ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΣΤΑ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ

Δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων και εφαρμογές στην Μηχανική Μάθηση

Μαρία-Αγγελική Πολλάλη
ΑΜ: 09320033

Επιβλέπων Καθηγητής
Χρήστος Κουκουβίνος
Καθηγητής του Εθνικού Μετσόβιου Πολυτεχνείου

Αθήνα, 2022

National Technical University of Athens

School of Applied Mathematical and Physical Sciences
MSc Mathematical Modeling in Modern Technologies and Financial Engineering

Latin Hypercube Sampling and applications in Machine Learning

Maria-Angeliki Pollali
AM: 09320033

Supervising Professor
Dr Christos Koukouvinos
Professor of National Technical University of Athens

Athens, 2022

Περίληψη

Αναμφίβολα ζούμε σε μια εποχή όπου τα δεδομένα κατακλύζουν την ζωή μας και μερικές φορές είναι ακατόρθωτο να τα αποθηκεύσουμε στην μνήμη του υπολογιστή μας (πρόβλημα των Μεγάλων Δεδομένων). Μια διέξοδο για την σύγχρονη Μηχανική Μάθηση και το πρόβλημα των Μεγάλων Δεδομένων αποτελεί η δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων. Συγκεκριμένα, φαίνεται ότι αν δειγματοληπτίσουμε από τα αρχικά δεδομένα μας με την μέθοδο των λατινικών υπερκύβων και στη συνέχεια εφαρμόσουμε κάποιο μοντέλο πρόβλεψης ή ταξινόμησης η ακρίβεια των αποτελεσμάτων μας θα είναι σχεδόν ίδια με το να χρησιμοποιούσαμε όλα τα αρχικά δεδομένα.

Η παρούσα διπλωματική εργασία δομείται ως ακολούθως: Στο πρώτο εισαγωγικό κεφάλαιο περιγράφεται το πρόβλημα των Μεγάλων Δεδομένων και κάποιες βασικές αρχές δειγματοληψίας. Στο δεύτερο κεφάλαιο θεμελιώνεται η δειγματοληψία με την μέθοδο των λατινικών υπερκύβων. Στο τρίτο κεφάλαιο αναλύεται η δομή και η λειτουργία της δειγματοληψίας με την μέθοδο των λατινικών υπερκύβων καθώς και των αλγορίθμων ταξινόμησης που χρησιμοποιήθηκαν στο πειραματικό στάδιο. Στο τέταρτο κεφάλαιο αναφέρεται αναλυτικά όλο το πειραματικό στάδιο και σχολιάζονται τα αποτελέσματα εφαρμογής της μεθόδου πάνω σε πραγματικά δεδομένα.

Λέξεις Κλειδιά— Λατινικοί Υπερκύβοι, Δειγματοληψία, Μηχανική Μάθηση, Τυχαία Δάση, Κ-Πλησιέστεροι Γείτονες

Abstract

Indisputably, we live in an era where data overwhelms our lives and sometimes it is impossible to store it in our computer memory (the Big Data problem). One way out for modern Machine Learning and the Big Data problem is the Latin Hypercube sampling method. It appears that if we sample from our original data using the Latin Hypercube method and then apply a prediction or classification model the accuracy of our results will be almost the same as if we used all the original data.

This thesis is structured as follows: The first introductory chapter describes the Big Data problem and some basic sampling principles. The second chapter establishes the sampling with the Latin hypercube method. The third chapter discusses the structure and operation of Latin hypercube sampling and the classification algorithms used in the experimental stage. In chapter four the whole experimental stage is reported in detail and results of applying the method on real data are commented.

Keywords— Latin Hypercube, Sampling, Machine Leaning, Random Forest, K-Nearest Neighbor

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κύριο Χρήστο Κουκουβίνο για την ανάθεση της παρούσας διπλωματικής εργασίας καθώς επίσης και για την πολύτιμη καθοδήγηση του διάρκεια των μεταπτυχιακών σπουδών μου.

Τις θερμές μου ευχαριστίες θα ήθελα να εκφράσω στον κύριο Σπύρο Παρασκευά για τη συνεχή υποστήριξη κατά την διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Τέλος θα ήθελα να εκφράσω τις βαθύτατες ευχαριστίες μου στην οικογένεια μου και ιδιαίτερα στους γονείς μου Γεώργιο Πολλάλη και Στυλιανή Νταντούτη για την υποστήριξη τους καθ' όλη τη διάρκεια των σπουδών μου και την αμέριστη συμπαράσταση που μου πρόσφεραν όλα αυτά τα χρόνια.

Περιεχόμενα

1 Εισαγωγή **21**

1.1 Τι είναι τα Μεγάλα Δεδομένα και το πρόβλημα των Μεγάλων Δεδομένων . . . 21

1.2 Τι είναι η δειγματοληψία και ποίοι είναι οι λόγοι που κάνουμε δειγματοληψία . 25

1.3 Μέθοδοι δειγματοληψίας 28

1.4 Δειγματοληψία με την μέθοδο των λατινικών υπερκύβων 32

2 Δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων **35**

2.1 Βασικοί ορισμοί και έννοιες 35

2.2 Βασικά χαρακτηριστικα της μεθόδου των Λατινικών Υπερκύβων 41

2.3 Σύγκριση της LHS με άλλες μεθόδους δειγματοληψίας 44

2.4 Ορθογώνιοι Σχηματισμοί 49

3 Λατινικοί Υπερκύβοι και το πρόβλημα των Μεγάλων Δεδομένων **53**

3.1 Η βιβλιοθήκη cLHS 53

3.2 Ταξινομητές και Σύνολα Δεδομένων 56

4 Πειραματικό Στάδιο, Συμπεράσματα και Μελλοντική Έρευνα **61**

4.1 Πειρματικό Στάδιο 61

4.2 Συμπεράσματα και Μελλοντική Έρευνα 73

A Παράρτημα με κώδικα αναπαραγωγής αποτελεσμάτων **77**

Ακρώνυμα

cLHS: Conditioned Latin Hypercube Sampling - Δειγματοληψία Λατινικού Υπερκύβου Υπό Όρους

K-NN: K-Nearest Neighbors - Κ-Πλησιέστεροι Γείτονες

LHD: Latin Hypeprcube Desing - Σχεδιασμός Λατινικού Υπερκύβου

LHS: Latin Hypercube Sampling - Δειγματοληψία Λατινικού Υπερκύβου

ML: Machine Learning - Μηχανική Μάθηση

OLH: Orthogonal Latin Hypercube - Οθρογώνιος Λατινικός Υπερκύβος

OLHD: Orthogonal Latin Hypercube Design - Ορθογώνιος Σχεδιασμός Λατινικού Υπερκύβου

RF: Ranfom Forest - Τυχαία Δάσοι

SOLH: Symmetric Orthogonal Latin Hypercube - Συμετρικός Ορθογώνιος Λατινικός Υπερκύβος

SRS: Stratified Random Sampling - Στρωματοποιημένη Τυχαία Δειγματοληψία

STS: Simple Random Sampling - Απλή Τυχαία Δειγματοληψία

Λίστα Σχημάτων

1.1 Κανονική Κατανομή 26
1.2 Απλή Τυχαία Δειγματοληψία 28
1.3 Συστηματική Δειγματοληψία 29
1.4 Διαστρωματική Δειγματοληψία 29
1.5 Δειγματοληψία σε Ομάδες 30
1.6 Βολική Δειγματοληψία 30
1.7 Δειγματοληψία με Ποσόστωση 31
1.8 Δειγματοληψία με Κρίση 31
1.9 Δειγματοληψία Χιονοστιβάδας 31

2.1 Σχεδιασμός με πολύ φτωχές ιδιότητες πλήρωσης χώρου 43
2.2 Σχεδιασμός με καλές ιδιότητες πλήρωσης χώρου 43
2.3 Τυχαιοποιημένος σχεδιασμός 44

3.1 Δένδρα αποφάσεων 56
3.2 Κ-Πλησιέστεροι Γείτονες 57
3.3 Iris Σύνολο Δεδομένων 58
3.4 Diabetes Σύνολο Δεδομένων 58
3.5 Wine Σύνολο Δεδομένων 59
3.6 Wifi Σύνολο Δεδομένων 59

4.1 Iris Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων 61
4.2 Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%) . . . 62
4.3 Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%) . . . 62
4.4 Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%) . . . 62
4.5 Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%) 63
4.6 Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%) . . . 63
4.7 Ακρίβεια Iris Σύνολο Δεδομένων 64
4.8 Wine Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων 64
4.9 Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων(80%) . . . 64
4.10 Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%) . . 65
4.11 Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%) . . 65
4.12 Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%) . . . 65
4.13 Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%) . . 66

4.14 Ακρίβεια Wine Σύνολο Δεδομένων 66
4.15 Diabetes Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων . . 67
4.16 Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%) 67
4.17 Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%) 67
4.18 Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβών (50%) 68
4.19 Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%) 68
4.20 Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%) 68
4.21 Ακρίβεια Diabetes Σύνολο Δεδομένων 69
4.22 Wifi Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων 69
4.23 Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων(80%) . . . 69
4.24 Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%) . . 70
4.25 Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%) . . 70
4.26 Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%) . . . 70
4.27 Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%) . . 71
4.28 Ακρίβεια Wifi Σύνολο Δεδομένων 71

Λίστα Πινάκων

4.1 Iris Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες) 72
4.2 Wine Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες) 72
4.3 Diabetes Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες) 72
4.4 Wifi Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες) 73

Κεφάλαιο 1

Εισαγωγή

1.1 Τι είναι τα Μεγάλα Δεδομένα και το πρόβλημα των Μεγάλων Δεδομένων

Τα τελευταία χρόνια υπάρχει ένα αυξανόμενο ενδιαφέρον για την Διαχείριση Μεγάλων Δεδομένων (ή αλλιώς Big Data) κινούμενο από πραγματικές ανάγκες της παραγωγής. Η πρώτη εμφάνιση του όρου έγινε το 1997 από τους επιστήμονες της Εθνικής Διοίκησης Αεροναυτικής και Διαστήματος (NASA). Ανέφεραν ότι αδυνατούσαν να αναπαραστήσουν γραφικά (visualization) τα σύνολα δεδομένων που κατείχαν, καθώς ήταν τόσο μεγάλα που ήταν ακατόρθωτο να τα αποθηκεύσουν στη κύρια μνήμη, στον τοπικό δίσκο και σε εξωτερικό σκληρό δίσκο. Έτσι δήλωσαν ότι αντιμετωπίζουν πρόβλημα Μεγάλων Δεδομένων. Οι τελευταίες τεχνολογικές εξελίξεις κυρίως στον τομέα των επικοινωνιών και των ολοκληρωμένων κυκλωμάτων έχουν δώσει την δυνατότητα να δημιουργηθούν μηχανισμοί παρακολούθησης των λειτουργιών ενός οργανισμού σε πολύ λεπτομερές επίπεδο. Η λεπτομερής αυτή ψηφιοποίηση των διαδικασιών παραγωγής έχουν καταστήσει μεγάλους οργανισμούς αλλά και εταιρείες μικρού μεγέθους ικανούς να παράγουν τεράστιους όγκους δεδομένων με πολύ ταχείς ρυθμούς. Τα δεδομένα αυτά κρύβουν πολύτιμη γνώση καθώς η ανάλυση τους μπορεί να οδηγήσει σε σημαντικές βελτιστοποιήσεις της παραγωγής αλλά και προβλήματα, αφού οι υπάρχουσες τεχνολογικές λύσεις για την διαχείριση δεδομένων δεν ανταποκρίνονται πλήρως στον όγκο αλλά και στην φύση τους. Ψηφιακά δεδομένα συναντώνται πλέον παντού: σε κάθε τομέα, σε κάθε οικονομία, σε κάθε οργανισμό και χρήση της ψηφιακής τεχνολογίας. Τα Μεγάλα Δεδομένα έλκουν όλο και περισσότερο το ενδιαφέρον των ηγετών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται να ωφεληθούν από την αξιοποίησή τους. Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή. Επίσης, τα μέσα εξόρυξης γνώσης από τα δεδομένα σημειώνουν σημαντική βελτίωση, καθώς τα διαθέσιμα λογισμικά για την εφαρμογή τεχνικών αυξανόμενης πολυπλοκότητας συνδυάζονται με την αυξανόμενη υπολογιστική ισχύ. Επιπλέον, η δυνατότητα παραγωγής, επικοινωνίας, μερισμού και πρόσβασης δεδομένων έχει εκτοξευθεί από την αύξηση του αριθμού των ατόμων, συσκευών και αισθητήρων, που συνδέονται σήμερα σε ψηφιακά δίκτυα. Το 2010, περισσότερα από

4 δισεκατομμύρια άνθρωποι, ή το 60% του παγκόσμιου πληθυσμού, χρησιμοποιούσαν κινητά τηλέφωνα, και περίπου 12% από αυτούς τους ανθρώπους είχαν έξυπνα τηλέφωνα, των οποίων η διείσδυση αυξάνεται κατά περισσότερο από 20% το χρόνο. Περισσότερα από 30 εκατομμύρια δικτυωμένοι κόμβοι αισθητήρων βρίσκονται πλέον στους κλάδους μεταφορών, αυτοκινητοβιομηχανίας, επιχειρήσεων κοινής ωφέλειας, καθώς και σε τομείς του λιανικού εμπορίου. Ο αριθμός αυτών των αισθητήρων αυξάνεται σε ποσοστό άνω του 30%. Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή δεδομένων. Αυτός είναι ο λόγος που τα μεγάλης κλίμακας δεδομένα έχουν γίνει πρόσφατη περιοχή των στρατηγικών επενδύσεων για τους οργανισμούς Πληροφορικής. Δεν είναι όμως μόνο οι οργανισμοί που παράγουν τεράστιους όγκους δεδομένων. Ακόμη και σε μικρότερη κλίμακα οργάνωσης, στο επίπεδο του ατόμου, η παραγωγή δεδομένων είναι πρωτόγνωρη. Οι περισσότεροι άνθρωποι διαθέτουν έναν ψηφιακό εαυτό, ως προβολή των δραστηριοτήτων τους στα κοινωνικά δίκτυα. Η Google εκτιμά ότι κάθε δύο μέρες το ψηφιακό υλικό που δημιουργείται από τους χρήστες είναι ισομεγέθες με το έντυπο υλικό που παρήγαγε η ανθρωπότητα από την αρχή της γραφής μέχρι το 2003. Έκρηξη στον όγκο των παραγόμενων δεδομένων παρατηρείται ακόμη στην επιστημονική έρευνα. Τομείς, όπως η ιατρική, η αστρονομία, η μετεωρολογία αλλά και η βιολογία χάρη στις νέες τεχνολογίες, τα νέα τηλεσκόπια, τους νέους και φτηνούς αισθητήρες και τα νέα μηχανήματα για την αποκωδικοποίηση DNA μπορούν και παράγουν όγκους δεδομένων που δεν είναι δυνατόν να αντιμετωπιστούν με τις υπάρχουσες υποδομές. Μάλιστα οι ρυθμοί αύξησης παρατηρούμε ότι είναι εκθετικής κατανομής. Έτσι προβλέπεται για τα επόμενα χρόνια μια ακόμη μεγαλύτερη "έκρηξη πληροφορίας". Θα λέγαμε όμως, ότι δεν υπάρχει ένα όριο μεγέθους δεδομένων πάνω από το οποίο αποκαλούνται "Μεγάλα Δεδομένα". Υπολογίζεται ότι σήμερα με το συγκεκριμένο όρο αναφερόμαστε συνήθως σε όγκους δεδομένων που κυμαίνονται από μερικά terabytes έως δεκάδες ή και εκατοντάδες petabytes (1.024 terabytes) ή exabytes (1.024 petabytes) ή zetabytes (1.024 exabytes).

- Το 2011, η ανθρωπότητα δημιούργησε πάνω από 1,2 τρισεκατομμύρια GB δεδομένων.
- Η Google λαμβάνει πάνω από 2.000.000 ερωτήματα αναζήτησης κάθε λεπτό.
- 72 ώρες βίντεο προστίθενται στο YouTube κάθε λεπτό.
- Υπάρχουν 217 νέοι χρήστες του Ίντερνετ κάθε λεπτό.
- Οι χρήστες του Twitter στέλνουν πάνω από 100.000 tweets κάθε λεπτό (που είναι πάνω από 140 εκατομμύρια ανά ημέρα).
- Εταιρείες, και οργανισμοί λαμβάνουν 34.000 "likes" σε κοινωνικά δίκτυα κάθε λεπτό.
- Διεθνή δεδομένα Εταιρειών προβλέπουν ότι η αγορά για την τεχνολογία των μεγάλης κλίμακας δεδομένων και υπηρεσιών θα φτάσει τα 16,9 εκατομμύρια δολλάρια.

Κάποιοι Ορισμοί που έχουν δοθεί για τα Μεγάλα Δεδομένα:
Η Gartner (η μεγαλύτερη επιχείρηση στον κόσμο που ασχολείται με την τεχνολογική έρευνα και συμβουλευτική), το 2012 έδωσε τον εξής ορισμό: "Τα Big Data είναι υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών". Στα "Μεγάλα Δεδομένα" συγκαταλέγονται όλες οι πληροφορίες των μέσων επικοινωνίας που είναι προσβάσιμες σε όλους μας και βρίσκονται στο Διαδίκτυο, δηλαδή φωτογραφίες, video και κείμενα, καθώς και όλα τα "κλειστά δεδομένα" των διαφόρων εταιριών αλλά και των κυβερνήσεων. Δηλαδή, η Gartner πρότεινε έναν ορισμό που περιλάμβανε τα τρία Vs (Volume, Velocity, Variety): τον όγκο, την ταχύτητα και την ποικιλία. Πρόκειται για έναν ορισμό που εστιάζει στο μέγεθος. Η έκθεση επισημαίνει το αυξανόμενο μέγεθος των δεδομένων, το αυξανόμενο ποσοστό παραγωγής τους και το αυξανόμενο εύρος των μορφών που εφαρμόζονται. Ο ορισμός αυτός έχει επαναληφθεί από τη NIST (National Institute of Standards and Technology) και διευρυνθεί από την IBM (International Business Machines) για να συμπεριλάβει και ένα τέταρτο V: την πιστότητα (Veracity).

Η Oracle αποφεύγει την χρήση των Vs για να καταλήξει σε έναν ορισμό. Αντίθετα, υποστηρίζει ότι τα μεγάλα στοιχεία είναι η δημιουργία αξίας από παραδοσιακές σχεσιακές βάσεις δεδομένων με στόχο τη λήψη επιχειρηματικών αποφάσεων, η οποία είναι εμπλουτισμένη με νέες πηγές μη δομημένων δεδομένων. Οι νέες αυτές πηγές περιλαμβάνουν blogs, social media, δίκτυα αισθητήρων, δεδομένα εικόνας και άλλες μορφές δεδομένων, τα οποία ποικίλλουν σε μέγεθος, δομή, μορφή και άλλους παράγοντες. Η Oracle υποστηρίζει ότι τα Μεγάλα Δεδομένα είναι το αποτέλεσμα από την ένταξη πρόσθετων πηγών δεδομένων για να αυξήσουν τις ήδη υπάρχουσες λειτουργίες. Αξίζει να σημειωθεί ότι ο ορισμός της Oracle εστιάζει στην υποδομή. Σε αντίθεση με ορισμούς που εκφράστηκαν από άλλους, η Oracle δίνει έμφαση σε μια σειρά από τεχνολογίες όπως: σχεσιακές και μη σχεσιακές βάσεις δεδομένων, εργαλεια διαχείρισης μεγάλων δεδομενών και γλώσσες προγραμματισμού. Έτσι, παρείχαν και έναν ορισμό και μια λύση για τα Μεγάλα Δεδομένα. Παρόλο που ο ορισμός αυτός είναι σχετικά πιο εύκολο να υιοθετηθεί σε σχέση με άλλους, υστερεί ωστόσο στην ποσοτικοποίηση. Σύμφωνα με τον ορισμό της Oracle δεν είναι σαφές ως προς το πότε ακριβώς ο όρος Μεγάλα Δεδομένα εντοπίζεται στην πράξη και παρέχει περισσότερο μία έννοια "θα τα καταλάβετε όταν τα δείτε".

Η Intel είναι μία από τις λίγες επιχειρήσεις που παρέχουν ποσοτικά στοιχεία στη βιβλιογραφία τους. Η Intel συσχετίζει τα Μεγάλα Δεδομένα με οργανισμούς που "δημιουργούν κατά μέσο όρο 300 terabytes (TB) δεδομένων εβδομαδιαίως". Αντί να δώσει έναν ορισμό όπως έκαναν οι προαναφερθέντες οργανισμοί, περιγράφει τα Μεγάλα Δεδομένα ποσοτικοποιώντας τις εμπειρίες των επιχειρηματικών εταίρων της. Επισημαίνει ότι οι οργανισμοί οι οποίοι μελετήθηκαν ασχολούνται εκτενώς με μη δομημένα δεδομένα και δίνουν έμφαση στη διεξαγωγή αναλύσεων των δεδομένων τους τα οποία παράγονται με ρυθμό 500 terabytes ανά εβδομάδα. Τέλος, ισχυρίζεται ότι ο πιο σύνηθες τύπος δεδομένων που συναντάται είναι οι επιχειρηματικές συναλλαγές που είναι αποθηκευμένες σε σχεσιακές βάσεις δεδομένων (σύμφωνα με τον

ορισμό της Oracle), και ακολουθούν τα έγγραφα, τα e-mail, τα blogs και τα social media.

Η Microsoft παρέχει ένα ιδιαίτερα περιεκτικό ορισμό: "Μεγάλα Δεδομένα" είναι ο όρος που χρησιμοποιείται όλο και περισσότερο για να περιγράψει τη διαδικασία εφαρμογής σημαντικής υπολογιστικής ισχύος - την τελευταία λέξη της Μηχανικής Μάθησης και της τεχνητής νοημοσύνης - σε μαζικά και εξαιρετικά πολύπλοκα σύνολα πληροφοριών. Ο ορισμός αυτός καθιστά σαφές ότι τα Μεγάλα Δεδομένα απαιτούν σημαντική υπολογιστική ισχύ. Η σημασία της υπολογιστικής ισχύος αναφέρθηκε και σε προηγούμενους ορισμούς, αλλά δεν ορίστηκε με ακρίβεια. Επιπλέον, ο ορισμός αυτός εισάγει δύο τεχνολογίες: την Μηχανική Μάθηση (Machine Learning) και την Τεχνητή Νοημοσύνη (Artificial Intelligence) που είχαν αγνοηθεί από προηγούμενους ορισμούς. Αυτό, ως εκ τούτου, εισάγει την ιδέα ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες που είναι ζωτικής σημασίας συστατικά του τελικού ορισμού.

Η GoogleTrends αναφέρει τους ακόλουθους όρους σε σχέση με τα Μεγάλα Δεδομένα: "ανάλυση δεδομένων, Hadoop, NoSQL, Google, IBM, και Oracle". Από αυτούς τους όρους μια σειρά από τάσεις είναι εμφανείς. Πρώτον, ότι τα Μεγάλα Δεδομένα είναι άρρηκτα συνδεδεμένα με την ανάλυση δεδομένων και την εξαγωγή γνώσης από τα δεδομένα. Δεύτερον, είναι σαφές ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες όπως φαίνεται και από τον το ορισμό της Microsoft. Τέλος, είναι προφανές ότι υπάρχει ένας αριθμός οργανισμών, κυρίως βιομηχανικών οργανισμών που σχετίζονται με Μεγάλα Δεδομένα. Όπως επισημαίνεται από το GoogleTrends, υπάρχει μια σειρά από τεχνολογίες που συχνά αναφέρονται ότι εμπλέκονται με τα Μεγάλα Δεδομένα. Αποθήκες μη δομημένων δεδομένων (NoSQL) όπως "Amazon, Dynamo, Cassandra, CouchDB, Mongo DB" κ.ά. παίζουν κρίσιμο ρόλο στην αποθήκευση μεγάλου όγκου μη δομημένων και ιδιαίτερα μεταβαλλόμενων δεδομένων. Για τη χρήση των χώρων αποθήκευσης δεδομένων NoSQL υπάρχει μια σειρά εργαλείων ανάλυσης και μεθόδων, συμπεριλαμβανομένων του στατιστικού προγραμματισμού, της μηχανικής μάθησης και την οπτικοποίηση πληροφοριών. Η εφαρμογή μίας από αυτές τις τεχνολογίες από μόνη της δεν είναι επαρκής για να αξιολογήσει τη χρήση του όρου Μεγάλα Δεδομένα. Αντίθετα, άλλες τάσεις δείχνουν ότι είναι ο συνδυασμός μιας σειράς τεχνολογιών και η χρήση σημαντικών συνόλων δεδομένων που εξηγούν τον όρο. Οι τάσεις αυτές δείχνουν τα Μεγάλα Δεδομένα σαν μια τεχνική κίνηση η οποία ενσωματώνει ιδέες, νέες και παλιές και σε αντίθεση με άλλους ορισμούς παρέχει λίγες αναφορές ως προς τις κοινωνικές και επιχειρηματικές επιπτώσεις.

Το 2014 η Wikipedia περιέγραφε τα Μεγάλα Δεδομένα ως "ένα ευρύτερο όρο για οποιαδήποτε συλλογή Σύνολο Δεδομένων τόσο μεγάλων και σύνθετων που είναι δύσκολο να επεξεργαστούν χρησιμοποιώντας χειροκίνητα εργαλεία ή παραδοσιακές εφαρμογές επεξεργασίας δεδομένων".

Δεκάδες είναι οι ορισμοί που δόθηκαν. Μελετώντας τους, μπορούμε να περιγράψουμε τον όρο ως Συλλογές Δεδομένων, τα οποία είναι δομημένα κατά ένα ποσοστό και αδόμητα στη πλειονότητά τους, και ο όγκος τους είναι τόσο μεγάλος που καθιστά πολύ δύσκολη την απο-

θήκευση, επεξεργασία και ανάλυσή τους με τη χρήση παραδοσιακών τεχνικών της επιστήμης της πληροφορικής. Τέλος, θα ήταν λάθος να προβούμε σε έναν ακριβή προσδιορισμό του όρου Μεγάλα Δεδομένα, αφού κάθε επιχείρηση χρησιμοποιεί την επιστήμη των Big Data για διαφορετικούς σκοπούς, επομένως είναι εμφανής η πολυδιάστατη φύση τους που σαφώς με το πέρασμα του χρόνου συνεχώς εμπλουτίζεται.

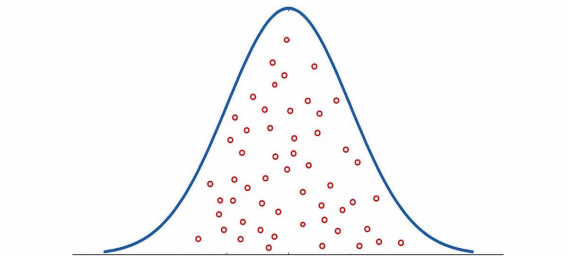
1.2 Τι είναι η δειγματοληψία και ποίοι είναι οι λόγοι που κάνουμε δειγματοληψία

Ας ξεκινήσουμε λοιπόν με ένα παράδειγμα από την καθημερινότητα και για να καταλάβουμε τους λόγους για τους οποίους η δειγματοληψία είναι ιδιαίτερα σημαντική. Κάθε δέκα χρόνια, η κυβέρνηση των ΗΠΑ διεξάγει απογραφή, δηλαδή καταμέτρηση κάθε ατόμου που ζει στη χώρα, όπως απαιτεί το σύνταγμα. Πρόκειται για ένα τεράστιο εγχείρημα. Το γραφείο απογραφής στέλνει μια επιστολή ή έναν εργαζόμενο σε κάθε αμερικανικό νοικοκυριό και προσπαθεί να συγκεντρώσει στοιχεία που θα επιτρέψουν την καταμέτρηση κάθε ατόμου. Αφού συγκεντρωθούν τα δεδομένα, πρέπει να επεξεργαστούν, να ταξινομηθούν και να αναφερθούν. Το όλο εγχείρημα απαιτεί χρόνια σχεδιασμού και δισεκατομμύρια δολάρια, γεγονός που εγείρει το ερώτημα: Υπάρχει καλύτερος τρόπος; Όπως αποδεικνύεται, υπάρχει. Αντί να έρχονται σε επαφή με κάθε άτομο του πληθυσμού, οι ερευνητές μπορούν να απαντήσουν στις περισσότερες ερωτήσεις με δειγματοληψία ατόμων. Στην πραγματικότητα, η δειγματοληψία είναι αυτό που κάνει το γραφείο απογραφής προκειμένου να συγκεντρώσει λεπτομερείς πληροφορίες για τον πληθυσμό, όπως το μέσο εισόδημα των νοικοκυριών, το επίπεδο εκπαίδευσης που έχουν οι άνθρωποι και το είδος της εργασίας που κάνουν οι άνθρωποι για να ζήσουν.

Αλλά τι ακριβώς είναι η δειγματοληψία και πώς λειτουργεί; Ένα δείγμα είναι ένα μικρό κομμάτι ή μέρος από κάτι που αντιπροσωπεύει ένα ευρύτερο σύνολο. Οι επιστήμονες συχνά συγκεντρώνουν δεδομένα από μια μικρή ομάδα (δείγμα) ως τρόπο κατανόησης ενός μεγαλύτερου συνόλου (πληθυσμός). Ακόμη και όταν ο πληθυσμός που μελετάται είναι τόσο μεγάλος όσο οι ΗΠΑ -περίπου 330 εκατομμύρια άνθρωποι- οι ερευνητές συχνά χρειάζεται να πάρουν δείγμα μόνο μερικών χιλιάδων ανθρώπων για να κατανοήσουν τους πάντες.

Οι ερευνητές μπορούν να κατανοήσουν με ακρίβεια εκατοντάδες εκατομμύρια ανθρώπους συλλέγοντας δεδομένα από μόλις μερικές χιλιάδες από αυτούς χάρις στους μαθηματικούς Valery Ivanovich Glivenko και Francesco Paolo Cantelli. Οι Glivenko και Cantelli (μεγάλοι ερευνητές της στατιστικής) μελέτησαν την θεωρία πιθανοτήτων. Στις αρχές της δεκαετίας του 1900, ανακάλυψαν ότι πολλές παρατηρήσεις που αντλούνται τυχαία από έναν πληθυσμό θα πάρουν φυσικά το σχήμα της κατανομής του πληθυσμού. Αυτό σημαίνει ότι, εφόσον οι ερευνητές επιλέγουν τυχαία δείγματα από έναν πληθυσμό και λαμβάνουν ένα δείγμα επαρκούς μεγέθους, τότε το δείγμα θα περιέχει χαρακτηριστικά που θα αντικατοπτρίζουν περίπου εκε-

ίνα του πληθυσμού. Αξίζει να αναφερθεί ότι στο Σχήμα 1.1 η μπλε γραμμή αντιπροσωπεύει μια κανονική κατανομή, γνωστή και ως καμπύλη καμπάνας. Κάθε κόκκινος κύκλος αντιπροσωπεύει μια παρατήρηση ή ένα άτομο που έχει ληφθεί δειγματοληπτικά από τον πληθυσμό. Εάν κάθε παρατήρηση επιλέγεται τυχαία, τότε το δείγμα θα αντανακλά φυσικά τις ιδιότητες του πληθυσμού. Χάρη σε αυτή την ιδιότητα των πιθανοτήτων, οι ερευνητές είναι σε θέση να κατανοήσουν μεγάλους πληθυσμούς με τη δειγματοληψία μικρών ομάδων από τον πληθυσμό



Σχήμα 1.1: Κανονική Κατανομή

Η τυχαία δειγματοληψία λαμβάνει χώρα όταν ο ερευνητής εξασφαλίζει ότι κάθε μέλος του πληθυσμού που μελετάται έχει ίσες πιθανότητες να επιλεγεί για να συμμετάσχει στη μελέτη. Είναι σημαντικό ότι ο πληθυσμός που μελετάται δεν είναι απαραίτητα όλοι οι κάτοικοι μιας χώρας ή μιας περιοχής. Αντίθετα, ο πληθυσμός μπορεί να αναφέρεται σε άτομα που μοιράζονται μια κοινή ιδιότητα ή χαρακτηριστικό. Έτσι, όλοι όσοι αγόρασαν ένα Ford τα τελευταία πέντε χρόνια μπορούν να αποτελούν πληθυσμό, το ίδιο και οι εγγεγραμμένοι ψηφοφόροι σε μια πολιτεία ή οι φοιτητές ενός πανεπιστημίου μιας πόλης. Ένας πληθυσμός είναι η ομάδα που οι ερευνητές θέλουν να κατανοήσουν.

Προκειμένου να κατανοήσουν έναν πληθυσμό χρησιμοποιώντας τυχαία δειγματοληψία, οι ερευνητές ξεκινούν με τον προσδιορισμό ενός πλαισίου δειγματοληψίας., ενός καταλόγου όλων των ατόμων του πληθυσμού που οι ερευνητές θέλουν να μελετήσουν. Για παράδειγμα, μια βάση δεδομένων με όλους τους αριθμούς σταθερών και κινητών τηλεφώνων στις ΗΠΑ είναι ένα πλαίσιο δειγματοληψίας. Μόλις ο ερευνητής έχει ένα πλαίσιο δειγματοληψίας, μπορεί να επιλέξει τυχαία άτομα από τον κατάλογο για να συμμετάσχουν στη μελέτη.

Ωστόσο, δεν είναι πάντα πρακτικό ή ακόμη και δυνατό να συγκεντρωθεί ένα πλαίσιο δειγματοληψίας. Παρ' όλα αυτά, υπάρχουν πολύ καλοί λόγοι για τους οποίους οι ερευνητές μπορεί να θέλουν να μελετήσουν άτομα σε κάθε μία από αυτές τις ομάδες. Όταν δεν είναι δυνατό ή πρακτικό να συγκεντρωθεί ένα τυχαίο δείγμα, οι ερευνητές συχνά συγκεντρώνουν ένα μη τυχαίο δείγμα. Ένα μη τυχαίο δείγμα είναι ένα δείγμα στο οποίο κάθε μέλος του πληθυσμού που μελετάται δεν έχει ίσες πιθανότητες να επιλεγεί στη μελέτη.

Επειδή τα μη τυχαία δείγματα δεν επιλέγουν τους συμμετέχοντες βάσει πιθανοτήτων, είναι

συχνά δύσκολο να γνωρίζουμε πόσο καλά το δείγμα αντιπροσωπεύει τον πληθυσμό που μας ενδιαφέρει. Παρά τον περιορισμό αυτό, ένα ευρύ φάσμα μελετών που διεξάγονται στον ακαδημαϊκό χώρο, τη βιομηχανία και την κυβέρνηση βασίζονται σε μη τυχαία δείγματα. Όταν οι ερευνητές χρησιμοποιούν μη τυχαία δείγματα, είναι σύνηθες να ελέγχουν τυχόν γνωστές πηγές δειγματοληπτικής μεροληψίας κατά τη συλλογή δεδομένων. Με τον έλεγχο των πιθανών πηγών μεροληψίας, οι ερευνητές μπορούν να μεγιστοποιήσουν τη χρησιμότητα και τη δυνατότητα γενίκευσης των δεδομένων τους.

Εν κατακλείδι, η δειγματοληψία στη στατιστική είναι η τεχνική της επιλογής ενός μέρους του πληθυσμού (το οποίο ονομάζεται δείγμα). Με την ορολογία πληθυσμός εννοούμε ένα πλήθος παρατηρήσεων ή μετρήσεων ο οποίος μπορεί να αποτελεί ένα πεπερασμένο ή άπειρο πλήθος στοιχείων (ονομάζεται μέγεθος του πληθυσμού και συμβολίζεται με N). Το πλήθος των στοιχείων ενός δείγματος ονομάζεται μέγεθος του δείγματος και συμβολίζεται με n.

Γιατί όμως είναι σημαντική η δειγματοληψία για τους ερευνητές και όχι μόνο; Σε ένα ερευνητικό έργο οι πόροι είναι περιορισμένοι ο χρόνος, τα χρήματα και οι άνθρωποι δεν είναι ποτέ απεριόριστα διαθέσιμα. Για το λόγο αυτό, τα περισσότερα ερευνητικά έργα στοχεύουν στη συλλογή δεδομένων από ένα δείγμα ανθρώπων και όχι από ολόκληρο τον πληθυσμό. Αυτό συμβαίνει επειδή η δειγματοληψία επιτρέπει στους ερευνητές να:

- Να εξοικονομήσουν χρόνο: Η επικοινωνία με όλους τους κατοίκους ενός πληθυσμού απαιτεί χρόνο. Και αναπόφευκτα, ορισμένοι άνθρωποι δεν θα ανταποκριθούν στην πρώτη προσπάθεια επικοινωνίας μαζί τους, πράγμα που σημαίνει ότι οι ερευνητές πρέπει να επενδύσουν περισσότερο χρόνο για την παρακολούθηση. Η τυχαία δειγματοληψία είναι πολύ ταχύτερη από την έρευνα όλων σε έναν πληθυσμό, και η λήψη ενός μη τυχαίου δείγματος είναι σχεδόν πάντα ταχύτερη από την τυχαία δειγματοληψία. Συνεπώς, η δειγματοληψία εξοικονομεί πολύ χρόνο στους ερευνητές.

- Να εξοικονομήσουν χρήματα: Ο αριθμός των ατόμων με τα οποία έρχεται σε επαφή ένας ερευνητής σχετίζεται άμεσα με το κόστος μιας μελέτης. Η δειγματοληψία εξοικονομεί χρήματα επιτρέποντας στους ερευνητές να συλλέξουν τις ίδιες απαντήσεις από ένα δείγμα που θα λάμβαναν από τον πληθυσμό. Η μη τυχαία δειγματοληψία είναι σημαντικά φθηνότερη από την τυχαία δειγματοληψία, επειδή μειώνει το κόστος που συνδέεται με την εύρεση ατόμων και τη συλλογή δεδομένων από αυτούς. Επειδή όλες οι έρευνες διεξάγονται με προϋπολογισμό, η εξοικονόμηση χρημάτων είναι σημαντική.

- Να συλλέγουν πλουσιότερα δεδομένα: Μερικές φορές, ο στόχος της έρευνας είναι να συλλέξει λίγα δεδομένα από πολλούς ανθρώπους (π.χ. μια δημοσκόπηση). Άλλες φορές, ο στόχος είναι να συλλεχθούν πολλές πληροφορίες από λίγους ανθρώπους (π.χ. μια μελέτη χρηστών ή μια εθνογραφική συνέντευξη). Όπως και να έχει, η δειγματοληψία επιτρέπει στους ερευνητές να θέτουν στους συμμετέχοντες περισσότερες ερωτήσεις και να συλλέγουν πλουσιότερα δεδομένα από ό,τι αν επικοινωνούσαν με όλους τους ανθρώπους ενός πληθυσμού.

1.3 Μέθοδοι δειγματοληψίας

Όπως ήδη αναφέρθηκε στην προηγούμενη παράγραφο, η δειγματοληψία είναι μια μέθοδος που μας επιτρέπει να αντλήσουμε πληροφορίες για τον πληθυσμό με βάση τα στατιστικά στοιχεία από ένα υποσύνολο του πληθυσμού (δείγμα), χωρίς να χρειάζεται να ερευνήσουμε κάθε άτομο. Υπάρχουν δυο βασικές μέθοδοι δειγματοληψίας: η Τυχαία Δειγματοληψία (probability sampling) και η μη Τυχαία Δειγματοληψία (non-probability sampling). Συγκεκριμένα:

- Τυχαία Δειγματοληψία. Στη Τυχαία Δειγματοληψία κάθε στοιχείο του πληθυσμού έχει ίσες πιθανότητες να επιλεγεί. Έτσι, δημιουργείται ένα δείγμα που είναι αντιπροσωπευτικό του πληθυσμού
- Μη Τυχαία Δειγματοληψία. Στη Μη Τυχαία Δειγματοληψία κάθε στοιχείο του πληθυσμού δεν έχει ίσες πιθανότητες να επιλεγεί. Έτσι, είναι προφανές ότι το δείγμα μπορεί να μην είναι αντιπροσωπευτικό για να εξαγάγουμε γενικά αποτελέσματα

Τυχαία Δειγματοληψία

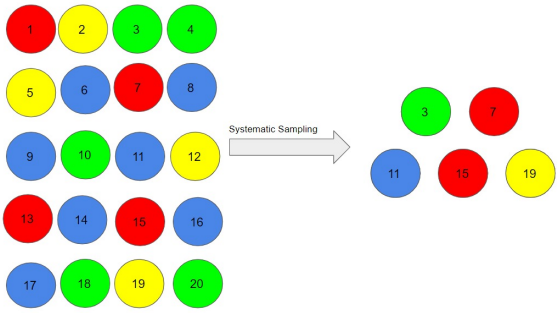
Υπάρχουν 4 κύριοι τύποι τυχαίας δειγματοληψίας: η Απλή Τυχαία Δειγματοληψία (Simple Random Sampling) , η Συστηματική Δειγματοληψία (Systematic Sampling), η Διαστρωματική Δειγματοληψία (Stratified Sampling) και η Δειγματοληψία σε Ομάδες (Cluster Sampling). Συγκεκριμένα:

- Απλή Τυχαία Δειγματοληψία: Κάθε άτομο επιλέγεται τυχαία και κάθε μέλος του πληθυσμού έχει ίσες πιθανότητες να επιλεγεί.

Σχήμα 1.2: Απλή Τυχαία Δειγματοληψία

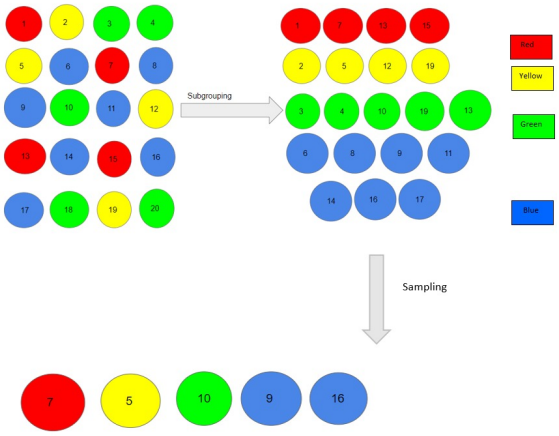
- Συστηματική Δειγματοληψία: Το πρώτο άτομο επιλέγεται τυχαία και τα υπόλοιπα επιλέγονται σύμφωνα με ένα σταθερό «διάστημα δειγματοληψίας». Η συστηματική δειγματοληψία είναι πιο βολική από την τυχαία δειγματοληψία. Ωστόσο, μπορεί να οδηγήσει σε μεροληψία εάν υπάρχει μοτίβο στο οποίο επιλέγουμε στοιχεία από τον πληθυσμό. Ένα, χαρακτηρικό παράδειγμα αυτής της δειγματοληψίας μπορεί να παρουσιαστεί στην κάτωθι εικόνα: Θεωρούμε μέγεθος πληθυσμούς n και θέλουμε να επιλέξουμε ένα δείγμα μεγέθους m. Τότε, για κάθε επόμενο άτομο που επιλέγουμε θεωρούμε την νόρμα $\frac{m}{n}$. Ας υποθέσουμε ότι ξεκινήσαμε με το άτομο με τον αριθμό 3 και θέλουμε ένα μέγεθος

δείγματος 5. Έτσι, το επόμενο άτομο που θα επιλέξουμε θα είναι σε διάστημα (20/5) = 4 από το 3ο άτομο, δηλαδή 7 (3+4), και ούτω καθεξής.



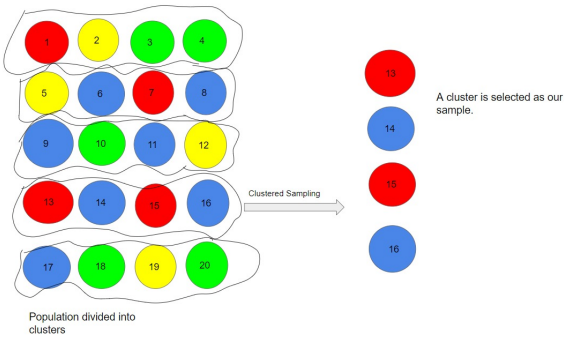
Σχήμα 1.3: Συστηματική Δειγματοληψία

- Διαστρωματική Δειγματοληψία: Χωρίζουμε τον πληθυσμό σε υποομάδες (που ονομάζονται στρώματα) με βάση διαφορετικά χαρακτηριστικά όπως το φύλο, η κατηγορία κ.λπ. Στη συνέχεια επιλέγουμε το(α) δείγμα(τα) από αυτές τις υποομάδες. Χρησιμοποιούμε αυτόν τον τύπο δειγματοληψίας όταν θέλουμε εκπροσώπηση από όλες τις υποομάδες του πληθυσμού. Ωστόσο, η τεχνική αυτή δειγματοληψίαα απαιτεί κατάλληλη γνώση των χαρακτηριστικών του πληθυσμού.



Σχήμα 1.4: Διαστρωματική Δειγματοληψία

- Δειγματοληψία σε Ομάδες: Χρησιμοποιούμε τις υποομάδες του πληθυσμού ως μονάδα δειγματοληψίας και όχι τα άτομα. Ο πληθυσμός χωρίζεται σε υποομάδες, γνωστές ως συστάδες, και μια ολόκληρη συστάδα επιλέγεται τυχαία για να συμπεριληφθεί στη μελέτη. Αυτός ο τύπος δειγματοληψίας χρησιμοποιείται όταν εστιάζουμε σε μια συγκεκριμένη περιοχή.

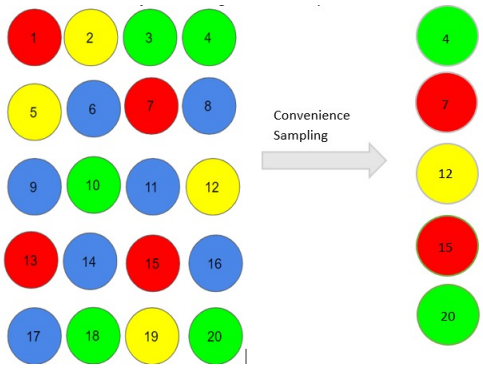


Σχήμα 1.5: Δειγματοληψία σε Ομάδες

Μη Τυχαία Δειγματοληψία

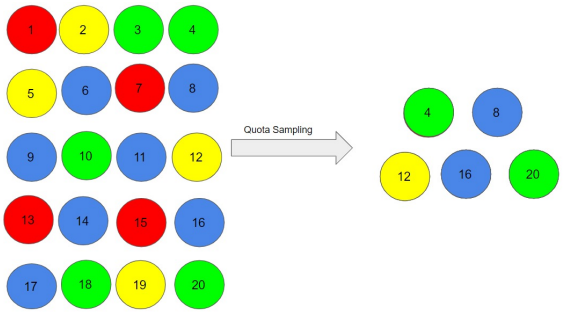
Υπάρχουν 4 κύριοι τύποι της Μη Τυχαίας Δειγματοληψίας: Βολική Δειγματοληψία (Convenience Sampling), Δειγματοληψία με Ποσόστωση (Quota Sampling), Δειγματοληψία με Κρίση (Judgment Sampling) και Δειγματοληψία Χιονοστιβάδας (Snowball Sampling).

- Βολική Δειγματοληψία: Τα άτομα επιλέγονται με βάση τη διαθεσιμότητα και την προθυμία τους να συμμετάσχουν. Η δειγματοληψία ευκολίας είναι επιρρεπής σε σημαντική μεροληψία. Για παράδειγμα, αν υποθέσουμε ότι τα άτομα με τους αριθμούς 4, 7, 12, 15 και 20 θέλουν να συμμετάσχουν στο δείγμα μας και, ως εκ τούτου, θα τα συμπεριλάβουμε στο δείγμα.



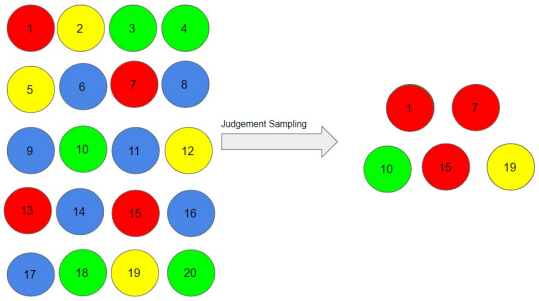
Σχήμα 1.6: Βολική Δειγματοληψία

- Δειγματοληψία με Ποσόστωση: Επιλέγουμε στοιχεία με βάση προκαθορισμένα χαρακτηριστικά του πληθυσμού. Στη δειγματοληψία με ποσόστωση, το επιλεγμένο δείγμα μπορεί να μην είναι η καλύτερη δυνατή εκπροσώπηση των χαρακτηριστικών του πληθυσμού που δεν λήφθηκαν υπόψη.



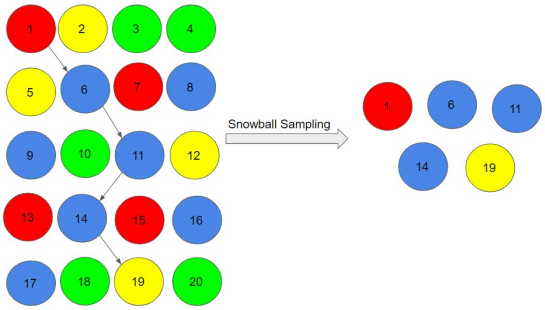
Σχήμα 1.7: Δειγματοληψία με Ποσόστωση

- Δειγματοληψία με Κρίση: Εξαρτάται από την κρίση των εμπειρογνωμόνων κατά την επιλογή των ατόμων που θα κληθούν να συμμετάσχουν. Η δειγματοληψία κατά ποσόστωση είναι επίσης επιρρεπής στη μεροληψία των εμπειρογνωμόνων και μπορεί να μην είναι απαραίτητα αντιπροσωπευτική.



Σχήμα 1.8: Δειγματοληψία με Κρίση

- Δειγματοληψία Χιονοστιβάδας: Ζητείται από τα υπάρχοντα άτομα να προτείνουν και άλλα γνωστά τους άτομα, έτσι ώστε το δείγμα να αυξάνεται σε μέγεθος όπως μια κυλιόμενη χιονόμπαλα. Αυτή η μέθοδος δειγματοληψίας είναι αποτελεσματική όταν είναι δύσκολο να προσδιοριστεί ένα πλαίσιο δειγματοληψίας. Υπάρχει σημαντικός κίνδυνος μεροληψίας επιλογής στη δειγματοληψία χιονοστιβάδας, καθώς τα άτομα που αναφέρονται θα έχουν κοινά χαρακτηριστικά με το άτομο που τα συστήνει.



Σχήμα 1.9: Δειγματοληψία Χιονοστιβάδας

1.4 Δειγματοληψία με την μέθοδο των λατινικών υπερκύβων

Η δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων είναι ένας τρόπος παραγωγής τυχαίων δειγμάτων από τις τιμές των παραμέτρων. Χρησιμοποιείται ευρέως στην προσομοίωση Monte Carlo, επειδή μπορεί να μειώσει δραστικά τον αριθμό των εκτελέσεων που απαιτούνται για την επίτευξη ενός λογικά ακριβούς αποτελέσματος.

Η δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων βασίζεται στον σχεδιασμό Λατινικού Τετραγώνου, ο οποίος έχει ένα μόνο δείγμα σε κάθε γραμμή και στήλη. Ένας "υπερκύβος" είναι ένας κύβος με περισσότερες από τρεις διαστάσεις. Το λατινικό τετράγωνο δηλαδή επεκτείνεται ώστε να λαμβάνονται δείγματα από πολλαπλές διαστάσεις και πολλαπλά υπερεπίπεδα.

Η μέθοδος πίσω από τη δειγματοληψία των λατινικών υπερκύβων

Η μονοδιάστατη δειγματοληψία Λατινικού Υπερκύβου περιλαμβάνει τη διαίρεση της αθροιστικής συνάρτησης πυκνότητας σε ίσα τμήματα και, στη συνέχεια, την επιλογή ενός τυχαίου σημείου δεδομένων σε κάθε τμήμα. Για να γίνει καλύτερα κατανοητό δίνεται το ακόλουθο παράδειγμα. Υποθέτουμε ότι ένα τυχαίο δείγμα με 100 σημεία δεδομένων. Πρώτον, διαιρούμε την συνάρτηση πυκνότητας σε 100 ίσα διαστήματα. Εάν η κατανομή ξεκινά από το 0 και τελειώνει με k, το πρώτο σημείο δεδομένων θα επιλεγεί από το διάστημα μεταξύ $(0, \frac{k}{100})$. Το δεύτερο σημείο δεδομένων θα προερχόταν από το διάστημα $(\frac{k}{200}, \frac{2k}{100})$, το τρίτο από το διάστημα $(\frac{2k}{100}, \frac{3k}{100})$ κ.ο.κ. Σε κάθε διάστημα θα επιλέγετε τυχαία ένα σημείο, δίνοντας 100 διαφορετικά σημεία.

Η δισδιάστατη δειγματοληψία Λατινικού Υπερκύβου δεν είναι πολύ πιο περίπλοκη και συνήθως εκτελείται με λογισμικό. Υποθέτοντας ότι δύο μεταβλητές, x_1 και x_2 είναι ανεξάρτητες, ακολουθείτε η μονοδιάστατη μέθοδος για να καταλήξουμε σε μονοδιάστατα δείγματα για x_1 και x_2 ξεχωριστά. Μόλις υπάρχουν δυο λίστες δειγμάτων, συνδυάζονται τυχαία σε δισδιάστατά ζεύγη. Για n-διάστατη δειγματοληψία Λατινικού Υπερκύβου χρησιμοποιείται η ίδια μέθοδος.

Γιατί δειγματοληψία με την μέθοδο των λατινικών;

Η δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων χρησιμοποιείται συνήθως για την εξοικονόμηση χρόνου επεξεργασίας στον υπολογιστή κατά την εκτέλεση προσομοιώσεων Monte Carlo. Μελέτες έχουν δείξει ότι μια καλά εκτελεσμένη δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων μπορεί να μειώσει τον χρόνο επεξεργασίας έως και κατά 50%.

Ένα άλλος τομέας, τον οποίο θα αναλύσουμε και στην συγκεκριμένη εργασία, που φαίνεται να παρουσιάζει καλά αποτελέσματα η μέθοδος των λατινικών υπερκύβων είναι ως μέθοδος δειγ-

ματοληψίας. Συγκεκριμένα, φαίνεται ότι αν δειγματοληπτίσουμε από τα δεδομένα μας με την μέθοδο των λατινικών υπερκύβων και στη συνέχεια εφαρμόσουμε κάποιο μοντέλο πρόβλεψης τα αποτελέσματα μας θα είναι σχεδόν ίδια με το να χρησιμοποιούσαμε όλα τα αρχικά δεδομένα.

Ενδιαφέρων αποτελεί και η ανάλυση μαζικών συνόλων δεδομένων με προσέγγιση για την εκτίμηση των παραμέτρων του πληθυσμού από ένα τεράστιο σύνολο δεδομένων. Η προτεινόμενη προσέγγιση μειώνει σημαντικά την απαιτούμενη ποσότητα μνήμης και η προκύπτουσα εκτίμηση είναι εξίσου αποδοτική με την ταυτόχρονη ανάλυση ολόκληρου του συνόλου δεδομένων (Li, Lin και Li (2012)).

Κεφάλαιο 2

Δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων

2.1 Βασικοί ορισμοί και έννοιες

Προκειμένου να κατανοήσει κανείς καλύτερα την χρήση των Λατινικών Υπερκύβων για την λύση προβλημάτων μεγάλων δεδομένων στην περιοχή της μηχανικής μάθησης στο κεφάλαιο αυτό δίνονται κάποιοι βασικοί ορισμοί και έννοιές για τους τόσο για τους λατινικούς υπερκύβους όσο και για τους λατινικούς σχεδιασμούς που είναι άρρηκτα συνδεδεμένοι.

Ορισμός 2.1.1: Ένας Σχεδιασμός Λατινικών Υπερκύβων είναι ένας $n \times k$ πίνακας, του οποίου κάθε στήλη είναι μια μετάθεση των $\{1, 2, ..., n\}$.

Η κύρια ιδιότητα ενός Σχεδιασμού Λατινικού Υπερκύβου (Latin Hypercube Design, LHD) είναι ότι επιτυγχάνει ομοιομορφία (uniformity) σε κάθε μια από τις m μεταβλητές (στήλες).

Παράδειγμα ενός 4×2 LHD:

$$\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 3 & 2 \\ 4 & 3 \end{bmatrix}$$

Συνήθως θεωρούμε ένα LHD στην κεντρική του μορφή (centered form). Δηλαδή, τα n στοιχεία κάθε στήλης θεωρούνται κεντραρισμένα ως προς το μηδέν και ισαπέχοντα (equally spread). Δηλαδή, θεωρούμε τα στοιχεία του συνόλου:

$$\{-\frac{(n-1)}{2}, -\frac{(n-3)}{2}, ..., 0, ..., \frac{(n-3)}{2}, \frac{(n-1)}{2}\}$$

αν το n είναι περιττός και

$$\{-\frac{(n-1)}{2}, -\frac{(n-3)}{2}, ..., -\frac{1}{2}, \frac{1}{2}, ..., \frac{(n-3)}{2}, \frac{(n-1)}{2}\}$$

αν το n είναι άρτιος.

Δηλαδή θεωρούμε τα στοιχεία:

$$\boxed{u_i = i - \frac{(n+1)}{2}, 1 \leq i \leq n}$$

Έτσι, στο προηγούμενο παράδειγμα, ο 4×2 LHD στην κεντρική του μορφή είναι:

$$\frac{1}{2}\begin{bmatrix} -3 & 3 \\ -1 & -3 \\ 1 & -1 \\ 3 & 1 \end{bmatrix}$$

Λήμμα 1: Είναι προφανές ότι:

1. $\sum_{i=1}^{n} u_i = 0$
2. $\sum_{i=1}^{n} u_i^2 = \frac{n(n^2-1)}{12}$

Απόδειξη:

1. $\sum_{i=1}^{n} u_i = \sum_{i=1}^{n}(i - \frac{(n+1)}{2}) = \sum_{i=1}^{n} i - \frac{n(n+1)}{2} = \frac{n(n+1)}{2} - \frac{n(n+1)}{2} = 0.$
2. $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}(i - \frac{(n+1)}{2})^2 = \sum_{i=1}^{n}(i^2 - 2i\frac{(n+1)}{2} + \frac{(n+1)^2}{4}) =$

   $\sum_{i=1}^{n} i^2 - (n+1)\sum_{i=1}^{n} i + \frac{n(n+1)^2}{4} = \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)n(n+1)}{2} + \frac{n(n+1)^2}{4} =$

   $\frac{n^3-n}{12} = \frac{n(n^2-1)}{12}$.

Ορισμός 2.1.2: Ένας LHD θα λέγεται ορθογώνιος, αν όλα τα ζεύγη των στηλών έχουν συσχέτιση μηδέν.

Λήμμα 2.1.1: Αν ο LHD δεν είναι στην κεντρική μορφή τότε δυο στήλες είναι ορθογώνιες όταν έχουν εσωτερικό γινόμενο ίσο με $\boxed{\frac{n(n+1)^2}{4}}$.

Απόδειξη: Αν x_1 και x_2 είναι δυο στήλες των LHD, τότε για να είναι ορθογώνιες πρέπει:

$Cov(x_1, x_2) = 0$,

Αλλά,

$Cov(x_1, x_2) = \sum_{i=1}^{n}(x_{i1} - \bar{x_1})(x_{i2} - \bar{x_2}) = \sum_{i=1}^{n} x_{i1}x_{i2} - \bar{x_2}\sum_{i=1}^{n} x_{i1} - \bar{x_1}\sum_{i=1}^{n} x_{i2} + n\bar{x_1}\bar{x_2}$,

Όμως,

$$\sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n} x_{i2} = \frac{n(n+1)}{2}$$

$$\bar{x_1} = \bar{x_2} = \frac{1}{2}\frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

,

Άρα:

$Cov(x_1, x_2) = \sum_{i=1}^{n} x_{i1}x_{i2} - \frac{(n+1)}{2} \cdot \frac{n(n+1)}{2} - \frac{(n+1)}{2} \cdot \frac{n(n+1)}{2} + \frac{n(n+1)(n+1)}{4} = \sum_{i=1}^{n} x_{i1}x_{i2} - \frac{n(n+1)^2}{4}$,

Άρα, $Cov(x_1, x_2) = 0 \Rightarrow \sum_{i=1}^{n} x_{i1}i2 = \frac{n(n+1)^2}{4}$.

Παράδειγμα 1: 4×2 LHD

$$\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 3 & 2 \\ 4 & 3 \end{bmatrix}$$

Οι δυο στήλες έχουν εσωτερικό γινόμενο 24. Για να είναι ορθογώνιες θα πρέπει να έχουν εσωτερικό γινόμενο $\frac{n(n+1)^2}{4} = 25$.

$$\frac{1}{2}\begin{bmatrix} -3 & 3 \\ -1 & -3 \\ 1 & -1 \\ 3 & 1 \end{bmatrix}$$

Στην κεντρική μορφή οι δύο στήλες έχουν εσωτερικό γινόμενο $-\frac{4}{4} = -1$

Παράδειγμα 2: 4×2 LHD

$$\begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 4 \\ 4 & 2 \end{bmatrix}$$

Έχουν εσωτερικό γινόμενο 25. Άρα είναι ορθογώνιες. Και στη κεντρική μορφή:

$$\frac{1}{2}\begin{bmatrix} -3 & 1 \\ -1 & -3 \\ 1 & 3 \\ 3 & -1 \end{bmatrix}$$

έχουν εσωτερικό γινόμενο 0.

Θεώρημα 2.1.1: Αν $n \equiv 2mod4$ δεν υπάρχει Οθογώνιος Σχεδιασμός Λατινικού Υπερκύβου (Orthogonal Latin Hypercube Design, OLHD με $k > 1$ στήλες.

Απόδειξη: Έστω ότι υπάρχει ένας τέτοιος σχεδιασμός. Τότε απο το κριτήριο ορθογωνιότητας, το εσωτερικό γινόμενο δυο τυχαίων στηλών θα είναι: $\frac{n(n+1)^2}{4}$. Αλλά, αφού $n = 4t + 2$, έχουμε:

$$\frac{n(n+1)^2}{4} = \frac{(4t+2)(16t^2+24t+9)}{4} = \frac{(2t+1)(16t^2+24t+9)}{4},$$

Το δεξιό μέρος της τελευταίας ισότητας δεν είναι ακέραιος αρθμός, και επομένως η ισότητα αυτή δεν ισχύει.

Θεώρημα 2.1.2: Αν $n \equiv 4mod8$ δεν υπάρχει OLHD) πρώτης τάξης με $k = 3$ στήλες, στον οποίο η αλληλεπίδραση οποιονδήποτε δυο στηλών να είναι ορθογώνια στην τρίτη στήλη.

Απόδειξη: Έστω $n \equiv 4mod8$ και έστω ακόμη ότι υπάρχει και τέτοιος σχεδιασμός. Συμβολίζουμε με:

$$\begin{bmatrix} a_1 & b1 & c1 \\ a_2 & b2 & c2 \\ ... & ... & ... \\ a_n & bn & cn \end{bmatrix}$$

τις 3 στήλες του σχεδιασμού αυτού που είναι μεταθέσεις των πρώτων n φυσικών αριθμών $\{1, 2, ..., n\}$.
Έστω x_1, x_2, x_3 είναι οι αντίστοιχες στήλες στην κεντρική μορφή. Αφού η αλληλεπίδραση οποιονδήποτε δυο στηλών του σχεδιασμού είναι ορθογώνια στην τρίτη στήη έχουμε ότι:

$$\boxed{\sum_{i=1}^{n} x_{i1}x_{i2}x_{i3} = 0}$$

ή ισοδύναμα $\sum_{i=1}^{n}(a_i - \bar{q})(b_i - \bar{q})(c_i - \bar{q})$ όπου, $\bar{q} = \frac{(n+1)}{2}$. Επομένως, έχουμε :

$\sum_{i=1}^{n} a_i b_i c_i = \bar{q}(\sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} a_i c_i + \sum_{i=1}^{n} b_i c_i) - \bar{q}^2(\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} c_i)$

Από το Λήμμα 2 (κριτήριο ορθογωνιότητας) έχουμε:

$$\boxed{\sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n} a_i c_i = \sum_{i=1}^{n} b_i c_i = \frac{n(n+1)^2}{4} = n\bar{q}^2}$$

και λαμβάνοντας υπόψιν ότι: $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = \sum_{i=1}^{n} c_i = \frac{n(n+1)}{2} = n\bar{q}$ παίρνουμε ότι:

$$\sum_{i=1}^{n} a_i b_i c_i = n\bar{q}^3 = \frac{n(n+1)^3}{8}.$$

Παρατηρούμε ότι, όταν $n \equiv 4 mod 8$, το δεξίο μέρος της τελευταίας ισότητας δεν είναι ακέραιος, και επομένως η τελευταία ισότητα δεν ισχύει.

Θεώρημα 2.1.3: Ένας δεύτερης τάξης OLHD με n γραμμές και $k = 4$ στήλες, στον οποίο η αλληλεπίδραση οποιονδήποτε δυο στηλών είναι ορθογώνια με την αλληλεπίδραση των υπολοίπων δύο στηλών, δεν υπάρχει όταν το n είναι άρτιος και δεν είναι πολλαπλάσιο του 16.

Απόδειξη: Υποθέτουμε ότι ένας τέτοιος σχεδιασμός υπάρχει και συμβολίζουμε τις 4 στήλες του:

$$\begin{bmatrix} a_1 & b1 & c1 & d1 \\ a_2 & b2 & c2 & d2 \\ ... & ... & ... & ... \\ a_n & bn & cn & dn \end{bmatrix}$$

Έστω ακόμα ότι x_1, x_2, x_3, x_4 είναι η αντίστοιχες στήλες στην κεντρική μορφή.
Αφού η αλληλεπίδραση οποιονδήποτε δύο στηλών είναι ορθογώνια με την αλληλεπίδραση των υπόλοιπων δυο στηλών, θα έχουμε:

$$\sum_{i=1}^{n} x_{i1} x_{i2} x_{i3} x_{i4} = 0 \Rightarrow \sum_{i=1}^{n} (a_i - \bar{q})(b_i - \bar{q})(c_i - \bar{q})(d_i - \bar{q}) = 0$$

Επομένως έχουμε
$\sum_{i=1}^{n} a_i b_i c_i d_i = \bar{q}(\sum_{i=1}^{n} a_i b_i c_i + \sum_{i=1}^{n} a_i b_i d_i + \sum_{i=1}^{n} a_i c_i d_i + \sum_{i=1}^{n} b_i c_i d_i) - \bar{q}^2(\sum_{i=1}^{n} a_i b_i +$

$\sum_{i=1}^{n} a_i c_i + \sum_{i=1}^{n} a_i d_i + \sum_{i=1}^{n} b_i c_i + \sum_{i=1}^{n} b_i d_i + \sum_{i=1}^{n} c_i d_i) + \bar{q}^2(\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} c_i) -$
$n\bar{q}^4$

Αλλά έχουμε:

$$\sum_{i=1}^{n} a_i b_i c_i = \sum_{i=1}^{n} a_i b_i d_i = \sum_{i=1}^{n} b_i c_i d_i = n\bar{q}^3 = \frac{n(n+1)^3}{8}$$

και

$$\sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n} a_i c_i = \sum_{i=1}^{n} a_i d_i = \sum_{i=1}^{n} b_i c_i = \sum_{i=1}^{n} b_i d_i = \sum_{i=1}^{n} c_i d_i = n\bar{q}^2 = \frac{n(n+1)^2}{2}$$

και

$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} d_i = \frac{n(n+1)}{2} = n\bar{q}$,

Οπότε έχουμε:

$\sum_{i=1}^{n} a_i b_i c_i d_i = \bar{q}(4n\bar{q}^3) - \bar{q}^2(6n\bar{q}^2) + \bar{q}^3(4n\bar{q}) - n\bar{q}^4 = 2n\bar{q}^4 - n\bar{q}^4 = n\bar{q}^4 = \boxed{\frac{n(n+1)^4}{16}}.$

Αν το n είναι άρτιος και το $n \neq 0 mod 16$ μπορούμε, να επαληθεύσουμε ότι το δεξίο μέρος της τελευταίας ισότητας δεν είναι ακέραιος, και επομένως η τελευταία ισότητα δεν ισχυεί.

Ορισμός 2.1.3: Ένας LHD λέγεται συμμετρικός αν για κάθε γραμμή d, $-d$ είναι επίσης μια γραμμή του σχεδιασμού.

Παράδειγμα 1: Οθογώνιος Λατινικός Υπερκύβος (Orthogonal Latin Hypercube, OLH) $(5, 2)$

$$\begin{bmatrix} 1 & -2 \\ 2 & 1 \\ 0 & 0 \\ -1 & 2 \\ -2 & -1 \end{bmatrix}$$

Ο 5×2 ορθογώνιος LHD είναι επίσης και συμμετρικος. Συμετρικός Οθογώνιος Λατινικός Υπερκύβος (Symmetric Orthogonal Latin Hypercube, SOLH) (5,2)

Παράδειγμα 2: $OLH(17, 4)$

$$\begin{bmatrix}
1 & 3 & 5 & 8 \\
3 & -1 & -8 & 5 \\
5 & -8 & 3 & 1 \\
8 & 5 & 1 & -3 \\
6 & 2 & -4 & 7 \\
2 & -6 & -7 & -4 \\
7 & -4 & 6 & -2 \\
4 & 7 & -2 & 6 \\
0 & 0 & 0 & 0 \\
-1 & -3 & -5 & -8 \\
-3 & 1 & 8 & -5 \\
-5 & 8 & -3 & -1 \\
-8 & -5 & -1 & 3 \\
-6 & -2 & 4 & -7 \\
-2 & 6 & 7 & 4 \\
-7 & 4 & -6 & 2 \\
-4 & -7 & 2 & 6
\end{bmatrix}$$

Ο 17×4 ορθογώνιος LHD είναι επίσης και συμμετρικός. SOLH (17,4)

2.2 Βασικά χαρακτηριστικα της μεθόδου των Λατινικών Υπερκύβων

Η δειγματοληψία Λατινικού Υπερκύβου (McKay, Conover και Beckman, (1979)) είναι μια μέθοδος δειγματοληψίας που μπορεί να χρησιμοποιηθεί για την παραγωγή τιμών εισόδου για την εκτίμηση των προσδοκιών συναρτήσεων μεταβλητών εξόδου. Λαμβάνεται η ασυμπτωτική διακύμανση μιας τέτοιας εκτίμησης. Αποδεικνύεται επίσης ότι η εκτίμηση είναι ασυμπτωτικά κανονική. Ασυμπτωτικά, η διακύμανση είναι μικρότερη από εκείνη που προκύπτει με απλή τυχαία δειγματοληψία, με το βαθμό μείωσης της διακύμανσης να εξαρτάται από το βαθμό προσθετικότητας της συνάρτησης που ολοκληρώνεται.

Ένα βασικό ερώτημα που προέκυψε με την ενασχόληση των ερευνητών με τα υπολογιστικά πειράματα (computer experiments) είναι το ακόλουθο: Θα μπορούσαν οι μέθοδοι δειγματοληψίας και μοντελοποίησης να εφαρμοστούν σε υπολογιστικά πειράματα· Ιδανικά, και ο πειραματικός σχεδιασμός και οι στρατηγικές μοντελοποίησης θα αναπτύσσονταν για να απευθύνονται στις ιδιαιτερότητες των υπολογιστικών πειραμάτων. Οι κλασσικοί πειραματικοί σχεδιασμοί χρησιμοποιούντα για την αντιμετώπιση μη-ντετερμινιστικών και σχετικά χαμηλών διαστάσεων φυσικών πειραμάτων. Σε τέτοιες περιπτώσεις, ιδέες όπως η τυχαιοποίηση έχουν

νόημα, και οι υποθέσεις του εκάστοτε μοντέλου μας κατευθύνουν στην τοποθέτηση των σημείων (points placement). Όμως, για τα υπολογιστικά πειράματα, μια επιθυμητή τεχνική δειγματοληψίας θα έπρεπε να είναι αρκετά προσαρμοστική ώστε:

- να παρέχει δεδομένα για τεχνικές μοντελοποίησης βασισμένες σε ετερόκλιτες στατιστικές υποθέσεις η καθεμία.
- να είναι ικανή να καλύπτει από μικρούς έως πολύ μεγάλους χώρους σχεδιασμού (χωρίς περιορισμούς πυκνότητας και τοποθεσίας δεδομένων).

Οι LHD εκπληρώνουν και τις δύο παραπάνω προϋποθέσεις. Ένας άλλος πιθανός λόγος που είναι τόσο διάσημη αυτή η μέθοδος σχεδιασμού είναι το γεγονός ότι οι Σχεδιασμοί Λατινικού Υπερκύβου μπορούν να βελτιωθούν χωρίς να χρειάζεται να λάβουμε υπόψη τις στατιστικές υποθέσεις του μοντέλου και μπορούμε να έχουμε στον σχεδιασμό όσα σημεία μπορεί κανείς να διαθέσει.

Ένας επιπλέον λόγος της χρησιμότητας των LHD έχει να κάνει με το ιστορικά δυσεπίλυτο πρόβλημα βελτιστοποίησης του σχεδιασμού πολύπλοκων περιπτώσεων μοντελοποίησης όπως η Gaussian process (Μια γκαουσιανή είναι μια στατιστική κατανομή X_t, $t \ T$, για την οποία κάθε πεπερασμένος γραμμικός συνδυασμός δειγμάτων ακολουθεί την πολυμεταβλητή Κανονική Κατανομή).

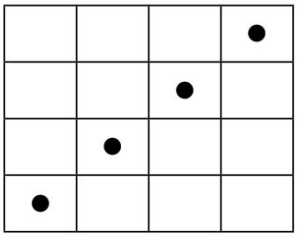
Μία άλλη βασική αιτία της δημοφιλίας της μεθόδου είναι η προσαρμοστικότητά της. Για παράδειγμα αν για κάποιο λόγο κάποιες διαστάσεις έπρεπε να εξαιρεθούν, ο τελικός σχεδιασμός που θα προέκυπτε παραμένει LHD (όχι ισοδύναμος με τον αρχικό αλλά και πάλι LHD). Άρα, αν δεν μπορούμε (οικονομικώς) να διαθέσουμε κι άλλο σύνολο δεδομένων για να γίνει εξαρχής κανονικός σχεδιασμός για κάποιον μικρότερο τομέα, τα υπάρχοντα δεδομένα μπορούν να ξαναχρησιμοποιηθούν χωρίς να είναι αναγκαία η μείωση των δειγματικών σημείων. Αυτό δε μπορεί να συμβεί σε μεθόδους όπως οι κεντρικοί σύνθετοι (central composite designs) και οι παραγοντικοί σχεδιασμοί (factorial designs). Σε αυτές τις μεθόδους, εφόσον εξαλειφθούν κάποιες διαστάσεις, τα σημεία «καταρρέουν» το ένα πάνω στο άλλο. Δυστυχώς το παραπάνω σκεπτικό δε μπορεί να εφαρμοστεί στην περίπτωση που προσθέσουμε νέες μεταβλητές στο πρόβλημα. Όταν προσθέτουμε μία ή περισσότερες διαστάσεις στο πρόβλημα αυτό πιθανότατα θα σήμαινε ότι τα σημεία του υπάρχοντος σχεδιασμού είχαν αυτές τις νέες διαστάσεις σε κάποιο συγκεκριμένο επίπεδό τους. Αυτό καταλύει την ιδιότητα της μη-κατάρρευσης των LHD. Σε αυτές τις περιπτώσεις το καλύτερο θα ήταν να χρησιμοποιήσουμε βελτιστοποίηση για να αποφανθούμε που θα έπρεπε να τοποθετηθούν τα νέα σημεία του σχεδιασμού.

Οι LHD δεν απεικονίζουν αυτόματα την ιδιότητα πλήρωσης χώρου. Ωστόσο, η εφαρμογή συγκεκριμένων κριτηρίων (αντικειμενικών συναρτήσεων) κατά τη διαδικασία σχεδιασμού διασφαλίζει ότι οι σχεδιασμοί λατινικού υπερκύβου εμφανίζουν αποτελέσματα πλήρωσης χώρου. Χαρακτηριστικό παράδειγμα LHD που δεν απεικονίζει αυτόματα την ιδιότητα πλήρωσης του

χώρου είναι το κάτωθι LHD(4,2). Θεωρούμε το

$$M = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \end{bmatrix}$$

και απεικονίζοντας τα παραπάνω σημεία παίρνουμε το παρακάτω σχεδιασμό ο οποίος είναι απόλυτα συσχετισμένος και δεν γεμίζει καλά τον χώρο. Τα σημεία πάνω στην διαγώνιο είναι πολύ κακά και έχουμε ταυτόσημες στήλες

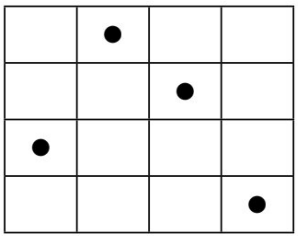


Σχήμα 2.1: Σχεδιασμός με πολύ φτωχές ιδιότητες πλήρωσης χώρου

Αντίθετα, παράδειγμα LHD που απεικονίζει αυτόματα την ιδιότητα πλήρωσης του χώρου είναι το κάτωθι LHD(4,2).Θεωρούμε το

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 3 \\ 4 & 1 \end{bmatrix}$$

και απεικονίζοντας τα παραπάνω σημεία παίρνουμε το παρακάτω σχεδιασμό όπου κάθε γραμμή και στήλη έχει ένα και μόνο ένα σημείο. Κάθε παράγοντας έχει 4 επίπεδα.

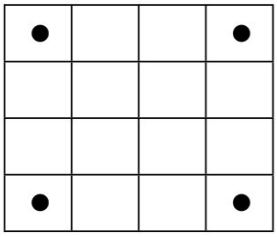


Σχήμα 2.2: Σχεδιασμός με καλές ιδιότητες πλήρωσης χώρου

Από την άλλη πλευρά, ένας παραγοντικός σχεδιασμός 2^2:

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 4 & 1 \\ 4 & 4 \end{bmatrix}$$

Σχεδιάζεται όπως φαίνεται στο κάτωθι σχήμα. Εάν το x_1 (ή το x_2) δεν είναι σημαντικό, η επανάληψη είναι άσκοπη για τα πειράματα υπολογιστών. Αρχή της αραιότητας των αποτελεσμάτων: μόνο λίγοι παράγοντες αναμένεται να είναι σημαντικοί.



Σχήμα 2.3: Τυχαιοποιημένος σχεδιασμός

2.3 Σύγκριση της LHS με άλλες μεθόδους δειγματοληψίας

Για να αναδειχτεί καλύτερα το πλεονέκτημα των LHD έναντι κάποιων άλλων γνωστών μεθόδων δειγματοληψίας (απλής τυχαίας και στρωματοποιημένης) θα χρειαστούμε τους ακόλουθους ορισμούς:

Ορισμός 2.3.1: Έστω ότι έχουμε N μονάδες ενός πληθυσμού και επιθυμούμε τυχαίο δείγμα $N_1, N_2, ..., N_n$ μεγέθους n. Η διαδικασία επιλογής ενός δείγματος από τα $\binom{N}{n}$ δείγματα ονομάζεται απλή τυχαία δειγματοληψία, αν κάθε ένα από αυτά έχει πιθανότητα ίση με $\frac{1}{\binom{N}{n}}$ να επιλεγεί.

Ορισμός 2.3.2: Έστω πληθυσμός μεγέθους N και έστω ότι αυτός μπορεί να διαιρεθεί σε k εσωτερικά ομοιογενείς υποπληθυσμούς μεγέθους $N_1, N_2, .., N_k$. Αν αυτοί είναι ξένοι μεταξύ τους ώστε να ισχύει ότι $N_1 + N_2 + .. + N_k = N$ (διαμέριση του πληθυσμού), οι υποπληθυσμοί αυτοί ονομάζονται στρώματα (strata).

Ορισμός 2.3.3: Έστω ότι από καθένα από τα στρώματα ενός πληθυσμού επιλέγεται ένα απλό τυχαίο δείγμα μεγέθους $n_i, i = 1, 2, \ldots, k$ ανεξάρτητα από τα άλλα. Το δείγμα μεγέθους $n = n_1 + n_2 + ... + n_k$ που προκύπτει από την ένωση των k ανεξάρτητων απλών τυχαίων δειγμάτων ονομάζεται Στρωματοποιημένο Τυχαίο Δείγμα (Stratified Random Sampling) και η διαδικασία επιλογής του ονομάζεται Στρωματοποιημένη Τυχαία Δειγματοληψία (Stratified

Random Sampling, SRS).

Ένας λόγος της δημοφιλίας των LHD αποδίδεται στη θεωρητική απόδειξη (McKay, Conover και Beckman (1979)) για τη μείωση της διασποράς κατά την αριθμητική ολοκλήρωση.

Έστω μια ντετερμινιστική συνάρτηση $Y = f(X)$, όπου f γνωστή, και η X έχει ομοιόμορφη κατανομή με τιμές από τον υπερκύβο $[0,1)^m$ και $Y \in R$ Θέλουμε την αναμενόμενη τιμή του Y, μ= $E(Y)$. Όταν δεν μπορούμε να υπολογίσουμε με αναλυτικό τρόπο το μ αναγκαστικά θα οδηγηθούμε σε προσεγγιστικές μεθόδους. Η απλούστερη μέθοδος είναι να δημιουργήσουμε ανεξάρτητα $X_1, X_2, \ldots, X_n$ από την ομοιόμορφη κατανομή στο $[0,1)^m$ και να εκτιμήσουμε το μ χρησιμοποιώντας την εκτιμήτρια:

$$\hat{\mu}_{srs} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Αν τώρα η $f(X) = X$ τότε αυτή είναι η εκτιμήτρια της μέσης τιμής του πληθυσμού. Αυτή θεωρούμε ότι είναι η εκτιμήτρια του μγια απλή τυχαία δειγματοληψία. Ας συμβολίσουμε με $\hat{\mu}_{srs}$ την εκτιμήτρια του μ στρωματοποιημένης δειγματοληψίας και έστω για απλότητα ότι $f(X) = X$ (άρα εκτιμούμε τον γενικό μέσο μ του πληθυσμού). Αυτή θεωρούμε ότι είναι η εκτιμήτρια του μγια απλή τυχαία δειγματοληψία. Ας συμβολίσουμε με $\hat{\mu}_{srs}$ την εκτιμήτρια του μστρωματοποιημένης δειγματοληψίας και έστω για απλότητα ότι $f(X) = X$ (άρα εκτιμούμε τον γενικό μέσο μτου πληθυσμού).

Έστω y_{ij} η τιμή του χαρακτηριστικού που μας ενδιαφέρει για τη μονάδα j του στρώματος i και X_{ij} η τιμή του χαρακτηριστικού της μονάδας j του δείγματος από το στρώμα i. Βάσει όσων έχουμε αναφέρει παραπάνω για την τυχαία στρωματοποιημένη δειγματοληψία, αν συμβολίσουμε ως $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} U_{ij}$ τον μέσο του i-οστού στρώματος, $i = 1, \ldots, k$ και μτον γενικό μέσο τότε θα ισχύει ότι:

$$\mu = \frac{1}{N}\sum_{i=1}^{k} N_i\mu_i.$$

Οι εκτιμήτριες κάθε στρωματικού μέσου μ_i θα είναι οι $\bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$ όπου συμβολίζουμε με X_{ij} την j-οστή μονάδα του i-οστού δείγματος. Αν συμβολίσουμε με $W_i = \frac{N_i}{N}$ $i = 1, \ldots, k$ το βάρος του κάθε στρώματος τότε μπορούμε να ορίσουμε την $\hat{\mu}_{STR}$ ως εξής: $\hat{\mu}_{STR} = \sum_{i=1}^{k} W_i\bar{X}_i$

Πρόταση 2.3.1: Ο εκτιμητής $\hat{\mu}_{STR}$ είναι αμερόληπτος εκτιμητής του μέσου μ, δηλαδή $E(\hat{\mu}_{STR})$ =μ, αν και μόνο αν $E(X) = \mu_i$ για κάθε $i = 1, 2, \ldots, k$

Ορισμός 2.3.4: Η αναλογική μέθοδος καταμερισμού στην περίπτωση στρωματοποιημένης

δειγματοληψίας είναι μια μέθοδος για τον προσδιορισμό των επιμέρους δειγματικών μεγεθών n_i λύνοντας το σύστημα εξισώσεων: $\{\frac{n_1}{n} = \frac{N_1}{N}, \frac{n_2}{n} = \frac{N_2}{N}, ..., \frac{n_k}{n} = \frac{N_k}{N}\}$. Προφανώς αν ξέρουμε το μέγεθος του πληθυσμού N και το μέγεθος του δείγματος n καθώς επίσης και τα μεγέθη $N_i, i = 1, ..., k$ των επιμέρους στρωμάτων μπορώ να προσδιορίσω κατά μοναδικό τρόπο τα μεγέθη n_i.

Με βάσει όλους τους παραπάνω ορισμούς μπορούμε να διατυπώσουμε το θεώρημα που συσχετίζει τις διασπορές των εκτιμητριών σε απλή τυχαία και τυχαία στρωματοποιημένη δειγματοληψία.

Θεώρημα 2.3.1: Αν θεωρήσουμε τις περιπτώσεις απλής τυχαίας δειγματοληψίας και αναλογικής στρωματοποιημένης δειγματοληψίας και επιπλέον υποθέσουμε ότι οι όροι $\frac{1}{N_i}$ είναι αμελητέοι τότε ισχύει:

$$Var(\hat{\mu}_{STR}) \leq Var(\hat{\mu}_{SRS})$$

Απόδειξη 2.3.1: Αρχικά ορίζουμε ως $f_i = \frac{n_i}{N_i} i = 1, \ldots, k$ το ποσοστό δειγματοληψίας για το στρώμα i στην περίπτωση της στρωματοποιημένης δειγματοληψίας και $f = \frac{n}{N}$ το ποσοστό δειγματοληψίας για την απλή τυχαία δειγματοληψία. Στην περίπτωση αναλογικής κατανομής ισχύει ότι: $\{\frac{n_1}{n} = \frac{N_1}{N}, \frac{n_2}{n} = \frac{N_2}{N}, ..., \frac{n_k}{n} = \frac{N_k}{N}\}$ ή ισοδύναμα $f_i = f \ \forall i = 1, ..., k$.

Επιπλέον, θέοτυμε ως: $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (U_{ij} - \mu)^2$ τη διασπορά του πληθυσμού και $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (U_{ij} - \mu)^2$ τη πληθυσμιακή διασπορά του στρώματος i.

Για ευκολία θα θεωρήσουμε την περίπτωση του προβλήματος όπου $f(X) = X$ άρα οι εκτιμήτριες $\hat{\mu}_{STR}$ και $\hat{\mu}_{SRS}$ είναι εκτιμήτριες του πληθυσμιακού μέσου μμε τισ δυο αντίστοιχες μεθόδους δειγματοληψίας.

Τέλος, θέτουμε χάριν συντομίας: $V_{SRS} = Var(\hat{\mu}_{SRS})$ και $V_{STR} = Var(\hat{\mu}_{STR})$

$V_{SRS} = Var(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} (\sum_{i=1}^{n} Var(X_i) + 2 \sum_{i<j} \sum Cov(X_i, X_j)) =$

$\frac{1}{n^2} (n\sigma^2 - 2 \sum_{i<j} \sum \frac{\sigma^2}{N-1}) = \frac{1}{n^2} (n\sigma^2 - \frac{2}{N-1} \frac{n(n-1)}{2} \sigma^2) = \frac{\sigma^2}{n} (\frac{N-n}{N-1}) = (\frac{1-f}{n} S^2)$ (1)

και

$V_{SRS} = Var(\sum_{i=1}^{k} W_i \bar{X}_i) = \sum_{i=1}^{k} W_i^2 Var(\bar{X}_i) = \sum_{i=1}^{k} W_i^2 \frac{S_i^2}{n_i} (1 - f_i)$

και επειδή στην περίπτωση της αναλογικής κατανομής έχουμε $n_i = \frac{nN_i}{N}$ θα πάρουμε ότι:

$V_{SRS} = \frac{1-f}{n} \sum_{i=1}^{k} W_i^2 S_i^2 = \sum_{i=1}^{k} \frac{W_i S_i^2}{n} - \sum_{i=1}^{k} \frac{W_i S_i^2}{N}$ (2)

Μπορούμε να γράψουμε ότι:

$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(U_{ij}-\mu)^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(U_{ij}-\mu_i)^2 + \sum_{i=1}^{k} N_i(\mu_i-\mu)^2 =$

$\sum_{i=1}^{k}(N_i-1)S_i^2 + \sum_{i=1}^{k} N_i(\mu_i-\mu)^2$

Και επειδή λόγω της υπόθεσης μπορούμε να αγνοήσουμε τους όρους $\frac{1}{N}$ και $\frac{1}{N_i}$ καταλήγουμε ότι:

$$S^2 = \sum_{i=1}^{k} W_i S_i^2 + \sum_{i=1}^{k} W_i(\mu_i-\mu)^2 \quad (3)$$

Τέλος, η (1) λόγω της (3) γίνεται:

$$V_{SRS} = \frac{1-f}{n}S^2 = \frac{1-f}{n}\sum_{i=1}^{k} W_i^2 S_i^2 + \frac{1-f}{n}\sum_{i=1}^{k} W_i(\mu_i-\mu)^2$$

και με την βοήθεια της (2):

$$V_{SRS} = V_{STR} + \frac{1-f}{n}\sum_{i=1}^{k} W_i(\mu_i-\mu)^2 \Rightarrow V_{SRS} \leq V_{STR}$$

Αν συμβολίσουμεσε αυτό το σημείο με $\hat{\mu}_{LHS}$ την εκτιμήτρια του μστο αρχικό μας πρόβλημα υπό την δειγματοληψία Λατινικού Υπερκύβου ισχύει το ακόλουθο θεμελιώδες θεώρημα.

Θεώρημα 2.3.2: Αν η σ.π.π. $f(x)$, $x \in S$ του τ.δ. X είναι μονότονη συνάρτηση ως προς κάθε μεταβλητή της, τότε ισχύει $Var(\hat{\mu}_{LHS}) \leq Var(\hat{\mu}_{SRS})$

Παρατήρηση 2.3.1: Έστω οτι X είναι μια κ-διάστατη τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας $f(x)$, $x \in S$.
Στην στρωματοποιημένη δειγματοληψία το δείγμα χωρίζεται με αυθαίρετο τρόπο σε στρώματα ενώ στην δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων η διαμέριση κατασκευάζεται με συγκεκριμένο τρόπο χρησιμοποιώντας διαμερίσεις του εύρους της κατανομής κάθε επιμέρους μεταβλητής X_i του X. Εξετάζουμε την περίπτωση που όλα τα X_i ειναι ανεξάρτητα μεταξύ τους.
Έστω ότι χωρίζουμε το εύρος καθενός απο τα $X_i, i = 1, ..., k$ σε N το πλήθος διαστήματα μεγέθους πιθανότητας $\frac{1}{N}$.
Το καρτεσιανό γινόμενο αυτών των διαστημάτων απαρτίζεται απο N^k κελιά το καθένα μεγέθους πιθανότηταςαπο N^{-k}.
Κάθε κελί μπορεί να παρασταθεί από ένα διάνυσμα κ συντεταγμένων $m_i = (m_{i1}, m_{i2}, ..., m_{ik})$, όπου m_{ij} είναι ο αριθμός του διαστήματος της μεταβλητής X_j στο i-οστό κελί.

Ένα δείγμα λατινικού υπερκύβου μεγέθους N, κατασκευάζεται από την τυχαία επιλογή N το πλήθος τέτοιων κελιών m_i με την προυπόθεση ότι για κάθε ϑ το σύνολο $\{m_{ij}\}_{i=1}^{N}$ είναι μια μετάυεση των ψηφίων 1,...,N. Παίρνουμε μια τυχαία παρατήρηση από κάθε κελί που επιλέξαμε τυχαία. Αν ϑεωρήσω τις δείκτριες μεταβλητές όπου $w_i = 1$ αν το κελί i βρίσκεται στο δείγμα ή 0 διαφορετικά έχουμε μια εκτιμήτρια του πληθυσμιακού μέσου μ:

$$\hat{\mu}_{LHS} = \frac{1}{N}\sum_{i=1}^{N^k} W_i X_i$$

Απόδειξη 2.3.2: Έχοντας την παραπάνω εκτιμήτρια ϑα ισχύει ότι:

$$Var(\hat{\mu}_{LHS}) = \frac{1}{N^2}\sum_{i=1}^{N^k} Var(w_i X_i) + \frac{1}{N^2}\sum_{i=1}^{N^k}\sum_{j=1_{j\neq i}}^{N^k} Cov(w_i X_i, w_j X_j) \quad (4)$$

Για τα w_i αποδεικνύονται εύκολα οι ακόλουθες σχέσεις:

- $P(w_i = 1) = \frac{1}{N^{k-1}} = E(w_i) = E(w_i^2)$ άρα η διασπορά ϑα είναι: $Var(w_i) = E(w_i^2) - E^2(w_i) = \frac{1}{N^{k-1}}(1 - \frac{1}{N^{k-1}})$

- Αν τα w_i και w_j αντστοιχούν σε κελία χωρίς καμία κοινή συντεταγμένη κελιού, τότε: $E(w_i w_j) = E(w_i w_j | w_j = 0)P(w_j = 0) + E(w_i w_j | w_j = 1) = \frac{1}{(N(N-1))^{k-1}}$

- Αν τα w_i και w_j αντστοιχούν σε κελία που έχουν τουλάχιστον μια κοινή συντεταγμένη, τότε $E(w_i w_j) = 0$

Υπολογίζουμε τώρα την ποσότητα: $Var(w_i X_i) = E(w_i^2)Var(X_i) + E^2(X_i)Var(w_i)$

Από το πρώτο μέρος της σχέσεης (4) είναι:

$$\sum_{i=1}^{N^k} Var(w_i X_i) = N^{-k+1}\sum_{i=1}^{N^k} E(X_i - \mu_i)^2 + (N^{-k+1} - N^{-2k+2})\sum_{i=1}^{N^k}\mu_i^2$$

όπου συμβολίζουμε ως μ_i τον μέσο του κελιού i.

Εφόσον ισχύει ότι: $E(X_i - \mu_i)^2 = N^k \int_i (x-\mu)^2 f(x)dx + (\mu_i - \mu)^2$

Άρα,

$$\sum_{i=1}^{N^k} Var(w_i X_i) = NVar(X) - N^{-k+1}\sum_{i=1}^{N^k}(\mu_i - \mu)^2 + (N^{-k+1} - N^{-2k+2})\sum_{i=1}^{N^k}\mu_i^2 \quad (5)$$

Και επιπλέον,

$$\sum_{i=1}^{N^k}\sum_{j=1_{j\neq i}}^{N^k} Cov(w_iX_i, w_jX_j) = \sum_{i\neq j}\sum \mu_i\mu_j E(w_iw_j) - N^{-2k+2}\sum_{i\neq j}\sum \mu_i\mu_j \quad (6)$$

Η (4) λόγω της (5), (6) μας δίνει τελικά:

$Var(\hat{\mu}_{LHS}) = \frac{1}{N}Var(X) - N^{-k-1}\sum_{N^k}^{i=1}(\mu_i-\mu)^2 + (N^{-k-1}-N^{-2k})\sum_{i=1}^{N^k}\mu_i^2 +$

$(N-1)^{-k+1}N^{k-1}\sum_R\sum\mu_i\mu_j - N^{-2k}\sum_{i\neq j}\sum\mu_i\mu_j$

όπου R συμβολίζουμε το σύνολο των $N^k(N-1)^k$ το πλήθος ζευγών (μ_i, μ_j) που αντιστοιχούν στα κελία με καία κοινή συντεταγμένη. Μετά από μερικές πράξεις και ξέροντας ότι: $\sum_{i=1}^{N^k} = N^k\mu$ η τελική μορφή της διασποράς θα είναι:

$$Var(\hat{\mu}_{LHS}) = Var(\hat{\mu}_{SRS}) + \frac{N-1}{N}(N^{-k}(N-1)\sum_R(\mu_i-\mu)(\mu_j-\mu)) \quad (7)$$

Από την (7) γίνεται κατανοήτό ότι:

$$Var(\hat{\mu}_{LHS}) \leq Var(\hat{\mu}_{SRS}) \Leftrightarrow \sum_R(\mu_i-\mu)(\mu_j-\mu) \leq 0. \quad (8)$$

Η σχέση (8) είναι ισοδύναμη με το ότι η συνδιασπορά μεταξύ των κελιών που δεν έχουν καμία κοινή συντεταγμένη είναι αρνητική (κάτι που εξασφαλίζεται από την υπόθεση της μονοτονίας της $f(x)x \in S$ ως προς κάθε μεταβλητή της μέσω του λήμματος του $Lehmann$).

Το Θεώρημα 3.2 μας λέει ότι υπό την ισχύ της μονοτονίας της συνάρτησης U η δειγματοληψία με την μέθοδο των Λατινικών Υπερκύβων είναι καλύτερη της απλής τυχαίας υπό την έννοια της ελάττωσης της διασποράς της εκτιμήτριας του δειγματικού μέσου.

2.4 Ορθογώνιοι Σχηματισμοί

Ένας Ορθογώνιος Σχηματισμός $OA(n, q, s, t)$ είναι ένας $n \times q$ πίνακας με στοιχεία επιλεγμένα από ένα σύνολο s διακεκριμένων συμβόλων διατεταγμένων έτσι ώστε, για κάθε επιλογή t στηλών του σχεδιασμού, καθένα απο τα s^t διανύσματα γραμμών να εμφανίζεται το ίδιο συχνά. Προφανώς, το s^t διαιρεί το n. Καλούμε n τον αριθμό των εκτελέσεων (παρατηρήσεων) του σχεδιασμού, q τον αριθμό των παραγόντων, s τον αριθμό των στάθμεων (επιπέδων) κάθε παράγοντα και, t την ισχύ strength του σχεδιασμού. Σε σχέση με τους παραγοντικούς σχεδιασμούς, κάθε $OA(n, q, s, t)$ ορίζει έναν κλασματικό παραγοντικό σχεδιασμό με n εκτελέσεις, q παράγοντες με s στάθμες ο καθένας και με ισχύ t. Η έννοια της ισχύος είναι άμεσα συνδεδεμένη με την αναλυτική τάξη του σχεδιασμού. Συγκεκριμένα, ένας ορθογώνιος σχηματισμός με ισχύ t ορίζει έναν κλασματικό παραγοντικό σχεδιασμό αναλυτικής τάξης $t+1$.

Ίδιο συχνά: αυτό ονομάζεται *index* $\lambda, (\lambda = \frac{n}{s^t})$

Παράδειγμα 1: $OA(4, 3, 2, 2)$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Έχουμε: $\lambda = \frac{n}{s^t} = \frac{4}{2^2} = 1$

Θεώρημα 2.4.1 Ανισότητες Rao Οι παράμετροι ενός $OA(n, q, s, t)$ ικανοποιούν τις κάτωθι ανισότητες για $u \leq 0$:

$n \geq \sum_{i=0}^{n} \binom{q}{i}(s-1)^i$ εάν $t = 2u$.

$n \geq \sum_{i=0}^{n} \binom{q}{i}(s-1)^i + \binom{q-1}{u}(s-1)^{u+1}$ εάν $t = 2u+1$.

Παράδειγμα 2: $OA(8, 4, 2, 3)$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Έχουμε: $\lambda = \frac{n}{s^t} = \frac{8}{2^3} = 1$. Επίσης, $n = 8$, $t = 3$: $q \leq \frac{(\frac{n}{s}-1)}{s-1} + 1 = \frac{\frac{8}{2}-1}{2+1} + 1 = 4$.

Παράδειγμα 3: $OA(8, 5, 2, 2)$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Έχουμε: $\lambda = \frac{n}{s^t} = \frac{8}{2^2} = 2$. Επίσης, $n = 8$, $t = 2$: $q \leq \frac{n-1}{s-1} = 7$

Παρόλο που οι ορθογώνιοι σχηματισμοί είναι ιδιαίτερα σημαντικοί δεν θα αναφερθούμε εκτενώς σε αυτούς στην συγκεκριμένη εργασία. Αξίζει να σημειωθεί ότι οι Georgiou et al.(2014) και Georgiou et al.(2018) κατασκευασαν σχεδιασμούς με καλές ιδιότητες πλήρωσης του χώρου που είναι πολύ χρήσιμοι στις εφαρμογές και παρέχουν ευελίξία τόσο για τον αριθμό των παρατηρήσεων (γραμμές) όσο και για τον αριθμό των παραγόντων(στήλες).

Κεφάλαιο 3

Λατινικοί Υπερκύβοι και το πρόβλημα των Μεγάλων Δεδομένων

Συνόψίζοντας λοιπόν όσα έχουμε παρουσιάσει παραπάνω μια διέξοδο στη σύγχρονη μηχανική μάθηση για το πρόβλημα των μεγάλων δεδομένων αποτελεί η δειγματοληψία με την μέθοδο των λατινικών υπερκύβων. Συγκεκριμένα, στα παρακάτω πειραματικά στάδια θα αποδειχθεί ότι εάν κάνουμε δειγματοληψία στα δεδομένα μας με την μέθοδο των λατινικών υπερκύβων η απώλεια που θα έχουμε στα αποτελέσματα μας θα είναι ελάχιστή.

3.1 Η βιβλιοθήκη cLHS

Το κύριο εργαλείο που θα χρησιμοποιήθει για τον πειραματικό αυτόν έλεγχο θα είναι η γλώσσα προγραμματισμού R.

Τι είναι η γλώσσα προγραμματισμού R;

- Είναι μία γλώσσα προγραμματισμού και περιβάλλον προγραμματισμού, εξειδικευμένο για υπολογισμούς στατιστικής φύσεως και οπτικοποίησης δεδομένων.
- Χαρακτηρίζεται και ως στατιστική γλώσσα προγραμματισμού δηλαδή σημαίνει ότι υποστηρίζει πάρα πολύ καλά όλα τα βήματα στη διαδικασία της στατιστικής ανάλυσης δεδομένων, τα οποία μπορούν να υλοποιηθούν (δηλαδή να προγραμματιστούν) με ακρίβεια, ευκολία και ταχύτητα.
- Έχει σχεδιαστεί με στόχο να κάνει πολύ εύκολα τα εξής: την ενσωμάτωση/ανάγνωση δεδομένων, τον καθαρισμό των δεδομένων (προεπεξεργασία) όσο περίπλοκος κι αν είναι αυτός, την εφαρμογή στατιστικών ελέγχων και μεθόδων πάνω στα δεδομένα, και την εξαγωγή συμπερασμάτων, τη δημιουργία και εφαρμογή στατιστικών/οικονομετρικών μοντέλων, την αξιολόγηση των στατιστικών μοντέλων και τη χρήση των στατιστικών μοντέλων για την αντιμετώπιση πραγματικών προβλημάτων.

Η κύρια βιβλιοθήκη πάνω στην οποία θα βασιστούν τα πειράματα μας είναι η cLHS: Conditioned Latin Hypercube Sampling. Κάτωθι παρουσιάζεται ο αλγόριθμος της cLHS (Budiman Minasny, Alex B. McBratney (2005)):

1. Χωρίζεται η κατανομή των ποσοστημορίων της X σε n στρώματα και υπολογίζεται η κατανομή των ποσοστημορίων για κάθε μία από τις μεταβλητές $q_j^i, ..., q_j^{n+1}, j = 1, ..., k$ και $i = 1, ..., n+1$. Υπολογίζεται το C, ο πίνακας συσχέτισης του πίνακα X.

2. Διαλέγονται n τυχαία δείγματα από το N, τα $x(i = 1, ..., n)$ είναι οι δειγματοληπτούμενες θέσεις και τα $r(i = 1, ..., N - n)$ είναι οι μη δειγματοληπτούμενες θέσεις. Υπολογίζεται το T, ο πίνακας συσχέτισης του πίνακα x.

3. Υπολογίζεται η αντικειμενική συνάρτηση. Για τις συνεχείς μεταβλητές είναι: $O_1 = \sum_i^n \sum_{j=1}^k |\eta(q_j^i \leq x_j < q_j^{i+1}) - 1|$, όπου $\eta(q_j^i \leq x_j < q_j^{i+1})$ είναι ο αριθμός των x_j που πέφτουν ανάμεσα στα ποσοστημόρια q_j^i και q_j^{n+1}. Για κατηγορικά δεδομένα, η αντικειμενική συνάρτηση είναι η κατανομή πιθανότητας για κάθε μία από τις κλάσεις $O_2 = \sum_{j=1}^c |\frac{\eta(x_j)}{n} - \kappa_j|$ όπου $\eta(x_j)$ είναι ο αριθμός των x που ανήκουν στην κλάσση j στο X. Για να εξασφαλιστεί ότι η συσχέτιση των δειγματοληπτικών μεταβλητών θα αναπαράγει τα αρχικά δεδομένα, προστίθεται μια άλλη αντικειμενική συνάρτηση: $O_3 = \sum_i^k \sum_{j=1}^k |c_{ij} - t_{ij}|$ όπου c είναι το στοιχείο του C, του πίνακα συσχέτισης του X, και t είναι το ισοδύναμο στοιχείο του T, του πίνακα συσχέτισης του x. Η συνολική αντικειμενική συνάρτηση είναι: $O = w_i O_1 + w_2 O_2 + w_3 O_3$ όπου w είναι το βάρος που δίνεται σε κάθε στοιχείο της αντικειμενικής συνάρτησης. Για γενικές εφαρμογές τοw τίθεται σε 1 για όλα τα στοιχεία της αντικειμενικής συνάρτησης.

4. Εκτελέιται ένα annealing schedule (Press et al., 1992): $Metro = exp[-\frac{\Delta O}{T}]$, όπου ΔO είναι η μεταβολή της αντικειμενικής συνάρτησης, και T είναι μια θερμοκρασία ψύξης (μεταξύ 0 και 1) η οποία μειώνεται κατά έναν παράγονταdσε κάθε επανάληψη.

5. Παράγεται ένας ομοιόμορφος αριθμός $rand$ μεταξύ 0 και 1, αν $rand < Metro$, αποδέχονται οι νέες τιμές, διαφορετικά απορρίπτονται οι αλλαγές.

6. Πραγματοποιούνται αλλαγές: Παράγεται ένας ομοιόμορφος αριθμός $rand$:
   If $rand < p$
   Επιλέγεται ένα δείγμα τυχαία από τα x και ανταλλάζεται με μια τυχαία τοποθεσία από τα μη δειγματοληπτούμα r.
   Else,
   Αφαιρούνται το(α) δείγμα(τα) από το x που έχει το μεγαλύτερο $\eta(q_j^i \leq x_j < q_j^{i+1})$ και αντικαθίστανται με τυχαίο(α) σημείο(α) από τα μη δειγματοληπτούμα r.
   End if
   Η τιμή του p κυμαίνεται μεταξύ 0 και 1 και δείχνει την πιθανότητα η αναζήτηση να είναι τυχαία ή να αντικαθιστά συστηματικά τα δείγματα που ταιριάζουν χειρότερα στα στρώματα.

7. Πήγαινε στο βήμα (3)

Επαναλαμβάνουμε τα βήματα (3)-(7) έως ότου η τιμή αντικειμενικής συνάρτηση πέφτει πέρα από ένα δεδομένο κριτήριο διακοπής ή για ένα καθορισμένο αριθμό επαναλήψεων.

Το τελικό δείγμα θα αντιπροσωπεύει έναν πραγματικό ή προσεγγιστικό λατινικό υπερκύβο του χώρου χαρακτηριστικών, όπου η κατανομή και η πολυμεταβλητή συσχέτιση θα διατηρούνται. Ο αλγόριθμος που περιγράφεται παραπάνω είναι κωδικοποιημένος στο Matlab (Mathworks, 2005). Η τεχνική αυτή έχει αρκετά πλεονεκτήματα, και συγκεκριμένα:

1. Συνεχείς καθώς και κατηγορικές μεταβλητές μπορούν να να ενσωματωθούν.
2. Η δειγματοληψία βασίζεται στην εμπειρική κατανομή των αρχικών δεδομένων, επομένως είναι μη παραμετρική.
3. Οι χωρικές συντεταγμένες μπορούν να ενσωματωθούν για να εξασφαλιστεί μια καλή διασπορά των σημείων δειγματοληψίας εάν αυτό απαιτείται και μπορούν να επιβληθούν πρόσθετοι περιορισμοί στην αντικειμενική συνάρτηση, π.χ. απόσταση από δρόμους και όρια χωραφιών.

Χρήση της βιβλιοθήκης γίνεται με την κάτωθι συνάρτηση που περιέχει συγκεκριμένα ορίσματα:

```
clhs(
  x,
  size,
  must.include,
  can.include,
  cost,
  iter,
  use.cpp,
  temp,
  tdecrease,
  weights,
  eta,
  obj.limit,
  length.cycle,
  simple,
  progress,
  track,
  use.coords,
  ...
)
```

3.2 Ταξινομητές και Σύνολα Δεδομένων

Εν συνεχεία οι δυο κύριοι ταξινομητές που θα χρησιμοποιήσουμε είναι:

- Τα Τυχαία Δάση (Random Forest)
- Οι Κ-Πλησιέστεροι Γείτονες (K-Nearest Neighbors (K-NN))

Ο Random Forest είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος Μηχανικής Μάθησης που κατοχυρώθηκε από τους Leo Breiman και Adele Cutler (στατιστικοί), ο οποίος συνδυάζει την έξοδο πολλαπλών δέντρων απόφασης για να καταλήξει σε ένα ενιαίο αποτέλεσμα. Η ευκολία χρήσης και η ευελιξία του έχουν τροφοδοτήσει την υιοθέτησή του, καθώς χειρίζεται τόσο προβλήματα ταξινόμησης όσο και παλινδρόμησης.

Δέντρα αποφάσεων (Decision Trees):
Δεδομένου ότι το μοντέλο Random Forest αποτελείται από πολλαπλά δέντρα αποφάσεων, θα ήταν χρήσιμο να ξεκινήσουμε περιγράφοντας εν συντομία τον αλγόριθμο των δέντρων αποφάσεων. Τα δέντρα αποφάσεων ξεκινούν με μια βασική ερώτηση, όπως: "Πρέπει να κάνω σερφ;". Από εκεί και πέρα, μπορούμε να θέσουμε μια σειρά ερωτήσεων για να καθορίσουμε μια απάντηση, όπως, "Είναι κύμα μεγάλης περιόδου;" ή "Φυσάει ο άνεμος στα ανοιχτά;". Αυτές οι ερωτήσεις αποτελούν τους κόμβους απόφασης στο δέντρο, λειτουργώντας ως μέσο διαχωρισμού των δεδομένων. Κάθε ερώτηση βοηθά ένα άτομο να καταλήξει σε μια τελική απόφαση, η οποία θα δηλώνεται από τον κόμβο φύλλου. Οι παρατηρήσεις που πληρούν τα κριτήρια θα ακολουθήσουν τον κλάδο "Ναι" και εκείνες που δεν πληρούν τα κριτήρια θα ακολουθήσουν την εναλλακτική διαδρομή. Τα δέντρα αποφάσεων επιδιώκουν να βρουν τον καλύτερο διαχωρισμό για την υποδιαίρεση των δεδομένων και συνήθως εκπαιδεύονται μέσω του αλγορίθμου CART (Classification and Regression Tree). Διάφορες μετρικές χρησιμοποίουνται για την αξιολόγηση των αποτελεσμάτων.

Σχήμα 3.1: Δένδρα αποφάσεων

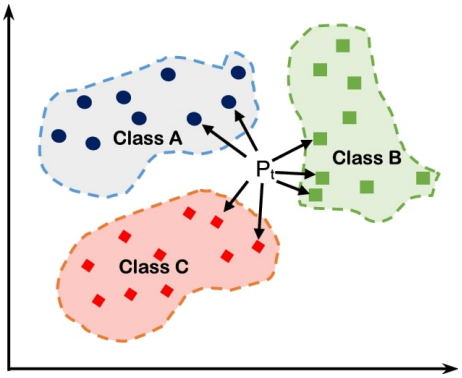
Στη στατιστική, ο αλγόριθμος Κ-Πλησιέστερων Γειτόνων (k-NN) είναι μια μη παραμετρική μέθοδος μάθησης με επίβλεψη που αναπτύχθηκε για πρώτη φορά από τους Evelyn Fix και Joseph Hodges το 1951 και αργότερα επεκτάθηκε από τον Thomas Cover (στατιστικοί).

Χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Και στις δύο περιπτώσεις, η είσοδος αποτελείται από τα k πλησιέστερα παραδείγματα εκπαίδευσης σε ένα Σύνολο Δεδομένων. Η έξοδος εξαρτάται από το αν το k-NN χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση:

- Στην ταξινόμηση k-NN, η έξοδος είναι μια ένταξη σε κλάση. Ένα αντικείμενο ταξινομείται με ψηφοφορία πληθικότητας των γειτόνων του, με το αντικείμενο να κατατάσσεται στην κλάση που είναι πιο κοινή μεταξύ των k πλησιέστερων γειτόνων του (το k είναι ένας θετικός ακέραιος αριθμός, συνήθως μικρός). Εάν $k = 1$, τότε το αντικείμενο απλώς κατατάσσεται στην κλάση αυτού του μοναδικού πλησιέστερου γείτονα.

- Στην παλινδρόμηση k-NN, η έξοδος είναι η τιμή της ιδιότητας για το αντικείμενο. Η τιμή αυτή είναι ο μέσος όρος των τιμών των κ πλησιέστερων γειτόνων.

Η k-NN είναι ένας τύπος ταξινόμησης όπου η συνάρτηση προσεγγίζεται μόνο τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι την αξιολόγηση της συνάρτησης. Δεδομένου ότι αυτός ο αλγόριθμος βασίζεται στην απόσταση για την ταξινόμηση, εάν τα χαρακτηριστικά αντιπροσωπεύουν διαφορετικές φυσικές μονάδες ή έρχονται σε πολύ διαφορετικές κλίμακες, τότε η κανονικοποίηση των δεδομένων εκπαίδευσης μπορεί να βελτιώσει δραματικά την ακρίβειά του.



Σχήμα 3.2: K-Πλησιέστεροι Γείτονες

Τέλος, τα 4 Σύνολα Δεδομένων πάνω στα οποία έγιναν τα πειράματα στο ακόλουθο κεφάλαιο είναι:

- Iris Σύνολο Δεδομένων
- Diabetes Σύνολο Δεδομένων
- Wine Σύνολο Δεδομένων
- Wifi Σύνολο Δεδομένων

Το Σύνολο Δεδομένων Iris δημοσιεύθηκε αρχικά στο UCL Machine Learning Repository. Είναι ένα μικρό Σύνολο Δεδομένων (150 γραμμές) και από το 1936 χρησιμοποιείται συχνά για τη δοκιμή αλγορίθμων μηχανικής μάθησης και οπτικοποιήσεων. Κάθε γραμμή του πίνακα

αναπαριστά ένα λουλούδι ίριδας, συμπεριλαμβανομένου του είδους του και των διαστάσεων των βοτανικών μερών του, του σέπαλου και του πέταλου, σε εκατοστά.

	Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Σχήμα 3.3: Iris Σύνολο Δεδομένων

Το Σύνολο Δεδομένων Diabetes προέρχεται αρχικά από το Εθνικό Ινστιτούτο Διαβήτη και Πεπτικών και Νεφρικών Νοσημάτων. Ο στόχος είναι να προβλεφθεί με βάση διαγνωστικές μετρήσεις εάν ένας ασθενής πάσχει από διαβήτη. Για την επιλογή αυτών των περιπτώσεων από μια μεγαλύτερη βάση δεδομένων τέθηκαν διάφοροι περιορισμοί. Συγκεκριμένα, όλοι οι ασθενείς εδώ είναι γυναίκες τουλάχιστον 21 ετών με καταγωγή από τους Ινδιάνους Pima. Θα μπορούσε να χαρακτηριστεί ως ένα μεσαίου μεγέθους Σύνολο Δεδομένων αφού αποτελείται από 758 γραμμές.

	t.pregnant <int>	plasma <dbl>	bl.press <int>	tr.thick <dbl>	serum.ins <dbl>	bmi <dbl>	diab <dbl>	age <int>	class <int>
1	6	148	72	35.00000	125	33.6	0.627	50	1
2	1	85	66	29.00000	125	26.6	0.351	31	0
3	8	183	64	29.15342	125	23.3	0.672	32	1
4	1	89	66	23.00000	94	28.1	0.167	21	0
5	0	137	40	35.00000	168	43.1	2.288	33	1
6	5	116	74	29.15342	125	25.6	0.201	30	0

Σχήμα 3.4: Diabetes Σύνολο Δεδομένων

Το Σύνολο Δεδομένων Wine αποτελείται από 13 διαφορετικές παραμέτρους του κρασιού, όπως η περιεκτικότητα σε αλκοόλη και τέφρα, η οποία μετρήθηκε για 178 δείγματα. Τα κρασία αυτά καλλιεργήθηκαν στην ίδια περιοχή της Ιταλίας, αλλά προέρχονταν από τρεις διαφορετικές ποικιλίες, επομένως υπάρχουν τρεις διαφορετικές κατηγορίες. Ο στόχος είναι να βρεθεί ένα μοντέλο που να μπορεί να προβλέψει την κατηγορία του κρασιού με βάση τις 13 μετρούμενες παραμέτρους και να βρεθούν οι κύριες διαφορές μεταξύ των τριών διαφορετικών κατηγοριών. Συνεπώς, το συγκεκριμένο Σύνολο Δεδομένων μπορεί να χρησιμοποιηθεί για προβλήματα ταξινόμησης

Το Σύνολο Δεδομένων Wifi συλλέχθηκε για την εκτέλεση πειραμάτων σχετικά με τον τρόπο με τον οποίο η ισχύς του σήματος Wifi μπορεί να χρησιμοποιηθεί για τον προσδιορισμό ενός

	Type <dbl>	Alcohol <dbl>	Sugar-free Extract <dbl>	Fixed Acidity <dbl>	Tartaric Acid <dbl>	Malic Acid <dbl>	Uronic Acids <dbl>	pH <dbl>	Ash <dbl>
1	1	14.23	24.82	73.1	1.21	1.71	0.72	3.38	2.43
2	1	13.20	26.30	72.8	1.84	1.78	0.71	3.30	2.14
3	1	13.16	26.30	68.5	1.94	2.36	0.84	3.48	2.67
4	1	14.37	25.85	74.9	1.59	1.95	0.72	3.43	2.50
5	1	13.24	26.05	83.5	1.30	2.59	1.10	3.42	2.87
6	1	14.20	28.40	79.9	2.14	1.76	0.96	3.39	2.45

Σχήμα 3.5: Wine Σύνολο Δεδομένων

εκ των τεσσάρων εσωτερικών δωματίων. Κάθε χαρακτηριστικό είναι η ισχύς του σήματος Wifi που παρατηρείται στο smartphone από όπου συλλέχθηκαν τα δεδομένα. Η μεταβλητή απόφασης είναι ένα από τα τέσσερα δωμάτια. Το δείγμα μας αποτελείται από 2000 γραμμές.

	wifi_signal_1 <int>	wifi_signal_2 <int>	wifi_signal_3 <int>	wifi_signal_4 <int>	wifi_signal_5 <int>	wifi_signal_6 <int>	wifi_signal_7 <int>	room <int>
1	-64	-56	-61	-66	-71	-82	-81	1
2	-68	-57	-61	-65	-71	-85	-85	1
3	-63	-60	-60	-67	-76	-85	-84	1
4	-61	-60	-68	-62	-77	-90	-80	1
5	-63	-65	-60	-63	-77	-81	-87	1
6	-64	-55	-63	-66	-76	-88	-83	1

Σχήμα 3.6: Wifi Σύνολο Δεδομένων

Κεφάλαιο 4

Πειραματικό Στάδιο, Συμπεράσματα και Μελλοντική Έρευνα

4.1 Πειρματικό Στάδιο

Random Forest: Το πειραματικό στάδιο με την χρήση του αλγορίθμου Random Forest ξεκίνησε με το Iris Σύνολο Δεδομένων. Αφού έγινε διαχωρισμός των δεδομένων(150) σε Δεδομένα Εκπαίδευσης (120) και Δεδομένα Ελέγχου (30) με δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων καθορίστηκε το μοντέλο εκπαίδευσης. Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου είναι τα κάτωθι. Δηλαδή το **86**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	4	6

Σχήμα 4.1: Iris Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων

Στη συνέχεια επί των Δεδομένων Εκπαίδευσης έγινε δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης δηλαδή καθορίστηκε με το 80% περίπου των δεδομένων εκπαίδευσης (100). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **83**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	5	5

Σχήμα 4.2: Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)

Επί των Δεδομένων Εκπαίδευσης έγινε και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 65% περίπου των δεδομένων εκπαίδευσης (80). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **83**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	5	5

Σχήμα 4.3: Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)

Επί των Δεδομένων Εκπαίδευσης έγινε ακόμα δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 50% περίπου των δεδομένων εκπαίδευσης (60). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **86**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	4	6

Σχήμα 4.4: Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)

Ύστερα, επί των Δεδομένων Εκπαίδευσης έγινε εκ νέου δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 33% περίπου των δεδομένων εκπαίδευσης (40). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **90**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	3	7

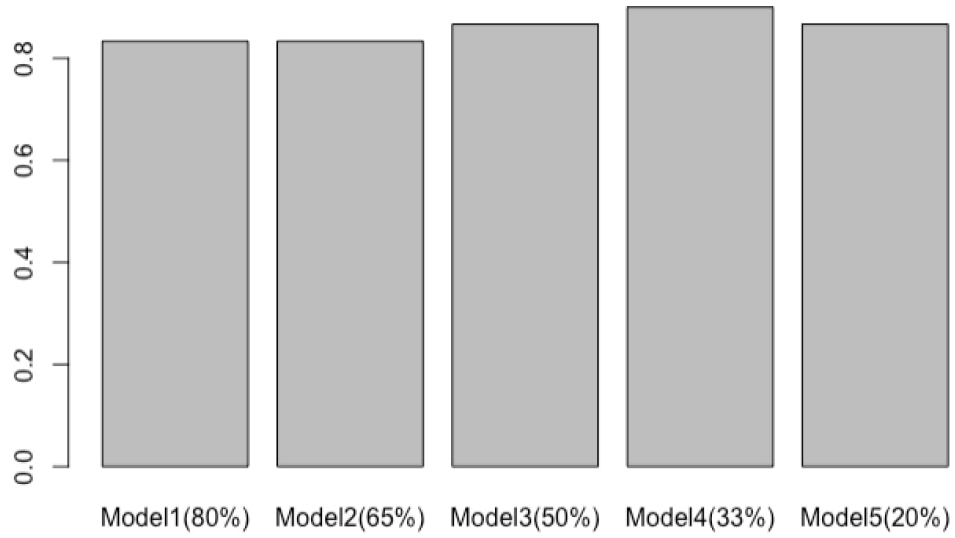
Σχήμα 4.5: Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)

Τέλος, επί των Δεδομένων Εκπαίδευσης έγινε για τελευταία φορά και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 20% περίπου των δεδομένων εκπαίδευσης (20). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **86**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	10	0
virginica	0	4	6

Σχήμα 4.6: Iris Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)

Συνοψίζοντας λοιπόν στο συγκεκριμένο Σύνολο Δεδομένων αν και αρχικά παρατηρείται μια μικρή πτώση της ακρίβειας κατάταξης των δεδομένων στη συνέχεια έχουμε με μικρότερη ποσότητα δεδομένων μεγαλύτερη ακρίβεια. Άρα, με την δειγματοληψία με την χρήση Λατινικών Υπερκύβων επιτυγχάνεται με μικρότερο Σύνολο Δεδομένων ελάχιστη απώλεια στην ακρίβεια κατάταξης. Στο κάτωθι σχήμα παρουσιάζονται συγκεντρωτικά όλα τα Ακρίβεια για τα 5 μοντέλα που προπονήθηκαν με δειγματοληψία με λατινικούς υπερκύβους.



Σχήμα 4.7: Ακρίβεια Iris Σύνολο Δεδομένων

Το πειραματικό στάδιο συνεχίστηκε με την χρήση του Wine Σύνολο Δεδομένων. Αφού έγινε διαχωρισμός των δεδομένων (178) σε Δεδομένα Εκπαίδευσης (150) και Δεδομένα Ελέγχου (28) με δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων καθορίστηκε το μοντέλο εκπαίδευσης. Τα αποτελέσματα στο Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των κρασιών κατατάχθηκαν στη σωστή κατηγορία.

	predicted		
observed	1	2	3
1	7	0	0
2	0	15	1
3	0	0	5

Σχήμα 4.8: Wine Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων

Στη συνέχεια επί των Δεδομένων Εκπαίδευσης έγινε δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης δηλαδή καθορίστηκε με το 80% περίπου των δεδομένων εκπαίδευσης (125). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των κρασιών κατατάχθηκαν στη σωστή κατηγορία.

	predicted		
observed	1	2	3
1	7	0	0
2	0	15	1
3	0	0	5

Σχήμα 4.9: Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων(80%)

Επί των Δεδομένων Εκπαίδευσης έγινε και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 65% περίπου των δεδομένων εκπαίδευσης (100). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των κρασιών κατατάχθηκαν στη σωστή κατηγορία.

	predicted		
observed	1	2	3
1	7	0	0
2	0	15	1
3	0	0	5

Σχήμα 4.10: Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)

Επί των Δεδομένων Εκπαίδευσης έγινε ακόμα δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 50% περίπου των δεδομένων εκπαίδευσης (75). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των κρασιών κατατάχθηκαν στη σωστή κατηγορία.

	predicted		
observed	1	2	3
1	7	0	0
2	0	15	1
3	0	0	5

Σχήμα 4.11: Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)

Ύστερα, επί των Δεδομένων Εκπαίδευσης έγινε εκ νέου δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 33% περίπου των δεδομένων εκπαίδευσης (40). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **100**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	1	2	3
1	7	0	0
2	0	16	0
3	0	0	5

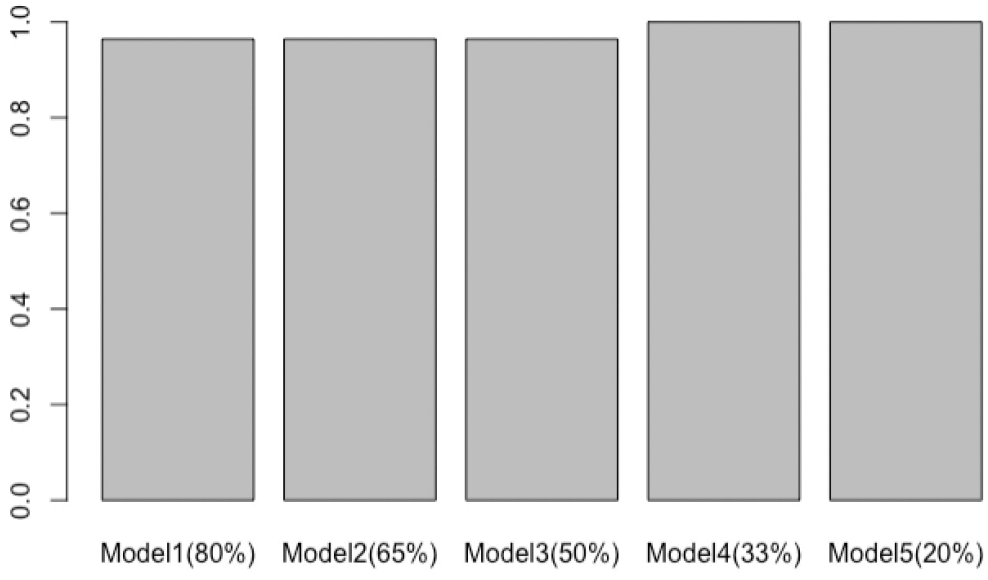
Σχήμα 4.12: Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)

Τέλος, επί των Δεδομένων Εκπαίδευσης έγινε για τελευταία φορά και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 20% περίπου των δεδομένων εκπαίδευσης (25). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **86**% των λουλουδιών κατατάχθηκαν στο σωστό είδος.

	predicted		
observed	1	2	3
1	7	0	0
2	0	16	0
3	0	0	5

Σχήμα 4.13: Wine Σύνολο δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)

Συνοψίζοντας λοιπόν στο συγκεκριμένο Σύνολο Δεδομένων παρατηρείται σταθερή ή ακόμα και μεγαλύτερη ακρίβεια στην κατάταξη των κρασιών σε σωστή κατηγορία. Άρα, με την δειγματοληψία με την χρήση Λατινικών Υπερκύβων επιτυγχάνεται με μικρότερο Σύνολο Δεδομένων ελάχιστη ή καθόλου απώλεια στην ακρίβεια κατάταξης. Στο κάτωθι σχήμα παρουσιάζονται συγκεντρωτικά όλα τα Ακρίβεια για τα 5 μοντέλα που προπονήθηκαν με δειγματοληψία με λατινικούς υπερκύβους.



Σχήμα 4.14: Ακρίβεια Wine Σύνολο Δεδομένων

Το τρίτο πείραμα έγινε στο Diabetes Σύνολο Δεδομένων. Αφού έγινε διαχωρισμός των δεδομένων (768) σε Δεδομένα Εκπαίδευσης (700) και Δεδομένα Ελέγχου (68) με δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων καθορίστηκε το μοντέλο εκπαίδευσης. Τα αποτελέσματα στο Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **79**% των ασθενών

κατατάχθηκαν στη σωστή κατηγορία(δηλαδήαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	46	7
1	7	8

Σχήμα 4.15: Diabetes Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων

Στη συνέχεια επί των Δεδομένων Εκπαίδευσης έγινε δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης δηλαδή καθορίστηκε με το 80% περίπου των δεδομένων εκπαίδευσης (525). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **79**% των ασθενών κατατάχθηκαν στη σωστή κατηγορία(δηλαδήαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	47	6
1	8	7

Σχήμα 4.16: Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)

Επί των Δεδομένων Εκπαίδευσης έγινεκαι πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 65% περίπου των δεδομένων εκπαίδευσης (510). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **82**% των ασθενών κατατάχθηκαν στη σωστή κατηγορία(δηλαδήαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	47	6
1	6	9

Σχήμα 4.17: Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)

Επί των Δεδομένων Εκπαίδευσης έγινε ακόμα δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 50% περίπου των δεδομένων εκπαίδευσης (380). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **82**% των ασθενών κατατάχθηκαν στη σωστή κατηγορία(δηλαδήαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	48	5
1	7	8

Σχήμα 4.18: Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)

Ύστερα, επί των Δεδομένων Εκπαίδευσης έγινε εκ νέου δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 33% περίπου των δεδομένων εκπαίδευσης (250). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **82**% των ασθενών κατατάχθηκαν στη σωστή κατηγορία(δηλαδήαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	49	4
1	8	7

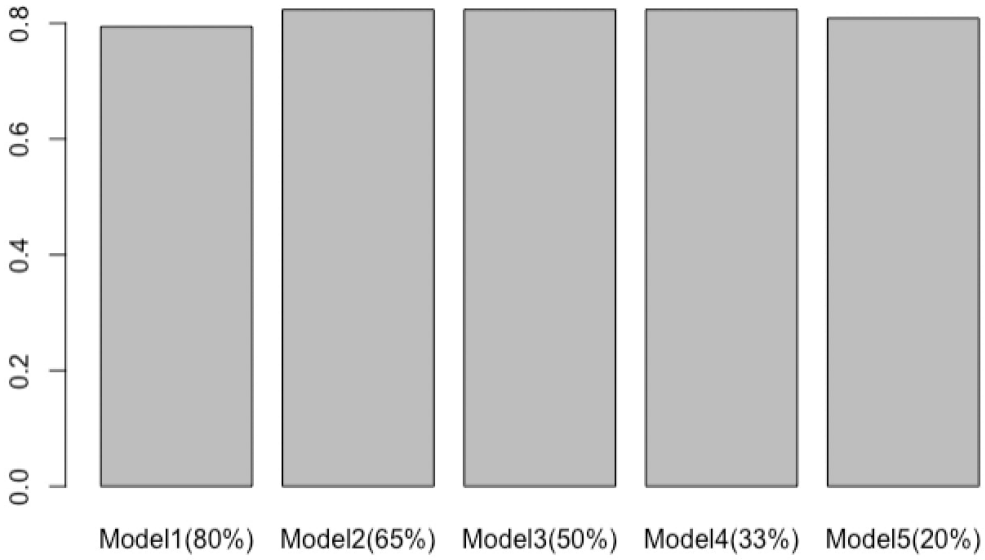
Σχήμα 4.19: Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)

Τέλος, επί των Δεδομένων Εκπαίδευσης έγινε για τελευταία φορά και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 20% περίπου των δεδομένων εκπαίδευσης (150). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **80**% των ασθενών κατατάχθηκαν στη σωστή κατηγορία(δηλαδή αν έχουν διαβήτη ή όχι).

	predicted	
observed	0	1
0	43	10
1	3	12

Σχήμα 4.20: Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)

Συνοψίζοντας λοιπόν στο συγκεκριμένο Σύνολο Δεδομένων παρατηρείται και πάλ σταθερή ή ακόμα και μεγαλύτερη ακρίβεια στην κατάταξη των ασθενών σε σωστή κατηγορία. Άρα, με την δειγματοληψία με την χρήση Λατινικών Υπερκύβων επιτυγχάνεται με μικρότερο Σύνολο Δεδομένων ελάχιστη ή καθόλου απώλεια στην ακρίβεια κατάταξης. Στο κάτωθι σχήμα παρουσιάζονται συγκεντρωτικά όλα τα Ακρίβεια για τα 5 μοντέλα που προπονήθηκαν με δειγματοληψία με λατινικούς υπερκύβους.



Σχήμα 4.21: Ακρίβεια Diabetes Σύνολο Δεδομένων

Τέλος, το τελευταίο πείραμα έγινε στο Wifi Σύνολο Δεδομένων. Αφού έγινε διαχωρισμός των δεδομένων(2000) σε Δεδομένα Εκπαίδευσης (1500) και Δεδομένα Ελέγχου (500) με δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων καθορίστηκε το μοντέλο εκπαίδευσης. Τα αποτελέσματα στο Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **97**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	92	10	0
3	1	1	125	0
4	1	0	1	129

Σχήμα 4.22: Wifi Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων

Στη συνέχεια επί των Δεδομένων Εκπαίδευσης έγινε δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης δηλαδή καθορίστηκε με το 80% περίπου των Δεδομένων Εκπαίδευσης (1125). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **97**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	93	9	0
3	1	1	125	0
4	1	0	1	129

Σχήμα 4.23: Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων(80%)

Επί των Δεδομένων Εκπαίδευσης έγινεκαι πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 65% περίπου των Δεδομένων Εκπαίδευσης (975). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **97**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	91	11	0
3	1	0	126	0
4	1	0	1	129

Σχήμα 4.24: Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)

Επί των Δεδομένων Εκπαίδευσης έγινε ακόμα δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 50% περίπου των Δεδομένων Εκπαίδευσης (750). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **97**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	92	10	0
3	0	1	125	1
4	1	0	1	129

Σχήμα 4.25: Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)

Ύστερα, επί των Δεδομένων Εκπαίδευσης έγινε εκ νέου δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 33% περίπου των Δεδομένων Εκπαίδευσης (495). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	89	13	0
3	0	2	125	0
4	1	0	1	129

Σχήμα 4.26: Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)

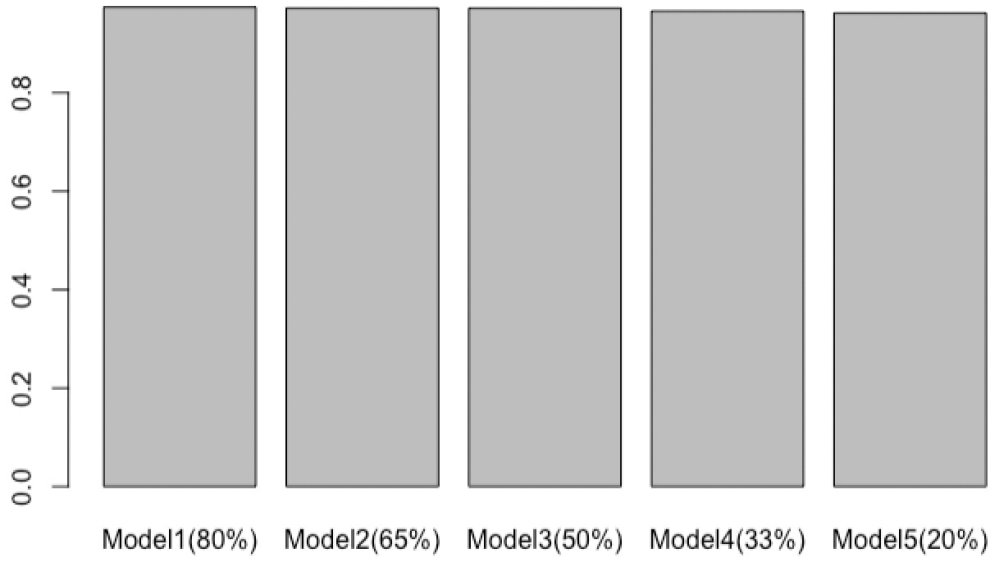
Τέλος, επί των Δεδομένων Εκπαίδευσης έγινε για τελευταία φορά και πάλι δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων. Το μοντέλο εκπαίδευσης καθορίστηκε τώρα με το 20%

περίπου των Δεδομένων Εκπαίδευσης (300). Τα αποτελέσματα στο κοινό Σύνολο Ελέγχου Δεδομένων είναι τα κάτωθι. Δηλαδή το **96**% των σημάτων Wifi κατατάχθηκαν στο σωστό δωμάτιο.

	predicted			
observed	1	2	3	4
1	140	0	0	0
2	0	90	12	0
3	1	1	124	1
4	1	0	3	127

Σχήμα 4.27: Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)

Συνοψίζοντας λοιπόν στο συγκεκριμένοΣύνολο Δεδομένων παρατηρείται και πάλι σταθερή ή κατά 1% μικρότερη ακρίβεια στην σωστή κατάταξη των δεδομένων. Άρα, με την δειγματοληψία με την χρήση Λατινικών Υπερκύβων επιτυγχάνεται με μικρότεροΣύνολο Δεδομένων ελάχιστη ή καθόλου απώλεια στην ακρίβεια κατάταξης. Στο κάτωθι σχήμα παρουσιάζονται συγκεντρωτικά όλα τα Ακρίβεια για τα 5 μοντέλα που προπονήθηκαν με δειγματοληψία με λατινικούς υπερκύβους.



Σχήμα 4.28: Ακρίβεια Wifi Σύνολο Δεδομένων

Κ-Πλησιέστεροι Γείτονες: Στους παρακάτω πίνακες εμφανίζονται συνοπτικά τα αποτελέσματα για κάθενα από τα παραπάνω σύνολα δεδομένων:

Iris Σύνολο Δεδομένων:

Πίνακας 4.1: Iris Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες)

Μοντέλο	Ακρίβεια
Iris Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων	93%
Iris Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)	96%
Iris Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)	96%
Iris Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)	96%
Iris Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)	96%
Iris Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)	96%

Wine Σύνολο Δεδομένων:

Πίνακας 4.2: Wine Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες)

Μοντέλο	Ακρίβεια
Wine Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων	71%
Wine Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)	71%
Wine Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)	67%
Wine Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)	75%
Wine Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)	64%
Wine Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)	67%

Diabetes Σύνολο Δεδομένων:

Πίνακας 4.3: Diabetes Σύνολο Δεδομένων (Κ-Πλησιέστεροι Γείτονες)

Μοντέλο	Ακρίβεια
Diabetes Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων	76%
Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)	76%
Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)	72%
Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)	73%
Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)	77%
Diabetes Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)	75%

Wifi Σύνολο Δεδομένων:

Πίνακας 4.4: Wifi Σύνολο Δεδομένων (K-Πλησιέστεροι Γείτονες)

Μοντέλο	Ακρίβεια
Wifi Σύνολο Δεδομένων χωρίς τη μέθοδο των Λατινικών Υπερκύβων	97%
Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (80%)	97%
Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (65%)	96%
Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (50%)	97%
Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (33%)	96%
Wifi Σύνολο Δεδομένων με τη μέθοδο των Λατινικών Υπερκύβων (20%)	96%

Είναι εμφανές ότι και σε αυτό το πειραματικό στάδιο όπου άλλαξε ο αλγόριθμος κατάταξης των δεδομένων με την δειγματοληψία με την χρήση των Λατινικών Υπερκύβων επιτυγχάνεται με μικρότεροΣύνολο Δεδομένων ελάχιστη ή καθόλου απώλεια στην ακρίβεια κατάταξης.

4.2 Συμπεράσματα και Μελλοντική Έρευνα

Εν κατακλείδι του πειραματικού σταδίου που παρουσιάσαμε στο προηγούμενο κεφάλαιο φαίνεται ότι η μέθοδος δειγματοληψίας με την χρήση των Λατινικών Υπερκύβων είναι μια καλή μέθοδος δειγματοληψίας για τα προβλήματα της μηχανικής μάθησης. Με αυτό τον τρόπο θα μπορούσε να πει κάποιος ότι «εξαλείφεται» το πρόβλημα των μεγάλων δεδομένων που θέσαμε στα πρώτα κεφάλαια. Τα 3 κύρια συμπεράσματα που μπορούμε να εξάγουμε είναι τα κάτωθι:

Συμπέρασμα 1

Από τα παραπάνω πειραματικά στάδια εξάγεται το συμπέρασμα ότι με την δειγματοληψία με τη μέθοδο των Λατινικών Υπερκύβων η απώλεια που έχουμε στα αποτελέσματα της μηχανικής μάθησης είναι ελάχιστη έως καθόλου. Μάλιστα, σε κάποιο ποσοστό δειγματοληψίας επί των αρχικών δεδομένων παρατηρούνται ακόμα και καλύτερα αποτελέσματα. Συγκεκριμένα, θα μπορούσαμε να πούμε σαν γενικό συμπέρασμα ότι αν κάνουμε δειγματοληψία έως και το 50% των δεδομένων μας (δηλαδή κρατήσουμε τα μισά από τα δεδομένα μας) τότε τα αποτελέσματα μας θα είναι καλύτερα, ίδια ή ελάχιστα χειρότερα.

Συμπέρασμα 2

Βέβαια αξίζει να σημειωθεί παρατηρώντας τα αποτελέσματα των παραπάνω πειραμάτων ότι σε ποσοστό άνω του 50% τα αποτελέσματα μας φαίνονται να χειροτερεύουν. Δηλαδή, δειγματοληπτώντας τα αρχικά δεδομένα μας σε ποσοστό μεγαλύτερο του 50% περιμένουμε πτώση της ακρίβειας του εκάστοτε μοντέλου.

Συμπέρασμα 3

Το τελευταίο αλλά και σημαντικότερο συμπέρασμα που εξαγάγουμε μετά τα πειραματικά αυτά στάδια είναι ότι η μέθοδος των Λατινικών Υπερκύβων αποτελεί ένα εργαλείο με το οποίο δύναται να «εξαλείψουμε» το πρόβλημα των μεγάλων δεδομένων. Έμπρακτα αυτό σημαίνει ότι αν έχουμε ένα μηχάνημα στην διάθεσή μας στο οποίο δεν δύναται να τρέξουν 1 εκατομμύριο δεδομένα και ούτε θέλουμε να χαλάσουμε κάποια χρήματα ώστε να πάμε σε κάποιο υπολογιστικό νέφος τότε αν κάνουμε δειγματοληψία με τη μέθοδο LHS θα καταφέρουμε να τρέξουμε το πρόβλημα μας με τα μισά δεδομένα στον υπολογιστή.

Μελλονική Έρευνα:

Σημαντική επέκταση της παρούσας εργασίας αποτελεί η διευρυνση του πειραματικού σταδίου σε περισσότερα Σύνολων Δεδομένων. Επιπλεον, σημαντική επέκταση θα ήταν και η λύση προβλημάτων πρόβλεψης αφού στην παρούσα εργασία λύνονται μόνο προβλήματα ταξινόμησης.

Για το πειραματικό στάδιο χρησιμοποιήθηκε το πακέτο της γλώσσας προγραμματισμού R, cLHS το οποίο εισήχθει από τους Minasny και McBratney, (2006). Είναι μια υπό συνθήκες χρήση της δειγματοληψίας με τη μέθοδο των Λατινικών Υπερκύβων. Η δημιουργία μιας βιβλιοθήκης που θα σέβεται όλες τις αρχές της δειγματοληψίας με τη μέθοδο των Λατινικών Υπερκύβων αποτελεί μια συνέχεια της συγκεκριμένης διπλωματικής.

Αποτελεσματικές στρατηγικές δειγματοληψίας που κλιμακώνονται ανάλογα με το μέγεθος του προβλήματος, τον υπολογιστικό προϋπολογισμό και τις ανάγκες των χρηστών είναι απαραίτητες για διάφορες αναλύσεις που βασίζονται στη δειγματοληψία, όπως η ανάλυση ευαισθησίας και αβεβαιότητας. Στη μελέτη των Sheikholeslami και Razavi, (2017), προτείνετε μια νέα στρατηγική, που ονομάζεται Προοδευτική Δειγματοληψία Λατινικού Υπερκύβου (Progressive Latin Hypercube Sampling - PLHS), η οποία παράγει διαδοχικά σημεία δείγματος, διατηρώντας προοδευτικά τις ιδιότητες κατανομής που μας ενδιαφέρουν (ιδιότητες Λατινικού Υπερκύβου, πλήρωση χώρου κ.λπ.), καθώς το μέγεθος του δείγματος αυξάνεται. Η χρήση της Προοδευτικής Δειγματοληψίας Λατινικών Υπερκύβων σε εφαρμογές της Μηχανικής Μάθησης και η σύγκριση των αποτελέσματων με την απλή Δειγματοληψία Λατικών Υπερκύβων αποτελεί προτεινόμενη μελλοντική έρευνα.

Τέλος, προτείνεται η χρήση της Δειγματαληψίας Λατινικών Υπερκύβων με συσχετισμένες μεταβλητές στα πλαίσια της Μηχανικής Μάθησης για την επίλυση προβλημάτων πρόβλεψης και ταξινόμησης.

Βιβλιογραφία

[1] Bergstra, J., and Bengio, Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research.* 13: 281-305 .

[2] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees.* Chapman Hall/CRC.

[3] Breiman, L. 2001. Random Forests. *Machine Learning.* 45:5-32.

[4] Georgiou, S.D., Stylianou, S., Drosou, K. and Koukouvinos, C. 2014. Construction of orthogonal and nearly orthogonal designs for computer experiments. *Biometrika.*101(3), 741–747.

[5] Georgiou, S., Koukouvinos, C., Liu, M. 2014. U-type and column-orthogonal designs for computer experiments. *Metrika.* Springer. 77(8) November.

[6] Gruijter, J.J. de, D.J. Brus, M.F.P. Bierkens, and M. Knotters. 2006. *Sampling for natural resource monitoring: statistics and methodology of sampling and data analysis.* Berlin: Springer-Verlag. p. 343.

[7] Hastie T., Tibshirani, R. and Friedman, J.H. 2009. *The Elements of Statistical Learning. Springer Series in Statistics.* Springer New York Inc., New York. 2nd edition.

[8] Iman, R. L., and Conover, W.J. 1980. *Small sample sensitivity analysis techniques for computer models, with an application to risk assessment, Communications in Statistics - Theory and Methods.* 9:17. 1749-1842.

[9] Katharopoulos, A. and Fleuret, F. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. *In Proceedings of the 35th International Conference on Machine Learning (ICML).* p 2525–2534.

[10] Kuhn, M. and Johnson, K. 2013. *Applied Predictive Modeling.* Springer New York.

[11] Li, R., Lin, D.K.J., Li, B. 2013. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry.* 29(5), 399-409

[12] Mathworks. 2005. Matlab Release 14. *The Mathworks Inc.* Natick, MA.

[13] McKay, M.D., Beckman R.J. and Conover W.J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 21: 239–245.

[14] Minasny, B. and McBratney A.B. 2002. The Neuro-m Method for Fitting Neural Network Parametric Pedotransfer Functions. *Soil Science Society of American Journal*. 66:352-361.

[15] Minasny, B., and McBratney A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences*. 32:1378-1388.

[16] Owen, A. B. 2013. *Monte Carlo Theory, Methods and Examples*. Art Owen.

[17] Pebesma, E.J., and G.B.M Heuvelink. 1999. Latin hypercube sampling of Gaussian random fields. *Technometrics*. 41, 303–312.

[18] Peng, H., Long F., and Ding C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 27(8):1226-1238.

[19] Viana, F. 2016. A Tutorial on Latin Hypercube Design of Experiments. *Quality and Reliability Engineering International*. 32: 1975-1985.

[20] Tang, B. 1993. Orthogonal Array-Based Latin Hypercubes. *Journal of the American Statistical Association*. 88(424).

[21] Stein M. 2012. Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics*. 29(2).

[22] Vinyals, O. and D. Povey. 2011. Krylov Subspace Descent for Deep Learning. *AISTATS*.

[23] Webster, R. and M. Lark. 2013. *Field Sampling for Environmental Science and Management*. Taylor and Francis Group. London.

[24] Yeh, I. C., and C.H. Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*. 36(2): 2473-2480.

[25] Zhang, Y. and Pinder, G.F. 2003. Latin-hypercube sample-selection strategies for correlated random hydraulic-conductivity fields. *Water Resources Research*. 39(8).

Appendix A

Παράρτημα με κώδικα αναπαραγωγής αποτελεσμάτων

```
############################
     IRIS DATASET
############################

library(purrr)
library(dplyr)
data(iris)
head(iris)


#Counting the rows of the data set


nrow(iris)


#Split to train and test with LHS


library(clhs)
train <- clhs(iris, size = 120, progress = FALSE,
              iter = 1000, simple = FALSE)


#train data
```

```
train=train$sampled_data
head(train)


#Keeping the rows that are sampled


rows<-rownames(train)


#test dataset


test <-iris[! rownames(iris) %in% rows,]
head(test)

##########################################
RF without LHS:
##########################################


# load the library
library(caret)

# define training control
train_control <- trainControl(method="cv", number=5)

x <- train[,1:4]
y <- train$Species

# train the model
model <- train(x=x,y=y ,
               trControl=train_control, method="rf")

# summarize results
print(model)



predicted_table <- predict(model, test[,-5])
```

```
table(observed = test[,5], predicted = predicted_table)



#Calcualte the accuracy


cm<-table(observed = test[,5], predicted = predicted_table)
n<-sum(cm)
diag<-diag(cm)
accuracy = sum(diag) / n
accuracy



##########################################################
With LHS in train data: from 120 keep 100
##########################################################



library(clhs)
train_1 <- clhs(train, size = 100, progress = FALSE,
                iter = 1000, simple = FALSE)



train_1=train_1$sampled_data
nrow(train_1)


#RF with LHS : 100 train data


# load the library
library(caret)

# define training control
train_control_1 <- trainControl(method="cv", number=5)

#data
```

```
x1 <- train_1[,1:4]
y1 <- train_1$Species

# train the model
model_1 <- train(x=x1, y=y1,
                 trControl=train_control_1, method="rf")

# summarize results
print(model_1)



predicted_table_1 <- predict(model_1, test[,-5])
table(observed = test[,5], predicted = predicted_table_1)


#Calcualte the accuracy


cm1<-table(observed = test[,5], predicted = predicted_table_1)
n1<-sum(cm1)
diag1<-diag(cm1)
accuracy_model_1 = sum(diag1) / n1
accuracy_model_1


############################################################
With LHS in train data: from 120 keep 80
############################################################



library(clhs)
train_2 <- clhs(train, size = 80, progress = FALSE,
                iter = 1000, simple = FALSE)



train_2=train_2$sampled_data
nrow(train_2)
```

```
#RF with LHS : 80 train data


# load the library
library(caret)

# define training control
train_control_2 <- trainControl(method="cv", number=5)


#data
x2 <- train_2[,1:4]
y2 <- train_2$Species

# train the model
model_2 <- train(x=x2, y=y2,
                 trControl=train_control_2, method="rf")

# summarize results
print(model_2)



predicted_table_2 <- predict(model_2, test[,-5])
table(observed = test[,5], predicted = predicted_table_2)


#Calcualte the accuracy


cm2<-table(observed = test[,5], predicted = predicted_table_2)
n2<-sum(cm2)
diag2<-diag(cm2)
accuracy_model_2 = sum(diag2) / n2
accuracy_model_2


##########################################################
With LHS in train data: from 120 keep 60
##########################################################
```

```
library(clhs)
train_3 <- clhs(train, size = 60, progress = FALSE,
                iter = 1000, simple = FALSE)



train_3=train_3$sampled_data
nrow(train_3)


#RF with LHS : 60 train data


# load the library
library(caret)

# define training control
train_control_3 <- trainControl(method="cv", number=5)

#data
x3 <- train_3[,1:4]
y3 <- train_3$Species

# train the model
model_3 <- train(x=x3, y=y3,
                 trControl=train_control_3, method="rf")

# summarize results
print(model_3)



predicted_table_3 <- predict(model_3, test[,-5])
table(observed = test[,5], predicted = predicted_table_3)



cm3<-table(observed = test[,5], predicted = predicted_table_3)
```

```
n3<-sum(cm3)
diag3<-diag(cm3)
accuracy_model_3 = sum(diag3) / n3
accuracy_model_3

##########################################################
With LHS in train data: from 120 keep 40
##########################################################



library(clhs)
train_4 <- clhs(train, size = 40, progress = FALSE,
                iter = 1000, simple = FALSE)



train_4=train_4$sampled_data
nrow(train_4)


#RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_4 <- trainControl(method="cv", number=5)

#data
x4 <- train_4[,1:4]
y4 <- train_4$Species

# train the model
model_4 <- train(x=x4, y=y4,
                 trControl=train_control_4, method="rf")

# summarize results
print(model_4)
```

```
predicted_table_4 <- predict(model_4, test[,-5])
table(observed = test[,5], predicted = predicted_table_4)



cm4<-table(observed = test[,5], predicted = predicted_table_4)
n4<-sum(cm4)
diag4<-diag(cm4)
accuracy_model_4 = sum(diag4) / n4
accuracy_model_4




##########################################################
With LHS in train data: from 120 keep 20
##########################################################



library(clhs)
train_5 <- clhs(train, size = 20, progress = FALSE,
                iter = 1000, simple = FALSE)



train_5=train_5$sampled_data
nrow(train_5)


#RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_5 <- trainControl(method="cv", number=5)
```

```
#data
x5 <- train_5[,1:4]
y5 <- train_5$Species

# train the model
model_5 <- train(x=x5, y=y5,
                 trControl=train_control_5, method="rf")

# summarize results
print(model_5)



predicted_table_5 <- predict(model_5, test[,-5])
table(observed = test[,5], predicted = predicted_table_5)



cm5<-table(observed = test[,5], predicted = predicted_table_5)
n5<-sum(cm5)
diag5<-diag(cm5)
accuracy_model_5 = sum(diag5) / n5
accuracy_model_5



accuracy<-c(accuracy_model_1,accuracy_model_2,
            accuracy_model_3,accuracy_model_4,accuracy_model_5)
barplot(accuracy)

############################
     WIFI DATASET
############################

colnames(wifi_localization) <- c("wifi_signal_1","wifi_signal_2",
"wifi_signal_3",
"wifi_signal_4","wifi_signal_5",
"wifi_signal_6","wifi_signal_7",
"room")
head(wifi_localization)
```

```
unique(wifi_localization$room)
nrow(wifi_localization)


#Split to train and test with LHS


library(clhs)
train <- clhs(wifi_localization, size = 1500, progress = FALSE,
iter = 1000, simple = FALSE)


#train data


train=train$sampled_data
head(train)




rows<-rownames(train)


#test dataset


test <-wifi_localization[! rownames(wifi_localization) %in% rows,]
head(test)


##########################################
RF without LHS:
##########################################


# load the library
library(caret)
```

```
# define training control
train_control <- trainControl(method="cv",
number=5)

x <- train[,-8]
train$room <- as.character(train$room)
train$room <- as.factor(train$room)
y <- train$room

# train the model
model <- train(x=x,y=y ,
trControl=train_control, method="rf")

# summarize results
print(model)



predicted_table <- predict(model, test[,-8])
table(observed = test[,8], predicted = predicted_table)



#Calcualte the accuracy


cm<-table(observed = test[,8], predicted = predicted_table)
n<-sum(cm)
diag<-diag(cm)
accuracy = sum(diag) / n
accuracy


##########################################################
With LHS in train data: from 1500 keep 1125 (apprx %75 )
##########################################################



library(clhs)
```

```
train_1 <- clhs(train, size = 1125, progress = FALSE,
iter = 1000, simple = FALSE)



train_1=train_1$sampled_data
nrow(train_1)


#RF with LHS : 525 train data


# load the library
library(caret)

# define training control
train_control_1 <- trainControl(method="cv",
number=5)

#data
x1 <- train[,-8]
train$room <- as.character(train$room)
train$room <- as.factor(train$room)
y1 <- train$room

# train the model
model_1 <- train(x=x1, y=y1,
trControl=train_control_1, method="rf")

# summarize results
print(model_1)



predicted_table_1 <- predict(model_1, test[,-8])
table(observed = test[,8], predicted = predicted_table_1)


#Calcualte the accuracy
```

```
cm1<-table(observed = test[,8], predicted = predicted_table_1)
n1<-sum(cm1)
diag1<-diag(cm1)
accuracy_model_1 = sum(diag1) / n1
accuracy_model_1


##########################################################
With LHS in train data: from 1500 keep 975 (65%)
##########################################################



library(clhs)
train_2 <- clhs(train, size = 975, progress = FALSE,
iter = 1000, simple = FALSE)



train_2=train_2$sampled_data
nrow(train_2)


RF with LHS : 510 train data


# load the library
library(caret)

# define training control
train_control_2 <- trainControl(method="cv",
number=5)


#data
x2 <- train_2[,-8]
train_2$room <- as.character(train_2$room)
train_2$room <- as.factor(train_2$room)
y2 <- train_2$room

# train the model
```

```
model_2 <- train(x=x2, y=y2,
trControl=train_control_2, method="rf")

# summarize results
print(model_2)



predicted_table_2 <- predict(model_2, test[,-8])
table(observed = test[,8], predicted = predicted_table_2)


#Calcualte the accuracy


cm2<-table(observed = test[,8], predicted = predicted_table_2)
n2<-sum(cm2)
diag2<-diag(cm2)
accuracy_model_2 = sum(diag2) / n2
accuracy_model_2


#########################################################
With LHS in train data: from 1500 keep 750 (50%)
#########################################################



library(clhs)
train_3 <- clhs(train, size = 750, progress = FALSE,
iter = 1000, simple = FALSE)



train_3=train_3$sampled_data
nrow(train_3)


RF with LHS : 380 train data
```

```
# load the library
library(caret)

# define training control
train_control_3 <- trainControl(method="cv", number=5)

#data
x3 <- train_3[,-8]
train_3$room <- as.character(train_3$room)
train_3$room <- as.factor(train_3$room)
y3 <- train_3$room

# train the model
model_3 <- train(x=x3, y=y3,
trControl=train_control_3, method="rf")

# summarize results
print(model_3)



predicted_table_3 <- predict(model_3, test[,-8])
table(observed = test[,8], predicted = predicted_table_3)



cm3<-table(observed = test[,8], predicted = predicted_table_3)
n3<-sum(cm3)
diag3<-diag(cm3)
accuracy_model_3 = sum(diag3) / n3
accuracy_model_3


##########################################################
With LHS in train data: from 1500 keep 495 (appr%33)
##########################################################



library(clhs)
train_4 <- clhs(train, size = 495, progress = FALSE,
```

```
iter = 1000, simple = FALSE)



train_4=train_4$sampled_data
nrow(train_4)


RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_4 <- trainControl(method="cv", number=5)

#data
x4 <- train_4[,-8]
train_4$room <- as.character(train_4$room)
train_4$room <- as.factor(train_4$room)
y4 <- train_4$room

# train the model
model_4 <- train(x=x4, y=y4, trControl=train_control_4, method="rf")

# summarize results
print(model_4)



predicted_table_4 <- predict(model_4, test[,-8])
table(observed = test[,8], predicted = predicted_table_4)



cm4<-table(observed = test[,8], predicted = predicted_table_4)
n4<-sum(cm4)
diag4<-diag(cm4)
accuracy_model_4 = sum(diag4) / n4
accuracy_model_4
```

```
##########################################################
With LHS in train data: from 1500 keep 300 (apprx 20%)
##########################################################



library(clhs)
train_5 <- clhs(train, size = 300, progress = FALSE,
iter = 1000, simple = FALSE)



train_5=train_5$sampled_data
nrow(train_5)


RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_5 <- trainControl(method="cv", number=5)

#data
x5 <- train_5[,-8]
train_5$room <- as.character(train_5$room)
train_5$room <- as.factor(train_5$room)
y5 <- train_5$room

# train the model
model_5 <- train(x=x5, y=y5, trControl=train_control_5, method="rf")

# summarize results
print(model_5)
```

```
predicted_table_5 <- predict(model_5, test[,-8])
table(observed = test[,8], predicted = predicted_table_5)



cm5<-table(observed = test[,8], predicted = predicted_table_5)
n5<-sum(cm5)
diag5<-diag(cm5)
accuracy_model_5 = sum(diag5) / n5
accuracy_model_5



accuracy<-c(accuracy_model_1,
accuracy_model_2,accuracy_model_3,
accuracy_model_4,accuracy_model_5)
barplot(accuracy)

############################
     DIABETES DATASET
############################
  {r setup, include=FALSE}
summary(data)
colnames(data) <- c("t.pregnant","plasma","bl.press",
                    "tr.thick","serum.ins","bmi","diab",
                    "age","class")
head(data,5)
tail(data,5)

class <- data$class
t.pregnant <- data$t.pregnant

expl_data <- data[,2:8]

#We assume that the zeros in the variable
#times.pregnant are not missing, hence we
#don't replace zeros with "NA" and Replace
#the zeros with "NA"


head(expl_data,10)
```

```
expl_data[expl_data==0] <- NA

df <- data.frame(t.pregnant,expl_data,class)
# Create the transformed dataframe
head(df,5)
# View the first 5 obs of the df
tail(df,5)
# View the last 5 obs of the df

cols <- character(nrow(df))
cols[] <- "black"
cols[df$class %in% c(0,1)] <- c("blue","red")
pairs(df[,-9],col=cols)
# We observe that there's no conspicuous
#discrimination available


library(naniar)
gg_miss_var(df)

vis_miss(df, sort_miss = TRUE)

library(VIM)
res1 <- summary(aggr(df,sortVar=TRUE))$combinations

res2 <- summary(aggr(df,prop=TRUE,combined=TRUE))$combinations
# The graph represents the pattern, with blue
#for observed and red for missing.

head(res1[rev(order(res1[,2])),])

matrixplot(df, sortby = 9)


marginplot(df[,c("tr.thick","serum.ins")])

library(finalfit)
expl <- c("t.pregnant","plasma","bl.press",
          "tr.thick","serum.ins","bmi","diab","age")
dep <- c("class")
```

```
ff_glimpse(df,dependent=dep,
           explanatory=expl,digits = 1)

df1 <- df
# The first dataset that has the
#missing values df1, which will be imputed

# Data imputation with mean/median has the following advantages:
# Easy and fast.
# Works well with small numerical datasets.
# Cons:
# Doesnt factor the correlations between
#features. It only works on the column level.
# Will give poor results on encoded
#categorical features (do NOT use it on categorical features).
# Not very accurate.
# Doesnt account for the uncertainty in the imputations.
# Since there are a few outliers
#in the variable tr.thick we will
#impute the missing values using
#the mean as follows:
df1$tr.thick[is.na(df1$tr.thick)] <- mean(df1$tr.thick,na.rm=T)
summary(df1$tr.thick)
which(is.na(df1$tr.thick))
# As expected the result is zero!

# Since there are loads of outliers
#in the variable serum.ins we will
#impute the missing values using the median as follows:
df1$serum.ins[is.na(df1$serum.ins)] <- median(df1$serum.ins,na.rm=T)
summary(df1$serum.ins)
which(is.na(df1$serum.ins))
# As expected the result is zero!

# Since there are loads of outliers
#in the variable bl.pressure we will
#impute the missing values using the median as follows:
df1$bl.press[is.na(df1$bl.press)] <- median(df1$bl.press,na.rm=T)
summary(df1$bl.press)
which(is.na(df1$bl.press))
# As expected the result is zero!
```

```
# Since there are loads of outliers
#in the variable bl.pressure we will impute the
#missing values using the median as follows:
df1$bmi[is.na(df1$bmi)] <- median(df1$bmi,na.rm=T)
summary(df1$bmi)
which(is.na(df1$bmi))
# As expected the result is zero!

# Since there aren't outliers in the
#variable plasma we will impute the missing
#values using the mean as follows:
df1$plasma[is.na(df1$plasma)] <- mean(df1$plasma,
                                      na.rm=T)
summary(df1$plasma)
which(is.na(df1$plasma))
# As expected the result is zero!

library(finalfit)
ff_glimpse(df1,dependent=dep,
           explanatory=expl,digits = 1)
# Indeed there aren't any missing values now!
which(is.na(df1))
# Second way to check there are no longer
#missing values

#Counting the rows of the data set


nrow(df1)


#Split to train and test with LHS


library(clhs)
train <- clhs(df1, size = 700, progress = FALSE,
              iter = 1000, simple = FALSE)


#train data
```

```
train=train$sampled_data
head(train)



rows<-rownames(train)


#test dataset


test <-df1[! rownames(df1) %in% rows,]
head(test)

#RF without LHS:
##########################################


# load the library
library(caret)

# define training control
train_control <- trainControl(method="cv",
                              number=5)

x <- train[,-9]
train$class <- as.character(train$class)
train$class <- as.factor(train$class)
y <- train$class

# train the model
model <- train(x=x,y=y , trControl=train_control,
               method="rf")

# summarize results
print(model)
```

```
predicted_table <- predict(model, test[,-9])
table(observed = test[,9], predicted = predicted_table)


#Calcualte the accuracy


cm<-table(observed = test[,9], predicted = predicted_table)
n<-sum(cm)
diag<-diag(cm)
accuracy = sum(diag) / n
accuracy


##########################################################
With LHS in train data: from 700 keep 525 (apprx %75 )
##########################################################



library(clhs)
train_1 <- clhs(train, size = 525, progress = FALSE,
                iter = 1000, simple = FALSE)



train_1=train_1$sampled_data
nrow(train_1)


#RF with LHS : 525 train data


# load the library
library(caret)

# define training control
train_control_1 <- trainControl(method="cv",
                                number=5)

#data
```

```
x1 <- train_1[,-9]
train_1$class <- as.character(train_1$class)
train_1$class <- as.factor(train_1$class)
y1 <- train_1$class

# train the model
model_1 <- train(x=x1, y=y1, trControl=train_control_1,
                 method="rf")

# summarize results
print(model_1)



predicted_table_1 <- predict(model_1, test[,-9])
table(observed = test[,9], predicted = predicted_table_1)


#Calcualte the accuracy


cm1<-table(observed = test[,9], predicted = predicted_table_1)
n1<-sum(cm1)
diag1<-diag(cm1)
accuracy_model_1 = sum(diag1) / n1
accuracy_model_1


##########################################################
With LHS in train data: from 768 keep 510 (65%)
##########################################################



library(clhs)
train_2 <- clhs(train, size = 510, progress = FALSE,
                iter = 1000, simple = FALSE)



train_2=train_2$sampled_data
```

```
nrow(train_2)


#RF with LHS : 510 train data


# load the library
library(caret)

# define training control
train_control_2 <- trainControl(method="cv",
                                number=5)


#data
x2 <- train_2[,-9]
train_2$class <- as.character(train_2$class)
train_2$class <- as.factor(train_2$class)
y2 <- train_2$class

# train the model
model_2 <- train(x=x2, y=y2, trControl=train_control_2,
                 method="rf")

# summarize results
print(model_2)



predicted_table_2 <- predict(model_2, test[,-9])
table(observed = test[,9], predicted = predicted_table_2)


#Calcualte the accuracy


cm2<-table(observed = test[,9], predicted = predicted_table_2)
n2<-sum(cm2)
diag2<-diag(cm2)
accuracy_model_2 = sum(diag2) / n2
accuracy_model_2
```

```
#########################################################
With LHS in train data: from 768 keep (50%)
#########################################################



library(clhs)
train_3 <- clhs(train, size = 380, progress = FALSE,
                iter = 1000, simple = FALSE)



train_3=train_3$sampled_data
nrow(train_3)


#RF with LHS : 380 train data


# load the library
library(caret)

# define training control
train_control_3 <- trainControl(method="cv",
                                number=5)

#data
x3 <- train_3[,-9]
train_3$class <- as.character(train_3$class)
train_3$class <- as.factor(train_3$class)
y3 <- train_3$class

# train the model
model_3 <- train(x=x3, y=y3, trControl=train_control_3,
                 method="rf")

# summarize results
print(model_3)
```

```
predicted_table_3 <- predict(model_3, test[,-9])
table(observed = test[,9], predicted = predicted_table_3)



cm3<-table(observed = test[,9], predicted = predicted_table_3)
n3<-sum(cm3)
diag3<-diag(cm3)
accuracy_model_3 = sum(diag3) / n3
accuracy_model_3


##########################################################
With LHS in train data: from 768 keep 250 (appr%33)
##########################################################



library(clhs)
train_4 <- clhs(train, size = 250, progress = FALSE,
                iter = 1000, simple = FALSE)



train_4=train_4$sampled_data
nrow(train_4)


#RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_4 <- trainControl(method="cv",
                                number=5)

#data
```

```
x4 <- train_4[,-9]
train_4$class <- as.character(train_4$class)
train_4$class <- as.factor(train_4$class)
y4 <- train_4$class

# train the model
model_4 <- train(x=x4, y=y4, trControl=train_control_4,
                 method="rf")

# summarize results
print(model_4)



predicted_table_4 <- predict(model_4, test[,-9])
table(observed = test[,9], predicted = predicted_table_4)



cm4<-table(observed = test[,9], predicted = predicted_table_4)
n4<-sum(cm4)
diag4<-diag(cm4)
accuracy_model_4 = sum(diag4) / n4
accuracy_model_4


#########################################################
With LHS in train data: from 768 keep 150 (apprx 20%)
#########################################################



library(clhs)
train_5 <- clhs(train, size = 150, progress = FALSE,
                iter = 1000, simple = FALSE)



train_5=train_5$sampled_data
nrow(train_5)
```

```
#RF with LHS : 40 train data


# load the library
library(caret)

# define training control
train_control_5 <- trainControl(method="cv",
                                number=5)

#data
x5 <- train_5[,-9]
train_5$class <- as.character(train_5$class)
train_5$class <- as.factor(train_5$class)
y5 <- train_5$class

# train the model
model_5 <- train(x=x5, y=y5, trControl=train_control_5,
                 method="rf")

# summarize results
print(model_5)



predicted_table_5 <- predict(model_5, test[,-9])
table(observed = test[,9], predicted = predicted_table_5)



cm5<-table(observed = test[,9], predicted = predicted_table_5)
n5<-sum(cm5)
diag5<-diag(cm5)
accuracy_model_5 = sum(diag5) / n5
accuracy_model_5



accuracy<-c(accuracy_model_1,
            accuracy_model_2,
```

```
              accuracy_model_3,
              accuracy_model_4,
              accuracy_model_5)
names.arg=c("Model1(80%)",
              "Model2(65%)",
              "Model3(50%)",
              "Model4(33%)",
              "Model5(20%)")
barplot(accuracy,names=names.arg)
```