ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογή αλγορίθμων βαθιάς μάθησης σε προβλήματα μεγάλου όγκου δεδομένων για πρόβλεψη και ταξινόμηση

της
Χριστοδούλου Ειρήνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Σταμούλης Γεώργιος

Λαμία, 2019

UNIVERSITY OF THESSALY

SCHOOL OF SCIENCE

INFORMATICS AND COMPUTATIONAL BIOMEDICINE

Investigate Deep Learning for big data prediction and classification tasks

Christodoulou Irene

Master thesis

Stamoulis Georgios

Lamia, 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ
ΚΑΤΕΥΘΥΝΣΗ

«ΠΛΗΡΟΦΟΡΙΚΗ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΑΣΦΑΛΕΙΑ, ΔΙΑΧΕΙΡΙΣΗ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΣΟΜΟΙΩΣΗ»

Εφαρμογή αλγορίθμων βαθιάς μάθησης σε προβλήματα μεγάλου όγκου δεδομένων για πρόβλεψη και ταξινόμηση

Χριστοδούλου Ειρήνη

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Σταμούλης Γεώργιος

Λαμία, 2019

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο [Εφαρμογή αλγορίθμων βαθιάς μάθησης σε προβλήματα μεγάλου όγκου δεδομένων για πρόβλεψη και ταξινόμηση] αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Η ΔΗΛΟΥΣΑ

Λαμία,

Υπογραφή

Εφαρμογή αλγορίθμων βαθιάς μάθησης σε προβλήματα μεγάλου όγκου δεδομένων για πρόβλεψη και ταξινόμηση

Χριστοδούλου Ειρήνη

Τριμελής Επιτροπή:

Σταμούλης Γεώργιος (επιβλέπων)

Δημητρίου Γεώργιος

Κοζύρη Μαρία

Επιστημονικός Σύμβουλος:

Παπαγεωργίου Ελπινίκη

Ευχαριστίες

Η ολοκλήρωση αυτής της διπλωματικής υλοποιήθηκε με την υποστήριξη ενός αριθμού ανθρώπων στους οποίους θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου. Πρώτα από όλους, θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια μου κ. Παπαγεωργίου Ελπινίκη, για την εμπιστοσύνη που μου έδειξε και για την δυνατότητα που μου έδωσε να πραγματοποιήσω την παρούσα διπλωματική εργασία καθώς και για την απαραίτητη καθοδήγηση που μου πρόσφερε και τις επιστημονικές της συμβουλές, συμβάλλοντας σημαντικά στην ολοκλήρωση της παρούσας διπλωματικής εργασίας. Επίσης, ευχαριστώ την οικογένεια μου ,για την στήριξή της σε όλο το χρονικό διάστημα της φοίτησής μου.

Λαμία, Φεβρουάριος 2019

Π Ε Ρ Ι Ε Χ Ο Μ Ε Ν Α

Περίληψη

Ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζουμε στην μηχανική μάθηση είναι το πρόβλημα των μη ισορροπημένων δεδομένων. Το πρόβλημα αυτό όσον αφορά την ταξινόμηση αναφέρεται στην πρόβλεψη μιας ή περισσοτέρων τάξεων/κατηγοριών (Classes) όταν υπάρχει εξαιρετικά χαμηλός αριθμός δειγμάτων σε μία από αυτές τις τάξεις/κατηγορίες.

Στα πλαίσια της παρούσας διπλωματικής εργασίας προσεγγίσαμε την επίλυση αυτού του προβλήματος με τεχνικές υπό- δειγματοληψίας και υπέρ- δειγματοληψίας και μέσω Ανάλυσης Κύριων Συνιστωσών (Principal component analysis (PCA)) μειώσαμε το μέγεθος των δεδομένων. Οι προτεινόμενες τεχνικές εφαρμόστηκαν για την επίλυση του προβλήματος με βάση την υπάρχουσα βιβλιογραφία και τα αποτελέσματα έδειξαν ακρίβεια ταξινόμησης έως 92%. Με τις κλασσικές τεχνικές ταξινόμησης, η τρίτη κλάση που είχε και το μικρότερο αριθμό δειγμάτων δεν μπορούσε καν να ταξινομηθεί και να δώσει ένα ικανοποιητικό αποτέλεσμα. Με τις προτεινόμενες τεχνικές και οι 3 κατηγορίες ταξινομήθηκαν με αποδεκτά ποσοστά ταξινόμησης, συγκρινόμενα με την υπάρχουσα βιβλιογραφία.

Επιπλέον, εφαρμόστηκαν προηγμένες τεχνικές νευρωνικών δικτύων, όπως η βαθιά προσέγγιση μάθησης, για να καταστεί δυνατή η πρόβλεψη φυσικού αερίου σε πραγματικά δεδομένα από τον Ελλαδικό χώρο. Στόχος της πρόβλεψης είναι η ακολουθία χρονικών παρατηρήσεων να συνεχιστεί και μετέπειτα με εστίαση στην πρόβλεψη της επόμενης ημέρας. Η προτεινόμενη μέθοδος εφαρμόστηκε, έχοντας ως σκοπό να φτάσουμε όσο το δυνατόν κοντά στον συγκεκριμένο στόχο. Τα αποτελέσματα έδειξαν, με την βοήθεια των κριτηρίων αξιολόγησης των μεθόδων προβλέψεων, τον συντελεστή συσχέτισης (R^2) να ανέρχεται έως και 98%.

Συμπεραίνουμε ότι, με βάση τα τελικά αποτελέσματα, οι τεχνικές υπό-δειγματοληψίας και υπέρ- δειγματοληψίας μπορούν να λύσουν το πρόβλημα των μη ισορροπημένων δεδομένων και η βαθιά μάθηση μπορεί να βοηθήσει ουσιαστικά σε προβλήματα πρόβλεψης, με στόχο την πρόβλεψη της επόμενης ημέρας.

Abstract

One of the major problems we face in Machine Learning is the class imbalance problem. This problem, in terms of classification, refers to the prediction of one or more classes when there is an extremely small number of samples in one of these classes/categories.

In the context of this thesis, we have approached solving this problem with both under-sampling and over-sampling techniques and through Principal Components Analysis (PCA) we have reduced the size of the data. The proposed techniques were applied to solve the problem based on the existing literature and the results showed a classification accuracy of up to 92%. The third class that had the smallest number of samples could not even be classified and give a satisfactory result with the implementation of conventional classification techniques. With the proposed techniques, all three classes were classified with acceptable classification rates, compared to the existing literature.

Additionally, advanced neural networks, such as deep learning, have been applied to enable natural gas to be predicted on real-world data from Greece. The objective of the prediction (ή forecasting) is to continue the sequence of time observations and then to focus on the next day's prediction/forecasting. The proposed method has been implemented with the aim of reaching as close as possible to that target. The results showed that the correlation coefficient (R^2) was up to 98% with the help of the forecasting criteria.

We conclude that, based on the final results, under-sampling and over-sampling techniques can solve the problem of the imbalance data and deep learning can actually help with forecasting problems with aim of the next day's prediction/forecasting.

1. Εισαγωγή

1.1. Τι είναι η Βαθιά Μάθηση (Deep Learning) και πως λειτουργεί.

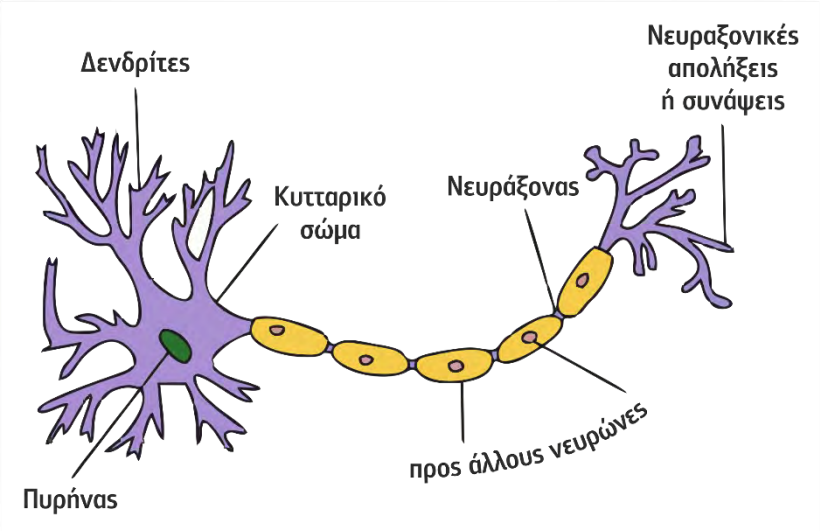
Η βαθιά μάθηση είναι ένα υποσύνολο της μηχανικής μάθησης όπου τεχνητά νευρωνικά δίκτυα, μαθαίνουν από ένα μεγάλο σύνολο δεδομένων. Η βαθιά μάθηση βασίζεται σε τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks), για την επίλυση διαφόρων προβλημάτων στην επεξεργασία δεδομένων, όπως η εικόνα, ο ήχος και το κείμενο. Τα μοντέλα βαθιάς μάθησης μπορούν να επιτύχουν την ακρίβεια της τεχνολογίας, μερικές φορές να υπερβαίνουν τις επιδόσεις σε ανθρώπινο επίπεδο.

Ένα βασικό πλεονέκτημα των δικτύων βαθιάς μάθησης είναι ότι συχνά συνεχίζουν να βελτιώνονται όσο αυξάνεται το μέγεθος των δεδομένων.

Ουσιαστικά η βαθιά μάθηση, υλοποιείται μέσω νευρωνικών δικτύων με περισσότερο από ένα απλό κρυφό στρώμα νευρώνων.

Εμπνευσμένο από τα νευρικά κύτταρα (νευρώνες) που αποτελούν τον ανθρώπινο εγκέφαλο, τα νευρικά δίκτυα περιλαμβάνουν στρώματα (νευρώνες) που συνδέονται μεταξύ τους σε παρακείμενα στρώματα.

[1]



Εικόνα 1 Σχηματικό διάγραμμα ενός τυπικού νευρώνα

[1] Πηγή Εικόνας: (ΓΕΩΡΓΟΥΛΗ 2015)

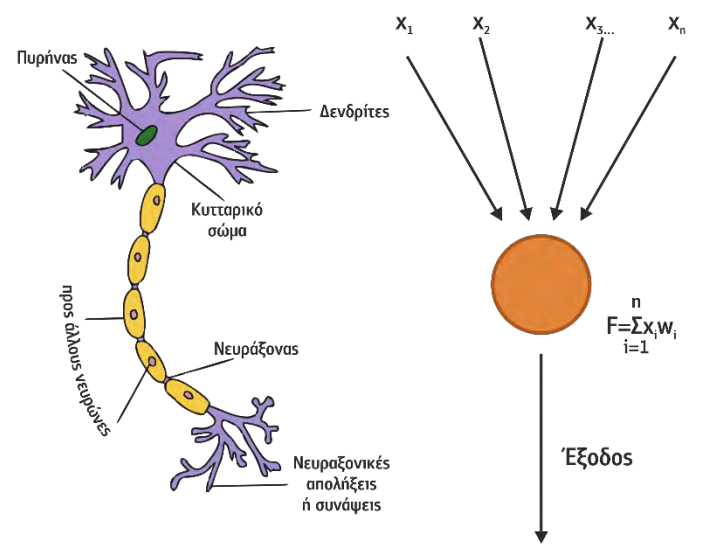
Είναι παρόμοιο με τη δομή και τη λειτουργία του ανθρώπινου νευρικού συστήματος (Εικόνα 1), όπου ένα πολύπλοκο δίκτυο διασυνδεδεμένων μονάδων υπολογισμού εργάζεται με συντονισμένο τρόπο για να επεξεργάζεται περίπλοκες πληροφορίες.

Ένας νευρώνας στον εγκέφαλο, λαμβάνει σήματα - έως και 100.000 - από άλλους νευρώνες, διαχωρίζεται από τα υπόλοιπα κύτταρα με μια μεμβράνη και έχει την ικανότητα να μεταφέρει ηλεκτρικά σήματα από το νευρώνα αυτόν προς τους υπόλοιπους νευρώνες με τους οποίους επικοινωνεί [1].

Σε ένα τεχνητό νευρωνικό δίκτυο, τα σήματα ταξιδεύουν επίσης μεταξύ των «νευρώνων». Αλλά αντί να μεταφέρει ένα ηλεκτρικό σήμα, ένα νευρωνικό δίκτυο εκχωρεί βάρη σε διάφορους νευρώνες. Ένας νευρώνας που έχει βάρη περισσότερο από έναν άλλο νευρώνα θα ασκεί περισσότερο επίδραση στην επόμενη στρώση των νευρώνων. Το τελικό στρώμα συγκεντρώνει αυτές τις σταθμισμένες εισροές για να βρει μια απάντηση .

Συνοπτικά, είναι παρόμοιο με τη δομή και τη λειτουργία του ανθρώπινου νευρικού συστήματος, όπου ένα πολύπλοκο δίκτυο διασυνδεδεμένων μονάδων υπολογισμού εργάζεται με συντονισμένο τρόπο για να επεξεργάζεται περίπλοκες πληροφορίες (Εικόνα 2).

[2]



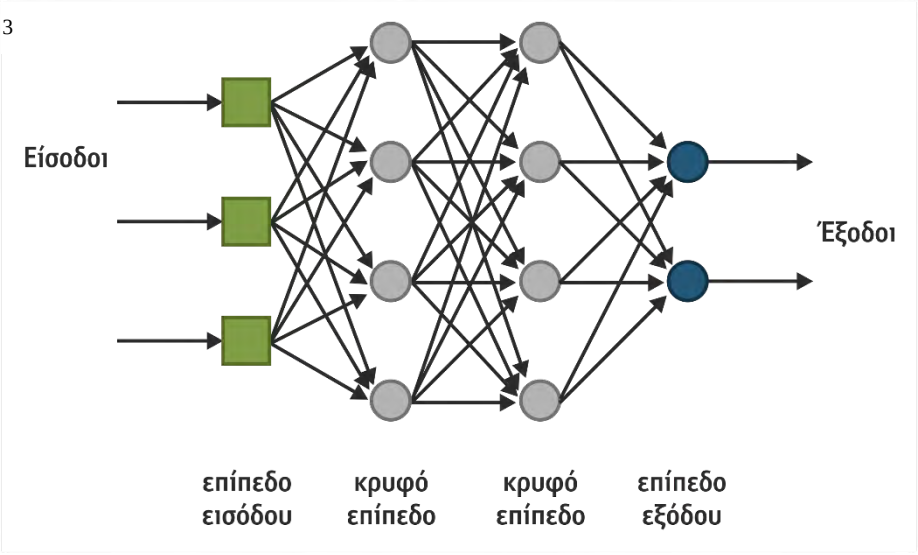
Εικόνα 2 Ο φυσικός νευρώνας σε σχέση με τον στοιχειώδη τεχνητό νευρώνα

[2]Πηγή Εικόνας: (ΓΕΩΡΓΟΥΛΗ 2015)

1.2. Αρχιτεκτονική Βαθιάς Μάθησης

Η Βαθιά Μάθηση χρησιμοποιεί αρχιτεκτονικές νευρωνικών δικτύων , και γι' αυτό τα μοντέλα βαθιάς μάθησης συχνά αναφέρονται ως δίκτυα νευρώνων .Ο όρος "βαθιά" συνήθως αναφέρεται στον αριθμό των κρυφών επιπέδων στο νευρικό δίκτυο.

[3]



Εικόνα 3 Νευρωνικά δίκτυα, τα οποία είναι οργανωμένα σε στρώματα που αποτελούνται από ένα σύνολο διασυνδεδεμένων κόμβων

Οι νευρώνες ομαδοποιούνται σε τρεις διαφορετικούς τύπους στρωμάτων:

1) Επίπεδο Εισόδου (Input Layer)
2) Κρυφό Επίπεδο (Hidden Layer(s)
3) Επίπεδο Εξόδου (Output Layer)

Το **επίπεδο** εισόδου λαμβάνει δεδομένα εισόδου και περνά τις εισόδους στο πρώτο κρυφό στρώμα.

Τα **κρυμμένα επίπεδα** λαμβάνουν ένα σύνολο βαρών από το επίπεδο εισόδου και παράγουν μια έξοδο μέσω μιας συνάρτησης ενεργοποίησης.

Το **επίπεδο εξόδου** επιστρέφει τα δεδομένα εξόδου.

[3]Πηγή Εικόνας: (ΓΕΩΡΓΟΥΛΗ 2015)

Ο τρόπος με τον οποίον είναι συνδεδεμένα τα επίπεδα, ο αριθμός των κρυφών επιπέδων και ο τύπος των νευρώνων που έχει το κάθε επίπεδο καθορίζουν την αρχιτεκτονική του Τεχνητού Νευρωνικού Δικτύου (ΤΝΝ).

Το " **Βαθιά** " στη βαθιά μάθηση αναφέρεται σε **περισσότερες από μία** κρυφές επιφάνειες .

2. Θεωρητικό Υπόβαθρο

2.1. Ταξινόμηση (Classification)

Η ταξινόμηση αποτελεί μια από τις βασικές εργασίες στο στάδιο της Εξόρυξης Δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου, το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η βασική ιδέα είναι η εξής: έχοντας ένα σύνολο από κλάσεις και ένα σύνολο δεδομένων με δείγματα, για τα οποία ξέρουμε σε ποια κλάση ανήκουν, στόχος της ταξινόμησης είναι η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να ταξινομήσει αυτόματα σε αυτές τις κατηγορίες νέα, άγνωστα, μη-ταξινομημένα δείγματα [2].

Συγκεκριμένα, δίνετε ένα σύνολο δεδομένων, το οποίο χωρίζεται σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης (training test) είναι ένα σύνολο από μεταβλητές, το οποίο έχει ένα σύνολο από γνωρίσματα, ένα από αυτά είναι η κλάση (Class). Πραγματοποιείται εύρεση ενός μοντέλου για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων. Ο τελικός στόχος είναι, νέες καταγραφές θα πρέπει να ανατίθενται σε μία κατηγορία με όσο μεγαλύτερη ακρίβεια γίνεται . Το σύνολο ελέγχου (test set) βοηθάει για να χαρακτηρίσουμε την ακρίβεια του μοντέλου.

Επομένως το σύνολο εκπαίδευσης χρησιμοποιείται για την κατασκευή του μοντέλου, και το σύνολο ελέγχου για την επικύρωση του συγκεκριμένου μοντέλου.

2.1.1. Αξιολόγηση Ταξινόμησης

Η απόδοση ενός μοντέλου μπορεί να εξαρτάται από πολλούς παράγοντες, εκτός του αλγορίθμου μάθησης, , όπως από:

- Την κατανομή των κλάσεων
- Το κόστος της λανθασμένης ταξινόμησης
- Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου [3].

Για την εκτίμηση της απόδοσης του μοντέλου χρησιμοποιούμε τα μέτρα (Metrics) και πιο συγκεκριμένα:

- Τον Πίνακα Σύγχυσης (Confusion Matrix)
- Την Αξιοπιστία (Accuracy)
- Τον Πίνακα Κόστους
- Την Ακρίβεια (Precision)

Πίνακας Σύγχυσης (Confusion Matrix)

Σε περιπτώσεις όπου οι κλάσεις δεν είναι ισομερώς κατανεμημένες ή όπου οι εσφαλμένες κατηγοριοποιήσεις διαφορετικών κλάσεων έχουν διαφορετικό κόστος, είναι σημαντική η εκτίμηση της ικανότητας πρόβλεψης του κατηγοριοποιητή για την κάθε κλάση. Ένας τρόπος παρουσίασης των επιδόσεων ανά κλάση ενός κατηγοριοποιητή είναι με τη χρήση του πίνακα σύγχυσης (confusion matrix). Ο Πίνακας Σύγχυσης είναι ένας δισδιάστατος πίνακας, όπου οι στήλες αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Στα κελιά του πίνακα αναγράφονται οι αληθινές θετικές, οι αληθινές αρνητικές, οι ψευδείς θετικές και οι ψευδείς αρνητικές προβλέψεις [4] .

	Πρόβλεψη Αρνητικής Κλάσης	**Πρόβλεψη Θετικής Κλάσης**
Πραγματική Αρνητική Κλάση	**tn**	**fp**
Πραγματική Θετική Κλάση	**fn**	**tp**

- **Αληθινές Θετικές Προβλέψεις** (true positive – tp): είναι το πλήθος των επιτυχών προβλέψεων για θετικές παρατηρήσεις
- **Αληθινές Αρνητικές Προβλέψεις** (true negative – tn): είναι το πλήθος των επιτυχημένων προβλέψεων για αρνητικές παρατηρήσεις .
- **Ψευδείς Θετικές Προβλέψεις** (false positive – fp): είναι το πλήθος των αποτυχημένων προβλέψεων για αρνητικές παρατηρήσεις.
- **Ψευδείς Αρνητικές Προβλέψεις** (false negative – fn): είναι το πλήθος των αποτυχημένων προβλέψεων για θετικές παρατηρήσεις.[4]

Αξιοπιστία (accuracy)

$$accuracy = sensitivity * \frac{pos}{pos + negat} + specificity * \frac{neg}{pos + negat} = \frac{tp + tn}{pos + negat}$$

Όπου pos είναι το πλήθος των θετικών παρατηρήσεων και negat είναι το πλήθος των αρνητικών παρατηρήσεων. Η αξιοπιστία (accuracy) ορίζεται ως το ποσοστό των ορθών θετικών προβλέψεων επί το ποσοστό των θετικών παρατηρήσεων συν το ποσοστό των ορθών αρνητικών προβλέψεων επί το ποσοστό των αρνητικών παρατηρήσεων ή ισοδύναμα ως το πλήθος των ορθών προβλέψεων προς το πλήθος των παρατηρήσεων [4].

Πίνακας Κόστους

C(Κόστος)	Πρόβλεψη Αρνητικής Κλάσης Class = Yes	Πρόβλεψη Θετικής Κλάσης Class = No
Πραγματική Αρνητική Κλάση Class = Yes	**Tn C(Yes|Yes)**	**Fp C(Yes|No)**
Πραγματική Θετική Κλάση Class = No	**Fn C(No|Yes)**	**Tp C(No|No)**

C(K) = TP x C(Yes|Yes) + FN x C(Yes|No) + FP C(No|Yes) + TN C(No|No)

Ακρίβεια (Precision)

$$\boldsymbol{precision = \frac{tp}{tp + fp}}$$

2.2. Πρόβλεψη

Η πρόβλεψη χρησιμοποιεί ορισμένες μεταβλητές για την πρόβλεψη άγνωστων ή μελλοντικών τιμών άλλων μεταβλητών [5]. Στόχος ενός μοντέλου πρόβλεψης είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών. Για παράδειγμα, η πρόβλεψη μπορεί να βασίζεται στη χρονολογική κατάταξη των δεδομένων.

Επιπλέον, η πρόβλεψη χρονολογικών σειρών είναι ένας σημαντικός τομέας της μηχανικής μάθησης [6].

2.2.1. Αξιολόγηση Πρόβλεψης

Για την επιλογή της κατάλληλης μεθόδου χρησιμοποιούνται τα κριτήρια αξιολόγησης των μεθόδων προβλέψεων. Τα κριτήρια αυτά βασίζονται στις τιμές των αποκλίσεων των προβλεπόμενων τιμών από τις αντίστοιχες πραγματικές τιμές της χρονοσειράς (Νικολάου 2007).

Για να προσδιορίσουμε την αξιοπιστία μιας συγκεκριμένης μεθόδου πρόβλεψης, θα πρέπει να μελετήσουμε τη διαχρονική συμπεριφορά των τιμών των σφαλμάτων

της πρόβλεψης. Αυτό γίνεται με την εφαρμογή διάφορων κριτηρίων, σύμφωνα με τα οποία αξιολογούμε τη χρησιμοποιούμενη μέθοδο πρόβλεψης. Κάθε ένα από τα κριτήρια αυτά ορίζεται από μία συγκεκριμένη συναρτησιακή σχέση των σφαλμάτων της πρόβλεψης και μπορεί να χρησιμοποιηθεί όχι μόνο για την αξιολόγηση μιας μεθόδου πρόβλεψης αλλά και για την επιλογή της " καλύτερης " μεταξύ δύο ή περισσοτέρων εναλλακτικών μεθόδων προβλέψεων (Μαργιά 2009). Τα κριτήρια αυτά είναι:

- Mean Squared Error [7]:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(Z(t) - X(t))^2$$

- Root Mean Squared Error [8]:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

- Mean Absolute Error [9]:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|Z(t) - X(t)|$$

- Mean Absolute Percentage Error [10]:

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{Z(t) - X(t)}{Z(t)}\right|$$

- Coefficient of Correlation [11]:

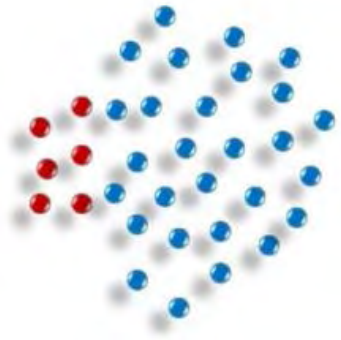
$$R = \frac{T\sum_{t=1}^{T} Z(t)\cdot X(t) \; - (\sum_{t=1}^{T} Z(t))(\sum_{t=1}^{T} X(t))}{\sqrt{T\sum_{t=1}^{T}(Z(t))^2 - (\sum_{t=1}^{T} Z(t))^2}\cdot\sqrt{T\sum_{t=1}^{T}(X(t))^2 - (\sum_{t=1}^{T} X(t))^2}}$$

3. Προβλήματα Ταξινόμησής και Αντιμετώπιση

3.1. Το πρόβλημα μη – ισορροπημένων δεδομένων

Στη μηχανική μάθηση όπου ο συνολικός αριθμός μιας κατηγορίας δεδομένων (θετικός) είναι πολύ μικρότερος από τον συνολικό αριθμό μιας άλλης κατηγορίας δεδομένων (αρνητική), θεωρείται πρόβλημα μη – ισορροπημένων δεδομένων. Στην Εικόνα 4 είναι ένα παράδειγμα μη ισορροπημένων δεδομένων, όπου οι κόκκινες κουκίδες δείχνουν τάξη μειονότητας ενώ οι μπλε κουκίδες υποδεικνύουν τάξη πλειοψηφίας.

[4]



Εικόνα 4 Απεικόνιση μη-ισορροπημένων δεδομένων

[4] Πηγή Εικόνας: https://www.datascience.com/blog/imbalanced-data

3.2. Γιατί είναι πρόβλημα

Το πρόβλημα αυτό μπορεί να παρατηρηθεί σε διάφορους κλάδους, όπως για παράδειγμα, στην ανίχνευση απάτης (ο αριθμός των απατών θα είναι πολύ χαμηλότερος από τις πραγματικές συναλλαγές), το κόστος της κακής κατάταξης μειονοτικής τάξης θα μπορούσε να είναι πολύ υψηλό. Αυτό σημαίνει ότι, εάν δεν εντοπίζονται σωστά οι περιπτώσεις απάτης, το μοντέλο δεν θα είναι χρήσιμο και επιπλέον είναι εξαιρετικά δαπανηρό για την εταιρεία ηλεκτρονικού εμπορίου. Εάν περάσει μια δόλια συναλλαγή, αυτό επηρεάζει την εμπιστοσύνη των πελατών και επιπρόσθετα χρηματικό κόστος. Επομένως, σκοπός είναι να αναγνωρίζονται όσο το δυνατόν περισσότερες δόλιες συναλλαγές.

Οι περισσότεροι αλγόριθμοι εκμάθησης μηχανών λειτουργούν καλύτερα όταν ο αριθμός των περιπτώσεων κάθε κλάσης είναι σχεδόν ίσος. Όταν ο αριθμός της κλάσης υπερβαίνει κατά πολύ την άλλη, προκύπτουν προβλήματα.

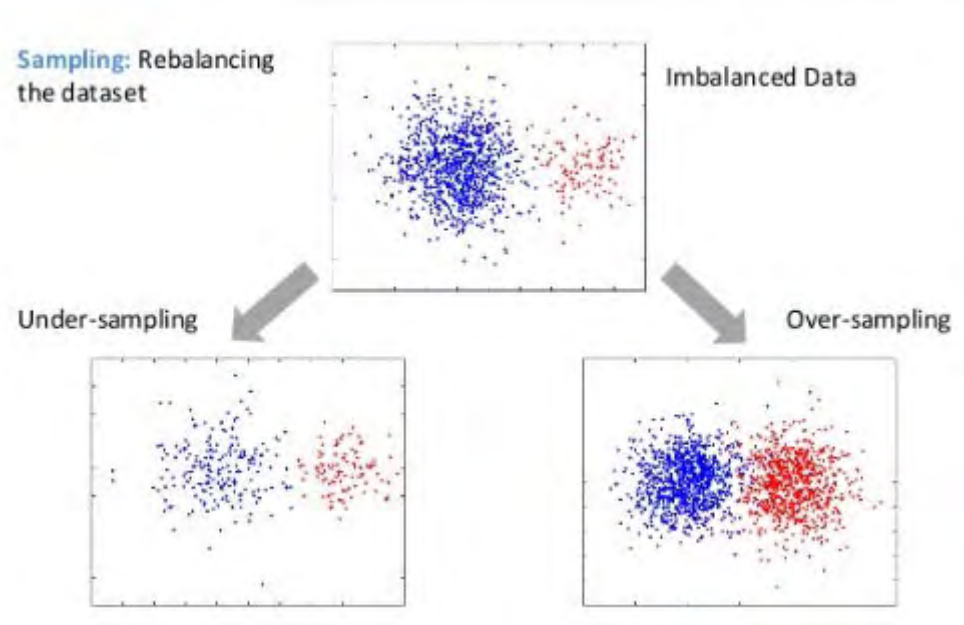
4. Τεχνικές Δειγματοληψίας για την Αντιμετώπιση προβλημάτων ταξινόμησης

Έχουν υπάρξει διάφορες προσεγγίσεις για την αντιμετώπιση προβλημάτων ταξινόμησης, όπως η μέθοδος MWMOTE (Majority Weighted Minority Oversampling TEchnique). Το MWMOTE αναγνωρίζει πρώτα τα δείγματα με λιγότερη πληροφορία της τάξεως μειονότητας και τους εκχωρεί βάρη, με βάση την ευκλείδεια απόσταση τους από τα πλησιέστερα δείγματα της τάξης πλειοψηφίας. Στη συνέχεια παράγει τα συνθετικά δείγματα από τα δείγματα της τάξης μειονότητας και τα ομαδοποιεί. Αυτό γίνεται με τέτοιο τρόπο ώστε όλα τα συνθετικά δείγματα να βρίσκονται μέσα σε κάποιο σύμπλεγμα μειονοτικών τάξεων (Barua, et al. 26 2012).

Επιπλέον, προσεγγίστηκε με την μέθοδο ADASYN (Adaptive Synthetic). Η βασική ιδέα της τεχνικής ADASYN είναι να χρησιμοποιεί μια σταθερή κατανομή για διάφορα δείγματα της μειοψηφίας ανάλογα με το επίπεδο δυσκολίας τους στην εκπαίδευση, όπου παράγονται περισσότερα συνθετικά δεδομένα για δείγματα μειονοτήτων και είναι πιο δύσκολο να εκπαιδευτούν σε σύγκριση με εκείνα τα

δείγματα μειονοτήτων που είναι ευκολότερο (Haibo, και συν. 2008) (Dr.D.Ramyachitra και P.Manikandan 2014).

Ακόμη, για την επίλυση του παρόν προβλήματος, χρησιμοποιήθηκε η τεχνική Random Forest, Στην συγκεκριμένη μέθοδο όταν προβλέπετε ένα νέο δείγμα, παράγεται μια πρόβλεψη από κάθε δέντρο και τα αποτελέσματα αυτά συνδυάζονται για να δημιουργήσουν μια ενιαία πρόβλεψη για ένα δείγμα μειονοτήτων (Chen, Liaw και Breiman 2004). Στην παρούσα εργασία ασχοληθήκαμε με τις τεχνικές υπέρ-δειγματοληψίας SMOTE και υπό-δειγματοληψίας NearMiss σε γλώσσα Python.



Εικόνα 5 Τεχνικές Δειγματοληψίας

[5] Πηγή Εικόνας: https://api.ning.com/files/vvHEZw33BGqEUW8aBYm4epYJWOfSeUBPVQAsgz7aWaNe0pmDBsjggBxsyq*8VU1FdBshuTDdL2-bp2ALs0E-0kpCV5kVdwu/imbdata.png

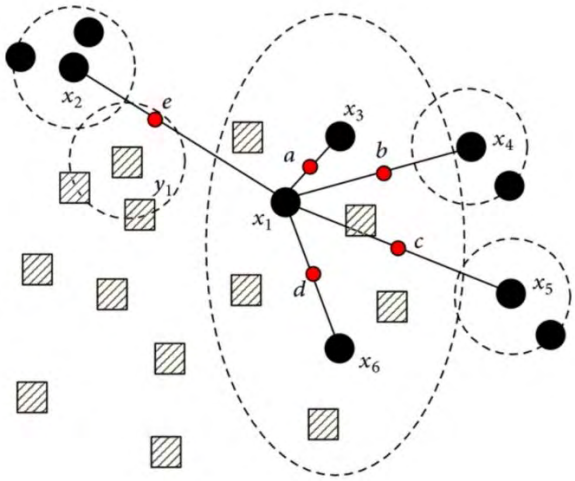
4.1. Τεχνικές υπέρ- δειγματοληψίας (Over-sampling) για προβλήματα ταξινόμησης

4.1.1. SMOTE (Synthetic minority over-sampling technique)

Η μέθοδος SMOTE δημιουργεί συνθετικά δεδομένα για να αυξήσει τον αριθμό των δειγμάτων στο σύνολο δεδομένων. Στόχος είναι η αύξηση της τάξης των μειονοτήτων έτσι ώστε το σύνολο δεδομένων να εξισορροπείται δημιουργώντας συνθετικές παρατηρήσεις βασισμένες στις υπάρχουσες παρατηρήσεις μειοψηφίας. Εφαρμόζει την προσέγγιση *KNN* (***k*-nearest neighbors**), με βάση τους *K* πλησιέστερους γείτονες, δημιουργεί τα συνθετικά δείγματα στον χώρο της μειοψηφίας. Ο αλγόριθμος παίρνει τα διανύσματα χαρακτηριστικών από τους πλησιέστερους γείτονές του και υπολογίζει την απόσταση μεταξύ αυτών των διανυσμάτων. Η διαφορά πολλαπλασιάζεται με τυχαίο αριθμό μεταξύ (0, 1) και προστίθεται στο χαρακτηριστικό. Ο αλγόριθμος SMOTE είναι πρωτοπόρος αλγόριθμος και πολλοί άλλοι αλγόριθμοι προέρχονται από το SMOTE (Chawla, et al. 2002).

Στην παρακάνω Εικόνα 6 βρίσκει τους 5 πλησιέστερους γείτονες στα σημεία δειγματοληψίας, τραβάει μια γραμμή σε κάθε ένα από αυτά, επιλέγει τυχαία δείγματα ανάλογα με το μέγεθος της υπέρ – δειγματοληψίας και παράγει νέα συνθετικά δείγματα κατά μήκος της γραμμής, μεταξύ του δείγματος μειοψηφίας και των 5 επιλεγμένων πλησιέστερων γειτόνων του.

6



Εικόνα 6

[6] Πηγή Εικόνας: https://www.researchgate.net/publication/287601878/figure/fig1/AS:316826589384744@1452548753581/The-schematic-of-NRSBoundary-SMOTE-algorithm.png

Η εκτέλεση υπέρ-δειγματοληψίας χρησιμοποιώντας SMOTE γίνεται μέσω της παρακάτω μεθόδου:

```
Class imblearn.over.sampling.SMOTE(sampling strategy=’auto’,
                                   random_state=None,
                                   k_neighbors=5,
                                   m_neighbors= 'deprecated',
                                   out_step='deprecated',
                                   kind='deprecated',
                                   svm_estimator='deprecated',
                                   n_jobs=1,
                                   ratio=None
                                   )
```

Περιγραφή μερικών παραμέτρων:

- **ratio**: Εάν είναι "auto", ratio (αναλογία) θα οριστεί αυτόματα για να εξισορροπηθεί το σύνολο δεδομένων. Αν δοθεί ένας ακέραιος αριθμός, ο αριθμός των παραγόμενων δειγμάτων είναι ίσος με τον αριθμό των δειγμάτων στην τάξη μειοψηφίας, πολλαπλάσιο από συγκεκριμένο ratio.
  - **k_neighbors:** αριθμός των πλησιέστερων γειτόνων που χρησιμοποιούνται για την δημιουργία συνθετικών δειγμάτων. **(default=5)**
  - **m_neighbors:** ο αριθμός των πλησιέστερων γειτόνων. Χρησιμοποιείται για να προσδιοριστεί εάν ένα δείγμα μειοψηφίας κινδυνεύει να θεωρηθεί «θόρυβος ». **(default=10)**
  - **kind:** επιλογή είδους:

'regular', 'borderline1', 'borderline2', 'svm'.

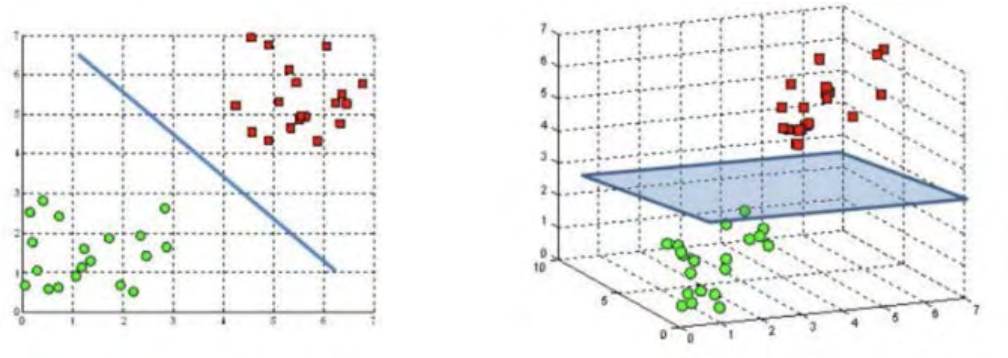
(default=’regular’) [12]

4.1.2. SMOTE SVM (Support Vector Machine)

Το SVM SMOTE είναι ένα αντικείμενο που παράγει συνθετικά δείγματα χρησιμοποιώντας τη μέθοδο SMOTE, αλλά μόνο για οριακά δείγματα. Ωστόσο, σε αντίθεση με το συνηθισμένο borderline smote, αυτή η μέθοδος χρησιμοποιεί φορέα υποστήριξης (support vector) ως δείγματα οριακής κατανομής (Saina και Purnami 2015).

Οι φορείς υποστήριξης της τάξης των μειονοτήτων βρίσκονται με την τοποθέτηση αντικειμένου ταξινόμησης SVM. Στη συνέχεια, επιλέγονται οι πλησιέστεροι γείτονες για κάθε φορέα υποστήριξης μειονοτήτων. Τα δείγματα μειοψηφίας που περιβάλλουν τα δείγματα της πλειοψηφίας, δηλαδή, όλοι οι πιο κοντινοί γείτονες που ανήκουν στην τάξη της πλειοψηφίας, θεωρούνται θόρυβος και δεν περιλαμβάνονται στη διαδικασία. Δείγματα το πολύ m / 2 NNs από την τάξη πλειοψηφίας θεωρούνται ασφαλή, ενώ τα δείγματα για τα οποία ο αριθμός των NNs από την τάξη πλειοψηφίας είναι μεγαλύτερος από m <2 (αλλά όχι m) θεωρούνται ότι βρίσκονται σε κίνδυνο. Συνθετικά δείγματα δημιουργούνται μέσω παρεμβολής για δείγματα σε κίνδυνο και παρεκβολή στα ασφαλές δείγματα.

Συνοπτικά, το SVM λειτουργεί με τη χαρτογράφηση των δεδομένων εκπαίδευσης σε ένα χώρο χαρακτηριστικών υψηλής διάστασης. Είναι ένας εποπτευόμενος αλγόριθμος εκμάθησης μηχανής, ο οποίος χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης. Σε αυτόν τον αλγόριθμο, κάθε στοιχείο δεδομένων απεικονίζεται ως σημείο σε ν-διάστατο χώρο (όπου ν είναι ο αριθμός των χαρακτηριστικών) με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Στη συνέχεια, πραγματοποιείται ταξινόμηση, βρίσκοντας το υπέρ-επίπεδο που διαφοροποιεί τις δύο κατηγορίες πολύ καλά. (Εικόνα 7)



Εικόνα 7 Υπέρ-επίπεδο σε 2D και 3D χαρακτηριστικό χώρο

Τα υπερ-επίπεδα είναι όρια απόφασης που βοηθούν στην ταξινόμηση των σημείων δεδομένων. Τα σημεία δεδομένων που εμπίπτουν σε κάθε πλευρά του υπερ-επιπέδου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Επίσης, η διάσταση του υπερ-επιπέδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 2, τότε το υπερ-επίπεδο είναι μόνο μια γραμμή. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 3, τότε το υπερ-επίπεδο γίνεται ένα δισδιάστατο επίπεδο (SRIVASTAVA και BHAMBHU n.d.) (Himani Bhavsar 2012).

4.1.3. BSMOTE (1&2) Borderline SMOTE of types 1 and 2

- Borderline SMOTE – type 1:

  Το bSMOTE1, παράγει συνθετικά δείγματα εφαρμόζοντας τον αλγόριθμο SMOTE, αλλά μόνο σε δείγματα που βρίσκονται κοντά στα σύνορα μεταξύ διαφορετικών τάξεων.

  Αρχικά επιλέγονται οι πλησιέστεροι γείτονες για κάθε δείγμα της μειονοτικής κατηγορίας. Τα δείγματα μειονοτήτων που περιβάλλουν εντελώς τα πλειοψηφικά δείγματα, δηλαδή: όλοι οι πιό πλησιέστεροι γείτονες ανήκουν στην τάξη πλειοψηφίας, θεωρούνται θόρυβος και δεν περιλαμβάνονται στην διαδικασία. Δείγματα, το πολύ m / 2 NNs από την τάξη πλειοψηφίας που θεωρούνται ασφαλή, επίσης δεν περιλαμβάνονται στην διαδικασία.

  Δείγματα για τα οποία ο αριθμός των NNs από την τάξη πλειοψηφίας είναι μεγαλύτερος από m <2 (αλλά όχι m) θεωρούνται σε κίνδυνο (κοντά στο όριο) και χρησιμοποιούνται για τη δημιουργία συνθετικών δειγμάτων. Δημιουργούνται νέα δείγματα μειοψηφίας κατά μήκος των γραμμών που συνδέουν τα δείγματα μειοψηφίας με τους πλησιέστερους γείτονες της μειονότητας.

- Borderline SMOTE – type 2:

  Η τεχνική του bSMOTE2 είναι παρόμοια με την τεχνική του bSMOTE1. Αυτό που διαφέρει το bSMOTE2 από το bSMOTE1 είναι ότι τα συνθετικά δείγματα δημιουργούνται από τους πλησιέστερους γείτονες μειοψηφίας, αλλά και από τους πλησιέστερους γείτονες της πλειονότητας. Ωστόσο, τα συνθετικά δείγματα που δημιουργούνται από τους γείτονες της πλειοψηφίας

δημιουργούνται πιο κοντά στο δείγμα μειοψηφίας, όταν δημιουργούνται από γείτονες μειονοτήτων (Hien M. Nguyen 2011) (Hui Han n.d.)[14].

4.2. Τεχνικές υπό-δειγματοληψίας (Under-sampling) για προβλήματα ταξινόμησης

4.2.1. NearMiss (1 & 2 & 3)

Η μέθοδος NearMiss είναι μια τεχνική υπο-δειγματοληψίας. Αντί να χρησιμοποιεί μέθοδο υπέρ-δειγματοληψίας στην τάξη των μειονοτήτων, χρησιμοποιεί μια απόσταση, αυτό έχει ως αποτέλεσμα η τάξη πλειονότητας να είναι ίση με την τάξη των μειονοτήτων.

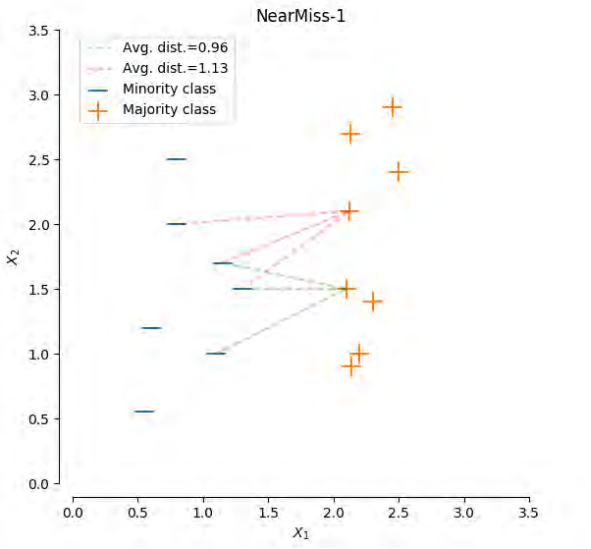
Η μέθοδος NearMiss υπολογίζει τις αποστάσεις μεταξύ όλων των περιπτώσεων της τάξης πλειοψηφίας και των περιπτώσεων της τάξης των μειονοτήτων. Στη συνέχεια επιλέγονται *k* περιπτώσεις της τάξης πλειοψηφίας που έχουν τις μικρότερες αποστάσεις από της τάξης μειονότητας.

Αν υπάρχουν *n* περιπτώσεις στην τάξη μειοψηφίας, η " nearest " μέθοδος θα έχει ως αποτέλεσμα *k* * *n* περιπτώσεις της τάξης πλειοψηφίας.

- NearMiss (1)

Η μέθοδος NearMiss-1 επιλέγει τα θετικά δείγματα για τα οποία η μέση απόσταση από τα πλησιέστερα δείγματα N της αρνητικής κλάσης είναι η μικρότερη. Στην Εικόνα 8, η τιμή *k είναι 3*, με βάση αυτή την τιμή, υπολογίζεται η μέση απόσταση σε 2 συγκεκριμένα δείγματα της τάξης πλειονότητας. Επομένως, στην περίπτωση αυτή το σημείο που συνδέεται με την πράσινη διακεκομμένη γραμμή θα επιλεγεί αφού η μέση απόσταση είναι μικρότερη[15].

[7]



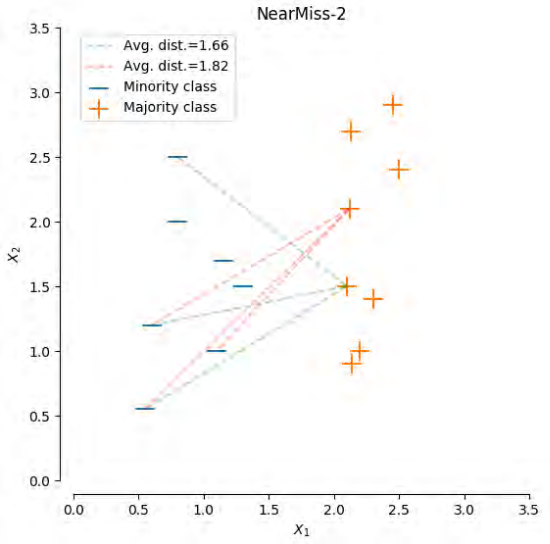
Εικόνα 8 Τεχνική υπό-δειγματοληψίας NearMiss-1

- NearMiss (2)

Η μέθοδος NearMiss-2 επιλέγει δείγματα από την τάξη πλειοψηφίας για τα οποία η μέση απόσταση από τους πιο μακρινούς γείτονες είναι η μικρότερη. Στο παρακάτω γράφημα, η τιμή *k είναι 3*, με βάση αυτή την τιμή, υπολογίζεται η μέση απόσταση σε 2 συγκεκριμένα δείγματα της τάξης πλειονότητας. Επομένως στην περίπτωση αυτή, το δείγμα που συνδέεται με την πράσινη διακεκομμένη γραμμή θα επιλεγεί αφού η απόσταση μεταξύ των 3 πιο μακρινών γειτόνων είναι η μικρότερη[15].

[7] Πηγή Εικόνας: https://imbalanced-learn.readthedocs.io/en/stable/under_sampling.html

[8]

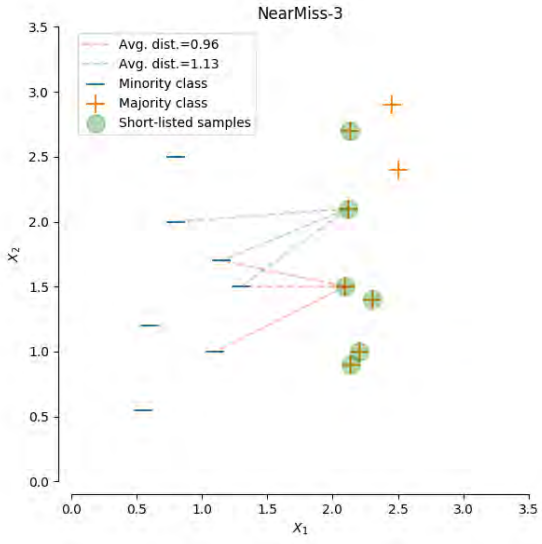


Εικόνα 9 Τεχνική υπό-δειγματοληψίας NearMiss-2

- NearMiss (3)

Στην μέθοδο NearMiss-3, χρησιμοποιείται ο πλησιέστερος γείτονας για τη δειγματοληψία δειγμάτων από την τάξη πλειοψηφίας () Στη συνέχεια, επιλέγεται το δείγμα με τη μεγαλύτερη μέση απόσταση από τους k πλησιέστερους γείτονες [15].

[9]



Εικόνα 10 Τεχνική υπό-δειγματοληψίας NearMiss-3

[8] [9] Πηγή Εικόνας: https://imbalanced-learn.readthedocs.io/en/stable/under_sampling.html

Η εκτέλεση υπό-δειγματοληψίας χρησιμοποιώντας τεχνική NearMiss γίνεται μέσω της παρακάτω μεθόδου:

Class **imblearn.under_sampling.NearMiss**(*sampling_strategy='auto',*
return_indices=False,
random_state=None,
version=1,
n_neighbors=3,
n_neighbors_ver3=3,
n_jobs=1,
ratio=None
) [13]

Περιγραφή μερικών παραμέτρων:

- ***n_neighbors:*** *ο* αριθμός πλησιέστερων γειτόνων. Υπολογίζει την μέση απόσταση από τα δείγματα μειονοτήτων.
- ***n_neighbors_ver3:*** το υποσύνολο των πλησιέστερων γειτόνων που επιλέγουν τα δείγματα

4.3. Διαφορές Υπέρ – δειγματοληψίας και Υπό – δειγματοληψίας

Η υπέρ-δειγματοληψία και η υπό-δειγματοληψία είναι αντίθετες και σχεδόν ισοδύναμες τεχνικές. Και οι δυο χρησιμοποιούνται όταν ο συνολικός αριθμός μιας κατηγορίας δεδομένων (θετικός) είναι πολύ μικρότερος από τον συνολικό αριθμό μιας άλλης κατηγορίας δεδομένων (αρνητική). Κάποιες από τις διαφορές τους είναι:

- η υπερ-δειγματοληψία αυξάνει το μέγεθος των δειγμάτων μειονότητας, ενώ η υπο-δειγματοληψία μειώνει το μέγεθος των δειγμάτων πλειονότητας.
- Με την μέθοδο της υπερ-δειγματοληψίας, τα συνθετικά δείγματα που δημιουργούνται, μπορεί να περιέχουν θόρυβο, το γεγονός αυτό μπορεί να συμβεί εάν οι διαφορετικές κατηγορίες (Classes) δεν χωριστούν καλά. Ενώ αντίθετα, η μέθοδος της υπο-δειγματοληψίας, σε μερικές περιπτώσεις εξαλείφουν το θόρυβο.

5. Principal component analysis – Ανάλυση Κύριων Συνιστωσών (PCA)

Το τεράστιο μέγεθος των δεδομένων στη σύγχρονη εποχή είναι ένα μεγάλο εμπόδιο για την εκτέλεση πολλών αλγορίθμων μηχανικής μάθησης. Ο κύριος στόχος της ανάλυσης PCA είναι να προσδιοριστούν τα πρότυπα στα δεδομένα και ο σκοπός, η ανίχνευση της συσχέτισης μεταξύ των μεταβλητών. Ο PCA είναι μέθοδος γραμμικού μετασχηματισμού.

Η μέθοδος PCA είναι ιδιαίτερα χρήσιμη όταν οι μεταβλητές ενός συνόλου δεδομένων συσχετίζονται σε μεγάλο βαθμό. Η συσχέτιση δείχνει ότι υπάρχει πλεονασμός στα δεδομένα. Λόγω αυτού του πλεονασμού, ο PCA μπορεί να χρησιμοποιηθεί για να μειώσει τις αρχικές μεταβλητές σε μικρότερο αριθμό νέων μεταβλητών (κύριες συνιστώσες) διερμηνεύοντας το μεγαλύτερο μέρος της διακύμανσης στις αρχικές μεταβλητές (Richardson 2009).

Η ακόλουθη εικόνα απεικονίζει το διάγραμμα δραστηριότητας 1 που δείχνει κάθε βήμα του PCA.

Διάγραμμα 1: Διάγραμμα ροής PCA

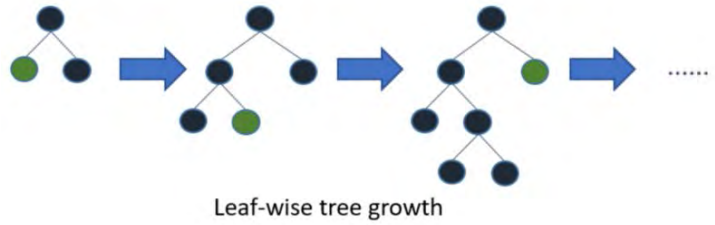
6. Ο αλγόριθμος lightGBM (Light Gradient Boosting Machine)

Ο αλγόριθμος lightGBM χρησιμοποιεί αλγόριθμους μάθησης βασισμένους σε δέντρο. Είναι σχεδιασμένο για να είναι κατανεμημένο και αποδοτικό με τα ακόλουθα πλεονεκτήματα:

- Ταχύτερη ταχύτητα εκπαίδευσης και υψηλότερη απόδοση.
- Χαμηλότερη χρήση μνήμης.
- Καλύτερη ακρίβεια.
- Υποστήριξη παράλληλης μάθησης και μάθησης GPU.
- Δυνατότητα χειρισμού δεδομένων μεγάλης κλίμακας.

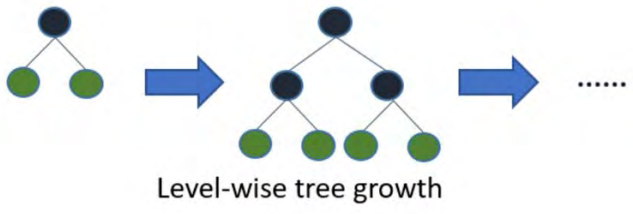
Ο αλγόριθμος lightGBM αναπτύσσει δέντρο κατακόρυφα ενώ άλλοι αλγόριθμοι βασισμένοι σε δέντρα δημιουργούν δέντρα κατά επίπεδο (βάθος), πράγμα που σημαίνει ότι ο lightGBM μεγαλώνει φύλλο – φύλλο (leaf-wise)(Εικόνα 11) ενώ άλλος αλγόριθμος μεγαλώνει σε επίπεδο (Εικόνα 12).

[10]



Εικόνα 11 Αλγόριθμος δέντρου lightGMB

[11]



Εικόνα 12 Αλγόριθμοι δέντρου

[10] [11] Πηγή Εικόνας: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

7. Περιγραφή δεδομένων

7.1. Δεδομένα Οστεοαρθρίτιδας

Τα δεδομένα συγκεντρώθηκαν από την βάση δεδομένων της Πρωτοβουλίας για την Οστεοαρθρίτιδα (Osteoarthritis Initiative (OI)).Η Πρωτοβουλία για την Οστεοαρθρίτιδα, είναι μια εθνική έρευνα που χρηματοδοτείται από τα Εθνικά Ινστιτούτα Υγείας (Τμήμα Υγείας και Ανθρωπίνων Υπηρεσιών). Τα δεδομένα συλλέχθηκαν από ερωτηματολόγιο που τέθηκε σε άντρες και γυναίκες με ή χωρίς συμπτώματα οστεοαρθρίτιδας.

Συγκεκριμένα, η μορφή των δεδομένων είναι ένας πίνακας τιμών, με 4793 γραμμές και 144 στήλες, που αποτελείται από 143 χαρακτηριστικά και 3 κατηγορίες (1,2,3).

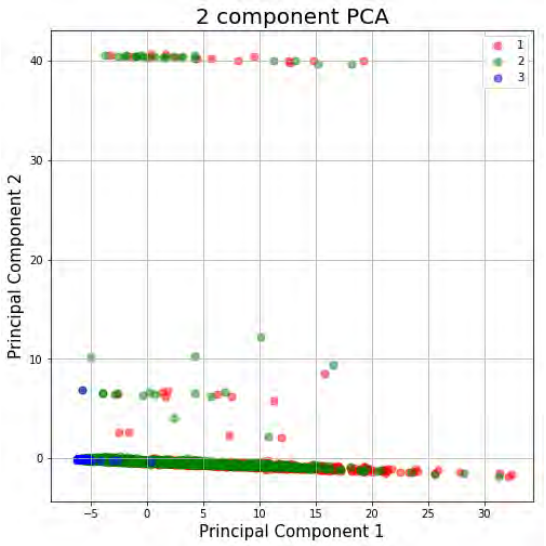
7.2. Προ-επεξεργασία δεδομένων

Ένα απαραίτητο κομμάτι της ανάλυσης του συνόλου δεδομένων αποτελεί και ο καθαρισμός των δεδομένων, κατά τον οποίο εντοπίζονται και διορθώνονται ή αφαιρούνται οι ανακριβείς και κατεστραμμένες τιμές στο σύνολο των δεδομένων (Μητσοτάκης 2018). Στην συγκεκριμένη περίπτωση όπου οι εγγραφές γίνονταν με μη αυτοματοποιημένο τρόπο και χωρίς συγκεκριμένη μορφολογία και τυποποίηση εμφανίστηκαν πολλά κενά πεδία. Όλα τα παραπάνω ως γνωστόν δύναται να επηρεάσουν αρνητικά την πορεία της ανάλυσης και να οδηγήσουν σε ψευδείς κανόνες κατά την διαδικασία της κατηγοριοποίησης. Τα πεδία των ελλιπών τιμών συμπληρώθηκαν με το μέσο όρο της στήλης που αντιπροσωπεύει το χαρακτηριστικό και περιείχαν τα κενά πεδία και από το αρχικό σύνολο δεδομένων αφαιρέθηκε το πεδίο ID των ασθενών λόγω της μη χρησιμότητάς του.

Επιπλέον, υλοποιήσαμε κανονικοποίηση (normalization) των δεδομένων σε N(0,1) και εφαρμόσαμε την μέθοδο Ανάλυση Κύριων Συνιστωσών (PCA), που αναλύσαμε στο Κεφάλαιο 5, για να μειώσουμε τις αρχικές μεταβλητές σε μικρότερο αριθμό νέων μεταβλητών (κύριες συνιστώσες) διερμηνεύοντας το μεγαλύτερο μέρος της διακύμανσης στις αρχικές μεταβλητές. (Γράφημα 1).

Τέλος, χωρίσαμε το σύνολο δεδομένων, σε σύνολο εκπαίδευσης (train set) 70% και σε σύνολο επικύρωσης (test set) 30%.

Όπως αναφέρθηκε παραπάνω, αρχική μορφή των δεδομένων είναι ένας πίνακας τιμών, που περιλαμβάνεται από 4793 γραμμές και 144 στήλες. Με την χρήση της μεθόδου Ανάλυση Κύριων Συνιστωσών (PCA) , η νέα μορφή που δίνεται στον πίνακα των δεδομένων τροποποιείται σε 4793 γραμμές και 2 στήλες (Πίνακας 1). Στο παρακάτω σχήμα παρουσιάζεται η αρχική και τροποποιημένη μορφή του πίνακα δεδομένων.



Γράφημα 1 Χρήση της μεθόδου Ανάλυση Κύριων Συνιστωσών (PCA)

Instance	F1	F2	F3	…	F143	F144
1	1	1	1	…	1	1
2	0	1	0	…	0	0
3	1	1	1	…	0	1
4	1	0	0	…	1	0
5	1	1	1	…	0	0
6	1	0	1	…	0	1
7	1	1	1	…	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4792	0	0	0	…	0	0
4793	0	0	0	…	0	0

PCA (2 Components)

with standardization

Instance	Component 1	Component 2
1	0,039	-0,34
2	-5,816	-0,11
3	0,19	-0,10
4	1,51	-0,49
5	329	-0,45
6	-5,14	-0,12
7	-5,32	-0,17
⋮	⋮	⋮
4792	6,77	-0,61
4793	-2,27	-0,30

Πίνακας 1: Η αρχική και τροποποιημένη μορφή του πίνακα δεδομένων, με την χρήση της μεθόδου Ανάλυση Κύριων Συνιστωσών (PCA)

8. Υλοποίηση Τεχνικών Δειγματοληψίας

8.1. Αποτελέσματα

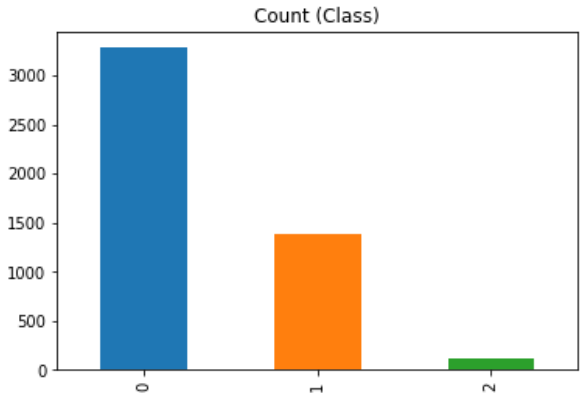
Στο παρόν κεφάλαιο γίνεται μια περιγραφή της διαδικασίας που απαιτήθηκε για την υλοποίηση των τεχνικών δειγματοληψίας που αναλύσαμε στο Κεφάλαιο 4.

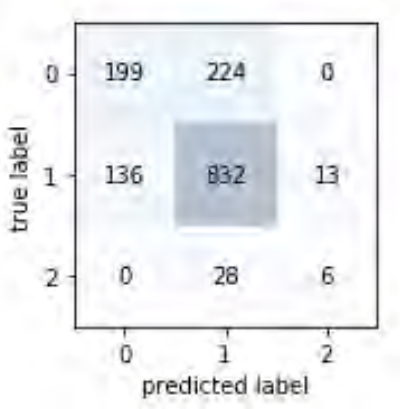
Αρχικά, θα εστιάσουμε την προσοχή μας στα δεδομένα που χρησιμοποιήσαμε και περιγράψαμε στο Κεφάλαιο 7. Όπως αναφέρθηκε,. η μορφή των δεδομένων είναι ένας πίνακας τιμών, με 4793 γραμμές και 144 στήλες, που αποτελούνται από 143 χαρακτηριστικά και 3 κατηγορίες (0,1,2).

Πιο αναλυτικά:

1. η κατηγορία (0) αντιπροσωπεύει 3281 συμμετέχοντες που δεν έχουν συμπτώματα οστεοαρθρίτιδας αλλά βρίσκονται στην ομάδα υψηλού κινδύνου,
2. η κατηγορία (1) αντιπροσωπεύει 1390 συμμετέχοντες, που έχουν συμπτώματα οστεοαρθρίτιδας και
3. η κατηγορία (2) αναφέρεται σε 122 συμμετέχοντες, οι οποίοι δεν έχουν συμπτώματα οστεοαρθρίτιδας και δεν βρίσκονται στην ομάδα υψηλού κινδύνου.

Η κατανομή των τριών κατηγοριών, απεικονίζεται στο Γράφημα 2.





Γράφημα 1 Η αρχική κατανομή των τριών (0,1,2) κατηγοριών

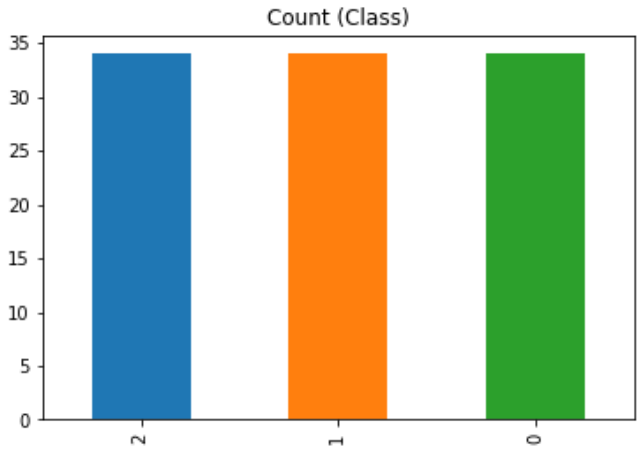
Από τα παραπάνω διαπιστώνουμε ότι, ο συνολικός αριθμός της κατηγορίας (2) είναι πολύ μικρότερος από τον συνολικό αριθμό των άλλων δύο κατηγοριών

(0,1). Το γεγονός αυτό, έχει ως αποτέλεσμα να δημιουργεί μεγάλο πρόβλημα ταξινόμησης. (ανισορροπίας τάξης). Συγκεκριμένα, η κατηγορία (2) που αντιπροσωπεύει συμμετέχοντες, οι οποίοι δεν έχουν συμπτώματα οστεοαρθρίτιδας και δεν βρίσκονται στην ομάδα υψηλού κινδύνου, λόγω του προβλήματος ανισορροπίας τάξης, να κατατάσσεται στην κατηγορία (1), και να κατηγοριοποιείται ως τους συμμετέχοντες, που έχουν συμπτώματα οστεοαρθρίτιδας με ακρίβεια ταξινόμησης 72,11%. Αποδεκτό ποσοστό ταξινόμησης αλλά χωρίς σωστή κατηγοριοποίηση των συμμετεχόντων.

Στην παρούσα διπλωματική εργασία, το συγκεκριμένο πρόβλημα, το προσεγγίσαμε με τις τεχνικές υπό-δειγματοληψίας (*NearMiss under-sampling*) και υπέρ-δειγματοληψίας (*SMOTE over-sampling*).

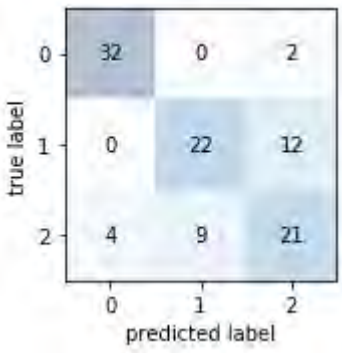
Αρχικά, με την υλοποίηση της τεχνικής **NearMiss-1** υπό-δειγματοληψίας, ο συνολικός αριθμός της κατηγορίας (0) από 423 μειώθηκε στα 34 δείγματα και αντίστοιχα της κατηγορίας (1) από 981 μειώθηκε σε 34 δείγματα. Τέλος, ο συνολικός αριθμός της κατηγορίας (2) παρέμεινε ίδιος (34 δείγματα), ως αποτέλεσμα να έχουμε ομοιόμορφή κατανομή δειγμάτων.

Classes	0	1	2
Before resampling	423	981	34
After resampling	34	34	34



Γράφημα 2 Η κατανομή των τριών (0,1,2) κατηγοριών, μετά την εφαρμογή τεχνικής Υπό-δειγματοληψίας

Από τα 34 δείγματα της κατηγορίας (0), τα 32 δείγματα ταξινομήθηκαν ως κατηγορία (0) και μόνο 2 δείγματα ταξινομήθηκαν ως κατηγορία (2). Αντίστοιχα, από τα 34 δείγματα της κατηγορίας (1), τα 22 δείγματα ταξινομήθηκαν ως κατηγορία (1) και 12 δείγματα ως κατηγορία (2).Επίσης, από τα 34 δείγματα της κατηγορίας (2), τα 4 δείγματα ταξινομήθηκαν ως κατηγορία (0), 9 δείγματα ως κατηγορία (1) και 21 δείγματα ως κατηγορία (2), με ακρίβεια ταξινόμησης *73,53%*.

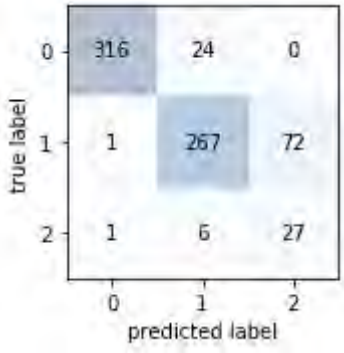


Confusion Matrix (1)
Accuracy 73.53%

(1) Αποτέλεσμα της χρήσης NearMiss-1 με ratio=1, n_neighbors =7 ,n_neighbors_ver3=5

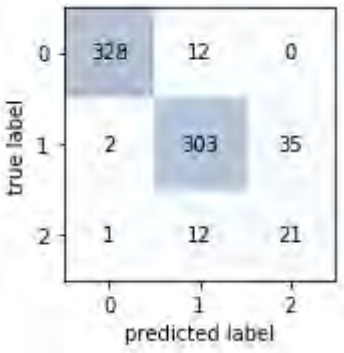
Στην παρούσα τεχνική, ως αριθμός πλησιέστερων γειτόνων (*n_neighbors*) ορίστηκε 7 να υπολογιστεί η μέση απόσταση από τα δείγματα μειονοτήτων και ως αριθμός 5 το υποσύνολο των πλησιέστερων γειτόνων (*n_neighbors_ver3*) που επιλέγουν τα δείγματα. Επιπλέον, με βάση την αναλογία (*ratio*) μειώθηκαν τα δείγματα της πλειοψηφίας.

Με την μέθοδο **NearMiss-2**, τα αποτελέσματα που πήραμε ήταν τα εξής:



Classes	0	1	2
Before resampling	423	981	34
After resampling	340	340	34

Confusion Matrix (2)
Accuracy 85.46%



Classes	0	1	2
Before resampling	423	981	34
After resampling	340	340	34

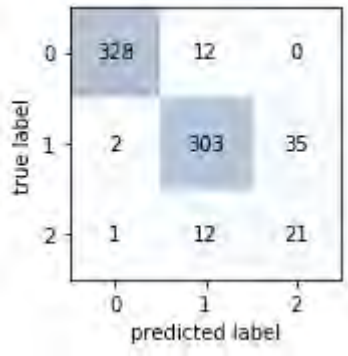
Confusion Matrix (3)
Accuracy 91.32%

(2)Αποτέλεσμα της χρήσης NearMiss-2 με ratio=0,1,n_neighbors=3,n_neighbors_ver3=3
(3)Αποτέλεσμα της χρήσης NearMiss-2 με ratio=0,1,n_neighbors=5,n_neighbors_ver3=3

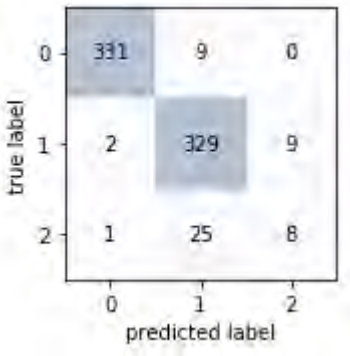
Στην δεύτερη (2) περίπτωση, τα αποτελέσματα έδειξαν ακρίβειά ταξινόμησης *85,46%*, ενώ αυξάνοντας τον αριθμό των πλησιέστερων γειτόνων (*n_neighbors*) στην τρίτη περίπτωση (3), κατά δύο (2), η ακρίβεια ταξινόμησης έδειξε καλύτερα αποτελέσματα, *91,32%*.

Συνεχίζοντας και αυξάνοντας τον αριθμό των πλησιέστερων γειτόνων (*n_neighbors*) κατά δύο (2), συνεχίζει να μας δίνει ακόμη καλύτερα αποτελέσματα ακρίβειας ταξινόμησης *93,57%*.

Η διαφορά είναι ότι στην περίπτωση (3), από τα 34 δείγματα της κατηγορίας (2), τα 21 δείγματα ταξινομήθηκαν ως κατηγορία (2), 12 δείγματα ταξινομήθηκαν ως κατηγορία (1) και μόνο 1 δείγμα ταξινομήθηκε ως κατηγορία (1). Ενώ αντίθετα, στην περίπτωση (3α), από τα 34 δείγματα της κατηγορίας (2), τα μόλις 8 δείγματα ταξινομήθηκαν ως κατηγορία (2), 25 δείγματα ταξινομήθηκαν λάθος στην κατηγορία (1) και 1 δείγμα ταξινομήθηκε ως κατηγορία (0).



Confusion Matrix (3)
Accuracy 91.32%



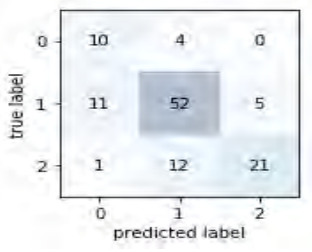
Confusion Matrix (3α)
Accuracy 93.56%

Classes	0	1	2
Before resampling	423	981	34
After resampling	340	340	34

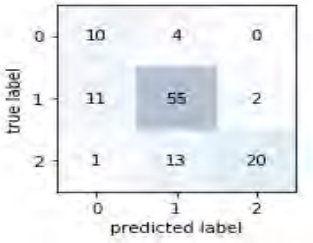
(3)Αποτέλεσμα της χρήσης NearMiss-2 με ratio=0,1,n_neighbors=5,n_neighbors_ver3=3

(3α)Αποτέλεσμα της χρήσης NearMiss-2 με ratio=0,1,n_neighbors=7,n_neighbors_ver3=3

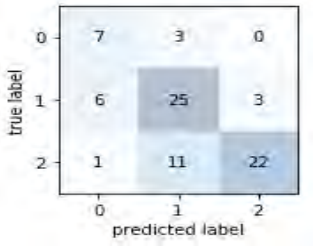
Με την μέθοδο **NearMiss-3**, τα αποτελέσματα που πήραμε ήταν τα εξής:



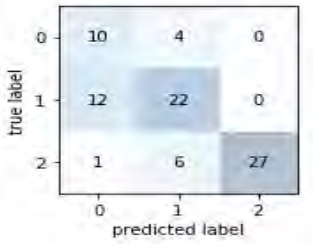
Confusion Matrix (4)
Accuracy 71.55%



Confusion Matrix (5)
Accuracy 73.28%



Confusion Matrix (6)
Accuracy 69.23%



Confusion Matrix (7)
Accuracy 71.95%

Classes	0	1	2
Before resampling	423	981	34
After resampling	14	68	34

Classes	0	1	2
Before resampling	423	981	34
After resampling	14	68	34

Classes	0	1	2
Before resampling	423	981	34
After resampling	10	34	34

Classes	0	1	2
Before resampling	423	981	34
After resampling	14	34	34

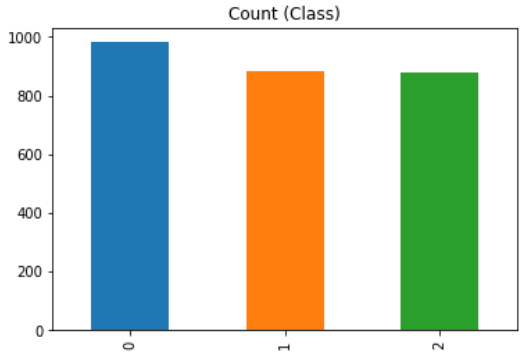
(4) Αποτέλεσμα της χρήσης NearMiss-3 με ratio=0,5,n_neighbors=3,n_neighbors_ver3=7

(5) Αποτέλεσμα της χρήσης NearMiss-3 με ratio=0,5,n_neighbors=5,n_neighbors_ver3=7

(6) Αποτέλεσμα της χρήσης NearMiss-3 με ratio=1,n_neighbors=3,n_neighbors_ver3=5
(7) Αποτέλεσμα της χρήσης NearMiss-3 με ratio=1,n_neighbors=7,n_neighbors_ver3=7

Συγκρίνοντας την περίπτωση τέσσερα (4) με την περίπτωση πέντε (5), βλέπουμε ότι στην περίπτωση πέντε (5), αυξάνοντας τον αριθμό των πλησιέστερων γειτόνων, από *n_neighbors=3* σε *n_neighbors=5*, είχαμε καλύτερη ακρίβεια ταξινόμησης φθάνοντας το *73.28%* έναντι το *71,55%*, με σταθερή και στις δύο περιπτώσεις (4,5), την αναλογία (*ratio*) στο 0,5 και το υποσύνολο των πλησιέστερων γειτόνων (*n_neighbors_ver3)* στο 7. Επιπλέον, η περίπτωση έξι (6), μας έδωσε ακρίβεια ταξινόμησης *69,29%*, ενώ η περίπτωση επτά (7), αυξάνοντας των αριθμό των πλησιέστερων γειτόνων (*n_neighbors*) και το υποσύνολο των πλησιέστερων γειτόνων (*n_neighbors_ver3)* και αφήνοντας σταθερή την αναλογία (*ratio*) στο 1 (και στις δύο περιπτώσεις (6,7)), μας έδωσε ακρίβεια ταξινόμησης *71,95%*.

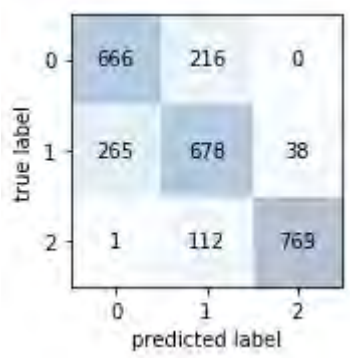
Συνεχίζοντας, με την υλοποίηση της τεχνικής **SMOTE** υπέρ- δειγματοληψίας και συγκεκριμένα με την τεχνική ***SMOTE –svm***, ο συνολικός αριθμός της κατηγορίας (0) από 423 αυξήθηκε σε 882 δείγματα, η κατηγορίας (1) παρέμεινε ίδια σε 981 δείγματα. Αντίστοιχα και της κατηγορίας (2) από 34 δείγματα αυξήθηκε σε 882 δείγματα, ως αποτέλεσμα να έχουμε ομοιόμορφή κατανομή δειγμάτων.



Γράφημα 3 Η κατανομή των τριών (0,1,2) κατηγοριών, μετά την εφαρμογή τεχνικής Υπέρ-δειγματοληψίας

Classes	0	1	2
Before resampling	423	981	34
After resampling	**882**	**981**	**882**

Επιπλέον, από τα 882 δείγματα της κατηγορίας (0), τα 666 δείγματα ταξινομήθηκαν ως κατηγορία (0), τα 216 δείγματα ως κατηγορία (1) και τα 34 δείγματα ως κατηγορία (2). Αντίστοιχα, από τα 981 δείγματα της κατηγορίας (1), τα 265 δείγματα ταξινομήθηκαν ως κατηγορία (0), τα 678 δείγματα ως κατηγορία (1) και 38 δείγματα ως κατηγορία (2). Τέλος, από τα 882 δείγματα της κατηγορίας (2), 1 δείγμα ταξινομήθηκε ως κατηγορία (0), τα 112 δείγματα ως κατηγορία (1) και τα 769 δείγματα ως κατηγορία (2) , με ακρίβεια ταξινόμησης *76,98%*.

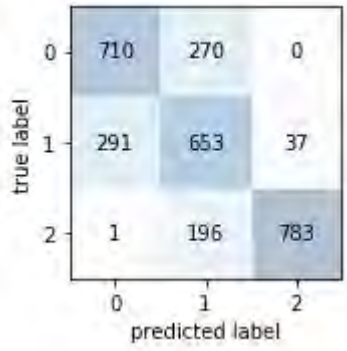


Confusion Matrix (8)

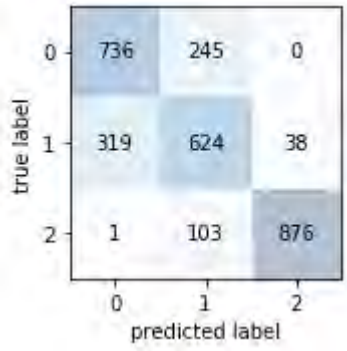
Accuracy 76.98%

(8) Αποτέλεσμα της χρήσης SMOTE-svm, ratio=0.9, k =3, m_neighbors = 3

Στην παρούσα τεχνική, ως αριθμός πλησιέστερων γειτόνων (*k*) ορίστηκε 3, με βάση τους τρείς (3) πλησιέστερους γείτονες επιλέγονται τα δείγματα μειοψηφίας και δημιουργούνται τα συνθετικά δείγματα και ως αριθμός 3 το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*) που επιλέγουν τα δείγματα μειοψηφίας. Επιπλέον, με βάση την αναλογία (*ratio*) αυξήθηκαν τα δείγματα της μειονότητας (δημιουργία συνθετικών δειγμάτων).



Confusion Matrix (9)
Accuracy 72.97%



Confusion Matrix (10)
Accuracy 76%

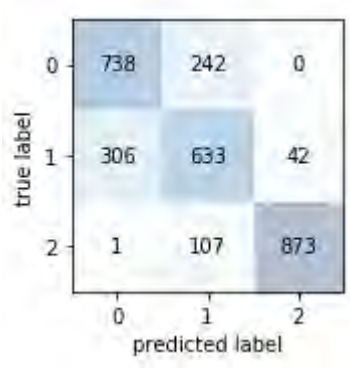
Classes	0	1	2
Before resampling	423	981	34
After resampling	980	981	980

Classes	0	1	2
Before resampling	423	981	34
After resampling	981	981	980

(9) Αποτέλεσμα της χρήσης SMOTE-svm, ratio=1, k =3, m_neighbors = 3

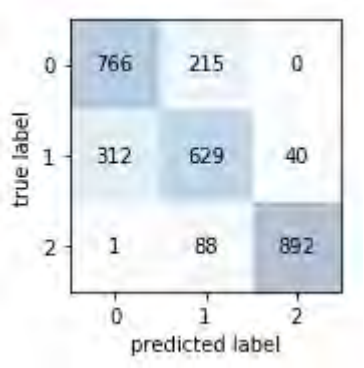
(10) Αποτέλεσμα της χρήσης SMOTE-svm, ratio=1, k =3, m_neighbors = 5

Στην ένατη (9) περίπτωση, αυξάνοντας την αναλογία (*ratio*) κατά 0,1 το ποσοστό ακρίβειας ταξινόμησης μειώθηκε κατά 4% σε σχέση με την περίπτωση οκτώ (8) ενώ στη δέκατη (10) περίπτωση αυξάνοντας κατά 2 το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*), παραμένοντας σταθερά η αναλογία (*ratio*) και ο αριθμός πλησιέστερων γειτόνων (*k*) , το ποσοστό ακρίβειας ταξινόμησης αυξήθηκε κατά 4% σε σχέση με την περίπτωση εννέα (9), πλησιάζοντας την περίπτωση οκτώ (8).



Confusion Matrix (11)

Accuracy 76.27%



Confusion Matrix (12)

Accuracy 77.71%

Classes	0	1	2
Before resampling	423	981	34
After resampling	980	981	981

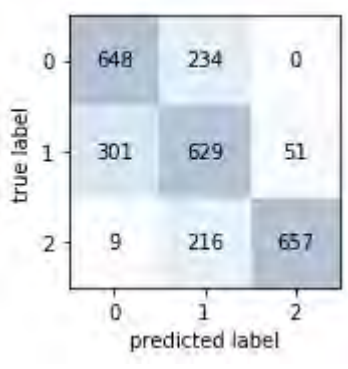
Classes	0	1	2
Before resampling	423	981	34
After resampling	981	981	981

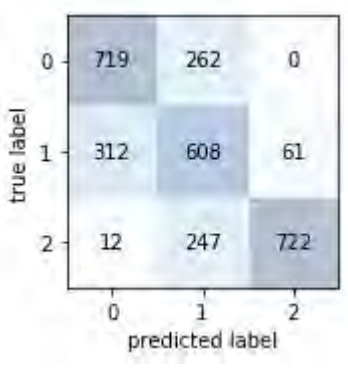
(11) Αποτέλεσμα της χρήσης SMOTE-svm, ratio=1, k =5, m_neighbors = 7

(12) Αποτέλεσμα της χρήσης SMOTE-svm, ratio=1, k =5, m_neighbors = 5

Συνεχίζοντας με την περίπτωση έντεκα (11) και με την περίπτωση δώδεκα (12), βλέπουμε ότι, αυξάνοντας τον αριθμό των πλησιέστερων γειτόνων (*k*) και το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*), παραμένοντας σταθερά η αναλογία (*ratio*), σε σχέση με τις περιπτώσεις εννέα (9) και δέκα (10), το ποσοστό ακρίβειας ταξινόμησης συνεχίζει και αυξάνεται, πλησιάζοντας στο *76,27%* και *77,71%*, αντίστοιχα.

Με την μέθοδο ***SMOTE-regular***, τα αποτελέσματα που πήραμε ήταν τα εξής:





Confusion Matrix (13)
Accuracy 70.46%

Classes	0	1	2
Before resampling	423	981	34
After resampling	882	981	882

Confusion Matrix (14)
Accuracy 69.62%

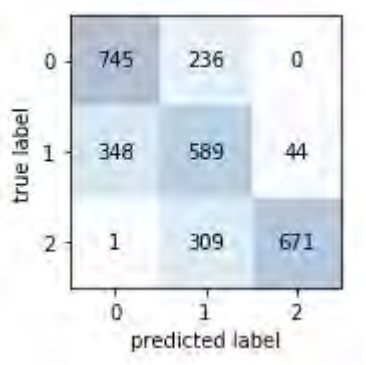
Classes	0	1	2
Before resampling	423	981	34
After resampling	981	981	981

(13) Αποτέλεσμα της χρήσης SMOTE-regular, ratio=1, k =3, m_neighbors = 3

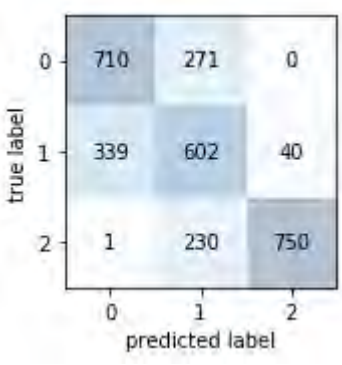
(14) Αποτέλεσμα της χρήσης SMOTE-regular, ratio=1, k =3, m_neighbors = 5

Στην παρούσα τεχνική και συγκεκριμένα στην περίπτωση δεκατρία (13), ως αριθμός πλησιέστερων γειτόνων (*k*) και ως υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*) ορίστηκε 3. Επιπλέον, με βάση την αναλογία (*ratio*) 1, αυξήθηκαν τα δείγματα της μειονότητας (δημιουργία συνθετικών δειγμάτων).Η συγκεκριμένη περίπτωση, έδωσε ποσοστό ακρίβειας ταξινόμησης *70,46%*, αυξάνοντας με συνθετικά δείγματα την κατηγορία (2), από 34 δείγματα σε 882 δείγματα, όπου από τα συγκεκριμένα 882 δείγματα, τα 9 δείγματα ταξινομήθηκαν ως κατηγορία (0), 216 δείγματα ως κατηγορία (1) και 657 ως κατηγορία (2). Αντίθετα με την περίπτωση δεκατέσσερα (14), αυξάνοντας το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*) σε 5, σε σχέση με το υποσύνολο των πλησιέστερων γειτόνων της κατηγορίας δεκατρία (13), το ποσοστό ακρίβειας ταξινόμησης μειώθηκε κατά λίγο, στο *69,72%*.

Με την μέθοδο ***SMOTE-borderline1***, τα αποτελέσματα που πήραμε ήταν τα εξής:



Confusion Matrix (15)
Accuracy 68.13%



Confusion Matrix (16)
Accuracy 70.06%

Classes	0	1	2
Before resampling	423	981	34
After resampling	980	981	981

Classes	0	1	2
Before resampling	423	981	34
After resampling	981	981	981

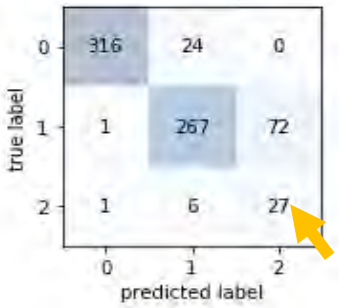
(15) Αποτέλεσμα της χρήσης SMOTE-borderline1, ratio=1, k =5, m_neighbors = 7

(16) Αποτέλεσμα της χρήσης SMOTE-borderline1, ratio=1, k =7, m_neighbors = 5

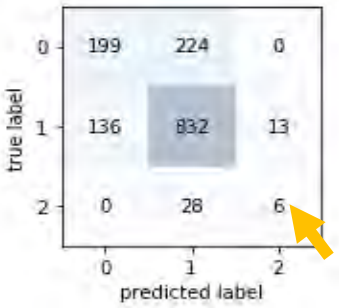
Η τεχνική ***SMOTE-borderline1,*** έδωσε ποσοστό ακρίβεια ταξινόμησης, *68,13%* στην περίπτωση δεκαπέντε (15) και *70,06%* στην περίπτωση δεκαέξι (16). Πιο αναλυτικά, στην περίπτωση δεκαπέντε (15) ως αριθμός πλησιέστερων γειτόνων (*k*) ορίστηκε 5, με βάση τους πέντε (5) πλησιέστερους γείτονες επιλέγονται τα δείγματα μειοψηφίας και δημιουργούνται τα συνθετικά δείγματα και ως αριθμός 7 το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*) που επιλέγουν τα δείγματα μειοψηφίας. Επιπλέον, με βάση την αναλογία (*ratio*) 1, αυξήθηκαν τα δείγματα της μειονότητας (δημιουργία συνθετικών δειγμάτων). Αντίθετα, στην περίπτωση δεκαέξι (16) ως αριθμός πλησιέστερων γειτόνων (*k*) ορίστηκε 7 και ως αριθμός 5 το υποσύνολο των πλησιέστερων γειτόνων (*m_neighbors*), με σταθερή την αναλογία (*ratio*) 1 και στις δύο περιπτώσεις (15,16).

8.2 Σύγκριση Αποτελεσμάτων

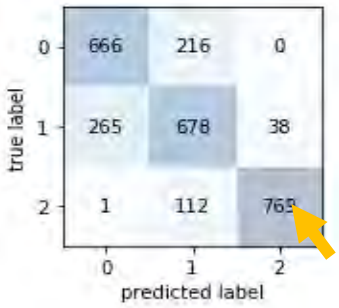
Και στις δύο τεχνικές (υπέρ-δειγματοληψία, *Confusion Matrix (17)*, υπό-δειγματοληψία, *Confusion Matrix (19)*), η κατηγορία (2) ταξινομήθηκε αρκετά καλά, σε σχέση με την αρχική της κατηγοριοποίηση *Confusion Matrix (18).*



Confusion Matrix (17): Χρήση τεχνικής υπό-δειγματοληψίας (Under-sampling), 85,43%



Confusion Matrix (18): Αρχική Μορφή, 72,11%



Confusion Matrix (19): Χρήση τεχνικής υπέρ-δειγματοληψίας (Over-sampling), 76,98%

Επιπλέον, με την μέθοδο της υπέρ–δειγματοληψίας (*Over-sampling*), τα συνθετικά δείγματα που δημιουργήθηκαν, μπορεί να περιείχαν θόρυβο, εάν πάρουμε υπόψιν μας την ακρίβεια ταξινόμησης, ενώ αντίθετα με την μέθοδο υπό-δειγματοληψίας, μπορεί να εξάλειψαν τον θόρυβο.

9. Χρονοσειρές

Η ανάλυση χρονοσειρών αποτελεί μια αναγκαία και από τις πλέον βασικές λειτουργίες της διοίκησης των επιχειρήσεων ενώ η πρόβλεψη χρονοσειρών αποτελεί μία απαραίτητη πηγή πληροφόρησης , η οποία υποστηρίζει τη λήψη αποφάσεων (Μαργιά 2009). Βασίζεται στην επεξεργασία διαθέσιμων δεδομένων και εξάγει την απαραίτητη πληροφορία προκειμένου να παρθούν οι κατά δύναμιν σωστές αποφάσεις. Με τον όρο χρονοσειρά (time series) εννοούμε μία σειρά από δεδομένα τα οποία συλλέγοντα διαχρονικά και εκφράζουν την εξέλιξη μιας μεταβλητής κατά τη διάρκεια ίσων χρονικών στιγμών ή περιόδων (Chan, 2002).

Πιο συγκεκριμένα, η χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων μιας μεταβλητής της οποίας οι τιμές αναφέρονται με βάση κάποια χρονική περίοδο π.χ. μέρα, εβδομάδα, μήνας, τρίμηνο κ.α. Μαθηματικά η χρονοσειρά ορίζεται από ένα δείγμα τιμών Y1,Y2,…,Yt της μεταβλητής Y για κάθε χρονική στιγμή t. Επομένως, το Y είναι μία συνάρτηση του t, και αυτό συμβολίζεται ως Y=f(t) ενώ η γραφική παράσταση της εν λόγω συνάρτησης παρουσιάζει την εξέλιξη της μεταβλητής Y στο χρόνο (Ζγέρα Φεβρουάριος 2016).

9.1. Συνεχείς – Διακριτές Χρονοσειρές

Οι χρονοσειρές διακρίνονται σε συνεχείς και διακριτές. **Συνεχής (continuous)** χρονοσειρές αποτελούν αυτές των οποίων η τιμή Y(t) παρατηρείται συνεχώς και χωρίς διακοπή, όπως για παράδειγμα η καταγραφή της θερμοκρασίας. Αντίθετα, **διακριτές (discrete)** χρονοσειρές αποτελούν αυτές των οποίων η τιμή Y(t) παρατηρείται για ορισμένα χρονικά διαστήματα και όχι συνεχώς, όπως για παράδειγμα η ημερήσια τιμή μίας μετοχής (Ζγέρα Φεβρουάριος 2016).

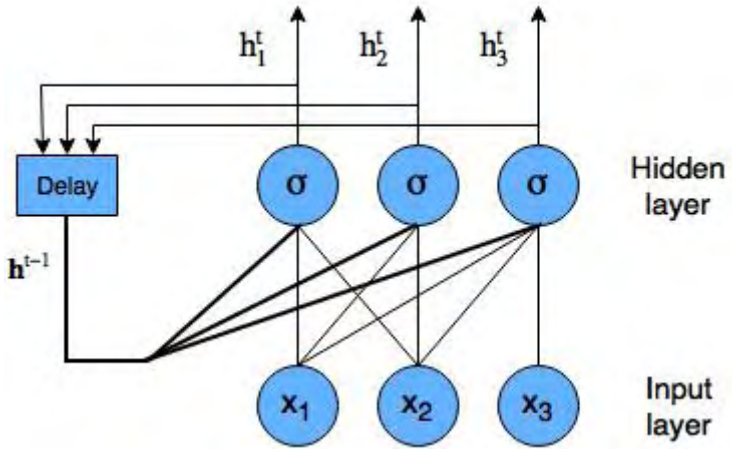
10. Μοντέλο Πρόβλεψης

Η επιλογή ενός κατάλληλου μοντέλου είναι εξαιρετικά σημαντική καθώς θα πρέπει να προσαρμοστεί με την χρονοσειρά για να χρησιμοποιηθεί για μελλοντικές προβλέψεις.

Για την καλύτερη πρόβλεψη και διεξαγωγή αποτελεσμάτων στην παρούσα διπλωματική εργασία υλοποιήσαμε το τεχνητό νευρωνικό δίκτυο για την πρόβλεψη χρονοσειρών (LSTM).

10.1. Το μοντέλο LSTM (Long Short-Term Memory)

Ένα δίκτυο LSTM είναι ένα είδος επαναλαμβανόμενου νευρικού δικτύου. Ένα επαναλαμβανόμενο νευρωνικό δίκτυο είναι ένα νευρωνικό δίκτυο που προσπαθεί να μοντελοποιήσει τη συμπεριφορά που εξαρτάται από το χρόνο ή τη σειρά - όπως οι τιμές των μετοχών, η ζήτηση ηλεκτρικής ενέργειας, καταναλώσεις φυσικού αερίου και ούτω καθεξής. Αυτό γίνεται με την επιστροφή στην έξοδο ενός στρώματος νευρικού δικτύου κατά τη χρονική στιγμή t στην είσοδο του ίδιου στρώματος δικτύου στο χρόνο t + 1 (Azzouni and Pujolle 2017). Όπως φαίνεται στην παρακάτω Εικόνα 1



Εικόνα 13 : Επαναλαμβανόμενο διάγραμμα νευρωνικού δικτύου με εμφανιζόμενους κόμβους

Γενικότερα, μπορούμε να πούμε ότι, ένα LSTM είναι κατάλληλο για την ταξινόμηση, τη επεξεργασία και την πρόβλεψη χρονολογικών σειρών δεδομένου χρονικού διαστήματος άγνωστου μεγέθους και διάρκειας μεταξύ σημαντικών γεγονότων.

11. Περιγραφή δεδομένων

Τα δεδομένα που συγκεντρώθηκαν με στόχο την επεξεργασία και ανάλυσή τους αφορούν πραγματικά δεδομένα από καταναλώσεις φυσικού αερίου στην Ελλάδα, για κάθε πόλη ξεχωριστά (Λάρισα, Τρίκαλα, Βόλος, Αθήνα, Θεσσαλονίκη, Καρδίτσα, Δράμα). Οι τιμές αυτές είναι σε μορφή πίνακα. Συγκεκριμένα, χρησιμοποιήθηκαν χρονοσειρές με ημερήσια δεδομένα ζήτησης φυσικού αερίου ανά μεμονωμένα σημεία εισόδου / εξόδου φυσικού αέριου, ανά κατηγορία καταναλωτών και ανά γεωγραφικό διαμέρισμα. Ενδεικτικά παρακάτω, απεικονίζεται δείγμα δεδομένων από την περιοχή Λάρισα.

Λάρισα

year	date	month	day	temp	L(d-1)	L(d-2)	L(d)
2013	1	1	2	8,7	2788,587	2806,28	2403,629
2013	2	1	3	7,2	2403,629	2788,587	2826,543
2013	3	1	4	7,3	2826,543	2403,629	3116,637
2013	4	1	5	4,5	3116,637	2826,543	3436,761
2013	5	1	6	5,4	3436,761	3116,637	3254,886
2013	6	1	7	7,1	3254,886	3436,761	3436,252
2013	7	1	1	4,2	3436,252	3254,886	4275,57
2013	8	1	2	1,9	4275,57	3436,252	4705,642
2013	9	1	3	0,2	4705,642	4275,57	5010,653
2013	10	1	4	2,7	5010,653	4705,642	4466,275

Πίνακας 2 : Δείγμα Δεδομένων Λάρισας

Στις στήλες *L(d-1), L(d-2)* και *L(d)* παρουσιάζονται οι πραγματικές μετρήσεις κατανάλωσης φυσικού αερίου για τρεις συνεχόμενες ημέρες του αντίστοιχου έτους, ημέρας, μήνα (στήλες *year*, *date*, *month*) με την αντίστοιχη θερμοκρασία (στήλη *temp*).

11.1. Προ-επεξεργασία δεδομένων

Ένα απαραίτητο κομμάτι της ανάλυσης του συνόλου δεδομένων αποτελεί και ο καθαρισμός των δεδομένων, κατά τον οποίο εντοπίζονται και διορθώνονται ή αφαιρούνται οι ανακριβείς και κατεστραμμένες τιμές στο σύνολο των δεδομένων (Μητσοτάκης 2018). Στην συγκεκριμένη περίπτωση, από το αρχικό σύνολο δεδομένων αφαιρέθηκαν τα πεδία *year, date, month, day* και *temp*. Στην παρούσα φάση, μας ενδιαφέρουν οι τιμές που παρήχθησαν τις συγκεκριμένες μέρες *(L(d-1), L(d-2), L(d))* με βάση τα πεδία (*year, date, month, day, temp*) και όχι οι τιμές αυτών συγκεκριμένων πεδίων.

12. Υλοποίηση LSTM

12.1. Αποτελέσματα

Στο παρόν κεφάλαιο γίνεται μια περιγραφή της διαδικασίας που απαιτήθηκε για την υλοποίηση του μοντέλου LSTM που αναλύσαμε στο Κεφάλαιο 10, για την πρόβλεψη κατανάλωσης φυσικού αερίου της επόμενης ημέρας για τις πόλεις Λάρισα και Τρίκαλα, Βόλος, Αθήνα, Θεσσαλονίκη, Καρδίτσα και Δράμα, σε γλώσσα Python.

Χρησιμοποιήσαμε τα δεδομένα των τριών (3) ημερών (L(d-2), L(d-1), L(d)) και προβλέψαμε τις τιμές της τέταρτης (4) ημέρας (L(d+1)). Στην συνέχεια συγκρίναμε τις προβλεπόμενες τιμές της τέταρτης (4) ημέρας (L(d+1)) με τις πραγματικές τιμές της τέταρτης (4) ημέρας (L(d+1)) και αξιολογήσαμε το μοντέλο με την βοήθεια των κριτηρίων αξιολόγησης των μεθόδων προβλέψεων και συγκεκριμένα με τον συντελεστή συσχέτισης (R2).

Πιο αναλυτικά, μετατρέψαμε τις αριθμητικές τιμές σε πιο «κατάλληλες», αριθμητικές τιμές, χρησιμοποιώντας την συνάρτηση MinMaxScaler(feature_range=(0, 1) .Αυτό που πετύχαμε είναι η κανονικοποίηση. Η κανονικοποίηση των δεδομένων πραγματοποιήθηκε ώστε να αντιμετωπιστούν δυσκολίες ορισμένων μεθόδων εξόρυξης. Για παράδειγμα, τα Νευρωνικά Δίκτυα λειτουργούν καλύτερα όταν οι τιμές εισόδου κυμαίνονται στην περιοχή [0,0..1,0].

Στην συνέχεια, ορίσαμε την παρακάτω συνάρτηση και δημιουργήσαμε ένα νέο σετ δεδομένων. Η συνάρτηση create_dataset, παίρνει δύο παραμέτρους, το σετ

δεδομένο μας και το look_back, που είναι ο αριθμός των προηγούμενων βημάτων χρόνου που θα χρησιμοποιηθούν ως μεταβλητές εισόδου για την πρόβλεψη της επόμενης χρονικής περιόδου.

```
def create_dataset(dataset, look_back):
    dataX, dataY = [], []
    for i in range(len(dataset)-look_back-1):
        a = dataset[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataset[i + look_back, 0])
    return numpy.array(dataX), numpy.array(dataY)
```

Έπειτα επιλέχθηκε το παρακάτω υπόδειγμα LSTM για την πρόβλεψη των τιμών χρονοσειράς για την τέταρτη (4)ημέρα (*L(d+1)*), ως το καλύτερο μ με την βοήθεια των κριτηρίων αξιολόγησης των μεθόδων προβλέψεων. Εφαρμόσαμε 2 ορατά στρώματα με 2 εισόδους, 150 κρυφά στρώματα (νευρωνικά δίκτυα) και 1 στρώμα εξόδου που μας δίνει την ενιαία πρόβλεψη τιμής

```
model = Sequential()
model.add(Dense(2, input_shape=(look_back, 1), activation='relu'))
model.add(LSTM(150, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(1, activation='relu'))

model.compile(loss='mean_squared_error', optimizer='adam',metrics=['mae','mape','mse','acc'])

history = model.fit(trainX, trainY,
                    epochs=200,
                    batch_size=100,
                    verbose=0,
                    callbacks = [keras.callbacks.EarlyStopping(monitor='val_loss',
                                 min_delta=0, patience=0, verbose=0, mode='auto')])
```

Στους παρακάτω συγκεντρωτικούς πίνακες (3-9), απεικονίζονται αναλυτικά,

- οι πραγματικές μετρήσεις κατανάλωσης φυσικού αερίου για πόλεις της Ελλάδας (Λάρισα, Τρίκαλα, Βόλος, Αθήνα, Θεσσαλονίκη, Καρδίτσα και Δράμα) των τριών ημερών (*L(d-2), L(d-1), L(d)*)
- οι αναμενόμενες (*Expected*) τιμές για την τέταρτη ημέρα (*L(d+1)*) και
- οι προβλεπόμενες (*Predicted*) τιμές για την τέταρτη ημέρα (*L(d+1)*).

Θεσσαλονίκη

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
11880708,97	11223296,42	10173186,98	12459988.52	9999858.00
10173186,98	11880708,97	12459988,52	13072803.18	12174465.00
12459988,52	10173186,98	13072803,18	12712448.36	12745342.00
13072803,18	12459988,52	12712448,36	10782432.75	12409700.00
12712448,36	13072803,18	10782432,75	13578651.37	10597548.00
10782432,75	12712448,36	13578651,37	18635176.63	13215989.00
13578651,37	10782432,75	18635176,63	21386352.20	17834918.00
18635176,63	13578651,37	21386352,2	21881362.46	20329144.00
21386352,2	18635176,63	21881362,46	19163100.34	20776598.00
21881362,46	21386352,2	19163100,34	15992568.03	18314496.00

Πίνακας 3

Βόλος

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
2601,028	2897,218	2401,32	2917.17	2357.46
2401,32	2601,028	2917,167	2892.94	2817.06
2917,167	2401,32	2892,939	2945.15	2795.05
2892,939	2917,167	2945,148	2855.35	2842.53
2945,148	2892,939	2855,348	3015.93	2761.01
2855,348	2945,148	3015,925	3946.83	2907.26
3015,925	2855,348	3946,827	3886.96	3797.94
3946,827	3015,925	3886,961	3928.65	3738.47
3886,961	3946,827	3928,651	3680.63	3779.85
3928,651	3886,961	3680,627	2997.18	3535.79

Πίνακας 4

Τρίκαλα

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
250,8414	251,9867	207,6641	246.38	210.12
207,6641	250,8414	246,3753	251.41	245.52
246,3753	207,6641	251,4148	230.94	250.15
251,4148	246,3753	230,9362	241.48	231.37
230,9362	251,4148	241,4796	258.31	241.03
241,4796	230,9362	258,3103	317.59	256.50
258,3103	241,4796	317,5921	367.39	311.47
317,5921	258,3103	367,3874	397.11	358.23
367,3874	317,5921	397,106	359.52	386.40
397,106	367,3874	359,5206	308.05	350.81

Πίνακας 5

Αθήνα

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
16571246	15604451	13957553	15242989.49	13854189.00
13957553	16571246	15242989	15202044.65	14958005.00
15242989	13957553	15202045	14483372.57	14922752.00
15202045	15242989	14483373	14476894.89	14304099.00
14483373	15202045	14476895	18338086.32	14298539.00
14476895	14483373	18338086	21490875.81	17660874.00
18338086	14476895	21490876	26925076.92	20499236.00
21490876	18338086	26925077	24812014.57	25599090.00
26925077	21490876	24812015	21883573.49	23584558.00
24812015	26925077	21883573	17118707.37	20858922.00

Πίνακας 6

Καρδίτσα

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
337,5599	342,8876	274,4727	331.41	275.95
274,4727	337,5599	331,4132	337.18	328.94
331,4132	274,4727	337,1788	305.65	334.23
337,1788	331,4132	305,6492	340.82	305.31
305,6492	337,1788	340,8156	370.98	337.58
340,8156	305,6492	370,9765	449.97	365.38
370,9765	340,8156	449,9732	523.85	437.65
449,9732	370,9765	523,8521	564.38	505.69
523,8521	449,9732	564,3764	504.33	543.13
564,3764	523,8521	504,3319	435.02	487.65

Πίνακας 7

Δράμα

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
474345,6	547736,7	506788,5	562432.90	497467.44
506788,5	474345,6	562432,9	586070.83	543728.06
562432,9	506788,5	586070,8	566606.07	563918.69
586070,8	562432,9	566606,1	536050.21	547269.12
566606,1	586070,8	536050,2	567818.89	521573.62
536050,2	566606,1	567818,9	715196.82	548300.12
567818,9	536050,2	715196,8	669813.71	680011.62
715196,8	567818,9	669813,7	652293.37	638079.75
669813,7	715196,8	652293,4	480505.24	622221.50
652293,4	669813,7	480505,2	548861.60	476228.91

Πίνακας 8

Λάρισα

L(d-1)	L(d-2)	L(d)	Expected(day+1)	Predicted(day+1)
2788,587	2806,28	2403,629	2826.54	2381.44
2403,629	2788,587	2826,543	3116.64	2777.92
2826,543	2403,629	3116,637	3436.76	3051.71
3116,637	2826,543	3436,761	3254.89	3354.10
3436,761	3116,637	3254,886	3436.25	3182.33
3254,886	3436,761	3436,252	4275.57	3353.62
3436,252	3254,886	4275,57	4705.64	4139.55
4275,57	3436,252	4705,642	5010.65	4545.39
4705,642	4275,57	5010,653	4466.28	4835.27
5010,653	4705,642	4466,275	3723.43	4319.09

Πίνακας 9

Για την αξιολόγηση του υποδείγματος χρησιμοποιήσαμε τα κριτήρια αξιολόγησης του μοντέλου πρόβλεψης. Βρήκαμε τις αποκλίσεις των προβλεπόμενων τιμών από τις πραγματικές τιμές και υπολογίσαμε τις τιμές των κριτηρίων R^2, MAE, RMSE, MAPE και MSE (για τις προβλεπόμενες τιμές). Τα αποτελέσματα που εξήχθησαν, απεικονίζονται στον Πίνακα 10.

Μετρικές Αξιολόγησης	Λάρισα	Τρίκαλα	Βόλος	Αθήνα	Θεσ/νίκη	Καρδίτσα	Δράμα
R^2	0,97	0,96	0,88	0,96	0,95	0,97	0,88
MAE	0,03	0,03	0,04	0,018	0,034	0,03	0,045
RMSE	0,05	0,05	0,07	0,03	0,05	0,05	0,06
MAPE	234,267	317,689	88,424	324,113	230,166	-	31,168
MSE	0,002	0,003	0,005	0,001	0,003	0,003	0,003

Πίνακας 10 Τιμές Κριτηρίων ου μοντέλου πρόβλεψης

Από τον πίνακα 10, απορρέει το συμπέρασμα πως οι προβλέψεις είναι κοντά στην πραγματική τους αξία και ότι είναι αμερόληπτες. Επιπλέον, ο συντελεστής συσχέτισης R^2 ανέρχεται ως και 97% στις πόλεις Λάρισα και Καρδίτσα, ακολουθούν τα Τρίκαλα και η Αθήνα με 96%, η Θεσσαλονίκη με 95% και με 88% η Δράμα και ο Βόλος.

Καταλήγοντας, μπορούμε να πούμε ότι, οι τιμές των κριτηρίων είναι αρκετά καλές και για τις εφτά (7) πόλεις της Ελλάδας (Λάρισα, Τρίκαλα, Βόλος, Αθήνα, Θεσσαλονίκη, Καρδίτσα και Δράμα) και δείχνουν ότι οι προβλεπόμενες τιμές με το μοντέλο LSTM πλησιάζουν αρκετά τις πραγματικές τιμές της χρονοσειράς.

13. Συμπεράσματα και Μελλοντικές Προοπτικές

Στην παρούσα διπλωματική εργασία μελετήσαμε το πρόβλημα μη–ισορροπημένων δεδομένων. Η προσέγγιση πραγματοποιήθηκε σε πραγματικά δεδομένα που συγκεντρώθηκαν από την βάση δεδομένων της Πρωτοβουλίας για την Οστεοαρθρίτιδα (Osteoarthritis Initiative (OI)).Ο συνολικός αριθμός της μίας κατηγορία ήταν πολύ μικρότερος από τον συνολικό αριθμό των άλλων κατηγοριών. Το συγκεκριμένο πρόβλημα το αντιμετωπίσαμε με τις τεχνικές υπέρ-δειγματοληψίας και υπό-δειγματοληψίας. Λόγω του συγκεκριμένου προβλήματος, καταλήξαμε στο γεγονός ότι ένας αλγόριθμος εκμάθησης μηχανής λειτουργεί καλύτερα όταν ο

αριθμός των περιπτώσεων κάθε κλάσης είναι σχεδόν ίσος. Όταν ο αριθμός της κλάσης υπερβαίνει κατά πολύ την άλλη, προκύπτουν προβλήματα.

Συμπληρώνοντας, μέσω της εφαρμογής από τις τεχνικές υπέρ-δειγματοληψίας και υπό-δειγματοληψίας καταλήξαμε στο συμπέρασμα ότι δεν μπορούν όλες οι τεχνικές να είναι κατάλληλες σε κάθε περίπτωση.

Επιπλέον, η επιλογή των διάφορων τεχνικών υπέρ-δειγματοληψίας και υπό-δειγματοληψίας σε σύνθετους τύπους δεδομένων πρέπει να γίνετε προσεκτικά ανάλογα με το τι είναι το ζητούμενο, ποιο είναι το δείγμα αλλά και με βάση το πώς είναι η δομή των δεδομένων που θα εξεταστούν.

Ακόμα, στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας που αφορά την ακρίβεια πρόβλεψης του φυσικού αερίου σε ημερήσια βάση, παίρνοντας υπόψιν μας τον συντελεστή συσχέτισης R^2 συμπεραίνουμε ότι οι προβλεπόμενες τιμές ήταν κοντά στις αναμενόμενες τιμές. Η προσέγγιση έγινε με Τεχνικό Νευρωνικό Δίκτυο και συγκεκριμένα με το LSTM (Long short-term memory) Τα αποτελέσματα της Λάρισας και της Καρδίτσας ήταν από τα καλύτερα, άνω του 0,9, φθάνοντας στο 0,97. Εντυπωσιακά, ήταν και τα αποτελέσματα για τα Τρίκαλα, την Αθήνα και την Θεσσαλονίκη, πλησιάζοντας το 0,96 και 0,95 αντίστοιχα. Τέλος, ο Βόλος και η Δράμα, είχαν αρκετά καλά αποτελέσματα αλλά όχι άνω του 0,9, το αποτέλεσμα κυμάνθηκε στο 0,88.

Κλείνοντας, σκοπός της πρόβλεψης δεν είναι να βγαίνει πάντα σωστή και να μαντεύει το αποτέλεσμα, πράγμα έως τώρα αδύνατο αλλά να μας δίνει τις περισσότερες φορές την ουσία της μελλοντικής εικόνας. Δηλαδή αν θα έχουμε μεγάλη ζήτηση φυσικού αερίου ή όχι και πόσο κατά προσέγγιση έτσι ώστε να είμαστε έτοιμοι για το απρόβλεπτο μέλλον.

Βιβλιογραφία

Azzouni, Abdelhadi, και Guy Pujolle. «A Long Short-Term Memory Recurrent Neural.» 8 June 2017.

Barua, Sukarna, Md. Monirul Islam, Xin Yao, και Kazuyuki Murase. «MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning.» *IEEE Transactions on Knowledge and Data Engineering*, 26 November 26 2012 : 405.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, και W. Philip Kegelmeyer. «SMOTE: Synthetic Minority Over-sampling Technique.» *Journal of Artificial Intelligence Research 16*, 2002: 321–357.

Chen, Chao, Andy Liaw, και Leo Breiman. «Using Random Forest to Learn Imbalanced Data.» July 2004.

Dr.D.Ramyachitra, και P.Manikandan. «IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW.» *International Journal of Computing and Business Research (IJCBR)*, 4 July 2014.

Haibo, He, Bai Yang, Garcia Edwardo A., και Li Shutao. «ADASYN: Adaptive synthetic sampling approach for imbalanced learning.» *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 26 September 2008.

Hien M. Nguyen, Eric W. Cooper, Katsuari Kamei. «Borderline over-sampling for imbalanced data classification.» 2011.

Himani Bhavsar, Mahesh H. Panchal. «A Review on Support Vector Machine for Data.» *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, December 2012.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. «Borderline-SMOTE: A New Over-Sampling Method in.» n.d.

Richardson, Mark. «Principal Component Analysis.» May 2009.

Saina, Hartayuni, και Santi Wulan Purnami. «Combine Sampling Support Vector Machine for Imbalanced.» *Procedia Computer Science 72*, 2015: 59 – 66 .

SRIVASTAVA, DURGESH K., και LEKHA BHAMBHU. «DATA CLASSIFICATION USING SUPPORT VECTOR.» *DURGESH K. SRIVASTAVA, 2LEKHA BHAMBHU*, n.d.

ΓΕΩΡΓΟΥΛΗ, ΚΑΤΕΡΙΝΑ. *ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ Μια Εισαγωγική Προσέγγιση*. ΣΕΑΒ, 2015.

Ζγέρα, Χριστίνα. *Μοντελοποίηση και Προβλέψεις: Ανάλυση των Τιμών του Αργού Πετρελαίου (WTI) με τη Χρήση Χρονοσειρών.* ΘΕΣΣΑΛΟΝΙΚΗ: ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ, ΦΕΒΡΟΥΑΡΙΟΣ 2016.

Μαργιά, Γεωργίας. «Ανάλυση και Πρόβλεψη Χρονοσειρών.» *Διπλωματική εργασία.* ΘΕΣΣΑΛΟΝΙΚΗ: ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ, 2009.

Μητσοτάκης, Στυλιανός. *Market Basket Analysis.* ΑΝΤΙΡΡΙΟ: Τ.Ε.Ι. Δυτικής Ελλάδας, 2018.

Νικολάου, Ευθύμιος Ι. «ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΓΡΑΜΜΙΚΩΝ, ΜΗ-ΓΡΑΜΜΙΚΩΝ ΚΑΙ ΝΕΥΡΟ-ΑΣΑΦΩΝ ΜΕΘΟΔΩΝ, ΓΙΑ ΤΗ ΒΡΑΧΥΠΡΟΘΕΣΜΗ ΠΡΟΒΛΕΨΗ ΠΑΡΑΓΩΓΗΣ ΕΝΕΡΓΕΙΑΣ ΑΠΟ ΑΙΟΛΙΚΑ ΠΑΡΚΑ.» *ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ.* Χανιά, Νοέμβριος 2007.

Ηλεκτρονική Βιβλιογραφία

1. http://repfiles.kallipos.gr/html_books/93/04a-main.html
2. https://repository.kallipos.gr/bitstream/11419/2970/1/02_chapter_05.pdf
3. https://slideplayer.gr/slide/2956833/
4. https://repository.kallipos.gr/bitstream/11419/1237/2/Kef._10.pdf
5. http://www.cs.uoi.gr/~pitoura/courses/dm/introspring11.pdf
6. https://repository.kallipos.gr/bitstream/11419/2966/1/02_chapter_01.pdf
7. https://en.wikipedia.org/wiki/Mean_squared_error
8. https://en.wikipedia.org/wiki/Root-mean-square_deviation
9. https://en.wikipedia.org/wiki/Mean_absolute_error
10. https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

11. https://en.wikipedia.org/wiki/Coefficient_of_determination
12. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
13. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.under_sampling.NearMiss.html
14. https://github.com/scikit-learn-contrib/imbalanced-learn/wiki/Over-sampling
15. https://imbalanced-learn.readthedocs.io/en/stable/auto_examples/under-sampling/plot_illustration_nearmiss.html