**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_CA2* |
| **Lecturer Name:** | *Marina Iantorno/John O’Sullivan*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Naomi Tunstead* |
| **Student Number:** | *sba22222* |
| **Assessment Due Date:** | *6th January 2023* |
| **Date of Submission:** | *6th January 2023* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Predictive Modelling on Cattle Prices in Ireland and France**

**MSc in Data Analytics**

Author: N. Tunstead

e-mail: [sba22222@student.cct.ie](mailto:sba22222@student.cct.ie)

Student ID: sba22222

**Abstract**

**Keywords:**

Table of Contents

[Introduction 6](#_Toc123769678)

[Motivation 6](#_Toc123769679)

[Background 6](#_Toc123769680)

[Machine Learning 6](#_Toc123769681)

[CRISP-DM Framework 7](#_Toc123769682)

[Data Processing & Analysis 8](#_Toc123769683)

[Business Understanding 8](#_Toc123769684)

[Determine the Business Objectives 8](#_Toc123769685)

[Research Questions 8](#_Toc123769686)

[Project Plan 8](#_Toc123769687)

[Data Understanding 9](#_Toc123769688)

[Appendices 14](#_Toc123769689)

[Appendix A 14](#_Toc123769690)

List of Figures

Figure2: CRISP-DM Framework 8

List of Tables

# Introduction

## Motivation

The price of cattle cattlees is an important factor in the livestock industry, as it affects the profitability of farmers and the cost of meat for consumers. Accurate prediction of cattle prices can help farmers make informed decisions about their operations and enable meat processors to plan their production and pricing strategies. In this thesis, we aim to develop machine learning models to predict the price of cattle cattlees in Ireland and France. We will gather and analyse data on various factors that may influence cattle prices, such as year, week, feed price and cattle category. By training and evaluating multiple machine learning models on this data, we hope to identify the most effective model for predicting cattle prices in these two countries. The results of this study may be of interest to a wide range of stakeholders in the livestock industry, including farmers, meat processors, and policymakers. ADD SENTIMENT ANALYSIS IN HERE plus reword

# Background

## Machine Learning

A subset of artificial intelligence (AI) called machine learning enables computers to solve a variety of problems more frequently and accurately than a human being. For instance, algorithms have been designed to detect tumors in patients more precisely than doctors. The objective of machine learning is to create software that can identify patterns and rules in data and use them in unexpected contexts. This enables machines to infer solutions, doing away with the necessity for explicit programming. Many of the issues facing the modern world are complicated, making it difficult to quickly encode them into a computer and fix them in a binary manner. Machines may handle problems in a variety of ways, and these ways are always being improved through research and analysis. The three divisions of machine learning are supervised learning, unsupervised learning, and reinforcement learning. The main focus of this thesis is the use of supervised learning techniques. The supervised learning techniques are the main emphasis of this project.

A unified algorithmic framework called machine learning was developed to discover computational models that, with little to no human involvement, accurately describe empirical data and the phenomena that underlie it. Figure 1 shows a typical machine learning framework and highlights the key phases in their respective blocks. In order to understand and build applications using a machine-learning framework, we employ the techniques provided in the blocks. A dataset needs to be obtained, cleaned, and separated into training and testing datasets before anything else. The data must subsequently be modified or normalized according to the task at hand. The predictive model is then fitted using the training data. The appropriate model parameters for the method are determined using the training data. Following that, statistical tests are used to evaluate the prediction model's results to see if they are statistically significant and if they are compatible with the hypothesis. The prediction model can then be used and improved with new information. To establish a more cohesive structure, this is depicted in Figure 2. For more detail, please see Appendix A.

Diagram

Description automatically generated

#### Figure1: Simple Machine Learning Workflow

## CRISP-DM Framework

Diagram

Description automatically generated

#### Figure2: CRISP-DM Framework

# Data Processing & Analysis

## Business Understanding

### Determine the Business Objectives Reword this

In this thesis, our aim is to predict the price of cattle cattlees in Ireland and France using machine learning models. Accurate prediction of cattle prices is of interest to a variety of stakeholders in the livestock industry, including farmers, meat processors, and policymakers.

For farmers, understanding the expected price of their cattle cattlees can help them make informed decisions about their operations, such as which breeds to raise and when to sell their animals. Meat processors also have a vested interest in accurate cattle price prediction, as it can help them plan their production and pricing strategies. Policymakers may also find it useful to have accurate cattle price predictions in order to track trends and make informed decisions about regulations and policies related to the livestock industry.

In order to achieve our goal of predicting cattle prices, we will gather data on various factors that may influence cattle prices, such as breed, age, weight, and grade. We will then use machine learning techniques to build predictive models that can take this data as input and output a predicted cattle price. By evaluating the performance of these models, we hope to identify the most effective model for predicting cattle prices in Ireland and France.

### Research Questions

This research will answer the following research questions (RQs):

RQ1 - What are the data requirements for predicting cattle prices for Ireland and France?

RQ2 - How should machine learning be applied to best predict cattle prices?

RQ3 - How can visualisation aid in delivering insight about cattle prices from this predictive model?

### Project Plan

To ensure a smooth and efficient project process, the team will divide the work into the following phases:

Phase 1: Research the current situation and analyse the structure of the data (dataset examination)

Phase 2: Implement code for data representation (data exploration and visualization)

Phase 3: Prepare the data for modeling (selection, cleaning, formatting, and any other necessary actions)

Phase 4: Select modeling techniques and build the model

Phase 5: Analyse the results from the previous phase. Repeat phases 4 and 5 as needed.

Phase 6: Create a report with the results.



#### Figure x: Gantt Chart of Project Plan

### Version Control

GitHub is a web-based platform that provides version control and collaboration tools for software development. Git Bash was used to pull Jupyter notebooks up to GitHub online where each version of the code was stored. The link to my GitHub can be found in Appendix B

# Data Understanding

Any machine learning project must include the data understanding phase. In order to comprehend a machine learning model's traits and properties, it requires investigating and analysing the data that will be used to train it. This stage is crucial since the accuracy and usefulness of the data can greatly affect how well the machine learning model performs.

It's critical to spot any potential issues or difficulties with the data during the data understanding step, such as missing numbers, outliers, or unbalanced classes. Identifying the links and patterns in the data is also crucial since it can help in choosing the right machine learning algorithms and evaluation measures. Gaining knowledge of the domain or issue under study can also help direct the creation of the machine learning model and facilitate the interpretation of the findings.

Overall, the data understanding step is critical for making sure the machine learning model is constructed on a firm foundation of high-quality and relevant data, which is necessary for obtaining good performance and delivering trustworthy results.

### Collecting the Data

The process of acquiring raw data for my research project was both rewarding and challenging. On the positive side, I was able to access a wide range of data sources that helped me to understand my research question more deeply. For example, I was able to obtain data from public databases such as Eurostat (*Database - Eurostat*, no date), FAO (*FAOSTAT*, no date) and European Commission(Development, no date). This research provided me with valuable insights into my topic even though I didn’t end up using some of them.

However, there were also some negative aspects to my data acquisition process. One challenge I faced was the time it took to locate and access the data I needed. Some data sources were more difficult to find than others, and some required me to go through a lengthy application process before I could access the data. Additionally, I encountered some technical difficulties when downloading and organizing the data, which required me to spend additional time troubleshooting and cleaning the data. For these datasets I could not investigate them any further for this reason which wasted valuable time.

Overall, the process of acquiring raw data for my research project was a valuable experience. While it had its challenges, the data I was able to obtain helped me to make significant progress on my research and contributed to the success of my project.

The data for the predictive model on cattle prices was collected from the European Commission Agri-Food Markets. (Development, no date) The licence/permission can be found in Appendix C For the sentiment analysis it was obtained from twitter. I created a twitter developer account was granted an API key in order to interact with twitter data. (*Developer Agreement and Policy – Twitter Developers*, no date)

### Describe Data

The df\_cattle dataset contained 6150 rows and 8 columns. The columns included Year (2022, 2021), Week (1 to 52), Begin Date (date of the Monday of every week), End Date (date of the Sunday of every week), Member State (Ireland, France), Category (Young cattle, Young bulls, Steers, Heifers, Cows, Calves slaughtered <8M, Bulls), Product (letter and number grade given to cattle based on quality of cattle) and Price (per cattle)

The df\_feed dataset contained 240 rows and 7 columns. The columns included Marketing Year (2022/2023, 2021/2022, 2020/2021), Reference Period (date of the Thursday of every week), Member State (Ireland, France), Product Name (Feed Barley), Market Name (Rouen, Dublin/North East/Midlands), Stage Name, Price (€/Tonne)

In Table 1, we can see the df\_cattle descriptive statistics. Important statistic to take note of are the mean and the median. For the ‘Year’, ‘Week’ and ‘Price’ variables we can see both the mean and the median are roughly equal. Therefore, the data in the ‘Year’, ‘Week’ and ‘Price’ columns are symmetrical meaning our data is normally distributed. However, if we look at Table 2 for the df\_feed data frame ‘Price (€/Tonne)’ variable the mean is 261 and the median is 240, which are not equal. Therefore, the data is skewed in the direction of the mean. This is important to note as this will impact the statistical inference.

The range of the data in each variable is also an important statistic. For df\_cattle the ‘Year’ and ‘Week’ variables have a relatively small range of 1 and 51, respectively. However, for both ‘Price’ in df\_cattle and ‘Price (€/Tonne)’ in df\_feed both have a large range of 549 and 258. This tells us the mean may be heavily influenced by outliers or extreme values, which could make it less representative of the data. Therefore, from now on we may need to use the median value for our measure of central tendency.

The standard deviation is a measure of the dispersion or spread of a dataset. It is calculated as the square root of the variance, which is the average of the squared differences between the values and the mean. Here we can see for the ‘Year’ and ‘Week’ variable in df\_cattle the standard deviation is 0.5 and 14.7 which are small. Indicating that the values are more concentrated around the mean. However, for the ‘Price’ variable in df\_cattle and ‘Price (€/Tonne)’ variable in df\_feed have significantly higher standard deviations of 71 and 70, indicating that the values are more spread out. It's worth noting that the standard deviation is sensitive to outliers, or values that are significantly larger or smaller than the rest of the data. If there are outliers in the dataset, the standard deviation may be artificially inflated and may not accurately reflect the dispersion of the data. In such cases, other measures of dispersion, such as the interquartile range, may be more appropriate.

The interquartile range (IQR) is a measure of dispersion or spread in a dataset. It is calculated as the difference between the 75th percentile and the 25th percentile and is a robust measure that is less sensitive to outliers than the range or standard deviation. The IQR tells us the range of the middle 50% of the values in a dataset. For the for the ‘Year’ and ‘Week’ variable in df\_cattle the IQR is 1 and 25 this is small meaning the middle 50% of the values are concentrated around the median, indicating that the data is more homogeneous. However, for the ‘Price’ variable in df\_cattle and ‘Price (€/Tonne)’ variable in df\_feed have an IQR of 99 and 103, indicating that there is a significant amount of variability in the data.

Text

Description automatically generated

#### Table 1: Descriptive Statistics on df\_cattle

Text, table

Description automatically generated

#### Table 2: Descriptive Statistics on df\_feed

In Figure 3, we can see that our target variable cattle prices for both Ireland and France not only have different distributions with France having outliners and Ireland being very compact and more normally distributed. There is also an interesting significant difference between the cattle category produced in each country. Not only does France produce

A picture containing graphical user interface

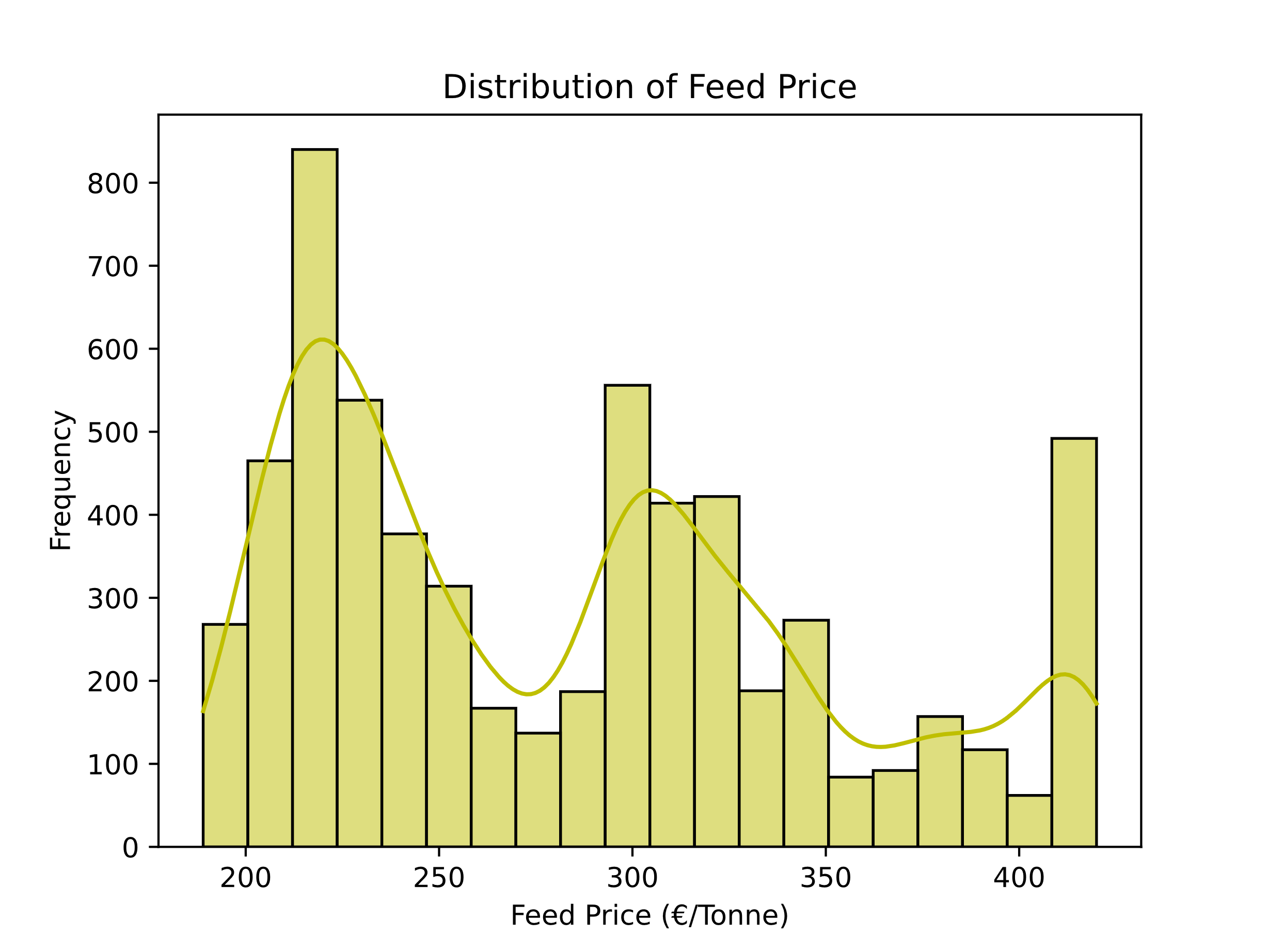
Description automatically generated

#### Figure 3: Violin plot of Distribution of target variable Cattle Price

Data Preparation

We observed that there were several columns we did not need. Such as in df\_cattle there were ‘End Date’ and ‘Product’, and in df\_feed there were ‘Year’, ‘Product Name’, ‘Market Name’ and ‘Stage Name’. None of these columns added any value to our analysis and therefore were dropped. We also noticed that the dates for both datasets did not match. We needed to resolve this in order for us to merge the two datasets. Using a lambda function we moved all the ‘Reference Period’ dates back three days so they are no longer on a Thursday but on a Monday so that the dates in both data frames matched. Note this may affect the accuracy of our analysis. We then renamed the ‘Begin Date’ column in df\_cattle to ‘Reference Period’ to match the column in df\_feed. This allowed us to complete a left join. Not only did this join both out datasets together it removed any rows in the df\_feed dataset we did not need. However, this left us with 317 missing values for ‘Feed Prices’.

There are several ways you can resolve this issue. Impute the missing values with the median, mean, backward filling, forward filling or even using machine learning such as MissForest. We choose to go with the simplest and most effective way, backward filling. We choose this method because it preserved the overall trend of the time series by replacing missing values with the most recent valid observation and kept the seasonality in the data.



#### Figure 4: Histogram of Distribution of Feed Price

Data preparation is a crucial step in the machine learning process, as it determines the quality of the data that is fed into the model. It involves a series of steps that are designed to clean, transform, and preprocess the data to make it suitable for analysis and modeling.

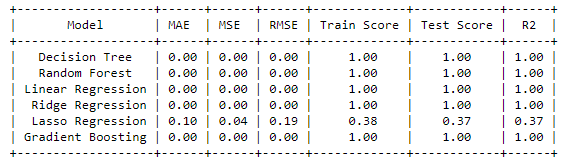
For example, consider a dataset that consists of customer reviews for a product. Before this data can be used for machine learning, it must be prepared. This may include tasks such as removing duplicates, handling missing values, standardizing text data, and encoding categorical variables.

Once these steps have been completed, the data is ready to be fed into a machine learning model. Without proper data preparation, the model may be unable to accurately learn from the data, leading to poor performance and unreliable results.

Table

Description automatically generated

#### Table x: Results Table for Each Model Before Grid Search CV was Applied



#### Table x: Results Table for Each Model After Grid Search CV was Applied

# References

*Database - Eurostat* (no date). Available at: https://ec.europa.eu/eurostat/web/main/data/database (Accessed: 5 January 2023).

*Developer Agreement and Policy – Twitter Developers* (no date). Available at: https://developer.twitter.com/en/developer-terms/agreement-and-policy (Accessed: 5 January 2023).

Development, E.C.-D.-G. for A. and R. (no date) *Agri-food data portal | Agricultural markets*. Available at: https://agridata.ec.europa.eu/extensions/DataPortal/agricultural\_markets.html (Accessed: 5 January 2023).

*FAOSTAT* (no date). Available at: https://www.fao.org/faostat/en/#data (Accessed: 5 January 2023).

*Legal notice* (no date). Available at: https://commission.europa.eu/legal-notice\_en (Accessed: 5 January 2023).

# Appendices

Appendix A: CRISP-DM Framework

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is a widely used method for structuring the process of data mining and machine learning. It consists of six main phases:

Business Understanding: This phase involves defining the problem or opportunity, understanding the context in which the problem exists, and identifying the goals and objectives for the project.

Data Understanding: In this phase, the data that will be used for the project is collected and explored. This includes understanding the structure of the data, identifying any issues or problems with the data, and generating some initial hypotheses about the relationships that may exist in the data.

Data Preparation: In this phase, the data is cleaned, transformed, and prepared for modeling. This may involve handling missing or incomplete data, normalizing data values, and creating new features or variables.

Modeling: This phase involves selecting and applying appropriate machine learning algorithms to the data in order to build predictive models.

Evaluation: In this phase, the performance of the model is assessed and evaluated using appropriate metrics and methods.

Deployment: This phase involves implementing the model in a production environment and monitoring its performance over time.

The CRISP-DM framework provides a structured approach for organizing the data mining process and helps ensure that all necessary steps are taken in order to build accurate and effective predictive models.

Appendix B: GitHub Link

https://github.com/sba22222

Appendix C: European Commission License/Permission  
“Unless otherwise indicated (e.g. in individual copyright notices), content owned by the EU on this website is licensed under the [license](http://creativecommons.org/licenses/by/4.0/). This means that reuse is allowed, provided appropriate credit is given and changes are indicated.”(*Legal notice*, no date)

Appendix: Word Count

X number of words without taking code, code comments, titles, references or citations into account