**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ  
«Βιοστατιστική και Επιστήμη Δεδομένων Υγείας»**

A close up of a sign

AI-generated content may be incorrect.

**Διπλωματική Εργασία**

**Πρόβλεψη πρόωρου τερματισμού κλινικών δοκιμών με μοντέλα μηχανικής μάθησης vs κλασσικής παλινδρόμησης,  
βάση ανάλυσης των χαρακτηριστικών σχεδιασμού τους**

**Diploma Thesis**

**Predicting early termination of clinical trials with machine learning vs. classical regression models, based on analysis of their design characteristics**

**Συγγραφέας**: Ανδρέου Ευγενία

**Επιβλέπων**: Σπύρος Δεναξάς

ΑΘΗΝΑ, 2025

Contents

[2. Summary 5](#_Toc208499375)

[3. Purpose – Description 7](#_Toc208499376)

[4. Introduction 7](#_Toc208499377)

[Clinical Trials 7](#_Toc208499378)

[Definition 7](#_Toc208499379)

[Advantages and Disadvantages of RCT 8](#_Toc208499380)

[Approvals and stages 8](#_Toc208499381)

[Phases 8](#_Toc208499382)

[ClinicalTrials.gov Registry 11](#_Toc208499383)

[Background 14](#_Toc208499384)

[Termination Reasons 14](#_Toc208499385)

[Previous Research 15](#_Toc208499386)

[5. Statistical Theory 19](#_Toc208499387)

[Hypothesis Testing 19](#_Toc208499388)

[p value 20](#_Toc208499389)

[Test Statistics for Logistic 20](#_Toc208499390)

[Confidence Intervals CL 21](#_Toc208499391)

[Type I, II Error 21](#_Toc208499392)

[Statistical Tests 22](#_Toc208499393)

[Mutual Information (MI) 22](#_Toc208499394)

[Akaike information criterion (AIC) 24](#_Toc208499395)

[Chi-Square Test 24](#_Toc208499396)

[Bias – Variance Tradeoff 28](#_Toc208499397)

[Quality of Fit 28](#_Toc208499398)

[Bias-Variance Trade-off 28](#_Toc208499399)

[Cross-validation (CV) 29](#_Toc208499400)

[k-fold cv 29](#_Toc208499401)

[Leave One Out Cross Validation (LOOCV) 30](#_Toc208499402)

[Validation Set approach 30](#_Toc208499403)

[Uses 30](#_Toc208499404)

[Drawbacks of CV 31](#_Toc208499405)

[Scaling (Normalization, Standardization) 31](#_Toc208499406)

[Models 32](#_Toc208499407)

[Logistic regression 32](#_Toc208499408)

[Generative Models 38](#_Toc208499409)

[K-Nearest Neighbours - KNN 41](#_Toc208499410)

[SVC 45](#_Toc208499411)

[Maximal Margin Classifier 45](#_Toc208499412)

[Support vector Machines – Non-Linear Boundaries 47](#_Toc208499413)

[Classification Trees 48](#_Toc208499414)

[Performance Metrics 54](#_Toc208499415)

[SHAP Plots 54](#_Toc208499416)

[Roc Curve – ROC 59](#_Toc208499417)

[Confusion matrix–Accuracy–Sensitivity-Specificity 60](#_Toc208499418)

[Confusion matrix 60](#_Toc208499419)

[Accuracy – Sensitivity – Specificity 60](#_Toc208499420)

[6. Methods 62](#_Toc208499421)

[Tools and Technologies 62](#_Toc208499422)

[Environment 62](#_Toc208499423)

[Libraries 63](#_Toc208499424)

[Data Sources 63](#_Toc208499425)

[Data Engineering - Feature Formats 64](#_Toc208499426)

[General 64](#_Toc208499427)

[Missing values 64](#_Toc208499428)

[Data Engineering 65](#_Toc208499429)

[Visualization Pre-Processing 69](#_Toc208499430)

[Dummies 69](#_Toc208499431)

[Interactions 69](#_Toc208499432)

[Models 70](#_Toc208499433)

[7. Results 71](#_Toc208499434)

[Descriptive Statistics – Visualization 71](#_Toc208499435)

[Selected Model 71](#_Toc208499436)

[Features Contribution 71](#_Toc208499437)

[8. Discussion 72](#_Toc208499438)

[Comparison with Literature - Strengths and Limitations 72](#_Toc208499439)

[Strengths 72](#_Toc208499440)

[Limitations 72](#_Toc208499441)

[Recommendations for Practice 73](#_Toc208499442)

[9. Appendices 73](#_Toc208499443)

[Python Code 73](#_Toc208499444)

[Data Dictionaries 73](#_Toc208499445)

[10. References 73](#_Toc208499446)

# Summary

**Πρόβλημα/Υπόβαθρο:**

Ο πρόωρος τερματισμός των κλινικών δοκιμών έχει ως αποτέλεσμα τη σπατάλη πόρων, ανθρώπινου δυναμικού αλλά και χρηματοδοτήσεων. Το αντίκτυπο είναι μεγαλύτερο σε προχωρημένες φάσεις, όπως η 2, 3 και 4, δεδομένου του μεγαλύτερου μεγέθους των παραπάνω πόρων.

Οι λόγοι μπορεί να οφείλονται μεταξύ άλλων, στα ίδια τα χαρακτηριστικά σχεδιασμού της έρευνας. Συνεπώς, ο τερματισμός μιας τόσο σύνθετης ερευνητικής διαδικασίας είναι πολύ παραγοντικός αλλά και μοναδικός για την εκάστοτε κλινική δοκιμή. Το σύμπλεγμα αιτιών καθίσταται συχνά δύσκολα ανιχνεύσιμο από τα κλασσικά μοντέλα πρόβλεψης, όταν αντίθετα τα μοντέλα μηχανικής μάθησης εκπαιδεύονται συγκεκριμένα στα εκάστοτε δεδομένα.

**Σκοπός:**

Το περιεχόμενο της παρούσας εργασίας είναι η ανάλυση των χαρακτηριστικών των κλινικών δοκίμων μέσω μοντέλων πρόβλεψης της πιθανότητας πρόωρου τερματισμού τους. Η αναγνώριση των χαρακτηριστικών αυτών μπορεί να συμβάλλει στη βελτίωση του σχεδιασμού των κλινικών δοκιμών και κατ’ επέκταση στην επαρκέστερη διαχείριση πόρων.

**Μέθοδοι:**

Για το σκοπό αυτό αναλύθηκαν οι κλινικές δοκιμές διαστήματος 2011-2024 από το <https://www.clinicaltrials.gov/> και <https://aact.ctti-clinicaltrials.org/> . Εκπαιδεύτηκαν μοντέλα μηχανικής μάθησης και κλασσικής παλινδρόμησης σε δεδομένα σχεδιασμού - αποτελέσματος και συγκρίθηκε η μεταξύ τους απόδοσή.

Συγκρίθηκαν μοντέλα κλασσικής στατιστικής όπως Logistic, Elastic Net, LDA, QDA, με μοντέλα μηχανικής μάθησης KNN, SVC, Gradient Boosting, Extreme Gradient Boosting μεταξύ τους μέσω ROC Curves.

**Αποτελέσματα:**

Το μοντέλο που υπερίσχυσε στις περισσότερες φάσεις ήταν το XGBoost. Τα αποτελέσματα των μεταβλητών εκτιμήθηκαν με SHAP plots, όπου ο μειωμένος αριθμός των συμμετεχόντων εμφανίζεται ως κυρίαρχο χαρακτηριστικό πρόβλεψης πρόωρου τερματισμού, ακολουθούμενο στις περισσότερες φάσεις από τις κακοήθειες ως πεδίου έρευνας των κλινικών δοκιμών< για κάθε φάση μιας κλινικής δοκιμής ξεχωριστά.

**Συζήτηση/Συμπεράσματα:**

Τα αποτελέσματα τη παρούσας εργασίας, συμφωνούν σε μεγάλο βαθμό με την προ υπάρχουσα βιβλιογραφία. Η επιπλέον συνεισφορά της εντοπίζεται στην ευρύτερη μελέτη των παραγόντων πρόωρου τερματισμού ανά φάση, εμπεριέχοντας κάθε φάση και πιθανό χαρακτηρισμό σχεδιασμού μιας κλινικής δοκιμής και συγκρίνοντας μεγαλύτερο αριθμό μοντέλων από ότι έχει πραγματευτεί η βιβλιογραφία. Τα αποτελέσματα αυτών των μελετών, μπορούν να συμβάλλουν στον προσεκτικότερο σχεδιασμό μιας κλινικής δοκιμής, σε σημεία που εμφανίζονται ως παράγοντες πρόωρου τερματισμού. Αποτέλεσμα αυτού είναι η βελτίωση διαχείρισης οικονομικών και ανθρώπινων πόρων αλλά και η ασφαλέστερη έκθεση των εθελοντών σε έρευνες με ουσιαστική συνεισφορά στις επιστήμες υγείας.

**Problem/Background:**

The premature termination of clinical trials results in the waste of resources, such as human resources and funding. The impact is greater in late phases, such as 2, 3 and 4, given the greater size of the above resources.

The reasons may be due, among other things, to the design characteristics of the research itself. Consequently, the termination of such a complex research process is very factorial but also unique to each clinical trial. The complex of causes often becomes difficult to detect by classical prediction models, when on the contrary, machine learning models are specifically trained on each data.

**Purpose:**

The content of this work is the analysis of the characteristics of clinical trials through prediction models of the probability of their premature termination. The identification of these characteristics can contribute to improving the design of clinical trials and, by extension, to more efficient resource management.

**Method:**

For this purpose, clinical trials from the period 2011-2024 were analysed from <https://www.clinicaltrials.gov/> and <https://aact.ctti-clinicaltrials.org/> . Machine learning and classical regression models were trained on design-outcome data, and their performance was compared for each phase of a clinical trials seperately.

**Results:**

The model that outperformed in most phases was XGBoost. Effects of the variables were estimated with SHAP plots, where the reduced number of participants was indicated as the main feature predicting early termination, followed in most phases by neoplasms as field of study of clinical trials.

**Discussion/Conclusions:**

The results of this study are greatly aligned with existing literature. Its additional contribution lies in the broader study of factors of early termination per phase, including every phase and possible characterization of a clinical trial design and comparing a larger number of models than has been compared in the literature. The results of these studies can contribute to a more careful design of a clinical trial, in points that appear as factors of early termination. This results to financial and human resource management improvement but also the safer exposure of volunteers to research with a substantial contribution to health sciences.

# Purpose – Description

Early termination of clinical trials (RCT) lead to great loss of funding and human (volunteers, administrators etc.) resources, especially in late phases were trials get more scaled. Despite the efforts of the responsible monitoring committees like FDA or the registries od clinical trials, there are still many phases initiated without meeting all criteria, wasting sources that could have been used elsewhere for meaningful scientific progress.

Many studies like this tried to provide explanations of trials’ early termination in order to avoid this problem and contribute to more efficient trial designs. The purpose of this study is to analyze trials from CLinicalTrials.gov registry to identify as many parameters as possible for early termination probability, Specifically, to detect trials’ design characteristics as responsible for early termination and also to define the best classifier for prediction of this probability. The results of the study aim to be combined with similar literature ones and provide a useful guide to responsible parties when a new trial is filed for initiation.

The chapters of this project are organized as follows:

* Introduction: Reference to ClinicalTrials.gov public database and its issues in trials registries. Some characteristics of an RCT and the main characteristics and purpose of each phase.
* Statistical Theory: Analysis of background knowledge for tests, metrics, models, and methods used in this study.
* Methods: Reference to methods tools and environments this study used for analysis. Detailed description of data sources, data engineering and data science analysis.
* Results: Results of model comparison, feature evaluation and metrics used for comparison.
* Discussion: Summary of key findings. Comparison of results and processes with background literature studies. Proposal of practical application of findings to optimization of trials’ design.
* Appendices: Additional information of code, dictionary of terms etc.
* References: Detailed Bibliography from text citation.

# Introduction

## Clinical Trials

### Definition

Clinical trials are designed to observe outcomes of human subjects under interventional experiment conditions controlled by the scientist. Randomized controlled trials, where participants are allocated at random (by chance alone) to receive one of several clinical interventions, are the ultimate evaluation of a healthcare intervention.[1]

### Advantages and Disadvantages of RCT

RCTs main advantage is the randomization which solves selection bias problems. Selection bias could be formed due to unknown confounders.

In contrast some biases that may occur are ‘misclassification-information bias’ where outcomes or exposure are incorrectly recorded, co-interventions (e.g., additional interventions received), contamination (e.g., subjects receiving the intervention outside the trial etc.). Another problem ‘volunteer bias’ (sub-type of ‘selection bias’) i.e., when the sample is not representative of the population from which it was drawn. This happens mostly due to restricted inclusion or exclusion criteria. More specifically, possible participants may be excluded due to eligibility criteria such as comorbidities, or other attributes like distance from study facility etc. Restricted criteria are defined to study design to achieve homogeneity of groups, control of confounders and thus ‘efficacy’ (measure of success of intervention under experimental conditions). On contrast, this leads to loss of generalizability to real world population, where comorbidities or other parameters may occur. The result is that RCT although offering very good ‘efficacy’ measures, cannot promise intervention’s ‘effectiveness’ (real world measure of success of intervention).

A well-structured RCT needs a well-defined and representative sample from population. Sample size needs to be sufficient for the study to achieve the appropriate power needed to detect a statistically significant difference. Studies’ outcomes must be reliable and meaningful measures. Above all else, ensuring safety is the top priority.[1]

### Approvals and stages

For a promising intervention in pre-clinical stage, the sponsor or/and investigator submits an investigational new drug (IND) application to FDA, with all necessary drug information and results till this stage. If approved phases I-III follow. Similarly, if those phases results are promising regarding not only efficacy but safety too, New Drug Application (NDA). Except FDA an external committee is also involved with this application. After final approval, phase IV follows, in which safety and effectiveness for the indicated population is monitored.[1]

### Phases

As mentioned above, safety is above all matters in an RCT, so the first of 4 phases purpose to test the safety and maximum tolerated dose (MTD) of a drug, the pharmacokinetics, pharmacodynamics and drug–drug interactions. [1] Participants are allocated at random to receive clinical interventions. RCT are considered the ultimate evaluation of a healthcare intervention.[2]

Below are explained all stages of an RCT in detail.

#### Pre-Clinical

Pre-Clinical stage includes animal experiments and evaluation of drug production and purity. It does not include human participants. Main concepts studied are: 1) Drug safety per dose, approximated to human quantities. 2) Pharmacodynamics (e.g., mechanisms, dose-response), 3) Pharmacokinetics (e.g., absorption, metabolism, drug interactions). Those data are all included in IND approval submitted to FDA for further drug investigation in humans. [1]

#### Early Phase 1 (‘Phase 0’)

Their main purpose is nor therapeutic nor diagnostic, but exploratory (e.g., screening studies, microdose). They include very limited to not at all human exposure. [3]

#### Phase 1 (‘Dose-Escalation’ or ‘Human Pharmacology’)

It is the first human experiment stage.[1]

In this stage are determined metabolism and pharmacologic aspects, side effects associated with increasing doses, and early indication of effectiveness.[3]

Thus, maximum tolerated dose (MTD) and doses before toxicity are determined. Subjects are followed closely for all dose levels for any toxic effects.

Although, the objective of this stage is dose and pharmacology testing, there is a misinterpretation from volunteers that it could be therapeutic. Some improvements regarding the consent form, could help solving more clearly this ‘therapeutic misconception’. Participants constitute a small number of healthy or/and diseased volunteers. This stage is open labeled. [1]

#### Phase 2 (‘Therapeutic Exploratory’)

They mostly test safety, pharmacokinetics, pharmacodynamics etc. An especial role these trials have is that they are also a supportive stage to proceed to phase III, as they answer crucial questions regarding doses, dose frequencies, administration routes, endpoints etc.

They may also study preliminary evidence of efficacy by comparing the drug with other controls (e.g., drug from published trials, standard therapies etc.), randomize and examine different dose arms (e.g., control arm).

Despite the possibility of studying all the above and that they are larger that phase I trials, they are also conducted with a small number of participants only with the disease/condition of interest. In combination with the safety concerns of early stages, this phase cannot serve any efficacy establishment but only support the continuity and planning of phase ΙΙΙ. That is the reason FDA usually requests for additional phase IV less common adverse events to be identified.

With the completion of the above early phases, a meeting may be conducted between sponsors, investigators and FDA to review preliminary data, IND or even manufacturing concerns, to decide if phase III will be initialized.[1]

There is another Phase specified as Phase 1/Phase 2, Trials that are a combination of phases 1 and 2. [3]In this study these phases were analyzed as separated per unique phase (e.g., Phase 1 and Phase 2) [3]

#### Phase 3 (‘Therapeutic Confirmatory’, ‘Comparative Efficacy’ or ‘Pivotal Trial’)

They are initiated after above preliminary effectiveness evidence. They are much larger that the above with a more diverse population. Their main topic is to identify efficacy and adverse events or else benefit-risk ratio[3][1].

However, it has no more than 300-3000 participants, which gives a statistical power to establish an adverse event no less than 1 in 100 participants [4].This means that less common adverse events are not probable to be identified in this stage, and this is the reason why occasionally FDA requests for more than one phase III clinical trials or additional phase IV trial after drug approval.

* Comparative Efficacy trials (‘superiority’ or ‘placebo-controlled trials’):

It is the most common type of phase III RCTs where an intervention is compared to a standard therapy or placebo. It is notable that many ‘placebo effect’ instances (health improvement by placebo administration). Some say it's due to better care under research conditions, while others say that participants with acute symptoms would either way improve as time passes even without intervention. However, this indicates that a promising indication from a study does not always establish efficacy, as placebo effect may occur to both interventions. Another problem in phase III is a debate regarding comparative design in cases like surgical placebo procedures.

* Equivalency trial (‘positive-control study’):

The objective of this type of study is to study whether the intervention of interest is similar or different from another comparator. In this type, a placebo is almost never used. The margins are prespecified by the investigator and although they are based on statistical evidence, clinical experience etc. guidance for margin specification remains poor.

* Non-inferiority trial:

This type is a variant of equivalency studies. Its objective is to exclude that drug of interest is less effective from comparator i.e., null hypothesis is there is difference.

Non-inferiority trials often appear major issues, such as biased results towards null hypothesis (no differences) or else false positive, due to incorrect design and analysis. It is notable that they are most prone to false positive than any other study. It is notable that a false positive for an equivalency study would be a false negative in a comparative efficacy study.

*Phase III analysis/balance methods*

* Randomization:

The massive positive of Phase III studies is randomization in treatment allocation. Its purpose is to eliminate confounders and systematic differences (bias) between groups, thus any difference is due to treatment and not confounders. A common method is to just randomly assign subjects per group like flipping a coin.

However, this could lead to imbalances in treatment assignments or covariates’ distribution. A solution is ‘block randomization’ in which the number of subjects per arm is equal and balanced after a specified block size. For example, in a trial with 2 arms, a block size of 4 subjects would have 2 positions in arm A and 2 positions in arm B.

Yet, this could also result in ‘unblinding’ issues. For example, in a block size of 2, investigator will understand that if the first subject belongs to treatment A the next belongs to treatment B.

* Stratification:

This method is usually used in combination with randomization. Stratification ensures that some prognostic factors of clinical importance are balanced within arm groups.

Though, in small sample sizes when block randomization and stratification had both been applied, the original intended balance may be lost. To avoid such problems, alternative randomization methods can be used such as minimization or dynamic allocation that reduce imbalances among multiple study arms.

* Blinding-Masking

Blinding aims to reduce ‘information bias’ of subjects’ outcomes. Blinding could be:

* ‘Single Blinding’: Subject unaware of the given treatments.
* ‘Double Blinding’: Subject and investigator unaware of the given treatment.
* ‘Triple Blinding’: Data analyst, subject, and investigator unaware of the given treatment.

Still, not all types of studies can be blinded. For example, drug delivery methods or expected adverse and toxicities could indicate which drug is provided.

Trial design can be customized to the above problems like ‘crossover’ where each subject is also its own control or ‘factorial’ where more than one treatment are simultaneously evaluated.

*Analysis methods*

* ‘Intention-to-treat’ analysis (or ‘analyzed as randomized’ rule):

Subjects are evaluated based on the arm they belong, no matter what treatment they receive. It is the primary analysis method and most used, as it has the benefits of randomization’s selection bias reduction.

* ‘As-treated’ or ‘per-protocol’ analysis:

Subjects are evaluated based on the treatment received, no matter in arm they belong to. It is used mainly as complementary analysis as it diminishes the randomization benefits and is prone to selection bias. Although in comparison to the above it seems not so useful method, it has much better results in the case that, there are ‘adherence’ or ‘contamination’ issues, where the above method fails. [1]

There is another Phase specified as Phase 2/Phase 3, Trials that are a combination of phases 2 and 3. [3]In this study these phases were analyzed as separated per unique phase (e.g., Phase 2 and Phase 3) [3]

#### Phase 4 (‘Therapeutic Use’ or ‘Post-Marketing’)

FDA may request phase IV trial after a drug approval. Their purpose is to identify less common adverse events and even evaluate cost, drug effectiveness in populations and doses similar or even different from the previous study.

After post-marketing phase IV many drugs require new black box

warnings post-marketing than pre-marketing phases (e.g, phase III) or many drugs even withdraw because of safety reasons.[1][3]

## ClinicalTrials.gov Registry

Clinical Trials.gov is a national registry of clinical trials, created from FDA in 2000 [5] due to a law of Food and Drug Administration Modernization Act of 1997 which required from National Library of Medicine (NLM) to create and manage a public database for investigational drugs [6][7]. Although today it includes clinical trials and observational studies, happening now, will happen, or completed/terminated, with corresponding information e.g., study names, sponsors, investigators, dates, conditions, intervention types, documents etc.

It includes studies from over 200 countries. Registry of studies should be done from Investigators and Sponsors [7]. Registration data are displayed as table containing sponsors, design features, sample eligibility criteria, study endpoints, study results, and other [6]. Especially, for FDA approved drugs results must be uploaded within one year of the study’s completion, but this does not apply for all studies. [5] [7]

Finally, Clinical Trials ‘expanded access’ provide the possibility to patients with serious illness, unable to participate in trials, to register for an unapproved treatment. [7]

Clinical trials registry dataset is a type of Electronic Health Record (EHR) data of participants, interventions, outcomes etc. EHR data, are digital records of medical information of any type. In general, healthcare datasets are complex and can be challenging to analyse due to their size and complexity.[8]

### Clinical Trials.gov Data issues

Many issues about inaccurate data registration at Clinical Trials.gov have been noticed. Most common problems are missing or misleading data (e.g., Intervention types, RP names or roles, enrolment size, dates, sites etc.)[9] [10] [11]and non-reporting results on time. However, most of this information are being mandatory to mention (e.g., sponsor’s name).

First field regarding RPs/PI personal information. For example, there have been noticed trials not mentioning one or both of responsible Parties (PR), Principal Investigator (PI) names, others not matching PRs’ number with corresponding roles etc. Additionally, many RP names are typed with various formats, within different fields of registry, or are even presented with personal information of fake non-existent person. Registries with more than one PI per trial, despite normally only one should bear responsibility. Even for multi-location trials each location’s PI registry should be distinguished from the main.[9] Another rarer issue is multiple primary outcome registries. [11] Further, problem is the wrongly labelled studies e.g., RCTs labelled as observational and via versa or even more general as ‘NA’ etc.[9] [11]

Another issue concerns the delayed or belated publication of studies and results.

Trials should register prospectively and upload main primary outcomes, before enrolment starts. Also, they should upload results within one year of study completion. However, not all obligations are met, as there are still trials not registering [9] or data uploaded later than they should, affecting investigators’ analysis by missing information etc., Another, serious data misleading is about enrolment sizes. Many studies did not mention pre-designed or actual sample sizes or mentioned extreme outlier values which was obviously mistaken. Moreover, in some the size referred to cluster samples, population proportion sizes etc. rather than actual participant’s size. Several, discrepancies between sample estimates during register and actual accruals (usually less sample achieved actually), something noticeable if compare enrolment registries to their text reports. [11]

Another issue that noticed in the precent work and complicates analysis, is the great amount of unstructured text data in ClinicalTrials.gov. Explanation summaries and descriptions are conducted through brief texts without formatting restrictions, although they include precious information e.g, termination reasons. Finaly, the most serious problem, is the discrepancies found between registries and their publication [10] [6] [11]. However, here ClinicalTrials.gov usually seemed to be more accurate. Specifically, publications studies were found not to mention part or whole number of adverse events (usually ClinicalTrials.gov were accurate) or actual efficacy, which is harmful for the health community. [6][10]

It is notable that on ClinicalTrials.gov registry documentations, is referred that some fields were made mandatory only after specific dates, while others are not mandatory even today as seen in icons below.[3]

|  |  |
| --- | --- |
| **\*** | Required |
| **\*§** | Required if Study Start Date is on or after January 18, 2017 |
| **[\*]** | Conditionally required |

For example, Study Documents **\*§** column from registry, has more than 100.000 null values out of 180.000 and more that were downloaded in precent analysis.

Due to the directions and pressure of the FDA and ICMJE the situation has improved considerably over the years [9], although it is important to be noted that ClinicalTrials.gov does not review the accuracy of registrations [7].

### Importance of Accurate Registration of Clinical Trials

The importance of accurate registry of clinical trials has many benefits, while the opposite may even have harmful consequences.

Because great funding is applied to RCT trials the appropriate registration of their design characteristics (especially, those leading to early termination), could be evaluated from registries’ statistical analysis. The detection of those characteristics could prevent the risk of wasted resources. Early termination is most costly on later phases where human and fund resources are greater. [12]

Clinical trials involve experiments implemented on human samples. They aim to health science evolution, and their results may have implications to future patient’s care [9]. The impact of clinical trials extends to individual patients by providing alternative therapies, and the society by increasing health care value.[1]

Especially, Randomized controlled trials (RCTs), which are fundamental for health science evolution and decision-making, should be reported transparently, regarding not only the treatment efficacy but also adverse events or safety risks that patients may occur e.g., death risk, hospitalization or abnormalities. As such it is harmful and thus unethical not to be reported accurately [10][1].

Furthermore, lots of studies are possibly to be published at journals with great impact on the clinical care community [6].

The reduction of literature bias will benefit prescribers, inform patients for new treatments and researchers doing systematic reviews [5], [9][5], without leading them to biased results, while improving replicability and reproducibility of studies.[13] All in all, such registries are of public accountability [9][14].

Finally, misleading records also cancels the participants' attempt to contribute to health knowledge, especially of those relying on the study results [14].

Moreover, public registration platforms, such as CLinicalTtials.gov, use the data for variety of analysis regarding the sample sizes, sponsor type (e.g., industry, on- industry), number of Principal Investigator (PI), sponsor or funder per trial, sources distribution for global studies, termination reasons, results, type of intervention, locations that trials are distributed globally, nations health burden, unmatched or missing registries within platforms for same trials etc. [9].

Due to the above, International Committee of Medical Journal Editors (ICMJE), decided in 2005 that trials should be registered in corresponding public platforms, such as [ClinicalTrials.gov](https://clinicaltrials.gov/) , before patients recruitment, if they need to qualify for publication [9][5]. Similar, requirements were decided from Food and Drug Administration FDA in 2007[5] which required all clinical trials, with one or more sites in the United States, to be registered to ClinicalTrials.gov.[2]

## Background

### Termination Reasons

Clinical trials depend on human volunteers and cost significant investments of human, physical, and financial resources. Given the above, an early terminated trial raises financial, ethical, and scientific concerns such as cost of resources

that could have supported other trial, ethical issues regarding volunteers whose participation may not contribute to meaningful scientific knowledge, scientific issues which lead to termination decision and finally safety concerns regarding efficacy and most important safety of provided drugs. [15]

RCTs initiation, involve extensive planning, cost, researchers’ and sponsors’ deliberation. Funding agencies (e.g., NIH) are trying to screen only quality studies before funding, by conducting many peer reviews, which are costly for enterprise and time consuming. Even though, still many trials achieve moving on but unfortunately, they end up early terminated [16][17]. Unfortunately, later phases have more complicated explanation for termination. [2]

Termination reasons seem to be difficult to obtain, even though registered outcomes.[16]

The most common reason is fully unsuccessful or low enrollment, especially in specific study fields(e.g., rarity of disease hardens the collection of sufficient sample size). [18] [19][15][2][17][20]. Larger sample size trials seem to have lower termination rates [18].

Field of study is also an indicator, with most common the cancer/neoplasms trials, followed by surgical, while pediatric trials were mentioned rarer [16][17][21][12]. Also, trials not accepting healthy volunteers were noticed to have greater rate of termination in one study [12] .Moreover, disease type was affecting the low accruals rates e.g., rarer diseases had lower enrollment. [2][21]

Poor accruals affect also the statistic power of a study, thus it is important to fulfil the pre-designed participant’s number. [20]. Lastly, number of sites impacts the enrolment outcome, as it is related to limitations of participants leaving nearby or not from sites etc.[12]

Another main explanation is drug efficacy, toxicity or more general the risk-benefit ratio [16] [15] [12][22][17].

Moreover, funding, type of sponsor etc. or even due to location of facilities [16].

Greater number of collaborators has been noticed to result in higher probability of termination[2]. Administrative explanations were also detected regarding protocols, investigators, drug supply or withdrawal etc. [15]

It is notable that ‘terminated’ trials refer to those who no longer treat participants, while ‘withdrawn’ refer to trials for which enrollment was never initiated, thus did not enroll any participant.[15]

### Previous Research

While all institutions and registries try to overcome early terminations and false registries to save human and funding resources, this topic has become popular to researchers who study reasons for RCTs early termination. Many technics have been used apart from classic models such as ML algorithms or even NLP processes to track differences indicating an early termination as soon as possible. [16]

Some of the research that dealt with the topic is:

[Follett L, Geletta S, Laugerman M (2019)](#_References) who studied only ‘completed’ and ‘terminated’ prior to 2015 trials from ClinicalTrials.gov. They used a combination of structured data offered from ClinicalTrials.gov public format with unstructured text data they from ‘description’ column. The study included transformation of ‘description’ column to a structured format by creating binary columns based on the presence of most frequently detected words in each category of studies. After, data engineering by removing common or unwanted characters-‘tokenization’, separating words, creating ‘tidy data frame’ of words-trials, inverse document frequencies, binarization etc. Words were assigned to one of two outcome categories, if some pre-specified frequency threshold was exceeded to this category trials while rarely detected the other category trials. Many words were evenly detected in both completed and terminated categories.

Finally, all data fitted with random forest models. Study resulted in specific words as model predictors, with the top five (based on their predictive importance), are ‘treat’, ‘chemotherapy’, ‘cancer’, ‘patients’, and ‘tumor’. Study was conducted in all phases and not separate per phase. [16]

[Geletta S, Follett L, Laugerman M (2019)](#_References) used NLP and ML algorithms to predict patterns of early trials’ termination on trials started prion 2015. Specifically, Latent Dirichlet Allocation (LDA) was used on unstructured texts data from descriptions of trials (e.g., Brief Summary). LDA uses Bayesian methods to model each document as a mixture of topics and each topic as a mixture of words, which are after corresponded to probabilities. Thus, a document can be represented by a vector of topic probabilities while each topic can be represented by a vector of word probabilities. Number of topics choice is mostly arbitrary.

Pre-processing of LDA included ‘tokenization’ as above study (data transformed to ‘tidy data frame’ so that there is one line per word -’token’ per clinical trial). Common English words and characters were also excluded. Document-term matrix is created by calculating the frequencies per trials.

Then, LDA was applied to the document-term matrix to create topic-word probabilities (β matrix) and document-topic probabilities (γ matrix).

Form all topics only 25 were used based on their probabilities.

After, those 25 topics were fitted with random forest classifier with 6 other structured data, to assess importance of the LDA topic probabilities. Another Random forest was fitted also for the 6 structured data only. Then a logistic regression was used for interpretation reasons. It was proven that both type of predictors gave better results, as structured alone lost in rare diseases.

For each topic, top 10 words in terms of the term-topic probabilities were evaluated and resulted in a top 5 of diseases fields that seemed to be most valuable indicators of termination (e.g., ‘surgery’, ‘surgical’ , ‘pain’ underlying surgical procedures. First topics were the surgical procedures, then dermatological, heart conditions, pregnancy and last HIV. [16]

[Kavalci E, Hartshorn A (2023)](#_References) studied only ‘completed’, ‘terminated’ or ‘withdrawn’ (no enrolment), clinical trials from ClinicalTrials.gov register for year 2012-2022 (before 2011 missing data were greater). This study combined two different public datasets, ClinicalTrials.gov and CHIA dataset (CHIA contains thousands of inclusion and exclusion criteria based on phase 4 studies). The first was used as source for RCT data (CSV format) and disease categorization features, while the second dataset to generate new set of eligibility criteria. ‘Text mining’ from unstructured text data with already structured data were combined.

Double phase characterizes studies were assigned to both phases e.g., phase 2/3.

Withdrawn trials were also characterized as ‘terminated’ in order to create a binary outcome.

Some examples of features used from ClinicalTrials.gov, is the number of primary and secondary outcomes, number of sites, termination by disease (e.g., neoplasm studies were most likely to fail) etc. Other regarding eligibilities such as gender, age, healthy volunteers, number of inclusion/exclusion criteria etc.

Especially, CHIA dataset was used to generate more complex eligibility criteria which were also used as search terms in text data. Those, data were transformed into binary a binary dataset of 12,864 features.

Other, important data and statistic handling: Missing data for numerical features were handled via Multiple Imputation by Chained Equations (MICE). Phases were studied at separate datasets each. Train test split 70:30 ratio. Imbalance dataset towards completed status (85:15 ratio) was handled with random under sampling on the train set only, in order for the test set to be representative of true status distribution. Feature election was used because eligibilities created enlarged the dataset and created noise. Firstly, elbow point of features vs model error plot was used to identify k number of features. Then, ANOVA F scores determined which features to keep.

For model evaluation fivefold cross validation was also used, because of small dataset occurring after under sampling. Dataset was split into 5 folds and each time a different fold was used as the test set. Thus, model is trained 5 times in total. Performance of the model was the average of 5.

Then features were analysed with logistic but also ensemble models like Random Forest and XGBoost to predict probability of termination. XGBoost model evaluated as the best performing.

After, two different types of XGBoost models were trained for all phases together. One with only study characteristics features, and the other with eligibility criteria search features and disease categorisation features additionally. McNemar test for paired nominal data was used to test the null hypothesis is that models perform the same.

For model interpretation (SHapley Additive exPlanations) SHAP plots for local overview were used. [12]

[Elkin M, Zhu X (2021)](#_References) studied interventional and observational studies from 2000 and after. Like above, features were created from the registries structured and unstructured data.

Features were formed in three categories:

1) ‘statistic features’ (e.g., number of collaborators, industry/non-industry sponsor, study design like masking type, eligibility criteria like gender, USA/non-USA location, number of countries etc.).

2) ‘Keyword features’ resulting from more unstructured data from ‘description’ field. Analysed with TF-IDF (term frequency-inverse document frequency). TF is the frequency of words in a document while IDF is the frequency of words in several documents. Based on IDF a term that occurs in many documents (e.g., ‘the’), is not a good discriminator and is given less weight.

3) ‘Embedding features’ to overcome the problem of sparse data for keywords appearing in small number, using Doc2Vec to generate vector representations of words.

Furthermore, five feature selection methods were used, including ANOVA, ReliefF, Mutual Information (MI), CIFE (Conditional Informative Feature Extraction) and ICAP (Interaction Capping). Dowdall method was implemented to combine and select features from all five methods. Dowdall favours feature with many first preferences, thus, If a feature is accidentally ranked to the bottom by a method, it will have little impact the aggregation value.

Finally, data were fitted to four classifiers, Neural Networks, Random Forest, XGBoost, and Logistic Regression. Three types of features were fitted and evaluated separately and combined. Highest metrics resulted from combination of features (all three together). Ensemble models like random forest and XGBoost outperformed logistic. Similar to first study above, cancer words were top detected to relate with early termination but also some rarer diseases (‘Mycosis’, ‘Fungoides’, and ‘Sezary’) which may reflect the difficulty of finding sufficient sample.[2]

[Williams R, Tse T, […]](#_References) studied ‘primary outcomes’ field from ClinicalTrials.gov for RCT studies to detect reasons of early termination and pattern that primary outcomes were reported on registry or public literature.

Study defined 3 types of explanation for trial termination:

1. Scientific reasoning was most common e.g., risk-benefit
2. Other than scientific study, e.g, low accrual as main in this category, administrative regarding to protocol or investigators, funding, product withdrawal, insufficient drug supply, etc.
3. Unknown-not provided [15]

[Gayvert K, Madhukar N, Elemento O (2016)](#_References) studied RCT provided drugs from ClinicalTrials.gov based on their features to predict toxicity related early termination of trial. In addition, were compared drugs that although been approved from FDA but failed in toxicity trials (FFT drugs).

Features regarding drug characteristics but also tissue expression (eg., genes etc.) where considered with additional data from DrugBank register. Examples of drug features are Hydrogen Bond Acceptor Count, Molecular Weight, Polar Surface Area etc. and examples of target tissue features are Target Liver Expression, Target Network Degree, Target LOF Mutation Frequency etc.

This study developed a new approach for predicting the odds of clinical trial outcomes using random forest (PrOCTOR). PrOCTOR integrates the two types of features mentioned above in order to distinguish FDA-approved drugs from FTT drugs.

Principal components analysis (PCA) was used due to high correlation of target expression features.

Study results showed that frequent adverse events are acceptable for drugs treating serious diseases such as cancer, which was mentioned to all above studies as well.

This because many drugs may have partly failed as FFT but alternative others may have failed to all similar requirements. Analysis also indicated that both categories of features were important indicators, while combination of other although uncorrelated features provided greater discriminative power as well. [22]

[Chapman S, Shelton B, […] Bhangu A (2014)](#_References) studied phase III/IV clinical trials from ClinicalTrials.gov database for year 2008-2009 for surgical procedures specifically, using logistic regression. Result were that surgical procedure trials were more likely to early terminate and appear low accrual rates as well. [21]

[Carlisle B, Kimmelman J, […] MacKinnon N (2015)](#_References) studied trials ‘completed’ or ‘terminated’ in 2011 from ClinicalTrials.gov registry, which had unsuccessful or fulfilled less than 85% of pre-designed enrolment. Other factors which resulted in poor enrolment, were greater number of eligibility criteria, non-industry funding, early phases, smaller number of sites, use of placebo etc. Statistical methods used, where chi squared test, backward feature selection and logistic model.[20]

Some of commonly used methods in most of the studies are: Choosing location this with most facilities, as main location, for multi-location trials. Evaluate models on test sets or parameter tunning with validation set cross validation and k-fold cv accordingly. Evaluation metrics always include balanced accuracy (imbalance towards ‘completed’), sensitivity, specificity and ROC curves. Imbalance of ClinicalTrials.gov datasets towards completed was usually treated with random under sampling.

# Methods

## Hypothesis Testing

Hypothesis testing explores how likely it is that the observed difference would be seen by chance alone if the null hypothesis were true.

The goal is not to accept or reject the null hypothesis but to rather to evaluate how likely it is for an observed difference to happen if the null hypothesis is true [23].

: Null Hypothesis assumes no difference between groups (e.g., in their mean or proportions). Any difference lies due to chance. [24]. It expresses a specified value of an unknown parameter of population , such as , where a known constant [25].

: Alternative Hypotheses assumes signiﬁcant diﬀerences between two groups or else that there is significant deviance of from .

The alternative hypothesis must be based on a logical explanation based on the problem background [24]. Alternative Hypothesis can be two-sided (or two tailed) such as , or one-sided (one tailed) such as depending on the problem (e.g., one tailed: assuming the true time results of a drug appear equals to 10, then ) [25].

Then a random sample of the evaluated populations is chosen. The acceptance or rejection of is based on the value of a given test statistic, which needs to have a defined distribution under the . Afterwards, values of statistic test are separated into two discrete spaces. In one region is accepted while in the other rejected with second space being called [25].

### p value

is the probability that an observed effect is simply due to chance, if is true[26] or else is the probability of seeing the observed difference, or greater, just by chance if the null hypothesis is true.

As a probability it can take any value between 0 and 1. Values close to 0 indicate that the observed difference is very unlikely to be due to chance and in contrast values near 1 indicate that observed difference is very likely to occur due to chance and random variation.

As an example, assume p value = 0.08 for a problem studied with logistic regression with given Odds ratio = 0.33. p value indicates that there is an 8% chance or else (8 out of 100 times) to observe such a difference of 66% risk reduction (or extremer) if no effect occurs, just by chance. [23]

Finally, is a measure of the strength of an association but does not provide any measure of the effect size and so it must not be used as the only guide for evaluating a result.[27]. More specifically the ‘significant’ or ‘non-significant’ result based on p value is obsolete. This is because, for e.g., p = 0.05 values, researchers will wrongly accept 5% of the times (5% of times the result occurs due to randomness), meaning to fall into a , but wrongly rejecting the hypothesis can induce a too. The wrong evaluation is more probable in small sample sizes, where significance may be missed due to small sample. Furthermore, p value should not be used for clinical decisions, as they offer no evidence for effect size, adverse events, etc.[23]

### Test Statistics for Logistic

For logistic regression statistic test is statistic that divides the difference in observed and proportions by the null standard error of sample’s . z statistic measures the number of null standard errors that the sample proportion falls from the null hypothesized proportion .

: sample size

: sample proportion

: mean of

: Standard error of (3.2)

: population proportion under (3.3)

: Null standard error of , that holds under

Note that as sample increases the standard error of samples proportion decreases and approximates populations proportion and distribution of p approximates the normal. [28]

### Confidence Intervals CL

Confidence intervals determine the range of all plausible values that a parameter can be assigned [28] or else a range of values within which it is likely that the

true population value lies.[23]

: The estimated standard error of p.

Note that this formula substitutes unknown population parameter with the sample proportion in the population standard error

: The standard normal percentile having right-tail probability equal to .

: a 95% conﬁdence interval for i.e., 95% of the times values of belong in this range of values. Also, .

We can be 95% conﬁdent that the population proportion is between the two values that arise from equation (3.4). [28]

### Type I, II Error

As the rejection of is based upon one value of the statistic test, two mistakes arise to make:

* : Is to reject while it is actually true. This error can occur when statistic test’s value belongs to the but should not.

A white rectangular box with black text

AI-generated content may be incorrect.

* : Is to accept it when it is actually untrue. This error can occur when statistic test’s value belongs outside the n but should not.[25]

[29]

It cannot be accurately specified what type of error occurs. However, the possibility of making can be computed through the . In contrast, possibility of Type II error cannot be computed as it is based on unknown true value of parameter. The unknown probability of is symbolized as its domain is 𝜃 values specified by the alternative hypothesis.

The probability of rejecting while actually true is called ‘power’ and defined as .

The critical region is chosen based on where the probabilities of Type I and II error could be minimum. The concurrent minimization of those two is practically not achievable. So values of , and which lead to small values of , should be chosen wisely in order to minimize those two errors [25].

## Statistical Tests

### Mutual Information (MI)

Mutual Information (MI) is widely use in neuroscience and machine learning. It is a measure which quantifies how much of the variation in one variable is predicted by another. In other words, it represents the degree to which learning the value of feature reduces the uncertainty about which category the outcome belongs to and thus it constitutes a measure of the informativeness of the feature.[30] It is a measure of the amount of information a random variable contains about another.

MI is a special case of , which is a measure of the distance between two distributions. More specifically, MI is the between the joint distribution and the product distribution . [31]

For MI to be explained, entropy definitions must be referred:

: measure of uncertainty that distribution is q when it is truly p:

: a measure of the uncertainty of a random variable:

:

Conditional entropy H(Y|X):

Thus, MI for a categorical variable and :

* Continues feature:

Computing mutual information for continues features is harder and usually needs data discretization.[32]

* Categorical feature :

[31]

Due to symmetry MI is the reduction in the uncertainty of X due to the knowledge of Y and via versa:

[31]

: prior Shannon entropy (uncertainty) of category for unknown value of feature

: conditional uncertainty of category for a known value of feature [30]

Thus, X provides information about Y as Y about X.

It is notable that MI of a random variable with itself is its entropy. The entropy is referred to as self-information:

Finally:

[31]

### Akaike information criterion (AIC)

AIC tests a model by how close the ﬁtted values are to the true values, as summarized by a certain distance between the two. The optimal model is that which tends to have the closest predicted and true outcome probabilities for a logistic regression. This model also minimizes the following:

AIC depends on loglikelihood and number of predictors. Increase in predictors increases AIC score and thus has a negative impact.

In case of same number of predictors are compared or even same predictors with different formatting (e.g., log, sqrt etc.), AIC part of equation that is compared is as the predictors part is the same for all compared models.[28]

### Chi-Square Test

**Chi square Distribution**

Chi-square test was proposed from Karl Pearson back in 1900 and is used for comparing unpaired group data, for categorical/nominal variables [33] [25]. Chi square test uses frequencies (i.e., the number of observations per category). As a non-parametric test, it doesn’t make population characteristics assumptions e.g., about distributions. However, non-parametric tests are generally less powerful to identify significant differences and reject null hypothesis [33][34] i.e., greater Type II error rate. [34]

Chi square test is used to test if observed frequencies differ significantly from the expected ones. Firstly, to recognize deviations of the sample observed frequencies from the theoretically expected ones (one-way tables) and secondly relationships between categorical variables (contingency tables). The closer to accept Ho i.e., closer the observed to the expected frequency, the smaller χ2 value will be and via versa. Also, as the df increase the χ2 value increases, as more squared differences are summarizing and the critical value for rejection of Ho also increases. [34] [35].

Chi square distribution is skewed, with skewness decreasing as the df increase, so as n 🡪 ∞ it approximates Ν(0,1) normal distribution [35].

N should be large enough for the chi square achieve approximation i.e., greater than sample 5.[36]. The area of a Chi Square distribution below 4 is same as of standard normal distribution below 2. [35].

**Degrees of freedom**

A standard normal deviate is a random sample from the standard normal

Distribution. The Chi Square is the distribution of the sum of squared standard normal deviates [35], thus all values are positive.[34] The degrees of freedom (df) equal the number of independent normal deviates e.g., only 1 independent normal deviate means a χ2(1) distribution with 1 df. [35]

**Chi square applications**

There are always two hypothesis H0 and H1. To reject Ho, it must be χ2 > c, where

c = critical value, defined from literature, based on df and a sig. level. Beyond c threshold null hypothesis is being rejected. [36][34]

Most common applications of chi square test are the two below.

* 1. Testing Goodness of Fit (One-Way Design)

Tests if observed data (from sample) differ signiﬁcantly from an expected population frequency, that can be theoretically determined even rationally or empirically.

One-way table example:

|  |  |  |
| --- | --- | --- |
| **Outcome Category** | **𝐸𝑖** | **𝑂𝑖** |
| 1 | 25 | 18 |
| 2 | 10 | 22 |
| 3 | 13 | 30 |
| 4 | 5 | 21 |
| 5 | 12 | 17 |
| 6 | 20 | 12 |

X = Random variable from which observations occur with mutual independent outcome.

Oi = (i.e., Ο1,Ο2, …,Ο𝑘) = observed frequencies (number of observations) that belong to set Ii (i.e., I1, …, Ik).

𝐸𝑖 = 𝑛𝑝i = expected frequency of set Ii (or ith category) based on H0.

pi = probability that a random observation’s outcome is from set Ii (or ith category).

𝑖 = 1, … , 𝑘, categories of outcomes.

df = k-1.

Degrees of freedom are calculated based on the number of categories k - 1. [34] as one category’s frequency can be calculated from the sum of all other frequencies and total number of sample/experiments occurring.

So df = number of categories – 1 (i.e., (e.g., 6 category outcomes df = 6 - 1 = 5). [37]

Oi = each observed frequency of corresponding category

n = total number of independent observations or total number of independent random experiments. [35]

* 1. Independency test (Two- Way Design)  
     (Contingency - frequency - cross tabulation tables)

Tests whether the distribution of frequencies in one variable is related to the distribution of frequencies to the second variable. Where if true (H1 accepted) the two of them are related.

H0: 2 Variables A,B are independent

H1: 2 Variables A,B are related

Issue: distribution of frequency of one variable is or is not associated with the distribution of frequency of the other.[34]

Degrees of freedom are calculated based on rows and columns of the table, where df =(R − 1)(C− 1).

R = number of rows

C = number of columns. [35]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***B1*** | ***B2*** | ***…*** | ***Bc*** | Sum A |
| ***A1*** | O11 | O12 | … | O1c | **O1j** |
| ***A2*** | O21 | O22 | … | O2c | **O2j** |
| ***…*** | … | … | … | … | **…** |
| ***Ar*** | Or1 | Or2 | … | Orc | **Orj** |
| Sum B | **Oi1** | **Oi2** | **…** | **Oic** | **N** |

The expected frequency for a cell in ith row and jth column:

Oi = Total for the ith row,

Oj = total for the jth column,

N = total number of observations.

Each observation contributes to only one cell (unpaired sample). thus, the sum of all cell frequencies is the same as the number of observations (sample) in the experiment

Ai = Variables with categories: 𝑖 = 1, …, 𝑟,

Bj = Variables with categories: 𝑗 = 1, …, 𝑐.

k = categories of combination i,j of variables A,B.

Ο𝑖,𝑗 = number of observations belonging in categories of Ai and Bj variables.

𝑝𝑖𝑗 = probability of an observation to belong in cell (𝑖,𝑗)

𝑝𝑖 = probability of an observation to belong in a category of A variable.

𝑝𝑗 = probability of an observation to belong in a category of B variable.

Under 𝐻0 ∶ 𝑝𝑖𝑗 = 𝑝𝑖 i.e., probability of belonging to one of ith categories of A variable, does not depend on probability of belonging to jth category of B variable.

**2x2 Tables Chi Square test**

|  |  |  |
| --- | --- | --- |
| A | B | **A+B** |
| C | D | **C+D** |
| **A+C** | **B+D** | **N** |

A, B, C, D = observed frequencies, Oi, in each cell

AD = OA × OD

BC = OB × OC

N = Total number of observations [34]

**Cramer V - Effect size of Chi Square**As the sample sizegets greater the more Chi square test can detect smaller effects, as the statistical power increases. However, the effect size of this measure is calculated through Cramér’s V(or Cramér’s phi) value.

φ ≤ 0,1 are considered to have small effect size, while

φ ≥ 0.3 are considered to have large effect size. [34]

## Bias – Variance Tradeoff

### Quality of Fit

To evaluate the quality of fit, there needs to be measured how well predictions match the true data values. The interest lies in evaluating the accuracy of predictions on unseen test data, rather than training data.

The quality of fit is evaluated through the accuracy of estimation , i.e., how well it performs when used to predict data. As for classification:

* *Training error rate:* The proportion of mistakes if estimator is used to the training set.

: Predicted label of ith observation by using

: True label of ith observation

: Indicator variable. if and if .

* *Test error rate:* It is associated with a set of test observations on which has been applied:

: Predicted label by using classifier to the test observation.

A good classifier is one for which the test error is smallest.[38]

* Overfitting: The model appears to be a good predictor on the training set while underperforming on unseen data of test set. This is due to its low bias and high variance in which the model may adapt too strongly to the data which includes noise. When overfitting, we say that model lacks ‘generalizability’.
* Underfitting: The model has high bias and lower variance. This is due to potential interrelationships between data features may have been ignored.[39]

### Bias-Variance Trade-off

Bias - Variance and their relationship is fundamental to the performance of supervised ML models.

Bias variance trade-off refers to finding the best balance between bias and variance to have a more generalizable model. Increased bias translates into underfitting, while increased variance translates into overfitting [39].

* Variance:

Refers to the amount by which would change if another data set was used (e.g., different training data). Ideally, should not change much within different sets. It gives model’s flexibility. Induces overfitting i.e., indicate false relationships due to noise. High variance model means that small changes of dataset (e.g., change any data point) result in great changes in estimate. More flexible models tend to have greater variance.

* Bias:

Refers to the *error* introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. It gives the model’s rigidity. High bias induces underfitting i.e., missing real relationships between features and the target. Low bias translates into that the model fits the data more closely. For example, a linear relationship may not fit well in real-life problems, so there will be bias in the estimation of .

* Flexibility:

A diagram of a diagram of a foot

AI-generated content may be incorrect.More flexible methods, have more variance and less bias. Flexibility is associated with degrees of freedom of the model, where less df means more restricted and robust model.[38]

The ultimate goal is to find the right balance between Bias-Variance (bias-variance trade-off)

[13][39]

## Cross-validation (CV)

The more complex and higher dimensional the data becomes, the more the model’s fit reflect the noise values in the data set. A sign of overfitting is that the model fits well to trained data but poorly to test data from the same initial sample .

k-fold has common use in which the data split into k different subsets (folds). The model is trained on every subset except one held-out subset in which the model is tested. CV needs to assure a total isolation of the test data. No test data must be part of selective (e.g., variables, parameters etc.) or training procedures. If not, then ‘information leaking’ is going to happen from test data’s information to fitting procedure, which results to inflated accuracy scores. [40]. Also, variable selection must not be done prior to cv, however, many studies miss this and have biased results.[41]

### k-fold CV

Firstly, dataset is randomly divided into k groups (folds), of approximately equal size.

Then the statistical method is fitted to folds and tested to the remaining held out fold (validation set). This process is repeated by having different fold as validation each time. From this process k number of error scores (number of misclassifications for binary outcome), that are averaged to the final error. [38][42]

The error rate calculated from each iteration of cv, is the probability of misclassification for classification problems. The error rate is a measure of predictive capacity of the classifier – model.

Under random sampling, it can be proved that it is cv is almost unbiased.[42]

Per fold error rate:

k- fold cv averaged error rate:

[42][38]

: true error (probability of misclassification) of a classifier designed on a sample of size n.

A chart of blue and green circles

AI-generated content may be incorrect.Bias is not too great as long as is small.[42]

[43]

### Leave One Out Cross Validation (LOOCV)

The extreme version of the k-fold CV approach in which k equals the total number of individual data entries n of dataset (e.g., training dataset). Instead of splitting the dataset of interest into k-folds, here only one data point is left out as test point each time for each cross-validation step.[39] So, LOOCV can be referred as a special case of k-fold that here equals to all data points .

LOOCV cv averaged error rate:

[42][38]

### Validation Set approach

Validation set s used for evaluation of a method, usually to find the performance of a model on unseen data. It includes simply dividing the initial dataset into two subsets, the ‘training set’ and ‘test set-validation set or hold-out set’. The model fitted on the training data and then predict values of unseen test data. Resulting test error rate is like above.

: test set data points

### Uses

CV is commonly used to overcome overfitting by model selection and models’ best parameters (parameter tuning) through grid search cross validation. Is also used for model assessment.[41] Another use is for estimation of test error rate, through training set, as training error rate is usually, much lower than test’s. The estimation of test error rate is done by the method described above. The error curve resulting from this process from the training data is similar to the one resulting from true test data error. [38]

### Drawbacks of CV

One pitfall cv may fall, is that when splitting data into folds, different subsets will be produced each time, so different tuning parameters may be chosen. For this problem ‘repeated grid search cv’ may be used, which repeats a specified times, and produces same amount of cv errors, so the minimal cv error parameter is chosen. [41]

Another one is the assumption of CV is that observations are independent. The existence of this assumption is critical to the validity of k-fold cross-validation e.g., it is violated for data from family members, where its members tend to have more similarities. [40][44]

## Scaling (Normalization, Standardization)

Scaling of features from a wide range of values into a more specific one, is crucial before applying ML models, especially in real world data, where even different units may be used. Scaling influences models performance and ensures no feature will over/under present than another. This ensures that no feature predominates among others e.g., in distance calculation of distance-based models like KNN. [13][45][46][39]

Most common methods are:

* **Standardization**: Transforms data to zero mean and unit variance
* Z-score   
  It is a standard method. It scales values to zero mean and unit variance, resulting in negative or positive values, based on if scaled data are below or above mean value of initial data. The drawback is that in presence of outliers, they influence mean and standard deviation, which are included in scaling formula. Note that outliers’ magnitudes i.e., extreme values, vary per feature and will their scaled range of values. Thus, some features’ predominance will still occur. This is why this standard used scaling method is not so efficient when comes to outliers in dataset.
* Robust Scaler   
  Because it uses the interquartile range (IQR between 0.25-0.75 percentile), it is useful to handle outliers. The benefit is that it is not influenced by large outliers.
* **Normalization**: binding values between two numbers, usually [0,1]
* Min-max Normalization   
  It is linear transformation, normalization method. Data are scaled within range [,], without losing relations among them. It results in a smaller standard deviation of scaled data. The drawback is that it is also prone to outliers.

* + Linear Scaling to Unit Range   
    Data are scaled within range [0,1]. Results in a smaller standard deviation of scaled data. In practice it is a special case of Min-Max scaling for , with similar drawbacks and benefits.
* Maximum Absolute Scaler   
  It is a commonly used normalization method. Brings values to common scale, improving comparison and performance. Resulting values range approximately from -1 to 1. More specifically, if only positive values included in feature, the range is [0, 1], while if only negative , the range is [-1, 0]. If both negative and positive values, the range is [-1, 1]. The benefit is that no negative data occur after scaling, which may not be justified from dataset. For positive only values of data, Maximum Absolute Scaler is also prone to outliers.
* Percentile transformation   
  It is a linear method of scaling used to compare values of data or evaluate where data values stand compared to other candidates. The benefit of this method is that scaled values have simple interpretation i.e., the percentage of observations with that value or below.
* SoftMax Scaling   
  It is a non-linear normalization method, especially useful for data not distributing evenly around their mean. Resulting values range in [0,1] interval.  
    
   with parameter. For small value of , is an approximately linear function.

There are other methods as well such as:

Robust z-Score (5,95) or (25,75) which are less sensitive to outliers, Power transformation, Tanh transformation etc.

[45][46][47][48]

## Class Imbalance

Class imbalance means that one of classes (two classes for binary outcome) is much greater in size than the other. This issue is common in real-world data that need classification e.g., spam classification of emails, where ‘not spam’ are obviously more than spam emails. Most of times interest lies on the minority class, rather than majority, thus deleting such low sample categories is not a correct approach. Most ML algorithms assume balanced data classes, so the boundaries are easily biased towards the majority class. Thus, models’ performance is decreased.

Two solution categories are implemented: data-level and algorithm -level. Data-level solution includes resampling, which could be under-sampling or over-sampling. However, even those arise issues regarding the models’ performance.

### Data-Level Approach

The advantage of resampling methods is that the datasets need pre-processing (e.g., resampling) only once, and that I ready to be used in whichever classifier.

Because of these resampling methods are commonly applied on training sets, with various alternatives. It is notable, that resampling should not be applied on test sets, where distribution must be representative of real-world population.

#### Under-Sampling

Under-sampling is the method where a random sample from majority class is selected to be kept (or else a random sample to be dropped). This means that incidences of data are deleted from majority class. Due to that training cost is also reduced. Under sampling problems is that useful features/information may be erased or minimized. Another issue is that variance of classifier is increased, as sample is reduced.

Under-Sampling is optimal for low complex data and does not induce great problems regarding overgeneralization. Another advantage lies in that it does not add new data and so the risk of creating false decision boundaries is limited.

There are two types of under-sampling, Controlled under-sampling and Cleaning under-sampling.

**Controlled under-sampling**

Controlled under-sampling allows a user-specified number of majority class to be kept.

RUS (Random under-sampling):

It is the simplest form of under-sampling. It includes randomly selecting a sample from the majority class, with or without replacement. It is the simplest method of under-sampling.

Its drawback is the information loss occurring, which is related with level f imbalance. The greater the imbalance, the more severe the information loss.

Near-Miss:

There are three types of Near-Miss.

Near-Miss-1 selects the majority-class sample with the smallest mean distance from the minority class to the nearest N sample.

Near-Miss-2 selects the majority-class sample with the smallest average distance from the minority class to the furthest N sample.

Near-Miss-3 is a two-stage algorithm that combines Near-Miss-1 and Near-Miss-2 and it’s more robust to noise.

**Cleaning under-sampling**

This technique does not allow for a user-determined sample size. Therefore, the resulting ratio of minority to majority class sample sizes can be different.

TL (Tomek Links):

It detects TLs in the majority class and removes them. Let d(.) be the distance between the two samples. If belong to different classes and there is no sample that satisfies the conditions , then the pair is a TL. Values classified as TLs are considered noise or borderline samples.

Cluster C (Cluster Centroids):

Uses k-means clustering to divide the majority class into clusters and balances the dataset using the cluster centroids.

CNN (Condensed Nearest Neighbour):

It iteratively determines the sample that should be removed by using 1-nearest neighbour rule.

Its purpose is to create the smallest subset of the majority class. It’s drawback is the noise sensitivity.

ENN (Edited Nearest Neighbours):

It is an extension of CNN, with the difference of using 3-nearest neighbour rule. Removes samples that do not match their neighbours. This

RENN (Repeated Edited Nearest Neighbours):

It is an extension of ENN and is repeating the algorithm multiple times.

ALL KNN:

It is an extension of RENN and it increasing the number of nearest neighbours at each

iteration.

OSS (One-Sided Selection):

It is an extension of ENN and applies TLs and the 1-nearest neighbour rule to remove noisy samples.

NCR (Neighbourhood Cleaning Rule):

It is and extension of ENN and focuses on cleaning the dataset rather than condensing it. If a value is classified as a majority class using 3-nearest neighbours, it is deleted. If the value is classified as a minority class, the majority class samples in the 3-nearest neighbours are deleted.

IHT:

Uses a classifier algorithm to remove samples with low predictive probabilities.

#### Over-Sampling

Oversampling adds instance copies of minority classes instances to the minority class or by generating synthetic data to the minority class.

Most common problem of oversampling is the overgeneralization. Overgeneralization is when minority class domain is introduced into the majority-class domain, especially in complex datasets. For example, minority and majority class overlap in some areas, oversampling interpolates samples from minority class even near boundaries, and thus a minority point may mistakenly fall into majority area. This makes training process for minority class more difficult instead of the opposite.

Moreover, oversampling complex data may result in the creation of unnecessary minority class samples (i.e., noisy samples). Similarly, complex data aggravate the problem.

Two above problems, cause the training process to be more complicated, rather than simpler, with unclear boundaries and thus, unstable classifier.

Because of this, oversampling is often combined with some ‘filter’ methods or even in combination with under-sampling by removing detrimental samples from data.

The oversampling methods that generate new random sample without considering the majority class, induces more problems for highly complex datasets. For this, other methods like Borderline SMOTE.

Despite, data complexity is a major issue, oversample with data filtering methods is considered a better choice for more complex data is some cases, while under- sampling may be better in other.

ROS (Random Oversampling)

It randomly samples currently available instances to minority class with replacements, i.e., it replicates available instances. The advantage is that the majority class does not overlap another class during training. However, repeated sampling can induce overfitting.

SMOTE (Synthetic Minority Over-Sampling Technique)

Most used method is SMOTE, which generates synthetic data to minority class.

First, it randomly selects an instance from minority class. After, the closest neighbours of this point of the same class are selected. Then a random data point from these neighbours is selected. Finally, interpolation is performed between the two data points to obtain a new minority class instance, with characteristics somewhere between the two data points.

ADASYN (Adaptive Synthetic Sampling)

It automatically determines the sample to be over-sampled for the minority class by considering the dataset distribution. A sample that needs to be over-sampled is determined based on the learning difficulty. The learning difficulty can be quantified through the ratio of values belonging to the majority class to those belonging to the minority class, for the k nearest neighbours.

BSMOTE (Borderline SMOTE)

It determines the best candidates for over-sampling in the entire dataset. It most uses the points that are close to the decision boundary, as it assumes that those far from the boundary may contribute little to classification success.

SVM SMOTE (Support Vector Machine SMOTE)

It is a subtype of Borderline SMOTE. As the naming implies it uses SVM algorithms to detect the decision boundaries

KMSMOTE (k-means SMOTE)

It is another variant of borderline SMOTE. K-means SMOTE employs k-means clustering before applying sampling. It groups samples together and creates new samples based on the clustering results.

Filtering SMOTE Methods

SMOTE-TL and SMOTE-ENN are the most typical filtering methods. They help in solving the drawback of simple SMOTE of generating noisy samples. Each method adds TLs or ENNs after applying SMOTE to obtain a cleaner space.[49]

### Algorithm-Level Approach

Algorithm-level solutions impose a bias on the minority class by changing the search technique of the algorithm. They are beyond the score of this study and will not be analysed.

[49][50]

## Sparse Data

A line with dots in the center

AI-generated content may be incorrect.Certain data patterns result to the ML estimates being inﬁnite or not existing and so does the coefficient. ML estimate of the effect is the log odds ratio for logistic regression. This occurs due to observing only successes or only failures over certain ranges of predictor values. It results to ‘perfect discrimination’ and happens for both quantitative and categorical predictors. ‘Perfect discrimination’ is when it is possible to predict the sample outcomes perfectly by knowing the predictor values (except possibly at boundary points between the two regions).

### Quantitative Predictors

The ML estimate is inﬁnite when the x predictor values for which y = 0 are completely below or completely above those values of x for which y = 1. For example, for x = 10, 20, 30, 40 () y = 0, while for x = 60, 70, 80, 90 () results to y = 1. This ‘perfect’ ﬁt has and . ) The likelihood function keeps increasing as increases towards ∞.

Most software fails to recognize when , and also report huge standard errors.

### Categorical Predictors

Let’s assume a contingency table between a predictor X and the dependent variable Y. When one of the cells counts is 0, that estimate is plus or minus inﬁnity. Sparse data occur when the table has many cells (i.e., categories or variables), most cell counts are small, and many may equal 0. These types of contingency tables are also said to be ‘sparse’.

A table with text on it

AI-generated content may be incorrect.A cell with 0 count is said to be ‘empty’. Although, in this cell’s sample probability is almost always positive. Such an empty cell is called a ‘sampling zero’. For each sampling zero cell the ML estimator and thus coefficient is infinite. This case can also cause bias in estimation of odds ratios.

In the example, cases 1 and 3 are sparse, as there are zero counts of success.

[28]

### Addressing the issue

Solutions to this problem could be to add a very small constant (e.g., )

For possibly inﬂuential observations, delete them or merge them to another category’s cell. More specifically, to ﬁt the model by excluding part of the data containing empty cells (e.g., exclude one category), or by combining that part with other parts of the data (e.g., merge one category into another), or by using fewer predictors.

## Regularization

Regularization is when the coefficient estimates are regularized i.e., constrained or else shrink towards zero, usually by adding some information to the process. Its purpose is to minimize the overfitting of an ML model by reducing the effect of noise. This way, models bias may increase slightly but the variance will be reduced. [51][39]

A case where regularization helps is when the number of predictors is large compared to the sample size. Another one is when covariates are correlated with each other. Overfitting and collinearity yield very unstable maximum likelihood estimates (MLEs) with some of the being even infinite. [51]

### Ridge

Ridge regression has effects on coefficient estimators by shrinking them towards zero, but never equal to zero, which can be a problem for high dimensional data.

: Shrinkage Penalty.[38]

: Fraction of Ridge estimators towards initial ones.

or else : . [38][52]

: Probability , given the value :

Log likelihood :

: Tunning parameter. Controls the size of shrinkage

: no coefficients shrink, and model is fitted as original

λ = ∞ : coefficients will approach zero

As increases, shrinkage increases, and decrease. Its value is usually selected through cross validation

Τhe estimate of Ridge coefficients is expected to be on average closer to the real value of than the unrestricted , i.e. for regression.[51]

: depends on λ value and scaling of predictors e.g., a greater scale of predictors would estimate much different on the outcome as the coefficients are also shrunk. [38]

Ridge regression shrinks correlated predictors’ coefficients towards each other, allowing them to borrow strength from each other. In the extreme case of k identical predictors, they all get identical coefficients with 1/kth the size that any single one would get if fit alone. It is ideal if there are many predictors, and all have non-zero coefficients.[53]

### Lasso

Lasso regularization works like Ridge, but here coefficients can get exact zero. Another difference of lasso is that is used. [38]

= : [54][52]

: Shrinkage Penalty

It is useful for high dimensional data where variable selection is needed and also it results in and easier interpretated model. [38]

In contrast to Ridge, lasso is indifferent to very correlated predictors and as it will pick one and ignore the rest. In the extreme case of k identical predictors however it fails. [53]

### Elastic Net

Elastic Net combined both Ridge and Lasso.

It is useful in high dimensional data and when correlated predictors occur.

: Shrinkage Penalty

* : Ridge regularization
* : Lasso regularization
* : In between Ridge and lasso regularization [56][52]

Or format:

## Models

### Binomial Distribution

Distribution results from number of Bernoulli trials which could have two possible outcomes referred as ‘success’ and ‘failure’, with ‘success’ usually labelled as the preferred outcome.

: Success probability for a given trial,

: The number of successes out of the n trials. So, has the binomial distribution with index n, parameter π.

trials are assumed to be:

* Identical: same probability of success for all
* Independent: the outcome of a trial is independent from other trials.

The binomial distribution for n trials with parameter π has mean and standard deviation:

The sampling distribution of the sample proportion p has mean and standard error:

The probability of outcome y for Y:

[28]

#### Significance test for Binomial proportion

The z statistic divides the difference between the sample proportion p and the null hypothesis value by the null standard error of p:

For null hypothesis :

[28]

#### Conﬁdence Intervals for a Binomial Proportion

A signiﬁcance test indicates whether a particular value for a parameter is plausible. Conﬁdence interval to determine the range of plausible values.

With a 100(1 − α)% conﬁdence interval for π, one can be 95% conﬁdent that the population proportion includes those values for the characteristic studied:

: estimated standard error of p.

: standard normal percentile having right-tail probability equal to .

For 95% conﬁdence,

Formula needs large to work well and is also poor for near 0 or 1.

[28]

#### The Need for Classification models

A Linear form of the relationship of X with a binary outcome y to be interpreted: [38]

: Probability of ‘success’.

Outcome Y is encoded as 0,1 for convenience.

A comparison of a number of graphs

AI-generated content may be incorrect.This relation cannot be explained through linear regression, because negative probabilities or greater than 1 occur, as seen below.

[38]

### Logistic Regression

#### Maximum Likelihood Method

Maximum Likelihood is the method used to fit the logistic regression model. It searches for coefficients that maximize the likelihood function:

From the equation it can be concluded that maximizing likelihood means finding those coefficients, for which each data point is assigned with a predicted probability that gives as close as possible the actual observed value for each data point i.e., assign value close to 0 for Y = 0 and close to 1 for Y = 1, when predicted

replaces coefficients in the above equation.[38]

#### Logistic Fit

It is a classification method that uses a logistic function for predicting probabilities to classify a binary dependent variable. The function yields a value of 0 or 1 which represents the negative (0) and the positive (1) outcome, by calculating an odds ratio probability based on the relationships between the independent variables (features) and the dependent variables (target).[39]

Logistic regression uses this transformation to interpret the binary outcome:

From the equation is notable that p(X) cannot take values equal or less than zero and equal or greater than 1, for whatever X is used. [38]

[39]A diagram of a logistic regression

AI-generated content may be incorrect.

##### Odds

By transforming above equation, we receive the Odds quantity: , which takes positive values [0, +∞). A unit increase in X multiplies odds by:

* Odds: [38]
* Logarithm of Odds (Log Odds/Logit): [28]

Log Odds : is linear to X. Thus, a unit increase in X changes the log odds by β1.

It must be noted clearly that relation between p(X) and X is not straight line, as is in simple linear regression. (i.e., it is not ). This is understandable when plot X ~ p(X) (fig above), that logistic gives an shaped curve. Thus, a unit change in X does not correspond to change in p(X) by β1 (but the log Odds by β1 as notes above). The rate of change of p(X) per unit of X depends on the value of X.

[38]

##### Interpreting logistic Output

Logit (log Odds) increases/decreases by β for every 1-unit change in X.

: Determines the rate of increase or decrease of change in logit with the corresponding change of X.

: Increase of logit. Ascending S curve.

: Decrease of logit. Descending S curve.

: Y is then independent of X. Horizontal straight line. Then, .

: The curve’s slope or else actual rate of change. For example, for

The slope approaches 0 as the probability approaches 1.0 or 0.

: median effective level, where 0 (outcome has a 50% chance).

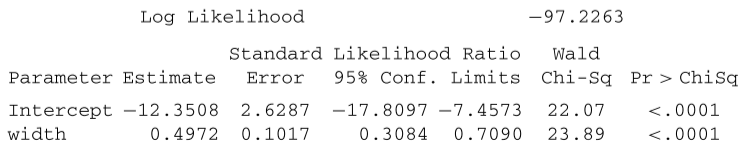
has slope approached zero(straight line).

Slope.

[28]

**Interpret Probability:**

For example, for a value of feature X, results to probability of success of 0.987 or 98.7% as seen below.



[28]

##### Interpretate Odds Ratio:

The odds multiply by for every 1-unit increase in x. Thus, the odds at level x + 1 equal the odds at x multiplied by .

Thus, for 1 unit increase of x:

Example for a specific value of :

For each a unit increase in x there is (1+0.64) = 64% increase in Odds.

[28]

##### Signiﬁcance Testing

Wald test statistic (z test statistic), following approximately standard normal distribution, measures the number of standard errors that the sample proportion falls from the null hypothesized proportion.

follows chi-squared distribution with df = 1.

The likelihood ratio test is more reliable. It compares the maximum of log-likelihood when β = 0 to the maximum of log-likelihood for β≠0. It follows chi-squared distribution with df=1.

[28]

##### Confidence Intervals

Wald conﬁdence interval for the parameter β coefficient and (confidence intervals for odds) accordingly:

[28]

#### Benefits & Drawbacks

Logistic regression is easy to interpret, and a well-studied model.

However, it has limitations regarding the sensitivity to noise and high dimensional data, especially if correlated. Also, as a parametric model it includes many assumptions, such as uniform (constant coefficients and linear relationship between features and outcome.[39]

### Generative Models

Generative models use another method to fit the classification, based on Bayes’ theorem/Bayes’ classifier if K classes assumed:

The idea is that it categorizes each observation x to the class where is largest.

This method is useful for cases where: 1. there is a serious separation between classes, where logistics coefficients become very unstable, and 2. if X follows an approximate normal distribution in each class with small sample size.

= prior probability that an observation x belongs to kth class.

= density function of X for an observation x which belongs to kth class. is large if probability of X ≈ x (i.e., belong to kth class) and small otherwise.

= posterior probability of an observation X = x belongs to the kth class, given the predictor value for this observation.

The idea is that instead of computing , to just plugin and to the equation (1). Specifically, estimations:

= By simply taking the fraction of ktvh class from a training set from a random sample of population.

= it is more difficult to estimate, and each of the LDA and QDA models make corresponding assumption. [38]

#### Linear Discriminant Analysis - LDA

LDA except all the above, is also a dimensionality reduction method. Its main goal is to convert high-dimensional data into a lower-dimensional space while optimizing the distinction between different classes. This property makes LDA especially ﬁtting for classiﬁcation tasks, as well as for extracting features and visualizing multifaceted, multi-class data.[8]

**Unique predictor, p = 1**

Assumptions of LDA to satisfy Bayes’ theorem:

* 1. = follows normal gaussian distribution
  2. = variance is equal among all classes.

Other assumptions:

* 1. = is unknown and simply assumed as the average of the training sample of kth class.
  2. = is unknown and simply assumed as a weighted average of the sample variances for each of the K
  3. classes.
  4. = for no other information, is simply assumed as the fraction of the number of observations in the kth category to the entire sample size.

Thus,

By replacing :

or the log, named Discriminant function which is linear to x :

which replacing for mean, variance assumptions is∶

is linear to x (linear discriminant analysis) and decision depends only on x. Decision boundaries are ‘lines’ themselves as well. In case i.e., equal prior probability to belong to each class, then classifies observation to class for which :

**Generalize for Predictors, p > 1**

Assumptions**:**

* 1. : predictors, follow multivariate gaussian (normal) distribution, with some correlation between them.
  2. : mean of X
  3. : is unknown p × p covariance matrix of X, common to all classes. Common covariance matrix makes LDA model linear to x.
  5. = are unknown means for each class.
  6. = are unknown.

Simple assumption to approximate those values happen also for p>1.

Decision boundaries number is based on pair of predictors. E.g., 3 predictors mean 3 boundaries or 3 separate spaces for each class k.

Error rates are a function of posterior probability threshold, where changes can differ sensitivity for specificity and via versa.

##### Benefits & Drawbacks

LDA in general is less flexible with its linear boundaries, with less variance. Moreover, assumption for common covariance is not actually happening to data, LDA will perform high bias. The fact that LDA depends on all data points, considering within class means and covariance matrices, makes it very vulnerable to data changes and thus less robust to outliers or noise. However, it is recommended when there is small training sample, as variance is reduced. [38]

#### QDA - Quadratic discriminant analysis

Generally, the concept of QDA is like LDA, using the Bayes’ theorem with plugged in values of , while differences lie to covariances and boundaries shape.

Assumptions:

* 1. : predictors, follow multivariate gaussian (normal) distribution, with some correlation between them.
  2. : mean of X
  3. : are unknown covariance matrices, different for each class k. Unequal covariance matrix makes QDA model quadratic to x.
  5. = are unknown means for each class.
  6. = are unknown.

QDA model has the drawback that many covariance matrices need to be computed. However, it is more flexible and better for great sample sizes, especially when variance of classifier is not of big concern. [38]

### k - Nearest Neighbours - KNN

KNN is a non-parametric supervised clustering ML method which was developed by Evelyn Fix and Joseph Hodges in 1951 and later optimized by Thomas Cover. KNN has various applications, especially in registering closest points, recognizing patterns, fraud detection etc.[58] It is used in data mining classification and disease detection. [8]

This model is instance based i.e., it does not use a function defined from training set, but it rather makes predictions by analysing the data in real-time when new data points appear, without necessarily having to go through a training phase.

It works based on number of neighbour data points, where k is equal to the

A diagram of a mushroom with red dots and yellow stars

AI-generated content may be incorrect.square root of the number of points and its distance (e.g., Euclidean) from a predefined point. In short, similar data points usually cluster nearby at space. A prediction for a query data point is based on the distance and thus similarity to the exciting neighbour data points. [58] [39]

[39]

**Parameters**

1. : Number of neighbour data points. Crucial tuning parameter. For large , computational demands increase. Choice is associated with dataset’s characteristics and impacts the accuracy.

* Small values (few neighbors) increase model flexibility and noise susceptibility i.e., overfitting.
* Large values (lots of neighbors) can underfit, while also being computationally more demanding. Boundaries are also smoother.

1. Distance Usual metrics :

* Euclidean : Most used.
* Manhattan : Better for high dimension data, for different scaling or sensitivity to specific predictors.
* Minkowski : Similar uses as Manhattan. [58]

It is notable that KNN is distance-based model, so scaling of data is a very important pre-processing step, as results can vary depending on the range of features values, especially, in real world datasets, where different units are often used. [45][13]

1. Neighbors:   
   All data points are ordered based on their distance and the number of closest data points of neighbors are becoming the neighbors.
2. Response:   
   For classification response is simply the most common occurring class of neighborhood. For regression could be similarly the mean, median etc. Note that test points will always take the same value as response, based e.g., on mean of neighborhood. [58]

#### Exact KNN

Exact KNN methods tend to be more accurate in finding the actual neighbor data points (i.e., actual close ones) and perform better where precision is important in the outcome e.g., health data. Exact KNN has two main types of methos : KNN search and KNN join for selecting the nearest neighbor’s data points. [58]

* KNN Search (Brute Force - BF or Exhaustive Search):   
  Finds nearest points of one point from the rest points. The process is repeated for all data points. However, due to calculating all distances for each data point, it is extremely computationally expensive in big datasets.   
  This method is better for problems which need speed and accuracy. In more detail:   
  : dataset  
  n: data points   
  : number of dimensions,   
  : a query (test) point,  
  : number of nearest neighbors  
  : a function that calculates the distance between point and query point .   
  : a set of k neighbor points.   
  KNN search finds a set of k neighbor points such that for each point p ∈ R , there is no other point where .
* KNN Join : Finds nearest points for each data point, but the process does not repeat as above per data point but happens for all data points as the same time. Better when uncovering hidden relationships or patterns needed and also computationally more efficient. This method is applied to k-means clustering, Outlier detection, KNN classification, Missing value computation etc. More detailed:  
  : dataset, : dataset  
  n: data points in   
  : number of dimensions  
  : number of nearest neighbors  
   : a function that calculates the distance between point

and query point   
: a set of k neighbor points.   
KNN Join find pairs of points such that for each , there are exactly k points with minimal distance. Or else, for each point in , we find its closest points in .[58]

#### Alternative methods and KNN models

Other alternative methods help in overcoming the drawbacks that exact ΚΝΝ cannot by itself.

* Computational efficacy:   
  Alternative KNN models like ‘Approximate Nearest Neighbors’ (ANN) for more complex data, are faster but with a cost of accurately choosing the neighbors, i.e., neighbors chosen might not be the actual nearest ones.
* Noise, Outlier:   
  Advanced distance metrics, weights (i.e., far neighbor points are more influential than close ones etc.). For example, variations of model K such as Adaptive KNN, Weight adjusted KNN and Fuzzy KNN to improve precision.
* Other:   
  e.g., k‑distance join, reverse KNN join, dynamic KNN join etc.[58]

Below methods are useful for computing the actual neighbor points accurately but also as fast as possible. Especially in high dimension B or R-tree indexing methods, perform purely.

1. Parallelization: divides the concurrent process of distance computing into different units. For example, this method is regarding device’s CPU way of procedure with command (e.g., = 1 work only on 1 core which is computationally slow, = 4 work on 4 cores, = -1 work on all available cores).
2. Dimensionality reduction (DR): Simply, decreases dimensionality e.g., through ‘principal components analysis’ - PCA.
3. Partitioning: Data is segmented through tree structures (e.g., R-tree, R\*-tree, Ball tree, KD-tree etc.) that perform pruning on part of data.

* Space based
* Data based [58]

#### Tree Structures of Exact KNN

Tree structures are a partitioning method, with the most common to be . Tree’s root is the whole data space which is the initial cell. Then data split at their median along the direction of a chosen coordinate. The process is repeated until split leaf cells are left with a threshold of data points. Final tree has a depth of about . Time that is consumed for this process is about

When a query point appears two methods of prediction may be chosen:

* Defeatist search: the query point is assigned down to the appropriate leaf and takes the value of the closest to it leaf point. Time consumption is more efficient at about or for constant . The great negative of this is that the ‘nearest point’ may be giving very wrong prediction i.e., may not be actual as similar as needed and also another leaf’s point may be the actual closest one to the query point. Unfortunately, failure rates may be unacceptable.
* Comprehensive search: Checks which other leaves may be needed to be taken into consideration for NN doners and returns the true closest neighbor point. Computationally it may take max time.

A big drawback of tree KNN is that their performance decreases a lot in high dimensions. For this problem, extra methods for dimension reduction are applied such as PCA as noted above. [59]

#### Benefits and Drawbacks

The highlights are that KNN makes no assumptions as a non-parametric model, and as mentioned before, it analyses data real-time, without necessarily need a training phase as it is distance based. It can also detect non-linear relationships.

Few factors are involved in its tunning, like such as selecting k-value, calculating distances etc, so KNN process is easy to understand and implement. Another advantage is that requires less training time than other ML models, however, it is time consuming in classification stage of test data, especially for large datasets. The presence of unrelated features makes process more difficult for KNN.[8]

Regarding other drawbacks is performs better with a smaller number of features. It also requires feature scaling because it is a distance-based model and finally it is sensitive to outliers. The performance can be altered based on the distance metric and as most supervised ML it is prone to overfitting. [39]

Computational requirements: A data point’s distance must be calculated from all the other data points. This escalates when it comes to bigger sample datasets [58]. Roughly, the time needed to find nearest neighbours is about time, which can be a serious time passing by with large n samples. [59]

* Noise, Outliers:   
  The fact that method is based on neighbors makes it susceptible to noise and outliers, in case a non-significant predictor or an extreme value in considered as neighbor.
* Accuracy:   
  By finding the efficient number and neighbor points precisely. Especially, in high dimensions, where distances lose their meaning, it is more difficult to find the actual near data points.
* ‘Curse of Dimensionality’ :   
  Model’s performance can decrease in high dimensions because distances become less meaningful.[58]

### Support Vector Machine - SVC

SVC has multidomain applications and is extremely popular to big data mining and pattern recognition.

The classification is based on hyperplane defining. SVC maximizes the minimum distance (margin) from either side of this hyperplane to the nearest point of the sample, as analyzed below. Finally, the largest possible margin hyperplane is chosen for the analysis. [39][8]

##### Hyperplane

A Hyperplane is a dimension flat subspace, which separates classes, in a p dimensional space or else a linear decision boundary. For example, for , it is a simple line separating two classes. For more than two classes multiple hyperplanes are used.

Generalizing for p the equation of a hyperplane is simply:

(7.1)

Any that satisfies this equation, is a point of the hyperplane and defines the hyperplane.

For training set of n training observations and

X lies on the hyperplane and defines it.

and X lies to one side of hyperplane

and X lies to the other side of hyperplane. Or more combined, ensuring all points are on the correct side:

Thus, for a test data point, the categorization is achieved based on the sign (positive or negative) the equation is assigned for a data point.

Thus for:

The farthest from zero is, the surer for the classification, as it belongs far from the boundaries space.

##### Maximal Margin Classifier – Hyperplane Margins

If data are perfectly separable, then there may be infinite hyperplanes, slightly rotated, that can separate them too. In order to stick with one hyperplane, the ‘maximal margin hyperplane’ is used among all the other hyperplane choices. More specific, two margins are defined in both sides of the hyperplane. These margins are defined by calculating the maximum distance M from each side of the hyperplane, from the closest to the hyperplane data points. So, we are looking for coefficients to satisfy the maximized value of M. The distance between margin and hyperplane is constant all along the hyperplane. Those closest to hyperplane data points that define where margins and thus hyperplane are oriented, are practically lying on the margin boundaries and are called ‘support vectors’. All in all, maximal margin hyperplane depends only on the close neighbour data points, that can ‘touch’ the margin boundaries, rather than the whole dataset.

Thus, any addition, removal or other change of support vectors or data points capable of being support vectors, will make the hyperplane orientation to change, when in contrast changes in the rest of the dataset, will not make any difference. This, however, implies overfitting issues may occur.

Finally, another problem arises when it comes to non-perfectly separable datasets, where the violation of margins and/or hyperplane is inevitable from some data points. This problem is solved by ‘soft margin classifiers’ or else ‘support-vector-classifiers’.

##### Support Vector – Soft margin Classifier

Support Vector classifier solves the above problems, by allowing misclassification and boundary violation to some data points close to margins and hyperplane. Those violators can be in the wrong side not only of margins but also hyperplane itself. This makes the model more robust and overcomes the problem of overfitting.

Similarly, with above M is the margin width and needs to be as large as it can be. Support vectors are now the data points, lying on margins or being inside them i.e. all data points which violated.

are ‘slap’ variables, which allow observations to violate boundaries.

For the training set:

Classification of a test data point:

* : , test point on correct side.
* : , test point is inside the margin.
* : , point on the wrong side of hyperplane.

##### Tuning parameter C

C parameter defines the number of violations allowed from the classifier and thus the bias-variance trade-off.

* , each , no violations allowed.
* , maximum number hyperplane violations allowed are equal to C.

The higher the C value, the wider the margins and the more the support vectors.

#### Non-Linear Boundaries SVC

Support vector machines are a support vector classifier, which use kernels to enlarge the feature space, and produce non-linear boundaries.

The mathematical solution to above functions includes the inner products of training observations, e.g., the inner product of two observations is

.

For observations and parameters, which are the trained coefficients per training observation and are only for support vector data points, the above function is formed as follows:

S is the collection of indices of support vectors.

and are estimated using number of n data point pairs.

The classification of test point x needs to compute all the inner products of that point with all the training points.

General form of function, where K is a kernel function, which quantifies the similarity of two observations. Te following most common kernels:

* Linear Kernel:  
  , where classifier is linear to x. It uses Pearson (standard) correlation, to quantify 2 observations similarity.
* Polynomial Kernel:  
  , . The greater the d the more flexible the boundaries.
* Radial Kernel:  
  . Works very local with close to boundaries points.

A diagram of a support vector machine

AI-generated content may be incorrect.The advantage of using a kernel rather than simply enlarging the feature space as we would do in a linear regression model (with polynomials, interactions etc.), is computational efficacy, as using kernels needs only to compute for all [38]

[39]

#### Benefits & Drawbacks

One of the highlights of the SVC is the ability to find non-linear relationships using a kernel function. This ‘kernel trick’ allows the data to be transformed into another dimension which enhances the dividing margin between the classes. Compared with other ML algorithms, SVC achieves higher performance with large datasets, higher accuracy. Training process is also conceptual simpler.

Despite being simpler, training is exceedingly slow and time consuming.

However, the limitation is that its prone to overfitting.[39]

It is also notable that SVC is prone to features values ranges, so scaling is a crucial pre-processing step. [13][8]

### Classification Trees

#### Recursive Binary Splitting

For predictors, Trees are separating this predictor’s space into distinct regions. Regions could have any shape, but usually, ‘boxes’ are chosen for interpretation facilitation. A split can result in two or a single region or leaf.

The process initiates from the top split, by considering all as first the first candidate for splitting, to finally choose the one that will give less Classification error rate (or RSS for regression) in both resulting regions. The first feature is the most important predictor.

The process continues similarly, from top to bottom. However, it is a greedy process as splits continue to be decided based on the rest of predictors and the already formed regions i.e., next regions are divisions of the previous ones. The process ends when a cut point is reached, e.g., a specific number of observations in leaves. [38]

Generally, a tree uses a structure that typically contains a root, internal nodes, branches, and leaves.

* Internal/Decision nodes:   
  Represents a tested feature. Node is where a ‘question’ is tested and where the predictors split.

e.g., Left: , Right: , the response measure for regression or classification accordingly.

* Branch node:

It represents the test result. It is where the outcome of tested ‘question’ is delegated

* A diagram of decision tree

  AI-generated content may be incorrect.Leaf/Terminal node:   
  Represents the class label assigned. It is where the class is finally assigned, at the bottom of the tree. Generally, it’s the final decision after all results of above attributes. The end results are a set fo rules that governs the path from root towards leaves.

Trees calculate probabilities in deciding on some courses of action. The training set is recursively divided into subsets based on feature values, so the data in each subset is purer than the data in the parent set. The classiﬁer can track the path from the root node to the leaf nodes and hold the sample’s category label for new unseen test data. [38][39][8]

Practically, each split leads to a Region , where the outcome values are always the same for all data points belonging in . For example, for classification this value is the most ‘common occurring class’ in this region, while for regression it usually is the mean value of the response for points in the region. In classification the interest lies not only to the most common class but also to the proportion of this class, among others. Finally, this process runs faster for less predictors (less dimensions p). [38]

Separating criteria

Classification:

* 1. Classification error rate:

Is the fraction of observation not belonging to kth class in a particular region and needs to be minimum. However, it is not used in practice.

: is the proportion of training observations from the kth class in the mth region.

* 1. Gini Index:   
     A measure of variance across classes and node purity as well. For close to zero or one. i.e., great purity in node. For very pure nodes (i.e., most of observations belong to kth class).   
     Gini Index has a small value for pure nodes..
  2. Entropy:  
     For near zero or one (pure node) entropy takes value near zero, i.e., entropy takes small value for pure nodes.

Regression:

= mean response or common label for all points in

= mean response or common label for all points in

Note that regression criteria is oriented to both nodes, while in classification per node. [38]

#### Pruned Trees

Simple trees are not typically used in ML, but alternative forms are (e.g., Gradient Boosting trees etc.)[39]. The closest alteration on a tree is pruning. Specifically, simple trees are at risk of overfitting, as the splits and leaves increase, i.e., bias is minimized on training data but the same isn’t always happening for test observations.

For this reason, pruned trees create less nodes and leaves in order to decrease variance with the trade of bias increase.

Cost complexity pruning

Initially, a very large unpruned tree is made. Then a series of other pruned subtrees are defined by pruning the initial . In order not to compute all the possible combination of , a parameter is defined, based on which a series of subtrees are created i.e., for each value of corresponds a subtree .

: number of terminal nodes (leaves)

: is a tunning parameter, that considers the bias-variance trade-off of complexity vs fitting to training set. It is usually selected through cross validation.

* : subtree equals the initial unpruned .
* : As increases less leaves remain i.e., more pruning happens.

Finally, after cross validation in the training set, the chosen subtree is the one for which the below obtains minimum value.

Reason why many subtrees are calculated, rather than one pruned based on a defined threshold (e.g., number of leaves to stop) is because tree structure is a greedy process as mentioned above. So, the pruning may have ended to a not so important split, while a more important may have followed, if continued.

#### Benefits & Drawbacks of Trees

Firstly, trees are fast, simple and easy to interpret, especially due to graphical representation not only regarding their ‘tree’ structure but also the regional spaces in cartesian planes. They can also outperform classic liner models for non-linear or complex relationships. Further, they are able to handle high-dimensional data.

Finally, decision tree supports incremental learning, which is immutable because of the alternative functions based on each internal node.

However, in general they are less accurate than best supervised methods.

Most important, they are very prone to data changes, i.e., they lack robustness or else suffer from high variance. For example, 2 different datasets or even 2 different random splits from the same data may result in different trees. Finally, as mentioned above, simple trees are also prone to overfitting. Lastly, it can be a lengthy process, particularly with large datasets or a high number of features. This is because the algorithm evaluates every potential split at each level of the tree, which can be computationally costly.

For the accuracy optimization, the below alternative trees’ methods have been defined, but with the trade of decreased interpretability. [38][8]

### Ensemble Models

Ensemble method combines several simpler models (‘weak predictors’) to result in a final model or prediction to achieve better performance. So, they combine the strength of other algorithms to complete training. [60][39]

One advantage of ensemble methods is that they perform in high predictive accuracy. They can also avoid overﬁtting problems that most unique algorithm supervised models arise. That is because, in case of a single ML algorithm can easily forecast ideally all the training data but with less accurate prediction for unseen data especially when using a small data size.

There are two standard ensemble learning algorithms: bagging and boosting algorithms. Bagging includes Random Forests, while Boosting includes Gradient Boosting Trees and XGBoost.[8]

#### Bagging Trees (Bootstrap)

Bagging means to randomly pull samples from the original data set to create a new data set of the same size. “Bagging” stands for Bootstrapping aggregation and is an ensemble method too. More specifically, initially it creates multiple models on subsets of data. Then those models’ predictions are ultimately combined to make the final prediction.[39]

Because of high variance and lower accuracy scores in simple trees, there have been defined other methods to solve those problems.

Random forests use the bagging method to define more than one trees, while also using the property of reducing variance by averaging this set of trees.

More specifically, bootstrapping is taking repeatedly random samples from the training set, separated random training samples are generated from the initial one. Thus, it is and for each of the training sets. [38]

A test point prediction in a classification problem, is decided by considering the class that each one of bagging trees gave and choose the most commonly occurring. [38]

Finally, the number of trees in not so crucial. As far as an efficient number of bagging trees is chosen, a greater one will not increase the accuracy or even overfit further. Thus, one positive aspect of bagging is the lower risk of overfitting.

Moreover, bagging trees offer the capability of summarizing predictor’s importances. For example, the total amount of Gini index decreased averaged over all B trees, by splits over a given predictor, can indicate and important predictor if this value is large enough. Similarly, happens with RSS for regression problems.

In contrast, due to more than one trees, they lose their interpretability efficacy compared to a single tree. [38]

#### Boosting

Boosting Trees differ from the above ensemble methods, in that trees grow sequentially, i.e., each tree grows based on the previous and that trees are fitted from the whole initial dataset, rather than bagging samples of it. [38]

Boosted Tree is an ensemble method that uses weak predictors (eg, decision trees) that can ultimately be boosted and lead to a better performing model (ie, the boosted tree). [39]In short, it combines a number several decision trees into a single predictor.[38] [60].

#### Random Forests – (RF)

RF is another ensemble method, which is an optimized form of bagging trees regarding the ‘tree decorrelation’. In contrast to boosting trees, RF differs to the splitting methods. Here, only a random sample of predictors are candidates (usually ) for each split and from those, only one is chosen. This way, bagging trees defined are not the same as one another and if for example, a very strong predictor occurs among others, he will not be present on average to (p − m)/p of the splits. This predictor will not prevail, and trees will not give similar results as it would happen in bagging trees. [16] Each decision tree will determine an outcome, and a majority “vote” approach is used to classify the data.[39] Averaging, uncorrelated trees lead to variance reduction, while with correlated ones this does not happen.

A diagram of a voting process

AI-generated content may be incorrect.It is called RF, since randomly generated decision trees are used to construct the final model.[39]

[39]

##### Benefits & Drawbacks

Using a part of predictors for fitting, is extremely useful when it comes to high dimensional datasets, or correlated predictors [38]. Random Forests are proven to be an extremely good performing model for classification, while able to detect non-linear relationships. They require less pruning than other algorithms and is also straightforward as it is simply a form of ensemble learning involving multiple classification trees.[16] Their ability of random sampling enhances generalizability and minimizes overfitting.

However, the number of trees and various internal parameters may hinder its performance. Additionally, it may be more time-consuming for great number of variables.[39]

#### Gradient Boosting Tree (GBDT) [61]

GBDT is subtype of Boosting trees. It is notable that GBDT works based not on parameters, but on functions’ space and that not but residuals are used.[38][60] Each new decision tree is trained based on compensating the prediction errors of the previous iteration’s tree [60].

More specifically, GBDT works by approximating the function , of the relation of , by computing of in an additive manner. Firstly, given the initial model, a decision tree is fitted to the residuals, which equal to true only for this iteration. After the first iteration, tree, each new tree is trained to minimize the following: [60]

: Residuals of previous iteration .

: loss function. Measure of prediction’s error.

: Prediction of current mth tree [60].

: matrix of features

: output vector

: Number of trees.

: , .

: Number of splits for terminal nodes.

: Initializing residuals for all in the training set.

: Learning rate. Constant regularization parameter, that limits the influence of trees ensembled to avoid overfitting. It takes small values (e.g., 0.001, 0.01) Boosting trees lean slowly’, so , slows the process, by determining how much trees change during the iterations.

Small values usually need greater . Generally, the slower the learning rate the better the performance .

: Initial function (e.g., that is going to be compensated.

: Tree fitted in the iteration.

: True function of relationship between

: Approximately computed function of

Then, the tree defined is added to the initial model to update the residuals by adding a shrunken version of the new tree, and this is repeated similarly.

Update previous tree [38][60][62]:

Update residuals [38]:

Final, model [60]:

Trees do not need to have many splits. Even one or two splits can be enough, as fitting small trees results in optimizing to areas that does not perform well. Although, a great number of trees can overfit. parameter is selected, as usual, through cross validation. [38]

##### Benefits and Drawbacks

Boosting Trees are very useful when it comes to unbalanced datasets. However, they include limited tunning parameters that makes them mor prone to overfitting compared to other ensemble methods (e.g., RF) [39]

#### Extreme Gradient Boosting - XGBoost

XGBoost is a regularized alternative of GMB. The difference lies in the computation of residuals per iteration , that is no more a loss function. Formula used for residuals is now:

: Regularization hyperparameters [60]

: Number of leaves in the tree

: norm of tree’s leaf weights [60][62]

XGBoost learns simpler trees with smoother weights, which leads to better generalization. Additionally, XGBoost employs Newton descent instead of gradient descent to optimize its trees, which leads to faster convergence. Finally, XGBoost also introduced a new feature split finding algorithm to speed up training. [60]

XGBoost gives very good performance in many problems, outperforming most of the times the commonly used models. This happens mainly, because it a scalable method to all scenarios, while also runs notable faster than them offering a computational relief even if ran on a common desktop. [62]

It is notable that Tree-based models, are better performing on wide features’ values range, with scaling not being so crucial pre-processing step. [13]

## Machine Learning vs Classic Statistic Models

### Machine Learning (ML) for Healthcare

Machine Learning (ML) refers to technics, of data-driven learning from experience, without being explicitly programmed. ML is one of the most used topics of AI. They are handling unstructured (text, image etc.), complex and bigger scale data quite fast and thus, they are arising in every field, one of which is healthcare that could benefit much from ML in unique challenges.

ML may complement and improve diagnostic and decision-making processes of healthcare professionals in patient care. By analysing data in real-time, recognizing patterns and thus identify diseases etc.

It could contribute to providing more efficient personalized treatments, based on patient’s medical history, lifestyle, diseases, genetic data, family history etc, as large volume or unstructured text or image (e.g., x-rays) data per patient can be handled with ML. Another benefit is in reaching conclusions faster and consequently, in sooner therapeutic intervention, reduction in adverse effects etc. Similarly, disease recognitions can be done faster before a condition reaches greater severity level. This is crucial in case of severe conditions where early detection is crucial, especially for neoplasms that early detection is hard with conventional means. The ability of image evaluation through ML contributes further to better prognosis of neoplasm. Additionally, it leads to deeper evaluation of therapy characteristics in rarer patient groups e.g., dosage, duration of therapy in children.

It can also be utilized is in organizational management. Specifically, in various types of data extraction from documents and Electronic Health Records (EHR) much faster than manually. This lead to identifying gaps in care and hidden risks. The benefit of handling various formats of data (unstructured, images etc.) fast establishes ML algorithms useful to organizational purposes. Healthcare and industry could use it for automation in clinical documentation, records administration etc.

Population Health management is another field ML finds application. Health Institutions already use ML algorithms to predict pandemic outbreaks, through web or social media real-time information. This is called ‘Crowdsourcing’ where large volumes of real-time social-media, webpage data etc, are analysed to predict a disease outbreak etc. ‘Crowdsourcing’ partially solves the problem of costly and time-consuming updating and maintaining health data but still needs optimization.

They also contribute to optimizing the expensive and demanding process of clinical trials designs, by computing best samples sizes, or even by being trained to recognize candidate participants through ‘Crowdsourcing’, detect design flaws and other criteria.

ML except supplementing healthcare decisions through analysis of complex data, also reduced human error, especially for automated processes where most errors occur.

It is notable that Electronic Health Records (EHR) increasing use, which include various formats of data (text, image, disease-coding etc.) has helped the integration of ML algorithms in healthcare and via versa. ML assists in reducing data errors occurring in EHR , while EHR formats of data are more efficiently handled from ML.

However, for ML algorithms to produce efficiently results, healthcare providers and researchers must be careful and precisely when they input data record. ‘Garbage-in/Garbage-out’ concept denotes the importance of high-quality input data or else, incomplete and erroneous values could wrongly train an algorithm. “Quality data” for ML training must be accurate, precise, complete, and generalizable (represent real-world population). Complete data is another issue arising in real-world data (e.g., due to visits missed etc.), which induces challenges for ML models too. Inout data format is a parameter for choosing the appropriate model to implement (e.g., Deep Networks are standard for image data classification).

[63][39][8]

### ML vs Classic Models

#### Supervised ML

Machine learning can be categorized into two categories: supervised learning and unsupervised. For the purposes of this study only Supervised Classification algorithms were applied. Supervised algorithms are given labelled data with known outcome, which are then categorised into training and testing data.

Classiﬁcation models are defined from these training data. These models can then be used to perform classiﬁcation on other unlabelled data. In short, models learn the patterns from training data and apply them to test data. [8]

#### ML vs Classic Models

A huge difference between ML and classic statistical models lies in their parametric or non-parametric form. In parametric models, parameters are fixed in number and models’ function has a known form-shape. In contrast non-parametric models adapt freely to data as they do not make such assumptions about the function. The most common assumptions regard distribution and shape i.e., linear function and normal distribution of data accordingly. Common parametric models for classification are logistic regression and Naïve Bayes, while non-parametric are KNN, SVC, Trees, Random Forests etc.

ML is already starting to replace classic methods in healthcare.

Some of the common classification models’ assumptions that ML algorithms overcome are:

Logistic regression may be limited due to not handling large number of features, especially if high correlated. Also, because logistic assumes uniform (constant coefficients) and linear relationship between features and outcome.

Naive Bayes uses a probabilistic approach based on Bayes theorem and thus, Bayesian logic. It assumes that the features are independent of each other.

[39]

## Performance Metrics

### SHAP Plots

Machine learning models steadily gain large applications on drug development, however their low interpretability of results, limits their implementation. SHAP plots is a feature-based plot method that enhanced ML models to better interpretability and thereby their trustworthy.

SHAP analysis is based on Shapley values, a concept from ‘collaborative game theory’. SHAP provides a fair distribution of a result payout among collaborative factors (e.g., features of a drug) where factors cooperate for a common purpose, without necessarily contributing equally.

For example, instead of ‘players’, let’s assume Feature A, Feature B and Feature C. Further, let’s assume a model’s outcome (e.g, outcome in percentage) instead of ‘game’, for which features contribute in order to be predicted.

Total response rate:

Feature A, Feature B, Feature C: 90%.

Each drug’s response rates:

Feature A: 40%

Feature B: 50%

Feature C: 60%

Paired drugs’ response rates:

Feature A, Feature B: 70%

Feature A, Feature C: 65%

Feature B, Feature C: 80%

Assumptions of SHAP:

• Efficiency: The sum of each drugs’ contributions equals the total response (e.g., 90%)

• Symmetry: If two Feature contribute always the same whichever the subset they belong (eg., pair or individually), they should receive an equal result payout.

• Additivity: If an intervention has multiple sub-samples and each has a separate payout result, then the contribution of each feature to the outcome is equal to the sum of contributions to each sub-sample. (e.g., two sub-populations of patients, the feature A contribution to the total population equals the sum of its contribution to each of the sub-populations).

• Null player: If a feature doesn't contribute to any subset (e., pair or individually), its share of the payout result is 0.

: Result payout

: ‘Player’ (e.g., Drug in this example)

SHAP value. One is assigned to each drug (‘player’ ) and corresponds to its contribution to the total outcome. SHAP values can be described as the weighted average of a drugs contributions across all possible subsets.

SHAP values satisfy above assumptions.

: Set of all drugs

: size of subset (eg., S = {Drug A, Drug B}) of size

: quantifies the contribution of drug j to subset S

: Weight of contribution.

: Weighted contribution of subset of size

sums all possible subset contributions without

In practice, other approximate methods are used to calculate SHAP values, as this formula would be calculating expensive.

Especially, in Pyhton SHAP package is used with e.g.,DeepExplainer, GradientExplainer etc.

: sample

: Model prediction for sample

: Average model prediction

: Sum of all SHAP values for sample i

Thus, for sample i, the SHAP values assigned to each feature describe how that feature contributed to the difference between the individual prediction of this sample from the average model prediction.

#### SHAP interpretation

Reported SHAP plots refer to test dataset. Different SHAP plots can illustrate explanation of feature contributions, locally i.e., on one data point or globally i.e., whole dataset.

* Bar Plot

Displays the mean absolute SHAP value for each feature, considering all data points i.e., global overview. Thus, it is a measure of feature importance but also ranks features top to bottom based on their mean absolute SHAP value or else impact on predictions.

Advantages:

Provides same units as model’s predictions, provides good interpretability.

A bar graph with numbers and symbols

AI-generated content may be incorrect.Disadvantages:

Lack of direction (eg., positive, negative) and monotonicity of impact.

* **Beeswarm Plot**

Displays SHAP value for each feature, considering all data points i.e., global overview, ordered by the mean absolute SHAP value. Thus, it is a measure of feature importance, which ranks features top to bottom based on their mean absolute SHAP value or else impact on predictions.

The dots are the SHAP values of each data point. Color range indicates feature values. High feature values are displayed as red, while lower as blue. Missing values, if any, are displayed as grey. Each dot is placed based on its SHAP value on x-axis.

Advantages:

The combination of dots’ color and place on x-axis indicates the direction of feature-prediction relationship, monotonicity, possible alike spreading or even outlier SHAP values.

Disadvantages:

A diagram of a graph

AI-generated content may be incorrect.Colour scale makes it difficult to surely explain the characteristics of relationship due to overlapping colored points, more complex relationships etc.

* **Scatter Plot**

Displays the relationship between one feature’s values (x-axis) and SHAP values (y-axis), considering all data points i.e., global overview but not for all features. On x-axis it may be displayed a histogram regarding feature value distribution.

Advantages:   
Relationship of feature-SHAP values trend indication (eg., linear/non-linear)

Interactions with other features if a vertical spread of SHAP values, or through coloring of points. The addition of trend lines makes patterns more obvious.

Can be used for k-fold validation, displaying all folds’ samples, and indicating different patterns within different samples of dataset. Each fold is assigned with a specific data point coloring.

A graph of value and values

AI-generated content may be incorrect.Disadvantages:  
Only one feature per plot.

* **Waterfall plot**

Displays SHAP values of a single data point but for all features i.e., it is a local overview. Displays features (y-axis) and SHAP values (x-axis). The value of that feature for that specific point is denoted in gray. SHAP values are displayed on the ‘arrow’ of each feature. Arrows are colored, with red color meaning that SHAP value increases the prediction , and blue if decreases it.

Advantages:

Outliers, SHAP value outliers, evaluate specific features.

A graph with numbers and a graph

AI-generated content may be incorrect.Disadvantages:  
Display only one feature.

#### SHAP for Classification

A close-up of a graph

AI-generated content may be incorrect.In classification models, SHAP plots may interpret: 1) probabilities, or 2) log-Odds. When probabilities are displayed, it is easier to interpret predictions, however, log-odds offer easier relationship interpretation between features (e.g., interactions etc.). Best practice is to evaluate both formats.

Example left probabilities are displayed on x-axis, while on the right log odds. [64]

### Roc Curve – PR Curve - AUC

#### Roc Curve - Receiver Operating Characteristic Curve

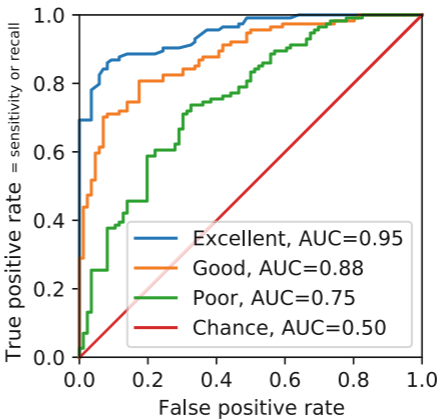
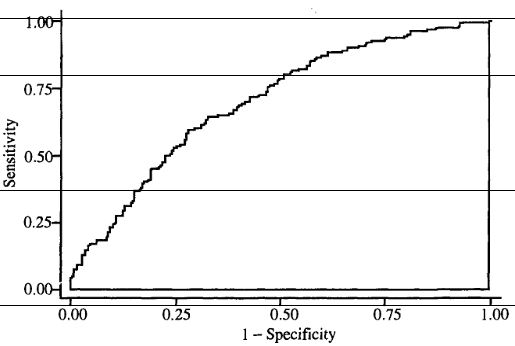
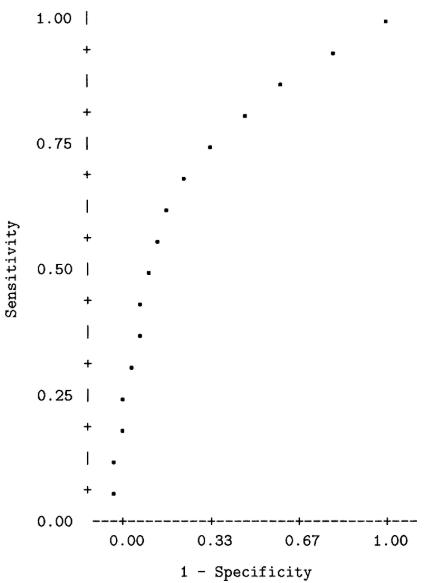
Receiver Operating Characteristic Curve (ROC) plots the probability of true signal detection (sensitivity) vs false signal detection (1-specificity) for a range of possible cut points. ROC curve is a completed description of a model’s accuracy.

Area Under the Curve (AUC) is a measure of the ability of the model to discriminate among the individuals who would achieve the outcome (Y=1) vs those who would not (Y=0). AUC values lie from 0 to. [65]

ROC curve is not the best choice for high imbalanced classes. When imbalances, changes in sensitivity or specificity, may induce great practical impact but not change ROC curve much. Alternatively, Precision-Recall Curve (PR-AUC) can be used, or other metrics. [66]

The closer the ROC curve is to the upper left corner, the higher the accuracy. At the upper left corner, sensitivity = 1 and 1 - specificity = 0 (false positive rate = 0), where AUC = 1.0. When 1 – specificity (x-axis) and sensitivity (y-axis) are 1:1, ROC Curve is drawn as a line of 45o, where AUC = 0.5 and indicates random guessing. [67]

The most used type is non-parametric ROC Curve (Empirical Roc Curve), as it does not consider data distribution (parametric Roc assumes normal distribution of outcome categories). By creating a 2x2 table for sensitivity and specificity on each cut-off value, creating the graph points, which are then just connected with a jagged line (rather than a smooth as in parametric). [67][28]



[43] [65][28]

**Cut points**

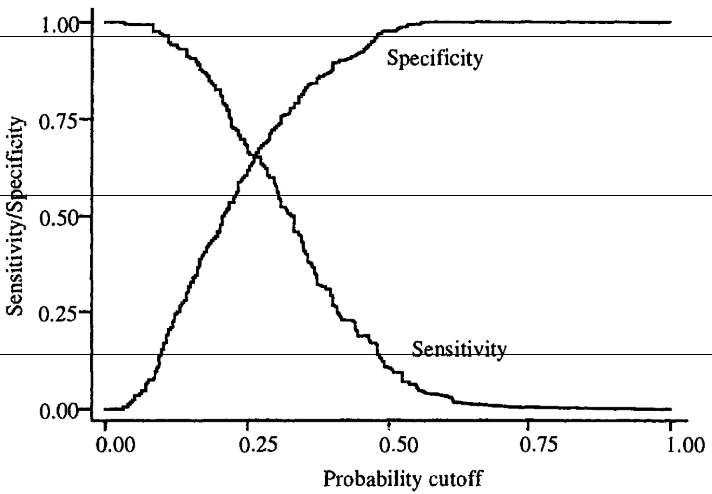
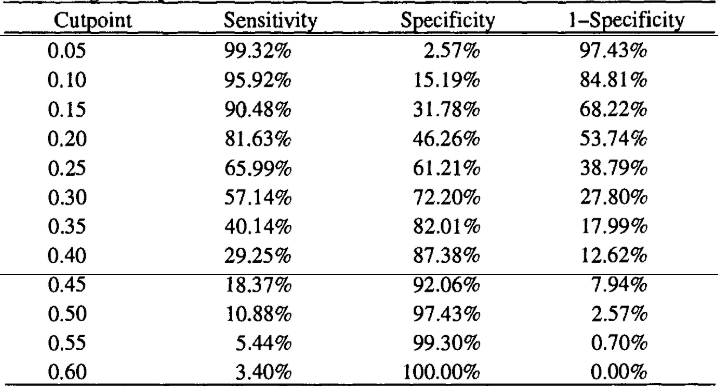
As far as it concerns cut points, the usually used in classification is and accordingly. The use of other cut points induces a trade-off between sensitivity and specificity and thus 1-specificity as the example seen below. [65]

If cut point is near 0, almost all predictions are , sensitivity is near 1 and

speciﬁcity is near 0, and the point for (1 – speciﬁcity, sensitivity) has coordinates near (1, 1). If cut point is near 1, almost all predictions are , sensitivity is near

0 and speciﬁcity is near 1, and the point for (1 – speciﬁcity, sensitivity) has coordinates near (0, 0).

One could choose another cut point which maximizes both sensitivity and specificity, however there are statistical benefits from the usual 0.5 value. By taking the sensitivity and 1- specificity for each cut point, a dot arises, and connecting those dots, result to the ROC curve [65]. ROC curve has a concave shape connecting the points (0, 0) and (1, 1) [28].



**AUC (ROC)**

The Area Under the Curve - AUC is the probability that a positive sample has a higher classiﬁcation score (as positive) than a negative sample. [66]  
  
The AUC is the probability that a subject achieving Y=1 outcome will have higher probability [65]. It is a measure of predictive power called ‘concordance index’. More specifically, let’s assume:

: all data points for which

: all data points for which

: pairs of each is paired with each

So, AUC is the proportion of those pairs for which (larger y) subject indeed had the higher the two probabilities (larger ). [28]  
A perfect classiﬁcation corresponds to an AUC of 1 and a random classiﬁcation to an AUC of 0.5. [66] [68]

AUC must be greater than 0.5 to be meaningful, and greater than 0.8 to be considered acceptable for a method evaluation. When comparing two or methods, this with greater AUC is considered to have the best diagnostic performance.[67]

* : Good. is generally, considered acceptable.
* : Fair
* : Poor
* : Fail
* : like random guessing [67]
* : Excellent. In practice it is rarely observed and if yes it may be due to perfect separation problems in data (sparse data). [65]

The AUC is often presented with a 95% CI because the data may be influenced by statistical errors. [67] The 95% CI provides the range of values in which the true value lies. Thus, one can be 95% sure that the CI includes the true value of AUC.[68]

#### PR-Curve - Precision-Recall Curve

PR-Curve focuses on the minority class so, it is a better option for imbalanced datasets[66] [69] [13] [70]. PR-Curve plots Precision (PPV) vs Recall (Sensitivity).[66] [70]

The baseline for PRC is not fixed but depends on the ratio of positives to negatives such as:

e.g., y = 0.09 if ratio P:N is 1:10. only for balanced datasets. In contrast for ROC curve baseline is fixed e.g, AUC = 0.5 for random guessing.

A graph of different colored lines

AI-generated content may be incorrect.In fact, AUC (PRC) is identical to the y-position of the PRC baseline.[70]

[43]

**Area Under Precision-Recall Curve (AUPRC)**

For AUPRC (or average precision) a perfect classiﬁcation corresponds to a value of 1, in contrast to ROC AUC, a random classiﬁcation does not necessarily correspond to 0.5. It depends on the prevalence ratio as mentioned before. [66]

### Confusion matrix

It is a common measure used for classification problems, as It summarizes the predictions made by a classification model organized into a table by class-confusion matrix. Specifically denoting the true positive, true negative, false-positive, and false-negative predictions. With these values various performance measures of model are calculate (eg, accuracy, sensitivity, specificity, etc).

A white rectangular object with black text

AI-generated content may be incorrect.For binary classification sample is divided into four categories, as seen below:

[45][43][39]

### Accuracy – Sensitivity – Specificity

A table of negative positive and negative negatives

AI-generated content may be incorrect.Given the below confusion matrix: [71]

True Positives (TP): Samples for which true and predicted labels are both 1.

True Negatives (TN): Samples for which true and predicted labels are both 0.

False Positives (FP): Samples for which true label is 0 and predicted label is 1.

False Negatives (FN): Samples for which true label is 1 and predicted label is 0.

* Accuracy (ACC):   
  Fraction of correct predicted outcomes to all predictions. It is one of the most used metrics for evaluating classifiers. It measures the ability of a predictor to correctly identify all samples, no matter if it is positive or negative.  
  Its benefits are simplicity and easy interpretation. The drawback of Acc is it cannot handle imbalanced datasets.

Sensitivity (SN) – Recall - True Positive Rate – (TPR):   
Sensitivity is the frequency/fraction of correctly predicted positive samples among all real positive samples. It measures the ability of a predictor to identify positive samples.  
It is also probability the diagnostic test is positive, given that subject is truly positive. It can handle imbalanced datasets.

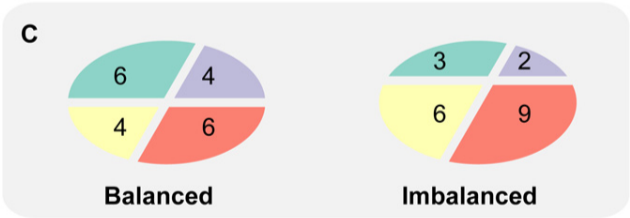
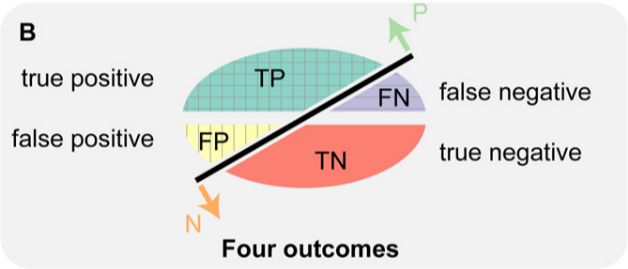
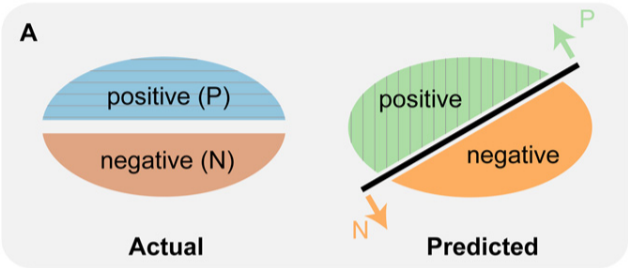
* Speciﬁcity (SP) – True Negative Rate (TNR) - (1 – FPR)   
  The fraction of correct predicted negative to all truly negative. It measures the ability of a predictor to identify negative samples. It is also the probability the diagnostic test is negative, given that subject truly does not have the outcome.
* Precision - Positive Predictive Value (PPV)   
  Fraction of correct predicted positive to all predicted positive. It can handle imbalanced datasets. Precision and Recall help to differentiate between error types. PPV does not take the false negatives into consideration.
* False Discovery Rate (FDR)  
  The fraction of false predicted positive among all predicted positive . FDR does not take the false negatives into consideration.
* False Positive Rate (FPR)  
  The horizontal axis is the FPR.
* Negative predictive value (NPV)   
  Fraction of correct predicted negative to all predicted negative.
* Balanced accuracy (BA-BAcc):   
  Accuracy metric that accounts for imbalanced samples. It solves most of the bias problems in an imbalanced dataset. It also considers the negative values.
* F1 score (F1):   
  Harmonic mean of PPV (precision) and sensitivity (recall). It balanced Precision and recall and is beneficial when there is a need to assess both false positives and false negatives by a single numerical representation. This measure, remove the true negatives from the calculation. Useful when one may care only about positive outputs. It is especially useful for imbalanced datasets, when true positives are considered as twice important

as the other samples.

* Matthew’s Correlation Coefﬁcient (MCC)  
  MCC is generally regarded as being one of the best balanced measures for imbalanced datasets. It also considers both negative and positives predictions.  
  MCC ranges from –1 to +1. A zero.
* : Indicates random guessing.
* : Indicates a perfect predictor.
* : Indicates a better predictor than random guessing.
* : Indicates a worse predictor than random guessing.
* : Indicates a ‘perfect’ predictor, which always predicts the opposite of the true label. So, the outputs should be interpreted with the opposite meaning where Y=1 it actually is 0.

Usually, the scores checked are sensitivity and speciﬁcity, for which the higher they are the better the diagnostic test.

[28][67][43][45][71][70][72]



[70]

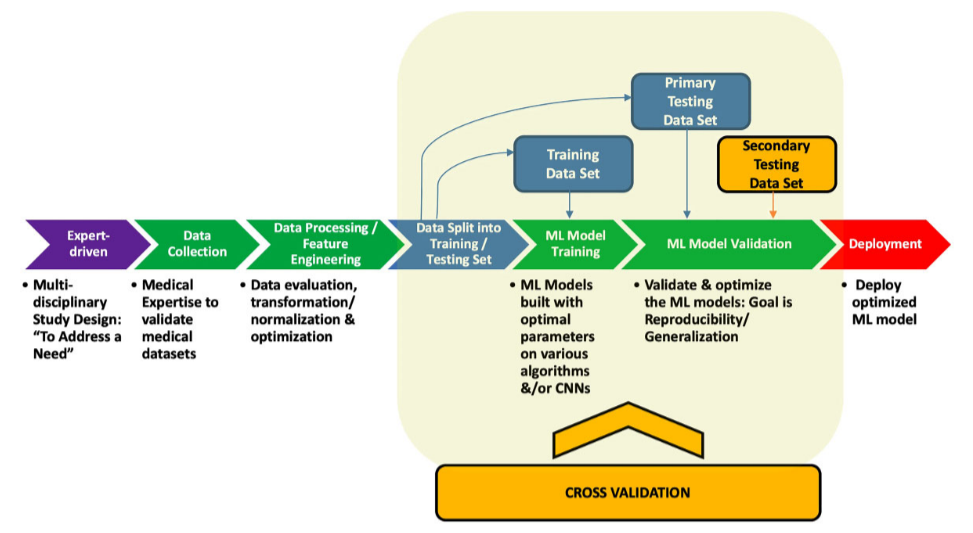
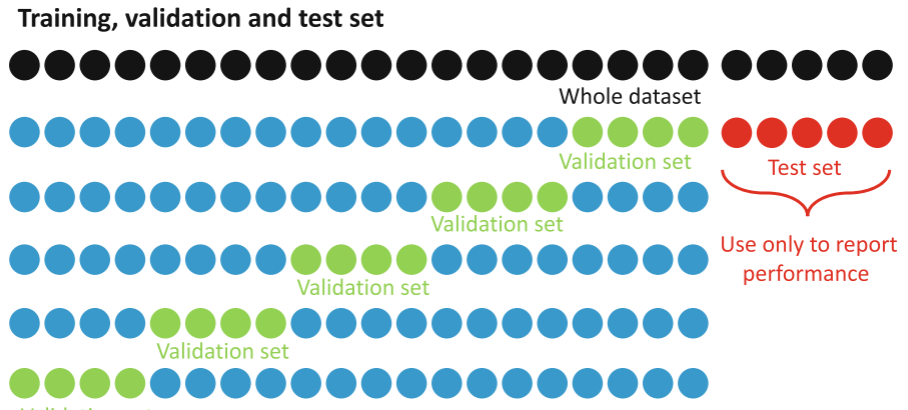
## Key Methodology Points

#### Pre-Modelling

* Primary Purpose:   
  Valid scientific question must be well defined as well as the eligibility criteria analysed should be such as to be reflect the actual ones from clinical practice. For example, models with too many features, may create unrealistic well performing predictors, without necessarily considering the actual useful features.[13]
* Representative Samples:   
  Samples should be representative of real-world distributions. More specifically, rarer categories, should not be ignored as it could lead to ‘spectrum bias’. They could, on the contrary, be encoded to an additional level e.g., ‘Other’, so distribution is preserved.[13]
* Sample Size:   
  Especially, for high dimensional data. Sample size should be large enough to achieve stable performance, especially in training sets, in order to avoid overfitting, noise from many irrelevant features and outliers. Unfortunately, sample size determination is still mostly empirical.[13]
* Missing Data:   
  Real world data commonly include missing data. Excluding all missing data could result to bias, sample reduction or even reduction in other features sample sizes. Many imputation methods could be used to replace nulls, or even models that are able to handle them e.g., XGBoost. [13]
* Scaling:   
  Real world data include wide ranges of feature values or even different units. This may result to features influencing more the model’s performance than others, especially in prone models like distance based KNN, SVC etc. Scaling is necessary for both train and test sets (whole dataset). The solution is Scaling (Standardizing, Normalizing) or utilizing a specific scale of feature e.g., logarithm.[13]
* Multicollinearity:   
  Collinearity happens when at least 2 variables are dependent one other, thus one can be linearly predicted from the other. Collinearity inflates variance of coefficients, thus, model ‘struggles’ to choose which feature coefficient to give most ‘weight’’, so changes in data lead to changes in coefficients estimations too. For example, linear models are especially prone to this problem. This way not only decreases model’s performance but also its interpretation as well. Some common solutions are bivariate analysis (only pair feature correlations) and Tolerance or Variance Inflation Factor – VIF for multicollinearity detection.[13]
* Imbalance Classes:   
  This happens when one classes sample is much higher than the others. This leads to bias prediction towards the majority class. This is why train sets should be resampled (over/under-resample) to restore similar distributions between classes. In contrast, test sets should not be resampled, as their distribution should be generalizable to real world’s.[13]
* Information Leakage (Data Snooping Bias):   
  It refers to transfer of data among train, test, validation sets or more generally, it happens when test set somehow influences the training process, [13] or it can be defined as the cases where information from the training set has “leaked” into the test set. [43] The problem arises due to problems in splitting or when a model is trained with features not available in unseen test data (e.g., train hospital data with discharge date, which only happens after the outcome event occurs) . Result is generalizability issues. For this reason, splitting must be always done before any transformation or process on data. Extremely high performing models should be suspicious of data leakage. [13]  
  The consequence is a rather falsely optimistic result. [43]  
  Common Causes of Data Leakage:
* Feature selection using the whole dataset.
* Dimensionality reduction using the whole dataset.
* Parameter selection using the whole dataset or the test set.
* Model or hyperparameter search using the whole dataset or the test set.
* For the same patient, if any of his data occurs in both train and test sets. For example, some of its visits are in the training set and some in the validation set [43]
* High Dimension data:   
  When great number features occur in dataset, especially if features are more than sample size. High-dimensional data can cause overfitting and generalizability problems, because many features are ‘learned’ for less data points and thus, any change in data provides different results. A solution to this is dimension reduction (e.g., PCA) or feature selection technics. However, those must not be done before Cross Validation, or else positive bias is induced, especially in high dimension data.[13][73]
* Randomness:   
  If randomness is not well implemented where needed (e.g., samples) then different results may occur even by using same data.[13]. Randomness helps in generalizability of method.[43]

#### Modeling

* Hyperparameter Tunning:   
  Hyperparameter optimization, lead to better performance in models. Overfitting of optimization can be avoided through Cross Validation or Regularization. [13]
* Overfitting (See Bias-Variance Tradeoff section):   
  Overfitting happens when models are trained on limited data points and fits too ‘close’ to them. The problem inflates if number of features is also large, creating noise. This way model is unable to predict on unseen data. Overfitting occurs mostly on train set. It can be eliminated by using more train set samples, by implementing correctly all other concepts discussed so far. [13]
* Generalization:   
  It is the ability of model to adapt to unseen data. Thus, the use evaluation and optimization processing should be on appropriate data sets. Specifically, hyper-parameter optimization should be done on validation set (through cv) in training data set, while model generalization and performance evaluation should be done on out of train set data i.e.., unseen test set. [13][43]



[43][39]

#### Post-Modeling

* Performance Metrics:   
  These should be chosen based on dataset, as not all metrics are appropriate for all dataset compositions. For example, accuracy is prone to bias for imbalanced class datasets. Alternative metrics for such datasets are balanced accuracy, no information rate, the Matthews correlation coefficient (MCC), F1 measure, ROC-AUC and PR-AUC (Precision–Recall AUC). If positive class is minority, precision and recall are good choices. [13]

If class imbalance occurs, ROC-AUC is not the best choice, and PR-AUC is preferred, as it focuses on the minority class. [66]  
Finally, in classification models, it is important to display the confusion matrix which provides broader insights of the models in terms of both overall and class-wise performance. [13]

* Comparison classic vs ML algorithms:   
  A completed evaluation of ML models, includes comparing them to classic statistical models such as logistic regression. This way ML models impact is better reflected.[13]

Interpretability: Some models are easier to interpret, such as logistic regression, trees etc. However, some ML models may outperform in some problems, but their greater complexity costs interpretability.   
Interpretability can be done locally for one feature or one data point or globally for whole features or dataset. [64][13]  
Common methods for interpretation are Shapley additive explanations (SHAP), permutation feature importance (PDP), activation atlass, local interpretable model-agnostic explanations (LIME).[13]

# Results

## Tools and Technologies

### Environment

Python version 3.12.10 through Visual Studio Code

pgAdmin4: version 9.0

All processes were managed with Python through Visual Studio Code and organized to jupyter notebook files, except part of processes regarding database source data, as explained below.

Jupyter files were organised based on the order of data engineering and analysis processes:

1. SQL

2. Data engineering

3. Visualization

4. Dummies-

5. Interactions

6. Models

7. Results

### Libraries

Python: Pandas, numpy, sklearn, statsmodels, sqlalchemy, pickle, matplotlib, seaborn, collections, geonamescache, scipy, itertools, patsy etc.

## Data Sources

* + - 1. [AACT Database](https://aact.ctti-clinicaltrials.org/downloads) of ClinicalTrials.gov:

AACT is a publicly available relational database that contains all information about every study registered in ClinicalTrials.gov. Content is downloaded from ClinicalTrials.gov daily and loaded into AACT. The Clinical Trials Transformation Initiative (CTTI) enhanced AACT in October 2016. Data are directly accessible in the cloud and can also be downloaded as a static copy. [74]

For this study, a static copy was downloaded in postgres.dmp format. The .dmp file provides SQL code for each table. Tables could be loaded by executing the corresponding SQL code. Specific tables were chosen to be studied, based on the features needed extraction. The use of database was the main source of data, for coherence and consistency of process, even if provided in structured format through CSV. Thus, database with combination to CSV, enhanced the creation of features, especially when one source provided a more unstructured format of a feature that the other.

For all data extracted, only ‘INTERVENTIONAL’ trials with start data >= ‘2011-01’ and study status 'COMPLETED', 'WITHDRAWN' or 'TERMINATED' were filtered (table columns: studies.study\_type, studies.study\_type,studies.overall\_status accordingly).

Tables from database was executed through on pgAdmin4 environment. After, Visual Studio Code was connected to database, and any data extracted or new tables created, were managed through this environment.[74]

* + - 1. [Clinical Trials.gov](https://clinicaltrials.gov/search) : Data downloaded in CSV format [7]. CSV was mainly used for comparison with database’s data, extraction of structured data that may not be provided in database source and also as the main source of trials according to their unique NCT numbers. It was not preferred as main source, because many columns provided data in unstructured text format, while in database source same data were structured.

**Filters applied:**

Study Status

* No longer looking for participants: Completed, Terminated
* Other: Withdrawn

Study Type

* Study Type: Interventional

Date Range

* Study starts: From: 0/01/2011
* Primary completion: To: 12/12/2024
  + - 1. [MeSH Terminologies Tree View](https://meshb.nlm.nih.gov/treeView): Disease, Conditions and phenomena categories. Terminologies and their codes were used in combination with corresponding tables of database. [75] ClinicalTrials.gov registry purposes if the use of NLM's Medical Subject Heading (MeSH)-vocabulary terms if possible, as the most appropriate.[3]

## Data Engineering

### Feature Formats

Three main data types/formats of features were analysed, with their columns renamed based on that:

* **Binary** : Suffix ‘\_Bin’ for easier recognition of data type/format.
* **Categorical**: Suffixes such as ‘\_Categ’ or ‘\_List’ for easier recognition of data type/format.   
  For, ‘\_Categ’ features, they all included two levels, thus no processing needed.   
  Especially for ‘\_List’ data that were initially unstructured text data that referred categorical features (e.g., ['CHILD', 'ADULT', 'OLDER\_ADULT'] for age feature). They were transformed into list element-data for convenience. This way two things were achieved, firstly, one trial record could belong to more than one levels of the same category, while data frame’s number of rows remained the same, without inflating the sample number. List element features were expanded and displayed as dummy variables for model and visual processes.
* **Continues**: Suffixes such as ‘\_counts’, ‘\_log’, ‘\_sqrt’ etc were added to naming for easier recognition of data type/format. Continues features were evaluated in different scales e.g., initial numeric, logarithmic, root etc. Continuous features were also examined as categorical with 2 levels (see ‘\_Categ’ above). The best format was chosen based on tests like AIC.

### Outcome Feature (Study Status)

There are three types of ‘closed’ clinical trials in ClinicalTrials.gov Registry:

* Completed: Study has concluded normally i.e., last participant has occurred.
* Terminated: Study stopped prematurely and will not resume. participants are no longer receiving intervention.
* Withdrawn: Study stopped prematurely, before enrolling any participant. [3]

‘Withdrawn’ category was dropped. This is because not enrolling any participants equals to study ‘Termination’. Thus, enrollment feature for these studies, is equivalent to study ‘Termination’ output as well. In other words, enrollment feature becomes identical to ‘Termination’ output i.e., a perfect predictor. This will produce bias for these studies.

Finally, the outcome feature is defined as a binary variable with levels encoded as 0: ‘Completed’ and 1: ‘Terminated’. Terminated encoded as the ‘positive’ label, as this study’s purpose is to evaluate the probability of early termination, rather than completion, thus this ενcoding is interpretively aligned with the purpose.

### Missing values

Missing values regarding numerical features, where dropped, as only a small amount was noticed compared to dataset (eg., only 54 rows for enrolment).

For categorical features, text mining was used to extract words from unstructured text data. For example, ‘Brief Summary’ column of csv source was widely used for text mining and null values filling. Similarly, happened for database text mining through different columns of tables.

Categorical missing data were many for some features, so dropping them would exclude other valuable features’ sample size. Because of this, missing values were transformed into additional or pre-existing levels of the corresponding categorical data such as ‘Other’ or ‘None’ (e.g., adverse events nulls meant no adverse occurred). For continues features, missing values were far too less and negligence towards the whole sample size, thus they were simply dropped. Imputation methods were not used, as most missing data appeared on categorical data rather than continues, where NLP methods are more practical but were not used in this study.

Many missing values contained in columns: regarding Study Documents (145000 out of almost 180000 records), Locations, Outcomes etc. Rows remaining null which could not be filled through text mining, were dropped as their number was small comparing the dataset. Method of missing values filling is not referring to each feature/column of ‘Features Pre-Processing’ field for convenience, as it is similar for all columns.

#### AACT Database Data Extraction

Specific tables were loaded by executing corresponding code from .dmp file. The choice of used tables was based on features needed extraction, that CSV format may not completely provide. From these tables others were created with the prefix ‘my\_’.

nct\_id column from studies table of AACT database, was used as the unique identifier of trials. It was joined with any table extracted or created from AACT and for any merge of AACT tables with csv tables etc. Thus, studies.nct\_id is selected for all below processes. It is not mentioned, for convenience, except the first table’s creation. Specifically, tables created from initial database tables and columns:

**my\_studies**: Tables: studies (Columns: nct\_id, brief\_title, official\_title, source, phase, number\_of\_arms, enrolment)  
For comparison to csv, number of arm groups, titles etc.

**my\_terminations**: Tables: studies (Columns: why\_stopped, overall\_status)  
Detection of termination related to accrual, funding, efficacy, toxicity etc. However, this table was not used.

**my\_conditions**:  Tables: mesh\_headings (Columns: qualifier, heading), conditions (Columns: name)

Conditions’ categories from [MeSH tree view](https://meshb.nlm.nih.gov/treeView) [75] was also used for this table. Detection of possible relation between disease/conditions’ category with probability of termination. Two levels of conditions from MeSH tree view were candidates for analysis, the general (e.g., C = Diseases) and the more detailed (e.g., C10 = Nervous System Diseases). Literature refers to specific study fields (e.g., neoplasms) associating with early termination, which were depicted from the detailed level. Thus, this level was chosen to be analysed.

**my\_covid**: Tables: browse\_conditions (Columns: mesh\_term)  
Separated trials to covid and non-covid. Evaluation of termination relation to covid. Defined a binary feature.

**my\_placebo**: Tables: design\_groups (Columns: group\_type)   
Separated trials to containing placebo and not. Evaluation of termination relation to placebo intervention. Defined a binary feature .

**my\_interventions**: Tables: browse\_interventions (Columns: mesh\_term)   
Table of detailed intervention (e.g., Rifampin). However, it was not used for analysis.

**my\_interventions\_types:** Tables: interventions (Columns: intervention\_type)Evaluation of type of intervention relation with termination eg., ['DRUG']

**my\_intervention\_methods**: Tables: interventions (Columns: intervention\_type, description)

Several words were searched through the columns of this table, in order to categorise the intervention method to ['Oral', 'Topical', 'Injection', 'Surgical'].

**my\_soc**: Tables: interventions (Columns: description), design\_groups (Columns: description, group\_type)

Separates trials to including standard care (SOC) treatment and not. Evaluation of termination to standard care. A binary feature was produced.

**my\_adverse**: Tables: reported\_event\_totals (Columns: event\_type, subjects\_affected)

Produced table with features regarding adverse level of severity e.g., ['Death', 'Other', 'Serious'] and number of adverse events.

**my\_adverse\_system**: Tables: reported\_events (Columns: organ\_system)

Produced table regarding organs affected from adverse events e.g., ['General', 'Nervous System', 'Skin']

**my\_designs**: Tables: designs (Columns: allocation, intervention\_model, primary\_purpose, masking, subject\_masked, caregiver\_masked, investigator\_masked, outcomes\_assessor\_masked)

Produced table regarding study design eg., allocation (randomized/non-randomized), masking (single, double etc.), intervention model (parallel, crossover etc.), primary purpose (treatment etc.). These data were also available in csv in same format.

**my\_eligibilities**: Tables: eligibilities (Columns: gender, healthy\_volunteers)

Produced table containing features of gender eg., [‘MALE’, ‘FEMALE’, ‘ALL’] and trials accepting healthy volunteers or not (binary feature).

**my\_outcomes**: Tables: outcomes (Columns: outcome\_type)

Produce table of outcomes eg., ['PRIMARY', 'SECONDARY', ‘OTHER’]

**my\_locations**: Tables: countries (Columns: name), facilities (Columns: country, city)

Produced table of facilities sites locations and number of them. Geonamescache library was used for this table too.

**my\_documents**: Tables: provided\_documents (Columns: has\_protocol, has\_icf, has\_sap)

Produced table of documents uploaded in clinical trials.gov e.g., ['Analysis Plan', 'Protocol', 'Consent Form']. Unfortunately, not all trials fill this field properly as there are many missing values, thus this feature was not used (145.000 plus missing values out of almost 180.000 rows in dataset). Field seems to be optional to fill, as those documents are necessary to a study design.

#### CSV Features Pre-Processing

All above features resulted from combination of csv with tables of database, except columns of csv being already in structured format (exceptions noted as ‘Used directly from csv.‘).

**Phases**: Used directly from csv. Phase stages combined such as Phase1/2 and Phase 2/3 that occur in dataset, where separated and analysed two as different rows/trials records, e.g., Phase 2/3 was transformed into Phase 2 and Phase 3 records.

**Age\_List**: CHILD, ADULT, OLDER\_ADULT

**Sex\_List**: ALL, FEMALE, MALE. Was transformed into list. Missing values were filled with text mining through ‘Brief Summary’ column.

**Funder\_Type\_List**: Used directly from csv. OTHER, NETWORK, INDIV, UNKNOWN, AMBIG, INDUSTRY, OTHER\_GOV, FED, NIH were the initial levels of categorical feature. Some of those levels had too little sample and produced sparse data errors in visualization models and analysis models. So, these levels were merged to greater ones as follows: 'OTHER':'FUNDER\_OTHER', 'NETWORK':'FUNDER\_OTHER', 'INDIV':'FUNDER\_OTHER', 'UNKNOWN':'FUNDER\_OTHER', 'AMBIG':'FUNDER\_OTHER', 'OTHER\_GOV':'GOVERM', 'FED':'GOVERM', 'NIH':'GOVERM', 'INDUSTRY': 'INDUSTRY'.

Also, text mining was applied on ‘Sponsors’ and ‘Collaborators’ columns to create another level for educational/university and hospital facilities, that replaced some rows of ‘'FUNDER\_OTHER’ and ‘GOVERM’ levels. Final, levels were defined as follows: EDU\_UNIV, INDUSTRY, HEALTH, FUNDER\_OTHER, GOVERM.

‘Sponsors’ and ‘Collaborators’ columns were not used further in the analysis.

**Datetime columns:**

Datetimes of dataset appeared into two formats: YYYY-MM-DD or YYYY-MM. All datetime data were converted to format : YYYY-MM. Were used for below feature creation.

**Completion\_Gap\_Counts**: Feature was created from Completion Date - Start Date columns in month counts. Also, evaluated as **Completion\_Gap\_Categ** format.

**Intervention\_Type\_List**: 'BEHAVIORAL', 'BIOLOGICAL', 'COMBINATION\_PRODUCT', 'DEVICE', 'DIAGNOSTIC\_TEST', 'DIETARY\_SUPPLEMENT', 'DRUG', 'GENETIC', 'INTERV\_OTHER', 'PROCEDURE', 'RADIATION'. Feature used as provided from dataset.

**Intervention\_Type\_Counts**: Number of ‘Intervention\_Type\_List’ column. Also evaluated as **Intervention\_Type\_Categ** format.

**Intervention\_Method\_List**: ['Injection', 'Oral', 'Surgical', 'Topical']. Created by merging database table with csv (see AACT database tables above field). Missing values were filled with text mining applied on ‘Brief Summary’ and ‘Interventions’ columns.

**Intervention\_Method\_Counts**: Number of Intervention\_Method\_List. Also evaluated as **Intervention\_Method\_Categ** format.

**Placebo\_Bin**, **Standard\_Care\_Bin**, **Healthy\_Bin**, **Covid\_19\_Bin**: Binary data for placebo intervention, standard care intervention, healthy volunteers’ acceptance, covid trials accordingly. Created by merging of database tables (my\_placebo, my\_soc, my\_eligibilities, my\_covid accordingly) and csv (see AACT database tables above field).

**Conditions\_List, Conditions\_Detail\_List**: Created by merging database table (my\_conditions) with csv (see AACT database tables above field). Two levels of information depth (see ‘AACT database’ field above): Conditions\_List is general (e.g., C = Diseases) Conditions\_Detail\_List is moredetailed (e.g., C10 = Nervous System Diseases). For this list element, the category ‘Phenomena and Processes-G’ and its deeper levels where removed from lists if list length was greater than 1, as the study interests in diseases and conditions, and not phenomena accompany diagnosis. Levels that had less sample size comparing to other levels, where imputed into other greater ones. This way more general categories were defined. This way sparsity of data due to zero sample size in some levels, that created problems to models fitting was avoided.

**Adverse\_List**: Severity level of adverse events. Created by merging database table (my\_adverse) with csv (see AACT database tables field). Missing data were considered as no adverse events occurring.

**Adverse\_Detail\_List:** Organ system that adverse was located.Created by merging database table (my\_adverse\_system) with csv (see ‘AACT Database’ field). Missing data were considered as no adverse events occurring.

**Adverse\_Counts**: Number of Adverse\_List. Created by merging database table (my\_adverse) with csv (see ‘AACT Database’ field). Was evaluated in different formats : **Adverse\_Counts\_Log, Adverse\_Counts\_Sqrt, Adverse\_Categ, Adverse\_Bin.**

**Allocation\_List**: RANDOMIZED, NON\_RANDOMIZED, NA\_RANDOMIZED (For one arm trials). Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Intervention\_Model\_List**: PARALLEL, SINGLE\_GROUP, CROSSOVER, SEQUENCIAL, FACTORIAL. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Masking\_List**: MASK\_NONE, SINGLE, DOUBLE, TRIPLE, QUADRUPLE. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Masking\_Detail\_List**: ['CARE\_PROVIDER', 'INVESTIGATOR', 'PARTICIPANT', 'OUTCOMES\_ASSESSOR']. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Arm\_Counts**: Created by merging database table (my\_studies) with csv (see ‘AACT Database’ field). Was evaluated also as **Arm\_Counts Log, Arm\_Counts\_Sqrt, Arm\_Categ**

**Country\_Counts**, **Continents\_List, Continent\_Counts, City\_Counts:** Created by merging database table (my\_locations) with csv (see ‘AACT Database’ field). Geonamescache library was used too. Also evaluated as **City\_Categ, Country\_Categ, Continent\_Categ** formats.

**Enrollment\_Counts**: regarding number of participants. Used directly from csv. Also, evaluated as **Enrollment\_Log**, **Enrollment\_Sqrt, Enrollment\_Categ**

**Phases\_List**: Five data frames df0, df1, df2, df3, df4, df5 for were produced based on phases, and analysed separately from this study. The number each data frame name ends is based on the phase (e.g., df1 for phase 1) except df5 referring to NA phase type which was named like this for convenience.

### Features not Used

**Study Title, official\_title, Other IDs, Acronym, Study URL, Brief Summary, nct\_id** Their purpose is to beunique identifiers of trials’ records, for registry organization reasons. They are given from administrators or in case of nct\_id from the public registry.

**Start Date, Completion Date, Primary Completion Date,** **First Posted, Results First Posted, Last Update Posted**Dates were not used for analysis, as they are not casual factor, but events happened during these periods are e.g., for year 2019 when covid studies started, covid trials were analysed. Especially, Completion date, could lead to ‘Snooping Bias’.

**Study Documents, Outcomes** (['OTHER\_PRE\_SPECIFIED', 'POST\_HOC',  
'PRIMARY', 'SECONDARY'])  
They included way too much missing data, they could not be utilized.

**Study Results, Primary Outcome Measures, Secondary Outcome Measures, Other Outcome** **Measures**They are not predesigned RCT characteristics but occurred after event of outcome. So, they cannot be part of predictor variables, or else  they could lead to ‘Data Snooping Bias’.

**Why Stopped**Why stopped feature from AACT database, was not used fo the probability of early termination, as it is associated only with outcome ‘Termination’ i.e., a perfect predictor.

## Visualization Pre-Processing

In this section plots and descriptive statistics were displayed per df of phase. Thus, pivot tables, chi-squared tests, count-plots, boxplots, kde-plots were displayed. Each feature was analysed through univariable logistic regression in relation to binary completion/termination outcome.

Additionally, in this section the best performed format of features were also chosen e.g., initial numeric, logarithmic etc. scale, through tests such as AIC. Tables for p-AIC an also p-values, confidence intervals from logistic models’ outputs were also displayed here.

All procedures were implemented through the creation of functions.

For achieving analysis per phase of trial, an input cell was placed on top of this step’s jupyter file, so one can input the data frame corresponding to trial want to analyse.

## Encoding

‘\_List’ data: List element data, were encoded through dummy creation. The problem was that list-element data cannot be used from models. Firstly, list-element rows were exploded to one row per element of list. However, this inflates sample size, as rows increase and duplicate for same nct\_id if more than one element occur in a list-row. For this reason, these data were then grouped by nct\_id, so number of rows remained the same as initial datasets (df0, df1, df2, df3, df4, df5).

‘\_Categ’ data: Were encoded as simple categorical data types with 2 levels. None of them had more than 2 levels, and no record could belong to more than one level.

‘\_Counts’, ’\_Log’, ’\_Sqrt’: Continues data were used as is and scaled to next level of analysis as mentioned below.

In this section, suffixes helped to choose which columns to binarize (0,1), make dummies when more than 2 levels or keep as in case of continuous data.

## Interactions

Interaction terms for Funder Type (e.g., INDUSTRY) and Intervention Type (e.g., DRUG) were created. As mentioned in literature, funder type is associated with trial termination, especially if industry (see Background field). However, this analysis purpose was to evaluate if funder type like industry (which most likely is pharmaceutical industry), considering intervention like drug (i.e., pharmaceutical product) results to greater probability of termination.

## Models

In this section, each trials phase was analyzed for early termination probability based on its characteristics. The data frames for each phase with shapes:   
df0: (2236, 130), df1: (26045, 155), df2: (28409, 155), df3: (16093, 155)

df4: (12813, 130), df: (173541, 160) containing all phases data frames.

Initially, each dataset was split into train and test set on with different proportion for each df. The train-test set proportion was decided based on background similar study that proposed 70:30 [12]. However, under-sampling used after train-test split (explained below), minimizes train set so that in some phases test set is greater.

Then train set only was Under sampled, due to great imbalance towards ‘Completed’ outcome. Test set should be left with the true distribution. Random Under sampling was chosen as the simplest method. In contrast, random over-sampling includes resampling methods to create greater ‘termination’ category (e.g., resampling minority class with replacement). This decision was also made from literature background [12]. In under-sampling methods, train set is minimized, as many rows are dropped from majority category. Combining the great sample sizes of data frames and the appropriate train-test split analogy per data frame, train samples remain efficient, with predictors by far less in size .

Scaling, used on train and test set. Scaling by using mean produced negative values in continues features, which is not justified from dataset values, so it was chosen to scale only by using standard deviation.

Finaly, models selected for performance comparison were:

* Classic Statistical Models:

Logistic Regression with Lasso (L1 penalty), Elastic Net Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA)

* Machine Learning Models:

K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Classification Pruned Tree, Random Forest, Gradient Boosting, Extreme Gradient Boosting (XGBoost).

Parameter tuning for models, was managed through k-fold cross validation with k=5 folds. Evaluation of models was done on test set, and metrics of accuracy, precision, recall, f1-score and AUC curves.

Furthermore, predictors contribution was evaluated through SHAP plots.

For achieving analysis per phase of trial, an input cell was placed on top of this step’s jupyter file, so one can input the data frame corresponding to trial want to analyse.

After pre-processing data engineering each phase was analysed separately through corresponding data frame.

## Descriptive Statistics – Visualization

In this section distribution of sample.

## Selected Model

XGBoost model outperformed at almost all phases (phase 1-phase 4 and non-applicable study phase), with exception of early phase 1 (named as phase 0 in this study) where QDA model was evaluated ss the best classifier. Model performance is in perfect alignment with background studies’ selected models. Models’ performance was evaluated with accuracy, precision, recall, f1-score and AUC curves.

## Features Contribution

Top feature contributor was for all phases the Enrolment for all phases. Another feature is Neoplasms disease RCTS, in mostly in early phases. Especially for phases 0 to 2, neoplasms undertake the second position in SHAP plots, while for phase 3 they much lower. Another, feature regards the Adverse event counts, which are present at the third position for phases 2-3 and lower for phase 4.

Feature contribution seems to be related with study phase. For example, adverse events are top contributor to corresponding phases 2-3, which main purpose is efficacy and drug safety, or later phases couldn’t initiate. Facilities and locations are another feature arising in different formats (e.g., city\_categ, country\_counts, continents etc.). Additionally, industry funder type was not a top contributor for this analysis, while mentioned by literature.  
Result of this study is aligning with background results of previous similar studies in most of the outcomes, as far as it concerns feature contribution.

Differences are observed mostly to disease categories or type of interventions. This study did not conclude any surgical or cardiovascular field disease as high contributors, as mentioned by literature, while intervention method of injection was in top 20 features in many phases and not surgical as mentioned from background studies.

# Discussion

## Comparison with Literature - Strengths and Limitations

Although, most of the results are aligned with background similar studies, some topics differ from this study either on models, results and data engineering methods. Features contribution and models’ performance are in line with background studies. Methods used were also similar. In general, this study analysed more features and evaluated more models, for all trial phases. So it is a more general approach not limited to specific topics or stages of n RCT.

### Strengths

More specifically, previous studies did not consider interactions such as funder type and intervention type (eg., drug-industry sponsor). Regarding adverse events, this study considered the type and the number of adverse events, but also the organ system located (e.g., ear), while in literature disease category is mostly referred and not adverse counts. The type of intervention method was analysed e.g., oral, topical, injection, surgical.

This analysis checked for different formats of a feature (e.g., enrolment, enrolment\_log etc.). Additionally, more models were evaluated than in literature but also ML to non-ML models were compared. Evaluation of ML models without considering the traditional methods would not reflect their actual impact [13] .In background studies mostly Random Forests and secondly XGBoost models are evaluated. Precent study had double purpose, first to compare classic statistical models with Machine Learning models and secondly to evaluate which model outperformed to prediction of termination probability.

Feature of Funder Type was analysed not only based provided categories (e.g., industry) but also two additional categories that were created for educational and hospital facilities, but also the interaction with intervention type (e.g., INDUSTRY x DRUG).

Finaly, all phases and even ‘Not Applicable’ type of phases were analysed separately, while in literature, most of the times, not all stages are evaluated.

### Limitations

The presence of missing values decreased sample size for many features and to features defined from others e.g., country\_counts as defined Countries\_Listetc.

Missing data were filled, where possible, while background studies usually dropped them based on a pre-specified field needed to have no nulls. However, current study did not use NLP text mining methods, which is mentioned in literature studies. What was used here, was based on regular expression text mining technics, where specific words were searched from unstructured text data (e.g., str.contains() command etc.). The drawback of regular expression text mining is that it is not possible to search for all appropriate words and expressions leading to the desired result. Another, deficiency of current study, is that ‘Why Dropped’ text from database source, was not analyzed, as this feature is always associated with ‘Termination’ outcome i.e., perfect predictor. However, it could be investigated to the ratios each of these reasons appear in ‘Termination’ category.

What need to be done additionally, is the analysis of enrollment feature as interaction to disease field, location etc. to conclude why low enrolment occurs as other studies have done.

## Recommendations for Practice

Literature seems to provide common outcomes, regarding the early termination reasons of RCTs. What needs to be done is further investigation those specific reasons, including interactions with other features, in order to evaluate them better. For example, low accrual is a common top contributor for trial termination, but the factors which lead to low accruals rates are not deeply evaluated. Similarly, for the second top contributor, which is neoplasm diseases, where one could expect to have greater completion rates.

All these need further investigation to detect deeper reasons leading to early termination, thus improving trials’ design or even communicating consent forms in a more efficient way to attract more volunteers, especially for crucial disease fields.

# References

initial value