A Comprehensive Study on Supply and Disappearance of Food Grains in USA

MSc Research Project
Data Analytics

Muskaan Kapoor
Student ID: x22105476

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Muskaan Kapoor
Student ID:	x22105476
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	14/12/2023
Project Title:	A Comprehensive Study on Supply and Disappearance of Food Grains in USA
Word Count:	7256
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Muskaan Kapoor
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comprehensive Study on Supply and Disappearance of Food Grains in USA

Muskaan Kapoor
x22105476
MSc Research Project in Data Analytics
National College of Ireland

Abstract

Agriculture stands as a pivotal sector in the global economy, serving as the primary source of sustenance for populations across the world, as a fundamental component of human survival, the availability and distribution of diverse food products hold paramount importance. This research is focused on analyzing the dynamics of food grain supply and its utilization within the United States, a nation recognized as a key player in the global agricultural market. The study is crucial for effectively managing the balance between the supply and demand of food grains such as Corn, Barley, Sorghum, and Oats, a task of critical importance for agricultural stakeholders, policymakers, and economic strategists. The research employs advanced machine learning methodologies, encompassing regression techniques such as Gradient Boosting, Bagging, Random Forest, AdaBoost, and K-Neighbours Regressor that delve into the patterns and trends of food grain demand. The findings from this study are anticipated to provide strategic insights into future demands for food grains, thereby facilitating informed decisions to ensure adequate supply, averting potential deficits or excesses in the market. Machine learning models can enhance supply chain visibility, improve pricing models to benefit both producers and consumers, develop predictive food waste models to curb losses, and aid breeding techniques for climate resilience.

Keywords— food grains, supply, disappearance, machine learning, regression models

1 Introduction

In the vast and dynamic arena of global agriculture, the United States emerges as a formidable entity, notably in the cultivation and distribution of key food grains. This nation's pivotal role in the global agricultural sector encompasses a broad spectrum of grains, including, but not limited to, corn, wheat, barley, and sorghum. These grains not only cater to domestic needs but also traverse international borders, feeding populations worldwide and anchoring the global food supply network. The supply and disappearance of food grains in the United States are critical factors in assessing the country's food security. The food availability data estimate the amount of food available for human consumption in the United States by measuring the supply of several hundred foods moving from production to marketing channels.

1.1 Background

For each commodity, the U.S. Department of Agriculture (USDA) calculates the residual of the total annual supply available by subtracting measurable uses, such as farm inputs, exports, ending stocks, and industrial uses. This data series also provides per capita availability data for hundreds of commodities, and it is a popular proxy for food trends and the only source of time series data on U.S. food availability in the country. The disappearance of food grains, or domestic disappearance from supply during a year, is estimated through supply and use balance sheets for each major commodity from which human foods are produced. It includes the aggregate of ending stocks, exports, food use, and an estimate for farm and industrial use Golan et al. (2004). Recent times have seen unprecedented disruptions in the global food grain supply chain. Factors such as the COVID-19 pandemic, geopolitical tensions like the Russia-Ukraine and Israel-Palestine conflict, and a series of natural calamities ranging from floods to earthquakes have considerably strained the global food grain logistics. The term ”disappearance” in this context refers to the utilization of these grains within the U.S. for various purposes, including human consumption, as well as feeding livestock and poultry, and in industrial applications.

1.2 Aim and Motivation

This study aims to understand the factors contributing to food loss and waste (FLW) at various stages of the food supply chain, scientists, governors, and policymakers can focus on future implications and develop strategies to minimize losses. By using historical data, the study seeks to identify and interpret patterns, trends, and anomalies in the production, consumption, and overall management of various food grains. The research will utilize advanced machine learning techniques and in-depth statistical analysis, with a focus on the Knowledge Discovery in Databases (KDD) process, to uncover deep insights and support informed decision-making. The primary motivation for this research is the critical need to ensure strong and efficient food grain supply systems, which is essential for national food security and the United States’ global food market leadership. Understanding the complexities of food grain supply and disappearance is also crucial for developing strategies to mitigate the impacts of supply chain disruptions, market fluctuations, and changing global food demands.

1.3 Research Question

There is limited research done on this topic at present therefore, this study will identify the most efficient way to calculate the supply and disappearance of food grains in the USA this will not only benefit the consumers but also the farmers, policymakers, economists, and government in improved decision making. The evaluated results can be further used to find the demand for Corn, Barley, Sorghum, and Oats.

RQ: Is machine learning the best approach to forecasting food grain (Corn, Barley, Sorghum, and Oats) trends in the U.S.A.? If yes, which model can generate the best results to estimate the food demand in the following years?
Sub-RQ: To what extent machine learning models can give accurate results in predicting food grain supply and demand in the USA? To address the primary research question and its related sub-research question, the following goals have been executed and accomplished.

In order to address the primary research question and its related sub-research question, the following goals have been executed and accomplished.

Objective 1: Examine and evaluate the literature on the supply and demand for food over the previous ten years.

Objective 2: Design an implemented structure that can predict the food grain forecasting.

Objective 3: Performing relevant data-preprocessing, cleaning and data transformation on the dataset to avoid any anomalies in the results.

Objective 4: Evaluation and Comparison of the built models to find best suited model for forecasting.

The structure of the report is as follows: Critical evaluation of the literature related to supply and disappearance of food grains that utilized similar approaches in Section 2. Section 3 and Section 4 will address KDD research methodology and design specification steps followed for this study. The detailed implementation techniques and models implemented are discussed in Section 5. Following that Section 6 will present the results and evaluation of the findings in the implementation section. Finally, Section 7 will discuss the conclusion and future scope following the acknowledgement and references.

2 Related Work

The adequate supply of food and consumption of grains have tremendous implications on the lives on an individual, communities and nations. It not only affects nutritional value of an individual but also global economy and food stability globally. The literature study will delve into the previous research, finding the past trends and patterns, shedding light on the impacts and the technological progress. The examination will encompass the different time frames and outcomes of the studies performed. The section below will consult a few studies that have used comparable methodologies because there isn't much research that directly relates to this one.

2.1 Related Work using Deep Learning and LTSM

The study Abraham et al. (2020) highlighted the importance of soyabean as a major source of feeding, ranking sixth in production among various agricultural crops. The main goal of the study is to predict soyabean harvest are, yield and production of the crop. The study utilised the dataset from 1961 to 2016, mainly focusing on Soyabean production in Brazil. The study does a significant comparison between Artificial Neural Networks and classical Time Series Analysis methods where ANN technique is employed for the prediction and the results are then compared with classical Time Series Analysis. The findings suggest that ANN is the best approach for predicting soyabean harvest area and production whereas classical linear function is more effective in predicting soyabean yield. It also recommends looking at hybrid systems, which combine fuzzy logic and neural networks with other techniques to get better results.

The paper by Muruganantham et al. (2022) presents a systematic literature review on the use of deep learning method for crop yield prediction using remote sensing data, it aims to understand the current state of research in this area and identify any gaps. The most common deep learning techniques found were convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, these were often combined into CNN-LSTM models to leverage the strengths of both approaches. Satellite remote sensing, especially with MODIS data, was the most important data source, vegetation indices like NDVI along with meteorological measurements were the most utilized input features. The study showed LSTM models are well-suited for incorporating temporal dynamics in the data, meanwhile CNNs can effectively extract geometric patterns from satellite images that correlate to crop yield.

On the other hand, research was conducted in 2020 Sabu and Kumar (2020) explores the use of predictive modelling techniques to forecast arecanut prices in Kerala, India. Arecanut is an important commercial crop, but its prices fluctuate significantly over time. The researchers employed three machine learning algorithms - random forest regression, support vector regression, and LSTM neural networks. Historical price data from 2005-2019 served as the model input. Various performance measures were used to evaluate and compare the methods including R-squared, RMSE, MAE, and MAPE. This research provides insights on leveraging AI to predict agricultural commodity prices. LSTM networks seem particularly well-suited for the complex, dynamic patterns in crop price data.

The article by Marndi et al. (2021) emphasize on using deep learning models to estimate crop production. The authors developed a LSTM-based recurrent neural network (RNN) to predict rice yields in Odisha, India using time-series data on previous year's production from 1990-2017. The study uses preprocessing steps like min-max normalization and splitting data before training the LSTM network. Various combinations of training and test splits were evaluated, with an 80-20 split found optimal. The proposed LSTM model achieved high accuracy, with a best test R2 of 0.985. This significantly outperformed a baseline multi- layer perceptron model. The time-series forecasting capabilities of LSTM are well-suited for modelling temporal crop yield patterns.

2.2 Food Supply and Demand using Time Series

The investigation Devi et al. (2021) utilizes wheat production data from 1980–81 to 2018–19, applying both the Box-Jenkins ARIMA model and Artificial Neural Network (ANN) methodologies. A hybrid approach combining these models is also employed. The investigation observed an increasing trend in the area, production, and yield of wheat crop in Haryana. The growth rate was positive across different sub-periods, with the highest growth in production and yield during the first sub-period (1980-1989). Among various ARIMA models tested, ARIMA (110) with drift was found to be the best for modelling wheat production. ANN also showed competence in forecasting the production behaviour.However, the hybrid model combining ARIMA and ANN outperformed both in terms offorecasting accuracy.

The research by Taylor et al. (2009) investigates methods for predicting wind power density. The study focuses on forecasting the probability density function of wind powerat five U.K. wind farm locations, looking ahead from one to ten days. The study used weather ensemble predictions from atmospheric models and statistical time series techniques for density forecasting of wind energy. The methods included applying autoregressive (AR) models, generalized autoregressive conditional heteroskedasticity (GARCH) models, and long-memory time series models to daily wind speed data. Additionally, neural networks were considered for calibrating atmospheric model predictions. The hybrid approach, involving calibration and smoothing of ensemble-based wind power density, yielded the best results. The study compared various methods using metrics like the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The calibrated ensemble model outperformed others, especially for short to medium-term forecasts

2.3 Supply and Demand of food grains using Machine Learning models

The study Nagar et al. (2021) examines best way to sustain the supply chain in the era of industry 4.0. The study highlight the benefits of machine learning for supply chain management (SCM) and provide insights into its various applications, it has been discussed the growing importance of machine learning in industries, particularly in the context of Industry 4.0, where it can help in forecasting, decision-making, and handling uncertainty. The study highlight the importance of SCM and demonstrate how machine learning can contribute to its effectiveness in various aspects. The paper is well-researched and offers valuable insights for both practitioners and researchers in the field of supply chain management and Industry 4.0.

Research conducted by Reddy and Kumar (2021) is focused on using machine learning (ML) strategies for predicting agricultural crop yields, emphasizing the growing need for accurate yield forecasts. The paper delves into a variety of ML models, including Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), and various algorithms including Support Vector Machine (SVM), Random Forest, K-Nearest neighbours (KNN), and Multinomial Logistic Regression were applied. It examines the effectiveness of these models in addressing the complexities and variability basics in agricultural data, the use of diverse ML techniques shows adaptability to different agricultural contexts and data types. Hybrid models and deep learning approaches particularly stand out for their advanced predictive capabilities, the SVM method achieved an accuracy of 97.77%, sensitivity of 96.55%, and precision of 99.24% whereas a modified CNN reported an RMSE of 1396.4, and an ANN-MLR model had RMSE values of 9.8% and 5.1% for different components.

Moreover, the other investigation aims to enhance the accuracy of Crop Yield Prediction (CYP) using machine learning algorithms, a relevant topic given the importance of agriculture in the Indian economy. The paper PS (2019) uses an agricultural dataset with 745 instances, splitting it into 70% for training and 30% for testing, which is a standard practice in machine learning research for model validation. The paper reports that the Random Forest algorithm achieved the highest accuracy, based on error analysis values across different feature subsets. However, specific details on how the accuracy of RF compares to other algorithms and the exact error analysis values or metrics used (suchas MAE, MSE, RMSE) are not explicitly mentioned. While the paper mentions the use of error analysis values for evaluating algorithms, specific metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE). The investigation does no mention of how the models handle potential overfitting or how they perform with unseen data which is one of the limitations.

The paper by Zelingher et al. (2021) assesses the impact of regional maize production variations on global maize prices over the years 1961-2018 using statistical and machine learning models. The study aims to identify most influential regions and quantify price sensitivity. The paper has used a dataset from 1961-2018 for 19 regional maize producing entities. Authors implemented linear models, classification trees, random forests, gradient boosting machines and have compared models on different evaluation metrics. The findings showed Ensemble methods RF and GBM overall performed the best with lowest error metrics like RMSE and highest AUC scores for classification. All models consistently identified North America as by far the most influential region affecting global maize prices through its production levels.

2.4 Research Gaps and Conclusion

Several machine learning, deep learning, and time series techniques are exhibited and compared with other approaches in earlier research projects; however, there were several limitations, such as the fact that commonly used features did not work for all approaches and did not only highlight the pattern of variance in the data. The study by Sabu and Kumar (2020) had limitations include data from only a single market, lack of exogenous factors like production estimates, and no extreme weather events during the study period. Future work can build ensembles, incorporate additional variables, and assess generalizability across crops whereas Muruganantham et al. (2022) research's challenges identified include difficulty in interpreting the neural network models, ensuring practical utility for farmers/policy makers, needing large diverse training data, and generalizability across locations and crop types. The study does not explicitly address how long-term climatic changes might impact wind power density forecasting Taylor et al. (2009). Research in this direction could be valuable for long-term planning and sustainability in wind energy. Certain models exhibited challenges like high computational costs, difficulties in handling large datasets, and reduced efficiency in specific scenarios. The effectiveness of these models can be context-dependent, and there might be challenges in generalizing the results across different crop types or environmental conditions.

Recent advancements in Deep Learning, Machine Learning, and Time Series analysis have significantly transformed the ability to get insights from the market trends about forecast future scenarios, optimize supply chains, and enhance agricultural yield predictions. As technological developments continue to evolve, further research in these areas holds the potential to significantly contribute to restoring food security and improving resource management strategies. The above section satisfies the first objective for the study.

Research Title	**Year of Research**	**Models Used**	**Evaluation Metrices Used**	**Performance of Models**
Hybrid linear time series approach for long term forecasting of crop yield Alam et al. (2018)	2018	ARIMA (210), ANN	Mean Absolute Percentage Error (MAPE)	ARIMA (210): 17.677%, ANN: 4.65%
Comparing Machine Learning Approaches for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts from Corn Production Romeiko et al. (2020)	2020	Linear Regression (LR), Support Vector Machine Regression (SVR), Artificial Neural Network (ANN), Gradient Boosted Regression Tree (GBRT), and Extreme Gradient Boosting (XGBoost)	Cross-validation (CV) correlation, R-squared, and mean-square-error (MSE)	CV: 0.65,0.80, 0.74, 0.87, 0.86 r- Squared: 20, 14, 18, 10, 9
An efficient approach for rice prediction from authenticated Block chain node using machine learning technique Nesarani et al. (2020)	2020	Random Forest, Multiple Linear Regression, Gradient Boosted Regression and Decision Tree Regression	Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Coefficient of Determination (R2)	RMSE value (0.623913) and High R2 value (0.941)
Automated food safety early warning system in the dairy supply chain using machine learning Liu et al. (2022)	2022	Naïve Bayes	Sensitivity, accuracy	74%, 87%

Table 1: Overview of past researches

Table 1 provides an overview of past research, showing the implementation of different technologies such as ARIMA, ANN, LR, SVR, Regression models in the same domain.

3 Methodology

This section of the document will discuss the methodology steps adopted in this research. In Data Mining and machine learning projects Knowledge Discovery in Databases (KDD) or CRISP-DM are widely used, this study is specifically focused on acquiring KDD technique because KDD is focused on uncovering the useful insights from the large datasets and it is most relevant approach for academic research. Below subsections and Figure 1 explains the different stages of KDD that are utilized in this research.

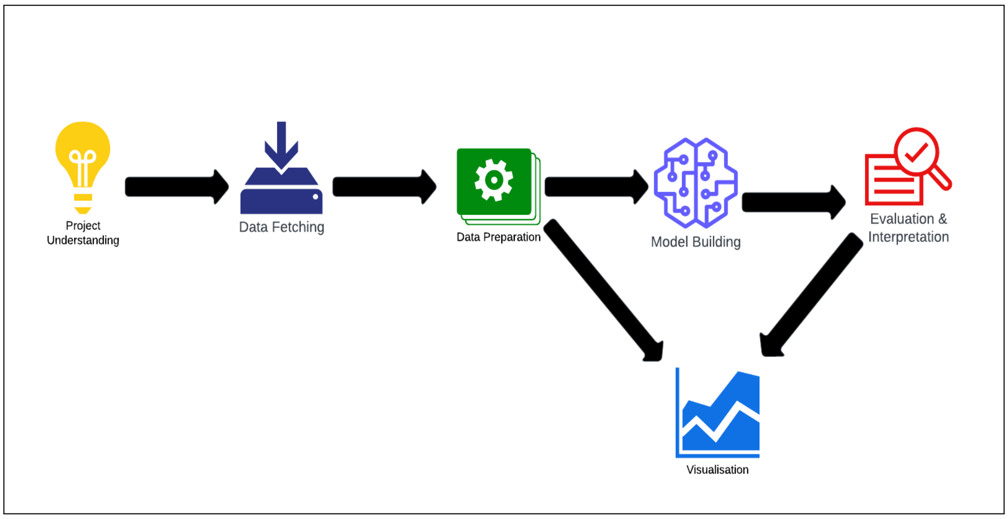


Figure 1: Research methodology

3.1 Knowledge Discovery in Database as a relevant approach

The Knowledge Discovery in Databases (KDD) process is follows a sequence of steps used to extract meaningful information from large and complex data sets. In data mining projects, KDD is crucial for transforming extensive data into actionable insights Fayyad et al. (1996). KDD makes it possible to modify the analysis process more, which is advantageous for projects that utilize machine learning, advanced statistical analysis, or other complex data analysis methods on the other hand, CRISP-DM is a more industry-focused framework that is commonly used for practical data mining projects with specific business objectives. The steps involved are Data Selection, Data Preprocessing or Data Cleaning, Data Transformation, Data Mining, Evaluation and Knowledge Integration.

3.2 Data Selection

Data selection is the process of identifying and choosing the most relevant and appropriate data from a larger dataset for analysis in a project. It involves determining which parts of the data is useful to the objectives of the project and which can be excluded. By focusing only on relevant data, it becomes easier to manage and process the information, leading to more efficient and effective data mining and discovers the patterns that are relevant to problems addressed. Irrelevant data can directly impact the evaluated results and sometimes can lead undergo nonessential processing and increases the complexity. For this study four separate datasets are obtained for 4 coarse grains such as oats, barley, corn, and sorghum for USA.

3.3 Data Preprocessing/ Data Cleaning

Data preprocessing and data cleaning are one of the most essential steps in a project which helps to transform raw data for analysis. Data preprocessing includes various steps such as formatting data that can be assessed easily while data cleaning is a subset of preprocessing which consists of removing null values, error or inconsistent data that can hamper the evaluation results. Data preprocessing and cleaning is done to improve the accuracy and efficiency. By cleaning the data, the inaccuracies are addressed, ensuring the data is consistent and reliable. The dataset obtained was very inconsistent so to avoid any discrepancies null values were moved, irrelevant data variables were removed, and column mapping was done.

3.4 Data Transformation

The process of data transformation is done to convert raw data into such a format that is easy to analyse. This can involve scaling features to a standard range, encoding categorical data into numerical formats, feature selection, feature engineering or creating attributes from existing ones. These transformations are important for aligning the data with the requirements of machine learning algorithms that will be used. Transformed data often it easier to visualize and understand complex datasets, leading to more meaningful and actionable insights. Data Transformation reduces the risk of error that can lead to wrong results. This step extracts a series of years from the original dataset. This series is crucial for the subsequent mapping and transformation processes, ensuring that each record is associated with the correct year for all the food grains.

3.5 Data Mining

Data Mining involves applying different algorithms and statistical methods to discover and analyse the data, aiming to find meaningful patterns within the dataset. The main aim of this step is to identify patterns and trends and detecting correlations, clusters, and anomalies in the data values. It helps in transforming the data into knowledge, aids in creating data driven decisions and scientific studies. This step delves with creating predictive models that can help in forecasting future trends. In this step different machine learning model are built standardisation and normalization is applied to four datasets to get better results. For this study Gradient Boosting Regressor, Bagging Regressor, Random Forest Regressor, AdaBoost Regressor, and K-neighbours Regressor are applied on the corn, barley, oats, and sorghum. Also feature selection will be adapted so that irrelevant features that can affect the analysis be removed.

3.6 Evaluation

This step includes evaluating the results obtained in the data mining steps to determine the validity and relevance of the analysis. It ensures that the patterns, relationships, and trends identified during data mining are not just random but are statistically and logically significant. This also includes considering how the findings can be applied in a practical context or what knowledge is gained from the findings. This step helps in confirming the accuracy of the results, which is necessary for making decisions based on the data analysis. The evaluation step provides feedback that can be used to refine the data mining process, adjust methodologies, or review data selection and preprocessing stages for better results.

3.7 Knowledge

The final step of KDD process is ‘Knowledge’ where patterns and models found during data mining can be used to obtain knowledge and integrate it with practical real time situations. This stage is about integrating these insights into decision-making processes, systems, and strategies. The knowledge gained is used to inform and develop policies, and actions that can address the identified issues or capitalize on opportunities. The evaluated results can benefit policymakers, economists to take decisions in the future based on the predictions. The application of this knowledge can also provide feedback for future data collection and analysis, creating continuous improvement and learning.

3.8 Conclusion

The above-mentioned research methodology for Supply and Disappearance of food grains was created to meet the requirements of this study. The project will follow the similar approach alike KDD methodology and all the stages will be applied on four datasets so to extract the best outcome to obtain the knowledge that can be used by agricultural stakeholders and economists.

4 Design Specification

This section of the document explains that the project utilizes two tier architecture which consists of 2 layers a business layer and a presentation layer as shown in Figure 2. The Business logic layer consists of appropriate data selection, cleaning and transforming the data for Exploratory Data Analysis, after that all the programming logic is done and then various evaluation metrices are applied whereas the Presentation Layer or client layer consist of visualizing the insights generated from the results in the form of graphs and charts. This study makes use of five machine learning models such as Gradient Boosting, Bagging, Random Forest, AdaBoost, and K-Neighbours Regressor and all the models are evaluated on different evaluation metrices such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared, Mean Squared Error (MSE), and Normalized Mean Square Error (NMSE). This section of the research meets Objective 2.

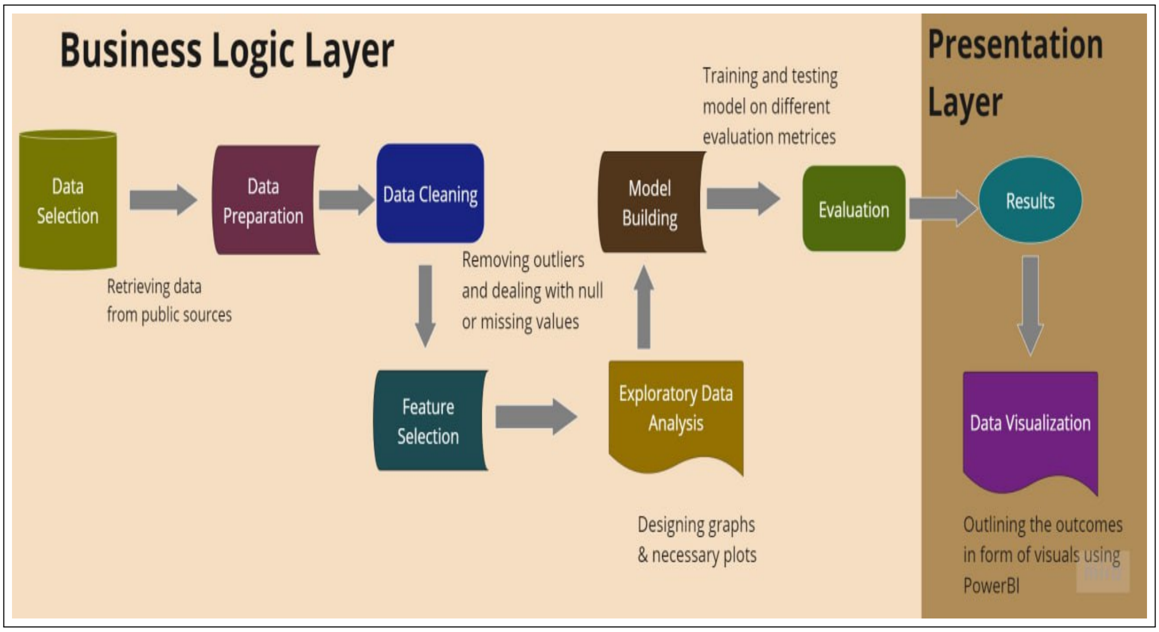

Business Logic Layer
Presentation Layer
Data Selection
Retrieving data from public sources
Data Preparation
Data Cleaning
Removing outliers and dealing with null or missing values
Feature Selection
Exploratory Data Analysis
Designing graphs & necessary plots
Model Building
Training and testing model on different evaluation metrices
Evaluation
Results
Data Visualization
Outlining the outcomes in form of visuals using PowerBI

Figure 2: Design Specification

5 Implementation

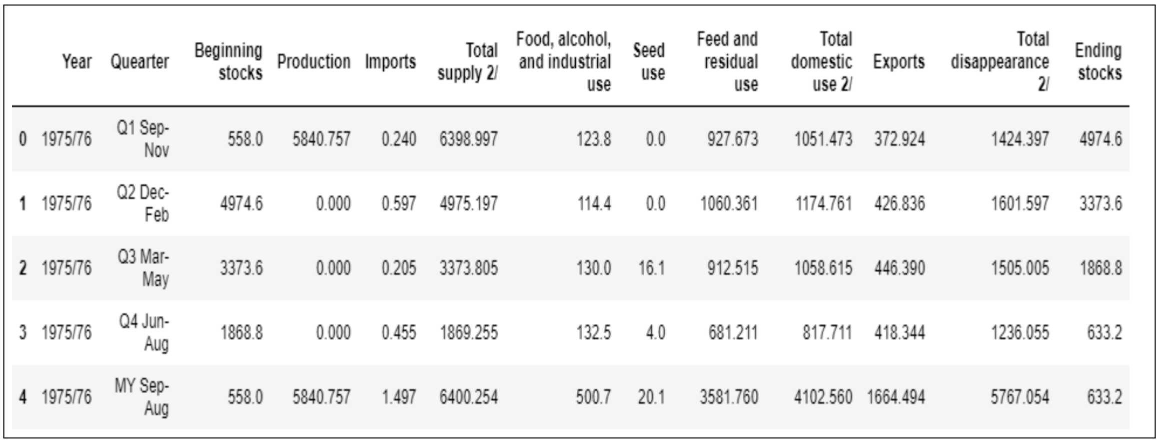
This section describes in depth implementation techniques that were taken in this research as shown in Fig.1. This section will show brief of the architecture followed, data cleansing, data transformation, exploratory data analysis of the transformed data, feature selection and models built, and next section will discuss the evaluation metrices applied on the models built to compare and analyse the best fit model for this study. This study had hardware and software requirements that helped in successfully completing the project on time. The hardware requirements are as follows Windows 10 operating system with a 256GB SSD, 16GB RAM, laptop, mouse, and wireless adaptor. Moreover, the model was built using Python programming language using jupyter notebook, and visualizations were done with the help of matplotlib, seaborn and Microsoft PowerBi.

5.1 Data Selection and Description of dataset

For this study, four datasets were obtained for 4 different crops such as corn, barley, oats, and sorghum. The dataset was publicly available on U.S. Department of Agriculture[1] website which is open to public for scientific studies. The dataset consolidated feed grains and foreign coarse grains but for this study feed grains are utilized. This incorporates the monthly feed data and annual feed yearbook, data is in weekly, quarterly, or annual format. The dataset is in zipped CSV format with collection of data from 1975 to 2023 and will be updated again on 12/13/2023. The dataset covers supply, beginning stocks, production, disappearance, industrial use, exports and ending stock value for all the feed grains as shown in Figure 3. The annual values for the last two years are projections or preliminary figures for the four datasets. Statistics on supply-use, both annual and

[1] https://www.ers.usda.gov/data-products/feed-grains-database/documentation/

quarterly, are based on marketing years. Calendar years are divided into marketing years, which are frequently written to incorporate both.



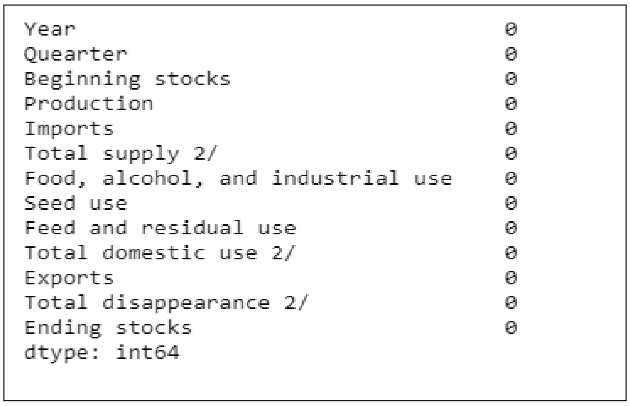
	Year	Quearter	Beginning stocks	Production	Imports	Total supply 2/	Food, alcohol, and industrial use	Seed use	Feed and residual use	Total domestic use 2/	Exports	Total disappearance 2/	Ending stocks
0	1975/76	Q1 Sep-Nov	558.0	5840.757	0.240	6398.997	123.8	0.0	927.673	1051.473	372.924	1424.397	4974.6
1	1975/76	Q2 Dec-Feb	4974.6	0.000	0.597	4975.197	114.4	0.0	1060.361	1174.761	426.836	1601.597	3373.6
2	1975/76	Q3 Mar-May	3373.6	0.000	0.205	3373.805	130.0	16.1	912.515	1058.615	446.390	1505.005	1868.8
3	1975/76	Q4 Jun-Aug	1868.8	0.000	0.455	1869.255	132.5	4.0	681.211	817.711	418.344	1236.055	633.2
4	1975/76	MY Sep-Aug	558.0	5840.757	1.497	6400.254	500.7	20.1	3581.760	4102.560	1664.494	5767.054	633.2

Figure 3: Head of Corn dataset

5.2 Data Preprocessing (Data Cleaning)

All the required libraries are imported using Python on jupyter notebook such as pandas, numpy, seaborn, matplotlib.pyplot for preprocessing stage, for data transformation StandardScaler, PCA from sklearn are imported and for model building sklearn.ensemble and sklearn.neighbors are utilized for all four datasets. Following, to load the data from the excel file ‘pandas’, this is done to target specific subset of data for the analysis. Once data loading was done then next step performed was column mapping to map the original names of columns to new names that would be easy to interpret and understand.

As a part of data cleaning step columns having missing values or ‘NaN’ values were removed. This step is important for maintaining data quality, as missing values could skew analysis results. Then years were assigned to the cleaned data to ensure that each record in the dataset has an associated year, which is critical for analysis. Then years from different yearbooks were extracted and applied to the transformed data, this step was done to maintain consistency in the time-related aspect of the data across different sheets. Lastly separate pandas DataFrames for different types of grains (like corn, sorghum, barley, oats) by applying the transformation function to different sheets. Moreover, basic functions such as shape (), size (), dtypes () and isnull () was applied on all the four datasets (Corn, Barley, Sorghum and Oats) to understand the structure and quality of the DataFrame. Shape and Size function were implemented to understand the size and volume of the dataset in terms of the number of the entries it contains whereas .isnull() was used to identify any missing values in the datasets so that those could removed in early stage and does not affect the future analysis. Figure 4 shows that the dataset does not consist of any null values. These steps are taken to prepare the data for an in-depth analysis of grain production and supply trends. Data cleaning, proper labelling of columns, and ensuring that each data point has a corresponding time frame are all critical steps in preparing data for trend analysis. By performing data preprocessing and data cleaning the third objective is satisfied.



```
Year                                 0
Quearter                             0
Beginning stocks                     0
Production                           0
Imports                              0
Total supply 2/                      0
Food, alcohol, and industrial use    0
Seed use                             0
Feed and residual use                0
Total domestic use 2/                0
Exports                              0
Total disappearance 2/               0
Ending stocks                        0
dtype: int64
```

Figure 4: Cleaned data for Barley

5.3 Data Transformation

In this step two key data transformation techniques are applied feature scaling and feature extraction; this is done to make the data more compatible with the algorithms being used and to enhance the overall effectiveness. In feature scaling the study adopts Standardization which ensures that features contribute equally to the model's performance and prevents features with larger scales from controlling the behaviour of the model. Furthermore, for feature extraction Principal Component Analysis (PCA) Maćkiewicz and Ratajczak (1993) is performed which is a dimensionality reduction technique used to reduce the number of features in a dataset while keeping the most important features only. By reducing dimensions, PCA helps in simplifying the dataset, improving model training efficiency, and enhancing model performance.

5.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a foundational approach, which aims to examine datasets by visualizing the attributes. This help in analysing the data to uncover the patterns, trends, and anomalies in the data. For this study visual exploration such as histograms, box plots and line charts are created that assist in finding patterns, trends, and outliers in the data. This heat map shows the pairwise correlation of all columns in the data frame and describes how each pair of variables closely relates to each other while highly correlated features can lead to multicollinearity, which might affect the performance of some models.

Figure 5 shows the heatmap visualization of the correlation matrix for Corn dataset. Fig shows that there's a strong negative correlation (-0.70) between 'Beginning stocks' and 'Production'. This could suggest that higher initial stocks may be associated with lower production in the same period, which could be indicative of production adjustments based on existing supplies. There's a very strong positive correlation (0.83) between 'Production' and 'Total supply 2/'. As expected, when production increases, the total supply also increases. 'Exports' have a strong positive correlation with 'Total disappearance 2/' (0.89). This indicates that exports are a significant factor in the total amount of grain disappearing from the domestic market, either due to being sent abroad or consumed. Interestingly, 'Ending stocks' have a slightly negative correlation with 'Total disappearance 2/' (-0.28) and 'Food, alcohol, and industrial use' (-0.51). This might imply that higher usage for these purposes can lead to lower ending stocks.

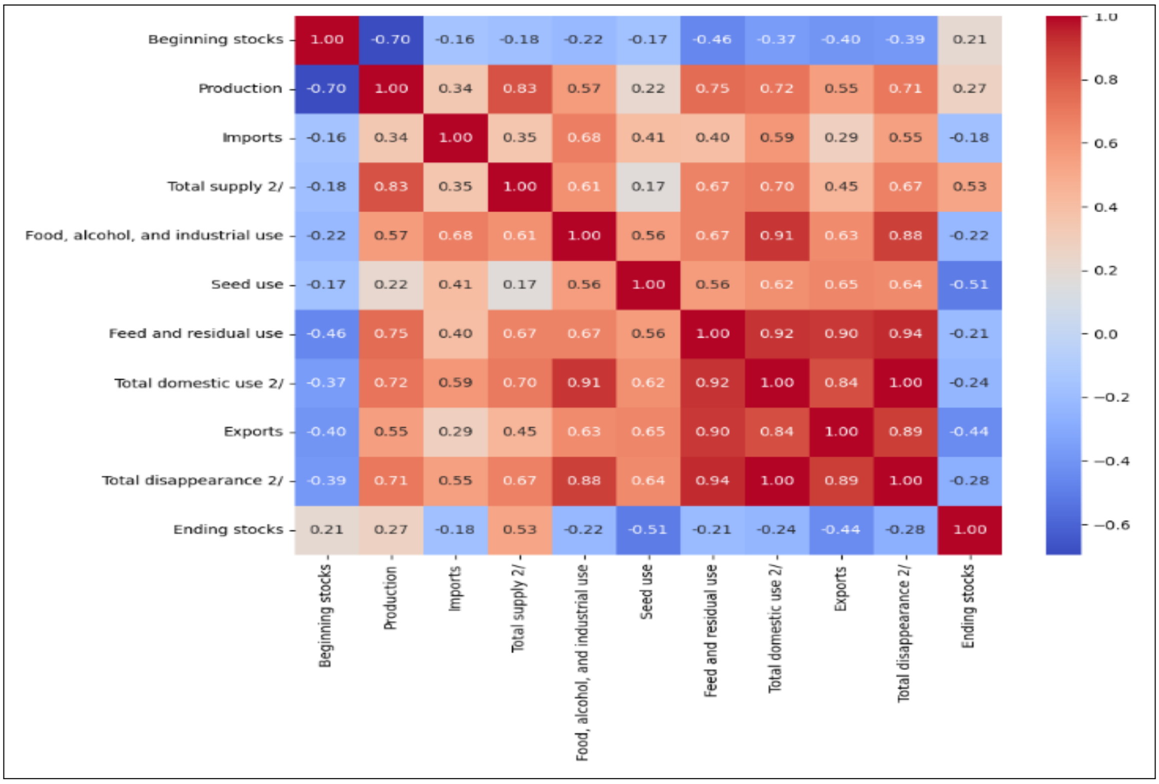


Figure 5: Correlation matrix for Corn

Figure 6 shows a visual representation of all numerical columns for the Sorghum dataset. The 'Supply Beginning Stocks' histogram suggests a right-skewed distribution, indicating that there are more instances with lower beginning stocks than higher ones. The 'imports' histogram appears to be highly skewed, with almost all values concentrated at the lower end, suggesting that imports are generally low or possibly sparse. The exports histogram is very sparse and skewed to the right, indicating that Sorghum has low export values. The ending stocks histogram is right-skewed, similar to the beginning stocks, suggesting that lower ending stocks are more common.

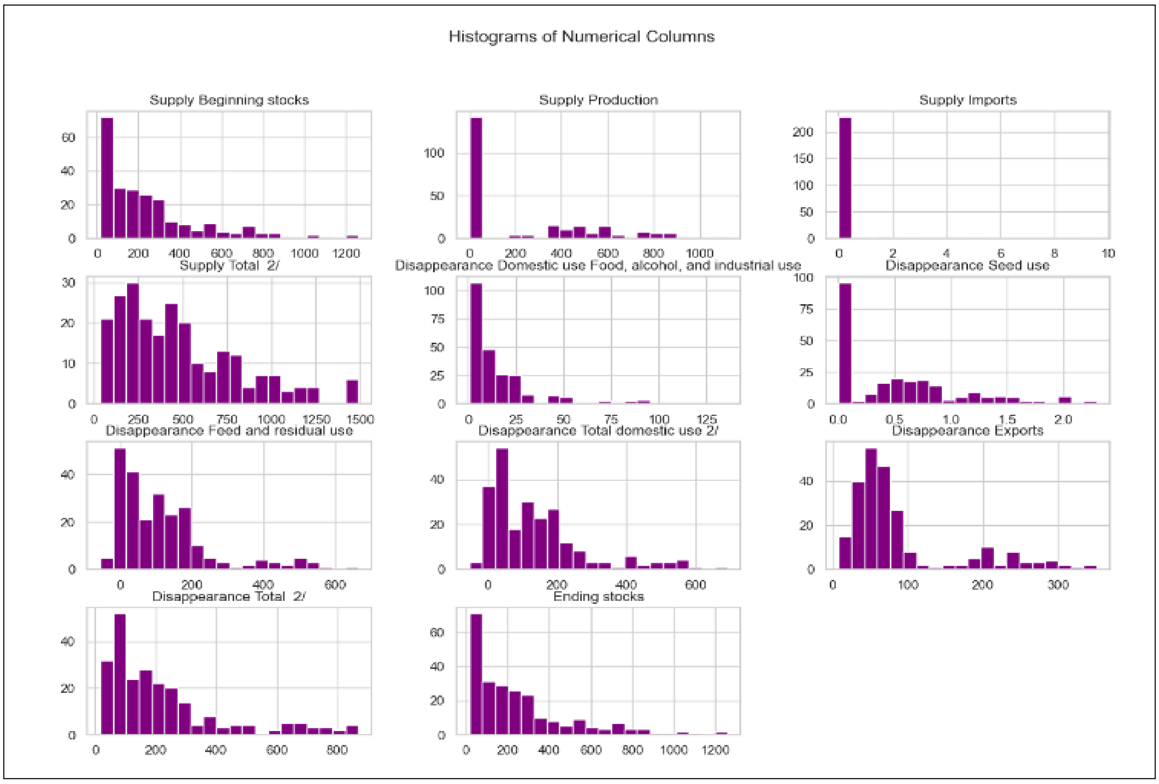


Figure 6: Histograms of numerical columns for Sorghum

Further progressing to Barley data 'Beginning stocks' indicate variability in the initial amount of barley stocks at the start of a period. A wide range suggest fluctuating beginning stocks from year to year while the outliers would indicate years with unusually high or low beginning stocks. A boxplot for 'Imports' with a smaller IQR but with outliers may suggest that barley imports are typically consistent, with a few exceptional years of high import volume as seen in Figure 7. The 'Total disappearance' boxplot reveals the overall consumption or usage pattern of barley; outliers on this plot could highlight years with unusual consumption patterns. The 'Domestic use' boxplot reflects the distribution of barley used domestically. A skewed boxplot indicates a consistent direction of deviation from the median, and outliers could point to years with extraordinary domestic use.

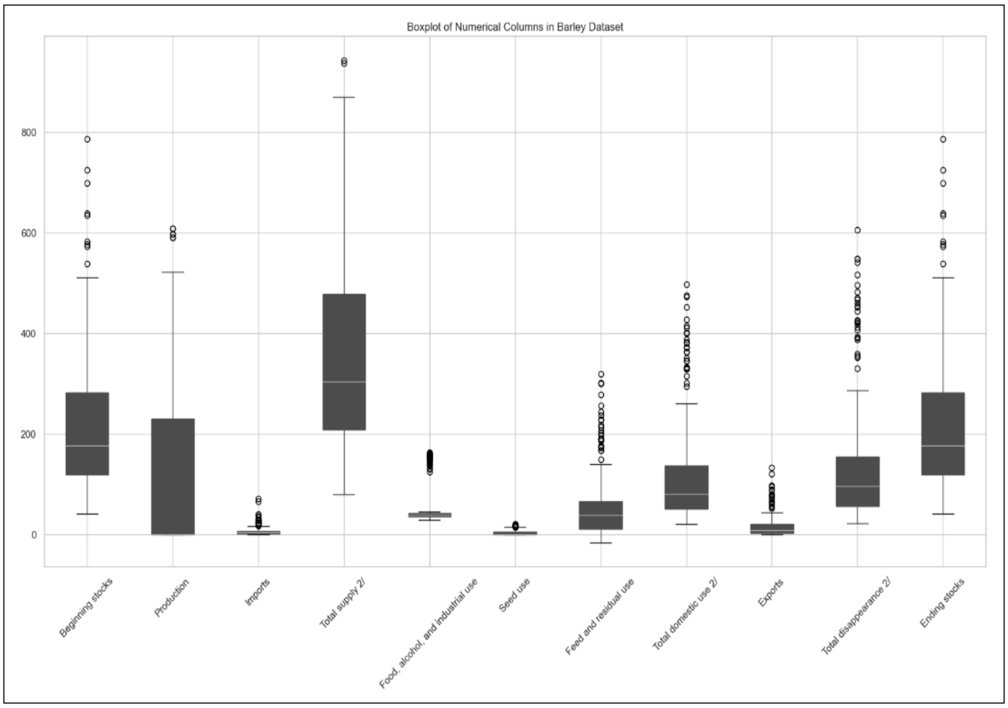


Figure 7: Boxplot for Barley

Moreover, the insights generated after performing EDA on oats dataset are as follows Figure 8:

a) The graph shows a general long-term decline in oats production over the years. This trend could indicate changes in agricultural practices, shifts in crop profitability, or changes in consumer demand.

b) The last part of the slope series shows a levelling off or a less steep decline, indicating a stabilization in production levels in recent years. This might be due to new varieties, improved farming techniques, or stable market conditions.

c) The shaded area suggests there has been significant year-to-year volatility in production. It could be influenced by factors such as variable weather conditions, market prices, or farming practices.

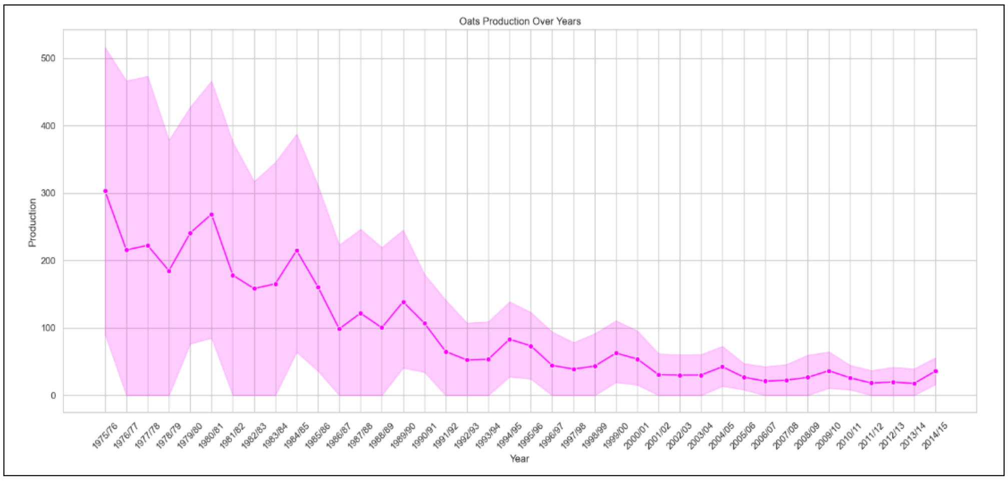


Figure 8: Oats production over years

While Figure 9 represent the initial years on the graph show significant volatility in barley exports, with sharp increases and decreases. This can suggest instability in the market or changes in production that had a direct impact on export capacity. From year 1975-1980 the trend shows a growing international demand after 1980s there is significant drop in the export which suggest there might be rising domestic consumption or increased competition from other producers. In 1990 there was a sudden increase in the export but later it dropped drastically.

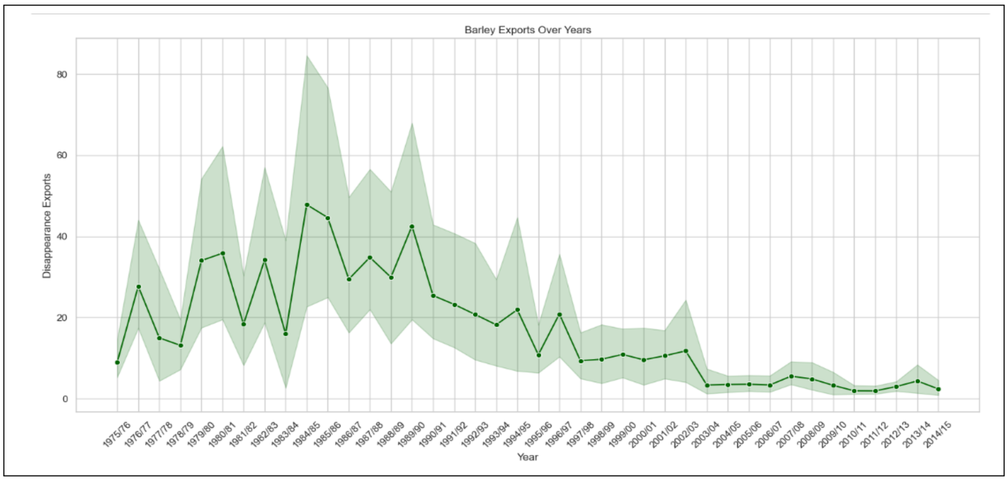


Figure 9: Barley Export over years

5.5 Model Building

This stage of implementation describes the models that were built to find the supply and disappearance of food grains. For this study five models are applied on all four datasets and out of those models the one with best predictive results can be used in future for forecasting. The first step of building a model is to split the data into training and testing sets where the training set is used to train the machine learning model, while the testing set is used to evaluate its performance. The training size is taken as 0.8 and testing size is 0.2. This helps in detecting issues like overfitting when a model performs well on the training data but poorly on new data. The goal is to build a model that not only learns the patterns in the training data but also applies these patterns effectively to new data. Before choosing any regression model for the study, 'lazypredict' library was utilised to automate the process of fitting multiple regression models to a dataset and comparing their performance Pandala (2020). This library is useful to explore which model performs best on the given dataset without manually coding each model's training and evaluation. Out of all the regression models Gradient Boosting Regressor, Bagging Regressor, Random Forest Regressor, AdaBoost Regressor, and K-neighbours Regressor are implemented on Corn, Barley, Sorghum and Oats datasets. After successfully building the models, each model will be evaluated on different evaluation metrics to obtain the best fitted model for the forecasting the result.

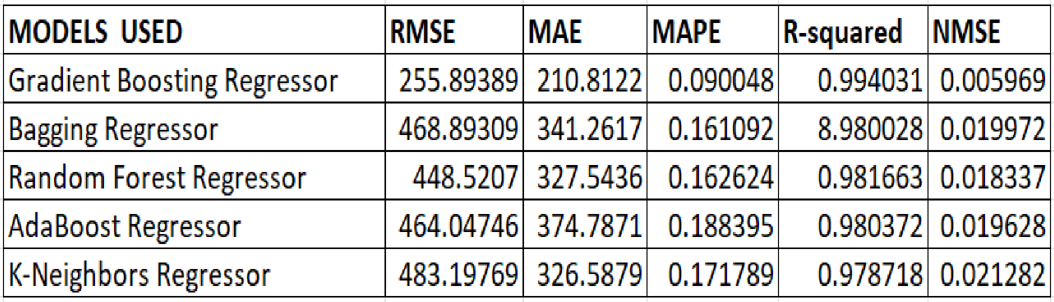
6 Evaluation

Once the model building step is done, then the models are evaluated on different evaluation metrices such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-Squared (R2), Normalized Mean Squared Error (NMSE). These metrices help in understanding the precision and robustness in different types of models, which is crucial for model selection and validation. MSE and RMSE are used because they emphasize larger errors more than smaller ones due to the squaring of each term. R2 is chosen because it provides a measure of how well the model will predict future. It indicates the "goodness of fit," which is useful for comparing models on the same dataset. NMSE is used as it normalizes the error term based on the variance of the dataset's actual values, providing a relative measure of performance that is independent of the data's scale.

Use visual aids such as graphs, charts, plots and so on to show the results.

6.1 Evaluation of Corn data

For Corn dataset The Gradient Boosting Regressor has the lowest RMSE and MAE as shown in Figure 10, indicating that on average, its predictions are closer to the actual values and lowest average error. All models have high R-squared values close to 1, suggesting a high proportion of the variance in the dataset. Lower the NMSE value better are the results, suggesting that the predicted values are closer to actual values. Therefore, Gradient Boosting Regressor performs best across all metrics, indicating it is the most accurate and consistent model for this dataset.

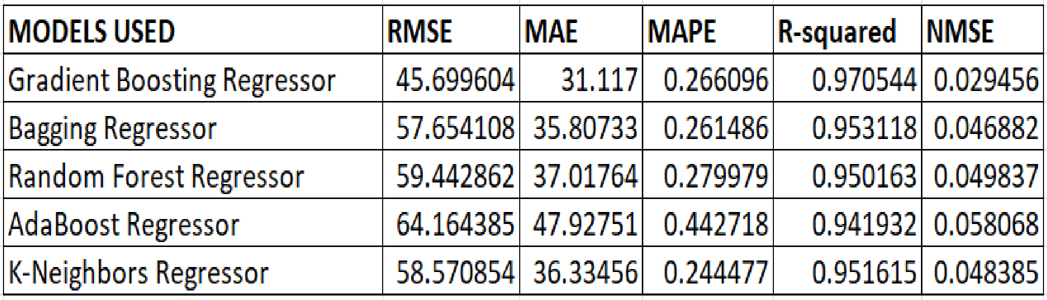


MODELS USED	RMSE	MAE	MAPE	R-squared	NMSE
Gradient Boosting Regressor	255.89389	210.8122	0.090048	0.994031	0.005969
Bagging Regressor	468.89309	341.2617	0.161092	8.980028	0.019972
Random Forest Regressor	448.5207	327.5436	0.162624	0.981663	0.018337
AdaBoost Regressor	464.04746	374.7871	0.188395	0.980372	0.019628
K-Neighbors Regressor	483.19769	326.5879	0.171789	0.978718	0.021282

Figure 10: Evaluation of Corn data

6.2 Evaluation of Sorghum data

Alike Corn as shown in Figure 11, the evaluated results shows that the Gradient Boosting Regressor stands out as the top-performing model across most metrics, particularly RMSE and R-squared, which indicates it is providing predictions that are both close to actual values and consistently capture a large proportion of the data's variance whereas if keeping percentage errors low is more important, the K-neighbours Regressor might be a better choice despite its slightly higher RMSE.

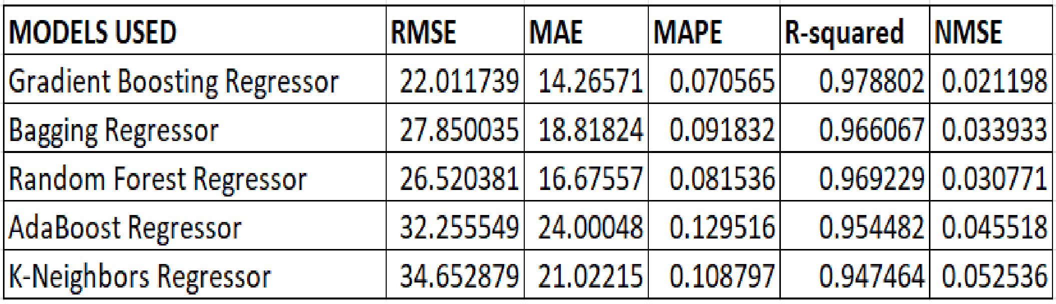


MODELS USED	RMSE	MAE	MAPE	R-squared	NMSE
Gradient Boosting Regressor	45.699604	31.117	0.266096	0.970544	0.029456
Bagging Regressor	57.654108	35.80733	0.261486	0.953118	0.046882
Random Forest Regressor	59.442862	37.01764	0.279979	0.950163	0.049837
AdaBoost Regressor	64.164385	47.92751	0.442718	0.941932	0.058068
K-Neighbors Regressor	58.570854	36.33456	0.244477	0.951615	0.048385

Figure 11: Evaluation of Sorghum data

6.3 Evaluation of Barley

The Gradient Boosting Regressor show emerges as the most accurate and reliable model among those tested. It not only has the lowest average error (MAE) and the smallest deviation in its predictions (RMSE) but also explains the highest proportion of variance in the target variable (R-squared). The Bagging Regressor and Random Forest Regressor show moderate performance, with slightly higher error metrics and lower R-squared values than the Gradient Boosting Regressor. The AdaBoost Regressor and K-neighbours Regressor have the highest error values across RMSE, MAE, and MAPE, and the lowest R-squared scores, suggesting they are less effective for the dataset as shown in Figure 12.

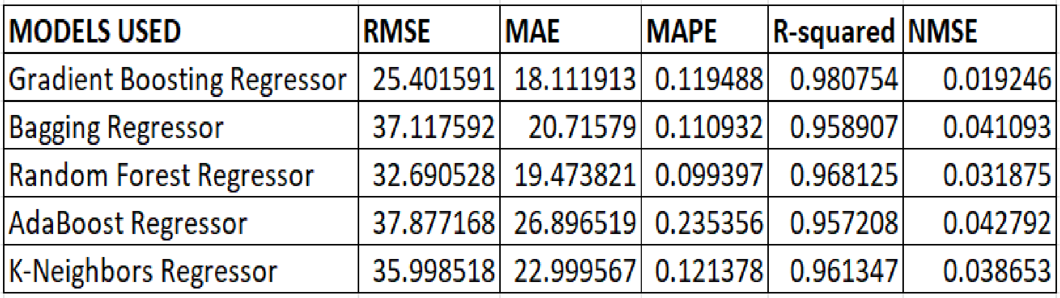


MODELS USED	RMSE	MAE	MAPE	R-squared	NMSE
Gradient Boosting Regressor	22.011739	14.26571	0.070565	0.978802	0.021198
Bagging Regressor	27.850035	18.81824	0.091832	0.966067	0.033933
Random Forest Regressor	26.520381	16.67557	0.081536	0.969229	0.030771
AdaBoost Regressor	32.255549	24.00048	0.129516	0.954482	0.045518
K-Neighbors Regressor	34.652879	21.02215	0.108797	0.947464	0.052536

Figure 12: Evaluation of Barley

6.4 Evaluation of Oats

For Oats dataset, Gradient Boosting Regressor has the lowest RMSE and MAE at 25.02 and 17.92 respectively, suggesting its predictions are, on average, closer to the actual values with the least variance and highest average accuracy among the models. The Bagging Regressor and Random Forest Regressor show the lowest MAPE (0.10), indicating that their errors are smaller relative to the actual values. The AdaBoost Regressor and Kneighbours Regressor exhibit higher error values and lower R-squared scores, indicating they are less effective for this dataset. The Gradient Boosting Regressor shows the lowest NMSE (0.02), indicating the smallest relative error compared to the others as shown in Figure 13.



MODELS USED	RMSE	MAE	MAPE	R-squared	NMSE
Gradient Boosting Regressor	25.401591	18.111913	0.119488	0.980754	0.019246
Bagging Regressor	37.117592	20.71579	0.110932	0.958907	0.041093
Random Forest Regressor	32.690528	19.473821	0.099397	0.968125	0.031875
AdaBoost Regressor	37.877168	26.896519	0.235356	0.957208	0.042792
K-Neighbors Regressor	35.998518	22.999567	0.121378	0.961347	0.038653

Figure 13: Evaluation of Oats

By performing different evaluation metrices and comparing the models Objective 4 for the research is met.

6.5 Discussion

Gradient Boosting Regressor has the lowest RMSE and MAE across all datasets, indicating its strong predictive accuracy and consistency. The variability in model performance across datasets suggests that data characteristics significantly influence model effectiveness. However, Random Forest and Bagging Regressors showed moderate performance with generally higher error metrics and slightly lower R2 scores across all datasets. AdaBoost and K-neighbours Regressors typically exhibited the highest error values and the lowest R2 scores, indicating less effective performance. The 4th dataset showed significantly higher RMSE and MAE for all models, suggesting either higher complexity or greater variability in the data. The first three datasets, while differing in their error magnitudes, showed similar patterns in model performance ranking. The error metrics in the 4th dataset for all models suggest a need for further data investigation, potential preprocessing, or considering alternative modelling approaches for oat's dataset.

7 Conclusion and Future Work

The research question stated in Section 1.3 for the study was successfully achieved as the findings of the study showed that machine learning specifically Gradient Boosting model is highly effective for forecasting trends for crops such as corn, barley, sorghum and oats. The model consistently exhibited lower error rates and strongly validates machine learning's efficacy in agricultural supply-demand forecasting that can help the policymakers, economists, producers to take data-driven decisions while random forest, bagging, and KNN showed reasonably good results. The study successfully satisfies all the objectives as Section 2 covers extensive literature review of the work done to examine food grain supply demand analysis over the past decade. Data Preprocessing, cleaning and transformation were carried out to handle missing values, dataset labeeling, and normalization. Appropriate machine learning models were designed and applied including Gradient Boosting, Adaboost, Random Forest and KNN regressor. The comparative evaluation of the each model's performance was sone using RMSE, MAE, R-Squared and NMSE to identify best suited model for future predictions.

Although the research successfully demonstrated the machine learning techniques in real world application area but it was limited to one country which can be expanded for other countries and major crops such as wheat and rice as well in future. Additional variables such as weather, soil properties and market fluctuations can further improve the forecast results. Future research can also experiment in hybrid models that can work on both historical data and temporal dynamics.

8 Acknowledgement

I would like to express my gratitude to National College of Ireland and School of Computing for their assistance throughout and my supervisor Mr. Teerath Kumar Menghwar for his constant support during the research. I am grateful of my family for their immense moral support and financial support throughout my masters program and beyond.

References

Abraham, E. R., Mendes dos Reis, J. G., Vendrametto, O., Oliveira Costa Neto, P. L. d., Carlo Toloi, R., Souza, A. E. d. and Oliveira Morais, M. d. (2020). Time series prediction with artificial neural networks: An analysis using brazilian soybean production, *Agriculture* **10**(10): 475.

Alam, W., Sinha, K., Kumar, R. R., Ray, M., Rathod, S., Singh, K. and Arya, P. (2018). Hybrid linear time series approach for long term forecasting of crop yield, *Indian J. Agric. Sci* **88**: 1275–1279.

Devi, M., Kumar, J., Malik, D. and Mishra, P. (2021). Forecasting of wheat production in haryana using hybrid time series model, *Journal of Agriculture and Food Research* **5**: 100175.

Fayyad, U. M., Haussler, D. and Stolorz, P. E. (1996). Kdd for science data analysis: Issues and examples., *KDD*, pp. 50–56.

Golan, E. H., Krissoff, B., Kuchler, F., Calvin, L., Nelson, K. E. and Price, G. K. (2004). Traceability in the us food supply: economic theory and industry studies, *Technical report*.

Liu, N., Bouzembrak, Y., Van den Bulk, L. M., Gavai, A., van den Heuvel, L. J. and Marvin, H. J. (2022). Automated food safety early warning system in the dairy supply chain using machine learning, *Food Control* **136**: 108872.

Maćkiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca), *Computers & Geosciences* **19**(3): 303–342.

Marndi, A., Ramesh, K. and Patra, G. (2021). Crop production estimation using deep learning technique, *Current Science* **121**(8): 1073.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H. and Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing, *Remote Sensing* **14**(9): 1990.

Nagar, D., Raghav, S., Bhardwaj, A., Kumar, R., Singh, P. L. and Sindhwani, R. (2021). Machine learning: Best way to sustain the supply chain in the era of industry 4.0, *Materials Today: Proceedings* **47**: 3676–3682.

Nesarani, A., Ramar, R. and Pandian, S. (2020). An efficient approach for rice prediction from authenticated block chain node using machine learning technique, *Environmental Technology & Innovation* **20**: 101064.

Pandala, S. (2020). Lazy predict.

PS, M. G. (2019). Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms, *Applied Artificial Intelligence* **33**(7): 621–642.

Reddy, D. J. and Kumar, M. R. (2021). Crop yield prediction using machine learning algorithm, *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, pp. 1466–1470.

Romeiko, X. X., Guo, Z., Pang, Y., Lee, E. K. and Zhang, X. (2020). Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production, *Sustainability* **12**(4): 1481.

Sabu, K. M. and Kumar, T. M. (2020). Predictive analytics in agriculture: Forecasting prices of arecanuts in kerala, *Procedia Computer Science* **171**: 699–708.

Taylor, J. W., McSharry, P. E. and Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models, *IEEE Transactions on Energy conversion* **24**(3): 775–782.

Zelingher, R., Makowski, D. and Brunelle, T. (2021). Assessing the sensitivity of global maize price to regional productions using statistical and machine learning methods, *Frontiers in Sustainable Food Systems* **5**: 655206.