Identification of Viable Dissolved Gas Analysis Subsets for Power Transformers

José Pedro Ribeiro Ferreira Pinto
Mestrado em Ciência de Dados
Departamento de Ciência de Computadores
2021

Orientador
Ricardo Teixeira Sousa, Investigador Auxiliar, LIAAD - INESC TEC – Laboratório de Inteligência Artificial e Apoio à Decisão – Instituto de Engenharia De Sistemas e Computadores, Tecnologia e Ciência

Coorientador
Alípio Mário Guedes Jorge, Professor Associado, Faculdade de Ciências da Universidade do Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

O trabalho coletivo e o apoio de inúmeras pessoas foi indispensável para desenvolver as tarefas necessárias para esta dissertação. Para o mostrar, estendo o meu agradecimento às seguintes pessoas.

Em primeiro lugar ao meu orientador Ricardo Sousa, que me guiou e acompanhou durante todo o processo, mostrando uma disponibilidade exemplar, especialmente neste período difícil durante o qual o trabalho foi desenvolvido.

Ao meu coorientador Alípio Jorge, que não só foi responsável pelo projeto e me ajudou no seu desenvolvimento, mas também me motivou como professor e me ajudou nos momentos em que mais precisei.

Ao meu colega Vítor Esteves, com quem trabalhei de perto nas tarefas em que este projeto se enquadra, nas quais nos mutuamente ajudamos.

À equipa do INESC TEC, que ajudou a definir os planos de trabalho e esclareceu alguns dos detalhes técnicos.

À equipa da Efacec, em particular ao Sérgio Tavares, que disponibilizou os dados, ajudou na sua interpretação e mostrou total disponibilidade para esclarecer quaisquer dúvidas e guiar o processo.

Por fim, e possivelmente mais importante, aos meus pais, que sempre me encorajaram, promoveram a curiosidade em mim que me fez seguir esta área e me apoiaram em todos os sentidos. Sem eles definitivamente esta tese nunca teria sido uma possibilidade.

Abstract

One of the key components of power systems is the Power Transformer (PT), being paramount for the functioning of modern electrical grids and, as such, always under constant usage for current industrialization activities. This permanent usage of the asset, coupled with higher than rated loads and cost reduction activities greatly increase failure rates and as a consequence can reduce the PT's lifespan. Given this, one of the most important problems in the PT field is the prediction and diagnosis of failures. Different hydrocarbon gases dissolved in PT oil, are generated by, and thus indicative of, many fault types, as such being one of the most important information sources for repair actions. Dissolved Gas Analysis (DGA) is the method by which the concentrations/ratios of these gases are measured, being one of the most important data obtention techniques whose output is employed by a myriad of different methods. However, DGA is very costly, increasing proportionally to the number of gases measured, requiring expensive materials, incurring lab costs for chemical analysis and making diagnosis operations slower, having to wait for the process to be performed for each gas. Thus said, we present a method for DGA data subset selection, with the intent of identifying the minimal set of gases that can be used by any existing approach with a negligible decrement in overall performance. Our approach integrates Machine Learning (ML) and subset selection techniques to provide as an output both the selected gas subsets and the models which can complete this incomplete set with the predicted values of the discarded DGA variables. We conclude this work with a thorough validation approach testing the obtained subsets and models in a set of common DGA interpretation methods.

Resumo

Um dos principais componentes dos sistemas energéticos é o Transformador de Potência (PT), fulcral para o funcionamento das modernas redes elétricas e, como tal, sempre em uso constante para as atividades atuais de industrialização. Este uso permanente do ativo, juntamente com cargas maiores do que os limites estabelecidos e atividades de redução de custos, aumenta em muito as taxas de falha e, como consequência, pode reduzir a vida útil do PT. Diante disso, um dos problemas mais importantes no campo de PTs é a previsão e diagnóstico de falhas. Diferentes gases de hidrocarbonetos dissolvidos no óleo do PT, são gerados e, portanto, indicativos de muitos tipos de falhas, como tal sendo uma das mais importantes fontes de informação para ações de reparo. A Análise de Gases Dissolvidos (DGA) é o método através do qual as concentrações/rácios destes gases são medidos, sendo uma das técnicas de obtenção de dados mais importantes, sendo estes dados utilizados por uma miríade de diferentes métodos. No entanto, a medição dos diferentes gases é muito cara, aumentando proporcionalmente ao número de gases medidos, exigindo materiais caros, incorrendo em custos laboratoriais para análises químicas e tornando as operações de diagnóstico mais lentas, tendo que aguardar a realização do processo para cada gás. Dito isto, apresentamos um método para seleção de subconjuntos de dados de DGA, com o intuito de identificar o conjunto mínimo de gases que podem ser usados por qualquer abordagem existente com um decréscimo insignificante no desempenho geral. A nossa abordagem integra aprendizagem máquina (ML) e técnicas de seleção de subconjuntos para fornecer como resultado os subconjuntos de gás selecionados e os modelos que podem completar este conjunto incompleto com os valores previstos das variáveis DGA descartadas. Concluímos este trabalho com uma abordagem de validação exaustiva testando os subconjuntos e modelos obtidos num conjunto de métodos comuns para interpretação de DGA.

Contents

List of Tables

List of Figures

Listings

Acronyms

2-ACF 2-acetylfuran

2-FAL 2-furaldehide

2-FOL 2-furfurol

5-HMF 5-hydroxymethylfurfural

5-MEF 5-methyl Furfural

ABC Artificial Bee Colony

AINC Artificial Immune Network Classification

ANN Artificial Neural Network

ANOVA Analysis Of Variance

C2H2 Ethylene

C2H4 Ethane

C2H6 Acetylene

CH4 Methane

CO Carbon Monoxide

CO2 Carbon Dioxide

CPU Central Processing Unit

CSI Classification Sampling Imputation

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DETC DE-energized Tap Changer

DGA Dissolved Gas Analysis

DNN Deep Neural Network

DP Degree of Polymerization

DT Decision Tree

EP Evolutionary Programming

FCM Fuzzy C-Means

FES Fuzzy Expert System

FIS Fuzzy Inference System

FL Fuzzy Logic

FRA Frequency Response Analysis

GA Genetic Algorithm

GBR Gradient Boosting Regressor

GBE Greedy Backward Elimination

GFS Greedy Forward selection

GLM General Linear Model

GPU Graphics Processing Unit

GRNN Generalized Regression Neural Network

H2 Hydrogen

HI Health Index

IFRA Impulse Frequency Response Analysis

IEC International Electrotechnical Commission

KNN K Nearest Neighbours

LASSO Least Absolute Shrinkage and Selection Operator

LDS Low-Dimensional Scaling

LLR Local Linear Regression

LLSSR Local Linear Semi-Supervised Regression

LSTM Long Short-Term Memory

MAE Mean Absolute Error

ML Machine Learning

MLP Multi Layer Perceptron

MSE Mean Squared Error

N2 Nitrogen

O2 Oxygen

OA Oil Quality Analysis

OLTC On Load Tap Changer

PCA Principal Component Analysis

PNN Probabilistic Neural Network

PT Power Transformer

PSO Particle Swarm Optimization

RBFNN Radial Basis Function Neural Network

RELM Regularized Extreme Learning Machine

ReLu Rectified Linear Unit

RF Random Forest

RMSE Root Mean Squared Error

RSI Regression Sampling Imputation

RUL Remaining Useful Life

SFS Sequential Forward Selection

SFRA Sweep Frequency Response Analysis

SVM Support Vector Machine

SVR Support Vector Regression

TDCG Total Dissolved Combustible Gas

TFRENN Transparent Fuzzy Rule Extraction from Neural Networks

WN Wavelet Network

WT Wavelet Transform

xGBM eXtreme Gradient Boosting Machine

Chapter 1

Introduction

The energy sector is, and has been for a long time, an ever expanding universe of inspiration, talent, innovation, but also concern. A lot of thought has been given to the way power is generated, transmitted and consumed, being the driving force of the current human society.

In this thesis we set our focus on a particular, and arguably extremely important, part of this sector, the Power Transformer (PT). While inside the spectrum of all its problems we aim at developing a general and efficient method for Dissolved Gas Analysis (DGA) gas subset selection, to reduce costs, error and downtime of PTs.

The remainder of this chapter will be divided into the following parts: (1) First, we provide further context to the area of our work, looking at some of the most common problems tackled in this field. (2) Then, we proceed to define our specific problem and the associated objectives of the work. (3) The employed methodology is then briefly explained. (4) A brief explanation of the contents of the chapters to come follows. (5) Finally, a summary of this chapter is provided.

1.1 Context

The way in which we interact with energy has always been the primary force shaping human society. From the first fires that allowed us to conquer the night, to the massive power plants that carve entire landscapes, we were always driven to obtain more power and control over it.

The energy sector is a global and highly competitive market place [18], where innovation, efficiency and human ingenuity are at the forefront of its every facet. In the most recent decades this has become

even more apparent, with a global trend of privatization and deregulation in this sector [17, 24], which has greatly increased an already fierce competition.

Amongst all the apparatus in the power systems, the PT is the most ubiquitous [18], and arguably after the generators themselves, the most important and studied. Competition has driven decisions to increase the load put on PTs, sometimes beyond the expected limits [17]. This, coupled with reduced maintenance as a cost saving measure creates a powerful recipe for PT failure. The fact that most currently operating PTs were installed in the 80s and are currently nearing or exceeding their expected operational life [42] further compounds the problem. All these factors combined lead to the expected and observed worldwide increase in PT failures [42], despite their very high reliability [18].

Other than the, somewhat obvious, consequences of PT failures that are the interruption of power transmission and distribution, electricity unreliability or total blackouts [11, 31], several other problems exist. These range everything from damage to the PT itself or the power grid, to fire or even explosions [18, 31]. Although, in case of failure, damage to only the PT is the best case scenario, this is extremely disrupting, as PTs are extremely costly pieces of equipment, sometimes accounting to as much as 60% of the total investment in a transformer station [34].

Given the great importance of PTs and the devastating consequences of their failure, several mechanisms exist to prevent problems, warn the user of their existence or to even automatically stop the transformer before it goes critical. Traditionally, to impede the occurrence of problems, preventative maintenance coupled with regular testing was used [17, 33]. However, these methods incur costs that, in such a competitive economy, simply cannot be afforded. In the end, any solution found is always compromising, weighting several conflicting factors (safety, cost, environmental concerns, etc.) [18]. These complex problems have been increasingly motivating the interest in research and development of data-driven decision making solutions.

1.1.1 Problems In This Field

From all the problems related to the management and monitoring of PTs the most commonly investigated is, by far, fault prediction and diagnosis [33]. Solving this problem implies finding the causes responsible for an identified PT fault. Whether it is a component malfunctioning, a failure of the grid, maloperation of the system or any other factor, identifying it is of paramount importance to resume regular activity as quickly as possible. Therefore, the reason as to why this particular problem is regarded of so much interest is that it allows PTs irregularities and failures to be corrected faster,

better and more cost effectively.

Of the various existing methods for solving this problem, the most common is DGA [11, 33], where different concentrations/ratios of gases in the insulator oil are measured [11, 27, 30]. The data obtained by DGA is then interpreted by a variety of different methods. Another approach for fault identification is Frequency Response Analysis (FRA), where energy of a set of standard frequencies is injected in the PT [33, 42], after which, the incurred perturbations are read from several other locations.

Remaining Useful Life (RUL) prediction and Health Index (HI) obtention are two very related problems which, due to their similarities, are usually approached as a singular one. In fact, while RUL pertains to predicting how long a transformer will last, while HI gives a single number that relates to the health of the transformer, or very closely, how long it will last [34, 38]. The results of these analyses can be used to monitor the PT condition and define optimal maintenance strategies [38].

Several other problems with far less prevalence in literature have been identified. These include:

- Load forecasting, where the electricity load put on system in the near future is predicted [28].
- Time to fault regression, where the prediction of remaining time until a fault occurs is performed [27].
- Power quality prediction, investigating the similarity of the existing power to its normal parameters [28].

For all of these, however, the literature is scarce, creating, as such, a large investigational space totally uncharted. This leads to the lack of standard practices in these domains, where each company or individual attempting to solve these problems does so in a different way with wildly different outcomes. This, to some extent, is present even for the most common methodologies and problems, such as DGA and FRA [27].

1.2 Problem Statement

Taking into consideration what we have presented, we opted to tackle an extremely important problem for which, strangely, no literature was found. This problem, proposed by Efacec, entails identifying viable (as small as possible) DGA gas subsets that could be used by any of the existing

or future DGA using methods without significantly impacting their efficacy, which is a problem of particular importance for several reasons. First, each measured gas incurs a considerable cost, needing expensive monitoring equipment, expendable resources, expert time, lab time and its associated costs, encouraging risky maintenance tactics. Furthermore, each gas also adds to the number of operations, increasing the chances of human error; while slowing down diagnosis tasks, which leads to lost revenue and impacts public perception (power instability is very damaging). Finally, with more expenditure of resources, the environmental impact is also increased. Therefore, using a proprietary, Efacec provided, DGA focused dataset we set out develop an approach that can efficiently tackle our selected problem of viable DGA gas subset identification.

1.3 Methodology

Due to the proprietary nature of our data and source code, we present an in depth description of the full data science pipeline tasks performed, with the intent of making our work as general and replicable as possible. First, a thorough dataset description and visualization process is presented. Then, all the preprocessing steps, parameters and selection process are detailed. Our full modeling methodology, where the goal is to restore the discarded variables with the remaining ones, follows. Finally, the modeling results, including a comparison of the predictions obtained from various DGA methods on the original and the restored subset data, are presented.

1.4 Document Structure

We will now provide a short summary of the contents of each chapter, their importance, as well as, how they integrate on the overall goals of this thesis.

1.4.1 Background

In chapter 2: Background we provide the conceptual and theoretical background required to understand the rest of the thesis. More emphasis and attention is paid to the material directly used in our work, however, shorter descriptions of related topics will be presented. Further material found to be useful or to go more in-depth in several areas will also be referenced for the interested reader.

1.4.2 State Of The Art

In chapter 3: State of the Art we present an overview of the gathered literature, encompassing both the topic directly at hand and several related ones. A focus has been placed on contents about PT data-mining problems, nevertheless, some similar issues, such as lifecycle management and diagnosis simulation, are also presented to give a better overview of the field.

1.4.3 Development

In chapter 4: Methodology we present our proposal for how to solve the selected problem, including the tested model structures, how each component interacts, the fundamentals that justify our solution and how our work differs from previous ones.

1.4.4 Tests And Validation

In chapter 5: Tests our test structure is presented, with an explanation of our dataset, the metrics used for evaluation, tried configurations, optimization algorithms and their parameters, among others.

1.4.5 Results

In chapter 6: Results the outcomes of our tests are explored, looking at how different parameters affected results, what models attained the best performance, showing training and prediction times and comparing to other works whenever possible.

1.4.6 Conclusion

In chapter 7: Conclusion we present the main findings of our work, whether our objectives were accomplished, the main difficulties, limitations and problems of our methodology, as well as, possible future work and improvements.

1.5 Summary

The energy sector has continuously seen a rise in importance and competition, with the PT being amongst the most important apparatus. The drive to reduce costs and downtime, while simultaneously improving reliability and efficiency has led to the introduction of many challenges. From amongst these, the most commonly tackled is fault diagnosis, while the most common way to approach it is using DGA.

With the weight given to DGA by the plethora of methods and types of problems that rely on it we present a method to obtain viable DGA gas subsets for any number of tasks, as well as, as methods to recreate the approximated values of the removed/ignored gases, with the intent of creating cheaper, faster and more effective alternatives for DGA data analysis.

For the rest of the thesis we will start by providing the necessary background materials to understand the contents, and then analysing the current state of the art in this field. We will follow by presenting our proposed methodology. The tests and the obtained results for our methods will subsequently be explained. Finally, we will present the conclusion bearing our main findings.

Chapter 2

Background

In this chapter we are going to present the required theoretical and practical knowledge required to understand the contents of this thesis. As this is being developed in the data science masters course, a basic understanding of common algorithms, terminology and methodology is expected from the reader. Therefore, a focus will be placed on the problem domain of Power Transformers (PTs), their data and problems. Nevertheless, a brief overview of the crucial parts of our methodology will be presented, namely of the topic of subset selection.

2.1 The Power Transformer

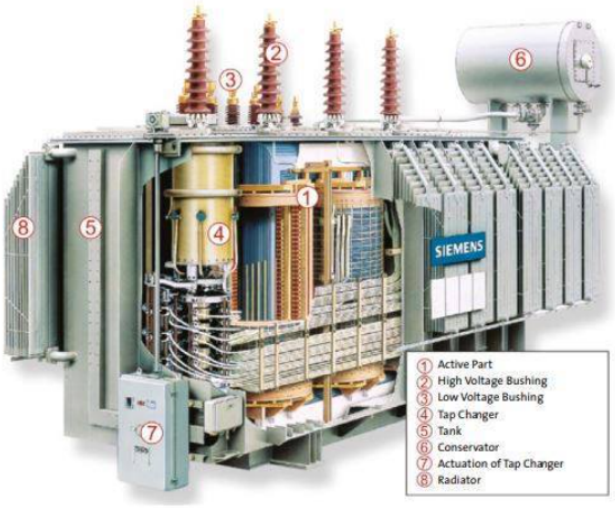

Figure 2.1: PT labeled with some of the most important components [29]

PTs are amongst the most widespread and crucial apparatus for the existing power distribution grids, being responsible for adapting voltage levels to the grids' needs [18]. Their inception dates back to the 1880s [16], upon which their humble beginnings rapidly expanded in number, complexity and importance. The most general definition of a PT defines a static machine used to transform, without changing frequency, power from one circuit to another [19]. The generation of electrical power is, in general, more efficient at low voltage levels [2], however, the opposite is true for the transmission of said power, with low voltages incurring more losses. This is a problem that the PT solves, by changing power characteristics, increasing voltage for transport and then decreasing it back for consumption [2].

Due to their importance PTs have been developed in many forms for many different uses, and characterized in different ways[1]. According to their usage, PTs can be step up or step down transformers; according to placement indoors or outdoors; according to their power characteristics three phase or single phase; with many more different categorizations existing [19].

In our work we focus on large power transformers near the power generators. These large transformers are typically three phase transformers[2], with their voltage range being typically from 600V to 5000V [9]. Of course, given their importance these devices are quite complex, having multiple subsystems, and quite expensive, accounting for the majority of the cost of PT stations [34]. These large transformers are known for their very high reliability, with a typical well designed and maintained PT having an expected operational life time ranging from 32 to 55 years [17, 18, 38]. Despite this, most currently operating PTs are reaching or exceeding their expected end of life, having been installed in the 1980s, with some even surpassing 60 years of operation [29, 33, 42].

2.1.1 Power Transformer Components

PTs are complex pieces of equipment constituted of many different components. Those that are typically found in any transformer include cores, windings, insulation materials, tap changers and bushings [5, 25, 29, 38]. From these some are replaceable, such as bushings, however most are not, such as tap changers, windings and insulating materials [5, 12, 29].

Some of these components are shown in figure 2.1, such as the bushings, tap changers and the

[1] The article at [15] was found to have a brief and very understandable explanation of PT basics, including some history, purpose and types.

[2] There are typically three phases in power systems, each having a sinusoidal voltage, with offsets between phases of about 120º [9, 28]

tank.

The windings are usually handmade out of copper and can be of different types according to the position of the core, such as shell of core type [29]. These are covered with an insulation paper, which provides greater mechanical and dielectric strengths [5, 29]. Greater than rated PT loads cause increased heat production in transformer windings, which can cause damage to the insulation and power losses [3]. These winding faults, in particular deformation, are difficult to identify before becoming very severe and are an active area of research [42].

The core is made out of steel, having great magnetic permeability which provides low resistance to the magnetic flux [29]. This low resistance is very important as PTs must create a path between windings for the magnetic flux to flow, while reducing power losses as much as possible [5, 29].

The tap changer is the component responsible for adapting the PT functioning to the necessities of the power grid [5, 29]. By changing the transmission ratio, fluctuations to the voltage levels caused by grid demand are compensated, with a change in grid demand having an opposite effect in the voltage levels [6, 29]. This is, if a higher demand is present the voltage will drop bellow optimal levels, with the tap changer making sure these go back to the desired ones. There are two main types/classifications of tap changers, with the DE-energized Tap Changer (DETC) only operating when the PT is not under load, while the On Load Tap Changer (OLTC) operates while the PT is supplying the grid [38].

Bushings are responsible for making the connection between the PT windings and the power grid through the PT tank. This component can bear higher than normal voltages, which is of extreme importance as its malfunctioning is known to be one of the main causes for the extreme PT failure by explosion [29].

There are a variety of thermal and electrical insulation materials in PTs, which provide it with greater dialectic strength [5, 29]. These include mineral oil, paper insulation and pressboard. They, in general, cannot be feasibly replaced during the PT lifetime, with as such, the PT life usually ending when one of these stops properly performing [29]. The PT oil functions both as an insulator and also as thermal cooling for the PT [3]. Paper insulation, on the other hand offers no cooling function and its condition, unlike for the oil, is very hard to directly assess, with in general, the Degree of Polymerization (DP) being used for its estimate [29]. New insulation paper starts with a DP of about 1200, ending its life when this value reaches less than 200 [20, 29].

With the PT being a critical device in the grid, whose unforeseen failure can cause major

problems, a variety of monitoring, warning and protective components are almost always present, with their sophistication increasing as the PT gets nearer to the generators or otherwise increases in importance [24]. These exist for measuring just about any important information concerning the PT current operating condition, with temperature, current, pressure and vibration being just some of the most common [24]. These PT protections and monitoring components include fuses, overcurrent relaying, differential relays, and pressure relays [24].

Interlinked with PT importance and distance to generators is their capacity, with smaller distances and greater importance being related with higher capacity. PTs of different capacities are associated with different monitoring components. PTs with less than 2500kVA usually employ fuses, while those between 2500 and 5000kVA use fuses or preferentially, due to their sensitivity, overcurrent relays [24]. For capacities between 5000 and 10000kVA, induction disc overcurrent relays are the most common [24]. Finally, for capacities over 10mVA, restraint percentage differential relays, as well as, temperature and pressure relays are usually present [24].

PT components for fault detection also greatly depend on the PT type. Sealed type PTs tend to use safety detectors [18]. Buchholz relays and oil thermometers are employed in PTs with conservator tank [18]. Relays for power tripping include Buchholtz relays (for BR gas protection), transformer current differential protection and overcurrent protection [17]. Buchholz and differential relays, however are known to only respond to severe failures [30].

2.1.2 Power Transformer Faults and Failures

As can be predicted by the large variety of fault warning and protection mechanisms, the topic of transformer faults is extremely important and complex. And it is growing more important as the PT failure rate has grown world wide [42]. This increase is caused by multiple compounding factors, including the aforementioned end of life of transformers, lower maintenance rate to decrease costs and increased loads placed on the grid [17, 24]. The amount of failures caused by transformer end of life is in fact higher than expected, due to the failure rate of transformers following what is sometimes referred to as a "bathtub" curve. This curve indicates that PTs have an higher failure rate early in life, due to manufacture or installation detects that make themselves known quickly, and late in life caused by material degradation [29, 39]. The end of life and replacement of PTs cause, as such, a double impact on failure rates, by first the old PTs failing, and then the new ones for replacement also presenting a high failure chance.

The tripping of warning and safety mechanisms is almost always unplanned and unforeseen, with about 70% of power distribution blackouts or milder disturbances being caused by short circuit faults [28, 33]. Other than short circuit, possible faults include overloading, oil leakages, electrical arcing, electrical corona and overheating of cellulose or oil [11, 17]. These can be categorised into the main groups of electrical, mechanical, or thermal and their causes internal or external [17].

There is a large amount of factors that contribute or directly cause PT failures and faults as they are constantly under electrical, mechanical, environmental, and thermal stresses, which negatively impact many components and reduce PT lifespan [11]. Despite this, the large majority of trippings do not happen due to actual transformer faults, with roughly 87% of them being caused by failures in the general grid, 10% by relay false positives, thus leaving only 3% up to the transformers themselves [17]. Some of these failures have causes external to the PT, including design or manufacture defects, poorly executed transport or installation, moisture, lightning strikes and earthquakes [17, 31]. Many other failures however, are caused by internal factors such as insulation deterioration, loss of winding clamping, overheating, oxygen accumulation, contamination in oil, partial discharge, winding resonance, switching operations, corona discharge and arcing, overload, short circuits and other thermal or electrical faults [11, 17, 18, 29, 31]. Each of these faults compounds and is worsened by transformer aging (loosened connections, oxidation of connectors or tap changer, increased vibrations and degradation of insulation), making the transformer less able to handle any abnormal events [17, 31]. Despite all the different faults and respective causes, most failures are, in the end, caused by insulation failure, with overheating being the most damaging factor [3, 17].

The consequences of a PT failure are varied. From the most obvious interruption of power transmission and distribution causing blackouts, brownouts, electricity unreliability or other outages, to environmental pollution and contamination, damage to the PT and grid, or even fire and possibly explosions [11, 18, 31].

2.1.3 Power Transformer Maintenance Strategies

In order to combat or totally prevent the multitude of possible failures, a series of different maintenance strategies are always employed. Each strategy has different advantages and disadvantages, but all help maximize the lifetime of PTs, while minimizing the risks associated with normal decay by aging and the impacts of any fault, should it occur [33]. In the end, maintenance strategies ensure that the PT is in working order almost at all times and that any problems are promptly resolved [29].

In recent years a trend of deregulation in the energy sector led to an increase in competition, which is making cost reductions in all areas increasingly important [17]. The one in particular that has seen the most changes is maintenance, with a reduction in complexity and frequency of maintenance operations. The topic of what action to take when a safety mechanism trips is increasing in importance, as shutting down for inspection can be expensive by reenergizing the transformer can lead to a catastrophic failure [17, 24].

Given this, although as we have already explained maintenance operations are crucial, a balance between optimal safety and cost savings must be maintained [29]. On one hand, performing frequent and complex maintenance routines is unfeasible due to cost constraints, while on the other, completely removing them or only performing corrective maintenance to fix a problem after it has occurred can incur even more costs due to PT degradation and consequentially possible need for replacement [29]. To further increase the difficulty of the task, the fluctuating lost revenue caused by PT unavailability, coupled with factors such as customer perception impacts, in case of power loss, make what is already a complex task into a nearly impossible one to perfect. Taking all this into consideration it becomes obvious that safety solutions are always compromising [18].

There are as such multiple types of maintenance strategies. These are corrective maintenance, preventive maintenance, with time based and condition based maintenance under its umbrella, predictive maintenance and risk-based maintenance [29]. Due to the aforementioned cost saving requirements there has been a trend of shifting from the traditional corrective and preventive maintenance strategies to the more intelligent predictive maintenance, with it continuously increasing in popularity [17, 33].

The simplest maintenance type is corrective maintenance, placing itself at an extreme of low maintenance costs but high risk [29, 36]. Here no PT monitoring, diagnosis or repair action is performed until after a failure has occurred. Due to this, any failure prevention is not performed and as a result PT lifespans suffer [1, 29]. Because of the high costs of PT repair and replacement the employment of this strategy on its own is not advised and is, as such, almost never actually used, despite presenting the lowest maintenance costs while no problems have yet occurred [1, 29].

Preventive maintenance, as its name suggests, is employed to prevent failures from occurring, with actions being performed at more frequent intervals than for corrective maintenance, thus increasing the ongoing maintenance costs but reducing the costs for transformer repair and replacement. As we mentioned previously, two subtypes can be identified, time based maintenance and condition based maintenance [29].

Time based maintenance is also very conceptually simple, with maintenance actions being undertaken at set intervals according to a schedule in a periodic fashion [29]. By employing this strategy, faults can be detected before becoming serious enough to cause failures and thus be corrected. It is, of course, possible that failures occur in between maintenance operations, with an increase in likelihood the higher the amount of time between them [1, 29]. This, as such, makes it crucial to find the right frequency for maintenance actions, creating a balance between risk prevention and expenses.

The second type of preventive maintenance, condition based maintenance, is slightly more complex. For this type, maintenance actions are only performed after a fault, hopefully still in early stages, is encountered [1, 29]. As action is only undertaken when strictly required the number of interventions is minimized [29]. On the other hand, problems that might be found during regular maintenance must now be identified using constant monitoring, which requiring specified instruments, data organizations and personnel can run very expensive [1, 29].

The second main type of maintenance is predictive maintenance. This strategy has much in common with preventive maintenance, with the goal of preventing problems while minimizing maintenance actions. This type's main difference from the previously detailed strategies is the way in which the requirement for maintenance actions is identified. Here the status of the PT is predicted with a variety of different methods, where machine learning and data driven approaches are common [8]. By employing complex tools the monitoring costs can sometimes be reduced and faults or failures predicted before other approaches would identify them [8].

The final type of maintenance approach is risk based maintenance, which is a hybrid approach where multiple other maintenance strategies can be employed depending on the perceived risk of a PT [29]. When the risk is deemed high, for example at start or end of life according to the "bathtub" failure pattern or when the PT is critical for operations, stricter preventive or predictive strategies can be employed [29, 39]. On the other hand, for low risk or less important PTs (like backup ones) the costs can be reduced, with time based maintenance using larger intervals between inspections, condition based acting only on bigger faults, predictive maintenance prescribing action only when most certain it is required or even relying only of corrective maintenance [1, 29]. For this strategy the methods employed must be carefully selected case by case, taking into consideration the multitude of factors that are deemed important [29].

2.2 Power Transformer Data

PT data comes from a variety of sources, in many forms and with different goals. Dissolved Gas Analysis (DGA) data is the most common, with Frequency Response Analysis (FRA), current and chemical data also being very prevalent. Other types include oil quality and degradation tests, furanic compound analysis, breakdown voltages, load losses, insulation resistance and other factory-floor tests [3, 14, 21]. For the detection of thermal failures better results tend to be achieved with DGA, furanic compound analysis, thermography and DP. Dielectric failures are commonly inspected with DGA and integrated sensors, while mechanical failure is more easily detected with FRA or leakage inductance. Finally, furanic compound analysis achieves the best results for general transformer degradation [18].

2.2.1 Dissolved Gas Analysis

DGA data is, as mentioned multiple times, by far the most used for fault diagnosis problems, being one of the primary indicators of transformer health and fault severity. This data is gathered through an analysis of PT oil, from which a variety of dissolved gases are obtained [11, 27, 33]. There is a large amount of relevant gases that can be measured. Different creation rates, concentrations, combinations and ratios are generated by, and thus indicative of, different fault types. The gases that are commonly measured are Hydrogen (H2), Methane (CH4), Ethylene (C2H2), Ethane (C2H4), Acetylene (C2H6), Oxygen (O2), Carbon Monoxide (CO), Carbon Dioxide (CO2) and Nitrogen (N2) [11, 27, 30].

The interpretation of these is not easy, with heavily skewed data, not linearly separable fault types and the fact that multiple faults can occur simultaneously, mixing multiple gases of different sources and possibly increasing amounts far beyond the values for which methods were developed [11, 26, 27]. Further compounding this difficulty, the fact that different PT configurations, operating conditions and maintenance can affect these gases means that is is no surprise that no universally accepted technique exists. Despite the increasing number of studies and alternatives, most DGA interpretation heavily relies on the experience of experts, thus varying widely amongst different organizations [27, 37].

Despite this, there are still some common, classical methods that see some widespread prevalence. These classical techniques include Duval's triangle, Roger's ratio, Doernenburg ratio, key gas method and the International Electrotechnical Commission (IEC) table (most of these have many variants) [11, 26, 30]. These methods have some important limitations, including not providing confidence or

severity values, being incapable of revealing trends and not giving an answer for some combinations of values [30, 37]. Although more modern Machine Learning (ML) based techniques have shown superior performance, being capable of overcoming some or all of these limitations, they are still not widely adopted [33].

2.2.1.1 Duval's triangle

Duval's triangle uses 3 gases CH4, C2H4 and C2H2 in a triangular coordinate system, where the relative concentration of these gases is utilized to delineate regions for different fault types [11]. Figure 2.2 shows the structure of the triangle.[3]

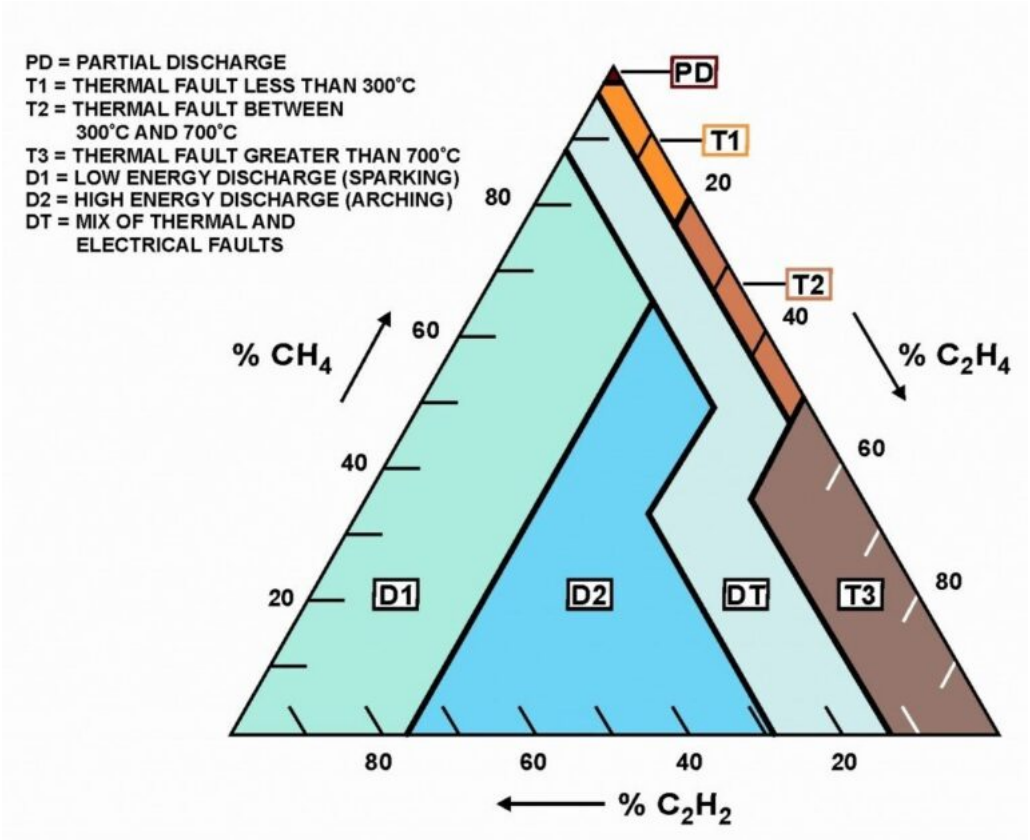

Figure 2.2: Labeled Duval's Triangle [35]

The triangle is divided into seven regions, each for a fault type. These faults can be divided into thermal, discharge and mixed types, and further divided by severity. The full detectable fault set is constituted by partial discharge, <300ºC thermal fault, 300ºC to 700ºC thermal fault, >700ºC thermal fault, sparking, arching and mixed thermal and electrical faults. [32]

[3]Although there are many images with identical information, this one was found to be the most clear.

Although, Duval's triangle is usually implemented with a projection from the triangular coordinate system, we decided to convert it into a simple set of rules, for which the understanding and implementation should be more straightforward.

PD - Partial discharge

$98\% \leq$ CH4

T1 - Thermal fault less than 300ºC

C2H2 $< 4\%$ and C2H4 $< 20\%$ and CH4 $< 98\%$

T2 - Thermal fault between 300ºC and 700ºC

C2H2 $< 4\%$ and $20\% \leq$ C2H4 $< 50\%$

T3 - Thermal fault greater than 700ºC

C2H2 $< 15\%$ and $50\% \leq$ C2H4

D1 - Low energy discharge (sparking)

$13\% \leq$ C2H2 and C2H4 $< 23\%$

D2 - High energy discharge (arching)

$13\% \leq$ C2H2 and $23\% \leq$ C2H4 $< 40\%$ OR

$29\% \leq$ C2H2 and $40\% \leq$ C2H4 $< 71\%$

DT - Mix of thermal and electrical faults

$4\% \leq$ C2H2 $< 13\%$ and C2H4 $< 50\%$ OR

$4\% \leq$ C2H2 $< 29\%$ and $40\% \leq$ C2H4 $< 50\%$ OR

$15\% \leq$ C2H2 $< 29\%$ and $40\% \leq$ C2H4

Given these developed rules, it should not be overly difficult to verify that they do define the regions of the triangle. One final note is that the selection between $\leq$ and $<$ was arbitrary, ensuring only mutually exclusive and complete regions. Under the assumption of infinite precision of measuring devices this should not be an issue, but in the event that measured concentrations fall exactly into the dividing lines special care should be had.

2.2.1.2 Rogers Ratio

Roger's ratio utilises ratios between the variables H2, CH4, C2H2, C2H4 and C2H6 to make predictions [26]. Although these gases are always used, the number of ratios and respective

interpretation can vary. In our research, the most observed number is 3, but 4 ratios were also observed [11, 22]. The number of different predicted faults has also been observed from 3 to 8. Given this, the most difficult version encountered, utilizing only 3 ratios for a total of 8 faults was selected for our work, and will be the only one detailed.

Table 2.1 contains the full description of this method, including not only the gases, ratios and predicted faults, but also the description of the fault meaning and information on the minimum thresholds for this method to be applicable in accordance to expected sensor sensitivity.

Table – Rogers Ratios for Key Gases

Code range of ratios		C_2H_2 / C_2H_4	CH_4 / H_2	C_2H_4 / C_2H_6	Detection limits and 10 x detection limits are shown below: C_2H_2 1 ppm 10 ppm; C_2H_4 1 ppm 10 ppm; CH_4 1 ppm 10 ppm; H_2 5 ppm 50 ppm; C_2H_6 1 ppm 10 ppm
<0.1		0	1	0	
0.1-1		1	0	0	
1-3		1	2	1	
>3		2	2	2	
Case	**Fault Type**				**Problems Found**
0	No fault	0	0	0	Normal aging
1	Low energy partial discharge	1	1	0	Electric discharges in bubbles, caused by insulation voids or super gas saturation in oil or cavitation (from pumps) or high moisture in oil (water vapor bubbles).
2	High energy partial discharge	1	1	0	Same as above but leading to tracking or perforation of solid cellulose insulation by sparking, or arcing; this generally produces CO and CO_2.
3	Low energy discharges, sparking, arcing	1 - 2	0	1-2	Continuous sparking in oil between bad connections of different potential or to floating potential (poorly grounded shield, etc); breakdown of oil dielectric between solid insulation materials.
4	High energy discharges, arcing	1	0	2	Discharges (arcing) with power follow through; arcing breakdown of oil between windings or coils, between coils and ground, or load tap changer arcing across the contacts during switching with the oil leaking into the main tank.
5	Thermal fault less than 150 °C (see note 2)	0	0	1	Insulated conductor overheating; this generally produces CO and CO_2 because this type of fault generally involves cellulose insulation.
6	Thermal fault temperature range 150-300 °C (see note 3)	0	2	0	Spot overheating in the core due to flux concentrations. Items below are in order of increasing temperatures of hot spots: small hot spots in core; shorted laminations in core; overheating of copper conductor from eddy currents; bad connection on winding to incoming lead or bad contacts on load or no-load tap changer; circulating currents in core—this could be an extra core ground (circulating currents in the tank and core); this could also mean stray flux in the tank. These problems may involve cellulose insulation which will produce CO and CO_2.
7	Thermal fault temperature range 300-700 °C	0	2	1	
8	Thermal fault temperature range over 700 °C (see note 4)	0	2	2	

Table 2.1: Roger's ratio table [35]

Here the utilized ratios are C2H2/C2H4, CH4/H2 and C2H4/C2H6. Depending on the value, for each of these ratios a code is attributed, with the combination of the 3 codes being used to select the fault. Unlike for Duval's triangle, the prediction of no fault is possible. The other predictions are low energy partial discharge, high energy partial discharge, low energy discharge (sparking), high energy discharge (arching), thermal fault less than 150ºC, thermal fault between 150ºC and 300ºC, thermal fault between 300ºC and 700ºC and thermal fault greater than 700ºC [22].

As such, overall, when compared to Duval's triangle, 5 instead of 3 gases are utilized, while partial discharge is divided into two (low and high energy), thermal faults below 300ºC are further divided, and mixed faults are removed.

2.2.1.3 Dornenburg's ratio

As Dornenburg's ratio was not utilized in the rest of our work due to its similarities with Roger's ratio, only a very brief description will be made. It utilizes 4 ratios between 5 different gases, H2, CH4, C2H2, C2H4 and C2H6; the same as for Roger's ratio, while only predicting 3 fault types, thermal, corona discharge and arcing. The ratios are C2H2/C2H4, CH4/H2, C2H6/C2H2 and C2H2/CH4 [11, 26].

2.2.1.4 International Electrotecnical Commission Table

Like Roger's ratio, there are many variants to the IEC Table. In fact, the IEC table stemmed from one variant of the Roger's ratio, removing the C2H6/CH4 ratio, that was found to be rarely useful [11]. Several improvements have been made during the years, to create these different versions [22, 26, 27].

Table 2.2 contains the ratios, respective codes and interpretation.

The similarities to Roger's ratio are clear, with almost the same set of faults, ratios and rules. Here 8 fault types, low energy partial discharge, high energy partial discharge, low energy discharge (arching), high energy discharge (arching), thermal fault less than 150ºC, thermal fault between 150ºC and 300ºC, thermal fault between 300ºC and 700ºC and thermal fault greater than 700ºC, are predicted from 3 ratios. These ratios are C2H2/C2H4, CH4/H2 and C2H4/C2H6, thus utilizing the same gases are Roger's ratio.

Gas Ratio	Value	Code
$X1=C_2H_2/C_2H_4$	X1<0.1	0
	$0.1 \leq X1 \leq 3$	1
	X1> 3	2
$X2= CH_4/H_2$	X2< 0.1	1
	$0.1 \leq X2 \leq 1$	0
	X2 > 1	2
$X3= C_2H_4/C_2H_6$	X3<1	0
	$1 \leq X3 \leq 3$	1
	X3 > 3	2

No.	Type of Fault	Code		
		X1	X2	X3
1	No Fault (F0)	0	0	0
2	Partial Discharge with low energy density (F1)	0	1	0
3	Partial Discharge with high energy density (F2)	1	1	0
4	Discharge (arc) with low energy (F3)	1/2	0	1/2
5	Discharge (arc) with high energy (F4)	1	0	2
6	Thermal faults of temperatures < 150 °C (F5)	0	0	1
7	Thermal faults of temperatures between 150 °C and 300 °C (F6)	0	2	0
8	Thermal faults of temperatures between 300 °C and 700 °C (F7)	0	2	1
9	Thermal faults of temperatures >700 °C (F8)	0	2	2

Table 2.2: IEC table [22]

2.2.1.5 Key Gas Method

The key gas method is the last DGA interpretation approach that will be detailed. It uses the concentration of a single gas per fault type, with high values indicating an existing fault [11]. Like Roger's ratio and the IEC table, there are many variants. Unlike them, these are far less formal, usually giving some guidance but no exact values or thresholds, thus requiring the knowledge and analysis of an expert. Despite this, a search for solid reference values was conduced, leading to those presented by Londo and Çelo [23].

Using these values, the amount of Total Dissolved Combustible Gas (TDCG) is first obtained by adding the concentrations of various gases, C2H2, C2H4, C2H6, CH4, H2 and CO, and the contribution (in percentage) from each gas is calculated. The final step is to identify cases where some gas passes a threshold. A CO concentration above 90% indicates overheated cellulose. A presence of H2 above 80% leads to corona in oil. Over 60% of C2H4 leads to overheated oil. For a concentration of C2H2 above 30% arcing in oil is predicted. Finally, in case none of these are

applicable, no fault is predicted by default [23].

It should be clear that for these rules there are several problems. First, a total of 6 gases are required to obtain TDCG, but only 4 are used any further. And second, is is possible that multiple faults are predicted simultaneously, leading, in the case only a single prediction is wanted, to different results depending on the order in which rules are applied.

2.2.2 Frequency Response Analysis

After DGA, FRA is one of the most common techniques, being noninvasive and typically used for the identification of mechanical faults, namely winding movement or deformation, which is otherwise difficult to detect. This method is usually performed at PT start of life to create a baseline which unlike DGA should not change, thus permitting faults to be detected at early stages [25, 31].

FRA works by injecting signals into the PT in one location, passing though and later being read at a different one. In accordance to the IEC 60076-18 standard these signals are mostly in the 20Hz to 2Mhz range. There are tipically 4 different configurations that are measured, end-to-end open-circuit test, end-to-end short-circuit test, capacitive inter-winding test and inductive inter-winding test [31, 42].

Like for DGA there are some variations in data. This time, however, it is more in terms of the way in which it is gathered. The first variant to appear was Sweep Frequency Response Analysis (SFRA), with newer ones, such as Impulse Frequency Response Analysis (IFRA) currently in use. There is, however, much like for DGA, no universal method for analysis and interpretation [42].

2.2.3 Other

There are a few other PT data types that were identified, such as current data and furan component analysis.

Furan compounds are generated from the degradation of the paper insulation in the PT, thus being indicative of its overall health, and include 2-acetylfuran (2-ACF), 2-furaldehide (2-FAL), 2-furfurol (2-FOL), 5-hydroxymethylfurfural (5-HMF) and 5-methyl Furfural (5-MEF) [14].

Current data tends to come in the form of multivariate time series, which can be raw waveforms (voltages and currents) or preprocessed waveforms (root mean square of value). There are typically

3 different series obtained simultaneously, corresponding to the 3 phases in power systems. As these series are related, a short-circuit in one of the phases has an impact on the others, but these quickly recover [9]. Furthermore, these time series can either have fixed or variable length depending on whether they are used for on-line (real time prediction) classification or on post-fault/off-line diagnosis [28].

2.3 Subset Selection

Subset selection, also called variable selection, attribute selection or feature selection is a process usually employed in preprocessing for removing less relevant data from a dataset, to improve modeling performance. It has a strong relation with dimentionality reduction, feature extraction and model selection, with these tending to try to achieve similar goals in different manners [7, 41].

In our work feature selection combined with dimentionality reduction is utilized in this standard way, with the goal of improving regression results. However, another way in which it is utilized is far different, with the intent not of improving results, but of identifying the subsets themselves, whose reduction does not greatly hinder performance.

2.3.1 Stepwise Feature Selection

For regression analysis, stepwise feature selection, also called stepwise regression or greedy selection, is the most popular approach. These methods work by adding the best feature (or deleting the worst feature) in a greedy fashion, at each round. Typically, the main issue is the selection of the stopping criteria, which is usually either cross-validation or a statistical test [7, 10]. In our work however, this is not a problem, as the goal is to identify a subset of each size and presenting the respective performances.

There are 3 types of stepwise feature selection. The first one, Greedy Forward selection (GFS) starts with no variables, adding one at a time until the stopping criterion is met. Greedy Backward Elimination (GBE), on the other hand starts with the full variable set and removes them one at a time. Finally, greedy bidirectional elimination combines both, at each step selecting whether to add or remove a variable, which obviously can lead into an infinite or overly long cycle [7, 10, 31].

2.3.2 Other

There are, of course, many other methods for subset selection[4], but as they do not constitute the core of our approach only a short mention will be performed.

The most basic method is exhaustive search, simply testing every single possible combination and comparing all subsets in some way to find the best.

A very common approach for subset selection involves the usage of ML methods from which variable importance can be extracted. Decision Trees (DTs), Random Forests (RFs) and variants are quite common for this given their efficiency and generally good performance [33].

Genetic Algorithms (GAs) have also seen some increasing interest, no doubt in part due to their conceptual appeal, trying to mimic aspects of the most widespread and effective optimization process known to Man, evolution. These work by randomly changing (mutating) subsets or combining some aspects (crossover) of several previously obtained [13].

Many other optimization processes are also quite common for this task, such as simulated annealing, inspired by metallurgy, or Particle Swarm Optimization (PSO) initially developed for the simulation of social/swarm behaviour [37].

2.4 Summary

During this chapter we have presented the main characteristics of the PT, its inception in the 1880s and the role of improving transmission performance. We have analysed the expected operational life, importance and cost. A detailing of PT their components, and respective roles was performed, followed by failures, their causes, consequences and maintenance actions that can be undertaken to mitigate them. We enumerated some sources of PT data. Most importantly DGA, with some of the most common methods applied to it, Duval's triangle, Roger's ratio, Dornenburg's ratio, IEC table and key gas method. FRA, furan compound analysis and current data were also looked at and explained. Finally, subset selection methods were explained, with particular importance given to the ones used in our work, GFS and GBE.

With the provided background the reader should now have enough of an understanding of all

[4]The Wikipedia page on feature selection was found to be of great help, containing far more content than what could be reasonably detailed here [40]

important methods, concepts and procedures utilised during the remainder of our work to follow without any issues. As such, we will now move to the detailing of the state of the art in the next chapter.

Chapter 3

State Of The Art

In this chapter we provide an in-depth overview of the current state-of-the-art techniques and methods related to our problem and subject area. We further elaborate on the topics presented in chapters 1 and 2. To provide a better organisation of the works and subjects presented and to ease comprehension, we divided the gathered material by subject matter, while further ordering it by publication year.

3.1 Fault Diagnosis And Prediction

As already stated, fault diagnosis and prediction is, by far, the most prominent problem tackled in literature. It consists on figuring out what factors are responsible for an occurred fault. The two most common approaches are Dissolved Gas Analysis (DGA) and Frequency Response Analysis (FRA), with the second being substantially less common.

3.1.1 Dissolved Gas Analysis

Miranda and Castro [26] developed their work with several challenging objectives. First, to improve the International Electrotechnical Commission (IEC) table for DGA based diagnostics. Second, the usage of data-mining methods, novel for this task. And third, making sure that the results are human understandable. To accomplish these goals, a model combining Artificial Neural Networks (ANNs) and a Fuzzy Inference System (FIS) was created, named Transparent Fuzzy Rule Extraction from Neural Networks (TFRENN). In this method the ANN finds the hidden relations in the data while the FIS transforms the output into a set of understandable rules. Six different methods were compared,

including, the new method, a variant, two IEC tables and two other methods. It was concluded that the new method outperformed all others, while giving a confidence value (not provided by most others) and providing results in cases the IEC tables could not.

In 2012 Mirowski and LeCun [27] performed a large review of machine learning techniques for DGA. A total of 15 methods were tested in two different problems. The first, consisted of a binary classification problem, to identify if a Power Transformer (PT) is fault prone or not (if it will fail in less than a selected, short amount of time). The second problem was regression to failure time, where the amount of time until a failure occurs is predicted. Other than DGA, some data on PT characteristics was used, including age, power and voltage, among others. The models tested include K Nearest Neighbours (KNN), C-45 Decision Trees (DTs), ANNs, Support Vector Machines (SVMs), Least Absolute Shrinkage and Selection Operator (LASSO), Local Linear Regression (LLR), Support Vector Regression (SVR), Low-Dimensional Scaling (LDS) and Local Linear Semi-Supervised Regression (LLSSR). It was concluded that for classification SVMs with a Gaussian kernel were the best, with ANNs with one hidden layer and logistic outputs in close second. For regression, LLR obtained the best results, with ANNs with one hidden layer and linear outputs in second.

Dhonge et al. [11] presents a summary of several classical DGA methods, such as, key gas method, Doernberg's Ratio, Roger's ratio, IEC ratio and Duval's triangle, as well as an ANN based approach. The ANN model used the same data as that of Duval's triangle, while achieving better results.

The work by Oliveira et al. [14] presents a study of data mining techniques for PT time-series failure prediction. The dataset used consists of DGA, Oil Quality Analysis (OA) and furans analysis. For prediction the models tested consisted of, two SVMs (1-class and 2-class variants), as well as, DTs, Random Forests (RFs) and Long Short-Term Memory (LSTM). The 2-class SVM obtained the best results overall, however, the LSTM improved with more time lags, gaining advantage at high values. It was concluded that the data sampling frequency was to low, as the degradation of the transformers was faster than the intervals. Given the results, with more data and longer series, the LSTM might become the best model.

Adrianto et al. [3] used Markov models to analyse the availability and reliability of PT oil, comparing the method with some widespread ones. Using DGA and breakdown voltages, the Markov method obtained 85% accuracy, while Duval's triangle, IEC and Roger's method, obtained 61%, 50% and 45% respectively. The main conclusions were that the new method obtained significant improvements over the others and that Total Dissolved Combustible Gas (TDCG) was of particular

importance.

3.1.2 Frequency Response Analysis

This work by Zhao et al. [42] presents a method for predicting winding fault type by using a variant of FRA, Impulse Frequency Response Analysis (IFRA). The data was obtained from a real PT, by modifying it to recreate the faults, while a SVM was used to analyse the data. The model obtained 80% accuracy, however, the dataset is quite small and no comparisons to other works or methods are provided, thus not allowing a true understand of the performance.

JinLon et al. [4] used FRA to predict winding deformations in PTs. A Regularized Extreme Learning Machine (RELM) was used to analyse the data and make predictions. This work's greatest weakness is the poor organization and sentence structure, which makes it extremely difficult to understand.

A non-invasive remote winding type recognition system is presented in the work by Mao et al. [25]. SVMs are used to analyse the FRA data of three PTs of 400, 275 and 13 kV in a majority vote wins system. Excellent results were obtained, given the very limited dataset. However, such limited amount of data simultaneously implies that the performance could be further improved and that the results cannot be fully trusted. This unreliability of the results is presented in the work itself, by a large variance in accuracy depending on the chosen train and test partitions.

3.1.3 Other Methods

In 2000, Mao [24] Implemented two novel approaches for fault diagnosis using simulated current data. The approaches are a Wavelet Transform (WT) direct decision-making logic model and a hybrid approach using an ANN and a WT. Although several types of faults were simulated, the objective of the models was only to differentiate between an internal fault and any other fault type. The generator itself was also developed in the scope of this work, thus, an in-depth overview of current simulation and the methods is provided. In the end, two different types of systems were tested, with both new approaches proving to be an improvement over the benchmark ANN. Furthermore, the hybrid approach, attained great results and generalization capabilities.

Morais et al. [28] present an overview of data mining techniques for power systems. Other than a short presentation of several works and methods, an investigation (training and testing) of several

models for time-series classification of short-circuits was performed. It was concluded that from all the methods, ANNs obtain the best overall results, however for very small time-series windows, J4.8 DTs can outperform them.

Gavrilovs and Vītoliņa [17] performed a study of PT failures and risk sources, enumerating failures, their sources and consequences. The main findings of the study were that only 3% of PT tripping is caused by an actual transformer failure, with 10% being relay false positives and 87% being faults in the general power system.

A review of artificial intelligence methods for fault diagnosis by Sun et al. [37] provides an extensive enumeration of existing methods, models and their variations. These models include Fuzzy Logic (FL), Fuzzy Expert Systems (FESs), Evolutionary Programming (EP), ANNs, Probabilistic Neural Networks (PNNs), Wavelet Networks (WNs), Genetic Algorithms (GAs), SVMs, Particle Swarm Optimization (PSO), Artificial Immune Network Classification (AINC), among others. Thus, this work provides a broad list of varied approaches, as well as a brief explanation of their workings, strengths, weaknesses and how they are commonly used.

A review of PT failure prediction methods was carried out by Ravi et al. [33]. From amongst the 160 initially gathered papers only 6 were found suitable, due to several criteria not being met, with duplicate papers, divergence from failure prediction and no presented accuracy being the most prominent. Such a low amount of papers might seem insufficient for a review, however, this gives important insight on the limitations of a lot of work submitted in this field. These limitation are, the lack of comparisons, standardised results, standard benchmarks, among others. The main conclusions were that, for DGA, ANNs obtained the best results, while for other areas the results are inconsistent and new methods must be found.

3.2 Remaining Useful Life and Health Index

In this work, Velásquez et al. [38] propose a new methodology for Health Index (HI) obtention, by using WNs. The data from 60 PTs was used and 12 machine learning techniques were employed to validate the usefulness of the HI obtained. The models used include, among others, SVMs, General Linear Models (GLMs), two eXtreme Gradient Boosting Machines (xGBMs), RFs and C5.0 DTs. The HI obtained by the new method was proven to be better than previous approaches, by allowing the tested methods to obtain better results.

In his thesis Nuno Morais [29] presents a new method for Remaining Useful Life (RUL) estimation using the Degree of Polymerization (DP) estimated by 2-furaldehide (2-FAL) analysis. A broad overview of PT components, failures and maintenance techniques is presented. A novel approach for correcting the drop in 2-FAL values after oil changes is developed to great results. A comparison of the results to several other methods was presented, with the new one attaining a significant efficacy improvement. The main limitations of this work are a less than optimal dataset and the overlooking of other available variables that might have allowed for better results.

Sarajcev et al. [34] present a novel approach for HI estimation using a hybrid model combining a Bayesian network with a Deep Neural Network (DNN). Using the results from the model an influence diagram is presented, to provide optimal maintenance scheduling. Unlike some other works, this one presents the HI in 5 discrete levels, from very bad to very good. Using both the results from the model and the influence diagram, a way to estimate costs or profits from a given choice (whether or not to perform maintenance) is provided. This work is limited by the model's lack of ability to use some complex data structures and an extremely small dataset.

3.3 Feature/Subset Selection

The usage of a neural fuzzy network was proposed by Naresh et al. [30] for DGA interpretation. Competitive learning was employed for variable selection, while a fuzzy rule base was created using subtractive clustering, subsequently being feed into the network. The model was applied to two datasets and the results were compared with several other methods, including, Roger's ration, Fuzzy C-Means (FCM), Radial Basis Function Neural Network (RBFNN) and Generalized Regression Neural Network (GRNN). The new model had the best results, followed by GRNN, with FCM proving to be the worst.

Chauhan et al. [9] in this masters dissertation from Thapar University perform a study of PT functioning under different operating conditions. These are, normal operation, magnetizing inrush, operation under external fault and over-excitation. The current analysis of PTs under these conditions was performed using several machine learning algorithms. Artificial Bee Colony (ABC) was used for feature selection, while RFs, SVMs, DTs and linear models were used for predition. RFs obtained the best results, while the linear model obtained the worst.

3.4 Other Problems

Georgilakis et al. [18], proposed a new stochastic petri net based methodology for simulating diagnosis and repair operations, allowing their respective duration estimations to be obtained. This method is mostly hand-crafted, with no automatically learned weights or structure from data. Thus, an in-depth overview of the diagnosis and repair process is also presented. The main limitations of this work are the slightly limited and inaccurate off-site section of the model, the fact the model cannot learn from data and the lack of modeling for some uncommon faults.

The work presented by Khalyasmaa et al. [21] develops a method for intelligent lifecycle management for power network equipment. A case study of a PT was conducted to validate the method. The data used is comprised of DGA, physical and chemical testing of oil and load losses, among others. The performance was measured in the task of predicting the future values for all the variables by using a tree boosting algorithm.

Nurmanova et al. [31] present an interpolation based approach for short-circuit severity prediction. Three interpolation methods, linear interpolation, natural cubic spline and piecewise Hermite interpolation, were used on FRA data. Sequential Forward Selection (SFS) was used for feature selection. In all tests cubic Hermite interpolation obtained the best results. No other works that attempt this task were presented, and as such, not comparisons were made.

In their work Das and Kempe [10] perform a comparison of different greedy feature selection methods, providing a deep mathematical background and comparing the results on a set of problems. Overall it was concluded that greedy algorithms perform well in practise even in cases for which the variables are heavily correlated.

3.5 Summary

In this section we reviewed and analysed related literature. Although effort was put into obtaining documents related to different problems and solutions in the PT field, our gathered literature is still very much dominated by fault diagnosis as a problem and the usage of DGA data as a solution. Some of our findings go in contradiction to those in the discussed reviews, namely, the fact that the most observed and successful data mining method was the SVM, not the reported ANNs. ANNs were however still very prevalent, with DTs and RF also being quite common. The recurring problems

found in the presented works pertained mostly to the datasets used. These were often too small, synthetic, with no real data used, not publicly available or lacking variety in terms of the systems tested.

Taking into consideration the strengths, weaknesses, the limitations and the underdeveloped areas we proceed in the next section to present our methodology, developments and planed contributions to the field.

Chapter 4

Development

The aim of this chapter is to present our proposed methodology for solving the selected problem, i.e identifying viable Dissolved Gas Analysis (DGA) gas subsets. With this goal in mind we developed a full data science pipeline. This contains, first, an extensive visualization of our dataset. Then, a variety of methods for the preprocessing of our data, including missing value imputation, distribution transformations and dimensional reduction, are employed. Using the cleaned data generated by the previous steps, our modeling approach, including the models tested and selection criteria, is explained. We also touch on the topic of the tested configurations (preprocessing steps, model parameters, etc). The details of this topic will, however, remain to be further elaborated upon on chapters 5 and 6.

4.1 The Dataset

Our dataset is as mentioned in chapter 1, proprietary, provided by Efacec, and therefore, it will not be made publicly available and specific data entries will not be shown or discussed. Nevertheless, aggregation metrics and measurements, as well as, different graphical analysis will be presented. Throughout this section we will present the type of variables available, their distributions and moments, simple relations between different variables, as well as, some more complex discovered by regression techniques.

4.1.1 Content

The dataset contains roughly 1000 entries split amongst 200 different Power Transformers (PTs), for an average of 5 entries per transformer. These entries will from now on be referred to as samples, although in some instances multiple entries appear for one PT at a given time, caused by measurements being taken in different locations of the PT. Each sample contains a total of 41 variables. From these 2 were removed due to containing either 0 (all missing) or 1 (no useful information) unique values, leaving 39 variables.

The variables present are therefore:

- Transformer specific information (8)
  - Reference - index/reference of power PT (unique per PT)
  - Power (kVA) - power rating (quilovoltampere (kVA)) of the PT
  - Tension (kV) - PT secondary voltage in kV
  - Manufacture Year - year the PT was manufactured
  - Oil Brand - brand of oil used in the PT
  - Oil Type - specific type of oil used (each oil type corresponds to one brand)
  - Oil Weight (ton) - total oil weight in tonnes
  - Disruptive Tension - maximal rated voltage before short circuit
- Sample data (31)
  - Sample information (2)
    - Collection Date - date and time of sample collection
    - Point of Collection - point/local of collection of the sample in the PT
  - Oil characteristics (10)
    - Oil Temperature - oil temperature of sample
    - Aspect - physical aspect of PT oil (clear vs cloudy)
    - Oil Density - PT oil density
    - Viscosity (40ºC) - oil viscosity at 40º
    - Interfacial Tension - tension of the membrane between oil and pure water
    - Acidity Index - oil acidity in mg KOH/g

    - Water Content - water content of the oil in %
    - Flash Point - Flash point of the oil in ºC (temperature at which vapors ignite when there is an ignition source)
    - Tang Delta (90ºC) - cosine of the phase angle (amount of current passing through the oil)
    - Color - numerical code of oil color (0.5 - yellow almost white, as it increases becomes redder and darker)
  - Particles present in the oil (6)
    - Sediments - percentage of sediments in oil
    - Mud - mud percentage in oil
    - Part4 - ?
    - Part6 - ?
    - Part14 - ?
    - Particle classification - ?
  - Furanic compounds (4)
    - 5 HMF - concentration of furanic compound 5-hydroxymethylfurfural (5-HMF)
    - 2 FOL - concentration of furanic compound 2-furfurol (2-FOL)
    - 2 FAL - concentration of furanic compound 2-furaldehide (2-FAL)
    - 5 MEF - concentration of furanic compound 5-methyl Furfural (5-MEF)
  - Oil DGA gases (9)
    - H2 - concentration of Hydrogen (H2)
    - CH4 - concentration of Methane (CH4)
    - C2H4 - concentration of Ethane (C2H4)
    - C2H6 - concentration of Acetylene (C2H6)
    - C2H2 - concentration of Ethylene (C2H2)
    - CO - concentration of Carbon Monoxide (CO)
    - CO2 - concentration of Carbon Dioxide (CO2)
    - O2 - concentration of Oxygen (O2)
    - N2 - concentration of Nitrogen (N2)

4.1.2 Distributions

In order to understand the distributions of the variables, as well as, to identify anomalies or interesting information, a variety of statistical moments were obtained and univariate graphical analyses performed. However, before these could be performed, a minimum amount of data cleaning was required. This data cleaning includes dividing Particle classification into 3 variables (Part1, Part2 and Part3) and changing values like "$<$ 0.5" or "$>$ 8" into floating point values. This processing represents the bare minimum required for any data manipulation and visualization.

4.1.2.1 Statistical Moments

An analysis of different statistical moments and data aspects (missing value counts, data types, etc.) was performed[1], from which we will detail and explain the most important takeaways.

First, all variables contain missing values, with some reaching as high as 85.4% of missing values, with a total of 98.1% of samples having at least one missing value. On the other hand, no DGA variable has missing values, which is crucial for our intended methodology.

Another important fact is that most of our variables are numeric, with only 4 being categorical. This, of course, makes the preprocessing and visualization steps easier, as categorical variables are harder to handle.

Comparing the unique value counts with total counts we concluded that the categorical variables do not contain an overly large number of different values for the dataset size, with the largest value being of 17, while the lowest of 3. It is worth noting that for this metric missing values were considered as a new unique value.

When inspecting the values of mean, standard deviation, minimum, maximum and quantiles, it became clear that all variables have extremely different scales and distributions. Further looking at the kurtosis and skewness we saw that almost all variables are very far from a normal distribution.

[1] A table containing the full contents of this analysis is present in the appendix

4.1.2.2 Graphical Analysis

A small amount of univariate graphical analyses were performed. Histograms and box plots were done for numerical variables, while for categorical variables bar plots were employed.

We will not be presenting an exhaustive explanation of each plot created, as such would be outside the scope of this thesis. We will therefore present only a few of the primary types of distributions found and address variables not directly shown by indicating to which type they belong.

From the histograms 5 types of distributions were found, which can be classified as heavily left skewed, left skewed, right skewed, normal and bimodal.

In figure 4.1 we can see an example of an heavily left skewed distribution, that of 5-HMF, where almost all values are concentrated in a single area (usually close or equal to 0) and only a few instances with far higher values, being severe outliers. These extreme points are however of great significance, as most PT related methods involve some sort of anomaly prediction/detection (rare faults, extreme events, etc.). The other variables with such a distribution in the dataset are 2-FOL, 2-FAL, 5-MEF, H2, CH4, C2H4, C2H6, C2H2, Acidity Index, Tang Delta (90ºC), Part4, Part6 and Part 14. Thus, all furanic compounds, as well as, some particles and most DGA variables present this type of distribution. These findings are in accordance with the values of skewness in the previously analysed moment summary table.

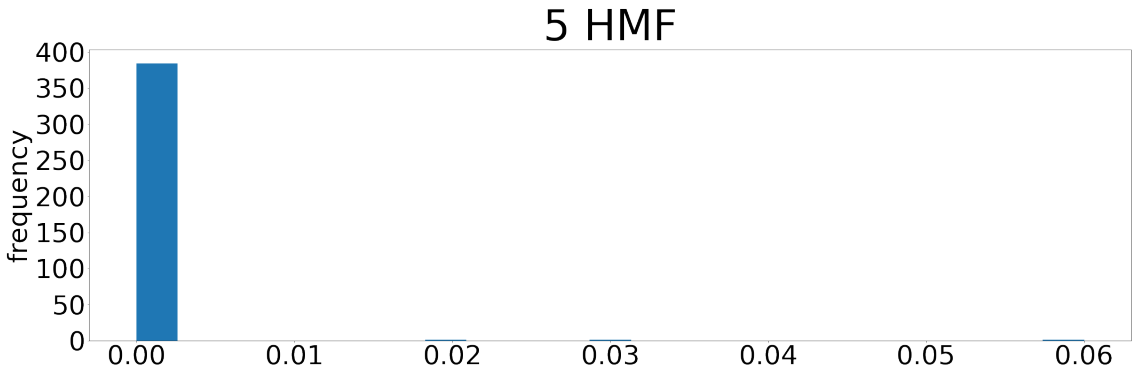

Figure 4.1: Histogram of the 5 HMF variable

For an example of the the left skewed distribution type we turn our attention to figure 4.2, containing the histogram of CO. The only difference between this kind of distribution and the previous one (heavily left skewed) is the degree of skewness, with this one having much less extreme values for this moment. The other variables with this distribution are Power (kVA), CO2, O2, Cor, Water Content, Sediments, Mud, Part1, Part2 and Part3.

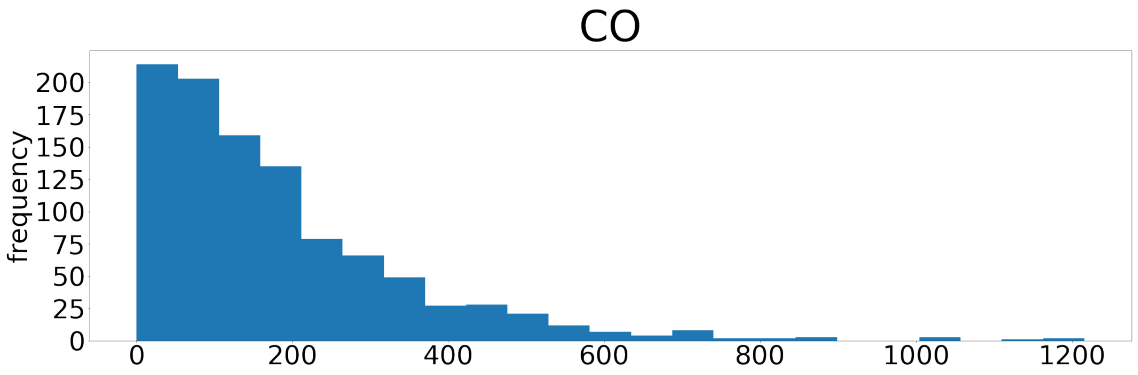

Figure 4.2: Histogram of the CO variable

The histogram of Oil density in figure 4.3, shows one example of the right skewed distributions, characterized by the skewness moment having the opposing sign to that of both the left skewed types. Others in this group are Tension (kV), Manufacture Year, Viscosity (40ºC) and Disruptive Tension.

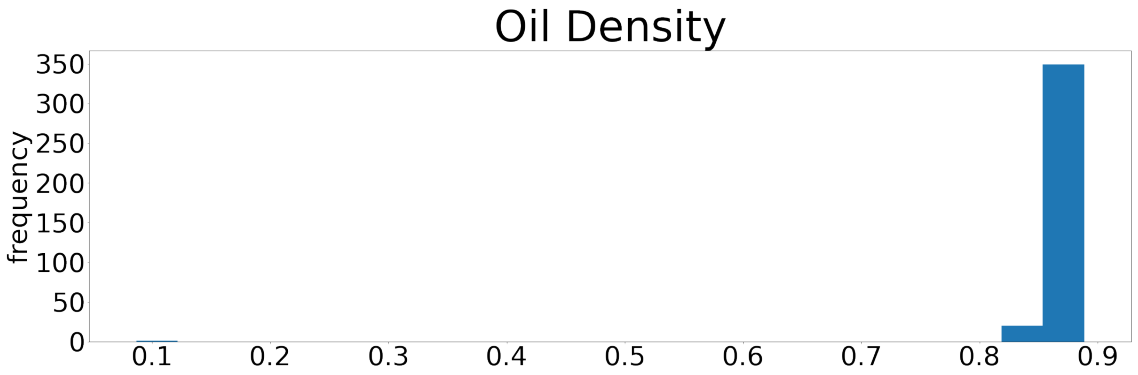

Figure 4.3: Histogram of the Oil Density variable

The normal distribution type contains variables whose distribution is roughly normal, presenting both a symmetrical and bell shaped histogram. In figure 4.4 we can see the histogram of Oil Temperature, presenting such a distribution. Other variables with this type are Reference, Power (kVA), Oil Weight (ton), Collection Date and Flash Point.

Finally, for the bimodal distribution type we can look at image 4.5, where the histogram of N2 is presented and the existence of two peaks can clearly be seen. Interfacial Tension is the only other instance of this distribution type in the dataset.

Much like for the histograms the box plots obtained can be divided into a few categories depending on their primary features. Unlike for the histograms however, only 4 groups exist, with the normal and bimodal distributions being indistinguishable in the box plots, and as such merged.

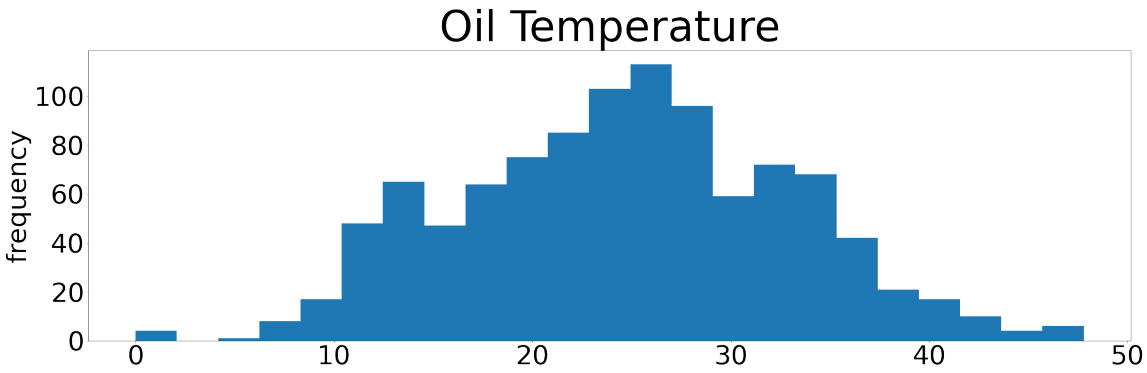

Figure 4.4: Histogram of the Oil Temperature variable

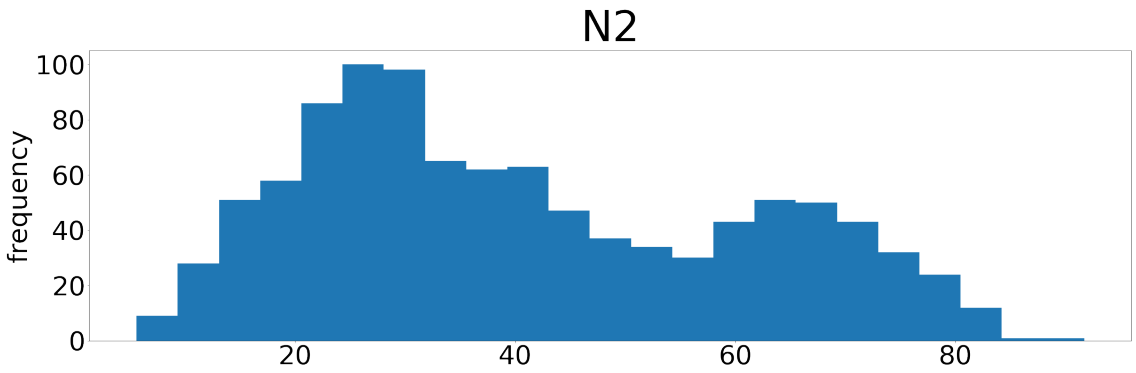

Figure 4.5: Histogram of the N2 variable

For consistency, the memberships of each group, as well as, the group names remain the same. The only exception to this is the merging of the normal and bimodal groups into a singular one. This new group which we have decided to call centered, contains as such the combined membership of the original two (bimodal and normal) which comprise it.

Taking this into consideration, for an example of the heavily left skewed distribution type we can look at figure 4.6 containing the box plot of 5-MEF. For this group, all the quantiles are merged into one point (at the minimum), with only a few points outside this area, presenting as extreme outliers.

Unlike for the previously analysed group, the left skewed one presents substantially more, but less severe, outliers and the quantiles no longer collapse into a single point. We can see this kind of distribution in figure 4.7, containing the box plot for CO.

The right skewed distribution type is very similar to the left skewed one, with the only difference being that the outliers appear below the quantile lines, instead of above. The box plot of Flash Point in figure 4.8 shows this very well.

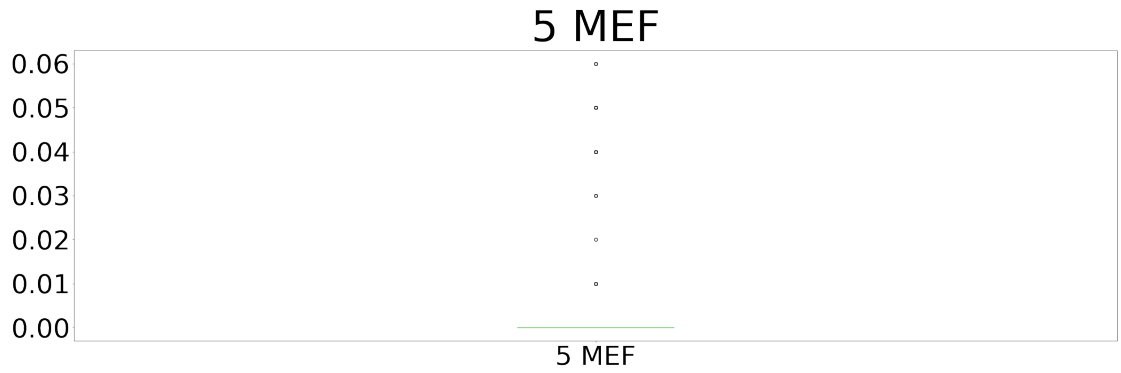

Figure 4.6: Box plot of 5 MEF

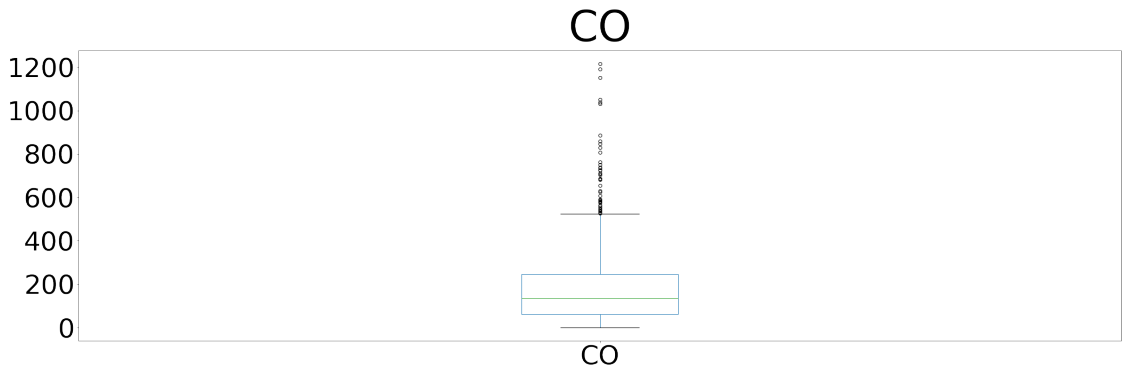

Figure 4.7: Box plot of CO

Finally, for the centered group we turn our attention to figure 4.9 containing the box plot of the variable Reference. Much like its name indicates, this distribution type is characterized by a symmetrical box plot with few outliers.

The final univariate graphical analysis we will be explaining is bar plots. These have only been applied to categorical variables, as although some transformations to numeric variables could be

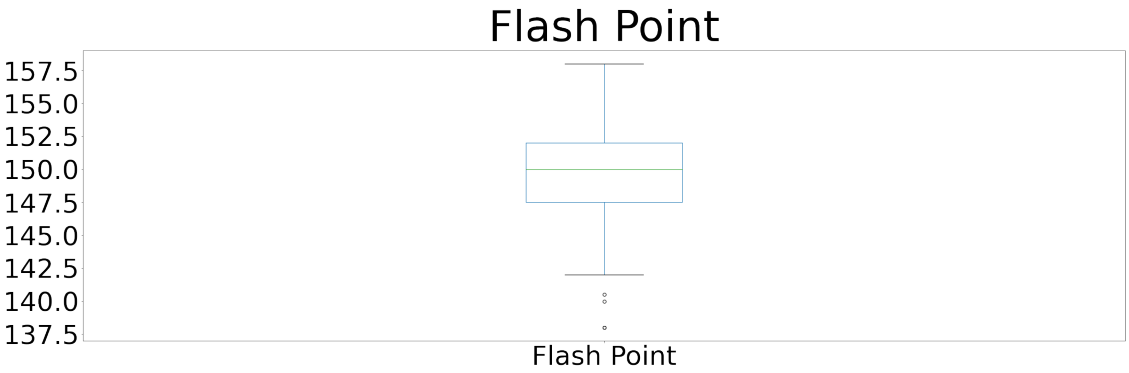

Figure 4.8: Box plot of Flash Point

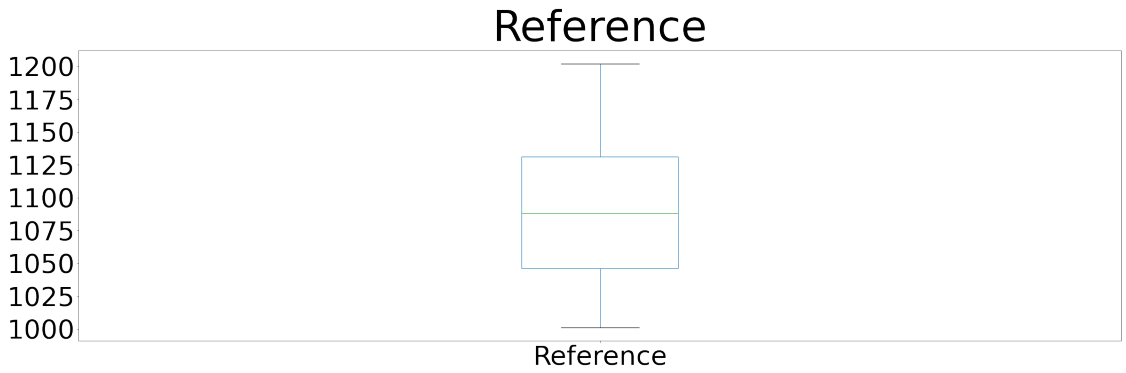

Figure 4.9: Box plot of Reference

performed in order for this kind of plot to be applicable, it was considered that no useful information could be extracted in such a way.

In figure 4.10 we can see the bar plot of Oil Brand. This is the only one that will be directly presented as all others have similar features. Here we can see a single dominant value, being far more common than all other value occurences combined.

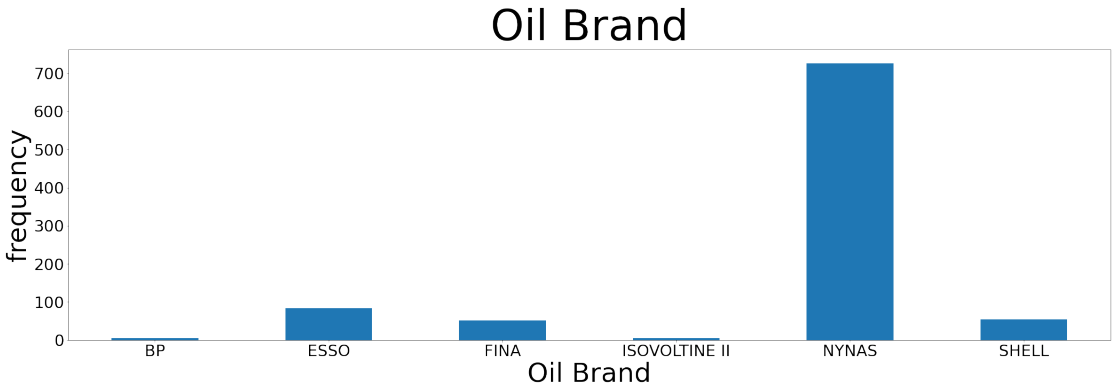

Figure 4.10: Bar plot of Oil Brand

4.1.3 Relations

The second important visualization step that was undertaken is the identification of relationships between pairs of variables variables. Understanding these relations not only helps with the subsequent preprocessing, namely in the feature selection process, but also allows us to better understand the final results and infer their meaning.

In order to do this a variety of bivariate analysis were performed and respective graphs and plots created. These include frequency and relative column frequency heatmaps, violin and box plots of

categorical variables split by class, kruskal test heatmaps and others. However, from all these, only two contain enough information density to be worth detailing in this thesis. Of course, all others did influence our methodology and result analysis process, but only in more subtle ways.

The first of these highly relevant analysis is a correlation heatmap. One such heatmap can be seen on figure 4.11, where, for compactness, only the variables with a value of correlation greater than 0.7 with any other were kept.

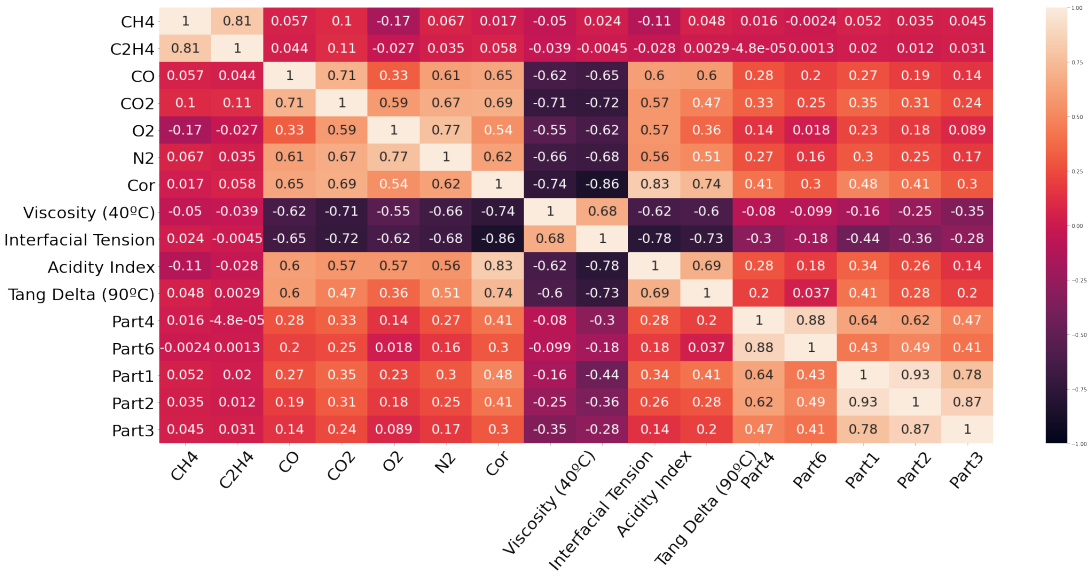

Figure 4.11: Correlation heatmap for variables with correlations greater than 0.7 with any other

Here we can identify the sets of variables that have high correlation with each other. From these, the first group consists of CH4 and C2H4. The second group consists of CO, CO2, O2, N2, cor, Viscosity (40º), Interfacial Tension, Acidity Index and Tang Delta (90º) (although some of these present negative correlation with others). The third group consists of Part4 and Part6. Finally, the last group consists of Part1, Part2 and Part3. The third and final group's members also present a moderate correlation with each other.

Given this, we expect that removing or somehow grouping features in the same groups might lead to better results in the modeling phase. It is also expected that for the DGA gases, when performing subset selection, some elements of these groups will be the first to be removed.

We would also like to note that some DGA variables are not present here, meaning that they have no high correlation with any variable. Thus we expect the variables H2, C2H6 and C2H2 to be kept in all but the smallest subsets, as predicting these is non trivial.

The second and final type of plot that we will be detailing in this section is the scatter plot.

Although these plots were created for each variable combination, given their sheer number (1521), only a small subset containing the most relevant insights will be explained.

Much like for the histograms, the scatter plots can be aggregated in groups with similar characteristics. For example, a member belonging to such a group can be seen on figure 4.12, containing the plot of CO vs CO2. Here the primary characteristic that we can notice is that the variance of one variable is proportional to the value of the other. Using such a relation we would expect to obtain good predictions for low values of both variables, but not so good predictions for the opposite case.

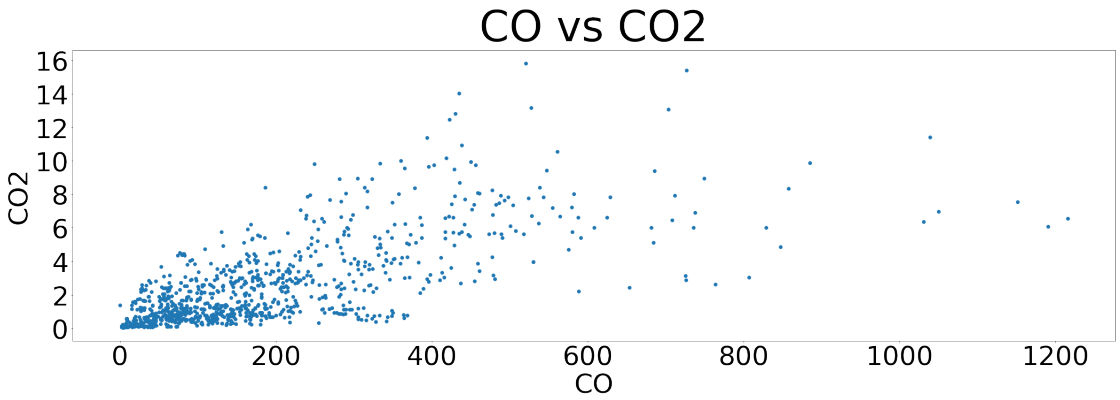

Figure 4.12: Scatter plot showing CO compared with CO2

The inverse effect from the previous relation can be clearly seen on figure 4.13, comparing O2 with N2, where the variance of one variable is inversely proportional to the value of the other. Of course, for this relation the conclusion regarding prediction effectiveness is opposite to the previous one. Meaning that we would expect better predictions for higher rather than lower values for both variables.

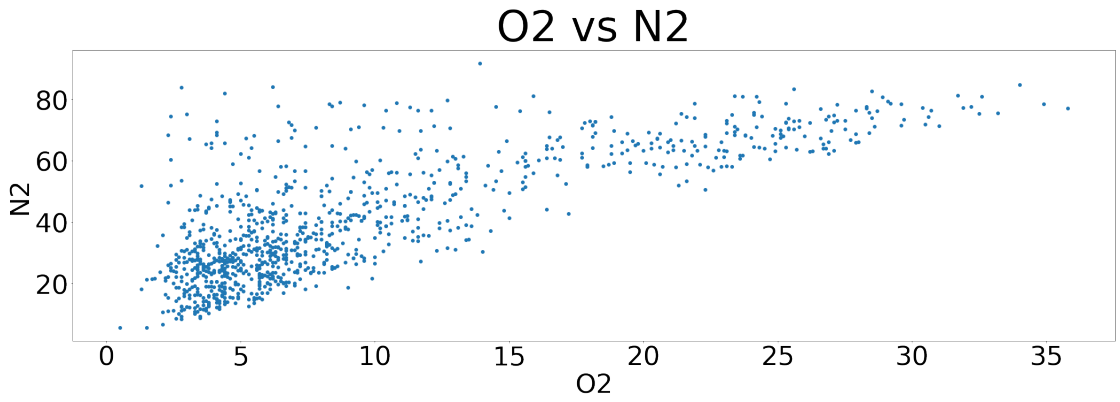

Figure 4.13: Scatter plot showing O2 compared with N2

Another important thing to note is that not all relations are linear. Other kinds of relations can be seen on figures 4.14 and 4.15. The first one shows what is clearly a polynomial relation between the variables CH4 and C2H4. On the other hand, the second plot, comparing Interfacial Tension and Acidity Index, presents an example of an exponential relation.

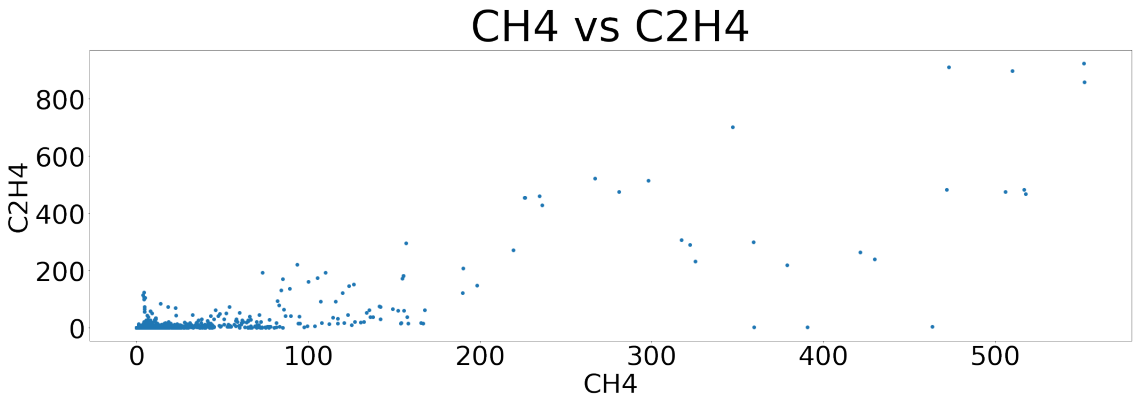

Figure 4.14: Scatter plot showing CH4 compared with C2H4

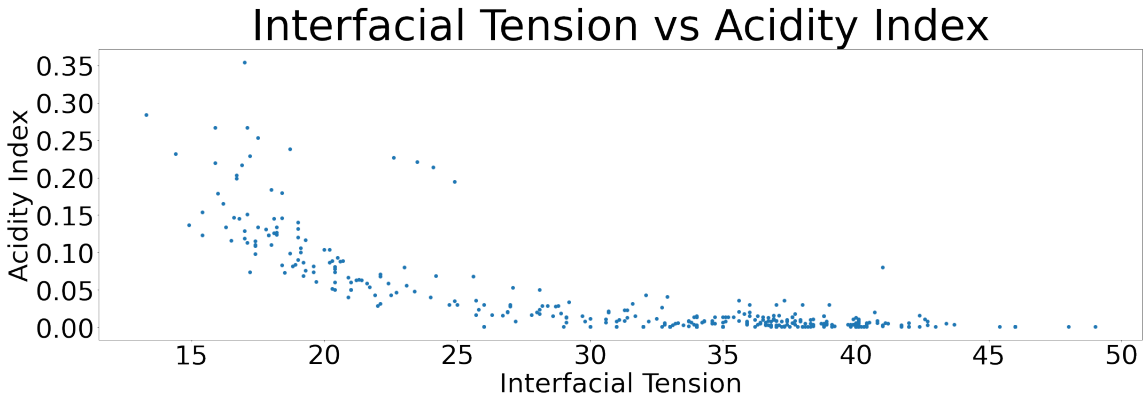

Figure 4.15: Scatter plot showing Interfacial Tension compared with Acidity Index

There are of course many instances where no discernible pattern can be spotted, with the points just creating an homogeneous cloud. No such examples will however be presented, as these are of little interest and did not guide our decision making.

4.1.4 Basic Regression

Having found the existence of obviously non linear relations on the dataset we employed simple regression algorithms to better identify and understand them. To accomplish this, two sets of regressions were performed. The first one, where the regression for all combinations of 2 variables (1 feature and 1 target) was performed. And a second, performed in a similar exhaustive manner with 3

different variables (2 features and 1 target).

For both, the set of regression models created consists of polynomial (up to degree 10), logarithmic, exponential and cubic spline. Although both the logarithmic and exponential models are not very involved, requiring only a transformation of either the features or the target, the polynomial and cubic spline models require some further explanation.

For the polynomial regression two methods were tested for choosing the optimal degree. These are k-fold cross validation and the F-test from Analysis Of Variance (ANOVA). By utilizing a visual graphical analysis the method employing the F-test was determined to obtain more suitable results and so only this one shall be explained in greater detail.

This method starts by using a degree of 0 (only a constant value feature), where the respective model is obtained. Then, the degree is increased by 1, upon which the new model is trained. This is followed by both the new model and the previous one being compared in the F-test. If the test provides a value smaller than a low threshold, the new model is saved and the next degree is tested in the same manner. On the other hand, if the value is higher, then the iteration stops and the model from the lower degree is selected as the best.

The mentioned threshold has to be decided, for which a very low value of 0.0001 was selected, as not only were we interested in the actual regression results but more importantly in insight, which with a smaller degree is easier to achieve as more interpretable results are obtained.

For the cubic spline model a few parameters have to be selected. These are the number of knots, their locations and the degree of each section. In order for the functions to be smooth, a minimum degree of 3 is required, thus for better interpretability (simpler model) we selected this value. Although selecting the knots manually for each combination of variables might lead to better results, this was not feasible and as such the knots were fixed on selected quantiles of the distribution. For this 2, 3 and 4 knots were tested. For the 2 knots they were positioned at the 33rd and 66th percentiles. The positions for the 3 knots are at the 25th, 50th and 75th percentiles. Finally, for 4 knots, the placements are at the 20th, 40th, 60th and 80th percentiles. Via graphical analysis, 3 knots were found to perform the best, and as such this value is used from this point on.

4.1.4.1 2D Regression

Having selected and parameterised the regression methods we proceeded with the obtention of the 2D regression models and respective plots. However, these plots contain a large amount of information and as such we will begin by explaining what each element means.

First, the title contains a few bits of important information. These are the two variables being plotted, the model type (polynomial, logarithmic, etc.), the R-squared of the regression and the model's mathematical formula.

The points are colored in accordance to their status as outliers. The red points are severe outliers that were removed before using regression. A few outlier detection methods were tested, from which the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) performed the best, and as such used.

In order to use DBSCAN, which is a clustering algorithm, a maximum distance value has to be defined. For this, we performed an optimization task, where the size of the second largest cluster was maximized. As this has to be done for each plot, a very fast and automated approach for optimization was needed. We, as such, employed a combination of random and grid search, where the limits for both were set to the minimums and maximums of the distributions. Furthermore, only a very limited amount of values could be tested, with the grid search using a set of equidistant points, while random search employed an identical amount of points sampled from an uniform distribution.

Once this optimization task is concluded, the points that are considered outliers are those that belong to any but the largest cluster, whose members are in turn considered ordinary points.

The blue line indicates the first regression model, that was trained without those severe (red) outliers. Unlike for the severe outliers where DBSCAN was employed, here the box plot outlier detection method obtained the best results. This is a much simpler method that defines outliers as those which are 1.5 times the inter quartile range from from the closest quartile of the distribution. The values of prediction error were used for this detection and the outlier points marked orange.

The second regression model, in orange, is created by discarding the outliers of the first one (orange and red points). Once again, the box plot outlier detection is performed, with new outliers (those not considered outliers in the first model) marked yellow, while those that no longer are outliers, marked blue. Finally, the points that at no time (before or during any regression) are considered outliers are marked green.

Having explained the contents of the regression figures we will now be analysing some of the most relevant regressions obtained. Of course there are many instances were the regression was very unsuccessfully, as well as many large sets of very similar and successful regressions. Only one element of each successful set of regressions will be explained, while all failed regressions will be ignored as these did not influence our decision making.

The first of several regression results can be seen on figure 4.16, which shows the regression of Tang Delta (90º) given 2-FAL. With an R-squared greater than 0.92 this regression is very good. If we refer back to the correlation heatmap we can see that this relation is not shown, meaning that it is a strictly non linear relation, for which a polynomial of degree 2 was found to be the best fit.

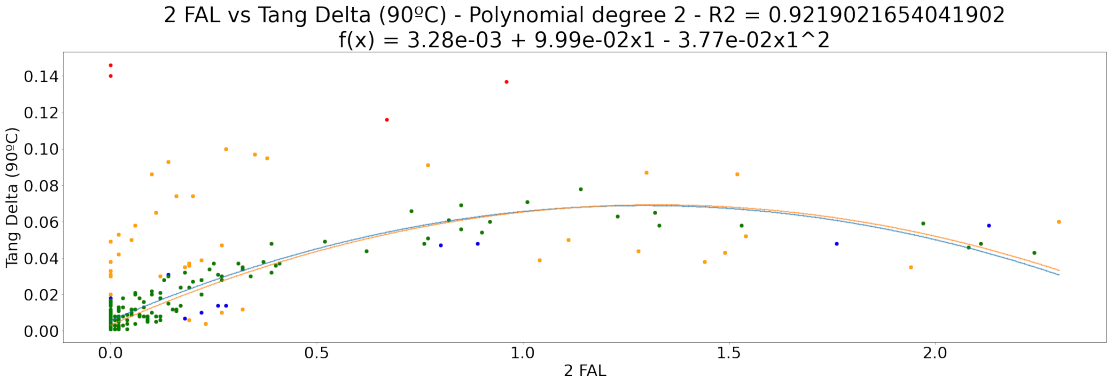

Figure 4.16: 2D regression of Tang Delta (90º) given 2-FAL

Another plot with very similar features is shown in figure 4.17 of CH4 given C2H6. Once again it shows a quadratic polynomial regression whose correlation is not presented in the heatmap. The most striking feature, however, is the collection of severe outliers. These and other similar points will have to be carefully handled in the next steps.

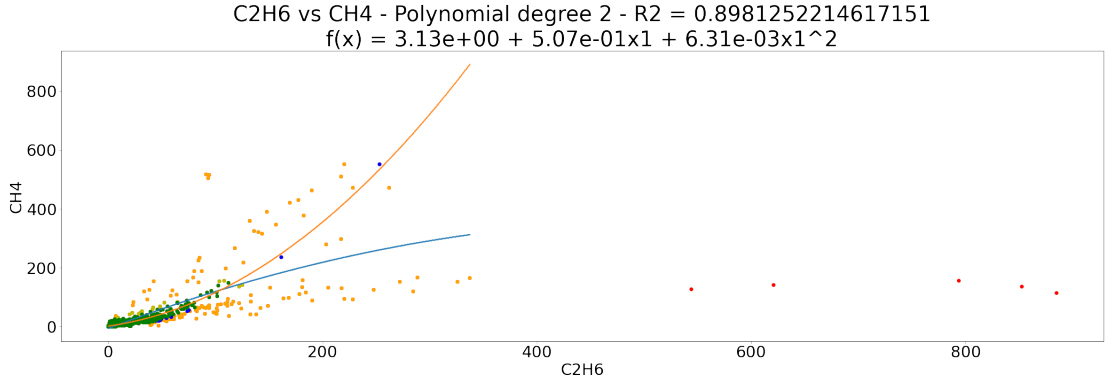

Figure 4.17: 2D regression of CH4 given C2H6

One plot that shows a quite different relation is that in figure 4.18, of C2H6 given CH4. Although

it presents the exact same variables as the previous plot, just swapped, this one tells a very different story. We can clearly see that the plot splits into two, implying the existence of two implicit models. The same also leads us to the conclusion that a third variable might cleanly distinguish these, improving the results by a good amount.

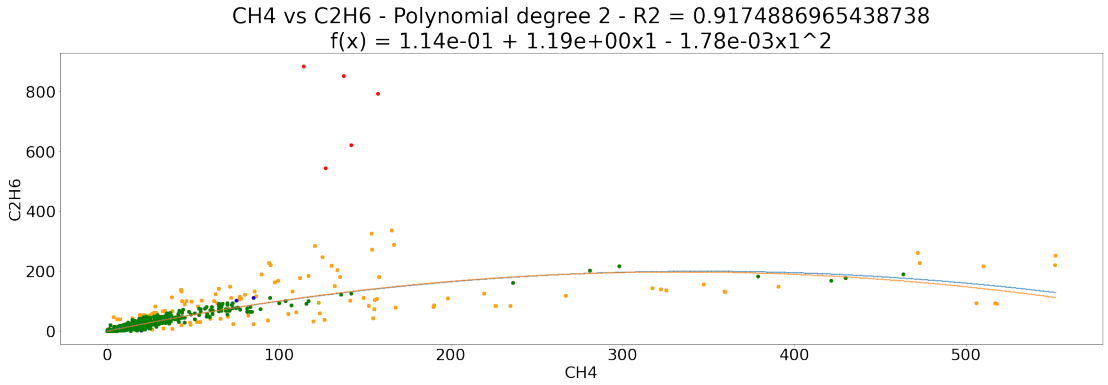

Figure 4.18: 2D regression of C2H6 given CH4

As pointed out in the previous subsection, it was expected that for N2 given O2 the regression results would be much worse in the higher values. We confirm our expectations by looking at figure 4.19, containing said regression.

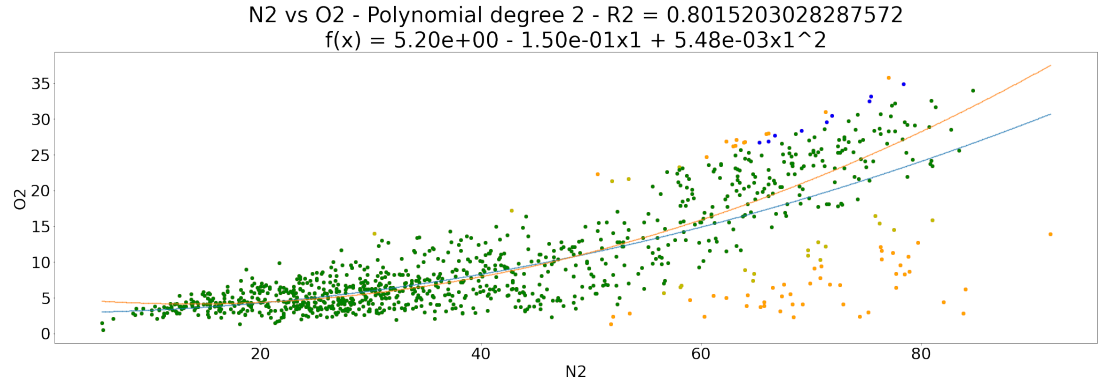

Figure 4.19: 2D regression of N2 given O2

4.1.4.2 3D Regression

Results like those shown in figure 4.18 give greater weight to the task of identifying relations with a larger number of intervening variables, and as such, regression using 3 variables was performed.

Both the methods and the meaning of the plot elements remain the same as those in the 2D regressions, with two exceptions. The first is the non creation of the title and as such the lack of all

its contents. The second is that, although the first model was created, the regression plane is not plotted (to reduce clutter).

All the 3D plots were created with 3D plotting software, and due to its nature getting a good idea of what is displayed from a single image is more difficult. Much like for the 2D regression, many plots were not analysed due to belonging to a group where all elements are very similar. From these groups, if the performance was found to be sufficient, exactly one element was presented. Given this, all examples with subpar performance were not included in our analysis.

Another factor that makes it so a plot will not be analysed is the lack of contribution from all variables. This happens when one feature and one target have a very strong relation, and as such, although the third variable is not relevant, the regression achieves great results.

For the first 3D plot we turn our attention to figure 4.20, containing the regression of Reference given Flash Point and Part14. For this one the model is a polynomial of degree 1, meaning a linear regression. We can also see that there are no severe outliers, only with one moderate outlier.

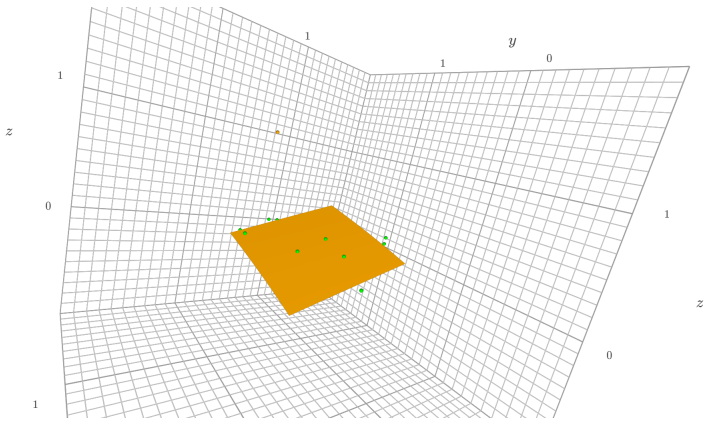

Figure 4.20: 3D regression of Reference given Flash Point and Part14

The second regression plot pertaining to the obtention of Manufacture Year by using Flash Point and Part14, much like in the previous plot, can be seen in figure 4.21. Here, the selected polynomial degree is 2 and with a R-squared greater than 0.95, the combination of these two features greatly improves prediction strength when compared to each individually, which achieved less than 0.7.

In figure 4.22 we can again see an example of a regression which achieves much better results than those by each feature individually. This figure shows the regression of Viscosity (90º) given Oil Weight and Part14.

With the previous 3 figures selected we can see a trend. That trend is that although Part14

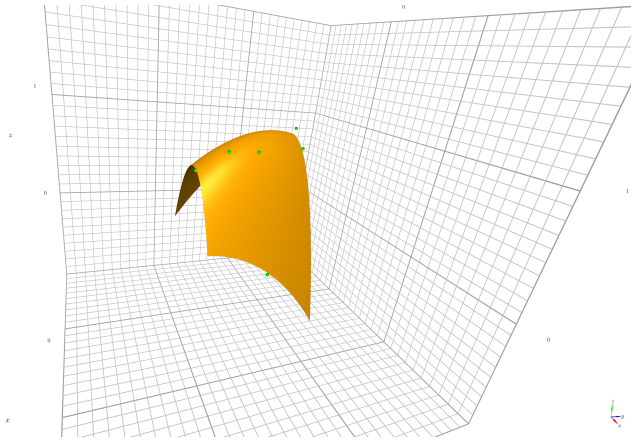

Figure 4.21: 3D regression of Manufacture Year given Flash Point and Part14

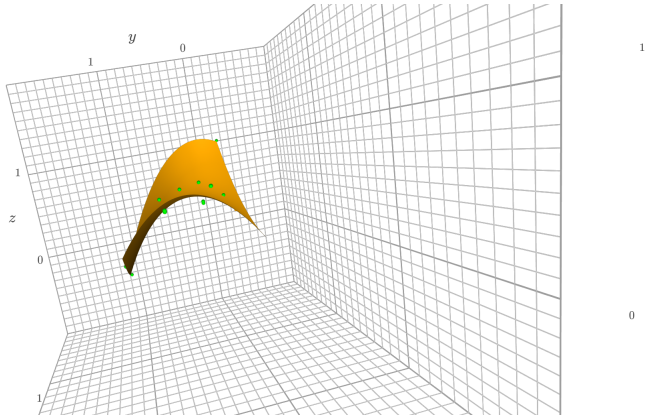

Figure 4.22: 3D regression of Viscosity (90º) given Oil Weight and Part14

individually does not allow for great regression results, by being combined with other variables it becomes much more useful. Although we will not present any more examples with Part14, many more of this happening exist.

For a different kind of distribution we can look at figure 4.23, where Viscosity (90º) and Part6 are used to predict Tang Delta (90º). Unlike the previous two quadratic plots, the signs of the second degree are different, as such creating a saddle surface.

Finally, we present figure 4.24, where C2H6 and Interfacial Tension are used to predict CH4. We remind the reader that in the 2D regressions we mentioned that of CH4 given C2H6, where we concluded that two underlying distributions are present and, therefore we would expect a third variable to be able to differentiate them. The aforementioned figure shows one example where a variable permits such a differentiation.

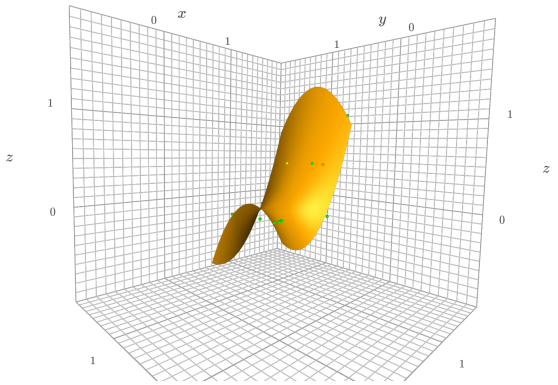

Figure 4.23: 3D regression of Tang Delta (90º) given Viscosity (90º) and Part6

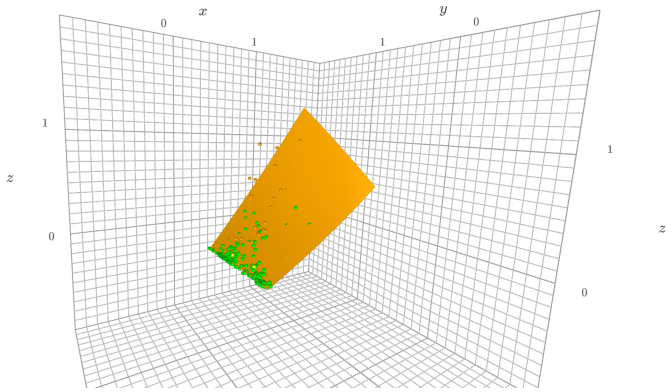

Figure 4.24: 3D regression of C2H6 given Interfacial Tension and CH4

4.1.5 Visualization Takeaways

Given the extensive visualization that was performed a recap of the main takeaways is warranted. We started by detailing the contents of the dataset, where we enumerate the existing variables and their meanings, finding the existence of transformer, sample, furanic compound, DGA, Oil and Particle variables.

Then we proceeded with an analysis of statistical moments, for which a table containing missing value data, means, modes, minimums, maximums, quantiles and other metrics was used. It was concluded that the variables have very different scales and distributions.

To better understand said distributions we performed univariate analyses, where histograms, box plots and bar plots were employed for the task. It was concluded that most variables are very unbalanced, with the numeric ones being mostly left skewed, while the categorical ones having very dominant values.

Having thoroughly studied the data distributions of individual variables we moved to the relations between them. For this we employed among many other plots, scatter plots and correlation heatmaps. From these we identified some groups of variables that have high correlation and as such would be useful in regressions. The finding that the variance of some variables is related with the values of others was a very important one. We also identified that some variable relations are intrinsically non linear.

Finally, in order to better explore these non linear relations we expanded our approach with several simple regression methods. First we did regressions with pairs of variables, but after analysing cases with multiple underling models we moved to regressions with 3 variables, hoping an extra variable would differentiate between models, which was found to be the case.

This visualization section was considered a requirement to properly explain and justify our decision making in the following sections, as well as to understand the final results and conclusions. As such, throughout the next sections we will also be indicating what parts of this visualization process are most responsible for the decisions being made.

4.2 Preprocessing

Having established our goal in chapter 1 and thoroughly described our dataset in the previous section of this chapter we will now explain our preprocessing approach.

Other than a few internal type corrections, the fist step was to handle missing and conditional string values. What we mean by conditional string values are those like "$<$ 0.05" or "$>=$ 8", which in order to be replaced correctly require us to apply a condition/restriction.

Given the small size of the dataset and the large amount of missing or conditional string values, the simplest method of just removing rows with these is not applicable, as such our focus is on imputation.

4.2.1 Missing Numerical Value Imputation

For numeric missing value imputation we tested 3 techniques of increasing complexity. These are mean imputation, regression imputation and a novel method that we developed, which we decided to call Regression Sampling Imputation (RSI).

The mean imputation is very simple. Here each missing value is just replaced with with the variables mean. The regression imputation uses a simple K Nearest Neighbours (KNN) model with all variables. Different values for the number of neighbours were tested and we settled on 5.

For the third type, RSI, as it is more complex and a new approach we will provide a more in depth explanation. Figure 4.25 shows a flowchart of the algorithm.

This algorithm works in the following way. First it is verified if there are missing values, and if there are, a set of regression models, containing all combinations of size n with the variable as a target, is created. We tested sizes 2 and 3, with 3 proving to be the best.

Then, we remove poor models, with those being the ones with an R-squared less than a given threshold. If no models remain we utilize the backup strategy of imputing with the mean.

We follow this by, for each missing value, sampling a random model, weighing each model proportionally to its respective R-squared. For the model selected we get the prediction and check if it is inside the variable's range.

As imputing with values outside the range will probably lead to bad results, we resample if that is the case, otherwise we impute with the obtained value. To prevent long or infinite resampling loops we limit these to 10 attempts, and if this limit is reached we impute the value with the mean.

This algorithm was developed with a few goals in mind. In contrast to the simple mean imputation it preserves much better the data distribution. The comparison with a regression imputation method that simply uses all variables at once is, on the other hand, slightly more nuanced.

When using a regression algorithm for imputation the data distribution and relations are preserved, however, a degree of bias caused by this regression is created. Such bias can be easily seen on continuous regression techniques such as polynomial regression, where in two dimensions all points will fall inside a line.

This can greatly hinder results and generalization, therefore, in order to combat this a common naive approach is to simply add random noise to the predictions. However, this noise is usually unrelated with the relations and distributions of the variables and, as such, also leads to other problems.

Taking these limitations into consideration we developed this method that while preserving the relations and distributions of variables simultaneously maintains a degree of variance caused by the differences in the regression models, thus creating noise that is in accordance to the dataset.

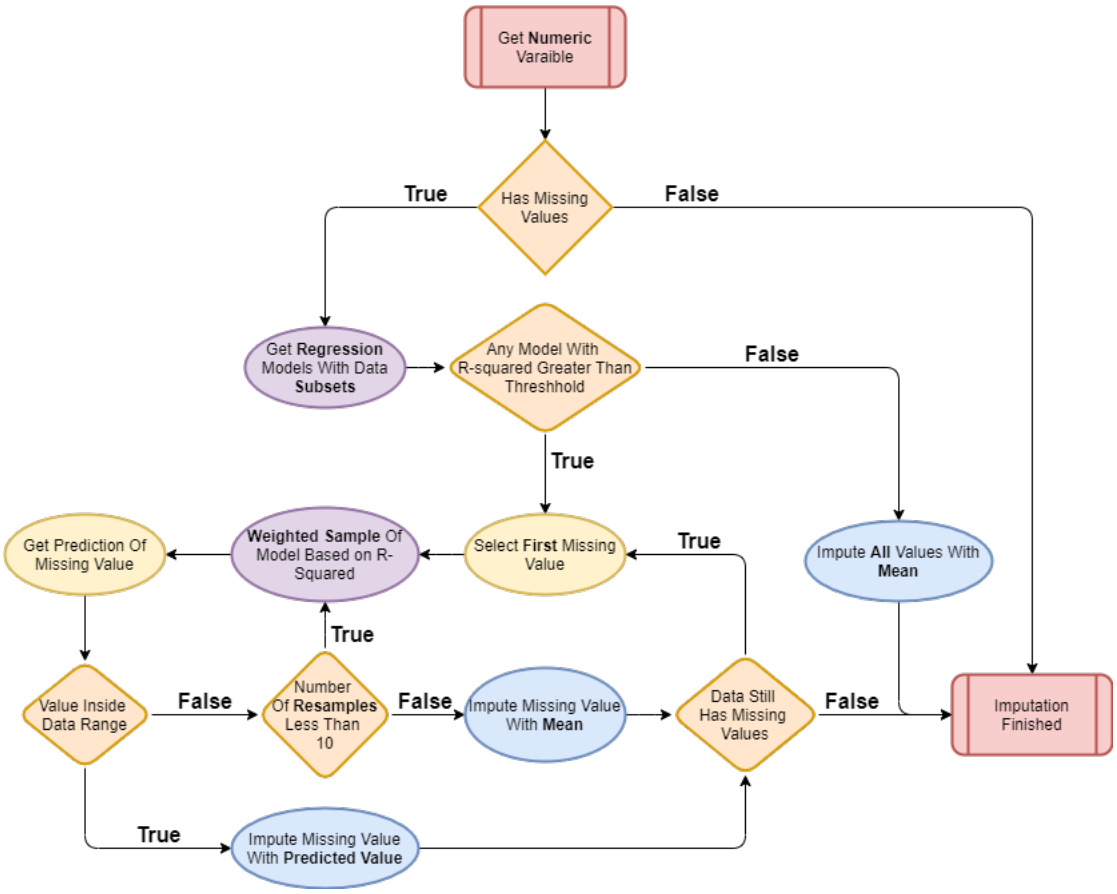

Figure 4.25: Flowchart of the regression sampling imputation algorithm

4.2.2 Conditional String Replacement

To replace the conditional string values mentioned at the beginning of this section, the same methods as those used for the numerical missing values were tested. That is, the methods of mean, regression and RSI.

The main divergence in the method for these types of values when compared to the previous ones is the restriction of the dataset to only the subset in which the condition is fulfilled. This is, if we have, for example, the value "< 1", we limit the dataset to the rows for which this variable has values less than one.

4.2.3 Missing Categorical Value Imputation

Once again, for the imputation of categorical variables a similar approach was used. Instead of the mean used for both numerical and categorical string imputation, we used the mode. Instead of our

RSI method we used Classification Sampling Imputation (CSI), which utilizes the same underling algorithm, with the regression models replaced with KNN. And, finally, the KNN that was used for regression was only slightly changed to be applicable to classification (by label encoding the classes and limiting neighbours to 1).

4.2.4 Transformations

Having imputed all missing values and transformed all variables to the correct types, we proceeded with some transformation of the dataset. Some of these transformation were employed to accommodate the modelling process, such as one hot encoding, while others to attempt to improve results, such as box cox.

The first transformation used is, the aforementioned, one hot encoding of the categorical variables. Very few machine learning algorithms can directly handle categorical variables, so this is a crucial step towards our goal. Due to its importance, this transformation was always used in all our models.

Because of the large amount of variables, further increased by the one hot encoding step, we looked at various methods for dimensional reduction. From these two promising ones were selected and subsequently tested.

The first one is Principal Component Analysis (PCA), for which the variables to be combined were selected by using their correlations. This selection works by finding the sets of variables that have high (absolute) correlations amongst themselves. In order to to do this a threshold value for the correlation has to be selected. Several values for this threshold for were tested, with lower threshold values leading to a greater dimensionality reduction.

Although several methods exist for the selection of the number of principal components to keep, we employed the simplest one, selecting only the best component. This was done as other methods did not seem promising and due to the limited amount of time available to try other approaches.

The second type of dimensionality reduction method involves using a regression method, from which the importance of each parameter is extracted via the weights for each variable. To obtain the model from which the variable importances will be gathered, the regression of the DGA variables values was performed. For this task many (although not all) regression methods can be used and therefore, due to its simplicity, short computational time and generally good results, the extra trees method was selected.

As seen in the previous section, the variables have extremely different distributions from each other. Adding to that, both proportional or inversely proportional relations were found between the variance of some variables and the values of others, which would most likely lead to the worsening of modeling results. To address both these problems the box cox transformation of these variables was performed, making their distributions closer to that of a normal distribution. This transformation also has two other advantages, with these being, the significant improvement of statistically based methods, while simultaneously tending to improve models computation times.

The last main issue which we aimed at solving by using transformations is that of the wildly different variable scales, which not only make it more difficult to understand results, but also are known to greatly decrease the performance of many models as well as worsen computational times. To solve this we employed the simple solution of normalizing the variables, thus scaling the their values to lay between 0 and 1.

Lastly, one thing that is important to note is that as the DGA variables are our targets, most of these transformations were not applied or were applied conditionally later during modeling. In fact, the dimentionality reduction methods were never used on these, and of course, as no DGA variable is categorical, neither did one hot encoding. Box cox and normalization, on the other hand, are the transformations that are employed at a later step, but only for the DGA variables that are not being predicted.

4.2.5 Validation

For the last part of the preprocesing section we will be looking at graphical results for some of the imputation and transformation steps, in order to give the reader a better understanding of what exactly is being done and what are the obtained results.

First we will analyse an example comparing all of the implemented numerical imputation methods. Figures 4.26, 4.27 and 4.28 show the results of each these methods on Interfacial Tension.

The first one is the mean, which results in exactly what we would expect, with all imputed entries having the same value. It is obvious that most information is not preserved, leading on one hand to the prevention of more complex algorithms from finding good relations, while on the other reducing the likelihood of overfitting.

The regression imputation method is presented on the second image, which displays imputed

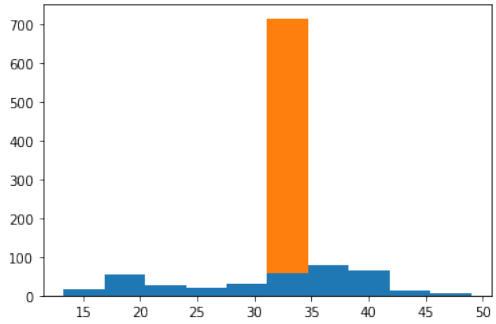

Figure 4.26: Histogram of Interfacial Tension mean imputation

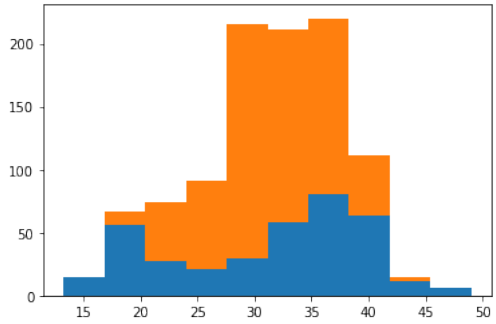

Figure 4.27: Histogram of Interfacial Tension regression imputation

values substantially more similar to those of the distribution. This of course maintains far more information, but it is also obvious that the distribution of the imputed values is still far from the original values.

The last numerical imputation image shows our RSI method, which maintains the distribution of the variable even more than the previous approaches. This does not necessarily mean it is better, since as only 2 features are being used, higher dimensional relations might not be found.

The last thing we would like to note regarding these results is the slightly left shifted imputed

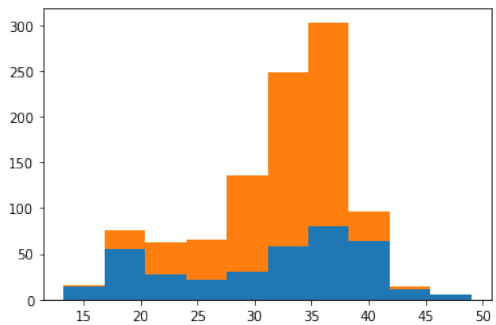

Figure 4.28: Histogram of Interfacial Tension regression sampling imputation

values for both regression and RSI when compared to the original data. Although at first glance it might seem that the results are not great, this shifting is both expected and proves the quality of the imputation. This is the case because in the dataset the missing values occur in earlier samples and as such on newer (at the sampling time) PTs. It is well known that PTs near to their beginning of life present lower values for many variables, such as DGA, furanic compounds or indeed the one presented.

We will not provide a detailed example of the conditional string replacement methods as the results for these are very similar to the one just seen, with just one notable exception. Unlike for missing values, these types of values appear in later samples, and as such, the shifting previously observed is reversed, with imputed values being slightly right shifted.

The categorical value imputation, once again, presents similar results to those from the numerical imputation and will not be detailed. In this case as there is no relation between transformer age and the categorical variables there is no deviation from the distributions.

From the set of performed transformations only a few will be further explored. One hot encoding, extra trees variable selection and normalization do not presenting any further interesting information, and as such, do not make the cut. This is due to the fact that one hot encoding just creates new categorical variables in a totally predictable manner; extra trees does not present any graphical models, while analysing its selected variables would not give any further insights; and normalization simply scales values, thus only the axis values would change.

This as such leaves, for the first transformation we will be looking at, PCA. For this we will be analysing a single component and the variables that comprise it. On figures 4.29 and 4.30 we can see the histograms of Cor and Acidity Index, while on 4.31 we can the the histogram of the respective principal component.

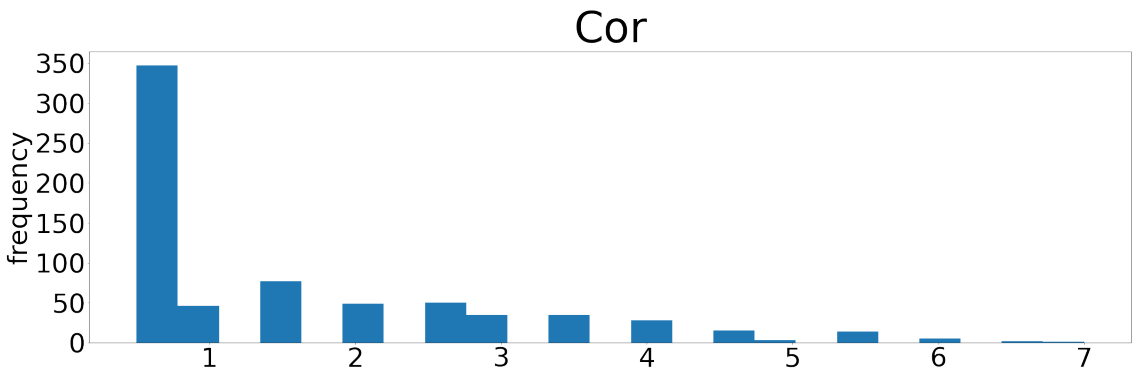

Figure 4.29: Histogram of the Cor variable

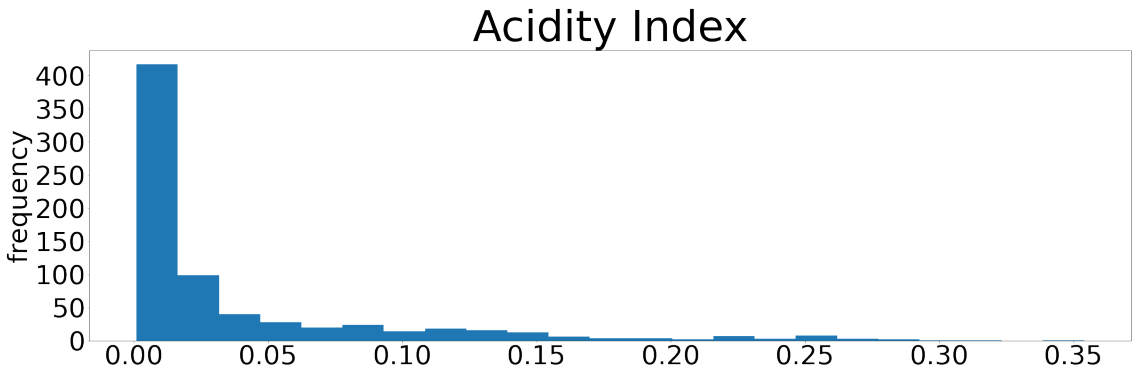

Figure 4.30: Histogram of the Acidity Index variable

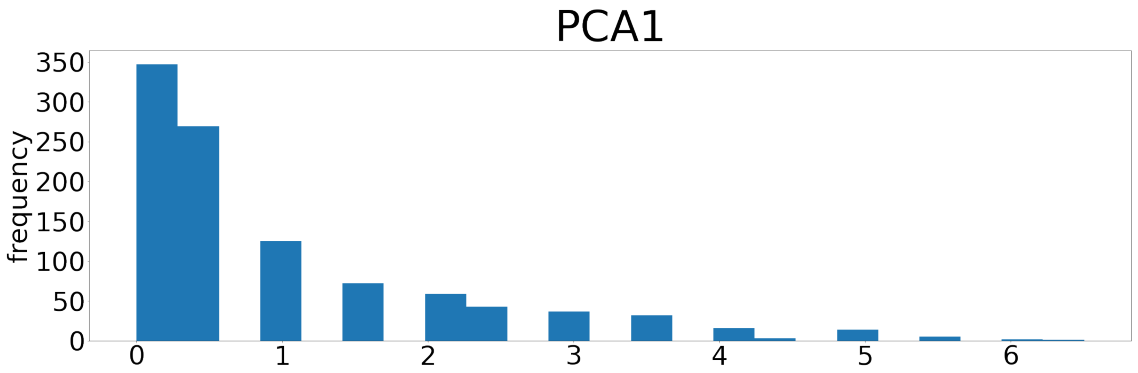

Figure 4.31: Histogram of the Acidity Index and Cor first principal component (PCA1)

Here it is clear that the two original variables have very similar distributions, although with different scales. Both follow a roughly exponential distribution curve, which is preserved, and in fact, amplified in the PCA component, making it easy to see how the main information from both variables is preserved.

For a good example of the effects of box cox transformations on our variables we can compare the figures 4.32 and 4.33, containing the original and the box cox transformed values of 2-FAL. This example is great as it allows us to see both what its success looks like and one of the limitations associated with this transformation. On the first figure we can see that most values are close or equal to 0, while a long tail extends beyond the value of 3. On the transformed variable we see that most of the distribution is now normal, however, since box cox does not transform values of exactly 0 in any way, a large amount of values outside the main body of the distribution is left. This, nevertheless, is not necessarily a problem, since as these values present little information some methods might be able to focus more on the more extreme and interesting cases, leading to better results.

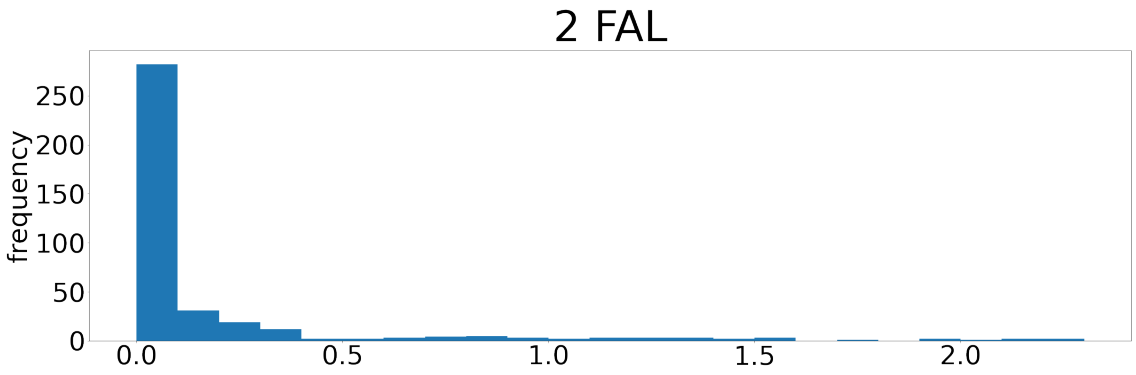

Figure 4.32: Histogram of the 2-FAL variable

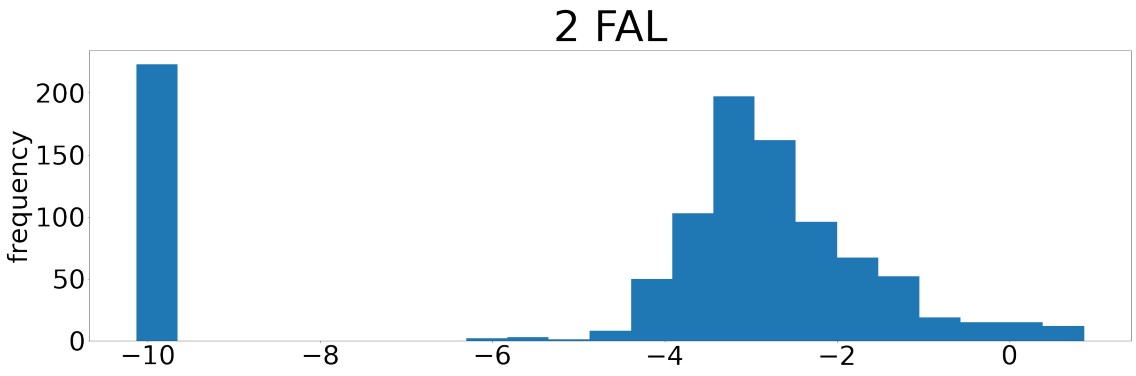

Figure 4.33: Histogram of the Box Cox transformed 2-FAL variable

4.3 Modeling

With the crucial visualization and preprocessing steps concluded and thoroughly detailed we now move to the core of our problem solving methodology, which is, of course, our subset selection approach and the generation, training and fine tuning of a myriad of machine learning algorithms.

4.3.1 Subset Selection

The first important component of our modeling approach encompasses the methods used for the selection of the subsets to test. As the amount of our DGA variables is 9, the total number of possible subsets, ignoring the complete set (a complete set would mean no reduction in variables) is 511. Although this number is not high when it comes to subset selection problems, as these grow exponentially with the number of variables, it is nevertheless, due to the large number of tests and lengthy computational times, unfeasible. Therefore, due to these limitations only a small amount of

all these can reasonably be tested in our limited time span.

A large number of subset selection methods exist, from which we selected two, with these being Greedy Forward selection (GFS) and Greedy Backward Elimination (GBE). These were selected as a middle ground between testing all combinations or only a hand picked few. They work by adding or removing a variable at a time, comparing each generated set and proceeding with the best. This is repeated until no variables remain in case of GBE or until all variables are present in case of GFS. Using this process the number of subsets that need to be tested is reduced from 511 to 45.

To better explain how the selected methods work we present in figure 4.34 a flowchart of one of them and its integration with the rest of the modeling approach. We will only analyse the chart for GBE, as GFS is equivalent with the exception that removal operations are replaced with addition.

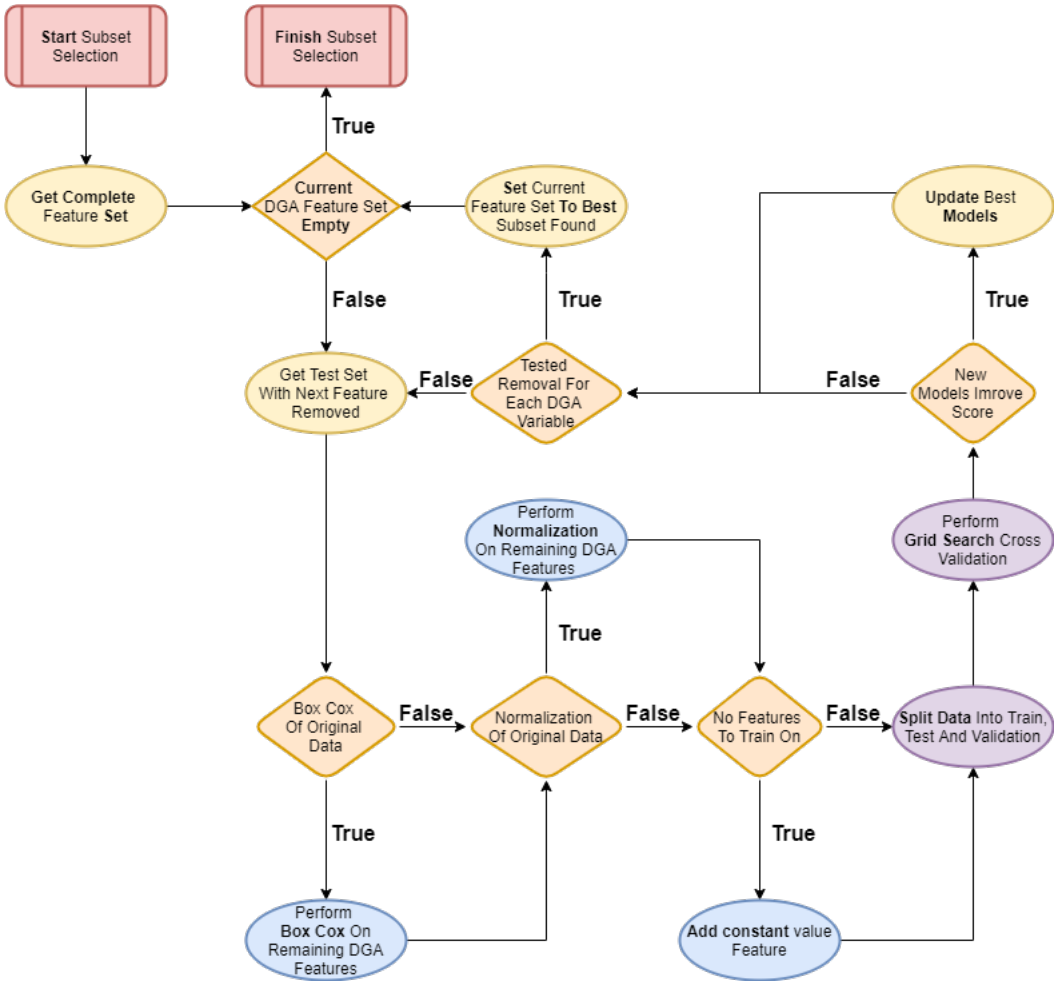

Figure 4.34: Flowchart of the GBE method

The first step in this method is to obtain the full feature set, which in this case contains all DGA variables and any support variable (Oils, Date, etc.). Then we obtain a new set with one of the DGA

features removed.

As we pointed out on the prepossessing section, the DGA variables were not immediately transformed, with this being delayed until a later moment. As at this point we know which DGA variables will be predicted and which will be used as features, now we perform the transformations of those that are not targets. This means that if the dataset was transformed via Box Cox or Normalization, so are now the new DGA features. An important step is to add a constant value in case no variable remains. This only happens in the last iteration if no support features are used.

Completing our set of DGA variable transformations we split the dataset into train, test and validation sets. We follow this by doing grid search cross validation, which will be detailed later, using the train and validation sets. Once this search process is complete we check this iteration's best model's performance using the test set and if it outperforms the best previously found we update our list containing the best models and subsets with the new model and its utilized subset.

Now, if there are still variables whose removal has not been tested we proceed with the next one, otherwise, we change our base feature set with the best one found this iteration. In case all DGA variables have now been removed, which of course does not happen in the first iteration, the selection process finishes and the best models and subsets for each subset size are returned.

Thus, to summarize, we start with the full DGA feature set and for each variable we move it from features to targets. Then we apply Box Cox, normalization or add a constant column as needed. This is followed by grid search cross validation to find the best models. Finally we update the set with the best subset found and do the next iteration with it.

4.3.2 Grid Search

In the previous part we referred that grid search cross validation is being used during the subset selection process. Although both grid search and cross validation are very common algorithms in data science and machine learning tasks we will still go into further detail on how this process is done, due to its central role on our methodology.

Grid search combs through a parameter space in a grid pattern, testing all combinations of the supplied parameters. In order to obtain the results for each parameter combination and compare them, we selected cross validation, which allows us to get robust performance metrics for our limited dataset. Cross validation splits the dataset into a number of folds, all with similar or equal sizes and

then utilizes one fold as a test set and all others combined as a training set. Each fold is selected in turn to be tested, thus obtaining multiple measures of models trained and then tested on slightly sets. The performance measures for all are averaged, therefore creating robust performance metrics while utilizing the dataset to its full extent.

In our case we set the number of folds to 5, not testing any other values as trying other parameters took precedence in importance, due to greater expected impacts on performance. As our objective is do identify general DGA gas subsets and respective models that can be used to reobtain discarded variables, the importance of each predicted DGA variable is equal, and with them having different scales, a scale independent performance metric was required for comparisons. Thus, we opted for the minimum of the R-squared values for all targets, fulfilling all the stated requisites and, as such, guiding the search to the improvement of the worst performing variable predictions.

In our grid search approach a train and validation set are supplied, with the train set being used for cross validation, while the validation is set aside for selecting amongst the different model types, 6 of which were tested in total.

Figure 4.35 shows an example of our grid search parameter data structure. We can see that it is a list, where each element contains the grid for one model. Each of these is itself a list, containing a set of individual parameter grids. Each grid set is done in such a way to avoid overlapping the search, thus preventing the testing of multiple equivalent parameter configurations, or to bound computational time.

To better explain this with one example we point out the first element of this data structure pertaining to the parameters of the Support Vector Regression (SVR) model, where 3 separate grids were created. The first and second elements were separated due to computational limitations, as the value of the hyperparameter C increases computation time far more for the linear kernel, thus needing to be limited. The third element was separated as the hyperparameters of degree and gamma only affect the polynomial kernel and as such, setting them for the others would lead to wasted time trying equivalent combinations.

The last thing that is important to address is reason as to why we used grid search rather than random search, which is known to achieve comparable results. Our motives are three fold. First, utilizing a static grid search more easily allows comparing results as these are less affected by randomness. The second reason is that random search would make it harder to pinpoint reasonable parameter ranges, as even using randomness the minima, maxima or distribution altering parameters

```
modelParameters = [
   [
        {
            "estimator__kernel": ["linear"],
            "estimator__C": [1,20,50,200,500,1000,1e4],
        },
        {
            "estimator__C": [1e0,1e1,1e2,1e3,1e4,1e5],
        },
        {
            "estimator__kernel": ["poly"],
            "estimator__degree": [1,2,3,4,5,7,10],
            "estimator__gamma": ["scale","auto"],
            "estimator__C": [1e0,1e1,1e2,1e3],
        }
    ],
    [
        {
            "estimator__hidden_layer_sizes": [(128,64),(256,128),(256,128,64)],
            "estimator__activation": ["tanh", "relu"],
            "estimator__batch_size": [32,64,128],
            "estimator__max_iter": [100,200],
        },
    ],
    [
        {
            "hidden_layer_sizes": [(32),(128,64),(128,64,32),(256,128)],
            "activation": ["tanh", "relu"],
            "batch_size": [8,16,32,64],
            "max_iter": [500],
        },
    ],
    [
        {
            "estimator__n_estimators": [100,500,1000,5000,10000],
            "estimator__max_features": ["sqrt","log2"],
        },
    ],
    [
        {
            "n_estimators": [100,500,1000],
            "max_features": ["sqrt","log2"],
        },
    ],
    [
        {
            "estimator__learning_rate": [0.01],
            "estimator__criterion": ["friedman_mse","mse"],
            "estimator__min_samples_split": [2,5,10,20],
            "estimator__min_samples_leaf": [1,2,5,10],
            "estimator__max_features": ["sqrt","log2"],
        },
    ],
]
```

Figure 4.35: Data structure of the hyperparameters for grid search cross validation

would need to be set. Finally, as many parameters affect computational time, randomly assigning them would make it too unpredictable, with for example almost all Multi Layer Perceptron (MLP) hyperparameters greatly affecting execution time, and as such, differences of up to hundreds of times between tests would not be out of the ordinary.

4.3.3 Models

The last thing that we need to address in this section is what models and respective hyperparameters are being trained and tested. The problem at hand is of a multi-output regression nature and as such we need models that can predict multiple targets.

This as such divides our models into two kinds. Those that have multiple output capabilities by default and those that do not. Although it might seem that models that do not allow multiple outputs are not suitable, there is a simple method to allow their usage. This method encompasses training a model for each output and then handling these in a group. The main limitation of this approach is that the models cannot take advantage of the relations between targets and as such might obtain worse results.

As mentioned in the previous subsection, a total of 6 models were tested. From these two have intrinsically multi-output capabilities, which are MLPs and Random Forests (RFs). From the remaining 4 models, two of them are the aforementioned ones but trained in the non multi-output manner referenced before. The other two models are the already mentioned SVR and the Gradient Boosting Regressor (GBR).

This model selection was chosen in order to cover a relatively large space of model types, with each of them being trained, storing information and making predictions in substantially different ways.

The final thing we would like to note is that the model types do not have to be the same for each subset size, with for example, the selected model that predicts all DGA variables (uses only support features) being a SVR, while the one that only predicts H2 being an MLP. The kinds of results, subsets and models obtained will, however, only be detailed in chapters 5 and 6.

4.4 Summary

Throughout this chapter we presented the bulk of our development work. We started by thoroughly describing the dataset and restating the problem. A set of graphical, statistical, univariate, bivariate, and regression analyses were performed. From these we identified the problems, limitations and strengths of the dataset.

Armed with this knowledge we proceeded with the crucial preprocessing steps. We started with value corrections, type corrections and missing value imputation, for which we developed a

novel imputation approach, RSI (and the categorical equivalent CSI). Then we performed data transformations, where we reduced and combined variables, changed their distributions and modified their scale.

With a full understanding of our dataset and its preparation completed, we started detailing our modeling approach, where the usage of GBE and GFS algorithms to find the DGA gas subsets was done. We followed with the implementation of grid search cross validation for selecting amongst models and their respective hyperparameters. And finally, we performed a description and selection of the model types, from which SVRs, MLPs, RFs, GBRs and two variants were employed.

All these steps required the selection of hyperparameters, parameters, settings or their respective ranges, as well as, the way in which these were tested, compared and finally selected will be presented in the following chapter.

Chapter 5

Tests And Validation

In order to obtain the best possible model performance, the fine tuning of the preprocessing steps and of the prediction methods employed is of great importance. Although this process usually does not heavily impact the size of the code base produced or the number of results presented, its presence is heavily felt in the amount of time taken setting up and testing a huge amount of preprocessing and model configurations. We will, as such, start this chapter by explaining our testing approach. However, with only the testing and fine tuning of the different aspects of our approach we still cannot be convinced of its efficacy. Thus, another extremely important step is the validation of the obtained results. We dedicate the latter part of this chapter to detail our validation steps and methods, which mostly fall within the testing of our obtained models on real problems.

5.1 Test Configurations

As we have presented on the previous chapter, the possible number of test configurations is huge, with multiple transformations that can be added or removed, several imputation steps, each with many parameters, a large amount of models and respective hyperparameters and more. Given this sheer amount of possible combinations, it simply is not feasible to test them all. As such, for the most part, each setting, parameter or hyperparameter is tested and tuned in isolation, with only a few instances where multiple are modified together to identify relations between them.

We have divided this section by the multiple types of settings that we had to tune and for better comprehension we have done so in the order in which it was done. Starting, as such, with the selection of the imputation method's hyperparameters.

5.1.1 Imputation Algorithms' Hyperparameters

Before any modeling can be done or even before it makes sense to perform any data transformations, the problem of missing values must be handled. Therefore, choosing the best hyperparameters for each of the implemented imputation algorithms was required. As mentioned in the previous chapter, 3 types of imputation methods for each of the imputations tasks were implemented. The simplest ones, mean (for numerical and conditional strings) and mode (for categorical), do not require any tuning, as such leaving us with two other methods to look at.

The simple regression imputation method, used for all but the categorical imputation task, is from the remaining two the simplest. It predicts one variable while using all others as features. For this K Nearest Neighbours (KNN) was employed. KNN in this context has two very important hyperparameters to tune, the number k of neighbours and the distance metric used. Each of these was tested independently. These tests were done via a graphical analysis, where most of the best and some worst results were looked at to identify over or under fitting. For the value of k, the range of values from 1 through 7 were tested, with the value 5 identified as the best. In the same way 3 distance metrics were tested, euclidean, manhattan and mahalanobis, with euclidean proving to be the best. These hyperparameters are only applicable for the case of numerical and conditional string imputation, with for the categorical imputation only a value k of 1 being applicable and distance metrics making no noticeable difference.

The second imputation method that requires hyperparameter selection is our novel Regression Sampling Imputation (RSI). We recall that this method, unlike the previous, predicts one variable from a subset of n others, selecting the models and subsets at random, weighing them by their R-squared performance. Furthermore, we point out that this method implements 4 different regression models, polynomial, exponential, logarithmic and cubic spline, from which only polynomial and cubic spline have tunable parameters. Much like for regression imputation, the selection was done via graphical analysis, with all parameters being selected by inspecting regression results. We started by selecting the value of n for the subset size. With n of 1, the total number of regression models to create is 1482[1], while for a value of 2, 27417. Further increasing this value turns the computation unfeasible, running into both time and memory limitations.

[1]The formula to calculate these values is 39×38Cn. With n=3 the obtained number is 329004.

5.1.1.1 Polynomial Imputation

For polynomial regression one parameter that has to be tuned is the maximum degree, while another is the method for selecting the best degree. Each method has itself other parameters that have to be selected.

The two tested methods for the polynomial degree selection were cross validation and a statistical F-test, like those found in an Analysis Of Variance (ANOVA) table. The initial maximum degree of the polynomial was set to 10. Cross validation has the parameter for the number of folds, but only 5 were used at any time. The F-test compares two models, taking into consideration both the performance and number of parameters. A low F-value indicates that the more complex model is statistically better. A threshold for this "low value" has to be set, were from the few tested a very low 0.0001 was selected, greatly reducing the degree of the obtained models. From both methods, the F-test obtained better results. Furthermore, given the high bounding power of this method, the selection of a maximum degree became unnecessary, and as such, only for computational purposes, it was left at the original value of 10.

5.1.1.2 Cubic Spline Imputation

For cubic spline, the parameters that need to be selected are the number of knots, their locations and the degree of each section between knots. Due to both the extreme amount of possible feature and target combinations for the regression, which impossibilitated the manual selection of the knots, and the wide range of possible distributions, a way to automatically select the knot locations was devised. This method places the knots at equidistant quantiles, thus ensuring that the most dense regions contain the highest model complexity. In order to reduce outfitting, while creating smooth regression curves a degree of 3 for each section was selected, with no others tested. Finally, the number of knots was selected, with the values of 2, 3 and 4 being tried and 3 proving to be the best.

5.1.2 The Initial Configuration

Once the selection of the parameters for the imputation process was completed and the process itself concluded, the initial fine tuning of the models hyperparameter ranges for testing was undertaken. But before we explain the processes by which such was done we must detail the initial configuration of the dataset. That is, to describe which imputation techniques, transformations and respective

parameters were selected.

At this point it bears mentioning that the modeling approach was conducted on two different datasets stemming from the original one. The first one is simply the original dataset with all the preprocessing steps applied to it. The second one, on the other hand, had it's support variables removed, thus containing only the Dissolved Gas Analysis (DGA) ones. The reasoning for this is twofold. First, testing a dataset without the support variables allows us to assess their degree of usefulness. Second, in order to prove the generality of this technique, only using a dataset with support variables that might in most circumstances be unavailable or inapplicable is not reliable. Thus, by using only DGA variables we are able to show that the results work when only the minimal amount of information (for it to make sense for our method to be applied) is present.

With this being said, the initial configuration was setup as follows. For the numerical and conditional string imputation, the RSI method was used, while for conditional imputation, Classification Sampling Imputation (CSI) was used. Principal Component Analysis (PCA) was employed with the components originating from the sets of variables that have correlation of 0.8 or greater amongst them selves. Thus, 3 components were created. The first one has Cor and Acidity Index. The second one has Part4 and Part6. And the third one has the one hot encoded columns for Shell oil brand and Diala D oil type. Extra trees dimentionality reduction was not applied, while Box Cox and normalization were.

We selected this initial configuration because we expected it to be the best one. However, tuning the model hyperparamters on one configuration leads to the bias towards it when testing others. Although as the parameters where only tuned for the full dataset and copied for the DGA only dataset, we would expect that this bias if far smaller for this second one, thus either validating or invalidating our preprocessing choices.

5.1.3 Initial Hyperparameter Search

With the initial dataset preprocessing configuration detailed, the selection and fine tuning of model hyperparameters proceeded. Before analysing the methods by which we selected the parameters, an implementation detail has to be explained. This is that the models employed all came from the python sklearn library, and thus, if a parameter is not referred or is somehow ignored, its value is that of the default.

Each model was tuned at a time, with each parameter itself also being selected independently.

Starting with all the default values, one promising hyperparameter was selected and a broad range of values for it tested. If the selected best values landed far from any extreme (minimum or maximum), the range was reduced, while if the extreme proved to be the best value, a new further value was added. This process was done iteratively until either the selected values fell totally inside our ranges and were on similar scales to those allowed or until computational times became excessive, either due to an extremely large amount of combinations to test, caused by a large grid, or due to the effect a parameter value has in increasing time or memory usage (like the Random Forest (RF) number of estimators parameter).

Regardless of the selected values if a hyperparameter was found to have only negligible effects on model performance it was once again removed from the hyperparameter test grid. On the other hand, if it was found to have significant results it was permanently added and, as such, always tested from then on in combination with others. The magnitude of its effects was obtained by looking at the test performance for each subset size, and by identifying changes in any of them.

In order to combat the previously mentioned preprocessing bias caused by this approach, the ranges were kept very broad and not very detailed. Thus, unless a dataset would require wildly different ranges for optimal hyperparameters, any improvements should still be noticeable and, as such, preprocessing configurations comparable.

Thus, applying this process to all models, the hyperparameters that were found to be worthwhile tuning were:

For Support Vector Regressions (SVRs):

- Kernel - From which linear, polynomial and rfb were found to be the only ones obtaining good results.
- C - A regularization parameter that greatly impacts computational time.
- Degree - Only applicable for polynomial kernel.
- Gamma - Kernel coefficient only applicable for polynomial and rfb kernel. Value different from default found to be useful only for polynomial.

For Multi Layer Perceptrons (MLPs):

- Hidden layer size - Size and number of hidden layers.
- Activation function - Function used in each neuron. Only the Rectified Linear Unit (ReLu) and tanh were found to be useful.

- Batch size - Minibatch size for training. Smaller values increase computational time.
- Maximum iterations - Maximum number of iterations before automatically stopping. Might stop earlier if the early stopping criteria is met.

For RFs:

- Number of estimators - Number of different trees trained. Proportionally impacts computational time and memory usage.
- Maximum features - Method for selecting maximum number of features per tree. Either logarithm or square root of total number o features.

For Gradient Boosting Regressors (GBRs):

- Learning rate - Greatly impacts computational time.
- Criterion - Method by which splits are compared. Only Friedman Mean Squared Error (MSE) and regular MSE were found to be useful.
- Minimum samples per split.
- Minimum samples per leaf.
- Maximum features - Method for selecting maximum number of features per tree. Either logarithm or square root of total number o features.

From this list of hyperparameters there are a few surprising things to note. Firstly, some of the MLPs hyperparameters that are known to be important and to usually improve results were found to not have a significant impact on performance. Those such as momentum and other parameters related to its functioning, learning rate and regularization terms. Then the RF and GBR methods, which both use Decision Trees (DTs) had quite different results in terms of the hyperparameters found to be relevant. With criterion, minimum samples per leaf or minimum samples per split, being found to need tuning in the later but not in the former.

One final thing that bears mentioning is that although both MLPs and RFs have two versions, one using their intrinsic multiple output capabilities and another not, the hyperparameters selected for each version were the exact same.

5.1.4 Preprocessing Fine Tuning

With a reasonable set of hyperparameters selected and a viable broad range for their values identified, the tuning of the preprocessing steps followed. But before describing our methodology and findings, it is important to mention how we compared different setups. As there are multiple subset sizes, each with their own regression metrics, no single value can be used for comparisons, since an improvement of the results for one size can lead to the detriment of another. As such, the comparison proceeded by selecting a minimum required R-squared, then identifying the size of the smallest subset that attains it. One set is considered better if its minimum size, achieving a R-squared greater than the selected threshold, is smaller. In case the sizes found are the same, the performances for each of these minimum sets are used as the comparison.

For the full dataset a threshold of 0.7 was selected as it is considered a high value for R-squared. On the other hand, the DGA only dataset is not capable of reducing the size as much, but achieves better results on the subsets of greater dimension. Due to this, a value of 0.9, which is very high, was selected. This, of course, means that the full and DGA only datasets might have different optimal preprocessing strategies. This, however is not a problem as both can be handled independently.

With the subject of result comparison addressed, our preprocessing fine tuning went as follows. We started with the full dataset, tuning individually, one at a time, each of the preprocessing steps. We began with the imputation steps, which are 3 in total, each having 3 different possible methods. For the numerical value imputation, the mean, regression and RSI methods were tested, with RSI proving the best. The same was done for the conditional string imputation, where, once again, RSI proved to be the best. For categorical imputation, the mode, classification and CSI strategies were tested, with CSI achieving the best results.

Then we moved to the tuning of the transformations. Starting with PCA, several alternatives were tested. The first alternative was to totally disable it. All others involve setting the correlation threshold for variable combination to various values, where 0.4, 0.5, 0.6, 0.7 and 0.8 tested. In the end, performing PCA with a threshold of 0.8 obtained the best performance.

The final 3 tested variations all involved disabling or enabling a transformation. Using the extra trees for dimentionality reduction was found not to be helpful, probably as the PCA was already doing most of the required dimentionality reduction. Box Cox presented the most interesting case. While disabling it substantially improved results for the higher subset dimensions, it however hindered those for smaller dimensions. Furthermore disabling Box Cox roughly doubled modeling times, from about

6 to 12 hours. Finally, disabling normalization was tried, however, an increase of over 100 times the computational time was observed and thus it was not possible to finish the model training.

As no change to the steps present in the original configuration improved results, continuing with more tests in the same way was not doable and, as such, a couple different approaches were performed. The first of which, involved changing all the imputation steps simultaneously from our sampling methods to simple regression and classification. Mostly the same outcomes were observed in respect to the previous approach, with the exception that for PCA the best threshold was decreased from 0.8 to 0.7. However, as these results were still worse than the original ones, they were discarded. The final performed experiment is very similar to the previous one, replacing the imputation steps with the mean and mode. Once more, the results were virtually identical, with the exception being that disabling the PCA and enabling the extra trees dimentionality reduction was found to be optimal. But, once again, the results were still worse that those from the original configuration.

For the DGA only dataset the same set of tests were performed, with exactly the same conclusions and thus will not be detailed. Thus, we reached the conclusion that our original preprocessing configuration was the best and, as such, we proceeded with further fine tuning of the models hyperparameters.

5.1.5 Final Model Hyperparameter Tuning

As the final part of our parameter tuning approach, in order to obtain the best results possible, we went back to model hyperparameter tuning. Unlike for the initial one, whose objective was to be both broad and fast enough to find the best preprocessing configuration, this time our goal is to greatly detail it, to find the truly best hyperparameters.

The approach taken in this final tuning is much alike the previous one, changing the ranges of the hyperparameters, one parameter and model at a time. One of the main differences is that the computational time limit was raised from a cap of roughly 6 hours to 72 hours (3 days). Thus, a more granular grid, where more costly hyperparameter configurations were permitted, was created. Another important difference is that unlike the previous hyperparameter tuning, this time no new hyperparameters were added, thus leaving only those previously found.

5.2 Validation

Up until this point the entirety of our results and performance comparisons rely on the value of R-squared obtained by the regressions. Although this is quite sufficient for model tuning and performance improvement, these results are not our final objective and due to some features of our approach are not totally reliable.

The attentive reader might have noticed in the previous section that the selection of the tuning parameters was in some part performed directly by comparing the results gathered from the test set, thus inducing some, however low, degree of bias and possible overfitting.

Another important aspect to note is that our work has a few different objectives. One is to identify subsets of DGA gases to guide and aid in the selection and allocation of resources for the most important sensing equipment and lab analysis. But another is to create models which with a limited set of DGA variables can construct a full set of them for usage in a variety of real problems. Thus, our work is intended in part as support to other modeling, predictive and machine learning activities.

Taking these two important objectives into consideration a somewhat extensive validation approach was developed. It first includes an outlier/extreme value analysis, and then the application of our models and respectively created datasets to a few relatively common DGA failure diagnosis techniques.

5.2.1 Outlier Validation

The vast majority of DGA problems include some sort of anomaly analysis, whether for failure prediction or diagnosis, abnormal events, power spikes or others. As such, the first step of our result validation approach involves verifying the integrity of the extremes obtained when compared with the original dataset. For this two separate tests were performed.

The first one looks at the regression results restricted to the points that are found to be outliers. This means, very simply, that the entries for which a variable is found to be an outlier in the original distribution are obtained an then the same metrics as for those used in regression are employed, thus obtaining the R-squared only for the points that are outliers in either the original or in the predicted distributions. In order to find and select these outliers the same method employed to find them in the original regression results is employed, which is also the same used for box plots.

The second test works in a very similar way to the first one, with the same outlier identification method. The primary difference is that instead of it being treated as a regression problem it is done as a binary classification problem. What this means is that the set of points is either marked as being an outlier or not and then a set of binary classification metrics is employed to compare the real and predicted outlier values for the DGA variables.

For both of these outlier validation approaches, the results obtained, as well as, the metrics that were employed and their meaning will be detailed in the next chapter.

5.2.2 Problem Validation

For the second set of validations tests we employed a group of common DGA failure diagnosis methods. Our dataset does not contain any real failure data, and as such a comparison with real values is not possible. Furthermore, the methods that we will be utilizing are usually only applicable when a fault has already been detected and not in general for any given sample. Although both mentioned factors indicate that these validation tests are not totally indicative of a real application, they should still be more than enough to get a very good approximation of the behaviour in actual operating conditions.

5.2.2.1 Duval's Triangle

As mentioned in chapter 3, the Duval's triangle method is amongst the most common fault diagnosis approaches for DGA data. In order to apply this method for our validation methodology, its predictions were obtained both for the original data and for the predicted data generated by our models. Then the Duval's predictions for each were compared with a multi class classification problem approach and its respective applicable metrics.

Although the implementation is quite simple, the details for the specific values and ranges to be used were quite difficult to come by, as this method is usually applied graphically rather than algorithmically. These details can, for the interested reader, be found on chapter 2.

5.2.2.2 International Electrotechnical Commission Table

Another quite common DGA fault diagnosis method is the International Electrotechnical Commission (IEC) table, where a set of gas concentration ratios is employed. As the methodology used is very

similar to that of the Duval's triangle, to avoid repetition it will not be detailed. Once again, the details regarding the value ranges in the IEC table can be found of chapter 2.

5.2.2.3 Rogers Ratio

Rogers ratio also belongs to this set of common fault diagnosis approaches. Like the IEC table it uses gas concentration ratios to make its predictions. Our methodology to handle this problem is identical to the previous ones.

5.2.2.4 Key Gas Method

The key gas method is the final validation approach tested, once more belonging to the DGA fault diagnosis set of approaches and with our usage of it being identical to the ones before. This one however, requires some further detailing due to the fact that for the most part it is less formal than the ones presented before. What this means is that there are no hard value and rules for the diagnosis and instead it relies on the expertise and graphical analysis of an individual. We, nevertheless, searched for reference values, which we remind are a rule of thumb rather than precise and well tested metrics. These reference values were obtained from the work of Londo and Çelo [23]. Here the value of Total Dissolved Combustible Gas (TDCG) is obtained by summing the concentrations of Hydrogen (H2), Methane (CH4), Ethane (C2H4), Acetylene (C2H6), Ethylene (C2H2) and Carbon Monoxide (CO). Then the percentage of contribution for this amount from each of these gases is obtained, and for values above certain limits a fault is predicted.

For a CO concentration above 90%, overheated cellulose is obtained as the fault. For a C2H4 concentration above 60%, overheated oil is predicted. A value for H2 greater than 80% gives corona in oil as the cause. Finally, for a concentration of C2H2 greater than 30%, arcing in oil is predicted. If none of these prove true, no fault is given.

We would finally like to note that the results for all the validation problems will be, much like for the outlier validation, detailed in the next chapter.

5.3 Summary

Throughout this chapter we presented our combination of fine tuning tests and configurations, as well as, the various result validation approaches implemented.

We have independently selected each parameter, model hyperparameter and preprocessing setting in order to reduce the total number of required configuration tests. We explained how we started by selecting the imputation method's parameters, then moved to the initial preprocessing configuration and hyperparameter range selection. We detailed how each preprocesing step was chosen, its impacts compared and how we finally obtained the final very detailed model hyperparameter grid for our grid search approach.

We followed the explaining of our testing methodology by the detailing of our result validation approach. An outlier analisys was performed, both as a regression problem and as a binary classification one, while a set of common DGA failure diagnosis methods were deployed to compare the obtained results for both the original and predicted data.

Throughout the next chapter we will be analysing the best results obtained by our test tuning approach and the respective validation metrics obtained.

Chapter 6

Results

In this chapter we present the totality of the results obtained from the best models in each of two datasets. First, we look at what was obtained for the full dataset, including the general regression performance metrics, the results for the outlier validation approach and finally those for the various problem validation steps employed. Then, utilizing an identical approach we perform the same analysis on the Dissolved Gas Analysis (DGA) only dataset.

6.1 Full Dataset

The full dataset contains, as was mentioned a few times in the previous chapters, the complete feature set from the original data, including furanic compounds, sample details and particle data. As our preprocessing was fully applied to this set, a few features might be missing due to the Principal Component Analysis (PCA) steps.

Before proceeding with the analysis of the obtained metrics we will here explain what were the selected models and subsets.

For the smallest subset of size 0:

- No DGA gas is present in the subset and thus all are predicted.
- A Random Forest (RF) with synthetic multi output capabilities was obtained.

For the subset of size 1:

- Hydrogen (H2) was added to the subset.

- The selected model was again a RF with synthetic multi output capabilities.

For the subset of size 2:

- Ethylene (C2H2) was added to the subset.
- The selected model was still a RF with synthetic multi output capabilities.

For the subset of size 3:

- Methane (CH4) was added to the subset.
- The selected model was once more a RF with synthetic multi output capabilities.

For the subset of size 4:

- Carbon Dioxide (CO2) was added to the subset.
- The selected model changed to a Gradient Boosting Regressor (GBR).

For the subset of size 5:

- Nitrogen (N2) was added to the subset.
- The selected model changed back to a RF with synthetic multi output capabilities.

For the subset of size 6:

- Acetylene (C2H6) was added to the subset.
- The selected model remained a RF, but this time with its standard multi output capabilities.

For the subset of size 7:

- Carbon Monoxide (CO) was added to the subset.
- The selected model changed back again to a RF with synthetic multi output capabilities.

For the subset of size 8:

- Ethane (C2H4) was added to the subset.
- The selected model was, this time, a Multi Layer Perceptron (MLP) with synthetic multi output capabilities.

Thus, for these results, Oxygen (O2) is always predicted and never a feature, while the Support Vector Regression (SVR) and MLP with default multi output capabilities are never found to be the best.

Now that we have detailed the obtained subsets and respective models we will start with the explanation and analysis of the obtained results.

6.1.1 Performance Metrics

The first set of result metrics we will analyse are those from the general regression. Although up to this point we have not detailed them, similar tables as to table 6.1 were employed in the test section to discern and compare different models and preprocessing configurations.

We would like to point the readers attention, once again to the aforementioned table, where some explanation of the contents and organization is in order. The first thing to note is the meaning of both axis. Each row of the table corresponds to a DGA gas subset of a given size, while each column corresponds to one of the variables, with the one exception to this rule being the last column, which simply contains the worst results for each row.

Each cell of this table contains a relatively high amount of information. Those whose content is just "N/A" imply that the variable was not predicted for the subset size, thus being used as a feature in the models. The other cells contain 3 different values separated by a bar. The first value corresponds to the R-squared, the second one to Root Mean Squared Error (RMSE), while the third one to Mean Absolute Error (MAE). This aggregation was done in order to greatly reduce the footprint of our results, as otherwise triple the number of tables would be required. Although only the R-squared was used for variable selection, the other two metrics permit us to have a better grasp of the actual amounts of error incurred by our models. The last structural element of our table that requires some explanation is the "worst" column, for which each of the several metrics was picked independently, where, for example, the R-squared might correspond to that of H2, while the RMSE corresponds to CO.

One further thing that we would like to point out before explaining the meaning of the exact values obtained is that our focus will lie entirely of the values of R-squared, and for this 2 important thresholds were defined. The first one is that of 0.7, which is generally considered a high value for this metric, while the second one is that of 0.9, which is an extremely high value.

For the analysis of the obtained values we will start by looking at the first two and last two DGA variables to be excluded from the models. The first variable to be excluded, and as such predicted is that of O2, which would be expected, given its relatively high correlation with other DGA and support variables, making it, as such, relatively trivial to predict. The second one, C2H4 was also somewhat

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	-1.354 | 40.555 | 20.892	0.346 | 77.92 | 32.294	0.31 | 98.289 | 27.262	0.475 | 43.085 | 25.007	-0.504 | 4.596 | 1.889
1	N/A	0.391 | 75.159 | 31.253	0.371 | 93.832 | 27.205	0.54 | 40.318 | 23.731	0.15 | 3.456 | 1.567
2	N/A	0.4 | 74.59 | 30.611	0.426 | 89.626 | 26.912	0.538 | 40.414 | 23.694	N/A
3	N/A	N/A	0.776 | 55.97 | 16.105	0.802 | 26.459 | 13.347	N/A
4	N/A	N/A	0.779 | 55.635 | 16.224	0.798 | 26.708 | 13.51	N/A
5	N/A	N/A	0.783 | 55.155 | 15.976	0.801 | 26.528 | 13.284	N/A
6	N/A	N/A	0.83 | 48.834 | 14.408	N/A	N/A
7	N/A	N/A	0.83 | 48.773 | 14.28	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	0.73 | 89.436 | 61.25	0.762 | 0.989 | 0.687	0.686 | 4.159 | 3.234	0.742 | 9.795 | 8.077	-1.354 | 98.289 | 61.25
1	0.745 | 86.954 | 59.922	0.767 | 0.978 | 0.675	0.689 | 4.138 | 3.211	0.755 | 9.555 | 7.935	0.15 | 93.832 | 59.922
2	0.753 | 85.607 | 59.188	0.761 | 0.991 | 0.679	0.703 | 4.039 | 3.141	0.76 | 9.459 | 7.817	0.4 | 89.626 | 59.188
3	0.774 | 81.897 | 56.241	0.794 | 0.921 | 0.623	0.761 | 3.626 | 2.869	0.762 | 9.421 | 7.769	0.761 | 81.897 | 56.241
4	0.77 | 82.56 | 56.484	N/A	0.771 | 3.546 | 2.833	0.765 | 9.353 | 7.761	0.765 | 82.56 | 56.484
5	0.785 | 79.926 | 54.233	N/A	0.853 | 2.844 | 2.219	N/A	0.783 | 79.926 | 54.233
6	0.785 | 79.805 | 54.653	N/A	0.858 | 2.797 | 2.19	N/A	0.785 | 79.805 | 54.653
7	N/A	N/A	0.87 | 2.671 | 2.076	N/A	0.83 | 48.773 | 14.28
8	N/A	N/A	0.908 | 2.25 | 1.632	N/A	0.908 | 2.25 | 1.632

Table 6.1: Table of the Full dataset general regression metrics. Results in the form R-squared | RMSE | MAE

expected, due to its high correlation with CH4. Although in this case, neither have obvious relations with any support variables, thus the removal of one implies that the other will be substantially harder to predict, thus making its pair CH4 one of the last ones to be excluded.

Conversely, H2 was both the last one to be removed and the one for which the worst results were achieved. This was observed in virtually all results, with this variable having by far the weakest relationships with any other variables. The second to last variable to be removed, C2H2 is in a similar situation to the previous. While not to such an extreme as H2, the strength of its relations with other variables is also very low. These results imply that unless some new support variable is found that is heavily related with these two variables, they will undoubtedly be the most likely candidates for inclusion in any DGA gas sampling set.

Two very important observations to make are that if we restricts ourselves to the very conservative 0.9 R-squared threshold, the size of the viable DGA gas subset found is only one less than the total. On the other hand, the other observation is that the more reasonable value of 0.7, allows us, if we are willing to take some more chances, to reduce the number of measured DGA gases from 9 to 3, which is an impressive reduction.

Although we will get to it in the next sections we would like to take this opportunity to already

point out that the rest of the results, for the most part, validate the conclusions that were here obtained.

One final thing that is important to note is that by using this method we found that the most contributing factor to performance was the selected subsets. This was far more influential than any preprocessing step or model and respective hyperparameters. It is due to this that the two subset selection approaches had wildly different performances, stemming from their differing obtained subsets.

6.1.2 Outlier Regression Validation

Once the model fine tuning was completed, where the selection of each step and parameter was done by comparing the previously shown performance tables, the first of several validation steps was performed.

On table 6.2 we can see the results for the outlier regression validation approach, which is structured identically to the one from the previous section, including the metrics and respective ordering. Only one thing is worth mentioning that did not happen previously. This is the presence of "nan" values for the metrics. These mean that for whatever reason they were not applicable. In this case it happens because of the inexistence of outlier values in N2.

The first thing we notice in this table is the very poor results for the "worst" column. These are quite concerning and somewhat unexpected, thus warranting further analysis. We see that these come exclusively from the O2 variable, which being the first one to be selected should exhibit the best performance. We can also notice that although all other variables perform slightly worse than what was seen in table from the previous section, which is to be expected, no other exhibits such a steep drop off in performance.

Given these factors we decided to analyse the outlier values of O2, upon which the culprit was immediately spotted. The problem lies in the fact that all but a couple values lie extremely close to 25, with only tiny variations. Thus the best model fit would be very close to a constant value. It is well known that R-squared measures the improvement of a model in relation to that obtained by the mean of the observations. As the best model would be close to just predicting the mean, almost no improvement can be obtained, thus making the initial value of 0.8 surprisingly high.

Taking this into consideration, the results are far from as bad as they first appear. Bar the case

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	0.88 | 79.84 | 203.01	0.62 | 40.14 | 77.25	0.75 | 30.1 | 76.21	0.84 | 40.36 | 59.17	0.95 | 3.38 | 10.42
1	N/A	0.65 | 37.33 | 73.74	0.79 | 28.24 | 69.38	0.85 | 38.41 | 57.68	0.94 | 3.56 | 11.81
2	N/A	0.67 | 36.99 | 72.37	0.79 | 27.86 | 68.83	0.86 | 37.22 | 55.71	N/A
3	N/A	N/A	0.91 | 21.7 | 48.7	0.92 | 23.93 | 39.75	N/A
4	N/A	N/A	0.95 | 20.89 | 34.26	0.96 | 19.01 | 26.98	N/A
5	N/A	N/A	0.91 | 20.82 | 47.1	0.92 | 23.2 | 39.71	N/A
6	N/A	N/A	0.91 | 21.25 | 47.31	N/A	N/A
7	N/A	N/A	0.93 | 20.78 | 44.1	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	0.39 | 111.98 | 144.45	0.57 | 1.03 | 1.29	-1.12 | 3.09 | 3.76	nan | nan | nan	-1.12 | 111.98 | 203.01
1	0.45 | 104.67 | 136.47	0.58 | 1.02 | 1.28	-1.03 | 3.08 | 3.68	nan | nan | nan	-1.03 | 104.67 | 136.47
2	0.47 | 103.85 | 134.7	0.58 | 1.01 | 1.28	-0.81 | 2.87 | 3.47	nan | nan | nan	-0.81 | 103.85 | 134.7
3	0.5 | 100.94 | 129.82	0.6 | 0.99 | 1.26	-0.46 | 2.54 | 3.12	nan | nan | nan	-0.46 | 100.94 | 129.82
4	0.68 | 77.39 | 104.53	N/A	-0.31 | 2.55 | 3.12	nan | nan | nan	-0.31 | 77.39 | 104.53
5	0.57 | 92.45 | 121.22	N/A	0.1 | 2.04 | 2.46	N/A	0.1 | 92.45 | 121.22
6	0.58 | 90.57 | 118.91	N/A	-0.59 | 2.71 | 3.25	N/A	-0.59 | 90.57 | 118.91
7	N/A	N/A	0.2 | 1.86 | 2.31	N/A	0.2 | 20.78 | 44.1
8	N/A	N/A	0.8 | 0.84 | 1.2	N/A	0.8 | 0.84 | 1.2

Table 6.2: Table of the Full dataset outlier regression metrics. Results in the form R-squared | RMSE | MAE

in which very accurate predictions of outlier values are required, which seems unlikely, these poor results seem otherwise inconsequential. In fact, we can turn to the MAE measures, where we see that at worst an error of 3 is expected, which given the measured values of 25 is relatively low, and in accordance to expectations.

Given these results the next validation analysis is even more crucial, as if similar performance is found our methodology would be in severe jeopardy.

6.1.3 Outlier Binary Classification Validation

On table 6.3 we can see the results for the outlier binary classification validation. The structure is once again very similar, with only the metrics used being changed as this is now a classification rather than a regression task.

Due to the nature of outliers, the classification here are undoubtedly unbalanced and, as such, although accuracy is still a useful metric it lacks the total required capabilities to properly describe

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	0.568 | 0.676 | 0.617 | 0.94	0.942 | 0.843 | 0.89 | 0.973	0.602 | 0.994 | 0.749 | 0.898	0.972 | 0.711 | 0.821 | 0.971	0.479 | 0.97 | 0.641 | 0.79
1	N/A	0.942 | 0.843 | 0.89 | 0.973	0.595 | 0.994 | 0.744 | 0.895	0.974 | 0.773 | 0.862 | 0.977	0.52 | 0.985 | 0.681 | 0.821
2	N/A	0.965 | 0.828 | 0.892 | 0.974	0.589 | 0.981 | 0.736 | 0.892	0.975 | 0.794 | 0.875 | 0.979	N/A
3	N/A	N/A	0.667 | 0.975 | 0.792 | 0.921	0.939 | 0.948 | 0.944 | 0.989	N/A
4	N/A	N/A	0.567 | 0.88 | 0.69 | 0.878	0.87 | 0.897 | 0.883 | 0.978	N/A
5	N/A	N/A	0.678 | 0.975 | 0.8 | 0.925	0.92 | 0.948 | 0.934 | 0.987	N/A
6	N/A	N/A	0.694 | 0.975 | 0.811 | 0.93	N/A	N/A
7	N/A	N/A	0.692 | 0.968 | 0.807 | 0.929	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	1.0 | 0.717 | 0.835 | 0.987	1.0 | 0.797 | 0.887 | 0.987	1.0 | 0.545 | 0.706 | 0.966	nan | nan | nan | 1.0	0.479 | 0.545 | 0.617 | 0.79
1	0.971 | 0.717 | 0.825 | 0.986	1.0 | 0.812 | 0.897 | 0.988	1.0 | 0.519 | 0.684 | 0.964	nan | nan | nan | 1.0	0.52 | 0.519 | 0.681 | 0.821
2	1.0 | 0.717 | 0.835 | 0.987	1.0 | 0.812 | 0.897 | 0.988	1.0 | 0.558 | 0.717 | 0.967	nan | nan | nan | 1.0	0.589 | 0.558 | 0.717 | 0.892
3	0.971 | 0.717 | 0.825 | 0.986	1.0 | 0.844 | 0.915 | 0.99	1.0 | 0.597 | 0.748 | 0.97	nan | nan | nan | 1.0	0.667 | 0.597 | 0.748 | 0.921
4	0.914 | 0.696 | 0.79 | 0.983	N/A	0.917 | 0.571 | 0.704 | 0.964	nan | nan | nan | 1.0	0.567 | 0.571 | 0.69 | 0.878
5	1.0 | 0.717 | 0.835 | 0.987	N/A	1.0 | 0.675 | 0.806 | 0.976	N/A	0.678 | 0.675 | 0.8 | 0.925
6	0.971 | 0.717 | 0.825 | 0.986	N/A	1.0 | 0.61 | 0.758 | 0.971	N/A	0.694 | 0.61 | 0.758 | 0.93
7	N/A	N/A	1.0 | 0.74 | 0.851 | 0.98	N/A	0.692 | 0.74 | 0.807 | 0.929
8	N/A	N/A	0.934 | 0.922 | 0.928 | 0.989	N/A	0.934 | 0.922 | 0.928 | 0.989

Table 6.3: Table of the Full dataset outlier binary classification metrics. Results in the form Precision | Recall | F1 | Accuracy

the results. Thus, the metrics of precision, recall and F1 were also included.

Unlike for R-squared the metrics of precision, recall, F1 and accuracy do not have any well defined thresholds for bad, good or great values. As such we have defined a set of thresholds that we found to be reasonable for the problem at hand. As this is a binary classification problem, we would expect a quite high accuracy and thus we set the minimum required value for this metric at 0.9. On the other hand, we do not wish to totally sacrifice the results for the minority class of the outliers and, as such, we selected a minimum value of 0.5 for all other metrics.

Upon initial analysis of the "worst" column we see that the minimum subset size that fulfils our requirements is 3. Due to the previously found results we turn our focus on the O2 variable, where we observe that it obtains in all cases the worst recall, while, on the other hand achieving amongst the best precision. This, of course, means that the model is predicting less extreme values than the real ones. Fortunately armed with the knowledge explaining these events and the fact that the latest results are sufficient, we can conclude that up to this point our approach is still viable.

6.1.4 Duval's Triangle

With a somewhat sub-par performance of our models in the outlier validation steps, the viability of our approach lies heavily on the results obtained for our set of problem validation methods, the first of which is the Duval's triangle. We can see the obtained metrics on table 6.4. Here we utilize a deterministic model to make predictions on the sets of variables, unlike for the previous sections, where we analyse the predictions for each individual variable. Due to this, the structure is slightly different, with each column now representing one metric instead of a variable.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.584	0.495	0.6	0.615	0.672	0.513	0.584	0.495
1	0.604	0.505	0.617	0.648	0.688	0.523	0.604	0.505
2	0.672	0.57	0.674	0.749	0.735	0.549	0.672	0.549
3	0.683	0.595	0.68	0.773	0.721	0.573	0.683	0.573
4	0.432	0.463	0.479	0.498	0.569	0.493	0.432	0.432
5	0.68	0.562	0.673	0.764	0.718	0.554	0.68	0.554
6	0.652	0.52	0.644	0.577	0.668	0.523	0.652	0.52
7	0.713	0.586	0.705	0.622	0.718	0.588	0.713	0.586
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6.4: Table of the full dataset Duval's triangle classification metrics.

As we are now dealing with a multiclass classification problem, simple precision or recall are no longer as applicable. Since these metrics compare one target variable to the rest of them, one metric for each class would need to be created. This, combined with the number of metrics and subsets would lead to an unfeasibly large a number of metrics to analyse. Thus, for compactness of the results, we aggregated these by using both macro and weighed averaging. We are, therefore left with a total of 7 metrics, with these being macro precision, weighed precision, macro recall, weighed recall, macro F1, weighed F1 and accuracy (which suffers from the same unbalance problems).

As the weighed version gives more importance to the most common classes than the macro version, in cases where the rare classes obtain worse results, we expect lower values for macro than weighed metrics, with the opposite case, of course, also being a possibility. On the other hand a possibility is that both good and bad results are present in different minority classes simultaneously, thus averaging the results. However, in practise this was never observed, in any of the problem validation approaches.

One final thing that bears mentioning is that unlike for binary classification, Duval's triangle predicts 7 different classes, and thus values of accuracy, precision, recall and F1 would be expected to decrease. Taking this into consideration, we lowered the minimum performance threshold of accuracy to 0.6, while that of all other metrics to 0.4.

In the initial analysis of the aforementioned table we see that these new thresholds have been reached for a minimum subset size of 1. One thing that is however important to note is that due to random variations in the models obtained the subset of size 4 performs worse. We note that Duval's triangle only uses C2H2, C2H4 and CH4, hence the perfect predictions for the subset of size 8 are a result of the only removed variable (O2) not having any impact on predictions and thus only the real original values being used on both instances.

The last thing that is important to note regarding these obtained results is that in all instances the macro metrics indicate worse performance than the weighed ones. What this means is that the minority classes attained, in general, worse results than the most common ones. This is an important observation, as none of the models trained directly on these classes but instead on the DGA values, which implies that the fault types were implicitly modeled. These suspicions are reinforced in the following sections.

6.1.5 International Electrotechnical Commission Table

The second of our problem validation approaches uses the International Electrotechnical Commission (IEC) Table in an identical manner to Duval's triangle in the previous section. As it is once again a multi class classification problem, where we compare this method's output for both the real and predicted DGA values, we once more create a table with a identical structure to the previous one, down to the selected performance metrics.

The IEC table predicts 9 different fault types (including no fault), unlike Duval's 7, thus we once again reduce the value threshold for the metrics. The one for accuracy was reduced to 0.55, while that for all other metrics was reduced to 0.35. These are of course, quite low values, but in order for the difficulty of this task to be on par, and thus comparable, with the rest of the approaches, this reduction is required. Since other than these thresholds, the structure, metrics and interpretation of the variables is the same as in the previous section, for brevity, we will not redescribe them.

Table 6.5 contains the obtained results for this method. The first thing to note is that using our defined minimum performance values, the subset size can be reduced to 2, which is worse than Duval's 1, but an improvement over that observed by all other methods. In fact, the overall results are improved over all cases so far, even though this is a more difficult problem.

There are two things that are still worth mentioning. The first is that the macro averaged metrics, again obtained worse results, thus implying that the minority classes are less well represented by our

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.623	0.326	0.646	0.348	0.704	0.354	0.623	0.326
1	0.688	0.337	0.687	0.33	0.697	0.362	0.688	0.33
2	0.766	0.362	0.751	0.363	0.761	0.378	0.766	0.362
3	0.898	0.527	0.896	0.551	0.902	0.531	0.898	0.527
4	0.826	0.488	0.835	0.539	0.855	0.494	0.826	0.488
5	0.899	0.541	0.896	0.577	0.899	0.543	0.899	0.541
6	0.902	0.519	0.896	0.551	0.901	0.58	0.902	0.519
7	0.919	0.595	0.915	0.687	0.925	0.634	0.919	0.595
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6.5: Table of the full dataset IEC table classification metrics.

model. The other observation is that much like for Duval's triangle, O2, which is the first variable to be removed, is not used by the IEC table, thus explaining the perfect results for the subset of size 8.

6.1.6 Rogers Ratio

Rogers ratio is a method that works in almost an identical way to the IEC table. Also utilizing ratios comprised of the same gases, while predicting all of the same fault types. As such, we once again use the same table structure as those of the previous subsections, but this time, due to no difference in number of classes we decided not to change the thresholds, thus leaving them at 0.55 for accuracy and 0.35 for all other metrics.

The first thing we see on table 6.6, containing the results for this validation problem, is that the performance is improved all across the board from the IEC table. This, of course, makes sense, due to the usage of the same gases and a simpler problem with one less class to predict. With this we conclude that great results can be obtained from only the support variables in this case (subset of size 0).

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.584	0.473	0.591	0.555	0.644	0.464	0.584	0.464
1	0.594	0.436	0.587	0.541	0.643	0.437	0.594	0.436
2	0.707	0.488	0.681	0.599	0.701	0.484	0.707	0.484
3	0.839	0.67	0.832	0.725	0.838	0.65	0.839	0.65
4	0.75	0.553	0.75	0.578	0.755	0.547	0.75	0.547
5	0.833	0.68	0.826	0.733	0.834	0.665	0.833	0.665
6	0.81	0.699	0.823	0.738	0.854	0.731	0.81	0.699
7	0.848	0.746	0.856	0.771	0.875	0.766	0.848	0.746
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6.6: Table of the full dataset Rogers ratio classification metrics.

Once again, the macro averaged metrics achieve worse results than the weighed ones, reinforcing our previous conclusions. Lastly, we like to remind the reader that the perfect scores for the subset of

size 8 stem from the non usage of O2 by this method.

6.1.7 Key Gas Method

The final of our validation problems is the key gas method, where the relative concentrations of 6 gases are used to predict 4 different fault types. This means that more gases are employed than by any other method so far, while the least amount of faults is predicted (not counting outlier methods). Given this, the thresholds were increased accordingly, with 0.7 for accuracy and 0.45 for all other metrics. We can see in table 6.7 that these requirements are fulfilled for the minimum subset size of 1.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.88	0.392	0.838	0.626	0.892	0.379	0.88	0.379
1	0.894	0.766	0.862	0.963	0.905	0.739	0.894	0.739
2	0.898	0.779	0.869	0.965	0.908	0.748	0.898	0.748
3	0.93	0.546	0.919	0.633	0.93	0.505	0.93	0.505
4	0.908	0.587	0.901	0.668	0.901	0.536	0.908	0.536
5	0.934	0.554	0.924	0.635	0.934	0.515	0.934	0.515
6	0.955	0.599	0.951	0.641	0.953	0.57	0.955	0.57
7	0.976	0.898	0.975	0.986	0.976	0.835	0.976	0.835
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6.7: Table of the full dataset key gas method classification metrics.

One worrying element of our table is the fact that the overall results do not decrease as the subset size does. This was, of course also observed in the previous subsections, but the deviations were small enough to be within the expected model variations and as such not mentioned. Upon a more detailed analysis we can see that the macro values in some instances present a larger difference to the weighed ones than ever before, thus implying terrible results for minority classes. An inspection of the predicted classes, obtained by this method, wielded the explanation as to why this is happening. This method is by far creating the most unbalanced predictions out of all analysed. In fact, over 95% of entries belonged to a single class. Due to this, a minor shift in the bias for our models greatly affects the recall for the minority classes, thus explaining the results. One presented hypothesis for this is that the way this method was implemented was somehow incorrect, however, no literature was found validating this, and as such, we decided to still present these results.

6.2 Dissolved Gas Analysis Only Dataset

The DGA only dataset contains, as its name implies, only DGA variables, thus containing no support variables and making the problems more difficult. Unlike for the full dataset, transformations, dimentionality reduction and missing value imputation are not applicable as preprocessing steps, thus only conditional string imputation was performed.

For brevity the results tables, respective metrics, thresholds and interpretation will not be explained as they have already been in the previous section. We, therefore, urge the reader to go back should any explanation be found lacking in detail. Much like in the previous section we will, before analysing the validation results, detail the obtained models and subsets, as these contain the most important changes from before.

For the smallest subset of size 0:

- No DGA gas is present in the subset and thus all are predicted.
- A GBR model was obtained.

For the subset of size 1:

- C2H2 was added to the subset.
- The selected model was a MLP with synthetic multi output capabilities.

For the subset of size 2:

- H2 was added to the subset.
- The selected model changed back to the GBR.

For the subset of size 3:

- C2H4 was added to the subset.
- The selected model was once more a GBR.

For the subset of size 4:

- N2 was added to the subset.
- The selected model changed to a RF with default multi output capabilities.

For the subset of size 5:

- CO was added to the subset.
- The selected model remained a RF but this time with synthetic multi output capabilities.

For the subset of size 6:

- CO2 was added to the subset.
- The selected model changed back again to a MLP with synthetic multi output capabilities. capabilities.

For the subset of size 7:

- C2H6 was added to the subset.
- A SVR model was selected.

For the subset of size 8:

- O2 was added to the subset.
- The selected model remained a SVR.

For this set of results CH4 is never predicted, unlike for the previous ones in which O2 has the privilege. There also is a much broader selection of models picked, with only MLP with standard multi outputs never being selected.

6.2.1 Performance Metrics

The first set of results is contained in table 6.8, with the general regression results that were utilized for parameter selection. The first thing to note is that regarding the "Worst" column, a minimum size of 6 is achieved while keeping all values of R-squared above 0.7.

The first and last two variables to be included are once again of particular importance, with C2H2 and H2 being the first ones, while CH4 and O2 are the last. Here the two most difficult to predict variables swapped order, with a couple of hypothesis explaining this. The first one, which is more likely, is the inherent randomness in the modeling process, which is validated by different orderings being obtained for very similar regressions. The second one is the C2H2 variable having stronger relations with support variables than with DGA variables, when compared to H2. Nevertheless, this was expected given the low strength of relations previously observed.

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	-0.25 | 29.55 | 24.539	-0.026 | 97.548 | 44.888	-0.012 | 119.018 | 42.523	-0.012 | 59.839 | 35.597	-0.743 | 4.948 | 4.246
1	-0.033 | 26.867 | 18.748	-0.037 | 98.078 | 43.472	-0.014 | 119.179 | 41.287	-0.014 | 59.893 | 35.452	N/A
2	N/A	0.074 | 92.705 | 43.436	0.074 | 113.869 | 39.609	0.125 | 55.637 | 33.624	N/A
3	N/A	0.884 | 32.793 | 18.895	N/A	0.677 | 33.807 | 21.212	N/A
4	N/A	0.858 | 36.307 | 17.222	N/A	0.788 | 27.361 | 17.061	N/A
5	N/A	0.909 | 29.026 | 13.101	N/A	0.706 | 32.26 | 16.203	N/A
6	N/A	0.9 | 30.455 | 12.694	N/A	0.719 | 31.504 | 15.653	N/A
7	N/A	0.965 | 17.926 | 7.641	N/A	N/A	N/A
8	N/A	0.974 | 15.495 | 6.433	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	-0.001 | 172.291 | 122.945	-0.002 | 2.03 | 1.646	-0.007 | 7.442 | 5.92	-0.002 | 19.313 | 16.596	-0.743 | 172.291 | 122.945
1	-0.0 | 172.246 | 120.477	0.024 | 2.003 | 1.529	0.02 | 7.342 | 5.327	-0.001 | 19.305 | 16.578	-0.037 | 172.246 | 120.477
2	0.125 | 161.09 | 114.241	0.069 | 1.957 | 1.517	0.049 | 7.232 | 5.493	0.093 | 18.374 | 15.684	0.049 | 161.09 | 114.241
3	0.371 | 136.57 | 94.346	0.134 | 1.887 | 1.328	0.167 | 6.77 | 4.783	0.271 | 16.469 | 13.635	0.134 | 136.57 | 94.346
4	0.595 | 109.654 | 69.327	0.508 | 1.423 | 1.0	0.785 | 3.439 | 2.392	N/A	0.508 | 109.654 | 69.327
5	N/A	0.644 | 1.21 | 0.844	0.855 | 2.828 | 1.861	N/A	0.644 | 32.26 | 16.203
6	N/A	N/A	0.86 | 2.776 | 1.768	N/A	0.719 | 31.504 | 15.653
7	N/A	N/A	0.95 | 1.664 | 1.209	N/A	0.95 | 17.926 | 7.641
8	N/A	N/A	N/A	N/A	0.974 | 15.495 | 6.433

Table 6.8: Table of the DGA only dataset general regression metrics. Results in the form R-squared | RMSE | MAE

The variable O2 went from the last variable to be predicted to the second to last. Here randomness is most certainly a factor, but unlike for the previous variables does not seem to be the most important. Its strong relations with Tang Delta (90ºC), Acidity Index, Cor and other support variables that are no longer present are more likely to have made its prediction more difficult. The last variable to be predicted (CH4) is removed much earlier than in the previous set of results, where it is the third one to be predicted. This however is not unexpected due to its already noted high correlation with C2H2 and no strong relations with any other variables. In fact we can see that its counterpart C2H2 is now kept much longer, with these two variables having virtually swapped positions. All other variables have also changed the order in which they are removed from the subset, but this can simply be attributed to the randomness in the sampling process.

One last thing to note is the improvement of the values obtained for the two largest subset sizes, when compared to those in the full dataset. We concluded that this is not due to random chance but instead a concrete result, as this was verified for almost all modeling configurations. The most likely cause is a small overfitting of the results for the full dataset when most of the DGA variables are employed. This could not be corrected due to the model hyperparameter ranges being selected equally for all subset sizes (individual selection was impossible due to time constraints) and tuning for the full dataset being done by improving smaller dataset sizes (the minimum achieving R-squared

greater than 0.7).

6.2.2 Outlier Regression Validation

Table 6.9 contains the set of results for the outlier regression validation task.

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	-0.12 | 241.98 | 745.12	-1.02 | 126.72 | 178.36	-0.0 | 30.26 | 83.47	-0.95 | 143.49 | 205.35	-0.0 | 6.47 | 29.01
1	-0.09 | 242.97 | 732.12	-1.01 | 128.47 | 177.95	0.01 | 29.29 | 82.94	-0.96 | 143.96 | 205.68	N/A
2	N/A	-0.24 | 90.5 | 135.58	0.07 | 53.12 | 122.03	-0.74 | 129.73 | 193.61	N/A
3	N/A	0.85 | 38.54 | 47.81	N/A	0.63 | 67.02 | 86.65	N/A
4	N/A	0.91 | 23.09 | 37.59	N/A	0.87 | 33.12 | 51.87	N/A
5	N/A	0.94 | 19.37 | 30.54	N/A	0.85 | 31.57 | 55.85	N/A
6	N/A	0.93 | 15.48 | 32.14	N/A	0.69 | 31.58 | 80.04	N/A
7	N/A	0.96 | 14.15 | 23.6	N/A	N/A	N/A
8	N/A	0.98 | 8.85 | 18.16	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	-8.29 | 530.77 | 561.86	-11.65 | 6.75 | 7.04	-48.22 | 17.94 | 18.12	nan | nan | nan	-48.22 | 530.77 | 745.12
1	-8.58 | 539.89 | 570.48	-8.17 | 5.38 | 5.99	-35.22 | 14.51 | 15.54	nan | nan | nan	-35.22 | 539.89 | 732.12
2	-6.46 | 472.11 | 503.25	-8.59 | 5.8 | 6.13	-33.99 | 14.73 | 15.28	nan | nan | nan	-33.99 | 472.11 | 503.25
3	-1.94 | 270.71 | 316.23	-3.69 | 3.65 | 4.29	-13.76 | 9.19 | 11.52	nan | nan | nan	-13.76 | 270.71 | 316.23
4	0.34 | 104.33 | 150.23	0.51 | 1.15 | 1.39	-0.28 | 2.34 | 2.93	N/A	-0.28 | 104.33 | 150.23
5	N/A	0.62 | 1.02 | 1.3	-0.02 | 1.99 | 2.6	N/A	-0.02 | 31.57 | 55.85
6	N/A	N/A	-0.45 | 2.43 | 3.41	N/A	-0.45 | 31.58 | 80.04
7	N/A	N/A	0.44 | 1.42 | 2.1	N/A	0.44 | 14.15 | 23.6
8	N/A	N/A	N/A	N/A	0.98 | 8.85 | 18.16

Table 6.9: Table of the DGA only dataset outlier regression metrics. Results in the form R-squared | RMSE | MAE

Much like for the full dataset, the O2 variable significantly hinders the performance shown in the "worst" column. The reason is the same as to the one for the full dataset, with all outlier values being extremely similar, making a high R-squared nearly impossible to achieve. As such, by employing the same reasoning regarding the analysis of the results we conclude that despite them being very poor, they do not single-handedly invalidate our approach. Nevertheless, considering this set of results in a void and using the our pre-established thresholds we conclude that the subset can only be reduced to size 8 while having high confidence, while size 7 can only be considered if we are willing to accept some more risks.

6.2.3 Outlier Binary Classification Validation

The results for our outlier binary classification validation task can be found on table 6.10.

Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	nan | nan | nan | 0.93	nan | nan | nan | 0.87	0.15 | 1.0 | 0.27 | 0.15	nan | nan | nan | 0.91	0.19 | 1.0 | 0.32 | 0.19
1	0.75 | 0.04 | 0.08 | 0.93	1.0 | 0.02 | 0.04 | 0.87	0.15 | 1.0 | 0.27 | 0.15	nan | nan | nan | 0.91	N/A
2	N/A	0.47 | 0.45 | 0.46 | 0.86	0.29 | 0.65 | 0.4 | 0.7	nan | nan | nan | 0.91	N/A
3	N/A	0.85 | 0.69 | 0.76 | 0.94	N/A	0.79 | 0.62 | 0.69 | 0.95	N/A
4	N/A	0.97 | 0.79 | 0.87 | 0.97	N/A	0.95 | 0.76 | 0.85 | 0.97	N/A
5	N/A	0.97 | 0.82 | 0.89 | 0.97	N/A	0.93 | 0.8 | 0.86 | 0.98	N/A
6	N/A	0.96 | 0.9 | 0.93 | 0.98	N/A	0.9 | 0.85 | 0.87 | 0.98	N/A
7	N/A	0.97 | 0.87 | 0.91 | 0.98	N/A	N/A	N/A
8	N/A	0.98 | 0.9 | 0.94 | 0.98	N/A	N/A	N/A
Subset Size	**CO**	**CO2**	**O2**	**N2**	**Worst**
0	nan | nan | nan | 0.96	nan | nan | nan | 0.94	nan | nan | nan | 0.92	nan | nan | nan | 1.0	nan | nan | nan | 0.15
1	nan | nan | nan | 0.96	1.0 | 0.02 | 0.03 | 0.94	nan | nan | nan | 0.92	nan | nan | nan | 1.0	0.15 | 0.02 | 0.03 | 0.15
2	nan | nan | nan | 0.96	nan | nan | nan | 0.94	nan | nan | nan | 0.92	nan | nan | nan | 1.0	0.29 | 0.45 | 0.4 | 0.7
3	1.0 | 0.22 | 0.36 | 0.96	1.0 | 0.22 | 0.36 | 0.95	0.74 | 0.3 | 0.43 | 0.94	nan | nan | nan | 1.0	0.74 | 0.22 | 0.36 | 0.94
4	0.97 | 0.72 | 0.82 | 0.99	0.96 | 0.7 | 0.81 | 0.98	0.98 | 0.65 | 0.78 | 0.97	N/A	0.95 | 0.65 | 0.78 | 0.97
5	N/A	0.93 | 0.78 | 0.85 | 0.98	1.0 | 0.7 | 0.82 | 0.98	N/A	0.93 | 0.7 | 0.82 | 0.97
6	N/A	N/A	0.86 | 0.64 | 0.73 | 0.96	N/A	0.86 | 0.64 | 0.73 | 0.96
7	N/A	N/A	0.9 | 0.81 | 0.85 | 0.98	N/A	0.9 | 0.81 | 0.85 | 0.98
8	N/A	N/A	N/A	N/A	0.98 | 0.9 | 0.94 | 0.98

Table 6.10: Table of the DGA only dataset outlier binary classification metrics. Results in the form Precision | Recall | F1 | Accuracy

Before going into detail on the actual performance values, one event that has not happened for the full dataset results is present here and requires some explaining. We can see that for the variables H2, CH4, C2H6, CO, CO2 and O2 some cells appear with "nan" values for precision, recall and F1. Unlike for N2, in which these values are caused by the inexistence of outliers in the original data, these new "nan" values are caused by a numerical instability in the calculation of these metrics when all the predictions are of the same class. In all instances that this is happening the majority (not outlier) class is being predicted. These results are however not meaningful and this format was, as such, kept to draw attention to these cases.

With this abnormal event explained, we turn our attention to the "worst" column, were we conclude that a minimum subset size of 4 can be obtained. A subset of size 3 would fulfill our requirement for an accuracy greater than 0.9, but not the one for all other metrics achieving better than 0.5 (both recall and F1 fail). Another important observation is that much like for the full dataset, O2 achieves good precision but the worst recall values, cementing its position as a difficult variable in which to achieve good outlier predictions.

6.2.4 Duval's Triangle

Table 6.11 contains the performance metrics for the Duval's triangle validation. Our initial analysis of the "worst" column indicates that, using the established thresholds for this task, a subset of size 3 can be achieved. We recall that for the full dataset O2, which is not used by this method, is the first to be removed, leading to a perfect score for subset of size 8. Here the same does not happen, with CH4 being the first removed variable, which is used by Duval's triangle.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.081	0.025	0.012	0.013	0.007	0.167	0.081	0.007
1	0.265	0.179	0.138	0.151	0.104	0.282	0.265	0.104
2	0.269	0.205	0.158	0.312	0.383	0.276	0.269	0.158
3	0.721	0.672	0.688	0.748	0.709	0.649	0.721	0.649
4	0.806	0.792	0.794	0.829	0.81	0.777	0.806	0.777
5	0.82	0.816	0.812	0.841	0.819	0.803	0.82	0.803
6	0.821	0.709	0.818	0.732	0.822	0.692	0.821	0.692
7	0.884	0.887	0.884	0.887	0.884	0.889	0.884	0.884
8	0.904	0.907	0.904	0.904	0.904	0.911	0.904	0.904

Table 6.11: Table of the DGA only Duval's triangle classification metrics.

The last important difference to note is regarding the macro versus weighed metrics. Unlike for the previous set of results, here there is no significant difference between these pairs of metrics, showing a better balance for prediction performance between majority and minority classes.

6.2.5 International Electrotechnical Commission Table

By analysing the results of the IEC table, contained in table 6.12, we conclude that for the our defined thresholds the minimum subset size attainable is 3.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.066	0.012	0.008	0.007	0.004	0.1	0.066	0.004
1	0.44	0.094	0.354	0.122	0.488	0.123	0.44	0.094
2	0.462	0.119	0.397	0.163	0.532	0.14	0.462	0.119
3	0.695	0.408	0.686	0.411	0.731	0.428	0.695	0.408
4	0.822	0.48	0.817	0.498	0.833	0.52	0.822	0.48
5	0.836	0.483	0.83	0.479	0.841	0.518	0.836	0.479
6	0.821	0.497	0.824	0.505	0.83	0.507	0.821	0.497
7	0.934	0.854	0.933	0.866	0.934	0.849	0.934	0.849
8	0.941	0.903	0.942	0.897	0.942	0.909	0.941	0.897

Table 6.12: Table of the DGA only IEC table classification metrics.

Unlike for the Duval's triangle results, a significant difference is found between the macro and weighed versions of the performance metrics, implying as such worse predictions for the minority classes.

6.2.6 Rogers Ratio

Table 6.13 contains the performance metrics for the Rogers ratio validation. Upon an initial analysis we conclude that a minimum size of 3 is achieved for our thresholds, much like for the previous set of validation tasks. Similarly to the results from IEC table, the macro metrics achieve worse results than the weighed ones.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.195	0.036	0.064	0.022	0.038	0.111	0.195	0.022
1	0.421	0.151	0.3	0.163	0.277	0.18	0.421	0.151
2	0.42	0.144	0.321	0.199	0.389	0.169	0.42	0.144
3	0.663	0.462	0.631	0.547	0.654	0.443	0.663	0.443
4	0.767	0.556	0.736	0.728	0.749	0.528	0.767	0.528
5	0.778	0.542	0.75	0.666	0.753	0.519	0.778	0.519
6	0.751	0.58	0.742	0.604	0.74	0.571	0.751	0.571
7	0.927	0.887	0.926	0.942	0.929	0.87	0.927	0.87
8	0.943	0.918	0.943	0.958	0.945	0.901	0.943	0.901

Table 6.13: Table of the DGA only Rogers ratio classification metrics.

6.2.7 Key Gas Method

The final set of results for the key gas method is present on table 6.14. Employing our defined thresholds we find that size 2 is the minimum achievable.

Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.861	0.309	0.797	0.287	0.742	0.333	0.861	0.287
1	0.859	0.231	0.796	0.215	0.742	0.249	0.859	0.215
2	0.864	0.642	0.802	0.621	0.747	0.667	0.864	0.621
3	0.863	0.668	0.811	0.777	0.813	0.678	0.863	0.668
4	0.911	0.83	0.895	0.942	0.911	0.792	0.911	0.792
5	0.947	0.91	0.942	0.981	0.95	0.871	0.947	0.871
6	0.961	0.938	0.958	0.98	0.962	0.908	0.961	0.908
7	0.984	0.977	0.984	0.983	0.984	0.972	0.984	0.972
8	0.993	0.99	0.993	0.991	0.993	0.989	0.993	0.989

Table 6.14: Table of the DGA only Key Gas classification metrics.

Similarly to most of the results, the macro metrics attain worse performance than the weighed ones. One important difference from the full dataset results is that the instability detected, with performance sometimes improving for lower subset sizes, is not present at all. This is likely the case due to the inexistence of support variables, thus making each subset size significantly different, while their presence might make the random fluctuations in the modeling more important than a single DGA variable.

6.3 Summary

Throughout this chapter we presented the models, subsets obtained and their respective performances for a set of 7 validation tasks. This was done for two different datasets, the full dataset and the DGA only dataset. The regression performance of the full dataset was superior, making greater dimentionality reduction possible. The worst set of results comes in the form of the outlier regression, for which both datasets performed very poorly. This, however, was found to not be very significant unless very accurate values of outliers are required for the task at hand.

Figures 6.1 and 6.2 show the overall performance for the full and the DGA only datasets respectively. Each of the lines presents the metrics obtained compared to their defined thresholds. What this means is that for each value, the respective threshold was subtracted, and thus values greater than 0 mean that the performance surpasses the established minimum requirement. As such, the black line at 0 represents all of the different thresholds, with all other values being shifted accordingly. Lastly, it bears mentioning that in each case only the worst statistic (different from the "worst" column) was used for each case (the one for which the greater subset size is required).

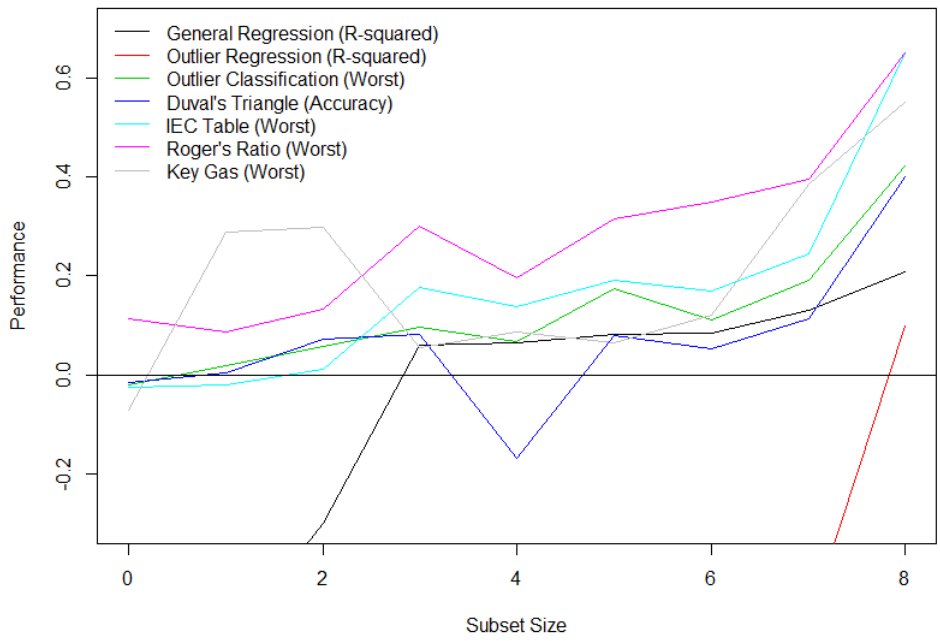

Figure 6.1: Plot for the full dataset of the various metrics compared with their thresholds.

From these it first becomes clear that the overall performance metrics for the full dataset are

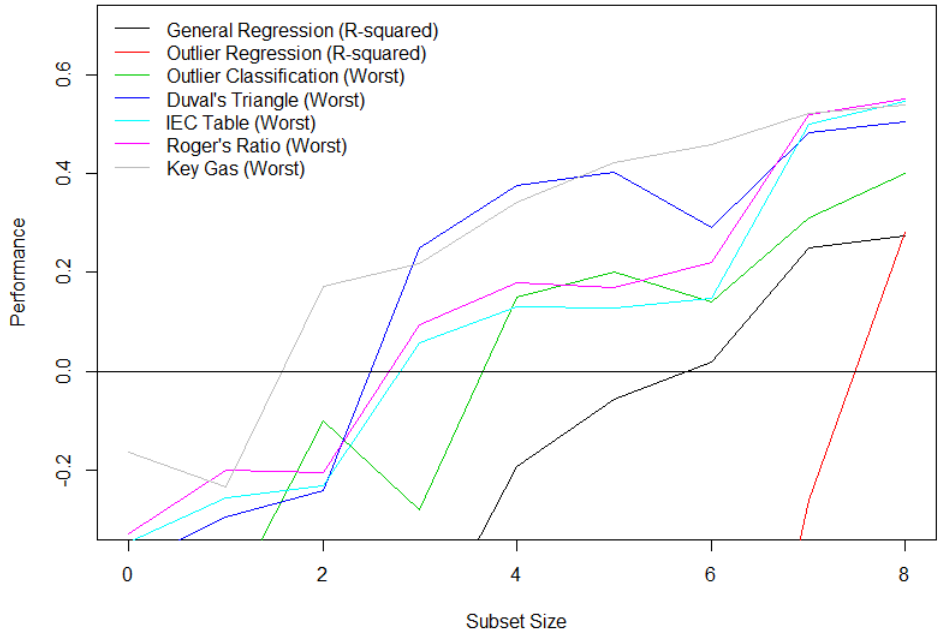

Figure 6.2: Plot for the DGA only dataset of the various metrics compared with their thresholds.

better, bar some outliers, such as the decrease in performance of Duval's triangle for size 4. Given this, for the full dataset there are 2 sizes of more interest. First, size 2, for which all but the regression metrics achieve the desired performance. And then size 3, where the performance values greatly increase and general regression joins the group of those that accomplish the desired results. Given the great increase in overall performance for just one more gas, we consider subset size 3 to be the most enticing. For the DGA only dataset, there are similarly 2 sizes of greater interest, 3 and 4. For size 3, 4 of the 7 validation tasks are up to par with the intended performance. For size 4, a great performance improvement is also observed, with one more classification task achieving our intended results. However, the increase is not as high as for the full dataset, so we conclude that size 3 is the most promising. Thus, for both datasets a subset of size 3 was found to be the best candidate.

Chapter 7

Conclusion

In this chapter we present the main conclusions and takeaways from our approach, reiterating interesting or innovative facets of our work. As this is, to our knowledge, the first work to attempt something of this kind there are still many limitations in our approach to be corrected, and future avenues to research.

We would like to start by stating our most important contribution. This is the identification of viable Dissolved Gas Analysis (DGA) gas subsets of size 3. These, while reducing the number of gases that need to be measured from 9 to 3, allow any number of models to obtain results almost identical to those from a real full 9 gas DGA dataset. In the end two viable subsets were obtained. Both have Hydrogen (H2) and Ethylene (C2H2), while one also has Methane (CH4) and the other Ethane (C2H4). Although the last 3 gases are commonly employed by classical fault prediction methods, H2, which proved to be one of the most important for our method is never used.

Another important result stems from our developed imputation approaches Regression Sampling Imputation (RSI) and Classification Sampling Imputation (CSI), which achieved better results than the limited set of common alternatives tested.

7.1 Limitations

Despite our frankly unexpectedly good results, our work is still accompanied with a large number of limitations. The first and most important one, that plagues many works in this field is the dataset. Although larger and more complete than many authors have access to, it is still insufficient in terms of

size and quality. Although it contains a good variety of variables and information, one in particular that is lacking is real failure data, which impossibilitates the usage of Machine Learning (ML) techniques for fault prediction in our validation steps. With only about 1000 entries and a large amount of missing data, which is split amongst many very different generators, the results were severely impaired.

Another limitation of our work involves the relatively small amount of subsets tested. Given that they present half of the intended output, testing as many subsets as possible is paramount. However, due to time and computational limitations, our subset selection algorithm only scratched the surface of possible combinations.

The last limiting factor of our work lies in the selection of hyperparameters for our models. Despite our best attempts the hyperparameter search grid was still relatively sparse and some combinations were not tested due to them greatly increasing computational times. The greatest offender of this was the Multi Layer Perceptron (MLP) model, where although larger layers, smaller batch sizes and higher iteration counts improved performance, these parameters had to be capped. This, in large part, occurred due to the modeling process being done exclusively on the Central Processing Unit (CPU).

7.2 Future Work

Most of our proposed and planned future work is aimed at correcting the limitations present on our work. Our first goal would be to find a larger dataset containing data about failures and utilizing it to both, try to achieve better performance, and to better validate our results with ML methods trained on predicting real failures.

With a greater amount of available time or computational resources (Our approach is heavily parallelizable) we pretend to further explore both the subset and hyperparameter search space. In particular we intend to grow the MLP models by employing deep learning libraries to fully utilize Graphics Processing Unit (GPU) computational performance improvements.

A different and interesting avenue for research that we plan to tackle is a further investigation of our RSI and RSI techniques, comparing them with a richer set of imputation techniques in a variety of problems and datasets.

Bibliography

[1] Ahmed EB Abu-Elanien and MMA Salama. Asset management techniques for transformers. *Electric power systems research*, 80(4):456–464, 2010.

[2] Thomas Ackermann, Göran Andersson, and Lennart Söder. Distributed generation: a definition. *Electric power systems research*, 57(3):195–204, 2001.

[3] Andrianto, Arif Nur Afandi, Aripriharta, Irham Fadlika, Setiadi Cahyono Putro, and Arrizal Haris Fajariawan. Analysis of maintenance scheduling transformer oil using Markov method. In *AIP Conference Proceedings*, volume 2228, page 030020. AIP Publishing LLC, 2020.

[4] JinLong Bai, Wei Wei, and HongZheng Mei. The Transformer Winding Deformation Based on Finite Element Method and Regular Limit Learning Machine. In *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, pages 814–819. IEEE, 2018.

[5] Mladen Banović and Jean Sanchez. Basics of power transformers. *Transformers Magazine*, 1(1): 12–15, 2014.

[6] MK Base. On-load tap-changers for power transformers. *Regensburg, Germany*, 2013.

[7] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *Machine Learning Proceedings 1994*, pages 28–36. Elsevier, 1994.

[8] Thyago P Carvalho, Fabrízzio AAMN Soares, Roberto Vita, Roberto da P Francisco, João P Basto, and Symone GS Alcalá. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.

[9] Ravi Shankar Chauhan and Amrita Guide Sinha. *Internal fault detection in three phase transformer using machine learning methods*. PhD thesis, 2015.

[10] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[11] Dipali D Dhonge, PS Swamin, and AG Thosar. 'Developing Artificial Neural Network (ANN) Model for Fault Diagnosis of Power Transformer Using Dissolved Gas Analysis (DGA). *International Journal of Scientific & Engineering Research*, 6(7), 2015.

[12] Nick Dominelli. Equipment health rating of power transformers. In *Conference record of the 2004 IEEE international symposium on electrical insulation*, pages 163–168. IEEE, 2004.

[13] Béatrice Duval, Jin-Kao Hao, and Jose Crispin Hernandez Hernandez. A memetic algorithm for gene selection and molecular classification of cancer. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 201–208, 2009.

[14] Eduardo Luís Meireles e Oliveira, Vera L Miguéis, Luís Guimarães, and José Luís Borges. Power Transformer Failure Prediction: Classification in Imbalanced Time Series. *U. Porto Journal of Engineering*, 3(2):34–48, 2017.

[15] Electrical4U. Electrical power transformer: Definition types of transformers. https://www.electrical4u.com/electrical-power-transformer-definition-and-types-of-transformer/, 2020. [Online; accessed 15-May-2021].

[16] Zoran Gajić. *Differential protection for arbitrary three-phase power transformers*. Department of Industrial Electrical Engineering and Automation, Lund University, 2008.

[17] Gerards Gavrilovs and Sandra Vītoliņa. Identification of Power Transformer's Failure and Risk Source. In *Proceedings of the 52st annual international scientific conference,(October)*, pages 1–4, 2011.

[18] Pavlos S Georgilakis, John A Katsigiannis, Kimon P Valavanis, and Athanasios T Souflaris. A systematic stochastic petri net based methodology for transformer fault diagnosis and repair actions. *Journal of Intelligent and Robotic Systems*, 45(2):181–201, 2006.

[19] Asif Islam. Detection of mechanical deformation in old aged power transformer using cross correlation co-efficient analysis method. 2012.

[20] Md Mominul Islam, Gareth Lee, and Sujeewa Nilendra Hettiwatte. A review of condition monitoring techniques and diagnostic tests for lifetime estimation of power transformers. *Electrical Engineering*, 100(2):581–605, 2018.

[21] Alexandra Khalyasmaa, Stanislav Eroshenko, Maksim Elaev, and Tran Duc Chung. Machine Learning Application for the High-Voltage Equipment Lifecycle Forecasting. In *2020 21st International Symposium on Electrical Apparatus & Technologies (SIELA)*, pages 1–4. IEEE, 2020.

[22] Shakeb A Khan, Md Danish Equbal, and Tarikul Islam. A comprehensive comparative study of dga based transformer fault diagnosis using fuzzy logic and anfis models. *IEEE Transactions on Dielectrics and Electrical Insulation*, 22(1):590–596, 2015.

[23] Leonidha Londo, Marialis Çelo, and Rajmonda Bualoti. Assessment of Transformer Condition Using the Improve Key Gas Methods. *International Journal of Engineering Research & Technology (IJERT)*, 4:48–55, 2015.

[24] Peilin Mao. *Power transformer fault diagnosis based on wavelet transform and artificial neural network*. PhD thesis, University of Bath, 2000.

[25] Xiaozhou Mao, Zhongdong Wang, Paul Jarman, and Andrew Fieldsend-Roxborough. Winding type recognition through supervised machine learning using frequency response analysis (FRA) data. In *2019 2nd International Conference on Electrical Materials and Power Equipment (ICEMPE)*, pages 588–591. IEEE, 2019.

[26] Vladimiro Miranda and Adriana Rosa Garcez Castro. Improving the IEC table for transformer failure diagnosis with knowledge extraction from neural networks. *IEEE transactions on power delivery*, 20(4):2509–2516, 2005.

[27] Piotr Mirowski and Yann LeCun. Statistical machine learning and dissolved gas analysis: a review. *IEEE Transactions on Power Delivery*, 27(4):1791–1799, 2012.

[28] Jefferson Morais, Aldebaro Klautau, Claudomir Cardoso, and Yomara Pires. *An overview of data mining techniques applied to power systems*. Citeseer, 2009.

[29] Nuno Filipe Osório Morais. Estimating the remaining lifetime of power transformers using paper insulation degradation. 2018.

[30] R Naresh, Veena Sharma, and Manisha Vashisth. An integrated neural fuzzy approach for fault diagnosis of transformers. *IEEE transactions on power delivery*, 23(4):2017–2024, 2008.

[31] Venera Nurmanova, Yerbol Akhmetov, Mehdi Bagheri, Amin Zollanvari, Gevork B Gharehpetian, and Toan Phung. A New Transformer FRA Test Setup for Advanced Interpretation and Winding

Short-circuit Prediction. In *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pages 1–6. IEEE, 2020.

[32] Christophe Perrier, Marielle Marugan, and Abderrahmane Beroual. Dga comparison between ester and mineral oils. *IEEE Transactions on Dielectrics and Electrical Insulation*, 19(5):1609–1614, 2012.

[33] Nanthiine Nair Ravi, Sulfeeza Mohd Drus, and Prajindra Sankar Krishnan. Data mining techniques for transformer failure prediction model: A systematic literature review. In *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 305–309. IEEE, 2019.

[34] Petar Sarajcev, Damir Jakus, and Josip Vasilj. Optimal scheduling of power transformers preventive maintenance with Bayesian statistical learning and influence diagramslabel. *Journal of Cleaner Production*, page 120850, 2020.

[35] Servkon. Duval triangle diagnostic method in the power transformers. https://servkon.com/the-duval-triangle-method-for-dga-diagnostic-in-power-transformers/, 2020. [Online; accessed 6-June-2021].

[36] Stephen D Smith. Periodic testing and maintenance of power transformers to extend life and improve reliability. In *1977 EIC 13th Electrical/Electronics Insulation Conference*, pages 156–158. IEEE, 1977.

[37] Huo-Ching Sun, Yann-Chang Huang, and Chao-Ming Huang. Fault diagnosis of power transformers using computational intelligence: A review. *Energy Procedia*, 14:1226–1231, 2012.

[38] Ricardo Manuel Arias Velásquez, Jennifer Vanessa Mejía Lara, and Andres Melgar. Converting data into knowledge for preventing failures in power transformers. *Engineering Failure Analysis*, 101:215–229, 2019.

[39] MVAJ Wang, A John Vandermaar, and K D_ Srivastava. Review of condition assessment of power transformers in service. *IEEE Electrical insulation magazine*, 18(6):12–25, 2002.

[40] Wikipedia. Feature selection — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Feature%20selection&oldid=1020695065, 2021. [Online; accessed 07-June-2021].

[41] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[42] Zhongyong Zhao, Chao Tang, Qu Zhou, Lingna Xu, Yingang Gui, and Chenguo Yao. Identification of power transformer winding mechanical fault types based on online IFRA by support vector machine. *Energies*, 10(12):2022, 2017.

Appendices

Feature	has NaN	NaN Count	NaN %	Dtype	count
Reference	False	0.0	0.0	float64	1020.0
Power (kVA)	True	9.0	0.878	float64	1020.0
Tension (kV)	True	9.0	0.878	float64	1020.0
Manufacture Year	True	9.0	0.878	float64	1020.0
Oil Brand	True	99.0	9.66	object	926.0
Oil Type	True	105.0	10.2	object	920.0
Oil Weight (ton)	True	9.0	0.878	float64	1020.0
Collection Date	False	0.0	0.0	float64	1020.0
Point of Collection	False	0.0	0.0	object	1020.0
Oil Temperature	False	0.0	0.0	float64	1020.0
5 HMF	True	638.0	62.2	float64	387.0
2 FOL	True	638.0	62.2	float64	387.0
2 FAL	True	638.0	62.2	float64	387.0
5 MEF	True	638.0	62.2	float64	387.0
H2	False	0.0	0.0	float64	1020.0
CH4	False	0.0	0.0	float64	1020.0
C2H4	False	0.0	0.0	float64	1020.0
C2H6	False	0.0	0.0	float64	1020.0
C2H2	False	0.0	0.0	float64	1020.0
CO	False	0.0	0.0	float64	1020.0
CO2	False	0.0	0.0	float64	1020.0
O2	False	0.0	0.0	float64	1020.0
N2	False	0.0	0.0	float64	1020.0
Cor	True	318.0	31.0	float64	707.0
Aspect	True	318.0	31.0	object	707.0
Oil Density	True	655.0	63.9	float64	370.0
Viscosity (40ºC)	True	875.0	85.4	float64	150.0
Interfacial Tension	True	656.0	64.0	float64	369.0
Disruptive Tension	True	109.0	10.6	float64	916.0
Acidity Index	True	294.0	28.7	float64	731.0
Water Content	True	109.0	10.6	float64	916.0
Flash Point	True	875.0	85.4	float64	150.0
Tang Delta (90ºC)	True	296.0	28.9	float64	729.0
Sediments	True	734.0	71.6	float64	291.0
Mud	True	739.0	72.1	float64	286.0
Part4	True	844.0	82.3	float64	181.0
Part6	True	844.0	82.3	float64	181.0
Part14	True	848.0	82.7	float64	177.0
Part1	True	844.0	82.3	float64	181.0
Part2	True	844.0	82.3	float64	181.0
Part3	True	844.0	82.3	float64	181.0
Total	True	1010.0	98.1	N/A	1020.0

Table 1: Data summary table - Part 1

Feature	unique count	mode	mode count	mean	std	min
Reference	202.0	1.09e+03	21.0	1.09e+03	53.2	1e+03
Power (kVA)	24.0	1.7e+05	254.0	1.82e+05	1.38e+05	0
Tension (kV)	6.0	400	574.0	313	102	10
Manufacture Year	51.0	2.01e+03	101.0	1.93e+03	349	0
Oil Brand	7.0	NYNAS	726.0	N/A	N/A	N/A
Oil Type	17.0	NYTRO TAURUS	371.0	N/A	N/A	N/A
Oil Weight (ton)	54.0	0	253.0	31	23.9	0
Collection Date	317.0	1.27e+09	18.0	1.36e+09	6.99e+07	1.26e+09
Point of Collection	5.0	Cuba Ponto Inferior (F)	604.0	N/A	N/A	N/A
Oil Temperature	314.0	28.2	13.0	24.7	8.26	0
5 HMF	5.0	0	384.0	0.000284	0.00355	0
2 FOL	3.0	0	386.0	0.00109	0.0213	0
2 FAL	78.0	0	226.0	0.171	0.407	0
5 MEF	8.0	0	371.0	0.00152	0.00808	0
H2	359.0	7	24.0	32.3	201	0
CH4	403.0	4	21.0	31.3	67.5	0
C2H4	238.0	0	121.0	20.8	83.5	0
C2H6	416.0	0	38.0	31.9	67.3	0
C2H2	134.0	0	701.0	3.79	29	0
CO	816.0	5	9.0	182	171	0
CO2	464.0	0.08	10.0	2.39	2.46	0.04
O2	255.0	4.4	22.0	9.78	7.49	0.5
N2	521.0	22.1	6.0	40.9	19.4	5.6
Cor	15.0	0.5	347.0	1.57	1.4	0.5
Aspect	3.0	L	701.0	N/A	N/A	N/A
Oil Density	190.0	0.871	16.0	0.867	0.0413	0.0869
Viscosity (40ºC)	150.0	9.93	2.0	9.55	0.834	7.33
Interfacial Tension	182.0	34	15.0	31.2	8.46	13.3
Disruptive Tension	368.0	86	17.0	79.7	12.4	23.3
Acidity Index	147.0	0.001	191.0	0.0382	0.06	0.001
Water Content	146.0	4	53.0	6.08	4.36	0.1
Flash Point	26.0	150	31.0	149	4.04	138
Tang Delta (90ºC)	90.0	0.001	246.0	0.015	0.0242	0.001
Sediments	26.0	0.009	99.0	0.0118	0.00485	0.001
Mud	16.0	0.002	121.0	0.00385	0.00284	0.001
Part4	178.0	233	2.0	1.88e+03	8.52e+03	41
Part6	173.0	32	3.0	567	4.22e+03	13
Part14	33.0	10	33.0	15	26.5	1.5
Part1	12.0	15	45.0	15.8	1.88	13
Part2	12.0	14	47.0	13.9	1.79	11
Part3	11.0	10	49.0	10.4	1.75	8
Total	3400.0	0	2580.0	5.68e+07	2.73e+08	0

Table 2: Data summary table - Part 2

Feature	25%	50%	75%	max	kurtosis	skew
Reference	1.05e+03	1.09e+03	1.13e+03	1.2e+03	-0.967	0.176
Power (kVA)	8e+04	1.7e+05	2e+05	4.5e+05	-0.194	0.966
Tension (kV)	220	400	400	400	-1.52	-0.436
Manufacture Year	1.98e+03	2e+03	2.01e+03	2.02e+03	26.8	-5.36
Oil Brand	N/A	N/A	N/A	N/A	N/A	N/A
Oil Type	N/A	N/A	N/A	N/A	N/A	N/A
Oil Weight (ton)	8.88	34	45	95	-0.432	0.366
Collection Date	1.29e+09	1.39e+09	1.44e+09	1.47e+09	-1.53	-0.0371
Point of Collection	N/A	N/A	N/A	N/A	N/A	N/A
Oil Temperature	18.8	24.8	30.4	47.8	-0.364	0.026
5 HMF	0	0	0	0.06	223	14.3
2 FOL	0	0	0	0.42	387	19.7
2 FAL	0	0	0.12	2.3	10.3	3.18
5 MEF	0	0	0	0.06	31.4	5.61
H2	6.3	11.8	21	3.24e+03	178	13.2
CH4	4	10.5	26	552	27.1	4.82
C2H4	1	2.1	6	924	59.8	7.14
C2H6	3.1	10.1	35	884	71.9	7.09
C2H2	0	0	0.4	553	304	17
CO	62	136	247	1.22e+03	6.02	2.02
CO2	0.64	1.42	3.25	15.8	3.69	1.77
O2	4.4	6.7	12.4	35.8	0.711	1.32
N2	25.3	36.4	58.1	91.7	-0.952	0.429
Cor	0.5	1	2.5	7	1.04	1.32
Aspect	N/A	N/A	N/A	N/A	N/A	N/A
Oil Density	0.867	0.871	0.872	0.888	348	-18.4
Viscosity (40ºC)	9.08	9.91	10.1	13.3	2.71	-0.489
Interfacial Tension	22.7	34	37.7	49	-1.06	-0.428
Disruptive Tension	72.4	82	89	100	0.749	-0.886
Acidity Index	0.001	0.012	0.04	0.354	5.48	2.33
Water Content	3.6	5	7.4	60.4	36.1	4.08
Flash Point	148	150	152	158	0.157	-0.305
Tang Delta (90ºC)	0.001	0.004	0.016	0.152	7.68	2.57
Sediments	0.009	0.01	0.013	0.03	2.09	1.45
Mud	0.002	0.003	0.005	0.016	3.64	1.84
Part4	154	307	679	8.97e+04	73.8	8.14
Part6	43.9	93	186	5.56e+04	164	12.6
Part14	9	10	12	272	56.4	6.93
Part1	14	15	17	24	3.48	1.53
Part2	13	14	15	23	4.04	1.4
Part3	9	10	11	18	2	1.17
Total	0.876	14.7	116	1.47e+09	19.3	4.61

Table 3: Data summary table - Part 3