Accuracy Control and Welding Distortion Prediction in a Deck Plate

Marcio Fleming

Master Thesis

presented in partial fulfillment
of the requirements for the double degree:
“Advanced Master in Naval Architecture” conferred by University of Liege
"Master of Sciences in Applied Mechanics, specialization in Hydrodynamics, Energetics and Propulsion” conferred by Ecole Centrale de Nantes

developed at West Pomeranian University of Technology, Szczecin
in the framework of the

“EMSHIP”
Erasmus Mundus Master Course
in “Integrated Advanced Ship Design”

Ref. 159652-1-2009-1-BE-ERA MUNDUS-EMMC

Supervisors:	**Prof. Remigiusz Iwańkowicz, West Pomeranian University of Technology, Szczecin, Poland** **Prof. Jean-David Caprace, Federal University of Rio de Janeiro, Brazil**
Internship Tutor:	**M.Sc. Nicole Schenk, Fr. Lürssen Werft GmbH & Co. KG, Germany.**
Reviewer:	**Prof. Hervé Le Sourne, Institut Catholique d'Arts et Métiers, France.**

West Pomeranian University of Technology, Szczecin
February, 2018

Blank page

DECLARATION OF AUTHORSHIP

I, Marcio Fleming, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Accuracy Control and Welding Distortion Prediction in a Deck Plate.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

8. I cede copyright of the thesis in favour of the University of West Pomeranian University of Technology

Signed: ……

Date: …….

Blank page

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Many industrial processes make use of welding to assemble structural parts. While this is standard procedure, the high temperatures encountered during the welding process can generate distortions in the base metal. These distortions negatively impact the parts by generating reworks in order to overcome them. This detrimental effect cannot be avoided, but it can be controlled. One way of performing such control is to add an allowance, which is an accepted amount of distortion. However, a prediction tool is elementary to determine the necessary tolerances.

Nowadays, the prediction methods can be grouped in three main approaches: experimental, computational (finite-element method) and machine learning. While the first two methods have been well studied, the machine learning approach is not as well understood.

The aim of this study is to explore machine learning algorithms such as neural networks and polynomial regressions in order to come out with a prediction model. Along with this, a best fitting study took place so that a simple formula for predicting metal distortion could be outlined.

In order to create some prediction models, the Knime software was used and main design parameters were gathered. The model's workflow has been organized so that several cases could be tested.

By using the workflow, results for eight variables were obtained. Nonetheless, the results were not satisfactory due to limitation of given data. Hence, the problem has its variables reduced to two, which increases the accuracy of the model.

Finally, it was possible to generate some models for the welding distortion prediction and it has been proven that these methods can be applied to such problems. Yet, they were not as accurate as the best fitting method, which means that more data is required so that the accuracy can be improved.

Keywords: Accuracy Control; Welding Distortion; Prediction.

1. INTRODUCTION

This first chapter introduces the welding distortion problem. Truly, most of processes in various industries make use of welding in order to assemble or join structural parts. The welding process of heating and cooling generates distortions which is an issue impossible to avoid but possible to be controlled. By trying to restrain such behaviors, the residual stresses are enhanced which can reduce the structural resistance.

Also, the usage of thinner plates causes the material to deform more than thicker plates. These distortions can be adjusted by the straightening process with the purpose of either correcting the appearance or the structural functionality. Nonetheless, the correction process generates extra cost due to reworks and it may also generate schedule delay that can infer in contractual penalties.

Having said that, it is desirable to have a prediction tool so that preventive actions can be taken and the welding distortion controlled within an acceptable limit.

The second chapter briefs the welding distortion problem along with the sorts of distortions and the parameters that have already been identified by other researchers. Even though, transverse, longitudinal, buckling, longitudinal bowing and angular were identified, only the transverse and longitudinal (in-plane distortions) are analyzed in this study due to the limitation of data available. In addition, three main group of influencer parameters have been outlined from the literature review: geometric parameters (design); material properties; and welding process parameters (manufacturing process).

Moreover, it was identified three main applications of welding prediction tools which are: formulas based on experimental setups; finite element models; and machine learning methods. Whereas there are several proposals for analytical and computational methods, there is little exploration of the machine learning methods. Another advantage of machine learning methods over analytical and computational methods is that it is less time demanding, coming to meet the time frame of this study. On top of that, a machine learning method can be reused to learn distortion for other elements whenever data is available requiring less time to be implemented. Hence, this study focused on offering a solution by making use of machine learning algorithms. Finally, the available methods of machine learning are commented and a chart presenting the distribution of found articles is displayed.

The third chapter describes the methodology used in order to describes the work environment, the actual measurement check-up, the selection of the methods to be analyzed, the database elaboration, historical data collection, the modelling process and the understanding of the contributing factors.

The chapters four discretizes the case of study by outlining the measurement procedure, the selection of a program in order to develop the model and a detailed modelling description.

Ultimately, chapters five and six present the results, analysis and conclusions. The aim of this study is the development of a prediction tool that can provide work-arounds for the welding distortions avoiding production problems that might arise during the manufacturing process and, additionally, to evaluate contributing factors.

2. LITERATURE REVIEW

The first section of this chapter details the welding distortion problem by commenting on how it arises, in which processes, what are the sorts of welding distortions and their influencer parameters. Additionally, the second section presents the available types of prediction tools that were find in the literature. As well, the third section details the machine learning methods. Finally, the fourth section presents an overview of the found researches and further state the motivation of selecting the machine learning approach.

2.1. Welding Distortion

Welding technology is greatly used in various areas in order to assemble structures of different usages due to its high productivity (Deng 2010) (Mahendramani and Swamy 2012) (Deng, Murakawa, and Liang 2007) (Deng and Murakawa 2008b). In addition, both low strength and high strength materials are used while producing the structural part in shipbuilding.

Whereas the low strength materials are cheaper, high-strength steels have been preferred to be used while producing steel structures in order to handle large amount of stress while having a better strength-to-weight ratio and to reduce topside weight, improve fuel consumption and enhance mission capability (Yang et al. 2014) (Deng and Murakawa 2008b). Moreover, the thinner structures are more likely to deform during welding since they have a lack of rigidity (Yang et al. 2014).

Welding distortion is a consequence of the non-uniform expansion and contraction of the welded material and the adjacent base material while the heating and cooling cycle of the welding occurs (Yang et al. 2014) (Deng, Liang, and Murakawa 2007) (Deng and Murakawa 2008a).

In addition, the residual stress occurs in a welded joint and this stress reacts to produce internal forces, provoking shrinkage or length deficiency of the main plates when comparing to the design dimensions (Kim et al. 2015).

As well, the welding distortion can lead to detrimental during the fabrication and service. Having said that, the welding distortion acts an initial imperfection of welded components (Yang et al. 2014). Any complex structure is subjected to welding deformation and the more complex the structure is, the greater the problems which will be inherent (Mahendramani and Swamy 2012).

As a matter of fact, it is impossible to avoid the welding-induced distortion during the assembly process (Yang et al. 2014) (TAJIMA et al. 2007) (Deng, Murakawa, and Liang 2008). On the other hand, it is possible to fabricate the structures with an acceptable level of accuracy to avoid the problems in the course of assembly (Deng, Murakawa, and Liang 2007). Thus, it is the designer and constructor duty to find the best balance between cost and acceptable limits in order to satisfy all stakeholders. As a result, the welding distortion reduces the fabrication accuracy while increasing the cost and working time to perform the necessary corrections (Yang et al. 2014).

One category of the distortions is the shrinkage which generates a difference between the dimensions of the actual parent metal and the dimensions of the design (Kim et al. 2015) (Deng, Murakawa, and Liang 2008). Additionally, the shrinkage induces to low quality in the production of ship blocks and reworking which decreases the productivity (Kim et al. 2015) (Deng and Murakawa 2008b). By that it can be inferred, there might be a delay in the project in case of the task being under the critical path and there is going to be an increase in the cost due to rework to adjust or to redo it.

Further, correcting unacceptable distortions is extremely costly and sometimes impossible. Excessive lateral distortion decreases the buckling strength of the structural members while under compressive loading (Mahendramani and Swamy 2012) (Deng, Liang, and Murakawa 2007) (Deng and Murakawa 2008b). The straightening process is used to reduce the deformations, mainly spot and line heating. However, it is mainly manual, costly and time-consuming (TAJIMA et al. 2007) (Deng, Murakawa, and Liang 2007) (Deng, Liang, and Murakawa 2007) (Deng and Murakawa 2008b).

Another problem is that, welding deformation and welding residual stresses are effects which oppose each other. Hence, while giving restraint to welding material in order to avoid deformation, the residual stress is increased. On the other hand, if the material is not restraining larger deformations will happen but less residual stress is decreased (Kim et al. 2015).

As a consequence, these initial imperfections can influence the structural behavior under variable loading and they can reduce the buckling strength of the structure (Yang et al. 2014). An additional issue, is that these imperfections lead to misalignment of the structural elements which will require straightening processes (Mahendramani and Swamy 2012).

Correspondingly, the welding shrinkage, distortion, and residual stresses are significant issues during the manufacturing process of welded structures made out of steel (Yang et al. 2014) (Mahendramani and Swamy 2012).

While constructing the ship hull, welding is greatly adopted to join stiffeners to plates, build subassemblies and blocks, and finally to join these blocks and assemble the ship hull (TAJIMA et al. 2007) (Deng, Murakawa, and Liang 2007). Usually, these blocks are all-welded, thin-plate structures (Deng, Murakawa, and Liang 2007). The shipbuilding construction can be categorized into the following stages:

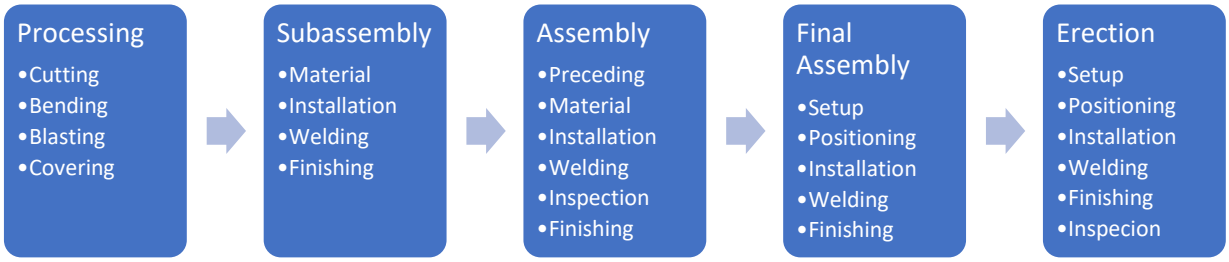

Figure 1 – Shipbuilding Construction - Flowchart

2.1.1. Sorts of Welding Distortion

Truly, the welding process is present in most of the processes and it is used to join the structural members. Furthermore, the welding distortions can be classified into:

- Transverse shrinkage (In-plane mode) (Yang et al. 2014) (Mahendramani and Swamy 2012) (Deng, Murakawa, and Liang 2007)

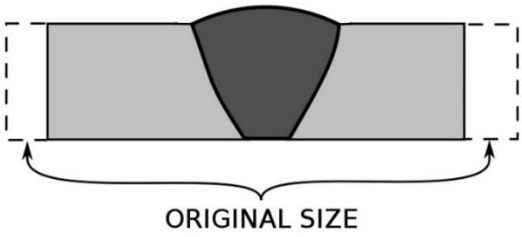

Figure 2 – Example of Transverse Shrinkage (Welding Defect 2017).

- Longitudinal shrinkage (In-plane mode) (Yang et al. 2014) (Mahendramani and Swamy 2012) (TAJIMA et al. 2007)

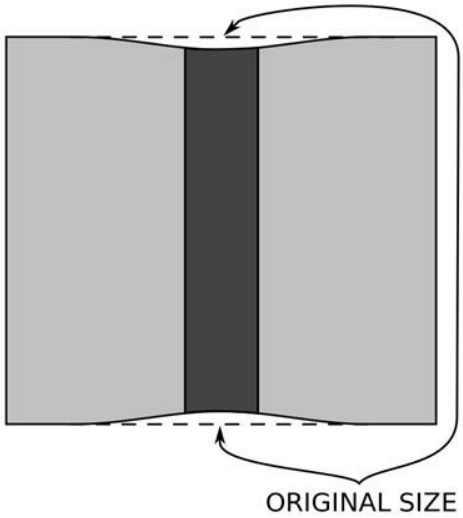

Figure 3 – Example of Longitudinal Shrinkage (Welding Defect 2017).

- Buckling (Out-of-plane mode) (Deng 2010) (Yang et al. 2014) (Mahendramani and Swamy 2012) (TAJIMA et al. 2007) (Deng and Murakawa 2008b)

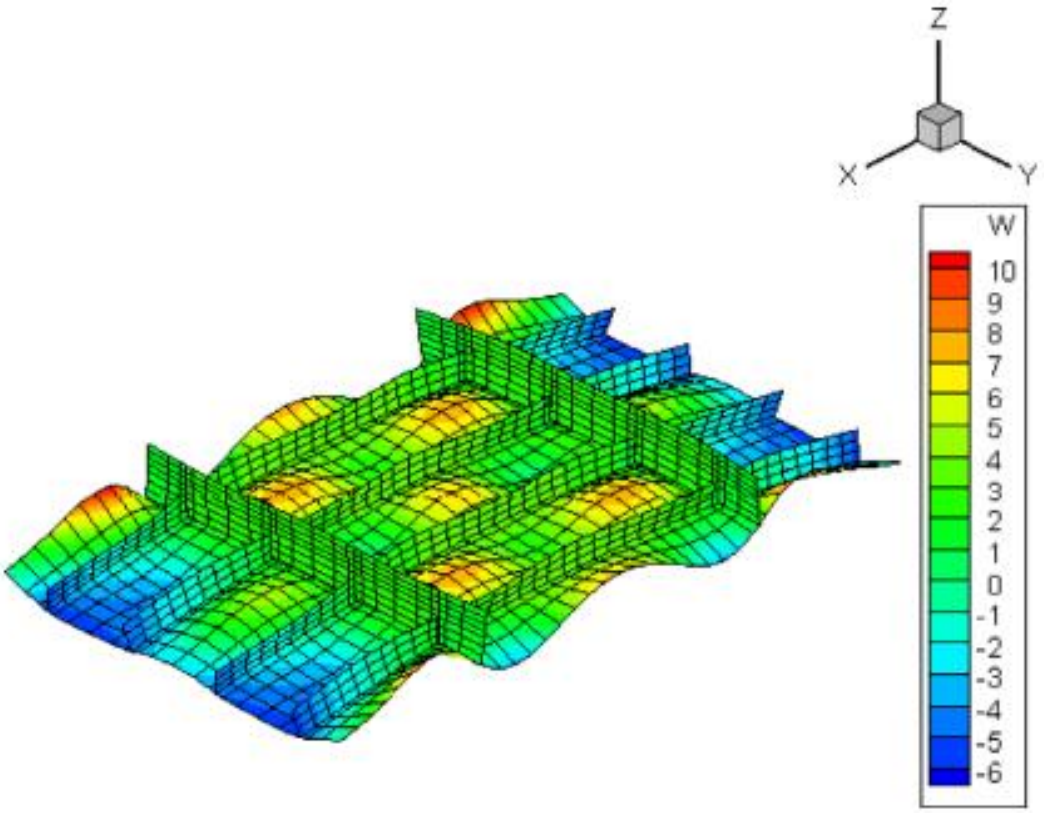

Figure 4 – Example of Buckling (Deng and Murakawa 2008b).

- Longitudinal bowing (Out-of-plane mode) (Yang et al. 2014)

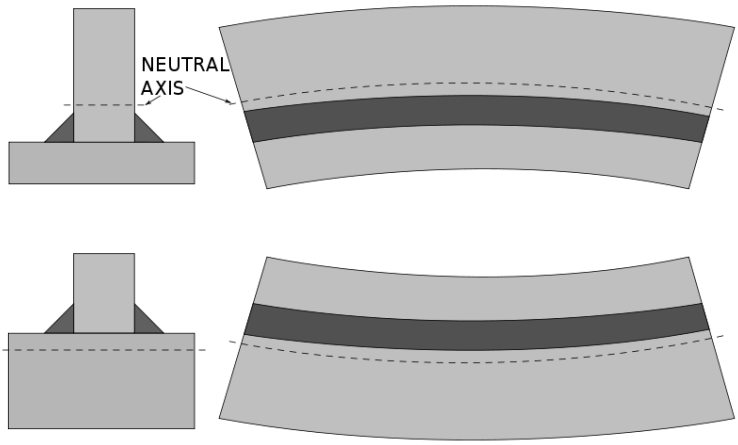

Figure 5 – Example of Longitudinal Shrinkage (Welding Defect 2017).

- Angular (Out-of-plane mode) (Yang et al. 2014) (Deng, Murakawa, and Liang 2007) (Deng, Liang, and Murakawa 2007)

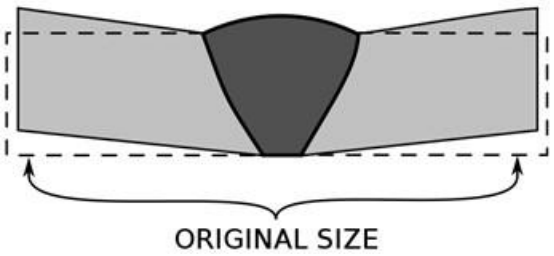

Figure 6 – Example of Angular Distortion (Welding Defect 2017).

2.1.2. Influencer Parameters

While reading the research papers, the parameters which are most likely to cause those deformations can be:

- geometric parameters (design) (Deng 2010) (Yang et al. 2014) (Mahendramani and Swamy 2012) (Deng, Murakawa, and Liang 2007);
    - dimensions of the structure (plate thickness; weld length and stiffeners' spacing)
    - type and size of welded joints
- material properties (Deng 2010) (Yang et al. 2014);
- welding process parameters (manufacturing process) (Deng 2010) (Yang et al. 2014) (Mahendramani and Swamy 2012) (Deng, Murakawa, and Liang 2007) (Deng, Murakawa, and Liang 2008);

- heat input;
- welding sequence;
- preheating;
- post heating;
- gaps;
- tack welds;
- the edge preparation;
- welding conditions;
- interpass temperature;
- the shape of penetration;
- positioning;
- welding procedure;
- and the degree of restraint during welding.

2.2. Prediction Tools

According to (Gray, Camilleri, and McPherson 2014), there are two major categories of tools to understand the mechanics of welding distortion in order to provide better strategies so that better control of this phenomenon can be achieved: Artificial Neural Networks and Computational Simulation.

However, there are two more other applications that can be quite useful for the matter, which are Fuzzy Logic and Feature Selection Method.

Finally, the Feature Selection and Artificial Neural Networks are sub-areas of machine learning which are closely related to statistics. Moreover, the fuzzy logic is based on the degrees of truth instead of probability prediction. As a matter of fact, these three methods can be seen as approaches that do not require fully physical discretization of the material and production processes. On the other hand, the computation simulation requires a good discretization of the model, including all physical phenomena or satisfactory approximation through analytical or numerical methods.

2.2.1. Artificial Neural Networks (ANNs)

The aim of an ANN is to infer functional relations between the observations and phenomena. When used for practical application, the objective could be to establish an empirical model that will relate the input of fabrication process to the likely deformation outcomes. As well, the process of establishing the relations is mainly statistical. Nonetheless, the methodology has the capability of deducing the interactions along the hidden layers and thus revealing influences which may not be promptly recognized when studying physical models.

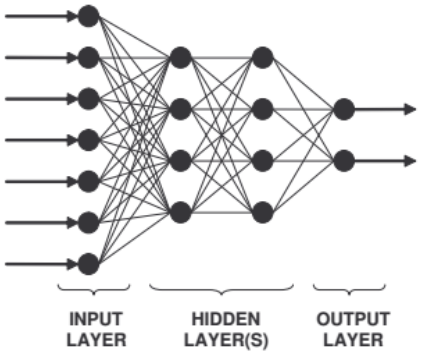

Figure 7 – Example of Artificial Neural Network (Caprace et al. 2007).

2.2.2. Fuzzy Logic

Moreover, another option is to use a fuzzy method which is consists of a system that provides a non-linear mapping between crisp input variables and crisp output ones and allow the use of linguistic expressions for the rules which define the input-output relationship (Caprace et al. 2009).

2.2.3. Feature Selection Method

According to (Chandrashekar and Sahin 2014), feature selection which can also be understood as variable elimination assists on understanding the data, reducing computational time, reducing the effect of curse of dimensionality and improving the predictor performance.
The aim of feature selection is to decide a subset of variables from input that can readily describe the input data while minimizing the effects from noise and irrelevant variables and still provide good prediction results.

Indeed, the variable elimination methods are broadly classified into: filter and wrapper methods. As the objective of this report is to briefly present the tools that can be used and as they are a lot of literature on the methods, the main concepts, some advantages and drawbacks are going to be shortly introduced.

2.2.3.1. Filter Methods

Filter approach makes use of ranking techniques, such as Pearson correlation criteria or mutual information (MI), in order to classify the important variables and remove variables below a threshold. They are used due to their simplicity and satisfactory success is reported for practical applications.

Some other favorable advantages that can be outlined are: it is computationally light and avoids overfitting and is proven to be good for certain databases; and they do not rely on learning algorithms.

On the other hand, the drawbacks are: the selected subset might not be optimal in that a redundant subset might be obtained; Some ranking methods do not discriminate the variables in terms of the correlation to other variables; Finding a suitable learning algorithm can also become hard since the underlying; and there is no ideal method for choosing the dimension of the feature space.

2.2.3.2. Wrapper Methods

Whereas the filter methods make use of a feature relevance criteria, the wrapper methods rely on the classification for obtaining a feature subset. Actually, wrapper methods utilize the predictor as a black box and the predictor performance as the objective function to evaluate the variable subset. They can be classified into Sequential Selection Algorithms and Heuristic Search Algorithms. The first one, begin with an empty set (full set) and add features (remove features) until the maximum objective function is obtained. A criterion is chosen which incrementally boosts the objective function until the maximum with the minimum number of features. On the contrary, the heuristic search algorithms evaluate different subsets to optimize the objective function. Different subsets are generated either by searching around in a search-space or by generating solutions to the optimization problem.

The main drawback of wrapper methods is the computational iterations required to obtain a feature subset. Another disadvantage is that these methods use classifier performance the objective function and by doing so they are prove to overfitting. In order to overcome the last downside, a separate holdout test set can be used to guide the prediction accuracy of the search.

2.2.3.3. Embedded Methods

Embedded methods target to shrink the computation time taken up for reclassifying different subsets which is performed in wrapper methods. In order to overcome it, the main way is to incorporate the feature selection as part of the training process.

2.2.3.4. Other Techniques

The techniques mentioned above are feature selection techniques using supervised learning. For instance, the output class labels of the data are known or could be derived. However, there are situations where operation details are unknown but their operational data is available. One of the mentioned methods is clustering techniques.

In addition, there are situations where there are partially known and unknown data and the semi-supervised learning can be applied. They consist of a mixture of unsupervised and supervised learning.

Another mentioned technique is ensemble feature selection where a single feature selection algorithm is run on different subsets of data samples obtained from bootstrapping method. The results are aggregated to obtain a final feature set.

2.2.4. Computational Simulation

On the contrary of the ANNs, the computational simulation relies on the real physical models in order to discretize the complex thermos-mechanical behavior induced by welding. This is being more used nowadays because of the rapid development of finite-element analysis (FEA) and the availability of more computing resources. However, the experimental results remain invaluable.

2.3. Machine learning

Machine learning can be defined as a field of research regarding automated large-scale data analysis (Barber 2012). Also, to better enlighten, the machine learning is a mimic of the human brain or biological systems where there is not a structured-defined algorithm stating the exact rules, but instead data is given to be learnt from. Hence, being possible to construct a better and useful approximation (Alpaydın 2014).

Truly, the machine learning includes many of the traditional areas of statistics, nonetheless, focusing on mathematical models and also prediction (Barber 2012).

Furthermore, there are a lot of methods within machine learning. Inclusive, sometimes, there might be difficult to strongly identify each one, since there can be a small-scale difference from one method to the other. In this section, we will define the main concepts and then target on the one the better suits our problem.

On the contrary, there are mainly three different categories of machine learning: supervised, unsupervised and semi-supervised. Additionally, there are some others which can be understood as one or more of them plus some additional features and they will be briefly discussed.

Finally, some known applications will be expressed in order to advance the details in the topic.

2.3.1. Supervised Learning

According to (Barber 2012), the supervised learning can be defined – Definition 13.1 - as follows. Given the set of data $D = \{(x^n, y^n), n = 1, \dots, N\}$ the task is to learn the relationship between the input x and output y such that, when given a novel input x^* the predicted output $y^*$is accurate. The pair (x^*, y^*) is not in D but assumed to be generated by the same unknown process that generated D. To specify explicitly what accuracy means one defines a loss function $L(y^{pred}, y^{true})$ or, conversely, a utility function $U = -L$.

There are two types: classification problem or regression problem. A classification problem occurs when the output is one of a discrete number of possible “classes”. For instance, the bank would like to analyze good and bad credit costumers when providing them a loan. The other problem, regression, is when the output is continuous – numbers. For example, we would like to predict the house’s selling prices based on their features.

2.3.2. Unsupervised Learning

According to (Barber 2012), the supervised learning can be defined – Definition 13.2 - as follows. Given the set of data $D = \{(x^n, y^n), n = 1, \dots, N\}$ is unsupervised learning, we aim to find a plausible compact description of the data. An objective is used to quantify the accuracy of the description. In unsupervised learning, there is no special prediction variable so that, from a probabilistic perspective, we are interested in modelling the distribution p(x). The likelihood of the model to generate the data is a popular measure of the accuracy of the description. This process is also known as density estimation in statistics.

2.3.3. Semi-supervised Learning

The semi-supervised learning is the mixture of both supervised and unsupervised learning. Hence, you will have partial data with output outlined and the other part will be only input data. By that, you will try to make use of the unsupervised learning to enhance the results that would be made only by the supervised learning with partial data.

2.3.4. Reinforcement Learning

According to (Alpaydın 2014), some applications, the output system is a sequence of actions. Cleary, a single action does not play a major role, but the sequence of right actions to achieve the goal – also known as policy. In addition, there is not a best action in any intermediate step. Yet, an action is considered good if it is part of a good policy. Hence, the machine learning program should be able to assess the policies and learn from past good action sequences in order to create a policy.

2.3.5. Deep Learning

In simple words, the deep learning is the implementation of the artificial neural network with a feature selection method but instead of removing the feature the deep learning will assign weights for the features not eliminating them.

2.3.6. Data Pre-processing

Which is the process of eliminating the out-of-range values, impossible data combinations, missing values, etc.

2.3.7. Online machine learning

In the online learning, the data keeps being updated subsequently when there are new data available. It may be for supervised or unsupervised context.

2.3.8. Dimensionality reduction

The dimensionality reduction can be divided into two main categories: feature selection and feature extraction.

2.3.8.1. Feature Selection

This is the process of assessing the features and them eliminating them in order to reduce the number of variables analyzed.

2.3.8.2. Feature Extraction

Feature extraction is the creation of additional features by combining existing ones and creating more meaningful features.

2.4. Research Paper

During the literature research, it was possible to find 51 articles related to welding distortion prediction. The Table 1 and Figure 8 below demonstrate their distribution among three main categories.

Table 1 – Welding Research Papers' Distribution

Classification	Counting	Percentage
Analytical/Experimental	14	27.5%
Computational	36	70.6%
Machine Learning	1	2.0%
Total	**51**	**100.0%**

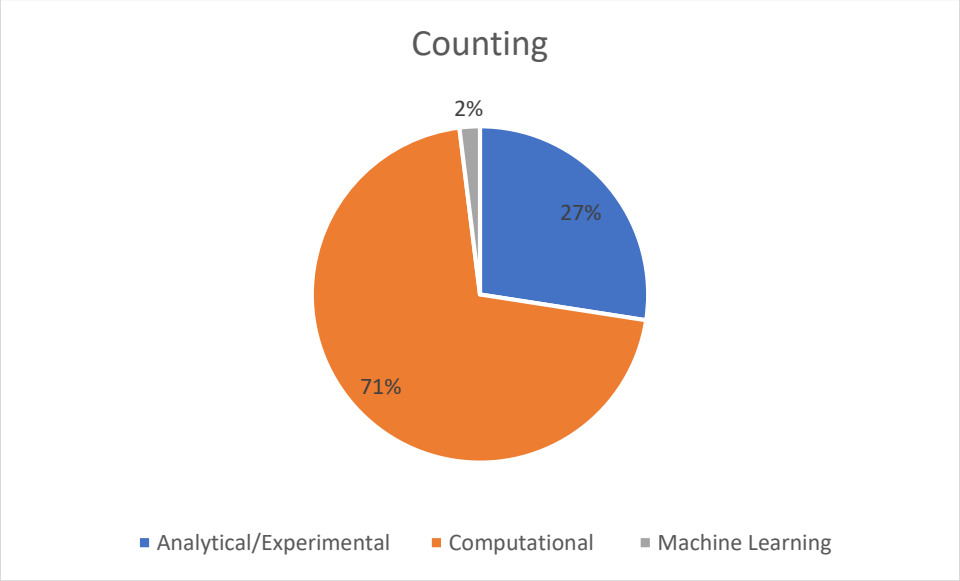

Figure 8 – Welding Research Papers' Distribution

By looking at them, it possible to see that most of the studies are dedicated to Computational methods. As well, when reading all papers is possible to see a trend from moving from analytical and experimental to Computational.

As a matter of fact, it is important to highlight that the welding distortion is a non-linear process which has a lot of influencer parameters. Due to data limitation, the only evaluated distortions in this study are the transverse and longitudinal distortions.

Additionally, the tool is going to be used by CAM/Nesting department which possesses design information only. Due to this fact and to limitation of time, the design parameters that can be retrieved were the parameters analyzed.

As the case of study is performed directly into a company which retains measurement information about their own process, computational and learning machine methods are preferred over the experimental ones.

As it had been said before, the computational method requires a considerable time to setup the model due to the need of explicitly stating all physical phenomena, and also to validate it. The internship period lasted four months and this time limitation could lead to a non-accurate solution. Truly, all methods are restrained within their limitation determined by the range of setups in which they are developed. Anyhow, the machine learning models have a strong capability of being reused with other set of data, adapting themselves with less effort on the setup procedure for new schemes. Another advantage of the machine learning is the validation process. The experimental and computational results might require additional subjects to be tested while the machine learning methods can separate part of the data to perform their validations.

As well, the company's intention is to explore even more the aspects of the welding distortion starting from the deck plate and evolving to each structural element.

Ultimately, due to the time frame, reusability, adaptation to new scenarios and exploration of new methods, the machine learning methods were adopted in order to explore solutions that could meet all requirements.

3. METHODOLOGY

Firstly, when starting the internship, a meeting with all stakeholders took place in order to confirm the expectations of the study subject and available resources and data in the company.
Next, a mapping of the process was developed in order to verify ensure that all possible design factors that can influence the process are covered.
Afterwards, a benchmark study of some available software to develop the prediction tool has been performed. Clearly, there are several learning methods available. Anyway, the ones included in this study were outlined.
Following, the database elaboration is detailed. The characteristics can be grouped in three major categories: block characteristics which are estimated by a personnel staff of the company; welding characteristics which are manually retrieved from the 2D drawing; and the welding characteristics which can be gathered directly from the 3D model.
Subsequently, the collection of the data would take place. Then, the prediction tool was modelled, and the contributing factors studied.
Finally, the results are explored, discussed and analyzed, and the conclusions are drawn.

3.1. Process Mapping

Firstly, a tour around the production facilities has been provided during the first days. Subsequently, the measurement team has been shadowed for 1-2 weeks where there was the opportunity to get to better know the processes. After this, the process has plotted in a flowchart which had been presented through some of the colleagues of various departments in order to assess it. After the compliance with the process, a brief description of the activities has taken place. Unfortunately, due to a policy, no photographs were taken during the internship period. The mapping of processes is detailed in Appendix I and as per non-disclosure agreement is restricted.

3.2. Verification/Implementation of monitoring tools

While accompanying the measurement crew, a checkup of the monitoring tools was done. However, there is only the verification of the distortions *in loco* and their records transferred into reports. The measurement team is limited and any change to process would require more resources which could require more time than the one proposed for the topic. Therefore, it was decided to make use of the existing system. Below find the description of the measurement process for longitudinal, transverse and angular distortions.

3.2.1. Measurement in (x,y)

The equipment used in order to perform the measurements in the "X, Y" direction is the Sokkia total station. Truly, this instrument consists of one optical scanning which is pointed to a prism or a cube where the point of the desired data is. As well, the equipment is also integrated with a small portable gadget which can allow the used to be away from the optical unit in order to trigger the measurement. Finally, the organization of measurements can be inputted inside this computer in order to expedite the service onsite, for instance: ship block number, target locations, etc.
When facing troublesome measurement positions, it is possible to measure additional points to create a new reference and then measure the aimed position. Below find an illustration of the equipment.
The procedure in order to measure the points goes as it follows. Firstly, the target points are setup into the software. When onsite, the best location in order to measure all points is selected. After that, the tripod is mounted, and the total station is attached to it. Next, the leveling of the total station must be done in order to achieve satisfactory accuracies. Following, the prism cube or prism stick which are the aims of the total station optical unit are placed in one of the positions:

- Bow – Portside (Vorne – Backbord)
- Bow – Centerline (Vorne – Mittellinie)
- Bow – Starboard (Vorne – Steuerbord)
- Stern – Portside (Hinter – Backbord)
- Stern – Centerline (Hinter – Mittellinie)
- Stern – Starboard (Hinter – Steuerbord)

Figure 9 – Measurement Device – X, Y direction – Sokkia Total Station (Product_cx_05.Jpg (375×310) n.d.).

The optical unit is pointed to the cube or stick, the operator indicates in the computer which is the point of measurement and the computer is triggered in order to obtain the measurement. The step is repeated until all desired points are collected. A representation of the main points is shown Figure 10.

As it had been mentioned before, different shapes or more complex structures might require extra points in order to give better view or even as coordinates in order to acquire main points, for instance: main plates divided into two heights, main plates with cut-off for doors or other openings, etc.

If there is no possibility to reach all points from the initial position, then reference points are placed on the walls of the workshop or nearby structures and then the equipment is unmounted and remounted in a more favorable location.

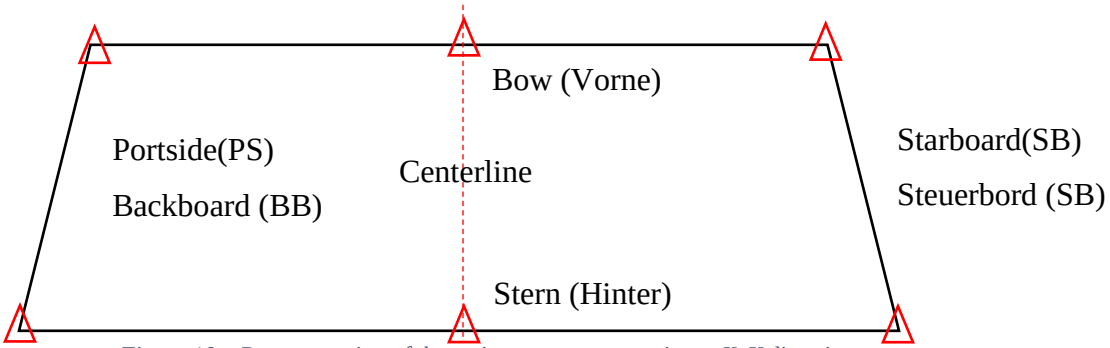

Figure 10 – Representation of the main measurement points – X, Y direction

Ultimately, the data is retrieved by a software in the computer and the data is transferred to an Excel spreadsheet which is used only for control and its named “Messprotokoll”. The retrieved information is used for the positioning approval and on-site corrections, not being saved as a databased nor becoming a feedback to the department which generates the allowance.
In addition, for each point, the table below is filled-in inside an Excel spreadsheet:

Table 2 – Input spreadsheet sample

	X	Y
Ist v.d. Schweißen (before welding)		
Ist n.d. Schweißen (after welding)		
Soll (Original DWG Target)		

3.2.2. Measurement in (z)

In addition, to check the discrepancies in z direction another equipment is used which consists of electronic laser device along with a stick with the receiver of the laser as shown in Figure 11. Some points (6-8) are marked along with the walls on the below level and they are checked. Manual corrections using hydraulic jack-ups, cutting processes or shimming pieces are used. After the corrections the block is released so another can be erected. The data is compiled into a report so named “Montageprotokoll”.

Figure 11 – Representation of the main measurement points – Z direction (61Z360TshtL._SL1001_.Jpg (1001×1001) n.d.).

3.3. Program Selection

As a matter of fact, there are a considerable number of tools available, which have been retrieved by searching online, in order to deal with machine learning which some examples are listed in the table below. As well, there is a nice review comparing the open source tools for data science which is written by (Wimmer and Powell 2016).

Table 3 – Software suites' samples

Free and Open Source	Proprietary with free and Open Source	Proprietary
R	Knime	Google Prediction API
Weka	Rapid Miner	MATLAB

As the company did not have any proprietary software for machine learning implementation, free and open source tools were considered in order to solve the proposed problem. Hence, the R, Weka, Knime and Rapid Miner were considered as feasible options and they are discussed below.

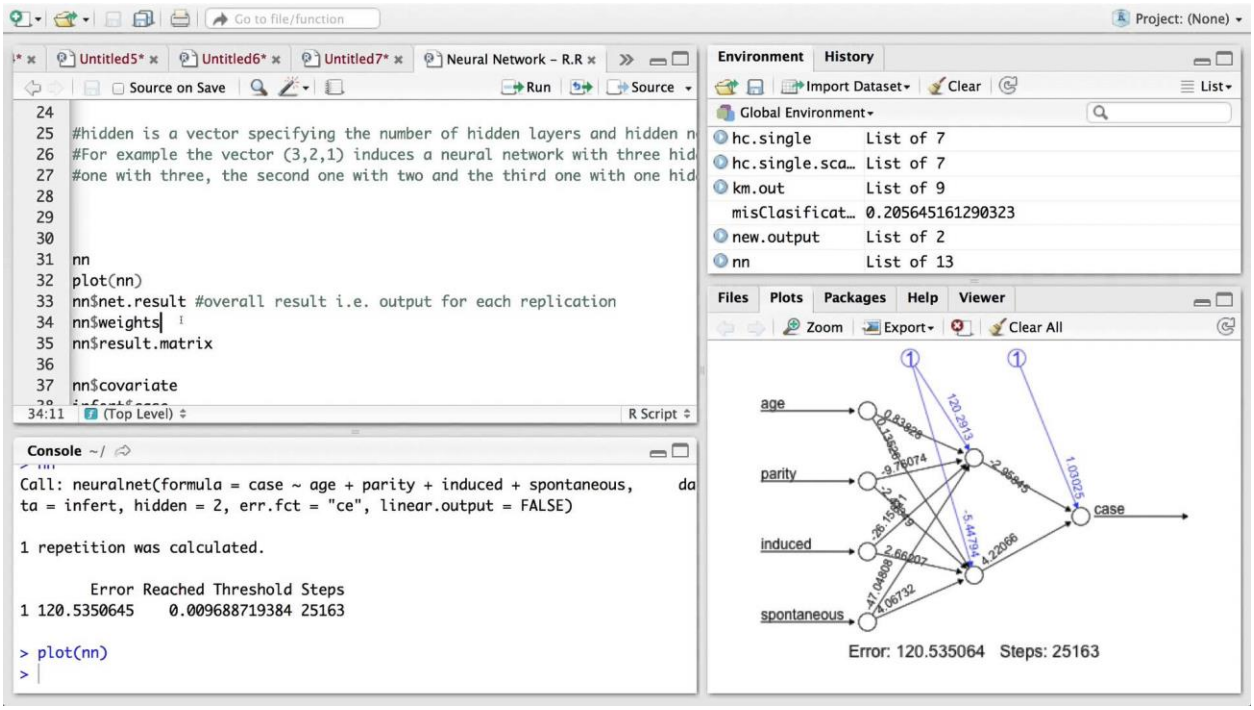

Figure 12 – Program Selection – R Studio – Screenshot (Maxresdefault.Jpg (1920×1080) n.d.)

Firstly, R is the simplest one, requiring a good level of programming skills in order to correctly setup a model. According to (Wimmer and Powell 2016), by default R does not provide visual features turning out to be difficult for a novice to create a workflow. Even though there is R Studio

graphical interface for R language, R language is still considered an interpreted language more than an environment. A view of the R studio is shown in Figure 12.

Secondly, Weka provides a more user-friendly environment by providing a more guided way to setup the model and being a collection of machine learning algorithms. As well, Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It also provides its own packages in order to reduce programing (Wimmer and Powell 2016). A screenshot of the GUI is presented in Figure 13.

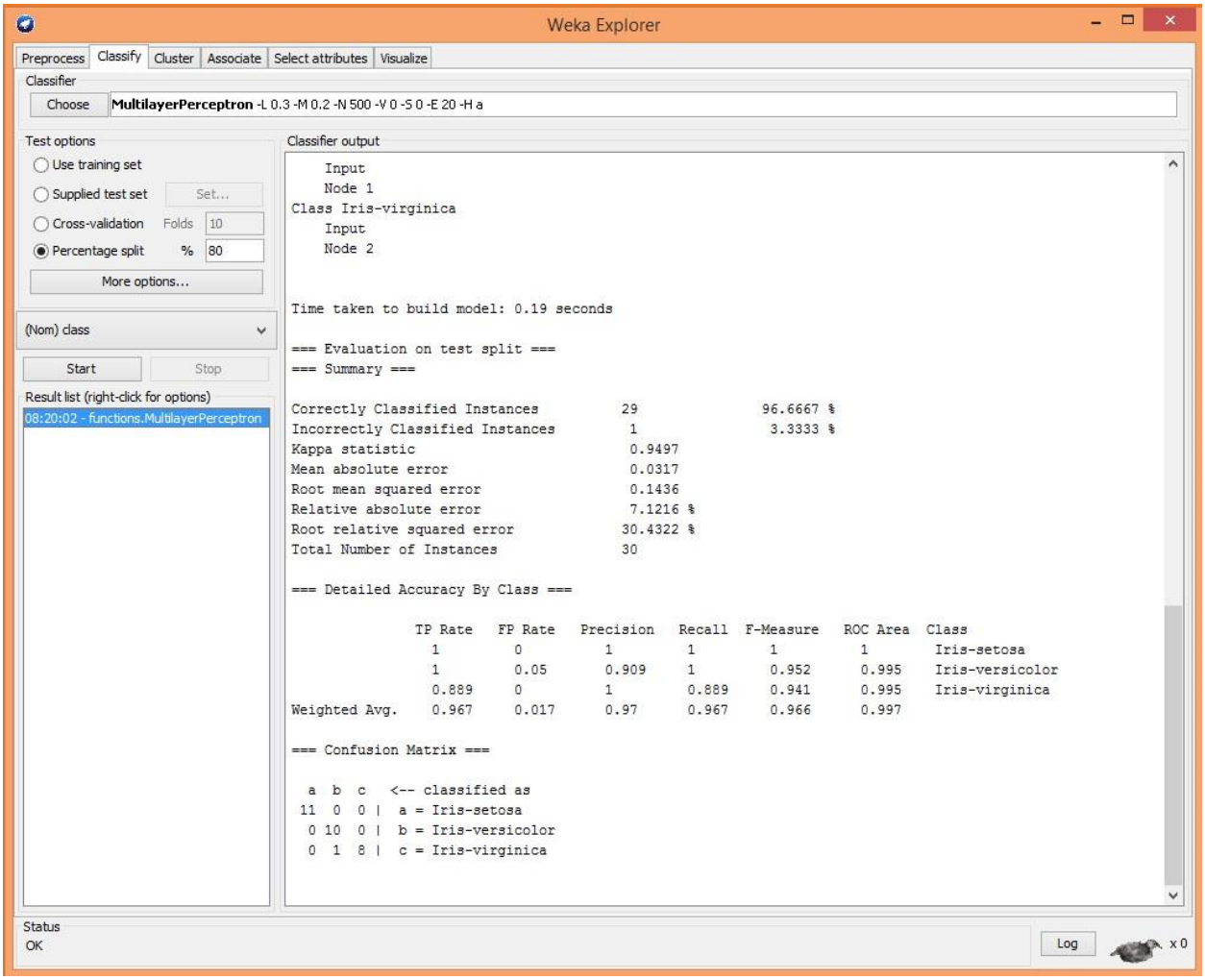

Figure 13 – Program Selection – Weka GUI – Screenshot (16_splitandstartrun.Jpg (1018×825) n.d.)

Thirdly, Rapid Miner has a lot of features to offer however some of them are limited in the free open source version. A positive point is the user-friendly workflow environment which allows you to create the flow without requiring high level of programming. A drawback of the free version is the limitation in memory access being 1 GB only (Wimmer and Powell 2016). Another downside

is the limitation of the community edition which grants you to work with only database up to 10,000 rows (Communications 2017). A glimpse of the software is provided in Figure 14.

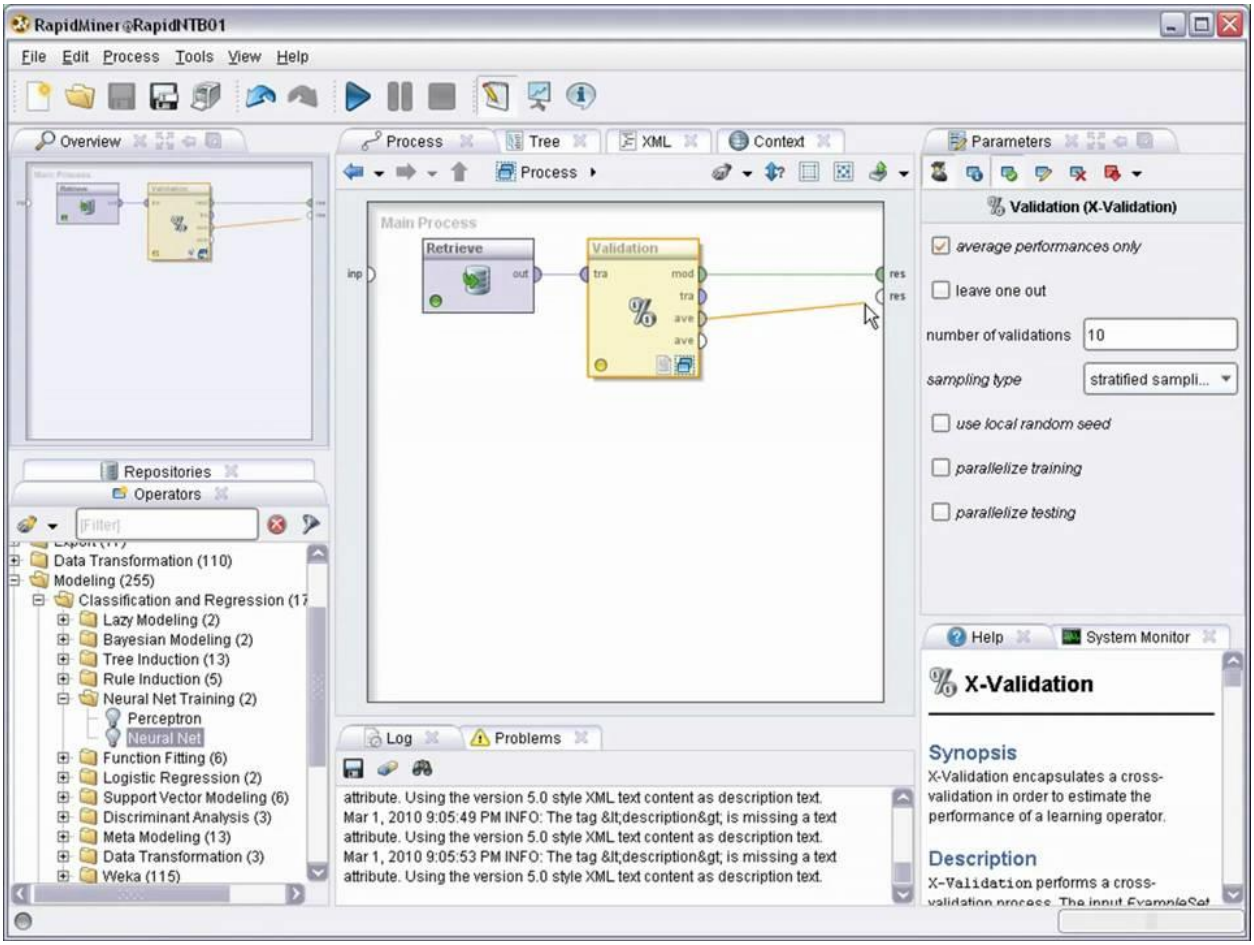

Figure 14 – Program Selection – Rapid Miner – Screenshot (Maxresdefault.Jpg (960×720) n.d.)

Finally, Knime is quite similar to Rapid Miner by offering a graphical workflow which facilitates the development of the model by a novice. In this sense, Knime has a slightly advantage on rapid miner by making use of a more colorful workflow with clear names for the nodes. Another advantage is the traffic lights for each node allowing the user to verify the flow between the model. In addition, Knime has integration with R language and Weka while Rapid Miner can only integrate with R language (Communications 2017). A representation of a workflow is demonstrated in Figure 15.

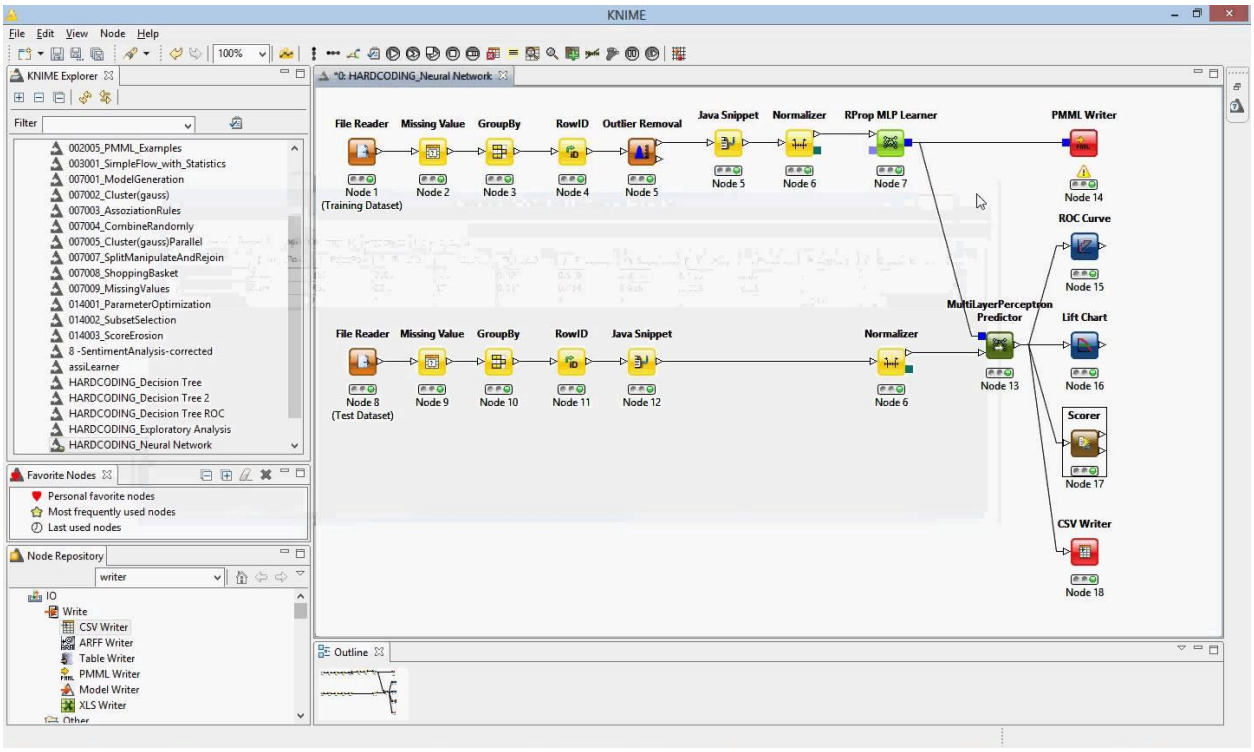

Figure 15 – Program Selection – Knime – Screenshot (Maxresdefault.Jpg (1440×900) n.d.)

A quick comparison matrix discussed by (Wimmer and Powell 2016) is partially presented below allowing a final overview of all software and their methods.

Table 4 – Open Source Tools – Comparison Matrix (Wimmer and Powell 2016)

Method	R	Weka	Rapid Miner	KNIME
K-means Clustering	Yes	Yes	Yes	Yes
Association Rule Mining	Yes	Yes	Yes	Yes
Linear Regression	Yes	Yes	Yes	Yes
Logistic Regression	Yes	Yes	Yes	Yes
Naïve Bayesian Classifiers	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes
Time Series Analysis	Yes	Yes	Some	Yes
Text Analytics	Yes	Yes	Yes	Yes
Big Data Processing	Yes	Yes	No	No
Visual Workflows	No	Yes	Yes	Yes

Ultimately, a free of charge tool should be selected which has driven us to open source solutions such as R, Weka, Rapid Miner and Knime. The first criterion in order to choose the software was the user-friendliness. Hence, Knime and Rapid Miner were the best options as they offer graphical

workflow facilitating the creation and application of the methods and not requiring high level of programming. The second criterion was the capability of processing data. While Rapid Miner limits the free version around 10,000 rows, Knime does not possess such restriction. The third criterion was the integration with other tools. Although both Rapid Miner and Knime offer the integration with R, just Knime explicit mention the integration with Weka. Having said that, the chosen software was Knime.

3.4. Selection of Method

Truly, Knime offers more than 1,500 modules for data science. Nonetheless, the selected methods were Neural Network, Linear Regression and Best Fitting.

As a matter of fact, the main advantage of the neural network is to deal with the nonlinearities which are resulted from a very nonlinear process such as welding.

As well, the company would like to verify their actual process and maybe to alter for a simple formula, therefore the linear regression also takes place.

Additionally, to complement the linear regression method, the best fitting method is developed.

3.5. Database Elaboration

As the proposed solutions should be oriented to be used by the CAM/Nesting team, main design parameters were preferred in order to elaborate the database. Therefore, three main sources of information were gathered to compose the database.

Firstly, the estimation of the main block characteristics which is developed by a personnel staff.

Secondly, the characteristics that could be outlined from the 2D drawings were analyzed and tabulated into a spreadsheet.

Thirdly, the characteristics which could be inferred from the 3D model database available from the company's design system.

Finally, all of the information was joined by the ship and section number resulting in 60 available blocks to be studied.

3.5.1. Block Characteristics

Block characteristics such as section number, main composition (Steel/Aluminum), length, width, height, volume and weight were estimated by one personnel and included here.

3.5.2. Welding Characteristics from 2D model – (Manual)

Later, each 2D drawing was verified and its elements classified into:

Transversal – Fillet weld of any element placed in the main plate which is positioned in the transversal direction (y); E.g. Frames and Carlings. Color: Light Blue

Longitudinal - Fillet weld of any element placed in the main plate which is positioned in the longitudinal direction (x); E.g. Stiffeners. Color: Green

Girder - Fillet weld of any element placed in the main plate which is positioned in the longitudinal direction (x) and has its sizing reminding a girder; Color: Pink

Butt Weld – Any butt weld. Color: Orange

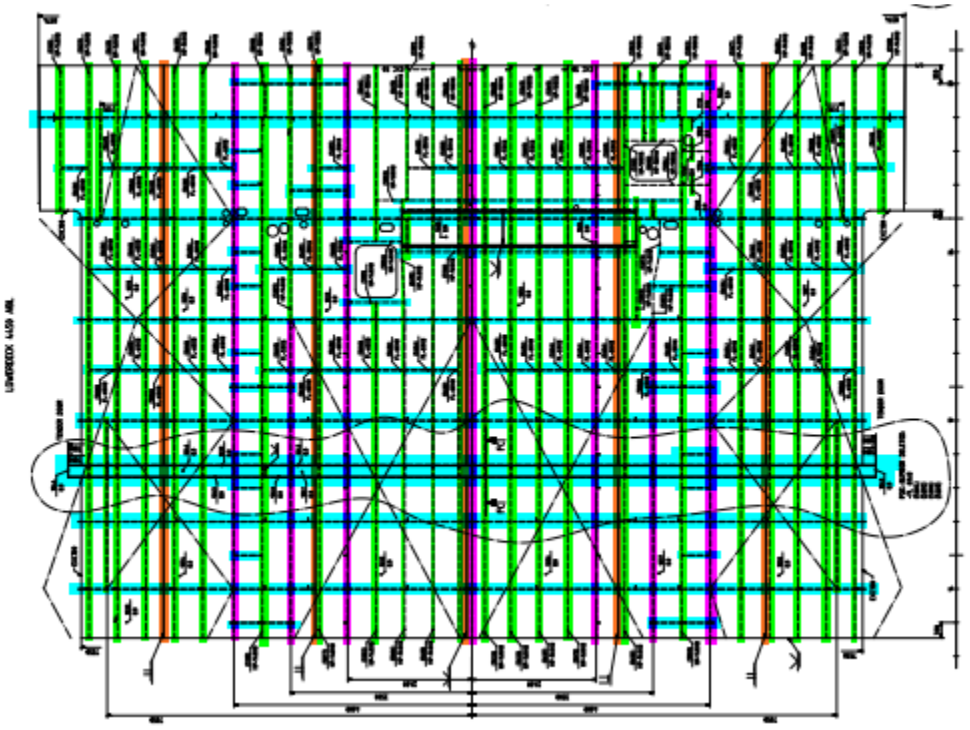

Figure 16 – Welding Research Papers' Distribution

In addition, after colors that would be representative, more than 50% of the plate, were counted as 1 complete element and this information was tabulated along with the spacing of girders, stiffeners

and frames. Moreover, the most representative thickness was included in the spreadsheet as well. Finally, as the spacing of girders and frames could be different between the blocks, an average value was taken for each section so that the machine learning methods could make use of these features.

3.5.3. Welding Characteristics from 3D model – (Automated)

This process was developed to outline the information of Plate, Girder, Stiffener and Frame automatically from the 3D drawing. It can express the similar information as in the topic above, but it cannot give the spacing. It had been created at first when the number of blocks were high, however when it dropped it was preferred to do it manually. On the other hand, it can still provide estimations of the welding length and weightage.

3.5.4. Collection of historical data

The process started around 2011 and its standards changed over the years. Hence, in order to collect all of those data, a common spreadsheet was used to gather all information and the data was collected from PDF and excel spreadsheets and, finally, some were directed retrieved from the total station files. The compilation and consolidations of these files into one database took around 2-3 months.

3.6. Modelling the prediction tool

The development of the model tool started in early September being improved to the most until December. During the development some study cases were assigned as it follows, at first all features were given to the model trying to predict the variables (x, y) in the for edges. As no good results were achieved, next the reduction of the variables (x, y) to one length and widthwise variables were done and tested again with all features. Better results were achieved, but with poor accuracy, hence the same model was tested but with one feature at each time Length and Width which were computed from the target points. Ultimately, a back-feature selection method was

applied to select the most relevant features and the model was run one more time with the selected features.

As a matter of fact, in order to enhance the learning of the outliers were removed by assessing the variation between the target and actual points and by using two different methods: the two-standard deviation which give us a 95% of confidence level; and the interquartile range.

In addition, a different proposal of using the principal component analysis (PCA) had taken place in order to reduce most of the features to two virtual features and the model was applied. Another difference between this attempt and the other models is that this model takes the outliers based on the PCA variables instead of considering the variation between the target and actual points.

3.6.1. General Overview

The model has been divided in three main categories “Loading Data”, “Pre-processing” and “Processing and Results” which can be visualized in the Figure 17 and Figure 18. After a first trial with all variables and considering X and Y direction, a second attempt reducing the points to measurements along X and Y direction were performed. In sequence, a trial with just the length and width variables took place in order to try to come up with a simple formula for the CAM/Nesting team. Finally, a back-feature selection method was implemented in order to outline the most significant geometric features in the design.

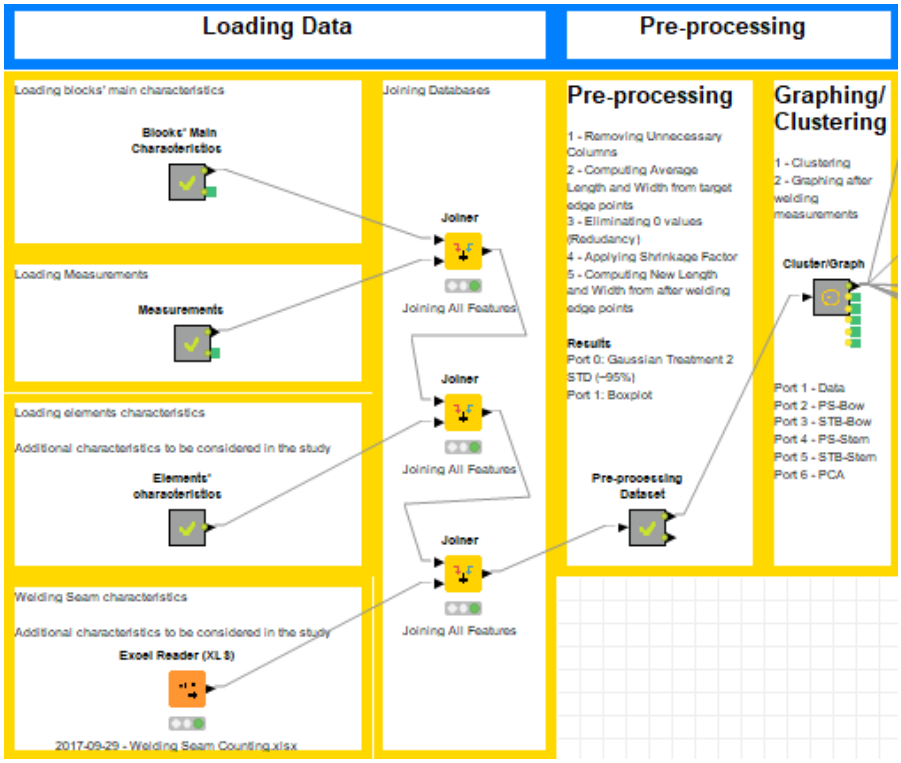

Figure 17 – KNIME Model – General Overview – Loading and Pre-processing

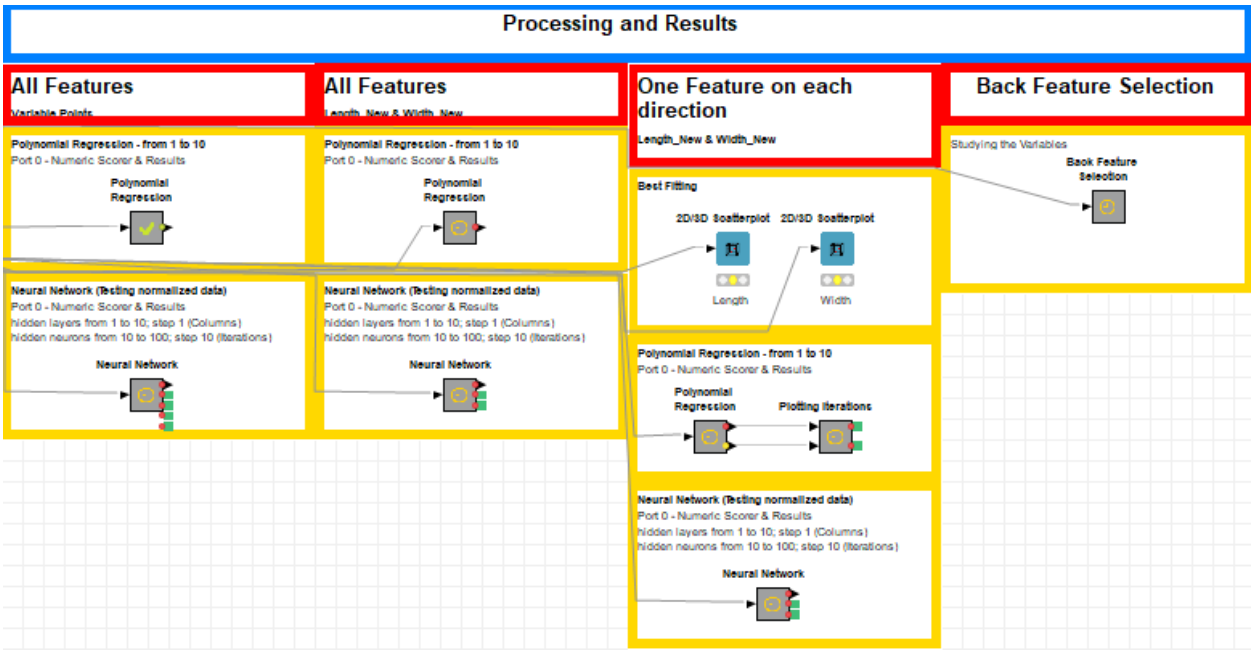

Figure 18 – KNIME Model – General Overview – Processing and Results

3.6.2. Loading Data

All data has been stored in Excel spreadsheets. Hence, the Excel Reader (XLS) node has been used so that the data could be load into Knime workflow. There are ten different spreadsheets that were loaded into Knime workflow.

3.6.2.1. Blocks' Main Characteristics

The first spreadsheet to be loaded was the blocks' main characteristics. This spreadsheet contains the ship number, the section number, main material used, average length, average width, estimated volume and estimated weight. These characteristics were estimated by a personnel staff from the Ship Construction department. The workflow representation is shown below in Figure 19 and Figure 20.

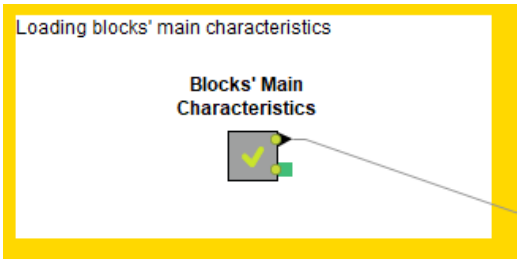

Figure 19 – KNIME Model – Loading Data – Blocks' Main Characteristics – General Workflow

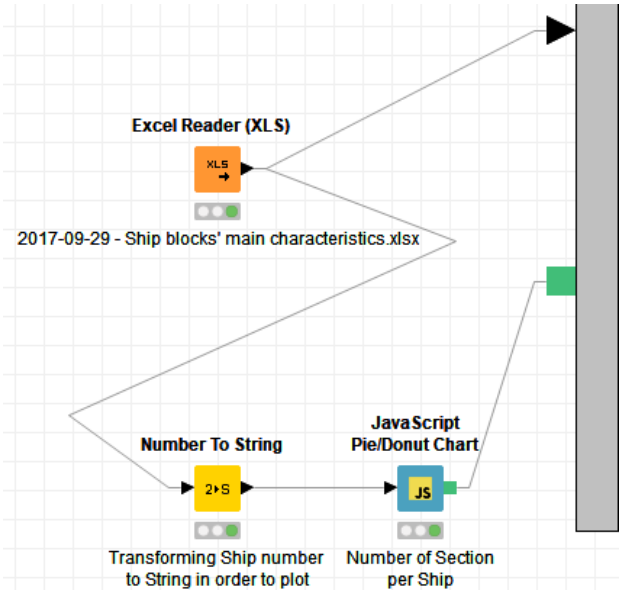

Figure 20 – KNIME Model – Loading Data – Blocks' Main Characteristics – Metanode

Additionally, there is a node converting the ship number into string so that a pie chart could be plot in order to demonstrate how many sections could be studied. Truly, the number of sections summed up is 452 and the distribution is demonstrated in Figure 21.

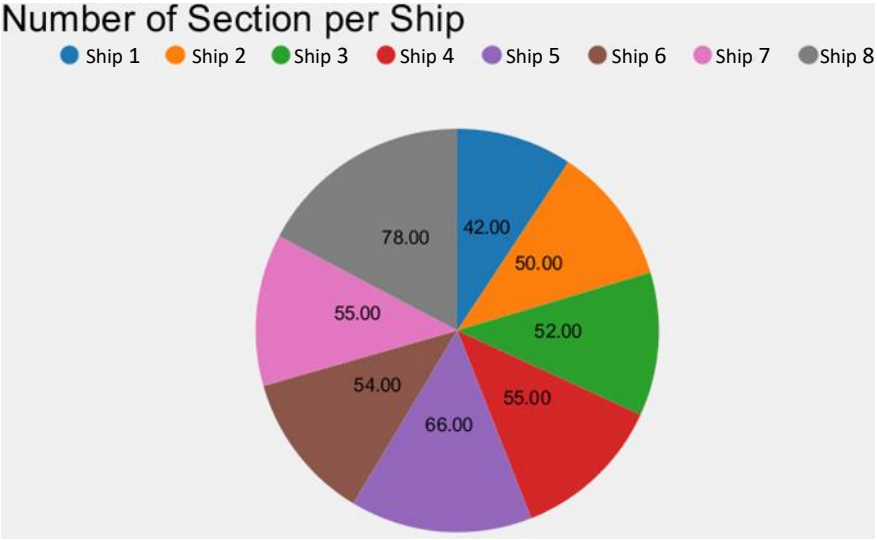

Figure 21 – KNIME Model – Loading Data – Blocks' Main Characteristics – Number of Sections per Ship

3.6.2.2. Measurements

The representation of the metanode in the workflow can be seen in Figure 22, Figure 23 and Figure 24. The Figure 22 represents the view in the general flow while Figure 23 and Figure 24 represent the inside setup of the metanode.

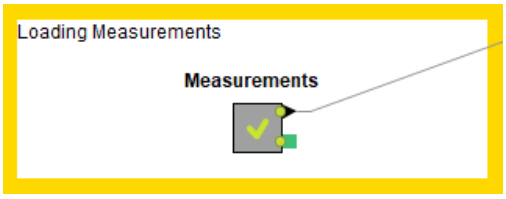

Figure 22 – KNIME Model – Loading Data – Measurements – General Workflow

Firstly, all measurements that where collected during the internship period are loaded into the workflow. Then, some numbers which KNIME recognized as string are converted back again into number (double) values in order to be used in the machine learning process.

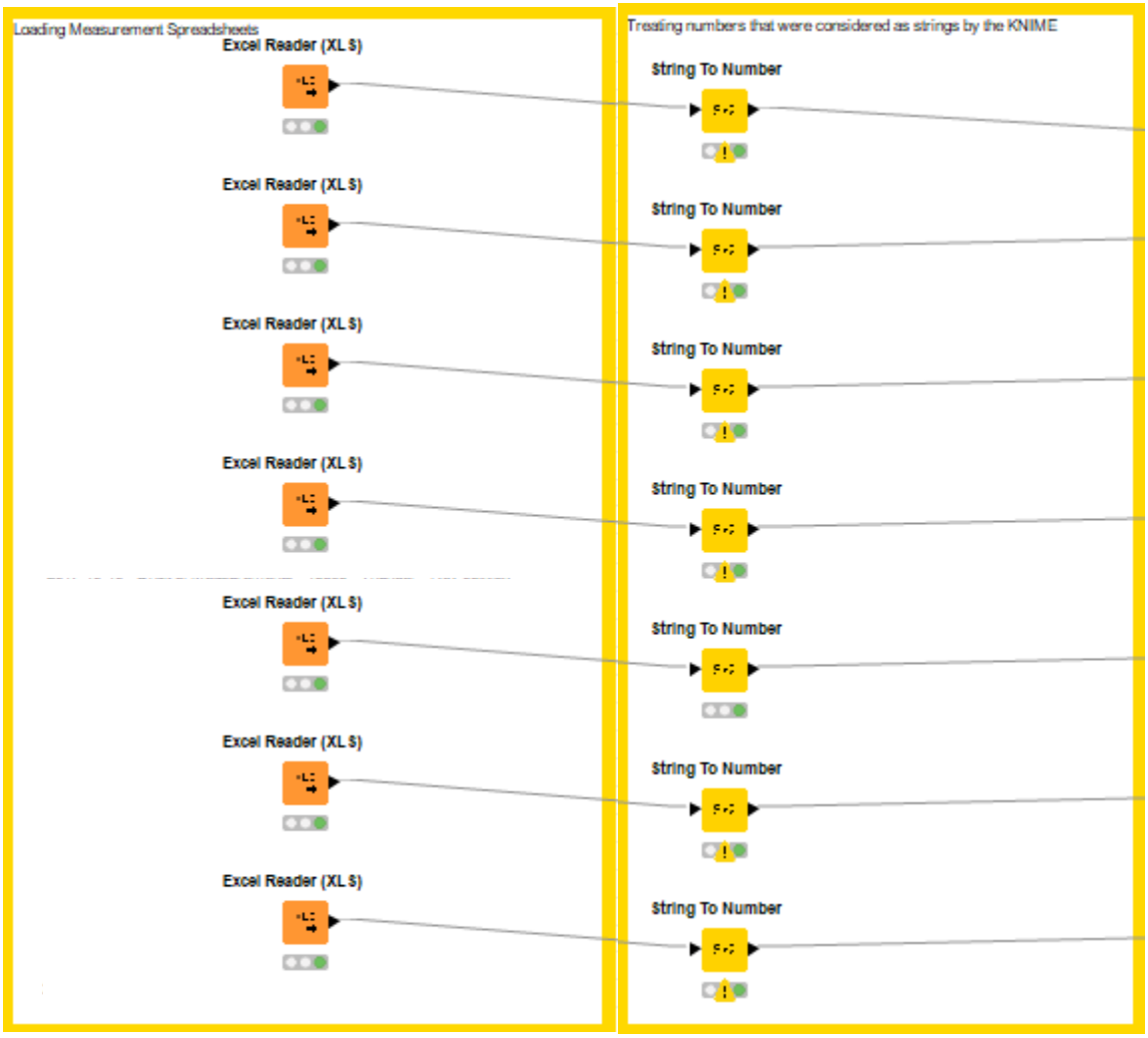

Figure 23 – KNIME Model – Loading Data – Measurements – Metanode – Loading and Converting Strings

Afterwards, there were some columns which were used to consolidate the data that were still being recognized by KNIME. Hence, they are excluded by the column filter node. In sequence, all tables are concatenated. Finally, along with the selection of good data and its forwarding to the general workflow, a plot with the data stratification is performed and it is shown in Figure 25.

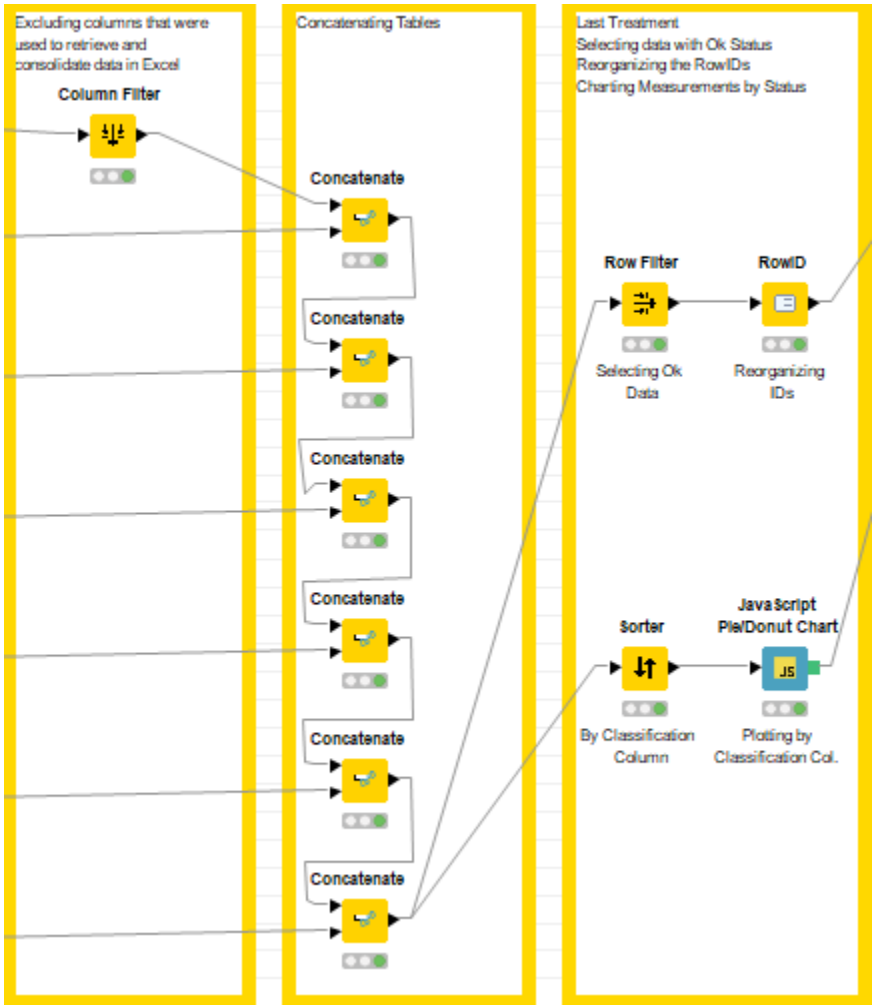

Figure 24 – KNIME Model – Loading Data – Measurements – Metanode – Eliminating Columns, Concatenating and Selecting Data

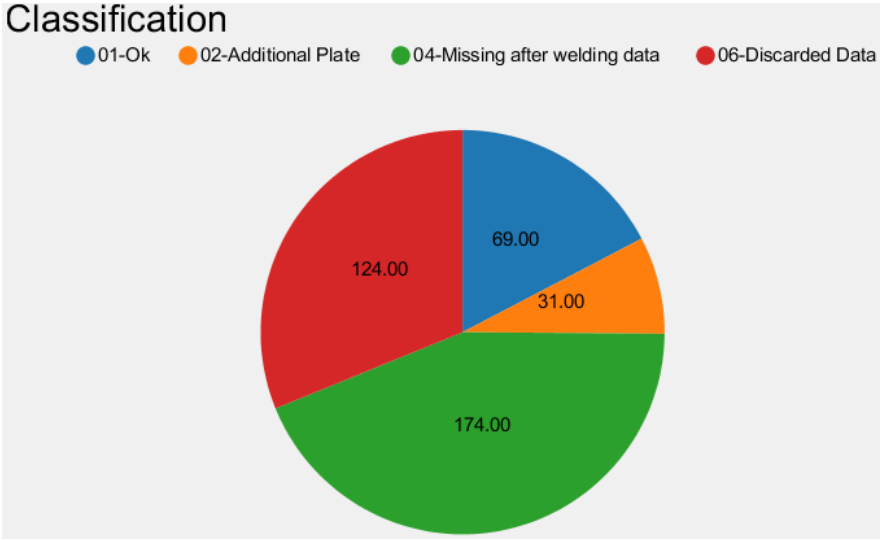

Figure 25 – KNIME Model – Loading Data – Measurements – Metanode – Number of Measurement by Classification

3.6.2.3. Elements' Characteristics

The elements characteristics is an exportation of all data available inside their current design system. Moreover, the features are: Ship, Section, Element, Layer, Geometry and Type of Material. Using these data, estimated welding length and weight were outlined based on each element existing inside a section. The layout of the elements' characteristics node is shown in Figure 26 and Figure 27.

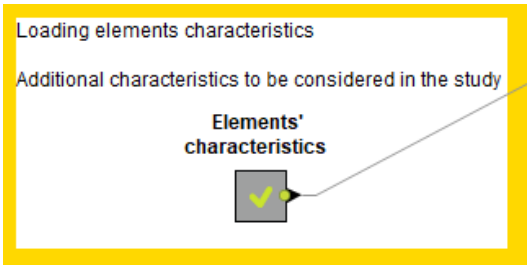

Figure 26 – KNIME Model – Loading Data – Elements' Characteristics – General Workflow

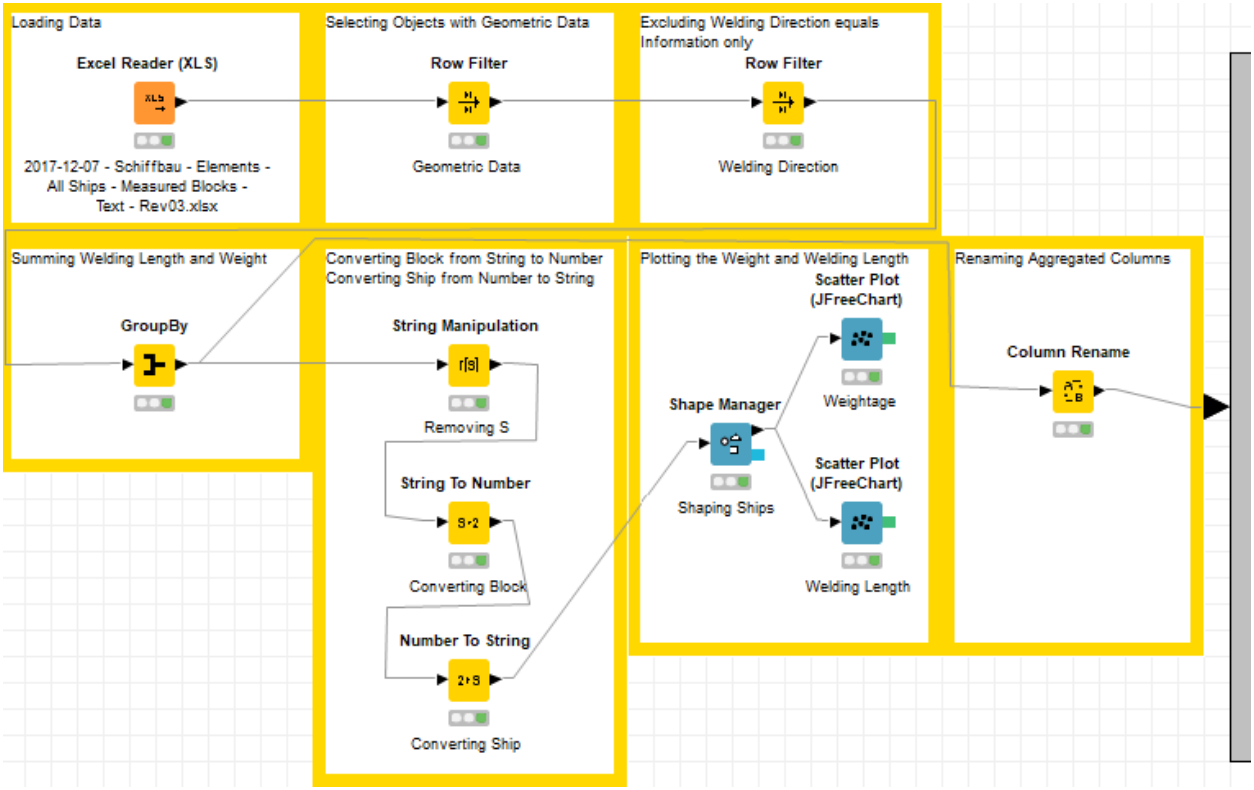

Figure 27 – KNIME Model – Loading Data – Elements' Characteristics – Metanode

As a matter of fact, this is the biggest database loaded into KNIME workflow, therefore the first action was to eliminate solely information rows from the database. This has been achieved by eliminating rows without any geometric data and no welding information. Then, the sum of the

estimated welding length and weight occurred followed by the renaming of the Sum columns in order to facilitate reading and coding at a later stage of the model.

Furthermore, the sums of the weightage and welding estimative were plotted according to the section and defining shapes for the dots representing each ship. The setup and plots are presented in Figure 28, Figure 29 and Figure 30.

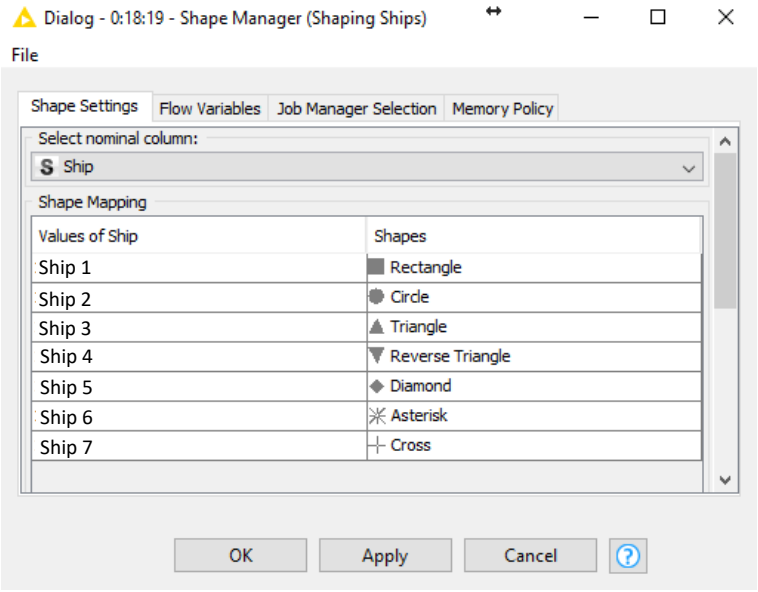

Figure 28 – KNIME Model – Loading Data – Elements' Characteristics – Metanode – Shape Manager

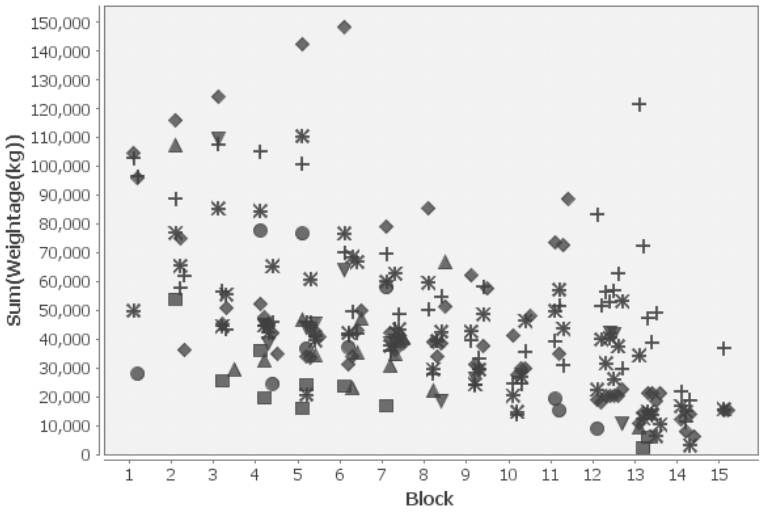

Figure 29 – KNIME Model – Loading Data – Elements' Characteristics – Metanode – Chart Plot – Sum of Weightage

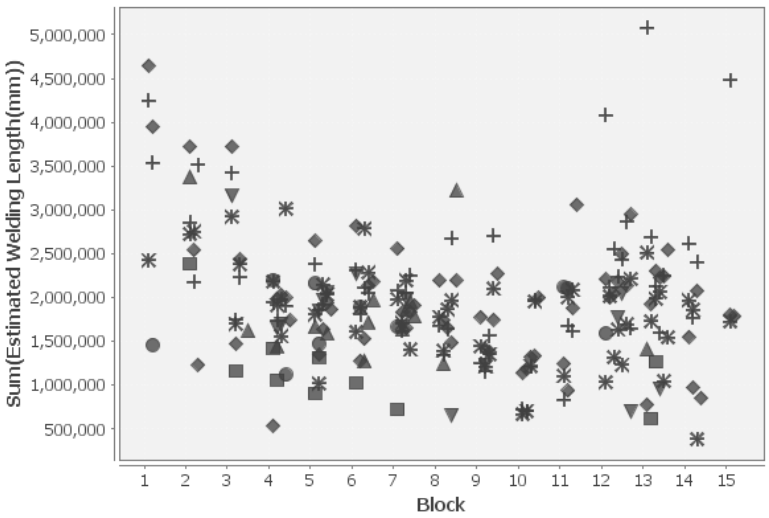

Figure 30 – KNIME Model – Loading Data – Elements' Characteristics – Metanode – Chart Plot – Sum of Weightage

3.6.2.4. Welding Seam

Based on least number of section which was provided by the measurements database (69), additional features based on the technical drawing of the main plate were stated. Features such as: Number of Transversal Elements, Number of Significant Welded Transversal Elements, Number of Longitudinal Elements, Number of Significant Welded Longitudinal Elements, Number of Girders, Number of Significant Welded Girders, Number of Butt Weld Seams, Number of Significant Weld Seams, Girder Spacing, Stiffener Spacing, Frame Spacing and Main Thickness of the deck plate. A representation of loading data in the general flow is displayed in Figure 31.

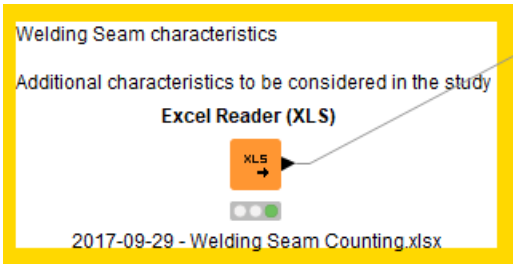

Figure 31 – KNIME Model – Loading Data – Welding Seam Counting – General Workflow

3.6.2.5. Joining

After loading all data, the joining process among 4 databases was performed and passed to the pre-processing process. An image is not displayed here as the representation of the joining process can be seen in Figure 17. The method used for joining the tables is Inner Join matching ship and section.

3.6.3. Pre-processing – Data

A general view of the pre-processing metanode and its components are demonstrated in Figure 32 and Figure 33.

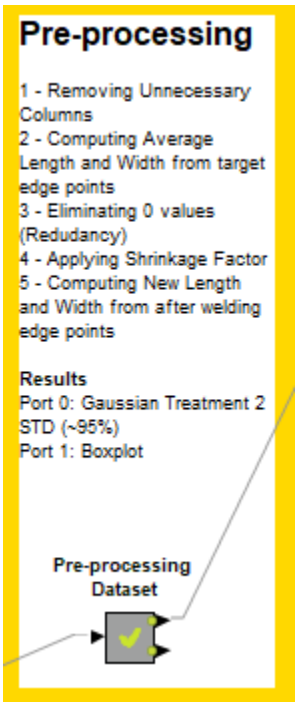

Figure 32 – KNIME Model – Pre-processing – Data – General Workflow

In this metanode, the initial removal of columns occurs at the early beginning to reduce the number of features to be studied which are considered as irrelevant, such as: information about "before welding" points which is an intermediate part of the process just before the butt welding and there are already some welded part into the sub-assemblies; Var 1 and Var 2 which are variations computed between target point and before welding, and, before welding and after welding due to the reason mentioned for the "before welding"; and finally string variables which are not accepted by the neural network learning method.

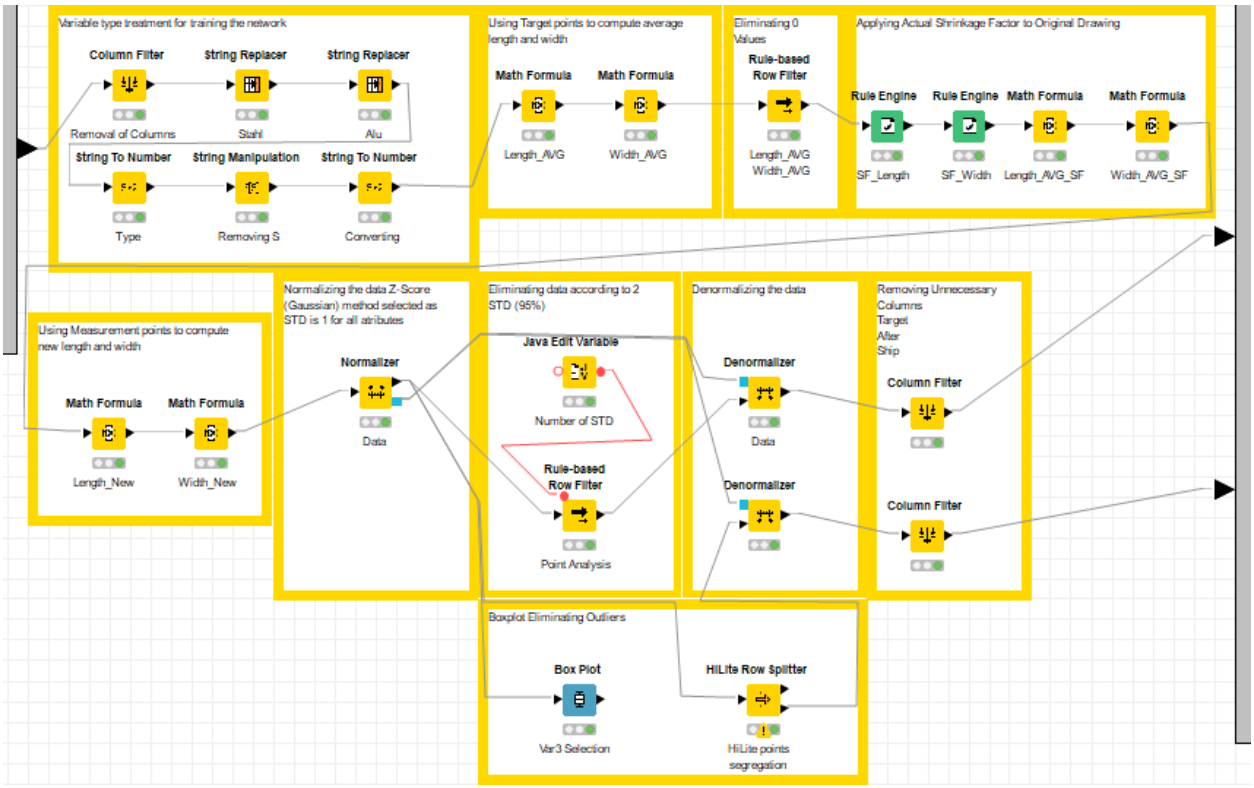

Figure 33 – KNIME Model – Pre-processing – Data – Metanode

As the training does not allow strings, the material type of the section is converted to integer number and the section variables have a character "S" removed and then converted to number. Figure 34 displays the nodes described so far.

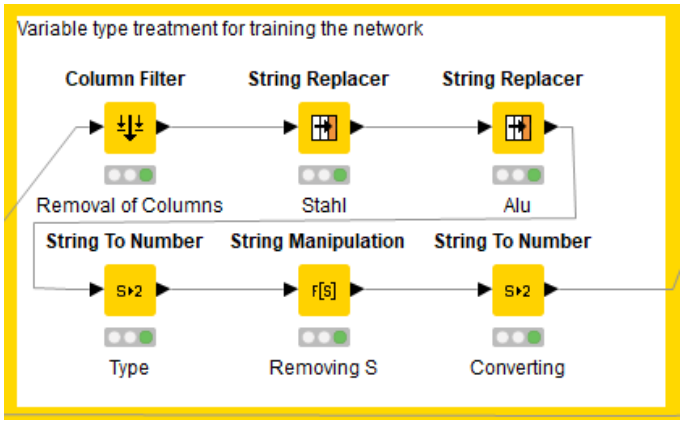

Figure 34 – KNIME Model – Pre-processing – Data – Treatment for training the network

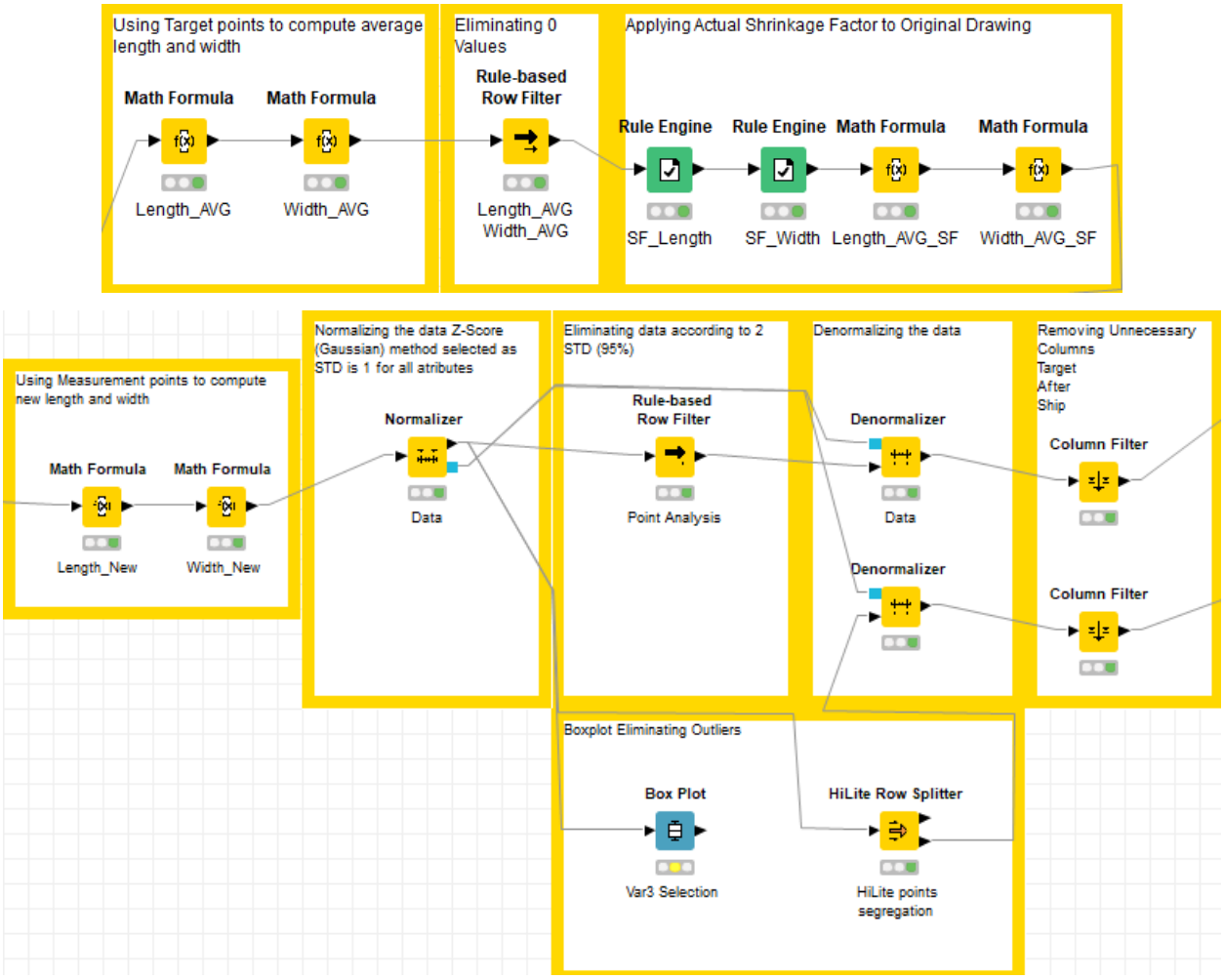

Figure 35 – KNIME Model – Pre-processing – Data – Rest of Operations

In order to reduce the number of learning variables and create an easier formula to be manipulated by the CAM/Nesting team the target points where reduced to average length and width. Next, there is a node in order to exclude null values. This node was created while developing the model when not all data was fully keyed in order to assure consistency of data. Nonetheless, now it does not impact anything in the model, but it is kept as the model can be used to assess different problems and it can assist in the development of new solutions.

Afterwards, the shrinkage factor is included in the database and applied to average length and width. On the same pace as average length and width, the reduction of the variables for the final position of the points has been created based on the total length and width and summing the variations.

Finally, two different statistical methods where applied: two standard deviations and the interquartile range. Whereas the two-standard deviation method eliminate the outliers reducing the

available data from 60 to 45 blocks, the interquartile range reduces from 60 to 48. Hence, the data which does not behave as most of the data is taken out of the database by using the variations as the criteria. The formulas applied to compute additional characteristics can be seen in Appendix II.

3.6.4. Pre-processing - Graphing and Clustering

The representations of the metanode are displayed in Figure 36 and Figure 37. While Figure 36 is the metanode in the general workflow, Figure 37 demonstrates the inner operations of this metanode.

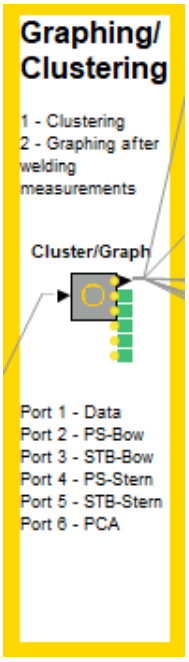

Figure 36 – KNIME Model – Pre-processing – Graphing and Clustering – General Workflow

Firstly, the node applies the principal component analysis (PCA) which is a statistical procedure to convert a set of observations of likely correlated variables into a set of values of linearly uncorrelated variables. Along with the PCA, k-Means clustering has been applied in order to identify any possible cluster. Subsequently, the discovered clusters have been shaped and the different material types highlighted.

Ultimately, as PCA and cluster information may vary according to each database to be studied, they have been removed before passing the data to the machine learning methods. In addition, scatter plots for all four corners and the PCA were plotted.

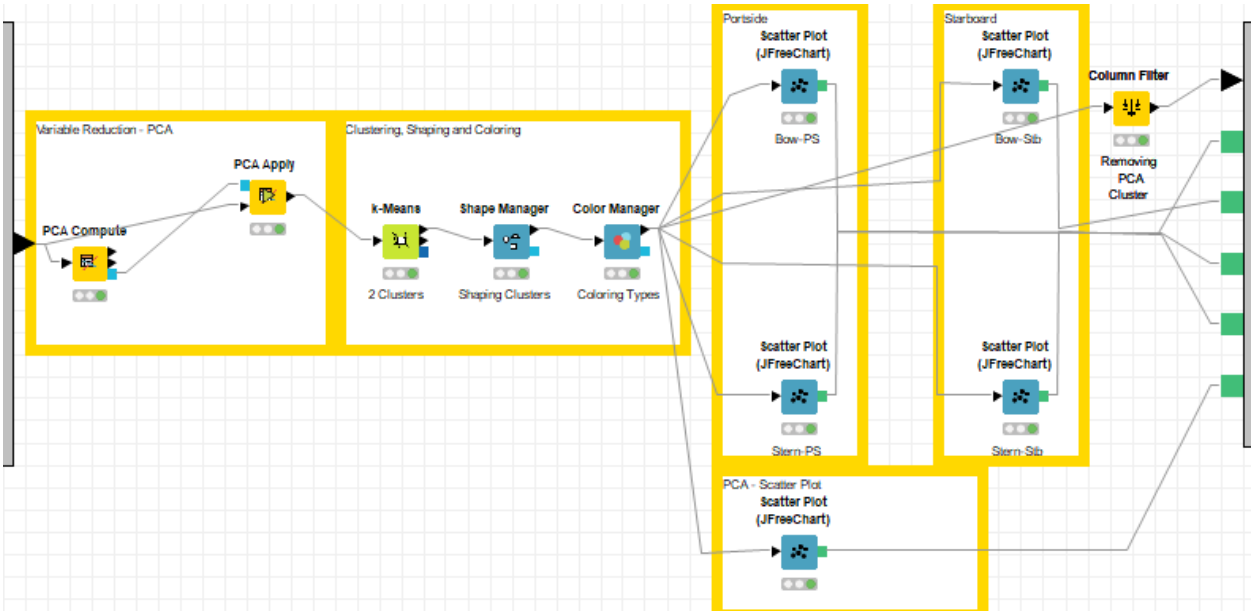

Figure 37 – KNIME Model – Pre-processing – Graphing and Clustering – Metanode

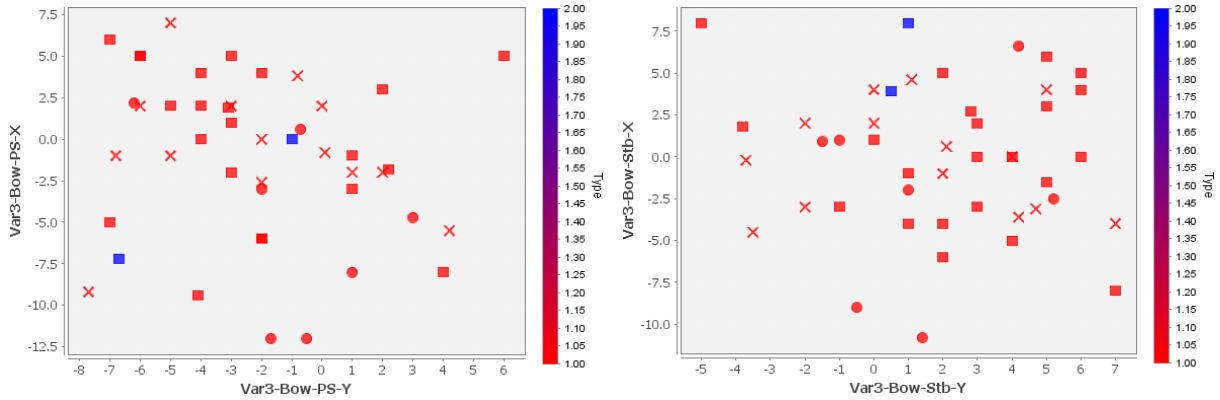

Figure 38 – KNIME Model – Pre-processing – Graphing and Clustering – Metanode – Bow Portside and Starboard Edges (Cluster 1 – Cross; Cluster 2 – Circle; Cluster 3 – Rectangle; Type 1 – Steel; Type 2 – Aluminum)

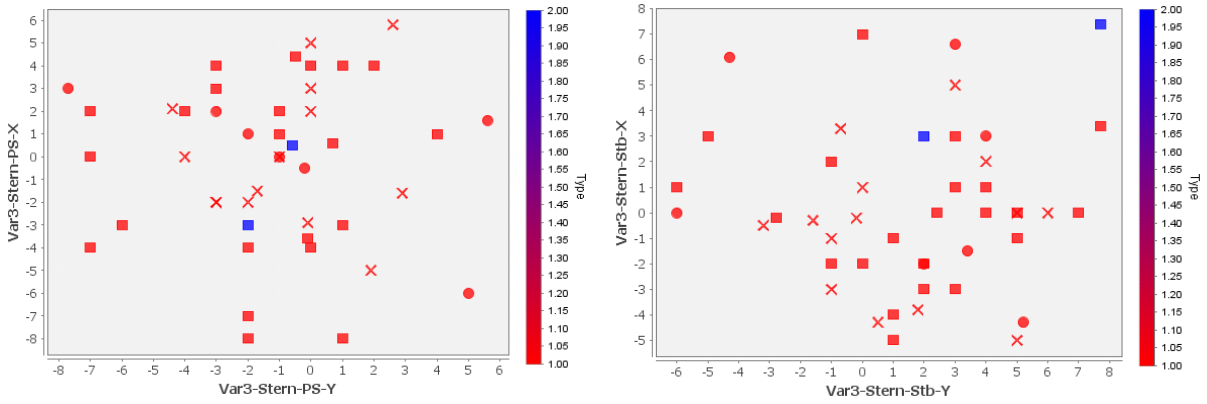

Figure 39 – KNIME Model – Pre-processing – Graphing and Clustering – Metanode – Stern Portside and Starboard Edge (Cluster 1 – Cross; Cluster 2 – Circle; Cluster 3 – Rectangle; Type 1 – Steel; Type 2 – Aluminum)

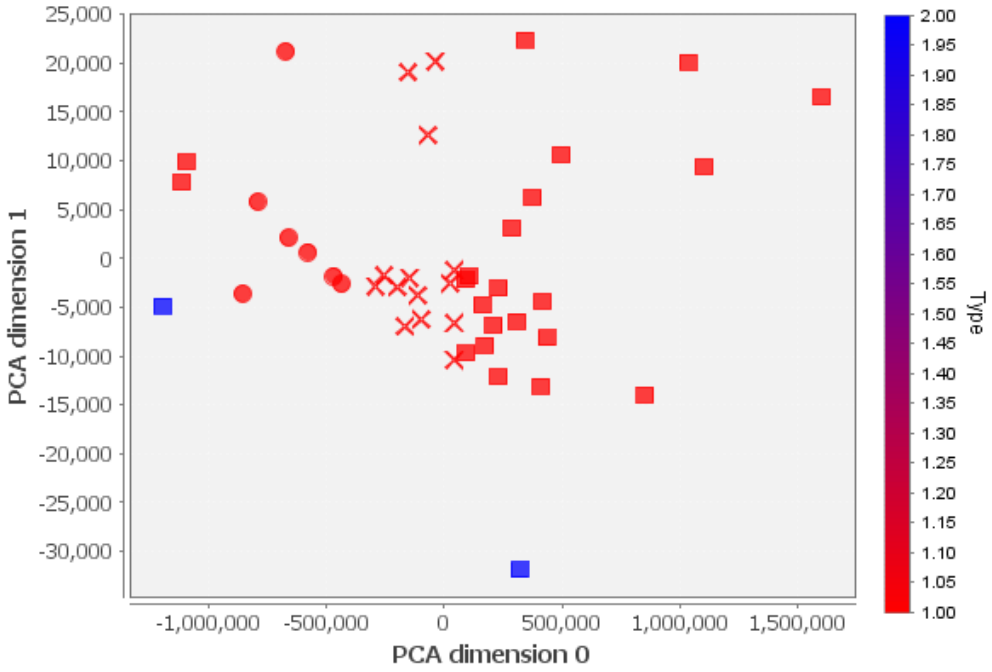

Figure 40 – KNIME Model – Pre-processing – Graphing and Clustering – Metanode – PCA (Cluster 1 – Cross; Cluster 2 – Circle; Cluster 3 – Rectangle; Type 1 – Steel; Type 2 – Aluminum)

3.6.5. Neural Network

A first graphical representation of the Neural Network metanode in the general flow can be seen in Figure 41. Further details of the metanode are presented from Figure 42 to Figure 50.

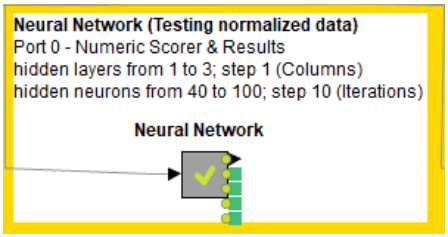

Figure 41 – KNIME Model – Processing – Neural Network – General Workflow

Truly, the neural network algorithm is composed by a number of hidden layers and hidden neurons. In general terms, while the hidden neurons are distributed in each layer providing the weightage values, the hidden layers constitute how many assessments are going to take place. In order to better assess our problem, the workflow has been organized so that many different setups could be verified. Clearly, the default value of hidden layers and neurons are 1 and 10. The interval loop start nodes are setups where the range and step of testing cases are assigned.

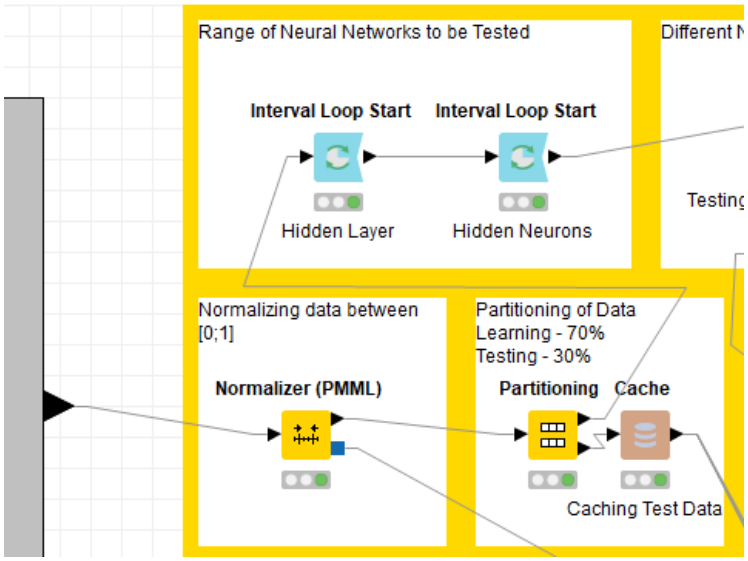

Figure 42 – KNIME Model – Processing – Neural Network – Setup

Before initializing the network, it is necessary to normalize the data in the model so that the learning algorithm takes place. For doing so, it is necessary to have only numbers as string variables are treated as classification problem. Later, it is elementary to divide the learning and setup, which in our case 70% of data was reserved for learning whereas 30% was retained for testing the model. As well, the data partitioning is draw randomly. After, a cache node was used to make faster access of data. From Figure 42 to Figure 45, the setup screenshots are displayed.

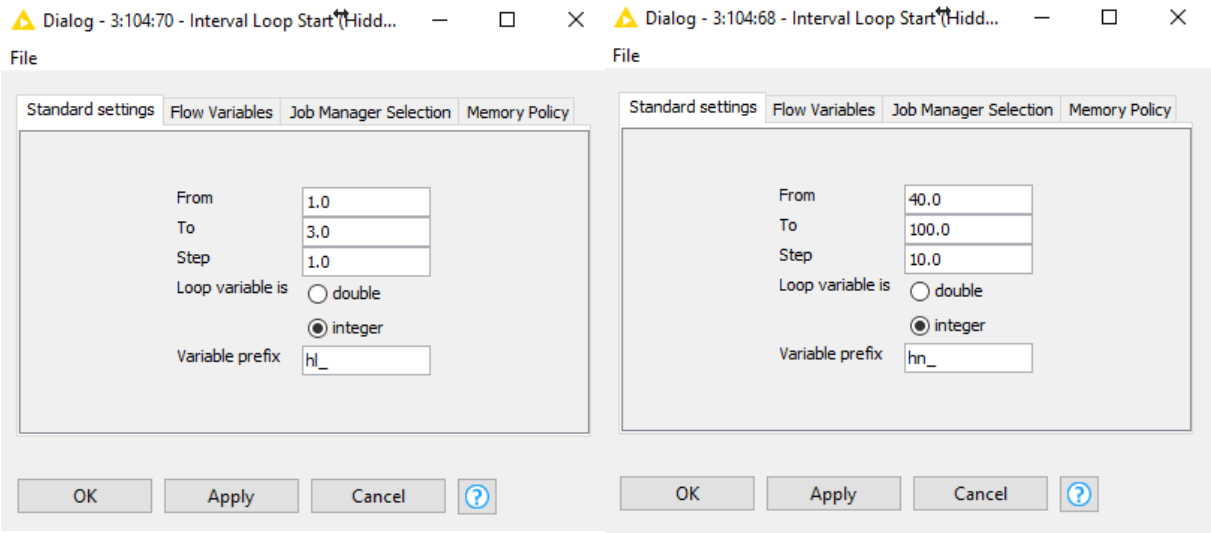

Figure 43 – KNIME Model – Processing – Neural Network – Setup – Hidden Layer and Hidden Neuron

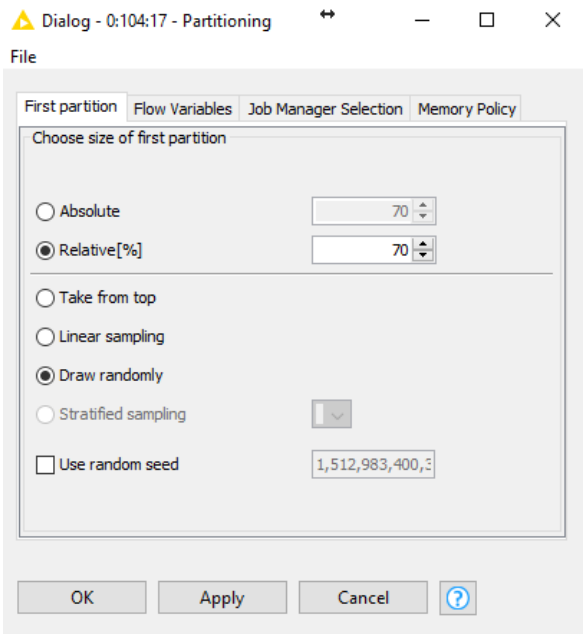

Figure 44 – KNIME Model – Processing – Neural Network – Setup – Data Partitioning

In sequence, there is another loop where we select the variables that the model should learn and test. Hence, for each variable the flow is going to select one variable, use it to learn, convert the model to cell and save it in a column.

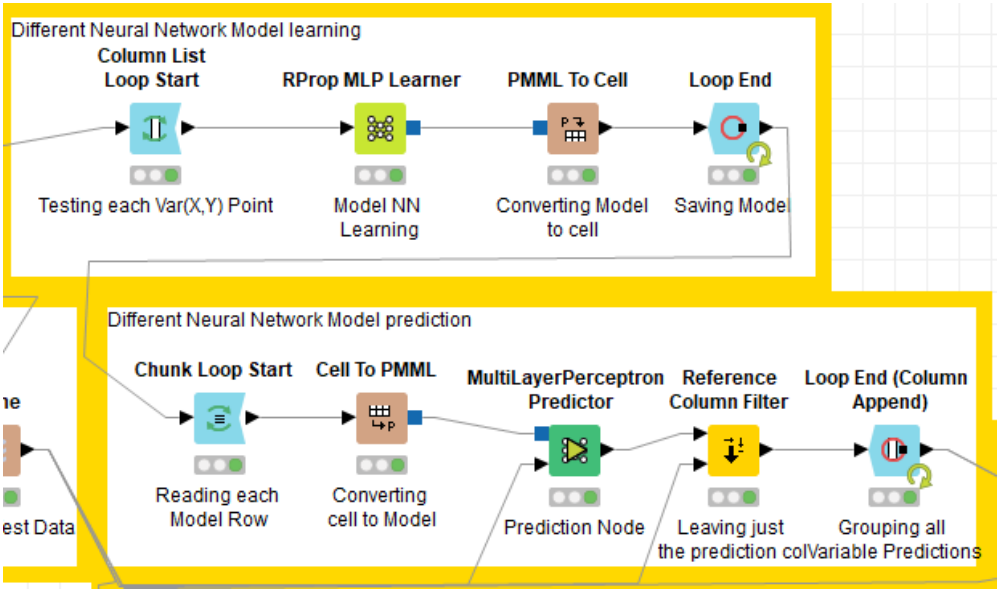

Figure 45 – KNIME Model – Processing – Neural Network – Learning and Prediction

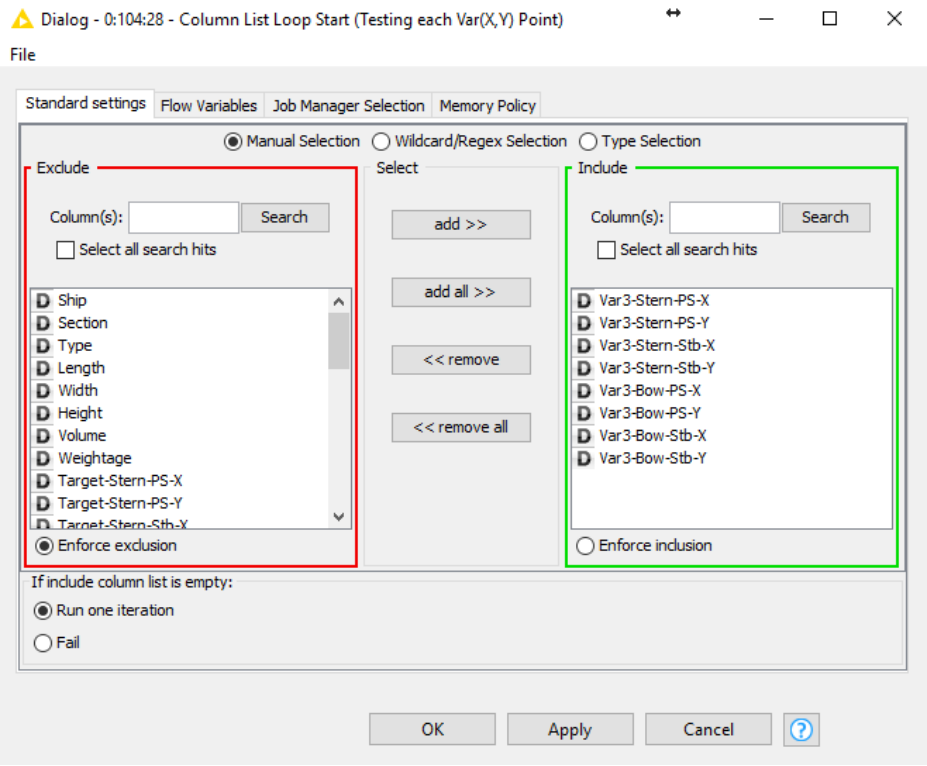

Figure 46 – KNIME Model – Processing – Neural Network – Variables to be Learned

At a later stage the model is going to get the column with all the models, create the prediction for each variable, remove all columns but prediction one and save all predictions into a table. The flow below actually denormalize the data in order to save the results and enable the plotting later on.

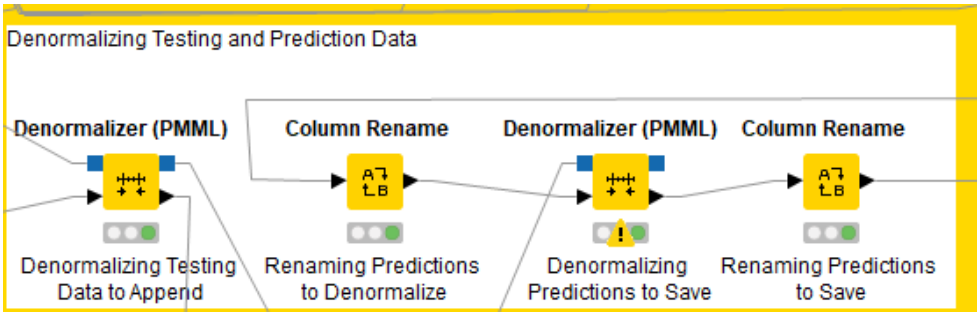

Figure 47 – KNIME Model – Processing – Neural Network – Denormalization to Save

Then a metanode with numeric scorer, which is presented in Figure 50, is used in order to test all variables with normalized data. Next, the results of the numeric scorer are joined with the denormalized data in order to save the statistical and denormalized predictions and the table is transposed before passing the results.

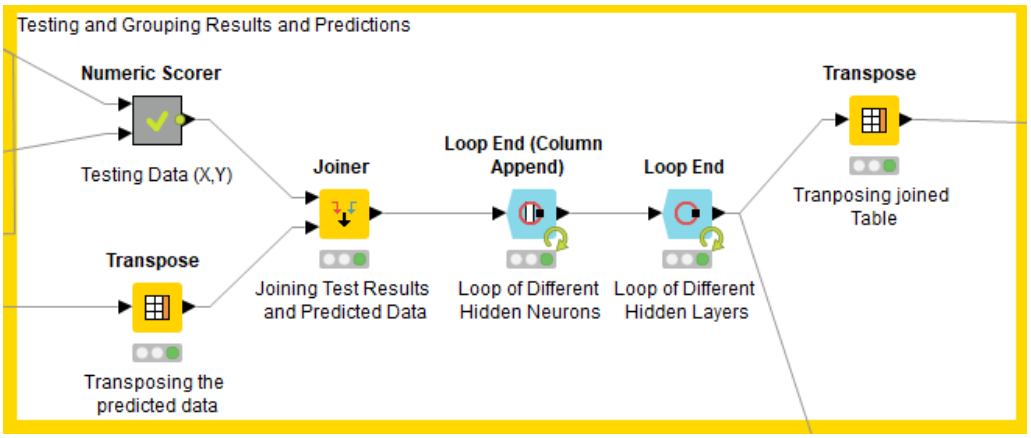

Figure 48 – KNIME Model – Processing – Neural Network – Joining Predictions and Numeric Test Results

Based on the results, a rule-based node selects an assigned iteration to plot the results obtained by the network, and also a joiner is used in order to provide the original value of deformations with the predictions.

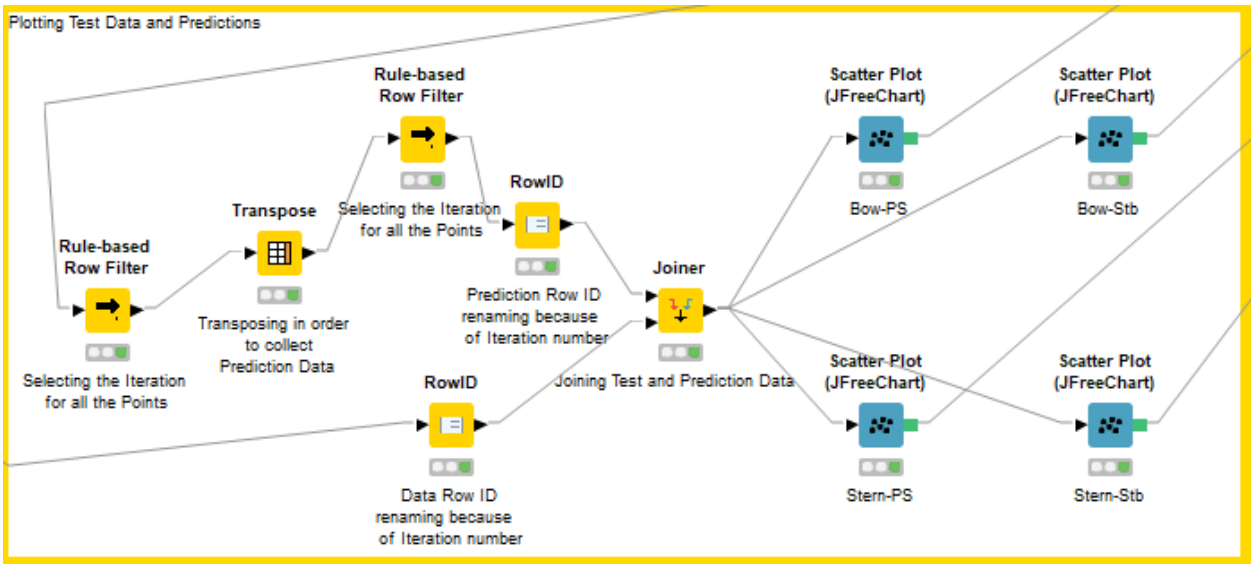

Figure 49 – KNIME Model – Processing – Neural Network – Joining Predictions and Numeric Test Results

The numeric scorer metanode displayed below was developed so all elected variables are tested and passed back to Neural Network metanode. Another representation of the assigned variables to be tested can be seen in Figure 51. It can be observed that they are the same as the ones assigned in the learning process.

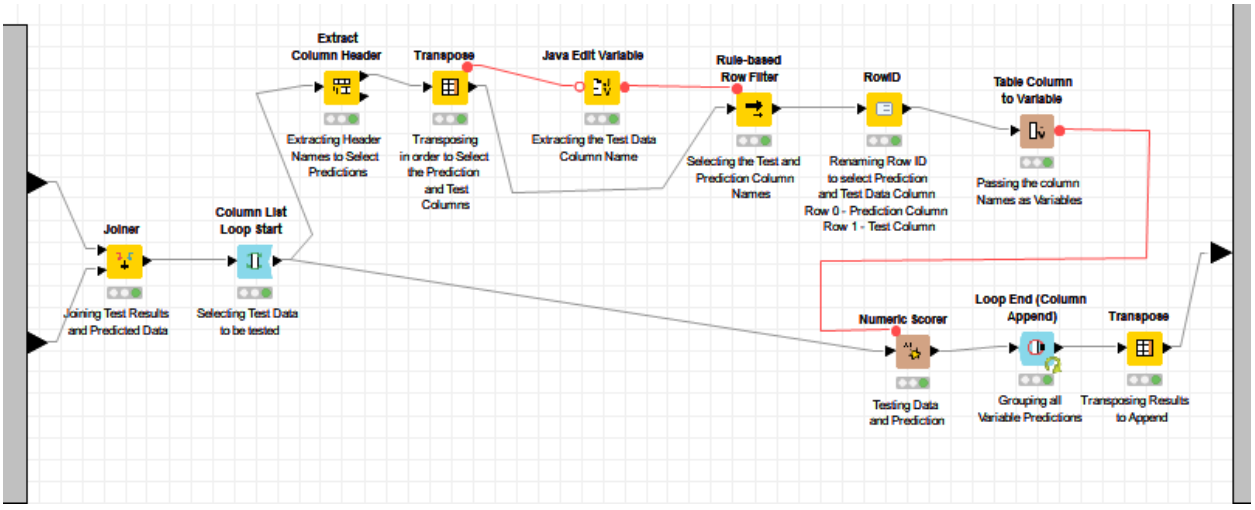

Figure 50 – KNIME Model – Processing – Neural Network – Numeric Test Metanode

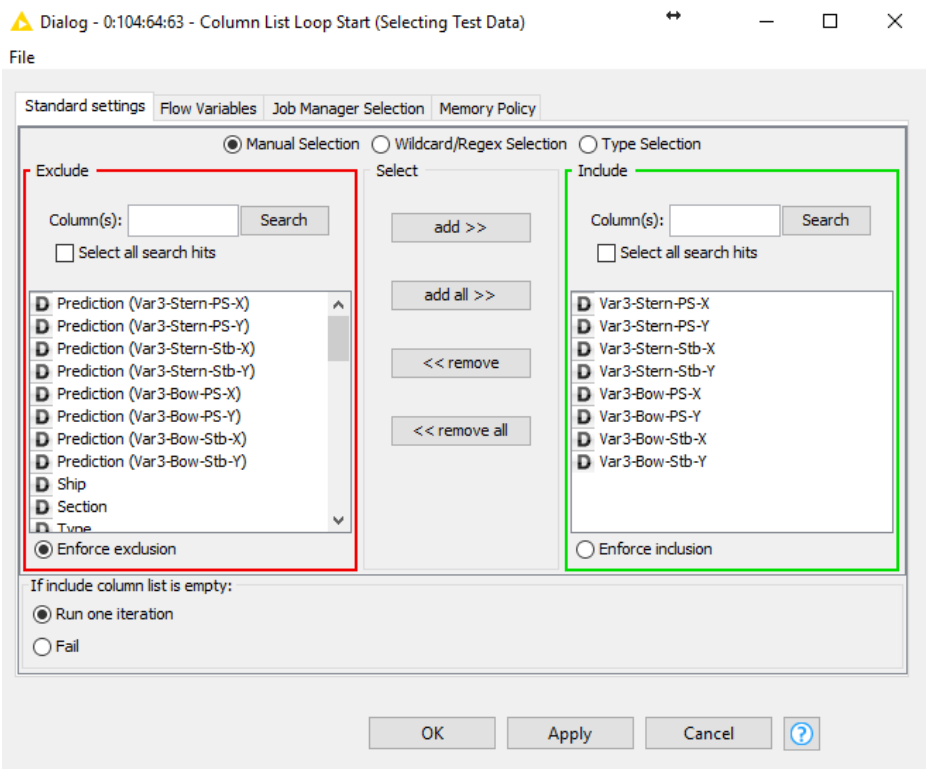

Figure 51 – KNIME Model – Processing – Neural Network – Numeric Test Metanode – Variables to be tested

3.6.6. Polynomial Regression

A first graphical image of the Polynomial Regression metanode in the general flow can be noticed in Figure 52. Moreover, the details of the metanode are given from Figure 53 to Figure 63.

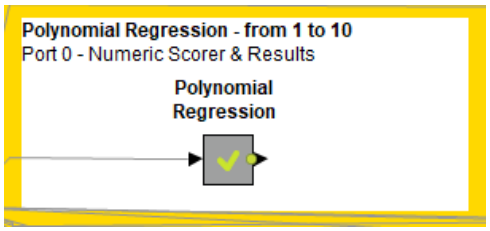

Figure 52 – KNIME Model – Processing – Polynomial Regression – General Workflow

Actually, the polynomial algorithm is composed by the number of degrees of the equation. Again, in order to better assess our problem, the workflow has been organized so that many different setups could be verified. Undoubtedly, the default degree value offered by Knime is two. The interval loop start node is where the range and step of testing cases are assigned.

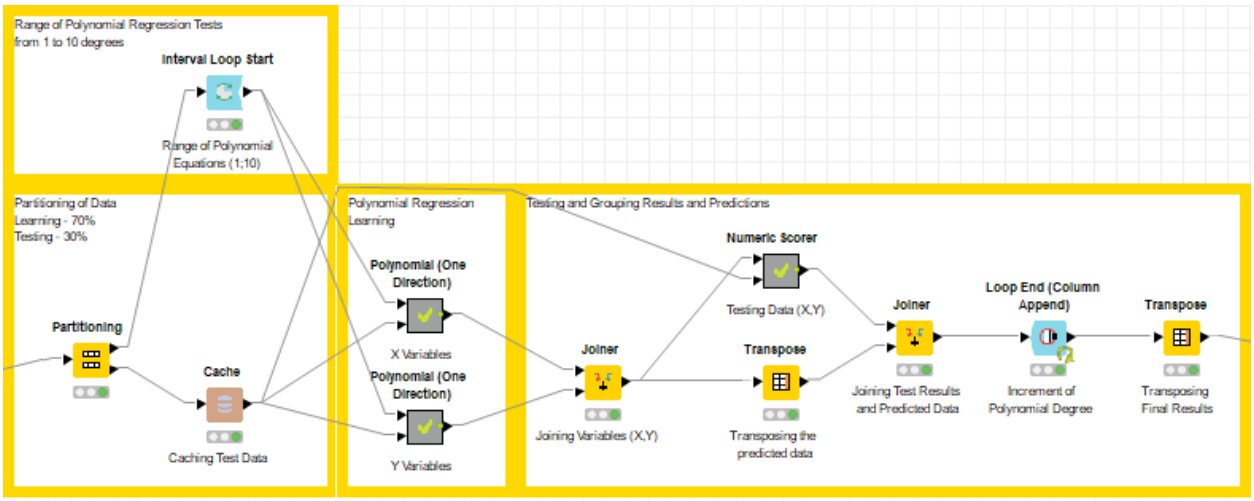

Figure 53 – KNIME Model – Processing – Polynomial Regression – Metanode

Unlikely Neural Network, before initializing the network, it is not mandatory to normalize the data in the model so that the learning algorithm takes place. On the other hand, it is still necessary to have only numbers. Next, it is elementary to divide the learning and setup which in our case 70% of data was reserved for learning whereas 30% was retained for testing the model. As well, the data partitioning is draw randomly. Then, a cache node was used to make faster access of data. In Figure 54 and Figure 55, the setup screenshots are displayed.

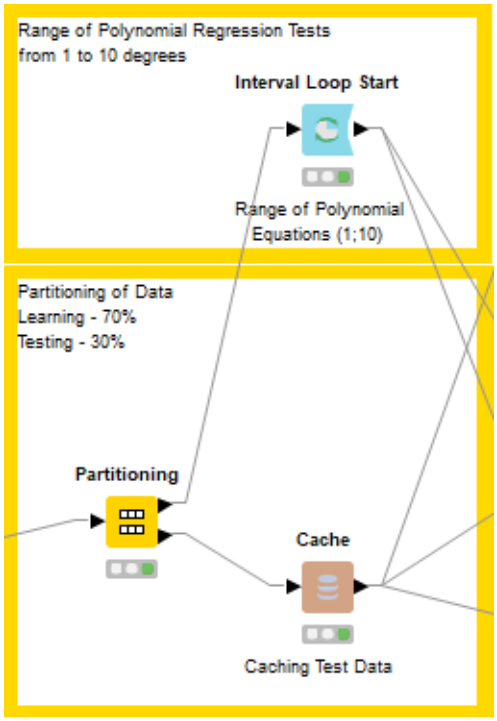

Figure 54 – KNIME Model – Processing – Polynomial Regression – Metanode - Setup

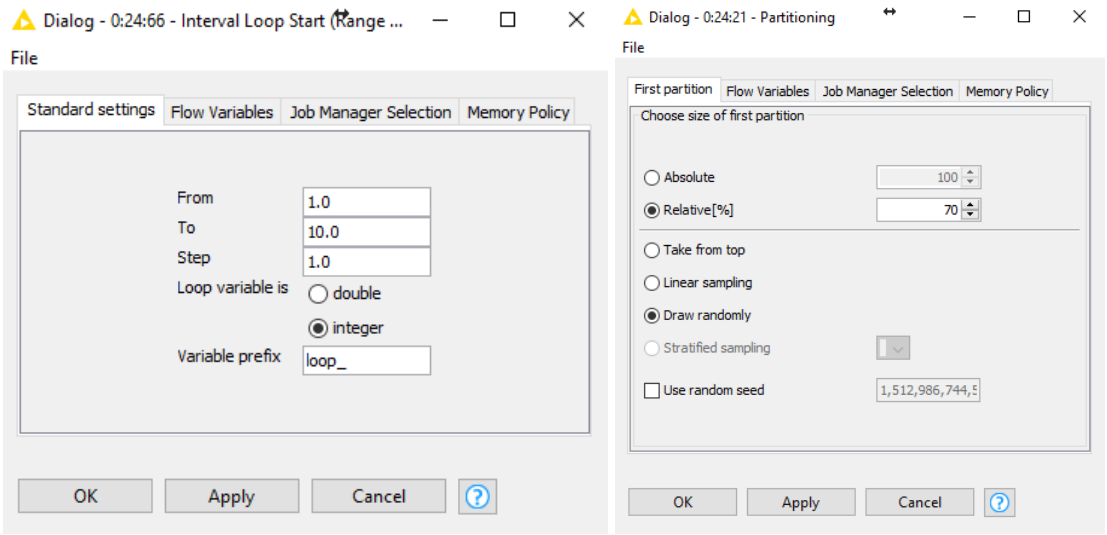

Figure 55 – KNIME Model – Processing – Polynomial Regression – Setup – Number of Degrees and Data Partitioning

In sequence, there are two other metanodes where the variables that the model should learn, and test are passed to. Once again, for each variable the flow is going to select one variable which have been divided into two directions “X” and “Y”, use it to learn, convert the model to cell and save it in a column. This way, at the end it would be possible to come out with two different formulas.

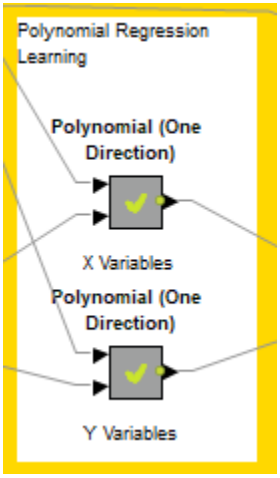

Figure 56 – KNIME Model – Processing – Polynomial Regression – Metanode – Learning and Predicting

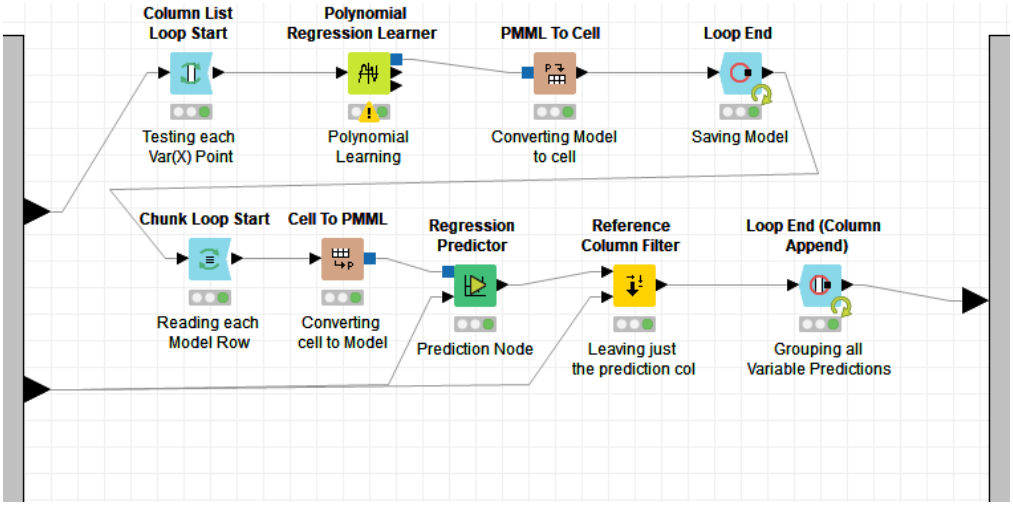

Figure 57 – KNIME Model – Processing – Polynomial Regression – Learning and Predicting – Metanode – X Direction

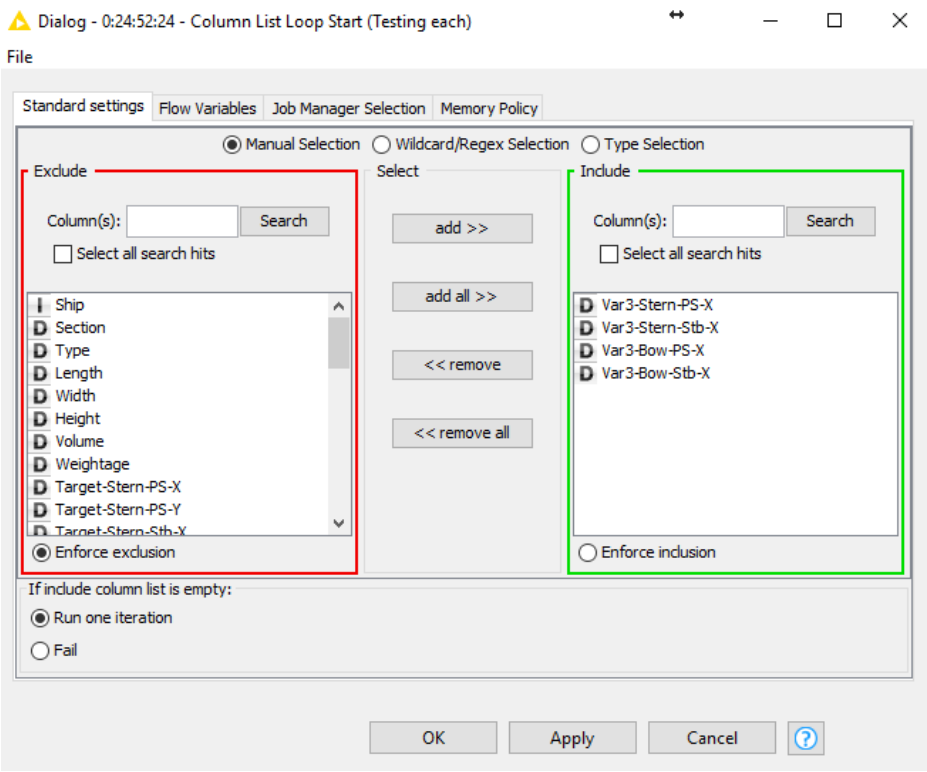

Figure 58 – KNIME Model – Processing – Polynomial Regression – Variables to be Learned – X Direction

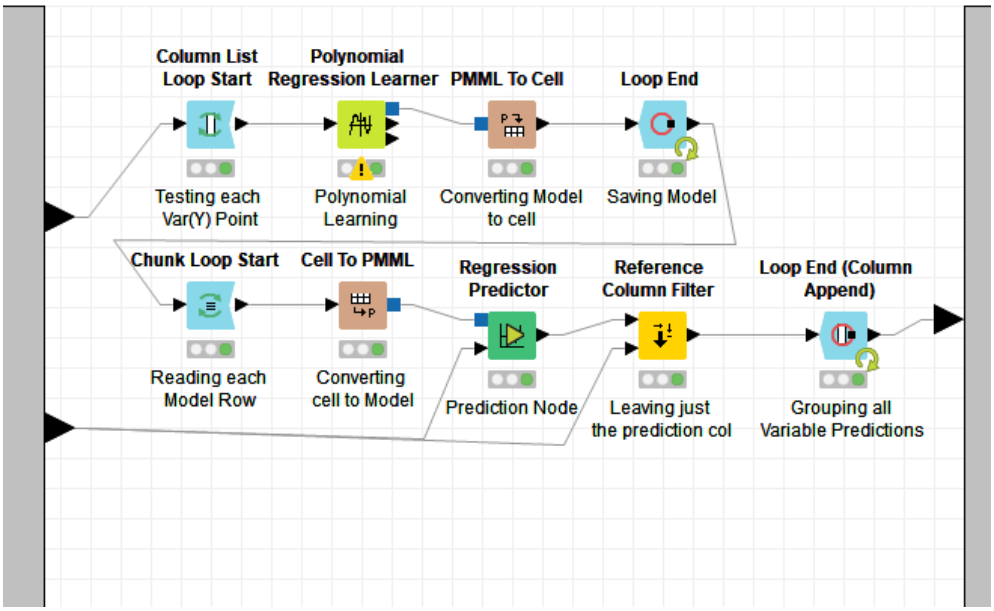

Figure 59 – KNIME Model – Processing – Polynomial Regression –Learning and Predicting – Metanode – Y Direction

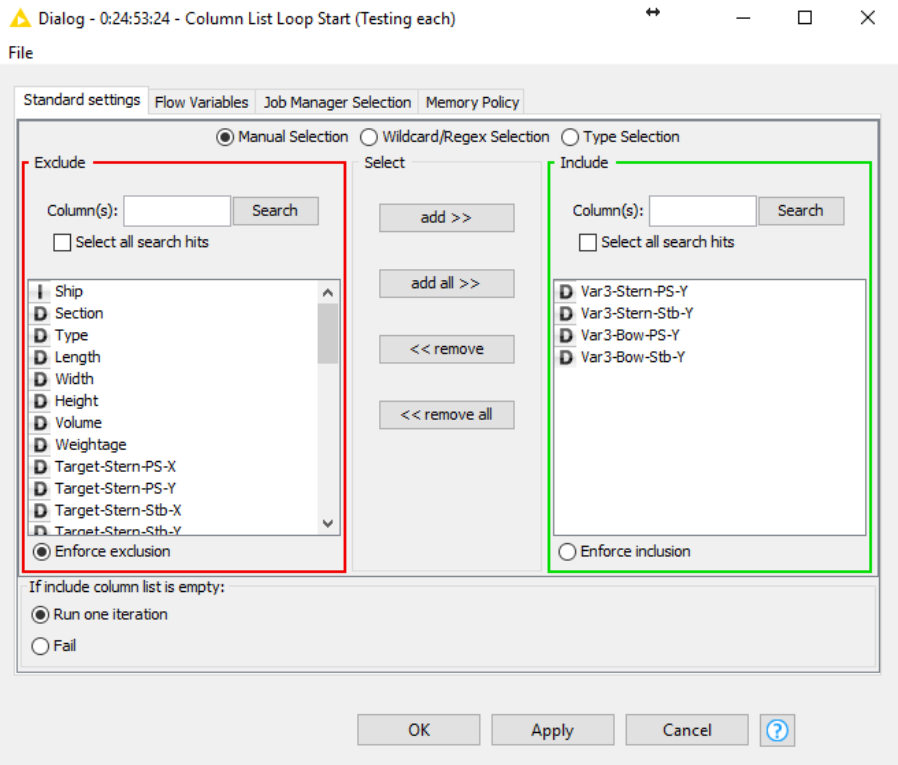

Figure 60 – KNIME Model – Processing – Polynomial Regression – Variables to be Learned – Y Direction

On the same pace as the neural network model, the actual values, the predictions and test results are joined in order to forward the values outside the metanode to the main workflow.

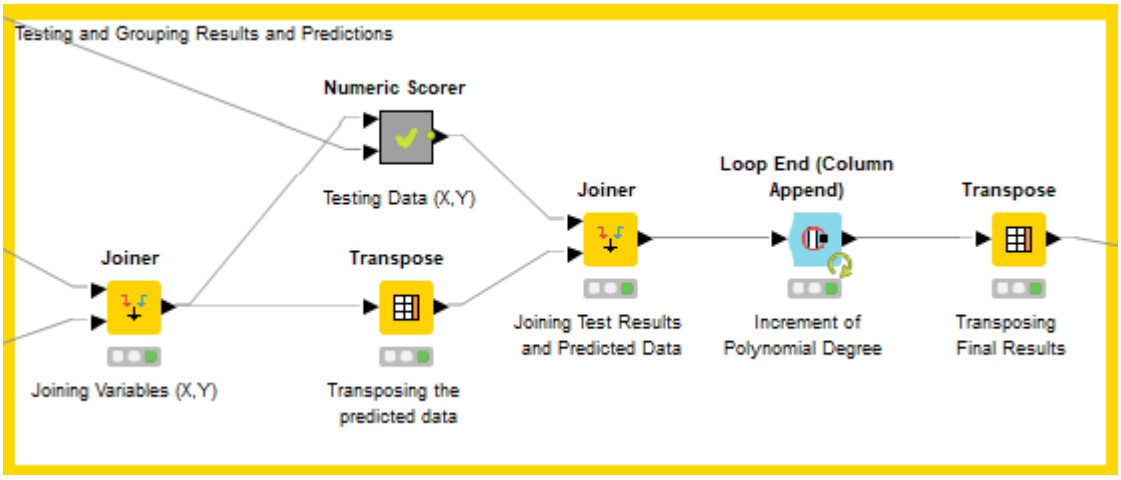

Figure 61 – KNIME Model – Processing – Polynomial Regression – Joining Predictions and Numeric Test Results

The numeric scorer metanode displayed below was developed so all elected variables are tested and passed back to Polynomial Regression metanode. Another representation of the assigned variables to be tested can be seen in Figure 63. It can be observed that they are the same as the ones assigned in the learning process.

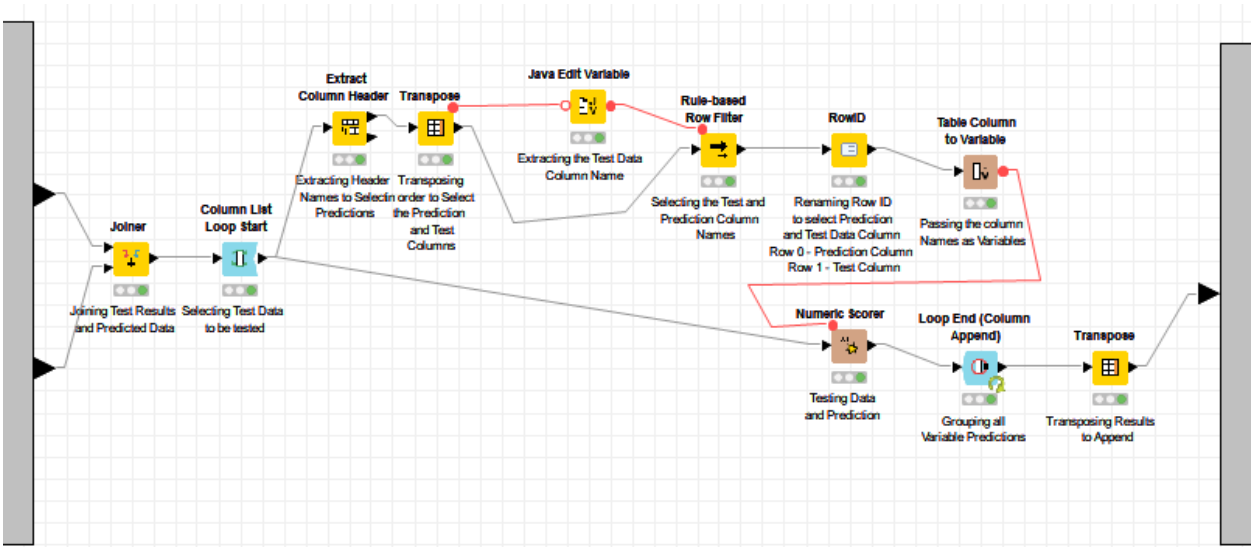

Figure 62 – KNIME Model – Processing – Polynomial Regression – Numeric Test Metanode

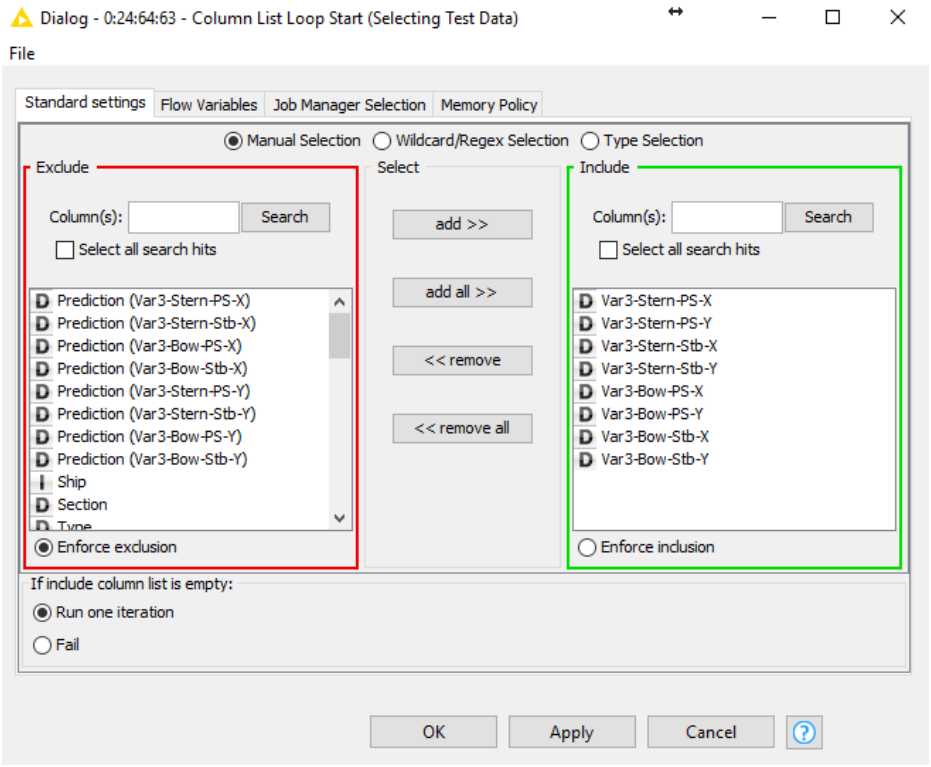

Figure 63 – KNIME Model – Processing – Polynomial – Numeric Test Metanode – Variables to be tested

3.6.7. Back Feature Selection

As a lot of variables has been used in order to study the problem, a back-feature selection has been implemented in order to verify the most significant variables to the problem. A display of the metanode in the general flow is shown in Figure 64.

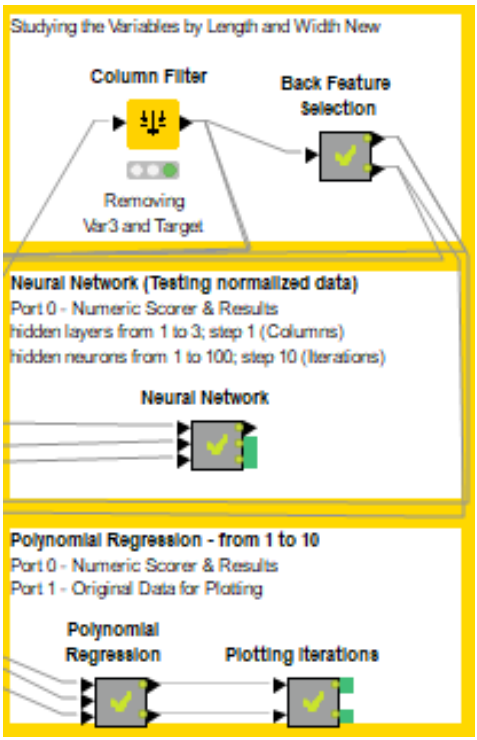

Figure 64 – KNIME Model – Processing – Back Feature Selection – General Workflow

A more detailed picture is presented in Figure 65. In true, the implementation of this method requires the usage of a prepared metanode model such as neural network but with one setup which in this case the number of hidden layers was 1 and the number of the hidden neurons 41.

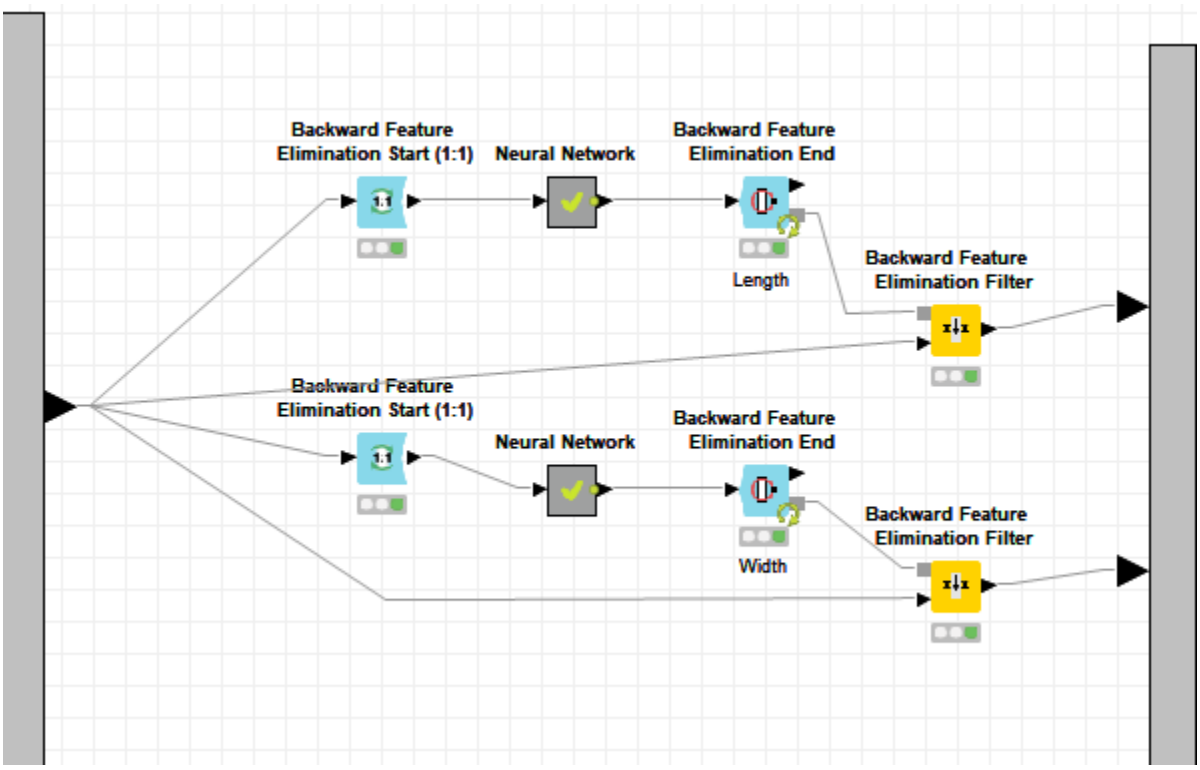

Figure 65 – KNIME Model – Processing – Back Feature Selection – Metanode

3.6.8. Best-fitting

The best fitting is actually the plot between the design values with the shrinkage factor on Y-axis with the actual shrink values on X-axis. A demonstration of the nodes in the general workflow is presented in Figure 66.

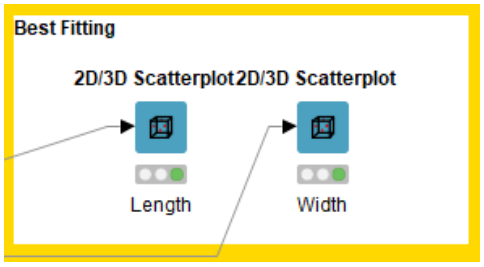

Figure 66 – KNIME Model – Processing – Best Fitting – General Workflow

4. RESULTS AND ANALYSIS

The results are analyzed by considering three main characteristics: the coefficient of determination (R^2), mean absolute error (MAE) and mean squared error (MSE). A brief explanation is given below.

In addition, the coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable(s) and its interpretation is that the values closest to 1 indicates that the fitted model explains all variability. It is important to highlight that R^2 can yield negative values when fitting non-linear functions to data.

Moreover, the mean absolute error is a measure of difference between two continuous variables and the lowest value should be preferred.

Furthermore, the mean squared error measures the difference between the estimator and what is estimated, and the lowest value should be promoted.

All in all, the first criterion used is the coefficient of determination, followed by the MAE and then MSE. Additionally, all plots generated by KNIME do not present the unit. Nonetheless, the unit should be considered as millimeters (mm).

4.1. Real Variation

Two representations of the variations are presented in Figure 67 and Figure 68.

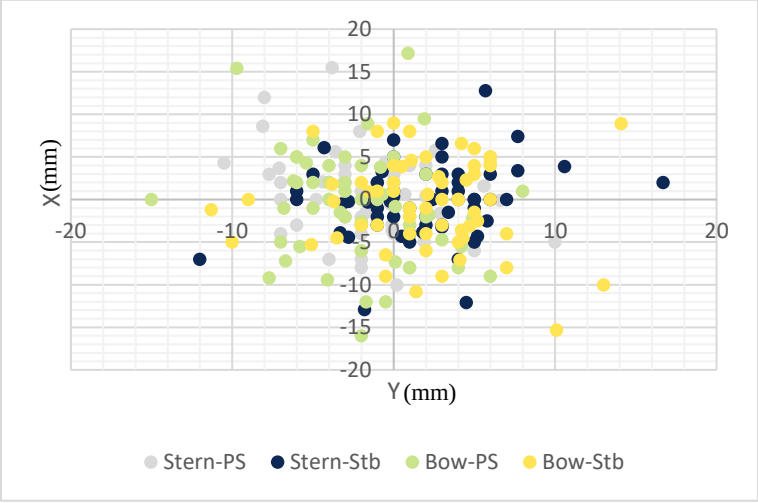

Figure 67 – Results – Real Variations – Target vs Actual points – All four corners together

From Figure 67 it is possible to visually verify that most of the variations occur inside the range of 10 mm. From the raw data, in the x-axis the maximum absolute variation is 17.2 mm, while in the y-axis 22 mm which is not displayed above. Nonetheless, the shrinkage factor for the Steel is 1.001 and for the Aluminum 1.002, and the average block length and width are respectively 11.39 m and 18.65 m indicating that the actual factor can handle most of the actual variations, but production problems might still arise.

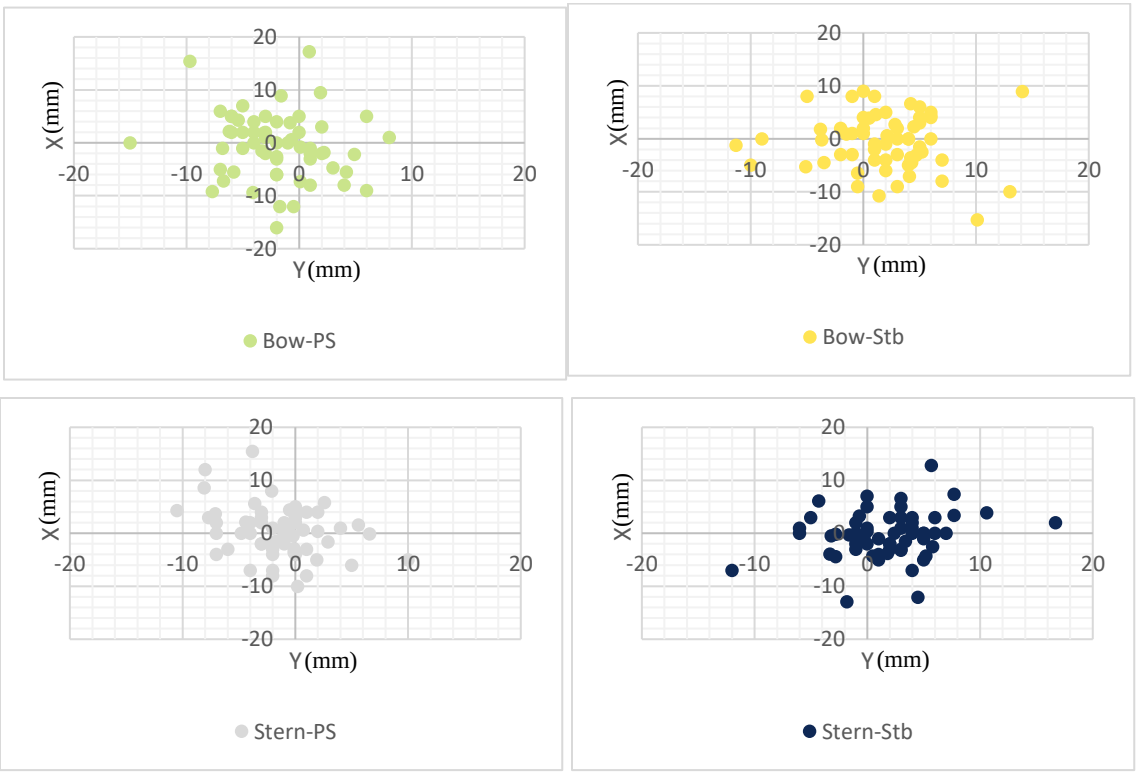

Figure 68 – Results – Real Variations – Target vs Actual points – All four corners separated

Figure 68 gives a better view of each corner. It must be reminded that the lengthwise has an excess of material which should be enough to avoid any production problem. On the other hand, on the widthwise direction production problems might occur as there would not be any excess of material. Ultimately, the comparison was considering the behavior of a single corner. If you take into account the behavior of two corners the effects on the arise of production problems can be even greater specially on Y-axis where there is not an excess of material on top of the shrinkage factor.

4.2. Polynomial Regression

4.2.1. All Features with eight variables

Although the method has been implemented, the attempt to run this trial failed because the number of features is way too high when compared to the number of data rows. Hence, more data is required in order to run this model.

4.2.2. All Features with two variables

On the same pace as the previous trial, with eight variables, the number of features is higher than the number of required data to run the model. Therefore, more rows or the reduction of the features are elementary so that a trial could be achieved.

4.2.3. One Feature with two variables and PCA Analysis

After running the model for four different methods – two-standard deviation, interquartile range using variation points when compared to the target and actual points to eliminate outliers and two-standard deviation, interquartile range using PCA to eliminate outliers varying from 1 to 10 degrees, the following charts from Figure 69 to Figure 72 could be outlined in which solely the best results were plotted until the first visual difference.

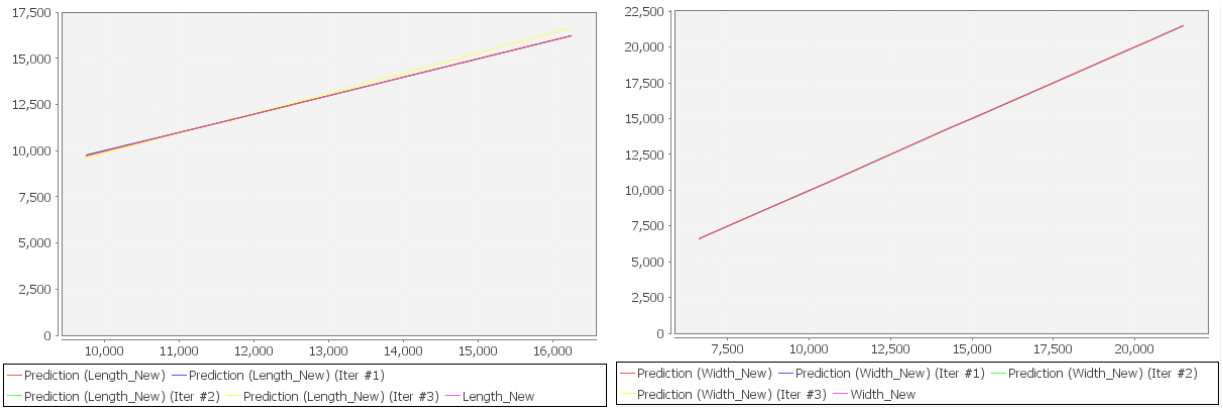

Figure 69 – Results – Polynomial Regression – One Feature with two variables – from 1 to 4 degrees – 2 STD

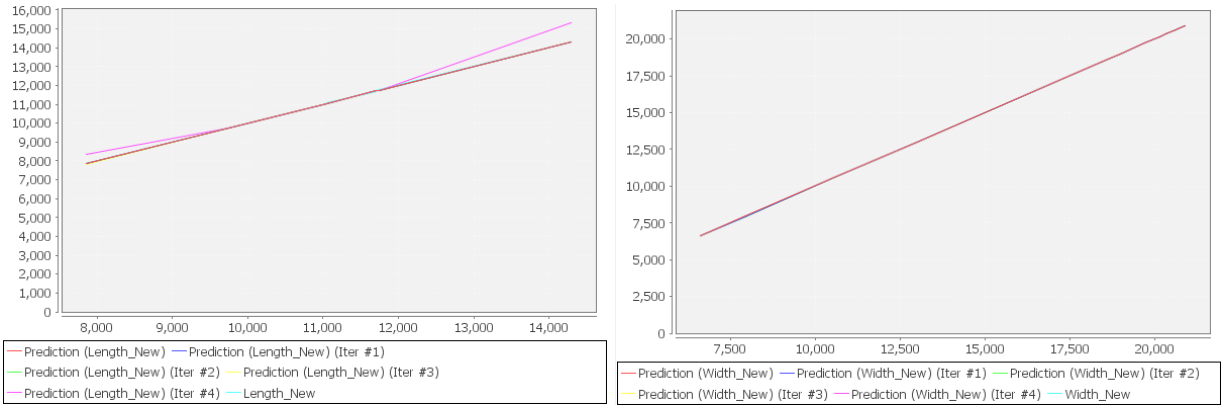

Figure 70 – Results – Polynomial Regression – One Feature with two variables – from 1 to 5 degrees – IQR

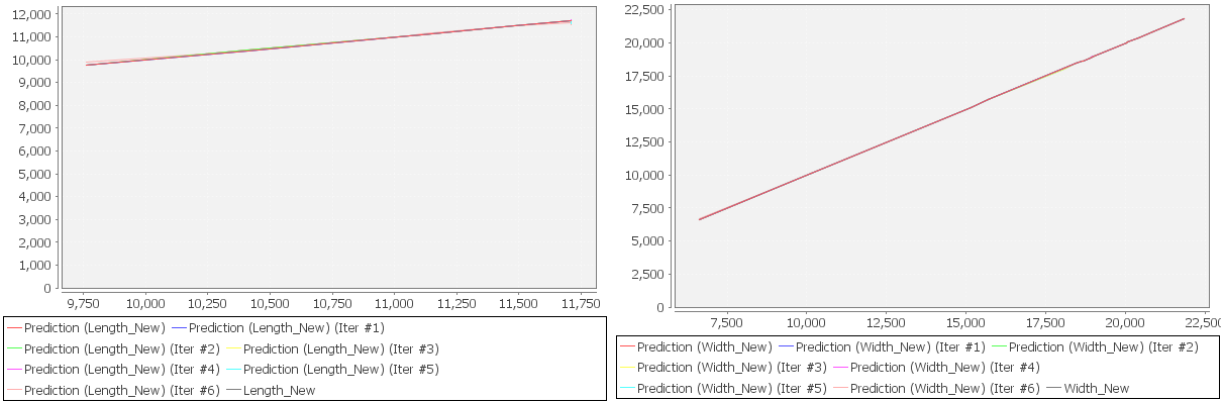

Figure 71 – Results – Polynomial Regression – One Feature with two variables – from 1 to 7 degrees – 2 STD – PCA

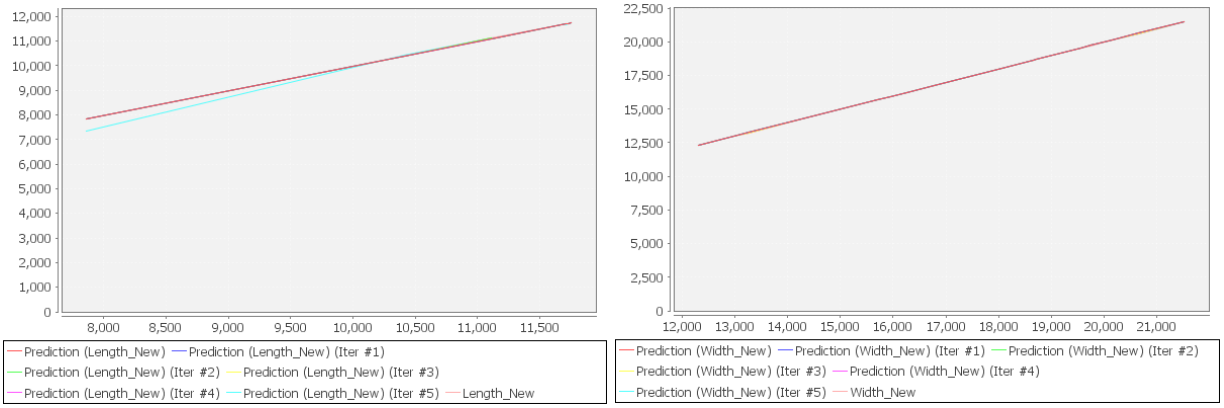

Figure 72 – Results – Polynomial Regression – One Feature with two variables – from 1 to 6 degrees – IQR - PCA

The use of IQR method increased the number of available data when compared to the two-standard deviation method which can be the reason why we could satisfactory achieve another number of degree. Analogously, when using PCA in order to eliminate the outliers more data was available to be evaluated and higher degrees could be achieved with the same method.

Clearly, the length was the variable that started to deviate first. This can be explained by being influenced mostly by the given features. Additionally, because its range is lower than the width, it has more variations inside a lower range. Ultimately, for the polynomial regression method, the best model was the two-standard deviations observing the variations in order to remove the outliers and its final statistical information is shown below.

Table 5 – Results – Polynomial Regression – One Feature with two variables – Best Result – 2 STD - 1 degree

Row ID	Prediction (Length_New)	Prediction (Width_New)	Number of Degrees	Average
R^2	1.00	1.00	1	1.00
MAE	3.23	3.94	1	3.58
MSE	16.84	25.00	1	20.92

4.3. Neural Network

4.3.1. All Features with eight variables

While the polynomial regression method could not handle the number of variables, the neural network could learn from the features and the results are demonstrated in Table 6. However, due to the number of non-linearities all coefficients of determination were negatives. Therefore, no setup for this model was not considered satisfactory and not being plotted or further discussed.

Table 6 – Results – Neural Network – All Features with eight variables – R^2

		Hidden Layers		
		1	**2**	**3**
Hidden Neurons	**40**	- 1.02	- 0.97	- 1.44
	50	- 1.23	- 1.15	- 1.59
	60	- 1.20	- 1.12	- 1.49
	70	- 0.95	- 1.21	- 0.75
	80	- 1.22	- 1.14	- 1.39
	90	- 1.14	- 1.30	- 0.44
	100	- 1.07	- 1.04	- 1.90

4.3.2. All Features with two variables

As it had been mentioned for the previous attempt, while the polynomial regression method could not handle the number of variables, the neural network could learn from the features and present results. Nonetheless, the coefficients of determination were improved by the reduction of variables to be studied. Thus, starting with the coefficient of determination and analyzing the errors, a best configuration was outlined for each outlier removal method.

Table 7 – Results – Neural Network – All Features with two variables – R^2 – 2 STD

Hidden Neurons \ Hidden Layers	1	2	3
40	0.682	0.727	0.575
50	0.731	0.764	0.582
60	0.900	0.593	0.699
70	0.811	0.725	0.486
80	0.795	0.630	0.502
90	0.781	0.719	0.527
100	0.741	0.775	0.484

Table 8 – Results – Neural Network – All Features with two variables – R^2 – IQR

Hidden Neurons \ Hidden Layers	1	2	3
40	0.861	0.539	0.550
50	0.857	0.690	0.733
60	0.765	0.801	0.836
70	0.913	0.742	0.902
80	0.836	0.898	0.803
90	0.519	0.830	0.813
90	0.736	0.735	0.812

In order to better assess the networks, just those with R^2 close to 0.9 were analyzed. Therefore, our case for the two-standard deviation is a neural network having 1 layer with 60 neurons. Moreover, for the interquartile range method, the following setups are examined: 1 layer with 70 neuros; 2 layers with 80 neurons; and 3 layers with 70 neurons.

Table 9 – Results – Neural Network – All Features with two variables – Best Results – Statistics – 2 STD

Statistics	Hidden Neurons	Hidden Neurons	Hidden Neurons	Hidden Neurons
	60	70	80	70
Method	**2 STD**	**IQR**	**IQR**	**IQR**
Number of Layers	**1**	**1**	**2**	**3**
R^2	0.9005	0.9126	0.8979	0.9019
MAE	0.0474	0.0398	0.0407	0.0415
MSE	0.0041	0.0045	0.0054	0.0043

Ultimately, it was chosen the following setup to be plotted: 1 layer with 60 neurons for 2 standard deviations method. In addition, it had been decided to display the following setup: 1 layers with 70 neurons.

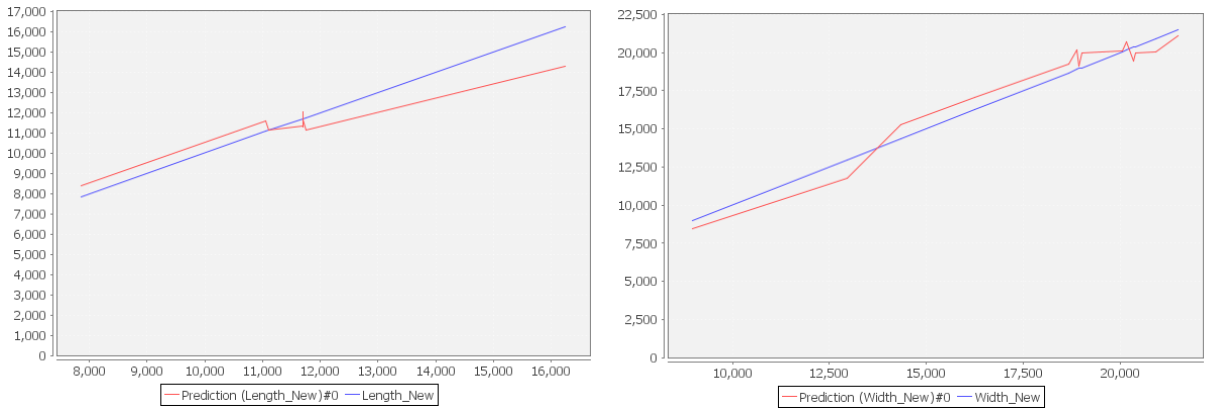

Figure 73 – Results – Neural Network – All Features with two variables – 1 layer and 60 neurons – 2 STD

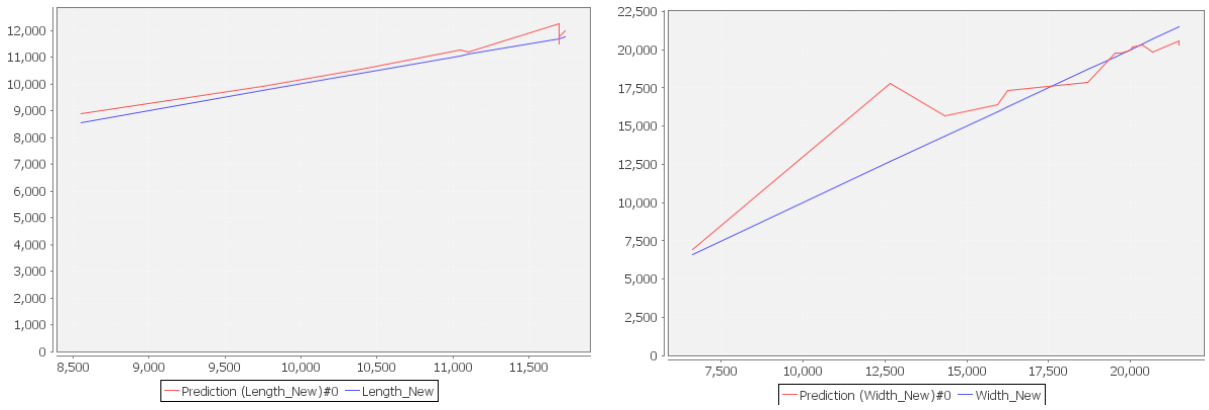

Figure 74 – Results – Neural Network – All Features with two variables – 1 layer and 70 neurons – IQR

4.3.3. One Feature with two variables and PCA Analysis

As it had been mentioned for the previous attempt, while the polynomial regression method could not handle the number of variables, the neural network could learn from the features and present results. Nonetheless, the coefficient of determination was improved by the reduction of variables to be studied. Thus, starting with the coefficient of determination and analyzing the errors a best configuration was outlined for each outlier removal method.

Table 10 – Results – Neural Network – One Feature with two variables – R^2 – 2 STD

Hidden Neurons \ Hidden Layers	1	2	3
1	0.750	0.634	0.597
11	0.745	0.829	0.769
21	0.837	0.732	0.725
31	0.974	0.895	0.804
41	0.979	0.904	0.755
51	0.981	0.767	0.781
61	0.952	0.941	0.810
71	0.977	0.868	0.737
81	0.947	0.870	0.677
91	0.932	0.841	0.762

Table 11 – Results – Neural Network – One Feature with two variables – R^2 – IQR

Hidden Neurons \ Hidden Layers	1	2	3
1	0.910	0.755	0.512
11	0.950	0.914	0.949
21	0.953	0.966	0.963
31	0.961	0.959	0.967
41	0.971	0.964	0.962
51	0.973	0.969	0.952
61	0.973	0.965	0.966
71	0.975	0.944	0.984
81	0.972	0.980	0.969
91	0.972	0.965	0.962

Table 12 – Results – Neural Network – One Feature with two variables – R^2 – 2 STD - PCA

Hidden Neurons \ Hidden Layers	1	2	3
1	0.455	0.472	0.716
11	0.917	0.953	0.916
21	0.923	0.974	0.965
31	0.917	0.917	0.913
41	0.964	0.885	0.978
51	0.970	- 5.009	0.892
61	0.972	0.970	0.957
71	0.965	0.830	0.728
81	0.955	0.948	0.948
91	0.921	0.952	0.931

Table 13 – Results – Neural Network – One Feature with two variables – R^2 – IQR - PCA

Hidden Neurons \ Hidden Layers	1	2	3
1	0.828	0.862	0.706
11	- 1.364	0.919	0.968
21	0.974	0.957	0.969
31	0.975	0.956	0.884
41	0.966	0.950	0.963
51	0.954	0.938	0.946
61	- 1.424	0.973	0.984
71	0.963	0.958	0.962
81	0.967	0.954	0.861
91	0.962	0.899	0.946

In order to better assess the networks, just those with R^2 above 0.95 were analyzed. Therefore, our initial ranges are neural networks with 1 layer and neurons varying from 31 to 71 for the two-standard deviation. As well, neural networks varying layers from 1 to 3 and neurons from 11 to 91 for the interquartile range method.

Table 14 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – 2 STD – 1 Layer

Statistics	Hidden Neurons				
	31	41	51	61	71
R^2	0.9743	0.9787	0.9807	0.9518	0.9772
MAE	0.0274	0.0234	0.0248	0.0311	0.0247
MSE	0.0017	0.0013	0.0013	0.0029	0.0014

Table 15 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – 1 Layer

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	0.9500	0.9526	0.9606	0.9715	0.9731	0.9727	0.9745	0.9717	0.9724
MAE	0.0212	0.0262	0.0227	0.0217	0.0204	0.0196	0.0205	0.0207	0.0198
MSE	0.0019	0.0019	0.0015	0.0011	0.0011	0.0010	0.0010	0.0011	0.0011

Table 16 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – 2 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9659	0.9593	0.9641	0.9691	0.9651	-	0.9797	0.9652
MAE	-	0.0144	0.0155	0.0149	0.0135	0.0156	-	0.0133	0.0174
MSE	-	0.0013	0.0015	0.0013	0.0011	0.0013	-	0.0008	0.0013

Table 17 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – 3 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9627	0.9666	0.9622	0.9524	0.9660	0.9844	0.9685	0.9625
MAE	-	0.0113	0.0134	0.0111	0.0167	0.0151	0.0092	0.0110	0.0108
MSE	-	0.0013	0.0013	0.0014	0.0017	0.0013	0.0006	0.0011	0.0013

Table 18 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – 2 STD – PCA – 1 Layer

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	-	-	0.9642	0.9701	0.9724	0.9646	0.9545	-
MAE	-	-	-	0.0237	0.0262	0.0249	0.0229	0.0269	-
MSE	-	-	-	0.0010	0.0015	0.0013	0.0011	0.0013	-

Table 19 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – 2 STD – PCA – 2 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	0.9529	0.9739	-	-	-	0.9702	-	-	0.9516
MAE	0.0333	0.0244	-	-	-	0.0201	-	-	0.0200
MSE	0.0027	0.0015	-	-	-	0.0011	-	-	0.0008

Table 20 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – 2 STD – PCA – 3 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9655	-	0.9783	-	0.9573	-	-	-
MAE	-	0.0269	-	0.0236	-	0.0281	-	-	-
MSE	-	0.0016	-	0.0012	-	0.0017	-	-	-

Table 21 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – PCA – 1 Layer

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9742	0.9749	0.9656	0.9542	-	0.9626	0.9675	0.9623
MAE	-	0.0234	0.0245	0.0263	0.0270	-	0.0246	0.0256	0.0244
MSE	-	0.0010	0.0012	0.0012	0.0020	-	0.0012	0.0011	0.0011

Table 22 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – PCA – 2 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9572	0.9555	-	-	0.9730	0.9576	0.9543	-
MAE	-	0.0268	0.0271	-	-	0.0220	0.0258	0.0238	-
MSE	-	0.0021	0.0018	-	-	0.0012	0.0015	0.0012	-

Table 23 – Results – Neural Network – One Feature with two variables – Best Results – Statistics – IQR – PCA – 3 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	0.9678	0.9694	-	0.9634	-	0.9837	0.9624	-	-
MAE	0.0207	0.0237	-	0.0268	-	0.0165	0.0218	-	-
MSE	0.0018	0.0018	-	0.0025	-	0.0012	0.0010	-	-

Finally, it was chosen the following setup to be plotted: 1 layer with 41 neurons for 2 standard deviations method. Moreover, the following configuration using PCA is displayed below, 3 layers and 41 neurons. In addition, it had been decided to display the following setup: 3 layers with 71

neurons. Furthermore, a representation with 3 layers and 61 neurons using PCA as outlier removal is shown below.

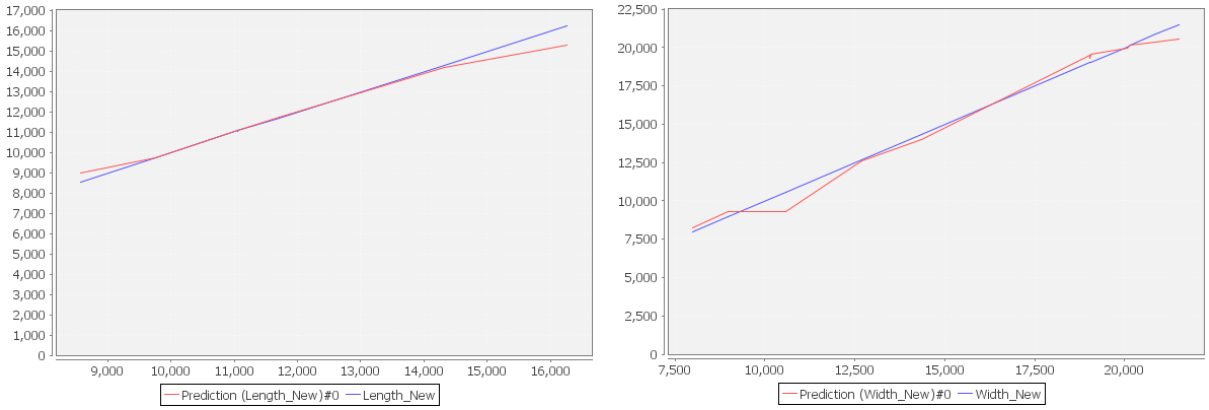

Figure 75 – Results – Neural Network – All Features with two variables – 1 layer and 41 neurons – 2 STD

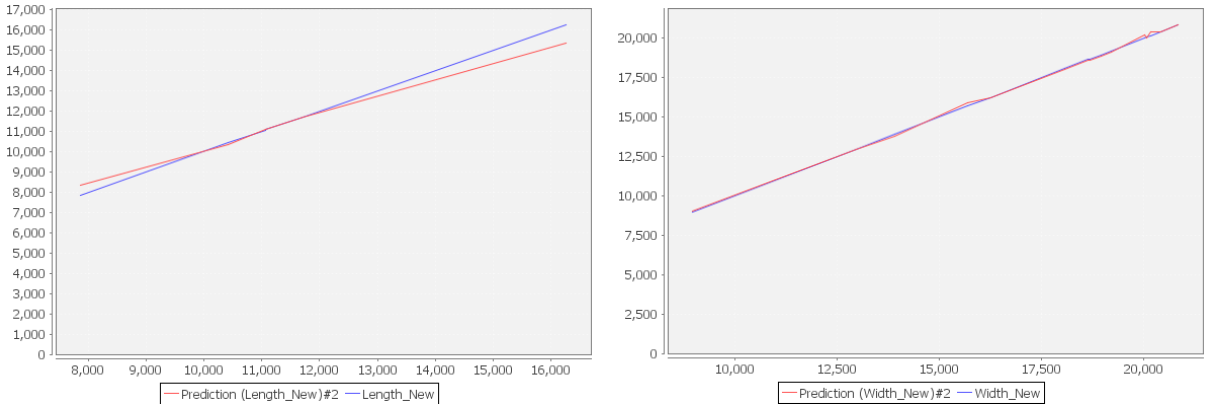

Figure 76 – Results – Neural Network – All Features with two variables – 3 layers and 71 neurons – IQR

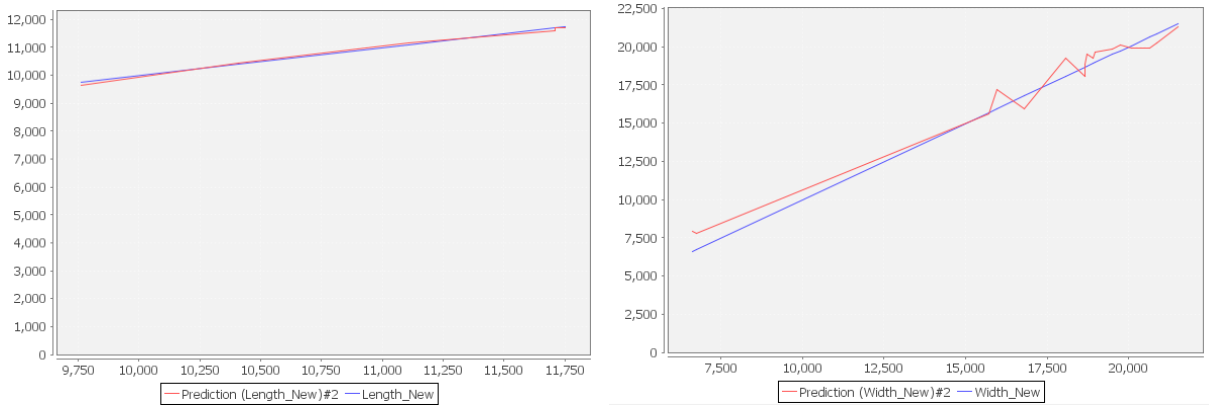

Figure 77 – Results – Neural Network – All Features with two variables – 3 layers and 41 neurons – 2 STD – PCA

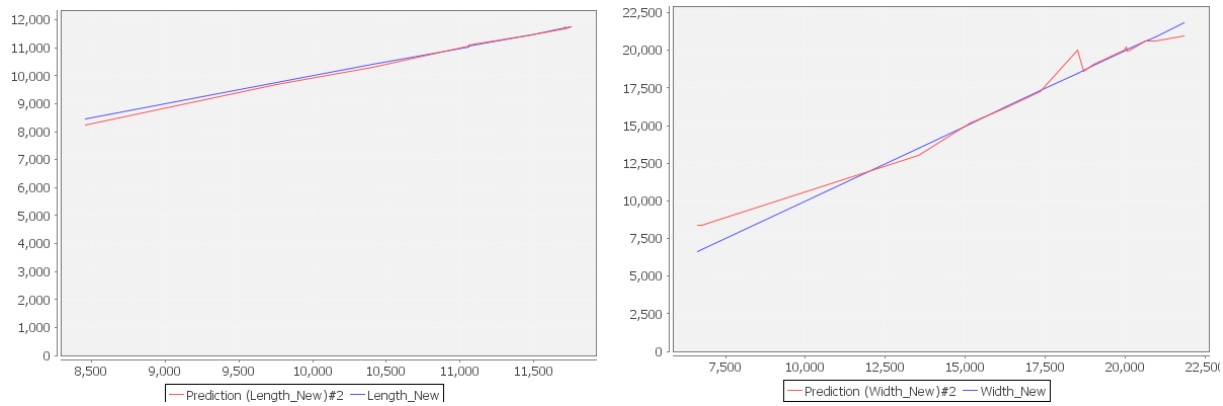

Figure 78 – Results – Neural Network – All Features with two variables – 3 layers and 61 neurons – IQR – PCA

It can be seen from the charts that the prediction curve detaches from the actual values. This can be explained by the reduced database which does not provide much information to be learn on those areas. Similarly, on the edges of the curve a gap is generated. This behavior indicates that the limits of learning database are being reached. Therefore, it is where the model is limited to and values further on that region will not have any reasonable value.

4.4. Selected Features

A back-feature selection was applied using a neural network model and both two-standard deviation and interquartile range methods. By using the results of the back-feature selection, a new round with a neural network and polynomial regression learning took place. The results of the applied methods are available in Appendix III. However, tree map charts based on the squared errors and number of features for length and width were plotted in Figure 79 and Figure 80.

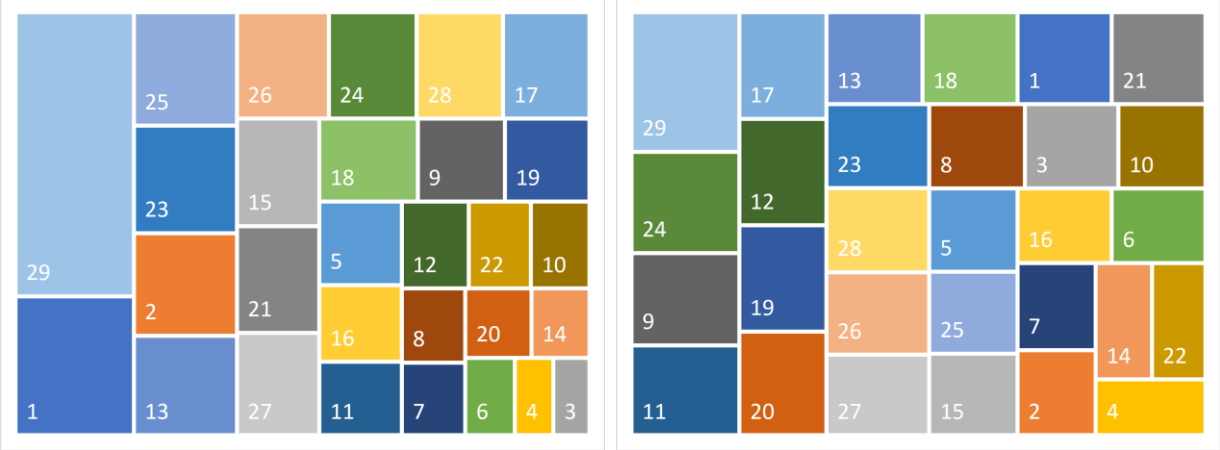

Figure 79 – Results – Polynomial Regression – Back Feature Selection with two variables – 2 STD

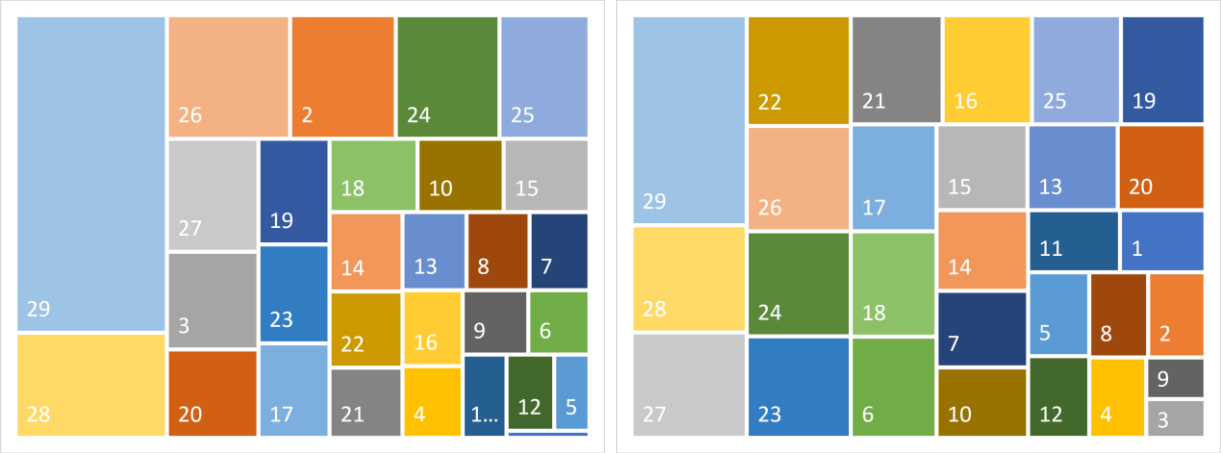

Figure 80 – Results – Polynomial Regression – Back Feature Selection with two variables – IQR

As the number of features are reduced the squared error decreases. Additionally, the table below contains the selected features in order to obtain the lowest squared errors for both outlier removal methods.

Table 24 – Results – Back feature selection – Selected Features

Direction	2-STD	IQR
Length	Length_AVG_SF; Elements_Transversal; Complete_Weld_Transversal	Length_AVG;
Width	Width_AVG_SF Spacing_Stiffener; Weightage_kg Thickness_Main_Plate	Width_AVG_SF; Length_AVG; Width_AVG;

4.4.1. Polynomial Regression

In the case of polynomial regression, even though the selected features were introduced the results were worse by comparing MAE and MSE. It is also noticed that the visual divergence starts in a lower degree. The best result of Polynomial regression with selected features is shown in Table 25.

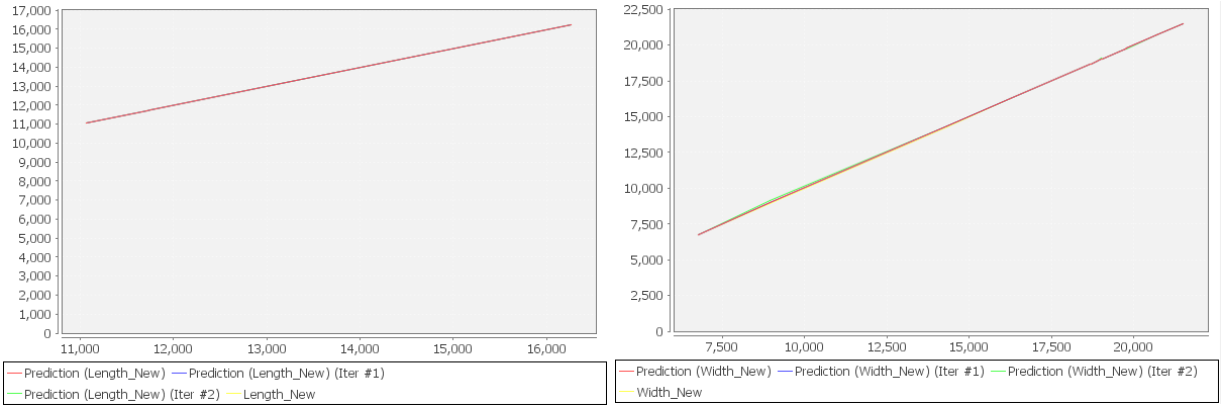

Figure 81 – Results – Polynomial Regression – Back Feature Selection with two variables – from 1 to 3 degrees – 2 STD

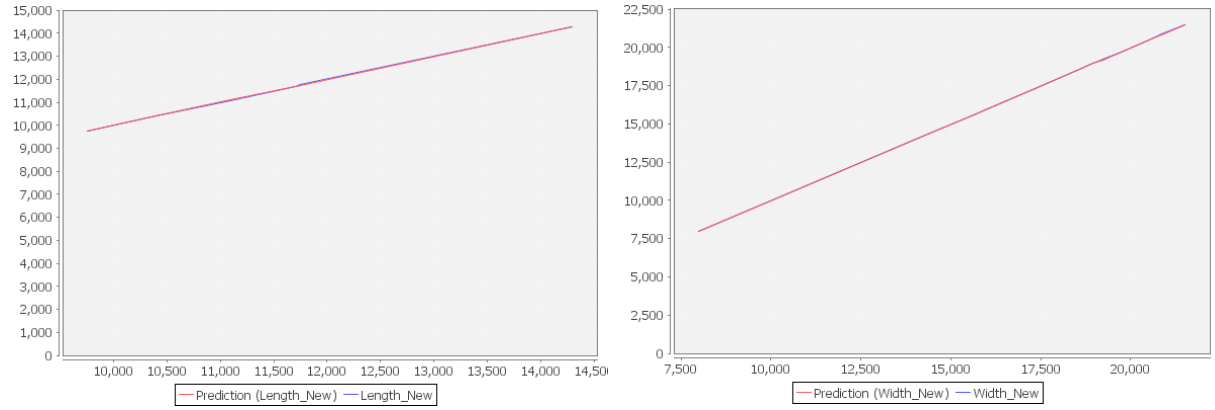

Figure 82 – Results – Polynomial Regression – Back Feature Selection with two variables – 1 degree – IQR

Table 25 – Results – Polynomial Regression – Back Feature Selection with two variables – Best Result – IQR - 1 degree

Row ID	Prediction (Length_New)	Prediction (Width_New)	Number of Degrees	Average
R^2	1.00	1.00	1	1.00
MAE	3.61	5.53	1	4.57
MSE	22.04	40.96	1	31.50

4.4.2. Neural Network

Once again, setups that present coefficient of determination above 0.95 were analyzed.

Table 26 – Results – Neural Network – Back Feature selection with two variables – R^2 – 2 STD

Hidden Neurons \ Hidden Layers	1	2	3
1	0.455	0.472	0.716
11	0.917	0.953	0.916
21	0.923	0.974	0.965
31	0.917	0.917	0.913
41	0.964	0.885	0.978
51	0.970	- 5.009	0.892
61	0.972	0.970	0.957
71	0.965	0.830	0.728
81	0.955	0.948	0.948
91	0.921	0.952	0.931

Table 27 – Results – Neural Network – Back Feature selection with two variables – R^2 – IQR

Hidden Neurons \ Hidden Layers	1	2	3
1	0.956	0.642	0.624
11	0.972	0.918	0.888
21	0.971	0.978	0.931
31	0.969	0.975	0.965
41	0.968	0.976	0.968
51	0.973	0.971	0.945
61	0.969	0.959	0.958
71	0.968	0.974	0.962
81	0.964	0.962	0.984
91	0.962	0.959	0.983

Table 28 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – 2 STD – 1 Layer

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	-	-	0.9642	0.9701	0.9724	0.9646	0.9545	-
MAE	-	-	-	0.0237	0.0262	0.0249	0.0229	0.0269	-
MSE	-	-	-	0.0010	0.0015	0.0013	0.0011	0.0013	-

Table 29 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – 2 STD – 2 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	0.9529	0.9739	-	-	-	0.9702	-	-	0.9516
MAE	0.0333	0.0244	-	-	-	0.0201	-	-	0.0200
MSE	0.0027	0.0015	-	-	-	0.0011	-	-	0.0008

Table 30 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – 2 STD – 3 Layers

Statistics	Hidden Neurons								
	11	21	31	41	51	61	71	81	91
R^2	-	0.9655	-	0.9783	-	0.9573	-	-	-
MAE	-	0.0269	-	0.0236	-	0.0281	-	-	-
MSE	-	0.0016	-	0.0012	-	0.0017	-	-	-

Table 31 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – IQR – 1 Layer

Statistics	Hidden Neurons									
	1	11	21	31	41	51	61	71	81	91
R^2	0.9565	0.9723	0.9707	0.9693	0.9675	0.9732	0.9691	0.9684	0.9637	0.9618
MAE	0.0285	0.0196	0.0204	0.0202	0.0201	0.0182	0.0171	0.0197	0.0191	0.0205
MSE	0.0015	0.0008	0.0009	0.0009	0.0009	0.0008	0.0008	0.0009	0.0009	0.0010

Table 32 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – IQR – 2 Layers

Statistics	Hidden Neurons									
	1	11	21	31	41	51	61	71	81	91
R^2	-	-	0.9785	0.9753	0.9758	0.9715	0.9589	0.9737	0.9623	0.9589
MAE	-	-	0.0090	0.0111	0.0104	0.0110	0.0171	0.0156	0.0160	0.0141
MSE	-	-	0.0004	0.0005	0.0005	0.0006	0.0010	0.0006	0.0009	0.0009

Table 33 – Results – Neural Network – Back Feature selection with two variables – Best Results – Statistics – IQR – 3 Layers

Statistics	Hidden Neurons									
	1	11	21	31	41	51	61	71	81	91
R^2	-	-	-	0.9646	0.9684	-	0.9580	0.9621	0.9844	0.9834
MAE	-	-	-	0.0126	0.0113	-	0.0106	0.0111	0.0079	0.0097
MSE	-	-	-	0.0007	0.0006	-	0.0008	0.0008	0.0003	0.0004

For the two-standards deviation the selected configuration is 3 hidden layers with 41 neurons whereas for the IQR the optimal setup is 3 hidden layers with 81 neurons.

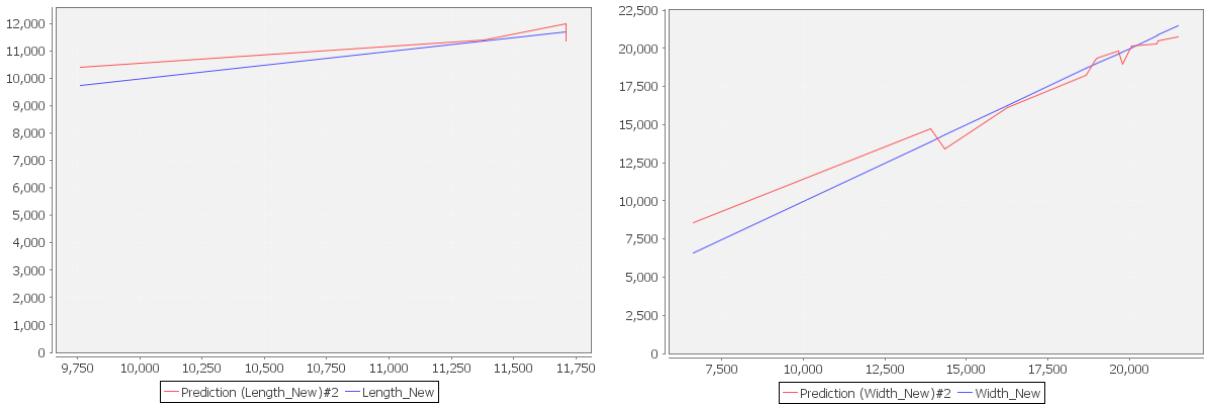

Figure 83 – Results – Neural Network – All Features with two variables – 3 layers and 41 neurons – 2 STD

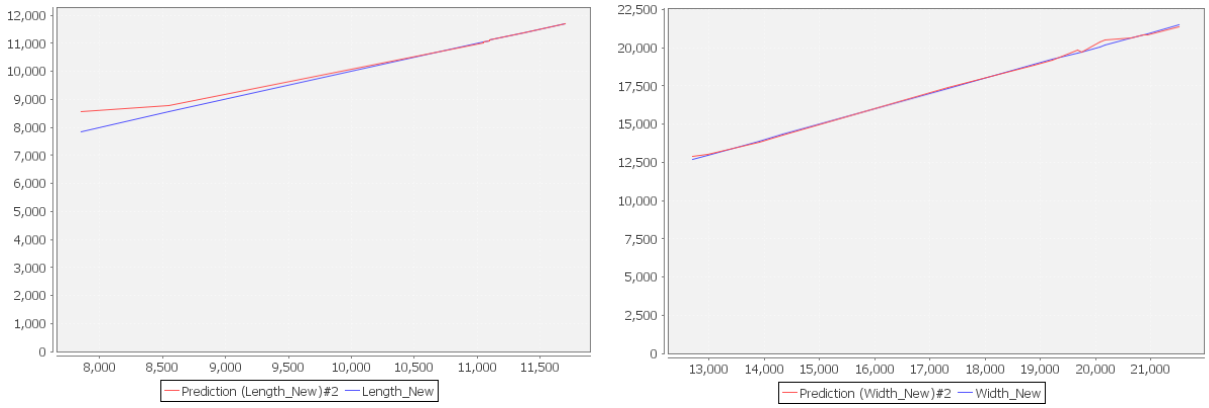

Figure 84 – Results – Neural Network – All Features with two variables – 3 layers and 81 neurons – IQR

4.5. Best Fitting

This method was applied in order to come out with simple formulae and to compare with the current shrinkage factor.

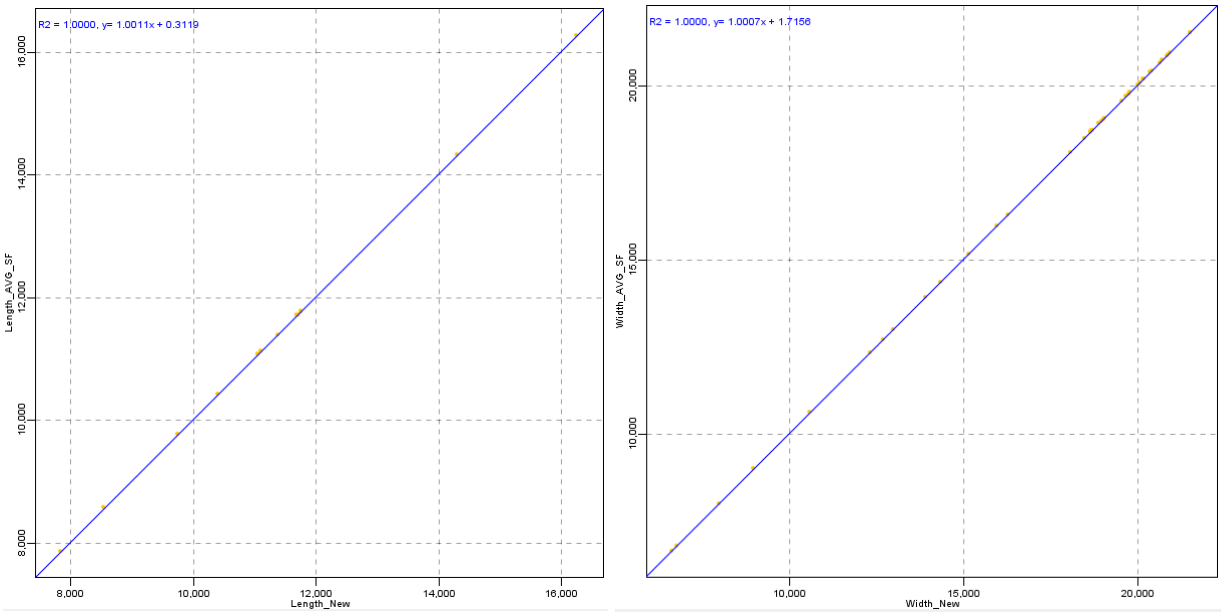

Figure 85 – Results – Best Fitting – Length_AVG_SF vs Length_New (Left) and Width_AVG_SF vs Width_New (Right) – 2 STD

$$Length_{AVG_{SF}} = 1.0011 \times Length_{New} + 0.3119$$

$$Width_{AVG_{SF}} = 1.0007 \times Width_{New} + 1.7156$$

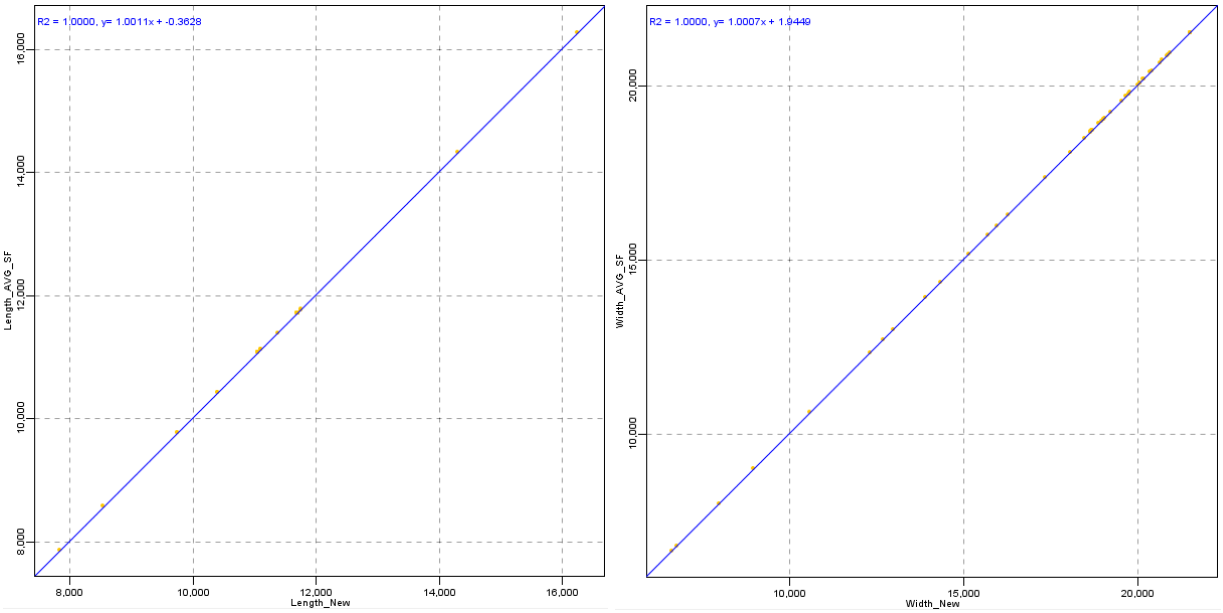

Figure 86 – Results – Best Fitting – Length_AVG_SF vs Length_New (Left) and Width_AVG_SF vs Width_New (Right) – IQR

$$Length_{AVG_{SF}} = 1.0011 \times Length_{New} - 0.3628$$

$$Width_{AVG_{SF}} = 1.0007 \times Width_{New} + 1.9449$$

The multiplications and constants are in millimeters. Considering that, the length formula could disregard the adjustment and by comparing the length factors from both 2 STD and IQR to the original it indicates the increase the factor by 0.0001.

As well, the average width of a block is 18.65m and if you divide the constant by this value, the factor could be increased by 0.0001, hence the formula factor could be taken as approximately 1.0008 which is lower than the actual shrinkage factor. Therefore, it would be possible to reduce it.

5. CONCLUSIONS

The welding distortions can be caused by three main categories: geometric parameters (design), material properties and welding process parameters (manufacture process). In this study, the geometric parameters were examined.

Firstly, all possible features were gathered starting from those available from the design system and then detailing them manually from technical drawings. These data were divided into: block main characteristics – estimated values by personnel; features retrieved from the system database; and features outlined from technical drawing. Subsequently, all possible measurements were joined to our data. The data treatment took around 4 months to be completed.

The selected program, KNIME, proved to be versatile for allowing the development of many different setups in an easy manner as detailed in chapter 3. It is a very straight-forwarded environment with a lot of different methods to analyze data. Even so, the development of the model took 3 months to achieve its final stage as it had been noticed space for improvement.

Two different outlier removal methods were used. By observing the selected setups, the IQR was the best one to work together with the neural network whereas the 2 STD was the best one to the polynomial regression method. It is believed that IQR had better results with neural network due to higher number of rows which allowed more data to be used in learning.

In this study the polynomial regression had better coefficients of determination when compared to the neural networks and, also, visually it could be seen a better behavior in the plot.

On the other hand, when coming to analyze non-linear behaviors including a lot of features, the polynomial regression cannot give any guess if your database is limited. In the same case, the neural network could give results even with limited data.

On top of that, the reduction of variables was mandatory in order to progress with the study. When presenting the 8 variables, our dataset was too limited and the nonlinearities too big solely giving negative coefficients of determination.

After the reduction of variables to be studied, initial results started to be achieved and they had been improved by the reduction of features. Additionally, it was found that number of transversal elements and the number of transversal welding seams influence the length while stiffener spacing, weight and thickness of main plate affect the width.

It had been noticed that even with the application of various methods that graphically no neural network was satisfactory even if the coefficient of determination and errors seemed to be low. The detachments along the way can be explained by the lack of data for the algorithm to learn and also due to overlearning in the case of setups with higher number of layer and neurons. Again, the gaps presented on the edges of the curves are due to the upper and lower boundaries of the database being reach by the predictor which can be improved by providing more data.

Having said that, a best fitting data comparison was performed finding that the factor should be slightly adjusted for both length and width according to the supplied data.

All in all, more data is elementary to perform firm conclusions and to further learn and test the model. KNIME is a powerful tool and the developed model can be adjusted and used to assess distortions in other elements such as sub-assemblies as long as the data is collected. Moreover, it is still necessary to verify how to integrate the PMML model with the design software used by the company in order to reduce workload of the nesting/CAM team. Finally, as another topic it could be assessed if the excess of material which is applied on top of the shrinkage factor could be reduced.

6. ACKNOWLEDGEMENTS

Firstly, I would like to thank both of my university supervisors Prof. Remigiusz Iwańkowicz and Prof. Prof. Jean-David Caprace for the given instruction. Without their advices on the path to be explored, this thesis could not be completed.

Additionally, Prof. Luciana Martimiano and Prof. Wagner Igarashi from State University of Maringá (UEM) who assisted me on selecting better references for the machine learning review and discussing different applications of machine learning algorithms.

Moreover, to M.Sc. Nicole Schenk who selected and supervised me during the internship. As well, all colleagues from Lürssen Werft who assisted me with learning German, understanding the company's production process and daily-living matters.

Also, I would like to extend my thanks to Leonardo Lopes, Leonardo Ruela and KNIME community with whom I discussed and came out with different ways to model the prediction tool. Without their assistance it would not be possible to learn the software in short-notice period.

Further, I would like to thank the friends either from ERASMUS or daily living in different countries. Your support on learning languages, on understanding various cultures and as friends is extremely appreciated.

Finally, I would like to state my special thanks to my family who supported me in my decisions and still sharing their experiences with me so that I can improve as a person and as a professional worker. I dedicate this thesis to all of them.

This thesis was developed in the frame of the European Master Course in "Integrated Advanced Ship Design" named "EMSHIP" for "European Education in Advanced Ship Design", Ref.: 159652-1-2009-1-BE-ERA MUNDUS-EMMC.

7. REFERENCES

16_splitandstartrun.Jpg(1018×825) N.d. https://jamesmccaffrey.files.wordpress.com/2015/07/16_splitandstartrun.jpg, accessed December 20, 2017.

61Z360TshtL._SL1001_.Jpg (1001×1001) N.d. https://images-na.ssl-images-amazon.com/images/I/61Z360TshtL._SL1001_.jpg, accessed December 20, 2017.

Alpaydın, Ethem 2014 Introduction to Machine Learning. 3. ed. Adaptive Computation and Machine Learning. Cambridge, Mass.: MIT Press.

Barber, D. 2012 Bayesian Reasoning and Machine Learning. Cambridge University Press.

Caprace, Jean-David, Francisco Aracil Fernandez, Nicolas Losseau, and Philippe Rigo 2009 A Fuzzy Metric for Assessing the Producibility of Straightening in Early Design. *In* The 14th International Conference on Computer Applications in Shipbuilding (ICCAS) Pp. 211–218. http://orbi.ulg.ac.be/handle/2268/21472, accessed August 28, 2017.

Caprace, Jean-David, Nicolas Losseau, Frederic Bair, et al. 2007 A Data Mining Analysis Applied to a Straightening Process Database. Ship Technology Research 54(4): 177–183.

Chandrashekar, Girish, and Ferat Sahin 2014 A Survey on Feature Selection Methods. Computers & Electrical Engineering 40(1): 16–28.

Communications, TRUE 2017 KNIME vs. RapidMiner. Cmotions. https://cmotions.nl/en/knime-vs-rapidminer/, accessed December 9, 2017.

Deng, De-An 2010 Theoretical Prediction of Welding Distortion in Large and Complex Structures. Frontiers of Materials Science in China 4(2): 202–209.

Deng, Dean, Wei Liang, and Hidekazu Murakawa 2007 Determination of Welding Deformation in Fillet-Welded Joint by Means of Numerical Simulation and Comparison with Experimental Measurements. Journal of Materials Processing Technology 183(2–3): 219–225.

Deng, Dean, and Hidekazu Murakawa 2008a Prediction of Welding Distortion and Residual Stress in a Thin Plate Butt-Welded Joint. Computational Materials Science 43(2): 353–365.

2008b FEM Prediction of Buckling Distortion Induced by Welding in Thin Plate Panel Structures. Computational Materials Science 43(4): 591–607.

Deng, Dean, Hidekazu Murakawa, and Wei Liang 2007 Numerical Simulation of Welding Distortion in Large Structures. Computer Methods in Applied Mechanics and Engineering 196(45–48): 4613–4627.

2008 Prediction of Welding Distortion in a Curved Plate Structure by Means of Elastic Finite Element Method. Journal of Materials Processing Technology 203(1–3): 252–266.

File:Ship Design Spiral.Jpg - MarineWiki N.d. http://www.marinewiki.org/index.php?title=File:Ship_design_spiral.jpg, accessed October 11, 2017.

Gray, Tom, D. Camilleri, and N. McPherson 2014 Control of Welding Distortion in Thin-Plate Fabrication: Design Support Exploiting Computational Simulation. Elsevier.

Kim, J, H Jeong, M Ji, et al. 2015 A Study on the Compensation Margin on Butt Welding Joint of Large Steel Plates during Shipbuilding Construction. IOP Conference Series: Materials Science and Engineering 88: 012040.

Mahendramani, G., and N. Lakshmana Swamy 2012 Effect of Included Angle in V-Groove Butt Joints on Shrinkages in Submerged Arc Welding Process. International Journal of Engineering Science and Technology (IJEST) 04: 1607–1613.

Maxresdefault.Jpg (960×720) N.d. https://i.ytimg.com/vi/8o3_kprq6ro/maxresdefault.jpg, accessed December 20, 2017.

Maxresdefault.Jpg (1440×900) N.d. https://i.ytimg.com/vi/4QWdkzg_aec/maxresdefault.jpg, accessed December 20, 2017.

Maxresdefault.Jpg (1920×1080) N.d. https://i.ytimg.com/vi/lTMqXSSjCvk/maxresdefault.jpg, accessed December 20, 2017.

Product_cx_05.Jpg (375×310) N.d. https://us.sokkia.com/sites/default/files/styles/product_gallery_full_normal_1x/public/cx-total-station-series/gallery/product_cx_05.jpg?itok=GHkwbr5d, accessed December 20, 2017.

TAJIMA, Yusuke, Sherif RASHED, Yasuhisa OKUMOTO, Yasuo KATAYAMA, and Hidekazu MURAKAWA 2007 Prediction of Welding Distortion and Panel Buckling of Car Carrier Decks Using Database Generated by FEA. JWRI 36: 65–71.

Welding Defect 2017 Wikipedia. https://en.wikipedia.org/w/index.php?title=Welding_defect&oldid=796756173, accessed September 12, 2017.

Wimmer, Hayden, and Loreen Marie Powell 2016 A Comparison of Open Source Tools for Data Science. Journal of Information Systems Applied Research 9(2): 4.

Yang, Y. P., R. Dull, H. Castner, T. D. Huang, and D. Fanguy 2014 Material Strength Effect on Weld Shrinkage and Distortion. Welding Journal 93: 421–s–430–s.

APPENDIX I – PRODUCTION PROCESS MAPPING

This appendix should not be distributed outside of the consortium as per de non-disclosure agreement (NDA – 10/01/2018).

APPENDIX II – MATHEMATICAL FORMULAS

Pre-processing – Length_AVG

($Target-Bow-PS-X$ - $Target-Stern-PS-X$)
+ ($Target-Bow-Stb-X$ - $Target-Stern-Stb-X$))/2

Pre-processing – Width_AVG

($Target-Stern-Stb-Y$ - if($Target-Stern-PS-Y$>=0,$Target-Stern-PS-Y$*-1,$Target-Stern-PS-Y$))+ ($Target-Bow-Stb-Y$ - if($Target-Bow-PS-Y$>=0,$Target-Bow-PS-Y$*-1,$Target-Bow-PS-Y$)))/2

Pre-processing – Length_AVG_SF

($Length_AVG$/1000 * SF_Length)*1000

Pre-processing – Width_AVG_SF

($Width_AVG$/1000 * SF_Width)*1000

Pre-processing – Length_New

(($After-Bow-Stb-X$ - $After-Stern-Stb-X$)+
($After-Bow-PS-X$ - $After-Stern-PS-X$))/2

Pre-processing – Width_New

(($After-Stern-Stb-Y$ - if($After-Stern-PS-Y$>=0,$After-Stern-PS-Y$*-1,$After-Stern-PS-Y$))+($After-Bow-Stb-Y$ - if($After-Bow-PS-Y$>=0,$After-Bow-PS-Y$*-1,$After-Bow-PS-Y$)))/2

APPENDIX III – BACK FEATURE SELECTION RESULTS

The best results are highlighted in yellow.

Table 34 – Results – Neural Network – Back Feature Selection with two variables – 2 STD – Length

RowID	Nr. of features	Squarred Error	Removed feature
0	0		Length_AVG_SF
1	1	0.1779	Elements_Transversal
2	2	0.1143	Complete_Weld_Transversal
3	3	0.0301	Elements_Girder
4	4	0.0322	Complete_Weld_Girder
5	5	0.0746	Height
6	6	0.0414	Spacing_Frame
7	7	0.0498	Complete_Weld_Longitudinal
8	8	0.0509	Length
9	9	0.0781	Length_AVG
10	10	0.0549	Complete_Weld_Butt
11	11	0.0654	Elements_Longitudinal
12	12	0.0629	SF_Length
13	13	0.1107	SF_Width
14	14	0.0435	Width_AVG_SF
15	15	0.0958	Weightage_kg
16	16	0.0685	Width_AVG
17	17	0.0993	Spacing_Stiffener
18	18	0.0881	Ship
19	19	0.0757	Elements_Butt
20	20	0.0497	Volume
21	21	0.0949	Width
22	22	0.0585	Type
23	23	0.1196	Section
24	24	0.1012	Width_New
25	25	0.1265	Weightage
26	26	0.1058	Spacing_Girder
27	27	0.0911	Thickness_Main_Plate
28	28	0.0996	Welding_Length
All		29	0.3643

Table 35 – Results – Neural Network – Back Feature Selection with two variables – 2 STD – Width

RowID	Nr. of features	Squarred Error	Removed feature
0	0		Width_AVG_SF
1	1	0.1519	Spacing_Stiffener
2	2	0.1168	Weightage_kg
3	3	0.1379	Thickness_Main_Plate
4	4	0.1073	Length
5	5	0.1285	Complete_Weld_Longitudinal
6	6	0.1225	Complete_Weld_Butt
7	7	0.1193	Volume
8	8	0.1415	Ship
9	9	0.1745	Welding_Length
10	10	0.1286	Elements_Butt
11	11	0.1698	Elements_Girder
12	12	0.1630	Type
13	13	0.1548	Width_AVG
14	14	0.1143	Length_AVG
15	15	0.1262	Height
16	16	0.1233	SF_Length
17	17	0.1632	Section
18	18	0.1539	Length_New
19	19	0.1626	Length_AVG_SF
20	20	0.1586	Elements_Transversal
21	21	0.1515	Weightage
22	22	0.1092	Spacing_Girder
23	23	0.1512	Elements_Longitudinal
24	24	0.1912	SF_Width
25	25	0.1269	Spacing_Frame
26	26	0.1487	Complete_Weld_Girder
27	27	0.1448	Complete_Weld_Transversal
28	28	0.1511	Width
All		29	0.2643

Table 36 – Results – Neural Network – Back Feature Selection with two variables – IQR – Length

RowID	Nr. of features	Squarred Error	Removed feature
0	0		Length_AVG
1	1	0.0076	Weightage_kg
2	2	0.1667	Complete_Weld_Transversal
3	3	0.1137	Spacing_Girder
4	4	0.0552	Elements_Transversal
5	5	0.0341	Complete_Weld_Girder
6	6	0.0503	Width_AVG_SF
7	7	0.0599	Length
8	8	0.0631	Elements_Girder
9	9	0.0541	Length_AVG_SF
10	10	0.0806	Weightage
11	11	0.0466	Thickness_Main_Plate
12	12	0.0466	Elements_Longitudinal
13	13	0.0638	Complete_Weld_Butt
14	14	0.0742	Complete_Weld_Longitudinal
15	15	0.0801	Height
16	16	0.0581	Section
17	17	0.0861	SF_Width
18	18	0.0823	Ship
19	19	0.0960	Width_AVG
20	20	0.1034	Elements_Butt
21	21	0.0654	Width_New
22	22	0.0711	Width
23	23	0.0899	Spacing_Frame
24	24	0.1639	Welding_Length
25	25	0.1419	SF_Length
26	26	0.1947	Volume
27	27	0.1320	Type
28	28	0.2039	Spacing_Stiffener
All	29	0.6154	

Table 37 – Results – Neural Network – Back Feature Selection with two variables – IQR – Width

RowID	Nr. of features	Squarred Error	Removed feature
0	0		Width_AVG_SF
1	1	0.0741	Length_AVG
2	2	0.0682	Width_AVG
3	3	0.0322	Thickness_Main_Plate
4	4	0.0643	Weightage
5	5	0.0720	Welding_Length
6	6	0.1215	Width
7	7	0.0980	Length_New
8	8	0.0700	Elements_Longitudinal
9	9	0.0347	Complete_Weld_Butt
10	10	0.0905	Elements_Butt
11	11	0.0800	Weightage_kg
12	12	0.0700	Spacing_Frame
13	13	0.1082	Spacing_Stiffener
14	14	0.1033	Complete_Weld_Transversal
15	15	0.1094	SF_Width
16	16	0.1368	SF_Length
17	17	0.1278	Height
18	18	0.1264	Complete_Weld_Girder
19	19	0.1294	Complete_Weld_Longitudinal
20	20	0.1055	Spacing_Girder
21	21	0.1404	Elements_Transversal
22	22	0.1615	Volume
23	23	0.1483	Elements_Girder
24	24	0.1529	Length_AVG_SF
25	25	0.1366	Type
26	26	0.1538	Section
27	27	0.1703	Ship
28	28	0.1734	Length
All		29	0.3397