ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΑΓΙΟΥ ΝΙΚΟΛΑΟΥ

Μοντελοποίηση Πρόβλεψη και Ανάλυση Παραγόντων του Χρόνου Αποφοίτησης

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εισηγητής: Δήμητρα- Σπυριδούλα Ευγενικού Α.Μ.009

Επιβλέπων: Δρ. Στυλιανός Παπαδάκης, Καθηγητής

2019

Technological Education Institute of Crete

School of Management and Economics

Department OF BUSINESS Administration
(AGIOS NIKOLAOS)

Modeling Forecasting and Analysis Of Graduation Time

DIPLOMA THESIS

Student: Dimitra – Spuridoula Eugenikou R.N.009

Supervisor: Dr. Stylianos Papadakis, Professor

Υπεύθυνη Δήλωση : Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του Τμήματος Δίοικηση Επιχειρήσεων Αγίου Νικολάου του Τ.Ε.Ι. Κρήτης.

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία ασχολήθηκε με την μελέτη και εφαρμογή των μοντέλων LinearRegression, Ridge(L2) ,Lasso(L1), Polynomial Regression και SVR τα οποία αποτελούν μεθοδολογίες των γραμμικών μοντέλων ,τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο (Μοναχόπουλος, 2016). Ο σκοπός της έρευνας είναι (Athanasiadis & Αθανασιάδης, 2015) η λειτουργία ενός μοντέλο να μπορεί να προβλέπει ποιές ιδιότητες ενός νεοεισερχόμενου φοιτητή (δεδομένου) επηρεάζουν τον χρόνο αποφοίτησης .Για να μπορέσει να επιτευχθεί (Athanasiadis & Αθανασιάδης, 2015) αυτό ήταν απαραίτητο να συλλεχθούν παλαιότερων φοιτητών ιδιότητες ούτως ώστε να εποπτευθεί η σωστή λειτουργία του κάθε μοντέλου, και να επιλεγεί το μοντέλο με την βέλτιστη απόδοση με το μικρότερο σφάλμα, όπου για τον σκοπό της συγκεκριμένης έρευνας επιλέχθηκε το Polynomial Regression .Συνοψίζοντας για την υλοποίηση των μοντέλων έγινε χρήση της γλώσσας προγραμματισμού Python και όλα τα αποτελέσματα από κάθε μοντέλο παρουσιάζονται και περιγράφονται μεσα στο κείμενο (Μοναχόπουλος, 2016).

Λέξεις Κλειδιά :Γραμμικά μοντέλα, πρόβλεψη, χρόνο αποφοίτησης, φοιτητής ,προγραμματισμός

ABSTRACT

The thesis essay focuses on studying and analyzing the theoretical models of LinearRegression, Ridge (L2), Lasso (L1), polynomial Regression and the SVR.These theoretical models are part of the Linear models (Μοναχόπουλος, 2016).The research analyses a model that can foresee, the parameters that influences the graduation time of a student (data) (Athanasiadis & Αθανασιάδης, 2015).In order to achieve that (Athanasiadis & Αθανασιάδης, 2015) have been selected data from students of previous years. In order to verify the correct function of each model. So can be chosen the optimal performance with minimum error, for this purpose has been chosen the Polynemial Regression .Summarizing, the Python language has been used for creating the models. The results of each model has been analyzed and presented in the essay (Μοναχόπουλος, 2016) .

KeyWords: Linear models, prediction, graduation time, student, programming

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ

Εφόσον ολοκληρώθηκε η πτυχιακή εργασία θα ήθελα να ευχαριστήσω τον επιβλέπων Καθηγητή μου Στέλιο Παπαδάκη καθώς μέσα από τα δικά του μαθήματα αγάπησα την επιστήμη των υπολογιστών . Επίσης να τον ευχαριστήσω για την βοήθεια και καθοδηγησή που μου πρόσφερε ώστε να ολοκληρωθεί η πτυχιακή εργασία .Επίσης θα ήθελα να ευχαριστήσω τον μπαμπά μου διότι σε όλη την διάρκεια με στήριξε και ήταν πάντα δίπλα μου. Τέλος την αφιερώνω στον επιβλέπων Καθηγητή και τον μπαμπά μου.

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΛΙΣΤΑ ΣΧΕΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

1.1 Δομή Πτυχιακής

Το υλικό της παρούσας διπλωματικής εργασίας κατανέμεται έξι κεφάλαια εκ των οποίων το πρώτο κεφάλαιο περιλαμβάνει εισαγωγικές έννοιες (Fotiou, D. & Fotiou, 2018) για την τεχνητή νοημοσύνη και τι μεθόδους μάθησης σ' ένα μοντέλο. Στο κεφάλαιο δύο αναλύονται τα μοντέλα που χρησιμποιήθηκαν για την υλοποίηση της έρευνας τα οποία είναι: Μηχανών Διανυσματικής Υποστήριξης (SVM), Ridge (L2), Lasso (L1) και Polynomial Regression. Επίσης στο τρίτο κεφάλαιο γίνεται ανάλυση των ιδιοτήτωνν που χρησιμοποιήθηκαν για την επεξεργασία του μοντέλου καθώς και των συνόλων δεδομένων που χρησιμοποιήθηκαν για την πρόβλεψή του μοντέλου (Τσαρμπόπουλος, 2016). Στο τέταρτο κεφάλαιο αναλύθηκαν οι ρυθμίσεις που χρειάστηκαν να εφαρμοστούν για να υλοποιηθεί το κάθε μοντέλο.Στο πέμπτο κεφάλαιο γίνεται αναφορά του προβλήματος παρατίθενται αναλυτικά πίνακες με τα αποτελέσματα του κάθε (Μαραγκάκης & Maragkakis, 2013) μοντέλου και τέλος στο έκτο κεφάλαιο όπου αναλύονται τα συμπεράσματα.

1.2 Τεχνητή νοημοσύνη

Τεχνητή νοημοσύνη είναι ένας τομέας της πληροφορικής και των μαθηματικών, ασχολείται με το πώς να δώσει στους υπολογιστές την δυνατότητα να ενεργήσουν έξυπνα(Newell, 1982). Η νοημοσύνη μπορεί να οριστεί ως η ικανότητα μάθησης, κατανόησης, επίλυση προβλημάτων και λήψη αποφάσεων, κ.τ.λ.π (Negnevitsky & Intelligence, 2005). Χωρίς αμφιβολία η τεχνητή νοημοσύνη έχει να κάνει με την προέλευση και τον τρόπο δημιουργίας από την ανθρώπινη συμπεριφορά (Fetzer, 1990). Η οποία χωρίζεται στη συμβολική τεχνητή νοημοσύνη, όπου προσπαθεί να υπάρξει μια έντονη ομοιότητα με την ανθρώπινη νοημοσύνη αλγοριθμικά με την χρήση συμβόλων και λογικών κανόνων υψηλού επιπέδου , όπως και στην υπο συμβολική τεχνητή νοημοσύνη, η οποία προσπαθεί να δημιουργεί την ανθρώπινη ευφυΐα χρησιμοποιώντας βασικά αριθμητικά μοντέλα που συνθέτουν επαγωγικά νοήμονες συμπεριφορές με τη συνεχόμενη αυτοοργάνωση απλούστερων δομικών συστατικών(Rizos & Ρίζος, 2011). Η ΤΝ περιλαμβάνει σήμερα μια τεράστια ποικιλία υποπεδίων, από το γενικό (μάθηση και αντίληψη) μέχρι συγκεκριμένο, αποδείξει μαθηματικών θεωρημάτων, οδήγηση ενός αυτοκινήτου σε έναν πολυσύχναστο δρόμο,διάγνωση ασθενειών, (Russell & Norvig, 1995) την ρομποτική, την αναγνώριση προτύπων, τα συστήματα ισχύος, τη βελτιστοποίηση, την επεξεργασία σήματος και τις κοινωνικές / ψυχολογικές επιστήμες, κ.τ.λ.π. (Kalogirou, 2001). Όπου με βάση τον επιθυμητό επιστημονικό στόχο (Μαραγκάκης & Maragkakis, 2013) αρκετοί επιστήμονες που ασχολούνται με τον τομέα της τεχνητή νοημοσύνης επιχειρούν να δημιουργήσουν λογισμικό ή πλήρες μηχανές όπου να επιλύουν με λογικά αποτελέσματα πραγματικά προβλήματα οποι-

ασδήποτε μορφής, αν και πολλοί θεωρούν ότι η απομίμηση της πραγματικής ευφυίας η ισχυρή ΤΝ , πρέπει να είναι ο τελικός στόχος (Rizos & Ρίζος, 2011).

1.2.1 Η ιστορική Αναδρομή της Τεχνητής Νοημοσύνης

Αρχικά να αναφερθεί ότι εδώ και αρκετά χρόνια γίνονται μελέτες για την λειτουργία του ανθρώπινου εγκεφάλου. Καθώς με την ανάπτυξη της τεχνολογίας που έχει υπάρξει ο άνθρωπος προσπαθεί να αντιγράψει τη λειτουργία του ανθρώπινου εγκεφάλου και τις νοητικές του διεργασίες .Στο αρχικό στάδιο για την προόδο των νευρωνικών δικτύων έγινε από τον νευροφυσιολόγο Warren McCulloch και τον μαθηματικό Walter Pitts . Ο McCulloch για την λειτουργία του ανθρώπινου εγκεφάλου και του νευρικού συστήματος πρόσφερε πάνω από 20 χρόνια στην έρευνα αυτή . Με την συνεργασία του Pitts δημοσίευσαν ένα πιθανό σενάριο για τη λειτουργία των νευρώνων και μετέπειτα ένα πρωταρχικό Νευρωνικό Δίκτυο (ΝΔ) , για την δημιουργία χρησιμοποιήθηκαν απλά ηλεκτρικά κυκλώματα (Ανδριανάκης & Andrianakis, 2008). Το 1950 ο μαθηματικός Άλαν Τούρινγκ, πατέρας της θεωρίας υπολογισμού και προπάτορας της τεχνητής νοημοσύνης, πρότεινε τη δοκιμή Τούρινγκ μία απλή δοκιμασία που θα μμπορούσε να εξακριβώσει αν μία μηχανή διαθέτει ευφυΐα. (Μαραγκάκης & Maragkakis, 2013). Παρόλα αυτά, το 1954 ο Martin Minsky ολοκληρώνει την διδακτορική του διατριβή με τίτλο «Θεωρία Νεύρο- Αναλογικής Ενίσχυσης Συστημάτων και οι Εφαρμογές τους στο Πρόβλημα του Εγκεφαλικού Μοντέλου» («Theory of Neural Analog Reinforcement System and its Application to the Brain-Model Problem»). Στη συνέχεια από τον ίδιο υπήρξε μια δημοσίευση με τίτλο «Βήματα προς την Τεχνητή Νοημοσύνη» («Steps Towards Artificial Intelligence»). Όπου γίνεται η πρώτη αναλυτική αναφορά στην τεχνητή νοημοσύνη και στα νευρωνικά δίκτυα όπως τα ξέρουμε έως σήμερα. Όμως «επίσημη» χρονιά έναρξης της έρευνας των νευρωνικών δικτύων, θεωρείται το 1959 όταν το Dartmouth Summer Research Project on Artificial Intelligence άρχισε επίσημα την ερεύνα στο πεδίο της τεχνητής νοημοσύνης. Επίσης σημαντικός παράγοντας ήταν η δημιουργία της γλώσσα προγραμματισμού LISP το 1958 από τον John McCarthy, ήταν η πρώτη ουσιαστικά γλώσσα προγραμματισμού η οποία βοήθησε στην δημιούργια εφαρμογών της Τ.Ν,στη συνέχεια υπήρξε η εμφάνιση των γενετικών αλγορίθμων το ίδιο έτος από τον Fridbreg και η παρουσίαση αναπτυγμένου νευρωνικού δικτύου perceptron το 1962 από τον Rosenblatt.(Βιαννιτάκη, Βασιλική, 2015). Στα τέλη του 1960 άρχισε η πτώση της Τ.Ν, όπου υπήρξε μια εποχή σχολιασμού καθώς και μείωση των κονδυλίων για ερευνητικά προγράμματα , εφόσον όλοι οι εξοπλισμοί του χώρου ήταν για την επίλυση σε αρκετά απλά προβλήματα . Στη διάρκεια της δεκαετίας του '70 δημιουργήθηκε ενδιαφέρον ξανά για τον τομέα διότι η ύπαρξει των εμπορικών εφαρμογών που εμφάνισαν τα εμπειρικά συστήματα , μηχανές Τ.Ν. με αποθηκευμένη γνώση για έναν εξειδικευμένο τομέα και δυνατότητα γρήγορης εξαγωγής αποτελεσμάτων , όπου λειτουργούν σαν ένα ειδικευμένο άνθρωπο στον αντίστοιχο κλάδο. Στη συνέχεια παρουσιάστηκε η γλώσσα λογικού προγραματισμού Prolog όπου πρόσφερε ώθηση στη συμβολική Τ.Ν, έπειτα στην αρχή της δεκαετίας του '80 ξεκίνησαν να δημιουργούνται πιο ισχυρά και με περισσότερες εφαρμογές νευρωνικά δίκτυα , όπως το πολυεπίπεδο perceptron και το δίκτυο Hopfield. Παράλληλα οι γενετικοί αλγόριθμοι και αάλλες παρεμφεροί μεθοδολογίες εξελίσονταν πλέον από κοινού, κάτω από την ομπρέλα του εξελικτικού υπολογισμού (Βιαννιτάκη, Βασιλική, 2015).

Χρόνος	Εξέλιξη
1950	Η δοκιμή Τουρινγκ όπου αναφέρει ο Άλαν Τούρινγκ είναι ότι προσπαθεί να ελέγξει την δυνατότητα μιας μηχανής να μπορέσει να συμβάλει ελεύθερα σε μια ανθρώπινη συζήτηση.
1951	Η αρχική δημιουργία προγραμμάτων για την τεχνητή νοημοσύνη που γράφτηκε στον υπολογιστή Ferranti Mark I, το ένα παίζει ντάμα και ήταν από τον Κρίστοφερ Στράκλι και το άλλο παίζει σκάκι όπου ο δημιουργός του ήταν ο Ντίτριχ Πρίνζ .
1956	Ο Τζον Μακάρθι αναφέρει την ονομασία «Τεχνητή Νοημοσύνη» που ήταν το κύριως αντικείμενο της ομιλίας του στο Ντάρτμουθ.
1958	Δημιουργός της συναρτησιακής γλώσσας προγραμματισμού Lisp ήταν ο Τζον Μακάρθι.
1965	Το εμπειρικό σύστημα Dendral όπου ήταν από τα πρώτα εμπειρικά προγράμματα άρχισε από τον Έντουαρτ Φάιγκενμπαουμ, στη προσπάθεια για την ολοκλήρωση δημιουργίας του συστήματος χρειάστηκαν δέκα χρόνια, με αποτέλεσμα την δομή των μοριακών οργανικών ενώσεων με την χρήση επιστημονικών.
1966	Άρχισε την λειτουργία του ένα από τα πρώτα εργαστήρια της μηχανικής μάθησης που κατασκευάστηκε στο Εδιμβούργο κάτω από την επίβλεψη τουΝτόναλντ Μίτσι και άλλους.
1970	Δημιουργήθηκ το Planner που χρησιμοποήθηκε στο SHRDLU, σε μία εξαιρετική αμοιβαία επίδραση ανάμεσα άνθρωπο και ηλεκτρονικό υπολογιστή.
1971	Άρχισε η κατασκευή του αλγορίθμου Boyer- Moore στο Εδιμβούργο.
1972	Ο Αλάν Κολμεροέρ κατασκευάσαι την γλώσσα προγραμματισμού Prolog.
1973	Στο Εδιμβούργο κατασκευάστηκε το Ρομπότ «Φρέντι», όπου είχε ένα προσαρμόσιμο σύστημα που εποπτευόταν από τον υπολογιστή.
1974	Η ερευνα του Τέντ Σόρτλιφ ήταν για το πρόγραμμα MYCIN, όπου επέδειξε ότι πλησιάζει για τα ιατρικά συμπεράσματα που στηρίζονται σε κάποιες αποφάσεις που μπορεί να λειτουργεί ακόμη στην περίπτωση ασάφειας. Το πρόγραμμα DENDRAL διαμόρφωσε την πορεία των εμπειρικών συατημάτων.

1991	Στο πόλεμο του περσικού κόλπου , έγινε η χρήση από τον Αμερικάνικο στρατό το πρόγραμμα DART, όπου εφαρμόστηκε για την ενίσχυση του προσωπικού και ανταπέδως τα 30 έτη μελέτης και έρευνας για την T.N .
1994	Ντίκμαννς και η Ντάιμλερ-Μπενζ παρουσίασαν σε πολυ σύχνα τους δρόμους του παρισίου, την ικανότητα ενός αυτοκινήτου να μπορεί να λειτουργεί ανεξάρτητο.
1997	Ο Deep Blue μια σκακιστικη μηχανή της εταιρείας IBM νίκησε τον διεθνώς ανγνωρισμένο πρωταθλητή του σκάκι Γκάρι Κασπάροφ.
1998	Η εταρεία Tiger Electronics έβγαλε στην αγορά το Φέρμπι, όπου ήταν από τα πρώτα μηχανήματα T.N. που ήταν τόσο κοντά στον άνθρωπο.
1999	Η Sony παρουσιάζει το AIBO, όπου είναι ένα από τα πρώτα κατοικίδια ρομποτ με T.N.
2004	Η υπηρεσία DARPA άρχισε ένα πρόγραμμα με όνομα DARPA Grand Challenge, όπου οι άνθρωποι που θα συμμετείχαν είχαν ως σκοπό να κατασκευάζουν αυτοκίνητο που να λειτουργεί ανεξάρτητα. Το έπαθλο ήταν ένα χρηματικό ποσό.

Πίνακας 1: Οι σπουδαίες στιγμές στην ιστορία της Τεχνητής Νοημοσύνης (Μαραγκάκης & Maragkakis, 2013).

1.3 Μηχανική Μάθηση

Ένας κλάδος της τεχντήτης νοημοσύνης είναι η μηχανική μάθηση (machine learning) όπου ασχολείται με αλγόριθμους και μεθοδολογίες που δίνουν την δυνατότητα στους υπολογιστές να «μαθαίνουν». Με βάση την μηχανική μάθηση μπορούν να δημιουργηθούν ευέλικτα προγράμματα στους υπολογιστές που να μπορούν με αυτοματοποιημένη ανάλυση δεδομένων και όχι με την παρέμβαση των ειδικών που τα προγραμμάτισαν . Η ανάλυση δεδομένων είναι ένα μεγάλο μέρος της μηχανικής μάθησης καθώς και της στατιστικής αυτό συμβαίνει διότι η μηχανική μάθηση έχει υιοθετήσει μεθοδολογίες της στατιστικής .

Η χρήση των αλγορίθμων μηχανικής μάθησης χρησιμποποιείται ανάλογα με το είδος του προβλήματος και έτσι αναλόγως εφαρμόζονται. Ορισμένοι αναφέρονται παρακάτω :

- Επιτηρούμενη μάθηση (supervised learning),έχοντας ένα αλγόριθμο που έχει δημιουργήσει μια συνάρτηση που επεξεργάζεται ένα σύνολο δεδομένων εισόδου σε γνώριμες εξόδους , καθώς επεξεργάζεται το σύνολο των δεδομένων στο μοντέλο που ο στόχος είναι να παρουσιάσει τη απόδοση θα έχει σε άγνωστα δεδομένα δηλαδή δεδομένα που δεν έχει ξανά επεξεργαστεί το μοντέ-

λο. Με την χρηση της επιτηρούμενης μάθησης μπορεί να διασταυρωθεί η σωστή απόδοση του μοντέλου σε καινούργια δεδομένα.

- Μη επιτηρούμενη μάθηση(unsupervised learning),ακολουθεί την εξής μεθοδολογία δημιουργείται ένα μοντέλο από τον αλγόριθμο χρησιμοποιώντας ένα σύνολο δεδομένων εισόδου που δεν λαμβάνει υπόψην τις τιμές εξόδους για την εκπαίδευση του μοντέλου.

- Ενισχυτική μάθηση (reinforcement learning), χρησιμοποιείται ένας αλγόριθμος που ερμηνεύει μια συγκεκριμένη λειτουργία για κάποιο δεδομένο.

Στην επιστήμη της στατιστικής υπάρχει ένας τομέας που καλείται θεωρία μάθησης και από που οι αλγόριθμοι μάθησης έχουν αντλήσει την λειτουργία τους (Athanasiadis & Αθανασιάδης, 2015).

1.3.1 Επεξεργασία των δεδομένων για εκπαίδευση ενός μοντέλου

Η εκπαίδευση ενός μοντέλου για να μπορέσει να έχει πιο αποδοτικά αποτελέσματα.Χρειάζεται σωστή προετοιμασία στα δεδομένα που θα επεξεργαστεί το μοντέλο ,όπου τα είναι τα ακόλουθα στάδια που αναφέρονται παρακάτω:

1. Προσδιορισμός των δεδομένων
2. Φιλτράρισμα των δεδομένων
3. Προεπεξεργασία των δεδομένων

Ο προσδιορισμός των δεδομένων τις περισσότερες φορές είναι το πιο σημαντικό στάδιο στη προπαρασκευή των δεδομένων.Είναι συνδεδεμένο με το τι θεωρείται σημαντικό στο πρόβλημα.Διότι ο πλήρης καθορισμός του προβλήματος βοηθάει στο να προσδιοριστούν πιά θα είναι τα δεδομένα εισόδου στο μοντέλο (Σουσούνης & Sousounis, 2011) .

Στο δεύτερο στάδιο της προετοιμασίας των δεδομένων γίνεται το φιλτράρισμα όπου ελέγχει αν τα δεδομένα θα παρουσιάσουν κάποια ασυνήθιστη συμπεριφορά από το

γενικό σύνολο αν παρατηρηθεί κάποια ιδιαίτερη συμπεριφορά βγαίνουν από το σύνολο εκπαίδευσης (Σουσούνης & Sousounis, 2011).

Στο τρίτο και τελευταίο στάδιο γίνεται η χρήση της κανονικοποίησης στα δεδομένα αυτή η μέθοδος είναι πολύτιμη όταν τα δεδομένα του μοντέλου διαφέρουν αρκετά ως πρός την τιμή τους. Για την κανονικοποίηση των δεδομένων θα πρέπει να χρησιμοποιείται μια τυπική περιοχή κανονικοποίησης .Όπου το προτιμότερο οι τιμές των δεδομένων να παίρνουν τιμές μεταξύ του 0 και 1 ή μεταξύ -1 και 1 (Σουσούνης & Sousounis, 2011).

Κατά την διαδικασία εκπαίδευσης ενός μοντέλου χρησιμοποιούνται τρία είδη δεδομένων:

- Εκπαίδευσης (Train data)
- Επαλήθευσης(Validation data)
- Εφαρμογής (Test data)

Τα δεδομένα εκπαίδευσης αναφέρονται τα δεδομένα εκείνα , τα οποία χρησιμοποιούνται για την εκπαίδευση του μοντέλου (Ανδριανάκης & Andrianakis, 2008).

Τα δεδομένα επαλήθευσης χρησιμοποιούνται για τον υπολογισμό και την παρακολούθηση του σφάλματος επαλήθευσης κατά την διάρκεια της εκπαίδευσης αυτό το σφάλμα μειώνεται όπως και το σφάλμα εκπαίδευσης (σφάλμα μεταξύ εξόδου και προσδοκώμενης εξόδου) (Σουσούνης & Sousounis, 2011).

Τα δεδομένα εφαρμογής ή αλλιώς τα δεδομένα ελέγχου ενός μοντέλου χρησιμοποιούνται ώστε να εξετάσουν την επίδοση του μοντέλου σε άγνωστα δεδομένα. Στο σημείο αυτό, να σημειωθεί ότι πάντοτε τα δεδομένα ελέγχου εμπεριέχονται στα δεδομένα εκπαίδευση (Ανδριανάκης & Andrianakis, 2008).

1.3.2Διασταυρούμενη Επικύρωση (Cross-Validation)

Η λειτουργία της διασταυρούμενης επικύρωση είναι ότι χρησιμοποιείται για την εκτίμηση ενός μοντέλου, δηλαδή έχει την ικανότητα να δείχνει την συμπεριφορά ενός μοντέλου για τα νέα δεδομένα (Παπαδάκης, 2016).

Διότι από ένα μοντέλο δεν θέλουμε να εξετάστεί μόνο η απόδοση που έχει στα δεδομένα εκπαίδευσης αλλά ουσιαστικά την απόδοση που θα έχει στα καινούργια δεδομένα σε δεδομένα δηλαδή που δεν έχει ξαναδεί ο αλγόριθμος (Παπαδάκης, 2016). Καθώς ένα μοντέλο μπορεί να έχει πολύ καλή απόδοση στα δεδομένα εκπαίδευσης ,αλλά να έχει πολύ άσχημη εκτίμηση στα καινούργια δεδομένα. Για να εξεταστεί η απόδοση του μοντέλου,χωρίζεται το σύνολο των δεδομένων του μοντέλου σε k υποσύνολα όπου από τα δεδομένα k-1 υποσύνολα είναι για την λειτουργία του μοντέλου τα οποία ονομάζονται δεδομένα εκπαίδευσης(train data) ,όμως το μοντέλο εκτιμάται στο υποσύνολο που είναι εκτός από την εκπαίδευση του μοντέλου το οποίο είναι το υποσύνολο με τα δεδομένα ελέγχου(text data) (Παπαδάκης, 2016)2016). Όμως η εκτίμηση του μοντέλου μπορεί να ήταν τελείως διαφορετική αν είχε χωριστεί αλλιώς το σύνολο των δεδομένων έτσι για να υπάρξει η καλύτερη εκτίμηση της γενίκευσης του μοντέλου (Διαμαντάρας, Κ, 2007) επαναλαμβάνεται η διαδιακασία k φορές έως ούτε να αξιολογηθούν όλα τα υποσύνολα (Παπαδάκης, 2016) .

Καθώς έχει ολοκληρωθεί η παραπάνω διαδικασία δηλαδή έχει γίνει ο διαχωρισμός του συνόλου των δεδομένων έχει αξιολογηθεί το κάθε k υποσύνολο έως k φορές τότε ο δείκτης επίδοσης μπορεί να υπολογιστεί απο τον μέσο όρο των k υποσυνόλων (Παπαδάκης, 2016) . Οπότε η χρήση της μέθοδου διασταυρούμενης επικύρωσης μπορεί να υπάρξει μόνο για να γίνει εκτίμηση της επίδοσης ενός μοντέλου σε νέα δεδομένα.

Εικόνα 6:K- fold cross-validation (Παπαδάκης, 2016).

ΚΕΦΑΛΑΙΟ 2

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine - SVM)

2.1 Εισαγωγή

Στο Support Vector Machine η μέθοδος ταξινόμησης αναπτύχθηκε από τον Vapnik και την ομάδα του (Youn, 2002) είναι ένας ταξινομητής επιβλεπόμενης μάθησης, είναι ταχύτατος εκπαιδεύεται εύκολα, χρησιμοποιείται σε γραμμικά και μη γραμμικά προβλήματα.
Συνήθως η βασική χρήση του είναι η κατηγοριοποίηση αλλά όμως χρησιμοποιείται και σε προβλήματα παλινδρόμησης . Μπορεί όμως να λειτουργήσει σε διάφορες εφαρμογές, κάποια από αυτές είναι η κατηγοριοποίηση κειμένων και αναγνώριση προσώπων.

2.2 Γραμμικά διαχωρίσιμα προβλήματα

Για να γίνει πιο σαφείς η λειτουργία του ταξινομητή SVM . Αρχικά θα εξετάσουμε την λειτουργία του σε γραμμικά διαχωρίσιμα προβλήματα, αν υποθέσουμε ότι έχουμε δύο κλάσεις C_0, C_1 που είναι γραμμικά διαχωρίσιμες και αποτελούνται από ένα σύνολο δεδομένων. Όπου αυτό δηλώνει ότι υφίσταται ένα διάνυσμα **w** και ένα κατώφλι w_0, είναι ως εξής:

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 \begin{cases} < 0, \alpha\nu\ \boldsymbol{x}\ \in C_0 \\ > 0\ , \alpha\nu\ \boldsymbol{x}\ \in C_1 \end{cases} \qquad (1)$$

Όπως παρατηρείται σε αυτό το πρόβλήμα ταξινόμησης δεν υπάρχει μία μόνο λύση αλλά για την ακρίβεια άπειρα – ζεύγη $(\boldsymbol{w}, w_0)$ μπορούν να διαχωρίσουν τις δύο κλάσεις (Διαμαντάρας, Κ, 2007)
Επομένως για να μπορέσει να γίνει σωστή ταξινόμηση μεταξύ των δύο κλάσεων,θα πρέπει να τεθεί ένα κριτήριο αξιολόγησης ,ονόματι περιθώριο ταξινόμησης γ, το οποίο υπολογίζεται από τα όρια των δύο κλάσεων .

$$\gamma = \gamma_0 + \gamma_1 \qquad (2)$$

ανάμεσα των δύο παρακάτω « περιθωρίων» όπου το γ_0 είναι το περιθώριο που ανήκει στην κλάση C_0 και το περιθώριο γ_1 ανήκει στην κλάση C_1

$$\gamma_0 = \min_{x\in C_0} \frac{|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||} = \min_{x\in C_0} \frac{-|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||},$$

$$\gamma_1 = \min_{x\in C_1} \frac{|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||} = \min_{x\in C_1} \frac{|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||}$$

Για να μπορέσει να περιγραφεί το κριτήριο ή αλλιώς κανόνας, και ώστε να γίνει σαφές η χρήση της λέξης «περιθωρίο» αν παρατηρήσουμε ότι η τιμή $\frac{|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||}$ είναι αντίστοιχη η ευθεία με την απόσταση του διανύσματος **x** από την γραμμική διαχωριστική επιφάνεια.

Όπου χρησιμοποιείται η εξίσωση του υπερεπιπέδου:

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 = 0 \qquad (3)$$

Επομένως οι αριθμοί γ_0,γ_1 είναι θετικοί και συμπίπτουν στην απόσταση της πιο κοντινής ιδιότητας (προτύπο) της διαχωριστικής επιφάνειας της κάθε κλάσης .Όσο τείνει να πλησιάζει η τιμή γ_0ή του γ_1 στο 0 είναι πιο οριακός ο διαχωρισμός των προτύπων για την κάθε κλάση , διότι τα πρότυπα που είναι οριακού μπορεί να εκτοπιστούν στην άλλη κλάση με κάποια μικρή μεταβολή. Με τα πρότυπα **x** της κλάσης C_0 όπως και τα πρότυπα x' της κλάσης C_1 για τα οποία πετυχένεται η μικρότερη απόσταση, δηλαδή, $\frac{|\boldsymbol{w}^T\boldsymbol{x}+w_0|}{||w||} = \gamma_0$και$\frac{|\boldsymbol{w}^T\boldsymbol{x}'+w_0|}{||w||} = \gamma_1$, ονομάζονται διανύσματα υποστήριξης(support vectors).Οπότε το περιθώριο ταξινόμησης γ, δίνει την εγγύηση ανάμεσα σε δύο κλάσεις συνολικό περιθώριο ταξινόμησης γ ,είναι ένα μέτρο εγγύτητας ανάμεσα σε δύο κλάσεις. Συνεπώς τα όρια ή αλλιώς τα διανύσματα υποστήριξης είναι τα σημεία της κάθε κλάσης που είναι πιο κοντά στην διαχωριστική ευθεία ή αλλιώς διαχωριστικό υπερεπίπεδο . Καθώς κάνοντας χρήση της εξίσωση του υπερεπιπέδου όπου το w_0 είναι υπεύθυνο για την ταυτόχρονη πορεία του υπερεπιπέδου .Το οποίο καλείται «κανονικό διαχωριστικό υπερεπίπεδο» όπου

- το κατώφλι w_0 το βάζει εντελώς στο κέντρο μεταξύ των δύο κλάσεων ,όπου γ_0=γ_1 και
- κλιμάκωση των **w** και w_0 είναι τέτοια ώστε

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 \begin{cases} \leq -1, \alpha\nu\ \boldsymbol{x}\ \in C_0 \\ \geq 1\ , \alpha\nu\ \boldsymbol{x}\ \in C_1 \end{cases} \qquad (4)$$

Δηλαδή ένα κανονικό διαχωριστικό υπερεπίπεδο σύμφωνα με τις ανισότητες της (4) έχουμε $\min_{x\in C_0}|\boldsymbol{w}^T\boldsymbol{x} + w_0| = \min_{x\in C_1}|\boldsymbol{w}^T\boldsymbol{x} + w_0| = 1, \kappa\alpha\iota\ \gamma_0 = \gamma_1 = \frac{1}{||\boldsymbol{w}||}, \acute{o}\pi o\upsilon$

$$\gamma = \frac{2}{||\boldsymbol{w}||} \qquad (5)$$

Οπότε, με τις ανισότητες που απεικονίζονται παραπάνω και μετά ζεύγη προτύπων, στόχων που υποθετικά διαθέτουμε $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_P, y_P)$ όπου $y_i = -1$ αν $\boldsymbol{x}_i \in C_0$ και $y_i = 1$ αν $\boldsymbol{x}_i \in C_1$.Ώσπου με την απλοποίηση των παραπάνω ανισοτήτων απεικονίζεται παρακάτω με την εξής μορφή:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1, \quad i = 1,2, \dots, P \qquad (6)$$

Τώρα θα χρειαστεί να οριστεί η ιδανικότερη λύση για ένα κανονικό διαχωριστικό υπερεπίπεδο (6) , όπου η μεγιστοποιήσει της γ (5) ισοδυναμεί με την μείωση της νόρμας $||\boldsymbol{w}||$ή $\left|\left|\boldsymbol{w}\right|\right|^2$.

Για να υπολογιστεί μετατρέπεται σε πρόβλημα βέλτιστου διαχωριστικού υπερεπιπέδου και με την χρήση των περιορισμών (6), και έτσι έχει δημιουργηθεί μια τετραγωνική συνάρτηση .

$$J(\boldsymbol{w}, w_0) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 \qquad (7)$$

Για να έχουμε μείωση στην τετραγωνική συνάρτηση (7) με την χρήση των περιορισμών (6) όπου χρησιμοποιώντας τους πολλαπλασιαστές Lagrange , έτσι έχουμε την συνάρτηση κόστους Lagrange , οπότε τοποθετείται στον κάθε περιορισμό (6) από ένας συντελεστής $\lambda_i, i = 1,2, \dots, P$

$$L(\boldsymbol{w}, w_0, \lambda_1, \dots, \lambda_P) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 - \sum_{i=1}^{P} \lambda_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1] \qquad (8)$$

Επιπλέον οι πολλαπλασιαστές Lagrange θα χρειαστεί να ικανοποιούν τις ανισότητες επειδή οι συγκεκριμένοι περιορισμοί εμφανίζονται με σύστημα ανισότητας

$$\lambda_i \geq 0, i = 1, \dots, N \qquad (9)$$

Η συνάρτηση L θα χρειαστεί να μειωθεί ως προς το **w** και το w_0, για να μπορέσει να αυξηθεί ως προς τα λ_i.Οπότε στο ιδανικό σημείο θα χρειαστεί να ισχύουν οι παρακάτω συνθήκες:

$$\frac{\partial L}{\partial w_0} = 0,\ \frac{\partial L}{\partial \boldsymbol{w}} = 0 \tag{10}$$

και

$$\lambda_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1] = 0,\ \ i = 1, \ldots . P \tag{11}$$

Ύστερα από την χρήση διάφορων πράξεων οι συνθήκες (10) μεταφράζονται ως εξής:

$$\sum_{i=1}^{P} \lambda_i y_i = 0 \tag{12}$$

$$\boldsymbol{w} = \sum_{i=1}^{P} \lambda_i y_i \boldsymbol{x}_i \tag{13}$$

Όπου η (13) μας δίνει την ιδανική τιμή του **w** όπου η ιδανική γραμμική διαχωριστική επιφάνεια είναι εξής:

$$g*(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_{\boldsymbol{0}} = \sum_{i=1}^{P} \lambda_i y_i \boldsymbol{x}_i^T\boldsymbol{x} + w_0 \tag{14}$$

Έπειτα μπορεί να υπολογιστεί η βέλτιστη πόλωση $w_{\boldsymbol{0}}$ με την χρήση (11) . Θεωρόντας ότι το $\boldsymbol{x}_i$ είναι ένα διάνυσμα υποστήριξης τότε:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_{i+} w_0) = 1$$

$$w_0 = \frac{1}{y_i} - \boldsymbol{w}^T\boldsymbol{x}_i \tag{15}$$

Για αριθμητικούς λόγους συνηθίζετε ο υπολογισμός για την αξιολόγηση της πόλωσης

w_0 , γίνεται από την μέση τιμή (15) σε όλα τα διανύσματα υποστήριξης, συγκεκριμένα

$$w_0 = \frac{1}{|I_{sv}|}\sum_{i\in I_{sv}}\left(\frac{1}{y_i} - \boldsymbol{w}^T\boldsymbol{x}_i\right) \quad (16)$$

ορίστηκαν συνολικά

$$I_{sv} = \{i: \boldsymbol{x}_i = διάνυσμα\ υποστήριξης$$

οι δείκτες των διανυσμάτων υποστήριξης .Από την (11) επίσης προκύπτει ένα σημαντίκο συμπέρασμα :

- οι μοναδικοί πολλαπλασιαστές λ_i που έχουν την δυνατότητα να μην είναι μηδενικοί είναι αυτοί για τους οποίους $y_i(\boldsymbol{w}^T\boldsymbol{x}_{i+}w_0) = 1$, συγκεκριμένα όπου συμπίπτονται με κάποιο διάνυσμα υποστήριξης $\boldsymbol{x}_i$. Το σύνολο που απομένει $\boldsymbol{x}_i$, χρειάζεται να έχει λ_i=0.

Οπότε σύμφωνα με την εξίσωση (13), η λύση του διανύσματος **w**, δίνει στα διανύσματα υποστήριξης μια γραμμικά θετική συγχώνευση. Πλεόν τα διανύσματα αποκτούν μια ιδιαίτερη σημασία .Μπορούν να γραφτούν:

$$\boldsymbol{w} = \sum_{i\in I_{sv}}^{P}\lambda_i y_i \boldsymbol{x}_i \quad (17)$$

$$g*(\boldsymbol{x}) = \sum_{i=1}^{P}\lambda_i y_i \boldsymbol{x}_i^T\boldsymbol{x}+w_0 \quad (18)$$

Επίσης για τα διανύσματα υποστήριξης ένα πρόβλημα είναι ότι δεν είναι γνωστά εξ αρχής, συνεπώς το πλήθος I_{sv} δεν είναι γνωστό παρά μόνο αφού λυθεί το πρόβλημα. Ωστόσο αν χρησιμοποιηθεί η εξίσωση (13) όπου αθροίζει όλα τα διανύσματα $\boldsymbol{x}_i$ αν και η μορφή αυτή έχει αρκετό πλεονασμό επειδή το πλήθος όλων των διανυσμάτων μπορεί να είναι μέγαλο αντιθέτως μπορεί το πλήθος των διανυσμάτων υποστήριξης να είναι αρκετά μικρό.
Με όλα τα παραπάνω έχουμε τα εξης:

$$\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \quad (19)$$

και

$$\sum_{i=1}^{P} \lambda_i \left[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 \right] = \sum_{i=1}^{P} \lambda_i y_i \sum_{j=1}^{P} \lambda_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i + w_0 \sum_{i=1}^{P} \lambda_i y_i - \sum_{i=1}^{P} \lambda_i$$

$$= \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_{i=1}^{P} \lambda_i$$

αντικαθιστώντας την εξίσωση του (8) παίρνει την παρακάτω συνάρτηση Lagrange

$$L(\lambda_1, \dots, \lambda_P) = \sum_{i=1}^{P} \lambda_i - \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \qquad (20)$$

Με την παραπάνω συνάρτηση L δεν εμφανίζονται ούτε το διάνυσμα **w** ούτε το κατώφλι w_0 πλέον οι άγνωστες παράμετροι που εμπεριέχονται είναι οι πολλαπλασιαστές λ_i .Οπότε αναζητείται η μεγιστοποίηση της (20) ως προς τα λ_i ή ισοδύναμα , την μείωση της συνάρτησης

$$L^d(\lambda_1, .., \lambda_P) = -L(\lambda_i, \dots, \lambda_P)$$

$$L^d(\lambda_1, .., \lambda_P) = \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_{i=1}^{P} \lambda_i \qquad (21)$$

οπότε το πρόβλημα γίνεται διαφορετικό αλλά ισοδύναμο πρόβλημα που ονομάζεται δυικό πρόβλημα . Επίσης θα πρέπει να αναφερθεί μια σημαντική λεπτομέρεια με ιδιαίτερες πρακτικές προεκτάσεις :η συνάρτηση κόστους L^d της μορφής (21) περιέχει P άγνωστες παραμέτρους τα πρότυπα και από τις δύο κλάσεις , καθώς το πρόβλημα περιέχει P^2 γινόμενα q_{ij} της μορφής $y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$, $i = 1, \dots, P, j = 1, \dots, P$, όπου πρέπει να υπολογιστούν και να αποθηκευθούν στη μνήμη.Όταν υπάρχει μεγάλο πλήθος προτύπων δημιουργούνται δύο σημαντικά προβλήματα: ο πλήθος P των αγνώστων είναι πολύ μεγάλο και έτσι έχει ως αποτέλεσμα ο αλγόριθμος του τετραγωνικού προγραμματισμού να είναι υπερβολικά αργός και το πλήθος P^2 των γινομένων q_{ij} που πρέπει να αποθηκευτούν στη μνήμη είναι τεράστιο. Οπότε αυτό το πρόβλημα δημιούργησε την ανάγκη αποδοτικότερων τρόπων επίλυσης του προβλήματος.Επίσης θα χρειαστεί να επισημανθεί ότι ένα πρόβλημα εφόσον δεν είναι γραμμικό η συγκεκριμένη μέθοδο δεν θα φέρνει ικανοποιητικά αποτελέσματα, γι αυτό έχουν υπάρξει μεθοδολογίες για την επίλυση διαφόρων προβλημάτων με ταξινομητή διανυσματικής υποστήριξης, όπως θα δούμε στην συνέχεια (Διαμαντάρας, K, 2007).

2.3 Μη γραμμικά διαχωρίσιμα προβλήματα

Υπάρχουν όμως κάποια προβλήματα που είναι μη γραμμικά διαχωρίσιμα στην περίπτωση αυτή δεν μπορεί να γίνει η χρήση της παραπάνω μεθόδους όπως είδη αναφέρθηκε, επειδή δεν θα απέδιδε για την σωστή κατηγοριοποίηση των περισσότερων δεδομένων. Κάνοντας κάποιες ρυθμίσεις στην μέθοδο της δίνετε η δυνατότητα να λειτουργεί αποδοτικά και για τα προβλήματα που είναι μη διαχωρίσιμα, ο παράμετροι που θα ορίσει ο χρήστης θα επηρεάζουν τον βαθμό επίδοσης. Δηλαδή μετατρέπουμε το αρχικό πρόβλημα με την εισαγωγή των μεταβλητών χαλαρότητας ξ_i για τον κάθε περιορισμό (6).

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i, \quad i = 1,..,P \tag{22}$$

$$\xi_i \geq 0$$

Επομένως στην περίπτωση που η μεταβλητή ξ_i είναι μεγαλύτερη από 1 τότε το δεδομένο $\boldsymbol{x}_i$ κατηγοριοποιείται σε λάθος κλάση. Οπότε τα δεδομένα έχουν ως αποτέλεσμα να ταξινομούνται λάθος:

$$\sum_{i=1}^{P} \xi_i > \textit{πλήθος προτύπων που ταξινομούνται λάθος}$$

Οπότε με την χρήση της συνάρτησης κόστους (8) δημιουργείται:

$$J_{ns}(\boldsymbol{w}, w_0) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\sum_{i=1}^{P}\xi_i \tag{23}$$

Η χρήση της παράμετρος C είναι για να δείχνει το πόσο μας ενδιαφέρει αν υπάρχουν λάθος κατηγοριοποίησης , αυτό το δίνει η τιμή που έχει καθορίσει ο ίδιος ο χρήστη. Δηλαδή όσο μεγαλύτερη τιμή έχει το C δίνεται ιδιαίτερη προσοχή στην σωστή κατηγοριοποίηση και το αντίστροφο.

Επιπλέον θα χρειαστεί να προσθέσουμε ότι οι μέθοδοι που έχουν αναφερθεί για τα γραμμικά και μη γραμμικά προβλήματα σε δύο κλάσεις είναι αποδοτικά , όμως υπάρχουν περιπτώσεις όπου διαχωρίζονται οι κλάσεις σε μη γραμμικές επιφάνειες (Διαμαντάρας, K, 2007), όπου εδώ θα χρειαστεί να γίνει χρήση άλλων μεθόδων.

2.4 Χρήση Συναρτήσεων Πυρήνα (Kernel functions)

Οι μέθοδοι που έχουν αναλυθεί παραπάνω για τον ταξινομητή διανυσματικής υποστήριξης (SVM) όπου μπορούν να δημιουργήσουν γραμμικές και μη γραμμικές διαχωριστικές επιφάνειες με στόχο των διαχωρισμό των κλάσεων. Όμως στην περίπτωση ενός προβλήματος όπου διαχωρίζονται οι κλάσεις σε μη γραμμικές επιφάνειες η παραπάνω μέθοδοι δεν μπορούν να είναι αποτελεσματική (Διαμαντάρας, Κ, 2007). Οπότε εφαρμόζοντας τους πυρήνες όπου αντιστοιχούν σε γραμμικά προβλήματα (Schölkopf, Herbrich, & Smola, 2001). Εφόσον κάνοντας την απαραίτητη μετατροπή όπου θα χρειαστεί ένας μη- γραμμικός μετασχηματισμός Φ(.) , που θα μετατρέπει τα μη- γραμμικά δεδομένα σε γραμμικά x→ Φ(x) ,y→ Φ(y). Επομένως στην περίπτωση αυτή ισχύουν ακριβώς τα ίδια με το παραπάνω εδάφιο (Διαμαντάρας, Κ, 2007).

Οπότε σε αυτή την περίπτωση μπορεί να εφαρμοστεί η συνάρτηση πυρήνα :

$$k(\boldsymbol{x},\boldsymbol{y}) = \Phi(\boldsymbol{x})^T\Phi(\boldsymbol{y}) \qquad (24)$$

Η συνάρτηση k ονομάζεται πυρήνας (Schölkopf et al., 2001). Εφαρμόζοντας αυτή την συνάρτηση απλοποιούνται οι πράξεις και είναι ιδιαίτερα σημαντικό στην περίπτωση που έχουμε διανύσματα πολλών μέχρι και άπειρων διαστάσεων. Οι συχνότερες συναρτήσεις πυρήνα είναι η εξής:

$e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ Γκασιουανή RBF
$[\boldsymbol{x}^T\boldsymbol{y}+\theta]^p$ Πολυωνυμική
$\tanh(a\boldsymbol{x}^T\boldsymbol{y}+\theta)$ Σιγμοειδής
$\frac{1}{\sqrt{\|x-y\|^2+c^2}}$ Αντίστροφη πολυτετραγωνική

Πίνακας 2:Συναρτήσεις πυρήνα (Διαμαντάρας, Κ, 2007)

Στο σημείο αυτό θα χρειαστεί να αναφερθεί, αν η συνάρτηση πυρήνα που θα εφαρμοστεί είναι η σωστή για τον διαχωρισμό των συγκεκριμένων κλάσεων μη γραμμικές επιφάνειες, εξετάζεται εφόσον έχει είδη χρησιμοποιηθεί η συνάρτηση πυρήνα στο πρόβλημα και εκτιμάται η αποδοτικότητα από την λύση που δίνει (Διαμαντάρας, Κ, 2007).

2.5 Παλινδρόμηση Με Μηχανές Διανυσματικής Υποστήριξης (Support Vector Regression -SVR)

Οι μηχανές διανυσματικής υποστήριξης εφαρμόζονται όπως αναφέρθηκε και παραπάνω σε προβλήματα κατηγοριοποίησης αλλά όμως και σε προβλήματα παλινδρόμησης . Τα προβλήματα παλινδρόμησης διαφέρουν από τα προβλήματα κατηγοριοποίησης στο ότι οι παρατηρήσεις ενός σύνολο εκπαίδευσης δεν έχουν μια διακριτή ετικέτα (±1), αντιθέτως συνδέονται με ένα οποιοδήποτε αριθμό του συνόλου (Γιαννούλη & Giannouli, 2014).

2.5.1 Απλή και πολλαπλή παλινδρόμηση

Η παλινδρόμηση (regression) είναι μια στατιστική τεχνική (Μιχαλοπούλου, Ε. & Michalopoulou, 2016) όπου αν υποθέσουμε την τοποθέτηση ενός υπερεπιπέδου μέσω ενός συνόλου σημείων εκπαίδευσης με ελάχιστο σφάλμα. Αυτό το σφάλμα παλινδρόμησης ονομάζεται κατάλοιπο ή υπόλοιπο, και ορίζεται ως η απόκλιση ανάμεσα στην αναμενόμενης και στην πραγματική τιμή των δεδομένων εκπαίδευσης του μοντέλου. Στόχος της γραμμικής παλινδρόμησης είναι να ελαχιστοποιήσει τα κατάλοιπα.Για να γίνει αυτό πιο κατανοητό, θα θεωρήσουμε ότι ένα σύνολο εκπαίδευσης παλινδρόμησης της μορφής:

$$D = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots, (\bar{x}_l, y_l)\} \subset R^n \text{x R} \tag{25}$$

Ας υποθέσουμε ότι η $\hat{f}(x)$ είναι ένα μοντέλο παλινδρόμησης στο D, τότε η ποσότητα

$$\rho_i = y_i - \hat{f}(\bar{x}_i) \tag{26}$$

για $(x_i, y_i) \in$ D ονομάζεται υπόλοιπο και μετρά τη διαφορά μεταξύ προβλεπόμενης και της πραγματικής τιμής. Όμως για να υπάρξει ένα τέλειο μοντέλο θα πρέπει να είναι μηδέν, δηλαδή, οι προβλεπόμενες τιμές θα πρέπει να ταιριάζουν απόλυτα με τις τιμές που παρατηρούνται στα δεδομένα εκπαίδευσης. Ωστόσο, είναι υπερβολικά αισιόδοξο να θεωρούμε ότι μπορούν να κατασκευάστουν τέτοια «τέλεια» μοντέλα για τα δεδομένα εκπαίδευσης σε μια ρεαλιστική πραγματικότητα. Ωστόσο, θα πρέπει να κατασκευαστεί μοντέλο, όπου τα κατάλοιπα να ελαχιστοποιούνται. Στην γραμμική παλινδρόμηση αυτό επιτυγχάνεται υπολογίζοντας το ελάχιστο άθροισμα των τετραγώνων των σφαλμάτων (γνωστή ως μέθοδος ελαχίστων τετραγώνων):

$$min \sum_{i=1}^{l} \rho_i^2 = min_{\hat{f}} \sum_{i=1}^{l} \left(y_i - \hat{f}(\bar{x}_i)\right)^2 \tag{27}$$

Όπου με $(x_i, y_i) \in$ D Παρατηρείται ότι η απόκλιση εξαρτάται από τον τύπο της συνάρτησης που επιλέγετε $\hat{f}$ για να υπολογιστούν οι αποκλίσεις. Όπου δίνετε ένα πρόβλημα

βελτιστοποίησης που επιτρέπει να υπολογιστεί το βέλτιστο γραμμικό μοντέλο παλινδρόμησης $\hat{f}^*$:

$$\hat{f}^* = argmin_{\hat{f}} \sum_{i=1}^{l} \left(y_i - \hat{f}(\bar{x}_i)\right)^2 \quad (28)$$

Εφόσον κατασκευάζονται γραμμικά μοντέλα μπορεί να ξανά γραφεί η παραπάνω εξίσωση ώς εξής:

$$(\bar{w}^*, b^*) = argmin_{\bar{w},b} \sum_{i=1}^{l} (y_i - \bar{w} \cdot \bar{x}_i + b)^2 \quad (29)$$

Οπότε το βέλτιστο γραμμικό μοντέλο είναι:

$$\hat{f}^*(\bar{x}) = \bar{w}^* \bar{x} - b^* \quad (30)$$

Έπειτα στην απλή γραμμική παλινδρόμηση όπου μεσο της παράπανω εξίσωσης είναι να μας δίνει το σφάλμα που προκύπτει από την αναμενόμενη και την πραγματική τιμή. Ένα βέλτιστο γραμμξμξμικό μοντέλο παλινδρόμησης κατασκευάζεται έτσι ώστε να μειώνονται τα σφάλματα. Επίσης στην πολλαπλή γραμμική παλινδρόμηση για να επιλυθούν πρέπει να τηρούν ορισμένες προυποθέσεις συγγραμικότητας (Γιαννούλη & Giannouli, 2014).

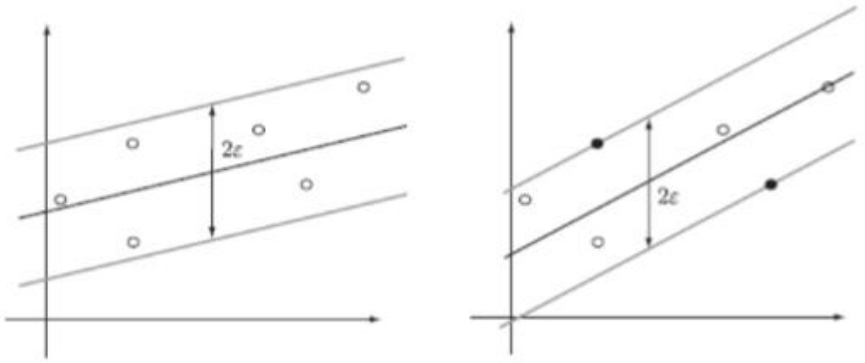
2.5.2 Παλινδρόμηση με μηχανές μέγιστου περιθωρίου

Ένα ισχυρό κίνητρο για την ανάπτυξη των διανυσμάτων υποστήριξης στα μοντέλα παλινδρόμησης είναι η απλή επέκταση από τη γραμμική στη μη γραμμική παλινδρόμηση χρησιμοποιώντας το τέχνασμα του πυρήνα. Για την ανάπτυξη των μηχανών διανυσματικής υποστήριξης στο πλαίσιο της παλινδρόμησης, αρχίζει με την πρωταρχική προσαρμογή των μηχανών μέγιστου περιθωρίου (Γιαννούλη & Giannouli, 2014). Στη περιίπτωση αυτή η βασική ιδέα είναι ότι ο ταξινομητής μηχανών μεγίστου περιθωρίου έχοντας το υπερεπίπεδο , ότι στόχος είναι να μεγιστοποιηθούν οι αποστάσεις των δεδομέμων από αυτό του υπερεπιπέδου (Γιαννούλη & Giannouli, 2014).
Ας υποθέσουμε ότι το σύνολο εκπαίδευσης της παλινδρόμησης είναι λοιπόν της μορφής:

$$D = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots, (\bar{x}_l, y_l)\} \subseteq R^n x R \quad (31)$$

Στην SVM παλινδρόμηση (regression)τα δεδομένα προσαρμόζονται σε ένα υπερσωλήνα πλάτους 2_ε , όπου $\varepsilon > 0$.Υπάρχουν πολλοί τρόποι προσανατολισμού του υ-

περσωλήνα αυτού, ώστε τα δεδομένα να βρίσκονται εντός του (Μιχαλοπούλου, Ε. & Michalopoulou, 2016). Όπου απεικονίζονται (Γιαννούλη & Giannouli, 2014) στο **Σχήμα 1**. Ουσιαστικά ο βέλτιστος αυτός προσανατολισμός επιτυγχάνεται όταν οι αποστάσεις των δεδομένων μεγιστοποιούνται από το υπερεπίπεδο που βρίσκεται στο κέντρο του σωλήνα (Μιχαλοπούλου, Ε. & Michalopoulou, 2016).



Σχήμα 1 Επίλυση προβλήματος παλινδρόμησης με χρήση SVM. Αριστερά παρατηρείται το μήκος 2^{ε} του υπερσωλήνα όπου τα όλα τα δεδομένα είναι έντος του και δεξιά παρατηρείται ο βέλτιστα προσανατολισμένος υπερσωλήνας μεγίστου περιθωρίου.

Άρα , παρατηρείται ότι ο στόχος είναι η μεγιστοποίηση ενός περιθωρίου . Είναι όμοιως με αυτόν στην περίπτωση γραμμικά διαχωριζόμενων δεδομένων όπου αναζητείτε το

$$J(\boldsymbol{w}, w_0) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 \tag{32}$$

αυτή τη φορά υπό άλλους περιορισμούς. Ώστε να επιτεχθεί ότι όλα τα δεδομένα $(x_i, \hat{y}_i)$ θα είναι εντός του υπερσωλήνα. Οπότε θέτονται η εξής συνθήκες:

$$\left.\begin{matrix} \mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{l} \leq \varepsilon \\ \hat{\mathrm{y}}_\mathrm{l} - \mathrm{y_i} \leq \varepsilon \end{matrix}\right\} \Leftrightarrow |\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{l}| \leq \varepsilon \; \forall \; \mathrm{i} = 1,2,\ldots,\mathrm{n} \tag{33}$$

όπου y_i η πραγματική τιμή και $\hat{y}_i = wx_i + b$ η εκτιμώμενη από το υπερεπίπεδο τιμή.

Στην παραπάνω περίπτωση όπου έχουμε υποθέση ότι είναι δυνατόν όλα τα δεδομένα να χωρέσουν, σ' ένα υπερσωλήνα πλάτους 2_ε . Όμως στην πραγματικότητα δεν είναι τόσο εφικτό να συμβεί. Οπότε σε αυτή την περίπτωση εισάγετε για κάθε δεδομένο που είναι εκτός τα όρια του υπερσωλήνα δηλαδή δεν ικανοποιήται ο περιορισμός $|y_i, \hat{y}_i| \leq \varepsilon \; \forall_i = 1,2,\ldots,n$, όπου μια χαλαρή μεταβλητή ξ_i^+ ή ξ_i^- ως ποινή. Αντίστοιχα

με το αν αρμόζει πάνω (ξ_i^+) ή κάτω (ξ_i^-) από τον σωλήνα. Η χρήση των χαλαρών μεταβλητών εξυπηρετούν ώστε να ενημερώσουν πόσο χρειάζεται να διορθωθούν τα δεδομένα ώστε να μετακινηθούν στο εσωτερικό του σωλήνα. Παρακάτω ορίζονται ως εξής:

$$\xi_i^+ \begin{cases} 0, & \alpha\nu\ y_i - \hat{y}_l \le \varepsilon \\ |y_i - \hat{y}_l| - \varepsilon, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \forall i = 1,2,\dots,n \tag{34}$$

$$\xi_i^- \begin{cases} 0, \alpha\nu\ \hat{y}_l - y_i \le \varepsilon \\ |y_i - \hat{y}_l - \varepsilon, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \forall i = 1,2,\dots,n \tag{35}$$

Εφόσον εφαρμόστηκαν οι χαλαρές μεταβλητές στο SVM παλινδρόμησης , χρειάζεται να συμπεριληφθούν στους περιορισμούς , καθώς και να προσθεθούν το αναλογό στην αντικειμενική συνάρτηση του προβλήματος βελτιστοποίησης. Εκφράζεται παρακάτω:

$$min\frac{1}{2}\left|\left|w\right|\right|^2 + C \cdot \sum_i^n(\xi_i^+ + \xi_i^-)\ \upsilon\pi\acute{o}\ \tau o\upsilon\varsigma\ \pi\varepsilon\rho\iota o\rho\iota\sigma\mu o\acute{\upsilon}\varsigma \begin{cases} y_i - \hat{y}_l \le \xi_i^+ + \varepsilon \\ \hat{y}_l - y_i \le \xi_i^- + \varepsilon \\ \xi_i^+ \ge 0, \xi_i^- \ge 0 \end{cases} \tag{36}$$

όπου = $i = 1,2,\dots,n$ (37)

Στην συνέχεια ακολουθείται παρόμοια διαδικασία όπως στην δυαδική ταξινόμηση. Εισάγωντας πολλαπλασιαστές Lagrange $\overrightarrow{\alpha^+} = (\alpha_1^+, \alpha_2^+, \dots, \alpha_n^+), \overrightarrow{a^-} = (a_1^-, a_2^-, \dots, a_n^-)$ και $\overrightarrow{\mu^+} = (\mu_1^+, \mu_2^+, \dots, \mu_n^+), \overrightarrow{\mu^-} = (\mu_1^-, \mu_2^-, \dots, \mu_n^-)$,με $a_i^+ \ge 0$, $\alpha_i^- \ge 0$ και $\mu_i^+ \ge 0, \mu_i^- \ge 0\ \forall i$. Ορίζεται Λαγκρανζιανή L_p , όπου παραγωγίζεται ως προς w, b, ξ_i^+και ξ_i^- θέτοντας τις παραγώγους ίσες με το μηδέν αντικαθιστούνται οι σχέσεις που προκύπτουν και ορίζεται η L_p και τέλος θα μεγιστοποιηθούν ως πρός a_i^+ και $\alpha_i^- \ge$ υπό τη συνθήκη $a_i^+ \ge 0$ και $\alpha_i^- \ge 0$. Με αυτήν την διαδικασία βρίσκονται οι τιμές των w και b όπου χρειάζεται για να οριστεί το SVM.

2.5.3 Παλινδρόμησης με μηχανές διανυσματικής υποστήριξης

Η παλινδρόμηση διανυσμάτων υποστήριξης αποτελεί μια εξειδίκευση αλγορίθμου SVM (Τσαρμπόπουλος, 2016). Είναι ένα απο τα σημαντικότερα πεδία της εφαρμογής η πρόβλεψη. Επίσης είναι ιδιαίτερα σημαντικό να αναφερθεί ότι η γραμμική παλινδρόμηση με μηχανές διανυσματικής υποστήριξης μπορεί να επεκταθεί σε μη γραμμική παλινδρόμηση με διάφορες μεθόδους όπως δημιουργώντας την βελτιστοποίηση Lagrangian (Γιαννούλη & Giannouli, 2014).

$$max_{\bar{a}} min_{\bar{x}} L(\bar{a}, \bar{x}) = max_{\bar{a}} min_{\bar{x}} (\varphi(\bar{x}) - \sum_{i=1}^{l} a_i g_i(\bar{x}))$$
(38)

υπο τους περιορισμούς

$$a_i \geq 0 \quad \text{για} \quad \text{κάθε} \quad i=1,....,I$$
(39)

Εδώ $g_i(\bar{x}) \geq 0$είναι οι περιορισμοί ανισότητας, οι μεταβλητές $\bar{\alpha}$ και $\bar{x}$ ονομάζονται δυικές πρωταρχικές μεταβλητές του προβλήματος βελτιστοποίησης.

Το αρχικό βήμα για την δημιουργία της βελτιστοποίησης Lagrange κατασκευάζονται περιορισμοί ανισότητας.

$$\xi_i + \varepsilon - y_i + \hat{f}(\bar{x}_l) \geq 0 \quad (40)$$
$$\xi_i' + \varepsilon - \hat{f}(\bar{x}_l) + y_i \geq 0$$
(41)
$$\xi_i \geq 0$$
(42)
$$\xi_i' \geq 0$$
(43)

Εφόσον έχουμε τέσσερις ομάδες με ανισοτικούς περιορισμούς θα χρειαστεί να εισάγουμε τέσσερις ομάδες δυικών μεταβλητών στην βελτιστοποίηση Lagrange. Αντικαθιστώντας την αντικειμενική συνάρτηση και τους περιορισμούς ανισότητας στην βελτιστοποίηση Lagrangian έχουμε τα ακόλουθα:

$$max_{\bar{a},\bar{a}',\bar{\beta},\bar{\beta}'}, min_{\bar{w},b,\bar{\xi},\bar{\xi}'} L\left(\bar{a},\bar{a}',\bar{\beta}\bar{\beta}',\bar{w},b,\bar{\xi},\bar{\xi}'\right)$$
$$= max_{\bar{a},\bar{a}',\bar{\beta},\bar{\beta}'}, min_{\bar{w},b,\bar{\xi},\bar{\xi}'} \left(\frac{1}{2}\bar{w}\cdot\bar{w}\right)$$
$$+ C\sum_{i=1}^{l}(\xi_i + \xi_i')$$
$$-\sum_{i=1}^{l} a_i\left(\xi_i + \varepsilon - y_i + \hat{f}(\bar{x}_l)\right)$$
$$-\sum_{i=1}^{l} a_i'\left(\xi_i + \varepsilon - \hat{f}(\bar{x}_l) + y_i\right) - \sum_{i=1}^{l}\beta_i\xi_i - \sum_{i=1}^{l}\beta_i'\xi_i')$$

υπό τους περιορισμούς

$$a_i, a_i', \beta_i, \beta_i' \geq 0$$

Για κάθε i=1,...,I και όπου $\hat{f}(\bar{x}) = \bar{w}\cdot\bar{x} - b$ (44)

Με δεδομένη μια λύση της βελτιστοποίησης της Lagrange

$max_{\bar{a},\bar{a}',\bar{\beta},\bar{\beta}'}\, n_{\bar{w},b,\bar{\xi}\bar{\xi}'} L\left(\bar{a},\bar{a}',\bar{\beta},\bar{\beta}',\bar{w},b,\bar{\xi},\bar{\xi}'\right) = L(\bar{a}^*,\bar{a}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)$ (45)

Γνωρίζοντας ότι θα ικανοποιούνται οι συνθήκες ΚΚΤ

$\frac{\partial L(\bar{a}^*,\bar{a}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial\bar{w}} = \bar{0}$ (46)

$\frac{\partial L(\bar{a}^*,\bar{a}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial b} = 0$ (47)

$\frac{\partial L(\bar{a}^*,\bar{a}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial\xi_i} = 0$ (48)

$\frac{\partial L(\bar{a}^*,\bar{a}'^*,\bar{\beta}^*,\bar{\beta}'^*,\bar{w}^*,b^*,\bar{\xi}^*,\bar{\xi}^*)}{\partial\xi_i'} = 0$ (49)

$\alpha_l^*\left(\xi_i^* + \varepsilon - y_i + \hat{f}^*(\bar{x}_l)\right) = 0$ (50)

$\alpha_l'^*\left(\xi_i'^* + \varepsilon - \hat{f}^*(\bar{x}_l) + y_i\right) = 0$ (51)

$\beta_i^*\xi_i^* = 0$ (52)

$\beta_i'^*\xi_i'^* = 0$ (53)

$\xi_i^* + \varepsilon - y_i + \hat{f}^*(\bar{x}_l) \geq 0$ (54)
$\xi_i'^* + \varepsilon - \hat{f}^*(\bar{x}_l) + y_i \geq 0$ (55)
$\xi_i^*, \xi_i'^* \geq 0$ (56)
$a_l, a_l' \geq 0$ (57)
$\beta_i^*, \beta_i'^* \geq 0$ (58)

Όπου κάθε i=1,...,I και $\hat{f}^*(\bar{x}) = \bar{w}^* \cdot \bar{x} - b^*$

Έχοντας ένα σύνολο εκπαίδευσης παλινδρόμησης

$$D = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots, (\bar{x}_l, y_l)\} \subseteq R^n x\ R,$$

Μπορεί να υπολογιστή το βέλτιστο μοντέλο διανυσματικής υποστήριξης για την παλινδρόμηση $\hat{f}^*(\bar{x}) = \bar{w}^* \cdot \bar{x} - b^*$με δυικό πρόβλημα.

$max_{\bar{\alpha},\bar{\alpha}^*} \varphi'(\bar{\alpha}, \bar{\alpha}') = max_{\bar{\alpha},\bar{\alpha}'} \left(-\frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} (a_i - a_i') + \sum_{i=1}^{l} y_i\ (a_i - a_i') - \varepsilon \sum_{i=1}^{l} (a_i + a_i') \right)$ (59)

Με τους περιορσμούς

$\sum_{i=1}^{l} (a_i + a_i') = 0$ (60)

$C \geq a_i, a_i' \geq 0$ (61)

Για κάθε i=1,...,I

$\bar{w}^* = \sum_{i=1}^{l} (a_i^* - a_i'^*)\, \bar{x}_l$ (62)

$b^* = \frac{1}{l} \sum_{i=1}^{l} \bar{w}^* \cdot \bar{x}_l - y_i$ (63)

Η παλινδρόμηση διανυσματικής υποστήριξης μπορεί να ερμηνεύσει (Γιαννούλη & Giannouli, 2014)
ζεύγη προτύπων (Διαμαντάρας, Κ, 2007) για την οποία ο συντελεστής $(a_i - a_i')$ είναι μη μηδενικός ως διάνυσμα υποστήριξης. Παρατηρείται ότι η λύση στο βέλτιστο μοντέλο παλινδρόμησης

$$\hat{f}^*(\bar{x}) = \bar{w}^* \cdot \bar{x} - b^* = \sum_{i=1}^{l}(a_i^* - a_i'^*)\bar{x}_l \cdot \bar{x} - \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i^* - a_i'^*)\bar{x}_l \cdot \bar{x}_j - y_j \quad (64)$$

εξαρτάται μόνο από τα διανύσματα υποστήριξης.Επίσης να αναφερθεί ότι η γραμμική παλινδρόμηση με μηχανές διανυσματικής υποστήριξης μπορεί να επεκταθεί σε μη γραμμική παλινδρόμηση με την εφαρμογή του τεχνάσματος του πυρήνα για τη βελτιστοποίηση και το μοντέλο. Όπου μπορεί να αντικατασταθεί το εσωτερικό γινόμενο στη βελτιστοποίηση και στο μοντέλο με μια κατάλληλη συνάρτηση πυρήνα για να επεκταθούν τα διανύσματα υποστήριξης παλινδρόμησης για μη γραμμικά σύνολα.

2.6 Αξιολόγηση Μοντέλου

Ο συνήθες στο στόχος στην πρόβλεψη σε ένα μοντέλο παλινδρόμησης είναι ο καθορισμός κάποιας συνάρτησης όπου θα δίνει όσο το δυνατόν πιο ακριβής την πρόβλεψη στο μοντέλο (Γιαννούλη & Giannouli, 2014). Επειδή ένα μοντέλο παλινδρόμησης με διανύσματα υποστήριξης έχει έναν αριθμό παραμέτρων που πρέπει να είναι συντονισμένες. Όπως ο πυρήνας k με τις αντίστοιχες παραμέτρους λ , και η σταθερά κόστους C. Για να μπορέσει να υπολογιστεί η ακρίβεια της πρόβλεψης στην παλινδρόμηση υπολογίζεται απο το σφάλμα παλινδρόμησης και δίνει την διαφορά μεταξύ της προβλεπόμενης και της πραγματικής τιμής των δεδομένων του μοντέλου (Γιαννούλη & Giannouli, 2014). Συνεπώς χρησιμοποιώντας το μέσο απόλυτο σφάλμα (MAE) όπου μας δίνει την διαφορά από την προβλεπόμενη και την πραγματική τιμή.Οπότε θα οριστεί μια συνάρτηση L_2 έχοντας τις παρατηρήσεις $(\bar{x}, y)$ενός μοντέλου $\hat{f}$,

$$L_2\big(y, \hat{f}(\bar{x})\big) = (y - \hat{f}(\bar{x})) \quad (65)$$

Υποθέτοντας ότι έχουμε ένα σύνολο εκπαίδευσης

$$D = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots, (\bar{x}_l, y_l)\} \subset R^n x\, R$$

Ορίζεται το μέσο απόλυτο σφάλμα υπολογίζεται στο D

$$MAE_D\left[\hat{f}_D[k,\lambda,\varepsilon,C]\right] = \frac{1}{l}\sum_{i=1}^{l} L_2\left(y_i, \hat{f}_D[k,\lambda,\varepsilon,C](\bar{x}_l)\right) \quad (66)$$

με $(\bar{x}_i, y_i) \in D$. Στην παραπάνω εξίσωση χρησιμοποιείται το μοντέλο $\hat{f}_D$ όπου ο δείκτης υποδεικνύει ότι κατασκευάστηκε χρησιμοποιώντας το σετ D. Ωστόσο το επιθυμητό αποτέλεσμα είναι όσο το δυνατόν λιγότερα σφάλματα για να υπάρχει μία πιο αξιόλογη πρόβλεψη. Γι αυτό χρησιμοποιούμε τα μέτρα αξιολόγησης επίδοσης για να έχουμε καλύτερη εποπτεία στην πρόβλεψη.

2.7 Παλινδρόμηση Κορυφογραμμής (Ridge Regression-L2)

Το φαινόμενο της πολυσυγγραμικότητας (multicollinearity) συμβαίνει όταν υπάρχει έντονη συσχέτιση ανάμεσα σε δύο ή περισσότερες μεταβλητές . Η εμφάνιση της πολυσυγγραμικότητας καταλήγει σε αυξημένα τυπικά σφάλματα των εκτιμήτριων ελαχίστων τετραγώνων , που έχει ως αποτέλεσμα να δυσκολεύει την εκτίμηση της επίδρασης των επεξεγηματικών μεταβλητών (ανεξάρτητων μεταβλητών) στην εξαρτημένη μεταβλητή (μεταβλήτη απόκρισης) . Επιπλέον σε τέτοιου είδους περιπτώσεις είναι δύσκολα να ανακαλυφθούν η στατιστικά σημαντικές μεταβλητές. Επίσης υπάρχουν ακραίες περιπτώσης της απόλυτης πολυσυγγραμικότητας όπου μια ανεξάρτητη μεταβλητή είναι γραμμικός συνδυασμός κάποιων ή όλων των άλλων ανεξάρτητων μεταβλητών .Σε τέτοιου είδους περιπτώσεις , η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί εφόσον έχει αφαιρεθεί μια μεταβλητή από το γραμμικά εξαρτημένο σύνολο. Καθώς,συχνά εμφανίζονται περιπτώσεις μεταβλητών χωρίς αυτές να είναι απόλυτα γραμμικά εξαρτημένες (Κάντα & Kanta, 2013).
Οπότε R_j^2 ο συντελεστής προσδιορισμού με εξαρτημένη μεταβλητή x_j και ανεξάρτητες όλες τις άλλες. Το R_j^2 εκφράζει κατά πόσο η x_j μπορεί να προβλεφθείαπό τις υπόλοιπες ανεξάρτητες μεταβλητές. Ο δείκτης 1-R_j^2 προσφέρεται σε ορισμένα προγράμματα στατιστικής ανάλυσης για εντοπισμό της πολυσυγγραμικότητας . Η $\frac{1}{1-R_j^2}$ είναι γνωστή ως παράγοντες μεγένθυσης διασποράς (variance inflation factor-VIF) δείχνει κατά πόσο αυξάνεται η διασπορά σ' ένα συντελεστή παλινδρόμησης όταν υπάρχουν συσχετίσεις των ανεξάρτητων μεταβλητών. Η μία κλασσική μέθοδος των ελαχίστων τετραγώνων δεν δίνει καλές εκτιμήτριες όταν υπάρχουν συσχετίσεις στα δεδομένα .Μία άλλη μέθοδος της επιλογής των μεταβλητών είναι ότι κατηγοριοποιεί τις μεταβλητές ως σημαντικές και μη σημαντικές . Η έντονη μεροληψία στην εκτίμηση μπορέι να δημιουργηθεί από την διαγραφή των μη σημαντικών. Βέβαια είναι καλύτερο να χρησιμοποιηθούν λίγο απ'όλες τις μεταβλητές απ'ότι να χρησιμοποιηθούν κάποιες εξ ολοκλήρου και κάποιες καθόλου.Συνεπώς αυτό κάνουν οι μεροληπτικές εκτιμήτριες (Κάντα & Kanta, 2013). Επίσης με την φιλοσοφία της Ridge Regression θεωρείται ότι είναι προτιμότερο να κρατηθούν κάποιες πληροφορίες απ' όλες τις μεταβλητές παρά να κρατηθούν μόνο μερικές από τις μεταβλητές και κάποιες να απορριφθούν εντελώς.Συνεπώς αυτή είναι η μέθοδος που ακολουθεί η κορυφογραμμή.

Η εκτιμήτρια με τη μέθοδο της κορυφογραμμής της παραμέτρου $\beta = (\beta_1, \beta_2, \dots \dots, \beta_n)$,προκύπτει

$$(X'X + \lambda I)\hat{\beta}^* = X'y \Rightarrow$$
$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y. \quad (67)$$

Όπου το Χ΄Χ είναι πίνακες ιδιοτιμών το λ είναι μια ρυθμιστική παράμετρος . Ισοδύναμα οι εκτιμήτριες των συντελεστών β_j υπολογίζονται έτσι ώστε να ελαχιστοποιούνται

$$\widehat{\beta} = argmin \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad (68)$$

υπό τον περιορισμό

$$\sum_{j=1}^{p} \beta_j^2 \leq t,$$

t είναι μια ρυθμιστική παράμετρος. Η εμφάνιση του όρου λI στην παραπάνω σχέση η του περιορισμού είναι αυτή που δικαιολογεί την έννοια της "ποινής" η μέθοδος στην εκτίμηση των συντελεστών . Επίσης θα αναφερθεί η λύση που δίνει η μέθοδος της κορυφογραμμής για ένα εύρος αποδεκτών τιμών .Μια αποδεκτή τιμή σημαίνει τιμή ότι η αντίστοιχη εκτιμήτρια έχει μικρότερο μέσο τετραγωνικό σφάλμα .Επομένως το μεσο τετραγωνικό σφάλμα των μελλοντικών προβλέψεων όπου μειώνονται αντίστοιχα. Εφόσον η εκτιμήτρια

$$\widehat{\beta} = \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{y} \quad (69)$$

είναι αμερόληπτη, γνωρίζοντας από την παραπάνω σχέση ότι η $\hat{\beta}_{ridge}$ είναι μεροληπτική. Καθώς η διασπορά $V(\hat{\beta}_{ridge})$είναι μικρότερη τιμή της V($\widehat{\beta}$).Οπότε το μέσο τετραγωνικό σφάλμα της $\hat{\beta}_{ridge}$μπορεί να είναι αρκετά μικρότερο της $\widehat{\beta}$. Σε ένα μοντέλο , το οποίο έχουν τυποποιηθεί όλοι οι συντελεστες έχουν περίπου ίσες διασπορές. Μπορεί δηλαδή να γίνει επιλογή μεταβλητών με βάση την απόλυτη τιμή τους.Επίσης μπορεί να διαγραφούν οι μεταβλητές που οι συντελεστές είναι αρκετά μικροί κατά την απόλυτη τιμή τους .Όπως είναι κατανοητό η μέθοδος της κορυφογραμμής μπορεί να χρησιμοποιηθεί τόσο ως μέθοδος συντελεστών σε ένα μοντέλο όσο και ως μέθοδο επιλογής μεταβλητών (Κάντα & Kanta, 2013).

2.8 LASSO (Least Absolute Shrinkage and Selection Operator-L1)

Στην μέθοδο της κορυφογραμμής όπου αναλύθηκε παραπάνω μας δίνει μοντέλα με λιγότερες μεταβλητές και πιο ερμηνευσιμα αλλα είναι εξαιρετικά ασταθής διαδικασία. Προτείνεται μια μέθοδος λεγόμενη Lasso.Όπου αυτή η μέθοδος μειώνει κάποιους συντελεστές και κάποιους άλλους τους θέτει ίσους με μηδέν .Καθώς αποτελεί και αυτή μια μέθοδο ποινής (Κάντα & Kanta, 2013).

Αν υποθέσουμε ότι έχουμε δεδομένα (x_i, y_i),όπου$x_i = (x_{i1}, \dots, x_{ip})$ και $y_i, i = 1,2, \dots, n.$ είναι οι τιμές που αντιστοιχούν στην i παρατήρηση σ' ένα συνήθες γραμμικό μοντέλο . Καθώς για την εκτίμηση της παραμέτρου $\beta = \beta_1, \beta_2, \dots, \beta_p$,με την μέθοδο Lasso η εκτιμήτρια ορίζεται ως εξής:

$$(\hat{a}, \hat{\beta}) = argmin\left\{\left(y_i - a - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2\right\} \qquad (70)$$

υπό τον περιορισμό

$$\sum_{j=1}^{p} \left|\beta_j\right| \leq t.$$

Όπου η παράμετρος t είναι μια ρυθμιστική παράμετρος.Η παραπάνω σχέση ως προς α είναι $\hat{a}$=$\bar{y}$. Χωρίς να επηρεαστεί αρνητικά η γενικότητα παρατηρείται ότι $\bar{y}$=0 επομένως να παραβλέψουμε το α . Οπότε με τις παραπάνω σχέσεις προκύπτει ότι οι εκτιμήτριες των συντελεστών ενός γραμμικού μοντέλου με τη μέθοδο Lasso δίνεται από την λύση του συστήματος

$$\hat{\beta} = argmin\left\{\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2\right\} \qquad (71)$$

υπο τον περιορισμό

$$\sum_{j=1}^{p} \left|\beta_j\right| \leq t.$$

όπου η παράμετρος $t \geq 0$ ελέγχει το ποσό της συρρίκνωσης που υφίστανται οι συντελεστές .

Ωστόσο αξίζει να σημειωθεί ότι οι μέθοδοι L1 και L2 αποτελούν ειδικές περιπτώσεις. Τέλος η Lasso μπορεί να εφαρμοστεί και στην περίπτωση γενικευμένων γραμμικών μοντέλων (Κάντα & Kanta, 2013).

2.9 Πολυωνυμική Παλινδρόμηση

Στην επιστήμη της στατιστικής, η πολυωνυμική παλινδρόμηση αποτελεί μία μορφή της γραμμικής παλινδρόμησης, στην οποία η σχέση ανάμεσα στην ανεξάρτητη μεταβλητή x και στην εξαρτημένη μεταβλητή y μοντελοποιείται μέσω ενός πολυωνύμου νιοστού βαθμού (Μοναχόπουλος, 2016). Ένα κοινό χαρακτηριστικό των μηγραμμικών μοντέλων παλινδρόμησης που δίνεται από γραμμικές συναρτήσεις της εξαρτημένης μεταβλητής y για την ανεξάρτητη μεταβλητή x είναι ότι οι συναρτήσεις είναι μονότονες ,αύξουσες ή φθίνουσες. Σε κάποιες περιπτώσεις η πολυωνυμική συνάρτηση κάποιου βαθμού k μπορεί να παράγει μια αποδεκτική προσέγγιση της πραγματικής συνάρτησης παλινδρόμησης. Όπου το μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης k όπου απεικονίζεται στην παρακάτω εξίσωση, θεωρόντας ότι τα σφάλματα παλινδρόμησης συνοδεύουν γκαουσιανή κατανομή με μέσο όρο 0 και διασπορά σ_0^2 (Ασπιρτάκης, Κουμπούλης, & Πετράκη, 2016).

Καθώς ένας πολυωνυμικός όρος, όπως ο τετραγωνικός ή ο κυβικός, μετατρέπει τη γραμμική παλινδρόμηση σε πολυωνυμική. Επειδή όμως οι τιμές αυτές ορίζονται σαν πολυωνυμικοί όροι στα δεδομένα εισόδου και **όχι** στους συντελεστές των συναπτικών βαρών(παράμετροι), το μοντέλο παραμένει γραμμικό. Αυτό μας επιτρέπει την κατασκευή μίας μη γραμμικής χαρακτηριστικής συνάρτησης, χωρίς να εμπλεκόμαστε σε ένα πιο πολύπλοκο μη γραμμικό μοντέλο. Για το λόγο αυτό, η πολυωνυμική παλινδρόμηση ορίζεται σαν μία ειδική κατηγορία των πολλαπλών γραμμικών μοντέλων (Μοναχόπουλος, 2016) .
Επίσης για την αξιολόγηση που γίνεται στις παραμέτρους χρησιμοποιώντας την μέθοδο ελαχίστων τετραγώνων, διότι σε ένα μοντέλο γραμμικής παλινδρόμησης οι ανεξάρτητες μεταβλητές είναι μη-γραμμικές στην συνάρτηση πολυωνυμικής παλινδρόμησης όμως οι παραμέτρους $\beta_0, \beta_1, \dots, \beta_k$ είναι γραμμικά, οπότε το σύνολο των σφάλματων με την χρήση ελαχίστων τετραγώνων είναι ένα διμετάβλητο μέρος διάστασεις n των (X,Y), έχει την μορφή(Ασπιρτάκης et al., 2016):

$$f(\beta_0,\beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_{1xi} + \beta_{2xi}^2 + \cdots + \beta_{kxi}^k\right)\right)^2 \qquad (72)$$

Όπου σε ένα μοντέλο πολυωνυμικής παλινδρόμησης η αξιολογηση των σφαλμάτων γίνεται με την εφαρμογή των ελαχίστων τετραγώνων όπου είναι $e_i = y_i - \hat{y}_i$, και $\hat{y}_i = \beta_0 + \beta_{1xi} + \beta_{2xi}^2 + \cdots + \beta_{kxi}^k$. Για την εκτίμηση της διασποράς των σφαλμάτων e_i ορίζεται ως εξής:

$$\sigma_e^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (73)$$

Τα μοντέλα πολυωνυμικής παλινδρόμησης συνήθως εξάγουν μία συνάρτηση, που ταιριάζει στο μοντέλο εισόδου, χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων ακριβώς όπως πραγματοποιεί η γραμμική παλινδρόμηση. Η μέθοδος αυτή ελαχιστοποιεί τη διασπορά που επιφέρουν οι συντελεστές εκτίμησης (παράμετροι). Η πρώτη σχεδίαση ενός πειραματικού μοντέλου πολυωνυμικής παλινδρόμησης, έκανε την εμφάνιση του το 1815, από τον Gergonne. Στον εικοστό αιώνα, η πολυωνυμική παλινδρόμηση έπαιξε σημαντικό ρόλο στην ανάπτυξη της παλινδρομικής ανάλυσης, δίνοντας μεγάλη έμφαση στην σχεδίαση και στα αποτελέσματα που σήμερα έχουμε επιφέρει. Πρόσφατα, η χρήση των πολυωνυμικών μοντέλων έχουν συμπληρωθεί από άλλες μεθόδους, χρησιμοποιώντας μη πολυωνυμικά μοντέλα, έχοντας κάποια πλεονεκτήματα σε συγκεκριμένα όμως προβλήματα (Ασπιρτάκης et al., 2016).

ΚΕΦΑΛΑΙΟ 3

3.1 Εισαγωγή

Για την αποτελεσματική μοντελοποίηση ενός προβλήματος ακολουθούμε την παρακάτω μεθοδολογία. Λαμβάνουμε υπόψη ότι δεν υπάρχει κάποια αναλυτική σχέση μεταξύ αιτίας και αποτελέσματος, αλλά μόνο ένα σύνολο παρατηρησιακών δεδομένων στη μορφή αιτία-αποτέλεσμα. Η αιτία κωδικοποιείται ώς ένα διάνυσμα ιδιοτήτων από τις οποίες θεωρούμε ότι εξαρτάται το αποτέλεσμα. Θεωρούμε δηλαδή το πρόβλημα ως ένα σύστημα το οποίο έχει εισόδους (ιδιότητες που εκφράζουν την αιτία ενός αποτελέσματος) εξόδους (αποτέλεσμα). Επίσης, το πραγματικό σύστημα επιτελεί την επεξεργασία των εισόδων για την παραγωγή του αποτελέσματος Η επεξεργασία αυτή είναι άγνωστη σε εμάς (μαύρο κουτί). Η διαδικασία της μοντελοποίησης έγκειται στη δημιουργία μιας παραμετρικής συνάρτησης (μοντέλο) η οποία θα έχει τις ίδιες εισόδους με το πραγματικό σύστημα και θα παράγει την ίδια έξοδο με το πραγματικό σύστημα για τα ίδια δεδομένα εισόδου. Αν αυτό είναι εφικτό τότε μπορούμε να αντικαταστήσουμε το πραγματικό σύστημα με το μοντέλο, το οποίο μοντέλο αναμένεται να προσομοιώνει πιστά το πραγματικό σύστημα. Το κέρδος από τη διαδικασία αυτή είναι η δυνατότητα πρόβλεψης για νέα άγνωστα δεδομένα την έκβαση των οποίων μπορούμε να προβλέψουμε με τη χρήση του μοντέλου, όπως επίσης και η δυνατότητα μελέτης σεναρίων.

Για την αποτελεσματική μοντελοποίηση θα πρέπει να ακολουθήσουμε μια σειρά από βήματα τα οποία περιγράφονται παρακάτω:

1. Προσδιορισμός των υποψήφιων ιδιοτήτων του συστήματος των εισόδων και της/των εξόδων του.
2. Εύρεση επαρκών και αντιπροσωπευτικών δεδομένων (περιπτώσεων-περιστατικών). Η αναζήτηση γίνεται από αρχειακές βάσεις δεδομένων του οργανισμού.
3. Προσδιορισμός των σημαντικών εισόδων του συστήματος από τα διαθέσιμα δεδομένα .
4. Επιλογή μοντέλου.
5. Αξιολόγηση και τελική επιλογή του μοντέλου που είναι καταλληλότερο για το συγκεκριμένο πρόβλημα.

3.2 Υποψήφιες ιδιότητες

Το πρώτο βήμα στη διαδικασία της μοντελοποίησης είναι ο προσδιορισμός των ιδιοτήτων που επηρεάζουν το αποτέλεσμα, το οποίο θέλουμε να μοντελοποιήσουμε Αρχικά προσδιορίζουμε τις ιδιότητες εκείνες που εμείς θεωρούμε ότι είναι σημαντικές, σύμφωνα με την εμπειρία μας, τη γνώση μας για το συγκεκριμένο πρόβλημα και τη διαίσθηση μας. Το βήμα αυτό δεν είναι δυνατό να πραγματοποιηθεί από μια υπολογιστική μηχανή και πρέπει να γίνει αποκλειστικά από τον άνθρωπο. Με βάση τα παραπάνω και λαμβάνοντας υπόψη και το γεγονός ότι για τις ιδιότητες που θα εντοπίσουμε θα πρέπει να υπάρχουν και διαθέσιμα δεδομένα καταλήξαμε στον παρακάτω πίνακα 3.

Α/Α	Ιδιότητα	Αιτιολόγηση (γιατί θεωρούμε ότι παίζει ρόλο)	Σχόλια	τύπος ιδιότητας
1	Φύλο	Στην Ελλάδα με το Σύνταγμα του 1975 θεμελιώθηκε η ισότητα των δύο φύλων, όπου ορίζεται "οι Έλληνες είναι ίσοι ενώπιον του νόμου κι έχουν ίσα δικαιώματα και υποχρεώσεις" καθιερώθηκε την χρονολογική περίοδο 1981-1989, καθώς ο σκοπός που καθιερώθηκαν οι νόμοι είναι η κατάργηση της ανισότητας απέναντι στο γυναικείο φύλο και η ύπαρξη ισότητας σε όλους τους τομείς (Αρσενίου, 2015). Επίσης στα προηγούμενα χρόνια η εκπαίδευση ήταν αποκλειστικό δικαίωμα μόνο των αντρών , όμως λόγω της ύπαρξης της κατάργησης της ανισότητας οι γυναίκες έχουν εμφανίση έντονη παρουσία και θέση στον τομέα της εκπαίδευσης. Καθώς σε αρκετές σχολές οι γυναίκες είναι που έχουν την καλύτερη απόδοση σε αντίθεση με τα αγόρια που όσο πάει ελαχιστοποιείται η ανάδειξη τους (Αρσενίου, 2015). Ωστόσο αρκετές έρευνες εμφάνισαν ότι ανάμεσα στα δύο φύλα στον επαγγελματικό τομέα υπάρχουν διαφορές, καθώς υφίστανται βαθιές παραδόσεις αρκετών χρόνων που επηρεάζουν τους ρόλους στην κοινωνία (Τσιλιμίγκρα, Stati, Στάτη, & Tsilimigkra, 2017). Επίσης αναλύσεις των διαφορών των φύλων κατά τη χρήση της αυτο-ρυθμιζόμενης μάθησης αποκάλυψε ότι τα κορίτσια ανέφεραν σημαντικά μεγαλύτερο ρεκόρ τήρησης και παρακολούθησης, από ό, τι τα αγόρια (Zimmerman & Martinez-Pons, 1990).		0,1 0=κορίτσι 1=αγόρι
2	Το απολυτήριο	Έρευνα δείχνει ότι οι μαθητές που έχουν ιστορικά καλύτερες επιδόσης στο επίπεδο της δευτεροβάθμιας εκπαίδευσης, έχουν μεγαλύτερη απόδοση στο επίπεδο της τριτοβάθμιας εκπαίδευσης ως φοιτητές , (Watson, Creed, & Patton, 2003) και ίσως αυτό να σηματοδοτεί ότι ο υψηλός βαθμός απολυτηρίου μπορεί να αποφέρει καλύτερους βαθμούς, αλλά θα πρέπει να τονιστεί ότι για να μπορέσει να εισαχθεί στην τριτοβάθμια εκπαίδευση θα χρειαστεί να έχει απολυτήριο λυκείου , καθώς οι νεοεισερχόμενοι φοιτητές για τα Α.Τ.Ε.Ι και Πανεπιστημία αποφασίζει την διαδικασία τοΥπουργείο Παιδείας (Κουδούνης, Μιχάλης, 2015).		0=όχι 1=ναι

3	Περιοχή	Για τα παιδιά που δεν έχουν καταγωγή από τα αστικά κέντρα υπάρχει μειωμένη απόδοση, όπως και χαμηλά ποσοστά εισαγωγής στα Α.Τ.Ε.Ι. και Πανεπιστήμια(Παπάνης & Βίκη, 2007b). Η καταγωγή από Αγροτικές οικογένειες φοιτητών, σε σύγκριση με τις αστικές, τα αντίθετα μέρη τους είχαν μοναδικές συνθήκες να αντιμετωπίσουν κατά τη λήψη αποφάσεων σχετικά με την εκπαίδευση και την επαγγελματική σταδιοδρομία. Μειωμένη δυνατότητα πρόσβασης στην τριτοβάθμια εκπαίδευση, περιορισμένα στην επαρχία τα σχολικά προγράμματα, περιορισμένη έκθεση στον κόσμο των επαγγελμάτων και μερικά πρότυπα έχουν σημαντικούς περιορισμούς . Δεν είναι έκπληξη το γεγονός ότι οι πρώτες μελέτες των μαθητών που έχουν φιλοδοξίες αποκάλυψε μια σημαντική διαφορά μεταξύ των φοιτητών από επαρχία και αστικών λύκειο στα σχέδια για να πάει σε κάποια μορφή μετά τη δευτεροβάθμια εκπαίδευση. Ένα μεγαλύτερο ποσοστό των αστικών φοιτητών, άνδρες και γυναίκες, που έχουν προγραμματιστεί για κολέγιο παρά από την επαρχία (Middleton & Grigg, 1959). Στον τομέα της εκπαίδευσης υπάρχουν περισσότερες ευκαιρίες για τα παιδιά που είναι από τα αστικά κέντρα συγκριτικά με τα παιδιά από την επαρχεία (Παπάνης & Βίκη, 2007b).		0=Επαρχία 1=Πόλη
4	Βαθμοί πρόσβασης,Δελτίο επιτυχίας	Οι εκπαιδευόμενοι της δευτεροβάθμιας εκπαίδευσης που τείνουν να έχουν αρκετά καλές επιδόσεις και να εισάγωνται με υψηλές βαθμολογίες στην τριτοβάθμια εκπαίδευση .Θεωρείται ότι θα έχουν και υψηλές φοιτητικές επιδόσεις (Ντούκα & Φραγκίσκου, 2016).		0= Κάτω από την βάση 1= Άνω από την βάση

5	Οικογενειακή κατάσταση	Το εκπαιδευτικό αλλά και το μορφωτκό επίπεδο των γονέων, δείχνει να επιδρά σηματικά στην πορεία της εκπαίδευσης ενός νέου, όπως και για την εισαγωγή του στην τριτοβάθμια εκπαίδευση (Χαλικιά & Κιτσαρά, 2016). Γι αυτό οι πολύ μορφωμένες οικογένειες είναι περισσότερο αποτελεσματικές στην ανάπτυξη των παιδιών τους σε αντίθεση με την κατώτερες σχολικές οικογένειες (Laosa, 1982). Διαπιστώθηκε από έρευνα ότι το μορφωτικό επίπεδο των γονέων σε φοιτητές του ΤΕΙ είναι στην πλειονότητα απόφοιτοι δημοτικού σχολείου (Μαύρος, 1998). Έπειτα οι οικογενειακοί ερευνητές έχουν συγκρίνει τα οικογενειακά αποτελέσματα μικρών και μεγάλων οικογενειών, καθώς αυξάνεται ο αριθμός των αδελφών,η εκπαιδευτική επίδοση μειώνεται (Downey, 1995). Επίσης τα παιδιά με δύο γονείς που εργάζονται δεν έχουν αρκετό χρόνο οι γονείς τους για την υγιή ανάπτυξη (35). Ειδικά ακόμη πιο δύσκολο για εκείνους που στηρίζουν τα παιδιά μόνοι τους. (Halpern & Murphy, 2013) Στη διάρκειας ζωής ενός παιδιού σε μια μονογονεϊκή οικογένεια. Από εμπειρικά συμπεράσματα δείχνουν ότι έχει αρνητική επίδραση στη ζωή και μειώνει το εκπαιδευτικό τους επίπεδο (Krein & Beller, 1988). Εξίσου τα παιδιά από διαζευγμένους γονείς μειώνει την ικανότητα εκπαιδευτικής επίτευξης ,διαζύγιο και χωρισμός συσχετίζονται θετικά με μειωμένη σχολική επίδοση. (Fagan & Churchill, 2012). Αρκετοί επιστήμονες από την Ελλάδα και το εξωτερικό επισημαίνουν ότι, όταν μια οικογένεια είναι κοινωνικά καταξιωμένη τόσο μεγαλύτερες επαγγελματικές προσδοκίες έχει ένας νέος. (Παπάνης & Βίκη, 2007a). Πιθανώς ένας από τους μεγαλύτερους παράγοντες που καθορίζουν την ακαδημαϊκή επιτυχία και γενικότερα τις επιδόσεις των παιδιών είναι η συμμετοχή των γονέων στην σχολική ζωή (Ζαγκλαρά, n.d.).		0=Χαμηλή έως βασική εκπαίδευση 1=Υψηλό εκπαίδευση

6	Οικονομική κατάσταση	Ωστόσο αρκετές μελέτες παρουσιάζουν ότι, οι νέοι που είναι από οικονομικά άνετες οικογένειες μπορούν, να ανταπεξέλθουν πιο εύκολα στην επιλογή του επαγγέλματος , και συνήθως εμφανίζουν πιο αποτελεσματικές εκπαιδευτικές (Χαλικιά & Κιτσαρά, 2016).Επίσης για τα αγόρια ενδέχεται ότι, το επάγγελμα και η οικονομική αμοιβή του πατέρα επηρεάζει τις επαγγελματικές φιλοδοξίες ενός παιδίου , δηλαδή όσο καλά αμειβόμενο και με κονωνικό κύρος επάγγελμα έχει ο πατέρας , το παιδί επιδιώκει το αντίστοιχο ενώ αν ο πατέρας έχει χαμηλό αμειβόμενο επάγγελμα ο γιός δεν θέτει μεγάλους επαγγελματικούς στόχους (Παπάνης & Βίκη, 2007b). Για τις μητέρες που δραστηριοποιούνται επαγελματικά και που έχουν ένα υψηλό εκπαιδευτικό επίπεδο , δίνει θετικό αντίκρισμα στις κόρες όπου εμφανίζουν υψηλές σχολικές και επιδιώκουν πιο ενεργά θέσεις στον επαγγελματικό τομέα (Παπάνης & Βίκη, 2007a). Ωστόσο διαπιστώνεται ότι για τα παιδιά που προέρχονται απο οικογένειες με οικονομικό και κοινωνικό επίπεδο συνήθως επιδρά για τους επαγγελματικούς στόχους (Χαλικιά & Κιτσαρά, 2016).		0=Χαμηλό εισόδημα 1=Υψηλό εισόδημα
7	Σειρά επιτυχίας	Οι ακαδημαϊκές και οι επαγγελματικές προσδοκίες για έναν νέο, ειναι ενας βασικος πυλωνας στην επιλογή αλλά και στην πορεία των σπουδών του στην τριτοβάθμια εκπαίδευση (Ζαγκλαρά, n.d.). Η επιθυμία των νέων για την εκπαίδευση είναι μεγάλη , διότι στοχεύουν να εισαχθούν στην τριτοβάθμια εκπαίδευση (Ζαγκλαρά, n.d.). Το Υπουργείου Παιδείας οδηγήθηκε στο αποτέλεσμα ότι μεγαλύτερο ποσοστό των μαθητών που εισέρχονται στην τριτοβάθμια εκπαίδευση δεν εισάγεται σε σχολές της πρώτης του προτίμησης αλλά σε σειρά από την 7η και μετά. Ειδικότερα για τους εισακτέους στα Τ.Ε.Ι., το παραπάνω φαινόμενο παρατηρείται σε μεγαλύτερο ποσοστό σε σχέση με τους εισακτέους σε Α.Ε.Ι. (Κουδούνης, Μιχάλης, 2015). Συνήθως η επιλογή της σχολής γίνεται με γνώμονα την επαγγελματική αποκατάσταση, οι σχολές που προτιμούν οι νέοι είναι αυτές που έχουν μεγάλες αμοιβές (Παπάνης & Βίκη, 2007b).		0=Επιλογή σειράς δέκα και πάνω 1=Επιλογή σειράς από εννέα και κάτω
	έξοδος του συστήματος	1. χρόνος αποφοίτησης σε εξάμηνα (regression) 2. 0 η 1 0=κανονική αποφοίτηση 1=καθυστερημένη αποφοίτηση.(classification)		Ακέραιος

Πίνακας 3: Υποψηφίων ιδιοτήτων

3.3 Εύρεση δεδομένων για τις υποψήφιες ιδιότητες

Τα δεδομένα τα οποία συλλέχθηκαν ήταν από το ΑΤΕΙ Κρήτης της σχολής Διοίκηση και Οικονομίας του τμήματος Διοίκηση επιχειρήσεων . Το αρχικό δείγμα των δεδομένων ήταν 446 ,όμως αποφασίστηκε για να υπάρξουν πιο αξιόπιστα αποτελέσματα ότι θα χρειαστεί από το δείγμα των δεδομένων όσοι έχουν αποφοιτήσει από 4 μέχρι 10 χρόνια . Όπου το κάθε δεδομένο αποτελείται από 10 ιδιότητες όμως δεν θα χρησιμοποιηθεί η ιδιότητα x8 όπου είναι ο βαθμός πτυχίου ,διότι στην περίπτωση που θα χρειαστεί να γίνει πρόβλεψη ενός νέου φοιτητή για το χρόνο αποφοίτησης του δεν θα είναι διαθέσιμος ο βαθμός πτυχίου επομένως θα χρησιμοποιηθούν 9 ιδιότητες.
Οπότε το σύνολο που θα επεξεργάστει το μοντέλο είναι 340, όπου τα 272 χρησιμοποιήθηκαν ώς δεδομένα εκπαίδευσης (train data) και τα 68 έως δεδομένα ελέγχου (test data) (Athanasiadis & Αθανασιάδης, 2015). Στον παρακάτω πίνακα απεικονίζονται οι ιδιότητες των δεδομένων που θα επεξεργαστούν.

'x1'	Φύλο
'x2'	Καταγωγή
'x3'	Υπηκοότητα
'x4'	Σειρά Προτίμησης
'x5'	Απολυτήριο Λυκείου
'x6'	Βαθμός Εισαγωγής
'x7'	Σειρά Επιτυχίας
~~'x8'~~	~~Βαθμός Πτυχίου~~
'x9'	Επάγγελμα Πατέρα
'x10'	Επάγγελμα Μητέρας

Πίνακας 4: Ιδιότητες δεδομένων

ΚΕΦΑΛΑΙΟ 4

Εφαρμογή Μοντέλων

4.1 Εισαγωγή

Στη διαδιακασία εκπαίδευσης του συστήματος η επιλογή του ιδανικότερου μοντέλου είναι αυτή που το μοντέλο θα δίνει την καλύτερη πρόβλεψη. Τα μοντέλα που χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος είναι πέντε τα οποία είναι LinearRegression, Ridge Regression(L-2), LASSO(L-1), Polynomial Regression και SVR. Οι παράμετροι του κάθε μοντέλου είναι διαφορετική ωστόσο θα αναλυθεί παρακάτω η χρήση και η εφαρμογή τους. Να σημειωθεί ότι η επιλογή του μοντέλου γίνεται μετά την επεξεργασία των δεδομένων. Στο παρόν κεφάλαιο θα αναλυθεί μόνο η μεθοδολογία του κάθε μοντέλου που χρησιμοποιήθηκε. Οπότε για την επιλογή του μοντέλου είναι η βέλτιστη απόδοση με το ελάχιστο σφάλμα που θα δώσει το κάθε μοντέλο (Athanasiadis & Αθανασιάδης, 2015).

4.2 Μοντέλο SVR

Κάνοντας χρήση του μοντέλου Μηχανών Διανυσματικής Υποστήριξης με μορφή παλινδρόμησης-Support Vector Machine(SVR) όπου αναζητείται η βέλτιστη απόδοση του μοντέλου. Για να επιτευχθεί η βέλτιστη απόδοση του μοντέλου ακολουθείται κάποια διαδικασία όπου γίνονται αρκετές δοκιμές και συγκρίσεις με της παραμέτρου του μοντέλου. Αρχικά επιλέγεται ο κατάλληλος πυρήνας όπου είναι ο κατάλληλος αναλόγως με το είδος 0του πρόβληματος, όπου στην προκειμένη περίπτωση έγινε η χρήση του πυρήνα RBF radial basis function . $k(x, y) = e^{-||x-y||^2/(2\sigma)^2}$ (Athanasiadis & Αθανασιάδης, 2015). Έχοντας κάνει την χρήση του πυρήνα rbf . Η επόμενη διαδικασία στον πυρήνα rbf είναι ότι χρειάζεται να υπολογιστεί η παράμετρος γ και C . Η C είναι η παράμετρος ποινής του όρου σφάλματος, και η γ είναι παράμετρος της rbf σε αυτό το σημείο γίνονται αρκετές δοκιμές μέχρι να καταλήξουμε ποιός είναι ο ιδανικός συνδυασμός που προσδιορίζει την βέλτιστη απόδοση του μοντέλου οι τιμές των παραμέτρων είναι C=2.0 και gamma=10.0. Καθώς εφαρμόζεται η μέθοδος cross validation καθώς αναζητώντας την βέλτιστη απόδοση του μοντέλου δεν γνωρίζουμε ποιές είναι οι ιδιότητες που μας δίνουν τις βέλτιστες τιμές. Ωστόσο μέσω τις επαναλήψεις βρίσκονται οι ιδιότητες που μας δίνουν τις βέλτιστες τιμές και χρησιμοποιούνται για την εκπαίδευση του συστήματος. Μια συνηθισμένη μέθοδος για να εκπαιδευτεί το σύστημα είναι ο χωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.Επίσης στο k-fold cross validation χωρίζεται σε k υποσύνολα , συγκεκριμένα 5. Έτσι ένα υποσύνολο εξετάζεται χρησιμοποιώντας τα υπόλοιπα k-1 υποσύνολο. Επομένως προβλέπεται ολόκληρο το σύνολο εκπαίδευσης από μια φορά. Η διαδικασία είναι απλή με έγκυρα αποτελέσματα, όμως απαιτεί αρκετό χρόνο λόγω των επαναλήψεων που απαιτούν οι συνδυασμοί .Έχοντας της κατάλληλες τιμές από της παραμέτρους γίνεται η εκπαίδευση του μοντέλου (Athanasiadis & Αθανασιάδης, 2015).

4.3 Μοντέλο LinearRegression

Το μοντέλο LinearRegression καθώς εκπαιδευέται ως ώστε να δίνει την ποιό ακριβής πρόβλεψη .Ακολουθείται κάποια διαδικασία για την εκπαίδευση του.Αρχικά χρειάζεται να οριστούν οι παράμετροι καθώς επιλέγονται ανάλογα με το πρόβλημα .Η παράμετρος που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου ήταν fit_intercept.Αρχικά η παράμετρος fit_intercept χρησιμοποιείται για τον υπολογισμό του μοντέλου και τείθεται σε True ή False όπου αν τεθεί True σημαίνει ότι γίνεται υπολογισμός με παρακολούθηση του μοντέλου αν τεθεί σε False στο μοντέλο δεν γίνεται υπολογίσμος με παρακολούθηση,στη δική μας περίπτωση η παράμετρος fit_intercept τέθηκε σε True διότι θέλουμε τον υπολογισμό της παρακολούθησης του μοντέλου.Η διαδικασία επιλογής των παραμέτρων καθώς και η ρύθμιση τους δεν απαιτούν αρκετές δοκιμές παραμόνο να γνωρίσουμε το είδος του προβλήματος που μας έχει ανατεθεί να επιλύσουμε.
Οπότε για να επιτευχθεί ο βασικός μας στόχος που είναι η πρόβλεψη να είναι πιο ακριβής. Επίσης αυτό που χρειάζεται για να γίνει μια ορθή εκπαίδευση του συστήματος χρησιμοποιείται η μέθοδος cross validation η οποία χωρίζει το σύνολο των δεδομένων σε k υποσύνολα (δηλαδή σε 5 σετ) δεδομένα εκπαίδευσης και δεδομένα ελέγχου ώστε να εκπαιδευτούν όλα και να διασταυρωθούν ποια δεδομένα δίνουν τις βέλτιστες τιμές .Έχοντας λοιπόν εφαρμόσει όλες της παραπάνω διαδικασίες μπορεί να γίνει η εκπαίδευση του μοντέλου .

4.4 Μοντέλο Ridge (L-2)

Εφαρμόζοντας το μοντέλο Ridge Regression (L-2) όπου χρησιμοποιείται ώστε να μπορέσει να δώσει μια όσο το δυνατόν πιο ακρίβης πρόβλεψης .Στο σημείο της υλοποίησης για την εκπαίδευση του μοντέλου έγινε η επιλογή των παραμέτρων alpha και fit_intercept που προσφέρει το μοντέλο.
Αρχικά με την χρήση της παραμέτρου alpha που σκοπό έχει να μειώνει την διακύμανση των εκτιμήσεων και παίρνει τιμές πάνω από το μηδέν ,χρειάστηκε να γίνουν αρκετές δοκιμές μεταξύ των τιμών μέχρι να βρεθεί η κατάλληλη όπου η ιδανική τιμή σε αυτό το πρόβλημα είναι alpha=0.001 στην συνέχεια η χρήση της παραμέτρου fit_intercept όπως αναφέρθηκε και στο παραπάνω μοντέλο μπορεί να τεθεί σε δύο ορίσματα True ή False που μας δίνουν την δυνατότητα να ορίσουμε αν θέλουμε να υπολογιστεί η παρακολούθηση στο μοντέλο.
Για την διασταύρωση της σωστής εκπαίδευσης του μοντέλου χρησιμοποιείται η μέθοδος cross validation όπου το kfold χωριζει το σύνολο των δεδομένων σε 5 υποσύνολα όπου τα 4 χρησιμοποιούνται ώς δεδομένα εκπαίδευσης και το υπόλοιπο 1 υποσύνολο για δεδομένα ελέγχου,όπως έχει αναφερθεί σε παραπάνω κεφάλαιο η δυνατότητα που μας δίνει το kfold μπορούν να εκπαιδευτούν όλα τα δεδομένα. Εφόσον έχουν γίνει οι απαραίτητες προετοιμασίες είναι έτοιμο το μοντέλο για την εκπαίδευση του.

4.5 Μοντέλο Lasso (L-1)

Χρησιμοποιώντας το μοντέλο Lasso με σκοπό να δίνει την πιο ακριβής πρόβλεψη του.Ωστόσο για να μπορέσει να επιτευχθεί θα χρειαστεί να γίνει εκπαίδευση στο μοντέλο δηλαδή να βρεθούν οι παράμετροι που θα χρησιμοποιηθούν. Όμως κάθε μοντέλο έχει ξεχωριστούς παραμέτρους έτσι και το μοντέλο Lasso. Αρχικά επιλέγεται η παράμετρος alpha και στην συνέχεια ρυθμίζετε αναλόγως με το ποια τιμή του alpha δίνει τα πιο βέλτιστα αποτελέσματα στην δική μας περίπτωση η βέλτιστη τιμή για την παράμετρο είναι alpha=0.001, αποφεύγετε να χρησιμοποιηθεί σε μηδενική μορφή για αριθμητικούς λόγους , η παράμετρος fit_interpect χρησιμοποιείται για να υπολογίζει την παρακολούθηση και τείθεται σε True ή False αναλόγως αν θέλουμε να υπολογιστεί η παρακολούθηση ή όχι.Εφόσον έχουμε τις ιδανικές τιμές στις παραμέτρους αυτό ακόμη που θα χρειαστεί είναι να αξιολογηθεί το μοντέλο πως επιδραεί σε καινούργια δεδομένα έτσι χρησιμοποιούμε το cross validation ,καθώς χρησιμοποιώντας το kfold όπου χωρίζει το σύνολο των δεδομένων σε k υποσύνολα ,όπου τα 4 απο τα 5 συνολικά υποσύνολα είναι τα δεδομένα εκπαίδευσης και το υπόλοιπο 1 είναι τα δεδομένα ελέγχου, συνεπώς αφού ολοκληρωθεί η διαδικασία και μοντέλο δίνει βέλτιστα αποτελέσματα μπορεί να χρησιμποποιηθεί σε καινουργια δεδομένα.

4.6 Μοντέλο Polynomial Regression

Με την χρήση του μοντέλου polynomial regression όπου αναζητάμε την πιο αποτελεσματική πρόβλεψη,όπως και στα προηγούμενα μοντέλα .
Για να μπορέσει να εφαρμοστεί το μοντέλο και να μπορεί να είναι αποδοτικό θα χρειαστεί να χρησιμοποιηθούν οι κατάλληλες τιμές στην παράμετρο, διότι και αυτό το μοντέλο χρησιμοποιεί τι δικές του παραμέτρους .Στο δικό μας μοντέλο χρειάστηκε να εφαρμοστεί μόνο η παράμετρος degree όπου είναι ο βαθμός του polynomial ,χρειάστηκαν αρκετές δοκιμές εώς ότε να βρεθεί ποιός είναι ο κατάλληλος βαθμος Polynomial που προσδίδει βελτίωση στο μοντέλο.Συνεπώς ο βαθμός polynomial που βελτιώνει την επίδοση του μοντέλου είναι με degree=8. Οπότε έχοντας το κατάλληλο βαθμό polynomial γίνονται οι δοκιμές μεταξύ των ιδιοτήτων. Εφόσον έχουν ρυθμιστεί οι τιμές εφαρμόσετε η μέθοδο cross validation χρησιμοποιώντας το kfold όπου χωρίζει το σύνολο των δεδομένων σε 5 υποσύνολα όπου τα 4 είναι τα δεδομένα εκπαίδευσης και το 1 χρησιμοποιείται ώς δεδομένο ελέγχου με αυτή την μέθοδο εκπαιδεύονται όλα τα δεδομένα. Καθώς ολοκληρωθούν όλες οι παραπάνω διαδικασίες μπορεί να ξεκινήσει η εκπαίδευση του μοντέλου.

ΚΕΦΑΛΑΙΟ 5

Η Εφαρμογή των μοντέλων

5.1 Εισαγωγή

Σε αυτό το σημείο θα αναλυθεί η υλοποίηση της εφαρμογής (Rizos & Ρίζος, 2011), που ως στόχο έχει να παρουσιάσει ποιοί παράγοντες επηρεάζουν τον χρόνο αποφοίτησης ενός φοιτητή. Θα αναφερθεί σε ποιό περιβάλλον ανάπτυξης δημιουργήθηκε ο κώδικας και με ποιά γλώσσα προγραμματισμού. Τα βήματα που ακολουθήθηκαν για την επεξεργασία των δεδομένων, τα μοντέλα που εφαρμόστηκαν για την επίλυση του προβλήματος και τα αποτελέσματα που προέκυψαν (Ανδριανάκης & Andrianakis, 2008) από κάθε μοντέλο.

5.1.1 Επεξεργασία δεδομένων

Για την υλοποίηση της εφαρμογής στη συγκεκριμένη διπλωματική εργασία χρησιμοποιήθηκαν τα μοντέλα LinearRegression, Ridge, Lasso, Polynomial και SVR (Τσαρμπόπουλος, 2016). Όπου για την απόδοση των μοντέλων χρειάστηκε να ακολουθηθούν ένα σύνολο βημάτων, ώστε να γίνει η σωστή προετοιμασία της επεξεργασία των δεδομένων (Athanasiadis & Αθανασιάδης, 2015).
Η διαδικασία υλοποίησης της εφαρμογής ξεκινάει με την καταχώρηση των δεδομένων σε ένα αρχείο excel, εκεί γίνονται οι απαραίτητες μετατροπές των δεδομένων . Στη συνέχεια το spyder αντλεί τα δεδομένα από το αρχείο excel. Καθώς επιλέγεται το μοντέλο που θα εφαρμοστεί,και έπειτα ακολουθείται η επεξεργασία τους . Όπως έχει ήδη αναφερθεί παραπάνω ο αριθμός των δεδομένων που χρειάστηκαν για την επεξεργασία είναι 340 ,όπου 272 είναι τα δεδομένα εκπαίδευσης
(train data) που χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος ,όπου η επικύρωση του συστήματος γίνεται (Athanasiadis & Αθανασιάδης, 2015) με τα δεδομένα ελέγχου (test data) όπου το κάθε σετ είναι 68.Επίσης γίνεται κανονικοποίηση των δεδομένων διότι έχει παρατηρηθεί βελτίωση της συμπεριφοράς του μοντέλου με αυτήν την τεχνική (Ανδριανάκης & Andrianakis, 2008) . Για να μπορέσουμε να καταλήξουμε σε κάποιο συμπέρασμα χρειάστηκαν αρκετές δοκιμές μεταξύ των παραμέτρων και των δεδομένων του κάθε μοντέλου. Ώσπου να βρεθούν ποια είναι τα δεδομένα που επηρεάζουν περισσότερο στην απόδοση του κάθε μοντέλου (Athanasiadis & Αθανασιάδης, 2015).

5.2 Περιβάλλον Ανάπτυξης του Κώδικα

Ο κώδικας της διπλωματικής αυτής εργασίας έχει γραφτεί σε γλώσσα προγραμματισμού Python και η εκτέλεση των προγραμμάτων έχει πραγματοποιηθεί (Τσαρμπόπουλος, 2016) στο spyder όπου είναι έκδοση ανοιχτού κώδικα του Ολοκληρωμένου Περιβάλλοντος Υλοποίησης (Integrated Development Environment ή IDE) (Τσαρμπόπουλος, 2016). Παρακάτω παρέχεται ο σύνδεσμος μέσω του οποίου

μπορεί να γίνει λήψη του συγκεκριμένου Περιβάλλοντος Υλοποίησης: (Τσαρμπόπουλος, 2016) https://www.spyder-ide.org/)

5.3 Αποτελέσματα

Χρησιμοποιώντας τα δεδομένα που έχουν περιγραφή και έχοντας κάνει χρήση ως μέτρο αξιολόγησης επίδοσης όπου επιλέγχηθηκε το μέσο απόλυτο σφάλμα (mae), καθώς και με την χρήση των μοντέλων που εφαρμόστηκαν LinearRegression, Ridge, Lasso. Polynomial Regression και SVR. Όπου έχουν αναλυθεί σε παραπάνω κεφάλαιο, αναζητείται ποιά ιδιότητα ή ποιός συνδυασμός ιδιοτήτων δίνει την βέλτιστη επίδοση στα μοντέλα. Τα αποτελέσματα απεικονίζονται στους παρακάτω πίνακες (Fotiou, D. & Fotiou, 2018).

Στον **πίνακα 5** απεικονίζονται τα αποτελέσματα με εφαρμογή του μοντέλου LinearRegression καθώς αναζητείται πια ιδιότητα των δεδομένων μας δίνει την καλύτερη επίδοση στο μοντέλο. Στα αποτελέσματα του **πίνακας 5** καθώς παρατηρείται ο μέσος όρος (μέσος όρος σφάλματος) της κάθε ιδιότητας όπου δείχνει την σημαντικότητα της για την καλύτερη επίδοση του μοντέλου . Τα παρακάτω απολέσματα δείχνουν ότι η ιδιότητα [**'x5 '**] με μέσο όρο σφάλματος=**1.4025933459896394** δίνει καλή επίδοση στο μοντέλο όλες οι υπόλοιπες είναι μη σημαντικές.Στη συνέχεια θα γίνει σύγκριση με την κάθε ιδιότητα ξεχωριστά.

	Μέσος όρος σφάλματος
'x1'	1.526241742920946
'x2'	1.5262935482940563
'x3'	1.5086776031936755
'x4'	1.4941012129983258
'x5'	**1.4025933459896394**
'x6'	1.4989145369640673
'x7'	1.4636185607014496
'x9'	1.5214881807488945
'x10'	1.53229206627136

Πίνακας 5:Όλες οι ιδιότητες με το μοντέλο LinearRegression

Στον **πίνακα 6** γίνεται σύγκριση της ιδιότητας [**'x5'**] όπου μας έδινε μέσος όρος σφάλματος=**1.4025933459896394** έτσι θα εξεταστει με τις υπόλοιπες ώστε να ελεγχεί αν και οι υπόλοιπες ιδιότητες προσδίδουν βελτίωση .Όπως παρατηρείται ο συνδυασμός των ιδιοτήτων[**'x5', 'x2']** βελτιώνει την επίδοση σε το μέσο όρο σφάλματος=**1.3936251904791233** όλοι οι υπόλοιποι συνδυασμοί δεν προσδίδουν βελτίωση στο μοντέλο.Οπότε με τον παραπάνω συνδυασμό θα γίνει σύγκριση αν και κάποια άλλη ιδιότητα επηρεάζει την επίδοση του μοντέλου.

		Μέσος όρος σφάλματος
'x5'	'x1'	1.4146382415362482
'x5'	**'x2'**	**1.3936251904791233**

'x5'	'x3'	1.4021668021426392
'x5'	'x4'	1.3937031804999236
'x5'	'x6'	1.4351750979993052
'x5'	'x7'	1.3947175692496312
'x5'	'x9'	1.405275538376156
'x5'	'x10'	1.4111821990791942

Πίνακας 6: Σύγκριση με τη κάθε ιδιότητα ξεχωριστά με το μοντέλο LinearRegression

Χρησιμοποιώντας τον συνδυασμό **['x5', 'x2']** όπου είχε δώσει βελτίωση στην επίδοση του μοντέλου με μέσο όρο σφάλματος=**1.3936251904791233** γίνεται συγκρίση του συνδυασμού με την κάθε ιδιότητα όπως παρατηρείται ο καινούργιος συνδυασμός [**'x5', 'x2', 'x4**'] δίνει στο μοντέλο μας καλύτερη επίδοση με μέσος όρος σφάλματος=**1.385488670138588** οι υπόλοιποι συνδυασμοί είναι μη σημαντική .Συνεχίζεται η σύγκριση των υπολοίπων ιδιοτήτων με τον καινούργιο συνδυασμό.

			Μέσος όρος σφάλματος
'x5'	'x2'	'x1'	1.4078639962879085
'x5'	'x2'	'x3'	1.389409637517067
'x5'	**'x2'**	**'x4'**	**1.385488670138588**
'x5'	'x2'	'x6'	1.4344728138842657
'x5'	'x2'	'x7'	1.389317756445114
'x5'	'x2'	'x9'	1.39462909514573
'x5'	'x2'	'x10'	1.4006242798114723

Πίνακας 7:Σύγκριση των υπόλοιπων ιδιοτήτων με τον συνδυασμό **['x5', 'x2']** με χρήση του μοντέλου LinearRegression

Στον **πίνακα 8** έχοντας τον συνδυασμό[**'x5', 'x2', 'x4**'] με μέσος όρος σφάλματος= **1.385488670138588** συνεχίζοντας τις συγκρίσεις παρατηρείται ότι εισάγωντας την ιδιότητα [**'x3'**] η επίδοση του μοντέλου βελτιώνεται και έτσι ο συνδυασμός των ιδιοτήτων [**'x5', 'x2', 'x4', 'x3**'] βελτιώνει την επίδοση του μοντέλου απο μέσο όρο σφάλματος =**1.385488670138588** σε **1.3837004938055832**, αντιθέτως όλες οι υπόλοιπες δεν δείχνουν απόδοση στο μοντέλο, θα συνεχιστεί η αναζήτηση σταθερά με τον συνδυαμό [**'x5', 'x2', 'x4', 'x3'**]συγκρίνοντας με τις υπόλοιπες ιδιότητες .

				Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x1'	1.3997568128472129
'x5'	**'x2'**	**'x4'**	**'x3'**	**1.3837004938055832**
'x5'	'x2'	'x4'	'x6'	1.4274506420880895
'x5'	'x2'	'x4'	'x7'	1.38587039803623
'x5'	'x2'	'x4'	'x9'	1.388334248900064
'x5'	'x2'	'x4'	'x10'	1.3926869942828297

Πίνακας 8 : Σύγκριση των ιδιοτήτων με τον συνδυασμό **['x5', 'x2' ,'x4']** με χρήση του μοντέλου LinearRegression

Στον **πίνακα 9** παρουσιάζονται τα αποτελέσματα με σύγκριση τον συνδυασμό **['x5', 'x2' ,'x4','x3']** όπου προστίθεται κάθε ιδιότητα με στόχο την ανάζητηση της βέλτιστης επίδοσης του μοντέλου.Όπως παρατηρείται από τα αποτελέσματα ο συνδυασμός

['x5', 'x2' ,'x4','x3', 'x7'] βελτιώνει την επίδοση του μοντέλου από μέσος όρος σφάλματος=**1.3837004938055832** σε μέσο όρο σφάλματος=**1.383029922090279** δίνει καλύτερη απόδοση στο μοντέλο .Συνεχίζουμε την βέλτιση επίδοση στο μοντέλο με τον συνδυασμό**['x5', 'x2' ,'x4','x3', 'x7'] .**

					Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x1'	1.396410717701627
'x5'	'x2'	'x4'	'x3'	'x6'	1.4276301060617453
'x5'	**'x2'**	**'x4'**	**'x3'**	**'x7'**	**1.383029922090279**
'x5'	'x2'	'x4'	'x3'	'x9'	1.3854258703098885
'x5'	'x2'	'x4'	'x3'	'x10'	1.3894826194583165

Πίνακας 9 : Συγκρίση των ιδιοτήτων με τον συνδυασμό **['x5', 'x2' ,'x4','x3']** με χρήση του μοντέλου LinearRegression

Όπως παρατηρείται στον **πίνακα 10** καμία άλλη ιδιότητα δεν προσδίδει επίδοση στο μοντέλο.Οπότε δεν θα χρειαστεί να γίνουν άλλες συγκρίσεις γι αυτό το μοντέλο , ο συνδυασμός που δίνει την καλύτερη επίδοση στο μοντέλο LinearRegression είναι **['x5', 'x2' ,'x4','x3', 'x7']** , όπου [**'x5']**=απολυτήριο λυκείου , [**'x2']**=καταγωγή, **['x4']**=σειρά προτίμησης, [**'x3**']=υπηκοότητα, **['x7']**=σειρά επιτυχίας.

						Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x7'	'x1'	1.3948769360341584
'x5'	'x2'	'x4'	'x3'	'x7'	'x6'	1.4470587379726256
'x5'	'x2'	'x4'	'x3'	'x7'	'x9'	1.3845738287643752
'x5'	'x2'	'x4'	'x3'	'x7'	'x10'	1.3886650196041113

Πίνακας 10 : Συγκρίση των ιδιοτήτων με τον συνδυασμό **['x5', 'x2' ,'x4','x3', 'x7']** με χρήση του μοντέλου LinearRegression

Στον **Πίνακα 11** παρουσιάζονται τα απολέσματα με την χρήση του μοντέλου Ridge (L2) .Καθώς εξετάζονται όλες οι ιδιότητες ξεχωριστά για να εντοπίσουμε πια ιδιότητα δίνει την καλύτερη επίδοση στο μοντέλο. Όπως παρατηρείται στον πίνακα η ιδιότητα [**'x5']** με μέσος όρος σφάλματος**=1.402599825115875** είναι η ιδιότητα που δίνει επίδοση στο μοντέλο μας. Έπειτα στην συνέχεια θα γίνει σύγκριση με την κάθε ιδιότητα ξεχωριστά.

	Μέσος όρος σφάλματος
'x1'	1.526241509313655
'x2'	1.5262928318039772
'x3'	1.508677893835768
'x4'	1.4941080381442187
'x5'	**1.402599825115875**
'x6'	1.4989071348675171
'x7'	1.4636228330451442
'x9'	1.5214878895258879
'x10'	1.5322915669991803

Πίνακα 11: Σύγκριση της κάθε ιδιότητα ξεχωριστά με την χρήση της μεθόδου Ridge(L2)

Έχοντας παρατηρήση στον **πίνακα 11** ότι η ιδιότητα που δίνει επίδοση στο μοντέλο είναι η **['x5']** με μέσος όρος σφάλματος=**1.402599825115875** οπότε θα γίνει σύγκριση και με τις υπόλοιπες ιδιότητες ώστε για να ελεγχεί αν επηρεάζει κάποια άλλη ιδι-

ότητα την επίδοση του μοντέλου.Όπως παρατηρείται στον **πίνακας 12** ο συνδυασμός που μας δίνει επίδοση στο μοντέλο είναι **['x5','x2']** με μέσο όρο σφάλματος=**1.393633382954695** . Με αυτό τον συνδυασμό θα συνεχίσούμε συγκρίσεις με τις υπόλοιπες ιδιότητες αναζητώντας ποιές ιδιότητες δίνουν καλή επίδοση στο μοντέλο

		Μέσος όρος σφάλματος
'x5'	'x1'	1.414644303692373
'x5'	**'x2'**	**1.393633382954695**
'x5'	'x3'	1.40217257125014
'x5'	'x4'	1.3937103681538916
'x5'	'x6'	1.435162764986171
'x5'	'x7'	1.3947188023035257
'x5'	'x9'	1.4052812490783668
'x5'	'x10'	1.4111885294332978

Πίνακας 12:Σύγκριση της ιδιότητας **['x5']**με τις υπόλοιπες ιδιότητες με χρήση του μοντέλου Ridge (L2)

Στον **πίνακα 13** χρησιμοποιώντας τον συνδυασμό του παραπάνω πίνακα ο οποίος ήταν ο **['x5','x2']** με μέσο όρο σφάλματος=**1.393633382954695** κάνοντας σύγκριση με όλες τις υπόλοιπες ιδιότητες με στόχο ποιές ιδιότητες δίνουν την βέλτιστη επίδοση στο μοντέλο μας. Όπως παρατηρείται στον **πίνακα 16** μόνο ο συνδυασμός **['x5', 'x2', 'x3']** με μέσο όρο σφάλματος=**1.3854956128957934** όλοι οι υπόλοιποι συνδυασμοί είναι μη σημαντικοί για την βελτίωση του μοντέλου.Θα συνεχίσουμε σταθερά με τον συνδυασμό **['x5', 'x2', 'x3']**.

			Μέσος όρος σφάλματος
'x5'	'x2'	'x1'	1.4078719860412037
'x5'	'x2'	'x3'	1.3894159111507958
'x5'	**'x2'**	**'x4'**	**1.3854956128957934**
'x5'	'x2'	'x6'	1.434458323792725
'x5'	'x2'	'x7'	1.3893201076419177
'x5'	'x2'	'x9'	1.3946375159965094
'x5'	'x2'	'x10'	1.4006311606225685

Πίνακας 13: Ο συνδυασμός των ιδιοτήτων **['x5', 'x2']** με χρήση Ridge (L2)

Όπως παρατηρείται στα αποτελέσματα του **πίνακα 14**, ότι μόνο η εισαγωγή της ιδιότητας [**'x3'**] προσδίδει βελτίωση του μοντέλου από μέσο όρο σφάλματος=**1.3854956128957934** σε **1.3837069728518974** ο οπότε έχοντας τον καινούργιο συνδυασμό **['x5, 'x2', 'x4', 'x3']** που προσδίδει επίδοση στο μοντέλο .Οπότε θα συνεχιστούν οι δοκιμές με τον καινούργιο συνδυασμό μεταξύ των ιδιοτήτων αναζητώντας την βέλτιστη επίδοση του μοντέλου.

				Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x1'	1.3997643840756446
'x5'	**'x2'**	**'x4'**	**'x3'**	**1.3837069728518974**
'x5'	'x2'	'x4'	'x6'	1.4274364646638642
'x5'	'x2'	'x4'	'x7'	1.3858722725588692
'x5'	'x2'	'x4'	'x9'	1.3883410610147955

'x5'	'x2'	'x4'	'x10'	1.392695920265953

Πίνακας 14: Ο συνδυασμός των ιδιοτήτων [**'x5', 'x2', 'x4']** με χρήση Ridge(L2)

Στον **πίνακας 15** γίνεται η σύγκριση της κάθε ιδιότητας ξεχωριστά με τον συνδυασμό **['x5', 'x2', 'x4', 'x3']** του παραπάνω πίνακα. Όπως παρατηρείται στον **πίνακα 15** η ιδιότητα [**'x7'**] βελτιώνει την επίδοση του μοντέλου από **1.3837069728518974** σε **1.3830328094956903** οι υπόλοιπες ιδιότητες δεν προσδίδουν βελτίωση .Οπότε συνεχίζουμε σταθερά της συγκρίσεις των ιδιοτήτων με τον συνδυασμό **['x5', 'x2', 'x4', 'x3', 'x7']** αναζητώντας τον συνδυασμό όπου δίνει την βέλτιστη επίδοση στο μοντέλο.

					Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x1'	1.3964167794896851
'x5'	'x2'	'x4'	'x3'	'x6'	1.427616012537912
'x5'	**'x2'**	**'x4'**	**'x3'**	**'x7'**	**1.3830328094956903**
'x5'	'x2'	'x4'	'x3'	'x9'	1.388353649716811
'x5'	'x2'	'x4'	'x3'	'x10'	1.389488879962245

Πίνακας 15: Ο συνδυασμός των ιδιοτήτων [**'x5', 'x2', 'x4', 'x3']** με χρήση Ridge(L2)

Στον **πίνακα 16** απεικονίζονται οι συγκρίσεις των ιδιοτήτων με τον συνδυασμό που μας δίνει την βέλτιστη επίδοση στο μοντέλο [**'x5', 'x2', 'x4', 'x7', 'x3']** . Όπως παρατηρείται στον **πίνακα 16** καμία ιδιότητα δεν βελτιώνει την επίδοση του μοντέλου.Ωστόσο δεν θα χρειαστεί να γίνουν άλλες συγκρίσεις για το μοντέλο Ridge. Οπότε καταλήγουμε για το μοντέλο Ridge ότι ο συνδυασμός που δίνει την βέλτιστη επίδοση στο μοντέλο είναι [**'x5', 'x2', 'x4', 'x7', 'x3']** ,δηλαδή [**'x5']** = απολυτήριο λυκείου**, ['x2']**=καταγωγή**, ['x4']**=σειρά προτίμησης**, ['x7']**=σειρά επιτυχίας**, ['x3']**=υπηκοότητα.

						Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x7'	'x1'	1.3948790117676968
'x5'	'x2'	'x4'	'x3'	'x7'	'x6'	1.4470297248774426
'x5'	'x2'	'x4'	'x3'	'x7'	'x9'	1.3845759569115113
'x5'	'x2'	'x4'	'x3'	'x7'	'x10'	1.3886677524705404

Πίνακας 16: Ο συνδυασμός των ιδιοτήτων [**'x5', 'x2', 'x4', 'x7', 'x3']** με χρήση Ridge(L2)

Στον **πίνακα 17** παρουσιάζονται τα αποτελέσματα που δίνει κάθε ιδιότητα ξεχωριστά με χρήση του μοντέλου Lasso όπως παρατηρείται η ιδιότητα που δίνει επίδοση στο μοντέλο είναι η [**'x5']** με μέσος όρος σφάλματος=**1.403227497518167** .Στην συνέχεια θα ακολουθήσουν συγκρίσεις με την ιδιότητα [**'x5']** με κάθε ιδιότητα ξεχωριστά

	Μέσος όρος σφάλματος
'x1'	1.5261792377760215
'x2'	1.526007534393452
'x3'	1.5087798178874463
'x4'	1.4951428707762056
'x5'	**1.403227497518167**
'x6'	1.4995454774658372
'x7'	1.4641250464430091

'x9'	1.5210389373510877
'x10'	1.532120990869617

Πίνακας 17:Τα αποτελέσματα με την κάθε ιδιότητα ξεχωριστά με χρήση του μοντέλου Lasso(L1)

Ο **πίνακας 18** παρουσιάζει τα αποτελέσματα της ιδιότητα [**'x5'**] με την κάθε ιδιότητα ξεχωριστά .Όπου παρατηρείται ότι ο συνδυασμός [**'x5', 'x2'**] βελτιώνει την επίδοση του μοντέλου απο μέσος όρος σφάλματος=**1.403227497518167** σε μέσος όρος σφάλματος=**1.3945515135187931** .Οπότε έχοντας τον καινούργιο συνδυασμό [**'x5', 'x2'**] συνεχίζουμε την αναζήτηση της βέλτιστης επίδοσης του μοντέλου.

		Μέσος όρος σφάλματος
'x5'	'x1'	1.4149456013329953
'x5'	**'x2'**	**1.3945515135187931**
'x5'	'x3'	1.4027085573385822
'x5'	'x4'	1.3948754458657957
'x5'	'x6'	1.4358898356074383
'x5'	'x7'	1.3951809847431982
'x5'	'x9'	1.405730127950707
'x5'	'x10'	1.4118705397304452

Πίνακας 18: Σύγκριση με τη κάθε ιδιότητα ξεχωριστά με χρήση του μοντέλου Lasso(L1)

Στον **πίνακα 19** παρουσιάζονται τα αποτελέσματα του συνδυασμού **['x5', 'x2']** όπου γίνονται συγκρίσεις των υπόλοιπων ιδιοτήτων, παρατηρείται στα αποτελέσματα ότι μόνο με την εισαγωγή της ιδιότητας [**'x4'**] βελτιώνεται η επίδοση του μοντέλου από μέσος όρος σφάλματος=**1.3945515135187931** σε μέσος όρος σφάλματος=**1.3866003728055092** ενώ οι υπόλοιπες ιδιότητες επιδεινώνουν την βελτίωση του μοντέλου.Επομένως συνεχίζουμε με τον καινούργιο συνδυασμό για την εύρεση της βέλτιστης επίδοσης.

			Μέσος όρος σφάλματος
'x5'	'x2'	'x1'	1.4084071234000555
'x5'	'x2'	'x3'	1.3900796620825793
'x5'	**'x2'**	**'x4'**	**1.3866003728055092**
'x5'	'x2'	'x6'	1.43511995428623
'x5'	'x2'	'x7'	1.3900202626889393
'x5'	'x2'	'x9'	1.3956519112666013
'x5'	'x2'	'x10'	1.4015718767084961

Πίνακας 19: Ο συνδυασμός των ιδιοτήτων **['x5', 'x2']** με χρήση Lasso(L1)

Χρησιμοποιώντας τον βέλτιστο συνδυασμό **['x5', 'x2', 'x4']** του **πίνακα 19 ,**όπου κάνοντας συγκρίσεις με την κάθε ιδιότητα ξεχωριστά τα αποτελέσματα που παρουσιαζονται στον **πίνακα 20** δείχνουν ότι μόνο η εισαγωγή της ιδιότητας [**'x3'**] βελτιώνει την επίδοση του μοντέλου απο μέσο όρο σφάλματος=**1.3866003728055092** σε μέσο όρο σφάλματος=**1.3846259013958024**.Συνεπώς έχοντας τον καινούργιο συνδυασμό συνεχίζουμε.

	Μέσος όρος σφάλματος

'x5'	'x2'	'x4'	'x1'	1.4006081132921007
'x5'	**'x2'**	**'x4'**	**'x3'**	**1.3846259013958024**
'x5'	'x2'	'x4'	'x6'	1.4282766574194876
'x5'	'x2'	'x4'	'x7'	1.3868455258490722
'x5'	'x2'	'x4'	'x9'	1.3892433064266392
'x5'	'x2'	'x4'	'x10'	1.3941333771411117

Πίνακας 20: Ο συνδυασμός των ιδιοτήτων **['x5', 'x2', 'x4']** με χρήση Lasso(L1)

Στον **πίνακα 21** έχοντας κάνει χρήση του βέλτιστου συνδυασμού του παραπάνω πίνακα **['x5', 'x2', 'x4', 'x3']** με συγκρίσει των ιδιοτήτων όπου που προσδίδουν την βέλτιστη επίδοση στο μοντέλο.Όπως παρατηρείται στον **πίνακα 21** η εισαγωγή της ιδιότητας [**'x7'**] βελτιώνει την επίδοση του μοντέλου από μέσο όρο σφάλματος=**1.3846259013958024** σε μέσο όρο σφάλματος= **1.3842424179623847** .Συνεχίζουμε με τον βέλτιστο συνδασμό.

					Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x1'	1.3970263817806254
'x5'	'x2'	'x4'	'x3'	'x6'	1.4284937831218838
'x5'	**'x2'**	**'x4'**	**'x3'**	**'x7'**	**1.3842424179623847**
'x5'	'x2'	'x4'	'x3'	'x9'	1.3863204487160814
'x5'	'x2'	'x4'	'x3'	'x10'	1.390737633630261

Πίνακας 21: Ο συνδυασμός των ιδιοτήτων **['x5', 'x2', 'x4', 'x3']** με χρήση Lasso(L1)

Στον **πίνακα 22** παρατηρείται από τα αποτελέσματα ότι στον συνδυασμό **['x5', 'x2', 'x4', 'x3', 'x7']** ότι η εισαγωγή καμίας ιδιότητας δεν προσδίδει απόδοση στο μοντέλο.Οπότε δεν θα χρειαστεί να γίνουν άλλες συγκρίσεις.Ο συνδυασμός που προσδίδει βελτίωση στην επίδοση του μοντέλου Lasso είναι **['x5', 'x2', 'x4', 'x3', 'x7']** όπου [**'x5'**]=απολυτήριο λυκείου,[**'x2'**]=καταγωγή,[**'x4'**]=σειρά προτίμησης,[**'x3'**]=υπηκοότητα και [**'x7'**]=σειρά επιτυχίας.

						Μέσος όρος σφάλματος
'x5'	'x2'	'x4'	'x3'	'x7'	'x1'	1.3958352253108388
'x5'	'x2'	'x4'	'x3'	'x7'	'x6'	1.4443689981939747
'x5'	'x2'	'x4'	'x3'	'x7'	'x9'	1.38576267526535
'x5'	'x2'	'x4'	'x3'	'x7'	'x10'	1.3898077387792362

Πίνακας 22: Ο συνδυασμός των ιδιοτήτων **['x5', 'x2', 'x4', 'x3', 'x7']** με χρήση Lasso(L1)

Στον **πίνακα 23** παρουσιάζονται τα αποτελέσματα που προέκυψαν με την εφαρμογή του μοντέλου SVR. Συγκρίνοντας τα αποτελέσματα με την κάθε ιδιότητα ξεχωριστά όπως παρατηρείται από τον μέσο όρο (Μέσος όρος σφάλματος) ότι από τις 9 οι 8 ιδιότητες [**x1**,**x2**,**x3**,**x4,x6**,**x7,x9**,**x10]** είναι μη σημαντικές για την επίδοση του μοντέλου ,αντιθέτως η [**x5**] είναι η ιδιότητα που δίνει επίδοση στο μοντέλο. Επομένως η [**x5**]=Απολυτήριο λυκείου είναι η πιο σημαντική ιδιότητα [**x5**] =

1.3471793099649028 έτη .Οπότε θα συνεχιστεί η σύγκριση τις [**x5**] με την κάθε ιδιότητα ξεχωριστά.

	Μέσος όρος σφάλματος
'x1'	1.54058823529417
'x2'	1.542397946066838
'x3'	1.4952941176471093
'x4'	1.5000753738064585
'x5'	**1.3471793099649028**
'x6'	1.5634615463150419
'x7'	1.4302843469449495
'x9'	1.5287902586350004
'x10'	1.5641493360585987

Πίνακας 23: Τα αποτελέσματα με την κάθε ιδιότητα ξεχωριστά με χρήση του μοντέλου SVR

Στον **πίνακα 24** εξετάζεται η ιδιότητα [**x5**] η οποία δίνει τα καλύτερα αποτελέσματα με την κάθε ιδιότητα ξεχωριστά για να ελεγχθεί αν και οι υπόλοιπες ιδιότητες βελτιώνουν την επίδοση του μοντέλου όπως παρατηρείται στον **πίνακα 24** ότι ο συνδυασμός [**x5, x3**] μεταβάλλει την επίδοση του μοντέλου από **1.3471793099649028**σε **1.3388003583045438**.Οι υπόλοιποι συνδυασμοί επιδεινώνουν την επίδοση του μοντέλου. Εν συνεχεία ακολουθείται (5) σταθερά με τον συνδυασμό [**x5**,**x3**] σύγκριση και με τις υπόλοιπες ιδιότητές.

		Μέσος όρος σφάλματος
'x5'	'x1'	1.4031607997907627
'x5'	'x2'	1.3602300050140852
'x5'	**'x3'**	**1.3388003583045438**
'x5'	'x4'	1.3661254571815604
'x5'	'x6'	1.427367373477166
'x5'	'x7'	1.3856230358267223
'x5'	'x9'	1.3873135565841206

'x5'	'x10'	1.3947048531546171

Πίνακας 24: Τα αποτελέσματα με σύγκριση με την κάθε ιδιότητα ξεχωριστά με το μοντέλο SVR

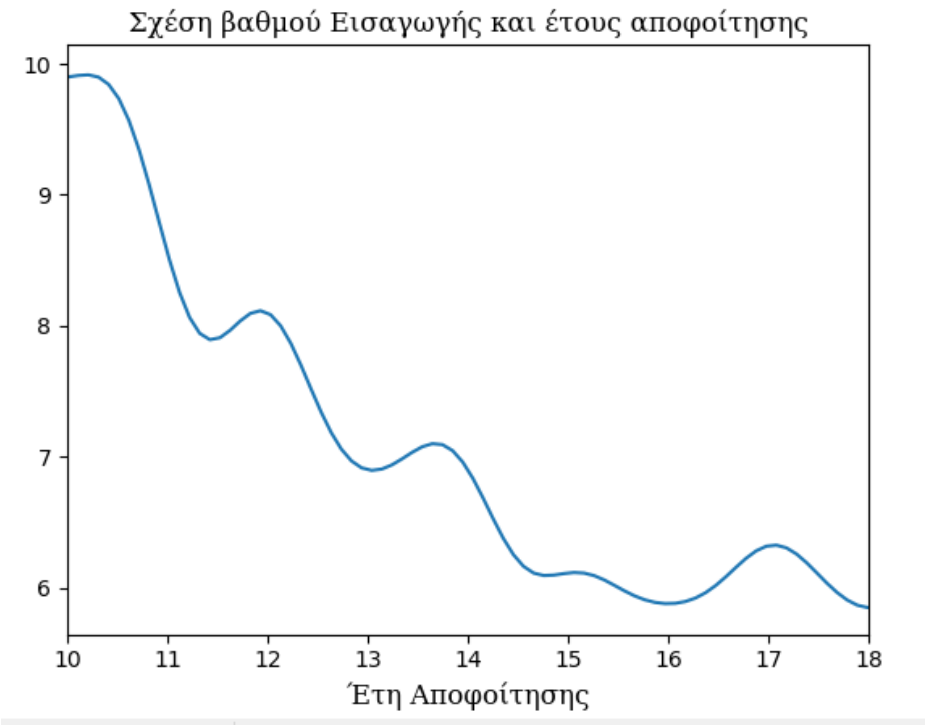
Έχοντας πάρει τον συνδυασμό [**x5, x3]** που έχει αποφέρει τα καλύτερα αποτελέσματα στον **πίνακα 24** , ελέγχεται ο συνδυασμός και με τις υπόλοιπες ιδιότητες όπως παρατηρείται στον **πίνακα 25** καμία ιδιότητα δεν προσδίσει επίδοση στο μοντέλο αντιθέτως επιδεινώνει την επίδοση .

			Μέσος όρος σφάλματος
'x5'	'x3'	'x1'	1.4113064692175576
'x5'	'x3'	'x2'	1.3491096995598337
'x5'	'x3'	'x4'	1.367433117928236
'x5'	'x3'	'x6'	1.4249405479142756
'x5'	'x3'	'x7'	1.4009740796832504
'x5'	'x3'	'x9'	1.3783350684343068
'x5'	'x3'	'x10'	1.402971317721572

Πίνακας 25:Τα αποτελέσματα με σύγκριση τον καλύτερο προηγούμενο συνδυασμό με την κάθε ιδιότητα ξεχωριστά με το μοντέλο SVR

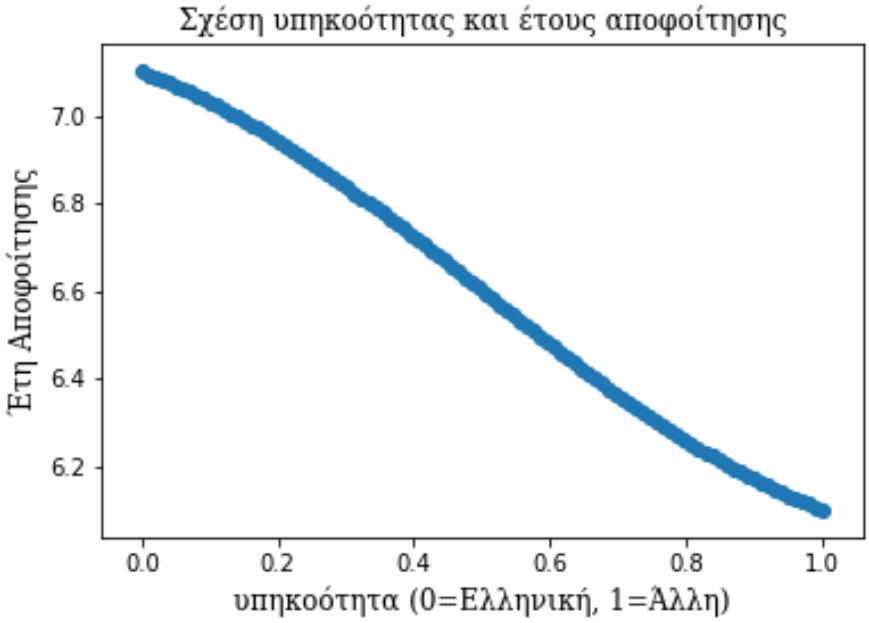
Οπότε αναζητώντας τον αρχικό μας στόχο ο οποίος ήταν ποιά ιδιότητα επηρεάζει το χρόνο αποφοίτησης ενός φοιτητή.Όπου έχοντας εφαρμόση το μοντέλο SVR με χρήση του μέσου απόλυτου σφάλματος(mae) όπως έχουμε είδη αναφέρει παραπάνω . Στην αναζήτηση για την βέλτιστη επίδοση του μοντέλου παρατηρήθηκε ότι ο καλύτερος συνδυασμός είναι ο [**x5,x3]** όλες οι υπόλοιπες ιδιότητες επιδεινώνουν την επίδοση του μοντέλου. Παρακάτω απεικονίζονται διαγραμματικά.

Στο **Διάγραμμα διασποράς 1** απεικονίζεται η σχέση του βαθμού εισαγωγής και με τα έτη αποφοίτησης και όπως σχηματίζεται στο διαγράμμα όσο μικρότερο βαθμό απολυτηρίου έχει ένας φοιτήτης τόσο αυξάνεται ο χρόνος φοίτησης ενώ όσο μεγαλύτερη βαθμολογία έχει τόσο τείνει να είναι πιό συνεπής στην διάρκεια των σπουδών .



Διάγραμμα Διασποράς 1: Σχέση βαθμού εισαγωγής και έτους αποφοίτησης

Όπως παρατηρείται και στο **Διάγραμμα διασποράς 2** η σχέση του χρόνου αποφοίτησης και της υπηκοοότητας δεν επηρεάζεται .Η ευθεία γραμμή του διαγραμμάτος δείχνει ότι ο χρόνος καθυστέρησης αποφοίτησης τείνει να μειώνεται όταν ο φοιτητής έχει διαφορετική υπηκοότητα από την ελληνική.



Διάγραμμα Διασποράς 2: Σχέση υπηκοότητας και έτους αποφοίτησης

5.4 Σύνοψη Αποτελεσμάτων

Σε αυτό το κεφάλαιο εξετάστηκαν τα μοντέλα (Σουσούνης & Sousounis, 2011) LinearRegression, Ridge(L2), Lasso(L1), Polynomial Regression και SVR αναζητώντας ποιές ιδιότητες των δεδομένων δίνουν βέλτιστη επίδοση στο μοντέλο. Επειδή οι συγκρίσεις των ιδιοτήτων των δεδομένων όπου έγιναν αρκετές δοκιμές μέχρι να εξαχθεί κάποιο ασφάλες συμπέρασμα για το ποιές ιδιότητές είναι αυτές που επηρεάζουν στον χρόνο αποφοίτησης ενός φοιτητή και με πιο μοντέλο. Ο έλεγχος που πραγματοποιήθηκε (Σουσούνης & Sousounis, 2011) ώστε να βρεθεί η ιδανικότερη επίδοση του κάθε μοντέλου απαιτήθηκε να γίνουν αρκετές δοκιμές ανάμεσα στις παραμέτρους που διαθέτει το κάθε μοντέλο.

Έπειτα μέσα από τους πίνακες απεικονίζονται τα αποτελέσματα για το κάθε μοντέλο ξεχωριστά (Athanasiadis & Αθανασιάδης, 2015).

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα

Μελετώντας θεωρητικά και πρακτικά από την σκοπιά των γραμμικών μοντέλων το πρόβλημα της επιλογής του μοντέλου και ιδιοτήτων που πραγματευτήκαμε στην παρούσα διπλωματική εργασία, εξάγωντας κάποια χρήσιμα συμπεράσματα όπου μελετάμε την επίδραση (Fotiou, D. & Fotiou, 2018) των ιδιοτήτων ενός δεδομένου (φοιτητή) στο πόσο επηρεάζει τον χρόνο αποφοίτησης ενός φοιτητή.

Η έρευνα ολοκληρώθκε με την εφαρμογή των 5 γραμμικών μοντέλων SVR, LinearRegression, Ridge(L2), Lasso(L1) και Polynomial Regression καθώς το καθε μοντέλο επεξεργάστηκε τις ιδιότητες των δεδομένων,έτσι σχετικά με τα αποτελέσματα (Κάντα & Kanta, 2013) οι δοκιμές και με τις συγκρίσεις των ιδιοτήτων που πραγματοποιήθηκαν (Σουσούνης & Sousounis, 2011) απο το συνόλο των δεδομένων που μελετήθηκε παρατηρήθηκε ότι συγκριτηκά με τα αποτελέσματα που μας έδωσε κάθε μοντέλο, την καλύτερη επίδοση την είχε το SVR ενώ η απόδοση των υπολοίπων διαφοροποιούνται ελάχιστα (Fotiou, D. & Fotiou, 2018) μεταξύ τους. Συμπερασματικά, δεδομένης της σημασίας της ακρίβειας στην πρόβλεψη που πρέπει να χαρακτηρίζει ένα μοντέλο (Κάντα & Kanta, 2013).

Όσον αφορά τα αποτελέσματα να θυμισουμε ότι οι αρχικές ιδιότητες που συλλέχθηκαν ήταν 10 αφαιρέθηκε η ['x8']=βαθμος πτυχιου διότι σε ένα νεοεισερχόμενο φοιτητή δεν διαθέτουμε τον βαθμό πτυχίου οπότε χρησιμοποιήθηκαν 9 ιδιότητες όπου το μοντέλο SVR αποφάσισε ότι μόνο 2 ιδιότητες επηράζουν τον χρόνο αποφοίτησης ενός φοιτητή οι οποίες είναι οι εξής: ['x5']=απολυτήριο λυκείου, ['x3']=υπηκοότητα, έχοντας αναφερθεί στους παράγοντες που επηρεάζουν στην ακρίβεια της πρόβλεψης (Athanasiadis & Αθανασιάδης, 2015). Επίσης θα χρειαστεί να αναφερθεί και ότι με την χρήση των γραμμικών μοντέλων μπορούν να επιλυθούν διάφορα είδη προβλημάτων όπως κατηγοριοποίησης αλλά και παλινδρόμησης όπως εφαρμόστηκε στην παρούσα έρευνα.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Athanasiadis, S., & Αθανασιάδης, Σ. (2015). *Βραχυπρόθεσμη Πρόβλεψη Ηλεκτρικού Φορτίου*.

Downey, D. B. (1995). When bigger is not better: Family size, parental resources, and children's educational performance. *American Sociological Review*, 746–761.

Fagan, P. F., & Churchill, A. (2012). The effects of divorce on children. *Marri Research*, 1–48.

Fetzer, J. H. (1990). What is Artificial Intelligence? In *Artificial Intelligence: Its Scope and Limits* (pp. 3–27). Springer.

Fotiou, D., & Fotiou, D. (2018). *Μπεϋζιανή επιλογή μεταβλητών με χρήση g prior στα κανονικά γραμμικά μοντέλα.*

Halpern, D. F., & Murphy, S. E. (2013). *From work-family balance to work-family interaction: Changing the metaphor*. Routledge.

Kalogirou, S. A. (2001). Artificial neural networks in renewable energy systems applications: a review. *Renewable and Sustainable Energy Reviews*, *5*(4), 373–401.

Krein, S. F., & Beller, A. H. (1988). Educational attainment of children from single-parent families: Differences by exposure, gender, and race. *Demography*, *25*(2), 221–234.

Laosa, L. M. (1982). School, occupation, culture, and family: The impact of parental schooling on the parent–child relationship. *Journal of Educational Psychology*, *74*(6), 791.

Middleton, R., & Grigg, C. M. (1959). Urban-Rural Differences in Aspira-tions. *Rural Sociology*, *24*, 247–254.

Μοναχόπουλος. (2016). *Μελέτη και υλοποίηση Deep Learning τεχνικών στον τομέα της υπολογιστικής όρασης (Doctoral dissertation).* (PhD Thesis).

Negnevitsky, M., & Intelligence, A. (2005). A guide to intelligent systems. *Artificial Intelligence, 2nd Edition, Pearson Education*.

Newell, A. (1982). *Intellectual issues in the history of artificial intelligence*.

Rizos, G., & Ρίζος, Γ. (2011). *Ευφυές σύστημα για το παιχνίδι σκάκι.*

Russell, S., & Norvig, P. (1995). Artificial Intelligence Prentice Hall. *Upper Saddle River, NJ*.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *International Conference on Computational Learning Theory*, 416–426. Springer.

Τσαρμπόπουλος, Δ. (2016). *Σχεδιασμός και υλοποίηση μεθοδολογίας βραχυπρόθεσμης πρόβλεψης τιμών μετοχών του ελληνικού χρηματιστηρίου με συνδυασμό εξελικτικών αλγορίθμων, μηχανών διανυσμάτων υποστήριξης και τεχνικής κυλιόμενου παραθύρου* (PhD Thesis).

Watson, M. B., Creed, P. A., & Patton, W. (2003). Career decisional states of Australian and South African high school students. *International Journal for Educational and Vocational Guidance*, *3*(1), 3–19.

Youn, E. S. (2002). *Feature selection in support vector machines* (PhD Thesis). University of Florida.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*(1), 51.

Ανδριανάκης, Ε. Α., & Andrianakis, E. (2008). *Εντοπισμός ατελειών σε επίπεδη πλάκα ομογενούς και σύνθετου υλικού με χρήση νευρωνικών δικτύων και μεθόδου πεπερασμένων στοιχείων.* (B.S. thesis).

Αρσενίου, Ι. (2015). *Εκπαίδευση, φύλο και τρόπος ένταξης στην αγορά εργασίας*.

Ασπιρτάκης, Γ., Κουμπούλης, Σ., & Πετράκη, Ε. (2016). *Απλή γραμμική παλινδρόμηση και εφαρμογές της.*

Βιαννιτάκη, Βασιλική. (2015). *Τεχνητή νοημοσύνη, ευφυείς πράκτορες και εφαρμογές στην πληροφορική υγείας.*

Γιαννούλη, Δ. Π., & Giannouli, D. P. (2014). *Εφαρμογές των Μηχανών Διανυσματικής Υποστήριξης σε Προβλήματα Ταξινόμησης και Παλινδρόμηση* (Master's Thesis).

Διαμαντάρας, Κ. (2007). *Τεχνητά Νευρωνικά Δίκτυα*.

Ζαγκλαρά, Π. (n.d.). Διατριβή: Κοινωνιοψυχολογικές διαστάσεις διαμόρφωσης εκπαιδευτικής πολιτικής: η πρόσβαση στην τριτοβάθμια εκπαίδευση - Κωδικός: 35409. Retrieved May 5, 2019, from http://thesis.ekt.gr/thesisBookReader/id/35409#page/1/mode/2up

Κάντα, Σ., & Kanta, S. N. (2013). *Μέθοδοι και κριτήρια επιλογής μοντέλου με ποινή (Bachelor's thesis).* (B.S. thesis).

Κουδούνης, Μιχάλης. (2015). *Το εκπαιδευτικό σύστημα της Ελλάδας και η εισαγωγή μαθητών στην τριτοβάθμια εκπαίδευση κατά τα έτη 2007 & 2008. 2007 & 2008*.

Μαραγκάκης, Γ., & Maragkakis, G. (2013). *Αλγόριθμοι μάθησης μηχανής για αποδοτική διαχείριση συστημάτων ηλεκτρικής ενέργειας*.

Μαύρος, Κ. (1998). *Κοινωνικό προφίλ των σπουδαστών ΤΕΙ: προέλευση, κοινωνικοοικονομικά χαρακτηριστικά, προσδοκίες, στάσεις, απόψεις: η περίπτωση του ΤΕΙ Αθήνας. Diss. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ). Σχολή Φιλοσοφική. Τμήμα Φιλοσοφίας, Παιδαγωγικής και Ψυχολογίας. Τομέας Ψυχολογίας, 1998.* (PhD Thesis). Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ). Σχολή Φιλοσοφική. Τμήμα Φιλοσοφίας, Παιδαγωγικής και Ψυχολογίας. Τομέας Ψυχολογίας,.

Μιχαλοπούλου, Ε., & Michalopoulou, E. (2016). *Μετα-ανάλυση με μηχανές διανυσμάτων υποστήριξης σε γονιδιακά δεδομένα*.

Ντούκα, Α., & Φραγκίσκου, Μ. (2016). *ΕΡΕΥΝΑ ΠΡΩΤΟΕΤΩΝ ΦΟΙΤΗΤΩΝ ΓΙΑ ΤΟ ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2010 - 2011*.

Παπαδάκης, Σ. (2016). TEI of Crete - eClass | ΔΕ500Θ - Τεχνητή Νοημοσύνη Στις... | Έγγραφα. Retrieved May 4, 2019, from https://eclass.teicrete.gr/modules/document/?course=DSA132

Παπάνης, Ε., & Βίκη, Α. (2007a). Η επίδραση του οικογενειακού περιβάλλοντος και του φύλου των μαθητών στη διαμόρφωση επαγγελματικών τύπων κατά την εφηβεία. Εμπειρική πανελλαδική έρευν. Retrieved May 5, 2019, from Ελληνική Κοινωνική Έρευνα -Greek Social Research website: http://epapanis.blogspot.com/2007/09/blog-post_4805.html

Παπάνης, Ε., & Βίκη, Α. (2007b). Κοινωνικοοικονομικοί παράγοντες και κίνητρα στην επιλογή επαγγέλματος. Retrieved May 5, 2019, from Ελληνική Κοινωνική Έρευνα -Greek Social Research website: http://epapanis.blogspot.com/2007/09/blog-post_5798.html

Σουσούνης, Μ.-Χ., & Sousounis, M.-C. D. (2011). *Συμβολή στον έλεγχο ανεμογεννητριών μονίμων μαγνητών με τεχνητά νευρωνικά δίκτυα* (B.S. thesis).

Τσιλιμίγκρα, Β., Stati, Ε., Στάτη, Ε., & Tsilimigkra, V. (2017). *Επιλέγοντας σπουδές στην κοινωνική εργασία σε περίοδο κοινωνικο-οικονομικής κρίσης: απόψεις φοιτητών/τριών του Τμήματος Κοινωνικής Διοίκησης και Πολιτικής Επιστήμης του Δημοκριτείου Πανεπιστημίου Θράκης*.

Χαλικιά, Φ.-Ε., & Κιτσαρά, Σ.-Α. (2016). *Παράγοντες που επηρεάζουν τις επιλογές σπουδών/ επαγγελματικές επιλογές των νέων: έρευνα σε φοιτητές του ΤΕΙ Δυτικής Ελλάδας*.

ΠΑΡΑΡΤΗΜΑ Α

Δηλώσεις μεταβλητών
```
a=0;
b=1;
```

Η συνάρτηση κάνει μετατροπή των δεδομένων σε κανονικοποιημένα δεδομένα.

```
def Transform(X,a,b):
    nx=a+(b-a)*(X-X.min(axis=0))/(X.max(axis=0)-X.min(axis=0))
    return nx
```

Η χρήση αυτού του κώδικα είναι που αντλεί τα δεδομένα από το excel

```
df=pd.read_excel("C:\\Users\\Dimitra-
Spuridoula\\Desktop\\ptyxiakh\\new_file_for_data\\data_corrected_dimitra.xlsx"),
sheename='data',header=1,sep=' ')
```

```
df=df[ ['x1', ' x2', 'x3', 'x4','x5','x6','x7', 'x9', 'x10','y1'] ]
```

Μετοναμαζετε η στήλη y1 σε y

```
df.rename(columns={'y1': 'y'}, inplace=True)
```

Ελέγχει την μετονομασία

```
print (df.columns)
```

Παίρνει απο την στήλη τους απόφοιτους απο 4 έως 10 χρόνια

```
df = df.query('y <=10 and y >=4 ')
```

Μπαίνουν στην λίστα οι ιδιότητες όπου γίνονται οι δοκιμές
```
lst=['x5']
```
Επιλέγει μονο τις τιμές X που υπάρχουν στη λίστα

```
X = df[lst].values

Παίρνει από το ευρετήριο μια ετικέτα μόνο
Y = df.loc[:,'y'].values

Καλείτε η συνάρτηση κανονικοποίησης
x=Transform(X,a,b)
Θέτουμε το Y σε y
y=Y

nfolds=5

kf = mykfold(x, nfolds)



cvscore=[]



for i in range(nfolds):

    x_trn, y_trn, x_tst, y_tst = kf.get_data_fold(i, x, y)




  Εδώ γίνεται χρήση των μοντέλων SVR,LinearRegression,Ridge και Lasso



    model=SVR(C=2.0,kernel='rbf',gamma=10.0)
    model=LinearRegression(fit_intercept= True)
    model=Ridge(alpha=0.001,fit_intercept=True)
    model=Lasso(alpha=0.001,fit_intercept=True)

    model.fit(x_trn,y_trn)



   Για όλα τα υπόλοιπα μοντέλα
    yhat=model.predict(x_tst)
```

Υπολογίζεται το μέσο απόλυτο σφάλμα (mae), μέσο τετραγωνικό σφάλμα(mse),τετραγωνικό μεέσο σφάλμα ρίζας(rmse)

```
mae=mean_absolute_error(y_tst,yhat)
rmse = sqrt(mse)
cvscore.append(mae)
```

Το cross-validation score μας δείχνει την βαθμολογία των πτυχών .

```
print ( 'crosval score: ', cvscore )
```

```
e=np.array(cvscore)
```

Εμφανίζονται τα αποτελέσματα.

```
print('Error(Years): ','min: ',e.min(),'max: ',e.max(),'μέσος όρος σφάλματος: ',e.μέσος όρος σφάλματος(),u"\u00B1",0.5*e.std())
```

```
print(lst,'-->',e.μέσος όρος σφάλματος())
```

Αυτή η κλάση προσεγγίζει το στοιχείο που δεν υπάρχει ή δημιουργήθηκε εκείνη την στιγμή

```
class Tree(dict):

    def __missing__(self, key):
```

```
        value = self[key] = type(self)()
        return value
```

Σε αυτή την κλάση ακολουθείται η διαδικασία μορφοποίησης του kfold

```
class mykfold:
    def __init__(self,x,nfolds=5):
        self.__nfolds=nfolds
        self.__nData=x.shape[0]
```

Εδώ η πράξη γίνεται για να βγάλει σε πόσες πτυχές πόσα δεδομένα πάνε

```
        self.__nrest=(self.__nData)%self.__nfolds

        self.__nDataTst=int(self.__nData/self.__nfolds)
        self.__nDataTrn=self.__nData-self.__nDataTst

        self.__index_original=[i for i in range(self.__nData)]

        self.__index_fold=Tree()

        self.__index_fold['nfolds']=self.__nfolds
```

```
        self.__index_fold[0]['trn']=self.__index_original[:(self.__nDataTrn-self.__nrest)]
        self.__index_fold[0]['tst']=self.__index_original[(self.__nDataTrn-self.__nrest):]

        self.__make_folds()
        self.__validate()
```

Η συγκεκριμένη συνάρτηση κάνει περιστροφή στη λίστα

```
    def __rotate_list(self,L,n):
        return L[n:] + L[:n]
```

Χρησιμοποιείται για την δημιουργία των πτυχών στα δεδομένα

```
    def __make_folds(self):
        L=self.__index_original
        for i in range(1,self.__nfolds):
            L=self.__rotate_list(L,self.__nDataTst)
            self.__index_fold[i]['trn']=L[:self.__nDataTrn]
            self.__index_fold[i]['tst']=L[self.__nDataTrn:]
```

Παίρνει τα δεδομένα τα οποία έχουν χωριστεί σε train set και test set

```
    def get_data_fold(self,foldid,x,y):
        itrn=self.__index_fold[foldid]['trn']
        itst=self.__index_fold[foldid]['tst']
        x_trn,x_tst=x[itrn],x[itst]
```

```
        y_trn,y_tst=y[itrn],y[itst]
        return x_trn,y_trn,x_tst,y_tst
```

Αυτή η συνάρτηση βάζει τους δείκτες στις μεταβλητές

```
    def get_index_fold(self,foldid):
        itrn=self.__index_fold[foldid]['trn']
        itst=self.__index_fold[foldid]['tst']
        return itrn,itst
```

```
    def __validate(self):
        sum=0
        for i in range(self.__index_fold['nfolds']):
            sum+=len(self.__index_fold[i]['tst'])
        assert(sum==self.__nData)
```

Μετράει πόσα δεδομένα έχει κάθε fold.

```
    def fold_lengths(self):
        sum=0
        fl=[]
        for i in range(self.__index_fold['nfolds']):
          l =len(self.__index_fold[i]['tst'])
```

```
        sum+=l
        fl.append(i)
    return fl
```