Predicting the amount of breast cap in broilers

Mantzaris, A.

Award date:
2016

Department of Mathematics and Computer Science
Information Systems Research Group

Predicting the Amount of Breast Cap in Broilers

Master Thesis

Antonios Mantzaris

Supervisors:

Dr. M. Razavian, TU/e, IE&IS
Dr. A. M. Wilbik, TU/e, IE&IS
D. Kuijpers, Marel Stork Poultry Processing BV

Eindhoven, 22 February 2017

Abstract

The current research study investigates the relationship between the breast cap weight of a broiler and visual measurements of its body (areas, distances from point to point and ratios of these distances) along with its total weight by using the CRISP-DM methodology and different types of data mining techniques. The techniques used are Multiple Linear Regression, Regression Trees, Neural Networks and Fuzzy Inference Systems. Moreover, it is examined which of these techniques is the most suitable for the present case. In addition, the most relevant predictors for the breast cap weight of the broiler are selected.

The analysis is based on data recordings generated by the systems of a poultry processor during several production days. For the present case, the Multiple Linear Regression performed better than the other methods demonstrating the ability of a regression equation to predict the breast cap weight in the broiler industry based on the total weight of the broiler and specific visual carcass traits. Last but not least, a Focus Group took place, at the latest part of the study case, during which a group of poultry experts reacted positive to the approach, confirmed the results and contributed to the exploration of further continuation of the present research.

Key words: CRISP-DM methodology, Multiple Linear Regression, Regression Trees, Neural Networks, Fuzzy Inference Systems, Breast Cap Weight, Predictors, Visual Measurements, Broiler, Focus Group

Preface

This thesis is made as a completion of the Master's program in Business Information Systems at Eindhoven University of Technology (TU/e). Several people have contributed academically, practically and with support to the conduction of this master thesis and the fulfillment of my master studies. I would therefore firstly like to thank my head supervisor in the university, Maryam Razavian, for her time, valuable input, and support throughout the entire period of my master thesis. Her constructive feedback and utterly motivational words helped me to develop my research skills, realise the project, and overcome any setback I may have faced during the project.

Furthermore, I would like to thank my second supervisor, Anna Wilbik, for taking the time of her busy schedule to meet with me when I mostly needed it. Her valuable comments on every deliverable and her immense knowledge in data mining led me to greatly expand my knowledge in the field in order to make this project possible. Also, a lot of appreciation is due for Mykola Pechenizkiy who accepted to be my third commitee member.

This project would also not be possible if it weren't for the experts at Marel Stork Poultry Processing. I thank my company supervisor, Dirk Kuijpers, for his valuable support and guidance throughout the project. He was always able to find some room on his agenda to meet with me and evaluate my work. I am also grateful to Ronnie van der Wijst, Jan-Pieter Feddema, and the people of the INNOVA team for taking the time to offer their contribution to my research.

At this point, I would like to thank my dearest friends Iva, Nino, Pawel and Hugo for all the great moments we shared over the last couple of years. I also appreciate them for all the love they gave in the most difficult times. Moreover, I want to thank all my Greek friends in Eindhoven for making my stay an incomparable experience to remember for the rest my life. Special thanks to Michalis, Konstantinos, Natalia, Charalambos, Georgia and Maria who I cherish deeply along with the times we spent together away from our homeland. Last but not least, my joy knows no bounds when expressing my gratitude to my best friends in Greece.

This project is dedicated to my family Alia, Kostas and Anastasia, for constantly believing in me and giving me their undeniable love and support. Without them I would have never embarked on such a beautiful journey and none of these would be possible.

Antonios Mantzaris, December 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The demand of poultry meat, together with the supply, has been increased the last decades all over the world and will continue to increase [1], since it is healthier and less costly than its substitutes, such as red meat, fish meat, etc. As a result, experts find it harder to predict which of the end-products will sell in order to plan production and orders accordingly. Thus, inaccurate forecasts are increasing, and along with them the costs of these errors. The current project puts a small stone towards the solution of such problems in the poultry industry. This introductory chapter sets the context and presents the motivation for the project. Moreover, the problem of the thesis is carefully explained and the proposed solution is provided. After the scope is presented, the research questions are stated and the approach to be followed is summarized. The final part of this chapter gives the structure and an overview of the content of the remainder of the report.

1.1 Context

Marel is a leading global provider of advanced equipment, systems and services to the poultry, fish, meat and further processing industries [2]. Marel Stork Poultry Processing (MSPP) is a brand of the Marel group and its main focus is on the poultry industry. MSPP is providing an Information System to its customers, which is called Innova and it is a support system for managing and controlling every information that is derived from the poultry production processes [3].

A customer of MSPP is referred to as poultry processor. A poultry processor, using the means provided by MSPP, can run more than one lines of production, which are managed by operators, and each line is able to process 13,500 broilers per hour. One day old broilers are transported from hatcheries to farms where they are raised. After approximately a month has passed and once the broilers are ready to be slaughtered, they are picked up by the poultry processor and are fed to its production lines. The live supply enters the poultry processor at the primary process and goes through a chilling tunnel for a couple of hours before the secondary process delivers the end product [4]. The end products of a poultry processor are then distributed to industrial and retail customers (Figure 1.1).

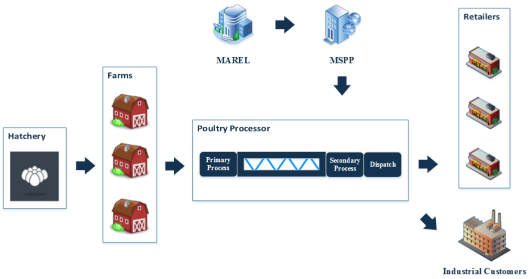


Figure 1.1: High-Level Poultry Process

In the beginning of every production day the poultry processor starts producing on prognosis and during the day retail orders are coming in, to which the poultry processor then adapts. The level of adaptation depends on the success rate of the prognosis [5].

1.2 Motivation and Problem Description

The broilers provided to the poultry processor, have an estimated weight and quality. During the primary process, the weight and quality of the broilers become more accurate due to clean measurements made in the process, and this accuracy is directly communicated to the production planning department. The operators, then, make the definitive decisions for the secondary process, at the latest, in the end of the chilling tunnel phase. The decisions are taken based on the raw material that the poultry processor has in stock in terms of weight. The operators calculate the average weight of a flock and assume that 20% to 25% of this weight corresponds to fillet weight based on experience. It is estimated that this method delivers a good amount of prediction accuracy (around 80%) but there is still place for improvement.

The accuracy achieved is not high enough in reality and may cause additional problems to occur because of it not reaching a satisfying level for the poultry processor to conduct its business in an optimal manner. The operators are usually taking decisions to produce with a chance of giveaway in order to be ensured that they will get to produce everything on time. The poultry processor mainly produces fixed batches, and the as-is method results in heavier or lighter end-products than the desired fixed weight, which in turn causes the poultry processor to sell them for less than they are actually worth, translating in loss for the poultry processor, or the opposite, acting adversely in customer satisfaction. In addition such a sub-optimal production planning may also lead to leftover of the end-product in the inventory of the poultry processor, which will need to be sold in the succeeding days.

During the chilling tunnel phase the operator shall have more precise information on the data that drives his whole production planning instead of taking decisions based on the estimated average weight of flocks. This method is not reliable enough since it prevents the operator from delivering more accurate production results and subsequently leads to considerable amounts of giveaway product as well as a mismatch between the supply fluctuation and the demand fluctuation. Increasing accuracy and offering timely prediction on the amount of fillet that each broiler contains will increase, in its turn, the efficiency in production and the accuracy in order fulfillment. This is possible by providing the operator with important knowledge on a highly valuable part of the broiler based on which he makes crucial production combinations and derives his final decisions.

The production planning process of the poultry processor guides the production process by creating different production plans, which focus on the product with the highest value, the fillet. Hence, the more accurate the prediction of the fillet weight in a broiler is, the more valuable information will be available to the operators to make the final production settings. However, the current research project will mainly deal with estimating the weight of the breast cap in a broiler rather than the fillet. Due to heavy human intervention in the filleting lines the data extracted from such lines are far less reliable than the data generated by an automated process (FIFO standard), that is the case in the breast cap cutting and weighing. Breast cap weight and fillet weight are highly correlated and an assumption is made that knowing the breast cap weight gives a valuable insight to the expected fillet weight.

Before providing the problem statement, it is crucial to understand the goals of the poultry processor. An interesting and descriptive approach to identify, present and better measure these goals is the GQM (Goal-Question-Metric) Methodology. GQM is a top-down approach which uses business goals to drive the identification of the right metrics and it consists of three levels [6]:

- Conceptual Level Goal defined for an object
- Operational Level Set of questions that characterizes the assessment or achievement of the goal

- Quantitative Level Set of metrics that is associated with the questions and it answers them in a measurable way

A graphic example of the GQM Methodology can be seen in the Figure 1.3.

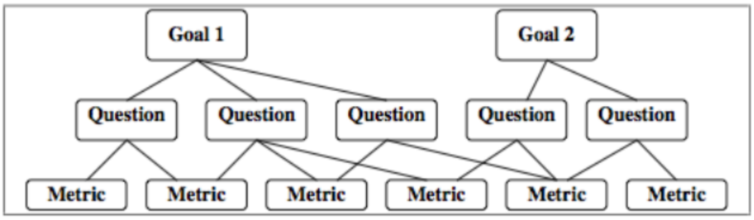


Figure 1.2: GQM Graph [7]

The application of the GQM methodology in the scope of the project can be seen in Figure 1.3. The most relevant goal to the project is the prediction accuracy, which is measured by the amount of error produced and directly or indirectly affects the remaining goals of the poultry processor. A detailed description of the metrics used to calculate accuracy is provided in the following chapter of the present thesis, whilst a list of the goals set by the poultry processor can be found in Appendix A along with their descriptions.

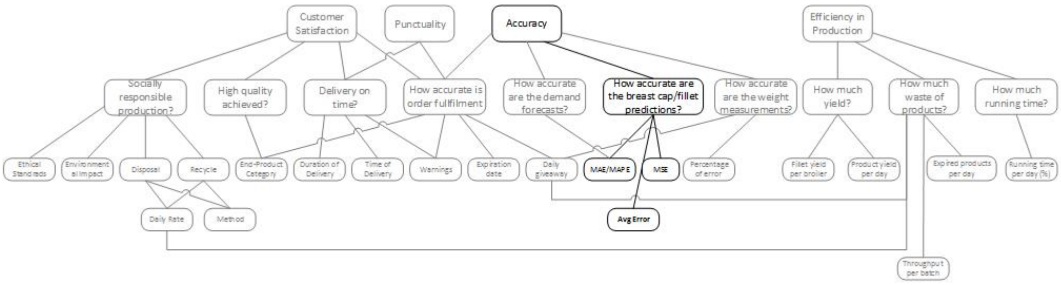


Figure 1.3: GQM on the Poultry Processor

Following the above observations, the problem addressed in this thesis can be summarized as follows:

Problem Statement:
The As-is method for calculating the amount of breast cap in a broiler is not an accurate enough approach to guide the operators in the planning department of the poultry processor towards taking optimal decisions in order to deliver the right amount of end-product at the right time.

1.3 Solution

The present graduation project will propose a comprehensive data mining approach which will provide the poultry processor with a way to extract accurate knowledge on the amount of breast cap a broiler holds using recorded specifications withdrawn from inline visual instruments and graders. This will presumably increase the efficiency of the production and will aid the operators to take crucial decisions in order to achieve as accurate as possible delivery of the right amount of end-product at the right time.

The goal of the project is to create a method to accurately predict the weight of a relevant broiler trait based on additional weight and visual measurements. The use of such an approach in practice will offer timeliness for the data extracted and the end-products while also considering the usefulness and usability of the knowledge produced for the people who will use it.

1.4 Scope

The scope of the current project is to identify methods and create models to reliably calculate the amount of breast cap weight in a broiler based on daily bird samples provided by the poultry processor, while also identifying patterns in a broilers body parts that affect the amount of its breast cap weight. The scope of this thesis project lies within the production planning department of the poultry processor as well as the development department of MSPP. Relevant data derived from previous projects conducted under the supervision of Marel are used as reference points and as a form of guidance throughout the course of the project.

1.5 Research Questions

In order to address the goals and purpose of the current project, several smaller scaled research questions were formulated, towards the main research question, serving as a step to step guide towards the solution of the main problem which was previously stated. To reach the desired results these questions were being processed and answered during the research. The research questions of the project are listed below:

Main Research Question

How can we offer an accurate estimation of the amount of breast cap in a broiler carcass?

Sub-Questions

RQ 1. What are the specifications of the current methods used for prediction purposes and what are their quality criteria?

RQ 2. What kind of data is available as input for the approach proposed?

RQ 3. What kind of methods/techniques shall be implemented in the available data?

RQ 4. How should the models created be evaluated?

RQ 5. How can we better predict the amount of breast cap in a broiler based on the available features?

RQ 6. How can the final model be integrated to the current operations?

1.6 Research Methodology

A data mining methodology is used for the completion of the project but initially a problem investigation is conducted. Thus, the first part of the research is getting acquainted with the company and the poultry production process, the theme of the project and the design science with a focus in requirements engineering. During this phase the problem statement is formulated followed by the research proposal including an indicative plan for the completion of the project.

With the completion of the first part, CRISP-DM (Cross Industry Standard Process for Data Mining) Methodology was followed. CRISP-DM is a data mining process model that describes commonly used approaches that expert data miners use to tackle business problems [8]. It borrowed ideas from the most important models created before 2000s and it is the groundwork for many later proposals. The CRISP-DM Special Interest Group (SIG) was set up with the aim of upgrading the CRISP-DM model to a new version better suited to the changes that have taken place in the business arena since the current version was formulated [9]. It is, currently, considered as the ”golden thread” that runs through almost every client engagement of this sort due to its powerful practicality, its flexibility and its usefulness when using analytics to solve business issues [10].

The CRISP-DM methodology is described into six phases to be carried out in a data mining project, as shown in Figure 1.4.

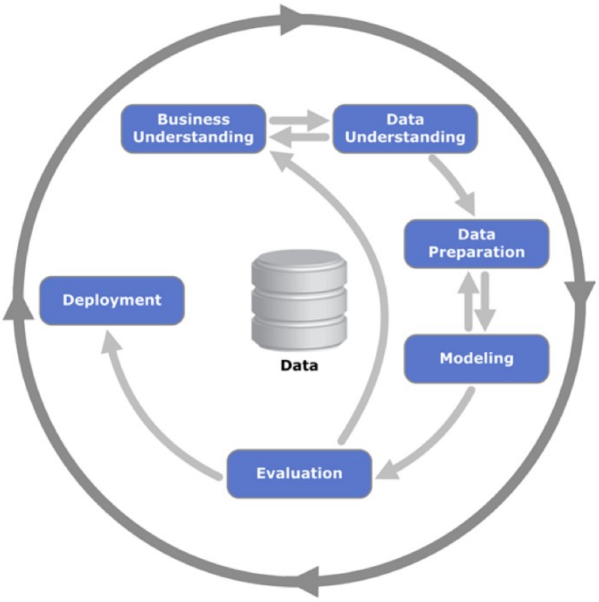


Figure 1.4: CRISP-DM Methodology[9]

Implementation details in each phase are given in the following description list.

Business Understanding
This initial phase includes orientation with the company environment and its production processes. Meetings with experts will take place in order to define the problem and to understand the project objectives and requirements from a business perspective. To carefully examine the most relevant goals to the project GQM methodology will be performed. The knowledge acquired in this phase will be converted into a data mining problem definition and a preliminary plan will be designed to achieve the list of objectives.

Data Understanding
The data understanding phase will begin with a visit to the poultry processor in order to learn more details about the production process and collect the initial data for analysis purposes. Afterwards, a close review of the data acquired will be conducted to enable familiarity with the data, identification of data quality problems, discovery of first insights into the data and possible detection of interesting subsets. Microsoft Excel will be the main tool used for that purpose due to it being simple and easy to configure.

Data Preparation
The data preparation phase will be implemented in Matlab environment. This phase covers all activities needed to construct the final dataset, as it will be fed into the models, from the initial raw data. Several problems between the relationships of different data sets and its variables will have to be overcome in order to merge all of the data into a single final set of data. Data will be transformed and cleaned in order to achieve the best outcome from running the models.

Modeling
In this phase, multiple linear regression will be applied with parameters calibrated accordingly to the input data set so that the output is the optimal one. The same procedure will be followed for the regression trees, neural networks and fuzzy inference systems models. This

will mean that some of the variables in the data set may need to be converted in different forms which will lead to a short revisit in the data preparation phase.

Evaluation

At this stage in the project, the models will be already built but before proceeding to the final deployment of the model, it is important to thoroughly evaluate the approach via available evaluation methods. 10-fold cross validation, mean squared error, mean absolute error and coefficient of determination will be used to choose the model instance that offers the best results. Reviewing the steps executed during model creation ensures that the business objectives are accomplished.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be presented in a way that the customer can use it (e.g. histograms, generated reports). A Focus Group shall also be prepared and organized internally in order to evaluate the high level goals of the end-result and to reach a final decision on the usability of the data mining results. Recommendations about further continuation of the research will be made in this phase as well. Last but not least, the last readjustments will be made, the documentation and presentation will be prepared and a final review on the entire project will be conducted.

1.7 Structure of the Report

For a systematic execution seven main phases of the project were identified: problem investigation, business understanding, data understanding, data preparation, modeling, evaluation and deployment. During each project phase there are activities which correspond to respective deliverables. Doing the literature study separately from the main phases of the research was considered to be not beneficial in such a data mining project so it was conducted in parallel. In this section it is explained how the work conducted during these phases is mapped in the report structure and under which chapter each research question is answered. A summarized version of the project outline can be found in Appendix B.

In this introductory chapter, the problem investigation and business understanding are presented while part of the first research question is also answered. Chapter 2 gives an extensive overview of literature in the field of poultry production, data mining and existing regression analyses. After the theoretical and research background, chapter 3 introduces the data mining modeling environment created for the purpose of the project while also answering the second and third research questions. In chapter 4, the training and evaluation of the models is performed and the first research question is being revisited as the fourth and fifth research question is also answered. The final chapter corresponds to the deployment of the model evaluating it in terms of usability and the answer to the last research question is provided. In the same chapter conclusions, limitations and future research are also presented.

Chapter 2

Background

Literature analysis enables to explore previous studies in the topic of research and integrate them on the project while familiarizing the reader with the area of research. To better understand the purpose of the current research, an introduction to the basics of poultry processing is vital. Moreover, the main concepts of data mining shall be addressed since the output of the research will emerge from the appropriate use of several data mining techniques. The chapter ends with a review on the available research on regression analysis in the context of poultry processing as a reference to what has already been done and which methods have been proven more effective in similar cases. Furthermore, the first Appendices present additional information from the literature study that could be considered interesting to the reader.

2.1 Poultry Processing

Poultry processing is a complex combination of biology, chemistry, engineering, marketing, and economics. While producing food for humans is the main goal of poultry processing, other equally essential goals include waste management, livestock feeds, and non-food uses of poultry. When considering the European marketplace, poultry products can range from a slaughtered carcass to a further processed product such as a chicken nugget or chicken sausage. An overview of the large poultry product range along with the different product characteristics is presented in Appendix C.

2.1.1 Modern Poultry Production Processes

Commercial poultry is extremely uniform in appearance and composition. Tightly managed breeding, incubation, rearing, and nutritional regimes have created a bird that is a virtual copy of its siblings. This uniformity has allowed poultry processing plants to develop into highly automated facilities with an efficiency that is unmatched by other livestock processors [11]. With line speeds of 13.500 broilers per hour, uniformity, automation, and efficiency are recurring themes and have been keys to the success of poultry processing.

2.1.1.1 Primary Process

The poultry production process starts from the live bird supply where trucks loaded with broiler crates arrive to the plant. Trucks are weighed together with the incoming supply to track weight information and the broilers are left in the air-conditioned garage for a two-hour rest phase to calm down. The room is lit solely with blue light, which cannot be seen by the broilers [12]. The unloading of the containers is a fully automated process where broilers slide onto a wide, slow-moving trampoline conveyor belt.Optical detectors monitor the containers to ensure that they are completely emptied. The emptied containers are cleaned and disinfected in a subsequent washing line and are prepared for the next transport. After the containers have been unloaded the trucks are cleaned and disinfected as well.

A particular concern of the modern poultry processors is to ensure that the slaughtering process will remain as stress-free as possible for the broilers. Thus, the animals are gently stunned via electricity or gas method, before the actual slaughtering process begins [13]. The stunning process in general provides a stress-free environment not only for the animals but also for the employees, which also helps in the minimization of dust development. At the end of the stunning tunnel is the so-called carousel. There the chickens are hung from an endless chain by their feet. Each broiler is set to a fixed position before a precise cut is performed to the neck of the hanged body.

After the broilers have been slaughtered, they are left for a short period of time to bleed out in order to pass through the scalders [14] and facilitate the subsequent feather plucking process, which is fully-automated. Onwards the head is removed completely and the hole created by the cut is trimmed. A final internal and external wash is necessary to remove possible debris, blood and fat clots [15].

It is necessary that the broilers go through inspections, before they are rendered ”ready to cook” [16], veterinarians and controllers of the local authority. Initially, the broilers are fed into a camera system, which compares the quality of the animals with certain quality standards of the poultry processing. The camera system is only able to determine quality in terms of a visual specimen of the outer body of a chicken. This phase includes checking the broilers for damage caused by the scalding and plucking, such as broken bones or colour differences and other defects on the skin. During that phase, an official controller also examines the breast side of the chicken.

Henceforth, the birds go through the machines that harvest their giblet and then through another inspection before entering the chilling tunnel. This time the official controller checks the carcass of the chicken, also from behind, along with the corresponding packet of giblets rendering the broilers suitable for the remaining processing. The giblets are then separated mechanically and the heart, liver and stomach is further processed or sold as a whole to foreign markets.

2.1.1.2 Intermediate Process

When the primary process has ended the broiler is measured once more for clean weight. The gutted chickens are hung from an overhead conveyor on chilling hooks. Before they pass into the chilling tunnel, an employee checks the carcass again to ensure it is in perfect condition. Chilling is a maturation phase during which the microbial growth is reduced and safety and duration on the end product is reassured [17]. In the chilling tunnel the chickens are chilled to below 2°C within 3 hours and it is considered as the intermediate process of the poultry processor’s production line, as it calms the meat from the primary process and prepares it for the secondary process.

2.1.1.3 Secondary Process

During the secondary process, only a small number of the slaughtered chickens leave the company as whole birds. Most of them are fed into the cut up process which is mostly automated, as a result of the uniform anatomy and technical developments done in the industry. Most of the meat is prepared for the self-service counters in the retail market.

Wing, breast, fillet, leg and thigh products are cut at various consecutive modules and are placed on conveyor belts. It takes between 6 to 8 minutes to cut up a chicken into the various pieces. The wing and leg pieces are weighed, portioned and manually placed in trays. In the case of breast caps and fillets, this work is done by the ”robot-batchers”, that portion the pieces in trays with gram-precision. The breast part of the chicken is cut and then distributed in different filleting lines depending on its weight measurement, where operators manually extract the fillet from the breast cap and place it on a conveyor belt towards the fillet grader and the robot-batcher [12].

Some of the products go through a marinating process before being packed. The marinating and further processing of the chicken along with the cut up process take place in chilled rooms. Maintenance of the cold chain as well as process and personal hygiene are decisive for perfect microbiology and therefore for the storage quality of the end products. Apart from hygiene clothing, the poultry processor also provides all operators with thermal clothing. Regular breaks, rotation at the workplaces and the ergonomic design of the activities are a few important factors that play a critical role in establishing optimum working conditions towards delivering maximum yield [12].

Transporting the trays from the cutting through to the packaging area is a fully-automated process. The temperature in the packaging area is 2°C, to ensure continuous and constantly cold chain. The trays are sealed with protective wrap and are then weighed and labelled. The batch number is printed on the label of the end product to ensure traceability. The sealed trays are packed into transport boxes, which are then stacked and moved by a robot onto a transport pallet. In the dispatch process, which marks the end of the secondary process, the goods are picked for the individual customers and are then loaded onto refrigerated trucks. The dispatch area and the interior of the truck are chilled to around 0°C. Each day, more than 500,000 packs leave the poultry processor plant to reach the refrigerated cabinets of the retail outlets on the next day [12].

A schematic view of the processes that take place in the production line of a modern poultry processor is shown in Figure 2.1.

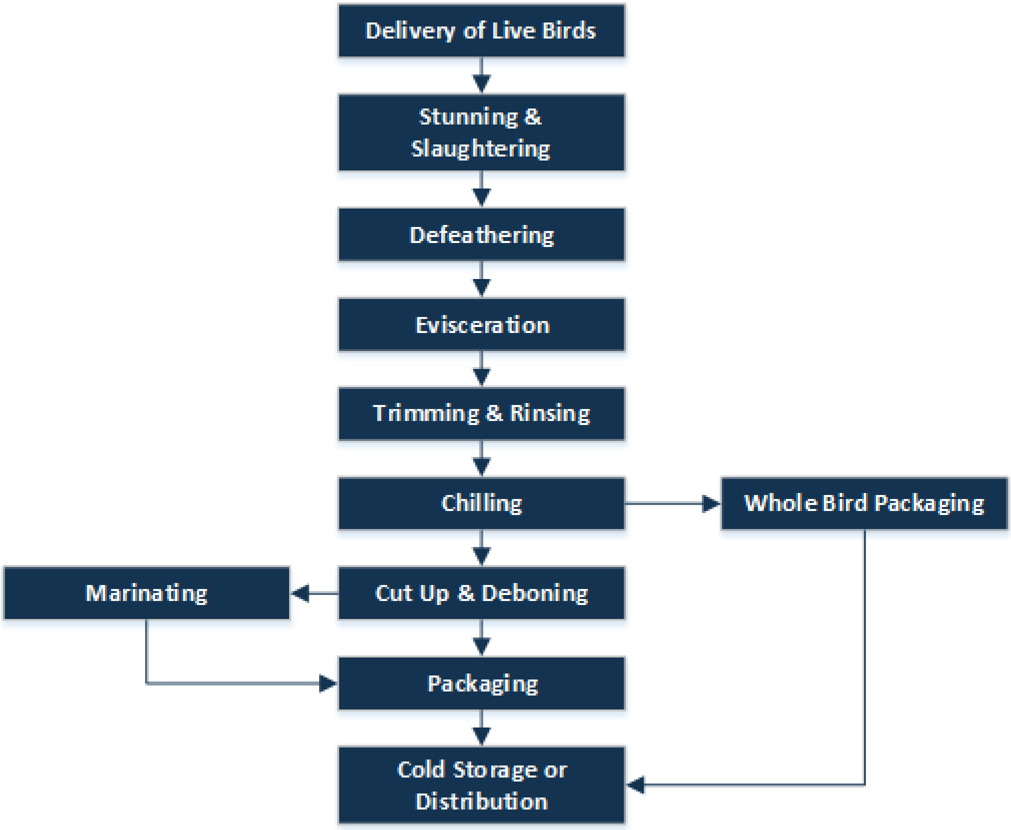


Figure 2.1: Poultry Processing [18]

A more detailed figure of the poultry production processes can be seen in Appendix D.

2.1.2 Production Planning Process

The production processes are guided by the planning department of the poultry processor by the creation of different production plans, as listed below:

- The rough plan is created on a weekly basis, it contains the forecasted amount of weight and the number of broilers that is needed for a whole production week. It is focused on the product with the highest value, the fillet, whereas in some occasions a rough plan is made for the thigh meat as well. Weather conditions, special events, history of the customer and general experience are crucial factors to decide on a final rough plan.
- The short plan is an estimation of the demands for retail and industry orders in kilograms for 2 or 3 days in advance
- The execution plan is constructed on a daily basis and it provides the information to start the production of the day.

Moreover, the production planning department has to deal with different type of orders coming in:

- Sales Orders - Retail customers (mainly two big supermarket chains) place their orders in the sales department and the orders are then emailed to the head of the planning department of the poultry processor in order to receive a confirmation. When the order is confirmed it is imported in the sales order list.
- Industry Orders - The same procedure as above is also followed for the industrial orders.
- External Bulk Orders - Bulk orders that arrive in the poultry processor from external sources. The same procedure as above is also followed for such type of orders.
- Internal Bulk Orders - Orders to fulfill the processing and complete the other order processes.
- Promotion Orders - Orders placed more than a month in advance, depending on the size of the promotion they represent. Promotion orders constitute a weekly order but they are translated on daily orders to facilitate the production process.
- Production Orders - Orders that are actually produced in a production day. As soon as real orders are coming in, the production orders are being updated.

Once all production orders are prepared, the poultry processor starts producing about 80% of the forecasted orders. By the end of the afternoon all the actual orders are received and depending on the performance of the forecast the remaining percentage of the orders is adapted. The trucks start leaving the plant about an hour after the last real order has arrived.

In many poultry processors large quantities of broilers are processed at each production day. At the same time they have to reach the highest level of order fulfillment, especially for the retail market, while managing the large amount of product flow, the costs, the giveaway and the production efficiency. All the aforementioned factors are the reason for the complexity in the production planning [19]. This complexity together with the lack of the right information at the right time hinders the production planning to perform optimally.

2.1.3 Relevance to the research

The equipment attached in the processes of a poultry processor is generating a massive amount of information that if used appropriately can give a whole new insight on several key factors that could improve the way it is operating. The information rendered significant for the remaining of the current research is the data generated from the information systems that are linked to the equipment of the poultry processor, the Innova IS and the IRIS system. To be able to extract knowledge from such data, it is important to also be familiar and have a clear understanding of the processes related to the data. In this project, the production process related to the data is the secondary process which comes right after the intermediate process, also known as the chilling tunnel, with a focus on the breast cap cut up lines and filleting lines. As it can be seen in the Figure 2.2, the broiler after exiting the chilling tunnel it goes through the IRIS system and is photographed from the front and back side while its weight is sent to the system and certain areas of the bird's body are calculated along with the coordinations of key points on its body. Then the broiler goes through a send device where its weight is sent to and after a few seconds its breast cap is getting cut and graded in a fully automatic process. The weight information of the breast cap of each bird is documented and the breast caps go to various filleting lines according to their weights through conveyor belts. There the operators are manually harvesting the fillets from the breast caps while also trimming residual parts and place the ready fillets onto conveyor belts towards their grading and batching.

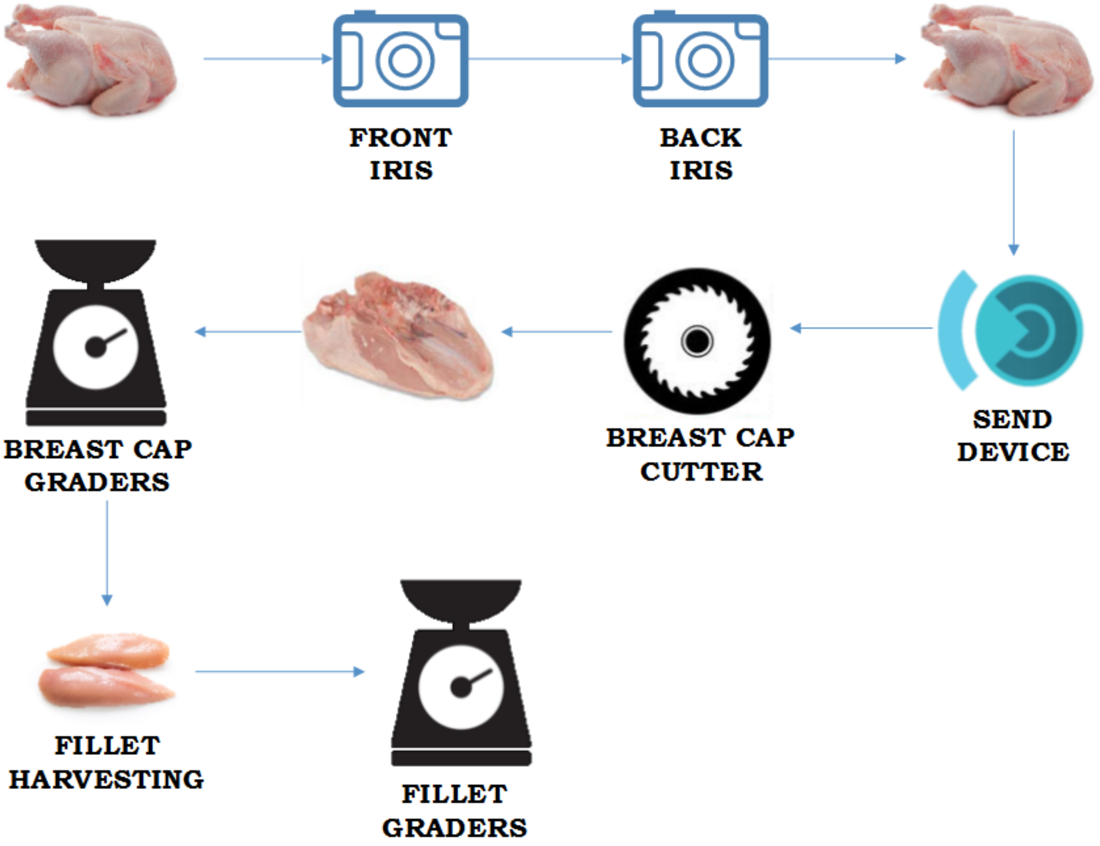


Figure 2.2: Process line used for data acquisition

The data collected by this process shall be used to extract knowledge and identify patterns between the different attributes of the broilers. This exploitation can definitely be performed with the application of data mining techniques.

2.2 Data Mining

Most people across the world are just now discovering the power of data mining [20]. There are numerous occasions where companies are required to analyze large amounts of data with the aim of rapid decision making. The same case applies in the current research project where data mining is used to facilitate this data analysis and to achieve adequate results [21]. Definitions and additional details that are interesting about data mining can be found in Appendix E.

2.2.1 Basics of data mining

Data mining is a pipeline containing many phases during which the initial data sets are collected, carefully examined, prepared and fed to predictive models. The models are prepared based on the finalized data set and are then validated before they can be deployed.

In this project, the initial data sets are event logs generated from different Information Systems. Data mining techniques, tools, and methods are going to be used to discover, monitor and improve real processes in the poultry processor by extracting knowledge from these event logs. An event log is a list of all activities performed in a process. In a poultry processor, information aboutthe broiler weight is recorded in such logs. These log files contain so called traces, in which activities are grouped per broiler. Besides activities, also the flock ID and timestamps of these activities are recorded. An example of an event log with multiple activities is displayed in the following Table. Each activity consists of one event.

Product ID	Timestamp	Process Info	Weight	Areas ...	Coords ...
8965628954364084225	09:48:16	0	1430	8685	267,151
8965549892607934489	09:48:17	3	0	113	0
8965628954364084426	09:48:18	0	1666	9264	285,166
8965628954364084879	09:48:18	4	1351	0	0

Table 2.1: Event Log

In the simplest case, data mining approaches find patterns or models in a single data table, while in some multi-relational data mining approaches, like in the case of the current research project, patterns or models are found from data stored in multiple tables or data sets.

A data mining process, in general, regardless of the area of application is comprised by the following main activities[20]:

- Business task: clarification of the business question behind the problem
- Data: provision and processing of the required data
- Modeling: analysis of the data
- Evaluation and validation: conducted during the analysis stage
- Application of data mining results and learning from the experience

A more detailed data mining process has been chosen for the completion of this project, which is called CRISP-DM methodology and has been already presented in detail through the introductory chapter as the project's research methodology. This methodology was selected as it is by far the most commonly used methodology by the industry data miners, according to a poll on a popular data mining website called KDNuggets, and it is considered the "de facto standard for developing data mining and knowledge discovery projects." [22] by data mining experts. Moreover, it is a straightforward approach including a stepwise guide that fits well to the current project while there is also detailed and easy to follow documentation which covers every aspect of importance. The modeling techniques that will be effectively implemented with the use of the aforementioned guide are explained in the next sections.

2.2.2 Regression Analysis

Regression is a statistical method that is used to asses the relationship between variables and to predict scores on a variable given scores on other variables.

Before embarking into regression, a small introduction in correlation is imperative since it is also mildly used in the current project. Correlation is a measure of association between two variables. Two variables are said to correlate if a change in one of them is accompanied by a predictable change in the other. The concept of correlation is commonly encountered in a range of techniques used in business forecasting and modelling. Correlation describes the association in a way that allows the researcher to easily interpret the strength of the association. Correlation is, in essence, standardized covariance and it is defined as the covariance divided by the standard deviations of a each variable [23].

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Dividing the standard deviations serves the purpose of rescaling the statistic so that the maximum and minimum values are always 1.0 and -1.0, respectively. Thus, a correlation of 0.6 means the same thing in terms of strength, regardless of the standard deviation of the variables.

There are many different types of correlation equations but for the purpose of the current project the focus is on the most commonly used one, the Pearson Product Moment Correlation, which summarizes the strength and direction of the association between two variables in a single number, the correlation coefficient, otherwise known as the Pearson coefficient or R score [24]. A positive coefficient means that the relationship is positive, while a negative one means that the relationship is negative. A zero correlation indicates no relationship between the two variables. The strength of the relationship between the variables is indicated by how close the coefficient is to +1 or -1 [25].

The common aspect of correlation and regression analysis is that both processes can estimate relationships among different variables. However, the researcher indulging in regression usually wants to find out the causal effect of one variable upon another. To explore such issues, data on the variables of interest is assembled and regression techniques are applied to estimate the effect of the causal variables upon the variable that they influence [26]. Regression analysis includes several techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable (target variable) and one or more independent variables (predictors). The most relevant ones with respect to the present research project are explained in detail below.

2.2.2.1 Multiple Linear Regression

The general purpose of the multiple linear regression is to learn more about the relationship between several independent variables and a dependent variable [27]. However, in order to thoroughly understand the concept of multiple linear regression, it is finer to firstly introduce the simple linear regression which is the simplest form of linear regression.

Simple linear regression can handle only one independent variable X and a dependent variable Y, which is expressed as a linear function of X [28]. The value y_i of variable Y, for every value x_i of variable X, is given by the following equation:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Parameters α and β express the linear dependence of the dependent variable Y from the independent variable X in the best way possible. Every pair of values α,β determines a different and separate linear relationship geometrically expressed by a straight line and the parameters of the equation are defined as follows:

- The constant term α has the value of y for x = 0.
- The factor β of x is the slope of the straight line, in other words the regression coefficient. It expresses the change of variable Y when variable X changes by one.
- The random variable ϵ_i represents the regression error and is defined as the difference between the actual value of Y and the predicted value of Y.

In case the dependent variable Y is linearly depended by more than one independent variables X(X_1,X_2,X_3,...,X_n) then we refer to the multiple linear regression, which is essentially several simple linear regressions combined. The equation that defines the relationship between these independent variables X and the dependent variable Y has the general form of:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni} + \epsilon_i$$

It can be observed that the equation of the multiple linear regression follows the pattern of the simple linear regression equation. Multiple linear regression, though, relies upon certain assumptions about the variables used. The most important assumptions to be satisfied are the following [29]:

Variables are normally distributed

Regression assumes that all the variables have normal distributions. Variables that don't meet this assumption can distort relationships and significance tests.

Linear relationship between the independent variables and the dependent variable

Multiple linear regression can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature. In such cases where the relationship between the independent variables and the dependent variable is not linear, the results of the regression analysis will under-estimate the true relationship and this underestimation will increase the chances of error appearance.

Variables are measured with no error

Many variables, that researchers are interested in, are also difficult to measure, making measurement error a significant concern. In simple regression models, unreliable measurements cause relationships to be under-estimated increasing the risk of errors.

No auto-correlation

Auto-correlation is present when consequent errors in a time series are correlated with each other. In that case estimates produced by the regression model are rendered inaccurate.

No or little multicollinearity

Independent variables shall be independent from each other in regression approaches. Given two or more of the independent variables correlate with each other, only one of them shall be retained in the final list of predictors while the rest shall be omitted. Moderate multicollinearity may not be problematic. However, severe multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model [30].

Homoscedasticity

The variance of the errors shall be the same across all levels of the independent variables. When this is not the case heteroscedasticity is indicated and it may lead to serious distortion of findings weakening the analysis thus increasing the possibility of errors.

The satisfaction of these assumptions usually means the reduction of the final data set by a certain percentage. As can be seen in Appendix F, a method that can further reduce the final data set while keeping only the truly relevant predictors is feature selection and it can be implemented in multiple linear regression through stepwise regression modeling among other ways. However, if any of the aforementioned assumptions is violated then the results yielded by the regression model may not be trustworthy resulting in inefficient or even misleading forecasts and scientific insights.

2.2.2.2 Regression Trees

A faster and easier method to interpret is the one of the decision trees [31]. A decision tree is a decision support methodology that uses a tree-like graph or a model of decisions and their possible consequences to make sense of a set of data. A decision tree for numerical data is known as a regression tree, and a decision tree for categorical data is known as a classification tree. Since this thesis uses numerical data from an industrial process, the decision tree analysis followed is the regression tree analysis.

A Regression tree is built through a process known as binary recursive partitioning. This is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch. Splitting in regression trees is made in accordance with the squared residuals minimization algorithm which implies that the expected sum variances for two resulting nodes should be minimized. This splitting rule is then applied to each of the new branches [32]. The process continues until the maximum tree is constructed, which means that splitting was made up to the last observations in the training set. As the maximum tree may turn out to be considerably big, in the case of regression trees, pruning may be required in order to remove insignificant nodes.

2.2.2.3 Neural Networks

Apart from the above regression techniques, a more general method of regression is also implemented in the current research project called neural network analysis. A neural network constitutes an architecture of a number of interconnected nodes, called neurons. Each neuron is characterized by inputs and outputs, and it locally implements a simple calculation. Every connection between two neurons is assigned with a weight value, which represents the knowledge that is stored in the network and determines its functionality. The output of each neuron is determined by its type, the interconnection with the rest of the neurons and other potential external inputs. Beyond a given operating capacity given by its initial design, the network usually develops an overall decent functionality through its learning [33]. The overall functionality of the neural network is mainly determined by:

- the network topology
- the neuron characteristics
- the training method
- the training data

The neural network architecture is described by the number of layers it consists of and its connections between the neurons. The most common type of a neural network architecture consists of a layer of "input" neurons connected to a layer of hidden neurons, which is then connected to a layer of output neurons, and it is called Multilayer Perceptron [34].

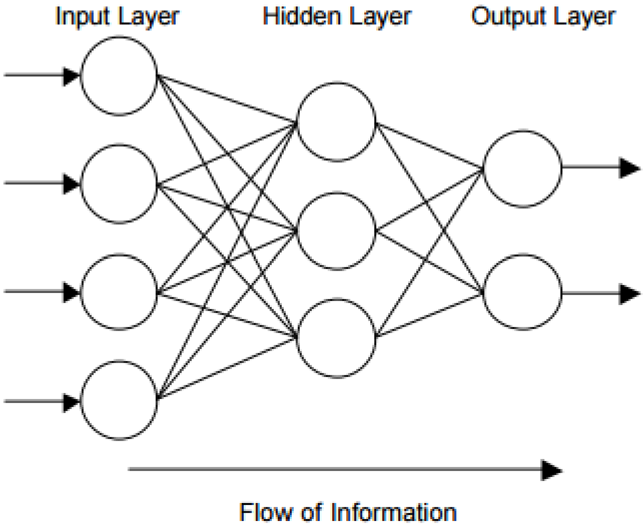


Figure 2.3: Simple Neural Network Architecture [34]

- The input neurons represent the data that is fed into the network.
- Each hidden neuron is determined by the activities of the input neurons and the weights on the connections between the input and the hidden neurons.
- The behaviour of the output neurons depends on the activity of the hidden neurons and the weights between the hidden and output neurons.

This simple type of network is interesting because it is self-organized since the hidden neurons are free to make their own representations of the input. The weights between the input and hidden neurons determine when each hidden neuron is active, and so by modifying these weights, a hidden neuron can choose what it represents. Furthermore, neural networks, with their ability to derive meaning from complicated or imprecise data, can be used to extract patterns that are too complex to be noticed by either humans or other data mining techniques.

2.2.3 Fuzzy Inference Systems

Except for the regression techniques another type of model, called fuzzy inference systems, is also used in the present research. This method allows to handle data vagueness in an intuitive and natural manner [35], mainly by the use of fuzzy rules. A fuzzy rule is a way of knowledge representation, which mimics the human way of thinking. The fuzzy sets of the dataset are combined with each other to create fuzzy rules that represent the knowledge that we have for the system. A common fuzzy rule is of the following type:

If x is A then y is B

The If part is referred to as the antecedents while the Then part is the consequents. A and B are the fuzzy sets and the degrees of membership in their elements is described by membership functions. The value x is the value of a predictor, which has been inducted into fuzzyfication, and y is the output of the system, which is provided by the inference system in a fuzzy form. Next, the fuzzy result is inducted into defuzzyfication and a crisp value is produced [36].

The fuzzy rule explained above expresses a rule whose output is a fuzzy set and it is a mamdani fuzzy rule in honor of Ebrahim Mamdani who was the first one to implement fuzzy logic [37]. A fuzzy rule though can have a lot of different form which are all extended versions of the simple mamdani fuzzy rule. A Mamdani type fuzzy inference system can be defined as a set of rules, given the input variables x = [x_1, ... , x_n] ∈ X and output variable y ∈ Y. More types of fuzzy inference systems are available but the most common ones are the mamdani and the sugeno types which are quite similar. Their primary difference is that in the Sugeno FIS there is no output membership function [38]. Instead the output is a crisp number computed by multiplying each input by a constant and then adding up the results.

If x is A then y is f(x),

where f(x) = ax + c.

2.2.4 Evaluation Methods

Evaluating a regression model substantially means deciding whether the relationships between variables are acceptable as descriptions of the data. The evaluation process can involve several analysis methods examining different aspects of a model.

2.2.4.1 Cross validation

Cross validation is a method which estimates the accuracy of a model by showing how its results will generalize to an independent data set. In most real life applications, there is a limited amount of data available, which leads to the idea of splitting the data. In cross validation, you decide on k folds, or partitions, of the data. In the case where $k = 10$, which is the number usually yielding the best error estimate, the data is split into ten approximately equal parts and nine tenths of the data is used for training while the remaining one-tenth for testing purposes. The procedure is repeated ten times so that in the end, every instance has been used exactly once for testing. This is called ten-fold cross validation. For every k, the error estimate shall be computed as well and the instance with the lowest error estimate is the most accurate one. Moreover, when the model has been estimated over some, but not all, of the available data, then the model using the estimated parameters can be also used to predict the data that were not used yet. In that case, computing the mean squared error of the data that were out of the sample and the mean squared error of the data that were in the sample will give a sign of the level of deficiency in the model based on the difference of the errors [39].

2.2.4.2 Coefficient of Determination

The coefficient of determination, denoted R^2, is a number ranging from 0 to 1 that indicates how well data fit a regression model in a straight line or curve. Its main purpose is the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model [40]. To compute the R-squared we also need to compute three sums of squares as explained below.

Let a data set have n values marked y_1,y_2...y_n (denoted as y_i), each associated with a predicted value f_1,f_2...f_n (denoted as f_i).

If $\overline{y}$ is the mean of the observed data:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured using the following formulas:

The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \overline{y})^2$$

The regression sum of squares, also called the explained sum of squares:

$$SS_{\text{reg}} = \sum_i (f_i - \overline{y})^2$$

The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

2.2.4.3 Mean squared error

The evaluation method of the mean squared error is the principal and most commonly used measure to validate the success of a numeric prediction as explained below [41].

Let the predicted values on the test instances be f_1, f_2 ... f_n ($f_i$refers to the numerical value of the prediction for the i_th test instance) and the actual values be y_1, y_2 ... y_n then the mean squared error is computed by

$$MeanSquaredError = \frac{(f_1 - y_1)^2 + ... + (f_n - y_n)^2}{n}$$

Taking the squared root of the mean squared error yields the root-mean-square error (RMSE), which gives more weight to the larger but infrequent errors.

2.2.4.4 Mean Absolute Error

Another metric commonly used to evaluate forecast performance for a time series with n elements is the mean absolute error (MAE) [42]. MAE is obtained by taking the absolute value of all bias values, and then taking the mean as shown in the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|$$

, where f_i is the prediction and y_i the true value.

In essence, MAE is the average of the absolute errors $|e_i| = |f_i - y_i|$. An advantage of the MAE is that it is easier to be interpreted since it is unambiguous and gives more insight about the average magnitude of the errors over an entire dataset without the effect of cancelling positive and negative errors that might occur using another metric, such as the mean error [43].

2.2.5 Relevance to the research

Every technique that is presented in this section is also implemented in the current research project. Each of the models trained is evaluated for data fitting while also the prediction accuracy delivered is measured. The mean squared error and the mean absolute error produced by the models are the key metrics that declare the best model with the lowest error and highest prediction accuracy for the present research along with the mean error. The mean error takes the sign of the error into account and shows how far the average plain error from zero. A comparison of MAE with RMSE will show the consistency of the error size. In case the difference is small then the large errors are infrequent and thus have low magnitude.

2.3 Review of Existing Analyses

Research has already been made regarding the weight estimation of the body composition of broilers. It is noticed that researchers undertaking such an investigation deploy interesting methodologies and techniques, and end up with interesting conclusions. The most relevant analyses are briefly presented in this section and facilitate the realisation of the current research.

Raji et al [44] investigated on the estimation of broiler carcass components weight and yield with sex being the fixed factor. The research team used 192 broilers separated in males and females which were processed and cut into component parts (breast, thigh, etc.) to measure the weight of their carcass and their individual body parts respectively. The data obtained was subjected into multiple linear regression and the best regression model for each dependent variable was determined based on the coefficient of determination (R^2). Their approach derived regression equations to estimate breast, thigh and fat weights from non invasive body measurements of the broilers. In addition, Pearson's correlation coefficient was used to determine the simple correlation between live weight of the bird, its body measurements and the target carcass components weight in grams and yield in percentage. The most relevant result of the analysis to the project showed that both male and female broilers have moderate to high and significant relationships between breast weight and live weight.

Another team of researchers [45] measured and weighed 179 broilers which were bred with different diets to reach certain weights. Then the chickens were scanned for total body electrical conductivity (TOBEC) before they were euthanized and frozen for further measurements. Again multiple linear regression was used with five independent variables in order to estimate the body weight from the body measurements already conducted. The analysis showed that pelvis width and body weight had the best R^2 value while adding also the breast width and the circumference of the chicken to the equation can further increase the R^2 improving the quality of fit.

Melo and his team [46] conducted physical and ultrasound measurements to broilers and derived prediction equations for estimating breast and abdominal fat weight and its proportions relative to live weight. Their best prediction models were determined using the stepwise procedure from Statistical Analysis Systems Institute [47]. The best prediction model for breast weight estimation was the simple regression of live weight. However, when the regression models did not include live weight the regression coefficients of physical and ultrasound measurements were mostly significant. Correlation coefficients were the highest for breast length and breast weight while breast width showed the lowest relation to its weight. Last but not least, the predicted values obtained by simple regression equation including live weight were not the ones showing the highest correlation coefficient with the observed values. These results suggest that live weight shall not be the only independent variable to be used in the prediction of breast weight.

Another research group from Belgium [48] performed in vivo measurements to twenty-four chickens and graded their breast meat weight. Correlations and coefficients of regression between breast meat weight and the various traits were estimated and the best models were determined using the stepwise multiple regression analysis [47]. In order to explain the variation of breast meat weight the models were also compared using R-squared. The results showed no difference between the sexes of the broilers and that thoracic circumference, live weight, keel length and thickness of the muscle were valuable indicators for estimating breast meat weight while keel angle and chest width appear to be poor predictors.

Olawumi [49] used broilers from a specific breed to perform his research using Pearson Correlation analysis. His results showed that carcass weight has significant positive phenotypic correlation with breast muscle weight at 0,921. He also proved Musa et al. findings [50], who observed significant positive association between breast muscle weight and leg muscle weight.

Munari et al [51] conducted different measurements on broilers and used them to perform the least-squares method analysis following the GLM procedure of Statistical Analysis Systems Institute [47]. The results presentation was provided in a correlation matrix and shows a correlation of 0.85 between breast fillet weight and breast meat weight while also high correlation was observed between the weight of the latter one with the chilled carcass weight (0,89).

The section concludes with the findings from an internal to Marel research project conducted by Saglibene [52]. The researcher investigated on the efficiency of the filleting lines in a poultry processor and to do that he personally handled and weighed 100 chickens to serve as input in his analysis. The data from the measurements was provided and when further examinatipn was conducted it showed that, on average, 58% of the breast cap weight corresponds to fillet meat weight.

2.4 Conclusion

Over the past years, several papers on the estimation of different broiler body parts based on several types of measurements have been published. This suggests that the research on the topic was and still is relevant, and as more research is done the more the relationship between different characteristics and measurements of a broiler is becoming clearer. As can be seen in the penultimate section of this chapter, a lot of interesting approaches were introduced in this particular field of research. From measuring the relationship of the thoracic circumference and the live whole weight of a broiler to examining the role of different sexes in such analyses, it seems that there are still many different ways to better determine the association of all the different measurements that can be performed to a broiler's body.

There is no apparent research found in the literature containing measurements from visual specimens of broilers and thus estimation of the relationship between the combination of different body areas of a broiler, the distances between key coordination points in a photographed broiler body and the total chilled carcass weight of a broiler with its breast cap weight. These new aspects of measurement may bring a whole new incentive that could lead to a series of more interesting approaches that could give new insights in the estimation of different body parts of a broiler solely through a visual specimen.

Chapter 3

Data Processing

In this chapter every task followed to derive the final dataset, which is used in the current research project, is thoroughly described. The tasks are allocated to two main phases, data understanding and data preparation, according to the aspects they are related to.

3.1 Data Understanding

Data understanding phase describes the activities that were performed starting from the collection of the data until the first essential insights into it. In this phase, the format of the data is raw, thus Microsoft Excel is used in order to make sense out of it and reach a clear understanding.

3.1.1 Data Collection

Figure 3.1 illustrates, in red circles, the points in the production line where data is recorded.

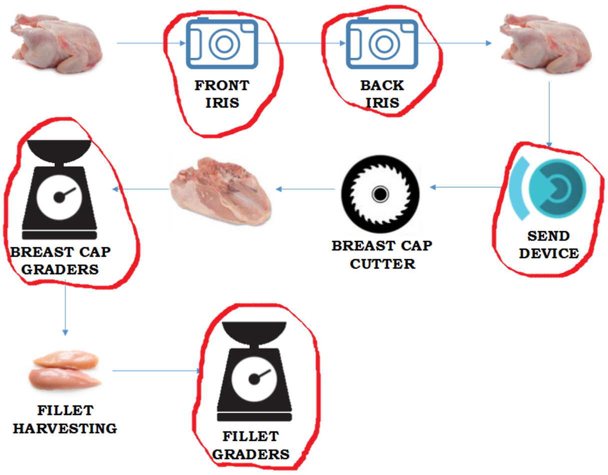


Figure 3.1: Process line used for data acquisition

3.1.1.1 Initial Dataset

The poultry processor uses Innova IS, which is able to log information generated by the equipment in the plant, as well as the IRIS system, which generates information on the areas of a broiler based on a visual specimen along with key coordination points on the image and collects the weight information of each broiler carcass as a whole. Subsequently, the initial data set consists of two parts, regarding the systems used to collect the data.

1. Weight measurements on the amount of breast cap in a broiler which are recorded by the in-line breast cap grader and take place right after the breast cap is cut.
2. Measurements on the whole weight of the broiler carcass which are recorded by a smart weigher and take place before the broiler enters the chilling tunnel. This weight data is fed in the IRIS system which calculates body areas and coordinations of certain key points on the photographed body of a broiler. The IRIS system is located at the beginning of the cut up line after the chilling tunnel.

The total weight of the broiler carcass is also fed in the send device which is placed just seconds before the actual cutting of the breast cap and it is used as a reference point in time between the IRIS data and the breast cap weight data. This system is placed near the breast cap cutter and at the same time it has a straightforward relationship with the IRIS system due to similar weight measurements aiding in the understanding of the relationships between the breast cap grader and the IRIS system.

3.1.1.2 Method

In order to conduct the actual collection of the data, a visit to the poultry processor plant was required for a production day. During the visit, a real time observation of the process is made possible, the time difference between key measurements are calculated and the relevant equipment, based on the requirements of the final dataset, is identified. Then, both the Innova IS, which handles the send device and the breast gap grader, and the IRIS system is accessed in order to start the data tracking process. The time window used for the logging of the data was 4 hours, which was considered enough time to collect a considerable amount of information for data understanding purposes. The visit to the poultry processor plant showed that the collection of the data for the current research can also be performed remotely since both Innova and IRIS can be accessed through virtual systems in order to manually enable and disable the logging.

3.1.2 Data Description

3.1.2.1 General Information

The raw data acquired from the send device and the graders through Innova is in a text format in which information of a measurement is represented in a line and every different attribute is separated by a symbol. IRIS systems are in the form of a database object which is then converted into a comma separated value file. The first logging of the data includes 8 different equipments recording several thousands of information each.

In the first recording, the log files of all four filleting lines in the poultry processor are also recorded. The real time observation of the filleting lines showed that they are not automated and that there is a heavy amount of human intervention. Breast caps are manually harvested for fillet by operators and at random times breast caps are removed, thrown away or added at free will, which mixes the sequence of the measurements and damages the reliability of the data. However, the data already recorded is manually examined and fed into Microsoft Excel for further analysis. The findings derived from the analysis confirmed the observations acquired in the poultry processor as unexplainable patterns are observed and it is made impossible to keep track of the fillet measurements and its timestamps in accordance with the breast cap, the send device or the IRIS measurements. Thus the data from the filleting lines is rendered as a subject for future research and non suitable for the current research project due to its unreliability. Further recordings of the log files followed with the omission of the filleting lines. Approximately 30.000 broilers are processed in the duration of recording session. A recording session usually begins at 10 am in the morning and ends at 4 pm in the afternoon. Details from one of such sessions are presented below.

- 31.873 log lines were extracted from the slave of the infeed breast cap grader.
- 45.126 log lines were extracted from the send device before the breast cap cutter.

- 47.397 log lines were extracted from the Back IRIS System in the Cut Up Line.
- 46.287 log lines were extracted from the Front IRIS System in the Cut Up Line.

It is interesting to observe the difference in the number of log lines. The reason for such a case is the fact that the breast cap grader records only measurement that have a weight amount above 0 as opposed to the IRIS and send device systems which also record measurements from empty shackles which result in rows with zero values. There is also a random possibility that the weight measurement of a breast cap is unusually high which indicates that two breast caps were weighed as one and thus corresponding to two broiler measurements in the other systems. Last but not least, small differences in the number of log lines may occur because the activation and deactivation of the recording in the systems is done manually in a sequential manner leading to more recordings for the first and less recordings for the last system to be enabled.

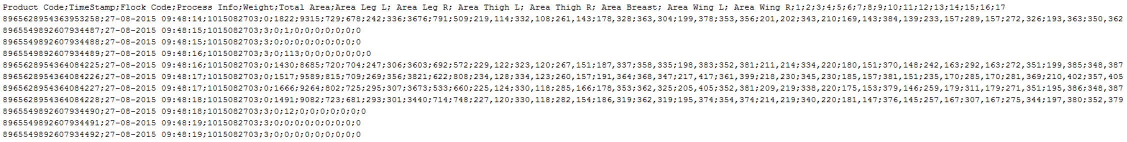
3.1.2.2 Data Attributes

The data attributes from the different measurement systems are presented in Table 3.1.

Breast cap grader	Fillet grader	IRIS system	Send device
Timestamp	Timestamp	Product Code	Timestamp
Grader Piece	Libra Weight	Timestamp	Product Code
Breast Cap Weight	Serial Number	Flock Code	Weight of the whole bird
Unit	Weight	Process Info	
Si Weight	Unit	Weight of whole bird	
Weight Quality	Piece Type	Total Area	
Weight Count	Output	Area of Left/Right Leg	
Output	Status	Area of Left/Right Thigh	
Length	Height	Area of Breast	
Batch ID	Length	Area of Left/Right Wing	
Status	Width	Coordinates of points (1 to 16 or 18)	
Material Number	Material Number		
	Key Activity		
	Batch ID		
	Run		
	Algorithm		
	Estimated Weight Accept Status		
	Accept Status		

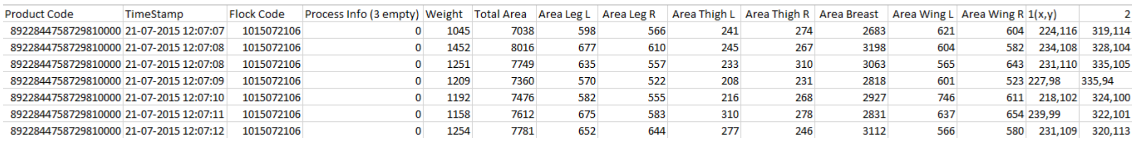
Table 3.1: Data Attributes

The reference values for every attribute can be seen on Table 3.2, 3.3, and 3.4. Samples of the raw format of the data (Figure 3.2, 3.3, and 3.4) are also provided at this part for easier realization of the actual information contained in the files.



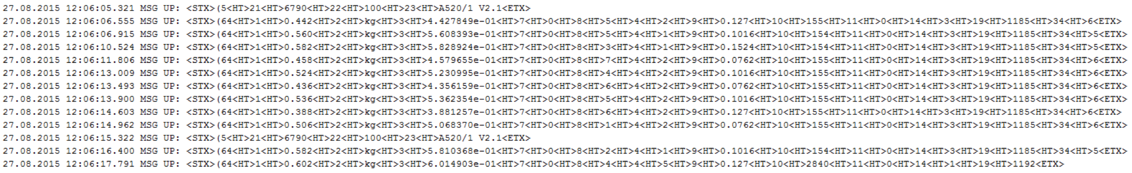
```
Product Code;TimeStamp;Flock Code;Process Info;Weight;Total Area;Area Leg L; Area Leg R; Area Thigh L; Area Thigh R; Area Breast; Area Wing L; Area Wing R;1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17
896562895436395325 8;27-08-2015 09:48:14;1015082703;0;1822;9315;729;678;242;336;3676;791;509;219,114;332,108;261,143;178,328;363,304;199,378;353,356;201,202;343,210;169,143;384,139;233,157;289,157;272,326;193,363;350,362
896554989260793448 7;27-08-2015 09:48:15;1015082703;3;0;1;0;0;0;0;0;0;0;0
896554989260793448 8;27-08-2015 09:48:15;1015082703;3;0;0;0;0;0;0;0;0;0;0
896554989260793448 9;27-08-2015 09:48:16;1015082703;3;0;113;0;0;0;0;0;0;0;0
896562895436408422 5;27-08-2015 09:48:16;1015082703;0;1430;8685;720;704;247;306;3603;692;572;229,122;323,120;267,151;187,337;358,335;198,383;352,381;211,214;334,220;180,151;370,148;242,163;292,163;272,351;199,385;348,387
896562895436408422 6;27-08-2015 09:48:17;1015082703;0;1517;9589;815;709;269;356;3821;622;808;234,128;334,123;260,157;191,364;368,347;217,417;361,399;218,230;345,230;185,157;381,151;235,170;285,170;281,369;210,402;357,405
896562895436408422 7;27-08-2015 09:48:17;1015082703;0;1666;9264;802;725;295;307;3673;533;660;225,124;330,118;285,166;178,353;362,325;205,405;352,381;209,219;338,220;175,153;379,146;259,179;311,179;271,351;195,386;348,387
896562895436408422 8;27-08-2015 09:48:18;1015082703;0;1491;9082;723;681;293;301;3440;714;748;227,120;330,118;282,154;186,319;362,319;195,374;354,374;214,219;340,220;181,147;376,145;257,167;307,167;275,344;197,380;352,379
896554989260793449 0;27-08-2015 09:48:18;1015082703;3;0;12;0;0;0;0;0;0;0;0
896554989260793449 1;27-08-2015 09:48:19;1015082703;3;0;0;0;0;0;0;0;0;0;0
896554989260793449 2;27-08-2015 09:48:19;1015082703;3;0;0;0;0;0;0;0;0;0;0
```

Figure 3.2: IRIS dataset (sample)



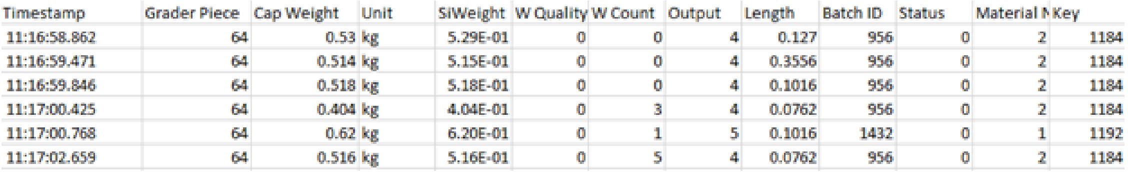
Product Code	TimeStamp	Flock Code	Process Info (3 empty)	Weight	Total Area	Area Leg L	Area Leg R	Area Thigh L	Area Thigh R	Area Breast	Area Wing L	Area Wing R	1(x,y)	2
8922844758729810000	21-07-2015 12:07:07	1015072106	0	1045	7038	598	566	241	274	2683	621	604	224,116	319,114
8922844758729810000	21-07-2015 12:07:08	1015072106	0	1452	8016	677	610	245	267	3198	604	582	234,108	328,104
8922844758729810000	21-07-2015 12:07:08	1015072106	0	1251	7749	635	557	233	310	3063	565	643	231,110	335,105
8922844758729810000	21-07-2015 12:07:09	1015072106	0	1209	7360	570	522	208	231	2818	601	523	227,98	335,94
8922844758729810000	21-07-2015 12:07:10	1015072106	0	1192	7476	582	555	216	268	2927	746	611	218,102	324,100
8922844758729810000	21-07-2015 12:07:11	1015072106	0	1158	7612	675	583	310	278	2831	637	654	239,99	322,101
8922844758729810000	21-07-2015 12:07:12	1015072106	0	1254	7781	652	644	277	246	3112	566	580	231,109	320,113

Table 3.2: IRIS data (Excel sheet)



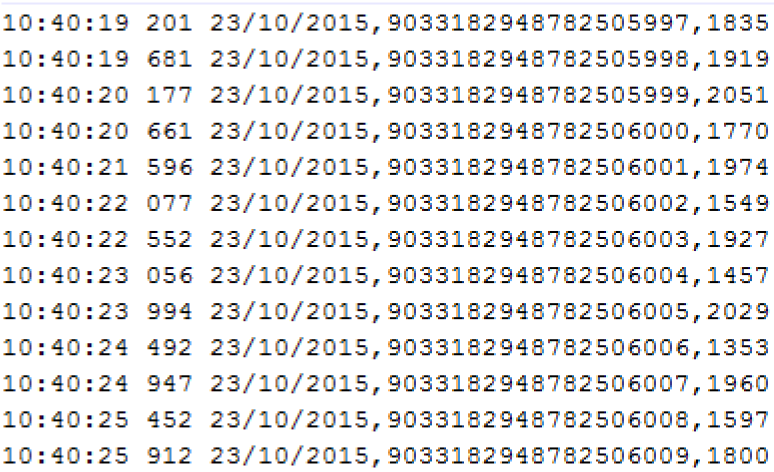
```
27.08.2015 12:06:05.321 MSG UP: <STX>(5<HT>21<HT>6790<HT>22<HT>100<HT>23<HT>A520/1 V2.1<ETX>
27.08.2015 12:06:06.555 MSG UP: <STX>(64<HT>1<HT>0.442<HT>2<HT>kg<HT>3<HT>4.427849e-01<HT>7<HT>0<HT>8<HT>5<HT>4<HT>2<HT>9<HT>0.127<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:06.915 MSG UP: <STX>(64<HT>1<HT>0.560<HT>2<HT>kg<HT>3<HT>5.608393e-01<HT>7<HT>0<HT>8<HT>5<HT>4<HT>1<HT>9<HT>0.1016<HT>10<HT>154<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>5<ETX>
27.08.2015 12:06:10.524 MSG UP: <STX>(64<HT>1<HT>0.582<HT>2<HT>kg<HT>3<HT>5.828924e-01<HT>7<HT>0<HT>8<HT>3<HT>4<HT>1<HT>9<HT>0.1524<HT>10<HT>154<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>5<ETX>
27.08.2015 12:06:11.806 MSG UP: <STX>(64<HT>1<HT>0.458<HT>2<HT>kg<HT>3<HT>4.579655e-01<HT>7<HT>0<HT>8<HT>7<HT>4<HT>2<HT>9<HT>0.0762<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:13.009 MSG UP: <STX>(64<HT>1<HT>0.524<HT>2<HT>kg<HT>3<HT>5.230995e-01<HT>7<HT>0<HT>8<HT>4<HT>4<HT>2<HT>9<HT>0.1016<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:13.493 MSG UP: <STX>(64<HT>1<HT>0.436<HT>2<HT>kg<HT>3<HT>4.356159e-01<HT>7<HT>0<HT>8<HT>6<HT>4<HT>2<HT>9<HT>0.0762<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:13.900 MSG UP: <STX>(64<HT>1<HT>0.536<HT>2<HT>kg<HT>3<HT>5.362354e-01<HT>7<HT>0<HT>8<HT>5<HT>4<HT>2<HT>9<HT>0.1016<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:14.603 MSG UP: <STX>(64<HT>1<HT>0.388<HT>2<HT>kg<HT>3<HT>3.881257e-01<HT>7<HT>0<HT>8<HT>6<HT>4<HT>2<HT>9<HT>0.127<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:14.962 MSG UP: <STX>(64<HT>1<HT>0.506<HT>2<HT>kg<HT>3<HT>5.068370e-01<HT>7<HT>0<HT>8<HT>1<HT>4<HT>2<HT>9<HT>0.0762<HT>10<HT>155<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>6<ETX>
27.08.2015 12:06:15.322 MSG UP: <STX>(5<HT>21<HT>6790<HT>22<HT>100<HT>23<HT>A520/1 V2.1<ETX>
27.08.2015 12:06:16.400 MSG UP: <STX>(64<HT>1<HT>0.582<HT>2<HT>kg<HT>3<HT>5.810368e-01<HT>7<HT>0<HT>8<HT>2<HT>4<HT>1<HT>9<HT>0.1016<HT>10<HT>154<HT>11<HT>0<HT>14<HT>3<HT>19<HT>1185<HT>34<HT>5<ETX>
27.08.2015 12:06:17.791 MSG UP: <STX>(64<HT>1<HT>0.602<HT>2<HT>kg<HT>3<HT>6.014903e-01<HT>7<HT>0<HT>8<HT>4<HT>4<HT>5<HT>9<HT>0.127<HT>10<HT>2840<HT>11<HT>0<HT>14<HT>1<HT>19<HT>1192<ETX>
```

Figure 3.3: Breast cap grader log file (sample)



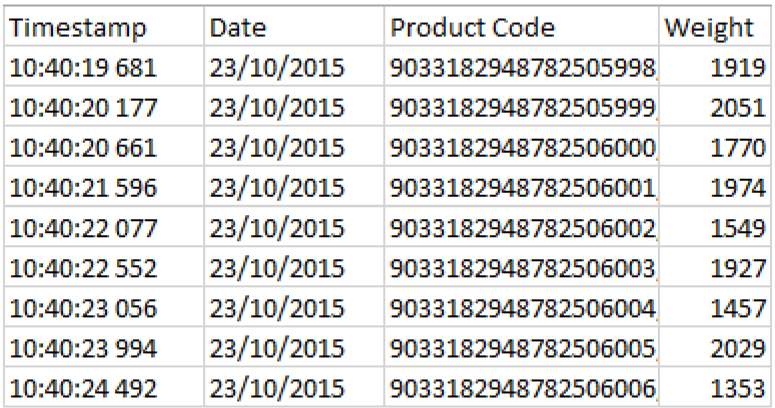
Timestamp	Grader Piece	Cap Weight	Unit	SiWeight	W Quality	W Count	Output	Length	Batch ID	Status	Material N	Key
11:16:58.862	64	0.53	kg	5.29E-01	0	0	4	0.127	956	0	2	1184
11:16:59.471	64	0.514	kg	5.15E-01	0	0	4	0.3556	956	0	2	1184
11:16:59.846	64	0.518	kg	5.18E-01	0	0	4	0.1016	956	0	2	1184
11:17:00.425	64	0.404	kg	4.04E-01	0	3	4	0.0762	956	0	2	1184
11:17:00.768	64	0.62	kg	6.20E-01	0	1	5	0.1016	1432	0	1	1192
11:17:02.659	64	0.516	kg	5.16E-01	0	5	4	0.0762	956	0	2	1184

Table 3.3: Breast cap data (Excel sheet)



```
10:40:19 201 23/10/2015,9033182948782505997,1835
10:40:19 681 23/10/2015,9033182948782505998,1919
10:40:20 177 23/10/2015,9033182948782505999,2051
10:40:20 661 23/10/2015,9033182948782506000,1770
10:40:21 596 23/10/2015,9033182948782506001,1974
10:40:22 077 23/10/2015,9033182948782506002,1549
10:40:22 552 23/10/2015,9033182948782506003,1927
10:40:23 056 23/10/2015,9033182948782506004,1457
10:40:23 994 23/10/2015,9033182948782506005,2029
10:40:24 492 23/10/2015,9033182948782506006,1353
10:40:24 947 23/10/2015,9033182948782506007,1960
10:40:25 452 23/10/2015,9033182948782506008,1597
10:40:25 912 23/10/2015,9033182948782506009,1800
```

Figure 3.4: Send device log file (sample)



Timestamp	Date	Product Code	Weight
10:40:19 681	23/10/2015	9033182948782505998	1919
10:40:20 177	23/10/2015	9033182948782505999	2051
10:40:20 661	23/10/2015	9033182948782506000	1770
10:40:21 596	23/10/2015	9033182948782506001	1974
10:40:22 077	23/10/2015	9033182948782506002	1549
10:40:22 552	23/10/2015	9033182948782506003	1927
10:40:23 056	23/10/2015	9033182948782506004	1457
10:40:23 994	23/10/2015	9033182948782506005	2029
10:40:24 492	23/10/2015	9033182948782506006	1353

Table 3.4: Send device data (Excel sheet)

The two images provided below show the visual specimen generated by the IRIS system for the front and the back part of the broiler respectively. The coordination points are shown in the image, with green and white color. The green color indicates the points that are software generated and thus are less reliable, while the white ones are always going to be on certain corners of the broilers body image.

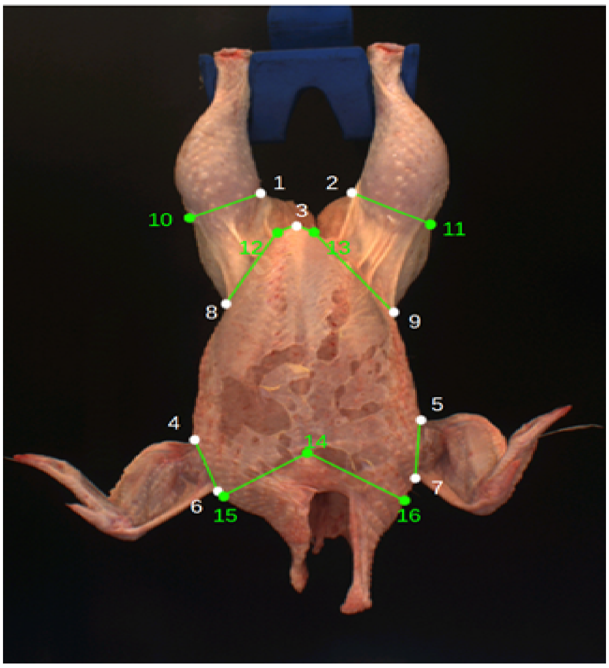


Figure 3.5: Image of the front side

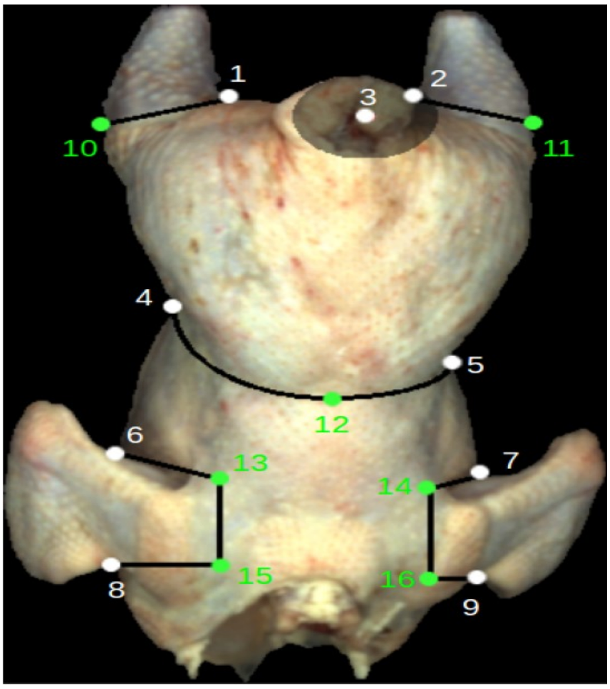


Figure 3.6: Image of the back side

3.1.2.3 Requirements Satisfaction

The data acquired are not all valuable to the scope of the project for satisfying the data mining goals that are already set. Extensive discussions and interviews with specialists in the company it is decided that the only relevant data that serves the purpose of the project for the prediction of the breast cap amount in a broiler is the following:

- Timestamp: The timestamp attribute generated helps towards the creation of a single dataset that contains the most important attributes from all the individual datasets. This attribute works as a reference point for each measurement and plays a defining role in achieving a link between the different systems since we know the time difference between each different measurement. Timestamps are also used to observe interesting potential patterns in the data collected. Once the link between the different datasets is achieved and the pattern recognition phase is completed, the timestamps are excluded from the final part of analysis.
- Weight Information: The total weight of the broiler carcass shall be considered as a key predictor to gain an insight on the breast cap weight of a broiler.
- Distance between coordination points: The distance between different points in a photographed broiler can show dependencies and specifications that would give additional knowledge about the capacity of a broiler without cutting it up. Thus, looking at the lengths and widths of different body parts may add valuable information in the relationship with the breast cap of a broiler, as seen in the literature study [48].
- Ratios: Since lengths and widths are considered valuable indicators for estimating breast cap amount in a broiler, it is only logical that their corresponding ratios or even width/length ration combinations may be relevant as well.
- Areas: The amount of two-dimensional space taken up by specific areas of a broiler carcass may give insight to which parts of the bird relate to its breast cap quantity.

3.1.2.4 Quality of the data

Initial findings show that data quality is high enough to proceed in further processing towards meeting the business needs. Data in each set is complete, with the exception of a small percentage of zero-values for data that is missing. Every dataset is available whenever it is required and is free from errors. Moreover, there is consistency among the data and all the definitions are understandable and potentially applicable to various present and future analyses [53].

3.1.3 Data exploration

In this section, data is explored and visualized in graphs generated by Excel. The first finding is that the IRIS data of the front part and the IRIS data of the back part of the broiler are identical in terms of weight and the time difference between them is 6-7 seconds. The relationship between the IRIS data and the send device is almost identical as can be seen in Figure 3.7.

There are some weight values that are no present in both the IRIS and the send device. In occurrences where the birds are rehanged in the production line or there is an error during the weight measurement these weight values are automatically replaced by extreme values for tracking purposes which the send device is configured to ignore. Other than that the measurements seem to follow a stable pattern and the identical relationship is easy to observe. The server time difference noted between the different systems is 4 minutes and 8-9 seconds with the send device ahead of IRIS. The real time difference in the process line is 1 minute and 28 seconds with the IRIS ahead of the send device.

During the production day, the operators in the poultry processor would take a 30 minute break for lunch and after 2 hours a 20 minute break for coffee causing the production to stop for the duration of the breaks. These production breaks can also be observed in the time gaps of the above scatter plot. The same pattern is followed in the breast caps as well.

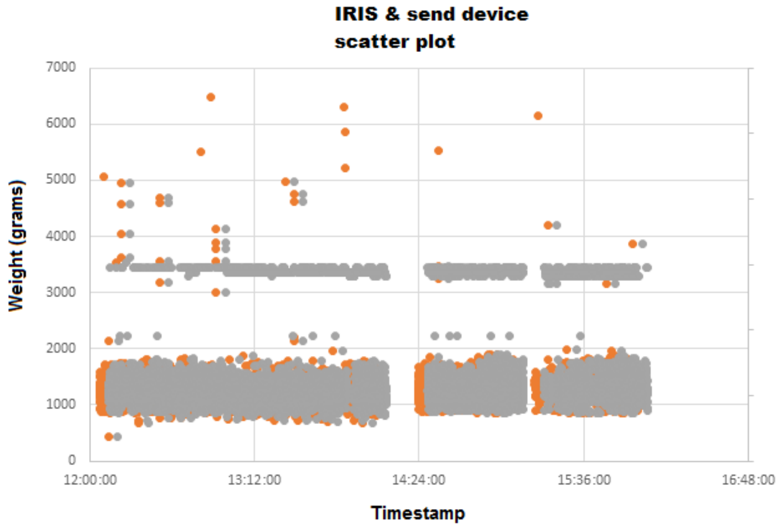


Figure 3.7: Scatter plot of IRIS (grey points) and Send device (orange points) data

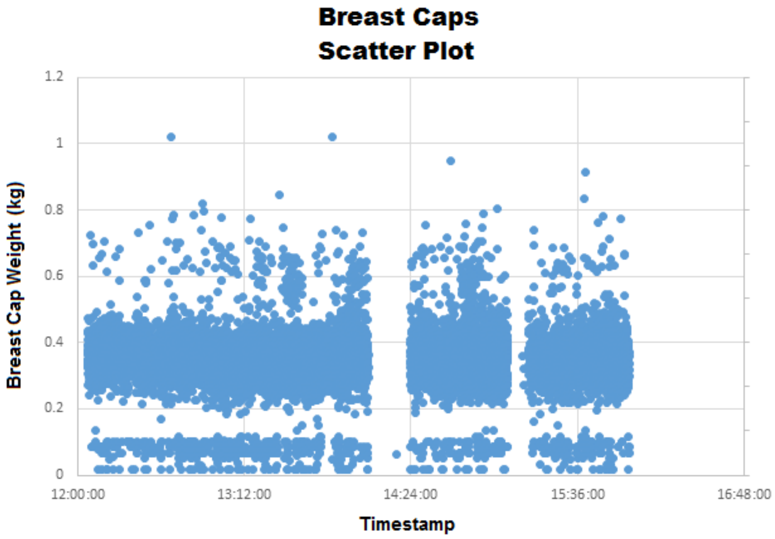


Figure 3.8: Scatter plot of breast cap data

Figure 3.8 shows more noise in the data. The extremely high values represent the merged measurements while the extremely low values represent small parts of meat or fat being detached from the breast cap during its fall in the conveyor belt and being graded separately. To examine better this type of noise a histogram was created as well. The breast cap data shall follow an almost perfect normal distribution which is the case in Figure 3.9.

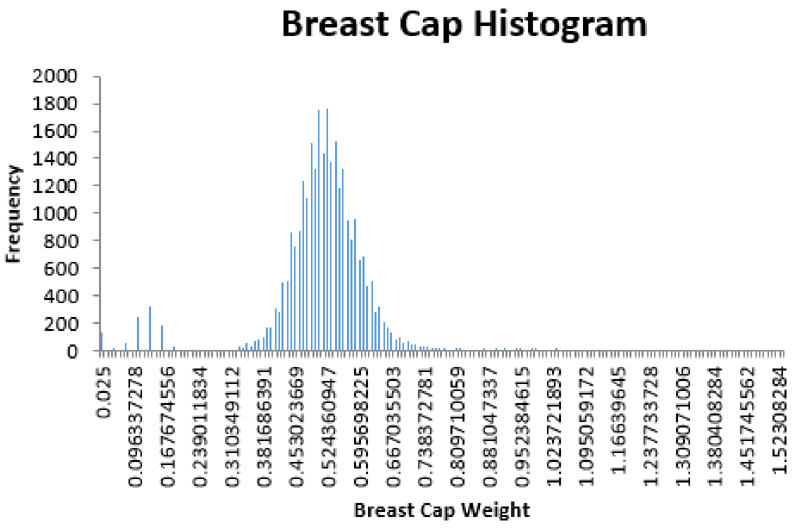


Figure 3.9: Histogram of breast cap weights

The histogram shows the frequency of the extremely low and high values which are away from the normally distributed values. Send device, IRIS and adjusted breast cap data are all represented in Figure 3.10 in a scatter plot where a time relationship is made apparent as well as a stability in the general pattern of the data.

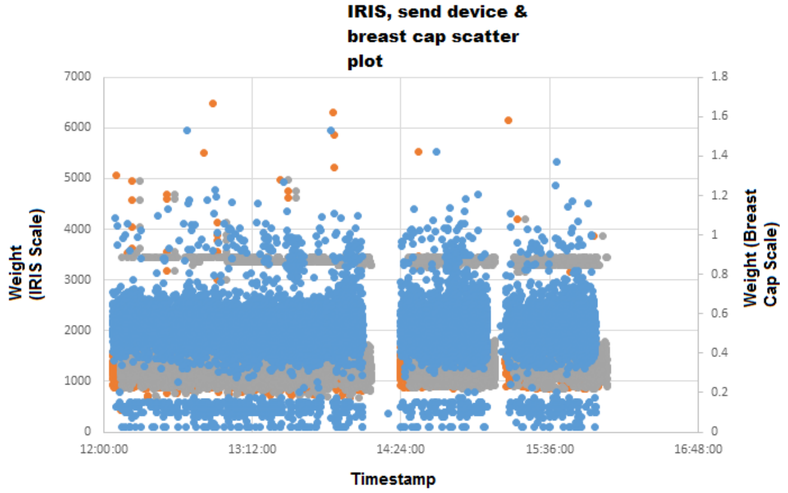


Figure 3.10: Total scatter plot

All the above observations raise doubts regarding the end-quality of the data, in terms of reliability, and are addressed and worked out in the next phase of the project, data preparation.

Another interesting finding in the data is the different broiler flocks used in the production line, as seen in Figure 3.11. Each flock supposedly includes a different group of birds destined for particular cut up lines. Knowing the weight specifications of a flock of birds is of great importance to the operators in the poultry processor because, subsequently, they have the appropriate information to decide what shall be produced by each flock.

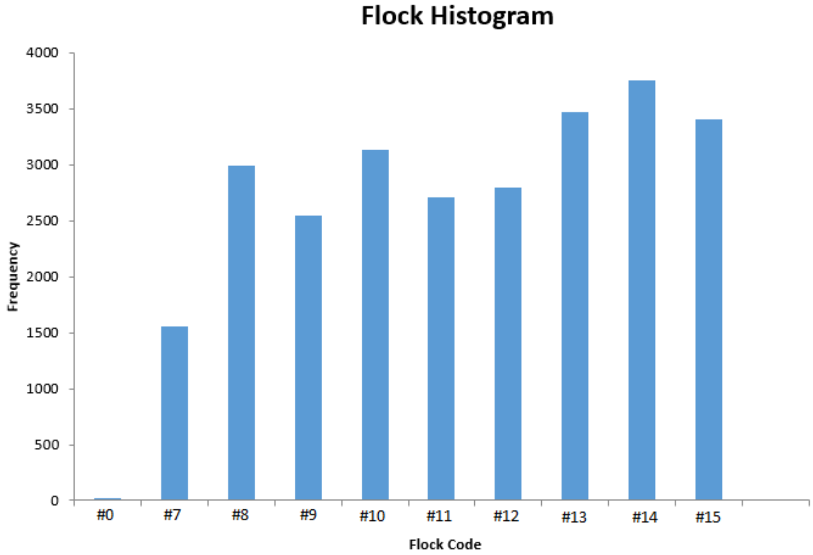


Figure 3.11: Histogram of broiler flocks

Figure 3.12 shows the scatter plot of the flocks with the timestamp which leads to the realisation that the flocks are fed in the production line sequentially with no flock interruptions noticed.

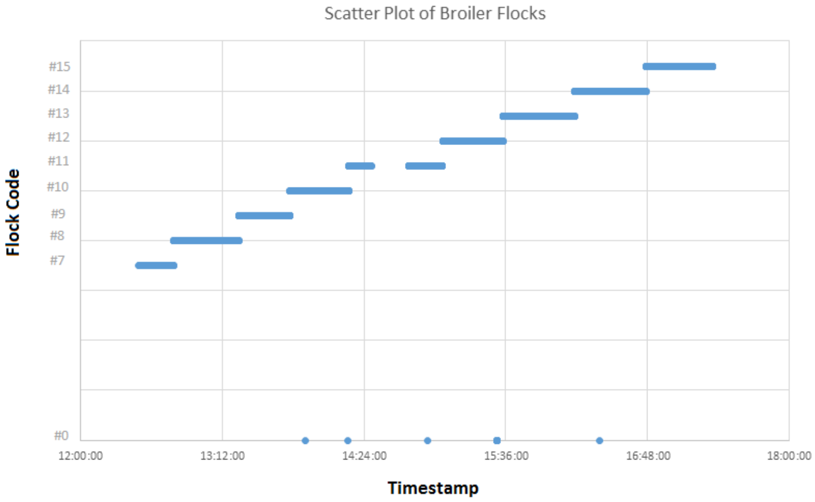


Figure 3.12: Scatter plot of broiler flocks

Therefore there is also the capacity to obtain predictive results in a flock level.

3.2 Data Preparation

In this section the steps taken towards deriving the final dataset are described along with the data that will be used for modeling purposes.

3.2.1 Format Data

The format of the raw data is not suitable to be fed in the models. Therefore during this step the initial data files are prepared for use in the consequent parts. Formatting tasks are performed to each type of log file with some minor differentiations.

Common actions performed

- The data of each log file is imported to tables where the delimiter of the attributes is the comma for the send device, the semi-colon for the IRIS data and <STX>, <HT>, and <ETX>for the breast cap grader.
- The numerical values of every attribute in the table is read as text in Matlab. Thus a reformat from text values to numerical values is applied.
- A name is given to each column according to the variable that its values represent.
- The timestamp, in the first column of each table, is converted into seconds, for easier handling throughout the remaining of the preparation, in the following manner:

$$timestamp = hours * 3600 + minutes * 60 + seconds$$

Breast cap grader log data

The weight of the breast caps, in the second column of the table, is converted into grams from kilograms by multiplying each row with 1000 so that it is in the same metric as in the other data sets.

IRIS system log data

The raw data of the IRIS comma separated file is of the same form for both the front and the back datasets. The headlines included in the first row of the raw data (Figure 3.1) are removed and the coordination points, which are of the form x,y, are separated into their x and y numeric parts in the table in order to handle them easier in the remaining of the data preparation phase.

3.2.2 Data Cleaning

After formatting the data, data cleaning is conducted in order to filter potential missing values, outliers, zero values, noise, etc., and handle them appropriately. As already mentioned in the data understanding, during a recording session the breast cap grader generates around 31.000 log lines while the other systems generate around 45.000 log lines.

Subset

As already mentioned in the data understanding, the production stops when operators are on a break and the production starts again after the break ends. Consequently, after a break, the first data recorded in the IRIS system is the first to be fed to the send device which corresponds to the first measurement in the breast cap grader. Based on this information as well as the absence of an accurate and standard way to relate the breast cap grader data with the other datasets, it is reliable to use a subset of all the data with the starting time right after a production break and the ending time right before the final production break occurs in the whole timespan of the recording. This decision is taken after discussions with experts of the Innova and IRIS systems which concluded that having a starting and an ending point that is surely accurate will also make the linking of the data sets easier and more reliable.

Zero Values

The IRIS (front and back) datasets and the send device dataset contain records with zero values in the weight attribute and subsequently in the rest ones as well. The percentage of the zero values in these datasets represent 35% of the total data. These values represent empty shackles in the cut up line, they offer no interesting information to the research or the data preparation phase, thus their corresponding rows are removed. Moreover, the column corresponding to the 12th coordination point of the back IRIS dataset contains only zero values indicating that the 12th coordination point is not recorded and thus this attribute was removed from the final dataset. Then, there are some measurements that contain zero values in particular coordination points which means that the broilers had broken or cut wings. These recordings are rarely the case and are eliminated as a whole as well. Another removal from the final dataset is the final column from both IRIS datasets (front and back) since it is empty and falsely generated by the IRIS system.

Noise

The breast cap grader dataset appears to have a lot of noise as shown in the data understanding phase. There are rows and more specifically values assigned on the attribute of breast cap weight that are not numerical or exceed greatly the reality. By examining the dataset more extensively using Microsoft Excel and talking to the experts in Marel, it is decided to remove all rows with weight values that are not numerical or are greater than 10000. The rows containing these kind of values correspond to 10% of the full dataset and do not represent actual breast cap measurements but contain information about the settings of the breast cap grader which are causing additional problems in linking the dataset with the rest ones while not offering any valuable information to the research project. In addition, extremely low values of breast cap weight indicate that small fat or meat parts of the breast cap would detach during its fall in the conveyor belt following the breast cap cut up and get graded as well. These values are dealt with during the merge of the tables along with breast cap values that represent two combined measurements in one. IRIS and send device data also contain extremely high weight values (to separate birds being rehanged and losing their original queue in the sequence) which are also dealt with in the merge of the tables.

Out of scope

Attributes that are rendered out of scope for the current research project are removed from the data sets. More specific, the modifications that were made are the following:

- Breast cap grader: Removed every attribute except from weight of breast cap and timestamp.
- IRIS Front & Back: Removed process information.

Further Cleaning

The following attributes were proven useful to the analysis but only until a certain level.

- The data of the send device is redundant, since the information contained are the same as in the IRIS dataset and is solely used for the purpose of better understanding and achieving a link between the IRIS dataset and the breast cap grader dataset. For that reason the send device dataset is completely removed from the final dataset after serving this purpose.
- The product code attribute in the send device and the IRIS front and back datasets is only used to achieve a link between them. After serving this purpose, this attribute iss also removed from the final dataset.
- The timestamp attribute in all the datasets is used as an additional aid to make a more reliable link between the datasets and potential further pattern recognition. After serving this purpose it is not included to the final dataset for the modeling phase.
- The flock code in the IRIS data is used to divide the data set in flocks to perform flock-wise predictions. However, it is not used as a predictor in any analysis.

3.2.3 Data Selection

Discussions with specialists in the Innova IS and the IRIS system took place in order to assess the meaning, the importance and the relevance of the big list of attribute definitions. Based on the aforementioned discussions, only some information is considered relevant for predicting breast cap weight in a broiler.

3.2.3.1 Attribute selection

From the data understanding phase as well as the data cleaning part of the data preparation phase, it is apparent that only some attributes are relevant to the research while the rest are redundant (Table 3.5).

Relevant Attribute	Quantity of attributes	Details
Weight of the broiler	1	The weight of the broilers carcass as fed in the IRIS system (front and back) and the send device.
Front areas of the broiler body	8	Total Area of the front side of the carcass / Area of the left leg / Area of the right leg / Area of left thigh / Area of right thigh / Area of breast / Area of left wing / Area of right wing.
Back areas of the broiler body	7	Total Area of the back side of the carcass / Area of the left leg / Area of the right leg / Area of thigh / Area of breast / Area of left wing / Area of right wing.
Front coordination points	16	Image points from 1 to 16 as shown on the visual specimen of the front part of the carcass.
Back coordination points	17	Image points from 1 to 18 (excluding 12) as shown on visual specimen of the back part of the carcass.
Weight of the breast cap	1	The weight of the broilers breast cap as it is graded after its cut up.

Table 3.5: Relevant Attributes to the Analysis

3.2.4 Data Construction

Some attributes are derived from existing ones in the IRIS data sets. The coordination points, before being excluded from the final dataset, are separated into their x and y values and the distance between some of these key pairs (Table 3.6) is calculated, is saved in a new attribute and is then included in the final dataset.

Front Widths	Back Widths
1 to 10	1 to 10
2 to 11	2 to 11
8 to 12	4 to 5
13 to 9	6 to 7
4 to 5	
8 to 9	
6 to 7	

Table 3.6: Derived Attributes - Width

Research indicates that breast length has a higher correlation with the breast weight than the breast width. This fact leads towards the inclusion of breast length distances in the dataset too, as can be seen in Table 3.7.

Front Lengths	Back Lengths
4 to 8	4 to 8
5 to 9	4 to 6
6 to 8	5 to 9
7 to 9	5 to 7
3 to 6	1 to 8
3 to 7	1 to 9
3 to 15-16	3 to 15-16

Table 3.7: Derived Attributes - Length

The euclidean formula is used to calculate the distance between points x1, y1 and x2, y2:

$$Distance = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Despite length and width, a combination of ratios is also calculated and included in the final dataset. An interesting ratio to consider is the one of the width and the height of the broiler carcass.

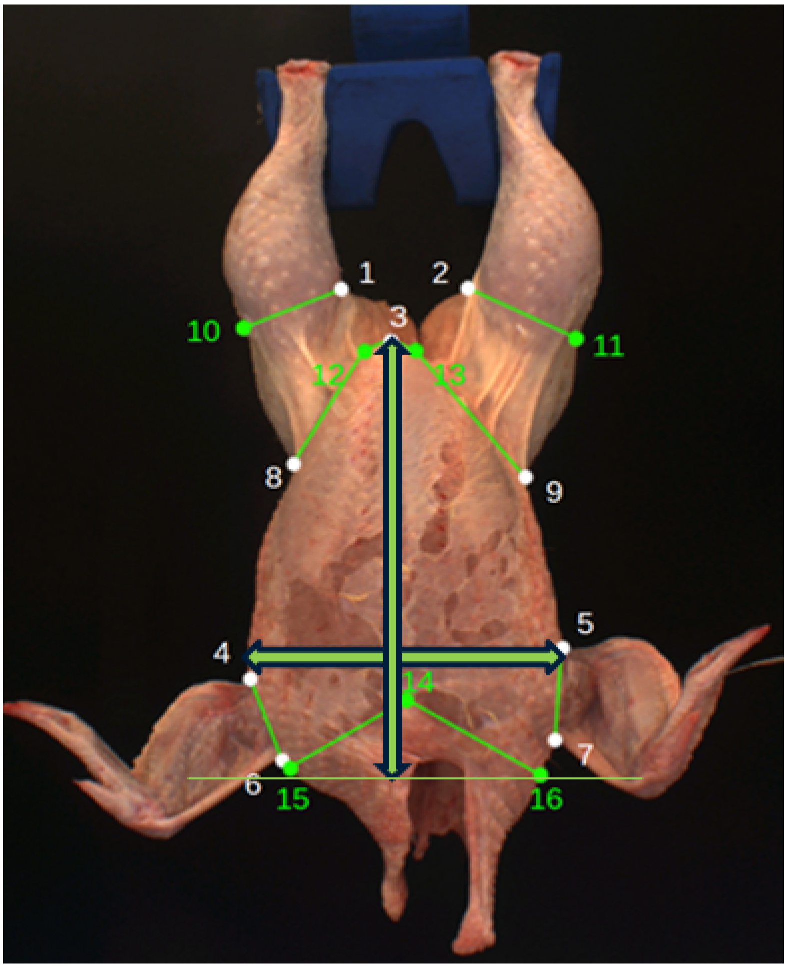


Figure 3.13: Ratio of Width & Height

As shown in Figure 3.13, the width of the broiler carcass is the distance from the left wing to the right wing (point 4 to point 5) while its height is the distance from the bottom of the breast (point 3) to the straight line that goes through points 15 and 16. The lower the height/width ratio, the bulker is the broiler.

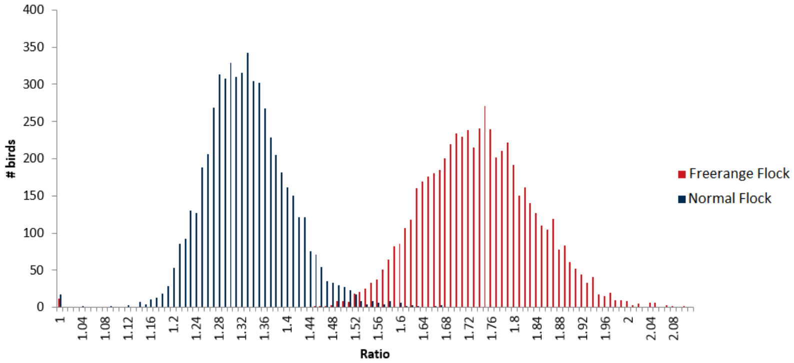


Figure 3.14: Comparable histogram of bulky and long bird flocks

An internal research of Marel showed that bulky and short broilers give a height/width ratio of 1.3 and long birds give a ratio of 1.75 [54]. Figure 3.14 shows the results of the research in a cumulative histogram of a normal (bulky) and a freerange (long) broiler flock. The short and bulky broiler flock is indicated in the blue part of the graph with a steep peek, while the long broiler flock is indicated in the red part of the graph with a less steep peek. A mixed flock would give either multiple peaks or a flat histogram.

That being said, the ratio values taken into account in the present research are shown in Table 3.8.

Front Ratios	Back Ratios
3 to 15-16 with 4 to 5	3 to 15-16 with 8 to 9
3 to 15-16 with 8 to 9	3 to 15-16 with 4 to 5
4 to 8 with 5 to 9	3 to 15-16 with 6 to 7
4 to 5 with 8 to 9	8 to 9 with 4 to 5
	8 to 9 with 5 to 6
	4 to 8 with 6 to 9

Table 3.8: Derived Attributes - Ratios

The histogram of the height/width ratio in the current dataset (Figure 3.15) shows that it consists of an average flock of birds which mostly contains neither bulky nor long birds but average sized ones.

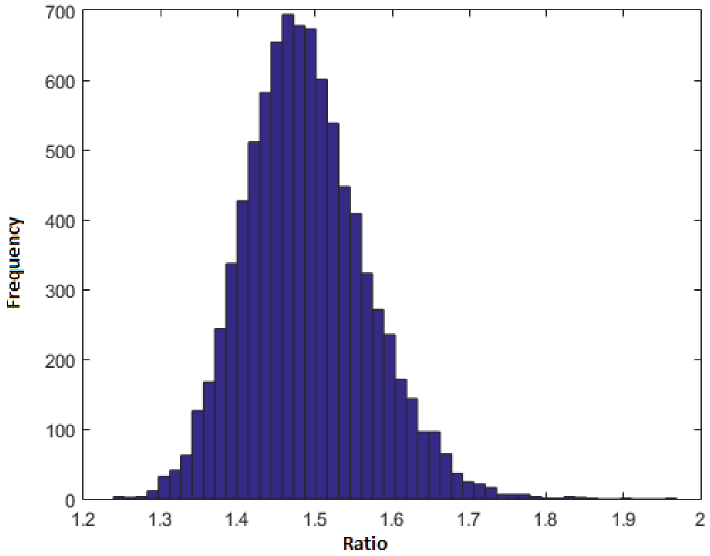


Figure 3.15: Comparable histogram of bulky and long bird flocks

3.2.5 Data Integration

All the processed data is merged into one final data set. Every row of it represents a different broiler and every column its characteristics (e.g. breast cap weight, total weight, front total area, distance from point 1 to 10, etc.). The methods that were followed are implemented in Matlab and they focus on merging the easy cases with straight forward one on one links between the different data tables. Further details are provided below.

Send device and breast cap grader

```
Data: Table of the send device & Table of the breast cap grader
Result: New merged table
initialization;
while not at the end of either of the two tables do
    if total weight >2705 then
        skip one row of the send device table;
    else if Breast cap weight <32% of the Total Weight then
        skip one row of the breast cap table;
    else if Breast cap weight >60% of the Total Weight then
        skip one row of the breast cap table;
        skip two rows of the send device table;
    else
        if 24 seconds <= (breast cap timestamp) - (send device timestamp) <= 28 seconds
        then
            merge the rows and add them as one to the new table;
        else
            skip one row of the breast cap table;
            skip two rows of the send device table;
        end
    end
end
```

Algorithm 1: Pseudocode - Send device with breast cap grader

Front and Back IRIS
The link is done based on the product code attribute which was the same for both the front and the back IRIS system dataset.

Final Dataset
The new table is horizontally concatenated with the merged IRIS data based on the common product code attribute between the send device, which is included in the new table, and the IRIS data.

3.2.6 Data Finalization

The final dataset contains, 10 thousand measurements. The final modifications made in the data do not change its meaning, but are required so that the modeling tools operate more effectively.

3.2.6.1 Final Adjustments

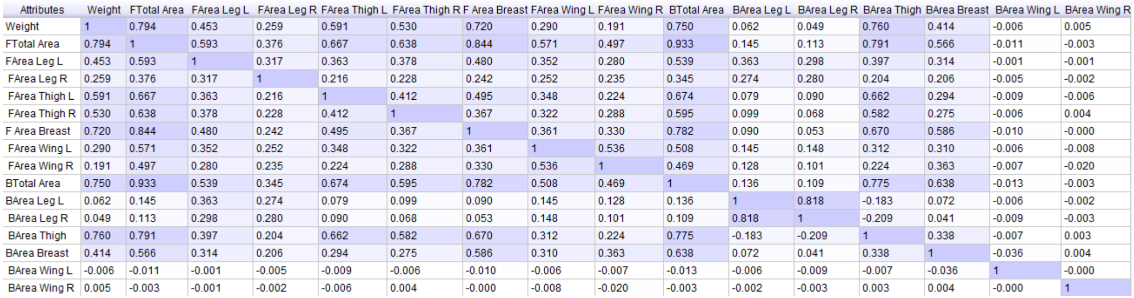
The to-be final dataset currently contains 8.817 rows but there are some issues observed in the data understanding phase that shall be addressed before the final dataset is fed to the models.

- There is a very small percentage of records with flock code #0, in which the Total Weight value is zero but the area and the coordination values are filled with real measurements.
- There is a couple of Area measurements which have extreme values. This is an abnormal case and it indicates error in the particular data values.
- A few of the coordination points have a "0,0" value. This is due to the fact that there are broilers going through the production line with broken or cut wings or legs, thus their corner points cannot be recognised and consequently calculated.
- Because of the aforementioned zero coordinations, their corresponding distance values are zero and as a result when taken into account for the ratio calculation they produce ratios of infinity values.

The above problematic values represent errors in the dataset and are removed since it cannot be used in the models. The final dataset is reduced to 8.769 rows with average total weight of 1.220 grams and Average breast cap weight of 496 grams.

3.2.6.2 Multiple Linear Regression Model

In the case of multiple regression some of the attributes may be considered relevant but still have to be excluded from the analysis because of multicollienarity issues, during which two or more attributes correlate with each other. To investigate on which of the attributes are highly correlated between one another Rapidminer is used a software tool that provides an integrated environment for machine learning, data mining, predictive analytics and business analytics. A correlation matrix is generated, with the help of the tool, to conduct an exploratory review on the relationships between the predictors. The results can be seen in the following table.



Attributes	Weight	FTotal Area	FArea Leg L	FArea Leg R	FArea Thigh L	FArea Thigh R	F Area Breast	FArea Wing L	FArea Wing R	BTotal Area	BArea Leg L	BArea Leg R	BArea Thigh	BArea Breast	BArea Wing L	BArea Wing R
Weight	1	0.794	0.453	0.259	0.591	0.530	0.720	0.290	0.191	0.750	0.062	0.049	0.760	0.414	-0.006	0.005
FTotal Area	0.794	1	0.593	0.376	0.667	0.638	0.844	0.571	0.497	0.933	0.145	0.113	0.791	0.566	-0.011	-0.003
FArea Leg L	0.453	0.593	1	0.317	0.363	0.378	0.480	0.352	0.280	0.539	0.363	0.298	0.397	0.314	-0.001	-0.001
FArea Leg R	0.259	0.376	0.317	1	0.216	0.228	0.242	0.252	0.235	0.345	0.274	0.280	0.204	0.206	-0.005	-0.002
FArea Thigh L	0.591	0.667	0.363	0.216	1	0.412	0.495	0.348	0.224	0.674	0.079	0.090	0.662	0.294	-0.009	-0.006
FArea Thigh R	0.530	0.638	0.378	0.228	0.412	1	0.367	0.322	0.288	0.595	0.099	0.068	0.582	0.275	-0.006	0.004
F Area Breast	0.720	0.844	0.480	0.242	0.495	0.367	1	0.361	0.330	0.782	0.090	0.053	0.670	0.586	-0.010	-0.000
FArea Wing L	0.290	0.571	0.352	0.252	0.348	0.322	0.361	1	0.536	0.508	0.145	0.148	0.312	0.310	-0.006	-0.008
FArea Wing R	0.191	0.497	0.280	0.235	0.224	0.288	0.330	0.536	1	0.469	0.128	0.101	0.224	0.363	-0.007	-0.020
BTotal Area	0.750	0.933	0.539	0.345	0.674	0.595	0.782	0.508	0.469	1	0.136	0.109	0.775	0.638	-0.013	-0.003
BArea Leg L	0.062	0.145	0.363	0.274	0.079	0.099	0.090	0.145	0.128	0.136	1	0.818	-0.183	0.072	-0.006	-0.002
BArea Leg R	0.049	0.113	0.298	0.280	0.090	0.068	0.053	0.148	0.101	0.109	0.818	1	-0.209	0.041	-0.009	-0.003
BArea Thigh	0.760	0.791	0.397	0.204	0.662	0.582	0.670	0.312	0.224	0.775	-0.183	-0.209	1	0.338	-0.007	0.003
BArea Breast	0.414	0.566	0.314	0.206	0.294	0.275	0.586	0.310	0.363	0.638	0.072	0.041	0.338	1	-0.036	0.004
BArea Wing L	-0.006	-0.011	-0.001	-0.005	-0.009	-0.006	-0.010	-0.006	-0.007	-0.013	-0.006	-0.009	-0.007	-0.036	1	-0.000
BArea Wing R	0.005	-0.003	-0.001	-0.002	-0.006	0.004	-0.000	-0.008	-0.020	-0.003	-0.002	-0.003	0.003	0.004	-0.000	1

Table 3.9: Correlation Matrix (Rapidminer)

The guide that was suggested by Evans [55] for the absolute value of the correlation coefficient derives the following categorization.

Correlation Coefficient	Interpretation
0,00 to 0,19	Very weak
0,20 to 0,39	Weak
0,40 to 0,59	Moderate
0,60 to 0,79	Strong
0,80 to 1,00	Very Strong

Table 3.10: Correlation Interpretation

Based on the correlation interpretation provided above, the attributes that have strong relationship between each other are identified and only one of them is kept in the adjusted final dataset for the multiple linear regression.

Attribute relationship (above 0.59)	Excluded Attributes
Total weight / Front total area / Front breast area / Back total area / Back area thigh Front area of left thigh / Front Total Area Back Area Thigh Front area of right thigh / Front Total Area Back area of left leg / Back area of right leg Back breast area / Back total area	Front total area Front breast area Back total area Back area thigh Back area of right leg

Table 3.11: Correlating Attributes

For the rest of the modeling techniques it is not problematic to have correlating attributes in the input dataset.

Chapter 4

Data Mining Approach

4.1 Modeling

This section describes the methods used to reach the goals of the project and assesses their individual results.

4.1.1 General Information

For the current research four different modeling techniques have been chosen to be developed. Before moving forward with the presentation of these modeling techniques, it is important to explain the test design and introduce some general assumptions that hold for each of the models used. These assumptions are listed below:

- The values of the measurements that constitute the final data set represent the real accurate values with no errors.
- The reliability of the evaluation results depends on the reliability of the data provided. Thus, the data sets are considered to be reliable and the method used to merge the data sets produces zero or negligible error.
- The final data set is a representative sample of the broilers commonly fed in the production lines of the poultry processor.

Test Design

Accuracy is the main parameter to measure the performance for the models that have been built [56]. The higher the accuracy means the better the performance of the model. While measuring the level of accuracy, several parameters are taken into account as well in order to obtain the optimum result:

Attribute selection

As already discussed in the data preparation phase, feature selection is performed to realize the optimum dataset that will give the best results in a certain approach. It is also interesting to observe how certain manual selections improve the accuracy of the models.

Remove error in the data and incomplete measurements

Incomplete data may decrease the accuracy of a model [57] as well as error in the data. For that reason it is decided that the erroneous and incomplete rows of data are to be removed from the final dataset and subsequently the training and test set.

10-fold cross validation

Cross validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Since k equals 10, the final dataset is repeatedly divided into training data and testing data randomly for ten times. Using cross validation an estimate of the expected model error can be obtained.

The accuracy of the training of the models is measured by the coefficient of determination, denoted R^2, which indicates how well the data fit the the model. The higher the R-squared the better the quality of the model trained. However, the value of the metric automatically increases when additional predictor variables are added to the model. To raise the reliability of the model the adjusted R-squared is also used to evaluate the training of the model together with the R-squared, since the adjusted R-squared takes into account the automatic increase of the metric and adjusts for the number of predictor variables in the model relative to the number of data points [58]. Moreover, there are cases where residual analysis is also performed, since the R^2 alone may not represent the real quality of the model.

The accuracy of the prediction made by the models, during the testing phase, is measured by the error that is produced when comparing the real testing data with the predicted data. Firstly, the average error of the observations is produced by taking the mean value of the positive and negative errors without regard to sign. Then the mean absolute error is calculated by measuring the average of the absolute value of the errors. Last but not least, root mean square error is also estimated by rooting the mean squared error (MSE) which takes the average of the squares of the errors. MSE is the second moment of the error as it also takes into account the variance of the true values.

In pursuance of selecting the fittest model, as far as model quality and prediction accuracy is concerned, the results for aforementioned metrics are compared with each other in the same model as well as with their outputs in different models.

4.1.2 Description and parameters

All the models of the current research project are developed with the use of MATLAB and each one of them holds different parameters. As can be observed in Appendix G, different parameters for the models are applied and tested in the same data set to derive the best performing models. The finest parameters for each modeling technique are provided in this section.

4.1.2.1 Multiple Linear Regression

Multiple linear regression is the most commonly used modeling technique when coming across similar data mining problems as the given one and it is thoroughly described in the Background chapter of the present thesis. However, the multiple multiple linear regression shall satisfy certain assumptions, which are also explained in the Background chapter, in order to provide reliable results. The steps towards reviewing the satisfaction of these assumptions is provided below:

- Scatter plots show that the relationship between the predictors and the independent variable is linear. For reference, the linearity between the total weight and the breast cap weight can be observed in Figure 4.1.

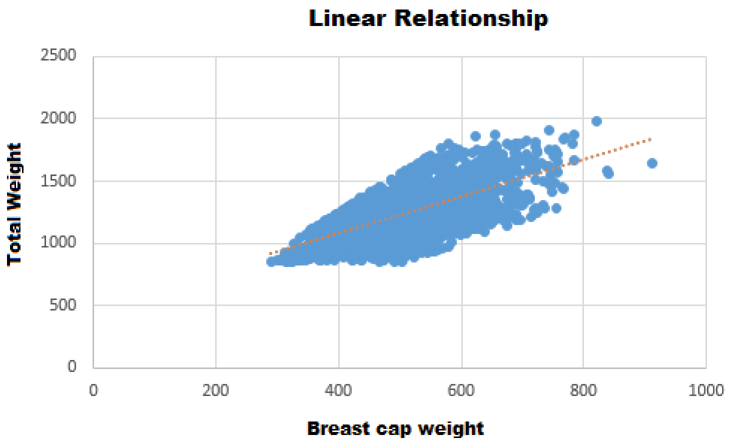


Figure 4.1: Plot of breast cap weight with total weight

- All variables in the data set are normally distributed after using Kolmogorov-Smirnov tests.
- As can be seen in the Data Preparation section, some groups of variables are highly correlated with each other and thus only one of them is kept in the final data set in order to achieve little multicollinearity.
- The values of every variable in the data set are independent from one another leading to no auto-correlation.
- There is partial homoscedasticity in the final dataset since the amount of error and noise is more or less the same for the most key variables in the dataset.

The final multiple linear regression model that is built represents the robust fit of the dependent variable to the predictors. A robust regression method is favored in this particular case because it is designed to be not overly affected by potential violations of the regression assumptions presented earlier on this stage nor by the outliers observed during the residual analysis. Removing the constant term, also known as intercept, from the fit further improved the quality of the model. The weight function delivering the best results is the 'talwar' function with tuning constant '2.795'. The formula of this weight function is presented below.

$$w = 1 * (abs(r) < 1),$$

where the value r is

$$r = resid/(tune * s * sqrt(1 - h)),$$

where "resid" is the vector of residuals from the previous iteration, "h" is the vector of leverage values from a least-squares fit, and "s" is an estimate of the standard deviation of the error term given by

$$s = MAD/0.6745,$$

where MAD is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution [59].

The model created is stratified 10-fold cross validated, which means that it was trained as well as tested ten times. In each iteration the 1/10th of the final dataset was held for testing purposes and the remaining was used for training. This process was repeated 10 times until all data had the chance to be held out from the training of the model in order to be used for the testing and vice versa. Additionally, feature selection was used to determine the best combination of predictors to be used in the final model. In the beginning the most predictive variable is chosen and in each consequent step one more predictive variable is to the model based on the increase of the coefficient of determination until not much more improvement is succeeded. In addition, for every variable it is checked that the T-statistic is less than -2 or more than 2 while the P-value is less than 0.05.

4.1.2.2 Regression Trees

Regression tree is an additional regression technique used in the current research which represents the model as a graphical tree. The assumptions made for the multiple linear regression hold for the regression trees as well. Regression trees assume the existence of a single numerical output, which is the breast cap weight in the current case, and one or more numerical predictors, which are again the same predictors used in multiple linear regression.

The tree model created in the present research represents a stratified 10-fold cross validated decision tree with binary splits for regression where each branching node is split based on the values of the predictors. Testing different parameters showed that the optimal maximum number of splits for the current data set is '25'. An estimation of the optimal sequence of pruned sub-trees is also taken into account when growing the regression tree while further pruning the tree did not offer significant improvements.

4.1.2.3 Neural Networks

Another attractive data mining option, selected for the modeling phase of the project, is the artificial neural networks. Importantly neural networks do not require any priori assumptions about the distribution of the data or the form of interactions between the variables leading to a non-trivial effort on the part of the analyst, especially when other approaches are also used [60]. Furthermore, neural networks offer high accuracy by approximating complex mappings and by being flexible with respect to noisy data. Other than that, they are also able to efficiently decide for occurrences that are not present in the training set establishing their capability of abstraction [61].

The neural network produced is a stratified 10-fold cross validated feedforward net with two hidden layers consisting of 20 neurons each (Figure 4.2). Different combination of hidden layers and neurons where applied before reaching the final decision.

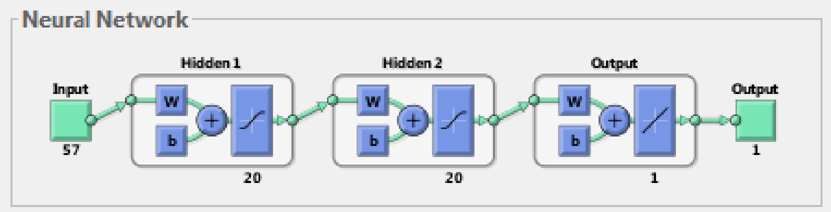


Figure 4.2: Neural Network Architecture

The function selected to train the neural network is the Levenberg-Marquardt function which is considerably faster than the rest of the functions and delivers the most favorable results.

4.1.2.4 Fuzzy Inference Systems

The last modeling technique chosen is the adaptive neuro-fuzzy inference system, also known as anfis. Fuzzy logic is more tolerant than other techniques on imprecise data, building the understanding of the imprecision into the process rather than dealing with it in the end. Specifically, anfis supports systems which satisfy the following properties:

- All output membership functions shall be linear.
- The fuzzy inference system shall be first or zeroth order.
- Different rules shall not share the same output membership function, which means that the number of output membership functions are equal to the number of rules.

The FIS structure complies with all of these constraints minimizing the occurrence of errors. However, the use of grid partitioning exponentially increases the number of rules when predictors are added to the model. This consumes a big amount of computational resources increasing drastically the running time and rendering the training of the model, which includes the full final data set, impossible. Instead, after further testing, subtractive clustering is used to generate a first order Sugeno-type Fuzzy Inference System structure with three rules. The model is then verified and the optimization capability of anfis is used to improve the model with 50 epochs using a combination of least squares and back-propagation gradient descent methods.

4.1.3 Assess Models

Based on the testing case and the optimal model parameters used (Appendix G), the models have yielded the mean evaluation results shown in Table 4.1, 4.2 and 4.3. The means represent the average values of the ten different training and testing iterations. Different groups of predictors are also used to measure accuracy and data fitting in different scenarios. Further comparisons of the final results for the best models of each data mining technique are presented in Appendix H.

Type of Data	Metric	Robust MLR	R-Tree	NN	FIS
Training sets	Mean Adjusted R-Squared	0,578	0,406	0,491	0,353
Testing sets	Mean Absolute Error	34,71	37,42	34,31	40,22
Testing sets	Root Mean Square Error	47,26	49,38	45,8	51,97
Testing sets	Mean Average Error	4,53	-0,23	0,15	-0,27

Table 4.1: Results with total weight excluded

Type of Data	Metric	Robust MLR	R-Tree	NN	FIS
Training set	Mean Adjusted R-Squared	0,651	0,448	0,464	0,465
Testing sets	Mean Absolute Error	35,31	35,52	35,16	35,16
Testing sets	Root Mean Square Error	48,56	47,13	46,59	46,59
Testing sets	Mean Average Error	5,13	-0,11	0,36	0,25

Table 4.2: Results with IRIS data excluded

Type of Data	Metric	Robust MLR	R-Tree	NN	FIS
Training set	Mean Adjusted R-Squared	0,701	0,477	0,529	0,508
Testing sets	Mean Absolute Error	32,54	34,99	32,91	33,45
Testing sets	Root Mean Square Error	45,97	46,77	44,51	45,64
Testing sets	Mean Average Error	4,7	0,72	-1,08	-1,53

Table 4.3: Results of the full data set

From the all the results several conclusions can be drawn:

1. Robust multiple linear regression model has the lowest mean absolute error.
2. Neural networks model has the lowest root mean square error
3. Robust multiple linear regression model has the highest quality from all the other models since it explains the largest amount of the variation in data
4. Average error differs from model to model without showing any correlation to the other metrics
5. The average error is the closest to zero and thus better performing in the regression tree model.

6. Every model performed considerably better when the total weight of the broiler is not omitted, with the exception of the neural networks models, which do not generate significant differences in their results.
7. Models with the highest R-squared deliver low error values as well.

4.2 Evaluation

In this part of the research, the best fitting model is used to deliver the final results and assess the degree to which it meets the business goals.

4.2.1 Evaluate Results

4.2.1.1 Approved Model

The robust multiple linear regression model is chosen as the final model which delivered the best performance among the models trained and tested in the modeling phase. The final model, which does not exclude the total weight of the broiler from the predictor variables, explains 70.1% of the variation in the data according to the R^2. The output of the coefficient of determination in the final model is quite high if we are to consider that broilers are live organizations the growth of which depends on a lot of different factors (feed, breed, weather, etc.) and thus are harder to predict.

A robust fit is more flexible to the outliers, thus R^2 alone is not a truly representative way of measuring the quality of the fit in this case since it also cannot determine whether the coefficient estimates and predictions are biased [30]. A residual analysis is performed to tackle the aforementioned issue. The case order residual plot along with the histogram over the residuals of the final regression model can be viewed in figures 4.3 and 4.4 respectively.

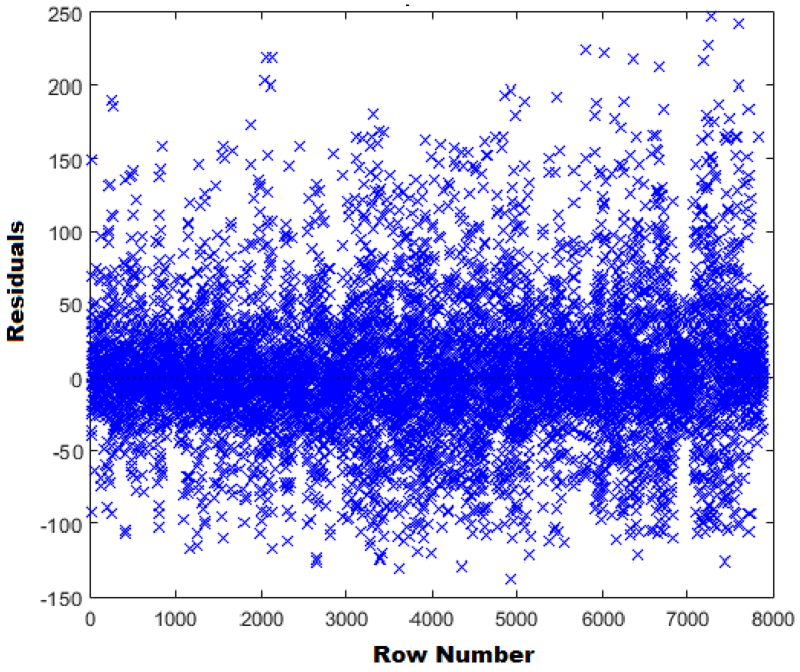


Figure 4.3: Histogram of residuals

Figure 4.3 shows that bias does not apply in the current case since residuals are randomly scattered without showing any systematic patterns. In addition, the histogram (Figure 4.4) reveals

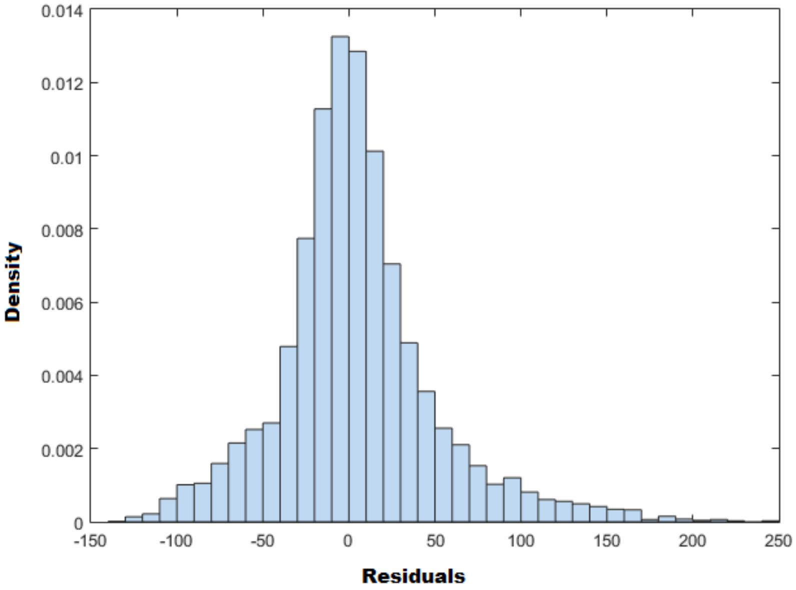


Figure 4.4: Histogram of residuals

that the residuals are normally distributed around zero which means that the expected value of the error term is very close to zero. Furthermore, when observing the aforementioned figures it is easy to detect infrequent residual outliers which justify the use of a robust approach. Last but not least, the mean average error and the root mean square error is also estimated on the training set to obtain a mean weight error value that will quantify the quality of the model. The mean average error for the training sets is 4 grams and the RMSE is 35.7 which translates to the real values being very close to the expected values of the set.

The best fitted model is the following:

$Cap\widehat{Weight} = (TotalWeight) * 0,318601093 - (FAreaLegL) * 0,081883973 - (FAreaLegR) * 0,025848802 - (FAreaThighL) * 0,106626944 - (FAreaThighR) * 0,062240903 + (FAreaBreast) * 0,022136921 - (BAreaLegR) * 0,051265842 - (BAreaThigh) * 0,039225353 + (BAreaBreast) * 0,066269489 + (fdi1_10) * 0,368654114 - (fdi4_5) * 2,238024856 - (fdi8_9) * 0,562674519 - (fdi6_7) * 2,531582068 + (bdi4_5) * 3,573616521 + (bdi6_7) * 1,741079649 - (fdi3_7) * 21,32025172 + (fbh) * 23,10761497 - (bdi4_6) * 0,503126373 - (bdi5_7) * 0,941819071 + (bdi2_9) * 0,246706679 - (bbh) * 1,368634807 - (frh_45) * 182,2434 + (brh_45) * 147,5330804$

4.2.1.2 Final Results

A new dataset, which is collected and created following the steps of the Data Processing phase, is fed into the final model to obtain the final results for the operators of the poultry processor. The final results represent data from 2 hours of production fed into the best trained model presented in the previous section. Figure 4.5 presents a comparative histogram of the real breast cap weight values together with the predicted ones where both values are normally distributed around their mean value without extreme dissimilarities.

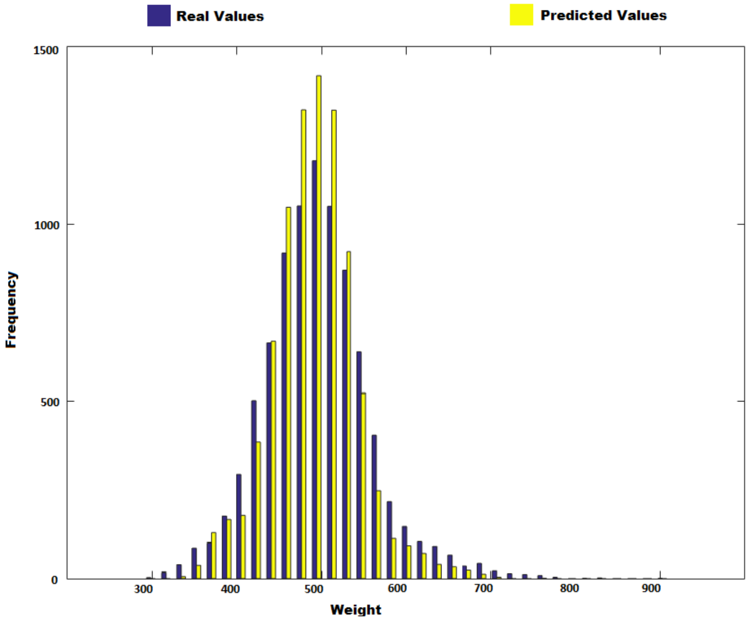


Figure 4.5: Histogram of real vs. predicted values

As can be seen in Table 4.4 the aforementioned model is compared with the benchmark to evaluate its performance and quality. For better understanding the comparison, several information have to be provided first regarding the real data set.

- 8769 broilers were used in the data set.
- The total weight of the broilers is 10689649 grams.
- The total weight of the breast caps is 434473 grams.
- The mean of the total weight of the broilers is 1219 grams.
- The mean of the breast cap weight is 495.5 grams.
- On average, about 58% of the breast cap weight corresponds to fillet weight [52].
- In the approach currently used by the poultry processor, it is estimated that 20% to 25% of the average of the total weight of a group of broilers corresponds to fillet weight [5].

By doing the appropriate calculations we derive with the results shown in Table 4.4.

Metric	**Final Model**	**Benchmark**
Total Predicted Breast Cap Weight	4.303.897	3.685.090–4.607.503
Total Predicted Fillet Weight	2.496.826	2.137.882–2.672.352
Percentage of Fillet	23,35%	20%–25%
Mean of Predicted Breast Cap Values	490,8	420–525
Mean Average Error	4,5	75,4–29,6
Mean Average Percentage Error	1%	16,3%–7%
Standard Deviation of Real Values	63,6	63,6
Standard Deviation of Predicted Values	50,2	N/A
Root Mean Square Error	46,2	N/A
Mean Absolute Error	32,5	N/A
Mean Absolute Percentage Error	6,5%	N/A

Table 4.4: Results of the final model

It is important to note the differences between the two approaches and realise the improved prediction accuracy that the data mining approach offers. Additional comparisons with other combinations of predictors can be found in Appendix J. Moreover, the datasets created and used in the research contain measurements from different flocks of birds which can be split up and separately fed in the final model to obtain flock-wise results (Appendix I).

4.2.2 Review process

The final data set was split into training and testing cases with the help of 10-fold cross validation. The 10-fold cross validation makes a loop of the data mining techniques and sends a certain randomly chosen part through the model as training data and the other part as testing data. In the models presented here 90% of the dataset was always used as training data and 10% as testing data in every loop.

Different combination of parameters were selected for each type of modeling technique and the best model was chosen by executing the following steps. First all the relevant variables were added into each model taking into account the assumptions of each one and removing variables that may violate these assumptions. Then the goodness of fit for different models was assessed with the use of R-squared following a residual analysis that was performed to check the consistency of the errors. Additional variables, that were proved irrelevant, were later excluded in order to raise the quality of the models.

At the same time the models were tested for prediction accuracy. The average error was initially computed to observe its distance from the zero value. Alarmed by the possibility of the infrequent large errors messing up with the error perception it was decided that more evaluation metrics should be used. The mean absolute error was later implemented in the solution which computes the average magnitude of errors in a set of forecasts without considering their direction. To adjust for the large but rare errors, the root mean squared error was also applied to the test cases which gives more weight to the large but infrequent errors. Finally a comparison between RMSE and MAE was made which showed that the RMSE was bigger than MAE in every model which indicates that large errors have indeed occurred but since the difference is not big the occurrences are not that frequent. From this comparison it can be concluded that the error size is not inconsistent and that the variation of the magnitude of errors is not high.

The final model selected is the robust multiple linear regression model which produced the best results in every metric in comparison with the other models. A new data set was fed into the final model to retest the prediction accuracy and obtain the end prediction results (Table 4.3). Then a histogram (Figure 4.4) of the real and the predicted values was constructed to have a more visualized result of the model which is easily interpreted by the operators in the poultry processor who will make the final decisions.

Because of the limited number and type of flocks used in the current research the solution might not hold up in its entirety if tested on special types of flocks. Thus, there may also be a case of the final model considered not suitable to generalize it to all flock types regardless of the other parameters that could characterize such a live organization as a broiler.

4.2.3 Determine next steps

The modeling and evaluation process have iterated multiple times producing satisfying enough results to finalize the approach and move forward with the project onto the deployment phase in order to understand and realise the usability of the data mining solution in the business context. The next steps of the current research are discussed and listed in the final chapter of the report.

Chapter 5

Discussions & Conclusions

5.1 Deployment

For the final phase of the research methodology a Focus Group is conducted in Marel Stork Poultry Processing.

5.1.1 Focus Group

The Focus Group [62] is a design of qualitative research used for evaluation purposes which also serves as an adjunct to quantitative data collection methods. The strength of such a research method is the possibility of collecting the values of the method at stake from different perspectives. In the present research the Focus Group took the form of a discussion session among experts with the purpose to confirm the results of the research project as well as explore on the usability of the final model. During the time of the session, the researcher gave a presentation explaining the context and the proposed solution to the participants. Later he asked questions and listened to what the experts had to say on the matter. The protocol of the aforementioned Focus Group is presented in Appendix L. For anonymity purposes the participants of the focus group will be referenced with the letter ”P” followed by a number, as listed in the Appendix.

5.1.2 Focus Group Outcome

Limitations

During the focus group P4 mentioned that in each measurement there is a small deviation from the real value which cannot be quantified making it hard to realise the impact of feeding more accurate measurements to the final model. As a result, the breast cap weight measurements recorded do not represent the reality, since they are not 100% accurate which render the final result numbers questionable to a certain degree.

Next Steps for the company

P3 noted that the attributes used to train the models seem logical and make sense. However, he added that another interesting path could be the use of 3D measurements as predictors since they could be more accurate and give more value to the final model. P1 also mentioned that a good idea is to create different models specifically for light and heavy birds.

As quoted by P2 “The approach of using large datasets and creating data mining models to acquire insights in different areas of the production line is quite a new thing for us and I think this is a nice approach which could be also used in other applications with different variables”. In other words, a similar model could be created with other predictors, such as the number of broken wings or feathers, concluding to the fact that there are countless possibilities for such a system.

Another interesting suggestion that surfaced from P2 is to examine which predictor combinations give more stable prediction results over the different flocks as well as observe if the same results apply for the production lines of different poultry processors. In addition, a challenging idea for further research, according to P1, would be to create a similar approach for the filleting lines that would predict the amount of fillet produced in a certain time span with a level of uncertainty. It is believed that such an assignment could even be considered as a graduation project for a following internship.

Moreover, P4 strongly supported an interesting scenario where the total weight of the bird could be accurately predicted based on the IRIS measurements in order to prevent the poultry processor from using smart weighers in the production line layout. This is also a promising idea that is circulating internally among the experts of the company and could bring benefit to the poultry processor. In that scenario, the poultry processor would solely use the IRIS system to distribute broilers with certain weights to certain production lines instead of the smart weigher. Because of the importance of the idea and its simplicity compared to the other ones, it was decided that it would be the immediate next step. Indeed, the first tentative steps towards that direction were already made, as seen in Appendix K, in consequence of the present Focus Group in order to produce an indication of the results that could be produced by the current final model in such a case.

Utility and Implementation of the approach in practice

When asked for the improvements that the new data mining approach delivers compared to the current approach used in the poultry processor, all the participants acknowledged that the new approach outputs better results. However P1 thinks that having a small of an increase in the accuracy using the approach presented does not bring much value in the end production performance, mainly because in the production planning the operators also find it important to have certain information and predictions 2 days in advance, like the number of broilers eventually needed for a production day, their volume and their supplier. The data mining approach presented cannot easily cover that area as well.

Another interesting topic that was set for discussion is the place in which such a model could be implemented. P5 stated that the approach presented has three purposes:

1. Give feedback to the farmers - The implementation has to be done in the beginning of the production so that there is no influence from the other processes.
2. Know what is coming - The model has to be placed in beginning of the production process and definitely before the chilling tunnel so that there is time to predict and calculate what is coming after the chilling process.
3. Have a theoretical target for the fillet yield in the filleting lines - Again the model has to be placed before the chilling tunnel.

It is commonly agreed that the second one is the most important one and also the one that requires the least amount of accuracy to be considered valuable. A reference to the expected fillet yield compared to the actual fillet yield has to be extremely accurate to bring value and that would require a very advanced and highly accurate system.

P2 also agreed to that and added that focusing on solely breast cap weight is not the ideal case but it was done in the current case for practical reasons since in the current case the poultry processor is mostly interested on the exact fillet weight.

Focus Group Conclusions

P3 added that the whole issue of the planning department is how to create a system that automatically does all the planning and scheduling when the right input is provided to it. He also concluded by saying that it is not yet clear what the output of the final model presented is going to be used for. So if the poultry processor installs it in the beginning of the production line to predict flock weights in order to decide early on which flocks are going to be used in which production batches then it will bring value. But making an individual bird prediction and then tell how much fillet can come up from this bird then it is not certain if the perfect accuracy, that is needed, can be reached.

In the end, the focus group participants decided that the research made is very interesting and a job well done as it gives a good insight on the countless and promising possibilities of using data mining approaches on the vast amount of data generated by the poultry processor.

5.2 Conclusion

A perfect model that is able to predict the exact value of a breast cap, let alone a fillet, is very hard to achieve if not impossible since it has to include several aspects of what makes a living organism grow more in certain parts of its body and less in others. Trying to achieve such a model would be very complex with variables that are difficult to measure and may differ among broilers, flocks and breeds. Therefore, it is believed that having a more simple model that is easy to understand and use, which at the same time predicts the value of a breast cap reasonably well, will also give a reasonably well depiction of how important different factors are to best predict the breast cap value in a broiler or even in a whole flock of birds.

5.2.1 Research Questions

In the beginning of the research project certain research questions were listed. The answers to these questions are summarized in this section.

5.2.1.1 RQ 1. What are the specifications of the current methods used for prediction purposes and what are the quality criteria?

This question is answered in the introduction part of the present report during the phase of Business Understanding. The predictions performed at the poultry processor are based on experience. The average weight of a flock of birds is calculated and an assumption is made that 20% to 25% of this weight corresponds to fillet weight. It is estimated that this method delivers a good amount of prediction accuracy (around 80%) but there is still place for improvement.

5.2.1.2 RQ 2. What kind of data is available as input for the approach proposed?

As seen in the data processing part of the report the data used as input for the data mining approach are categorized into predictors and target variables. The target variable in this case is the breast cap weight of the broiler which is estimated based on a group of predictors consisting of the total weight of the broiler's carcass, areas of the broiler, distances between certain coordination points in the image of a broiler's carcass as well as the ratio of the height and width of the broiler's breast cap. The flock code of the broilers was also used to separate the data sets flock-wise.

5.2.1.3 RQ 3. What kind of methods/techniques shall be implemented in the available data?

This question is answered in the data mining approach chapter. The data mining methods used in the data set are multiple linear regression, regression trees, neural networks and fuzzy inference systems. All the methods were parametrized to offer the best possible results and after several testing iterations the best performing model was chosen as the final model for deployment.

5.2.1.4 RQ 4. How shall the models created be evaluated?

In the fourth chapter of the report it is stated that the training and testing of the models is implemented with 10-fold cross validation. To measure the quality of the trained models r-squared was used along with an observation of the residuals plots and the variance of the real and predicted values. Mean average error was used to measure the prediction accuracy of the models along with mean absolute error to realise the actual error made in a flock of birds on average without regard to the sign of the error. However, root mean squared error was also used because infrequent large errors were observed and this metric gives more weight to the large errors. Comparing the mean absolute error with the root mean squared error show a small difference in their values which indicates a small variation in the magnitude of the errors.

5.2.1.5 RQ 5. How can we better predict the amount of breast cap in a broiler based on the available features?

The robust multiple regression formula provided in the fourth chapter of the thesis (4.2.1.1), gives the carcass features that are most relevant to predicting the amount of breast cap in a broiler. Appendix M gives a description of the variable names used in the models. It was noted that areas offered slight improvements to the results but still their contribution to the prediction was more than the contribution of the ratios. However, when comparing the results of the different techniques used, it was apparent that the highest predictors are:

- The total weight of the broiler
- The distance from point 4 to point 5 on the front side
- The distance from point 6 t point 7 on the back side

5.2.1.6 RQ 6. How can the final model be integrated to the current operations?

During the focus group that took place in Marel Stork Poultry Processing, it was agreed that such a model shall be installed in the primary process to predict the breast cap weights flock-wise in order to have prediction of what is coming after the chilling tunnel for the distribution of the flocks in the filleting production lines.

5.2.1.7 Main RQ. How can we offer an accurate estimation of the amount of breast cap in a broiler carcass?

The answers to all the sub-questions that are listed above lead to the answer of the main research question. A robust multiple linear regression formula is generated to predict the breast cap weight of a broiler based on a group of predictors supplied by the IRIS system. Such a model shall be implemented before the chilling tunnel of the production so that the operators will have time to make calculations in order to take the right decisions before the secondary process begins. The results will be presented to the operators in the form of a histogram with the predicted breast cap weight values, similar to the ones on Appendix H, as well as individual predictions corresponding to each separate bird.

5.2.2 Limitations

It is important to note the limitations of the research involved in this thesis. One of the limitations in this research project is the reliance of the final results in the sample used to train the models. To mitigate this limitation, flocks of birds of the most common breed with a weight range from 900 grams to 2000 grams were use which made it possible to investigate on a more generalised prediction approach. However, the results of such an approach may be limited as there may be questions regarding to the extent of generalisation, since the approach may not behave as adequate on extremely specialized flocks of birds. Such scenarios were unable to be tested due to further limitations in the communication with the poultry processor.

Another limitation of this study is the inherent reliability of the data sets used. As previously discussed in the focus group, every measurement, which is generated from the systems used in the poultry processor, contains a small error in terms of deviation from the real value of that same measurement. Moreover, the method followed to merge individual broiler measurements between the different systems is of a questionable reliability. Since there is no apparent way of merging all the data sets into a final data set in a tangible manner, timestamps were the main factor used to link the measurement of a broiler in one system with the measurements of the same broiler in the remaining systems. To further minimize the effect of such a limitation, an assumption is made that all the data sets used represent the real measurements and correspond to the correct broiler. Moreover, outliers that are made apparent in the final data set are dealt with.

As an addition to the previous limitation, the logging of the measurements was mostly done remotely without being able to monitor whether operators handling the production process intervene to the product sequence between systems or not. As a result, it was impossible to track the broilers and their parts, nor the possible manual interventions occurring throughout the process that would disorder part of the product sequence. Such a problem affects the consistency and reliability of the data generated in the semi-automated filleting lines. Being aware of the aforementioned limitation gave the green light to turn the focus of the research to the breast caps of the broiler as opposed to the fillets.

Finally, time and resource limitations did not allow the study to examine the behavior of the final model in various poultry processors nor in flock of broilers with special characteristics, such as light/heavy weight or particular breed.

5.2.3 Future Work

There is a considerable variety of promising research directions that could be pursued in order to further expand the approach proposed in the present research. One such direction would be to investigate on how a similar data mining approach could be implemented on the filleting lines of a poultry processor. A system performing prediction of the fillet weight in a time-span with an amount of uncertainty shall be developed in such a case where heavy human intervention is involved and disturbs the consistency of the data recorded.

Furthermore, raising the accuracy of the measurements generated by the systems of the poultry processor will also increase the prediction accuracy of the model. In addition, the implementation of a system that logs three-dimensional measurements of the broiler's carcass could provide even more accurate and relevant predictors for the final model.

Another interesting direction would be to review the present methodology in different poultry processors as well as flocks of special breeds of broiler to observe possible changes in the behavior of the results. Last but not least, the vast amount of data generated by the systems in the poultry processor give the opportunity to investigate on endless possibilities that through data mining could aid or even enable the poultry processor to reach specific business goals.

Bibliography

[1] T. P. Robinson and F. Pozzi, *Mapping supply and demand for animal-source foods to 2030.* FAO, 2011. 1

[2] Marel, "Marel Homepage, url = http://www.marel.com,." 1

[3] Marel, "Innova Services Brochure," 2013. 1

[4] Marel, "The World of Poultry Processing," 2013. 1, 60

[5] R. V. D. Wijst, "Interview," 2015. 2, 44

[6] H. Dow, "Goal Question Metric and Software Quality," 2007. 2

[7] V. Basili, G. Caldiera, and H. Rombach, "The Goal Question Metric Approach," 1994. 3

[8] P. Chapman, J. Clinton, and R. Kerber, *The CRISP-DM User Guide*, 1999. 4

[9] P. Chapman, *Crisp-DM 1.0: Step by Step Data Mining Guide.* SPSS Inc.,SIG, 2000. 4, 5

[10] "What is the CRISP-DM Methodology?." 4

[11] A. R. Sams, *Poultry Meat Processing.* CRC Press, 2001. 7

[12] Landgeflugel, "Modern Poultry Processing." 7, 8, 9

[13] D. Fletcher, "Slaughter technology," *Poultry Science*, 1999. 8

[14] J. M. Gonzlez-Alvarado, E. Jimnez-Moreno, R. Lzaro, and G. G. Mateos, "Effects of Cereal, Heat Processing, and Fiber on Productive performance and digestive traits of broilers," *Poultry Science*, 2007. 8

[15] S. Barbut, "Poultry: Slaughter Line Operation," *Encyclopedia of Meat Science*, 2004. 8

[16] I. Guerrero-Legaretta, *Handbook of Poultry Science and Technology: Primary Processing.* Wiley, 2010. 8

[17] C. James, C. Vincent, T. I. de Andrade Lima, and S. J. James, "The Primary Chilling of Poultry Carcasses A Review," 2005. 8

[18] K. Wemp and D. Stone, "Lesson 9: Poultry Processing." 9

[19] S. J. H. M. Rijks, "Applying Business Process Redesign Heuristics in an Artifact Centric Process Modeling Approach: A case study in the poultry processing industry," Master's thesis, Eindhoven University of Technology, 2014. 10

[20] A. Ahlemeyer-Stubbe and S. Coleman, *A Prcatical Guide to Data Mining for Business and Industry.* Wiley, 2014. 11, 12, 61

[21] C. Aggarwal, *Data Mining: The Textbook.* Springer, 2015. 11, 61

[22] O. Marban, G. Mariscal, and J. Segovia, *A Data Mining and Knowledge Discovery Process Model.* I-Tech, 2009. 12

[23] R. Brown, *Fundamentals of Correlation and Regression.* Reagan Brown, 2012. 12

[24] G. Yule and M. Kendall, *An Introduction to the Theory of Statistics.* Charles Griffin and Co, 1968. 13

[25] S. Dowdy and S. Wearden, *Statistics for Research.* Wiley, 1983. 13

[26] A. O. Sykes, "An Introduction to Regression Analysis," 1992. 13

[27] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied Regression Analysis: A Research Tool.* Springer, 1998. 13

[28] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers.* John Wiley and Sons, Inc., 2003. 13

[29] J. W. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test," 2002. 13

[30] J. Frost, "What Are the Effects of Multicollinearity and When Can I Ignore Them?," 2013. 14, 42

[31] L. Pekelis, "Classification And Regression Trees : A Practical Guide for Describing a Dataset," 2013. 14

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2008. 14

[33] H. K. D. H. Bhadeshi, "Neural Networks in Materials Science," 1999. 15

[34] M. Sordo, "Introduction to Neural Networks in Healthcare," 2002. 15

[35] D. Dubois and H. Prade, "Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets," 2010. 16

[36] R. J. Marks, "An Introduction to Fuzzy Inference," 2000. 16

[37] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," 1975. 16

[38] K. Guney and N. Sarikaya, "Comparison of Mamdani and Sugeno fuzzy inference system models for resonant frequency calculation of rectangular microstrip antennas," 2009. 16

[39] A. Sylvain, "A survey of cross-validation procedures for model selection," 2009. 16

[40] N. R. Draper and H. Smith, *Applied Regression Analysis.* Wiley-Interscience, 1998. 17

[41] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011. 17, 61

[42] R. Hyndman and K. A., "Another look at measures of forecast accuracy," 2005. 17

[43] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," 2005. 17

[44] A. O. Raji, J. U. Igwebuike, and I. D. Kwari, "Regression models for estimating breast, thigh and fat weight and yield of broilers from non invasive body measurements," 2010. 18

[45] J. Latshaw and B. L. Bishop, "Estimating Body Weight and Body Composition of Chickens by Using Noninvasive Measurements," 2001. 18

[46] J. E. Melo, M. M. Motter, L. R. Morao, M. J. Huguet, Z. Canet, and M. C. Miquel, “Use of in-vivo measurements to estimate breast and abdominal fat content of a free-range broiler strain,” *Animal Science*, 2003. 18

[47] S. I. Inc., *SAS/STAT 9.1 User’s Guide.* SAS Publishing, 2004. 18, 19

[48] J. M. Larivire, C. Michaux, V. Verleyen, C. Hanzen, and P. Leroy, “Non-Invasive Methods to Predict Breast Muscle Weight in Slow-Growing Chickens,” 2009. 18, 25

[49] S. O. Olawumi, “Phenotypic correlations between live body weight and carcass traits in arbor acre breed of broiler chicken,” *International Journal of Science ad Nature*, 2013. 19

[50] H. H. Musa, G. H. Chen, J. H. Cheng, B. C. Li, and D. M. Mekki, “Study on Carcass characteristics of chicken breeds raised under the intensive condition,” *International Journal of Poultry Science*, 2006. 19

[51] G. C. Venturini, V. A. R. Cruz, J. O. Rosa, F. Baldi, L. El Faro, M. C. Ledur, J. O. Peixoto, and D. P. Munari, “Genetic and phenotypic parameters of carcass and organ traits of broiler chickens,” 2014. 19

[52] M. Saglibene, “Design of an efficiency measurement system for the filleting lines of Marel Stork Poultry Processing,” Master’s thesis, Wageningen University, 2015. 19, 44

[53] E. Vannan, “Quality Data An Improbable Dream?: A process for reviewing and improving data quality makes for reliable and usable results,” 2001. 25

[54] J.-P. Feddema, “Interview,” 2015. 33

[55] J. D. Evans, *Straightforward Statistics for the behavioral sciences.* Brooks/Cole Publishing, 1996. 36

[56] A. Wilbik, “Business Intelligence Lecture Notes.” 37

[57] E. Acuna and C. Rodriguez, *The treatment of missing values and its effect in the classifier accuracy.* University of Puerto Rico, 2004. 37

[58] H. Theil, “Economic Forecasts and Policy,” 1961. 38

[59] M. Inc., “fitlm: Create linear regression model.” 39

[60] S. Yashpal and S. C. Alok, “Neural Networks in Data Mining,” *Journal of Theoretical and Applied Information Technology* , 2009. 40

[61] A. Vesely, “Neural Networks in Data Mining,” 2003. 40

[62] D. L. Morgan, “Focus Groups as Qualitative Research,” 1996. 46

[63] Landgeflugel, “Basic Product Range - Edition 1.” 58

[64] S. Tuffery, *Data Mining and Statistics for Decision Making.* Wiley, 2011. 61

[65] D. S. Yates, D. S. Moore, and D. S. Starnes, *The Practice of Statistics.* Freeman, 2008. 61

[66] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning.* Springer, 2013. 63

[67] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley, “Application of high-dimensional feature selection: evaluation for genomic prediction in man,” 2015. 63

[68] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” 2003. 63

[69] R. R. Hocking, "The Analysis and Selection of Variables in Linear Regression," 1976. 63

[70] J. Mark and M. A. Goldberg, "Multiple regression analysis and mass assessment: A review of the issues. ," 2001. 63

[71] W. H. Greene, *Econometric Analysis.* Prentice Hall, 2012. 64

Appendix A

Poultry Processor Goals

Efficiency in production
Maximize end-product yield, decrease total running time of the production, minimize waste of resources as well as giveaway, etc.

Quality
Broilers are categorized into A, B and C depending on their presentation, meat quality and their limb condition (Ronnie van der Wijst, 2015). During the production process quality standards need to be met and quality checks performed to define the product quality (Marel, 2015).

Minimizing Costs
Reduce costs of the poultry processors in any way that would not harm the way it produces and generally conducts business.

Customer Satisfaction
Customers who are not satisfied submit their complaints to the poultry processor and if the poultry processor does not comply it translates with reduction on customer satisfaction. The poultry production process involves considerable amounts of waste disposal which needs to be recycled if not limited. Animal welfare organizations may complain for the way broilers are handled and the poultry processor needs to make sure these organizations are satisfied to the point that they wont affect customer satisfaction.

Punctuality
Right amount of the end-product at the right time. It also translates to order fulfillment.

Traceability
Trace and track products throughout the production process.

Improving Information Sharing
Better communication of important information between people and different departments of the poultry processor.

Innovation
New innovative ways to improve the processes of the poultry processor.

Competitiveness
Gain competitive advantage from other players in the poultry industry.

Accuracy in measurements
There is a small error in the accuracy of every measurement recorded by the poultry processor in the as-is situation. Improving the accuracy of the measurements will improve the production process as well.

Accuracy in predictions

This goal involves accuracy on the forecasts of the demand as well as accuracy on predicting the amount of fillet weight that is being produced. The latter is the goal the drives the current research.

Appendix B

Project Outline

Project Outline		
Phase	Activities	Report Chapter
Problem Investigation	Problem Statement Research Proposal Marel Stork Poultry Processing Orientation	Introduction
Business Understanding	Discussions with experts Business Goals and KPIs Data Mining Goals Project Plan	
Data Understanding	Data Collection Data Description Data Exploration Data Quality Verification	Data Processing
Data Preparation	Data Cleaning Data Transformation Data Preprocessing	
Modeling	Selection of Modeling Techniques Experimental Setup Model Creation Model Assessment	Data Mining Approach
Evaluation	Evaluation Results Interpretation Review	
Deployment	Conclusions Focus Group Limitations Future Work	Discussions & Conclusions

Literature Study

Table B.1: Project Outline

Appendix C

Poultry Product Range

Whole or Parts	Bone Removal	Preparation	Flavor	Post Processed
Breast (whole and split)	Bone in	Canned	Asian	Bacon
Breast chops	Boneless	Cooked	BBQ	Bologna
Breast cutlets		Breading	Buffalo style	Bratwurst
Breast scaloppini		Deep fried	Cajun	Breakfast sausages
Breast strips		Dry roasted	Citrus	Burgers
Breast tenderloins		Frozen	Dijon mustard	Dinner sausages
Drumsticks		Grilled	Hickory	Ham
Ground meat (lean/flat ratios)		Marinated	Honey	Kebab
Necks		Ready-to-cook	Honey smoked	Luncheon meat
Thighs		Roasted	Honey-pepper	Meatballs
Whole birds		Rotisserie like	Lemongrass	Nuggets
Wings		Smoked	Maple	Pate
		Sun dried with gravy	Mesquite	Patties
			Smoke	Salami
			Teriyaki	Sausages
			Zesty Italian	Sausage rolls
				Summer sausages

Table C.1: Poultry Product Characteristics [63]

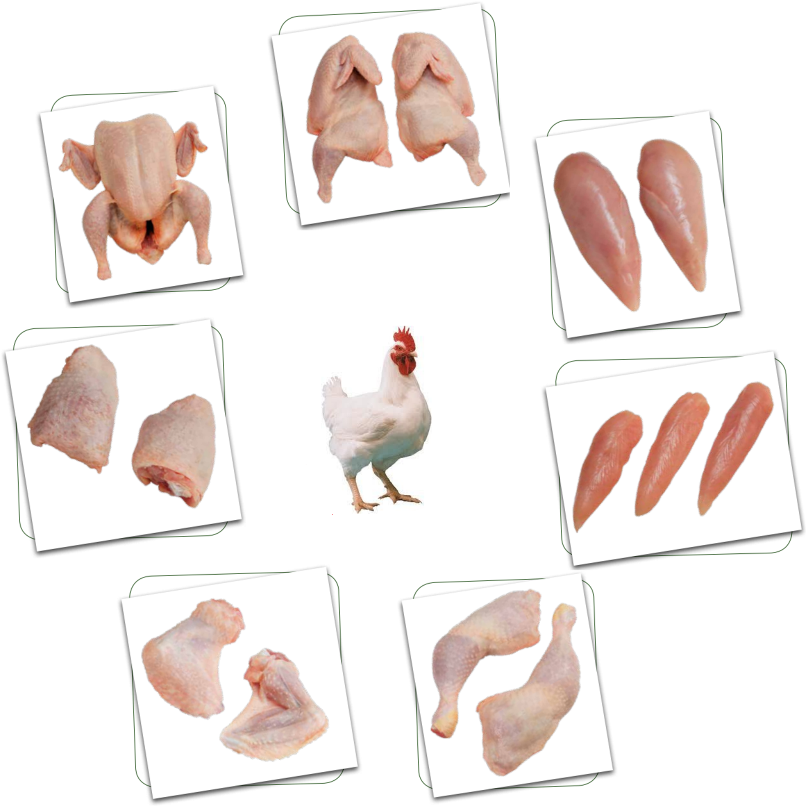


Figure C.1: Most Common Poultry Products

Appendix D

Poultry Processing - Detailed Process Steps

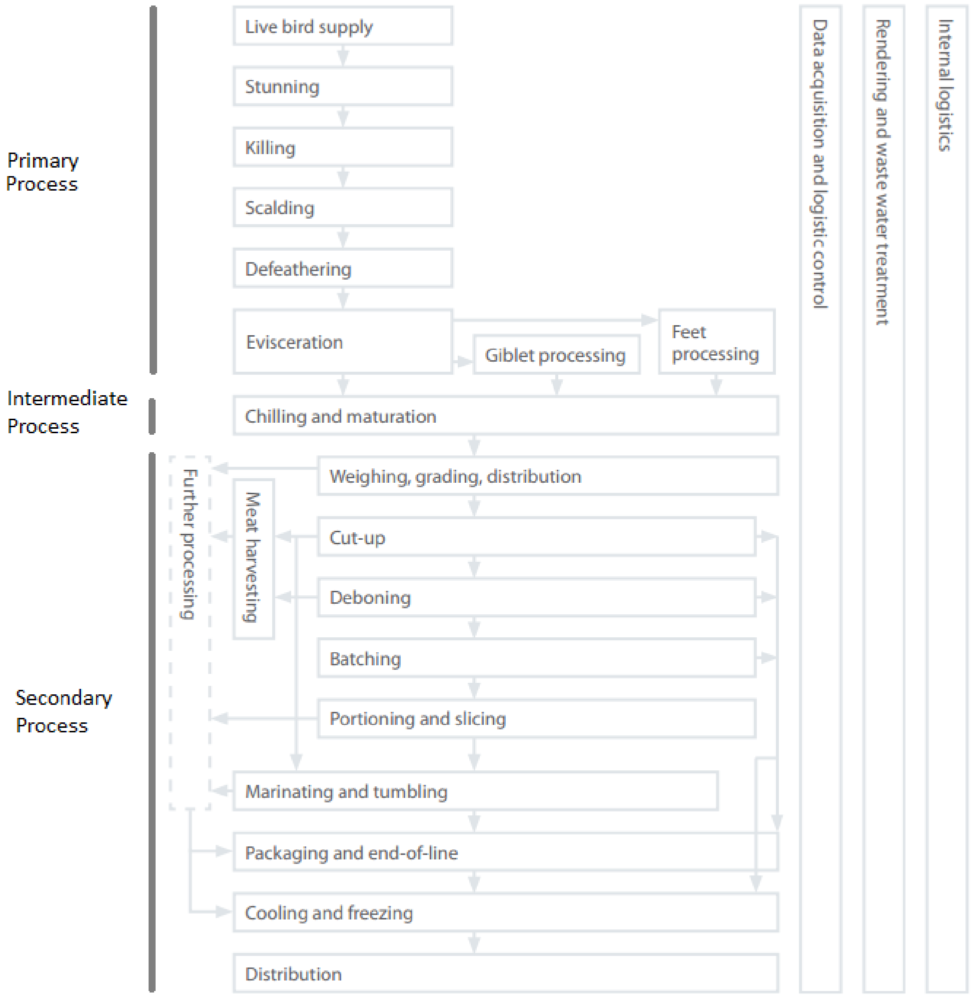


Figure D.1: Detailed Poultry Processing Steps [4]

Appendix E

Data Mining Definition

In the modern age, virtually all automated systems generate some form of data either for diagnostic or analysis purposes. Moreover, the World Wide Web (WWW) overwhelms us with information and meanwhile, most of the choices that people make are recorded [41]. The amount of data generated seems ever-increasing with no end in sight. However, computers make it possible to store a deluge of data that would otherwise be trashed [21]. This burst of data hides potentially useful information that if made explicit or taken advantage of it could lead to a better understanding in different aspects of our world [41]. Data mining provides the means to make sense of all this data.

There are more than one definitions for the term "data mining". Each one slightly different and less or more descriptive than the other ones. For the purpose of the current project the following definition is selected:

"***Data mining*** *is the set of methods and techniques used for exploring and analysing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies [64] (also known as patterns).*"

In a nutshell, data mining is the art of extracting knowledge from data, thus making it both descriptive and predictive:

- the descriptive (or exploratory) techniques are designed to bring out information that is present but buried in a mass of data. In the descriptive analytics, data is explored by looking at summary statistics or proportions for categorical variables, and then behavioral relation between variables can be observed by carrying out cross tabulations. A cross tabulation is a table usually in the form of a single variable such as gender in the rows against the target such as buy or not buy in the columns of the table and can lead to tests of importance of the single variable. If the target is independent of the single variable, then the single variable is not likely to be very important. If the target is related to the single variable, then it is potentially important in helping us to achieve the target [20].

- the predictive (or explanatory) techniques are designed to anticipate new information based on the present information. In the predictive analytics, we can carry out further analyses like regressions, decision trees, neural networks, etc. These analyses can be carried out with simple random samples of the data as well as with a stratified random sample. Simple random sample is a subset of individuals (sample) chosen from a larger set (population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals [65]. However, stratified random sample is a probability sampling procedure in which the target population is first separated into mutually exclusive, homogeneous subsets (strata), and then a simple random sample is selected from each subset (stratum). The samples selected from the various strata are then combined into a single sample.

The present project follows predictive techniques since the exact relationship between breast caps, fillets and other parts of the broilers body is unknown and extraction of this new information is attempted by building data mining models following certain techniques, such as regression, neural networks, etc. Data mining models that are properly built output the essentials of the useful information while contributing in the further reduction of the data quantity.

In general there are two main approaches in data mining:

E.0.3.1 Supervised learning

In supervised learning the training data (observations, measurements, etc.) is accompanied by labels indicating the class of the observations and the new data is classified based on the model that was built on the training set. The most well-known tasks associated with supervised learning are classification, association and regression. The latter will be used in the current research project.

E.0.3.2 Unsupervised learning

In unsupervised learning the class labels of the training data are unknown. The possibility of the existence of classes or cluster in the data is based on the set of measurements, observations, etc. The most well-known tasks associated with unsupervised learning are clustering and association.

Appendix F

Stepwise Feature Selection

Most of the times the dataset to be fed into the models consists of too many attributes and it may also be unclear which ones are the most relevant to derive the finest model possible. In such cases, feature selection is followed. Feature selection is the process of selecting a subset of relevant features (predictors) for use in model construction. It is an interesting and preferable approach since it simplifies the models for easier interpretation [66], it reduces the training times and enhances generalisation by also reducing data overfitting. The data usually contain many variables that may be redundant and can thus be removed without incurring much loss of information [67]. A variable may be relevant but in the presence of another relevant variable it may be redundant due to the strong correlation with the latter one [68].

Stepwise feature selection is a process implemented in regression models which picks the relevant predictors automatically. In the current research this takes the form of a sequence of adjusted R-squared tests [69]. In detail, the stepwise approach used is forward selection during which the process starts with no variables in the model and tests the addition of each variable using the adjusted R-squared as model comparison criterion. The variable that improves the model the most is added and the process is repeated until none other improves the model in a significant amount.

A way to test for errors in such stepwise regression models is to assess the model against a set of test data that are not used to train the model [70]. A technique called cross validation, which operates in a similar manner, is used for the present research and is explained in the following section.

Appendix G

Training and Testing

Multiple Linear Regression

- Non-Generalized Linear Model

Generalized linear model is suitable for non normal distributed variables. In the current case all the variables are normally distributed and non-generalized linear model is investigated instead, as seen in Table G.1. The latter model represents the least squares fit of the independent variable to the rest of the data, the predictors. A least squares estimator is selected because it is unbiased and among unbiased estimators it has the lowest variance, as mentioned in the Gauss-Markov theorem [71]. The inclusion of an intercept in the model slightly increases performance.

Metric	Value
Adjusted R-Squared	0,489
Root Mean Square Error	45,71
Mean Absolute Error	33,44
Mean Average Error	3,1

Table G.1: Non-Generalized Regression

- Stepwise Linear Regression Model

Forward stepwise linear regression is also used to determine a final model. The model trained contains an intercept and linear terms for each predictor. Predictors are added in the model when they increase the adjusted R-squared by at least 0,01 and removed when the increase is less than 0.001. Because of the total weight being such a high predictor for the breast cap the above limits had to be set so that more predictors, which are not strongly correlated to the total weight, are taken into account for the final model to increase the quality of the model. The list of the predictors selected for the final model are:

Total Weight, Area of Right Thigh, Front Distance from 4 to 5, Front Distance from 6 to 7, Back Distance from 6 to 7 and Front Ratio of the height with the Distance from 4 to 5.

The results delivered by the model are presented in Table G.2.

Metric	Value
Adjusted R-Squared	0,491
Root Mean Square Error	45,64
Mean Absolute Error	33,42
Mean Average Error	2,6

Table G.2: Stepwise Regression

- Robust Linear Regression

For the robust regression different weight functions, such as ”bisquare”, ”logistic”, ”fair”, etc., are tested with various different tuning parameters. A model with no weight function is also tested with the ordinary least squares function. The best results are produced with the ”talwar” weight function as mentioned in the main part of the report. The comparison of the results between the different weight functions are presented in Table G.3.

Type of Data	Metric	andrews	bisquare	cauchy	fair	logistic	welsch
Training sets	Mean Adjusted R-Squared	0,624	0,617	0,613	0,598	0,607	0,62
Training sets	Root Mean Square Error	37,2	38,9	38,8	39,1	39	39,1
Testing sets	Mean Absolute Error	32,9	33,1	32,75	32,89	32,79	32,78
Testing sets	Root Mean Square Error	46,19	46,22	45,93	45,82	45,92	46,12
Testing sets	Mean Average Error	4,19	4,21	2,99	2,44	2,61	3,71

Table G.3: Robust regression results

The model using ”andrews” weight function has the highest r-squared and rmse rendering it the fittest model of them all. Table G.4 compares the robust regression models which use ”andrews” and ”talwar” weight functions with a robust model which uses no weight function.

Type of Data	Metric	andrews	talwar	ordinary least squares
Training sets	Mean Adjusted R-Squared	0,624	0,701	0,483
Training sets	Root Mean Square Error	37,2	33,5	45,6
Testing sets	Mean Absolute Error	32,9	32,54	33,49
Testing sets	Root Mean Square Error	46,19	45,97	45,6
Testing sets	Mean Average Error	4,19	4,17	2,1

Table G.4: Final robust regression results

- Best Multiple Linear Regression Model

It is apparent that the robust model is performing better than the other ones. Additionally, when comparing the fit of the residuals in Figure G.1, it is apparent that the residuals are nearly all closer to the straight line for the robust fit.

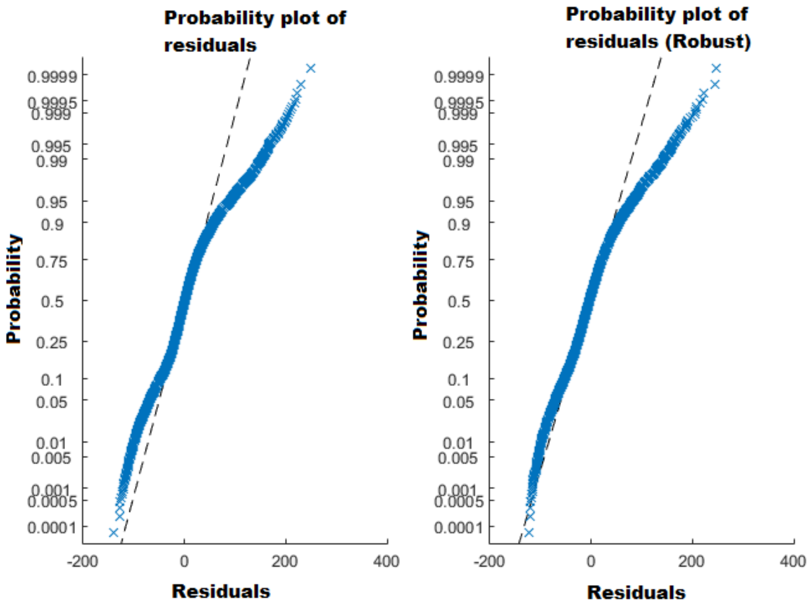


Figure G.1: Plot of residuals for non-robust and robust fit

Regression Tree

- Regression tree without pruning

Different value for the maximum splits of the tree are tested from 1 until 30. Splitting less than 10 times makes the tree overly simple while more than 15 times makes the tree large and unreadable without further pruning as can be seen in Figure G.1 and G.2. Figure G.3 illustrates the optimal tree output.

Variables of the model outputs (as introduced in Figure 3.5 and 3.6)

– x1 = Total Weight
– x18 = Front distance from 4 to 5
– x20 = Dront distance from 6 to 7
– x24 = Back distance from 6 to 7
– x32 = Front distance from 7 to 9

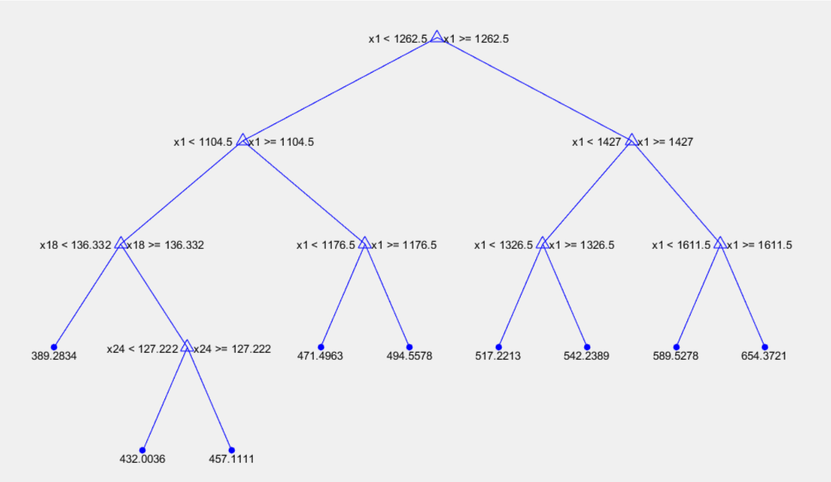


Figure G.2: Regression tree with 8 maximum splits

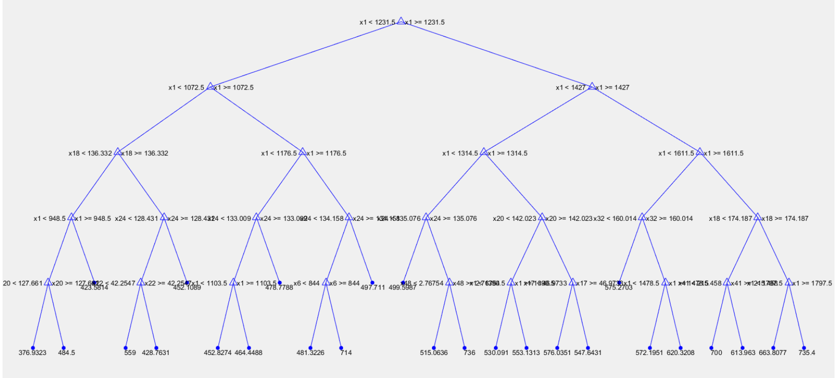


Figure G.3: Regression tree with 25 maximum splits

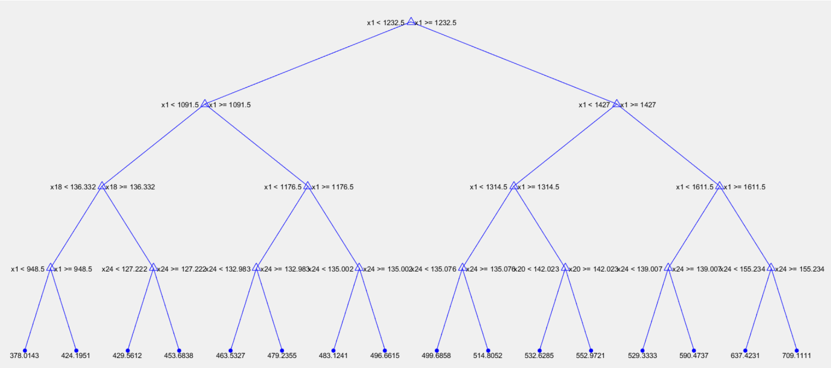


Figure G.4: Regression tree with 15 maximum splits

Table G.5 compares the different regression tree results.

Type of Data	Metric	8 splits	25 splits	15 splits
Training sets	Mean Adjusted R-Squared	0,448	0,44	0,455
Testing sets	Mean Absolute Error	35,67	35,34	35,07
Testing sets	Root Mean Square Error	47,35	47,46	46,95
Testing sets	Mean Average Error	1,19	1,31	1,08

Table G.5: Regression tree results (no pruning)

It is easy to understand that 15 maximum splits in the tree improves the performance of the tree.

- Regression tree with pruning

Same procedure is followed for pruned regression trees. This time different levels of pruning are tested as well which makes it possible to raise the number of maximum splits while keeping the tree in an optimal size. The best combination achieved can be seen in Figure G.4 and it has prune level 11 with 24 maximum splits.

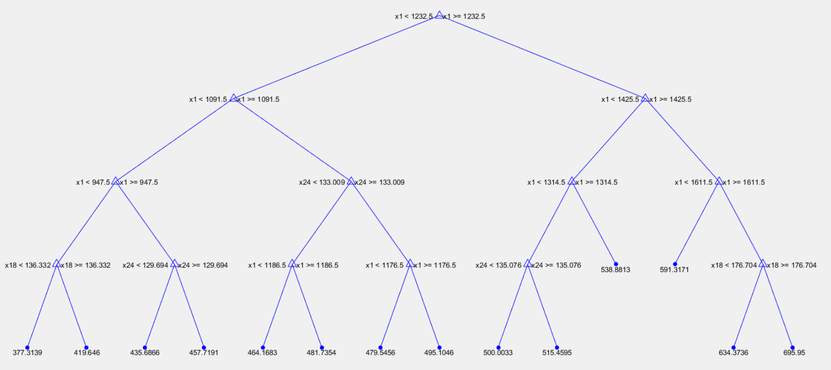


Figure G.5: Regression tree with 25 maximum splits

The results are listed in Table G.6.

Type of Data	Metric	Level 11 w/ 24 max splits
Training sets	Mean Adjusted R-Squared	0,477
Testing sets	Mean Absolute Error	34,99
Testing sets	Root Mean Square Error	46,77
Testing sets	Mean Average Error	0,72

Table G.6: Best pruned regression tree results

Neural Networks

Several training functions, that update weight and bias values via specific methods, are used to develop the final neural network model. These are listed below, according to the method they use to update weight and bias values:

- **Levenberg-Marquardt(lm)**: often the fastest backpropagation algorithm.
- **BFGS Quasi-Newton(bfg)**: alternative to the conjugate gradient methods for fast optimization but requires more storage and computation in each iteration than the conjugate gradient algorithms.
- **Resilient Backpropagation(rp)**

- **Scaled Conjugate Gradient(scg)**
- **Conjugate Gradient with Powell/Beale Restarts(cgb)**
- **Fletcher-Powell Conjugate Gradient(cgf)**
- **Polak-Ribire Conjugate Gradient(cgp)**
- **One Step Secant(oss)**: requires less storage and computation per epoch than the BFGS algorithm and slightly more storage and computation per epoch than the conjugate gradient algorithms. It can be considered a compromise between full quasi-Newton algorithms and conjugate gradient algorithm.
- **Variable Learning Rate Backpropagation(gdx)**: combines adaptive learning rate with momentum training.

All training functions are tested with 1 hidden layer of 10 neurons to derive the results documented in Table G.7.

Type of Data	**Metric**	**lm**	**bfg**	**rp**	**scg**	**cgb**	**cgf**	**cgp**	**oss**	**gdx**
Training sets	Mean Adjusted R-Squared	0,512	0,465	0,484	0,476	0,483	0,491	0,461	0,459	0,314
Testing sets	Mean Absolute Error	33,1	34,6	33,89	34,14	33,83	33,45	34,95	34,95	39,51
Testing sets	Root Mean Square Error	44,93	46,59	45,84	46,09	45,83	45,49	46,78	46,87	52,53
Testing sets	Mean Average Error	0,94	-0,18	-0,85	1,21	-0,31	-0,21	-0,11	-0,14	1,8

Table G.7: Neural Netowrk Training Functions

Several combinations of hidden layers and neurons are also tested on the best performing learning algorithm, Levenberg-Marquardt. The results presented in Table G.8 are based on the number of neurons on each of the two layers.

Type of Data	**Metric**	**10s**	**20**	**10/10**	**10/20**	**20/10**	**20/20**
Training sets	Mean Adjusted R-Squared	0,512	0,501	0,511	0,512	0,516	0,529
Testing sets	Mean Absolute Error	33,1	33,27	33,28	33,19	33,21	32,91
Testing sets	Root Mean Square Error	44,93	44,91	44,95	44,90	44,86	44,41
Testing sets	Mean Average Error	0,94	5,45	1,88	2,51	-0,45	-1,08

Table G.8: Neural Netowrk Training Functions

Although the differences are minor, it can be observed from Table G.8 that the best performing model is the one which contains 2 hidden layers with 20 neurons on each one of them. Regression plots are displayed in Figure G.6 to verify the best trained neural network model.

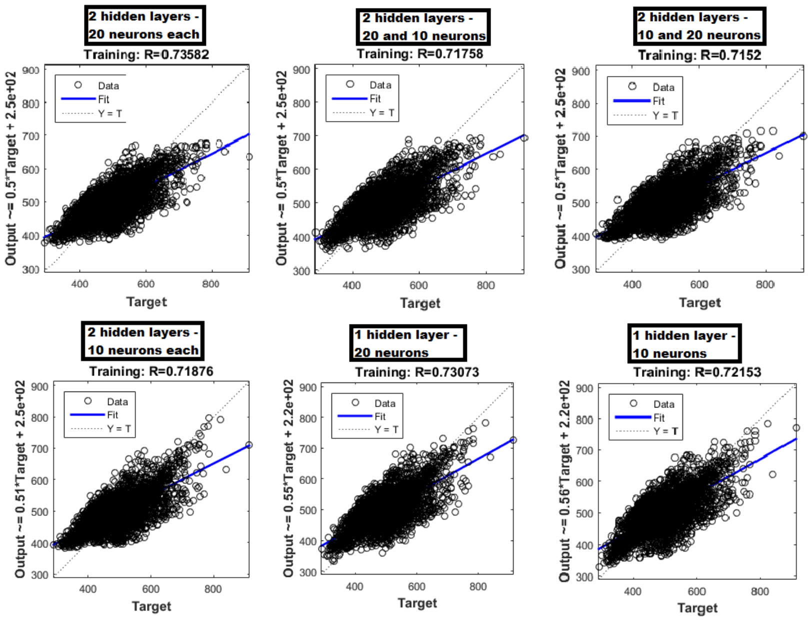


Figure G.6: Regression plots for different neural networks

Fuzzy Inference Systems

- Grid Partition (genfis1)

The use of grid partition raises the running time of the training exponentially with the add of each predictor. The maximum number of predictors to render the model trainable by the systems used in the present research is 3. The total weight of the broiler, the front distance from 4 to 5 and the back distance from 6 to 7 are the only predictors used to train a model using genfis1. These predictors appear to be higher predictors than the rest ones, after discussions with experts as well as the results of previously tested models from ther other data mining techniques tested. The results for different type and number of membership functions can be seen in Table G.9 and G.10.

Type of Data	Metric	Gaussian	Generalized bell-shaped	Trapezoidal-shaped	Triangular-shaped
Training sets	Mean Adjusted R-Squared	0,34	0,34	0,335	0,337
Training sets	Root Mean Square Error	44,27	44,28	44,31	44,3
Testing sets	Mean Absolute Error	42,43	42,45	42,65	42,55
Testing sets	Root Mean Square Error	58,83	58,87	59,07	58,99
Testing sets	Mean Average Error	14,62	14,64	14,65	14,7

Table G.9: FIS results for 3 membership functions

As expected 9 rules are created for each of the models.

Type of Data	Metric	Gaussian	Generalized bell-shaped	Trapezoidal-shaped	Triangular-shaped
Training sets	Mean Adjusted R-Squared	0,351	0,352	0,347	0,344
Training sets	Root Mean Square Error	44,14	44,13	44,22	44,19
Testing sets	Mean Absolute Error	42,07	42,11	42,23	42,27
Testing sets	Root Mean Square Error	58,37	58,33	58,52	58,67
Testing sets	Mean Average Error	14,84	14,64	14,73	14,72

Table G.10: FIS results for 5 membership functions

25 rules are generated for each of the models. In both cases, the generalized bell-shaped membership function delivers the best results with minor differences. The final results, as shown

in the tables above, are very poor and the running time increases with every additional predictor, which makes it even harder to approve the use of grid partitioning for the current research.

- Clustering

One of the clustering methods used to build the FIS is the subtractive clustering (genfis2), which is a fast, one-pass method that does not perform any iterative optimization. The model type for the FIS structure returned is a first-order Sugeno model with three rules. To further optimize the model anfis is used with 50 epochs.

The other clustering method used is the fuzzy c-means clustering(genfis3). This method creates a structure of sugeno or mamdani. Table G.11 compares the results of the different clustering FIS techniques.

Type of Data	**Metric**	**genfis2**	**genfis3-mamdani**	**genfis3-sugeno**
Training sets	Mean Adjusted R-Squared	0,508	0,266	0,497
Testing sets	Mean Absolute Error	33,45	43,28	33,55
Testing sets	Root Mean Square Error	45,64	56,78	45,69
Testing sets	Mean Average Error	-1,53	-2,56	0,91

Table G.11: Clustering FIS results

Appendix H

Data Mining Techniques Comparison

The error histograms of the best models from each data mining technique are presented in Figures H.1, H.2, H.3 and H.4.

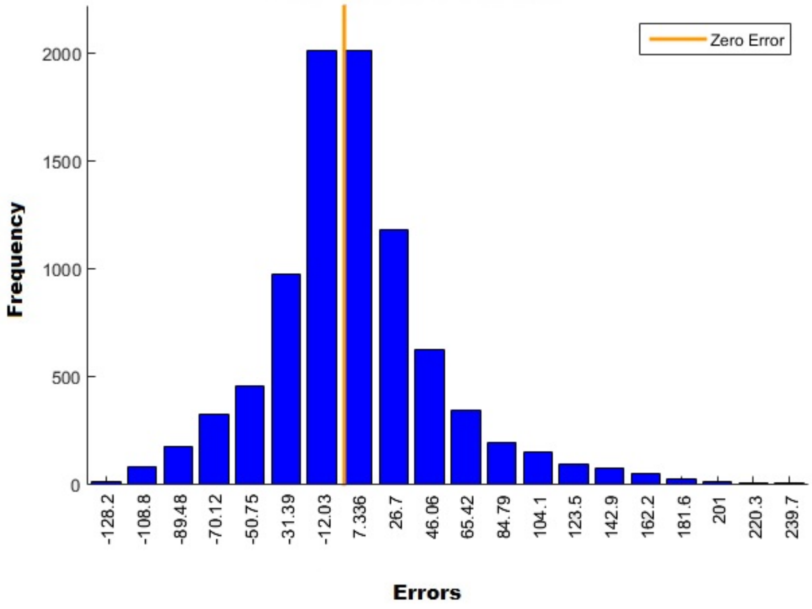


Figure H.1: Errors histogram for robust multiple linear regression

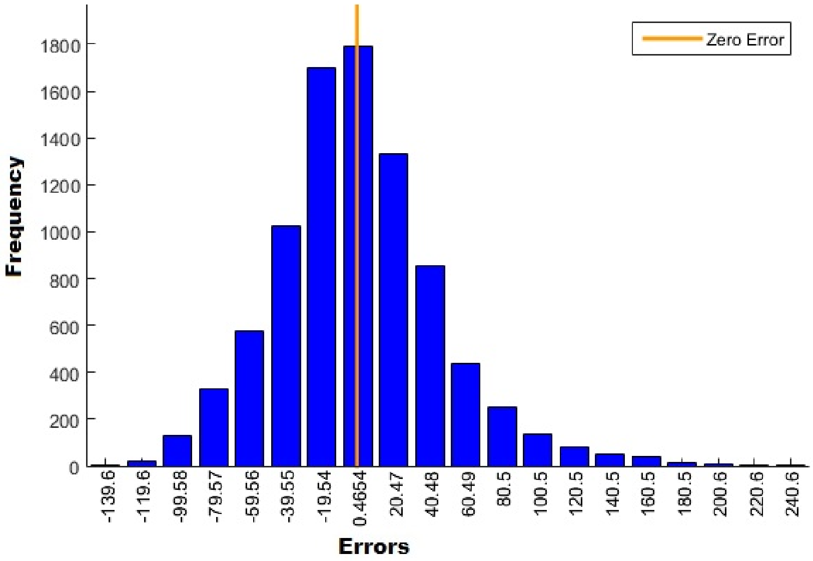


Figure H.2: Errors histogram for regression trees

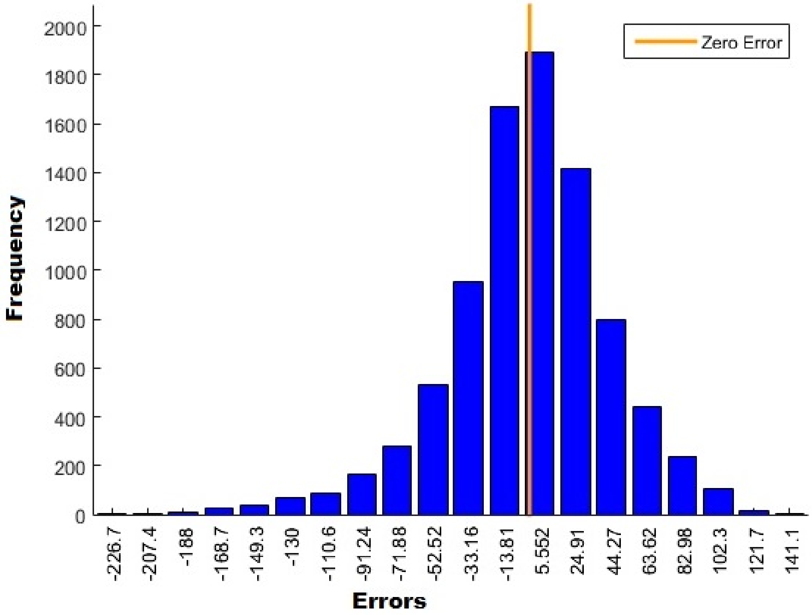


Figure H.3: Errors histogram for neural networks

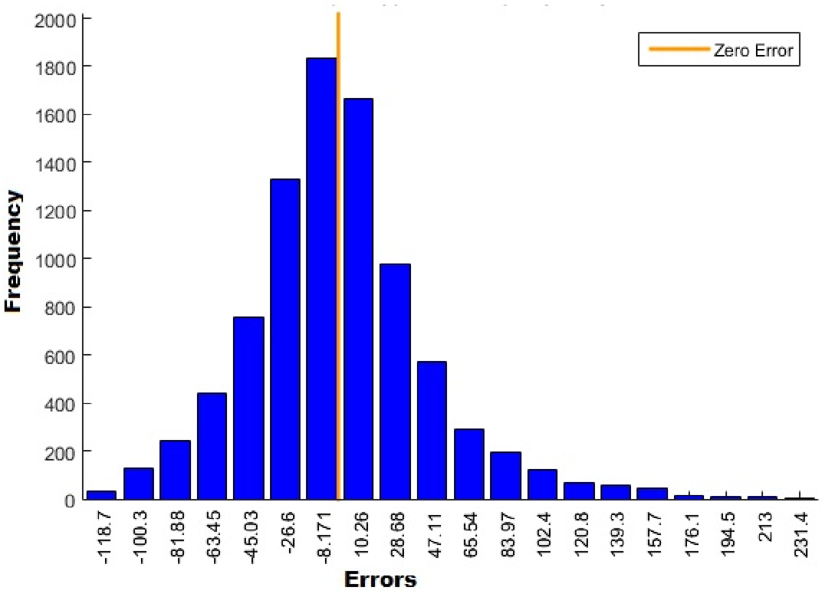


Figure H.4: Errors histogram for fuzzy inference systems

Figure H.1 shows that the amount of errors, for the robust regression, is higher for the values close to the orange line and less at the sides, which means that most errors done are close to zero. The same thing can be observed for the other techniques as well. However, the robust regression errors are more concentrated to the orange line.

Another comparison to be made is the histograms of real values with the predicted values of the breast cap weights in Figures H.5, H.6, H.7 and H.8.

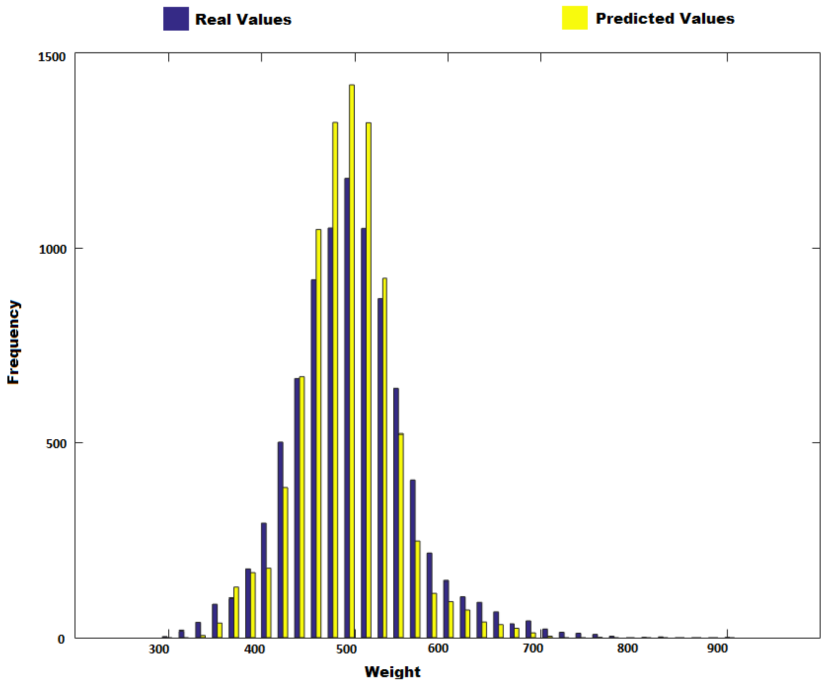


Figure H.5: Histogram of real and predicted values for robust regression

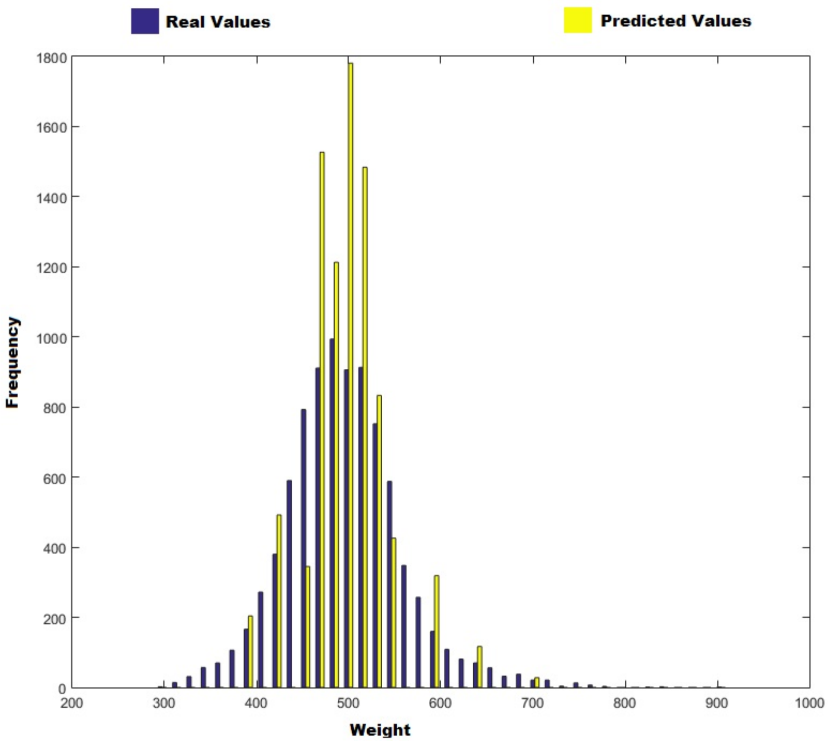


Figure H.6: Histogram of real and predicted values for regression trees

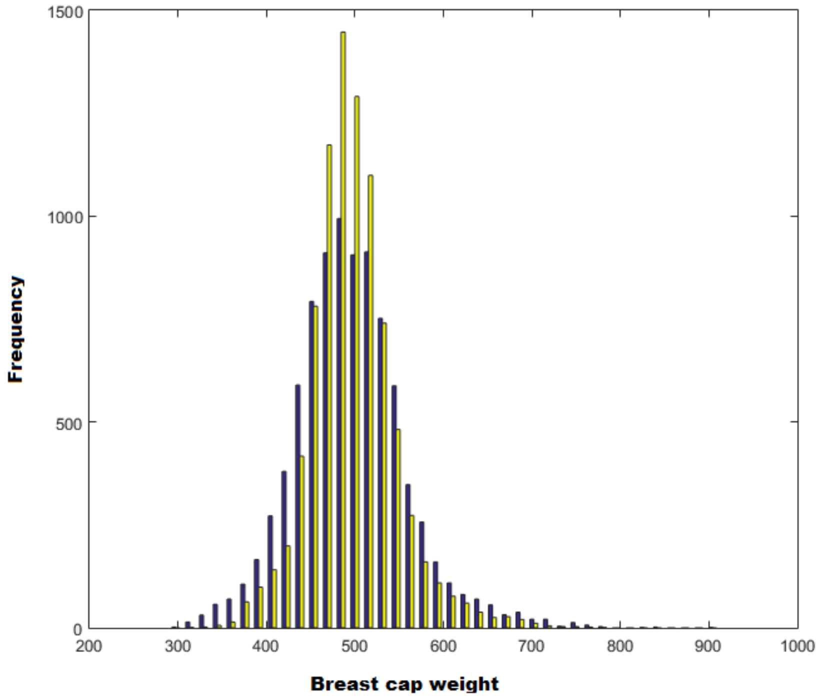


Figure H.7: Histogram of real and predicted values for neural networks

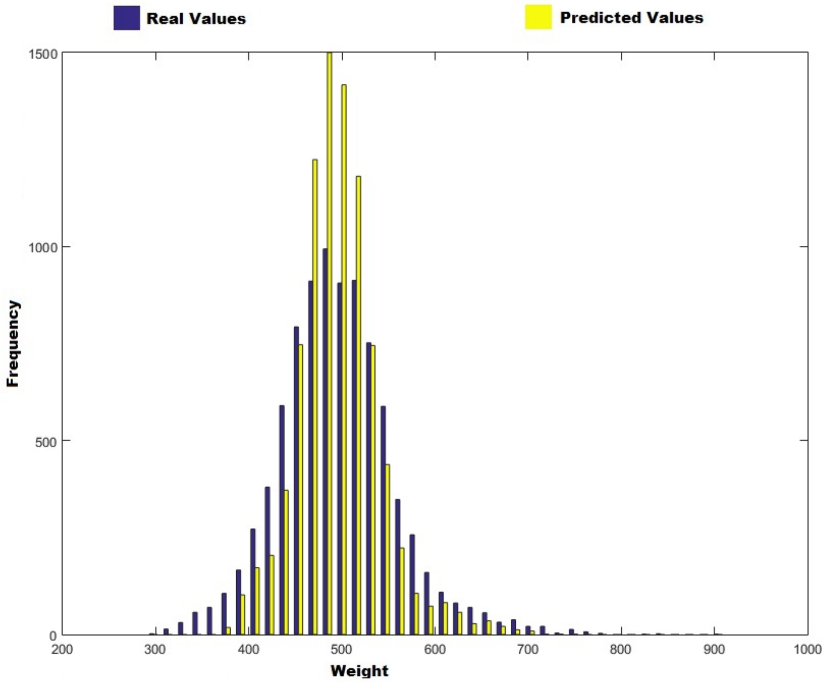


Figure H.8: Histogram of real and predicted values for fuzzy inference systems

Here it is easy to see that prediction of the breast cap weight is closer to the real values for the robust regression model than for the other three models. However, the prediction of the neural networks model follows the pattern of the real values sufficiently as well.

To better compare the predictions for each model separate histograms for the breast cap weights are provided. Figure H.9 illustrates the histogram of only the real breast cap values and Figures H.10, H.11, H.12 and H.13 illustrate the histogram of only the predicted breast cap values for each of the separate models.

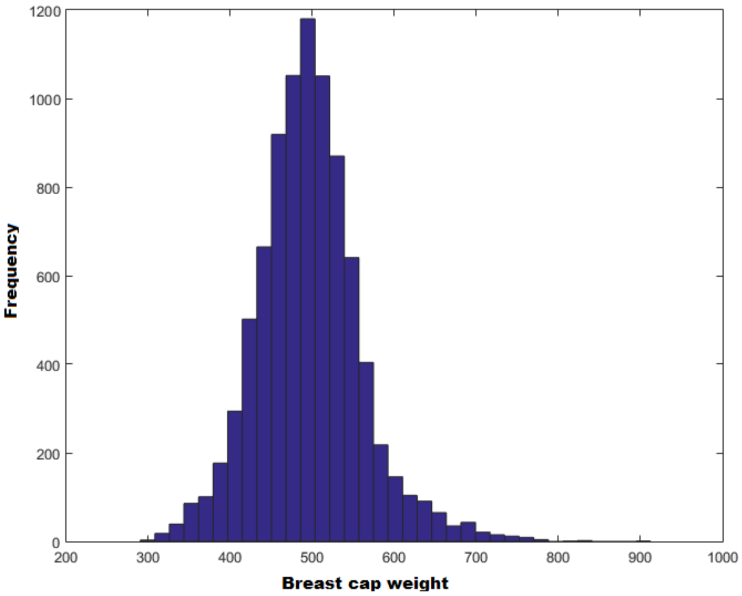


Figure H.9: Histogram of real breast cap weight values

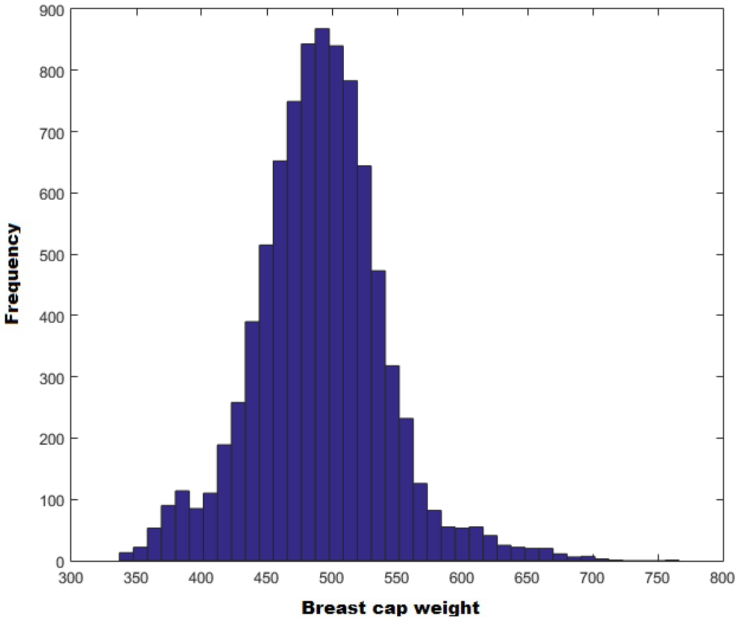


Figure H.10: Histogram of predicted values - Robust regression

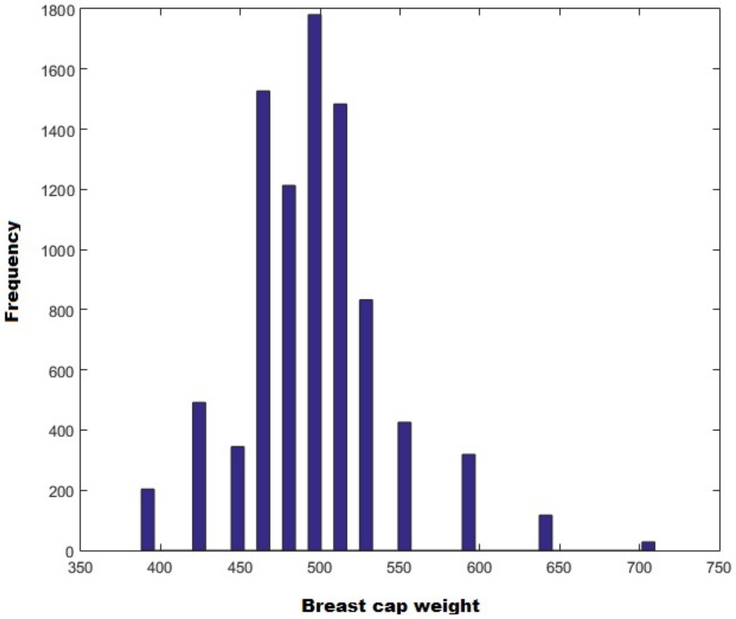


Figure H.11: Histogram of predicted values - Regression trees

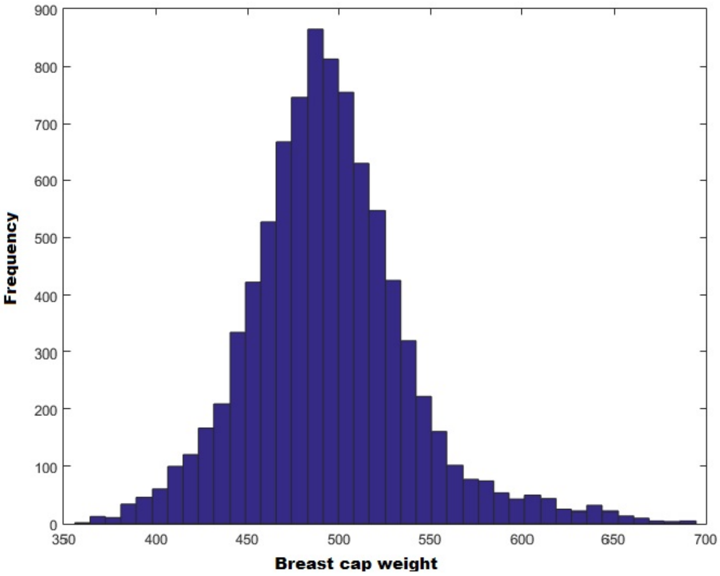


Figure H.12: Histogram of predicted values - Neural Networks

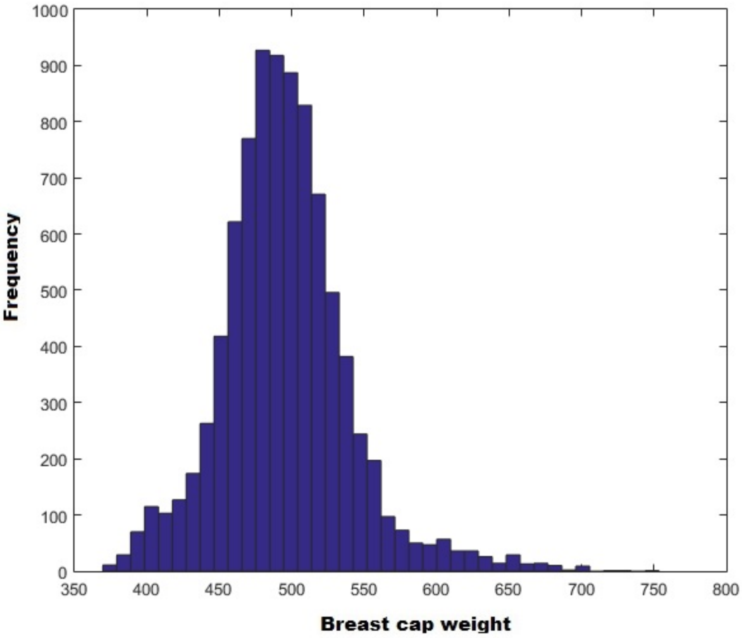


Figure H.13: Histogram of predicted values - Fuzzy Inference Systems

With the exception of the regression trees histogram (Figure H.11), figures H.10, H.12 and H.13 are pretty similar to Figure H.9. However the robust regression histogram, again, shows a slightly closer relationship with the real values histogram.

Appendix I

Flockwise prediction results

The results obtained when feeding separate flocks of the final data set into the trained robust regression model with predictors from the IRIS data and the total weight of the broiler.

Metric	Selected final model
Root mean square error	35,26
Mean absolute error	22,59
Mean absolute percentage error	4,6%
Mean average error	1,5
Sum real breast cap weight	555.308
Sum predicted breast cap weight	553.632
Mean real breast cap weight	494,9
Mean predicted breast cap weight	493,4

Table I.1: Results for Flock 7

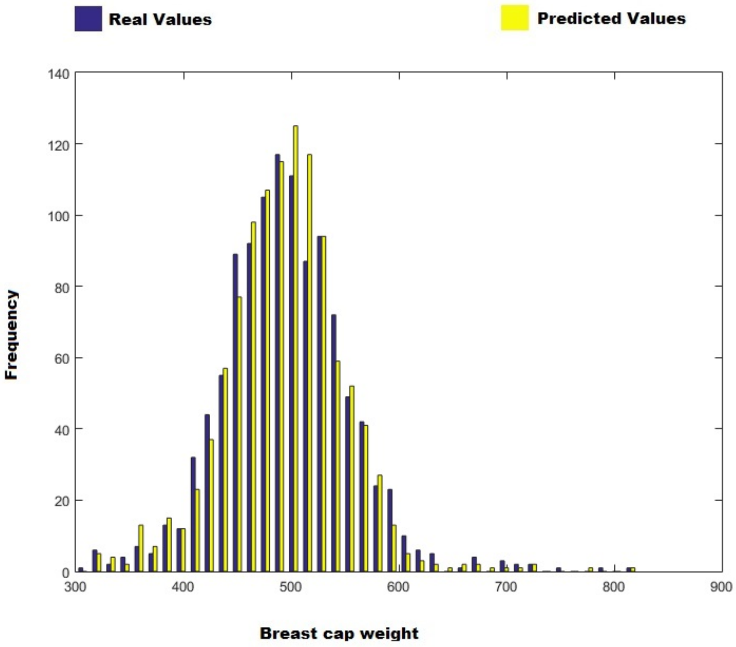


Figure I.1: Histogram of real vs predicted values (Flock 7)

Metric	**Selected final model**
Root mean square error	39,42
Mean absolute error	26,31
Mean absolute percentage error	5,6%
Mean average error	2
Sum real breast cap weight	1.099.922
Sum predicted breast cap weight	1.095.194
Mean real breast cap weight	472,3
Mean predicted breast cap weight	470,3

Table I.2: Results for Flock 8

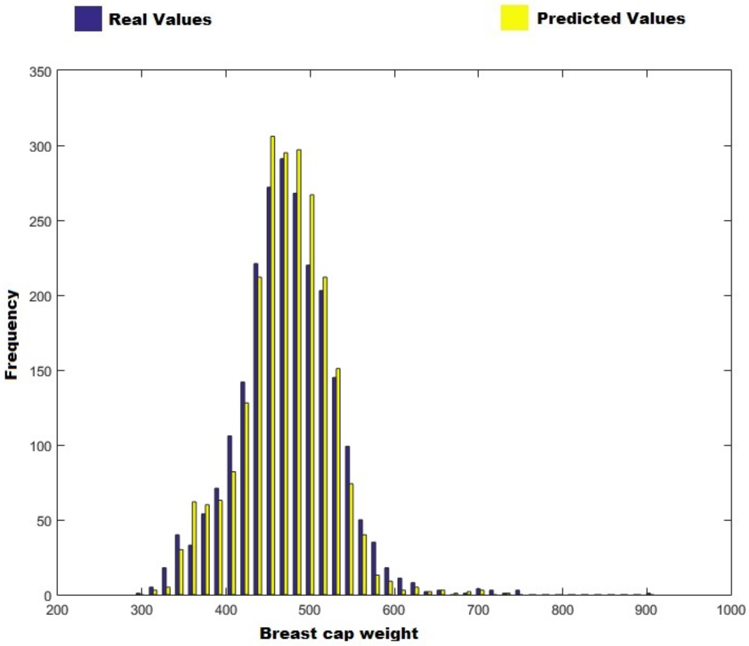


Figure I.2: Histogram of real vs predicted values (Flock 8)

Metric	**Selected final model**
Root mean square error	49,23
Mean absolute error	37,73
Mean absolute percentage error	8%
Mean average error	3,2
Sum real breast cap weight	905.176
Sum predicted breast cap weight	899.010
Mean real breast cap weight	477,4
Mean predicted breast cap weight	474,2

Table I.3: Results for Flock 9

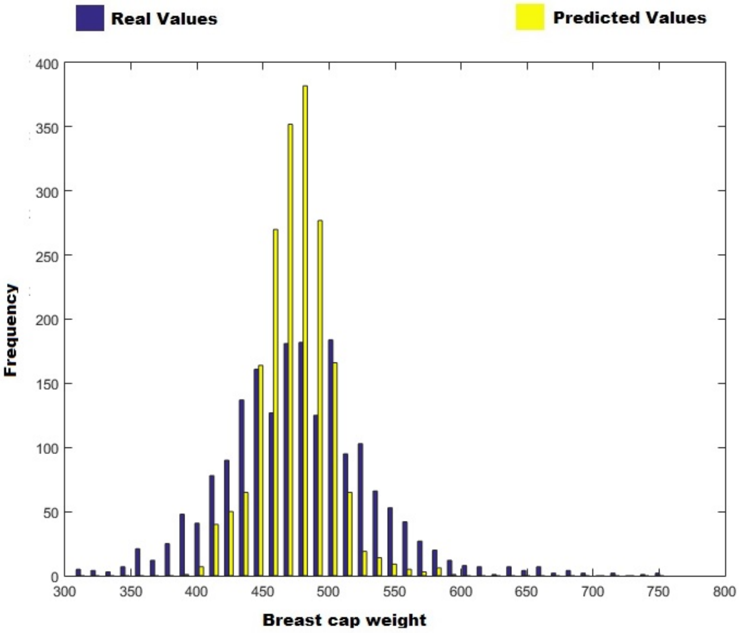


Figure I.3: Histogram of real vs predicted values (Flock 9)

Metric	Selected final model
Root mean square error	45,4
Mean absolute error	33,9
Mean absolute percentage error	6,6%
Mean average error	3,1
Sum real breast cap weight	1.330.436
Sum predicted breast cap weight	1.322.281
Mean real breast cap weight	511,5
Mean predicted breast cap weight	508,4

Table I.4: Results for Flock 10

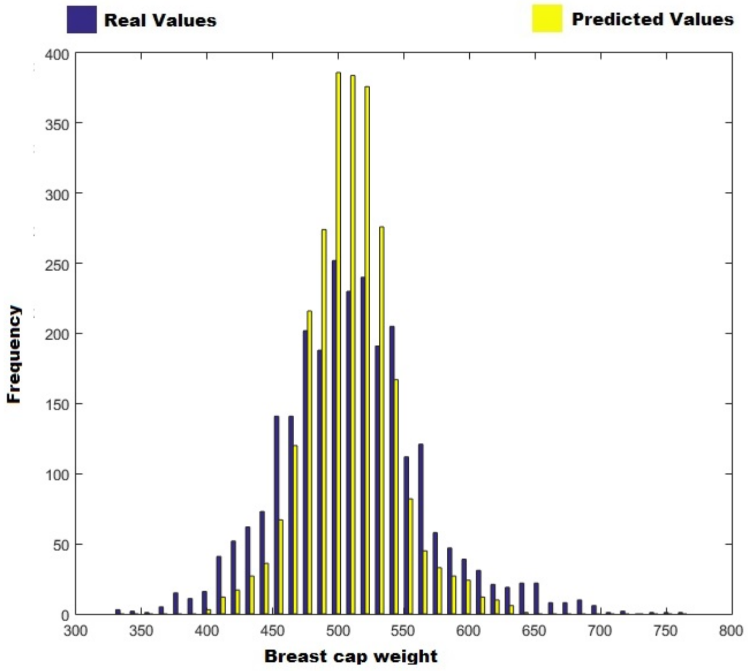


Figure I.4: Histogram of real vs predicted values (Flock 10)

Metric	Selected final model
Root mean square error	56
Mean absolute error	39,9
Mean absolute percentage error	7%
Mean average error	6,4
Sum real breast cap weight	453.888
Sum predicted breast cap weight	448.624
Mean real breast cap weight	552,8
Mean predicted breast cap weight	546,4

Table I.5: Results for Flock 11

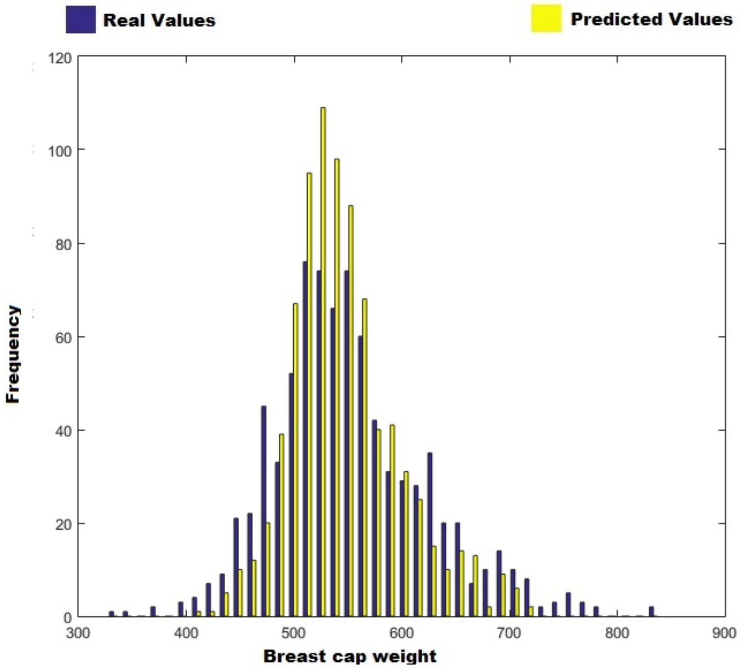


Figure I.5: Histogram of real vs predicted values (Flock 11)

Appendix J

Current approach vs Data mining approach

The final data set holds the information of 8.769 broilers with a mean total weight of 1.219 grams. The total weight of all broilers is 10.689.649 grams and the 4.344.740 grams correspond to the breast cap weight. The mean of the breast cap weight is 495.4 grams. If 58% of the breast cap weight corresponds to fillet weight, as derived from the internal project to Marel described in the literature study, then a comparison of the different prediction approaches can be realised. Table I.1 serves exactly that purpose. The current approach takes the average of a group of birds and assumes that 20%-25% corresponds to fillet weight. The data mining approaches are the robust regression models with different set of predictors.

Metric	**Weight + IRIS**	**Only Weight**	**Only IRIS**	**Current Approach(20%-25%)**
Total breast cap weight	4.304.872	4.298.440	4.301.365	3.685.090 4.607.503
Fillet weight	2.496.826	2.493.095	2.494.792	2.137.882 2.672.352
% of fillet weight	23,35%	23,32%	23,33%	20% - 25%
Mean average error	4,17	6,1	5	75,4 - 29,6
Mean Absolute Percentage Error	6,5%	7,8%	7,1%	NA
Mean Percentage Error	0,92%	1,24%	1,02%	16,3% - 7%
Mean Predicted Breast Cap Weight	490,9	489,3	490,4	420 - 525

Table J.1: Comparison of the prediction approaches

It is obvious that the robust multiple linear regression model with the total weight and the iris data as predictors performs a lot better than the current approach already implemented with higher accuracy.

Appendix K

Relationship of the IRIS data with the total weight of the broiler

To provide the company with an indication to its immediate next step regarding this project, another multiple regression model was trained, with the front and back IRIS measurements as predictors and the total weight of the bird as target variable. The results produced were impressive and can be seen in Table K.1.

Metric	**Non-Linear Regression Model**
R-squared	0.946
Mean absolute error	26,53
Mean absolute percentage error	2,2%
Root mean squared error	37,4
Mean average error	-0,4
Sum real total weight	10.689.649
Sum predicted total weight	10.693.301
Mean real total weight	1.219
Mean predicted total weight	1.219,4

Table K.1: Results for IRIS - Weight

The accuracy can be better observed in Figure K.1 where real and predicted weight values are included in the same histogram.

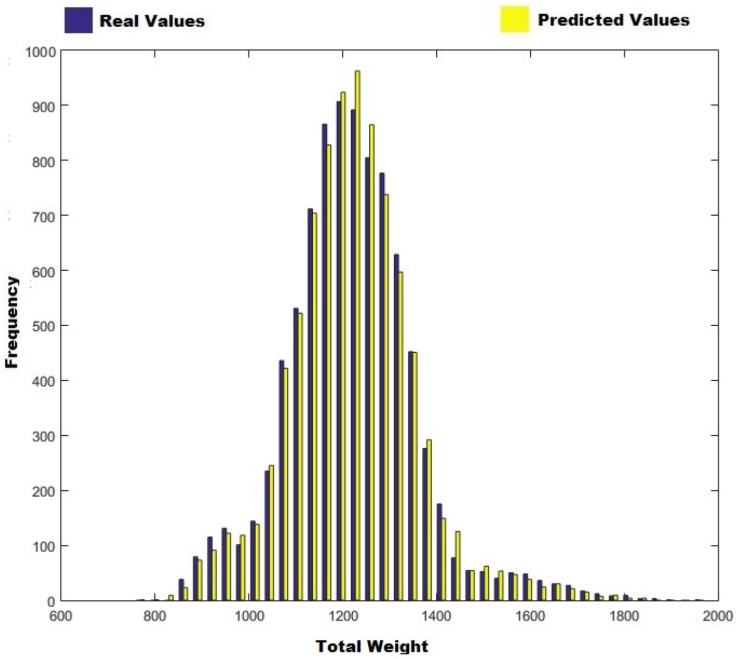


Figure K.1: Histogram of real vs predicted values (Total Weight)

Figure K.2 illustrates the histogram of solely the predicted total weights.

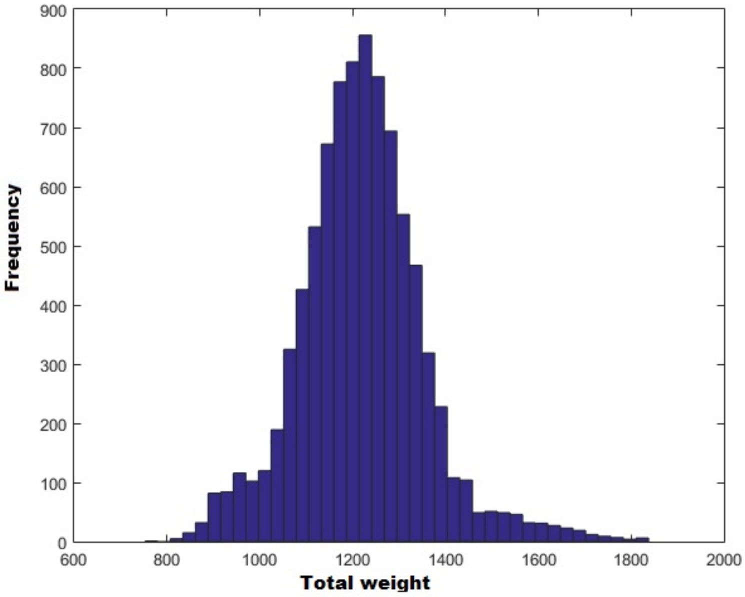


Figure K.2: Histogram of predicted values (Total Weight)

In Figure K.3 a histogram of the prediction errors is provided, where it is easy to observe that the errors are concentrated very close to the orange line, also known as the zero error line.

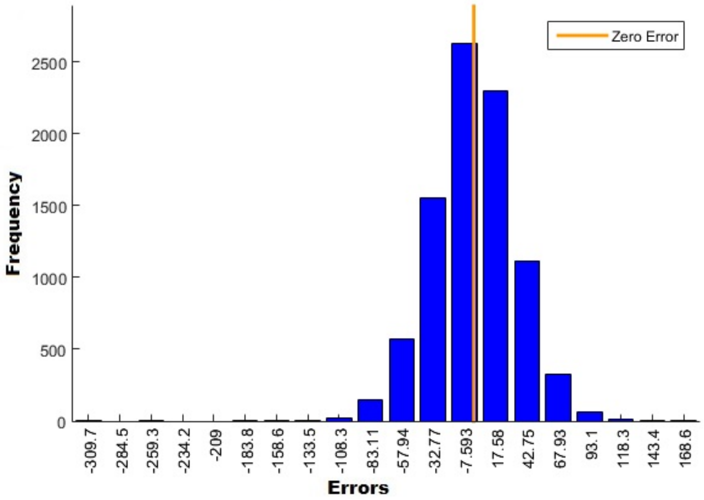


Figure K.3: Histogram of errors (Total Weight

Appendix L

Focus Group Protocol

L.1 Introductory Information

Below is the protocol for the Focus Group conducted in Marel Stork Poultry Processing. The purpose is to confirm the results of the research project as well as to explore on the usability of the final model. The focus group is recorded and run by a moderator and an assistant moderator, and it lasts for 2 hours.

Participants

- P1. Engineer 0-Series: Overall supervision of the discussion and support to the researcher.
- P2. Measurement and Control Engineer: Expertise on the IRIS system and relevant perspective towards the usability, and implementation of the model in the plant due to his frequent interactions with the production line of the poultry processor.
- P3. Poultry Logistics Specialist: Verification of the results of the model and comparison with the already used approach.
- P4. Manager in software development and smart handling: Head of the department that will provide the model to the poultry processor. Ability to offer an opinion in any of the subjects set for the present Focus Group.
- P5-P8. Experts from the INNOVA team: Experts working on various other projects of the company that will offer a different perspective to the table regarding mostly future research.

L.2 Main Part

Good morning and welcome to our session. Thank you for taking the time to talk with us about predicting the breast cap in broilers. My name is Antonis Mantzaris and I have been working with Marel Stork Poultry Processing to understand how certain data generated by a poultry processor can be of essential use to the company.

You were invited to this session because you are considered a relevant stakeholder in providing feedback on this topic. There are no wrong or right answers but rather different points of view. During the time well be here, I will ask you some questions, and I will listen to what you have to say. I will try to not participate that much, so please feel free to respond to each other and to speak directly to others in the group sharing your opinion even if it differs from what other people have said. Keep in mind that were just as interested in negative comments as positive comments and at times the negative ones are the most helpful. We want to hear from all of you that's why I may sometimes be encouraging someone, who has been quiet, to talk, or asking someone to hold off for a few minutes.

Youve probably also noticed the recording device. Were recording the audio of this session because we dont want to miss any of your comments. People often say very helpful things in such discussions and we cant write fast enough to get them all down. You may be assured of complete confidentiality and that your names are not going to be used in the report. The report will be delivered to Marel and a confidential version of it will be published in the library of TU/e.

I will kick things off by giving a short presentation about the solution we are offering. Please feel free to stop me if you have a question.

[Presentation]

Do you have any more questions?

Lets begin with introductions. Please tell us your first name, your background and what is your working position.

Now that we know a little bit about you, Id like you to think and discuss on some questions I have prepared for today.

Questions

1. What do you think of the approach presented?
   - Is this level of accuracy enough for you?
   - Does it bring value?
   - What is missing?
   - Is it better than what you had before?
2. How would it help the planners?
   - How this approach will be able to give an insight on the fillets?
   - Can it be used on fillets on a future time?
3. What should be the next step?
   - When is the best time to have this model in the process?
   - When do we collect the data?
   - When calculate?
4. In what way will the approach affect the order punctuality?
   - Right order at the right time?
   - Reduction of giveaway?

Do you have any other comments?

Thank you again for taking the time to participate in this discussion.

Appendix M

List of variables

Target Variable

- CWeight = The weight of the breast cap

Predictors

- FWeight = The total weight of the broiler carcass.
- FTotalArea = The total area the broiler from the front side, minus the wings.
- FAreaLegL = The area of the left leg of the broiler from the front side.
- FAreaLegR = The area of the right leg of the broiler from the front side.
- FAreaThighL = The area of the left thigh of the broiler from the front side.
- FAreaThighR = The area of the right thigh of the broiler from the front side
- FAreaBreast = The area of the breast of the broiler from the front side.
- FAreaWingL = The area of the right wing of the broiler from the front side.
- FAreaWingR = The area of the right wing of the broiler from the front side.
- BTotalArea = The total area the broiler from the back side, minus the wings.
- BAreaLegL = The area of the left leg of the broiler from the back side.
- BAreaLegR = The area of the right leg of the broiler from the back side.
- BAreaThigh = The area of a thigh of the broiler from the back side
- BAreaBreast = The area of the breast of the broiler from the back side.
- BAreaWingL = The area of the right wing of the broiler from the back side.
- BAreaWingR = The area of the right wing of the broiler from the back side.
- fdi1_10 = The distance from point 1 to point 10 on the front side.
- fdi2_11 = The distance from point 2 to point 11 on the front side.
- fdi8_12 = The distance from point 8 to point 12 on the front side.
- fdi9_13 = The distance from point 9 to point 13 on the front side.
- fdi4_5 = The distance from point 4 to point 5 on the front side.
- fdi8_9 = The distance from point 8 to point 9 on the front side.
- fdi6_7 = The distance from point 6 to point 7 on the front side.
- fdi4_8 = The distance from point 4 to point 8 on the front side.
- fdi5_9 = The distance from point 5 to point 9 on the front side.
- fdi5_6 = The distance from point 5 to point 6 on the front side.
- fdi5_7 = The distance from point 5 to point 7 on the front side.
- fdi3_7 = The distance from point 3 to point 7 on the front side.
- fdi6_8 = The distance from point 6 to point 8 on the front side.
- fdi7_9 = The distance from point 7 to point 9 on the front side.
- fdi3_6 = The distance from point 3 to point 6 on the front side.
- fbh = The height of the breast of the broiler on the front side.

- frh_45 = The distance from point 1 to point 10 on the front side.
- frh_89 = The ratio of the height of the breast with the distance from point 8 to point 9 on the front side.
- fr59_48 = The ratio of the distance from point 5 to point 9 with the distance from point 4 to point 8 on the front side.
- fr45_89 = The ratio of the distance from point 4 to point 5 with the distance from point 8 to point 9 on the front side.
- bdi1_10 = The distance from point 1 to point 10 on the back side.
- bdi2_11 = The distance from point 2 to point 11 on the back side.
- bdi4_5 = The distance from point 4 to point 5 on the back side.
- bdi8_9 = The distance from point 8 to point 9 on the back side.
- bdi6_7 = The distance from point 6 to point 7 on the back side.
- bdi4_8 = The distance from point 4 to point 8 on the back side.
- bdi5_9 = The distance from point 5 to point 9 on the back side.
- bdi5_7 = The distance from point 5 to point 7 on the back side.
- bdi1_6 = The distance from point 1 to point 6 on the back side.
- bdi2_9 = The distance from point 2 to point 9 on the back side.
- bdi1_8 = The distance from point 1 to point 8 on the back side.
- bdi4_6 = The distance from point 4 to point 6 on the back side.
- bbh = The height of the breast of the broiler on the back side.
- brh_45 = The ratio of the height of the breast with the distance from point 4 to point 5 on the back side.
- brh_89 = The ratio of the height of the breast with the distance from point 8 to point 9 on the back side.
- brh_67 = The ratio of the height of the breast with the distance from point 6 to point 7 on the back side.
- br59_48 = The ratio of the distance from point 5 to point 9 with the distance from point 4 to point 8 on the back side.
- br45_89 = The ratio of the distance from point 4 to point 5 with the distance from point 8 to point 9 on the back side.