**# 🌾 Project Overview:**

**### Agricultural Producer Price: Definition and Context**

Agricultural producer price, commonly referred to as the farm-gate price, is the amount farmers receive for their products at the point of production, before the goods enter downstream stages such as transportation, storage, processing, or retail distribution. By excluding post-farm costs, this measure reflects the farmer’s actual income per unit of crop or livestock sold. It is therefore a direct indicator of farm profitability and a cornerstone metric for assessing the economic well-being of agricultural producers.

**### Economic and Policy Significance**

Producer prices serve as a critical signal in the agricultural economy. They influence decisions at multiple levels:

- **\*\*Farm-level:\*\*** guiding production choices, crop allocation, and marketing strategies.

- **\*\*Policy-level:\*\*** informing agricultural support schemes, food security planning, and trade policy design.

- **\*\*Market-level:\*\*** shaping supply chain dynamics and investment flows in the sector.

Because they are directly tied to farm income, fluctuations in producer prices have cascading effects on food security, market stability, and the sustainability of rural livelihoods.

**### Forecasting Challenges**

Despite their importance, producer prices are difficult to forecast. The drivers are numerous, interdependent, and often non-linear, including:

- **\*\*Economic factors:\*\*** inflation, demand shifts, input costs, and trade flows.

- **\*\*Environmental variability:\*\*** climate change, extreme weather events, and resource availability.

- **\*\*Policy and socio-political influences:\*\*** subsidies, tariffs, geopolitical instability, and institutional interventions.

These factors vary not only across countries but also across commodities and time periods, introducing additional complexity.

**### Project Objective**

This project seeks to forecast agricultural producer prices using multi-dimensional datasets from FAOSTAT. The analysis integrates economic, environmental, geographic, input-related, and policy variables to address two central questions:

1. Which factors most strongly drive producer price dynamics across crops and regions?

2. How accurately can future prices be forecasted using available data and advanced modeling approaches?

**### Intended Impact**

The outputs of this work are intended for a wide audience, including farmers, agricultural scientists, economists, and policymakers. By producing reliable forecasts, the project aims to:

- Improve decision-making at both farm and policy levels,

- Enhance resilience of agricultural markets, and

- Support long-term strategic planning in food and agriculture.

**# 📊 Data:**

To build a robust forecasting model for producer prices, we’ll need to assemble not only the historical price series itself but a rich set of contextual predictors that capture the many factors driving these prices.

Fortunately, FAOSTAT offers a centralized, open-source platform where most of these datasets are readily available. Below is a summary of the datasets we integrated into our final modeling dataset, along with the rationale for including each:

1. **\*\*Historical Producer Prices (Target variable):\*\***

Annual farm-gate prices for each crop or commodity, by country and year.

2. **\*\*Crop Production:\*\***

Includes production volumes, yield, area harvested, production indices, and production values.

*\*Why:\** Variability in production affects supply-side dynamics, which directly impact producer prices.

3. **\*\*Input Usage:\*\***

Includes consumption and trade related data for fertilizers and pesticides.

*\*Why:\** Input availability and costs influence production costs and profitability, thereby shaping pricing trends.

4. **\*\*Trade:\*\***

Includes import/export quantities and values, trade indices, and indicators such as the self-sufficiency ratio and agricultural trade openness index by country and year.

*\*Why:\** International trade dynamics and global market integration often drive domestic price fluctuations.

5. **\*\*Macro-Economic Indicators:\*\***

Encompasses GDP, gross fixed capital formation (GFCF), and value added by agriculture, disaggregated by country and year.

*\*Why:\** National economic health affects both consumer demand and the financial capacity of producers to invest in agriculture.

6. **\*\*Policy & Institutional Data:\*\***

Includes data related to government expenditure and foreign direct investment (FDI) in agriculture and related fields

*\*Why:\** Policy decisions and institutional investments can incentivize or deter production, thereby influencing market supply and pricing.

7. **\*\*Climatic & Environmental Variables:\*\***

Includes temperature changes on land, greenhouse gas emissions from agriculture, nutrient balance, land use changes, and other environmental indicators.

*\*Why:\** Environmental stressors and climate variability have a direct impact on agricultural productivity and yield stability.

8. **\*\*Demographics and Employment:\*\***

Covers rural and urban population distributions as well as employment statistics within the agricultural sector.

*\*Why:\** Demographic trends influence labor availability and consumption demand, both of which are critical to understanding producer price movements.

**### Data Handling**

Each dataset was individually downloaded from the FAOSTAT website in its raw CSV format. The raw data files were then cleaned and transformed using the Pandas library in Python. This preprocessing step involved operations such as renaming columns, filtering relevant rows, data transformation, and converting data types to ensure consistency across datasets.

*\*The Python notebook files used to clean the individual raw FAOSTAT datasets can be found [*here*](https://github.com/Leonidus1995/farmer-prices-forecasting/tree/main/files\_data\_cleaning).\**

In parallel, a relational database was set up locally using PostgreSQL. For each cleaned dataset, a corresponding table was created in the PostgreSQL database. The cleaned data were then loaded into their respective tables using the SQLAlchemy and Pandas libraries.

Once all datasets were stored in the database, the individual tables were integrated using a series of left joins, primarily based on common keys such as area, year, and item (where applicable). These join operations were executed using SQLAlchemy in combination with Pandas.

The final integrated dataset—comprising all contextual variables aligned with historical producer prices—was then exported as a CSV file. This consolidated file served as the input for further data wrangling, feature engineering, and model development.

*\*All files related to database creation, data loading, and data integration could be found [*here*](https://github.com/Leonidus1995/farmer-prices-forecasting/tree/main/database\_files).\**

**# 🛠️ Pre-processing and Feature Engineering:**

The integrated dataset, created by merging individual FAOSTAT datasets, initially contained 392,856 rows and 116 features. Among these, 22 features had more than 50% missing values, and 10 of them had over 90% missingness, making them unsuitable for reliable analysis. Meanwhile, 75 features had missing values under 20%, which are more manageable.

![Before pre-processing](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/feat\_dist\_pre\_clean.png)

Top 10 Features with >90% Missing Data:

1. laying – 97.66%

2. milk\_animals – 96.27%

3. terms\_of\_trade – 96.19%

4. import\_market\_concentration\_index – 95.50%

5. export\_market\_concentration\_index – 95.46%

6. self\_sufficiency\_ratio – 95.23%

7. import\_dependency\_ratio – 95.23%

8. revealed\_comparative\_advantage\_index – 94.95%

9. yield\_or\_carcass\_weight – 92.30%

10. producing\_animals\_or\_slaughtered – 92.25%

These 10 features were dropped due to the impracticality of imputing such large gaps without introducing bias.

**### Strategy for Handling Features with 10–90% Missing Values**

We identified 36 features with missing data between 10% and 90%. Our goal was to preserve as many of them as possible by identifying whether the missingness was concentrated in specific countries or items. Since the dataset contains 166 countries and 202 items, we examined how many of these each feature covered.

For example, the column "fdi\_ag\_forest\_fish" has approximately 63% missing values. Upon inspection, it has data for only 101 countries but covers almost all items, suggesting that the missingness is primarily due to the absence of country-level data.

**### Item-Level Filtering**

We first performed item-level filtering to reduce missingness across key features. The goal was to preserve the 36 features with moderate to high missingness (10–90%) by identifying and removing items that contributed most to the data gaps.

The dataset originally included 202 items. We applied a two-step strategy to filter out problematic items:

1. **\*\*Removal of Animal-Related Items:\*\***

We first removed 43 items related to animals and animal products. This decision was based on the observation that several of the top 10 features with >90% missingness (e.g., laying, milk\_animals, yield\_or\_carcass\_weight, and producing\_animals\_or\_slaughtered) were exclusively associated with animal-related items. Retaining such items would not meaningfully contribute to analysis and would increase overall sparsity.

2. **\*\*Removal of Items with Systematic Missingness:\*\***

We further removed 25 additional items that consistently lacked data across the 36 features we aimed to preserve. These items had minimal overlap with the retained features and provided little usable information.

By applying this filtering strategy, we retained 134 of the original 202 items, significantly improving data coverage across the selected features. This step was critical in reducing row-level missingness and ensuring more robust model training on the cleaned dataset.

**### Country-Level Filtering**

In addition to filtering items, we also performed country-level filtering to reduce missingness across key features. The goal was to retain as many countries as possible while ensuring data quality.

The dataset originally included 166 countries. We applied a two-step strategy to filter out problematic countries:

1. **\*\*Focus on Features with Sufficient Data:\*\***

We began by concentrating on features that had data for 100 to 131 countries. To prevent excessive country-level data loss, any features with fewer than 100 countries were dropped.

2. **\*\*Removal of Countries with Extensive Missingness:\*\***

We observed that 71 countries had complete missingness in at least one of the 2 crucial features ('credit\_to\_ag\_forest\_fish', 'afs\_employment\_share\_in\_total\_employment'). To strike a balance between retaining countries and ensuring data quality, we removed any country missing all values in both of these 2 key features.

By applying this filtering strategy, we retained 149 of the original 166 countries, significantly improving data coverage across the selected features.

**### Dataset After Initial Cleaning**

The dataset was reduced to 211,006 rows and 97 features. We removed the top 17 features with highest proportion of missing data (> 60%).

- 76 features now have missing values under 10%.

- 90 features have missing values under 40%.

This structured reduction minimized data loss while improving the overall quality and reliability of the dataset for modeling.

![After pre-processing](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/feat\_dist\_post\_clean.png)

*\*Detailed step-by-step information about the data cleaning and pre-processing could be found [*here*](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/pre\_processing.ipynb).\**

**## Missing Data Assessment for effective Imputation Strategies**

After data cleaning, the six features with the highest proportion of missing data are:

- govt\_expenditure\_on\_ag\_forest\_fish (50.00%),

- producer\_price (45.32%),

- afs\_employment\_share\_in\_total\_employment (42.63%),

- credit\_to\_ag\_forest\_fish\_share\_total\_credit (41.29%),

- credit\_to\_ag\_forest\_fish (41.29%), and

- aoi\_credit\_to\_ag\_forest\_fish (41.29%)

Among these, the missing values in the **\*\*govt\_expenditure\_on\_ag\_forest\_fish\*\*** column

can be partially addressed by leveraging the following related variables:

**\*\*agri\_orientation\_index\_govt\_expenditure, ag\_forest\_fish\_share\_in\_total\_gdp, and total\_govt\_expenditure\*\***.

The Agricultural Orientation Index (AOI) for governmental expenditure is defined as:

AOI = (Agriculture's share of government expenditure) / (Agriculture's share of GDP)

where,

Agriculture's share of government expenditure =

(govt\_expenditure\_on\_ag\_forest\_fish) / (total\_govt\_expenditure)

Hence,

govt\_expenditure\_on\_ag\_forest\_fish =

(agri\_orientation\_index\_govt\_expenditure) ×

(ag\_forest\_fish\_share\_in\_total\_gdp) ×

(total\_govt\_expenditure)

Applying this approach reduced missingness in govt\_expenditure\_on\_ag\_forest\_fish from 50.00% to 33.36%. **\*\*However, the imputed values diverged substantially from observed data, introducing considerable bias. Consequently, we reverted to the original values and explore alternative strategies for handling the remaining missing data.\*\***

Next, the other columns with high missing data have data for 118, 123, 129 countries out of 149 total. In order to reduce the amount of missingness in these columns, we tried to find the common countries that have missing data for all these columns.

Unfortunately, there was limited overlap among these columns in terms of countries with missing data.

**\*\*Moreover, due to the high sparsity in the 'producer\_price' column, we opted to remove it from the dataset. Instead, we will focus on forecasting the 'producer\_price\_index', which offers more consistent coverage across countries and years.\*\***

Additionally, we excluded FDI-related columns from the dataset owing to their substantial missingness. However, to retain some measure of foreign investment relevance, we derived a new feature: value\_added\_aff\_per\_total\_fdi, calculated as the ratio of value\_added\_ag\_forest\_fish to total\_fdi\_inflows for each country-year pair. Both input variables have relatively high data availability, making the resulting ratio more reliable. This new feature serves as an indicator of how the agriculture, forestry, and fisheries sectors contribute to the national economy in relation to foreign direct investment.

Since no further improvement in data completeness is achievable through external sources or transformations, we now proceed to apply data imputation techniques to address the remaining gaps. Before initiating imputation, it is essential to incorporate indicator columns that categorize countries based on geographic (e.g., region, sub\_region) and economic attributes (e.g., least\_developed\_country, european\_union\_country, low\_income\_food\_deficit\_country, etc.).

These groupings are critical for informed imputation, as they enable us to estimate missing values using data from countries with similar geographic and economic profiles. For instance, imputing missing values for an African country using data from European nations would be inappropriate due to the stark differences in regional and developmental contexts. Group-based imputation ensures that such contextual nuances are respected, improving both the validity and reliability of the imputed values.

At this stage, the merged dataset contains 211,006 rows and 106 columns.

![Missingness correlation heatmap](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/heatmap\_top40.png)

The heatmap shows pairwise correlations in missingness across the top 40 variables.

A value closer to 1 means those variables often missing together.

![Matrix plot for missingness](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/matrixplot\_top30.png)

![Heatmap temporal missingness](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/heatmap\_missing\_top30.png)

The matrix plot and temporal heatmap above effectively reveal the missingness patterns among the top 30 variables with the highest proportion of missing data. Notably, 6 of the top 9 variables are missing nearly all data during the first decade (1990–2000) and year 2024, including:

- 'afs\_employment\_share\_in\_total\_employment'

- 'total\_employment\_afs'

- 'agri\_orientation\_index\_govt\_expenditure'

- 'govt\_expenditure\_on\_ag\_forest\_fish'

- 'total\_govt\_expenditure'

- 'area\_temporary\_crops'

Imputing a large and consistent block of missing values, especially spanning a decade, poses a significant risk of introducing bias and unrealistic trends.

**\*\*To address this, we decided to proceed with two parallel datasets. The first retains all 106 features but restricts the time span to 2001–2023, thereby avoiding the need to impute the substantial early-decade gaps in the six most problematic features. The second dataset excludes these six features entirely, allowing us to preserve the full temporal coverage from 1991 to 2023 without introducing unreliable imputations.\*\***

*\*Detailed step-by-step information about the missing data assessment process could be found [*here*](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/data\_imputation.ipynb).\**

**## Data Imputation: Dataset-1 (2001-2023)**

After excluding the decade 1991–2000, the dataset was reduced from 211,006 to 155,474 rows. This filtering substantially lowered the extent of missingness in the feature set. Among the 106 columns, 92 now contain fewer than 15% missing values, while the remaining 14 range between 20% and 32%. Overall, the level of missing data is moderate and well-suited for imputation strategies designed for time-series forecasting.

![Feature distribution in dataset-1](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/feature\_distribution\_dataset\_1.png)

Here are the top 14 features with greater than 20% missing values:

![top missing features in dataset-1](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/top\_missing\_cols\_dataset\_1.png)

Out of the 155,474 rows, 521 lack values for the target variable, producer\_price\_index (PPI). These records cannot simply be discarded, as doing so without care would break time-series continuity and introduce inconsistencies. To address this, we first examined missingness patterns within each country–item pair. This exploration revealed systematic gaps that guided our filtering strategy.

We removed country–item pairs that met any of the following criteria:

1. The target variable (PPI) is entirely missing for the series.

2. More than 50% of the target variable values are missing across the series.

3. More than 50% of item-related values (area\_harvested, production, yield) are missing.

4. Missing target variable values for one or both of the most recent years (2022–2023).

5. No records at all for the years 2022–2023.

The dataset originally contained 6,753 unique country–item pairs. Applying the above rules led to the removal of 630 pairs that failed to meet the criteria, primarily due to incomplete PPI or item-related data. The resulting dataset thus includes 147 countries and 130 items, comprising 137,356 rows.

**### Imputation of item-independent columns:**

Before imputing, we followed two principles:

1. **\*\*Scope by dependency:\*\*** Impute country-only variables separately from country-item variables, using the appropriate time series (country level vs. country–item level).

2. **\*\*Avoid look-ahead bias:\*\*** Perform imputation only using the training window (2001–2021), without using future observations to fill past gaps.

We first imputed the country-only variables (item-independent).

Our imputation strategy is tailored to each variable’s distribution and

temporal pattern. When a variable’s time series is approximately constant or linear,

simple methods—such as linear or spline interpolation may suffice. In contrast,

if the series exhibits abrupt or random spikes, these simpler approaches are

likely inadequate. In those cases, more advanced, model-based methods

(e.g., machine-learning imputation) are warranted to leverage both

(i) cross-sectional information from related variables (“horizontal” signals) and

(ii) the variable’s own temporal structure (“vertical” signals).

Additionally, when variables show similar trends across countries within the same

region or sub-region, information from one country can be used to impute missing values in another.

These patterns are best diagnosed visually, so we will first plotted the

country-level variables (independent of item) and explore the data to guide method selection. Following is the subset of item-independent columns that were explored visually:

![feature visualization plot-1 dataset-1](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/feature\_visualization\_plot1\_dataset\_1.png)

From the plots, area- and population-related variables exhibit approximately

constant (at most linear) trends over time. By contrast, variables related to

fertilizers, trade (imports/exports), and finance are markedly more volatile

and do not follow a stable linear pattern.

Moreover, the dataset contains five columns related to temperature change, each with relatively low missingness (2–3%). These gaps were imputed using the mean temperature change across countries within the same sub-region. Because climate trends are typically regionally consistent, sub-regional averages provide a robust and straightforward basis for filling missing values. For example, countries within Eastern Europe are expected to show broadly similar temperature shifts over time, making this approach appropriate.

Once the temperature change variables were imputed, we addressed shorter gaps of one or two consecutive missing values appearing at the start, middle, or end of time series. In these cases, simple methods such as linear interpolation, last observation carried forward (LOCF), or next observation carried backward (NOCB) were applied. Even if the broader trend is nonlinear, assuming linearity across such small gaps is unlikely to introduce meaningful bias into the analysis.

Next, during exploratory analysis, we observed that the columns 'nitrogen\_production' and 'phosphorus\_production' contain numerous zero values for certain countries. This pattern is plausible, as smaller economies may not produce these nutrients domestically and instead rely on imports. To handle missing values in these columns, we established the following imputation rules:

Rule 1: If the first non-missing value in a series is 0, all preceding missing values are set to 0.

Rule 2: If the last non-missing value in a series is 0, all subsequent missing values are set to 0.

Rule 3: If more than 50% of the non-missing values in a series are 0, all missing values in that series are set to 0.

After imputing smaller gaps and easier filling, the remaining challenge was to address the large gaps of missing values in the dataset, which required more advanced imputation strategies. Since the data was structured as a multi-panel time series, one possible approach would be to apply time series models to impute missing values within each individual series. However, there are several limitations to this option:

1. **\*\*Short length of series:\*\*** Each time series contains only 21 time points, which is quite limited for reliable time series modeling.

2. **\*\*High degree of missingness:\*\*** Some series are missing values across many time points, and in certain cases, an entire series for a variable is absent. This leaves very few data points to infer from.

3. **\*\*Lack of autocorrelation:\*\*** Even for series with a reasonable number of observations, time series modeling is not always viable. Autocorrelation must be significant for past values to help predict missing ones. If no autocorrelation is present, imputation essentially reduces to filling values with the mean, which is not informative.

Given these challenges, supervised machine learning models that leverage both cross-sectional information (data across different countries or units) and within-panel information (data within the same series) offer a more promising direction for imputing missing values. This dual perspective allows the model to borrow strength from related variables and countries when individual time series are too sparse.

As shown in the autocorrelation plot (example) below, the absence of meaningful autocorrelation in a sample of series further underscores the limitations of time series–only methods and highlights the need for cross-sectional ML-based approaches.

<p align="center">

<img src="https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/dataset\_1\_autocorrelation\_plot1.png" width="30%">

<img src="https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/dataset\_1\_autocorrelation\_plot2.png" width="30%">

<img src="https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/plots/dataset\_1\_autocorrelation\_plot3.png" width="30%">

</p>

In short, the time series has weak autocorrelation overall, with only a small

short-term dependency at lag 1 (and maybe lag 2). After that, it behaves more

like noise. This makes classical time series models (like ARIMA) less effective for imputation, since they rely on strong, sustained autocorrelation.

**\*\*To impute large gaps of missing data across the dataset, we employ the LightGBM model.\*\***

- LightGBM effectively captures complex, cross-sectional relationships in heterogeneous, multi-country datasets. It can learn across related variables and panels, using information from correlated features and similar regions to improve imputation accuracy.

- The model’s ability to handle non-linear relationships makes it well-suited for economic and trade data, which rarely follow simple trends.

- It also accommodates mixed data types- categorical, continuous, and temporal, without heavy preprocessing.

- Furthermore, LightGBM’s built-in tolerance for missing values and high scalability enable efficient, large-scale modeling of interconnected country-level panels rather than isolated time series.

**### LightGBM-Based Imputation of Missing Predictors**

Out of the total 94 item-independent predictor columns, 55 numeric (continuous) columns contained missing values and were imputed using a LightGBM-based column-wise modeling approach.

**\*\*Step 1: Column Ordering by Missingness\*\***

To maximize information flow across predictors, columns were ordered from lowest to highest missingness. Columns with fewer missing values were imputed first so that subsequent models-trained for more incomplete columns-could leverage these newly imputed features as additional inputs.

**\*\*Step 2: Model Training and Validation Setup\*\***

Imputation was performed one column at a time.

For each target column:

- The model was trained on all rows from 2001–2021 where the target column had non-missing values.

- A 10% validation subset was randomly sampled from these non-missing rows to evaluate imputation accuracy.

- The remaining 90% of complete rows were used to train the LightGBM model.

**\*\*Step 3: Evaluation Metrics\*\***

Model accuracy on the validation set was evaluated using six complementary metrics: RMSE, MAE, R², nRMSE\_mean, nRMSE\_std, and MAPE (%).

- RMSE (Root Mean Squared Error) penalizes large deviations, emphasizing high-impact errors.

- MAE (Mean Absolute Error) reflects the typical absolute error, less influenced by outliers.

- R² quantifies the proportion of variance explained by the model, capturing how well trends are reproduced.

- nRMSE\_mean (RMSE normalized by mean) and nRMSE\_std (RMSE normalized by standard deviation) provide scale-independent error measures, indicating relative error magnitude compared to central tendency and dispersion, respectively.

- MAPE (%) expresses the average error as a percentage of actual values, offering intuitive interpretability but may be inflated for columns with many near-zero values.

Together, these metrics provide a balanced view of model performance across diverse variables differing in scale, variability, and distribution.

**\*\*Step 4: Performance Screening\*\***

While LightGBM performed strongly across most predictors, its accuracy declined for certain variables- especially emission-related columns.

To maintain imputation reliability, a conservative performance threshold was applied:

- R² < 0.7, or

- MAPE > 50%, or

- nRMSE\_mean > 0.5

Columns failing any of these criteria were flagged as problematic. In total, 15 columns were excluded from LightGBM imputation and would be handled using alternative models better suited for their distributions.

**\*\*Step 5: Post-Imputation Validation\*\***

After imputation, the distributions of the 40 successfully imputed columns were inspected to ensure coherence with the original variable patterns. For example, variables defined as non-negative (e.g., fertilizer use, production) were verified to have no negative imputed values.

To enforce this constraint programmatically, the LightGBM objective function was chosen dynamically based on the target variable’s range:

```python

# Choose appropriate LightGBM objective

y\_min = y\_train.min()

if y\_min >= 0:

obj = "tweedie" # suitable for non-negative continuous targets

else:

obj = "regression" # suitable for real-valued targets

```

**### Imputation of Remaining 15 Item-Independent Columns**

To handle missing values in the remaining 15 item-independent columns, three different models were evaluated- LightGBM, Tabular Variational Autoencoder (TVAE), and K-Nearest Neighbors (KNN) Imputer. The goal was to identify the most accurate approach for each column based on a common validation framework. Following were the

15 columns to be imputed:

```{python}

['emission\_share\_agri\_waste\_mgt', 'total\_fdi\_inflows',

'emission\_share\_farmgate', 'emission\_share\_land\_use\_change',

'emission\_share\_energy\_use', 'emission\_share\_crops',

'emission\_share\_pre\_and\_post\_production',

'value\_added\_aff\_per\_total\_fdi',

'emission\_share\_end\_to\_end\_agrifood',

'emission\_share\_ipcc\_agriculture', 'total\_pesticide\_export\_value',

'phosphorus\_production', 'potassium\_agri\_use',

'emission\_share\_livestock', 'aoi\_credit\_to\_ag\_forest\_fish']

```

**\*\*Data Preparation\*\***

Rows with complete data across all 15 target columns were pooled to create training and validation sets. To prevent data leakage, all rows corresponding to the year 2023 were excluded from training.

From the pooled data, one row per country was randomly selected to form the validation set (112 rows total), while the remaining 1,931 rows were used for training the imputation models.

**\*\*Model Summaries\*\***

- **\*\*LightGBM:\*\*** A gradient boosting model trained to predict missing values column-by-column using available features.

- **\*\*TVAE (Tabular Variational Auto Encoder):\*\*** A deep generative model that learns the joint probability distribution of all features rather than predicting missing values directly. It can sample new rows that are statistically consistent with the training data.

TVAE was trained with conditional inputs to guide sampling. Starting with *\*area\** as the conditional column, 85 of 112 validation rows were successfully generated. Using *\*sub\_region\** as a condition increased this to 110 rows, and finally, conditioning on *\*region\** (fewer category levels) produced all 112 rows.

- **\*\*KNN Imputer:\*\*** A non-parametric approach that imputes missing values using the average of the k nearest rows, determined by Euclidean distance across non-missing features. Each missing cell is imputed independently.

**\*\*Model Comparison\*\***

All three models were evaluated on the validation set using four metrics- RMSE, MAE, R², and nRMSE\_std.

The Mean Absolute Percentage Error (MAPE) was not used, as many columns contain zero or near-zero values, which distort percentage-based errors.

**\*\*Key Results\*\***

- KNN Imputer achieved the best performance across most columns, particularly all emission-share variables and other continuous targets.

- LightGBM performed best only for the column potassium\_agri\_use.

- TVAE, while conceptually powerful, was not competitive for direct point imputation in this context.

**### Imputation of Item-Dependent Columns:**

The imputation of item-dependent columns was performed separately from the item-independent variables. Out of 106 total columns, 10 were identified as item-dependent, including the target variable *\*producer\_price\_index\**:

```{python}

['export\_quantity', 'export\_value', 'import\_quantity', 'import\_value',

'area\_harvested', 'production', 'yield',

'gross\_production\_value', 'gross\_production\_index', 'producer\_price\_index']

```

**\*\*Data Preparation\*\***

The columns *\*gross\_production\_value (GPV)\** and *\*gross\_production\_index (GPI)\** capture similar information. While GPV provides absolute production values in monetary terms, GPI expresses the same trend as an index relative to 2014–2016, making it a rescaled version of GPV.

To avoid multicollinearity, *\*gross\_production\_value\** was removed from the dataset.

Exploratory checks revealed that several items contained entire columns of missing values, especially for trade-related variables (imports and exports). Imputing full time series for these cases would be unreliable. Therefore, items missing more than 50% of values in import/export columns were removed, resulting in the exclusion of 30 items.

There were 8 columns that were to be imputed:

```{bash}

['production', 'area\_harvested', 'gross\_production\_index', 'yield',

'import\_value', 'import\_quantity', 'export\_value', 'export\_quantity']

```

**\*\*Imputation Model\*\***

Missing values in these columns were imputed using LightGBM, trained column-wise.

For each column:

- Training used all data from 2001–2021 where the column had valid (non-missing) values.

- 10% of those rows were reserved as a validation set.

- The model used the Tweedie objective, suitable for skewed, non-negative data (common in agricultural and trade variables).

**\*\*Evaluation\*\***

Model performance was evaluated using the following metrics:

RMSE, MAE, R², nRMSE\_mean, nRMSE\_std, and MAPE (%).

LightGBM performed consistently well across all item-dependent columns.

**\*\*Final Dataset\*\***

After completing the imputation process, the final fully imputed dataset contained:

- 117,207 rows

- 103 columns

This dataset was then used for downstream modeling and analysis.

*\*Detailed step-by-step information about the missing data imputation process for the item-dependent columns could be found [*here*](https://github.com/Leonidus1995/farmer-prices-forecasting/blob/main/dataset\_1\_third.ipynb).\**

**# 📊 Exploratory Data Analysis**

**## Distribution Analysis of Predictor Variables**

- Exploring the distributions of predictor variables revealed substantial right-skewness across most agronomic, trade, economic, and financial indicators, reflecting the dominance of a few large agricultural and economic systems.

- Applying transformations such as log(1+x) and Yeo-Johnson generally stabilized variance and improved symmetry, while a QuantileTransformer was more effective for variables with extreme zero inflation or outliers, such as emission shares.

- Several features displayed multimodal patterns, indicating natural clustering among countries (e.g., small, mid-tier, and large economies).

- Variables containing both positive and negative values, such as total\_fdi\_inflows and nutrient balance metrics, required decomposition into separate positive, negative, and binary indicator components to preserve information and achieve stable distributions.

- Structural zeros (common in trade, nutrient production, and pesticide use) were treated as genuine characteristics of the data rather than noise.

Overall, these distributional insights guided transformation choices, helped control outlier influence, and informed feature engineering to ensure more balanced and interpretable modeling inputs.

**## Relationship Between Target and Predictors**

- Analysis of the producer price index (PPI) trends revealed regional differences in price dynamics over time. Countries in Asia and to a lesser extent Africa and the Americas showed a pronounced upward fluctuations in PPI in the recent years, for major crops, indicating stronger relative price growth compared to earlier years.

- In contrast, Europe and Oceania exhibited relatively stable PPI patterns with limited fluctuations.

- It is important to interpret these increases as relative growth within regions, not as evidence of higher absolute price levels. Farmers in Asia and Africa have seen faster price rises relative to their own historical baselines, rather than higher overall prices than those in developed regions.

**# 🤖 Modeling:**

**# 🧪 Model Evaluation:**

**# 🚀 Results / Deployment:**