ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΤΟΝ ΔΙΑΧΥΤΟ ΥΠΟΛΟΓΙΣΜΟ

ΚΥΡΚΙΡΗΣ ΕΥΣΤΡΑΤΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κωνσταντίνος Κολομβάτσος
Επίκουρος Καθηγητής

Λαμία 10 Φεβρουαρίου έτος 2023

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

AUTOMATED MACHINE LEARNING IN DIFFUSION COMPUTATION

KIRKIRIS EFSTRATIOS

FINAL THESIS

ADVISOR

Kolomvatsos Konstantinos
Assistant professor

Lamia 10 February year 2023

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος πχ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (πχ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 10/2/2023

Ο Δηλών

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Table of Contents

ΠΕΡΙΛΗΨΗ

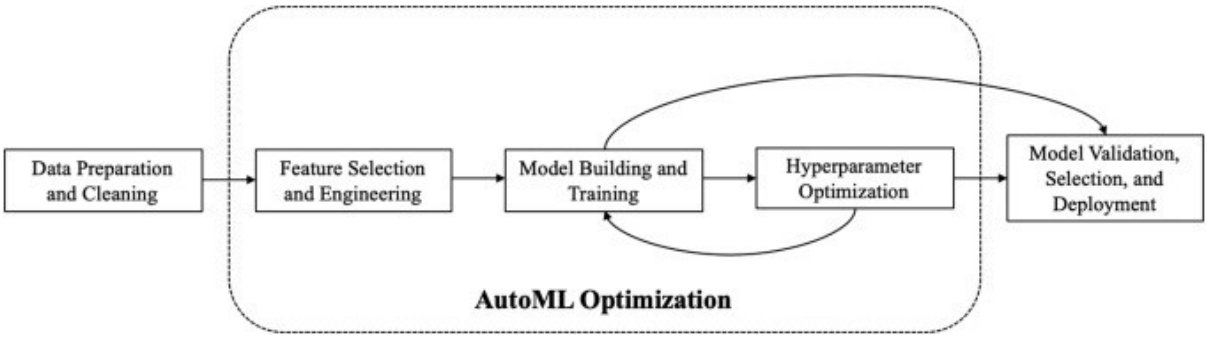
Ο στόχος αυτής της εργασίας είναι, να ενθαρρύνει την χρίση της αυτοματοποιημένης μηχανικής μάθησης από τους αρχάριους και έμπυρους επιστήμονες της μηχανικής μάθησης με το να τους παρέχει μοντέλα με ευκολία και σε μικρό χρονικό διάστημα, πάνω στον τομέα του διάχυτου υπολογισμού. Αρχικά, θα μελετήσουμε κάποιες γενικές έννοιες για το, τι είναι η αυτοματοποιημένη μηχανική μάθηση και τι προσπαθεί να επιτεύξει, πως λειτουργεί, και μερικά παραδείγματα βιβλιοθηκών και προγραμμάτων που παράγουν μοντέλα αυτόματα. Επίσης, θα μελετήσουμε τον τομέα του διάχυτου υπολογισμού με την περιγραφή του τρόπου λειτουργίας του και διάφορων παραδειγμάτων χρήσεων αυτών των μοντέλων. Έπειτα, αναφέρω την διαδικασία κατά την οποία χειρίστηκα τα άρθρα τα οποία χρησιμοποίησα από το google scholars καθώς, και μερικά παραδείγματα εφαρμογών του συνδυασμού της αυτοματοποιημένης μηχανικής μάθησης και του διάχυτου υπολογισμού πάνω σε διάφορους τομείς όπως, της ιατρικής και της αστρονομίας. Τέλος, ακολουθεί ένα πείραμα για το αν τα αυτοματοποιημένα μοντέλα μπορούν να ανταγωνιστούν τα μοντέλα που παράχθηκαν με χειρωνακτικό τρόπο, κάνοντας ομαδοποίηση σε φωτογραφίες από τηλεσκόπια.

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

(1.1 Αυτοματοποιημένη Μηχανική Μάθηση)

(Ενότητα 1.1.1 Βασικές Έννοιες)

Ο τομέας της αυτοματοποιημένης μηχανικής μάθησης είναι ένας σύγχρονος τομέας της μηχανικής μάθησης ο οποίος άρχισε να κάνει τα πρώτα βήματα του την περίοδο του 2013. Πολλοί οργανισμοί άρχισαν να στρέφονται σε αυτόν τον τομέα λόγο της συνεχούς αυξανόμενης ζήτησης της μηχανικής μάθησης, που με την σειρά αυτής, προκάλεσε την αύξηση της ζήτησης εμπείρων επιστημόνων. Όμως, η κάλυψη του ανθρώπινου δυναμικού δεν μπορεί να θεωρηθεί ως ένας από τους κύριους παράγοντες για την ανάπτυξη αυτού του τομέα. Αν κάποιος διαβάσει μερικά παραδείγματα πάνω σε αυτόν τον τομέα θα καταλάβει πως, ο στόχος της αυτοματοποιημένης μηχανικής μάθησης είναι, να κάνει ευρέως διαθέσιμη την χρήση της μηχανικής μάθησης από εταιρίες και απλούς χρήστες, χωρίς να χρειάζονται εξειδικευμένες γνώσεις πάνω σε αυτόν τον τομέα ή γενικά, πάνω στον τομέα των μεγάλων δεδομένων και της μηχανικής μάθησης. Επιπλέον, αυτός ο τομέας μπορεί να βοηθήσει και τους αναλυτές δεδομένων με το να, αυτοματοποιεί ορισμένες ενέργειες η οποίες θεωρούνται επίπονες όπως, η παράγωγη και τροποποίηση των μοντέλων, κερδίζοντας έτσι, μερικές εβδομάδες ή μήνες εργασίας με το να παράγουν αυτά τα μοντέλα σε περίοδο ημερών.



Εικόνα 1.1.1.α: Διαδικασία παραγωγής μοντέλου μηχανικής μάθησης και περιγραφή των τομών που μπορούν να αυτοματοποιηθούν.

Περιγραφή εικόνας:
Καθαρισμός και προετοιμασία των δεδομένων (Data preparation and cleaning). Σε αυτό το στάδιο ο αναλυτής θα βρεθεί αντιμέτωπος με την διαχείριση των δεδομένων. Αν και αυτό είναι το πρωταρχικό στάδιο, θα πρέπει να φροντίσει ότι τα δεδομένα θα πρέπει να είναι αξιοπιστία και σε άψογη κατάσταση καθώς, αυτά θα κρίνουν την απόδοση του μοντέλου. Κάποιες βασικές ενέργειες είναι:

- Καθαρισμός από ελλιπείς ή μηδενικές τιμές..
- Μορφοποίηση και αναγνώριση του είδους των δεδομένων.
- Αναγνώριση των καθοριστικών ετικετών που θα χρησιμοποιηθούν.

- Αναπροσαρμογή των δεδομένων που βρίσκονται εκτός κάποιας κλίμακας έναντι κάποιων άλλων με σκοπό, την αποφυγή θορύβου. Μπορούμε να εφαρμόσουμε διάφορους αλγορίθμους όπως, Min-Max, τροποποίηση κατά δεκαδικά κτλ.

Επιλογή και χειρισμός των χαρακτηριστικών (feature selection and engineering). Σε αυτό το στάδιο, ο αναλυτής θα επιλέξει τις ετικέτες ή θα δημιουργήσει καινούριες με βάση τις γνώσεις του στο πρόβλημα και τα δεδομένα που έχει στη κατοχή του. Έπειτα, θα τις τροφοδοτήσει στο μοντέλο.

Κατασκευή και εκπαίδευση μοντέλου (Model Building and Training). Σε αυτό το στάδιο, θα επιλέξουμε το είδος του μοντέλου ανάλογα με το είδος του προβλήματος μας. Για παράδειγμα, το πρόβλημα μπορεί να χαρακτηριστεί ως πρόβλημα ομαδοποιήσεις ή παλινδρόμησης. Έπειτα, θα πρέπει να εκπαιδεύσουμε το μοντέλο και να πάρουμε την απόδοση του. Αυτό είναι ίσος και το ποιο χρονοβόρο στάδιο την μηχανικής μάθηση καθώς, η επιλογή του σωστού μοντέλου χειρίζεται την κατασκευή πολλαπλών μοντέλων.

Αναπροσαρμογή των υπερ-ετικετών (Hyper-parameter Optimization). Σε αυτό το στάδιο, θα πρέπει να αναπροσαρμόσουμε τον αλγόριθμο του μοντέλου μας ώστε να πάρουμε την καλύτερη δυνατόν απόδοση. Για παράδειγμα, αλλαγή του K στον αλγόριθμο K-means.

Εκτίμηση, Επιλογή και Εφαρμογή μοντέλου (Model Validation, Selection and Deployment). Τώρα, αφού έχουμε το μοντέλο στα χέρια μας, ήρθε η ώρα να το εφαρμόσουμε στην πράξη. Βέβαια, το μοντέλο μας τίθεται ακόμα σε δοκιμή καθώς, μπορεί να διαπιστώσουμε τυχόν διαρροές προβλέψεων ή ελλιπείς εφαρμογή στο πρόβλημα μας. Έτσι, μετά από τις τελευταίες δοκιμές θα δοκιμάσουμε να δώσουμε τρόπους όπου οι χρήστες θα μπορούν να δοκιμάσουν το μοντέλο.

Όπως αναφέρθηκε παραπάνω, υπάρχουν πολλές λειτουργίες οι οποίες μπορούν να αυτοματοποιηθούν, συγκεκριμένα αυτές που ειδικεύονται στην δημιουργία και στην καλυτέρευση του μοντέλου αλλά, αυτές οι αυτοματοποιήσεις δεν μπορούν να θεωρηθούν πάντα ως οι ποιο βέλτιστες. Αυτό συμβαίνει διότι, δεν μπορούμε πάντα να δημιουργήσουμε ένα “μαύρο κουτί” από το οποίο, θα δημιουργούμε όλα τα μοντέλα μας. Τα προβλήματα τις καθημερινότητας τα οποία τίθενται προς επίλυση, πολλές φορές είναι περίπλοκα και διαφορετικά μεταξύ τους, πράγμα που σημαίνει ότι, ένας μονάχα αλγόριθμος δεν μπορεί να τα λύσει όλα. Επιπλέον, η χρίση ενός μονάχα αλγορίθμου καθιστά την εμφάνιση bias ως ένα γεγονός το οποίο δεν μπορούμε να γνωρίζουμε με μια πρώτη ματιά. Συγκεκριμένα, η αυτοματοποιημένη επιλογή των ετικετών που θα δίναμε στο μοντέλο και τα επιλεγμένα χαρακτηριστικά του μοντέλου, είναι κάποιοι από τους βασικούς λόγους που μπορούν να οδηγήσουν στην εμφάνιση bias. Εν κατακλείδι, ο κλάδος της αυτοματοποιημένης μηχανικής μάθησης δεν έχει καταφέρει να αυτοματοποιήσει πλήρως την μηχανική μάθηση, αν και κάποιοι οργανισμοί προσπαθούν να την κάνουν πλήρως αυτοματοποιημένη, τα περισσότερα προβλήματά που παρουσιάζονται τυχαίνουν πολλές φόρες να λύνονται με την βοήθεια των επιστημόνων.

(Ενότητα 1.1.2 Μοντέλα)

Όπως αναφέρθηκε, ο τομέα την αυτοματοποιημένης μηχανικής μάθησης, είναι ένας σχετικά σύγχρονος τομέας, πράγμα που σημαίνει ότι, ο τομέας βρίσκεται στα πρώτα στάδια ή ίσως, στα μέσα στάδια ανάπτυξης και εφαρμογής. Αυτό έχει ως αποτέλεσμα, τα μοντέλα της αυτοματοποιημένης μηχανικής μάθησης τα οποία αναπτύχθηκαν από ορισμένες εταιρίες, να διαφέρουν ως προς τον σκοπό τους, την διαθεσιμότητα τους άλλα και στην απόδοση τους. Για παράδειγμα, η Google παρέχει την δυνατότητα δημιουργίας μοντέλων μεσώ ενός cloud API αφού του έχουμε δώσει τα δεδομένα. Τέλος, άλλοι οργανισμοί όπως η Datarobot, προσπαθούν να αυτοματοποιήσουν πλήρως την διαδικασία της μηχανικής μάθησης, από την συλλογή των δεδομένων και των καθαρισμό τους μέχρι και την παραγωγή του μοντέλου.

Google Cloud's AutoML [1]

Μιας και η Google είναι μια από τις εταιρίες με την μεγαλύτερη μηχανή αναζήτησης, είναι λογικό κάποιος να υπόθεση πως, μια εταιρία τέτοιων μεγεθών θα κατέχει στη διάθεση της μεγάλους όγκους δεδομένων. Αυτό την φέρνει σε μια εύκολη θέση για την δημιουργία των μοντέλων καθώς, αξιόπιστα μοντέλα πρέπει να εκπαιδευτούν πάνω σε μεγάλους όγκους δεδομένων. Όμως, αυτό λειτουργεί και ως ένα μειονέκτημα διότι, οι τεράστιοι όγκοι δεδομένων θα πρέπει πρώτα να προετοιμαστούν και έπειτα να τροφοδοτηθούν στο μοντέλο, και αν μιλάμε για ένα περίπλοκο μοντέλο, η εκπαίδευση και η τροποποίηση, καθιστά την δημιουργία του μοντέλου χρονοβόρα. Έτσι, η Google για να διευκολύνει αυτήν την διαδικασία αλλά και να ενθαρρύνει καινούριους χρήστες ώστε να εμπλακούν στον τομέα την μηχανικής μάθησης, αποφάσισε να δημιουργήσει το δικό της αυτοματοποιημένο μοντέλο με όνομα "Vertex AI".

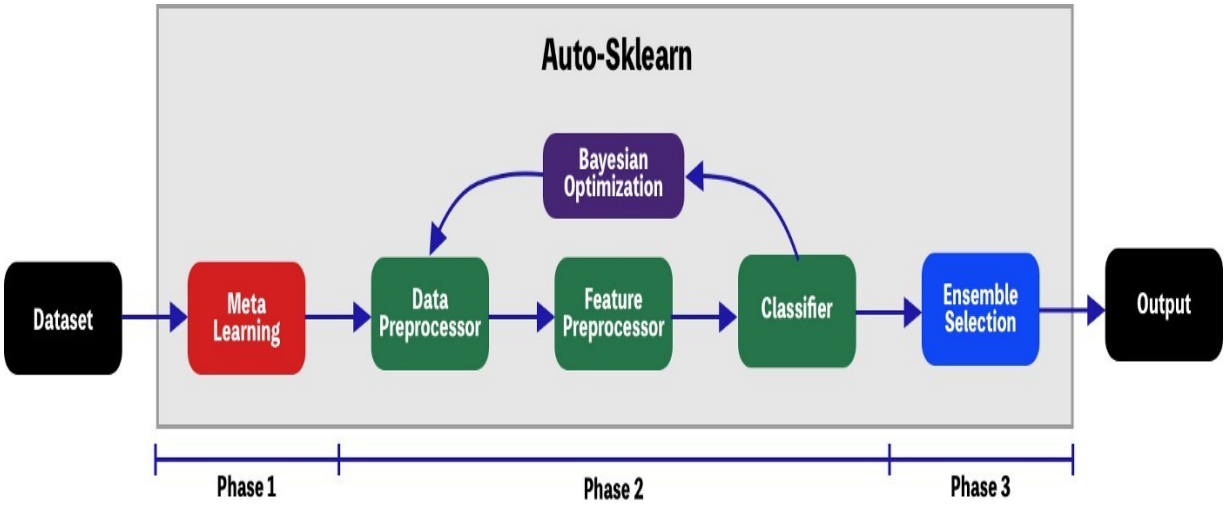
Αρχικά, το μοντέλο μπορεί να χρησιμοποιηθεί μεσώ του Google Cloud. Από εκεί και πέρα, το μόνο που έχουμε να κάνουμε είναι να του δώσουμε τα δεδομένα μας σε οποιαδήποτε μορφή κι' αν είναι, φωτογραφίες, κείμενα, στήλες κτλ. Αν για παράδειγμα, έχουμε ένα πρόβλημα ομαδοποίησης σε δεδομένα μέσα σε στήλες, θα πρέπει να έχουμε καθαρίσει τα δεδομένα και να έχουμε βάλει ετικέτες στις στήλες. Πρέπει να σημειωθεί πως, αν και η διαδικασία της επιλογής των ετικετών είναι αυτοματοποιημένη, τις περισσότερες φορές ο αλγόριθμος δεν μπορεί να διακρίνει με ευκρίνεια τις ετικέτες που ίσως είναι προφανές σε εμάς, οπότε θα πρέπει να ορίσουμε προφανείς ετικέτες σύμφωνα με το πρόβλημα μας για να πάρουμε ακριβείς αποτελέσματα. Έπειτα, ο αλγόριθμος θα επιλέξει τις ετικέτες και θα εφαρμόσει τον καλύτερο αλγόριθμο ομαδοποίησης. Τέλος, θα μας δώσει τα αποτελέσματα ως ένα μοντέλο το οποίο θεωρεί ότι είναι το ποιο καλύτερο για τις προβλέψεις του προβλήματος.

Open-source auto-sklearn [2][3]

Η sklearn είναι μια από τις από τις ποιο διαδεδομένες βιβλιοθήκες της python οι οποία προσφέρει πολλές λειτουργίες στον τομέα της μηχανικής μάθησης. Μια από τις πολλές υπό βιβλιοθήκες που προσφέρει η sklearn είναι και η auto-sklearn από την οποία, μπορούμε να δημιουργήσουμε μοντέλα με τους τρόπους τους οποίους έχουμε αναφέρει.

Δημιουργήθηκε το 2015 ως ένα ανοικτό κομμάτι κώδικα με κύριο δημιουργώ τον Matthias Feurer και δημοσιεύτηκε ως ένα ερευνητικό άρθρο με όνομα "Efficient and Robust

Automated Machine Learning". Η auto-sklearn σε αντίθεση με άλλα μοντέλα, χρησιμοποιεί τον αλγόριθμο του Bayesian Optimization από τον οποίο προσπαθεί να βρει την κατάλληλη διαδικασία παραγωγής, δηλαδή την επιλογή των κατάλληλων ετικετών, είδους και χαρακτηριστικών, για την δημιουργία του καλύτερου μοντέλου. Βέβαια, αυτή η διαδικασία καθιστάτε πολύ πολύπλοκη και χρονοβόρα καθώς, για την εύρεση της κατάλληλης διαδικασίας παραγωγής θα πρέπει να γνωρίζουμε ποια είναι η χειρότερη, η μέση και η καλύτερη. Αυτή την γνώση μπορούμε να την αποκτήσουμε μέσω της τεχνικής του Grid Search το οποίο, δημιουργούμε πολλαπλά μοντέλα με διαφορετικά χαρακτηριστικά, απ' τα οποία θα επιλέξουμε το καλύτερο. Όμως, αυτή η τεχνική παρουσιάζει μερικά προβλήματα σε ότι αφορά περίπλοκα μοντέλα διότι, θα πρέπει να στραφούμε σε ένα ιδικό κομμάτι των μοντέλων τα οποία θα περιέχουν σταθερές τιμές. Έτσι, για να λυθεί αυτό το πρόβλημα τις εύρεσης της καλύτερης διαδικασία παραγωγής, δημιουργήθηκαν οι γενετικοί αλγόριθμοι η οποίοι χρησιμοποιούνται στο αλγόριθμο TPOT και βρίσκουν την καλύτερη διαδικασία παραγωγής σε λιγότερο χρόνο αλλά, χρησιμοποιούν προκαθορισμένα δεδομένα.

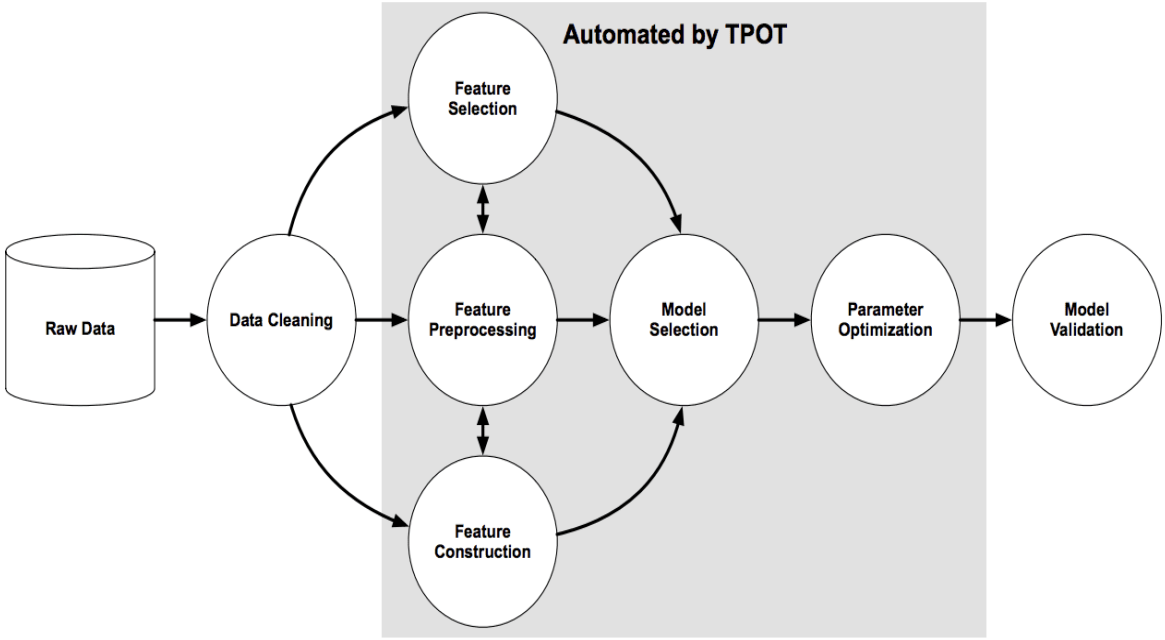


Εικόνα 1.1.2.α: Διαδικασία παραγωγής ενός αυτοματοποιημένου μοντέλου με την χρήση της Auto-sklearn.

Περιγραφή εικόνας:
Αρχικά, το στάδιο 1 θα προσπαθεί να βρει τις κατάλληλες τιμές από τα καλύτερα μοντέλα έτσι ώστε, να μειώσουμε το μέγεθος της έρευνας. Ύστερα, αυτές οι τιμές θα χρησιμοποιηθούν στο στάδιο 2 για να παραχθούν κάποια bayesian models τα οποία, θα επιλέξουν την καλύτερη διαδικασία παραγωγής για την δημιουργία του μοντέλου στο στάδιο 3. Θα σημειωθεί πως, αφού στο στάδιο 1 και 2 δεν εμπλέκεται κανένας αλγόριθμος, μας δίνεται η δυνατότητα να διαμορφώσουμε τα χαρακτηριστικά αυτών ανάλογα με το πρόβλημα μας.

TPOT [4][5]
Το μοντέλο TPOT είναι μια βιβλιοθήκη της python που όπως και οι auto-sklearn, προσφέρει την δυνατότητα δημιουργίας μοντέλων με το να βρίσκει την κατάλληλη διαδικασία παραγωγής για τα δεδομένα μας. Σε αντίθεση με την auto-sklearn, το TPOT χρησιμοποιεί

μια μορφή ενός δένδρου για να περιγράψει ποιες λειτουργίες πρέπει να εκτελεστούν όπως, επεξεργασία δεδομένων και το είδος των μοντέλων και παραμέτρων, για να παραχθεί η καλύτερη διαδικασία παραγωγής. Επίσης, έναντι του Bayesian Optimization, το TPOT χρησιμοποιεί μια μορφή γενετικών αλγορίθμων από τους οποίους, προσπαθούν να αναπροσαρμόσουν αυτόματα την δομή των δεδομένων και το είδους των μοντέλων που θα χρησιμοποιηθούν ώστε, να παραχθεί ένα supervised μοντέλο με την καλύτερη απόδοση σε ότι αφορά το classification. Τέλος, ως προ την επιλογή της καλύτερης διαδικασίας παραγωγής, οι γενετικοί αλγόριθμοι έχουν πάρει την γενικοί τους ιδέα από το κανόνα του Δαρβίνου. Συγκεκριμένα, θα παράγουμε κάποια μοντέλα τα οποία θα έχουν μια τιμή από μια fitness function. Έπειτα, θα αναπροσαρμόσουμε τις τιμές ώστε να βρίσκονται μεταξύ του 0 και 1 και το άθροισμα να είναι 1 και θα επιλέξουμε n τυχαία μοντέλα όπου, το fitness value να είναι μεγαλύτερο του n και ύστερα θα προβούμε σε συζεύξεις των μοντέλων για να παράγουμε mutations.



Εικόνα 1.1.2.β: Διαδικασία παραγωγής μοντέλου με αυτοματοποιημένων τρόπο με την χρήση του αλγορίθμου TPOT.

H20 [6] [7]

Η H2O είναι ακόμα μια βιβλιοθήκη η οποία μας παρέχει την δυνατότητα παραγωγής μοντέλων. Αυτή η βιβλιοθήκη μπορεί να χρησιμοποιηθεί όχι μόνο στην python αλλά και σε άλλες γλώσσες όπως, Scala, R αλλά, μπορεί επίσης να χρησιμοποιηθεί μέσω ενός GUI μέσα σε servers η οποίοι τρέχουν spark ή AWS. Ακόμα, εκτός από τον τομέα της μηχανικής μάθησης, η H20 προσπαθεί να επεκταθεί και στον τομέα των νευρωνικών δικτύων, δημιουργώντας νευρωνικά δίκτυα με έναν αυτοματοποιημένο τρόπο. Αυτό το επιτυγχάνει με το να τρέχει αντίγραφα των στοιχείων που έχουμε δόση στο δίκτυο σε έναν ή όλους τους κόμβους ταυτόχρονα και έπειτα, μεταβάλει τις τιμές των νευρώνων του δικτύου με τις

μέσες τιμές που πήραμε.

Αναφορικά περιγράφονται οι λειτουργείς της H20:

- Παρέχει λειτουργίες οι οποίες εξηγούν τους αλγορίθμους που χρησιμοποιεί η βιβλιοθήκη για να παράξει ένα μοντέλο ή μια ομάδα μοντέλων.

- Αφού επιλέξει τις ετικέτες τις οποίες κρίνει ως τις ποιο χρήσιμες από τα δεδομένα που του έχουμε παραδώσει, προβαίνει στην εκπαίδευση των μοντέλων με βάση κάποιων αλγορίθμων όπως ο GBMs και DNNs.

- Έπειτα θα πρέπει να επιλέξει τα καλύτερα μοντέλα. Αυτό θα το κάνει μεσώ 2 ομάδων που θα δημιουργήσει και θα περιέχουν, στην πρώτη όλα τα μοντέλα με την καλύτερη απόδοση και στην δεύτερη, τα μοντέλα με την καλύτερη απόδοση ανά αλγόριθμο.

- Τέλος, επιστρέφει όλα τα μοντέλα από τις δυο ομάδες ανακατανεμημένα προς την απόδοση.

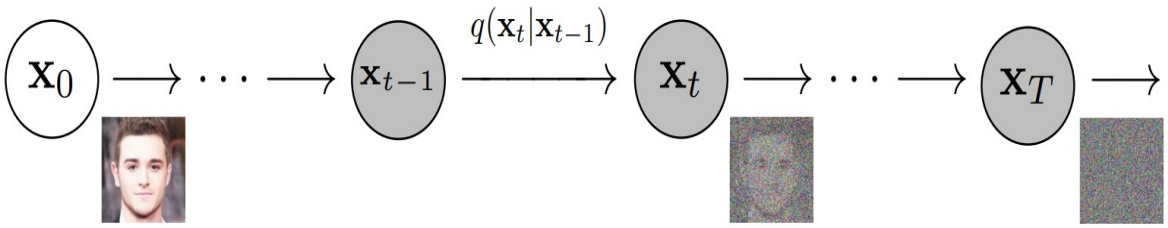
(1.2 Διάχυτος Υπολογισμός)

(Ενότητα 1.2.1 Βασικές Έννοιες)

Ο διάχυτος υπολογισμός και ειδικότερα τα διάχυτα μοντέλα, έχουν ως κύριο σκοπό να δημιουργούν δεδομένα από την αρχή, βασιζόμενα σε δεδομένα τα οποία είχαν εκπαιδευτεί. Αυτό το επιτυγχάνουν με το να, κάνουν προσθήκη Γκαουσιανού θορύβου στα δεδομένα μέχρι αυτά μην αναγνωρίζονται και έπειτα, προσπαθούν να τα επαναφέρουν ξανά στην αρχική τους κατάσταση με την χρήση ενός τυχαίου “σπόρου”. Τα διάχυτα μοντέλα μπορεί να θεωρηθούν ως μια αναβάθμιση των μοντέλων παραγωγής δεδομένων καθώς, μερικά μοντέλα όπως τα GANs και τα VAEs, παρουσιάζουν μερικά προβλήματα στην σταθερότητα των δεδομένων παραγωγής ή μπορεί να στηρίζονται στο surrogate loss για να τα παράγουν. Σε αντίθεση με αυτά τα μοντέλα, τα διάχυτα μοντέλα στηρίζονται στην θεωρεία της μη-ισορροπημένης θερμοδυναμικής όπου, εφαρμόζοντας μια μαρκοβιανή αλυσίδα σταδιακής διοχέτευσης του θορύβου, αυτή η διαδικασία αποκτά μια σταθερότητα. Επιπλέον, αξίζει να σημειωθεί πως, τα διάχυτα μοντέλα δεν χρειάζονται περισσότερη εκπαίδευση για την καλυτέρευση της προσθήκης του θορύβου, κάτι το οποίο είναι αναγκαίο σε κάποια άλλα μοντέλα όπως τα VAEs. Όμως, η χρήση της μαρκοβιανής αλυσίδας λειτουργεί ως ένα διπλό ξίφος καθώς, τα διάχυτα μοντέλα εξακολουθούν να είναι βραδύτερα στην παραγωγή των δεδομένων σε ότι αφορά τα GANs. Τέλος, κάποια άλλα μοντέλα όπως το Dall-E 2, Stable Diffusion και ChatGBT, έχουν αρχίσει να χρησιμοποιούνται όλο ένα και περισσότερο στην καθημερινή και επαγγελματική μας ζωή, με την παραγωγή κειμένων, φωτογραφιών και σκηνών από ταινίες, με το να τους δίνουμε μια περιγράφει για το τι θέλουμε να πάρουμε. Από την πρακτική γωνιά, αυτά τα μοντέλα μπορεί να μας διευκολύνουν ή ακόμα και, να καλυτερεύουν την ζωή ορισμένων ανθρώπων

που πάσχουν από δυσλεξία και άλλες νοητικές ασθένειες αλλά, αυτά τα μοντέλα εξακολουθούν να μην έχουν τις λεπτομέρειες τις οποίες θα παίρναμε από την εργασία των ανθρώπων.

Όπως αναφέρθηκε παραπάνω, τα διάχυτα μοντέλα για να παράγουν καινούρια δεδομένα, πρέπει να περάσουν από μια μορφή εκπαίδευσης. Συγκρινόμενα, αυτή η διαδικασία χωρίζετε σε 2 φάσης, στην σταδιακή προσθήκη θορύβου μέσω μιας μαρκοβιανής αλυσίδας και στην οπισθοδρομική αφαίρεση του θορύβου όπου, ο θόρυβος μετατρέπετε σε μορφές δεδομένων. Στην πρώτη φάση, το μοντέλο με την σταδιακή προσθήκη γκαουσιανού θορύβου στα δεδομένα, προσπαθεί να δημιουργήσει μια πλήρως αλλοιούμενη μορφή των αρχικών δεδομένων.



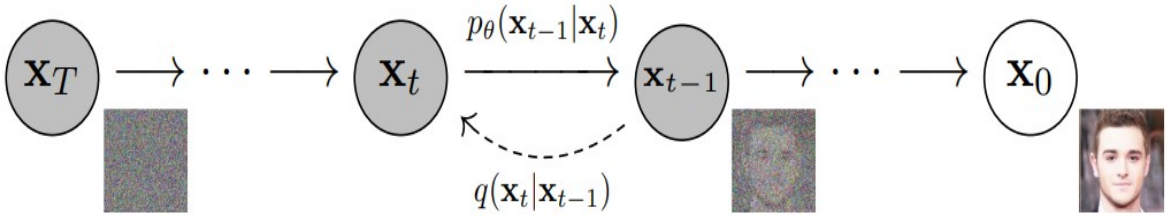
Εικόνα1.2.α: Διαδικασία προσθήκης γκαουσιανού θορύβου με την χρήσης της μαρκοβιανής αλυσίδας.

Η γκαουσιανή συνάρτηση χρησιμοποιείτε, καθώς και άλλες γκαουσιανές σχέσεις λόγο του ότι, η μικρές σταδιακές προσθήκες θορύβου μας δίνουν μερικά πλεονεκτήματα σταθερότητας και παραμετροποίησης της διαδικασίας:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Εικόνα 1.2.β: Μαρκοβιανή συνάρτηση προσθήκης γκαουσιανού θορύβου.

Στο επόμενο στάδιο, θα πρέπει να πάρει τον πλέον γκαουσιανό θόρυβο και να αρχίσει να παράγει δεδομένα τα οποία υποθέτουμε ότι είναι ίδια με τα αρχικά.



Εικόνα 1.2.γ: Οπισθοδρομική διαδικασία αφαίρεσης γκαουσιανού θορύβου.

Αυτό το καταφέρνει με το να μαθαίνει την κοινή κατανομή του γκαουσιανού θορύβου πάνω στα δεδομένα μετά την απαλοιφή αυτού, γνωρίζοντας βέβαια την προηγούμενη κατάστασή, δηλαδή την κατάσταση πριν την αλλοίωση των δεδομένων:

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Εικόνα 1.2.δ: Συνάρτηση πιθανότητας εύρεσης των χρονικών παραμέτρων των γκαουσιανών μεταβάσεων.

Ως προς τα παραδείγματα εφαρμογών των διάχυτων μοντέλων, πολλά άρθρα στο google scholars αναφέρονταν σε ιατρικούς τομείς με κύρια δεδομένα τις φωτογραφίες. Για παράδειγμα, μια αναζήτηση σε άρθρα μετά το 2018 με τις συγκεκριμένες λέξεις “diffusion machine learning", παίρνουμε 16.7 χιλιάδες αποτελέσματα σχετικά με τον διάχυτο υπολογισμό και την μηχανική μάθηση. Αν κάποιος πρόσθεση την λέξει “MRI”, θα λάβει 10.3 χιλιάδες αποτελέσματα σχετικά με τους μαγνητικούς τομογράφους, και 13.4 χιλιάδες σε όλο τον τομέα της ιατρικής. Ένα κύριο γεγονός αυτής της εξειδίκευσης είναι ότι, οι περισσότερες νοητικές ασθένειες αναγνωρίζονται είτε με φυσικό τρόπο, ωτοψίες και τα φυσικά αποτελέσματα στους ασθενείς, ή με μηχανικό τρόπο όπως οι μαγνητικοί τομογράφοι. Όμως, τα μηχανήματα μπορεί να μην παρέχουν πάντα την μέγιστη ακρίβεια λόγο της έλλειψης ευκρίνειας στις φωτογραφίες που παίρνουμε. Με αυτό το γεγονός, μπορούμε να χρησιμοποιήσουμε μοντέλα διάχυτου υπολογισμού τα οποία, μετά από λίγα λεπτά επεξεργασίας των φωτογραφιών, θα μπορέσουν να δώσουν μια δεύτερη γνώμη στους ιατρούς με το να τους παρέχουν φωτογραφίες με μεγαλύτερη ευκρίνεια. Επιπλέον, αυτά τα μοντέλα είναι σε θέση να δώσουν ακριβής αποτέλεσμα ακόμα και αν δεν έχουμε πολλά δεδομένα καθώς, τα αρχικά δεδομένα για το πρόβλημα μας τείνουν να έχουν μεγαλύτερο βάρος έναντι των δεδομένων που μπορεί να προσθέσουμε αργότερα στο μοντέλο. Από την άλλη μεριά, αυτά τα μοντέλα τείνουν να παρουσιάζουν ορισμένα προβλήματα ως προς το πως θα επιλέξουν τα αποτελέσματα. Αυτό το πρόβλημα χαρακτηρίζεται ως κατάρρευση των ορίων επιλογών όπου, το μοντέλο με το πέρασμα του χρόνου θα χρειάζεται όλο ένα και λιγότερα στοιχεία για να πάρει τις αποφάσεις του, πράγμα που μειώνει την ακρίβεια επιλογής των σωστών αποφάσεων. Επιπλέον, ένας άλλος παράγοντας που μπορεί να επηρεάσει την διαδικασία επιλογής του μοντέλου είναι ο θόρυβος που μπορεί να δημιουργηθεί από το μοντέλο ή από εξωτερικούς παράγοντες όπου, το μοντέλο μπορεί να αποφασίσει βασιζόμενο στα στοιχεία που θα του δώσουμε αργότερα, χωρίς να κοιτάξει τα προηγούμενα. Το ερώτημα του, αν το μοντέλο θα επηρεαστεί από την επιρροή μας, είναι ένα ερώτημα που χρειάζεται περαιτέρω μελέτη καθώς, μερικές έρευνες δείχνουν ότι ο θόρυβος παραμένει θόρυβος, είτε δημιουργηθεί από εμάς είτε από το μοντέλο. Εν κατακλείδι, τα διάχυτα μοντέλα μπορούν να προσφέρουν μεγάλη βοήθεια στους επιστήμονες που ειδικεύονται στην διάγνωση των νοητικών ασθενειών με το να προσφέρουν μια δεύτερη γνώμη αλλά, λόγο τις μεγάλης πολυπλοκότητας αυτού του τομέα, τα μοντέλα θα χρειαστούν να περάσουν από πολλά πειράματα και επεξεργασίες για να παρθούν σε θέση να λάβουν την πλήρη ευθύνη της διάγνωσης.

ΚΕΦΑΛΑΙΟ 2 Βιβλιογραφική Επισκόπηση

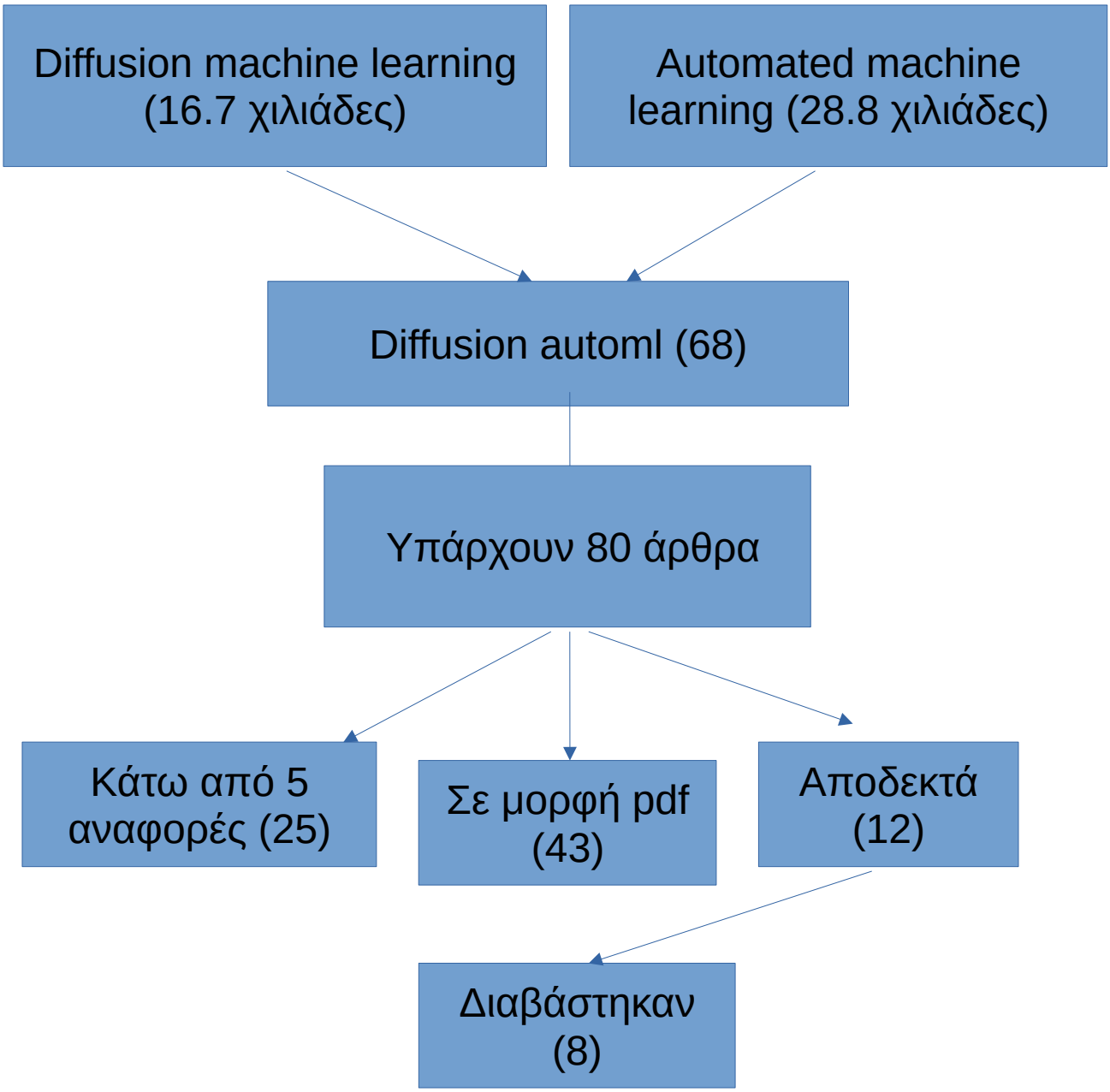
Η βιβλιογραφία την οποία επέλεξα να κάνω την έρευνα επάνω στον διάχυτο υπολογισμό και την αυτοματοποιημένη μηχανική μάθηση, έγινε μεσώ του google scholars και επέλεξα τα άρθρα με βάση μερικών κρητικών:

- Φίλτρα: since 2018 | by relevance | review articles. Αυτά τα φίλτρα επιλέχθηκαν διότι, επέλεξα να επικεντρωθώ στα ποιο σύγχρονα άρθρα.

- Τα άρθρα τα οποία δεν είναι διαθέσιμα, δηλαδή βρίσκονταν σε μορφή PDF, απορρίφθηκαν για λόγους ασφάλειας ή ήταν διαθέσιμα υπό πληρωμή. Αν η περιγραφή της έρευνας ήταν αρκετά κατανοητή για τους στόχους και τα αποτελέσματα της έρευνας, το άρθρο επιλέχθηκε για αναφορά ή για περιγραφή.

- Άρθρα τα οποία είχαν κάτω από 5 αναφορές απορρίφθηκαν για λόγους αξιοπιστίας.

Το παρακάτω διάγραμμα περιγράφει την διαδικασία επιλογής:



Ο παρακάτω πίνακας παρουσιάζει συνοπτικά κάποιες από τις τεχνολογίες οι οποίες υιοθετούνται στα προαναφερόμενα άρθρα της βιβλιογραφίας.

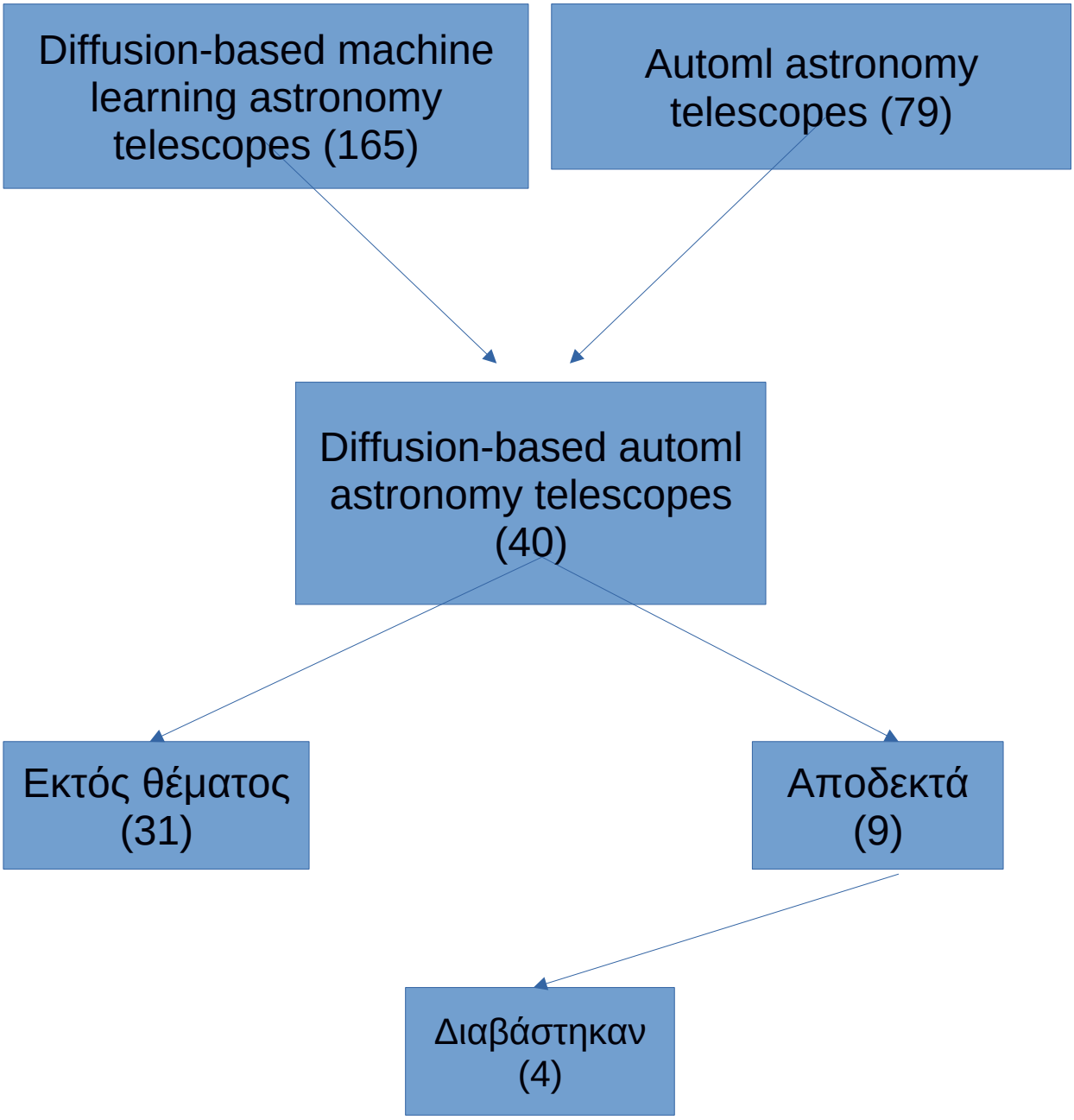
Άρθρο	**Τεχνολογία**	**Πεδίο Εφαρμογής**
[8]	Αυτοματοποιημένη μηχανική μάθηση	-
[9]	Κατανεμημένα μοντέλα δεδομένων	-
[11]	Κατανεμημένα μοντέλα δεδομένων	Ιατρική

[14]	Κατανεμημένα μοντέλα δεδομένων	-
[10]	Κατανεμημένα μοντέλα δεδομένων	Ιατρική
[15]	Κατανεμημένα μοντέλα δεδομένων	-

Για μια ποιο ειδικότερη αναφορά παραδειγμάτων για το πως η αυτοματοποιημένη μηχανική μάθηση και ο διάχυτος υπολογισμός μπορεί να χρησιμοποιηθεί σε διάφορους τομείς, θα αναφέρω παραδείγματα εφαρμογών στους τομείς τις ιατρικής καθώς, οι περισσότερες εφαρμογές ειδικεύονται πάνω σε αυτόν τον τομέα. Πώς όμως ο διάχυτος υπολογισμός μπορεί να συνδυαστεί με την αυτοματοποιημένη μηχανική μάθηση; Όπως γνωρίζουμε, η αυτοματοποιημένη μηχανική μάθηση προσπαθεί να δημιουργήσει την καλύτερη διαδικασία παραγωγής ενός μοντέλου με την μεγαλύτερη ακρίβεια, και αφού εμπλέκουμε τον διάχυτο υπολογισμό, πολύ πιθανό είναι να χρησιμοποιήσουμε φωτογραφίες ως δεδομένα. Ένα πείραμα εφαρμογής είναι, η παραγωγή ενός μοντέλου χρησιμοποιώντας τον αλγόριθμο TPOT για την εύρεση των αλλαγών που έχει υποστεί ο εγκέφαλος κατά την διάρκεια της γήρανσης [11]. Χρησιμοποιώντας τον αλγόριθμο TPOT για την αυτοματοποίηση της παραγωγής του μοντέλου, το πείραμα κατάφερε να δώσει μοντέλα τα οποία μπορούσαν να διακριβώσουν τις διαφορές τις γήρανσης σε ένα μεγάλο εύρος των ηλικιών. Επιπλέον, τα μοντέλα κατάφεραν να δώσουν κάποια επιπλέον στοιχεία τα οποία μπορεί να οδηγήσουν σε αλλαγές του εγκεφάλου, τα οποία στοιχεία δεν ήταν εξαρχής αναγνωρίσιμα. Ως προς την μεθοδολογία, τα δεδομένα τα οποία τροφοδοτήθηκα ήταν περιοχές του εγκεφάλου τα οποία ύστερα διασπάστηκαν σε δεδομένα προς εκπαίδευση και σύγκριση. Έπειτα ο αλγόριθμός TPOT, έκανε τις κατάλληλες ενέργειες όπως, επιλογή των χαρακτηριστικών που τους είχαν ορίσει και την παραγωγή κάθε γενιάς των μοντέλων και μετασχηματισμό αυτών, για να επιλέξει το καλύτερο μοντέλο από κάθε γενιά. Τέλος, με την παραγωγή της τελευταίας γενιάς, το TPOT θα επιστρέψει κάποια δεδομένα, μέσα σε αυτά είναι η διαδικασία παραγωγής με το μικρότερο μέσο ποσοστό λάθους. Αυτή η διαδικασία παραγωγής έπειτα χρησιμοποιήθηκε για να παράξει επιπλέον μοντέλα.

Για την εφαρμογή του διάχυτου υπολογισμού και της αυτοματοποιημένης μηχανικής μάθησης στον τομέα που θα επικεντρωθεί το πείραμα, δηλαδή στον τομέα της αστρονομίας, λήφθηκαν υπόψιν όλα τα άρθρα τα οποία βρίσκονται εντός θέματος λόγο του ότι, υπάρχει ένας μειωμένος αριθμός άρθρων που αναφέρονται στην χρήση αυτών των δυο τομέων στην αστρονομία.

Αποτελέσματα εφαρμογής των δυο τομέων στην αστρονομία:



Μερικά παραδείγματα άρθρων είναι:

Άρθρο	**Τεχνολογία**	**Διεργασία**
[13]	Κατανεμημένα μοντέλα δεδομένων	Παρακολούθηση ουράνιων σωμάτων
[16]	Αυτοματοποιημένη μηχανική μάθηση	Παρακολούθηση θέσεων των ουράνιων σωμάτων

Ειδικότερα, ο τομέας της αστρονομίας έχει ως κύριο σκοπό, την κατανόηση του σύμπαντος για το πως λειτουργεί ή ακόμα, και πως δημιουργήθηκε με το να μελετά ορισμένα διαστημικά σώματα υπό απόσταση. Τα κύρια όργανα μελέτης αυτού του τομέα είναι αισθητήρες κοσμικής ακτινοβολίας, τα οποία μπορούν να καταγράψουν ακτίνες γάμα από εκατοντάδες έτη φωτός και τηλεσκοπία όπως το James Webb. Από αυτά τα δυο εργαλεία, θα επικεντρωθώ κυρίως στα τηλεσκόπια καθώς, από αυτά μπορούμε να αποκτήσουμε φωτογραφικά δεδομένα τα οποία μπορούμε να επεξεργαστούμε με τα διάχυτα μοντέλα.
Ως προς την χρήση της αυτοματοποιημένης μηχανικής μάθησης και του διάχυτου υπολογισμού σε αυτόν τον τομέα, τα άρθρα τα οποία υπάρχουν στο google scholars είναι αρκετά ελάχιστα, 165 αποτελέσματα με “machine learning diffusion-based astronomy telescopes” και 40 με “automl astronomy telescopes” και με μια μικρή ματιά στις πρώτες 8 σελίδες παρατήρησα ότι, μόνο 9 από αυτά στρέφονται πάνω στον τομέα της αυτοματοποιημένης μηχανικής μάθησης, πράγμα που σημαίνει ότι, η χρήση της αυτοματοποιημένης μηχανικής μάθησης στον τομέα της αστρονομίας είναι ένα καινούριο εργαλείο. Μερικά παραδείγματα εφαρμογών στον τομέα της αστρονομίας είναι, με τα δεδομένα τα οποία έχουμε συλλέξει, είτε από τηλεσκόπια είτε από διάφορους αισθητήρες, μπορούμε να απαλείψουμε τον θόρυβο που μπορεί να έχει προκληθεί από κοσμικές ακτινοβολίες ή από το ίδιο το υλικό. Επιπλέον, αν θέλουμε να μελετήσουμε κάποια συγκεκριμένα ουράνια σώματα όπως τους πλανήτες, μπορούμε να αφαιρέσουμε όλα τα αστέρια για την διευκόλυνση της μελέτης. Ακόμα, η δημιουργία ενός αλγορίθμου με όνομα “GaMPEN” [12], κατάφερε να βρει το μέγιστο μέγεθος καθώς και την γενική μορφή που θα έχει ένας γαλαξίας, χωρείς να γνωρίζουμε πολλά στοιχειά. Τέλος, ένας ακόμα αλγόριθμος βασιζόμενος στον διάχυτο υπολογισμό με όνομα “DeCiSpOT” [13], κατάφερε να έρθει κοντά στην ακρίβεια του “Kalman filter” σε ότι άφορα την αναζήτηση και παρακολούθηση των ουράνιων σωμάτων, χρησιμοποιώντας μικρότερο αριθμό αισθητηρίων και δεδομένων.

ΚΕΦΑΛΑΙΟ 3 Διεξαγωγή Πειράματος

(3.1 Στόχος)

Ο στόχος του πειράματος είναι, η δημιουργία ενός μοντέλου με αυτοματοποιημένο τρόπο το οποίο θα έχει την καλύτερη δυνατή ακρίβεια σε σύγκριση με ένα άλλο μοντέλο που παράχθηκε χειρωνακτικά. Συγκεκριμένα, το μοντέλο θα δέχεται μερικές φωτογραφίες από ένα τηλεσκόπιο και θα επιστέφει εάν μέσα σε αυτές τις φωτογραφίες αναγνωρίζει αστέρι ή έναν γαλαξία.

(3.2 Διεξαγωγή)

(Ενότητα 3.2.1 Περίληψη Υλοποίησης)

Αρχικά, θα πρέπει να δούμε τα δεδομένα μας [17]. Θα βρεθούμε αντιμέτωπη με 4000 φωτογραφίες οι οποίες είναι ήδη κατηγοριοποιημένες σε ομάδες, galaxy και star και έχουν

resolution: 64x64 και είναι άχρωμες. Για παράδειγμα:

galaxy

Αυτό μας βοηθά ως προς την επεξεργασία των δεδομένων καθώς, δεν χρειαζόμαστε να ορίσουμε τις ομάδες ή ένα συγκεκριμένο resolution ή να εμπλέξουμε τα χρώματα στα RGB options. Έπειτα, χώρισα τις φωτογραφίες σε training και validation data frames και τέλος, τις πέρασα στο μοντέλο που μου έβγαλε η ImageClassifier function του autokeras. Με αυτή την απλή διαδικασία, λάβαμε ένα μοντέλο βαθιάς μηχανικής μάθησης το οποίο εκτελεί ομαδοποίηση των φωτογραφιών με ακρίβεια 76%.

(Ενότητα 3.2.2 Επισκόπηση Κώδικα)

```
import tensor flow as tf
from autokeras import ImageClassifier

if __name__ == "__main__":
    # Εισάγουμε τα δεδομένα για εκπαίδευση
    train_df = tf.keras.utils.image_dataset_from_directory('path', batch_size= 2,
validation_split=0.2, subset="training", seed=123)

>>Found 3986 files belonging to 2 classes.
>>Using 3189 files for training.

    # Εισάγουμε τα δεδομένα για επιβεβαίωση
    val_df = tf.keras.utils.image_dataset_from_directory('path', batch_size= 2,
validation_split=0.2, subset="validation", seed=123)

>>Found 3986 files belonging to 2 classes.
>>Using 797 files for validation.


    # Επιλογή του μοντέλου
    model = ImageClassifier(num_classes=2, project_name="image_classifier",
max_trials=1, directory='path', objective="val_loss", overwrite=False)

>>Search: Running Trial #1
>> Value |Best Value So Far |Hyperparameter

>> vanilla |? |image_block_1/block_type
```

>> True |? |image_block_1/normalize

>> False |? |image_block_1/augment

>> 3 |? |image_block_1/conv_block_1/kernel_size

>> 1 |? |image_block_1/conv_block_1/num_blocks

>> 2 |? |image_block_1/conv_block_1/num_layers

>> True |? |image_block_1/conv_block_1/max_pooling

>> False |? |image_block_1/conv_block_1/separable

>> 0.25 |? |image_block_1/conv_block_1/dropout

>> 32 |? |image_block_1/conv_block_1/filters_0_0

>> 64 |? |image_block_1/conv_block_1/filters_0_1

>> flatten |? |classification_head_1/spatial_reduction_1/reduction_type

>> 0.5 |? |classification_head_1/dropout

>> adam |? |optimizer

>> 0.001 |? |learning_rate

```
# Εισαγωγή των δεδομένων στο μοντέλο
model.fit(x=train_df, y=None, epochs=2, callbacks=None, validation_data=val_df)
```

```
>>Epoch 1/2
>>1595/1595 [==============================] - 344s 216ms/step - loss: 0.8074
- accuracy: 0.7604 - val_loss: 0.5879 - val_accuracy: 0.7604

>>Epoch 2/2
>>1595/1595 [==============================] - 341s 214ms/step - loss: 0.5755
- accuracy: 0.7645 - val_loss: 0.5612 - val_accuracy: 0.7604

>>Trial 1 Complete [00h 11m 34s]
>>val_loss: 0.5611721873283386
```

(3.3 Αποτέλεσμα)

Καταλήξαμε στην δημιουργία ενός μοντέλου με:

Epoch 1/2
loss: 0.8074
accuracy: 0.7604
val_loss: 0.5879
val_accuracy: 0.7604

Epoch 2/2
loss: 0.5755
accuracy: 0.7645
val_loss: 0.5612
val_accuracy: 0.7604

Total elapsed time: 00h 11m 34s

Το μοντέλο RESNET34_ARCH [18] με 8 hidden layers και 25 epochs, χρησιμοποιήθηκε ως σύγκριση για το πείραμα [19]. Ως, προς τα αποτελέσματα, κατέληξε στα έξεις:

Epoch 1/25
loss: 6.7431
accuracy: 0.7406
val_loss: 382981111808.0000
val_accuracy: 0.7544

Epoch 2/25
loss: 0.5482
accuracy: 0.7638
val_loss: 3472421.2500
val_accuracy: 0.7544
.
.
.
Epoch 25/25
loss: 0.1768
accuracy: 0.9329
val_loss: 0.4900
val_accuracy: 0.7870

Συνολικός χρόνος ολοκλήρωσης των epochs: 60s

Ως προς θέμα χρόνου και ακρίβειας, το μοντέλο που παράχθηκε χειρωνακτικά είναι καλύτερο σε ότι αφορά το τελικό μοντέλο, πράγμα που είναι λογικό. Το αυτοματοποιημένο μοντέλο κατάφερε με μόνο 2 epochs να έρθει σε ακρίβεια του 76% ενώ, το άλλο με 25

epochs, κατάφερε να τελειώσει με ακρίβεια 95%. Όσο περισσότερα epochs, τόσο καλύτερη θα είναι η ακρίβεια του μοντέλου, κάτι το οποίο είναι λογικό. Ως προς τον χρόνο παραγωγής, τα 2 epochs για την κατασκευή ενός αυτοματοποιημένου μοντέλου, ολοκληρώθηκαν σε 11 λεπτά, και με την χρήση επαναλήψεων για την εύρεση ενός άλλου μοντέλου, ο χρόνος ολοκλήρωσης των epochs φτάνει τα 43 λεπτά. Αυτή, η αύξηση του χρόνου μπορεί να εξηγείτε στην αλλαγή του μοντέλου το οποίο θα παραχθεί καθώς, ο χρόνος παραγωγής διαφέρει από μοντέλο σε μοντέλο. Επιπλέον, η μικρή ποσότητα VRAM της κάρτας γραφικών μπορεί να εξηγήσει αυτούς τους χρόνους παραγωγής διότι, ο τομέας της μηχανικής μάθησης χρειάζετε μεγάλη επεξεργαστική δύναμη από τις κάρτες γραφικών. Αξίζει να αναφερθεί ότι, το RESNET34 αναφέρεται ως ένα per-trained μοντέλο όποτε, το μόνο που έχει να κάνει είναι, να μάθει τα χαρακτηριστικά των δεδομένων, τα οποία μπορούν να μεταφερθούν από dataset σε dataset μαζί με τα βάρη και τα biases των κόμβων. Αυτοί είναι μερικοί λόγοι που μπορεί να εξηγήσουν αυτήν την τεράστια διαφορά χρόνου παραγωγής μεταξύ αυτών των μοντέλων αλλά, θα πρέπει να εκτελεστούν περισσότερα πειράματα για την απόκτηση ενός καλύτερου μοντέλου. Τέλος, αν επικεντρώσουμε την προσοχή μας στα 2 πρώτα epochs, το αυτοματοποιημένο μοντέλο είχε καλύτερη ακρίβεια στο μοντέλο και στην μεταβλητή αλλά, παρατηρείται περισσότερο epoch loss και περισσότερος χρόνος παραγωγής.

ΚΕΦΑΛΑΙΟ 4 Συμπεράσματα

Όπως διαπιστώθηκε από το πείραμα, με έναν απλό τρόπο καταφέραμε να αποκτήσουμε ένα μοντέλο το οποίο καταφέρνει να ομαδοποιήσει της φωτογραφίες από τα τηλεσκοπία. Βεβαία, αυτό το μοντέλο μπορεί να χαρακτηριστεί ως ένα μοντέλο μέτριας απόδοσης αλλά, για έναν χρήστη ή επιστήμονα ο οποίος έχει εισαγωγικές γνώσεις στον τομέα της μηχανικής μάθησης και θέλει να κάνει κάποιες γενικές αναλύσεις, το μοντέλο που θα παραχθεί θα είναι ικανοποιητικό για τις απαιτήσεις του. Όμως, θα το επαναλάβω για ακόμα μια φορά, η αυτοματοποιημένη μηχανική μάθηση δεν πρόκειται να αντικαταστήσει τα μοντέλα τα οποία έχουν παραχθεί από τους ειδικούς σε αυτήν την χρονική και τεχνολογική περίοδο. Η μέτρηση της ακρίβειας και μόνο, δεν είναι μια τιμή η οποία μπορεί να κρίνει την αξία του μοντέλου. Η αναγνωρίσει των χαρακτηριστικών μέσα σε ένα μεγάλο όγκο δεδομένων και οι υποδομές συλλογής και διευκρίνισης της αξιοπιστίας αυτών των δεδομένων, είναι κάποια από τα χαρακτηριστικά που η αυτοματοποιημένη μηχανική μάθηση δεν μπορεί ακόμα να μιμηθεί. Από την άλλη μεριά, η αυτοματοποιημένη μηχανική μάθηση μπορεί να αντικαταστήσει τους "ψεύδο-επιστήμονες δεδομένων" οι οποίοι μπορεί να μην κατέχουν ή αρνούνται να εμβαθύνουν τις γνώσεις τους πάνω στον τομέα της μηχανικής μάθησης και αυτό έχει ως αποτέλεσμα, η αξία τους στο εργατικό δυναμικό να μειώνετε με το πέρασμα του χρόνου. Έτσι, ο τομέας της αυτοματοποιημένης μηχανικής μάθησης μπορεί να ληφθεί ως ένα εργαλείο, που όπως κάθε εργαλείο, η καλή αξιοποίηση και κατανόηση αυτού του εργαλείου είναι μια καλή δεξιότητα μιας και, η αυτοματοποιημένη μηχανική μάθηση γίνεται όλο ένα και ποιο δημοφιλή με το πέρασμα του χρόνου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. https://cloud.google.com/vertex-ai/docs/beginner/beginners-guide
2. https://towardsdatascience.com/auto-sklearn-an-automl-tool-based-on-bayesian-optimization-91a8e1b26c22
3. https://machinelearningmastery.com/auto-sklearn-for-automated-machine-learning-in-python/
4. https://machinelearningmastery.com/tpot-for-automated-machine-learning-in-python/
5. https://www.geeksforgeeks.org/tpot-automl/
6. https://www.geeksforgeeks.org/automl-using-h2o/
7. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html
8. https://towardsdatascience.com/automated-machine-learning-d8568857bda1
9. https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da
10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4928591/
11. https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.25028
12. https://iopscience.iop.org/article/10.3847/1538-4357/ac7f9e/meta
13. https://www.spiedigitallibrary.org/journals/optical-engineering/volume-58/issue-4/041607/Diffusion-based-cooperative-space-object-tracking/10.1117/1.OE.58.4.041607.full
14. https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/
15. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
16. https://www.aanda.org/articles/aa/abs/2022/05/aa42998-21/aa42998-21.html
17. https://www.kaggle.com/datasets/divyansh22/dummy-astronomy-data

18. https://www.kaggle.com/code/akzenith/auc-score-0-95-resnet-from-scratch

19. https://www.kaggle.com/datasets/pytorch/resnet34