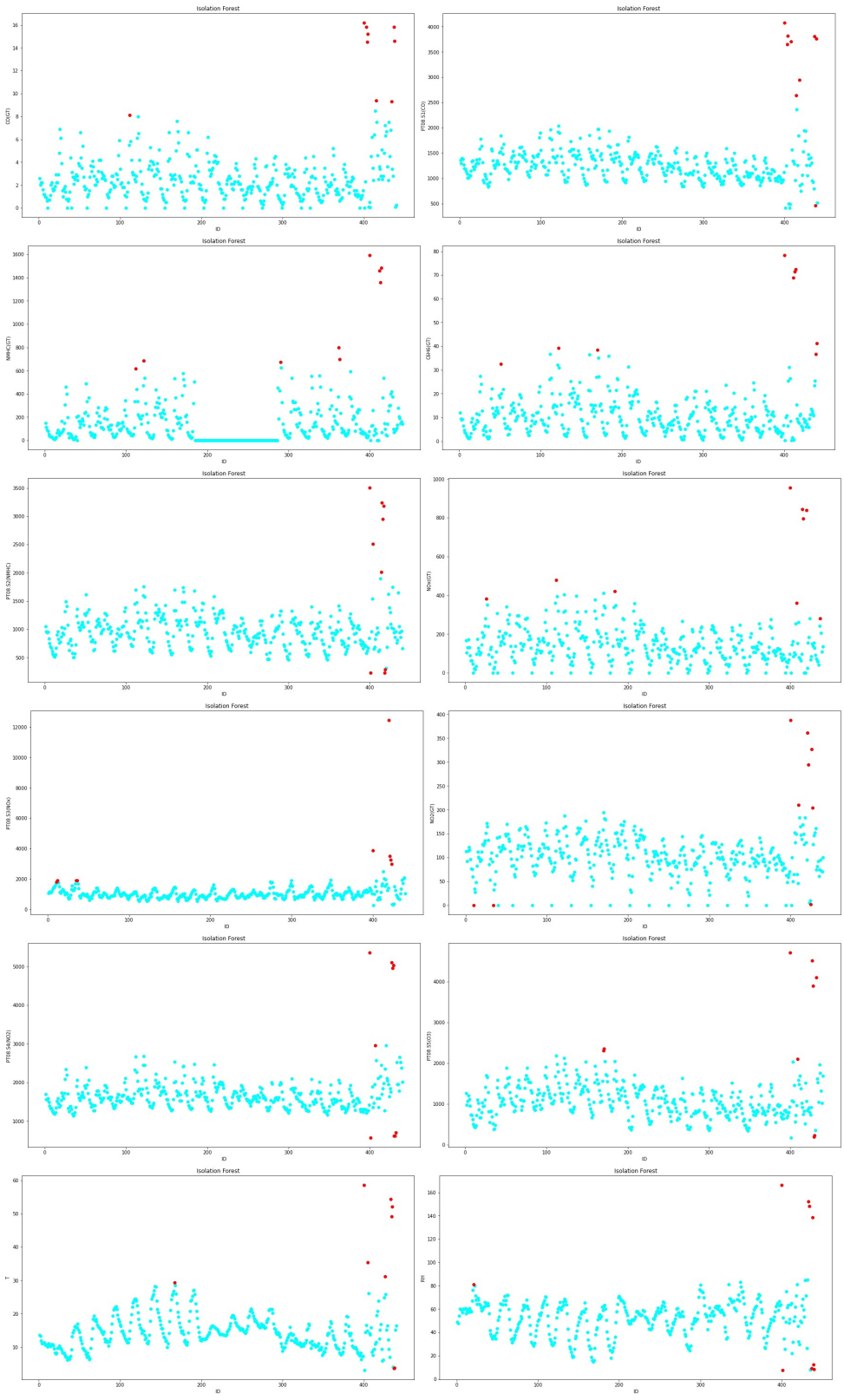
Σχήμα 4.37: Γραφήματα ακραίων τιμών Elliptical Envelope στο Air Quality



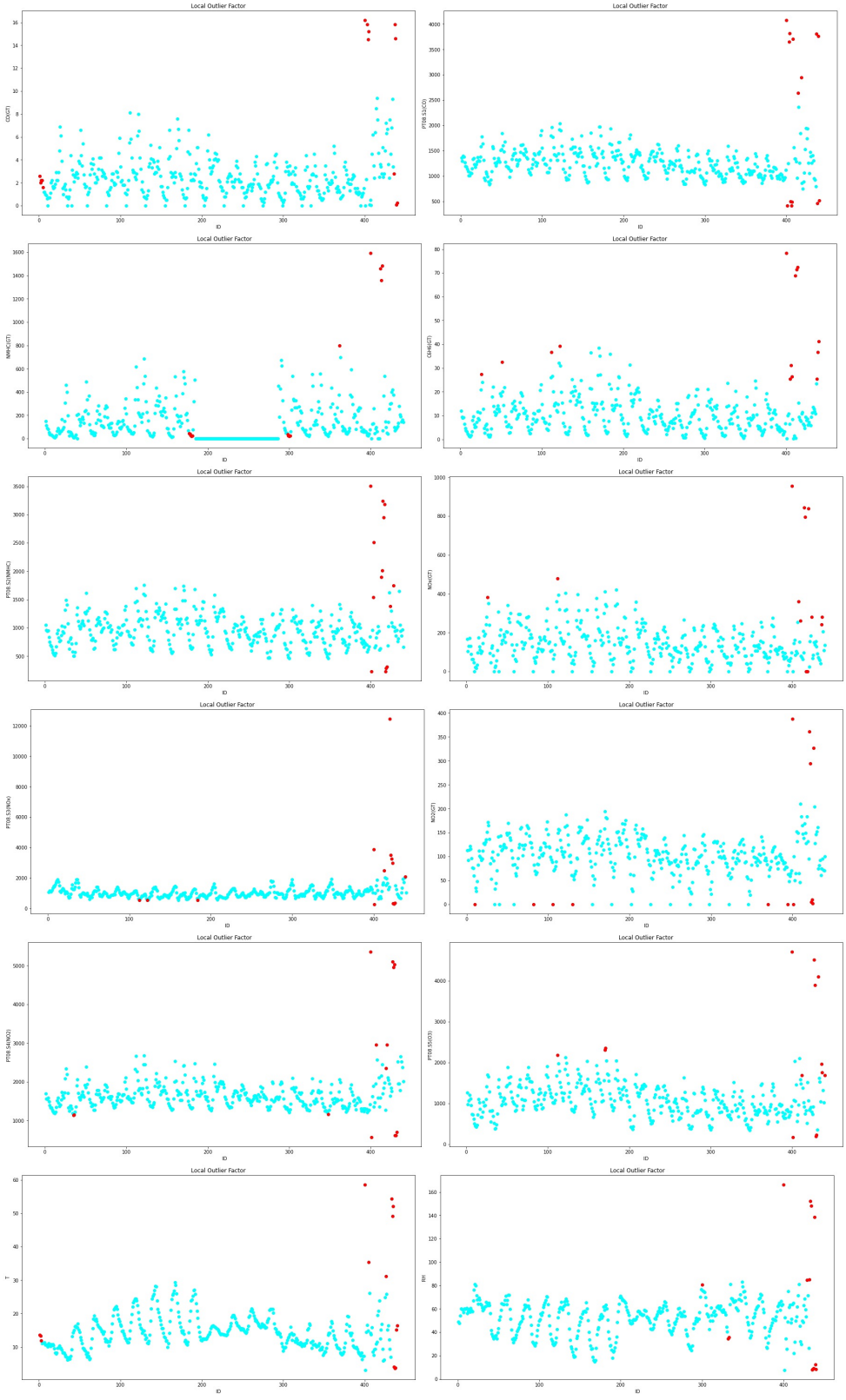
Σχήμα 4.38: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Air Quality



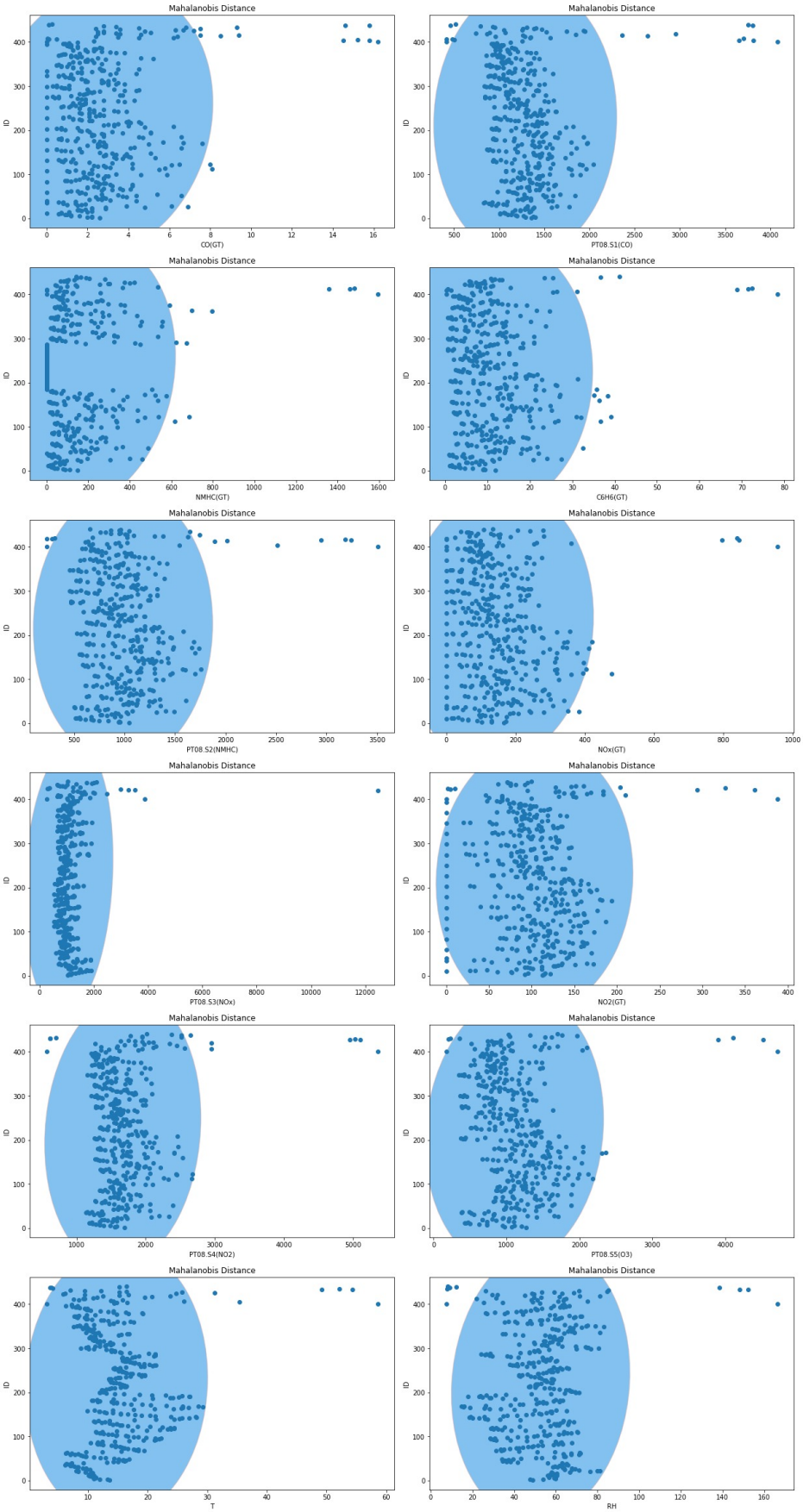
Σχήμα 4.39: Γραφήματα ακραίων τιμών Isolation Forest στο Air Quality



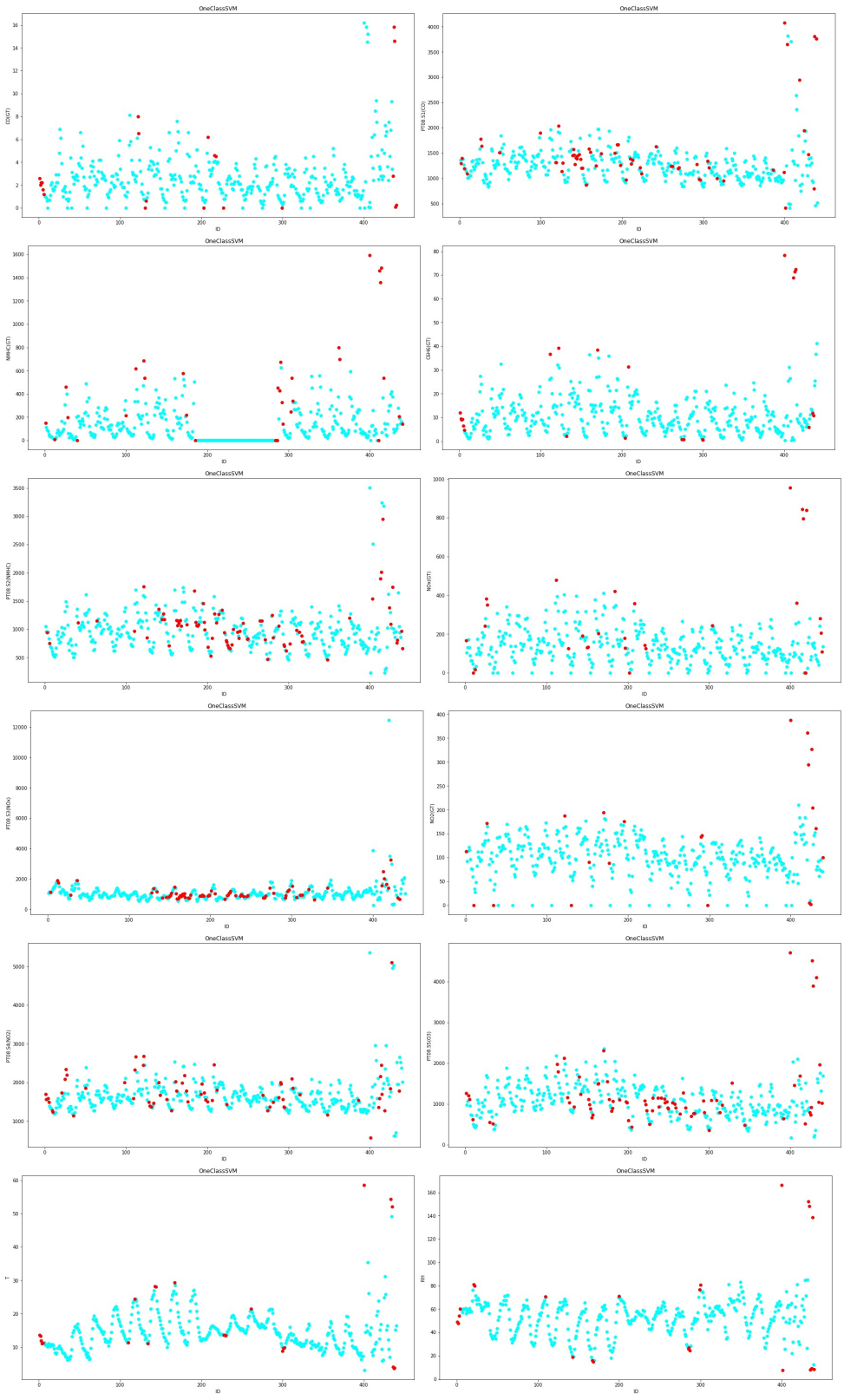
Σχήμα 4.40: Γραφήματα ακραίων τιμών Local Outlier Factor στο Air Quality



Σχήμα 4.41: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Air Quality



Σχήμα 4.42: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Air Quality



4.7.2 Forest Fires

Το συγκεκριμένο σετ δεδομένων Forest Fires αφορά την πρόβλεψη της έκτασης καμένης γης λαμβάνοντας υπόψη τις υπόλοιπες μεταβλητές. Οι μεταβλητές που έχει το σετ είναι οι: ID, X, Y, month, day, FFMC, DMC, DC, ISI, temp, RH, wind, rain και area όπου είναι η στήλη στόχος. Στα Σχήματα 4.43, 4.44, 4.45, 4.46, 4.47, 4.48 και 4.49 παραθέτονται τα αποτελέσματα μετά την ανίχνευση των ακραίων τιμών (10%) του κάθε αλγορίθμου. Επιπλέον στους Πίνακες 4.29 και 4.30 παρουσιάζονται τα αποτελέσματα από τις προβλέψεις για 10% ακραίες τιμές. Τα ποσοστά είναι ιδανικά και τηρούν τις προϋποθέσεις που τέθηκαν. Παρατηρούμε ότι τα ποσοστά του Random Forest είναι για ακόμα μία φορά υψηλότερα από αυτά του SVM. Επίσης ο αλγόριθμος OneClassSVM έχει τα χαμηλότερα ποσοστά ακριβείας από όλους τους άλλους αλγορίθμους. Στους Πίνακες 4.27 και 4.28 παρουσιάζονται τα αποτελέσματα για 5% ακραίες τιμές, όπου και σε αυτή τη περίπτωση είναι ιδανικά. Γενικώς παρατηρούμε ότι όλα τα ποσοστά που προέρχονται από το σετ δεδομένων με 5% ακραίες τιμές είναι μεγαλύτερα από αυτά με 10%. Στον αλγόριθμο DBscan με SVM το ποσοστό του σετ δεδομένων αφού αφαιρέθηκαν οι ακραίες τιμές έχει μία μικρή απόκλιση σε σύγκριση με το αρχικό, κάτι το οποίο δεν επιζητείται.

Πίνακας 4.27: Ποσοστά του Forest Fires για 5% με Random Forest

Forest Fires data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	77.11%	50.41%	75.46%
Elliptical Envelope	77.11%	46.75%	75%
GMM	77.11%	58.90%	68.39%
Isolation Forest	77.11%	51.79%	75.62%
LOF	77.11%	56.56%	72.91%
Mahalanobis distance	77.11%	55.56%	69.66%
One Class SVM	77.11%	58.94%	58.83%

Πίνακας 4.28: Ποσοστά του Forest Fires για 5% με Support Vector Machine

Forest Fires data set with SVM	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	35.94%	22.22%	36.46%
Elliptical Envelope	35.94%	22.17%	34.07%
GMM	35.94%	26.94%	33.59%
Isolation Forest	35.94%	24.59%	34.90%
LOF	35.94%	26.25%	35.58%
Mahalanobis distance	35.94%	29.29%	34.48%
One Class SVM	35.94%	24.02%	27.23%

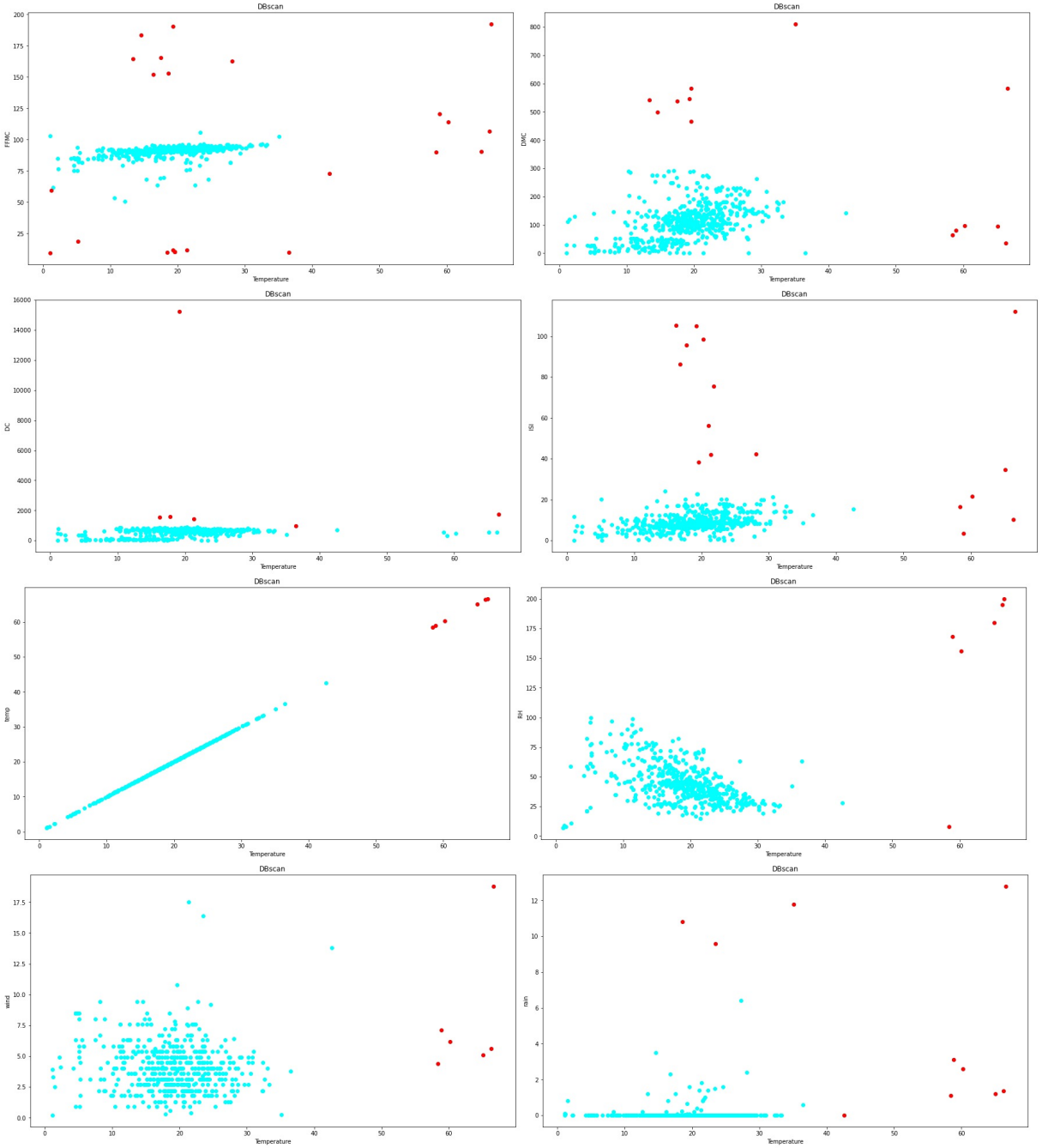
Πίνακας 4.29: Ποσοστά του Forest Fires για 10% με Random Forest

Forest Fires data set with RF	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	77.11%	42.18%	70.31%
Elliptical Envelope	77.11%	37.22%	74%
GMM	77.11%	38.50%	67.25%
Isolation Forest	77.11%	28.61%	72.67%
LOF	77.11%	51.90%	69.87%
Mahalanobis distance	77.11%	37.33%	67.71%
One Class SVM	77.11%	34.03%	55.63%

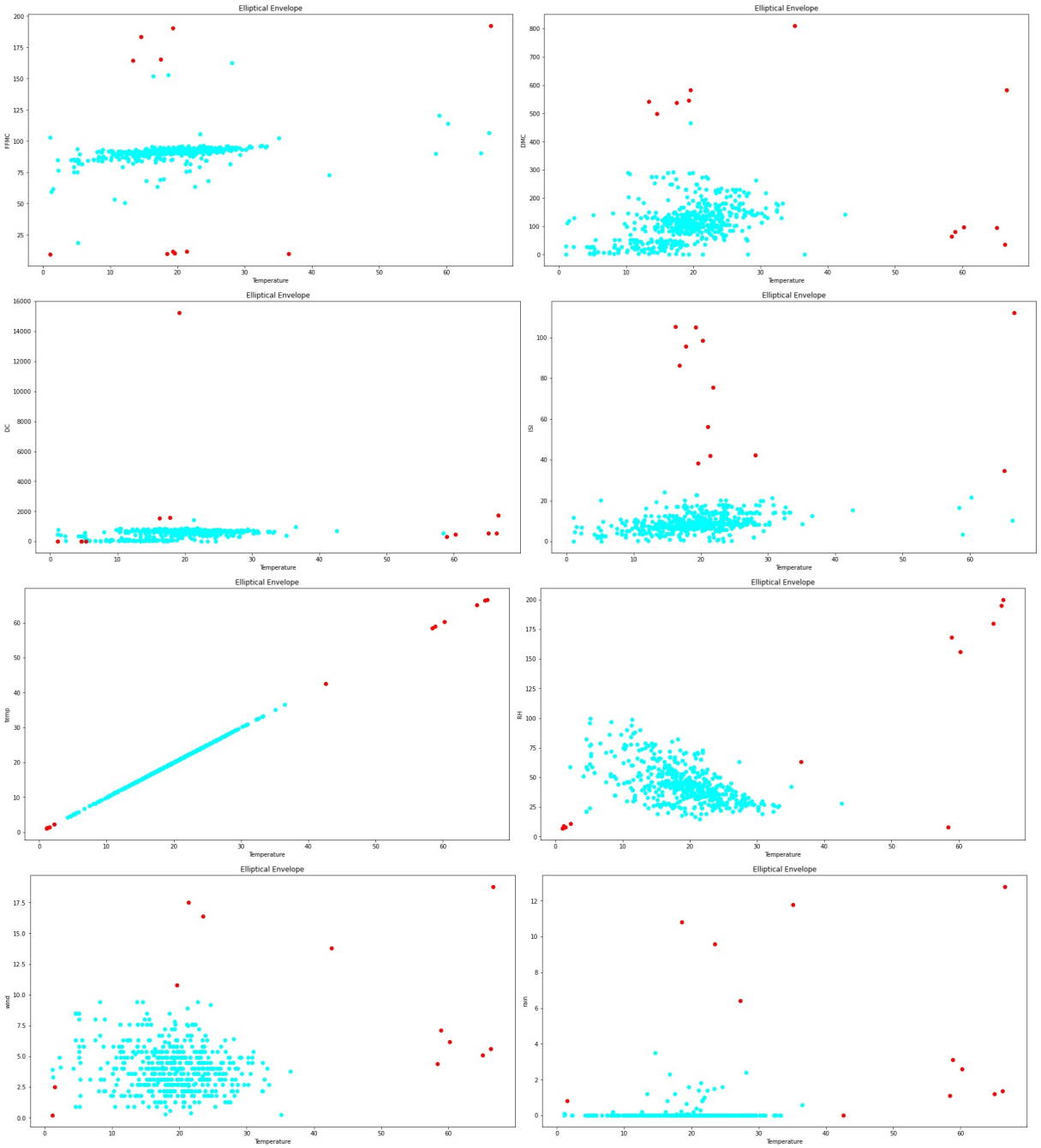
Πίνακας 4.30: Ποσοστά του Forest Fires για 10% με Support Vector Machine

Forest Fires data set with SVM	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	35.94%	15.43%	33.17%
Elliptical Envelope	35.94%	19.08%	30.53%
GMM	35.94%	13.89%	30.19%
Isolation Forest	35.94%	15.20%	33.05%
LOF	35.94%	18.29%	34.73%
Mahalanobis distance	35.94%	13.20%	32.81%
One Class SVM	35.94%	14.74%	24.25%

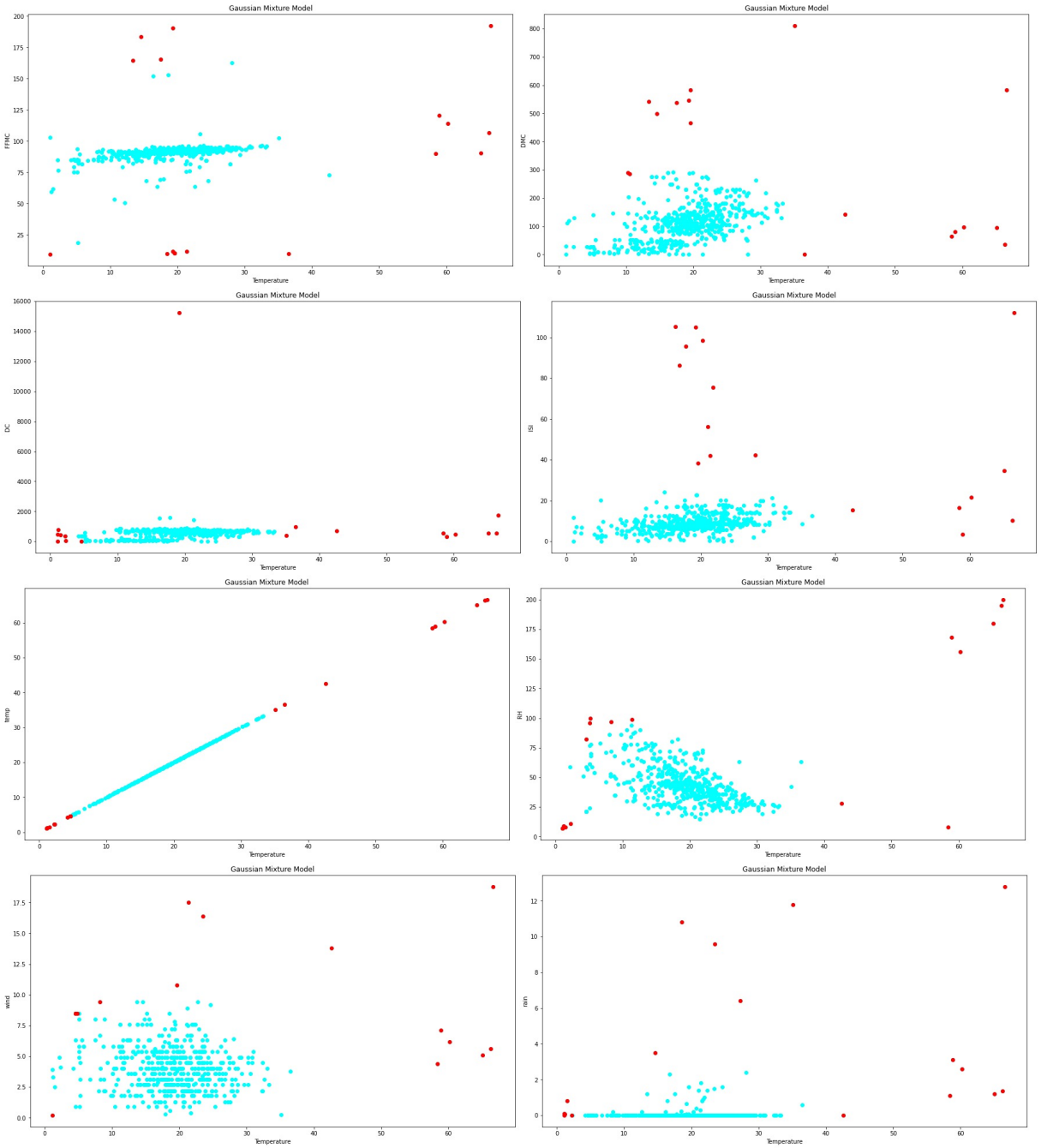
Σχήμα 4.43: Γραφήματα ακραίων τιμών DBscan στο Forest Fires



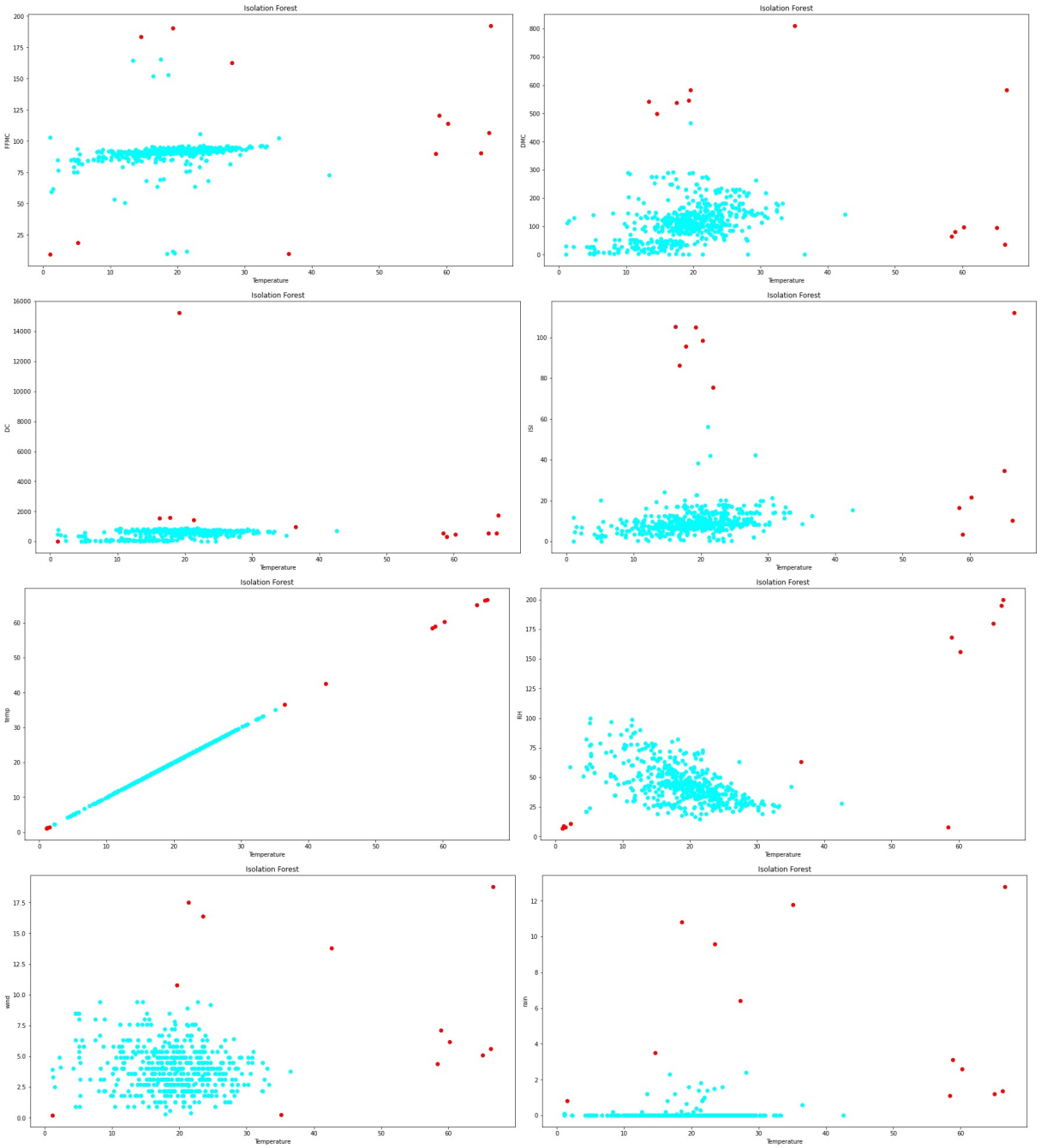
Σχήμα 4.44: Γραφήματα ακραίων τιμών Elliptical Envelope στο Forest Fires



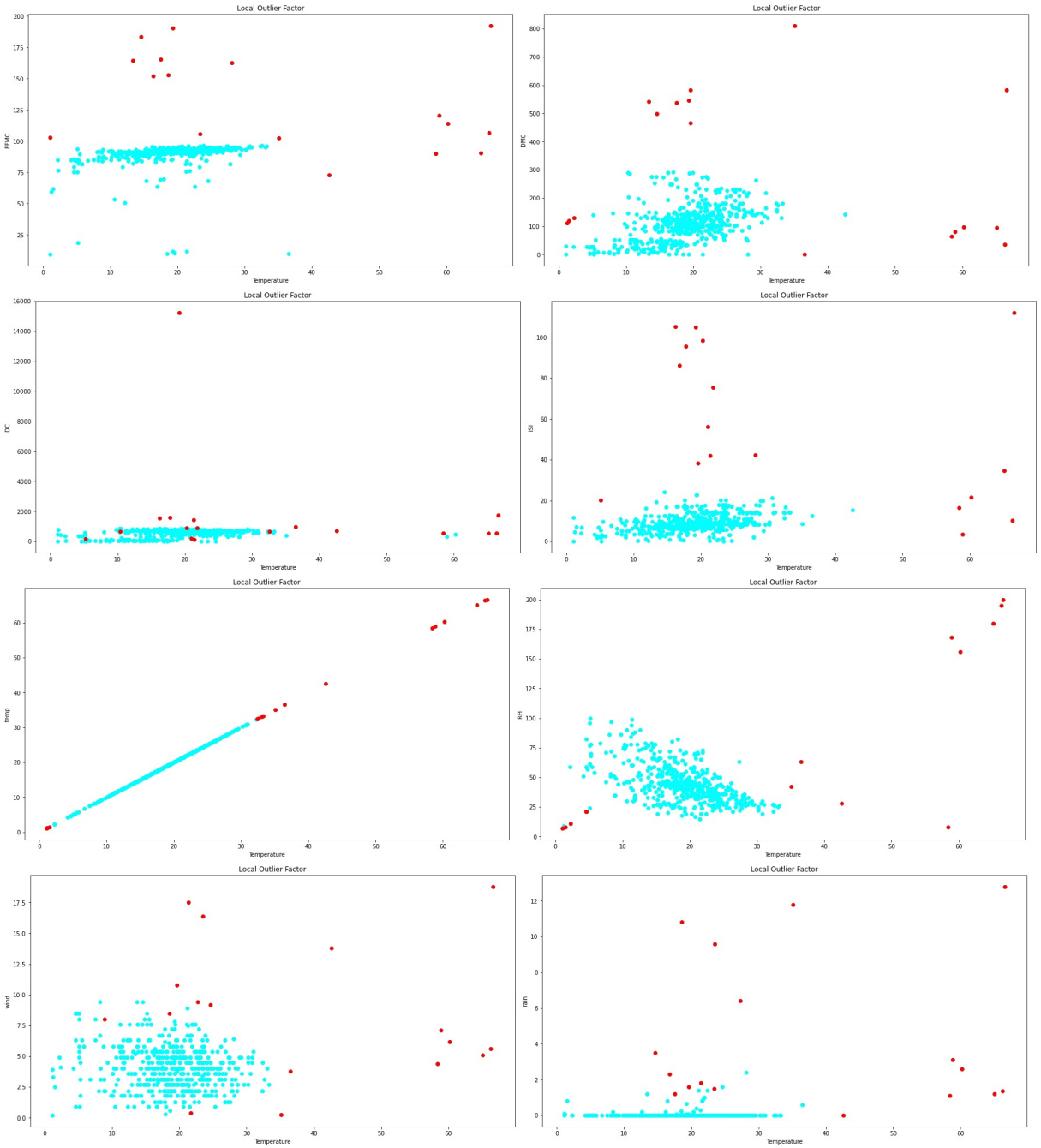
Σχήμα 4.45: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Forest Fires



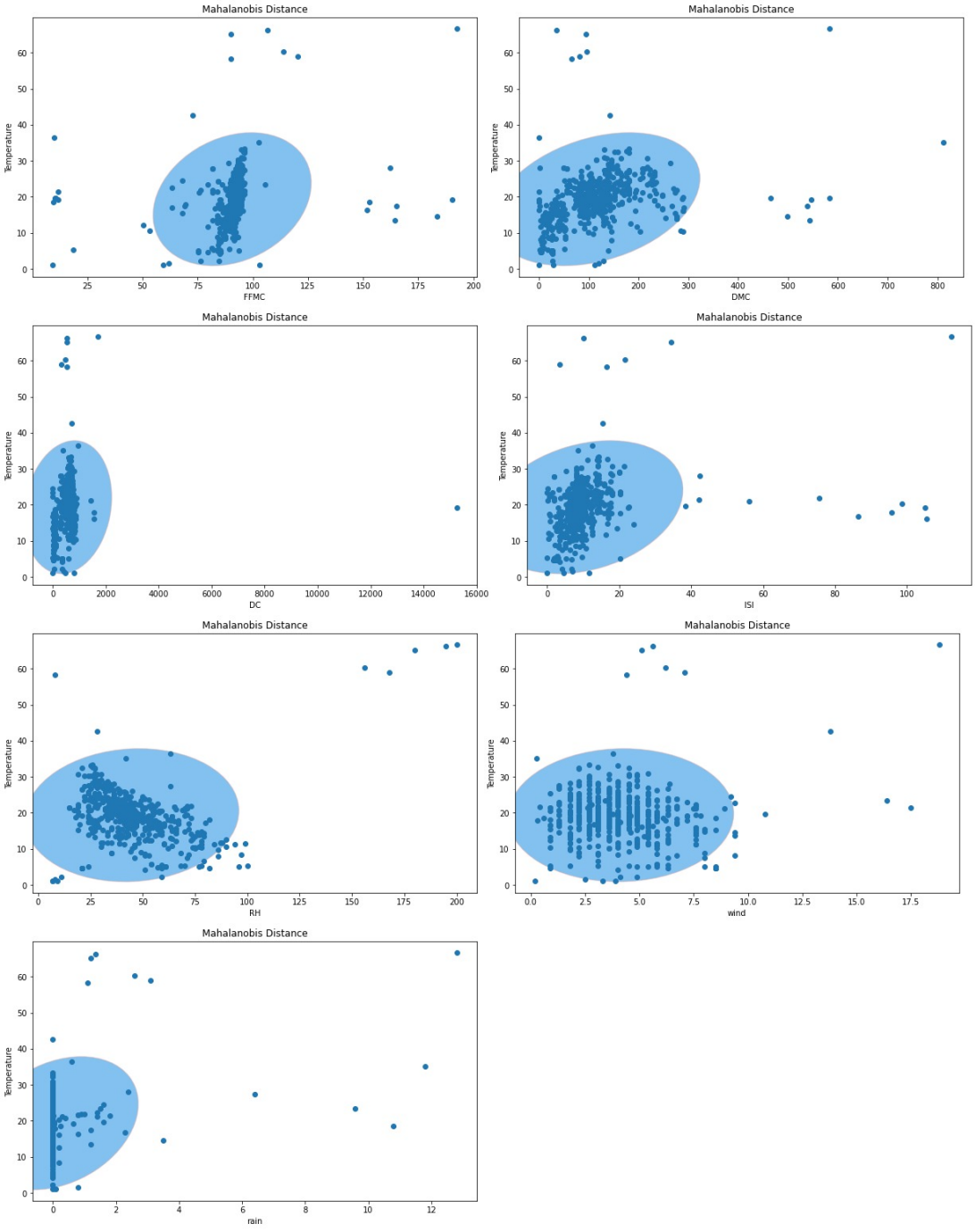
Σχήμα 4.46: Γραφήματα ακραίων τιμών Isolation Forest στο Forest Fires



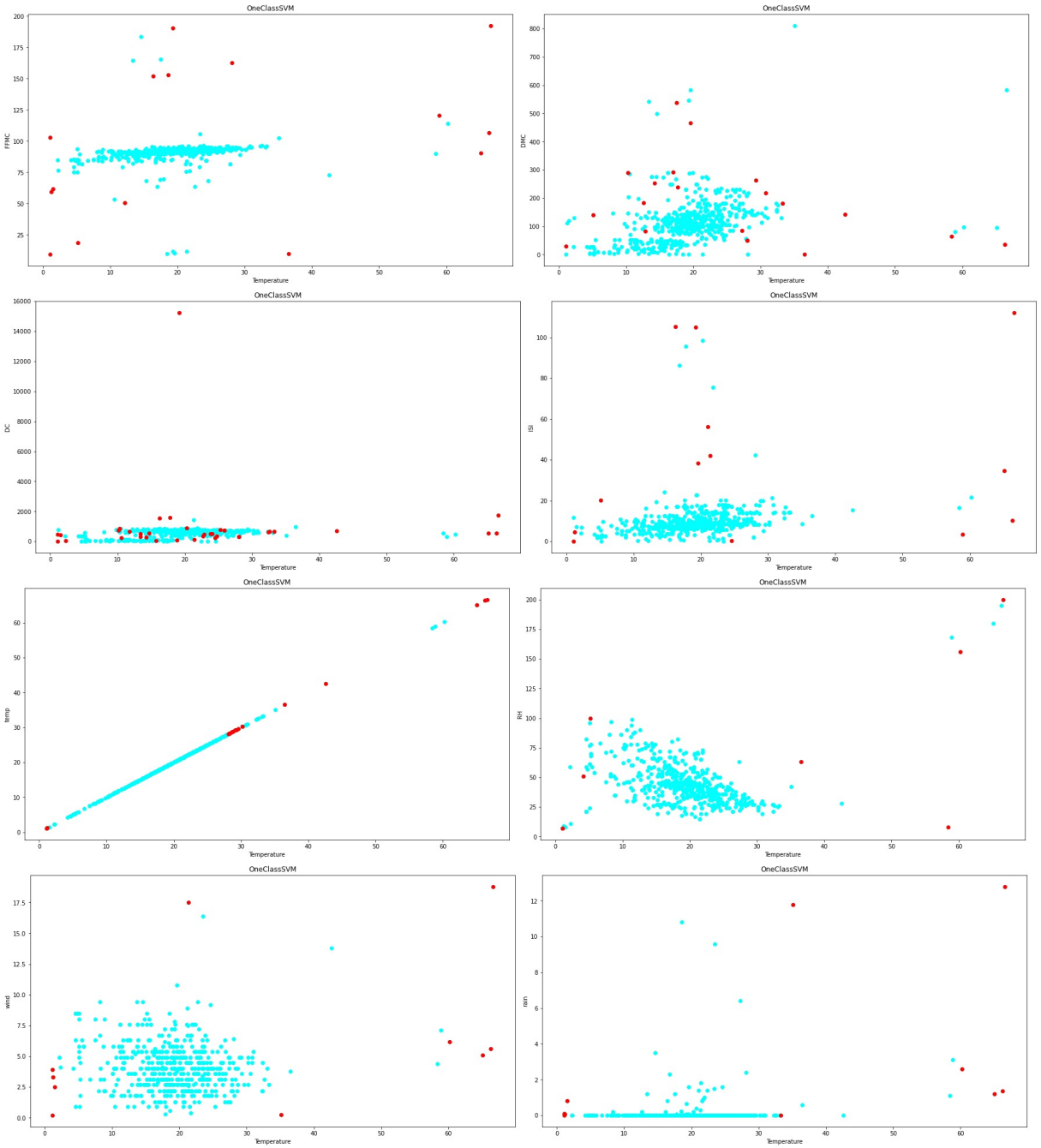
Σχήμα 4.47: Γραφήματα ακραίων τιμών Local Outlier Factor στο Forest Fires



Σχήμα 4.48: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Forest Fires



Σχήμα 4.49: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Forest Fires



4.7.3 QSAR Bioconcentration classes

Το σετ δεδομένων QSAR Bioconcentration έχει ως στόχο την πρόβλεψη τι είδους είναι κάθε χημική ουσία ανάλογα την τιμή που παίρνει, βασιζόμενη σε κάποιες μεταβλητές. Οι στήλες του σετ αυτού είναι οι: ID, nHM, piPC09, X2Av, MLOGP, ON1V, N-072, B02[C-N], F04[C-0] και Bioconcentration Factor(BCF) όπου είναι η στήλη στόχος και εμπεριέχει τις τιμές κάθε χημικής ουσίας. Στα Σχήματα 4.50, 4.51, 4.52, 4.53, 4.54, 4.55 και 4.56 παρουσιάζονται οι ακραίες τιμές (10%) που έχουν βρεθεί από τον κάθε αλγόριθμο. Επιπλέον στους Πίνακες 4.33 και 4.34 παρατηρούμε τα αποτελέσματα από τις προβλέψεις των αλγορίθμων με 10% ακραίες τιμές. Τα ποσοστά των αποτελεσμάτων είναι τα ιδανικά καθώς το μεγαλύτερο ποσοστό πρόβλεψης παρουσιάζεται στο αρχικό σετ δεδομένο, ακολουθούμενο από το ποσοστό του σετ αφού αφαιρεθούν οι ακραίες τιμές και τέλος το ποσοστό με τις ακραίες τιμές. Επίσης παρατηρούμε ότι ο GMM αλγόριθμος και στις δύο περιπτώσεις του Random Forest και SVM έχει το μεγαλύτερο ποσοστό ακριβείας αφού αφαιρεθούν οι ακραίες τιμές. Επιπλέον για άλλη μία φορά ο OneClassSVM έχει τη χαμηλότερη διακύμανση ποσοστού ακριβείας στα σετ δεδομένων αφού εισαχθούν και αφαιρεθούν οι ακραίες τιμές. Στους Πίνακες 4.31 και 4.32 παρουσιάζονται τα ποσοστά ακριβείας της πρόβλεψης με 5% ακραίες τιμές. Τα αποτελέσματα είναι ιδανικά εκτός του αλγορίθμου LOF στην περίπτωση Random Forest όπου τα ποσοστά ακριβείας είναι μεγαλύτερα με τις ακραίες τιμές στο σετ δεδομένων, από ότι αφού αφαιρεθούν αυτές. Επιπλέον όπως είναι φανερό στις περιπτώσεις που έχουμε 5% ακραίες τιμές τα ποσοστά είναι ως επί το πλείστον μεγαλύτερα από αυτά του 10%.

Πίνακας 4.31: Ποσοστά του QSAR classes για 5% με Random Forest

QSAR data set with RF	Original data	With 5% outliers	Without outliers
DBscan	81.46%	54.23%	75.60%
Elliptical Envelope	81.46%	58.18%	71.95%
GMM	81.46%	46.16%	71.31%
Isolation Forest	81.46%	54.76%	74.23%
LOF	81.46%	56.69%	56.51%
Mahalanobis distance	81.46%	47.80%	68.48%
One Class SVM	81.46%	47.62%	56.87%

Πίνακας 4.32: Ποσοστά του QSAR classes για 5% με Support Vector Machine

QSAR data set with SVM	Original data	With 5% outliers	Without outliers
DBscan	79.90%	51.56%	70.36%
Elliptical Envelope	79.90%	57.88%	70.06%
GMM	79.90%	48.12%	72.03%
Isolation Forest	79.90%	51%	73.77%
LOF	79.90%	48.25%	53.51%
Mahalanobis distance	79.90%	51.14%	68.40%
One Class SVM	79.90%	53.70%	56.51%

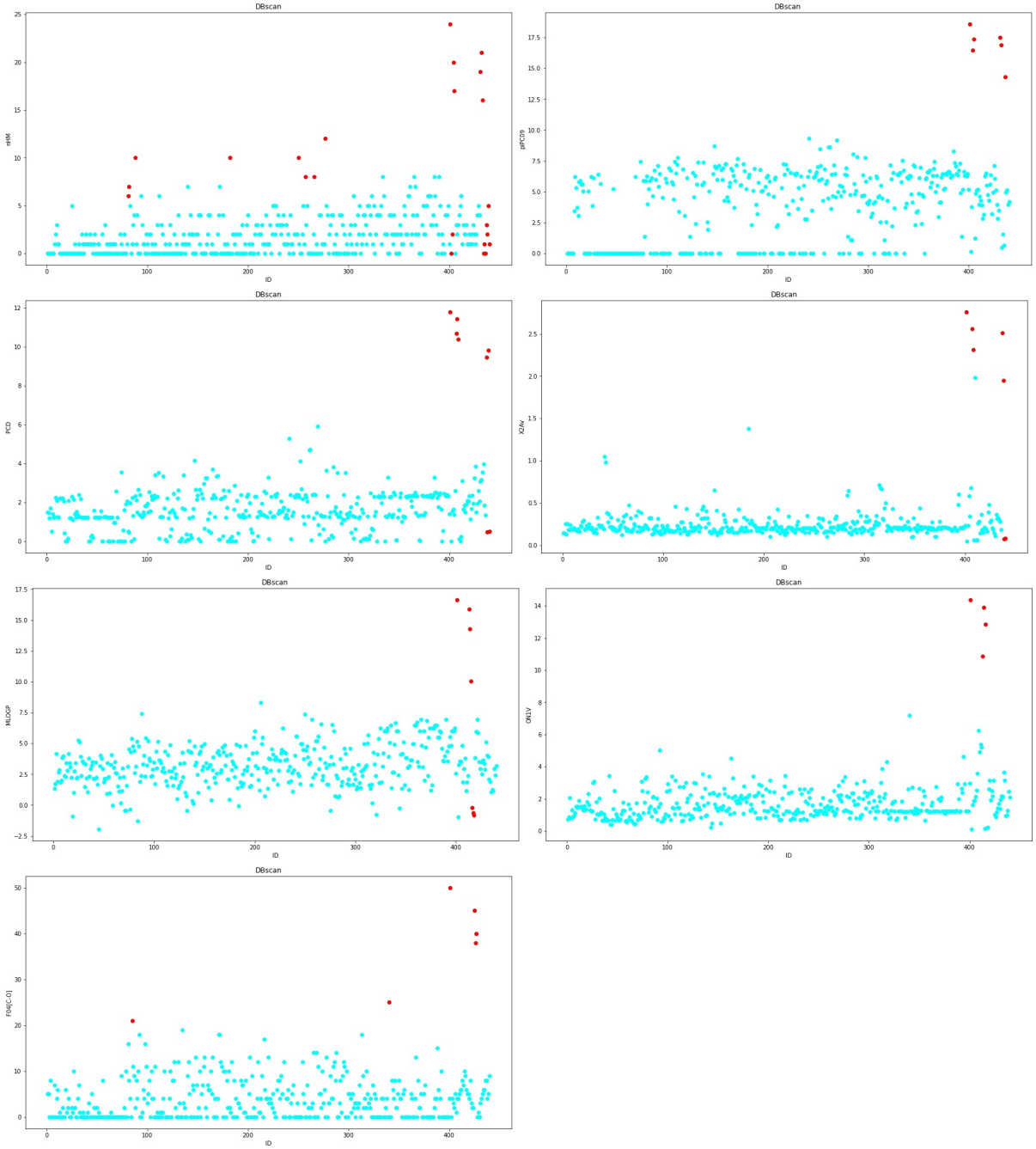
Πίνακας 4.33: Ποσοστά του QSAR classes για 10% με Random Forest

QSAR data set with RF	Original data	With 10% outliers	Without outliers
DBscan	81.46%	24.01%	59.74%
Elliptical Envelope	81.46%	26.23%	52.55%
GMM	81.46%	36.92%	62.68%
Isolation Forest	81.46%	33.81%	40.75%
LOF	81.46%	32.21%	56.43%
Mahalanobis distance	81.46%	26.98%	44.58%
One Class SVM	81.46%	34.97%	41.12%

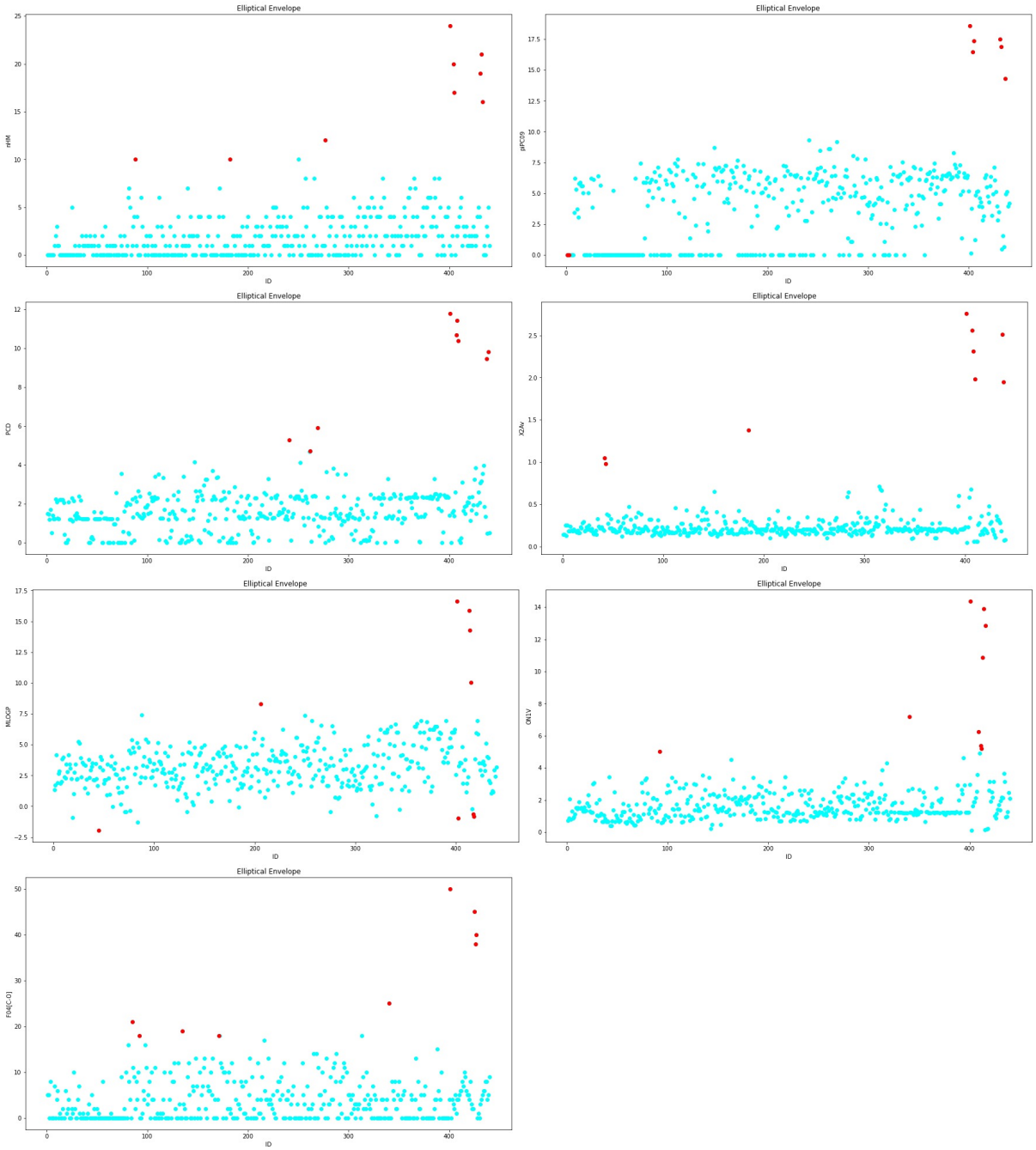
Πίνακας 4.34: Ποσοστά του QSAR classes για 10% με Support Vector Machine

QSAR data set with SVM	Original data	With 10% outliers	Without outliers
DBscan	79.90%	30.22%	57.42%
Elliptical Envelope	79.90%	32.26%	57.33%
GMM	79.90%	33.51%	60.01%
Isolation Forest	79.90%	34.46%	48.59%
LOF	79.90%	22.15%	56.09%
Mahalanobis distance	79.90%	34.85%	49.25%
One Class SVM	79.90%	23.42%	38.26%

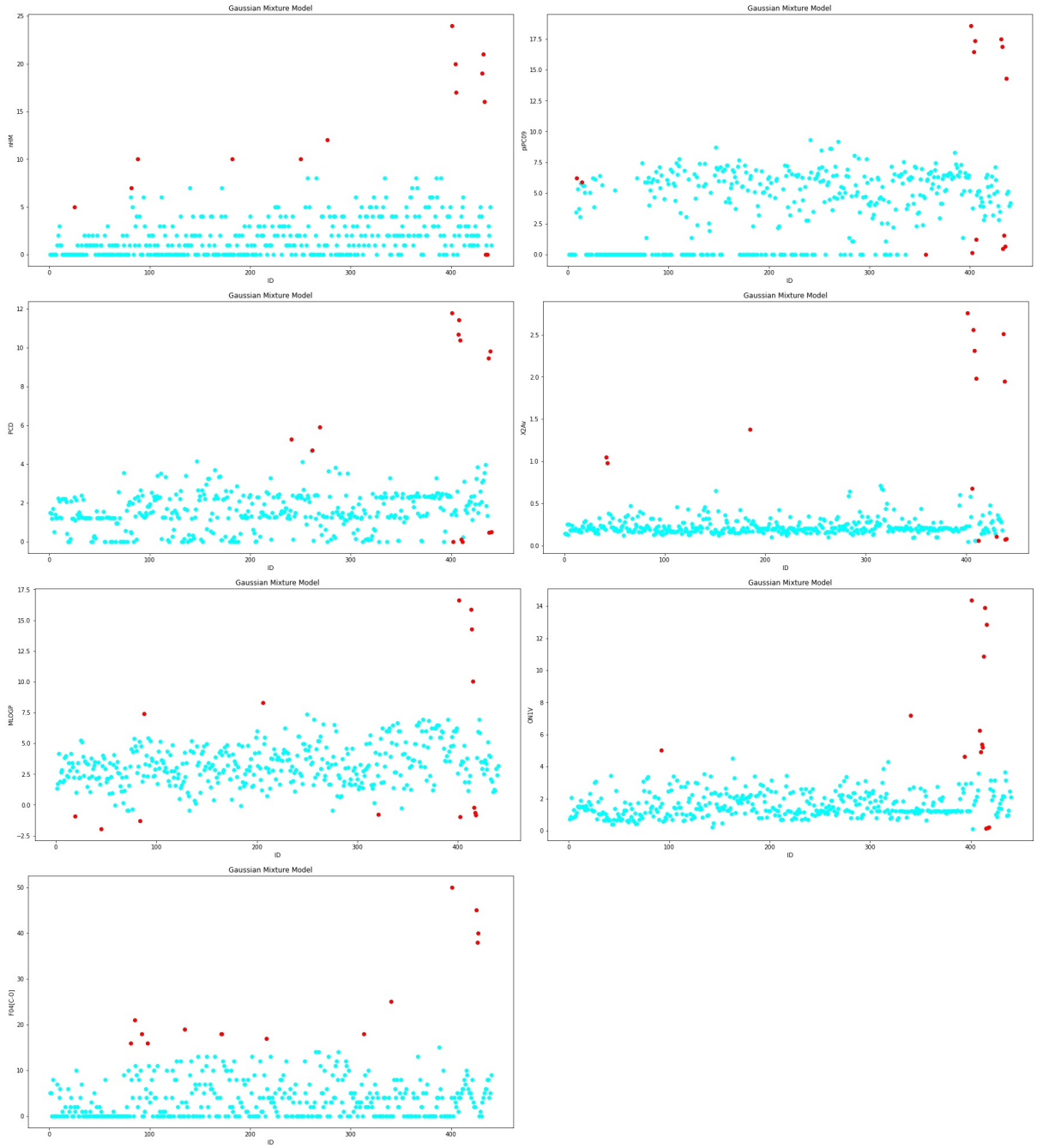
Σχήμα 4.50: Γραφήματα ακραίων τιμών DBscan στο QSAR Bioconcentration classes



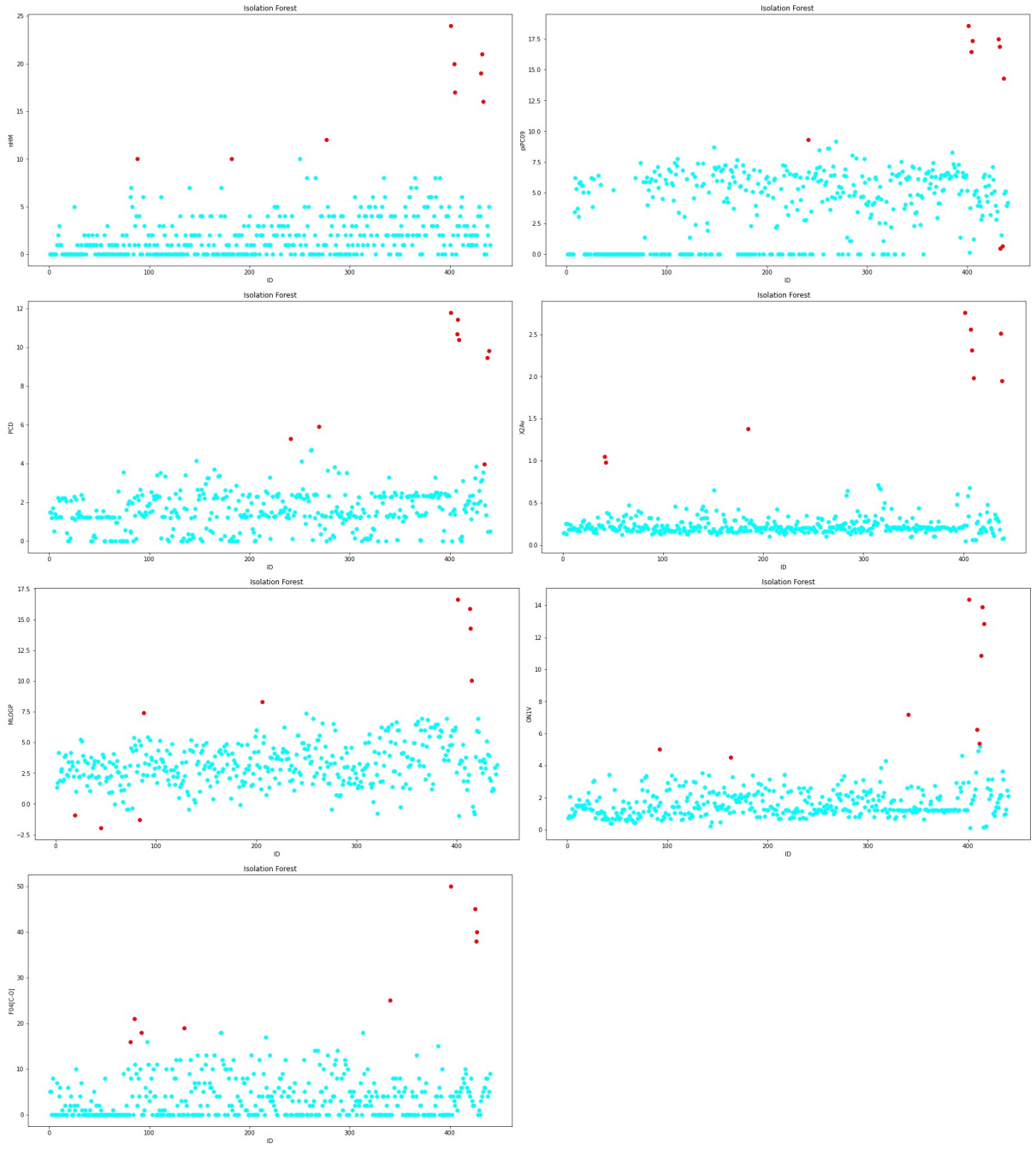
Σχήμα 4.51: Γραφήματα ακραίων τιμών Elliptical Envelope στο QSAR Bioconcentration classes



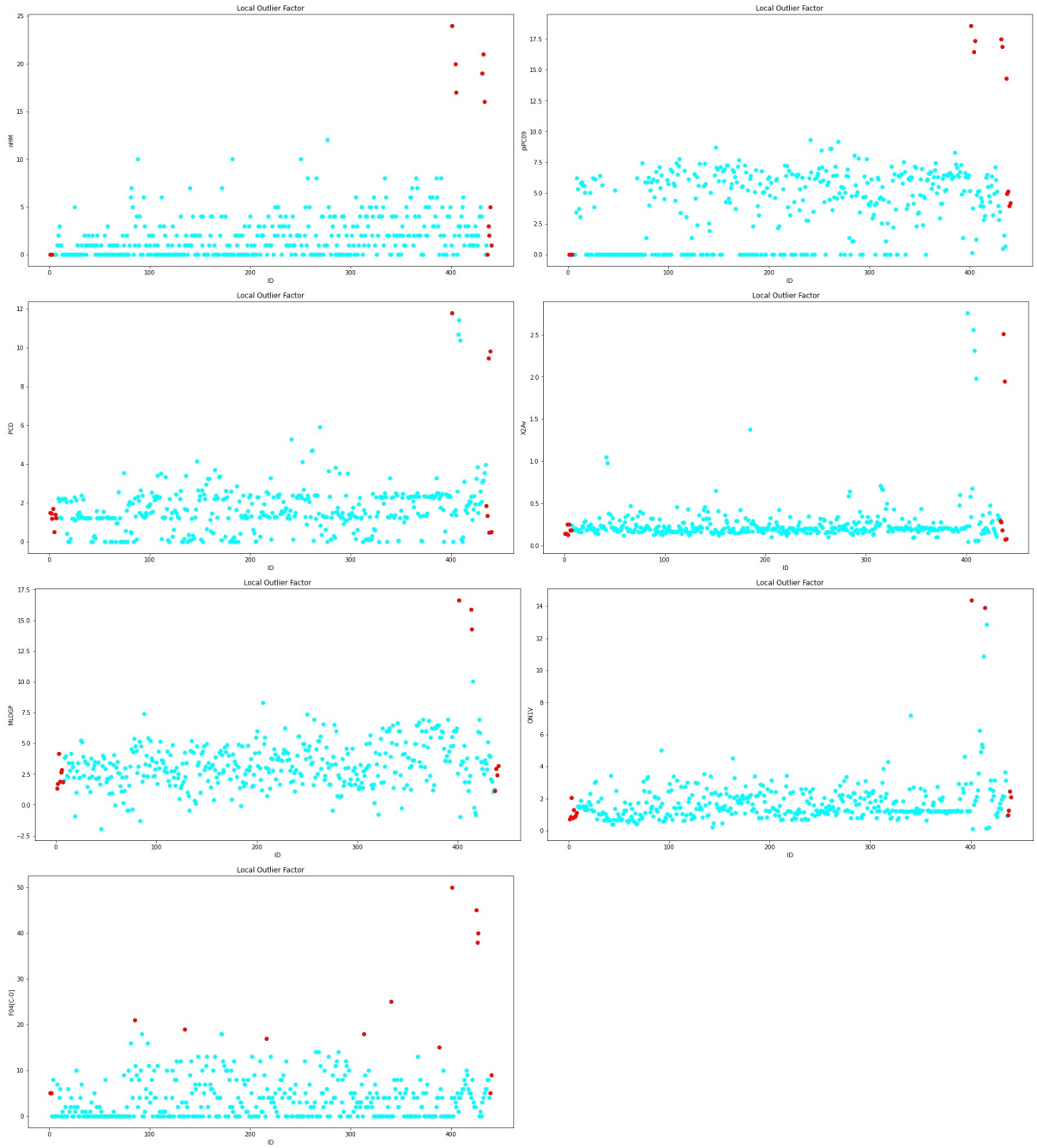
Σχήμα 4.52: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο QSAR Bioconcentration classes



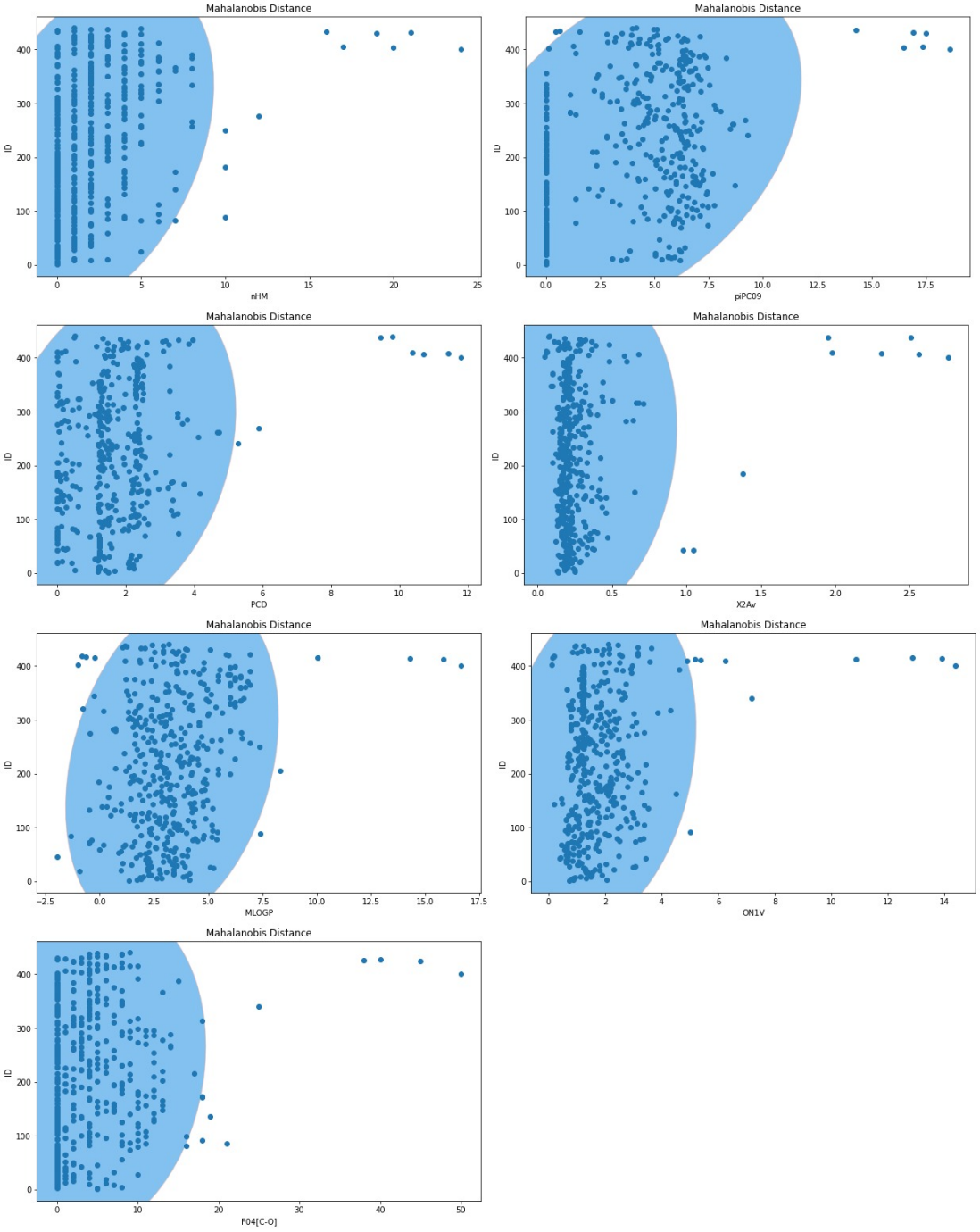
Σχήμα 4.53: Γραφήματα ακραίων τιμών Isolation Forest στο QSAR Bioconcentration classes



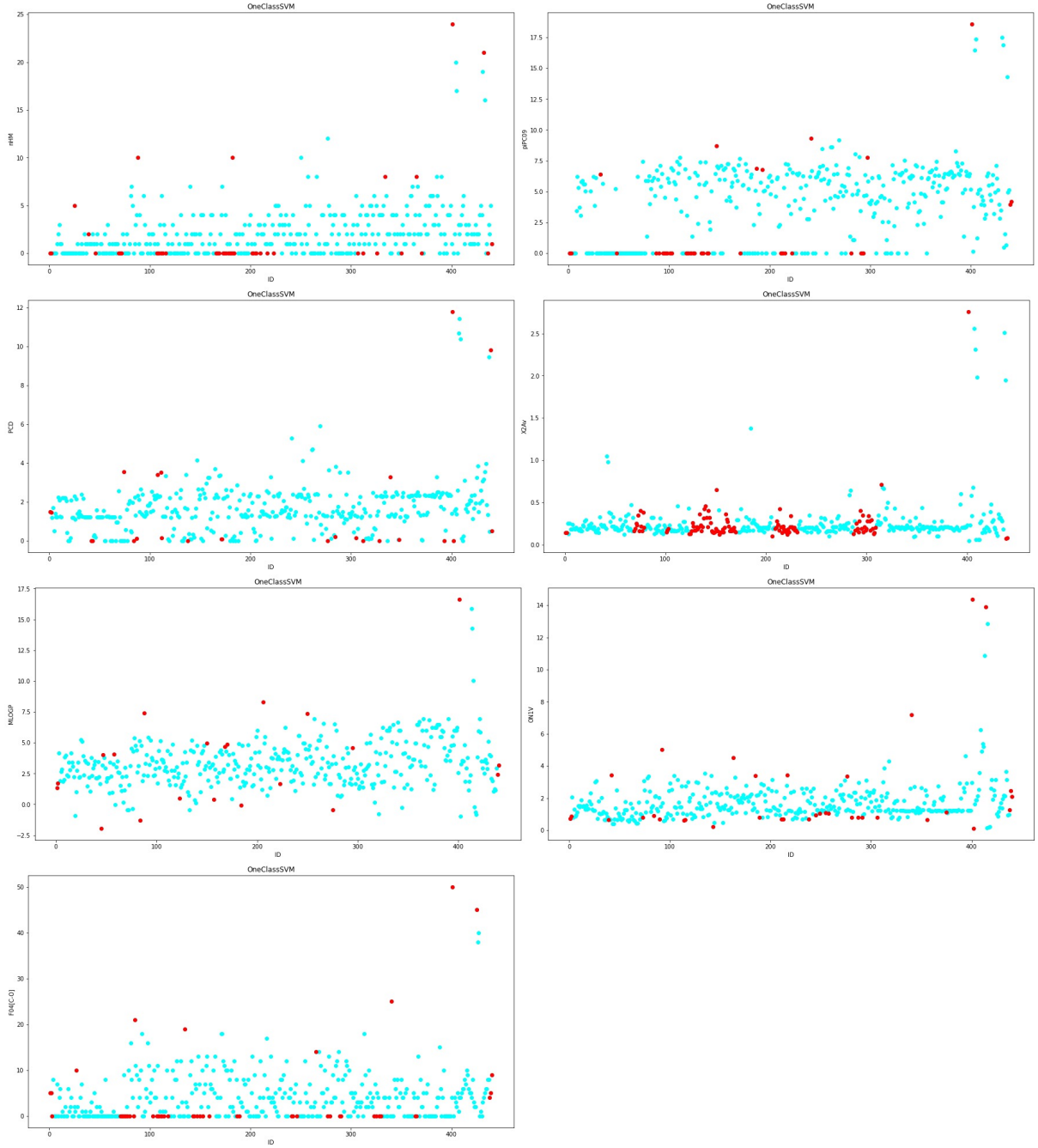
Σχήμα 4.54: Γραφήματα ακραίων τιμών Local Outlier Factor στο QSAR Bioconcentration classes



Σχήμα 4.55: Γραφήματα ακραίων τιμών Mahalanobis Distance στο QSAR Bioconcentration classes



Σχήμα 4.56: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο QSAR Bioconcentration classes



4.7.4 QSAR aquatic toxicity

Το σετ δεδομένων QSAR aquatic toxicity έχει ως στόχο την πρόβλεψη του ποσοστού τοξικότητας των υδάτων για τα ψάρια Pimephales promelas. Το σετ έχει ως μεταβλητές τις: ID, TPSA(Tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, C-040 και quantitative response (LC50) όπου είναι ο στόχος για την πρόβλεψη της τοξικότητας. Στα Σχήματα 4.57, 4.58, 4.59, 4.60, 4.61, 4.62 και 4.63 παρουσιάζονται οι ακραίες τιμές (10%) που ανιχνεύτηκαν από τον κάθε αλγόριθμο. Επίσης στους Πίνακες 4.33 και 4.34 παραθέτονται τα ποσοστά πρόβλεψης των αλγορίθμων Random Forest και SVM με 10% ακραίες τιμές. Με τον αλγόριθμο Random Forest όλα τα αποτελέσματα είναι επιθυμητά, ενώ στον SVM παρατηρούμε ότι τα ποσοστά απέχουν από τα ιδανικά, καθώς σε πολλές περιπτώσεις τα ποσοστά του σετ δεδομένων αφού αφαιρεθούν οι ακραίες τιμές, είναι μεγαλύτερα από του αρχικού σετ δεδομένων. Κάτι τέτοιο γίνεται λόγω της ιδιορρυθμίας του αλγορίθμου και της πολυπλοκότητας υπολογισμού της πρόβλεψης. Συνεπώς σε αυτό το σετ ο Random Forest παρουσιάζει καλύτερα αποτελέσματα. Στους Πίνακες 4.35 και 4.36 παρατηρούμε τα ποσοστά των αλγορίθμων με 5% ακραίες τιμές. Σε όλες τις περιπτώσεις των αλγορίθμων με Random Forest, εκτός του LOF, τα αποτελέσματα είναι τα επιθυμητά. Σε αντίθεση ο SVM παρουσιάζει αποκλίσεις από τα επιθυμητά αποτελέσματα, καθώς τα ποσοστά του σετ δεδομένων αφού έχουν αφαιρεθεί οι ακραίες τιμές είναι μεγαλύτερα από αυτά του αρχικού σετ. Όσο αφορά τον Random Forest δεν υπάρχουν μεγάλες μεταβολές μεταξύ των ποσοστών 5% και 10%. Σε αντίθεση στον SVM αλγόριθμο παρατηρούμε ότι τα ποσοστά με 5% ακραίες τιμές είναι υψηλότερα από αυτά του 10%.

Πίνακας 4.35: Ποσοστά του Aquatic toxicity για 5% με Random Forest

Aquatic data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	55.04%	36.62%	52.08%
Elliptical Envelope	55.04%	34.40%	38.04%
GMM	55.04%	29.27%	47.10%
Isolation Forest	55.04%	35.57%	49.45%
LOF	55.04%	31.61%	52.60%
Mahalanobis distance	55.04%	33.46%	46.76%
One Class SVM	55.04%	32.82%	42.49%

Πίνακας 4.36: Ποσοστά του Aquatic toxicity για 5% με Support Vector Machine

Aquatic data set with SVM	Original data	With 5% outliers	Without outliers
DBscan	11.29%	8.06%	11.73%
Elliptical Envelope	11.29%	7.15%	11.73%
GMM	11.29%	4.90%	11.89%
Isolation Forest	11.29%	5.57%	9.52%
LOF	11.29%	5.42%	16.26%
Mahalanobis distance	11.29%	8.68%	10.98%
One Class SVM	11.29%	5.98%	5.08%

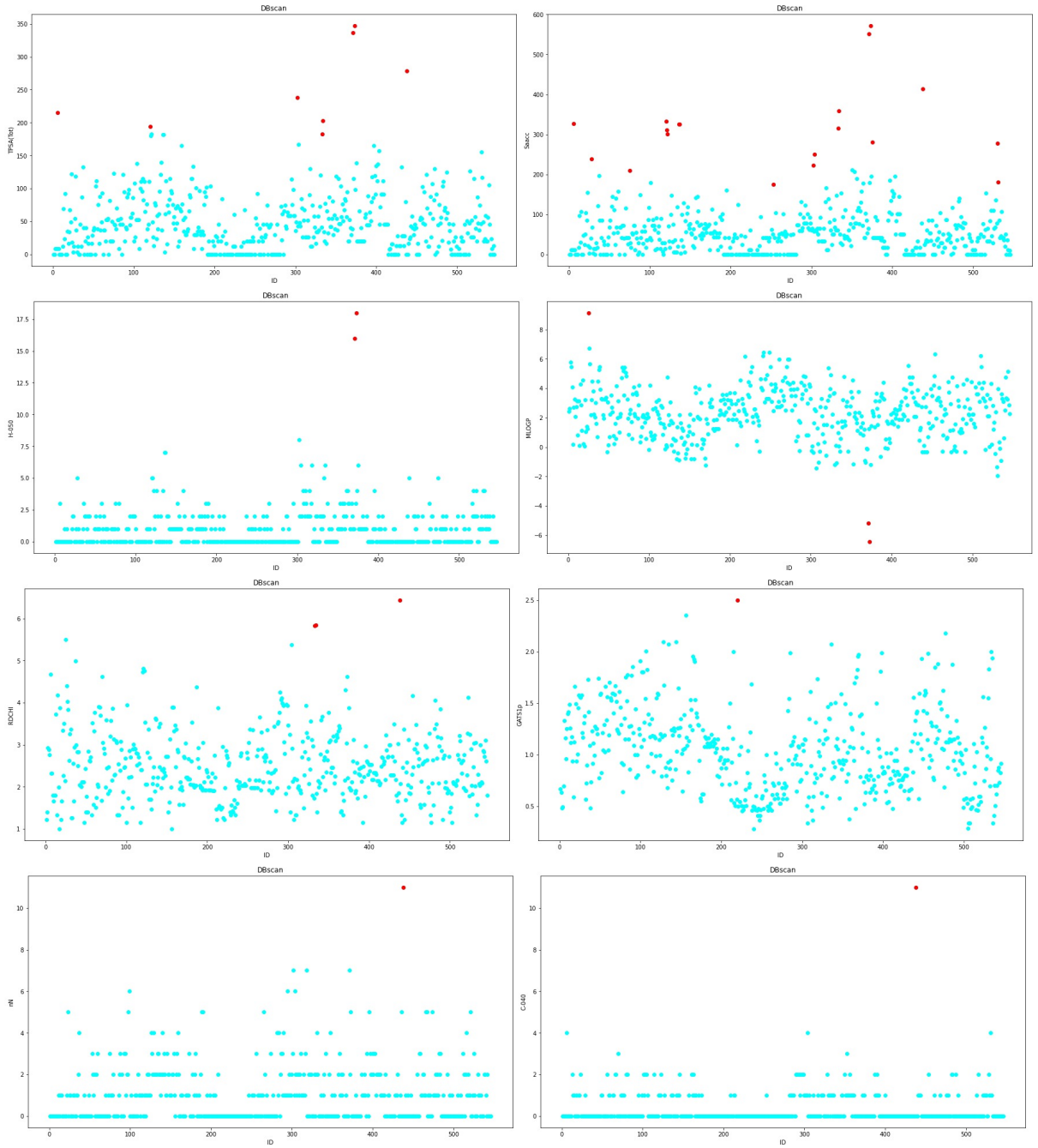
Πίνακας 4.37: Ποσοστά του Aquatic toxicity για 10% με Random Forest

Aquatic data set with RF	Original data	With 10% outliers	Without outliers
DBscan	55.04%	34.14%	46.82%
Elliptical Envelope	55.04%	38.70%	51.73%
GMM	55.04%	28.37%	50.16%
Isolation Forest	55.04%	35.62%	49.38%
LOF	55.04%	28.99%	40.47%
Mahalanobis distance	55.04%	40.15%	48.99%
One Class SVM	55.04%	24.20%	29.26%

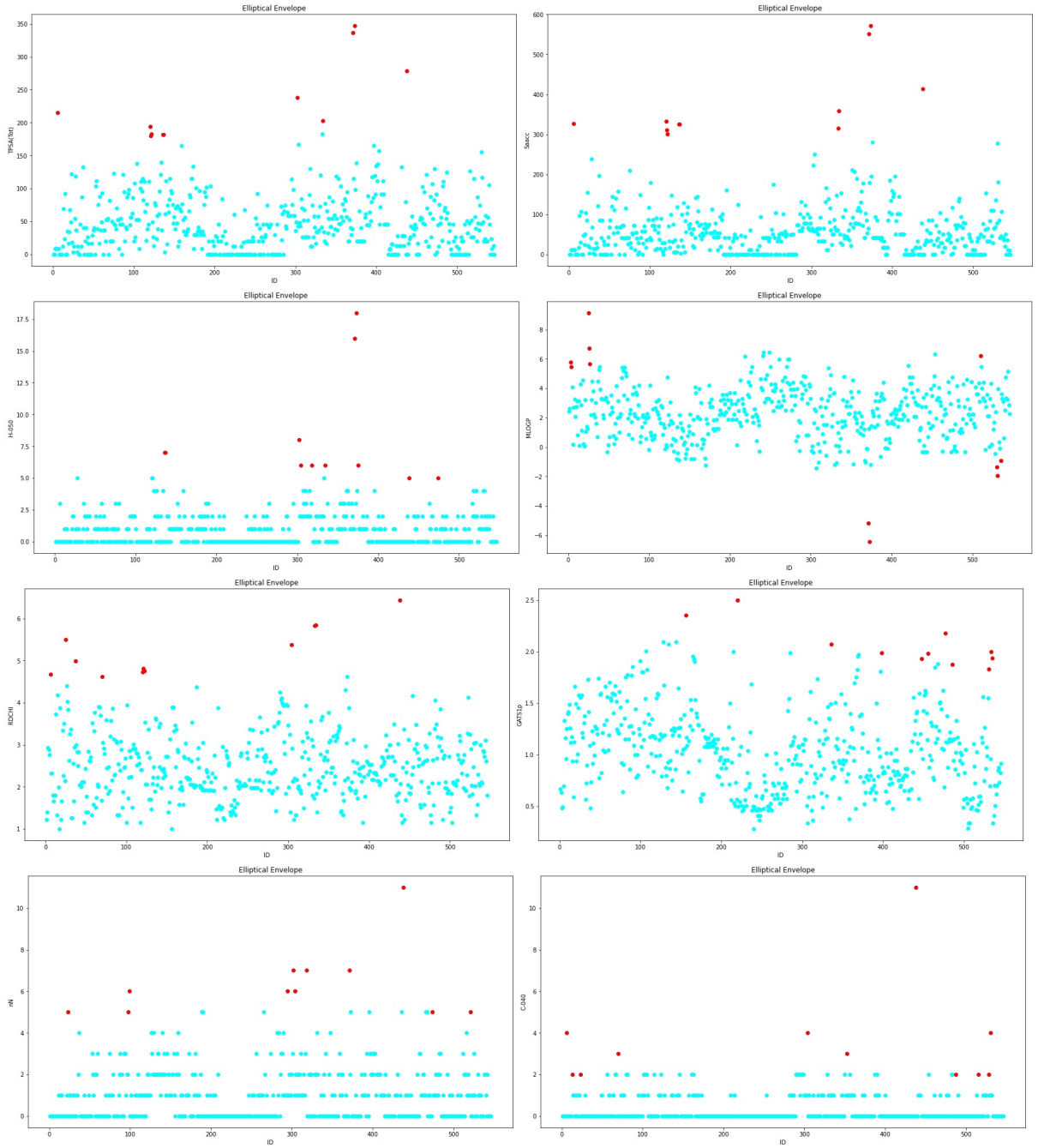
Πίνακας 4.38: Ποσοστά του Aquatic toxicity για 10% με Support Vector Machine

Aquatic data set with SVM	Original data	With 10% outliers	Without outliers
DBscan	11.29%	3.10%	11.75%
Elliptical Envelope	11.29%	5.13%	11.26%
GMM	11.29%	4.62%	13.15%
Isolation Forest	11.29%	4.71%	13.68%
LOF	11.29%	3.28%	11.78%
Mahalanobis distance	11.29%	6.64%	14.91%
One Class SVM	11.29%	3.73%	3.76%

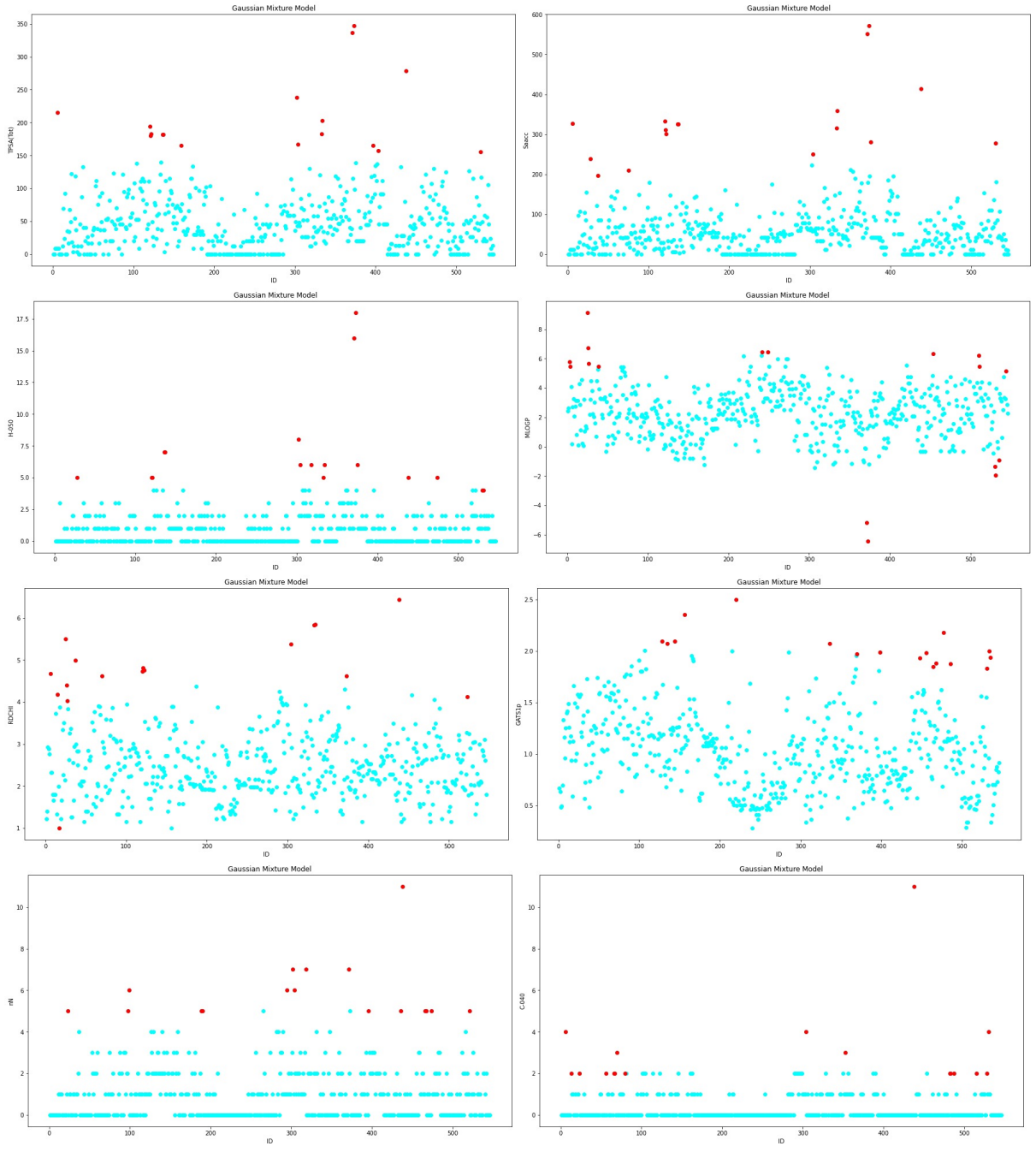
Σχήμα 4.57: Γραφήματα ακραίων τιμών DBscan στο Aquatic toxicity



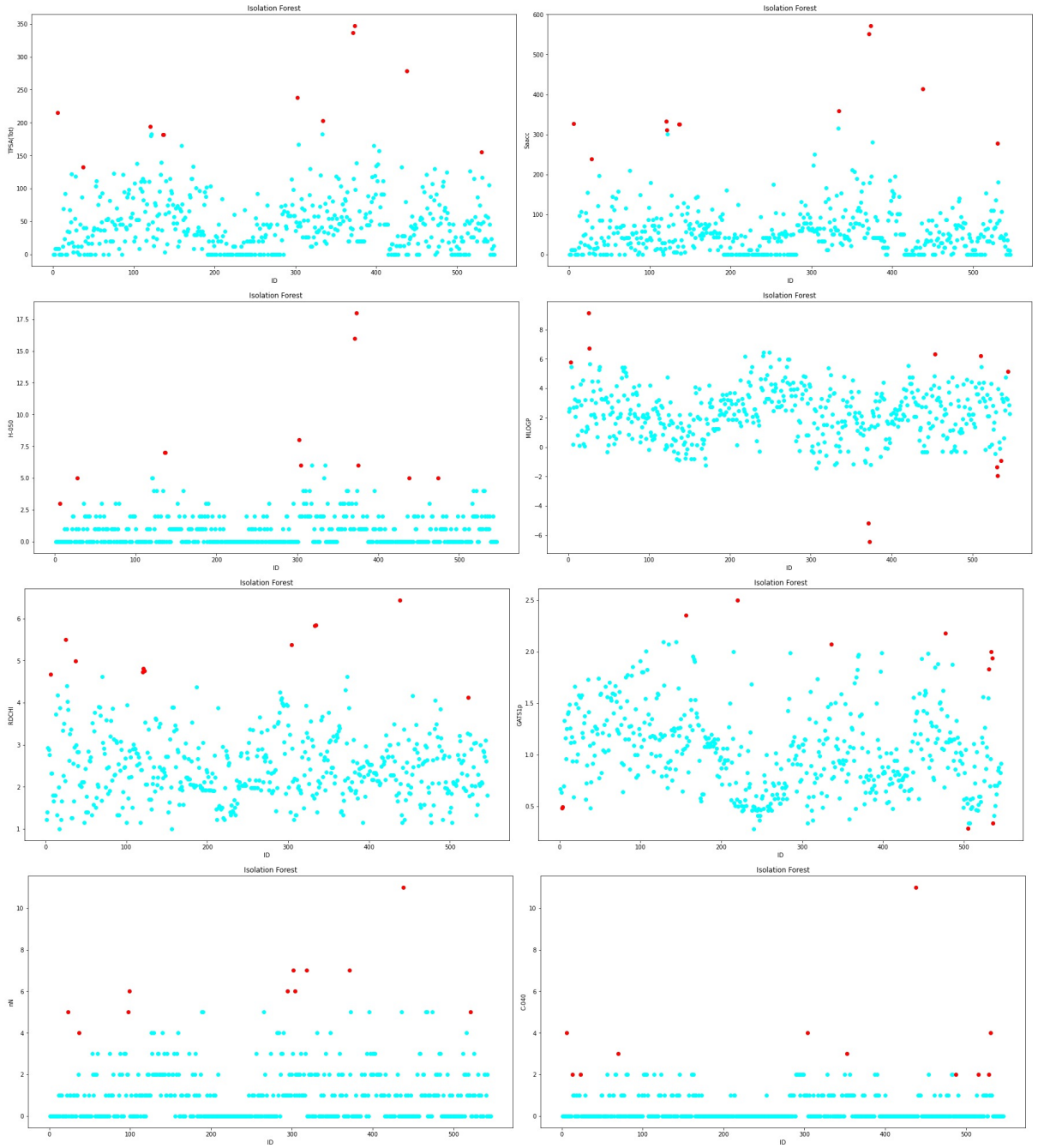
Σχήμα 4.58: Γραφήματα ακραίων τιμών Elliptical Envelope στο Aquatic toxicity



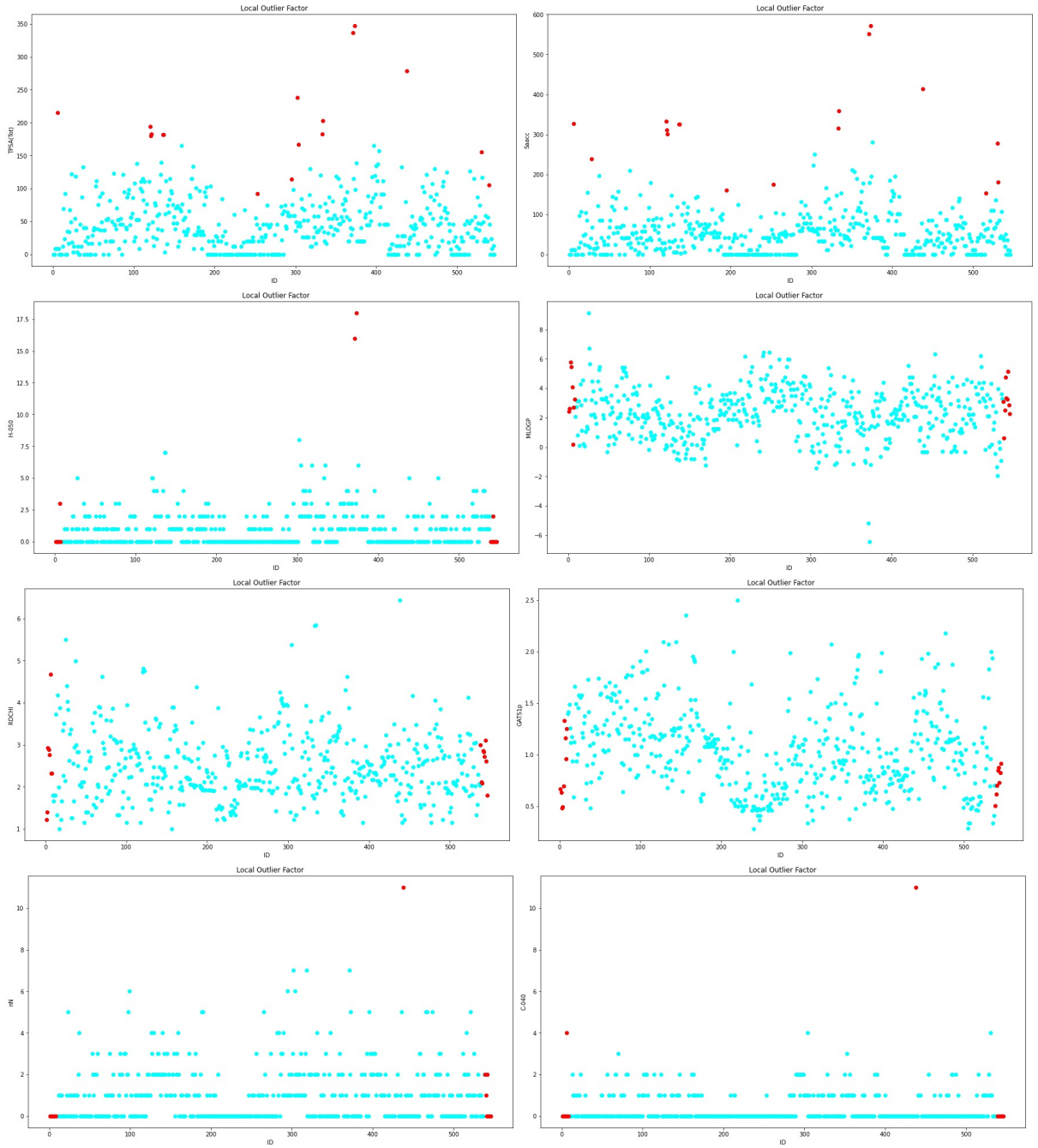
Σχήμα 4.59: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Aquatic toxicity



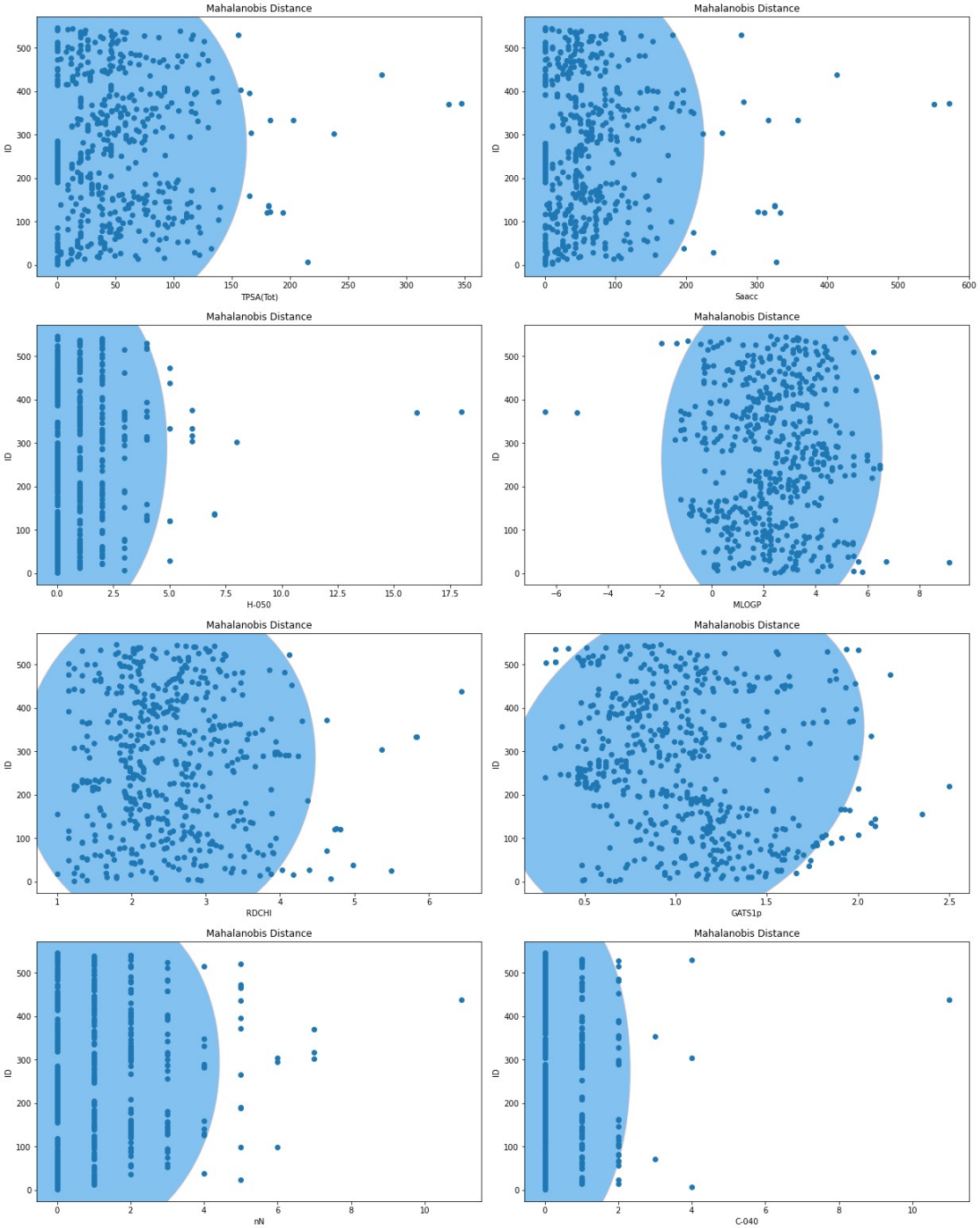
Σχήμα 4.60: Γραφήματα ακραίων τιμών Isolation Forest στο Aquatic toxicity



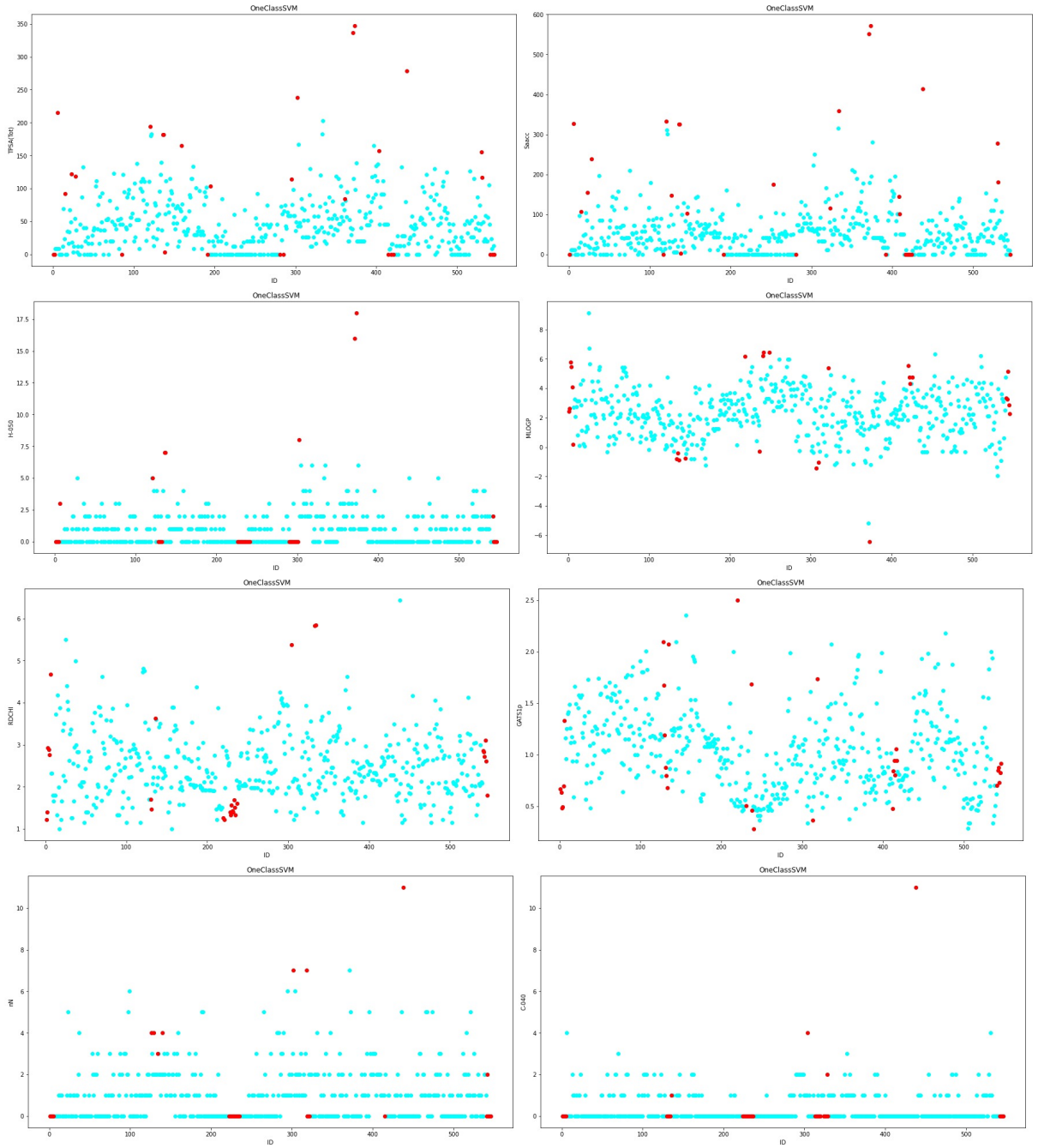
Σχήμα 4.61: Γραφήματα ακραίων τιμών Local Outlier Factor στο Aquatic toxicity



Σχήμα 4.62: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Aquatic toxicity



Σχήμα 4.63: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Aquatic toxicity



4.7.5 Productivity Prediction of Garment Employees

Το σετ δεδομένων Productivity Prediction of Garment Employees αφορά την πρόβλεψη της απόδοσης των υπαλλήλων μίας εταιρείας ρούχων. Οι μεταβλητές που έχει το σετ είναι οι: ID, quarter, day, team, targeted productivity, smv, wip, over time, incentive, idle time, idle men, number of style change, number of workers και actual productivity που είναι η μεταβλητή στόχος. Στα Σχήματα 4.64, 4.65, 4.66, 4.67, 4.68, 4.69 και 4.70 παρουσιάζονται οι ακραίες τιμές (10%) που βρίσκει ο εκάστοτε αλγόριθμος. Επιπροσθέτως στους Πίνακες 4.41 και 4.42 αναφέρονται τα ποσοστά ακριβείας της πρόβλεψης για τους αλγορίθμους μηχανικής μάθησης Random Forest και SVM αντίστοιχα με 10% ακραίες τιμές. Όπως είναι αντιληπτό τα δεδομένα και στις δύο περιπτώσεις είναι ιδανικά και συμπίπτουν με τα αναμενόμενα. Παρ' όλ' αυτά τα ποσοστά του Random Forest είναι πολύ υψηλότερα από του SVM κάτι που οφείλεται στην ιδιαιτερότητα του δεύτερου αλγορίθμου. Στους Πίνακες 4.39 και 4.40 παρουσιάζονται τα αποτελέσματα για 5% ακραίες τιμές. Παρατηρούμε ότι στην περίπτωση που χρησιμοποιείται ο αλγόριθμος Random Forest τα ποσοστά με 5% ακραίες τιμές είναι υψηλότερα από αυτά του 10%, ενώ αντίθετα με τη χρήση του SVM αλγορίθμου τα ποσοστά που παρουσιάζει το σετ δεδομένων με 5% ακραίες τιμές είναι χαμηλότερα από αυτά του 10%. Παρ' όλ' αυτά και στις δύο περιπτώσεις του Random Forest και του SVM τα αποτελέσματα είναι ιδανικά. Η παρουσία αρνητικών τιμών στα ποσοστά του SVM οφείλεται στην ιδιαιτερότητα του σετ δεδομένων, καθώς και στην πολυπλοκότητα του αλγορίθμου.

Πίνακας 4.39: Ποσοστά του Productivity για 5% με Random Forest

Productivity data set with RF	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	87.60%	52.86%	83.44%
Elliptical Envelope	87.60%	58.27%	85.85%
GMM	87.60%	55.16%	84.30%
Isolation Forest	39.64%	31.60%	85.13%
LOF	87.60%	56.34%	64.77%
Mahalanobis distance	87.60%	63.47%	84.90%
One Class SVM	87.60%	57.96%	73.28%

Πίνακας 4.40: Ποσοστά του Productivity για 5% με Support Vector Machine

Productivity data set with SVM	**Original data**	**With 5% outliers**	**Without outliers**
DBscan	10.19%	-6.64%	8.01%
Elliptical Envelope	10.19%	-3.40%	7.84%
GMM	10.19%	-3.89%	7.75%
Isolation Forest	10.19%	-4.45%	4.42%
LOF	10.19%	-3.51%	2.47%
Mahalanobis distance	10.19%	-4.86%	8.02%
One Class SVM	10.19%	-4.71%	8.61%

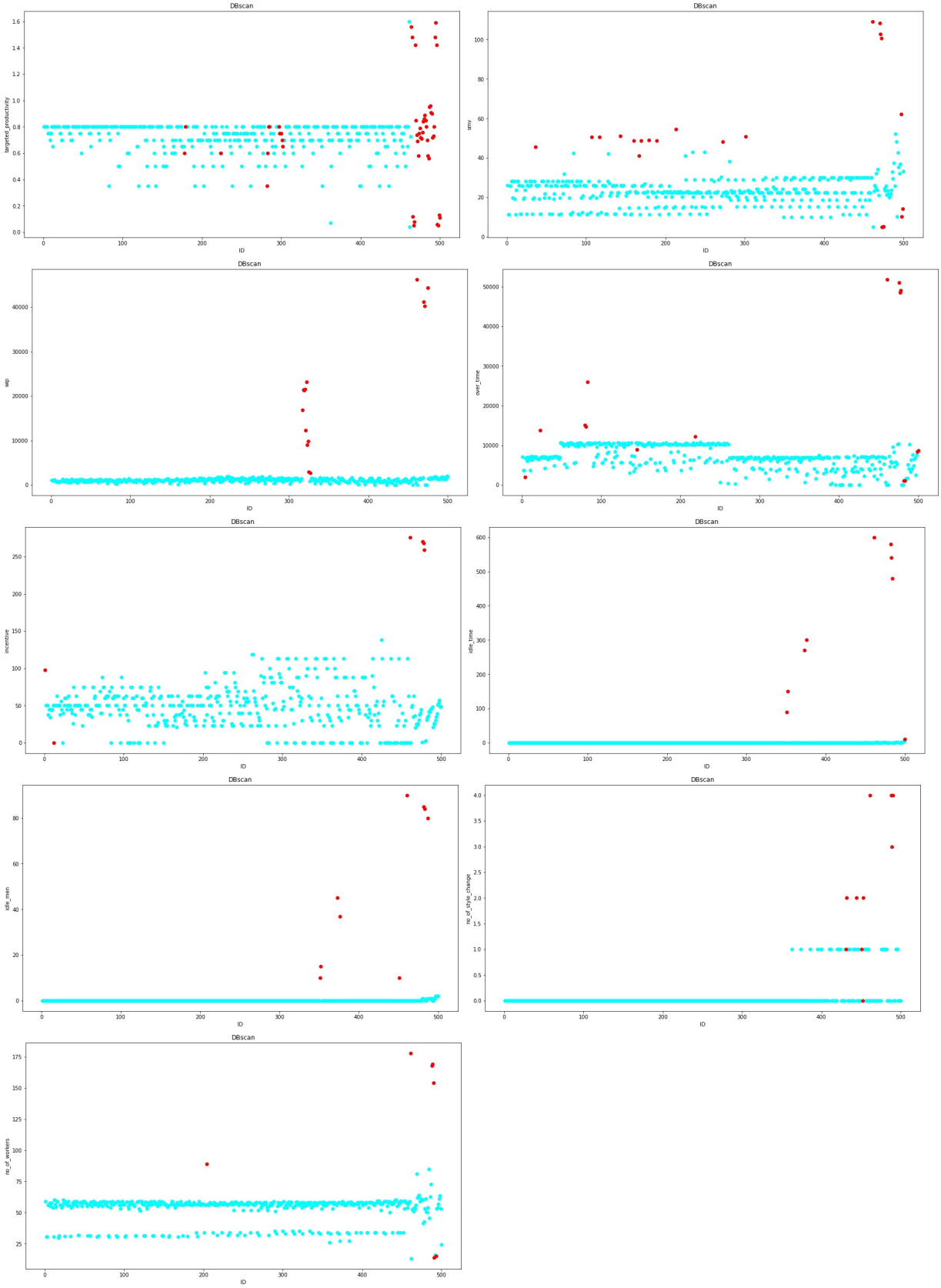
Πίνακας 4.41: Ποσοστά του Productivity για 10% με Random Forest

Productivity data set with RF	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	87.60%	43.41%	84.22%
Elliptical Envelope	87.60%	41.17%	61.90%
GMM	87.60%	38.61%	79.56%
Isolation Forest	87.60%	31.60%	59.94%
LOF	87.60%	37.59%	64.29%
Mahalanobis distance	87.60%	41.43%	82.12%
One Class SVM	87.60%	44.43%	50.87%

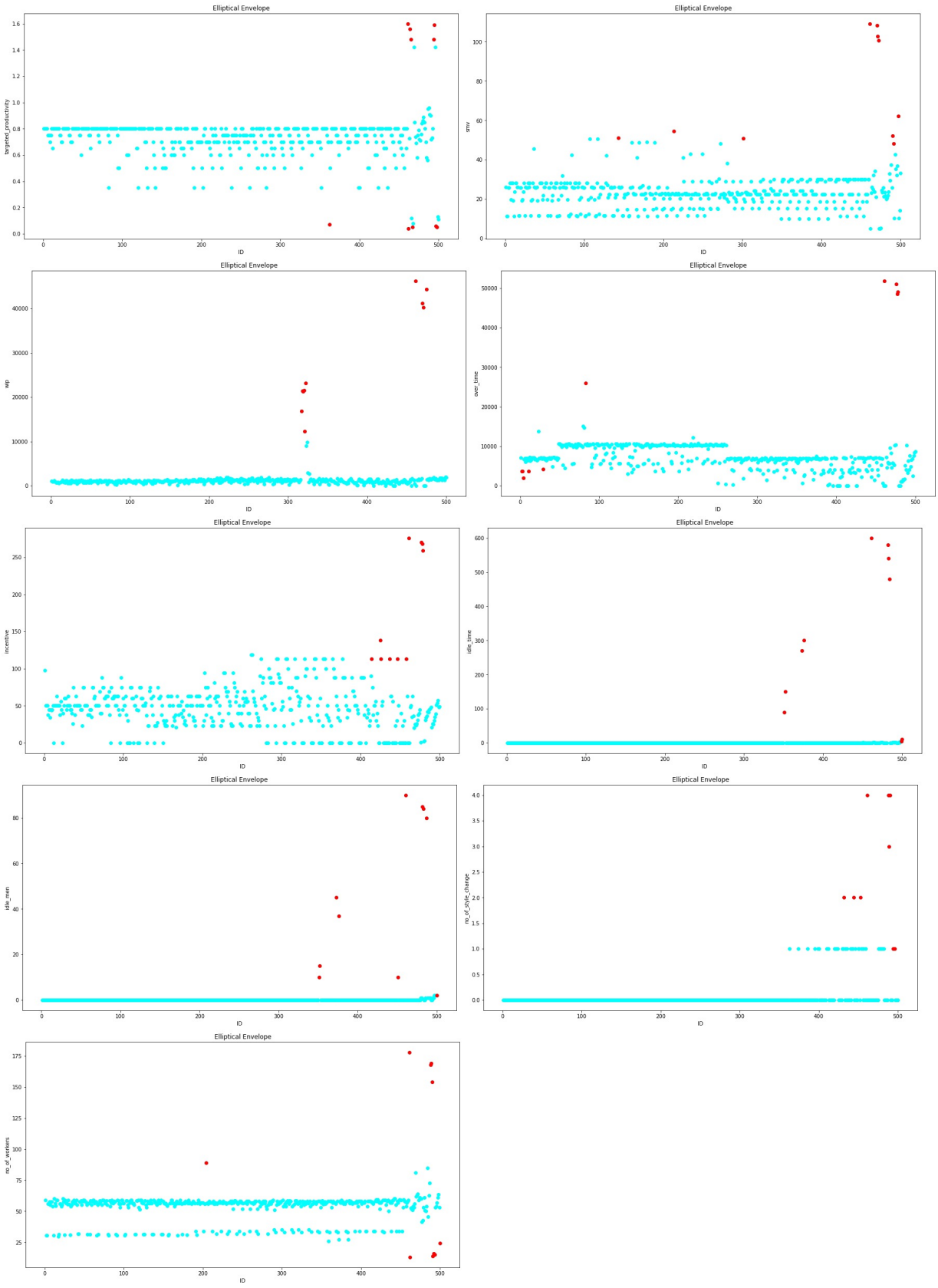
Πίνακας 4.42: Ποσοστά του Productivity για 10% με Support Vector Machine

Productivity data set with SVM	**Original data**	**With 10% outliers**	**Without outliers**
DBscan	10.19%	-3.86%	8.01%
Elliptical Envelope	10.19%	-3.82%	3.42%
GMM	10.19%	-0.03%	4.24%
Isolation Forest	10.19%	-3.63%	1.64%
LOF	10.19%	-3.71%	2.08%
Mahalanobis distance	10.19%	-3.74%	7.52%
One Class SVM	10.19%	-3.82%	3.09%

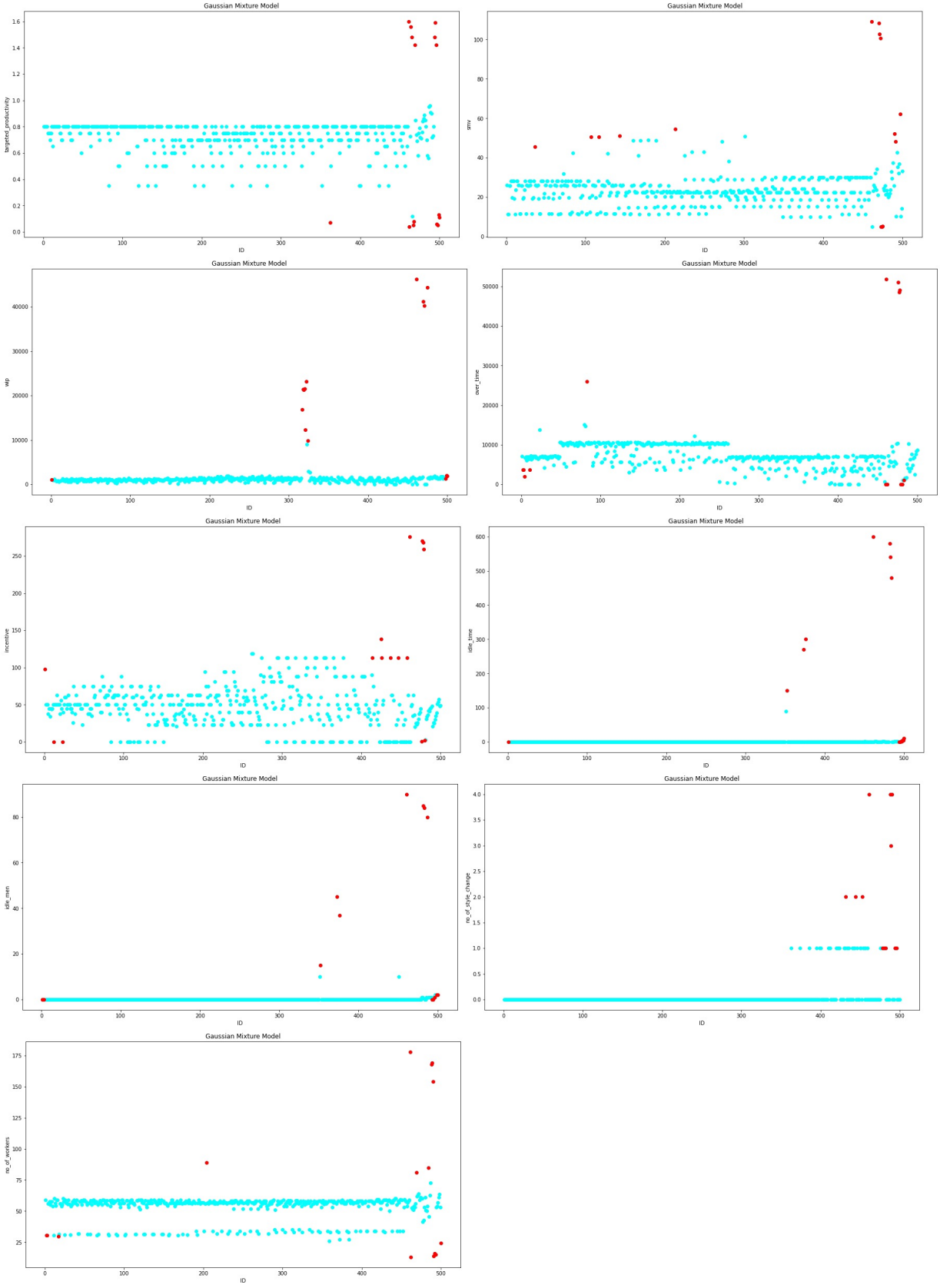
Σχήμα 4.64: Γραφήματα ακραίων τιμών DBscan στο Productivity Prediction of Garment Employees



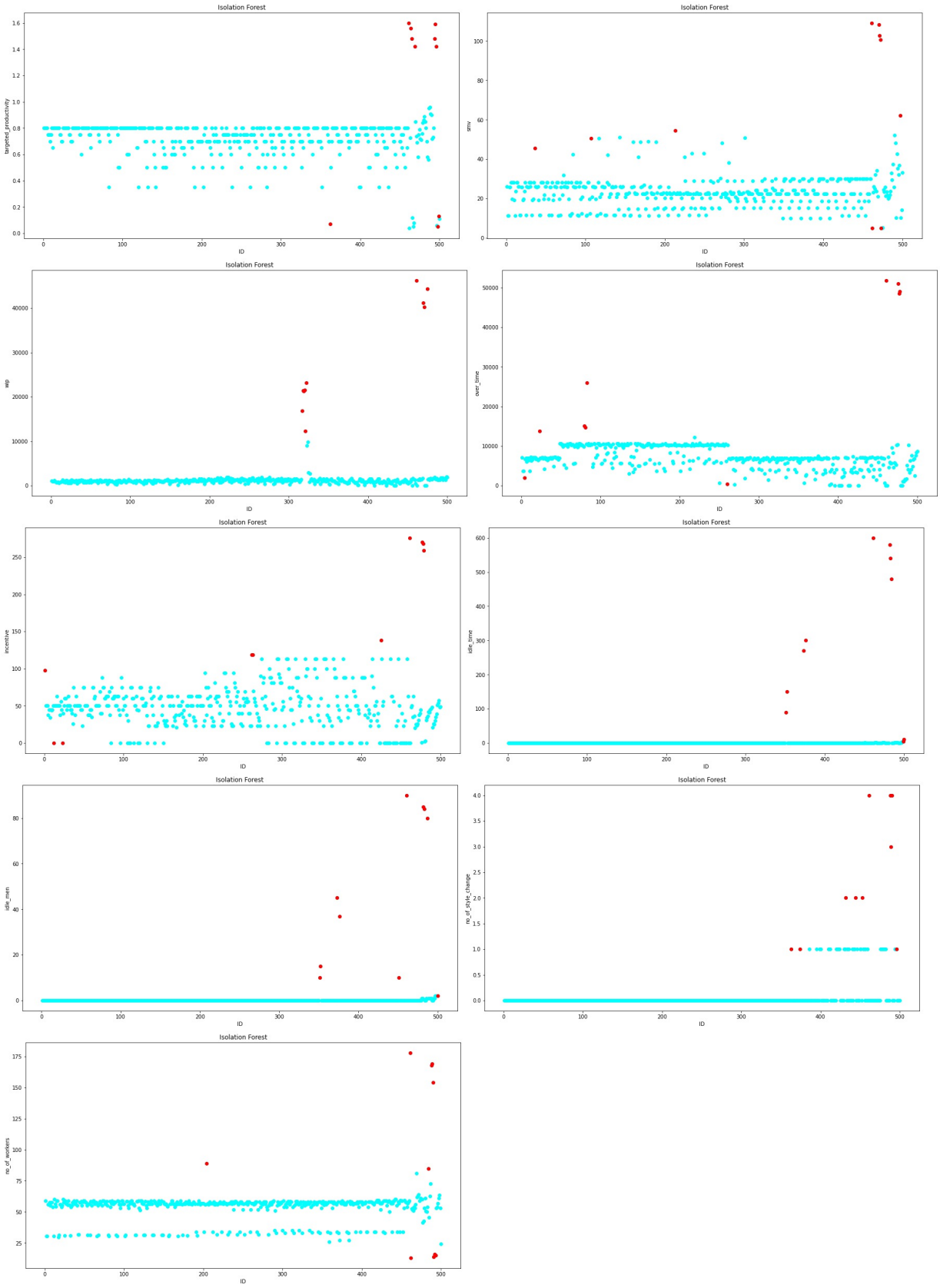
Σχήμα 4.65: Γραφήματα ακραίων τιμών Elliptical Envelope στο Productivity Prediction of Garment Employees



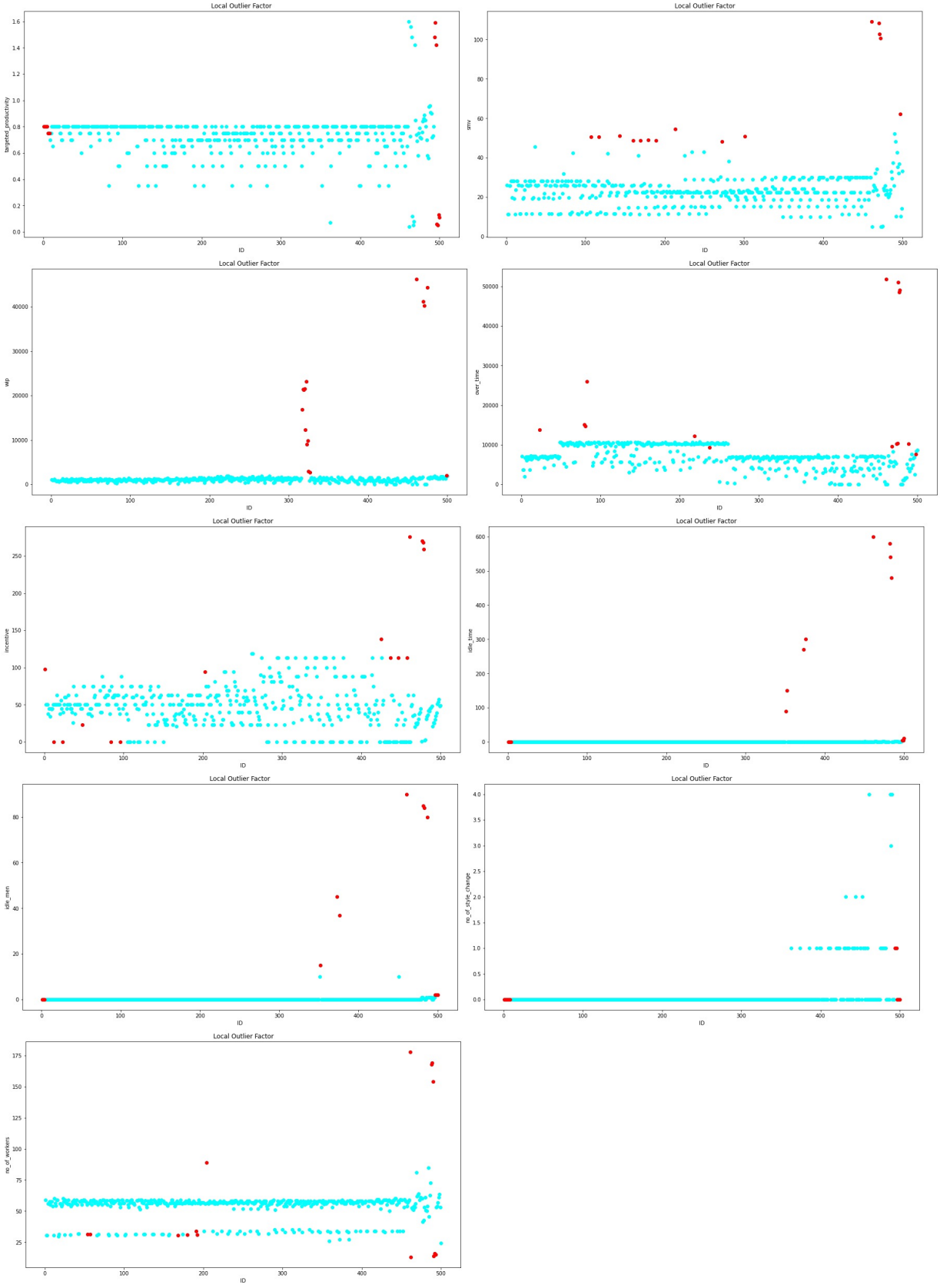
Σχήμα 4.66: Γραφήματα ακραίων τιμών Gaussian Mixture Model στο Productivity Prediction of Garment Employees



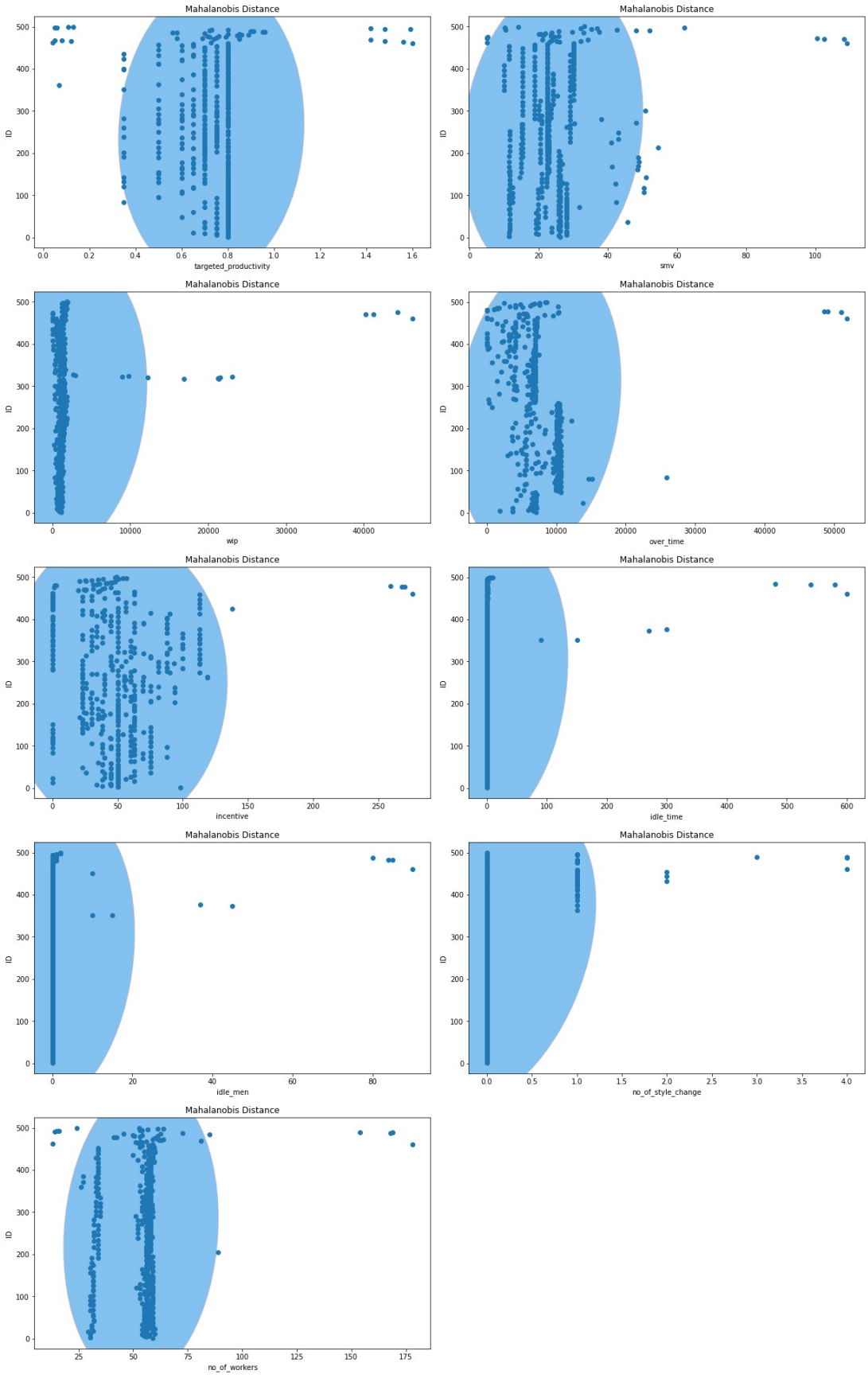
Σχήμα 4.67: Γραφήματα ακραίων τιμών Isolation Forest στο Productivity Prediction of Garment Employees



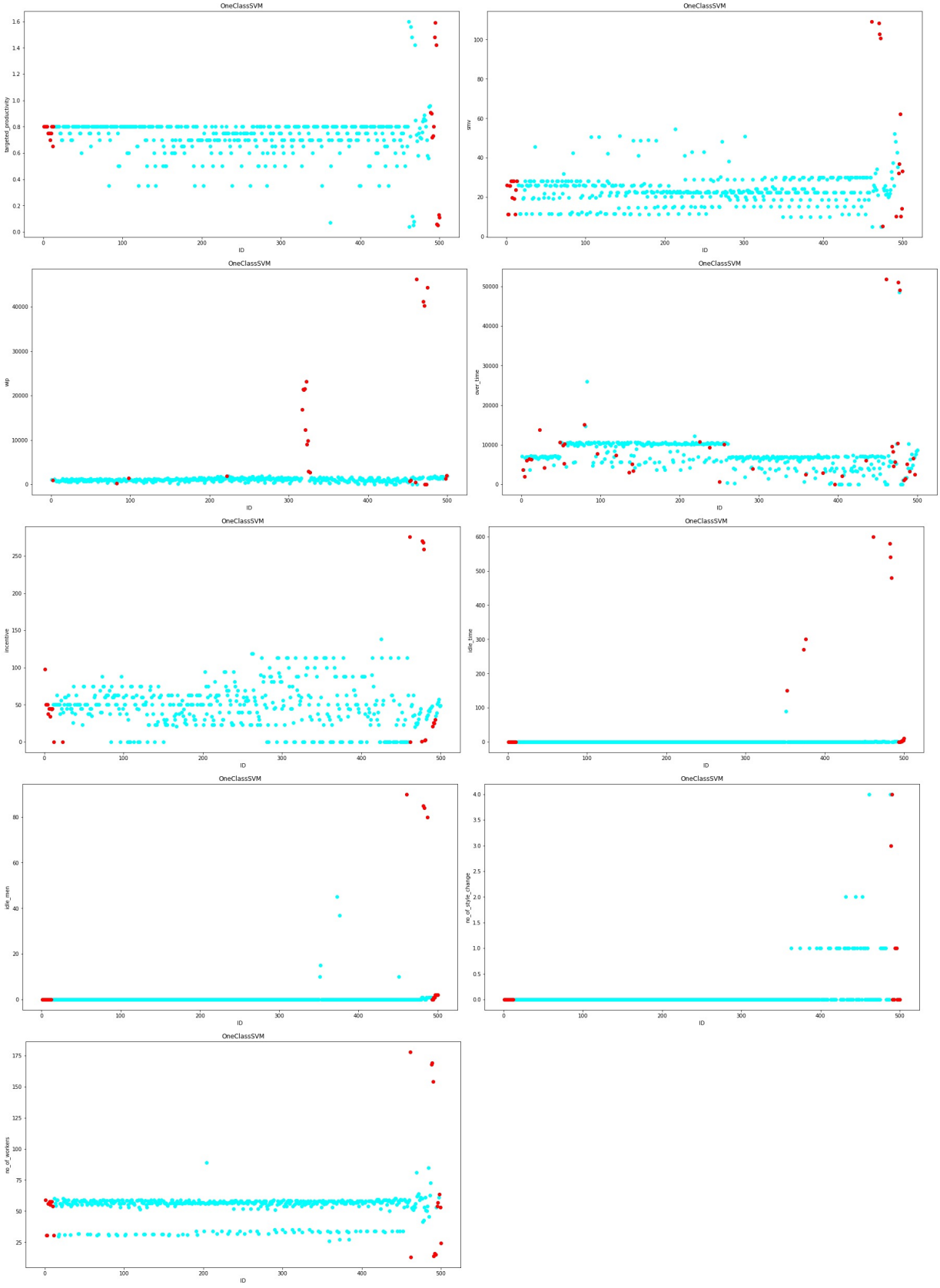
Σχήμα 4.68: Γραφήματα ακραίων τιμών Local Outlier Factor στο Productivity Prediction of Garment Employees



Σχήμα 4.69: Γραφήματα ακραίων τιμών Mahalanobis Distance στο Productivity Prediction of Garment Employees



Σχήμα 4.70: Γραφήματα ακραίων τιμών One Class Support Vector Machine στο Productivity Prediction of Garment Employees



Κεφάλαιο 5

Συμπεράσματα

Η παρούσα διπλωματική εργασία αφορά τη σύγκριση ορισμένων μεθόδων εύρεσης ακραίων τιμών. Συγκεκριμένα, οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν οι DBscan, Elliptical Envelope, Gaussian Mixture Model, Isolation Forest, Local Outlier Factor, Mahalanobis distance και One Class SVM. Επιπλέον για τη λήψη των ποσοστών ακριβείας της πρόβλεψης των αλγορίθμων χρησιμοποιήθηκαν δύο αλγόριθμοι μηχανικής μάθησης οι Random Forest και Support Vector Machine. Τα δέκα σετ δεδομένων που χρησιμοποιήθηκαν αναφέρονται στους Πίνακες 4.1 και 4.2 και προορίζονται για προβλέψεις πειραμάτων με αλγόριθμους μηχανικής μάθησης.

Χωρίζονται σε δύο κατηγορίες των πέντε σετ, της κατηγοριοποίησης και της παλινδρόμησης, αντίστοιχα. Τα στάδια σύγκρισης των αλγορίθμων είναι τρία. Το πρώτο στάδιο αναφέρεται στη σύγκριση των αλγορίθμων με είσοδο τα δεδομένα των αρχικών σετ, χωρίς καμία παρέμβαση. Στο δεύτερο στάδιο εισάγονται οι ακραίες τιμές στα σετ δεδομένων της τάξεως του 5% και 10%, οι οποίες έχουν εύρος [$x/2$, $2 \times y$], αν υποθέσουμε ότι το εύρος του αρχικού σετ δεδομένων είναι [x,y]. Στο τελικό τρίτο στάδιο γίνεται η σύγκριση των αλγορίθμων αφού έχουν ανιχνευθεί και απαλειφθεί οι ακραίες παρατηρήσεις. Σύμφωνα με τη θεωρία οι προβλέψεις θα πρέπει να κατατάσσονται ως εξής: υψηλότερα ποσοστά θα πρέπει να παρουσιάζει το πρώτο στάδιο που έχει το αρχικό σετ δεδομένων, έπειτα να ακουλουθεί το σετ δεδομένων όπου έχουν αφαιρεθεί οι ακραίες τιμές και στο τέλος να βρίσκονται τα ποσοστά που προήλθαν από το σετ με τις εισαγόμενες ακραίες τιμές.

Οι προσδοκόμενες προβλέψεις πραγματοποιήθηκαν ως επί το πλείστον και τηρούσαν τις προϋποθέσεις της θεωρίας. Παρ' όλ' αυτά σε περιπτώσεις όπως οι Πίνακες 4.14, 4.38, 4.35 και 4.36 παρουσίασαν ποσοστά τα οποία απείχαν αρκετά

από τα αναμενόμενα με αποτέλεσμα η τρίτη περίπτωση των σετ δεδομένων με τις διαγραμμένες ακραίες τιμές να έχει υψηλότερη ακριβεία από ότι τα αρχικά σετ δεδομένων. Επιπροσθέτως στην περίπτωση των Πινάκων 4.5, 4.6, 4.3 και 4.4 παρατηρούμε ότι τα σετ με την ύπαρξη ακραίων τιμών παρουσιάζουν υψηλότερα ποσοστά ακριβείας σε σύγκριση με τις άλλες δύο περιπτώσεις. Επιπλέον κατά το πλείστον των περιπτώσεων στα σετ δεδομένων με 5% ακραίες τιμές τα ποσοστά της δεύτερης περίπτωσης, δηλαδή με την ύπαρξη ακραίων τιμών στο σετ δεδομένων, ήταν υψηλότερα από αυτά των σετ δεδομένων με 10% ακραίες τιμές. Το αντίστροφο συνέβει στην τρίτη περίπτωση, δηλαδή στα σετ δεδομένων κατά τα οποία έχουν διαγραφεί οι ακραίες παρατηρήσεις, τα σετ με 5% ακραίες τιμές είχαν χαμηλότερα ποσοστά ακριβούς πρόβλεψης από τα σετ με 10% ακραίες τιμές.

Συνολικά ο αλγόριθμος με τα χαμηλότερα ποσοστά επιτυχίας της πρόβλεψης ήταν ο One Class SVM, ιδιαίτερα στην τρίτη σύγκριση δηλαδή των σετ δεδομένων χωρίς ακραίες τιμές. Τόσο στην κατηγορία παλινδρόμησης όσο και στην κατηγοριοποίησης ο DBscan παρουσίασε από τα πιο υψηλά ποσοστά πρόβλεψης. Τέλος, σταθεροί στη μεγαλύτερη διάρκεια των πειραμάτων εμφανίστηκαν οι αλγόριθμοι Elliptical Envelope, Gaussian Mixture Model, Isolation Forest και Local Outlier Factor.

Συνεπώς καταλήγουμε στο συμπέρασμα ότι ο τρόπος λειτουργίας του κάθε αλγορίθμου καθώς και οι ιδιαιτερότητες των σετ δεδομένων σε συνδυασμό με τη λανθασμένη προεπεξεργασία τους, μπορούν να οδηγήσουν σε λανθασμένα συμπεράσματα καθώς και σε μη αποδοτικά πορίσματα. Παρ' όλ' αυτά η διπλωματική εργασία αποδεικνύει ότι με τον σωστό χειρισμό των δεδομένων καθώς και την επιλογή των κατάλληλων αλγορίθμων η απαλοιφή των ακραίων τιμών συμβάλει στη βελτίωση της ακριβούς πρόβλεψης των δεδομένων.

Βιβλιογραφία

[1] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 38, p. 100306, November 2020.

[2] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in *Information and Decision Sciences* (S. C. Satapathy, J. M. R. Tavares, V. Bhateja, and J. R. Mohanty, eds.), (Singapore), pp. 511–518, Springer Singapore, 2018.

[3] S. Kaliyaperumal, M. Kuppusamy, and S. Arumugam, "Labeling methods for identifying outliers," *International Journal of Statistics and Systems*, vol. 10, pp. 231–238, October 2015.

[4] W.-R. Chen, Y.-H. Yun, M. Wen, H.-M. Lu, Z.-M. Zhang, and Y.-Z. Liang, "Representative subset selection and outlier detection via isolation forest," *Analytical Methods*, vol. 8, no. 39, pp. 7225–7231, 2016.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, December 2008.

[6] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *2007 IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, 2007.

[7] M. Alshawabkeh, B. Jang, and D. Kaeli, "Accelerating the local outlier factor algorithm on a GPU for intrusion detection systems," in *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units - GPGPU '10*, ACM Press, 2010.

[8] M. Ashrafuzzaman, S. Das, A. A. Jillepalli, Y. Chakhchoukh, and F. T. Sheldon, "Elliptic envelope based detection of stealthy false data injection attacks in smart grid control systems," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1131–1137, 2020.

[9] R. D. Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 1–18, January 2000.

[10] J. Yang, T. Deng, and R. Sui, "An adaptive weighted one-class SVM for robust outlier detection," in *Proceedings of the 2015 Chinese Intelligent Systems Conference*, pp. 475–484, Springer Berlin Heidelberg, November 2015.

[11] Y. Zhang, *New Advances in Machine Learning*. IntechOpen, 2010.

[12] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–239, May 1996.

[13] D. D. and S. S. Babu, "Methods to detect different types of outliers," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, IEEE, March 2016.

[14] J. Jiang, G. Han, L. liu, L. Shu, and M. Guizani, "Outlier detection approaches based on machine learning in the internet-of-things," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 53–59, 2020.

[15] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019.

[16] J. Han, M. Kamber, and J. Pei, "12 - outlier detection," in *Data Mining (Third Edition)* (J. Han, M. Kamber, and J. Pei, eds.), The Morgan Kaufmann Series in Data Management Systems, pp. 543–584, Boston: Morgan Kaufmann, third edition ed., 2012.

[17] L. Li, R. J. Hansman, R. Palacios, and R. Welsch, "Anomaly detection via a gaussian mixture model for flight operation and safety monitoring," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 45–57, March 2016.

[18] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems," *Springer Link*, vol. 29, p. 228–237, July 2003.

[19] A. Apsemidis, S. Psarakis, and J. M. Moguerza, "A review of machine learning kernel methods in statistical process monitoring," *Computers & Industrial Engineering*, vol. 142, 2020.

[20] H. C. Mandhare and S. R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, June 2017.

[21] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, 10 2004.

[22] C. Cassisi, A. Ferro, R. Giugno, G. Pigola, and A. Pulvirenti, "Enhancing density-based clustering: Parameter reduction and outlier detection," *Information Systems*, vol. 38, pp. 317–330, May 2013.

[23] C.-H. Lin, K.-C. Hsu, K. R. Johnson, M. Luby, and Y. C. Fann, "Applying density-based outlier identifications using multiple datasets for validation of stroke clinical outcomes," *International Journal of Medical Informatics*, vol. 132, p. 103988, December 2019.

[24] S.-y. Jiang and Q.-b. An, "Clustering-based outlier detection method," in *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 2, 2008.

[25] M. Jiang, S. Tseng, and C. Su, "Two-phase clustering process for outliers detection," *Pattern Recognit. Lett.*, vol. 22, no. 6/7, pp. 691–700, 2001.

[26] D. Yu, G. Sheikholeslami, and A. Zhang, "FindOut : Finding outliers in very large datasets," *Knowledge and Information Systems*, vol. 4, pp. 387–412, September 2002.

[27] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: a wavelet-based clustering approach for spatial data in very large databases," *The VLDB Journal The International Journal on Very Large Data Bases*, vol. 8, February 2000.

[28] X. Su, Y. Lan, R. Wan, and Y. Qin, "An outlier detection algorithm based on arbitrary shape clustering," in *Advanced Data Mining and Applications* (R. Huang, Q. Yang, J. Pei, J. Gama, X. Meng, and X. Li, eds.), (Berlin, Heidelberg), pp. 627–635, Springer Berlin Heidelberg, 2009.

[29] M. Celik, F. Dadaser-Celik, and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*, IEEE, June 2011.

[30] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of spatial outlier by using improved z-score test," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, April 2019.

[31] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, p. 1, December 2020.

[32] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the mahalanobis distance," *Journal of Experimental Social Psychology*, vol. 74, pp. 150–156, January 2018.

[33] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description - ODD '13*, ACM Press, 2013.