



**INVESTIGATING THE DEPENDENCE ON MACHINE SIZE OF THE  
ENERGY CONFINEMENT IN TOKAMAKS USING  
DATA-DRIVEN METHODS**

Master Thesis  
presented by

**Karina Chiñas Fuentes**

Thesis Supervisors

Prof. Dr. Geert Verdoolaege  
PhD Student Joseph Hall

Erasmus Mundus Program on  
Nuclear Fusion Science and Engineering Physics

July 8, 2023





European Master of Science in  
Nuclear Fusion and Engineering Physics

**INVESTIGATING THE DEPENDENCE ON MACHINE SIZE OF THE  
ENERGY CONFINEMENT IN TOKAMAKS USING  
DATA-DRIVEN METHODS**

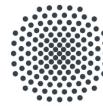
Master Thesis  
presented by

**Karina Chiñas Fuentes**

Thesis Supervisors

Prof. Dr. Geert Verdoolaege  
PhD Student Joseph Hall

July 8, 2023



Universität Stuttgart





# **Declaration in lieu of oath**

I am aware that plagiarism is not consistent with academic and research ethics. I declare in lieu of oath that all the work presented in this report, which is not cited in the text or by references, is my own work.

City, country, the

GHENT, BE.

  
\_\_\_\_\_  
Karina Chiñas Fuentes

# **Copyright Agreement**

I hereby grant the FUSION-EP consortium the non-exclusive right to publish this work.

I declare that this work is free of copyright claims of third parties.

City, country, the

GHENT, BE.

  
\_\_\_\_\_  
Karina Chiñas Fuentes



## ACKNOWLEDGEMENTS

I am infinitely grateful to the FUSION-EP cohort and the European Education and Culture Executive Agency for selecting me to participate in this Erasmus master's programme. This intercultural and academic experience empowered me to reach my full potential.

I am also grateful to Prof. Dr Geert Verdoollaeghe and PhD Student Joseph (Joe) Hall, especially to Joe for being patient, encouraging me to follow my ideas (even with the risk of not finding something fruitful), helping me to discern when it was time to stop the analysis and start a new one, and always teaching me with laughs.

I am, of course, also grateful to all my professors in this master's programme, especially to Dr Alf Köhn-Seemann, whose lecture on Fusion Research at Universität Stuttgart plays a vital role in this thesis.

I would also like to thank my family – gracias abuelito y abuelitas, gracias mamá, gracias papá y gracias hermanas, les amo – and friends who always supported me and checked my well-being. Thank you, Carlos Iván Espejel Velázquez, for helping me find this programme and always believing in me. I want to especially thank Daniel Hernández Mota, my bachelor's peer, who always answered my questions regarding random forests and helped me to gain deeper insight into machine learning-related topics. And to Luis Carlos Herrera Quesada, my master's peer, who helped me understand complicated subjects, particularly transport, turbulence, and stellarator-related topics.

All this work is dedicated to you.

Finalmente, quiero agradecer a todas las personas que se preocupan por y cuidan el medio ambiente, especialmente a las personas más vulnerables y a los indígenas de México. Este trabajo también va dedicado a ustedes.

Thank you, gracias, dankeschön, dankuwel, ευχαριστώ πολύ.



## CONTENTS

ACKNOWLEDGEMENTS . . . . .	iii
ABSTRACT . . . . .	vii
NOTATION . . . . .	ix
1 INTRODUCTION . . . . .	1
1.1 Energy Confinement Time . . . . .	2
1.2 What is H-mode? What are ELMs? . . . . .	5
2 INFLUENTIAL POINTS AND DATA ANALYTICS . . . . .	9
2.1 Regression Diagnostics . . . . .	9
2.2 Random Sampling . . . . .	12
2.3 Multicollinearity . . . . .	14
2.3.1 Assessing Dependencies in Datasets . . . . .	17
2.4 Model Comparison . . . . .	19
2.5 Conclusions . . . . .	21
3 MACHINE LEARNING ALGORITHMS . . . . .	23
3.1 Feature Selection . . . . .	23
3.1.1 All Variables of Interest . . . . .	23
3.1.2 Entropy Variables . . . . .	25
3.1.3 Research Variables . . . . .	28
3.1.4 Low Multicollinearity Variables . . . . .	29
3.2 Classification . . . . .	29
3.2.1 Imbalanced Data . . . . .	30
3.2.2 Gaussian Process and Random Forest . . . . .	33
3.2.3 Feature Importance . . . . .	38
3.3 Results . . . . .	39
3.3.1 Predictions . . . . .	41
3.4 Other Applications in Fusion . . . . .	41
4 TOKAMAK ANALYSES . . . . .	43
4.1 The Physics . . . . .	43
4.1.1 Instabilities in a Tokamak . . . . .	43
4.1.2 Insights within the Dataset . . . . .	45
4.2 Variables of Importance and Tokamak Characteristics . . . . .	48
4.2.1 Dynamic Analyses with Random Sampling . . . . .	50
4.3 The Engineering Perspective . . . . .	52
5 CONCLUSIONS . . . . .	55
A DERIVATION OF RESEARCH VARIABLES . . . . .	59
REFERENCES . . . . .	65



## ABSTRACT

Accurately determining the energy confinement time of fusion devices is essential in the decision-making process for constructing such machines – methods to study this parameter include simulation codes and empirical scaling laws. The International Thermonuclear Experimental Reactor (ITER) used the 1998 energy confinement scaling law to design its reactor through the global H-mode confinement database. The study yielded that the major radius – a cost-related parameter – scaled with a factor of two. However, a reevaluation indicates a considerably weaker dependency when considering the latest database update. In other words, the influence of the major radius on the energy confinement time reduces.

The primary objective of this thesis is to contribute to understanding the factors behind the reduced significance of machine size in energy confinement. The project divides into three essential parts. Firstly, the focus is on identifying the critical observations within the new database that influence the relevance of this parameter the most. Statistical tests are then applied to examine whether data-related issues, such as collinearity, contribute to the observed effects. The second part employs two machine-learning algorithms, namely Gaussian process and random forests, to identify the variables that characterize the decrease in the influence of the major radius. These variables may not necessarily belong to the energy confinement time scaling law. In the final part, the insights gained from the preceding chapters are related to tokamak physics, which includes revisiting the theory of instabilities and exploring engineering aspects that characterize each tokamak in the dataset.

The results indicate that decreased influence on machine size is not attributed solely to data-related issues. Instead, four main parameters emerge as predictors of whether this decrease in influence will affect an observation. These parameters include the normalised ion gyroradius, the normalised ion collision frequency, the safety factor, and the normalised plasma beta. When the best algorithm was employed to predict outcomes in new machines, namely ITER and SPARc, it was observed that their size significantly impacted their energy confinement time. Additionally, spherical machines exhibit distinct clusters within these four variables, displaying notable differences from less compact machines. However, the dataset only comprises 3.8% of spherical machines, indicating a need for more extensive data on spherical tokamaks to further enhance the project’s outcomes. Moreover, it was noted that most of the observations classified as unaffected could surpass the Greenwald limit. Finally, it was observed that distinguishing between observations that influence the regressor associated with the major radius makes it possible to mitigate the variability in predicting the energy confinement time when using ordinary least squares.

Understanding the factors contributing to the reduced significance of machine size in energy confinement has significant implications for designing and optimising future fusion devices, leading to more efficient and cost-effective solutions. Furthermore, the need for additional data on spherical tokamaks underscores the importance of expanding research efforts to refine the predictive models for future fusion devices.

**Keywords:** Global H-mode confinement database, Energy confinement time, Tokamak, Major radius, Gaussian process, Random forest, Data analytics.



## NOTATION

### ACRONYMS AND CONNOTATIONS

ELMs	Edge Localized Modes (also referred as ELMy)
H-mode	High-confinement mode
IAEA	International Atomic Energy Agency
ITER	International Thermonuclear Experimental Reactor
L-mode	Low-confinement mode
LCFS	Last Closed Flux Surface
MHD	Magnetohydrodynamics
OLS	Ordinary Least Squares
Shot	Single experimental run or discharge of a tokamak device
Subset	Reduction in number of rows, when referring to a dataset
MCL	Multicollinearity
DS, ds	Dataset
w.r.t.	with respect to

### DATABASES

DB2	Global H-confinement mode, version 2.8 (also as DB2.8 or DB2P8)
DB5	Updated DB5, version 5.2.3
STDB5	Standard criteria applied to DB5 and ELMy subset
new_DB5	Subset of STDB5, without DB2 observations
decreasing_DB5	Subset of new_DB5
small_ds	Subset of decreasing_DB5, without DB2 observations
big_ds	Subset of decreasing_DB5, without DB2 observations
clean_DB5	Modified STDB5, with a treatment in missing values and reduction of columns

### GENERAL MATHEMATICS

$\sim$	close to a real number
$\subset$	Subset
$T$	Transpose
$\mathbb{E}[\cdot]$	Expected value of one input $\cdot$

$\mathcal{N}$	Normal distribution
$\mathcal{O}(\cdot)$	Big-O notation, denoting the complexity of an algorithm
$\nabla$	Vector differential nabla-operator
$\mathbf{I}_n$	$n \times n$ identity matrix
$Z \sim$	Distribution of random variable $Z$
$\mathcal{GP}$	Gaussian Process
$\sigma$	Standard Deviation
$\sigma(z)$	Any sigmoid function
$k(\mathbf{x}, \mathbf{x}')$	kernel function of two inputs $\mathbf{x}, \mathbf{x}'$
$\text{Cov}[\mathbf{x}, \mathbf{x}']$	Covariance of two inputs $\mathbf{x}, \mathbf{x}'$
$\text{Var}[\cdot]$	Variance of one input $\cdot$
$M$	Total number of columns in dataset
$n$	Number of rows/observations in a dataset
$\mathbf{X}$	$n \times (M + 1)$ matrix representing a database
$x_i$	$i$ -th row of $\mathbf{X}$
$\mathbf{y}$	Target variable, an $n$ -dimensional column vector
$\hat{\beta}$	Parametric vector, an $(M + 1)$ -dimensional column vector
$h_i$	Leverage of the $i$ -th observation
$r_i$	Studentized residual of the $i$ -th observation
$\varepsilon$	Gaussian noise, an $n$ -dimensional column vector
$E$	Entropy of a dataset
$S_{ij}$	Similarity between the $x_i$ and $x_j$ observations

### TOKAMAK-RELATED AND PHYSICS SYMBOLS

$\mu_0$	Vacuum magnetic permeability	$4\pi \times 10^{-7} \text{ Hm}^{-1}$
$q$	Elementary charge	$1.602176634 \times 10^{-19} \text{ C}$
$\bar{n}_e$	Averaged electron density	$10^{19} \text{ m}^{-3}$
$\kappa_a$	Elongation of the LCFS	$\text{m}^3$
$a$	Minor radius of the tokamak	$\text{m}$
$R_{geo}$	Major radius of the tokamak	$\text{m}$
$P_{fus}$	Fusion power	MW
$P_{l,th}$	Thermal power lost due to the transport through the LCFS	MW
$P_\alpha$	Energy carried by the alpha particles contributing to D-T fuel	MW

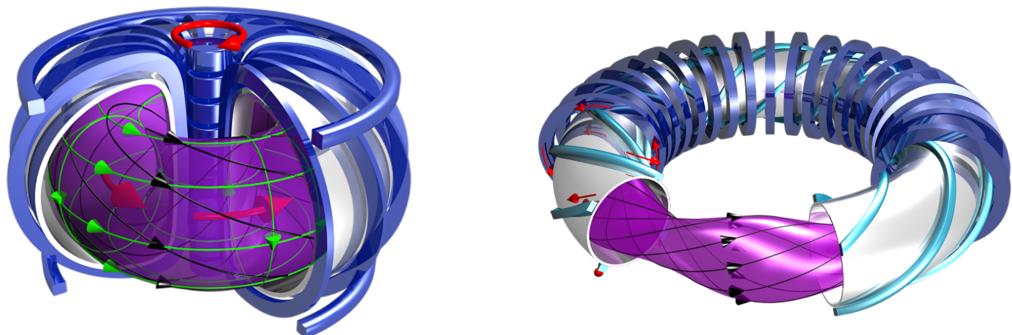
$P_\Omega$	Ohmic heating	MW
$P_{aux}$	Auxiliary heating power	MW
$P_{rad}$	radiated power	MW
$B_\theta$	Poloidal magnetic field	T
$B_t$	Toroidal magnetic field	T
$I_p$	Plasma current	MA
$\langle \sigma v \rangle$	Fusion reaction rate	$\text{m}^3 \text{s}^{-1}$
$\tau_{E,th}$	Energy confinement time	s
$\hat{T}$	Total averaged temperature of the plasma	eV
$\omega_c$	Ion cyclotron frequency	Hz
$M_{eff}$	Effective atomic mass of the plasma	amu
$\beta_t$	Plasma pressure normalized to $B_t$ ( $= 2\mu_0 \bar{n}_e \hat{T}/B_t$ )	-
$\epsilon$	Inverse aspect ratio ( $= a/R_{geo}$ )	-
$\gamma_{rad}$	Parametrization constant for defining radiative losses $\in [0, 1]$	-
$\nu_*$	Ion collision frequency normalized to bounce frequency of trapped particles.	-
$\rho_*$	Ion gyroradius normalized to $a$ .	-
$Q$	Fusion Gain ( $= \hat{P}_{fus}/P_{aux}$ )	-
$q_{95}$	Plasma safety factor ( $= \epsilon B_t/B_\theta$ ) at the 95% poloidal flux surface	-
$q_{cyl}$	$q_{95}$ with cylindrical approximation	-
$\alpha_x$	Regression coefficient of the $x$ variable in the scaling law for $\tau_{E,th}$	-
$\chi_x$	Regression coefficient of the $x$ variable in the scaling law for $\omega_i \cdot \tau_{E,th}$	-
$\alpha_R \sim \cdot$	Regression coefficient $\alpha_R$ of $R_{geo}$ is close to a number $\cdot$	-



# I

## INTRODUCTION

Around 1919, Rutherford conducted numerous experiments at the University of Manchester, which resulted in the first artificial nuclear reaction performed on Earth. Later, scientists started dreaming of utilizing this energy for electricity production. Initially, linear magnetic confinement devices were studied but quickly discarded due to their instabilities. Then, toroidal devices were analysed. It was noted that a strong twisted magnetic field along the torus is crucial for circular confinement to avoid charge accumulation, which terminates confinement in this configuration. These particular field lines can be created in two devices: tokamaks and stellarators. Stellarators use external coils, some of which revolve helically around the plasma. While tokamaks use an induced net toroidal plasma current [1], [2]. More on the characteristics of Tokamaks are found in Chapter 4. Figure 1.1 shows the main difference between these devices.



**Figure 1.1:** In both images, the toroidal coils are shown in dark blue, the red arrows represent a current, and the black arrows represent the required rotational transform of the magnetic field lines. LEFT: tokamak showing the poloidal magnetic field lines, as green arrows, resulting from the presence of the induced current. RIGHT: stellarator having helical coils, in light blue [2]. Images from [3].

In more modern times, between these two devices, tokamaks are the popular option for fusion research due to their simple geometry. Other techniques implement laser pulses, like the research done at the National Ignition Facility [4]; however, these devices are not discussed in this project. The International Thermonuclear Experimental Reactor (ITER) is a megaproject composed of 35 nations dedicated to investigating tokamak physics and helping to be one step closer to employing fusion energy for peaceful usage [5].

In 1991, the International Atomic Energy Agency (IAEA) released a document indicating the conceptual design of ITER, where they mentioned two ways of predicting the energy confinement time: by codes modelling the transport in the plasma and by constructing an empirical scaling law, with the latter being the preferred option [6]. Today, this method remains the dominant approach when investigating the energy confinement time of any tokamak, mainly because a direct estimate of this variable is still unavailable, even with the significant advancements in theoretical physics in tokamaks [7], [8].

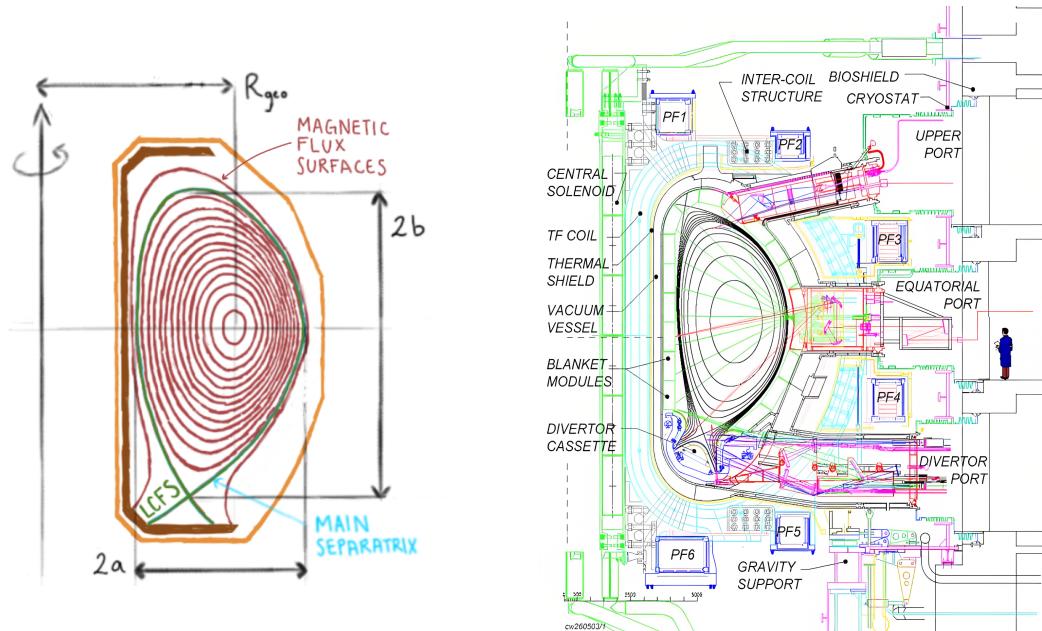
## 1.1 Energy Confinement Time

Since the plasma experiences a decrease in turbulent transport and improved confinement is achieved during the high-confinement mode (H-mode), this technique is expected to be the nominal inductive operation for ITER; with edge localised modes (ELMs) – small periodic bursts of energy and particles ejected from the plasma that impact the plasma-facing components [9], [10].

The expression of the energy confinement time  $\tau_{E,th}$  [s] results from an empirical standard power law conformed by eight plasma parameters: plasma current  $I_p$  [MA], toroidal magnetic field  $B_t$  [T], central line-averaged electron density  $\bar{n}_e$  [ $\cdot 10^{19} \text{ m}^{-3}$ ], thermal power lost due to the transport through the last closed flux surface (LCFS)  $P_{l,th}$  [MW], major radius  $R_{geo}$  [m], elongation of the LCFS  $\kappa_a = V/(2\pi R_{geo}\pi a^2) = b/a$  (with  $V$  [ $\text{m}^3$ ] being the plasma volume inside the LCFS,  $a$  [m] the minor radius of the tokamak, and  $2b$  [m] the height between the upper and lower plasma edge), inverse aspect ratio  $\epsilon = a/R_{geo}$ , and the effective atomic mass of the plasma  $M_{eff}$  [9]; see Figure 1.2 for a visual description of the plasma geometrical parameters. The scaling law of the energy confinement time is:

$$\tau_{E,th} = \alpha_0 \cdot I_p^{\alpha_I} \cdot B_t^{\alpha_B} \cdot \bar{n}_e^{\alpha_n} \cdot P_{l,th}^{\alpha_P} \cdot R_{geo}^{\alpha_R} \cdot \kappa_a^{\alpha_\kappa} \cdot \epsilon^{\alpha_\epsilon} \cdot M_{eff}^{\alpha_M}. \quad (1.1)$$

Each accompanying coefficient  $\alpha_x$  is estimated by applying the logarithm to Eq. (1.1) and then using the ordinary least squares (OLS) regression technique. There is another version that considers the triangularity  $\delta$ , another geometrical parameter that characterizes the plasma cross-section of the LCFS; however, I will not consider this parameter as it turns out to have a weak dependence for the dataset of interest [7].



**Figure 1.2:** LEFT: A drawing of a simple poloidal cross-section of a tokamak is used to depict its geometrical parameters; open field lines ending in an open divertor. More on flux surfaces in the following section. RIGHT: ITER poloidal cross-section illustrates a more complex plasma configuration. Image from [11]. It is worth noting that there is the possibility of a second separatrix at the top and that the open field lines could end in a closed divertor.

The data used to estimate the regression coefficients has been collected since 1989 by the H-mode Database Working Group. This data is referred to as the Global H-mode Confinement Database, and since 2001, it has followed a framework established by the International Tokamak Physics Activity (ITPA). Version 2.8,

ELMy subset, of this database (DB2.8<sup>1</sup>) was utilised to estimate ITER's physics and its design. DB2.8 is characterised by nine devices: ASDEX, ASDEX-UPGRADE (AUG), Alcator C-MOD, DIII-D, JET, JFT-2M, JT-60U, PBXM, and PDX. It contains 1310 data points [7]. When computing the OLS to Eq. (1.1), using the flags within DB5 to retrieve DB2, one obtains the following regression parameters

$$\tau_{E,th} = 0.05 \cdot I_p^{0.78} \cdot B_t^{0.32} \cdot \bar{n}_e^{0.43} \cdot P_{l,th}^{-0.67} \cdot R_{geo}^{2.22} \cdot \kappa_a^{0.39} \cdot \epsilon^{0.58} \cdot M_{eff}^{0.18}. \quad (1.2)$$

Nowadays, one can work with the updated version of the global H-mode confinement database DB5 version 2.3, released in 2021. This database is characterised by having 14,153 observations of nineteen different devices: ASDEX, AUG, Alcator C-MOD, COMPASS, DIII-D, JET, JFT-2M, JT-60U, MAST, NSTX, PBXM, PDX, START, T-10, TCV, TDEV, TEXTOR, TFTR, TUMAN-3M [12]. When considering the ELMy subset of DB5 with the standard criteria (STDB5), composed of 6,252 observations and 18 devices (without TUMAN-3M), one obtains that the regression parameters of Eq. (1.1) are the following:

$$\tau_{E,th} = 0.08 \cdot I_p^{1.13} \cdot B_t^{0.07} \cdot \bar{n}_e^{0.17} \cdot P_{l,th}^{-0.68} \cdot R_{geo}^{1.45} \cdot \kappa_a^{0.28} \cdot \epsilon^{0.06} \cdot M_{eff}^{0.21}. \quad (1.3)$$

The standard criterion is the same applied to previous databases and it is characterised by shots having [7]:

- (i) no pellet fueling,
- (ii) no strong internal transport barriers,
- (iii) no excessive magnetohydrodynamics activity near the  $\beta_t$ -limit<sup>2</sup>,
- (iv) steady energy content,
- (v) limited radiative power,
- (vi) limited fast particle energy content, and
- (vii) minimum safety factor ( $q_{95} \geq 2.5$ ).

Considering this, when comparing Eq. (1.2) with Eq. (1.4) one can notice that, except for  $P_{l,th}$ , all the regression parameters change notably, this is due to the complexity and uncertainty in the database. This observation has motivated the use of robust regression techniques, such as those rooted in Bayesian probability and the geodesic least squares regression technique [7], [13].

One might think that the most significant influence on  $R_{geo}$  may come from the presence of spherical tokamaks (MAST, NSTX, and START), as this type of device was not part of DB2. However, when removing the spherical tokamaks from STDB5, the scaling law of the energy confinement time gets the following parameters:

$$\tau_{E,th} = 0.09 \cdot I_p^{1.22} \cdot B_t^{-0.04} \cdot \bar{n}_e^{0.15} \cdot P_{l,th}^{-0.67} \cdot R_{geo}^{1.28} \cdot \kappa_a^{0.11} \cdot \epsilon^{-0.01} \cdot M_{eff}^{0.23}, \quad (1.4)$$

resulting in a further decrease in  $\alpha_R$ .

Two quick observations follow from this: (i) the new information suggests that the energy confinement time for the ITER scenario decreases, and (ii) the major radius of a tokamak is not as influential on the energy confinement time, as the DB2.8 suggests. The latter can be of particular interest since the cost of a machine scales with  $B_t^2 \cdot R_{geo}^3$  [14].

---

<sup>1</sup>For ease of reading, this font will be used to refer to databases only.

<sup>2</sup>More explained in Chapter 4.

A similar technique to obtain information on the energy confinement time is with the scaling law of the dimensionless physics variables. This implies the application of the Buckingham  $\Pi$ -Theorem, which states that a function that expresses a physical law has the property of generalised homogeneity, meaning that it does not depend on the units of measurement [15]. The method was first applied to tokamak physics by Kadomtsev and a perspective on anomalous transport described by the Fokker-Planck equation was implemented by Connor and Taylor [7], [16]. The result is the following:

$$\omega_i \cdot \tau_{E,th} = \chi_0 \cdot B_t^{\chi_B} \cdot \rho_*^{\chi_\rho} \cdot \beta_t^{\chi_\beta} \cdot \nu_*^{\chi_\nu} \cdot q_{95}^{\chi_q} \cdot \kappa_a^{\chi_\kappa} \cdot \epsilon^{\chi_\epsilon} \cdot M_{eff}^{\chi_M} \quad (1.5)$$

where dimensionless is achieved by multiplying the ion cyclotron frequency  $\omega_i = qB_t/M_{eff}$  [16]. Here,

$$\rho_* = \rho_i/a \quad (1.6)$$

is the ion gyroradius  $\rho_i$  normalised to the minor radius,  $\nu_*$  is the ion collision frequency  $\nu_{ii}$ ,

$$\nu_* = \nu_{ii} \left( \frac{m_i \cdot R_{geo}^3}{q \hat{T} a^3} \right)^{1/2} \cdot q_{cyl} R_{geo}, \quad (1.7)$$

normalised to the bounce frequency of the trapped particles,  $q_{cyl}$  is the safety factor at the 95% poloidal flux surface with cylindrical approximation,

$$q_{cyl} = \frac{2\pi}{\mu_0} \cdot \frac{a^2 \kappa_a B_t}{R_{geo} I_p}; \quad (1.8)$$

where  $I_p$  is in A, and  $m_i$  is the ion mass. By considering the Kadomtsev constraint, one can get the dimensionless regression parameters  $\chi_x$ , from the engineering regression parameters  $\alpha_x$  as [14], [16]:

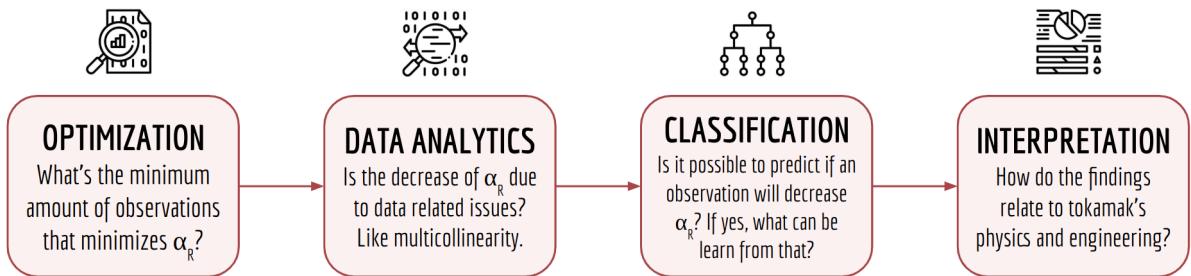
$$\begin{aligned} \chi_B &= 0, & \chi_q &= \frac{\alpha_R - 4\alpha_I + 3\alpha_P - 2\alpha_n}{5(1 + \alpha_P)}, \\ \chi_\rho &= \frac{2(-3\alpha_R - 3\alpha_I - 9\alpha_P + \alpha_n)}{5(1 + \alpha_P)}, & \chi_\kappa &= \frac{\alpha_k + \alpha_P}{1 + \alpha_P}, \\ \chi_\beta &= \frac{\alpha_R + \alpha_I + 8\alpha_P + 3\alpha_n}{5(1 + \alpha_P)}, & \chi_\epsilon &= \frac{2\alpha_\epsilon - 3\alpha_R + \alpha_I - 5\alpha_P + 2\alpha_n}{2(1 + \alpha_P)}, \\ \chi_\nu &= \frac{-\alpha_R - \alpha_I - 3\alpha_P + 2\alpha_n}{5(1 + \alpha_P)}, & \chi_M &= \frac{5\alpha_M + 3\alpha_R + 3\alpha_I + 4\alpha_P - \alpha_n - 5}{5(1 + \alpha_P)}, \end{aligned} \quad (1.9)$$

which facilitates the shift from the engineering scaling law to the dimensionless scaling law. It is in fact preferable to obtain the estimates of  $\chi_x$  in this way than direct estimation. This is because  $\rho_*$ ,  $\beta_t$ , and  $\nu_*$  are considerably affected by error bars and there is considerable collinearity between the intercept,  $\epsilon$ , and  $\beta_t$  [13].

Various intriguing analyses can be conducted on the energy confinement time scaling law with the updated database. For instance, one can focus on the correlation among variables and how this affects the interpretability of the results. OLS characterizes by assuming that there is no multicollinearity in the

variables and that their uncertainty is negligible concerning the response variable [17]. Nevertheless, this is not true for the global H-mode confinement database [13]. Furthermore, work has been done to obtain the regressor parameters  $\alpha_x$  using a method that does not neglect the uncertainty in the predictor variables [7]. Others have investigated the energy confinement time with non-linear or non-power scaling forms [18], [19]. However, this project uses data analysis and machine learning algorithms to focus on the dominant causes influencing the decrease of  $\alpha_R$ . The aim is to use these algorithms to inspect all the columns available in DB5 and determine if any provide insight into explaining the change in  $\alpha_R$ .

The proposed workflow involves three steps: first, identifying a subset of records corresponding to shots that contribute to the decrease in  $\alpha_R$ ; second, applying data-driven algorithms to this subset to determine common characteristics; and third, analyzing the results of step two for interpretability in terms of tokamak physics. Figure 1.3 illustrates this idea.



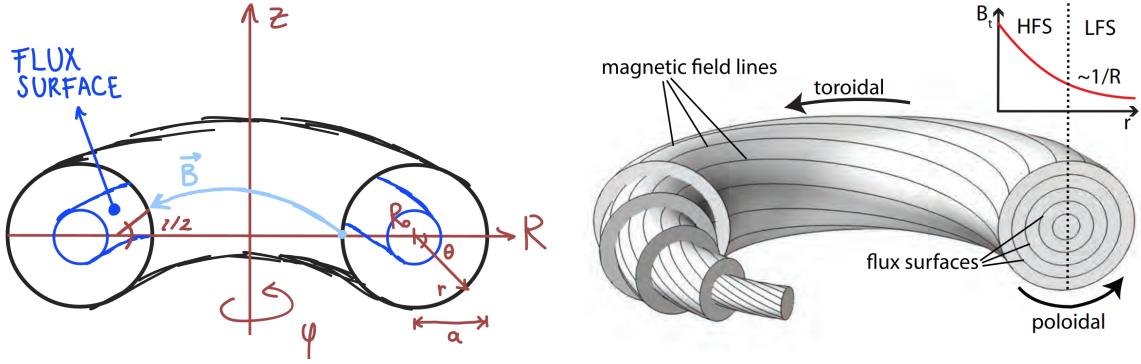
**Figure 1.3:** Proposed workflow to study the change of the regression parameter  $\alpha_R$  in the updated global H-mode confinement database, standard-ELMy subset.

The report's structure follows: Chapter 1 concludes with a brief description of the phenomena described by the database. Chapter 2 provides an overview of the analyses conducted to identify the minimum subset of variables contributing to the most significant reduction in  $\alpha_R$ . Additionally, this chapter encompasses various regression diagnostics employed to characterize different datasets. Chapter 3 focuses on the classification analyses, including the underlying theory, and examines the most influential variables in the algorithms' learning process. This chapter also incorporates the information obtained from Chapter 2. Chapter 4 briefly overviews tokamak physics, focusing on stability theory. Then, the physics inside a dataset is discussed. Following the investigation, data analysis is applied to investigate the main characteristics, including materials, within the observations that influence  $\alpha_R$ . Finally, Chapter 5 summarizes the main conclusions derived from this project. Appendix A shows the theory needed to understand statements made in Chapter 3.

## 1.2 What is H-mode? What are ELMs?

As briefly mentioned in the previous section, it is required to have twisted magnetic field lines<sup>3</sup> to confine the plasma effectively. Another vital requirement for fusion is the existence of magnetic flux surfaces, as alpha-particles could take about  $10^5$  turns in a device to deliver their energy to the plasma, and they can be confined within them [1]. Furthermore, the shape of the flux surface indicates the density and temperature profiles. Flux surfaces are achieved by creating open twisted lines [1]. They also relate closely to the safety factor. For a better understanding, let us review the magnetic field configuration in a Tokamak, shown in Figure 1.4.

<sup>3</sup>Here, a magnetic field line is a curve everywhere tangential to the direction of the magnetic field.



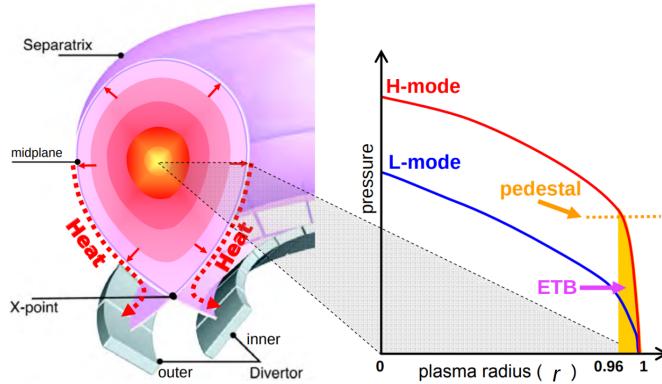
**Figure 1.4:** LEFT: Drawing showing the two main coordinates implemented to study magnetic confinement in tokamaks. The confinement region is for  $r < a$ , the LCFS is at  $r = a$ , and the scrape-off layer (SOL) is at  $r > a$ . RIGHT: another perspective of magnetic fluxes. HFS and LFS stand for high and low field sides, respectively.

Image from: Fig. 1.3 in [20]

The rotational transform  $\iota$  serves as a radial-dependent indicator of the twists in magnetic field lines. If  $\iota$  is a rational number, the magnetic field lines close after  $n_t$  toroidal turns and  $n_p$  poloidal turns. Conversely, if  $\iota$  is irrational, a flux surface will be covered ergodically by a single field line. Some instabilities result from having a rational  $\iota$  [1]. The safety factor is defined as the number of toroidal transits required for one full poloidal transit, also being the inverse of the rotational transform. In a linear tokamak approximation, the safety factor is expressed as  $q_s(r) = r B_\varphi / (R_{geo} B_\theta)$ . The radial change of the rotational transform is known as magnetic shear,

$$s = \frac{r}{q_s(r)} \cdot \frac{dq_s(r)}{dr}, \quad (1.10)$$

which is the degree of change of the twist in the magnetic field. High magnetic shear is generally advantageous for plasma confinement as it reduces transport losses by creating *transport barriers* at the edge, or *pedestal* of the plasma [1]. Edge transport barriers (ETB) are narrow radial regions with a strongly reduced particle and thermal diffusivity resulting in decreased local fluctuation levels [1]. The result is referred to as high-confinement mode; when these ETB are not observed, the plasma is at low-confinement mode. Figure 1.5 shows the change in plasma pressure as a function of the radial position for both L- and H-mode. The result of the H-mode is a considerable increase in the confinement time by two [1].

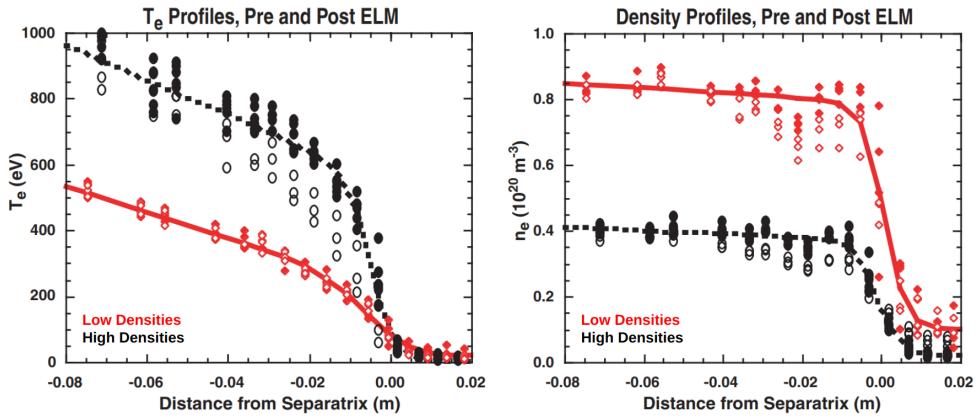


**Figure 1.5:** Formation of edge transport barriers based on pressure profile. They can also be observed in temperature and density profiles [1]. The inner side is also the HFS. Adapted image from: Fig. 1.4 in [20].

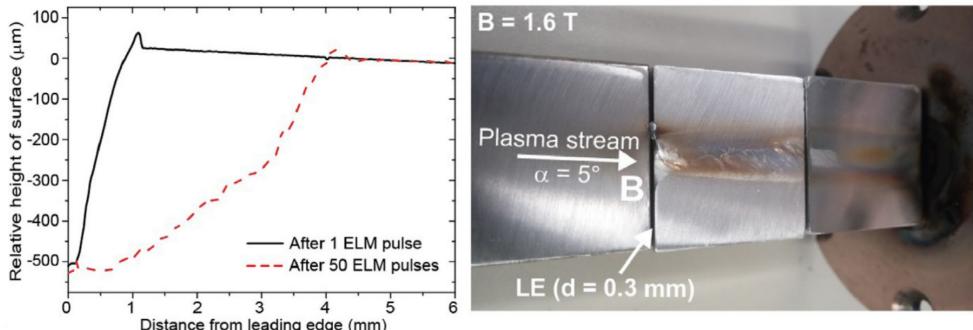
However, the transition from L-mode to H-mode is not solely attributed to magnetic shear. The pedestal

temperature also plays a crucial role in developing ETB, as well as increasing the fusion gain of a device [1]. In fact, this transition remains not fully understood, and despite extensive research on theories and scaling laws, the fundamental physics mechanisms behind it have yet to be revealed [21].

Despite the remarkable improvement in performance, the H-mode is not temporally stable. The edge transport barrier relaxes quasi-periodically, causing a flattening of the edge profiles [1]. The relaxation process is caused by losing energy and expelling particles into the SOL in a quasi-periodic manner – these phenomena are referred to as ELMs [20]. This can result in erosion of the plasma-facing components when the expelled energy is too high [1]. Figure 1.6 shows the electron temperature (left) and density (right) profiles at H-mode for two different situations. Notice how the profiles flatten after the ELM has occurred. Moreover, Figure 1.7 presents the melting of plasma-facing components in ASDEX-Upgrade caused by ELMs.



**Figure 1.6:** DIII-D electron temperature (left) and density (right) profiles at H-mode for high and low densities. Full-coloured symbols belong to observations before and ELM type-I happened, and non-coloured symbols belong to observations after said ELM happened [22]. Adapted images from: Fig. 7 in [22], presented by [1].

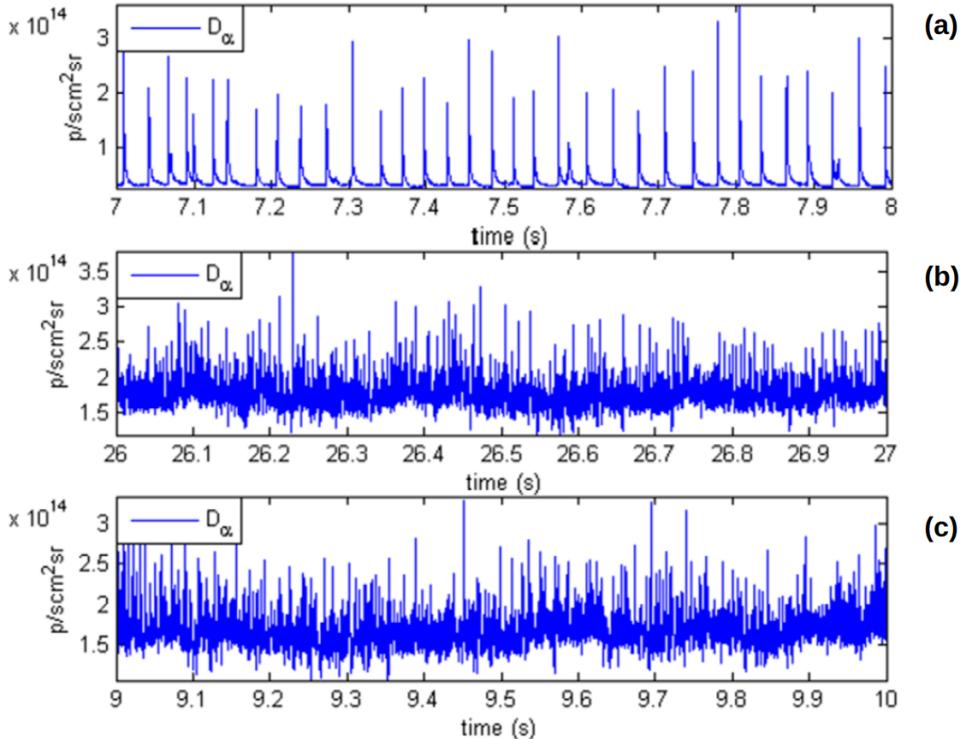


**Figure 1.7:** ASDEX-Upgrade melt studies caused by ELMs on plasma-facing divertor plates. LEFT: plot of erosion severity given the number of ELM pulses and spatial profile. Right: image of the eroded divertor plate [23]. Adapted images from: Fig. 7 in [23], presented by [1].

To observe ELMs, high spatial and temporal resolution diagnostic systems are required, as they can have sizes between 2-5 cm and lifetimes of 20-50  $\mu\text{s}$  [24]. According to their characteristics, like repetition frequency and energy loss, ELMs can be classified; some of the most common are [20], [25], [26]

- (i) **Type I** – the most dangerous: they occur in fully developed H-mode plasmas, and 5% to 10% of the pedestal's energy is lost in a crash. The higher the size of the vessel, the higher the loss. Their frequency range is 10-250 Hz; it increases with heating power and drops with increasing current. Figure 1.8 (a) shows the presence of this ELM as an emissivity signal in JET.

- (ii) **Type II**: these characterize for having several hundreds of Hz with rather small energy loss. They tend to happen in highly collisional plasmas (large density and low temperature), making them unsuitable for the fusion devices' goals. Figure 1.8 (b) shows a signal for this ELM in JET.
- (iii) **Type III**: they tend to be observed right away after the L-H transition happened; their frequency decreases with increasing heating power, – contrary to **type I** ELMs. Figure 1.8 (c) shows a signal for this ELM in JET.
- (iv) **Type V**: they also lead to significant energy loss; but, form at the same values of collisionality as **type III**. It only differs from the previous type by its repetition rate w.r.t. the increase of the heating power. However, sometimes **type III** and **V** are not distinguished since **type V** are more common in spherical data; sometimes both are presented as **type III/V**.



**Figure 1.8:** Examples of ELMs registered in JET as Balmer  $D_\alpha$  emission, for (a) type I, (b) type II, and (c) type III. Adapted image from: Fig. 1 in [25]

To prevent ELMs from being too harmful, they can be triggered through pellet injection, i.e. deuterium–neon [27], so that small controlled ones are created instead of having one extensive ELM. It is also possible to inject nitrogen or noble gases at the divertor to reduce the heat load in the plasma-facing components. However, there is also the scenario of ELM-free H-mode, the tradeoff of this confinement regime is the impurity accumulation at the plasma's core, and they are not stationary [1], [25].

# II

## INFLUENTIAL POINTS AND DATA ANALYTICS

This chapter aims to derive an effective method that identifies the observations that substantially influence  $\alpha_R$ . Previously, various algorithms, such as gradient descent, Markov chains, and simulated annealing, were implemented by [28]. The results showed subset sizes ranging from 880 to 1700 with  $\alpha_R$  values between 1.1 and 1.2. Nonetheless, I have assessed the impact of influential points and outliers on the new records in STDB5 using conventional regression diagnostics instead of subjecting the dataset to complex optimisation algorithms.

### 2.1 Regression Diagnostics

It is possible to determine the influential points in a regression technique, and even a single regressor, through concepts in regression diagnostics known as DFBETAs, residuals, and leverage [29]. These analyses are explained and evaluated in the standard database in this section.

Consider a dataset  $\mathbf{X}$  in the form of an  $n \times (M + 1)$  matrix<sup>1</sup>, where  $n$  is the total number of observations and  $M$  is the total number of variables used to predict the target variable  $\mathbf{y}$ , an  $n$ -dimensional column-vector. OLS is a regression technique that assumes the target variable,

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad (2.1)$$

is a parametric function dependent on arbitrary parameters  $\beta$ , a  $(M + 1)$ -dimensional column-vector, and Gaussian noise  $\varepsilon$ , an  $n$ -dimensional column-vector often referred to as irreducible error [30]. With OLS, one can predict the target variable  $\hat{\mathbf{y}}$  with the optimised parameters  $\hat{\beta}$  such that  $\hat{\mathbf{y}} - \mathbf{y} = \varepsilon$  is as small as possible. Viewed as a linear algebra problem, one can find the closest vector to  $\mathbf{y}$  using Euclidean distances. The solution to this problem is the orthogonal projection of  $\mathbf{y}$  onto a subspace  $\mathcal{R}^n$  spanned by the columns of  $\mathbf{X}$  [30]. Therefore,

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (2.2)$$

Now, Eq. (2.2) can be used to estimate the regression parameters  $\alpha_R$  when applying the logarithm to Eq. (1.1); namely,

$$\begin{aligned} \log(\tau_{E,th}) = & \log(\alpha_0) + \alpha_I \cdot \log(I_P) + \alpha_B \cdot \log(B_t) + \\ & \alpha_n \cdot \log(\bar{n}_e) + \alpha_P \cdot \log(P_{l,th}) + \alpha_R \cdot \log(R_{geo}) + \\ & \alpha_\kappa \cdot \log(\kappa_a) + \alpha_\epsilon \cdot \log(\epsilon) + \alpha_M \cdot \log(M_{eff}); \end{aligned} \quad (2.3)$$

---

<sup>1</sup>The +1 is to account for the intercept  $\alpha_0$  in Eq. (1.1)

thus,  $\hat{\beta} = [\log(\alpha_0), \alpha_1, \dots, \alpha_M]^T$ . Notice that placing the intercept in an exponential function is necessary when expressing the scaling law as in Eq. (1.1).

Considering this, it is possible to assess the leverage and residual of each point. Leverage is a measure dependent on the mean of a predictor variable and how far an observation is from that mean. In other words, an observation will have high leverage if it is unusually far from the rest of the observations [31]. Mathematically, the leverage  $h_i$  of the  $i$ -th observation is

$$h_i \equiv \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T, \quad (2.4)$$

with  $\mathbf{x}_i$  denoting the  $i$ -th row-vector of the dataset  $\mathbf{X}$  [29]. It is common to state that  $h_i$  is high if  $h_i > (2M + 2)/n$  [31]. The residual is a measure that states how much a prediction is different from the actual target value. There are various ways to estimate this, but here the studentised deleted residual  $r_i$  is utilised. The idea is to obtain  $\hat{\mathbf{y}}_{i(-i)}$ , which denotes the prediction on the  $i$ -th observation based on a regression model estimated without that observation; then,

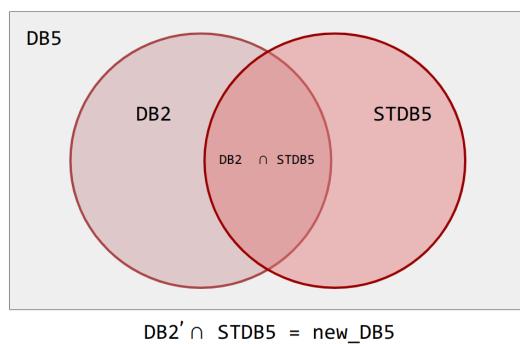
$$r_i \equiv \frac{1}{s_i} (\mathbf{y}_i - \hat{\mathbf{y}}_{i(-i)}), \quad (2.5)$$

where  $s_i^2$  is the estimated variance of  $\mathbf{y}_i - \hat{\mathbf{y}}_{i(-i)}$  [32]. With this, it is common to see that the  $i$ -th observation is said to have a high residual if  $|r_i| > 3$ . However, one might want to be more strict and reduce the threshold to a more appropriate value to what is being observed with the data [31].

One can also get an insight into the influential points over specific variables by computing the DFBETAS. The idea is to look at the change in regression coefficients when the  $i$ -th row is removed and compare the  $j$ -column with the original dataset [29], [33]. The estimate is defined as

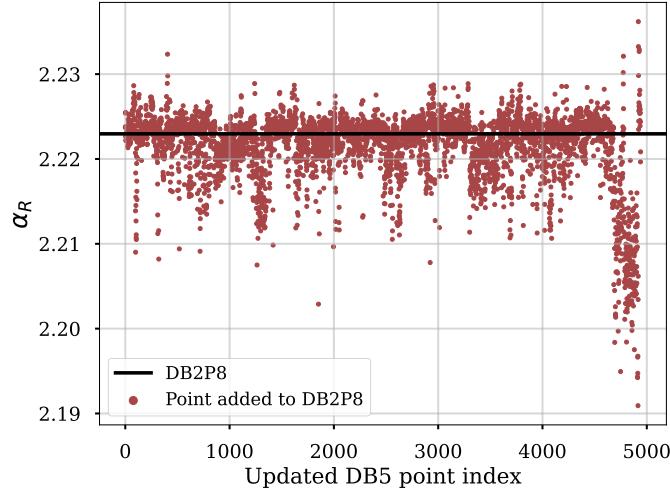
$$\text{DFBETAS}_{ij} \equiv \frac{\beta_j - \beta_{j(-i)}}{s_i \sqrt{\mathbf{X}^T \mathbf{X}_{jj}^{-1}}}; \quad (2.6)$$

and, it is common to consider that the  $i$ -th point is influential on the  $j$ -variable if  $\text{DFBETAS}_{ij} > 2/\sqrt{n}$ ; however, this is not a perfect threshold, and one can adjust it according to the data [33]. Now, it is useful to look at Figure 2.1 showing the Venn Diagram of the databases that will be worked through the project.



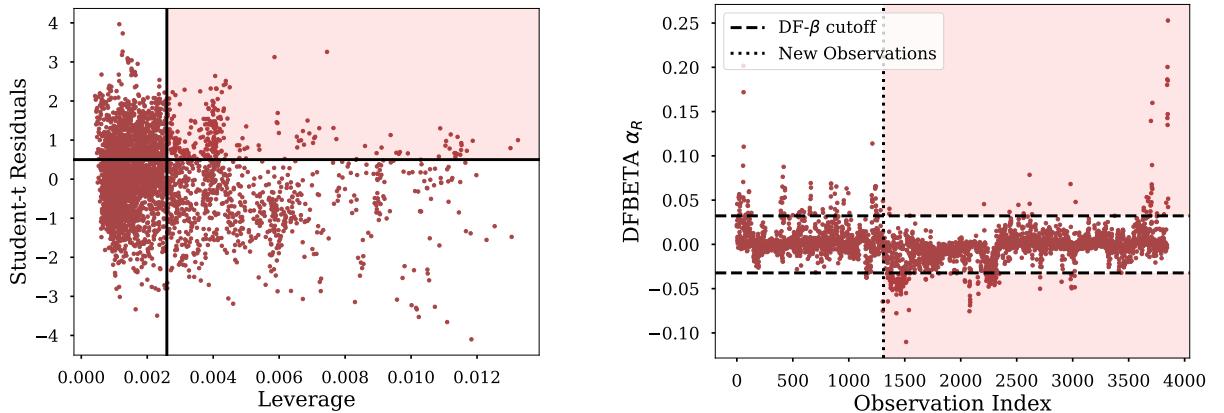
**Figure 2.1:** Venn Diagram showing the databases' naming for an easier interpretation.

To determine which observations will undergo regression diagnostics, one new point at a time was selected from new\_DB5 and added to DB2 to then perform OLS. With this, it is possible to identify the individual points that decreased  $\alpha_R$ . The result is shown in Figure 2.2.



**Figure 2.2:** Variation of  $\alpha_R$  given the individual presence of a point in new\_DB5. The black line shows the value of  $\alpha_R$  when only DB2 is considered. This black line is referred to as *baseline*. The points below the baseline are referred to as *decreased\_DB5*. When taking only the *decreased\_DB5* with DB2,  $\alpha_R = 1.05$ .

Thus, one can focus on the observations below the baseline; I will refer to these points as *decreased\_DB5*. This database is characterised by 2,544 observations from 12 devices: START, NSTX, MAST, JET, D3D, AUG, AUG with ITER-like-walls (AUGW), JET with ITER-like-walls (JETILW), JT60U, TFTR, COMPASS, and CMOD; with shots done from 1989 to 2017. Figure 2.3 shows the results of applying the regression diagnostics to *decreased\_DB5* along with DB2.



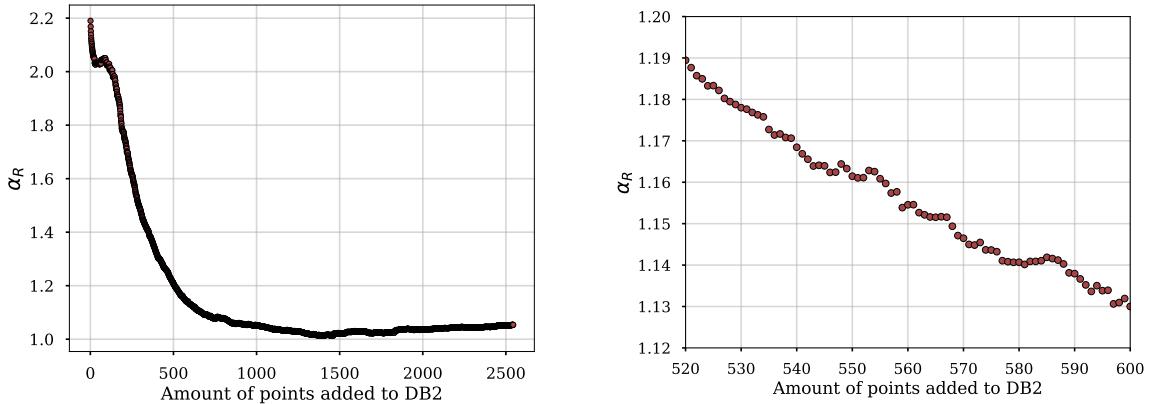
**Figure 2.3:** LEFT: Influential points to all regressors shown in the pink box according to the abovementioned concept. The horizontal and vertical black lines show the cutoff of the Student-t residuals and leverage, respectively. RIGHT: Influential points to  $\alpha_R$  shown in the pink areas according to the DFBETAs. DB2 along with *decreased\_DB5* were studied. See [34] for code.

In this section, it was observed that influential points, as defined earlier, can play a crucial role in determining the observations that increase  $\alpha_R$ ; however, not the ones that decrease it. This statement is illustrated when computing the OLS model on two subsets of data points based on their positions in the leverage-residual plot (Figure 2.3, left): one in the pink area and the other in the white area; the resulting values of  $\alpha_R$  were 1.86 and 1.03, respectively. Similarly, when we analysed the DFBETAs plot (Figure 2.3, right), the values of  $\alpha_R$  for the pink and white areas were 1.46 and 1.15, respectively. The results

suggest that conventional regression diagnostics may not be entirely reliable in identifying the points that significantly impact the reduction of  $\alpha_R$  in the confinement time scaling law, especially when using DFBETAs.

## 2.2 Random Sampling

Another attempted approach, which proved to be the most efficient in achieving the research goal, is random sampling. The idea behind this method is to add the points in `decreasing_DB5` to `DB2` cumulatively without removing the previously added points. This routine allows observing the impact of point grouping on the value of  $\alpha_R$ . Figure 2.4 presents this procedure's result and a close-up view of it.



**Figure 2.4:** LEFT: Change in  $\alpha_R$  due to the cumulative behavior of points in `decreasing_DB5`. RIGHT: close-up of the left figure showing that although the added points individually decrease  $\alpha_R$ , collectively they increase it.

It is possible to observe that some of the points increase  $\alpha_R$  although the overall trending is the reduction of it. From this, one can label the points that increase and decrease  $\alpha_R$  in `decreasing_DB5` and then: (i) do a random sampling for all points in the dataset with a higher tendency in picking the decreasing points, and (ii) do a random sampling only of the decreasing points within `decreasing_DB5` to then subject them to OLS, along with `DB2`. The sampling is performed multiple times with different sampling subsets; each subset contains unique observations, known as bootstrap without replacement [30]. Algorithm 1 depicts the general idea of what was done for both situations.

---

**Algorithm 1** Random sampling to get representative subsets of influential points.

---

**Require:** `seeds` = list with 500 different seeds.

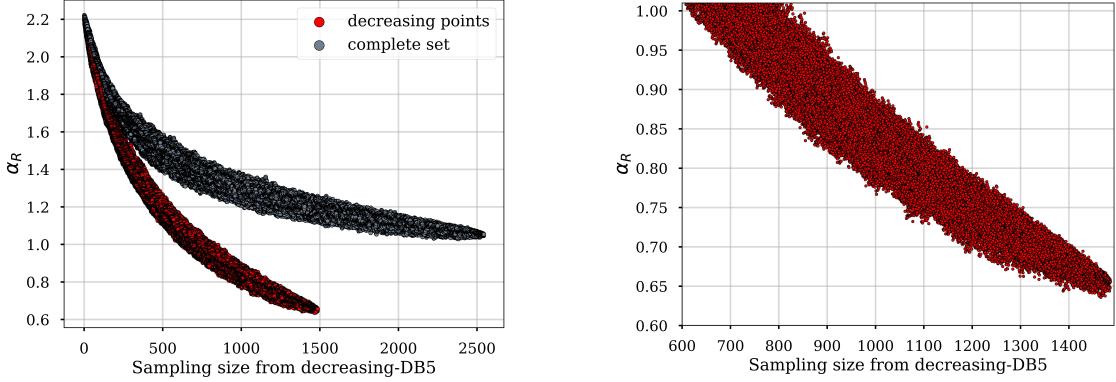
```

for seed in seeds do
    for subset_size in range(1, length(dataset)) do
        subset ← sampling(decreased_DB5, with random_seed = seed)
        data ← subset + DB2
        coefficients ← OLS(data)
        alpha_R ← coefficients[position = 5]
        plot(alpha_R vs subset_size)
    
```

---

This approach generates representative groups that provide different estimates of  $\alpha_R$  based on the subset size. Figure 2.5 shows the resulting groups and a close-up of the points of interest after implementing Algorithm 1.

One can quickly appreciate that the random sampling of the decreasing points in `decreasing_DB5` established smaller values of  $\alpha_R$  with smaller subset sizes than the random sampling on all points in the dataset. Some of the results that might be of interest to study are shown in Table 2.1<sup>2</sup>. From this table, two subsets of interest will be considered from the subsequent analyses: `small_ds` and `big_ds`, which are characterised by having the smallest subset size for a considerably low  $\alpha_R$ , and lowest  $\alpha_R$  but large dataset, respectively.

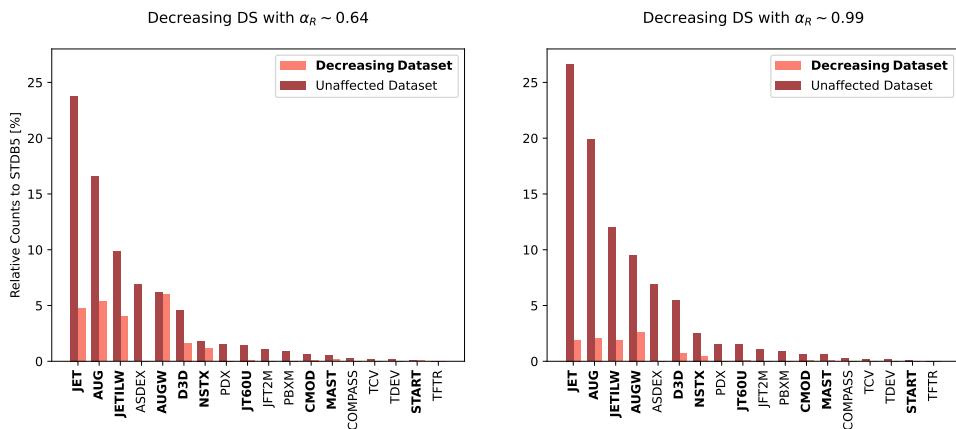


**Figure 2.5:** LEFT: Change of  $\alpha_R$  when sampling different subset-sizes from `decreasing_DB5`. Each of the subsets is added to DB2 to then compute OLS and obtain the respective  $\alpha_R$ . RIGHT: close-up to the points of interest:  $\alpha_R < 1$ .

**Table 2.1:** Characteristic results of random sampling from the decreasing points in `decreasing_DB5`.

	Smallest Subset for $\alpha_R < 1$	Smallest $\alpha_R$ for subset < 1400	Smallest $\alpha_R$ found
$\alpha_R$	0.9998	0.6439	0.6357
Subset size	618	1388	1466
Observations decreasing $\alpha_R$	9.88%	22.20%	23.45%
Dataset Name	<code>small_ds</code>	-	<code>big_ds</code>

Figure 2.6 shows the tokamaks present in `small_ds` (right) and the `big_ds` (left). Notice how, in both situations, spherical tokamaks play a role in influencing  $\alpha_R$ . It is also interesting how ASDEX with tungsten walls (AUGW) has a relative amount of observations decreasing and not affecting  $\alpha_R$  in `big_ds`, while the opposite is observed in `small_ds`.



**Figure 2.6:** LEFT: Counts of tokamaks influencing  $\alpha_R$  based on `big_ds`. RIGHT: Counts of tokamaks influencing  $\alpha_R$  based on `small_ds`.

<sup>2</sup>Although the observations in DB2 were added to estimate their respective  $\alpha_R$ , the resultant named dataset only contains the new observations in STDB, none of them belong to DB2.

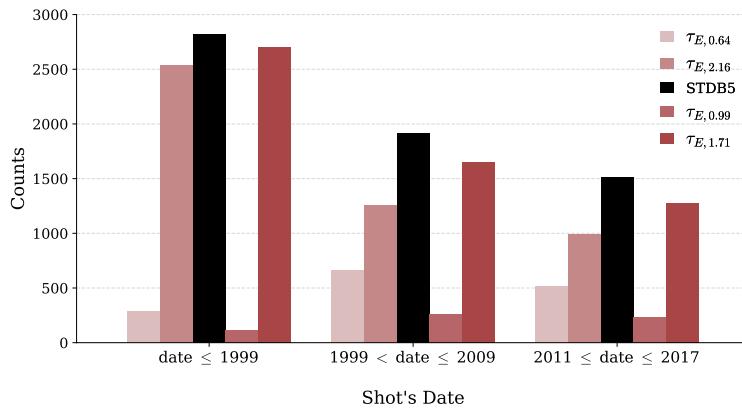
The question is: out of all the dataset subsets shown in Figure 2.5 (right), which one will provide the most informative insights into the factors causing the decrease in  $\alpha_R$ ? If I were to consider all of them, I would need to study a total of 382,495 subsets. However, to streamline the analysis, I will focus initially on the smallest value found of  $\alpha_R$  and the smallest subset size for which  $\alpha_R < 1$  since this research aims to discern which columns can yield information into what is decreasing  $\alpha_R$ . Table 2.2 shows the corresponding parameters for estimating  $\tau_{E,th}$  based on Eq. (2.3) when utilising the obtained datasets, with the addition of DB2, and when removing them from STDB5.

**Table 2.2:** Change of parameters in the energy confinement time scaling law based on the obtained datasets. All parameters were estimated with observations in DB2 for a fairer comparison.

	Scaling	$\alpha_0$	$I_p$	$B_t$	$\bar{n}_e$	$P_{l,th}$	$R_{geo}$	$\kappa_a$	$\epsilon$	$M_{eff}$
From big_ds	$\tau_{E,0.64}$	0.10	1.42	0.02	-0.14	-0.53	0.64	-0.002	-0.56	0.13
	$\tau_{E,2.16}$	0.06	0.78	0.24	0.41	-0.75	2.16	0.5	0.79	0.22
From small_ds	$\tau_{E,0.99}$	0.10	1.24	0.11	-0.02	-0.57	0.99	0.14	-0.19	0.12
	$\tau_{E,1.71}$	0.08	1.01	0.12	0.27	-0.71	1.71	0.34	0.32	0.23

Notice that the subset leading to the smaller  $\alpha_R$  diminishes the influence of elongation. Further analyses regarding the influence of parameters on the model within each subset will be conducted in the subsequent sections.

To finish comparing these two subsets, Figure 2.7 displays the number of shots per dataset according to their execution dates. It is evident from the graph that a significant majority of the observations influencing  $\alpha_R$  were collected after the 2000s. Moreover, there is no available data from the year 2010; this is not only for STDB5 but also for STDB5.



**Figure 2.7:** Bar plot showing the dates for which the registers were made within each dataset.

## 2.3 Multicollinearity

One encounters a multicollinearity issue when a regression algorithm, such as OLS, struggles to distinguish the effects of one variable from another on the target variable [35]. This issue arises due to significant inter-correlations among the predictor variables [36]. Strong multicollinearity has two main effects: an increase in the standard errors of each coefficient in the model and numerical instability [29], [35]. These consequences might lead to unstable parameter estimation, unreliable models, and weak predictive ability [37]. In the case of extreme multicollinearity, small changes in the data will produce wide swings

in parameter estimates, but for low or moderate cases, it does not necessarily imply a severe issue [37], [38]. However, evaluating the extent of multicollinearity in a dataset will determine whether the model effectively explains how the target variable (in our case,  $\tau_{E,th}$ ) is influenced by each variable ( $I_p$ ,  $B_t$ ,  $\bar{n}_e$ ,  $P_{l,th}$ ,  $R_{geo}$ ,  $\kappa_a$ ,  $\epsilon$ , and  $M_{eff}$ ) [35]. To determine the presence of multicollinearity in the diverse available datasets, I will consider the variance inflation factor (VIF), condition index, and variance decomposition.

### Variance Inflation Factor (VIF)

This widely used method measures the degree of collinearity in an independent variable of a linear regressor by indicating how much the estimated variance of the independent variable has increased w.r.t. the case where no multicollinearity exists [38]. The method consists in computing the coefficient of determination  $R_k^2$  of the  $k$ -th parameter when regressing it against the remaining explanatory variables [29], [37]. Generally speaking the coefficient of determination  $R^2$  is computed as [29]

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2.7)$$

where  $\bar{y}$  is the mean value of the target vector; this expression i.e. belongs to Eq.(2.1) model. All the other symbols remain with the same interpretation as discussed above. The VIF associated to the  $k$ -th regressor parameter is therefore [37]

$$VIF_k = \frac{1}{1 - R_k^2}. \quad (2.8)$$

In the ideal case,  $VIF_k = 1$ , indicating no collinearity between the  $k$ -th regressor parameter and the other variables in the model. When  $VIF_k \geq 10$ ; then, one might have a multicollinearity issue [37]. It is possible to observe how much a parameter's variance has increased due to VIF through the following equation

$$Var[\beta_k] = \frac{(1 - R^2) \cdot \sum (y_i - \bar{y})^2}{n - M - 1} \cdot \frac{1}{\sum x_k^2} \cdot VIF_k, \quad (2.9)$$

where  $\sum x_k^2$  denotes the sum of the squared deviations for the  $k$ -th independent variable around its mean [38].

### Condition Index

To compute the condition index of all regressor parameters in  $\mathbf{X}$ , it is necessary to (i) scale  $\mathbf{X}$  to have equal (unit) column length and (ii) perform the singular-value decomposition to the scaled (non-centred) matrix [29]. The decomposition is performed as  $\mathbf{X} = \mathbf{UDV}^T$ ; where<sup>3</sup>  $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{I}_M$ , and  $\mathbf{D}$  contains the non-negative *singular values, or principal components*,  $\mu$  and it is diagonal [29]. The obtained singular values are the square roots of the eigenvalues of the covariance matrix; thus, they provide information on the amount of variance explained by each principal component in the data [29]. Furthermore, they are ordered based on the amount of variance they explain. When a high condition number (not to be confused with the condition index) is observed, variables associated with higher principal components

---

<sup>3</sup>The shape the elements are:  $\mathbf{X}^{n \times M} = \mathbf{U}^{n \times n} \mathbf{D}^{n \times M} \mathbf{V}^{M \times M}{}^T$ . In this context,  $M$  also accounts for the intercept; hence,  $M = 9$  when assessing the energy confinement scaling law.

are considered of interest when explaining multicollinearity patterns within that dataset [29], [39]. The condition number of the matrix  $\mathbf{X}$  is computed as

$$c(\mathbf{X}) = \frac{\max(\mu)}{\min(\mu)} \geq 1. \quad (2.10)$$

The condition index  $\eta_k$  of the  $k$ -regressor parameter is obtained as [29]

$$\eta_k = \frac{\max(\mu)}{\mu_k} \quad k = 1, \dots, M. \quad (2.11)$$

Notice that the largest condition index is the condition number of the matrix  $\mathbf{X}$ . These two values are of great interest when assessing multicollinearity in a dataset because if  $c(\mathbf{X}) \sim 5$  or 10, a dataset might have weak dependencies. A dataset might have moderate to strong dependencies if  $c(\mathbf{X}) \sim 30$  or 100. Furthermore, one can also think there are as many near dependencies as high condition indexes [29].

### Variance Decomposition and Associated $\Pi$ Matrix

Over the presence of high condition indices, it is recommended to perform the variance decomposition over the variance-covariance matrix  $\mathbf{V}(\hat{\beta})$  of the least-squares estimator  $\hat{\beta}$  to determine the variables involved in multicollinearity [40]. This is done by decomposing  $\mathbf{V}(\hat{\beta})$  as [29], [39]

$$\mathbf{V}(\hat{\beta}) = \sigma^2 \left[ (\mathbf{U} \mathbf{D} \mathbf{V}^T)^T (\mathbf{U} \mathbf{D} \mathbf{V}^T) \right]^{-1} = \sigma^2 \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T. \quad (2.12)$$

For the  $k$ -th component of  $\hat{\beta}$ ,

$$Var[\beta_k] = \sigma^2 \sum_j \frac{V_{kj}^2}{\mu_j^2}. \quad (2.13)$$

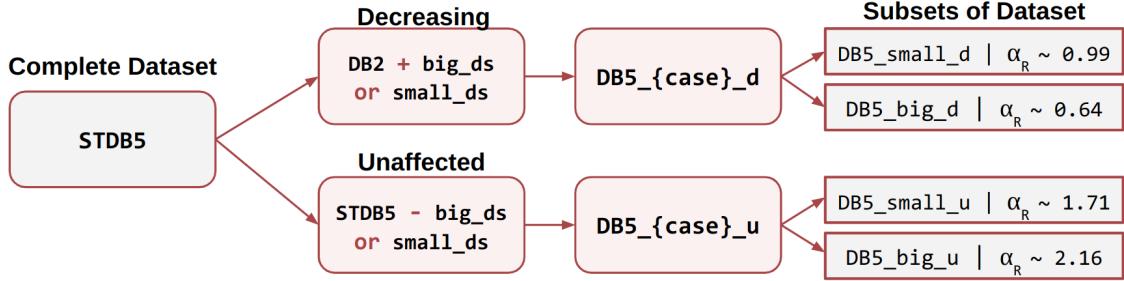
Here,  $\sigma^2$  is the common variance of the components of  $\varepsilon$  in Eq. (2.1),  $\mu_j$ 's are the singular values, and  $V_{kj}$  is the  $(k, j)$  element in  $\mathbf{V}$  [29]. Eq. (2.13) allows identifying unusually high proportions of variance between two or more coefficients by having the singular values in the denominator. From this reasoning, it follows the idea of defining the  $k, j$ -th *variance-decomposition proportion*  $\Pi_{jk}$  as the proportion of variance corresponding to the  $k$ -th regressor, related by the  $j$ -th component of its decomposition in Eq. (2.13) [29]. This influence is observed through the  $\Pi$  matrix, which is calculated as

$$\Pi_{jk} = \frac{\phi_{kj}}{\sum_{j=1}^M \phi_{kj}}, \quad (2.14)$$

where  $\phi_{kj} = V_{kj}^2 / \mu_j^2$ . The ideal case, where no multicollinearity exists in the dataset, would behave as  $\Pi_{ij} = \delta_{ij}$ , a Kronecker delta. In other words, the  $Var[\beta_k]$  would not be affected by near dependency among the columns of  $\mathbf{X}$  [29]. The degradation of a regression estimate due to multicollinearity is evinced when a singular value  $\mu_j$  associates with two or more coefficients with considerable proportions. The regressors that isolate a singular value indicate that they do not have a collinear dependence concerning others [29].

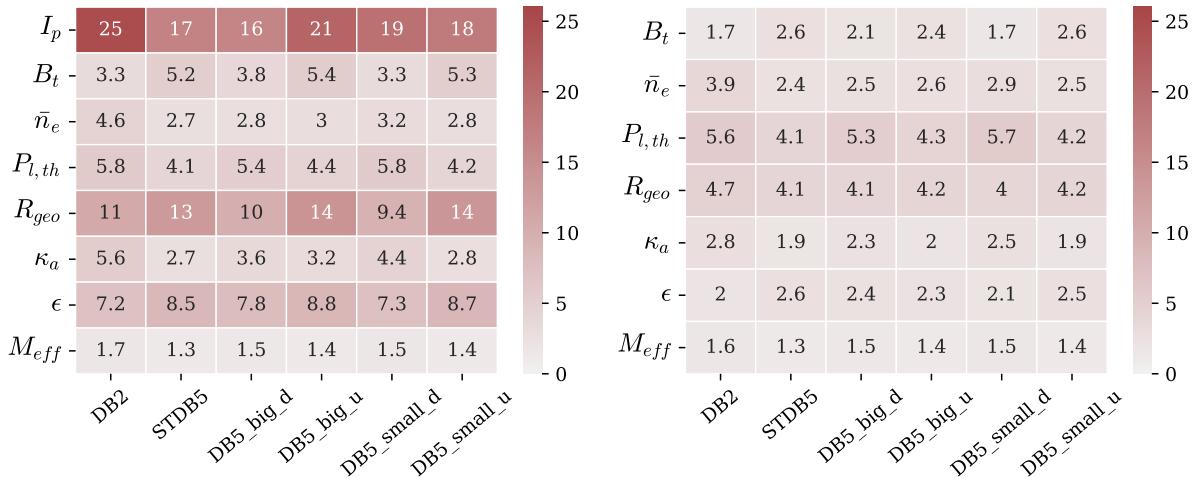
### 2.3.1 Assessing Dependencies in Datasets

In this subsection, I will present the results of conducting the aforementioned multicollinearity tests on four datasets: DB2, STDB5, `small_ds`, and `big_ds`. For a comprehensive analysis of `small_ds` and `big_ds`, these datasets have been divided into two subsets: decreasing and unaffected. Each subset contains the respective observations for its specific case. The corresponding labels, and visual representation, for each case are displayed in Figure 2.8.



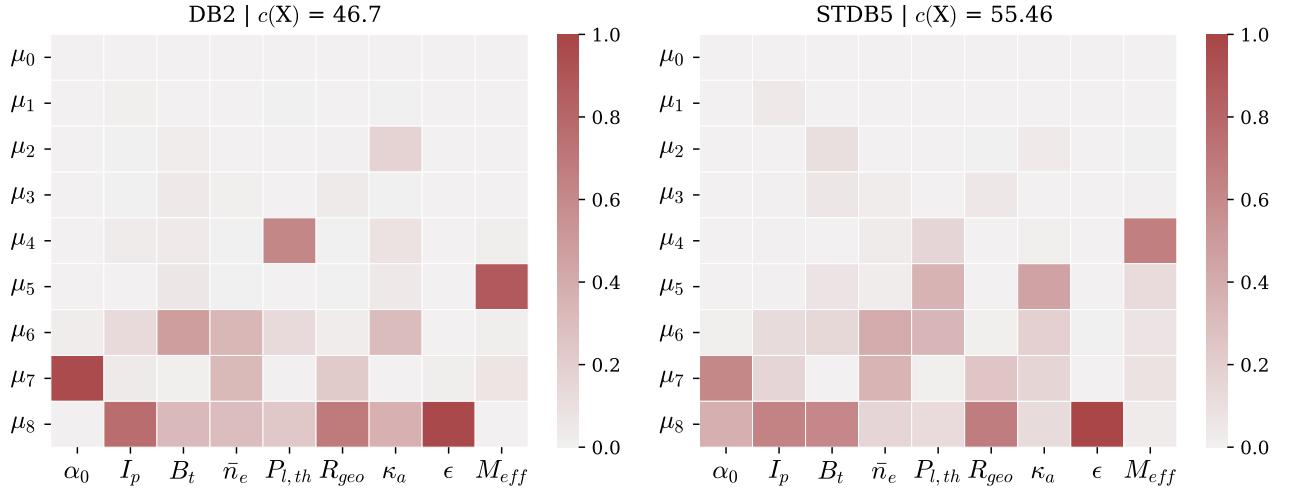
**Figure 2.8:** Diagram showing the split of `small_ds` and `big_ds` into creating two datasets containing the observations that decrease or do not affect  $\alpha_R$ . Notice that resultant subsets are the same that were used to estimate the parameters shown in Table 2.2.

The obtained VIF values for all regressors, categorised by dataset, are presented in Figure 2.9 (left). According to this metric, it is evident that in all datasets, the plasma current gets the most variance inflation due to multicollinearity, followed by the major radius. When the plasma current is excluded from the dataset, all variables exhibit notoriously lower VIF values. However, the variable with the highest VIF value shifts from the major radius to the thermal power loss; see Figure 2.9 (right). Nevertheless, this does not imply that the plasma current should be removed from the model; this is to demonstrate how variance inflation of other variables is affected by the presence of a specific independent variable.

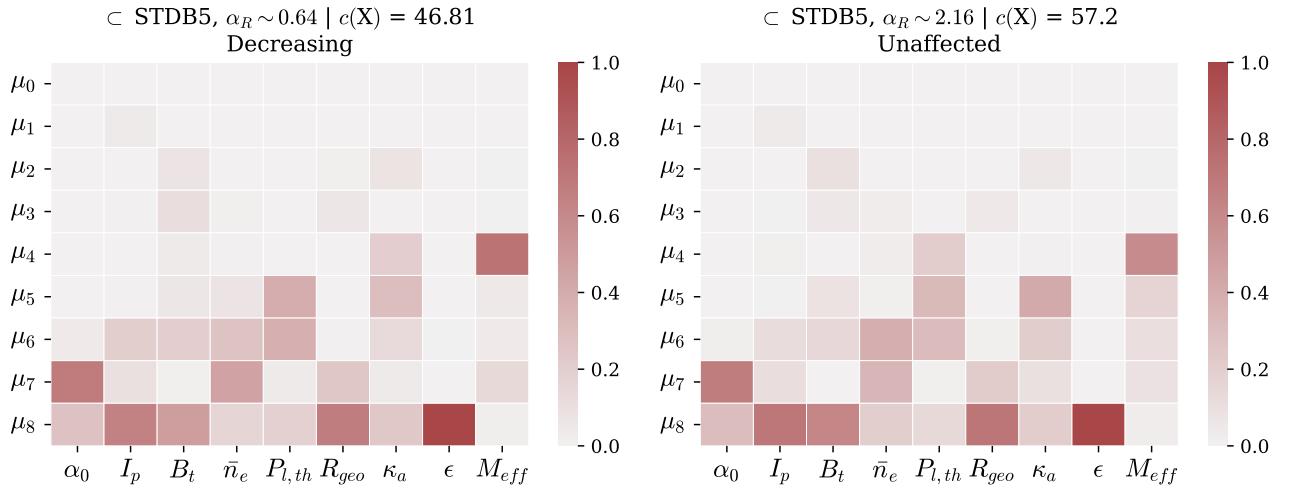


**Figure 2.9:** LEFT: VIF scores of all regressors. RIGHT: VIF scores of regressors after removing the plasma current. It is considered that a high  $VIF \geq 10$ .

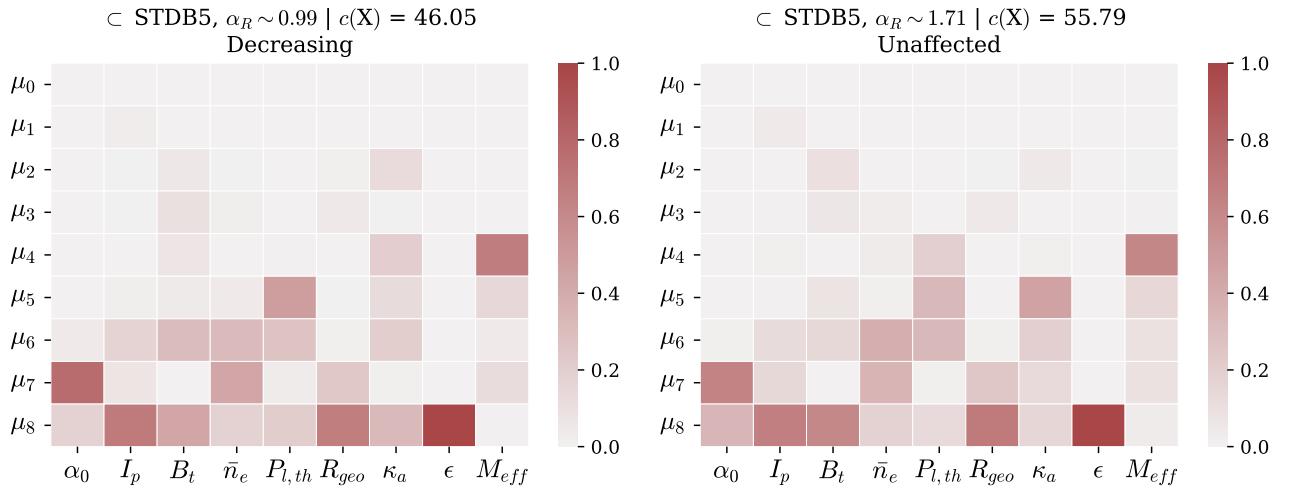
The condition number and the  $\Pi$  matrix was also computed for all datasets, and all cases showed a high condition number, meaning that they are likely to present multicollinearity issues. Moreover, all datasets showed three condition indexes ranging [10, 20], one condition index ranging [25, 35], and one (also their condition number) ranging [46, 58].



**Figure 2.10:**  $\Pi$  matrix for DB2 (left) and STDB5 (right) with their corresponding condition number  $c(\mathbf{X})$ , for all variances associated to the regressor variables in Eq. (2.1).



**Figure 2.11:**  $\Pi$  matrix for observations that decrease  $\alpha_R$  (left) and does not affect  $\alpha_R$  (right), based on `big_ds` with their corresponding condition number  $c(\mathbf{X})$ , for all variances associated to the regressor variables in Eq. (2.1).



**Figure 2.12:**  $\Pi$  matrix for observations that decrease  $\alpha_R$  (left) and does not affect  $\alpha_R$  (right), based on `small_ds` with their corresponding condition number  $c(\mathbf{X})$ , for all variances associated to the regressor variables in Eq. (2.1).

Figure 2.1 shows the results implemented to DB2 and STDB5. In both datasets, the effective mass showed an isolated principal component, meaning that this variable is not collinear with the others. Almost the same behaviour is observed for the intercept, except that in STDB5, its variance is almost split in half by two principal components. All other variances share various principal components. However, the singular value  $\mu_8$  accounts for more than 60% of  $Var(I_p)$ ,  $Var(B_t)$  and  $Var(\epsilon)$ , with the latter being 99% in both DB2 and STDB5. Figure 2.11 show the results for the two subsets resulting from `big_ds`. In this subset, the variance of the effective mass is isolated to mostly one principal component. A similar pattern is observed for  $\mu_8$ , where it explains a significant portion of the variance in  $I_p$ ,  $B_t$ ,  $\bar{n}_e$ ,  $P_{l,th}$ ,  $R_{geo}$ ,  $\kappa_a$ , and primarily  $\epsilon$ . Notably, in the unaffected dataset compared to the decreasing dataset, the influence of  $\mu_8$  increases. Figure 2.12 presents the results of the analyses conducted on `small_ds`. A similar pattern is observed, although the increase in influence for  $\mu_8$  is not as significant as in the previous case. However, it is worth noting that the percentage of variance explained by  $\kappa_a$  drops from approximately 30% in the decreasing dataset to around 15% in the unaffected dataset.

Overall, similar patterns and behaviours are observed across all datasets, including the total number of high  $\eta_k$  present in the data. It is worth mentioning that the inclusion of spherical tokamaks in the dataset has a slight mitigating effect on multicollinearity within STDB5. Removing these devices from the dataset results in  $c(\mathbf{X}) = 57.93$ .

## 2.4 Model Comparison

Multivariate models are updated when more data is available. One can use statistical tests to compare models and assess the performance of the multivariate model given the new data, removal or addition of variables [41]. In this section, I will use the F-test to compare nested models; these are models where one is a particular case of the other, for instance, a reduced model from a full model [42]. Here, our entire model is described by Eq. (2.3), and the reduced model is the removal of at least one of its variables. To compute the F-statistic, one needs the sum of square error (SSE), being

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (2.15)$$

one can then compute its value for the specific case to obtain

$$\Delta SSE = SSE_r - SSE_f. \quad (2.16)$$

Here, the subscripts  $r, f$  denote the reduced and the full model, respectively. The mean square error (MSE) is the SSE divided by the number of observations  $n$  in the dataset. The F-statistic, or F-value, is then

$$F = \frac{1}{\Delta M} \cdot \frac{\Delta SSE}{MSE_f}, \quad (2.17)$$

where  $\Delta M$  is the number of coefficients being tested, also referred to as the numerator degrees of freedom  $\nu_1$ . The denominator degrees of freedom is  $\nu_2 = n - M$ . These two are important when computing the P-value [42]. The P-value, denoted as  $Pr(F_{\nu_1, \nu_2} \geq F)$ , represents the probability of observing the obtained

F-value based on the F-distribution with degrees of freedom  $\nu_1$ , and  $\nu_2$ . Usually, it is considered that having  $F \geq 2.5$ , or  $P \leq 0.05$ , is enough to consider that the reduced model is statistically significantly different than the full model and that it performs better. This type of analysis is also referred to as ANOVA (analysis of variance) test [43]. Table 2.3 show the F- and P-values for DB2, and STDB5. Table 2.4 show the same results for the two subsets that follow from `small_ds`, and `big_ds`, as explained in Figure 2.8.

The results indicate that, as anticipated, no variable should be eliminated when using the DB2 dataset. However, for STDB5, removing the inverse aspect ratio may be acceptable without significantly degrading the model's performance. For the other subsets of STDB5, it is observed that removing the elongation and toroidal magnetic field may not harm the model when utilizing STDB5\_big\_d. In the case of STDB5\_small\_d, only the averaged-line electron density should be considered for removal. For the remaining unaffected datasets, all variables should be retained. Furthermore, when the elongation is removed from the subset with  $\alpha_R \sim 0.64$ , the  $Pr(>F)$  value increases from 0.32 to 0.42. When both  $\kappa_a$  and  $B_t$  are removed from the same subset, all the other variables maintain a  $Pr(>F)$  value of 0. Interestingly, none of the subsets suggests that removing the major radius may not degrade the model.

**Table 2.3:** ANOVA test applied to DB2, and STDB5.

		DB2		STDB5	
Removed		F	Pr(>F)	F	Pr(>F)
$I_p$		961.58	0.0	4809.57	0.00
$B_t$		138.99	0.0	25.79	0.00
$\bar{n}_e$		494.28	0.0	332.97	0.00
$P_{l,th}$		2674.97	0.0	10024.60	0.00
$R_{geo}$		2022.43	0.0	3221.36	0.00
$\kappa_a$		94.67	0.0	94.70	0.00
$\epsilon$		131.14	0.0	2.96	0.09
$M_{eff}$		36.51	0.0	157.27	0.00

**Table 2.4:** ANOVA test applied to subsets of STDB5 that result in cases where observations decrease or do not affect  $\alpha_R$  w.r.t.  $\alpha_R$  in DB2.

Removed	Decreasing $\alpha_R \sim 0.64$		Unaffected $\alpha_R \sim 2.16$		Decreasing $\alpha_R \sim 0.99$		Unaffected $\alpha_R \sim 1.71$	
	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)
$I_p$	4165.11	0.00	2370.15	0.0	2149.90	0.000	3757.01	0.0
$B_t$	1.011	0.32	300.46	0.0	18.89	0.000	69.15	0.0
$\bar{n}_e$	85.69	0.00	1971.23	0.0	1.23	0.269	752.88	0.0
$P_{l,th}$	2017.40	0.00	14937.77	0.0	1609.65	0.000	11206.25	0.0
$R_{geo}$	294.38	0.00	7450.69	0.0	471.16	0.000	4404.12	0.0
$\kappa_a$	0.002	0.97	350.36	0.0	10.45	0.001	144.63	0.0
$\epsilon$	136.93	0.00	507.59	0.0	11.20	0.001	77.07	0.0
$M_{eff}$	25.50	0.00	182.73	0.0	15.86	0.000	179.70	0.0

## 2.5 Conclusions

In this chapter, I have demonstrated how bootstrapping sampling can find representative subsets that decrease  $\alpha_R$  without complex optimization algorithms. Conventional regression diagnostics were also implemented but failed to discern the observations that reduced  $\alpha_R$ . The smallest  $\alpha_R$  found through sampling was 0.6357, with a subset size 1466. It is worth noting that the order of points being added to STDB5 from decreasing\_DB5 to create Figure 2.4 (left) does not influence the overall result, but it does change the shape of the plot. Furthermore, I have examined the presence of multicollinearity among different subsets and investigated whether it is possible to remove a variable from the model presented in Eq. (2.3) without detrimental effects, given the updated observations. The analyses revealed that  $R_{geo}$  plays a crucial role in estimating  $\tau_{E,th}$ , and it was found that multicollinearity may not be solely responsible for the decrease in  $\alpha_R$  when utilizing STDB5; this is evident from the fact that the division of the dataset into decreasing and unaffected subsets did not exhibit a significant reduction in multicollinearity; this absence of variability within the split datasets emphasizes the need to explore potential physical factors contributing to the decrease in  $\alpha_R$ .



# III

## MACHINE LEARNING ALGORITHMS

Machine learning refers to computer programs that learn and make decisions or predictions based on extracting useful information from data. What turns data analysis into machine learning is that the algorithm *learns* from the data in an automated manner [30]. Learning algorithms can guide research when mathematical expressions are unavailable or data is limited. For instance, in drug discovery, these algorithms identify molecules with specific properties, such as binding affinity to a target protein, even when the underlying mechanism of action is not well understood [44]. Astronomy [45], and neuroscience [46] are other fields where machine learning is implemented to recognize patterns and correlations in data, leading to new areas of research or refinement of existing theories. In this chapter, I will explain how two supervised machine learning algorithms can be used to investigate the causes that lead to a decrease in  $\alpha_R$ , based on the subsets obtained in the previous chapter.

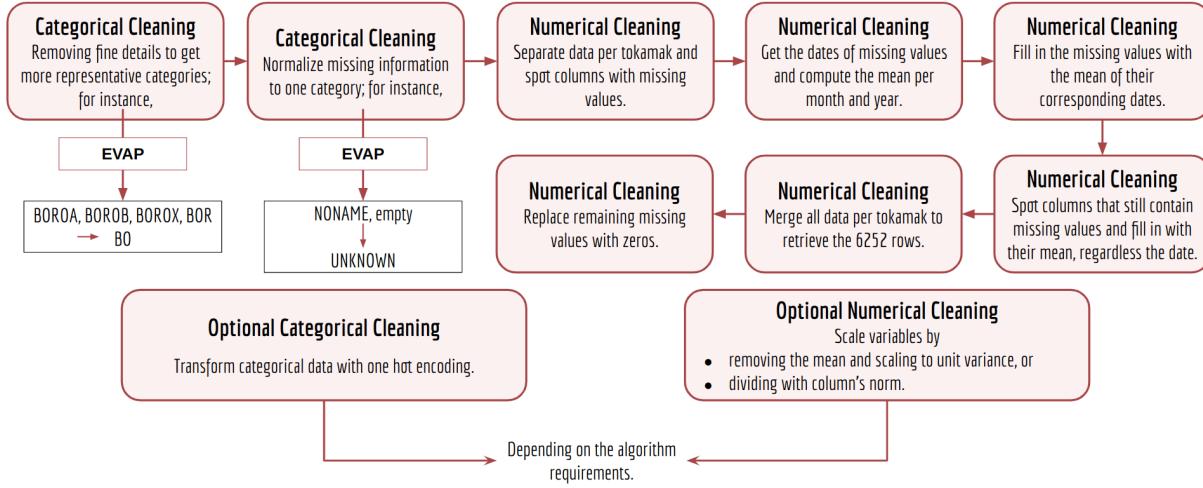
### 3.1 Feature Selection

One of the challenges when implementing any machine learning algorithm is feature selection, the process of deciding which columns of a dataset should be used so that the algorithm can learn better and faster from them. Moreover, identifying and extracting the right features becomes more complex as the number of columns in the dataset increases since adding or removing them could reduce the algorithm's learning quality [30]. To address this issue, I have (i) identified which of all the variables available in DB5 can help understand why there is a reduction in  $\alpha_R$ , (ii) utilised the concept of entropy in information theory to identify which of these variables contain more information, (iii) considered relevant research to decide on a set of variables that will yield more insight into the reduction of  $\alpha_R$ , and (iv) studied the different resultant subsets for multicollinearity.

#### 3.1.1 All Variables of Interest

DB5 has 192 columns containing information on the tokamaks, including divertor and wall materials, plasma characteristics, heating mechanisms, instabilities, impurities, power, and more [12]. However, since I am using the standard version, not all columns have complete values, and not all are relevant. Tables 3.1 to 3.7 show the description [12] of 47 variables that I have considered to be interesting for this project, based on the knowledge I have gained through this master programme. In all tables, **bold features** represent the variables provided by the research group, not found in [12], and underlined features are categorical data.

To clean STDB5, I have applied the pipeline shown in Figure 3.1. The cleaned dataset, which includes the 47 variables of interest and all rows from STDB5, will be referred to as `clean_DB5`. In subsequent references to `small_ds` or `big_ds` (see Table 2.1), their cleaned versions will be used.



**Figure 3.1:** Pipeline showing the criteria to clean STDB5. The one-hot encoding creates a binary vector, for each category in the column, with 1 if the observation has that observation and 0 otherwise [47].

**Table 3.1:** Variables of interest regarding plasma characteristics.

Feature	Description	Feature	Description
BEIMHD	Beta Shafranov	MEFF	Effective atomic mass
<b>BETASTAR</b>	Plasma pressure normalised to $B_t$	<b>NUSTAR</b>	normalised Ion collision frequency
<b>CONFIG</b>	Plasma configuration	<b>PREMAG</b>	Flag for startup: with or without pre-magnetization current
DWDIA	Time rate of change of the total plasma stored energy	Q95	Plasma safety factor at the 95% poloidal flux surface
EPS	Inverse aspect ratio	<b>RHOSTAR</b>	normalised Ion Gyroradius
IP	Plasma current	TORQ	Torque on plasma due to NBI
KAREA	Plasma elongation	VOL	Plasma volume
<b>LCOULOMB</b>	Coulomb Logarithm $\log_e(\Lambda) = 30.9 - \log_e(\bar{n}_e^{0.5}/\hat{T})$	WMHD	Total plasma energy (MHD equilibrium)

**Table 3.2:** Variables of interest regarding heating.

Feature	Description	Feature	Description
AUXHEAT	Type of auxiliary heating	<b>PELLET</b>	Pellet material if injected
ECHMODE	Mode of ECRH waves	<b>PICRH</b>	ICRH power absorbed by the plasma
ENBI	Neutral beam energy weighted by power	<b>PNBI</b>	Total NBI power minus shine through
<b>ICScheme</b>	ICRH heating scheme	<b>POHM</b>	Total Ohmic Power
PECRH	ECRH power absorbed by the plasma	<b>PRAD</b>	Total radiated power as measured by Bolometer

**Table 3.3:** Variables of interest regarding impurities.

Feature	Description
<u>EVAP</u>	Evaporated material to cover the vessel
ZEFF	Line average plasma effective charge, from Bremsstrahlung
ZEFFNEO	Plasma effective charge, from neo-classical resistivity

**Table 3.4:** Variables of interest regarding tokamaks' characteristics.

Feature	Description	Feature	Description
AMIN	Horizontal plasma minor radius	<u>LIMMAT</u>	Limiters' material
BT	Vacuum toroidal magnetic field at $R_{geo}$	<u>TOK</u>	Tokamak's name
DIVMAT	Material of divertor tiles	<u>WALMAT</u>	Walls' material

**Table 3.5:** Variables of interest regarding power loss and ELMs.

Feature	Description	Feature	Description
BEIMHD	NBI power that is lost from the plasma through charge exchange and unconfined orbits	<u>ELMTYPE</u>	Type of ELMs
PLTH	Estimated loss power corrected for charge exchange and unconfined orbit losses	<u>ELMFREQ</u>	ELM frequency

**Table 3.6:** Variables of interest regarding temperature.

Feature	Description
TAV	Total volume average temperature
TEV	Total volume averaged electron temperature
TIV	Total volume averaged ion temperature

**Table 3.7:** Variables of interest regarding particles.

Feature	Description	Feature	Description
NEL	Central line average electron density	<u>OMEGACYCL</u>	Ion Cyclotron Frequency $\omega_i = qB_t/M_{eff}$
NESOL	Electron density in scrape-off layer	<u>WFFORM</u>	Total fast ion energy due to NBI
WFICFORM	Total fast ion energy due to ICRH estimated from approximate formula		

While all these variables may contain valuable information, it does not necessarily mean that an algorithm will optimally perform when utilising all of them. There are instances where it is desirable to reduce the number of features and only include the most representative ones. This process is often referred to as feature engineering. Additionally, it has been observed that as the number of features in a model increases, the number of required observations exponentially grows in order to achieve satisfactory performance. This phenomenon is commonly known as the "curse of dimensionality" [48]. Therefore, it is crucial to identify a subset of the relevant features that can reduce the model's complexity and improve its overall performance. The attempts to identify a subset are explained in the subsequent subsections.

### 3.1.2 Entropy Variables

By considering features as random variables, it is possible to use the concept of entropy in information theory to estimate the amount of order or information in a feature [49]. The entropy of a random variable  $Z$  is defined as [50]:

$$E(Z) = - \sum_z p(Z = z) \cdot \log(p(Z = z)). \quad (3.1)$$

Here,  $p(Z = z)$  represents the prior probability of  $z$ . The units of entropy depend on the logarithm being used, if it is  $\log_2$  then the units are bits, which represent the number of bits<sup>1</sup> required to characterize the random event [52]. However, there is no specific expression for the probabilities of all the features of interest, it is possible to approximate Eq. (3.1) using the similarity of observation, which directly depends on the distance between them [49]. To estimate the similarity between two observations  $S_{ij}$  for numerical data, the Euclidean distance is implemented as

$$S_{ij} = \exp(-\gamma \cdot D_{ij}), \quad D_{ij} = \left[ \sum_{k=1}^M \left( \frac{x_{ik} - x_{jk}}{\max(F_k) - \min(F_k)} \right)^2 \right]^{1/2}, \quad (3.2)$$

where  $D_{ij}$  is a  $n \times n$  matrix. Here,  $F_k$  represents the column vector belonging to the  $k$ -th feature, out of the total  $M$  features. And,  $x_{ik}$  (or  $x_{jk}$ ) is the  $i$ -th (or  $j$ -th) row, and the  $k$ -th column in  $\mathbf{X}$ . Finally,  $\gamma$  is a parameter which can be tuned according to the problem at hand. It is common to set  $\gamma = 0.5$  [49]. For categorical features, the similarity is computed through the Hamming distance<sup>2</sup> as [49]:

$$S_{ij} = \frac{1}{M} \sum_{k=1}^M \delta_{ij}(x^k); \quad \text{with } \delta_{ij}(x^k) = \begin{cases} 1, & \text{if } x_i^k = x_j^k \\ 0, & \text{if } x_i^k \neq x_j^k \end{cases}. \quad (3.3)$$

From this, it is possible to evaluate the entropy of a dataset<sup>3</sup> as [49]:

$$E = - \sum_{\substack{i,j=1 \\ i \neq j}}^N \left[ S_{ij} \log(S_{ij}) + (1 - S_{ij}) \log(1 - S_{ij}) \right]. \quad (3.4)$$

To estimate the importance of a feature's presence in a dataset in terms of information, one can compute the entropy of the dataset with the missing  $k$ -th feature  $E_{-F_k}$  and compare this instance to another one missing a different feature. For example, if  $E_{-F_1} > E_{-F_2}$ , one can say that feature 1 is more important than feature 2, as the former imposes more structure in the database [49]. The idea is shown in Algorithm 2 to rank the feature importance of `small_ds` and `big_ds` was implemented.

---

**Algorithm 2** Ranking features' importance based on entropy.

---

**Require:** `entropy_features` = empty one-dimensional array of size  $M$

```

for feature in columns(clean_dataset) do
    data  $\leftarrow$  clean_dataset without feature
    entropy_k  $\leftarrow$  get_entropy(data)
    entropy_features.append(entropy_k)
ordered_features  $\leftarrow$  sort_values(entropy_features)

```

---

<sup>1</sup>If  $\log_e$ , units are nats. If  $\log_{10}$ , ban. And, if  $10\log_{10}$ , deciban (db) [51].

<sup>2</sup>In case the reader is interested, one can simply compute  $\delta_{ij}(x^k)$  in Python with:

`np.frompyfunc(lambda x,y: x==y, 2, 1).reduce(np.array(np.meshgrid(F_k, F_k)))`

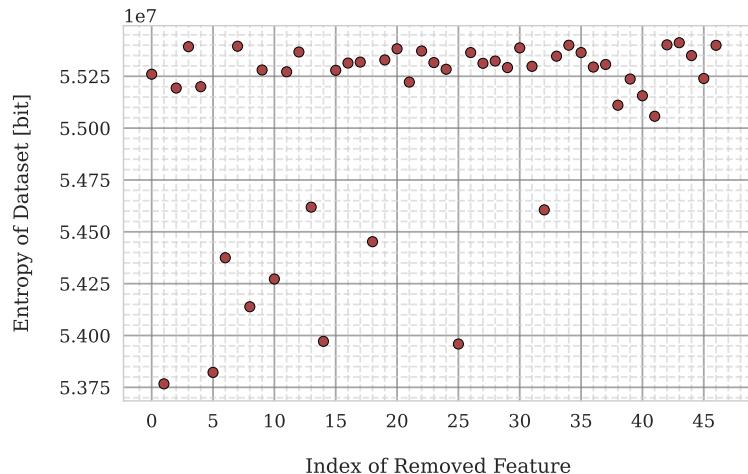
<sup>3</sup>In the case of  $S_{ij} = 0$ , it is considered that  $0\log(0) = 0$ .

The result of this method is shown in Figure 3.2. Notice that the entropy associated with the dataset for all variables, with one missing column, results in high values of bits; this is just a reflection that the dataset itself is extraordinarily disordered with high levels of uncertainty [53]. Moreover, the information contributed by the categorical data is less than the numerical data; this can also be observed in Figure 3.2, along with Table 3.8. When implemented in Python, the result is the following when printing the ordered features from the most important to the less important.

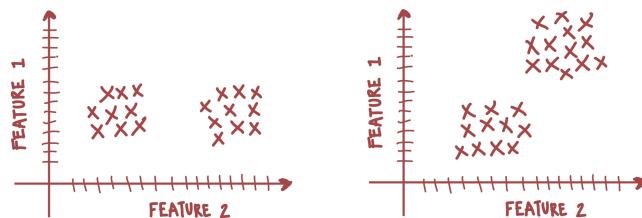
```
>>> print(ordered_features)

['WFICFORM', 'WFFORM', 'RHOSTAR', 'ZEFFNEO', 'DWDIA', 'BETASTAR', 'POHM', 'NEL',
 'NUSTAR', 'EPS', 'TAV', 'PFLOSS', 'WMHD', 'Q95', 'MEFF', 'PLTH', 'LCOULOMB',
 'OMEGACYCL', 'KAREA', 'PICRH', 'TIV', 'PRAD', 'TEV', 'PNBI', 'PECRH',
 'ELMFREQ', 'IP', 'ENBI', 'AMIN', 'ZEFF', 'TORQ', 'NESOL', 'BT', 'BEIMHD',
 'VOL', 'TOK', 'WALMAT', 'EVAP', 'PREMAG', 'LIMMAT', 'DIVMAT', 'ELMTYPE',
 'ECHMODE', 'ICSHEME', 'PELLET', 'CONFIG', 'AUXHEAT']
```

Meaning that the less important in providing the information is the type of auxiliary heating used during the shot, and the most important is the total fast ion energy due to ICRH. It is also interesting to note that most categorical data provide little information. From this, we can take the first 15 most important variables regarding entropy and subject that subset to analysis. In this case, the *entropy variables* will be WFICFORM, WFFORM, RHOSTAR, ZEFFNEO, DWDIA, BETASTAR, POHM, NEL, NUSTAR, EPS, TAV, PFLOSS, WMHD, Q95, and MEFF.



**Figure 3.2:** Result of applying Algorithm 2 to the 47 variables of interest in `big_DB5`; the same output resulted when applying it to `small_DB5`, despite having 848 fewer observations which may have made it more ordered. Table 3.8 shows the labels of the index of the removed feature. In [54] is the code I developed to create this plot.



**Figure 3.3:** Simple drawing showing the comparison between two situations to understand feature dependence for finding ordered groups.

Note that the presence of certain variables can overshadow or enhance the importance of others. For example, a single feature can contain ordered groups by itself, as shown in Figure 3.3 (left), while other features may need to be combined to obtain an ordered group, as shown in Figure 3.3 (right).

While the Entropy method considers dependence on other features, it does not evaluate the best subset of features. Due to the many possible feature subsets that can be assessed from the 47 variables of interest, alternative methods have been implemented to identify relevant subsets that provide insights into the factors affecting the decrease in  $\alpha_R$ .

**Table 3.8:** Numbering of variables in Figure 3.2.

0 - AMIN	12 - EPS	24 - PECRH	36 - TEV
1 - AUXHEAT	13 - EVAP	25 - PELLET	37 - TIV
2 - BEIMHD	14 - ICSSCHEME	26 - PFLOSS	38 - TOK
3 - BETASTAR	15 - IP	27 - PICRH	39 - TORQ
4 - BT	16 - KAREA	28 - PLTH	40 - VOL
5 - CONFIG	17 - LCOULOMB	29 - PNBI	41 - WALMAT
6 - DIVMAT	18 - LIMMAT	30 - POHM	42 - WFFORM
7 - DWDIA	19 - MEFF	31 - PRAD	43 - WFICFORM
8 - ECHMODE	20 - NEL	32 - PREMAG	44 - WMHD
9 - ELMFREQ	21 - NESOL	33 - Q95	45 - ZEFF
10 - ELMTYPE	22 - NUSTAR	34 - RHOSTAR	46 - ZEFFNEO
11 - ENBI	23 - OMEGACYCL	35 - TAV	

### 3.1.3 Research Variables

During the last year of my master, I attended a two-week workshop at CEA Cadarache that required my team and me to estimate the toroidal magnetic field  $B_t$  [T] and the radius  $R_{geo}$  [m] of a tokamak given the target fusion power  $P_{fus}$  [MW] and fusion gain  $Q$ . The solution to this problem requires the creation of a system of non-linear equations [55], such as:

$$R_{geo}^3 B_t^4 = \frac{P_{fus} C_\beta^2 q_{95}^2}{C_{fus} C_I^2 \kappa_a \epsilon^4 \beta_N^2}, \quad (3.5)$$

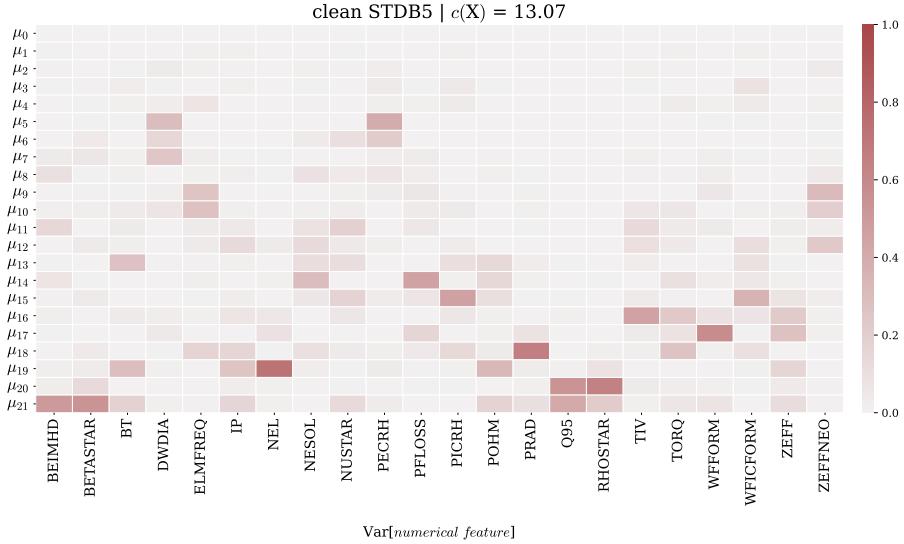
and

$$R_{geo}^{-\gamma_R} \cdot B_t^{-\gamma_B} = C \cdot \Gamma(Q) \cdot \alpha_0 \cdot P_{fus}^{\alpha_p} \cdot M_{eff}^{\alpha_M} \cdot \kappa_a^{\alpha_k} \cdot \epsilon^{\gamma_\epsilon} \cdot n_N^{\alpha_n} \cdot q_{95}^{-\gamma_l} \cdot \beta_N, \quad (3.6)$$

where  $\Gamma(Q)$ ,  $\beta_N$ ,  $n_N$ ,  $\gamma_R$ ,  $\gamma_B$  and the constants  $C$ ,  $C_I$ ,  $C_{fus}$ ,  $C_\beta$  are defined through the derivation (see Appendix A) as properties of the tokamak are established. The main takeaway of this research, related to this project, is that one can get an expression where the energy confinement time  $\tau_{E,th}$  scales with four engineering variables: line average electron density  $\bar{n}_e$ , temperature of the plasma  $\hat{T}$ , outer radius  $R_{geo}$ , and toroidal magnetic field  $B_t$ . Hence, I can consider this set of variables as a subset of interest. Additional to these variables, I also consider  $\rho_*$ ,  $v_*$ , and  $\beta_t$ , as these are the variables that fundamentally govern the transport at the plasma's core, according to Kadomtsev's analysis [8], [14]. Therefore, I have decided that the *research variables* are: NEL ( $\bar{n}_e$ ), TAV ( $\hat{T}$ ), BT ( $B_t$ ), RHOSTAR ( $\rho_*$ ), NUSTAR ( $v_*$ ), and BETASTAR ( $\beta_t$ ).

### 3.1.4 Low Multicollinearity Variables

To assess the amount of multicollinearity in the variables of interest, I have implemented the same techniques explained in the previous chapter to `clean_DB5`. The result is shown in Figure 3.4.



**Figure 3.4:**  $\Pi$  matrix and condition number for all numerical features of interest.

The VIF of PLTH and PNBI were 4200 and 3600, respectively. To derive a subset of variables with low multicollinearity, I systematically removed one variable at a time and recalculated the condition number along with the VIF of each variable until they stopped being too high. To reach a condition number of 13.07 and a maximum VIF of 14, the following variables were removed: PLTH, AMIN, LCOULOMB, KAREA, EPS, MEFF, TAV, ENBI, PNBI, OMEGACYCL, TEV, VOL, WMHD. Hence, the resultant variables with low multicollinearity are:

```
>>> print(low_multicollinearity_features,)

[ 'BEIMHD', 'BETASTAR', 'BT', 'DWDIA', 'ELMFREQ', 'IP', 'NEL', 'NESOL', 'NUSTAR',
  ↵ 'PECRH', 'PFLOSS', 'PICRH', 'POHM', 'PRAD', 'Q95', 'RHOSTAR', 'TIV', 'TORQ',
  ↵ 'WFFORM', 'WFICFORM', 'ZEFF', 'ZEFFNEO']
```

made of 22 numerical features.

## 3.2 Classification

As mentioned, machine learning algorithms are trained to learn from data to make predictions or decisions. If the algorithm predicts numerical data, it is known as regression, such as OLS. On the other hand, if the algorithm predicts classes or categorical data, it is known as classification. Both regression and classification are forms of supervised machine learning, meaning that the data must include a variable output column  $y$  for the algorithm to learn from the input columns  $X$ . The algorithm aims to learn how the input relates to the output to correctly predict  $y$  for new inputs it has never seen [30]. Given this information, it is possible to use the results from Chapter 2 to add a new column to `clean_DB5` containing the labels that do not affect or decrease  $\alpha_R$ , labelled as 0 and 1, respectively. Observations belonging to either `small_ds` or `big_ds` (depending on the case being studied) will be labelled with 1.

Numerous algorithms are available for classification, some of which may or may not be well-suited for the problem. Specific algorithms, referred to as "data-hungry", require large amounts of data and tend to perform well in fields such as social media, where a vast amount of data is generated daily. However, these algorithms can be challenging to implement in academic research as acquiring more data is typically expensive or unattainable [56], [57]. On the other hand, some algorithms perform exceptionally well with a small amount of data because they require the inverse of a particular matrix, which can be computationally expensive [30].

### 3.2.1 Imbalanced Data

For this project, I have adopted the step-by-step framework proposed by [58], designed to handle imbalanced data classification problems. The framework focuses on selecting a classifier that performs well in predicting the minority class, which in our case is the class that decreases  $\alpha_R$ . The author suggests first focusing on choosing the correct metric and then using that to test various algorithms.

Various metrics evaluate different aspects of the algorithm. For example, a particular metric can enhance the performance in predicting all classes, the majority or the minority class; selecting the wrong metric means selecting an incorrect algorithm for solving a specific problem [58]. Since we need to predict two labels, 0 and 1, we have a binary classification problem; various metrics for this type of problem can be obtained through a "confusion matrix" tool (see Figure 3.5). Table 3.9 shows some of the most common metrics obtained through a confusion matrix.

The utilisation of the confusion matrix requires the data to be split into two sections: one for training the algorithm and the other for validating its performance on unseen observations. Typically, the validation dataset comprises 20% to 50% of the complete dataset, depending on the amount of available data [30]. For this project, I created a validation dataset by stratified random<sup>4</sup> sampling 30% of clean\_DB5, with its corresponding labels.

		Labeled as decrease [1]	Labeled as unaffected [0]	
		True Positive (TPs)	False Positive (FPs)	Precision
Predicted Decrease	Labeled as decrease [1]	True Positive (TPs)	False Positive (FPs)	$\frac{TPs}{TPs + FNs}$
	Labeled as unaffected [0]	False Negatives (FNs)	True Negatives (TNs)	
Predicted Unaffected	Labeled as unaffected [0]	False Negatives (FNs)	True Negatives (TNs)	
		Recall	$\frac{TPs}{TPs + FNs}$	

**Figure 3.5:** Confusion matrix for a binary class and two metrics. This is a tool to obtain different metrics to evaluate an algorithm, here showing for precision and recall. The minority class is referred to as positive. See Table 3.9 for more metrics.

It is essential to mention that all classifiers use a threshold, whose values  $\in [0, 1]$ , to make decisions or predictions. The choice on this threshold directly impacts the algorithm's performance; if it is too high, the classifier will seldom predict the positive class, and vice versa [30].

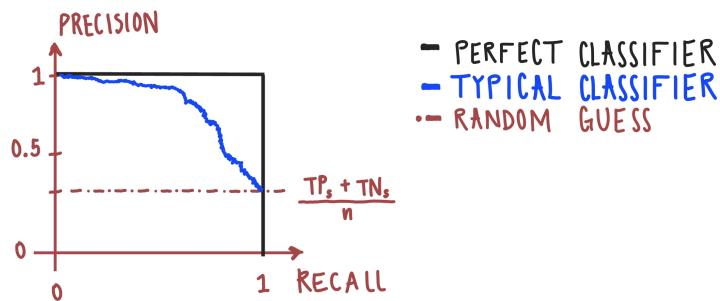
---

<sup>4</sup>This means that there is the same amount of classes in the training dataset as in the validation dataset.

**Table 3.9:** Some of the obtainable metrics from a confusion matrix [30], [58].  
Remember that  $n$  is a dataset's total number of observations.

Metric	Formula	Description
Misclassification Rate	$\frac{FNs + FPs}{n}$	Fraction of predictions being incorrect
Accuracy	$\frac{TNs + TPs}{n}$	Complement of misclassification rate
Precision	$\frac{TPs}{TPs+FPs}$	Fraction of predicted positives actually being positive
Recall	$\frac{TPs}{TPs+FNs}$	Fraction of actual positives correctly predicted
Fall-out	$\frac{FPs}{FPs+TNs}$	Probability of false alarm
Specificity	$\frac{TPs}{FPs+TNs}$	Compliment of fall-out
False discovery rate	$\frac{FPs}{FPs+TPs}$	Fraction of incorrect positive predictions
False negative rate	$\frac{FNs}{FNs+TPs}$	Fraction of actual positive incorrectly classified
False omission rate	$\frac{FNs}{FNs+FPs}$	Fraction of incorrect negative relative to tall incorrect classifications
Prevalence	$\frac{FNs + TPs}{n}$	Proportion of actual positive instances in the dataset
$F_1$ -score	$\frac{2 \cdot precision \cdot recall}{precision + recall}$	Harmonic mean of precision and recall
$F_\beta$ -score	$\frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2(precision + recall)}$	Used to account that recall is $\beta$ -times as important as precision

When dealing with imbalanced or asymmetric problems, it is recommended to focus on metrics such as  $F_1$ -score, precision, and recall [58]. Imbalanced data refers to a dataset with more instances of a particular class than others. An asymmetric problem arises when a false negative is more severe than a false positive or vice versa. For example, consider the diagnosis of a rare disease. In both situations, training an algorithm with accuracy or misclassification rate would lead to disastrous results [30]. As mentioned in Table 3.9,  $F_1$ -score is the harmonic mean of the precision and recall curve, which means it is a metric that represents the performance on the algorithm predicting the minority class in a single metric [30]. Another way to merge these two metrics is through the precision-recall curve (PRC), which is used to optimize the threshold of a classifier [58]. Figure 3.6 shows how to interpret these curves.



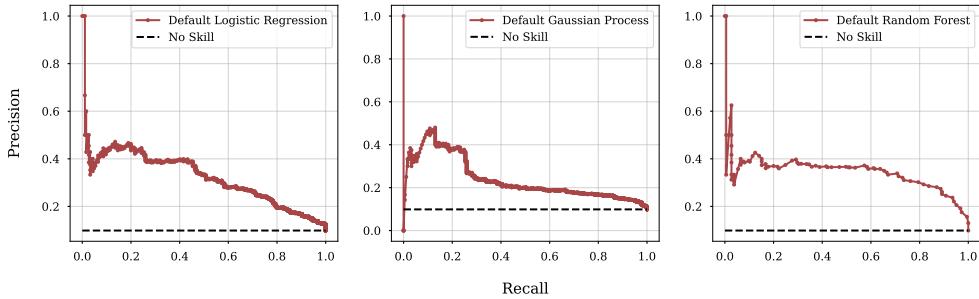
**Figure 3.6:** Drawing of a precision-recall curve (PRC) for three classifiers, as the threshold of the classifier decreases. The closer the area under the curve is to one, the better the classifier.

After having decided on the metric, now it is time to revise different classifiers and decide which one is worth our attention. For a quick comparison, I have decided to consider

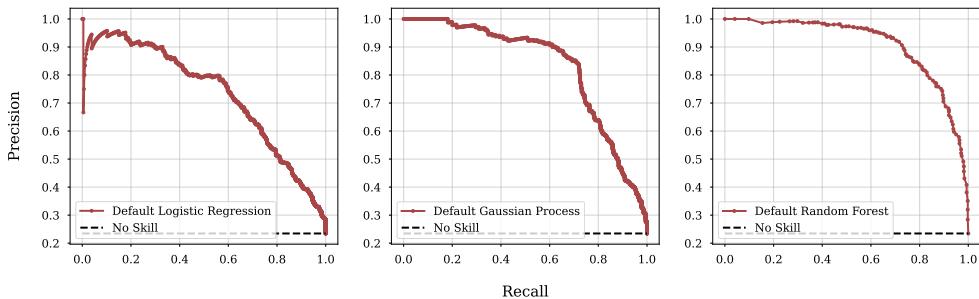
- (i) a linear parametric model: logistic regression,
- (ii) a non-linear probabilistic model: Gaussian process,
- (iii) a non-parametric ensemble: random forest.

These three algorithms share the property of not requiring a large amount of data to perform well, but there are significant differences between them. Algorithm (i) differs from (ii) and (iii) by modelling the output as a linear combination of weighted inputs. Both (ii) and (iii) can handle linear and non-linear combinations of the inputs. The main difference between (ii) and (iii) is that (ii) is a probabilistic model that assumes the observations are drawn from a multivariate normal distribution, whereas (iii) makes no strong assumptions<sup>5</sup> about the data and makes decisions based on information gain [30].

One can quickly implement these algorithms using scikit-learn<sup>6</sup>, a Python library designed to execute various machine learning algorithms [60]. For a quick comparison, I have implemented these three models using the default settings of scikit-learn<sup>7</sup> and all variables of interest with one-hot encoding applied to categorical data and scaling numerical values. results of their PRC are shown in Figure 3.7 when considering `small_ds` and Figure 3.8 when considering `big_ds`.



**Figure 3.7:** PRCs of all variables of interest and default settings in scikit-learn for logistic regression, Gaussian process and random forest; utilizing `small_ds` to label observations. A model with no skill is shown with dashed lines, which means it is a model that classifies based on random guesses. In order, the  $F_1$ -score obtained per model were: 0.16, 0.31, and 0.26.



**Figure 3.8:** PRCs of all variables of interest and default settings in scikit-learn for logistic regression, Gaussian process and random forest; utilizing `big_ds` to label observations. A model with no skill is shown with dashed lines, which means it is a model that classifies based on random guesses. In order, the  $F_1$ -score obtained per model were: 0.66, 0.77, and 0.81.

The worst performance is on the linear classifier, regardless of the type of dataset being used. On the other hand, Gaussian process and random forest perform appropriately for `big_ds`, only; this means that `small_ds` is a case of severely imbalanced data and might require an algorithm that specializes in this

<sup>5</sup>In general, all machine learning algorithms assume that the input data is independent and identically distributed, this is known as the i.i.d. assumption.

<sup>6</sup>This library uses CPU power, a library that implements GPU with Nvidia graphic cards is cuML [59].

<sup>7</sup>Version 1.2.2

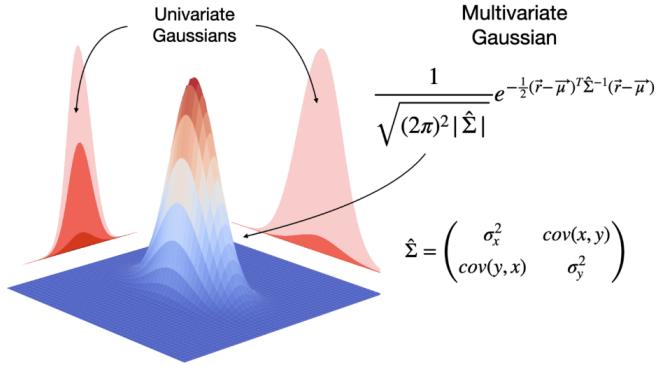
type of situation, like the one-class classifier [58]. For now, I will focus on the analysis of `big_ds` using Gaussian process (GP) and random forest (RF).

### 3.2.2 Gaussian Process and Random Forest

GP and RF are two machine learning algorithms that can be used both for regression and classification. The main difference between these two methods is that one uses a stochastic approach, while the other uses information gain from various models for making decisions [30]. I will explain in more detail the difference between them.

#### Gaussian Process

A collection of finite random variables whose joint distribution is a Gaussian distribution can be considered a Gaussian process. In other words, it is a method for modelling data using multivariate Gaussian distributions [61]. Figure 3.9 shows the case of a multivariate Gaussian distribution for two features contained in an arbitrary vector  $\vec{r}$ , with mean  $\vec{\mu}$  and covariance matrix  $\hat{\Sigma}$ .



**Figure 3.9:** . Illustration of a bivariate Gaussian distribution. Image from: [62].

GP is a collection of random variables representing the value of a real process  $f(\mathbf{x})$ , at location  $\mathbf{x}$ , which is fully defined by its mean function  $m(\mathbf{x})$  and covariance function  $k(\mathbf{x}, \mathbf{x}')$ , expressed as [30], [63]

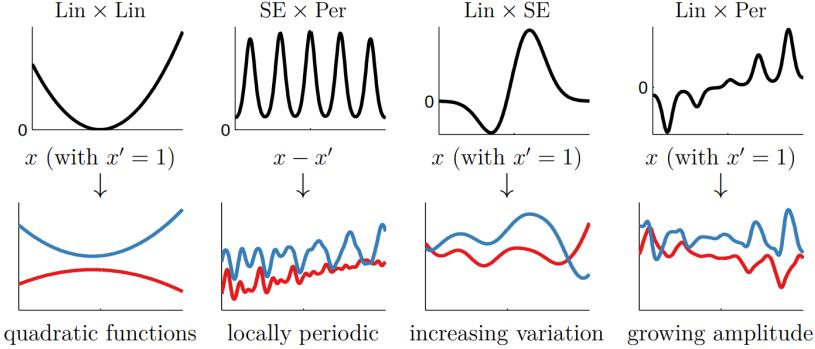
$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (3.7)$$

where

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})) \cdot (f(\mathbf{x}') - m(\mathbf{x}'))] = \text{Cov}[f(\mathbf{x}), f(\mathbf{x}')], \end{aligned} \quad (3.8)$$

with  $\mathbf{x}'$  denoting a different observation. It is common to set the mean  $m(\mathbf{x})$  to zero, as it is possible to add an extra term in the covariance function to represent its uncertainty [64], and to treat the covariance function as kernels [30]. The heart of GPs lies in the kernel and not in  $f(x)$  per se. The latter is more of a tool that allows deriving the path for inference. In fact, the problem of learning GPs is the problem of learning the optimal properties of the kernel. This is because the kernel states the relationship of given data to new data [63], [64].

Formally speaking, a kernel is a positive-definite function of two inputs  $\mathbf{x}, \mathbf{x}'$  that are Euclidean vectors, but they could also represent categorical inputs, graphs, images, or text. Another way to interpret a kernel is by treating them as mathematical objects specifying the similarity of a function evaluated on different objects, as shown in Eq. (3.8) [64]. By adding or multiplying kernels with different characteristics, it is possible to create a new kernel that encompasses all of these properties; examples of this are shown in Figure 3.10.



**Figure 3.10:** Some of the most common kernels, linear (Lin), squared exponential (SE), and periodic (Per); combined to form a new kernel. The top row of plots represents  $k(\mathbf{x}, \mathbf{x}')$  and the bottom plots are two samples of a GP with the specified kernels above; this type of sampling is known as sampling the *prior* [61].

Image from: Fig. 2.2 in [64].

Once the mean and the covariance function have been specified, one proceeds to compute the predictive distribution  $p(\mathbf{y}' | \mathbf{x}', \mathbf{X}, \mathbf{y})$  for the unknown target variable  $\mathbf{y}'$ , given the new input  $\mathbf{x}'$  and the data used to train the algorithm  $\mathbf{X}$ , with their corresponding labels  $\mathbf{y}$ ; namely,

$$p(\mathbf{y}' | \mathbf{x}', \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{y}' | \mu_*, \Sigma_*), \quad (3.9)$$

which represents the posterior after observing the data and turns to be a GP too. Here,

$$\begin{aligned} \mu_* &= \mathbf{k}_* K^{-1} \mathbf{y} \\ \Sigma_* &= k(\mathbf{x}', \mathbf{x}') - \mathbf{k}_* K^{-1} \mathbf{k}_*^T, \end{aligned} \quad (3.10)$$

where  $K^{-1}$  is the inverse of covariance matrix  $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ , and  $\mathbf{k}_*$  is the row-vector  $\mathbf{k}_* = (k(\mathbf{x}_1, \mathbf{x}') \dots k(\mathbf{x}_n, \mathbf{x}'))$ . Then, one finds the optimal hyperparameters of the kernel by maximizing the marginal likelihood  $p(\mathbf{y} | \mathbf{X})$ , also Gaussian, through an optimization algorithm, like gradient descent. Broadly speaking, this is the general procedure to implement GPs for regression, an algorithm with a complexity  $\mathcal{O}(n^3)$  [61]. Nevertheless, when working GPs for classification, not everything is analytically tractable since the labels are discrete and cannot have a Gaussian likelihood. In this situation, one "squashes" the output of a regression model into a class probability through what is known as a response function; for instance,

$$p(y = 1 | \mathbf{x}) = \sigma(y f(\mathbf{x})), \text{ and, } p(y = 0 | \mathbf{x}) = 1 - \sigma(y f(\mathbf{x})), \quad (3.11)$$

where  $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$  and  $\sigma(z)$  can be the sigmoid function  $\sigma(z) = [1 + \exp(-z)]^{-1}$  or the cumulative density function of a standard normal distribution  $\sigma(z) = \Phi(z) = \int_{-\infty}^z \mathcal{N}(x|0, 1) dx$ , known as the probit regression [63], [65]; this depends on the problem at hand. Here, we take the sigmoid function. After

defining the model through the link function, one can obtain the unnormalised posterior as Eq. (15.33) in [65]:

$$\ell(f) = \log p(y | f) + \log p(f | X) = \log p(y | f) - \frac{1}{2} f^T K^{-1} f - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi, \quad (3.12)$$

which can be normalised and analytically approximated by Laplace approximation [63], [65], [66]. One then gets the posterior predictive distribution as Eq. (15.48) in [65],

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbb{E}[f_*], \text{Var}[f_*]),$$

where  $\mathbf{x}_*$  is the best point used for  $f$ ; and  $f_* = f(\mathbf{x}_*)$ . Then, the hyperparameters of the kernel are estimated with a gradient-based optimization algorithm, such that the marginal likelihood is maximised [66]; namely, Eq. (15.51) in [65],

$$\log p(y | \mathbf{X}) \approx \log p(y | \hat{f}) - \frac{1}{2} \hat{f}^T K^{-1} \hat{f} - \frac{1}{2} \log |K| - \frac{1}{2} \log |K^{-1} - \nabla \nabla \log p(y | f)|, \quad (3.13)$$

where  $\hat{f}$  is the maximum a posteriori estimation algorithm<sup>8</sup> (MAP) applied to  $f$ . These steps are applied to obtain the GP classifier method, shown in Algorithm 3; where  $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$ ,  $\mathbf{I}_n$  is the  $n \times n$  identity matrix, and Cholesky refers to the Cholesky decomposition, used to efficiently inverse matrices [63], [65].

---

**Algorithm 3** Gaussian Process Classifier, based in Algorithm 15.2 in [65]. The algorithm I have implemented is *GaussianProcessClassifier()* of scikit-learn [68], which also follows this structure.

---

**Require:** MAP using iteratively reweighted least squares.

```

 $f \leftarrow 0$ 
repeat
   $W \leftarrow -\nabla \nabla \log p(y | f)$ 
   $B \leftarrow \mathbf{I}_n + W^{1/2} K W^{1/2}$ 
   $L \leftarrow \text{Cholesky}(B)$ 
   $b \leftarrow Wf + \nabla \log p(y | f)$ 
   $a \leftarrow b - W^{1/2} L^T \backslash (L \backslash (W^{1/2} Kb))$ 
   $f \leftarrow Ka$ 
until convergence
 $\log p(y | X) \leftarrow \log p(y | f) - \frac{1}{2} a^T - \sum_i \log L_{ii}$ 
% Perform prediction
 $\mathbb{E}[f_*] \leftarrow \mathbf{k}_*^T \nabla \log p(y | f)$ 
 $v \leftarrow L \backslash (W^{1/2} \mathbf{k}_*)$ 
 $\text{Var}[f_*] \leftarrow k_{**} - v^T v$ 
 $p(y_* = 1) \leftarrow \int \sigma(z) \mathcal{N}(z | \mathbb{E}[f_*], \text{Var}[f_*]) dz = 0$ 

```

---

The algorithm takes  $\mathcal{O}(n^3)$  for fitting, and  $\mathcal{O}(n^2 n')$  for prediction, where  $n'$  is the total number of new observations to be classified [65]. Extensive details on the derivation of this algorithm are found on pp. 525-528 [65], and pp. 33-48 [63].

## Random Forest

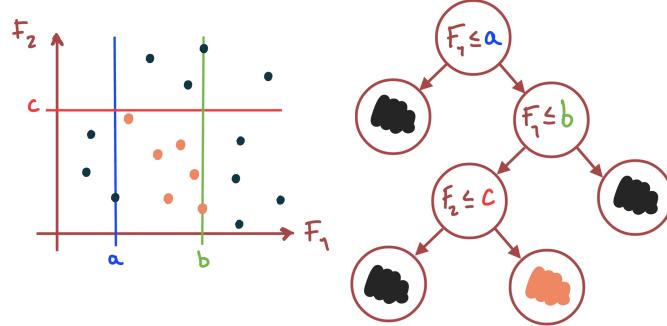
*Ensemble learning* is a method that involves constructing a new model by combining multiple basic

---

<sup>8</sup>A probabilistic framework to estimate probability densities. A good description of MAP is found in [67].

models, each trained in slightly different ways. By doing so, the ensemble leverages the models' different outputs to define the input-output relationship. [30]. RF is an instance of ensemble learning using decision trees with different randomised samples.

A decision tree is a rule-based model that structures a graph called a tree. Various disjoint regions are created to divide the input space, with each region having a fixed value to predict the output [30]. A simple tree for a classification problem is shown in Figure 3.11 using two numerical features.



**Figure 3.11:** Simple drawing illustrating a plot for two numerical features and their corresponding binary decision tree. Here,  $a$ ,  $b$ , and  $c$  represent real numbers creating decision boundaries, shown as lines in the plot. The top node of a tree is called the *root node*, and its subsequent branches are referred to as child-internal nodes. Nodes with a condition are known as *parent nodes*, while the nodes that follow them are their *child nodes*. A node that does not have any further branches is called a *leaf node*, which leads to a classification outcome [30].

Drawing is inspired by [69].

Observations may overlap with the decision boundary, as shown in Figure 3.11 for  $b$ ; in such situations, it can be assigned to the majority class in its vicinity, corresponding to regions with a higher risk of misclassification. Furthermore, it is worth mentioning that if the number of leaf nodes is equal to the number of observations in the data set, then it is said that the tree is fully grown, meaning that the model mimics the training data. It has lost its generalisation capability, known as *overfitting* the model [30]. To prevent this, one can set a stopping criterion.

The challenge in decision trees lies in determining the constants that establish optimal decision boundaries for a given dataset, considering the infinite number of available options. The recommended approach for deriving these constants is employing *greedy algorithms* [30]. Specifically, the splitting rules are generated sequentially, one at a time, starting from the root node. This process is known as the recursive binary splitting algorithm [30], which has been implemented in various versions, including CART (classification and regression trees) [70], ID3 (Iterative Dichotomiser 3) [71], and C4.5 [72].

To create a binary decision tree for a binary classification problem, one begins by designing the root node; this implies iterating over all features and assessing which ones give the best split. There are three different criteria to measure the quality of a split through their amount of *impurity*; hence, the lower the impurity, the better [30], [73]. The different metrics are shown in Figure 3.12. Mathematically, each of the criteria for computing the impurity per node  $i(N)$  is defined as [30], [73]

- misclassification error:

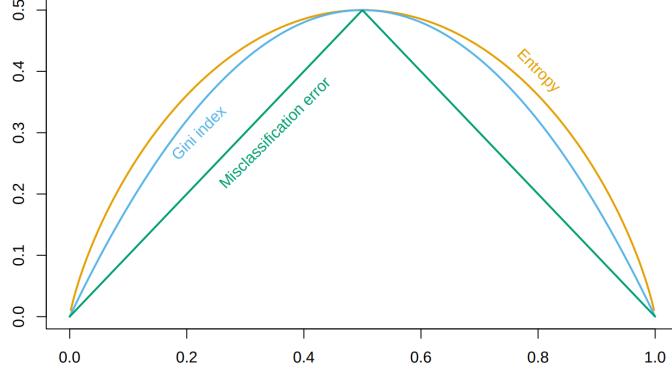
$$i(N) = 1 - \max(\pi_1, 1 - \pi_1), \quad (3.14)$$

- Gini index:

$$i(N) = 2\pi_1(1 - \pi_1), \quad (3.15)$$

- and, entropy:

$$i(N) = -\pi_1 \log(\pi_1) - (1 - \pi_1) \log(1 - \pi_1). \quad (3.16)$$



**Figure 3.12:** Measure of node impurity  $i(N)$ , vertical axis, versus the proportion of the first class  $\pi_1$  in a given region denoted by a split, horizontal axis. Image from: Fig. 9.3 in [74].

Once the optimal root node has been determined, the subsequent splits are constructed based on the selected criteria and the preceding node, until the stopping criterion is met. The general idea of how to implement a binary decision tree classifier is shown in Algorithm 4.

---

#### Algorithm 4 Basic principles in a Decision Tree Classifier

---

**Require:** Stopping condition  $stopping\_cond$

**Require:** Training dataset  $\mathcal{D}$  and New observations  $\mathcal{D}'$

**Model Construction:**

Split data considering all observations in  $\mathcal{D}$

Compute quality of split according to the selected impurity metric

**while**  $stopping\_cond$  is not fulfilled and further splits are possible **do**

    Go through all possible splits considering previous ones

    Pick the split that minimizes impurity of the node

    Check for  $stopping\_cond$

Store tree conditions in  $T$

**Prediction:**

**for** each row in  $\mathcal{D}'$  **do**

    Traverse tree  $T$  based on the conditions

    Get prediction for the row

---

One drawback of decision trees is their high sensitivity to the training data, even when the tree is not fully grown. To address this issue, an improvement can be achieved by creating multiple trees trained on the same dataset but with the introduction of randomness. These ensembles of randomised trees are known as random forests.

A common practice to create a random forest is to start with a specified number of trees, typically ranging from 50 to 1000. The number of trees corresponds to the number of bootstrap samples, with replacement, that will be drawn from the dataset. Each subset generated through bootstrapping is then used to train an individual tree. Subsequently, a new observation can be passed through all the trees, and the majority vote across the ensemble determines the final classification. This process is referred to as aggregation. The overall procedure, combining bootstrapping and aggregation, is known as bagging. It is worth noting that other models can be constructed following the same reasoning by utilising different base models [30], [73].

For an ensemble to outperform a single model, two key requirements must be met: (i) each classifier within the ensemble should be uncorrelated, and (ii) the individual classifiers should have an error probability below 0.5 [73].

An additional source of randomness can be introduced to enhance the decorrelation among trees. Instead of considering different tree sizes while using all features, one can fully grow each tree but utilise different randomised subsets of features to develop them. It is worth noting that the latter requires less computational power than the former [73], but the optimal implementation depends on the dataset at hand; for instance, if there are several features with similar information, it is advisable to implement the latter technique.

### Algorithm Tuning

For an algorithm to achieve optimal performance, tuning its hyperparameters according to the specific dataset is crucial. The advantage of Gaussian process is that it automatically tunes its hyperparameters while calculating the maximum likelihood [61]. Conversely, the hyperparameters of Random Forests, such as the number of trees to be created or the number of features to be considered per split, are tuned using a grid search methodology [30]. A grid search involves exploring a range of different values for the hyperparameters and evaluating the algorithm's performance for each combination. The hyperparameters that yield the best performance are then selected as the optimal choices [30], [61]. Furthermore, it is worth mentioning that, for better performance, Gaussian process benefits from scaling the numerical values before its implementation, while Random forests do not require said preprocessing [61], [73].

### 3.2.3 Feature Importance

There is a way to assess the most significant features for the algorithm to learn. The idea is based on randomly shuffling the values of a single feature and subjecting the modified dataset to the model. The model's performance is then evaluated using a specific metric, such as precision or recall. If the model's performance significantly worsens compared to the original dataset, it is considered that the respective feature played an essential role in learning [75]. It is important to note that if there are two correlated features and one of them is randomly shuffled, the model can still access their properties through the other correlated variable, thereby masking the importance of the shuffled feature [76].

---

**Algorithm 5** Feature importance algorithm by permutation [77].

---

**Require:** Trained model *model*

**Require:** Training or validation dataset  $\mathcal{D}$

*rep*  $\leftarrow$  total amount of repetitions (here I use 200)

*m*  $\leftarrow$  metric score based on  $\mathcal{D}$  (here I use  $F_1$  score)

**for** each feature  $F_k$  in  $\mathcal{D}$  **do**

**for**  $j$ -th repetition in range(*rep*) **do**

        randomly shuffle  $F_k$  to get a new dataset  $\tilde{\mathcal{D}}$

$\tilde{m}_{kj} \leftarrow$  metric\_score( $\tilde{\mathcal{D}}$ ) for the  $j$ -th shuffle using the *model*

    Compute importance  $i_k$  for  $F_k$  with corresponding standard deviation  $\sigma_k$

$$i_k = m - \frac{1}{rep} \sum_{j=1}^{rep} \tilde{m}_{kj}, \quad \sigma_k = \sqrt{\frac{\sum_j (\tilde{m}_{kj} - \text{mean}(m_k))^2}{rep}}$$

Rank features based on greater importance.

---

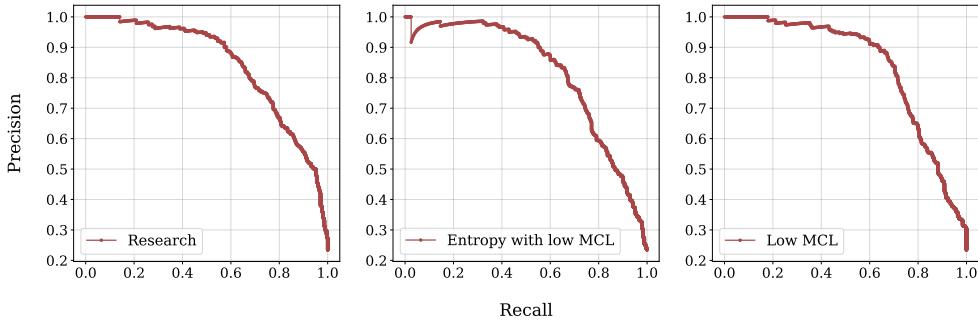
### Entropy with Low Multicollinearity Variables

Before presenting the results of the feature importance for each algorithm and subset of features, I have ex-

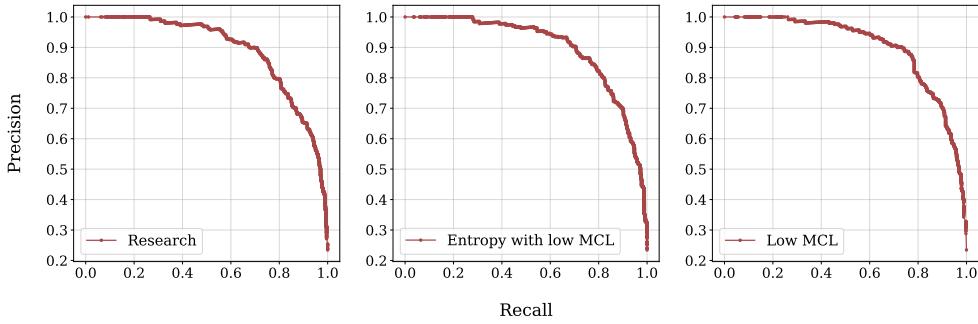
amined the multicollinearity in the entropy variables. It yielded  $c(\mathbf{X}) \sim 30$ . Consequently, when analysed with an RF classifier, I strategically eliminated the features with high collinearity and low importance. The resulting set of features demonstrates low multicollinearity and high importance. These features include WFICFORM, WFFORM, RHOSTAR, ZEFFNEO, DWDIA, BETASTAR, NUSTAR, PFLOSS, and Q95. Collectively, they exhibit a multicollinearity value of  $c(\mathbf{X}) = 7.92$ . Henceforth, this subset of features will be referred to as the *entropy with low multicollinearity*. It is worth mentioning that the research variables have  $c(\mathbf{X}) = 6.97$ .

### 3.3 Results

Up to this point, two algorithms, Gaussian process and random forest, can be used to assess three different subsets of variables: research, entropy with low multicollinearity, and low multicollinearity. Consequently, we have six different models to examine the variables that may provide insights into the reduction of  $\alpha_R$  in STDB5. Additionally, I have added the categorical features to the entropy with low multicollinearity and research features for a complete analysis. Figure 3.13 depicts the PRCs for the three cases utilizing Gaussian Process, while Figure 3.14 illustrates the PRCs for tuned Random Forests.

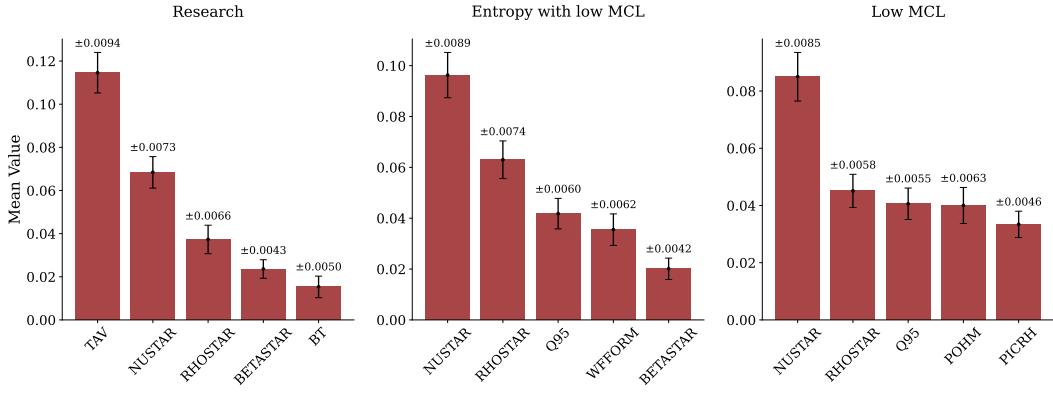


**Figure 3.13:** PRCs for GP classifiers with three different subsets of features being used. In order, the  $F_1$ -score obtained per model were: 0.74, 0.73, and 0.77.

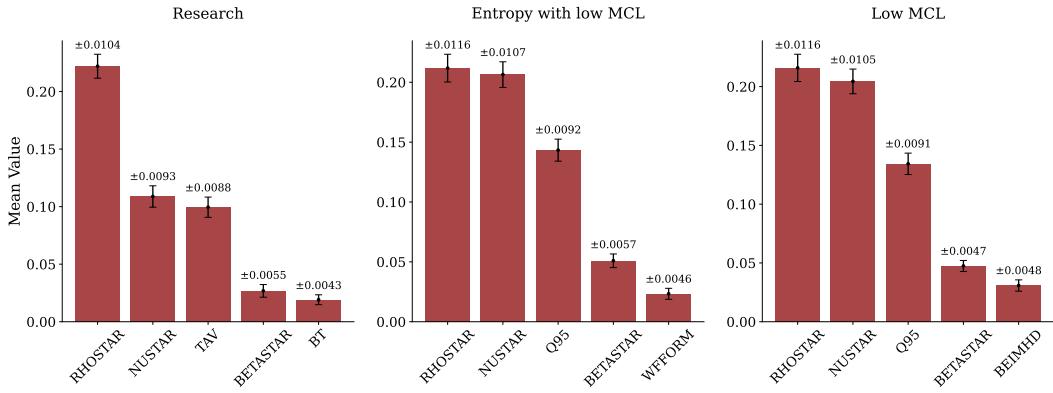


**Figure 3.14:** PRCs for RF classifiers with three different subsets of features being used. In order, the  $F_1$ -score obtained per model were: 0.80, 0.81, and 0.82.

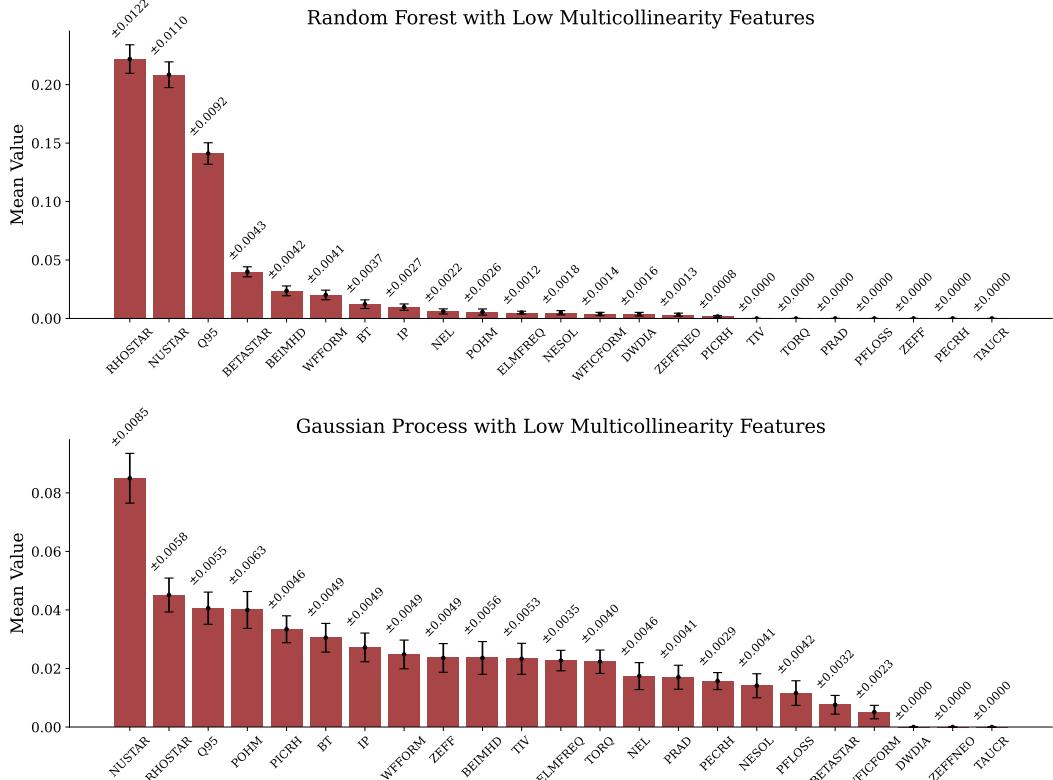
It is evident that RF consistently outperforms GP in all scenarios. Nevertheless, all algorithms effectively learned from the various datasets to make predictions that reduce  $\alpha_R$ , exhibiting acceptable performance. The Random Forest model utilizing the subset of variables with low multicollinearity proved the most effective. Figure 3.15 displays the resulting first five feature importance for GP, while Figure 3.16 illustrates the feature importance for RF across the three subsets of variables. All analyses were made with 200 rounds of shuffling; the codes I developed for these analyses are found in [78], [79].



**Figure 3.15:** GP's most important features with standard deviations.



**Figure 3.16:** RF's most important features with standard deviations.



**Figure 3.17:** Complete feature importance for the best subset of features, for GP (bottom) and RF (top).

According to all algorithms, the normalised collision frequency and the normalised ion gyroradius emerge as crucial factors in predicting observations that lead to a decrease in  $\alpha_R$ . The normalised plasma beta ranks as the third most significant feature. The following vital features include the safety factor, average temperature, toroidal magnetic field, and others, which may vary. Notably, none of the categorical features played a significant role in the learning process of the algorithms. Figure 3.17 show the complete feature performance for the low multicollinearity variables.

Interestingly, GP uses most of the features to make predictions, while RF takes less than half of the subset to outperform GP. The subsequent chapter will thoroughly examine and interpret these variables to identify commonalities and establish connections with tokamak physics.

### 3.3.1 Predictions

Conveniently, the RF performs well with  $\rho_*$ ,  $\nu_*$ ,  $q_{95}$ , and  $\beta_t$  since these are parameters that are relatively easy to find in other devices. If the model is reduced to these features, the performance of the RF does not decrease drastically. When only these variables are kept, the precision and  $F_1$ -score are 0.8, with a recall of 0.79. Table 3.10 shows the central values for ITER and SPARC (a compact tokamak being developed by the MIT Plasma Science & Fusion Center [80]).

**Table 3.10:** Values of main features for two new toroidal machines. Values provided by the research group, except for SPARC's  $q_{95}$ , that was retrieved from [81]. Here,  $\tau_{E,th}$  is the expected value.

	$I_p$	$B_t$	$\bar{n}_e$	$P_{l,th}$	$R_{geo}$	$\kappa_a$	$\epsilon$	$M_{eff}$	$\rho_*$	$\beta_t$	$\nu_*$	$q_{95}$	$\tau_{E,th}$
<b>ITER</b>	15	5.3	1.03	87	6.2	1.8	0.32	2.5	0.002	2.24	0.014	3	3.5
<b>SPARC</b>	8.7	12.2	3.1	25	1.85	1.97	0.31	2.5	0.003	1.20	0.03	3.2	0.77

When their respective  $\rho_*$ ,  $\nu_*$ ,  $q_{95}$ , and  $\beta_t$  are put into the RF, both predictions result as **unaffected**; this could be interpreted as the machines being better described by the unaffected scaling law, presented in Table 2.2; namely,

$$\tau_{E,2.16} = 0.06 \cdot I_p^{0.78} \cdot B_t^{0.24} \cdot \bar{n}_e^{0.41} \cdot P_{l,th}^{-0.75} \cdot R_{geo}^{2.16} \cdot \kappa_a^{0.5} \cdot \epsilon^{0.79} \cdot M_{eff}^{0.22}. \quad (3.17)$$

When using this scaling for predicting ITER and SPARC's  $\tau_{E,th}$ , one finds 1.12 s and 0.27 s, respectively; both values with more than 60% error, considering the approximation error formula with the table's value being the true value.

## 3.4 Other Applications in Fusion

Before delving into the tokamak analyses relative to the essential features, it is worth mentioning some of the various novel applications of machine learning in fusion and nuclear technology. This section is inspired by a lecture presented by Dr Cristina Rea<sup>9</sup> at the Joint ICTP-IAEA School on AI for Nuclear, Plasma and Fusion Science [82].

Another application of multi-machine analysis is in disruption prediction. Disruptions are current-driven instabilities that abruptly terminate plasma confinement and can cause significant damage to the machine,

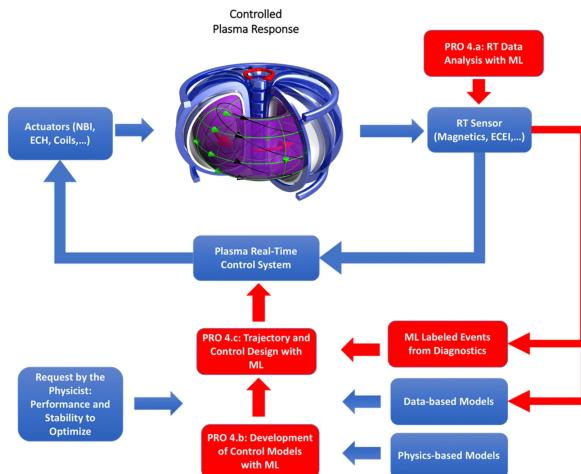
<sup>9</sup>The current group leader in Disruption Studies at MIT Plasma Science & Fusion Center.

making their avoidance crucial [83]. Random forests [84] and Deep-learning algorithms<sup>10</sup> can be employed to develop disruption prediction algorithms, even with limited data, and transfer knowledge from existing devices to new ones [85], [86]. These algorithms can be deployed in real-time environments to safeguard the machine [87], [88]. Additionally, the learning can be adapted to various operational regimes [89]. The authors in [87] also demonstrate that time-sequence non-disruptive data is specific to the machine, while time-sequence disruption data contain device-independent knowledge.

The edge plasma conditions have a significant impact on the operation and performance of a device. The advancement of this field has relied on the use of gyrokinetic codes and fluid simulations. However, some codes are still in the developmental stage, and others lack crucial physics. Despite various equation adaptations for effects such as blob dynamics or heat fluxes affecting plasma-facing components, no global code fully aligns with observations [90]. The authors in [90] have demonstrated the feasibility of constraining a physics-informed deep-learning framework with partial differential equations to learn and diagnose unknown turbulent fields effectively. The resulting learning exhibits consistency with the Braginskii two-fluid theory and provides an advanced way to design magnetised-turbulent plasmas diagnostics. They successfully employed deep learning to unveil the dynamics of unknown turbulent plasmas based on partial observations. Physics-informed algorithms are also employed in stellarators. In W7-X, neural networks are utilised to solve the heat equation and calculate the heat flux on its divertor tiles [82]. However, one of the current limitations of physics-informed learning algorithms is that they necessitate more robust algorithms and computational frameworks due to their involvement in highly non-convex optimisation problems [82].

Gaussian Process and nonlinear gyrokinetic simulations are employed to obtain a simultaneous evolution of energy sources, enabling the prediction of steady-state solutions for electron temperature, ion temperature, and electron density channels in SPARC [91]. This approach reduces computational costs and avoids the need for common simplifications, such as neglecting magnetic fluctuations or assuming a fixed density profile [91].

It is evident that data-driven machine learning methods are powerful tools that can address significant challenges in nuclear engineering and numerous other fields, including robotics, industrial processes, and medical research, leading to transformative advancements in these domains. These methods have effectively bridged the gap between practical implementations and theoretical models [92]. To conclude this chapter, I want to highlight the envisioning of the essentiality of machine learning algorithms in control rooms of nuclear devices. This concept is depicted in Figure 3.18.



**Figure 3.18:** Schematic of the role played by machine learning (ML) algorithms implemented in real-time situations within control rooms. Adapted image from: Fig. 8 in [92], presented by [82].

<sup>10</sup>A good introduction to deep-learning is found in [85].

# IV

## TOKAMAK ANALYSES

### 4.1 The Physics

In this chapter, I provide a brief overview of tokamak instabilities to then relate it to the classification found in Chapter 2. The discussion follows with a differentiation between the classification points and an interpretation through a dimensionless scaling parameter. In continuation, I analyse the dataset concerning the main features and present dynamic analysis. Finally, I conclude this chapter with a review from an engineering perspective based on plasma-facing components.

From this point onward, the comparative plots that distinguish between the decreasing and unaffected cases will only include the new observations in STDB5, excluding any observations from DB2, unless otherwise stated. This approach provides a more transparent comparison of the observations influencing  $\alpha_R$ .

#### 4.1.1 Instabilities in a Tokamak

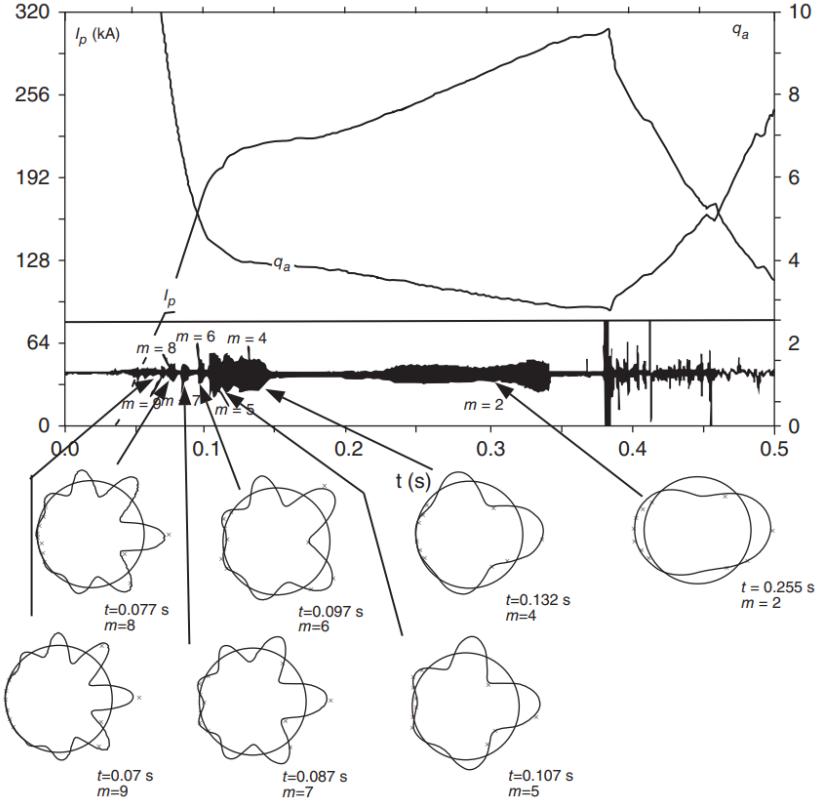
Recall that high magnetic shear is beneficial within tokamaks; this also implies that problems arise when the safety factor takes small values. This can happen when the poloidal magnetic field is large due to strong gradients in the plasma current profile. These issues can lead to long-wavelength instabilities, such as *kink instabilities*. From stability analysis – under the assumption of circular cross-sections of the plasma – it follows that

$$\frac{n_p}{n_t} < q_s(a) = \frac{2\pi a^2}{\mu_0 R_{geo}} \cdot \frac{B_\varphi}{I_p}, \quad (4.1)$$

is a condition that must be fulfilled to avoid long-wave instabilities [1]. Due to the fast current ramps during the initialisation of a tokamak, this is the stage where kink instabilities are more likely to occur. Figure 4.1 shows the different kink instabilities as their poloidal turns change. Notice that the figure does not consider the case  $n_p = 1$ . When  $n_p = 1$ , one obtains  $q_s(a) > 1$  an upper limit for the toroidal plasma current, known as the Kruskal-Shafranov limit, in the form of

$$I_p < \frac{2\pi a^2 B_\varphi}{\mu_0 R_{geo}} \cdot \frac{1}{2} \rightarrow [q_s(a) > 2], \quad (4.2)$$

where the  $1/2$  factor was experimentally determined. When this limit is not fulfilled, the plasma will likely end up in a disruption [1].  $\beta_t$  is another limited parameter to maintain equilibrium in fusion plasmas. Although this limit is more crucial to stellarators than tokamaks, it is worth looking at.



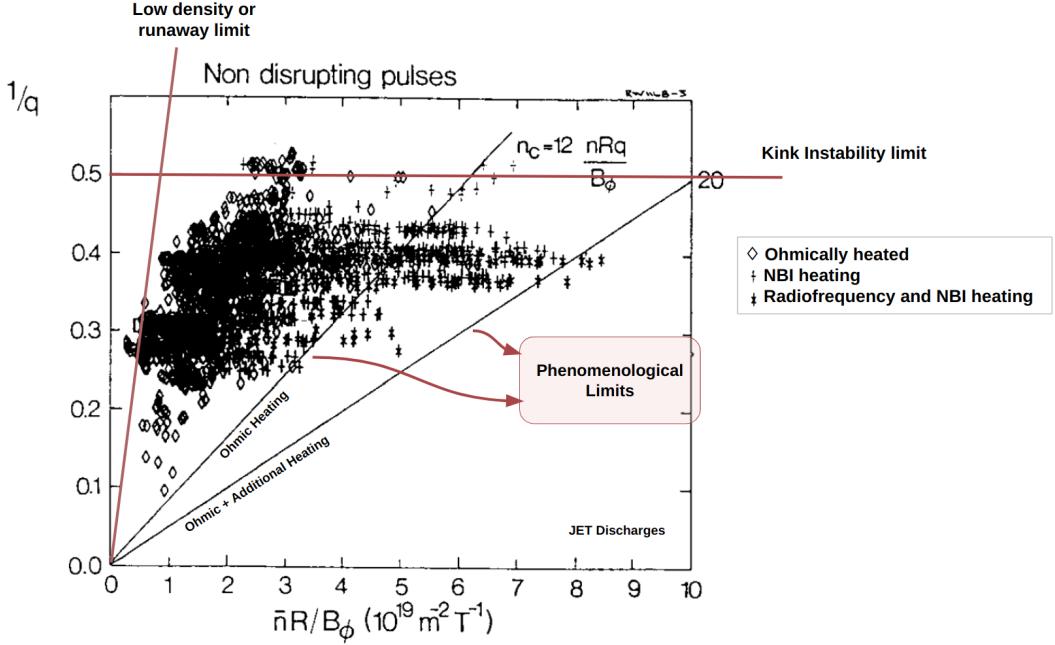
**Figure 4.1:** Appearance of long-wave instabilities during the current ramp up in ASDEX. Here  $m$  represents the poloidal turns. Image from: Fig. 4.4 in [93].

The phenomenon known as the *Shafranov shift* occurs when the concentric flux surfaces shift outward w.r.t.  $R_{geo}$  [93]. This shift arises due to the creation of twisted magnetic field lines, resulting in a poloidal current called the Pfirsch–Schlüter current, which, in turn, generates a vertical magnetic field. As a result, the plasma's shape is modified [1]. The Shafranov shift increases with  $\beta_t$ ; as  $\beta_t$  increases, the plasma pressure rises, increasing the diamagnetic current; this necessitates an increase in the Pfirsch–Schlüter current to maintain the confinement. To avoid instabilities, the shift must be lower or equal to the minor radius [1]. Through what is known as the Grad–Shafranov Equation, it is possible to estimate a limit for  $\beta_t$  [93]. From this analysis, it follows that the maximum stable  $\beta$ , is [1]

$$\beta_{eq}(r) \approx \frac{1}{2} \frac{a}{R_{geo}q_s^2(r)} = \frac{1}{2} \epsilon q_s^{-2}(r). \quad (4.3)$$

From this, it is possible to notice that spherical tokamaks are less restrictive to this limit due to their high aspect ratio.

One final limitation I will talk about, encountered only by tokamaks, is the *Greenwald limit*, an experimentally observed soft-operational limit to the electron density that ends in disruption when surpassed [14], [94]. The explanation of this limit is focused on the mechanism that leads to strong edge cooling or on collisionality-enhanced turbulent transport [14]. The stable parameter ranges are usually summarised in a plot known as the Hugill diagram that compares the Murakami parameter  $\bar{n}_e R_{geo}/B_t$  vs the inverse of the safety factor [1]. Other plots for stability analysis compare the plasma current with the density [95]. Figure 4.2 shows a Hugill plot for JET published in 1989.



**Figure 4.2:** Demonstrative Hugill plot for old stable shots performed by JET. Coloured lines are demonstrative and not computed with mathematical expressions. The low-density limit happens when there are not enough collisions to ionize the gas and start a plasma[1]. The phenomenological limits are the Greenwald limits which discern two different scenarios for heating. Adapted image from: Fig. 4 in [96], presented in [1].

The Greenwald limit can be estimated as [14]:

$$n_{GW} = \frac{10 \cdot I_p}{\pi a^2}. \quad (4.4)$$

One can obtain the limit by taking the maximum density with its corresponding registered  $q_{95}$ ,  $B_t$ , and  $R_{geo}$  and plot a line with inverse slope  $\bar{n}_e R_{geo} q_{95} / B_t \cdot 10^{19} \text{ m}^{-2} \text{T}^{-1}$  [97]. Nevertheless, it is worth mentioning that the Swiss Plasma Center at EPFL recently demonstrated that this is not a concern for big devices, like ITER, as they can use twice the amount of Hydrogen than expected without the risk of disruption [98].

#### 4.1.2 Insights within the Dataset

##### Dimensionless Scaling Law and Turbulence

There is a way to gain insight into the physics of a dataset through specific values taken by parameter regressors. One variable that exhibits an interesting attribute is  $\chi_\rho$ , the regressor parameter of the dimensional scaling law, found in Eq. (1.5). The significance of this variable stems from the fact that Eq. (1.5) can be expressed in terms of the Bohm time  $\tau_B = a^2 / D_B$ , where  $D_B \propto \hat{T} / B_t$  denotes the Bohm diffusivity [7]. Considering that  $\tau_B \propto \omega_i^{-1} \rho_*^{-2}$ , the dimensionless scaling can be formulated as

$$\tau_{E,th} \propto \tau_B \cdot \rho_*^{2+\chi_\rho} \cdot \beta_t^{\chi_\beta} \cdot v_*^{\chi_v} \cdot q_{cyl}^{\chi_q} \cdot R_{geo}^{\chi_R} \cdot \kappa_a^{\chi_\kappa} \cdot \epsilon^{\chi_\epsilon} \cdot M_{eff}^{\chi_M}. \quad (4.5)$$

In the case where  $\chi_\rho \sim -2$ , the scaling is referred to as *Bohm scaling*, as the macroscopic plasma dimensions represent the turbulence scale length instead of the gyroradius. When  $\chi_\rho \sim -3$ , the scaling is

referred to as *gyro-reduced Bohm* as the gyroradius is proportional to the Bohm-normalised diffusivity [7]. This implies that transport in new devices can be assessed with fundamental dimensionless variables by matching  $\nu_*$  and  $\beta_t$ , but not through  $\rho_*$ ; i.e. ITER  $\rho_*$  values are 5-8 times lower than present tokamaks [8]. Cases where  $\rho_*$  is considerably low, are associated with strongly magnetised plasmas [99]. The standard assumption is that

$$\omega_i \cdot \tau_{E, the} \propto \rho_*^{-(2+\chi'_\rho)} \cdot F(\nu_*, \beta_t, \{p_i\}), \quad (4.6)$$

with  $0 \leq \chi'_\rho \leq 1$ , with 0 corresponding to the Bohm scaling, and 1 to gyro-Bohm scaling. Here,  $F$  represents a function depending on the other dimensional variables and some engineering variables  $\{p_i\}$ . The characteristic turbulence scale length  $\ell \approx \rho_*^{\chi'_\rho} \cdot a^{1-\chi'_\rho}$  [8]. For instance, for H-mode, it is observed that  $\chi'_\rho = 0.9 \pm 0.3$ , where plasma transport is dominated by collisionless instabilities and low- $\beta$ . On the contrary, L-mode discharges display Bohm scaling [8], [100].

Table 4.1 shows the values  $\chi_\rho$ ,  $\chi'_\rho$  for different datasets, as described in Table 2.2, and characteristic turbulence scale length  $\ell$  for ITER and SPARC, based on Table 3.10.

**Table 4.1:** Estimate of parameters  $\chi_\rho$ ,  $\chi'_\rho$ , and characteristic turbulence scale length  $\ell$  for new devices.

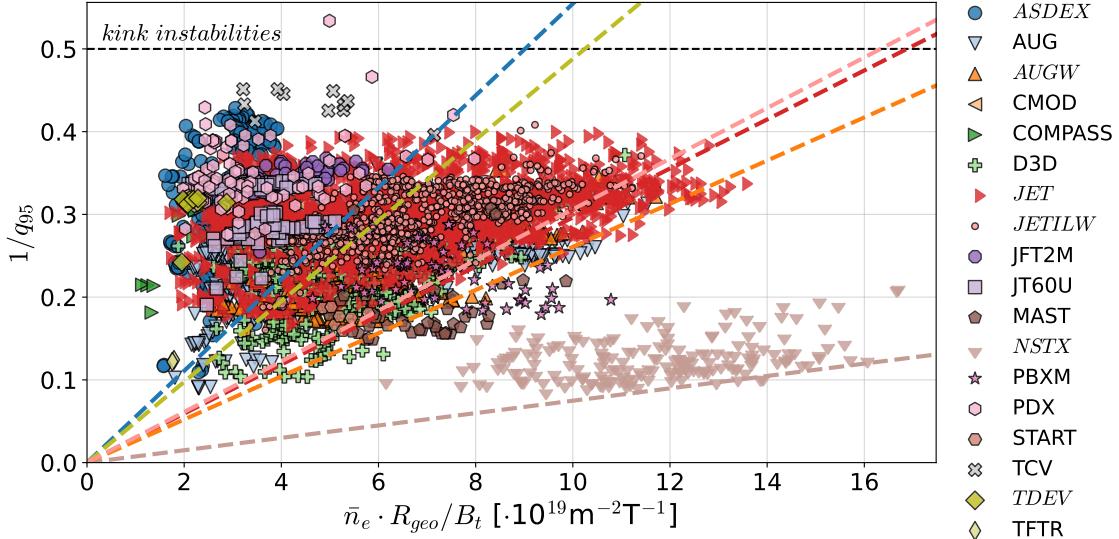
		big_ds		small_ds		
		DB2	STDB5	$\tau_{E,0.64}$	$\tau_{E,2.16}$	$\tau_{E,0.99}$
$\chi_\rho$		-3.09	-1.80	-1.31	-2.63	-1.5
$\chi'_\rho$		1.09	0.2	0.69	0.63	-0.5
ITER	$\ell$ [m]	0.0012	7.99	227.43	0.026	64.99
SPARC	$\ell$ [m]	0.0017	1.67	22.33	0.019	8.46
						1.15
						0.37

From Table 4.1, it is possible to notice that the only subset of observations that fall out of the standard assumptions is the decreasing observations in `small_ds`; this suggests that it lacks physics representation. Both subsets of observations created by `big_ds` follow the H-mode values range, with the unaffected records being close to a gyro-Bohm scaling. STDB5 and the unaffected observations both in `small_ds` and `big_ds` are closer to a Bohm scaling, while DB2 is the only subset with a clear tendency to a gyro-Bohm scaling. Furthermore, the decreasing observations in `big_ds` lead to the most extensive characteristic turbulence scale length, while DB2 gave the smaller ones. It is worth mentioning that the gyroBohm scaling is the situation expected from many turbulent theories and is the scaling leading to favourable predictions for confinement time [7].

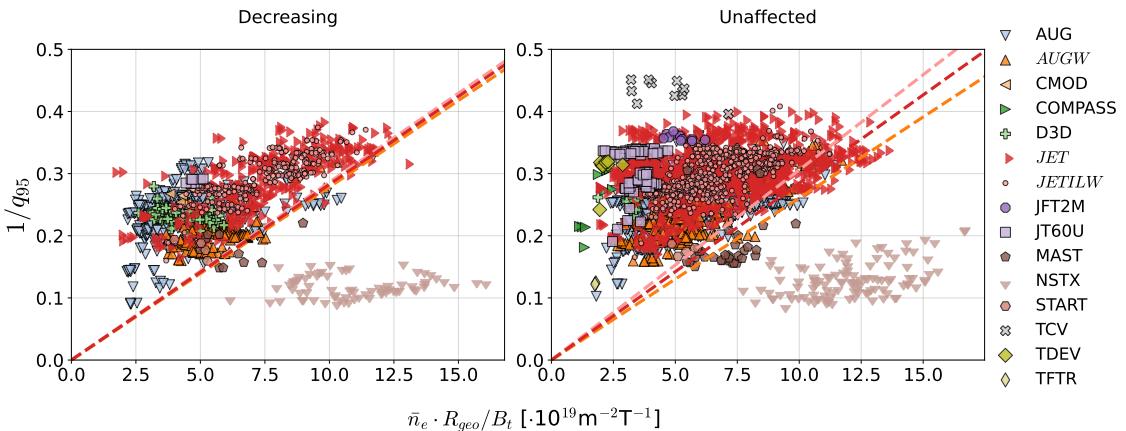
### Hugill Plots

Now that the previous analysis has helped to determine that the classification created by `big_ds` is more pertinent in terms of physics-related analysis over `small_ds`, let us have a look at the density limits of the datasets with physical relevance. Figure 4.3 shows the Hugill plot of STDB5 along with the Greenwald limit of selected tokamaks: ASDEX, AUGW, JET, JETILW, NSTX, and TDEV. This is to stress the fact that not all devices will have the same density limit. Notice how different is the density limit for NSTX compared to TDEV; maybe this is expected since the former is a spherical tokamak and the latter is not. However, notice how different the limit is between ASDEX and AUGW. The ITER-like devices, namely AUGW and JETILW, present close density limits. Interestingly, JET surpassed this limit multiple times, while AUGW and JETILW did a few times. Finally, PDX is the only registered device with a stable shot above the kink instability limit.

Figure 4.4 shows comparative Hugill plots for the new observations in STDB5, w.r.t. DB2, for the case of observations that decrease and do not affect  $\alpha_R$ . This comparison shows that ITER-like devices and JET decreasing observations have similar Greenwald limits compared to unaffected observations, w.r.t. the influence over  $\alpha_R$ . Furthermore, most observations in JET that surpass its density limit are within the unaffected classification, as well as the ones for AUGW, despite being fewer.



**Figure 4.3:** Hugill plot for STDB5 showing the density limit for kink instabilities and the Greenwald limit for the devices shown in italics. The reader can also find an interactive plot in HTML format by downloading the file in [\[101\]](#); it allows the user to select the tokamaks and limits to be displayed.



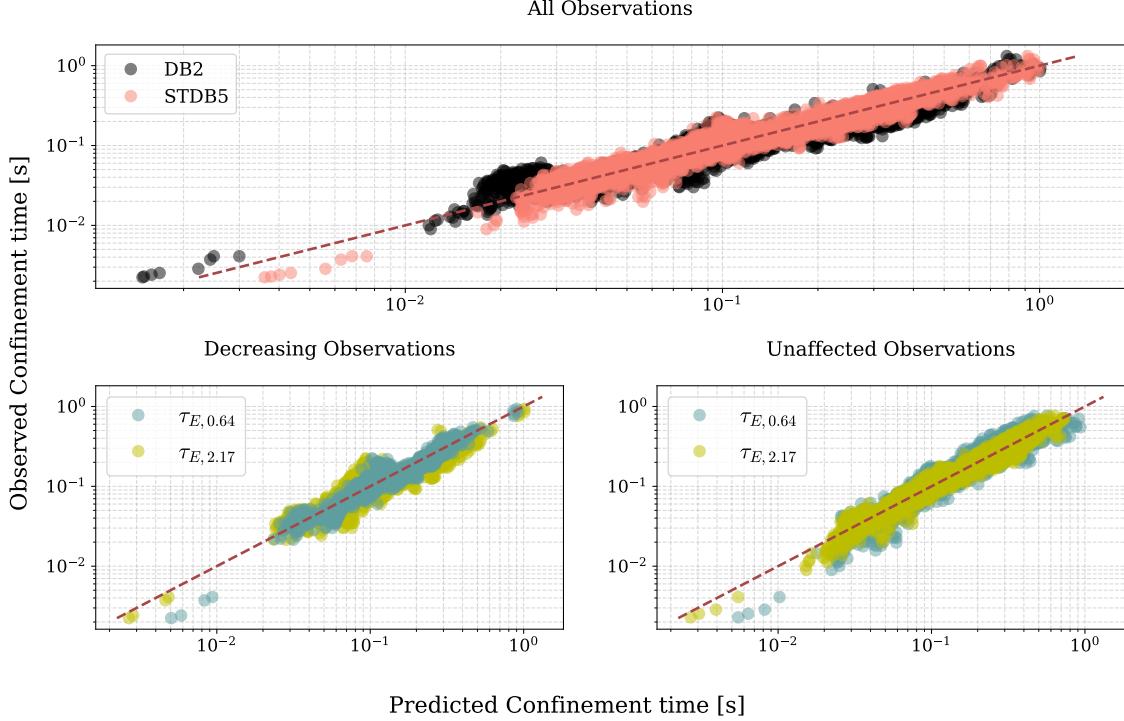
**Figure 4.4:** Comparative Hugill plot for new observations in STDB5 showing the density limit for kink instabilities and the Greenwald limit for ITER-like devices and JET; for the case of observations that decrease  $\alpha_R$  (left) and do not affect it (right) based on `big_ds`. If the reader is interested, the code I developed to create the Hugill plots is found in [\[102\]](#).

From this analysis, I believe it would be interesting to take the time-series data of all the unaffected observations that surpass the Greenwald density limit and try to find if they have similarities and subject those to physics and plasma-facing components interpretation.

### Variability in Confinement Time

Let us have a look at the variability on the predictions created by the scaling laws of DB2, DB5, and the resultants datasets from `big_ds` as considered in Figure 2.8. This is shown in Figure 4.5. From the

plot, it is possible to observe that the predictions done with the complete STDB5 generate less variability compared to DB2. Furthermore, When subjecting the data labelled as decreasing to the `big_ds`'s resultant scaling laws, the variability in predictions is less in  $\tau_{E,0.64}$  compared to  $\tau_{E,2.17}$ ; and vice versa. This suggests that it is beneficial distinguishing between observations that influence or not the decrease in  $\alpha_R$  when making predictions.



**Figure 4.5:** Comparative plots of measured and predicted  $\tau_{E,th}$ , with different scaling laws. The labelling in observations is based in `big_ds`. One can find the developed code for this figure in [103].

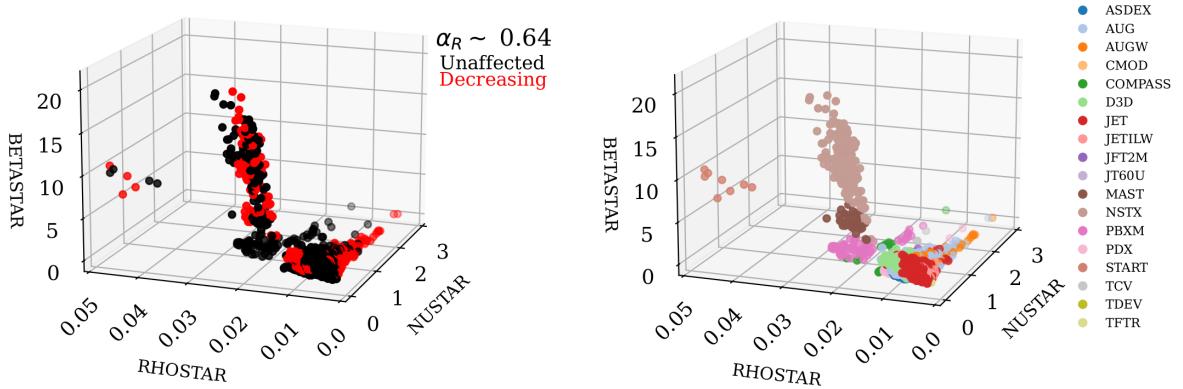
## 4.2 Variables of Importance and Tokamak Characteristics

According to the previous chapter, the normalised ion collision frequency (NUSTAR), the normalised ion gyroradius (RHOSTAR), the plasma safety factor at the 95% poloidal flux surface (Q95), and the normalised plasma beta (BETASTAR) are some of the main features that one should pay attention when determining which observations influence the reduction on  $\alpha_R$ . These are also four of the seven parameters in the dimensionless form of the energy confinement time  $\omega_i \tau_{E,th}$ . So, let us further revise these quantities. Table 4.2 shows the descriptive statistics of these four variables and Figure 4.6 shows a 3-dimensional scatterplot for  $\rho_*$ ,  $v_*$ , and  $\beta_t$ ; this table discerns from spherical data, which makes 3.77% of STDB5, to depict how different spherical values are from non-spherical ones.

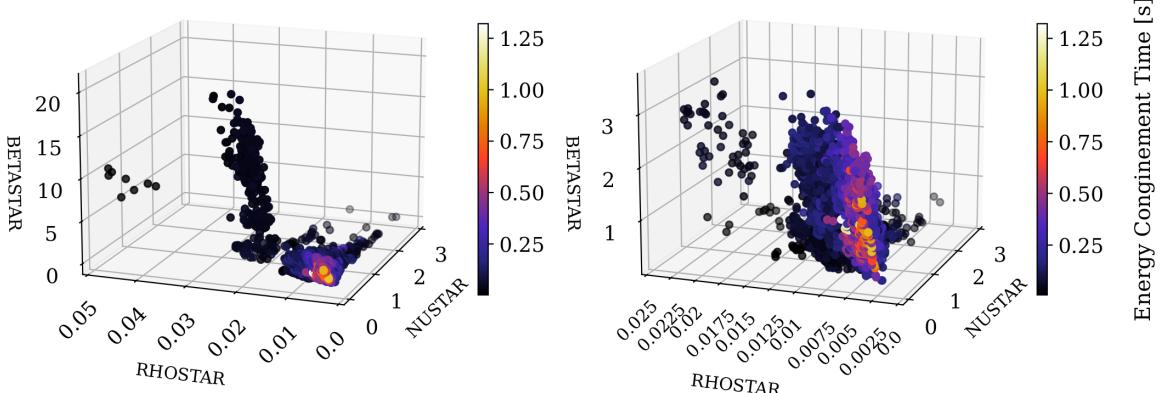
**Table 4.2:** Descriptive statistics for the main variables in STDB5 split into two cases: without spherical tokamaks (left) and spherical tokamaks only (right). This highlights the difference among the ranges in features like the normalised plasma beta, confinement time, and normalised ion collision frequency.

STDB5 ( without   with only ) Spherical Tokamaks					
	TAUTH	RHOSTAR	NUSTAR	BETASTAR	Q95
<b>mean</b>	0.18   0.03	0.01   0.02	0.19   0.07	1.350   11.73	3.83   7.46
<b>std</b>	0.15   0.01	0.0   0.0	0.21   0.02	0.56   3.57	0.88   1.61
<b>min</b>	0.01   0.00	0.00   0.01	0.00   0.02	0.21   4.48	1.87   3.26
<b>25%</b>	0.08   0.03	0.00   0.02	0.06   0.06	0.90   9.06	3.21   6.26
<b>50%</b>	0.13   0.04	0.01   0.02	0.12   0.07	1.250   11.93	3.64   7.48
<b>75%</b>	0.25   0.04	0.01   0.02	0.23   0.09	1.750   14.19	4.25   8.62
<b>max</b>	1.32   0.06	0.02   0.05	2.89   0.13	3.520   20.68	11.13   11.31

Overall, it is evident that spherical tokamaks generally exhibit significantly higher normalised plasma beta values compared to non-spherical devices, as explained in Eq. (4.2). Additionally, it is worth noting that START is the only one capable of achieving  $\rho_* > 0.03$  among the devices considered. This range of observations corresponds to cases that impact  $\alpha_R$  and cases that do not affect  $\alpha_R$ . However, spherical tokamaks cannot achieve large  $\nu_*$  values compared to other devices like ASDEX-U, ASDEX-W, and COMPASS. Nevertheless, how do these tendencies relate to the energy confinement time? Figure 4.7 shows the  $\tau_{E,th}$  variation based on these three parameters.



**Figure 4.6:** 3-dimensional scatter-plot of  $\rho_*$ ,  $\nu_*$ , and  $\beta_t$  showing the observations that influence  $\alpha_R$  (left) and its corresponding tokamak (right).



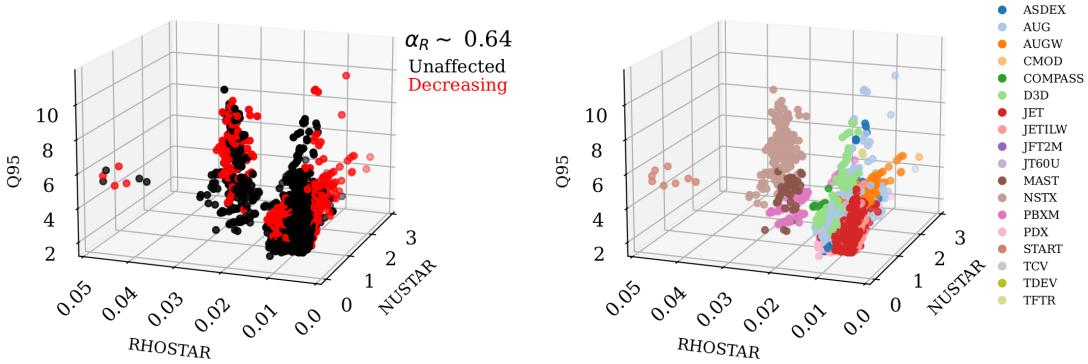
**Figure 4.7:** normalized plasma beta, with normalized ion gyroradius and normalized collisionality plotted along with  $\tau_{E,th}$  for the complete STDB5 (right) and the case where spherical devices has been removed (right).

It is reasonable to expect that a larger value of  $\beta_t$  would lead to a longer  $\tau_{E,th}$ . However, it is interesting that devices with higher  $\beta_t$  exhibit the lowest energy confinement times. However, as already discussed, several variables influence  $\tau_{E,th}$ , some of which can be observed experimentally, while others are theoretically estimated. Additionally, there are unknown variables, commonly referred to as latent variables. Examples of latent variables include plasma rotation and the electron-ion heating ratio [7].

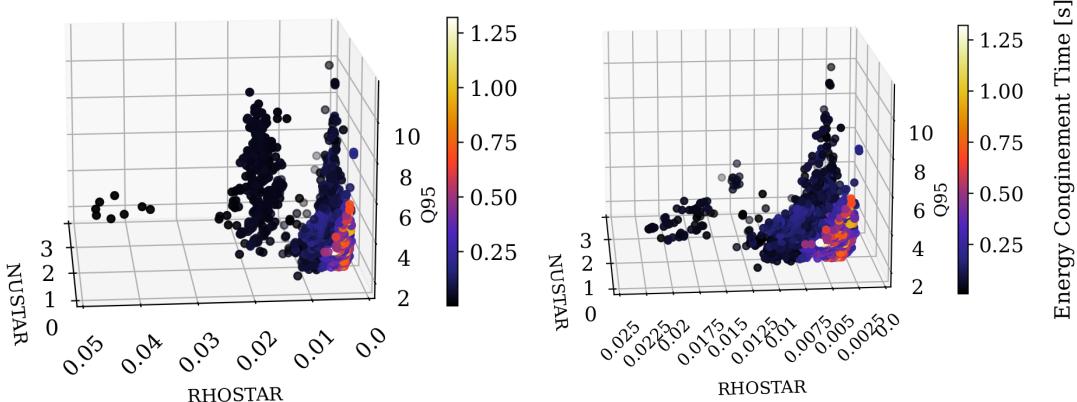
Let us now investigate the three main features in the RF:  $\rho_*$ ,  $\nu_*$ , and  $q_{95}$ . Figure 4.8 shows the 3-dimensional plot for these variables with their labels (left) and their respective tokamak (right).

In contrast to the previous case, when considering  $\beta_t$  instead of  $q_{95}$ , there is no perfect separation between the spherical and toroidal devices. In the vicinity of the cluster containing MAST and NSTX observations, there are toroidal shots that slightly overlap with them, belonging to PBXM and COMPASS.

For completeness, Figure 4.9 illustrates the variation of the energy confinement time given these three parameters, along with an additional azimuthal rotation in the plot.



**Figure 4.8:** 3-dimensional scatter-plot of  $\rho_*$ ,  $\nu_*$ , and  $q_{95}$  showing the observations that influence  $\alpha_R$  (left) and its corresponding tokamak (right).



**Figure 4.9:** Safety factor, with normalised ion gyroradius and normalised collisionality plotted along with  $\tau_{E,th}$  for the complete STDB5 (right) and the case where spherical devices have been removed (right).

#### 4.2.1 Dynamic Analyses with Random Sampling

Now that the key features have been identified, it is possible to retrieve some of the sample subsets that influence  $\alpha_R$ , shown in Figure 2.5 (right), and plot the change of the labelling in these variables as  $\alpha_R$

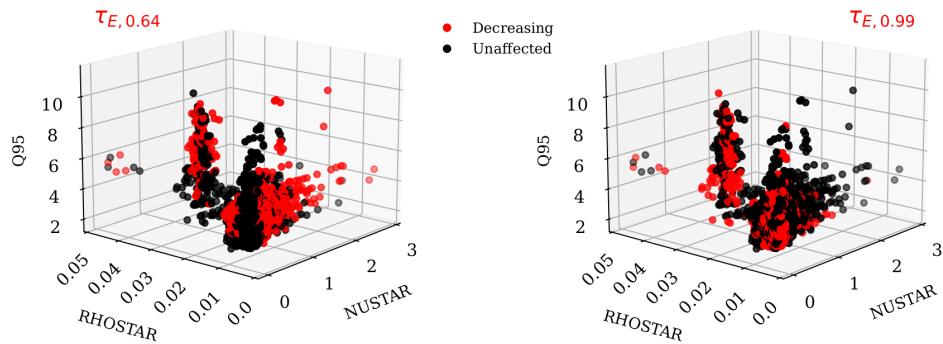
decreases. Figure 4.10<sup>1</sup> shows a comparison between the `big_ds` (left) and the `small_ds` (right) for  $\rho_*$ ,  $\nu_*$ , and  $q_{95}$ . By clicking [this link](#)<sup>2</sup>, you can see the dynamic evolution of the influence on  $\alpha_R$ , as well as the rotation of the plot. From this, it is possible to observe that as  $\alpha_R$  decreases, a clearer separation between the affected and unaffected observations becomes more pronounced.

As shown in Table 4.2, and Figure 4.8, there is a significant difference between the spherical and toroidal devices, so let us have the same analysis by removing the spherical data. Figure 4.11<sup>3</sup> shows the equivalent to Figure 4.10, but with removed spherical data. By clicking [this link](#)<sup>4</sup>, you can see the dynamic evolution of the influence on  $\alpha_R$ , as well as the rotation of the plot. Having both perspectives makes it easier to describe the observations that decrease  $\alpha_R$  in terms of  $\rho_*$ ,  $\nu_*$ , and  $q_{95}$ .

Now, let us perform the same analysis, but we use  $\beta_t$  instead of  $q_{95}$ . Figure 4.12<sup>5</sup> shows the comparison between the labels of the decrease and unaffected observations, now considering  $\beta_t$ . Again, the labelling based on `big_ds` gives a clearer structure to the influencing observations compared to the `small_ds`. Figure 4.13<sup>6</sup> shows the same comparison, with the removal of spherical tokamaks.

One of the questions I asked in Chapter 2 was: "out of all the dataset subsets shown in Figure 2.5 (right), which one will provide the most informative insights into the factors causing the decrease in  $\alpha_R$ ?". It is now evident that, for supervised learning, the dataset with the smallest  $\alpha_R$  obtained provides more information than the smallest subset size containing a significant decrease in  $\alpha_R$ . Furthermore, this is the labelling that creates datasets that comply with the standard assumptions when investigating the  $\omega_i \cdot \tau_{E,th}$  scaling; as shown in the previous section.

There is a branch in machine learning known as *unsupervised* learning. This approach does not give labelled data to the algorithm for it to learn. One example is the search for characteristic clusters within the data. Initially, in this project, I attempted to implement unsupervised learning before supervised learning using all the features of interest. Nevertheless, the results obtained were too complex to interpret. Now that I have gained insights into the main features, it is possible to explore characteristic clusters. In this situation, the differentiation made by the `small_ds` may prove more beneficial than that made by the `big_ds`, primarily due to the former containing fewer observations and thus less noise. However, this aspect requires further investigation.



**Figure 4.10:** Comparison of between the observations that influence  $\alpha_R$  based on `big_ds` (left) and the `small_ds` (right).

<sup>1</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRQ.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRQ.gif)

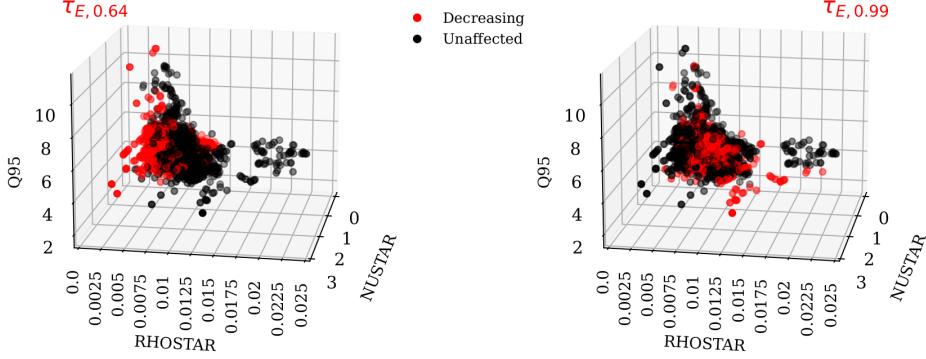
<sup>2</sup>[github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRQ.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRQ.gif)

<sup>3</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRQ\\_noSPHE.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRQ_noSPHE.gif)

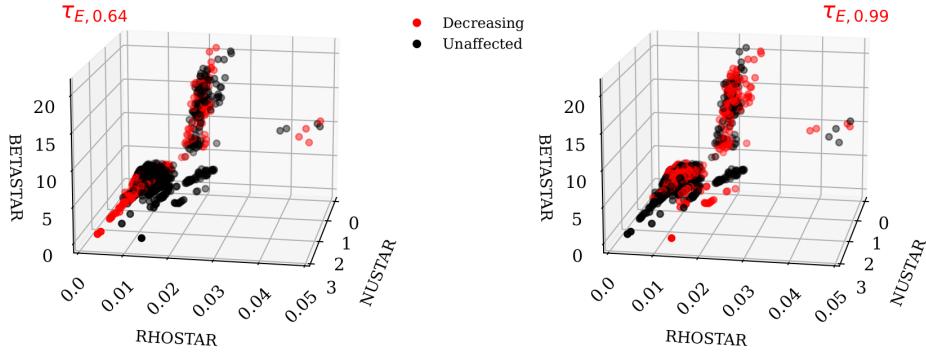
<sup>4</sup>[github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRQ\\_noSPHE.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRQ_noSPHE.gif)

<sup>5</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRB.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRB.gif), and dynamic evolution:[github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRB.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRB.gif)

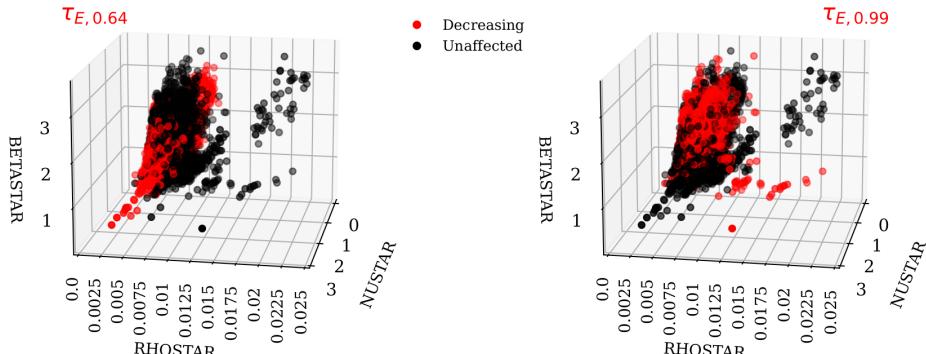
<sup>6</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRB\\_noSPHE.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roCoNRB_noSPHE.gif), and dynamic evolution:[github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRB\\_noSPHE.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roDynNRB_noSPHE.gif)



**Figure 4.11:** Comparison of between the observations that influence  $\alpha_R$  based on `big_ds` (left) and the `small_ds` (right), for  $\rho_*$ ,  $\nu_*$ , and  $q_{95}$ . Here, observations for MAST, NSTX, and START have been removed.



**Figure 4.12:** Comparison of between the observations that influence  $\alpha_R$  based on `big_ds` (left) and the `small_ds` (right), for  $\rho_*$ ,  $\nu_*$ , and  $\beta_t$



**Figure 4.13:** Comparison of between the observations that influence  $\alpha_R$  based on `big_ds` (left) and the `small_ds` (right), for  $\rho_*$ ,  $\nu_*$ , and  $\beta_t$ . Here, observations for MAST, NSTX, and START have been removed.

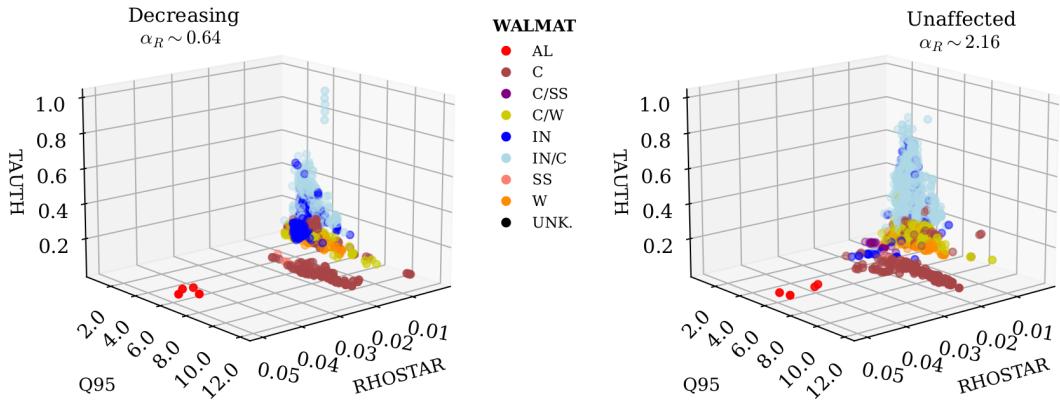
### 4.3 The Engineering Perspective

In 1997, JET achieved fusion plasmas utilising graphite walls and a 1:1 mixture of deuterium and tritium. However, there has been a trend towards replacing carbon walls with metallic walls, commonly referred to as "ITER-like walls." For example, JET's ITER-like walls are composed of Beryllium and tungsten, with the divertor entirely made of tungsten. The motivation for these replacements lies in graphite's issues with erosion and high fuel retention [104]. Although graphite has a higher damage threshold compared to Beryllium, it is not susceptible to chemical erosion and has a low atomic number, which is

crucial for mitigating undesired Bremsstrahlung radiation [1], [104]. Tungsten is incorporated to improve Beryllium's damage threshold as it has a high melting point ( $3,422\text{ }^{\circ}\text{C}$ ) [104]. However, it also has a high atomic number, which increases Bremsstrahlung radiation and leads to substantial losses due to electron-ion collisions [1]. Overall, the design and analysis of plasma-facing components and material science constitute an active research field that continues to influence new projects.

In this section, I will briefly analyse the materials that characterize the decreasing and unaffected observations concerning  $\alpha_R$ . Firstly, let us examine how each case relates to the energy confinement time and the main features found in the previous chapter.

Figure 4.14<sup>7</sup> illustrates the wall materials concerning the energy confinement time, safety factor, and normalised gyroradius. In both cases, the observations involving walls made of Inconel (an alloy of nickel, chromium, and iron [105]), or Inconel combined with Carbon demonstrate higher confinement times. The devices associated with these observations include COMPASS, D3D, JET, and JETILW, with both subsets encompassing data from 1994 to 2016. The divertors of these devices are constructed using various materials. Carbon is utilised in COMPASS, D3D, and JET, with the latter combining it with Beryllium. JETILW employs only Tungsten. It is worth noting that these observations do not exhibit high safety factors compared to others in both datasets. AUG and AUGW employ Tungsten walls, with their divertors composed of Carbon with Tungsten and Tungsten, respectively. CMOD, JFT2M, MAST, PBXM, and PDX feature stainless steel walls. ASDEX, TCV, and TDEV have walls made from a combination of Carbon, while AUG, JT60U, MAST, NSTX, and TFTR have walls composed purely of Carbon. Finally, the isolated observations of Aluminium belong to START, with its divertor made of Inconel. In these registers, the only device that contains Molybdenum in its divertor is CMOD.

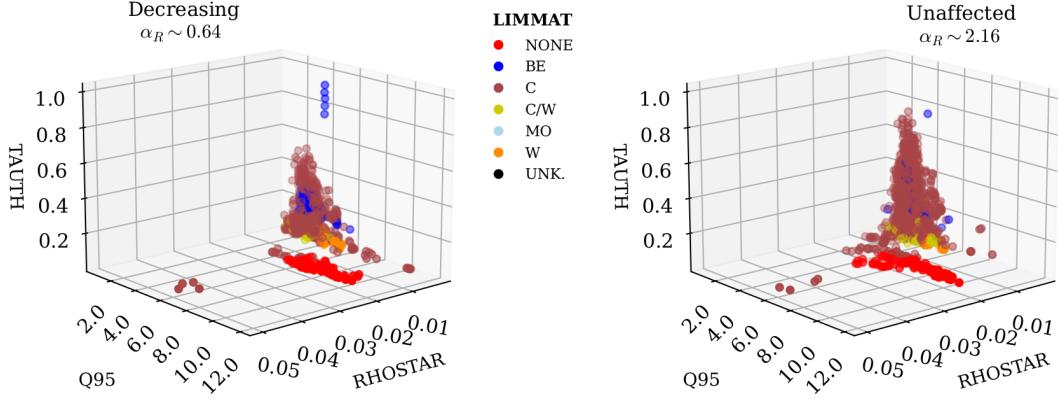


**Figure 4.14:** Wall material present in the decreasing (left) and unaffected (right) dataset based in big\_ds for the two of the main features of interest. The labels are: Aluminium (AL), Carbon (C), Inconel (IN), stainless steel (SS), Tungsten (W), and unknown (UNK.)

In tokamaks, limiters are strategically placed within the vacuum vessel to protect the machine against plasma. However, this concept was later replaced by the invention of divertors [1], [106]. On the other hand, JET incorporates divertors and small wall-guard limiter tiles. These tiles are individually attached to the wall, allowing for easy removal when replacement or inspection is required. Their design also proves useful for researching Tritium accumulation, which is a current concern for ITER [107]. Figure 4.15<sup>8</sup> shows the 3-dimensional plot for  $\tau_{E,th}$ ,  $q_{95}$ , and  $\rho_*$  with their corresponding limiter material.

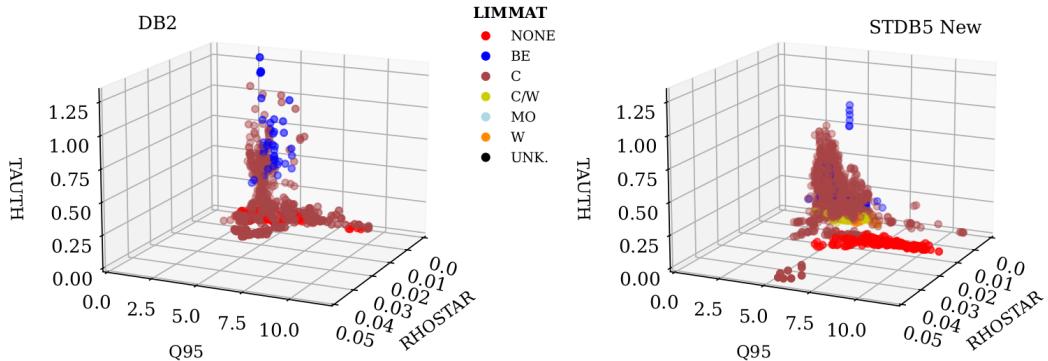
<sup>7</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roWALNRT.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roWALNRT.gif)

<sup>8</sup>Click to see its rotation: [github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roLIMNRT.gif](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/data/GIFs/roLIMNRT.gif)



**Figure 4.15:** Limiter material present in the decreasing (left) and unaffected (right) dataset based in `big_ds` for the two of the main features of interest. The labels are: Beryllium (BE), Carbon (C), Molybdenum (MO), Tungsten (W), and unknown (UNK).

Figure 4.14 along with Figure 4.15 provides a better perspective on the main plasma-facing components that contribute to higher energy confinement time, with  $q_{95} \sim 3$  and low  $\rho_*$ . The highest recorded energy confinement time is 0.93 s, observed in 1990 for JET in the decreasing dataset. However, the highest  $\tau_{E,th}$  recorded in STDB5 is 1.32 s, also in 1990 (coming from DB2) for JET, classified as unaffected; this is depicted in Figure 4.16, where the comparison is made now between datasets.



**Figure 4.16:** Limiter material present in the DB2 (left) and new shots, w.r.t. DB2, in STDB5 (right).

Despite that the devices with W materials do not achieve high energy confinement times compared to the C materials, it is worth noting that W-like devices sustain better density limits compared to C-like devices; this is observed with ASDEX and AUGW in STDB5's Hugill plot, Figure 4.3.

To close this section, it is worth mentioning one of the current hot topics in material science and fusion research, which is the development of limiters and *divertorlets* using Lithium and Gallium, referred to them as "liquid metals" [108]. Implementing these materials will enable the machine to withstand higher heat fluxes while simultaneously enhancing plasma performance [109].

# V

## CONCLUSIONS

The motivation for this project originated from the quest to comprehend the observed decrease in the dependence on the scaling law for energy confinement time in tokamaks relative to the major radius. During the design of the International Thermonuclear Experimental Reactor, the value of the regressor parameter was  $\alpha_R \sim 1.97$ , while the updated analysis now yields a value of  $\alpha_R \sim 1.47$ . Given the significance of this parameter, both in terms of design and anticipated cost, understanding this decrease is of utmost importance. Statistical tests and physical analysis were implemented in the updated database in this project to dissect the issue.

First, the most influential points on the regressor parameter  $\alpha_R$  were searched using typical regression diagnostics and random sampling. The former method helped identify observations that increased  $\alpha_R$  rather than decreased. On the other hand, the random sampling method focused on observations that led to a decrease in  $\alpha_R$ , revealing several subsets that reduced the parameter to 0.64. This revealed that the responsible records that reduced  $\alpha_R$  occupied from 9.88% to 23.5% of the database, depending on the intensity of the decrease. From this analysis, only two subsets, out of 382,495 resultant subsets, were selected for further investigation through statistical tests and physical interpretation. These subsets were referred to as `small_ds` and `big_ds`, whose observations create  $\alpha_R \sim 0.99$  and  $\alpha_R \sim 0.64$ , respectively; the compliment of these observations – the unaffected observations – create  $\alpha_R \sim 1.71$  and  $\alpha_R \sim 2.16$ , respectively.

Afterwards, statistical tests were implemented to assess the variables' multicollinearity in the energy confinement time scaling law. The variance inflation factor indicated that the plasma current and the major radius have the most impact by multicollinearity in all datasets, namely DB2, STDB5, `small_ds`, and `big_ds` (with their resultant labelled subsets). Upon removing this variable, the variance inflation factor decreased significantly for all other variables across all datasets. Furthermore, the condition number, condition index, and variance decomposition analyses were applied to all datasets, confirming the presence of significant multicollinearity caused by most of the parameters in the regressor. Notably, no substantial reduction in interdependence was observed when distinguishing between observations that influence and those that do not influence  $\alpha_R$ . These findings indicate that factors other than data-related issues may contribute to the observed decline in  $\alpha_R$  with the updated database.

To conclude the statistical analysis, the F-test to compare nested models was implemented in all datasets. This test helps to address whether removing a variable will significantly harm a mode. As anticipated, no variables could be removed without adversely affecting the scaling for the DB2 and the unaffected observations derived from `big_ds` and `small_ds`. However, this was not observed for the inverse aspect ratio for STDB5. The decreasing observations in `big_ds` yield a safe removal of the elongation and the toroidal magnetic field. At the same time, the decreasing observations in `small_ds` yield the deletion of the line-averaged electron density is not harmful to the model. None of the scenarios demonstrated a safe removal of the major radius.

The research continues by applying supervised machine learning algorithms to predict observations influencing  $\alpha_R$ . For that, all columns in STDB5 were analysed. Some relevant features within DB5, like the bootstrap current fraction or the neutral pressure in the main chamber, were empty in STDB5. Therefore,

only 47 variables were considered for further investigation. The objective subsequently shifted towards selecting a subset of features for a learning algorithm to predict whether an observation would influence negatively  $\alpha_R$  or not. The decision not to include all variables was based on two primary reasons: (i) the variables exhibited a significant degree of multicollinearity when considered together, and (ii) increasing the number of variables in a learning algorithm makes it more challenging to enhance performance. Therefore, entropy analysis, research, and multicollinearity were the approaches to finding subsets of interest. Then, Gaussian process and random forests were the selected classification algorithms to predict observations. The results yield that random forest performed best with a subset of variables derived from low multicollinearity and high information gain. Thus, this algorithm was utilised to predict new machines, like ITER and SPARC. The algorithm yields that the expected observations on these machines follow a scaling law that does not decrease  $\alpha_R$ . However, the predicted energy confinement time differs considerably from the expected value. The prediction analysis then concludes with a review of the novel applications of machine learning in fusion research.

The features that demonstrated the most impact in  $\alpha_R$  were: the normalized ion gyroradius, the normalized ion collision frequency, the safety factor, and the normalized plasma  $\beta$ . Therefore, the analysis continues revising the theory and how tokamak instabilities relate to these variables, especially the kink instabilities and the density limit. Before analysing the density limit in the datasets, first, the standard assumption with the normalized gyroradius when exploring the scaling of  $\omega_i \cdot \tau_{E,th}$  was revised. This approach identifies whether a dataset follows a Bohm or a gyroBohm scaling. The results yield that almost all datasets follow a Bohm scaling except for DB2 and unaffected records in `big_ds`, which follow a gyroBohm scaling. The decreasing dataset given by `small_ds` was the only case that falls outside the assumption range. Moreover, the decreasing dataset provided by `big_ds` gives the most extensive characteristic turbulence scale length, while DB2 provides the smaller. From this, it was concluded that the classification followed by the `small_ds` lacks representative physics.

After reaching the previous conclusion, the study continued by plotting various Greenwald limits for all machines. There, it was noted that spherical devices might exhibit considerably different limits compared to devices with smaller aspect ratios and that ITER-like devices have relatively similar limits compared to other machines. Furthermore, JET is the only device with stable shots that surpasses the Greenwald limit. When the same analysis is performed in the subsets that decrease and do not affect  $\alpha_R$ , the results show that most of the observations that surpass the Greenwald limit belong to the unaffected classification. Also, the limits from the decreasing observations for the ITER-like devices are almost the same, while the unaffected observations are slightly different. Based on these findings, it can be inferred that employing machine learning algorithms on time-series data from unaffected shots that exceed the Greenwald limit can facilitate the identification of underlying similarities among these shots.

Then, the study continues with the variability in the observation versus prediction of the energy confinement time. The results indicated that STDB5 provided predictions closest to the observed values compared to DB2. Additionally, the decreasing scaling derived from `big_ds` yielded predictions closer to the observations labelled as decreasing, while the unaffected scaling showed better prediction performance for unaffected observations. These findings suggest that distinguishing between observations that influence  $\alpha_R$  can enhance the prediction of the energy confinement time when using ordinary least squares in a scaling law.

The research then proceeded to visualize the main variables in a 3-dimensional and dynamic manner, comparing both spherical and non-spherical tokamaks. These visualizations revealed two significant observations: (i) The decreasing and unaffected labelling generated by `big_ds` demonstrated a more distinct separation than `small_ds`, explaining the lower performance of classifiers when using the `small_ds`. And (ii) spherical tokamaks exhibited prominent clusters in the variables of interest, distinguishing them from non-spherical ones. This finding suggests that conducting the complete analysis again while excluding the spherical tokamaks could potentially improve the predictions of observations that influence  $\alpha_R$ .

and the energy confinement times for new devices like ITER and SPARC.

Finally, the project concludes with a brief analysis of how plasma-facing components relate to the variables of interest and how the decreasing or unaffected labelling characterizes them. This analysis shows that the plasma-facing materials with the largest energy confinement times are classified as decreasing. These components are distinguished by their use of Carbon with Inconel wall material, Beryllium in their limiters, and Beryllium with Carbon in their divertor. These particular observations belong to JET. While ITER-like devices achieve similar values as expected in ITER, they do not reach the same level of energy confinement time as those obtained by JET.

Overall, this project shows how to obtain characteristic observations that influence any regressor parameter given a dataset and the regressor model. For the case of the energy confinement time, many variables relevant to tokamak physics were studied, regardless if they were numerical, categorical, or had missing information. This comprehensive approach is undertaken to acknowledge the significant influence of numerous factors on the plasma state during a plasma discharge, such as the heating system, wall material, pressure chamber, collisionality, and other relevant variables. However, the categorical data had little relevance when predicting the influence of the major radius on the energy confinement time. This project could benefit from spherical data, as less than 5% represents it in STDB5.; this could help decide whether separating the spherical data from the non-spherical data is necessary when further investigating the observations that negatively influence  $\alpha_R$ . And it would also help resolve whether spherical devices should have their own scaling law for the energy confinement time.

This research presents a new approach that could be pursued to enhance the accuracy of predicting the energy confinement time in new devices by distinguishing between observations that decrease the major radius's impact on the energy confinement time. Since the major radius plays a crucial role in determining the performance and cost of a device, this distinction becomes particularly significant. Undoubtedly, more work must be done to refine the predicting algorithms.



# A

## DERIVATION OF RESEARCH VARIABLES

In this chapter, I will provide a step-by-step explanation of how to derive Eq. (3.5) and Eq. (3.6) to gain a deeper understanding of why the *research variables* are of particular interest. The derivation is based on an attempt to estimate the impact of the scaling law on tokamak design, using a zero-dimensional analysis. These mathematical derivations heavily draw upon the ideas and research findings presented in [14], which serves as the primary reference for this section. However, it is important to note that the content presented here is not intended to replicate or mimic the material found in [14]. Instead, the purpose of this derivation is to offer a more accessible and reader-friendly approach, especially for those without a strong background in tokamak physics. Here, the temperature and any other energy are expressed in keV. Furthermore, flat profiles of current and density are considered.

### A Brief Analysis on the Design of a Tokamak

When having the target fusion power  $P_{fus}$  and fusion gain  $Q$ , for instance,  $P_{fus} = 500$  MW and  $Q = 10$ ; as well as prescribed geometric coefficients, it is possible to figure out the optimal major radius  $R_{geo}$  and toroidal magnetic field  $B_t$  for best stability. The first thing one should think about is the possible risks that might terminate the plasma confinement; as already mentioned, one of them is the Greenwald limit  $n_{GW}$ , estimated as:

$$n_{GW} = \frac{10 \cdot I_p}{\pi a^2}.$$

Now, if the length of the poloidal cross-section is defined as  $L_\theta = 2\pi a \cdot \sqrt{(1 + \kappa_a^2)/2}$  it is possible to approximate the plasma current as [14]:

$$I_p = \frac{1}{\mu_0} L_\theta B_\theta, \quad (\text{A.1})$$

where  $B_\theta$  is the poloidal magnetic field and can be obtained from the safety factor  $q_{95}$ . One can use these definitions, and the fact that  $a = \epsilon R_{geo}$ , to have an expression for the plasma current in terms of the toroidal magnetic field; namely,

$$I_p = C_I \cdot \frac{\epsilon^2}{q_{95}} \cdot B_t R_{geo}; \text{ with, } C_I = \frac{2\pi}{\mu_0} \sqrt{(1 + \kappa_a^2)/2} \cdot 10^{-6}. \quad (\text{A.2})$$

With this, it is possible to get an expression for the Greenwald Limit in terms of  $R_{geo}$  and  $B_t$ , instead of  $I_p$  and  $a$ , such that

$$n_{GW} = \frac{10}{\pi} \cdot \frac{C_I \epsilon^2 B_t R_{geo}/q_{95}}{\epsilon^2 R_{geo}^2} = \frac{10}{\pi} C_I \cdot \frac{B_t}{q_{95} R_{geo}}. \quad (\text{A.3})$$

From this, one gets the normalised density  $n_N = \bar{n}_e/n_{GW}$ . Thus,

$$\bar{n}_e = \frac{10}{\pi} C_I n_N \frac{B_t}{q_{95} R_{geo}}. \quad (\text{A.4})$$

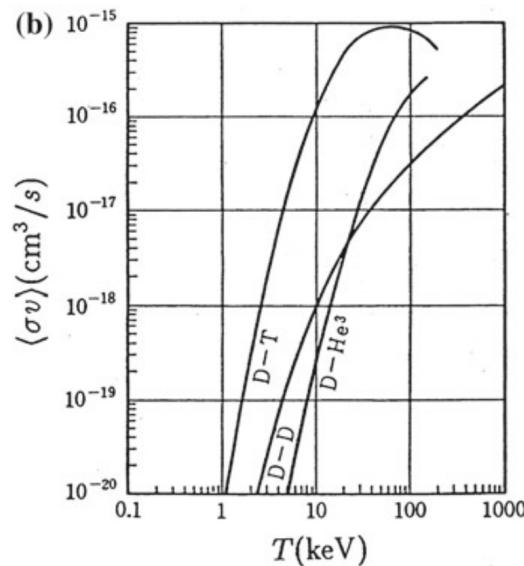
Now, let's discuss MHD instabilities. This is a complex and extensive topic; but, for this section, it is sufficient to understand that these instabilities can occur due to perturbations that cause the displacement of the plasma, such as the *Rayleigh–Taylor interchange instability* or kink instabilities [110]. If the perturbation propagates perpendicular and slightly parallel to the magnetic field, the plasma will exhibit *ballooning instabilities*, which are pressure-driven instabilities that can result in the loss of plasma confinement [1], [110]. One can reduce the chance of observing MHD instabilities when the tokamak complies with the  $\beta$ -limit, as mentioned in Chapter 4. For this analysis, it is convenient to express this limit in the form of a percentage; namely, [14]:

$$\beta\% = 100 \cdot \beta_t \leq \beta_{eq} = g \cdot \frac{I_p}{a B_t}, \quad (\text{A.5})$$

where, and  $g$  is a proportionality constant, often considered between 2-4, due to the Troyon limit [111]. From this, the normalised plasma-beta  $\beta_N$  is defined,

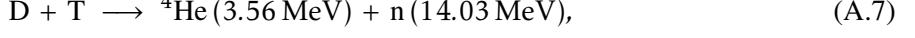
$$\beta\% = \beta_N \cdot \frac{I_p}{a B_t} \quad (\text{A.6})$$

such that the stability limit can be read as  $\beta_N < g$  [14]. Let us now define the type of species that will be in the tokamak. Figure A.1 shows the highest fusion reaction rate  $\langle \sigma v \rangle$  versus the plasma temperature for different elements.



**Figure A.1:** Feasibility of obtaining fusion power depending on the species in the plasma at a given temperature. These lines are obtained when convoluting a Maxwellian distribution with the respective fuel's fusion cross section [1]. The lines shown are for Deuterium-Tritium (D-T), (Deuterium-Deuterium) D-D, and Deuterium-Helium-3 (D-He<sup>3</sup>). Image from: Fig. 1.2 (b) in [110].

Since D-T is more likely to undergo fusion at lower temperatures, compared to other fusion fuels, it is a popular option for nuclear fusion research. Nevertheless, Tritium is an expensive and difficult element to work with [1], which is why some companies prefer to research D-He<sup>3</sup> [112]. ITER will be working with D-T [9]; hence, these are the species that we are considering in this study. Its fusion reaction reads



with a total energy release of  $E_{DT} = 17.59 \text{ MeV}$ , per fusion reaction. The ratio of the total energy transferred by the alpha particles  $\lambda = 17.50/3.56 \approx 4.94$  [14]. Knowing the density of the deuterium and tritium ( $n_D$  and  $n_T$ , respectively), one can approximate the  $P_{fus}$  through Figure A.1 and

$$P_{fus} = n_D n_T \langle \sigma v \rangle_{DT} E_{DT} V_t, \quad (\text{A.8})$$

where  $V_t = 2\pi^2 \kappa_a R_{geo} a^2$  is the volume of the tore [14]. By assuming that  $n_D = n_T = \bar{n}_e/2$ , replacing the expression of  $V_t$ , and approximating  $\langle \sigma v \rangle_{DT} \approx 1.1810 \times 10^{-24} \hat{T} \text{ m}^3 \text{s}^{-1}$  through a polynomial fit for temperatures in the 10.3-18.5 keV range [14], one can get that

$$P_{fus} = C_{fus} \kappa \epsilon^2 R_{geo}^3 (\bar{n}_e \hat{T})^2, \quad (\text{A.9})$$

with  $C_{fus} = 17.59 \cdot q \pi^2 \cdot 1.18 \cdot 10^{-24} \cdot 10^{2.19}/2 \approx 1.63 \cdot 10^{-3}$ . To get the expression of  $P_{fus}$  in terms of  $R_{geo}$  and  $B_t$  instead of the plasma pressure  $\bar{n}_e \hat{T}$ , one can use the definition of  $\beta\%$  and  $\beta_N$ . First, we express  $\beta\%$  in a convenient form, such as:

$$\beta\% = C_\beta \frac{\bar{n}_e \hat{T}}{B_t^2}; \text{ with } C_\beta = 4 \cdot 100 \cdot \mu_0 \cdot 10^{19} \cdot 10^3 q \approx 0.805. \quad (\text{A.10})$$

Therefore,  $\bar{n}_e \hat{T} = C_\beta^{-1} C_I \beta_N B_t^2$ . Now, we use this to re-express Eq. (A.9) as

$$P_{fus} = \frac{C_{fus} C_I^2}{C_\beta^2} \cdot \frac{\kappa_a \epsilon^4}{q_{95}^2} \beta_N^2 R_{geo}^3 B_t^4. \quad (\text{A.11})$$

We have derived Eq. (3.5). To derive Eq. (3.6), it is necessary to think about the power balance in equilibrium at steady state. This means that the total plasma heating sources  $P_{hss}$  is equal to the power loss  $P_{loss}$ ; namely,

$$P_{hss} = P_\Omega + P_\alpha + P_{aux} = P_{loss} = P_{rad} + P_{l,th}. \quad (\text{A.12})$$

Here,  $P_\Omega$  is the ohmic heating (often neglected),  $P_\alpha$  is the energy carried by the alpha particles that are fully contributing to the D-T fuel,  $P_{aux}$  is the auxiliary heating power,  $P_{rad}$  is the radiated power, and  $P_{l,th}$  is the power lost due to transport through the LCFS [14]. In order to have an explicit expression for the radiative losses, one needs to consider specific characteristics of the tokamak, like the plasma facing components [113], [114]. It is for this reason, that is preferable to parametrize the expression of

the radiative losses with a coefficient,  $0 \leq \gamma_{rad} \leq 1$ , as  $P_{rad} = (1 - \gamma_{rad})P_{hss}$ . From this, one gets from the power balance that

$$P_{hss} = (1 - \gamma_{rad})P_{hss} + P_{l,th}, \quad (\text{A.13})$$

meaning,

$$P_{l,th} = \gamma_{rad} (P_\Omega + P_\alpha + P_{aux}). \quad (\text{A.14})$$

If one neglects the ohmic heating, and makes use of  $P_{fus} = \lambda P_\alpha$  and  $Q = P_{fus}/P_{aux}$ , then, one obtains the power loss due to transport in the form of

$$P_{l,th} = \gamma_{rad} P_{fus} \left( \frac{1}{Q} + \frac{1}{\lambda} \right). \quad (\text{A.15})$$

It is now time that we make use of the energy confinement time, which is defined as [110]

$$\tau_{E,th} = \frac{1}{P_{l,th}} \cdot 3V_t \hat{T} \bar{n}_e, \quad (\text{A.16})$$

where it has been assumed that the electron temperature is equal to the ion temperature [14]. By making use of the explicit expression of the torus volume, Eq. (A.9), and Eq. (A.15) one gets

$$\bar{n}_e \hat{T} \tau_{E,th} = 6\pi^2 \left[ \gamma_{rad} C_{fus} \left( \frac{1}{\lambda} + \frac{1}{Q} \right) \right]^{-1}. \quad (\text{A.17})$$

We now take the scaling law of the energy confinement time, shown in Eq. (1.1),

$$\tau_{E,th} = \alpha_0 \cdot I_p^{\alpha_I} \cdot B_t^{\alpha_B} \cdot \bar{n}_e^{\alpha_n} \cdot P_{l,th}^{\alpha_P} \cdot R_{geo}^{\alpha_R} \cdot \kappa_a^{\alpha_\kappa} \cdot \epsilon^{\alpha_\epsilon} \cdot M_{eff}^{\alpha_M},$$

to multiply it with  $\bar{n}_e \hat{T}$ , replace  $I_p$  with Eq. (A.2), and  $P_{l,th}$  with Eq. (A.15); namely,

$$\bar{n}_e \hat{T} \tau_{E,th} = \alpha_0 \cdot \hat{T} \cdot \bar{n}_e^{\alpha_n+1} \left( \frac{C_I}{q_{95}} \epsilon^2 B_t R_{geo} \right)^{\alpha_I} \cdot B_t^{\alpha_B} \cdot R_{geo}^{\alpha_R} \cdot \kappa_a^{\alpha_\kappa} \cdot \epsilon^{\alpha_\epsilon} \cdot M_{eff}^{\alpha_M} \cdot \left[ \gamma_{rad} P_{fus} \left( \frac{1}{\lambda} + \frac{1}{Q} \right) \right]^{\alpha_P} \quad (\text{A.18})$$

which then can be simplified by replacing  $\bar{n}_e$  with  $n_N$ , and  $\bar{n}_e \hat{T}$  with  $\beta_N$  to then equate with Eq. (A.17); such as:

$$[C \cdot \Gamma(Q)]^{-1} = \alpha_0 \cdot R_{geo}^{\gamma_R} \cdot B_t^{\gamma_B} \cdot P_{fus}^{\alpha_P} \cdot M_{eff}^{\alpha_M} \cdot \kappa_a^{\alpha_\kappa} \cdot \epsilon^{\gamma_\epsilon} \cdot n_N^{\alpha_n} \cdot q_{95}^{-\gamma_I} \cdot \beta_N, \quad (\text{A.19})$$

Which is Eq. (3.6). Here, the exponents  $\gamma_x$  are:

$$\begin{aligned}\gamma_I &= 1 + \alpha_n + \alpha_I, & \gamma_R &= \alpha_R + \alpha_I - \alpha_n, \\ \gamma_\epsilon &= 1 + \alpha_\epsilon + 2\alpha_I, & \gamma_B &= \alpha_B + \alpha_n + \alpha_I + 2.\end{aligned}\tag{A.20}$$

Moreover,

$$\Gamma(Q) = \left[ \gamma_{rad} \cdot \left( \frac{1}{\lambda} + \frac{1}{Q} \right) \right]^{\alpha_p+1}, \text{ and, } C = \left( \frac{10}{\pi} \right)^{\alpha_n} \cdot \frac{C_I^{\gamma_I} \cdot C_{fus}}{6\pi^2 C_\beta}. \tag{A.21}$$

It is possible to refine the analysis by considering the alpha-particle dilution, different ion and electron temperatures, or different density profiles [14]. However, for the purpose of this chapter, it is not necessary to implement these changes. By taking into account Eq. (A.11), and Eq. (A.18) it is clear how the relation reduces to an expression involving the engineering research variables, with  $\beta_t$ , the prescribed geometric coefficients and effective mass.



## REFERENCES

- [1] A. Köhn-Seemann, *Fusion research*, Wintersemester, Universität Stuttgart, 2022.
- [2] D. A. Hartmann, “Stellarators,” *Fusion Science and Technology*, vol. 49, no. 2T, pp. 43–55, Feb. 2006. doi: [10.13182/FST06-A1103](https://doi.org/10.13182/FST06-A1103).
- [3] W. Picot, “Magnetic fusion confinement with tokamaks and stellarators,” *Bulletin of the International Atomic Energy Agency*, vol. No. 2/2021, Jun. 2021, IAEA. [Online]. Available: <https://www.iaea.org/bulletin/magnetic-fusion-confinement-with-tokamaks-and-stellarators> (visited on 05/13/2023).
- [4] Lawrence Livermore National Laboratory, *How NIF Works*, [Online], Retrieved June 15, 2023, 2021. [Online]. Available: <https://lasers.llnl.gov/about/how-nif-works>.
- [5] ITER, “What is ITER?,” 2020. [Online]. Available: <https://www.iter.org/proj/inafewlines>.
- [6] IAEA, “ITER conceptual design report,” *International Atomic Energy Agency, Vienna*, 1991. [Online]. Available: [https://inis.iaea.org/collection/NCLCollectionStore/\\_Public/22/064/22064645.pdf](https://inis.iaea.org/collection/NCLCollectionStore/_Public/22/064/22064645.pdf).
- [7] G. Verdoollaeghe *et al.*, “The updated ITPA global H-mode confinement database: Description and analysis,” *Nuclear Fusion*, vol. 61, Jan. 2020. doi: [10.1088/1741-4326/abdb91](https://doi.org/10.1088/1741-4326/abdb91).
- [8] I. P. E. G. on Confinement, Transport, I. P. E. G. on Confinement Modelling, Database, and I. P. B. Editors, “Chapter 2: Plasma confinement and transport,” *Nuclear Fusion*, vol. 39, no. 12, p. 2175, Dec. 1999. doi: [10.1088/0029-5515/39/12/302](https://doi.org/10.1088/0029-5515/39/12/302). [Online]. Available: <https://dx.doi.org/10.1088/0029-5515/39/12/302>.
- [9] IAEA, *ITER Technical Basis*. Vienna: IAEA, 2002. [Online]. Available: <https://www-pub.iaea.org/mtcd/publications/pdf/iter-eda-ds-24.pdf>.
- [10] A. W. Leonard, “Edge-localized-modes in tokamaks,” *Physics of Plasmas*, vol. 21, no. 9, Sep. 2014. doi: [10.1063/1.4894742](https://doi.org/10.1063/1.4894742). [Online]. Available: <https://www.osti.gov/biblio/1352343>.
- [11] R. Barnsley, “Radiation hardness in ITER diagnostics,” 2014, ITER Diagnostics Division. [Online]. Available: <https://www.slideserve.com/cybele/radiation-hardness-in-iter-diagnostics>.
- [12] G. Verdoollaeghe *et al.*, “The updated ITPA global h-mode confinement database: Description and analysis,” *Princeton Plasma Physics Laboratory, Princeton University*, 2021. [Online]. Available: <http://arks.princeton.edu/ark:/88435/dsp01m900nx49h>.
- [13] G. Verdoollaeghe, “Robust estimation of tokamak energy confinement scaling through geodesic least squares regression,” in *26th IAEA Fusion Energy Conference*, 2016. [Online]. Available: <https://nucleus.iaea.org/sites/fusionportal/Shared%20Documents/FEC%5C%202016/fec2016-preprints/preprint0759.pdf>.
- [14] Y. Sarazin *et al.*, “Impact of scaling laws on tokamak reactor dimensioning,” *Nuclear Fusion*, vol. 60, no. 1, p. 016010, Oct. 2019. doi: [10.1088/1741-4326/ab48a5](https://doi.org/10.1088/1741-4326/ab48a5). [Online]. Available: <https://dx.doi.org/10.1088/1741-4326/ab48a5>.
- [15] J. Bakarji, J. Callahan, S. L. Brunton, and J. N. Kutz, “Dimensionally consistent learning with buckingham pi,” *arXiv preprint arXiv:2202.04643*, 2022. [Online]. Available: <https://arxiv.org/abs/2202.04643>.
- [16] V. Ostuni, “Impact de la composition du plasma et du rapport d’aspect sur les performances globales du tokamak,” *Doctoral School of Physics and Material Sciences (Marseille)*, 2022, thèse de doctorat, in partnership with CEA Cadarache (Bouches-du-Rhône). [Online]. Available: <https://www.theses.fr/2022AIXM0345>.

- [17] S.-E. Skjelbred, *Introductory econometrics*, Summer Semester, University of Oslo, 2015. [Online]. Available: <https://www.uio.no/studier/emner/sv/økonomi/ECON4150/v15/lectures/ols-assumptions.pdf>.
- [18] A. Murari, E. Peluso, P. Gaudio, and M. Gelfusa, “Robust scaling laws for energy confinement time, including radiated fraction, in tokamaks,” *Nuclear Fusion*, vol. 57, no. 12, p. 126 017, Sep. 2017. doi: [10.1088/1741-4326/aa7bb4](https://doi.org/10.1088/1741-4326/aa7bb4). [Online]. Available: <https://dx.doi.org/10.1088/1741-4326/aa7bb4>.
- [19] O. Kardaun, “The tortuous route of confinement prediction near operational boundary - improvement of analysis based on iterhdb4/ldb3 database,” International Atomic Energy Agency (IAEA), IAEA-CN-149 IAEA-CN-149, 2006, p. 142. [Online]. Available: [http://www-pub.iaea.org/MTCD/Meetings/PDFplus/2006/cn149\\_BookOfAbstracts.pdf](http://www-pub.iaea.org/MTCD/Meetings/PDFplus/2006/cn149_BookOfAbstracts.pdf) (visited on 04/01/2023).
- [20] F. M. Laggner, “Inter-elm pedestal structure development in asdex upgrade,” Ph.D. dissertation, Technische Universit at Wien, Wien, Jun. 2017. [Online]. Available: [https://pure.mpg.de/rest/items/item\\_2452969\\_5/component/file\\_2452970/content](https://pure.mpg.de/rest/items/item_2452969_5/component/file_2452970/content).
- [21] U. Stroth *et al.*, “Progress from asdex upgrade experiments in preparing the physics basis of iter operation and demo scenario development,” *Nuclear Fusion*, vol. 62, no. 4, p. 042 006, 2022. doi: [10.1088/1741-4326/ac207f](https://doi.org/10.1088/1741-4326/ac207f).
- [22] A. Loarte *et al.*, “Characteristics of type i elm energy and particle losses in existing devices and their extrapolation to iter,” *Plasma Physics and Controlled Fusion*, vol. 45, no. 9, p. 1549, Aug. 2003. doi: [10.1088/0741-3335/45/9/302](https://doi.org/10.1088/0741-3335/45/9/302).
- [23] R. Pitts *et al.*, “Physics conclusions in support of iter w divertor monoblock shaping,” *Nuclear Materials and Energy*, vol. 12, pp. 60–74, 2017, Proceedings of the 22nd International Conference on Plasma Surface Interactions 2016, 22nd PSI. doi: <https://doi.org/10.1016/j.nme.2017.03.005>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352179116302885>.
- [24] M. Endler, I. Garc ía-Cort é s, C. Hidalgo, G. F. Matthews, A. Team, and J. Team, “The fine structure of elms in the scrape-off layer,” *Plasma Physics and Controlled Fusion*, vol. 47, no. 2, p. 219, Jan. 2005. doi: [10.1088/0741-3335/47/2/002](https://doi.org/10.1088/0741-3335/47/2/002).
- [25] A. Murari *et al.*, “Extensive statistical analysis of elms on jet with a carbon wall,” *Plasma Physics and Controlled Fusion*, vol. 56, no. 11, p. 114 007, 2014. doi: [10.1088/0741-3335/56/11/114007](https://doi.org/10.1088/0741-3335/56/11/114007).
- [26] V. Solokha, G. Kurskiev, A. Yashin, *et al.*, “Classification of Edge Instabilities at Globus-M2 Tokamak,” *Plasma Physics Reports*, vol. 49, no. 4, pp. 419–427, Apr. 2023. doi: [10.1134/S1063780X23600184](https://doi.org/10.1134/S1063780X23600184).
- [27] O. Vallhagen, I. Pusztai, M. Hoppe, S. Newton, and T. F ü l ö p, “Effect of two-stage shattered pellet injection on tokamak disruptions,” *Nuclear Fusion*, vol. 62, no. 11, p. 112 004, Nov. 2022. doi: [10.1088/1741-4326/ac667e](https://doi.org/10.1088/1741-4326/ac667e).
- [28] E. Villalobos, B. Bauyrzhan, and L. Steyn, “Scaling law: DB5 vs DB2.8,” *Universiteit Gent*, 2022.
- [29] D. Belsley, E. Kuh, and R. Welsch, *Regression Diagnostics*. Canada: Wiley-Interscience, 2004.
- [30] A. Lindholm, N. Wahlstr ö m, F. Lindsten, and T. B. Sch ö n, *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. [Online]. Available: <https://smlbook.org>.
- [31] F. A. Viernes, “Linear regression models and influential points,” *Towards Data Science*, Mar. 2021. [Online]. Available: <https://towardsdatascience.com/linear-regression-models-and-influential-points-4ee844adac6d>.
- [32] C. Shalizi, “Outliers and influential points,” *Carnegie Mellon University*, 2015, Statistics Data Science. [Online]. Available: <https://www.stat.cmu.edu/~larry/=stat401/lecture-20.pdf>.
- [33] J. Goldstein-Greenwood, “Detecting influential points in regression with DFBETA(S),” *University of Virginia Library*, Jul. 2022, Research Data Scientist. [Online]. Available: <https://data.library.virginia.edu/detecting-influential-points-in-regression-with-dfbetas/>.

- [34] K. Chiñas-Fuentes, *Regression analysis for decreased db5*, Accessed: 14 April, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Optimization/Diagnostics/Regression\\_Analysis\\_for\\_Decreased\\_DB5\\_LATEX\\_IMAGES.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Optimization/Diagnostics/Regression_Analysis_for_Decreased_DB5_LATEX_IMAGES.ipynb).
- [35] A. F. Siegel, *Practical Business Statistics*, Eighth. Academic Press, 2022. doi: <https://doi.org/10.1016/C2019-0-00330-5>.
- [36] N. Shrestha, “Detecting multicollinearity in regression analysis,” *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39–42, 2020. doi: [10.12691/ajams-8-2-1](https://doi.org/10.12691/ajams-8-2-1).
- [37] J. Cheng, J. Sun, K. Yao, M. Xu, and Y. Cao, “A variable selection method based on mutual information and variance inflation factor,” *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 268, p. 120652, 2022. doi: <https://doi.org/10.1016/j.saa.2021.120652>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386142521012294>.
- [38] R. M. O’Brien, “A caution regarding rules of thumb for variance inflation factors,” *Qualitative & Quantitative*, vol. 41, pp. 673–690, Oct. 2007, Published on March 13, 2007. doi: [10.1007/s11135-006-9018-6](https://doi.org/10.1007/s11135-006-9018-6).
- [39] D. Liao and R. Valliant, “Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data,” *Survey Methodology*, vol. 38, pp. 189–202, Jan. 2012.
- [40] J. H. Kim, “Multicollinearity and misleading statistical results,” *Korean Journal of Anesthesiology*, vol. 72, no. 6, pp. 558–569, Jul. 2019, Published online on July 15, 2019. doi: [10.4097/kja.19087](https://doi.org/10.4097/kja.19087).
- [41] S. Šašić, T. Veriotti, T. Kotecki, and S. Austin, “Comparing the predictions by nir spectroscopy based multivariate models for distillation fractions of crude oils by f-test,” *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 286, p. 122023, 2023. doi: <https://doi.org/10.1016/j.saa.2022.122023>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386142522011714>.
- [42] M. Inlow, *Applied regression and time series analysis, f-tests and nested models*, Rose-Hulman Institute of Technology, Feb. 2014. [Online]. Available: <https://www.rose-hulman.edu/class/math/inlow/Math485/ftests.pdf>.
- [43] T. Haslwanter, “An introduction to statistics with python, with applications in the life sciences,” *Switzerland: Springer International Publishing*, 2016. [Online]. Available: <http://www.springer.com/de/book/9783319283159>.
- [44] I. Wallach, M. Dzamba, and A. Heifets, *Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery*, 2015. arXiv: [1510.02855 \[cs.LG\]](https://arxiv.org/abs/1510.02855).
- [45] S. Dieleman, K. W. Willett, and J. Dambre, “Rotation-invariant convolutional neural networks for galaxy morphology prediction,” *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, Apr. 2015. doi: [10.1093/mnras/stv632](https://doi.org/10.1093/mnras/stv632). eprint: <https://academic.oup.com/mnras/article-pdf/450/2/1441/3022697/stv632.pdf>. [Online]. Available: <https://doi.org/10.1093/mnras/stv632>.
- [46] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick, “Detection of prodromal alzheimer’s disease via pattern classification of magnetic resonance imaging,” *Neurobiology of aging*, vol. 29, no. 4, pp. 514–523, 2008. doi: [10.1016/j.neurobiolaging.2006.11.010](https://doi.org/10.1016/j.neurobiolaging.2006.11.010).
- [47] J. Brownlee, *Why one-hot encode data in machine learning?* MachineLearningMastery.com, Jul. 2017. [Online]. Available: <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>.
- [48] L. Chen, “Curse of dimensionality,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 545–546. doi: [10.1007/978-0-387-39940-9\\_133](https://doi.org/10.1007/978-0-387-39940-9_133). [Online]. Available: [https://doi.org/10.1007/978-0-387-39940-9\\_133](https://doi.org/10.1007/978-0-387-39940-9_133).
- [49] M. Dash and H. Liu, “Feature selection for clustering,” *Arizona State University*, Jun. 2020. [Online]. Available: <https://www.public.asu.edu/~huanliu/papers/pakdd00clu.pdf>.

- [50] V. B. Vaghela, K. H. Vandra, and N. K. Modi, “Entropy based feature selection for multi-relational naïve bayesian classifier,” *Journal of International Technology and Information Management*, vol. 23, no. 1, 2014. doi: [10.58729/1941-6679.1062](https://doi.org/10.58729/1941-6679.1062). [Online]. Available: <https://scholarworks.lib.csusb.edu/jitim/vol23/iss1/2>.
- [51] D. J. MacKay, *Information theory, inference and learning algorithms*, Fourth. Cambridge university press, 2014. [Online]. Available: <http://www.inference.org.uk/mackay/itila/>.
- [52] J. Brownlee, *A gentle introduction to information entropy - machinelearningmastery.com*, MachineLearningMastery.com website, Retrieved June 1, 2023, 2020. [Online]. Available: <https://machinelearningmastery.com/what-is-information-entropy/>.
- [53] S. Vajapeyam, “Understanding shannon’s entropy metric for information,” *Arxiv*, Mar. 2014. [Online]. Available: <https://doi.org/10.48550/arXiv.1405.2061>.
- [54] K. Chiñas-Fuentes, *Entropy for all variables*, Accessed: 26 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Entropy/All\\_Variables.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Entropy/All_Variables.ipynb).
- [55] K. Chiñas-Fuentes, N. Das, J. Gallego, D. Kirhetov, L. Steyn, and P. Manas, “Dimensioning of a tokamak,” *IRFM CEA Cadarache*, 2023.
- [56] A. Adadi, “A survey on data-efficient algorithms in big data era,” *Journal of Big Data*, vol. 8, no. 1, p. 24, 2021. doi: [10.1186/s40537-021-00419-9](https://doi.org/10.1186/s40537-021-00419-9).
- [57] M. S. Rahman and H. Reza, “A systematic review towards big data analytics in social media,” *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 228–244, 2022. doi: [10.26599/BDMA.2022.9020009](https://doi.org/10.26599/BDMA.2022.9020009).
- [58] J. Brownlee, *Step-by-step framework for imbalanced classification projects - machinelearningmastery.com*, Accessed May 07, 2023, 2020. [Online]. Available: <https://machinelearningmastery.com/framework-for-imbalanced-classification-projects/>.
- [59] NVIDIA Corporation, “CuML 23.04.00 documentation,” RAPIDS Documentation and Resources. [Online]. Available: <https://docs.rapids.ai/api/cuml/stable/#>.
- [60] L. Buitinck *et al.*, “API design for machine learning software: Experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.
- [61] T. Dhaene, *Machine learning, gaussian processes*, Faculty of Engineering and Architecture. Winter Semester. Lecture 8., Universiteit Gent, 2022.
- [62] *Visualizing a multivariate gaussian*, HashPi, Retrieved June 4, 2023, 2021. [Online]. Available: <https://www.hashpi.com/visualizing-a-multivariate-gaussian>.
- [63] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006. [Online]. Available: [www.GaussianProcess.org/gpml](http://www.GaussianProcess.org/gpml).
- [64] D. K. Duvenaud, “Automatic model construction with gaussian processes,” Ph.D. dissertation, University of Cambridge, Jun. 2014. [Online]. Available: <https://www.cs.toronto.edu/~duvenaud/thesis.pdf>.
- [65] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, Massachusetts; London, England: MIT Press, 2012.
- [66] D. J. Rosevear and A. de Waal, “Gaussian processes applied to class-imbalanced datasets,” in *Proceedings of the 59th Annual Conference of SASA*, 2017, pp. 41–48. [Online]. Available: <https://www.researchgate.net/publication/321533851>.
- [67] J. Brownlee, *A gentle introduction to maximum a posteriori (map) for machine learning - machinelearningmastery.com*, MachineLearningMastery.com, [Accessed: May 25, 2023], Nov. 2019. [Online]. Available: <https://machinelearningmastery.com/maximum-a-posteriori-estimation/>.

- [68] scikit-learn, *scikit-learn.gaussian\_process.GaussianProcessClassifier*, Accessed May 15, 2023, 2015. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.gaussian\\_process.GaussianProcessClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessClassifier.html).
- [69] S. Dutta, *Decision Tree Classification Clearly Explained!* YouTube video, Accessed: May 19, 2023, Jan. 2021. [Online]. Available: <https://youtu.be/ZVR2Way4nwQ>.
- [70] L. Breiman, *Classification and Regression Trees*, 1st. Routledge, 1984. [Online]. Available: <https://doi.org/10.1201/9781315139470>.
- [71] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986. doi: [10.1007/BF00116251](https://doi.org/10.1007/BF00116251). [Online]. Available: <https://doi.org/10.1007/BF00116251>.
- [72] S. L. Salzberg, “C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993,” *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994. doi: [10.1007/BF00993309](https://doi.org/10.1007/BF00993309). [Online]. Available: <https://doi.org/10.1007/BF00993309>.
- [73] T. Dhaene, *Machine learning, decision trees and random forests*, Faculty of Engineering and Architecture. Winter Semester. Lecture 9., Universiteit Gent, 2022.
- [74] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [75] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.
- [76] *Permutation feature importance*, scikit-learn website, 2023. [Online]. Available: [https://scikit-learn.org/stable/modules/permuation\\_importance.html](https://scikit-learn.org/stable/modules/permuation_importance.html).
- [77] scikit-learn, *Permutation feature importance*, 2023. [Online]. Available: [https://scikit-learn.org/stable/modules/permuation\\_importance.html](https://scikit-learn.org/stable/modules/permuation_importance.html).
- [78] K. Chiñas-Fuentes, *Random forest and gaussian process for three different subsets*, Accessed: 26 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Classification/Sklearn\\_Tuning/ML\\_Classifiers\\_in\\_big\\_ds.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Classification/Sklearn_Tuning/ML_Classifiers_in_big_ds.ipynb).
- [79] K. Chiñas-Fuentes, *Best models*, Accessed: 26 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Classification/Sklearn\\_Tuning/Best\\_Models.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Classification/Sklearn_Tuning/Best_Models.ipynb).
- [80] D. Chandler, *MIT-designed project achieves major advance toward fusion energy*, Accessed on June 12, 2023, Sep. 2021. [Online]. Available: <https://news.mit.edu/2021/MIT-CFS-major-advance-toward-fusion-energy-0908>.
- [81] A. J. Creely *et al.*, “Overview of the sparc tokamak,” *Journal of Plasma Physics*, vol. 86, no. 5, p. 865 860 502, 2020. doi: [10.1017/S0022377820001257](https://doi.org/10.1017/S0022377820001257).
- [82] C. Rea, *Machine learning for magnetic fusion science part ii*, Joint ICTP-IAEA School on AI for Nuclear, Plasma and Fusion Science, Trieste, May 2023. [Online]. Available: <https://indico.ictp.it/event/10174/session/30/contribution/55/material/slides/0.pdf>.
- [83] A. H. Boozer, “Theory of tokamak disruptions,” *Physics of Plasmas*, vol. 19, no. 5, pp. 058 101–058 101, 2012. doi: [10.1063/1.3703327](https://doi.org/10.1063/1.3703327).
- [84] W. Hu, J. Hou, Z. Luo, J. Li, and et al., “Prediction of MARFE movement in density-limit disruptive plasmas on EAST using Random Forest,” *Chinese Physics B*, May 2023. doi: [10.1088/1674-1056/acd2b0](https://doi.org/10.1088/1674-1056/acd2b0).
- [85] J. Brownlee, “What is deep learning?” *Machine Learning Mastery*, Aug. 2020. [Online]. Available: <https://machinelearningmastery.com/what-is-deep-learning/>.
- [86] A. Murari, E. Peluso, M. Lungaroni, R. Rossi, and M. Gelfusa, “Investigating the physics of tokamak global stability with interpretable machine learning tools,” *Applied Sciences*, vol. 10, no. 19, p. 6683, Sep. 2020. doi: [10.3390/app10196683](https://doi.org/10.3390/app10196683). [Online]. Available: <http://dx.doi.org/10.3390/app10196683>.

- [87] J. Zhu, C. Rea, K. Montes, R. Granetz, R. Sweeney, and R. Tinguely, “Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks,” *Nuclear Fusion*, vol. 61, no. 2, p. 026 007, Dec. 2020. doi: [10.1088/1741-4326/abc664](https://doi.org/10.1088/1741-4326/abc664). [Online]. Available: <https://dx.doi.org/10.1088/1741-4326/abc664>.
- [88] JET Contributors *et al.*, “Disruption prediction with artificial intelligence techniques in tokamak plasmas,” English, *Nature Physics*, vol. 18, no. 7, pp. 741–750, 2022, Publisher Copyright: © 2022, Springer Nature Limited. doi: [10.1038/s41567-022-01602-2](https://doi.org/10.1038/s41567-022-01602-2).
- [89] J. Zhu *et al.*, “Scenario adaptive disruption prediction study for next generation burning-plasma tokamaks,” *Nuclear Fusion*, vol. 61, no. 11, p. 114 005, Oct. 2021, Published 7 October 2021. doi: [10.1088/1741-4326/ac28ae](https://doi.org/10.1088/1741-4326/ac28ae).
- [90] A. Mathews, M. Francisquez, J. W. Hughes, D. R. Hatch, B. Zhu, and B. N. Rogers, “Uncovering turbulent plasma dynamics via deep learning from partial observations,” *Phys. Rev. E*, vol. 104, p. 025 205, 2 Aug. 2021. doi: [10.1103/PhysRevE.104.025205](https://doi.org/10.1103/PhysRevE.104.025205). [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.104.025205>.
- [91] P. Rodriguez-Fernandez, N. Howard, and J. Candy, “Nonlinear gyrokinetic predictions of sparce burning plasma profiles enabled by surrogate modeling,” *Nuclear Fusion*, vol. 62, no. 7, p. 076 036, May 2022, Published 13 May 2022. doi: [10.1088/1741-4326/ac64b2](https://doi.org/10.1088/1741-4326/ac64b2).
- [92] D. Humphreys, A. Kupresanin, M. Boyer, and et al., “Advancing fusion with machine learning research needs workshop report,” *Journal of Fusion Energy*, vol. 39, no. 1, pp. 123–155, Sep. 2020. doi: [10.1007/s10894-020-00258-1](https://doi.org/10.1007/s10894-020-00258-1). [Online]. Available: <https://doi.org/10.1007/s10894-020-00258-1>.
- [93] H. Zohm, *Magnetohydrodynamic stability of tokamaks*. John Wiley & Sons, 2015.
- [94] D. Morozov, “Greenwald density limit and power balance in tokamaks,” *Journal of Physics: Conference Series*, vol. 941, p. 012 009, Dec. 2017. doi: [10.1088/1742-6596/941/1/012009](https://doi.org/10.1088/1742-6596/941/1/012009).
- [95] M. Greenwald, “Density limits in toroidal plasmas,” *Plasma Physics and Controlled Fusion*, vol. 44, 2002.
- [96] A. Tanga *et al.*, “The control of plasma parameters to avoid high current disruptions in JET,” in *IEEE Thirteenth Symposium on Fusion Engineering*, 1989, 1271–1273 vol.2. doi: [10.1109/FUSION.1989.102444](https://doi.org/10.1109/FUSION.1989.102444).
- [97] H. Q. Liu *et al.*, “Operational region and sawteeth oscillation in the east tokamak,” *Plasma Physics and Controlled Fusion*, vol. 49, no. 7, p. 995, Jun. 2007. doi: [10.1088/0741-3335/49/7/004](https://doi.org/10.1088/0741-3335/49/7/004).
- [98] ANS-Nuclear-Cafe, “EPFL researchers update fusion’s “Greenwald limit”,” *NuclearNewswire*, Jun. 2022. [Online]. Available: <https://www.ans.org/news/article-4026/epfl-researchers-update-fusions-greenwald-limit/>.
- [99] P. J. Catto, “Practical gyrokinetics,” *Journal of Plasma Physics*, vol. 85, no. 3, p. 925 850 301, 2019. doi: [10.1017/S002237781900031X](https://doi.org/10.1017/S002237781900031X).
- [100] J. E. Maggs, T. A. Carter, and R. J. Taylor, “Transition from bohm to classical diffusion due to edge rotation of a cylindrical plasma,” *Physics of Plasmas*, vol. 14, no. 5, p. 052 507, 2007. doi: [10.1063/1.2722302](https://doi.org/10.1063/1.2722302).
- [101] K. Chiñas-Fuentes, *Density limit analyses: Hugill plots*, Accessed: 26 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Defence/HugillPlot\\_STDB5.html](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Defence/HugillPlot_STDB5.html).
- [102] K. Chiñas-Fuentes, *Hugill plots*, GitHub, Accessed: 29 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Tokamak\\_Interpretation/HugillPlots.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Tokamak_Interpretation/HugillPlots.ipynb).
- [103] K. Chiñas-Fuentes, *Analysing the variability in the energy confinement time based on classification*, Accessed: 26 June, 2023. [Online]. Available: [https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Tokamak\\_Interpretation/Tau\\_Eth.ipynb](https://github.com/Chinnasf/FUSION-EP-Master-Thesis/blob/main/Tokamak_Interpretation/Tau_Eth.ipynb).

- [104] S. Brezinsek, *Jet iter-like wall*, Fz-juelich.de, Retrieved June 10, 2023, Sep. 2022. [Online]. Available: <https://www.fz-juelich.de/en/iek/iek-4/forschung/jet-iter-like-wall>.
- [105] Collins Dictionary, *Definition of Inconel*, Collinsdictionary.com, Retrieved June 11, 2023, Jun. 2023. [Online]. Available: <https://www.collinsdictionary.com/dictionary/english/inconel>.
- [106] Max-Planck-Institut für Plasmaphysik, *Limiter*, Ipp.mpg.de, Retrieved June 11, 2023, 2022. [Online]. Available: <https://www.ipp.mpg.de/1766018>.
- [107] C. Pardanaud *et al.*, “D retention and material defects probed using raman microscopy in JET limiter samples and beryllium-based synthesized samples,” *Physica Scripta*, vol. 96, no. 12, p. 124 031, Sep. 2021. doi: [10.1088/1402-4896/ac267b](https://doi.org/10.1088/1402-4896/ac267b). [Online]. Available: <https://dx.doi.org/10.1088/1402-4896/ac267b>.
- [108] J.-C. Yang *et al.*, “Experimental investigation of the flowing lithium limiter. part 1. the spreading characteristics of lithium on solid substrate without an external magnetic field,” *Fusion Engineering and Design*, vol. 189, p. 113 489, 2023. doi: <https://doi.org/10.1016/j.fusengdes.2023.113489>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092037962300073X>.
- [109] A. Fisher, Z. Sun, and E. Kolemen, “Liquid metal “divertorlets” concept for fusion reactors,” *Nuclear Materials and Energy*, vol. 25, p. 100 855, 2020. doi: <https://doi.org/10.1016/j.nme.2020.100855>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352179120301228>.
- [110] K. Miyamoto, *Plasma Physics and Controlled Nuclear Fusion* (Springer Series on Atomic, Optical, and Plasma Physics, 38), eng, 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg : Imprint: Springer, 2005.
- [111] F. Troyon, R. Gruber, H. Saurenmann, S. Semenzato, and S. Succi, “Mhd-limits to plasma confinement,” *Plasma Physics and Controlled Fusion*, vol. 26, Jan. 1984. doi: [10.1088/0741-3335/26/1A/319](https://doi.org/10.1088/0741-3335/26/1A/319).
- [112] Helion, “Fusion energy,” 2022. [Online]. Available: <https://www.helionenergy.com/fusion-energy/>.
- [113] T. Pütterich, E. Fable, R. Dux, M. O’Mullane, R. Neu, and M. Siccinio, “Determination of the tolerable impurity concentrations in a fusion reactor using a consistent set of cooling factors,” *Nuclear Fusion*, vol. 59, no. 5, p. 056 013, 2019. doi: [10.1088/1741-4326/ab0384](https://doi.org/10.1088/1741-4326/ab0384).
- [114] D. Reiter, G. Wolf, and H. Kever, “Burn condition, helium particle confinement and exhaust efficiency\*,” *Nuclear Fusion*, vol. 30, no. 10, p. 2141, 1990. doi: [10.1088/0029-5515/30/10/012](https://doi.org/10.1088/0029-5515/30/10/012).