





Article

The Impact of Exogenous Variables on Soybean Freight: A Machine Learning Analysis

Karina Braga Marsola ^{1,*}, Andréa Leda Ramos de Oliveira ¹, Matheus Yasuo Ribeiro Utino ², Paulo Mann ³ and Thayane Caroline Oliveira da Conceição ¹

¹ Agroindustrial Logistics and Commercialization Laboratory, School of Agricultural Engineering, University of Campinas, Av. Cândido Rondon, 501, Campinas 13083-875, SP, Brazil

² Institute of Mathematical and Computer Sciences, University of São Paulo, Av. Trab. São Carlsense, 400, São Carlos 13566-590, SP, Brazil

³ Institute of Mathematics and Statistics, Rio de Janeiro State University, Av. São Francisco Xavier, 524, Rio de Janeiro 20550-013, RJ, Brazil

* Correspondence: kbraga@unicamp.br

Abstract: Predicting road freight prices is a challenging task influenced by multiple factors. Understanding which variables have the greatest impact is essential for building more accurate models, and consequently for enhancing the competitiveness of Brazilian soybeans in the global market. This study aims to evaluate the influence of different exogenous variables on soybean freight prices and to analyze how this influence varies across different distance ranges. To achieve this, a combination of machine learning techniques was applied to a comprehensive dataset containing information on freight costs, regional characteristics, production, fuel prices, storage, and commercialization. The results indicate that distance is the most significant variable in determining freight costs, directly reflecting operational expenses such as fuel consumption and labor costs. Additionally, macroeconomic factors such as the exchange rate and export volume play a crucial role, highlighting the global context of Brazil's soybean exports. Stratified analysis by distance ranges reveals distinct patterns; short-distance freight is predominantly related to domestic markets, while medium- and long-distance freight are strongly linked to export logistics.



Academic Editors: Jozef Gašparík and Davor Dujak

Received: 21 October 2024

Revised: 9 January 2025

Accepted: 17 January 2025

Published: 28 January 2025

Citation: Marsola, K.B.; Oliveira, A.L.R.d.; Utino, M.Y.R.; Mann, P.; Conceição, T.C.O.d. The Impact of Exogenous Variables on Soybean Freight: A Machine Learning Analysis. *Sustainability* **2025**, *17*, 1067. <https://doi.org/10.3390/su17031067>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: agricultural logistics; classification; freight price determinants; regression; road freight

1. Introduction

Brazil's efficiency in agricultural sectors such as soybeans, corn, sugar, orange juice, coffee, and meat is highly recognized on the international stage. This recognition is primarily attributed to productivity gains in the field, technological innovations, and continuous investments in research [1]. Products such as soybeans have complex supply chains influenced by factors such as climate, seasonality, price fluctuations, equipment availability, logistical congestion, transportation delays, ownership of the cargo, and requirements related to sustainability and product quality [2].

The main challenge faced by the Brazilian agricultural sector is the infrastructure necessary for the movement and flow of agricultural products [3]. Logistical functions and the costs associated with transportation are critical factors that directly impact soybean exports [4]. Brazil's transportation sector has faced significant structural challenges, which are largely attributed to a lack of integrated planning in infrastructure development [5].

Due to the lack of an adequate rail and waterway network, the road transportation system is the main mode for transporting agricultural products in Brazil, which restricts

the adoption of a more efficient multimodal transport system [6,7]. Logistical efficiency and low transportation costs are essential in order for Brazilian agriculture to maintain its competitiveness internationally, especially compared to other commodity-producing and commodity-exporting countries [8].

In the European Union, long-distance deliveries typically span around 600 km, with most freight being transported over distances between 300 km and 999 km and only a few routes exceeding 1000 km [9]. In Canada, which covers 9.9 million km², railways account for 55% of freight transport, while in the United States, with an area of 9.8 million km², rail transport makes up 53% of total freight movement [10,11]. In contrast, Brazil presents a very different scenario. The country's freight transportation system is heavily dependent on road transport, with a distribution across road, rail, and waterways that differs significantly from other countries of similar size. Brazil's infrastructure includes 1.564 million kilometers of roads (only 13% paved), 30.6 thousand kilometers of railways (of which only one-third are commercially active), and 41.7 thousand kilometers of navigable waterways (with only 19.5 thousand kilometers being economically viable) [12]. With a land area of 8.5 million km², in 2024, road transport handled 50% of agricultural bulk cargo, while 33% was transported by rail and 17% by waterways [13]. A notable example is the road route spanning over 1500 km that connects Mato Grosso, one of Brazil's largest soybean-producing states, to the port of Santos, a key export hub. These extensive distances directly affect domestic logistics costs, including road transportation and vessel wait times at ports, which contribute significantly to the final cost of soybeans [4,14].

Freight price forecasting plays a fundamental role in the commodities trade as well as in price analysis for the agricultural sector. Traditionally, research has focused on production and yield forecasting [15–18] and on price forecasting in agricultural product markets [19–21]. Machine learning (ML) methods such as Random Forest and SVM are widely used in agriculture to improve the accuracy of yield forecasts and anomaly detection, contributing to better management of agricultural systems [22].

However, when evaluating production, we must consider its critical factor, namely the selling price. One of the critical components in price formation is agricultural freight, which directly influences the final cost of commodities. Despite this, few studies have advanced analysis of the multidimensionality of variables that impact the price formation process [23]. Therefore, understanding and predicting variations in freight costs becomes essential in supporting negotiations and promoting more accurate price analyses, in turn contributing to strategic decision-making in the sector. In this sense, as discussed by Sarker [24], the application of ML techniques offers an effective means of analyzing exogenous variables such as distance and seasonality as well as of identifying patterns that influence road freight prices, enabling greater accuracy in predictions and decision-making in the logistics and agricultural sectors.

Machine learning models such as KNN, LightGBM, and Logistic Regression have demonstrated great efficacy in handling large datasets with temporal and spatial variability [25], making them suitable for evaluating the costs of road freight such as soybean transportation. Furthermore, the application of supervised techniques such as Random Forest and Decision Tree allows for the capture of complex patterns in supply and demand, productivity, and transportation factors [19].

Analyzing and comparing different ML algorithms has become a common focus in the literature. For example, Kulkarni et al. [26] evaluated KNN, Random Forest, XGBoost, and LightGBM to predict freight costs, identifying the most influential factors and determining which model offers the best accuracy. Similarly, Tsolaki et al. [27] used Logistic Regression, Decision Tree, Random Forest, and XGBoost to model transportation costs in various scenarios, considering vehicle routing, transportation demand, and route optimization.

Previous studies have also adapted predictive techniques from stock market analysis, addressing the prediction of soybean freight prices as either a regression or classification task [28]. As highlighted by prior research, framing this problem as a classification task has been shown to “provide better information for decision-making” [28]. Building on this foundation, our study explores both regression and classification approaches to predict soybean freight prices. Through our experiments, we demonstrate that classification offers greater explainability, whereas regression lacks this critical advantage.

From other perspectives, Das et al. [29] and Wu et al. [30] investigated the application of deep learning techniques such as graph neural networks and sentiment analysis to forecast stock prices using social media data, an approach that could also be adapted for predicting soybean road freight price trends. Their studies integrated historical price data with sentiment analysis extracted from textual sources, demonstrating how machine learning techniques can effectively combine diverse data types to improve prediction accuracy.

Other works have employed time series approaches. For instance, Fan et al. [31] introduced an innovative method for predicting soybean futures prices by utilizing a Long Short-Term Memory (LSTM) model with dual-stage attention augmented by sequence decomposition and feature expansion. Sequence decomposition is executed using the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique, which effectively extracts patterns and removes noise. Dual-stage attention is then applied to capture the spatiotemporal relationships between the input features and the target sequence. Similarly, another study [32] employed LSTM models to classify soybean future prices as high or low. They focused on the classification perspective rather than on predicting exact price values, thereby reducing the impact of noise when relying on regression.

However, regardless of the technique employed—whether traditional machine learning methods or deep learning approaches—not all studies have utilized a diverse range of input features. As noted by Silva et al. [28], many prominent works have primarily relied on environmental variables such as climate-related factors [33], the type of vehicle and cargo weight [34], or a limited set of features [26]. In contrast, our study incorporates multiple data sources as model inputs, including regional characteristics, production volumes, fuel prices, storage capacities, and other relevant factors, thereby offering a more comprehensive framework for prediction.

In addition to machine learning-based approaches, econometric models for freight and demand forecasting have also been explored. For example, ARIMA, ARIMAX, and SARIMAX have been compared to Artificial Neural Networks (ANN) in studies using European market data, particularly on the Netherlands-to-Italy route. Findings indicate that Multi-Layer Perceptron (MLP) neural networks outperform in freight rate predictions, while ARIMA models excel in demand forecasting due to lower prediction errors [35].

In contrast, our approach proposes the use of classical ML models to ensure explainability through feature importance analysis, thereby avoiding the high computational costs of deep learning models. By adopting more efficient models, we aim to provide clear insights into the factors influencing predictions while balancing model accuracy with computational feasibility. This approach allows for a deeper understanding of how different features impact the results, which is crucial in practical applications where explainability is essential. Moreover, in the context of increasingly complex and interconnected global supply chains, real-time logistics management supported by advanced decision-support systems plays a pivotal role in improving operational efficiency and reducing costs, directly contributing to global trade competitiveness and sustainability [36].

Therefore, the goal of this research is to assess whether the price of soybean road freight is influenced by a set of associated exogenous variables and to determine how the

influence of these variables varies across different distances and models. Our hypothesis is that by utilizing a dataset that includes not only macroeconomic variables, it will be possible to predict the price of grain road transportation and identify association patterns in freight behavior. For this, we use eight ML methods for regression and classification: Decision Tree, ExtraTrees, KNN, LightGBM, Logistic Regression, Random Forest, Passive-Aggressive, and XGBoost. Additionally, we leverage AI explainability techniques to assess the importance of each variable in order to gain deeper insights into the influence of exogenous variables on the predictions.

2. Materials and Methods

2.1. Dataset

The data used in this study were obtained from official sources provided by the Federal Government and research institutes. Data collection was conducted on a monthly basis, covering the period from 2015 to 2019. The guiding principle for constructing the database was the recording of freight values by month. Each record contains information about the freight cost for transporting soybeans from an origin municipality to a destination municipality considering a specific distance and a specific month of a given year.

It is important to note that the freight records exhibit particularities related to the seasonality and variability of routes, which are typical of soybean transportation by road; therefore, the presence of a freight record in a specific month does not imply its repetition in subsequent months, reflecting the dynamics of the market and the variability in transportation demand and supply.

The data for each variable were initially organized across one or more files, which were then consolidated into a single dataset. To achieve this unification, data cleaning and outlier identification were essential. The first step in the unification process involved compiling the different databases into a single detailed format, while also addressing data corrections such as improper formatting, duplicate and/or ambiguous values, and missing values. The rationale behind the selection of these variables is detailed in Table 1.

Data points were classified into four scenarios (all data points, three price ranges: low, medium, and high, with 85,280 records, which are available in the Supplementary Materials, Table S1) based on specific thresholds for historic freight values, following Equation (1). Table 1 lists the input variables and the reasoning behind their choice, and additionally includes descriptions of each variable and the number of occurrences for each classification. This distinction is necessary because freight prices behave differently depending on the distance traveled [37,38].

$$\text{Categorical Freight(freight value)} = \begin{cases} \text{Scenario 1. All data points} \\ \text{Scenario 2. Low} & \text{if freight value} < 60 \\ \text{Scenario 3. Medium} & \text{if } 60 \leq \text{freight value} < 100 \\ \text{Scenario 4. High} & \text{if freight value} \geq 100 \end{cases} \quad (1)$$

Table 1. Overview of the (exogenous) variables in each group along with their corresponding motivations and supporting references.

Groups	Variables	Motivations	References
Freight	Freight Price, Distance, Origin, Destination, Month and Year	Evaluate the relationship between the distance traveled and the cost of road freight transportation	Kengpol et al. [39], Márquez and Cantillo [40]
Region	Origin State, Destination State, Origin Municipality and Destination Municipality	Analyze the impact of transport corridors on freight prices.	Péra et al. [41]

Table 1. Cont.

Groups	Variables	Motivations	References
Production	Planted Area by Municipality, Planted Area by State, Municipality Harvested Area, State Harvested Area, Municipality Production, State Production, Average State Yield and Municipality Yield, Municipality Production Value and Harvest Period	Assess how regional productivity levels, productive potential, and the seasonality of transport demand influence the pricing of transport freight.	Melo et al. [42], Cicolin and Oliveira [43]
Fuel	Maximum, Average and Minimum Price of Diesel, Maximum, Average and Minimum Price of Ethanol, Maximum, Average and Minimum Price of Gasoline	Examine how operational transport factors and fluctuations in diesel prices impact the overall cost of road freight transportation.	Filippi and Guarnieri [44], Teixeira et al. [45], Wetzstein et al. [46]
Storage	Storage Capacity by Origin State and Storage Capacity by Destination State	Analyze how the capacity of grain storage facilities at both origin and destination points influences the pricing trends of freight transportation.	Melo et al. [42], Cicolin and Oliveira [43]
Commercialization	International Price (CBTO), Soybean Price (Parity), National Market, Crushing Capacity Industry by Origin State, Crushing Capacity Industry by Destination State, Average Monthly Exchange Rate, Diesel Imports, Monthly Export Tonnage by Origin State and Yearly Export Tonnage by Origin State	Investigate how factors such as international and national market dynamics, crushing capacities, exchange rates, diesel oil imports, and export volumes influence the freight pricing of agricultural products.	Asai et al. [47], Sonaglio et al. [48]

2.2. Splitting the Data for Training and Testing

Let $\mathcal{T} = \{\mathcal{X}, \mathcal{Y}\}$ be our desired classification or regression task \mathcal{T} , composed of the single original preprocessed dataset \mathcal{X} , where each pair of elements (\mathbf{x}, y) for $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ constitutes a data point. In this context, $\mathbf{x} \in \mathcal{X}$ contains the independent variables, while $y \in \mathcal{Y}$ represents the dependent variable. To train the machine learning model \mathcal{M} for predicting freight, we partition the dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, with size N , using cross-validation. Specifically, the dataset \mathcal{D} is divided into K folds (i.e., partitions), each of approximately equal size, which are stratified according to the distribution of the target variable y within \mathcal{D} .

In this process, fold \mathcal{D}_i is used to evaluate the model based on predefined metrics and trained on the complementary dataset $\mathcal{D} \setminus \mathcal{D}_i$. Afterwards, the average of all folds \mathcal{D}_i is used as the final metric. This cross-validation approach helps to mitigate bias and provides a more robust estimation of model performance compared to single-split methods. Moreover, when dealing with classification tasks, the stratified K -fold approach is employed to ensure that the class distribution is approximately maintained across all folds, which can provide a better representation of the data. We used $K = 5$ for our experiments.

2.3. Preprocessing

Preprocessing techniques such as filling in missing values, data normalization, and representing categorical features as numerical vectors were employed to enable use of the dataset by the machine learning algorithms and improve the performance of the models.

2.3.1. KNN Imputer

To fill in the missing values, we used a KNN Imputer. This method selects the K closest neighbors of the element \hat{y} to be filled based on some distance metric $d(x, y)$. Afterwards, a

measure is computed as mean of the the K closest neighbors and used to fill in the missing values, following Equation (2):

$$\hat{y} = \frac{1}{K} \sum_{i=1}^K y_i \quad (2)$$

where \hat{y} is the imputed value and y_i represents the values of the K nearest neighbors.

The process of filling missing values is vital for machine learning models, as most of them cannot process data with missing values. We employed $K = 5$ and used the Euclidean distance as the distance metric.

2.3.2. Z-Score Normalization

The dataset \mathcal{X} is composed of d -dimensional feature vectors \mathbf{x}_n . Due to the potential variability in the values across the i -th dimension of different data points, normalization becomes essential for effective operation with machine learning models. To address this, we normalize the feature set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, producing a transformed set of features $\mathcal{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_N\}$, where each feature is rescaled for consistency across dimensions as follows:

$$\mathbf{x}'_n = \frac{\mathbf{x}_n - \boldsymbol{\mu}}{\sigma} \quad (3)$$

where \mathbf{x}'_n is the z-score normalized feature vector and \mathbf{x}_n represents the original vector.

The vectors $\boldsymbol{\mu}$ and σ denote the mean and standard deviation, respectively, for each i -th dimension across all feature vectors in \mathcal{X} .

This operation is particularly useful for allowing different features to be compared with each other, helping to prevent features with higher values from overshadowing the others. Furthermore, z-score normalization results in dimensionless values, facilitating interpretation and analysis of the data.

2.3.3. One-Hot Encoding

Let $\mathcal{X}_{:,j}$ denote j -th feature across all data points, where each x_{nj} represents the value of the j -th feature for the n -th data point. Suppose that the j -th feature $\mathcal{X}_{:,j}$ is categorical and takes values from a finite set of m distinct categories $C = \{c_1, c_2, \dots, c_m\}$ across all data points. We apply one-hot encoding to this feature to transform it into a set of binary vectors, allowing the machine learning models to handle the categorical data.

For each distinct category $c_k \in \mathcal{X}_{:,j}$, we create a new feature vector $\mathcal{X}_{:,k}$ for each data point n containing only binary values, where a 1 appears in the position corresponding to the category present in x_{nk} and 0 appears elsewhere. After that, the dataset contains k new binary features, with the j -th feature being removed. This process is executed for each categorical feature that is present in the dataset. This transformation ensures that the categorical feature $\mathcal{X}_{:,j}$ is converted into a numerical form suitable for machine learning models.

2.4. Models

Given the distinct nature of the tasks at hand, namely, classification and regression, we employ eight classical machine learning models to address both tasks effectively: K-Nearest Neighbors (KNN), Logistic Regression, Passive-Aggressive, Decision Tree, Random Forest, ExtraTrees, LightGBM, and XGBoost. The primary motivation for using these models is, first, to accurately classify freight price data points as low, medium, or high, second, to predict the freight value through regression, and finally, to identify the most influential independent variables for classification. We achieve the first goal by training on historical data, and rely on explainability techniques to attain the third.

KNN is a simple yet powerful model that classifies instances based on their proximity to other data points. For classification, the most frequent class among the nearest neighbors is selected, while for regression the average value of the neighbors is computed. While straightforward, KNN excels at capturing local patterns in the data without making strong assumptions about the distribution, which sets it apart from traditional analytical methods that rely on fixed functional forms.

Logistic Regression is a probabilistic classifier that models the relationship between input features and the probability of an instance belonging to a particular class. It is a foundational algorithm in machine learning, providing interpretable coefficients and a robust framework for binary and multinomial classification. Unlike traditional statistical methods, logistic regression allows for more flexible handling of feature interactions and nonlinear decision boundaries, making it more adaptable to real-world data.

The Passive-Aggressive model is a linear model that excels in online learning scenarios, adjusting its weights dynamically based on prediction accuracy. It remains passive when making correct predictions, but aggressively updates weights when errors are made, allowing it to quickly adapt to changes in the data. This iterative learning process is especially useful in streaming or dynamic environments where data distributions change over time, representing an advantage over traditional analytical models that often rely on static assumptions.

Decision Tree is a nonparametric approach to classification and regression that splits data according to learned thresholds. This model's ability to visualize decision boundaries enhances interpretability and provides valuable insights into how different variables contribute to predictions. In contrast to traditional analytical models, Decision Tree models are more flexible and do not assume any predefined functional form, which makes them ideal for modeling complex nonlinear relationships.

Ensemble models such as Random Forest, Extra Trees, LightGBM, and XGBoost build upon the Decision Tree algorithm by aggregating multiple trees to form a more robust and generalized model. These techniques significantly reduce the risk of overfitting by introducing randomness and increasing model diversity. Moreover, they provide a wealth of hyperparameters that can be tuned to improve model performance, surpassing the limitations of traditional analytical methods that often struggle with complex high-dimensional datasets. These ensemble methods offer higher accuracy and stability, especially in the presence of noisy or imbalanced data, making them superior to many traditional statistical approaches.

All of these machine learning models overcome several limitations of traditional analytical methods, which often rely on rigid statistical assumptions such as normal distributions or linearity in relationships between variables. For example, many statistical methods require data to meet specific conditions, such as homoskedasticity or independence, which are not always present in real-world datasets. Additionally, models such as Linear Regression and ANOVA can become ineffective when dealing with nonlinear relationships or complex interactions between variables, something that machine learning models, particularly those based on decision trees and ensembles, can efficiently capture. Unlike traditional methods, machine learning models do not require these assumptions, and are able to handle high-dimensional data, large data volumes, and even noisy data, providing greater flexibility and accuracy in their predictions. This enables them to overcome the limitations of traditional methods, offering a more robust and adaptive approach to complex forecasting and classification tasks.

2.5. Hyperparameter Tuning

The machine learning models have hyperparameters that directly impact performance, making their selection crucial for both classification and regression tasks. The optimization method employed in this paper was the Tree-structured Parzen Estimator (TPE), which utilizes Bayesian techniques to efficiently select hyperparameters even for complex and high-dimensional search spaces [49]. Compared to traditional search methods such as random search or grid search, the TPE reduces the number of iterations required to find hyperparameters.

To test a certain combination of hyperparameters, a metric m is computed to evaluate the quality of the parameters, resulting in a discrete distribution of the parameter values concerning the metric m . Afterwards, it is possible to estimate the probability density using Kernel Density Estimation (KDE), obtaining two distributions: one $l(x)$ that is below a threshold γ and another $g(x)$ that is above the threshold. The first distribution represents promising parameters, while the second represents less promising ones; thus, the goal is to maximize the ratio between $l(x)$ and $g(x)$ in order to identify promising search spaces.

This process is iterative; testing a new hyperparameter alters the distribution based on previous results, improving the estimate of the probability density of quality hyperparameters. During our experiments, 20 iterations were conducted. For each experiment with a chosen set of hyperparameters, we computed a metric m to evaluate the iteration. We selected the model based on the best iteration, measured by the best metric m .

Table 2 presents the hyperparameter tuning details and corresponding values along with the best parameters identified for each model in both classification and regression tasks.

Table 2. Hyperparameter tuning parameters and best values for classification and regression tasks for KNN, Logistic Regression, Passive-Aggressive, Decision Tree, Random Forest, ExtraTrees, LightGBM, and XGBoost.

Model	Tuning Parameters	Values	Best Classification	Best Regression
KNN	K	[3, 500]	14	13
	weights	uniform, distance	distance	distance
	metric	cityblock, cosine, euclidean	cityblock	cityblock
Logistic Regression	C	$[10^{-5}, 10^5]^1$	$2.11 \cdot 10^3$	-
	penalty	None, l1, l2	l1	-
Passive Aggressive	C	$[10^{-4}, 10^1]^1$	0.01	$1.28 \cdot 10^{-4}$
	tol	$[10^{-5}, 10^{-1}]^1$	$4.32 \cdot 10^{-5}$	$2.85 \cdot 10^{-4}$
	loss	hinge (classification) sqhinge (classification) epsilon (regression) sqepsilon (regression)	hinge	sqepsilon
	criterion	gini (classification) entropy(classification) log loss (classification) squared error(regression)	gini	squared error
Decision Tree	max depth	[2, 10]	9	7
	min samples split	[2, 10]	8	7
	min samples leaf	[1, 4]	3	2

Table 2. Cont.

Model	Tuning Parameters	Values	Best Classification	Best Regression
Random Forest	n estimators	[50, 500]	471	218
	criterion	gini (classification) entropy(classification) log loss (classification) squared error(regression)	gini	squared error
	max depth	[2, 10]	10	10
	min samples split	[2, 10]	8	8
	min samples leaf	[1, 4]	1	3
Extra Trees	n estimators	[50, 500]	183	207
	criterion	gini (classification) entropy(classification) log loss (classification) squared error(regression)	entropy	squared error
	max depth	[2, 10]	10	10
	min samples split	[2, 20]	20	20
	min samples leaf	[1, 20]	12	1
	max features	[0.5, 1.0]	0.64	0.87
XGBoost	n estimators	[50, 500]	321	263
	max depth	[2, 10]	8	6
	max leaves	[2, 5]	0	0
	learning rate	[0.01, 0.3]	0.29	0.04
	colsample bytree	[0.5, 1.0]	0.92	0.64
	lambda	[0.0, 10.0]	2.12	5.92
	alpha	[0.0, 10.0]	1.81	7.25
	gamma	[0.0, 10.0]	1.83	5.16
LightGBM	n estimators	[50, 500]	130	348
	max depth	[2, 10]	9	6
	max leaves	[2, 31]	23	11
	learning rate	[0.01, 0.3]	0.23	0.17
	colsample bytree	[0.5, 1.0]	0.68	0.95
	lambda	[0.0, 10.0]	6.86	9.70
	alpha	[0.0, 10.0]	2.31	7.75
	min child samples	[1, 10]	7	10
	min split gain	[0.0, 5.0]	0.08	0.92

¹ In log scale. Sqhinge is squared hinge, epsilon is epsilon insensitive, and sqepsilon is squared epsilon insensitive.

2.6. Feature Importance

In the context of AI, explainability focuses on providing insights into why a model makes a particular decision. If we can successfully determine the reasons behind a model's prediction of the freight value as high, medium, or low, we can gain valuable insights into how specific variables influence these outcomes. This understanding allows us to prioritize one dimension over another when necessary; for example, if certain variables, such as distance or weight, have a stronger impact on the prediction of high freight costs, then we could adjust these factors to optimize pricing decisions. By leveraging explainability techniques, we can make more informed decisions, potentially favoring one set of variables

over another depending on their influence on the final classification. Here, we used the feature importance, specifically using the permutation approach to determine the top K features of the dataset.

First, we trained a machine learning model \mathcal{M} on all d features of the dataset and evaluated its performance using a specified metric. Next, to assess the importance of each feature, we considered a specific feature $\mathcal{X}_{:,j}$. A permutation π was applied to this feature such that each element x_{ij} is mapped to another element x_{kj} according to the bijector function $\pi(x_{ij}) = x_{kj}$.

After randomly shuffling the values of the rows along a single feature j , we evaluated the previously trained model on this newly corrupted dataset to check the impact of shuffling the data. If the metric for the permuted dataset was worse than the one for the original dataset, this result indicates that feature j is important for the problem. On the other hand, if the metric for the corrupted dataset is the same or better than on the original dataset, then feature j is not of great relevance. This process was repeated for all features and for $n_repetitions$ iterations, providing greater consistency in the selection of the most important features. In this work, we used $n_repetitions = 5$.

3. Evaluation Metrics

3.1. Classification

To evaluate the classification problem, we used the accuracy, precision, recall, and F1-score.

3.1.1. Accuracy

Accuracy measures the proportion of correct predictions made by the model in relation to all predictions, as represented by Equation (4):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

3.1.2. Precision

Precision measures the proportion of true positives correctly detected relative to the total number of actual positive values, as represented by Equation (5).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

3.1.3. Recall

Recall measures the proportion of true positives correctly detected by the model, as represented by Equation (6).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

3.1.4. F1-Score

The F1-score is the harmonic mean of recall and precision, as represented by Equation (7).

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

3.2. Regression

To evaluate the regression problem, we used the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MdAE), and R^2 .

3.2.1. MSE

The MSE measures the average of the squares of the errors, which are the differences between predicted and actual values, as represented by Equation (8).

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

3.2.2. RMSE

The RMSE measures the square root of the average of the squares of the errors, which are the differences between the predicted and actual values, as represented by Equation (9).

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (9)$$

3.2.3. MAE

The MAE measures the average of the absolute values of the errors, which are the differences between predicted and actual values, as represented by Equation (10).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

3.2.4. MdAE

The MdAE measures the median of the absolute values of the errors, which are the differences between predicted and actual values, as represented by Equation (11).

$$\text{MdAE} = \text{median}(|y_i - \hat{y}_i|) \quad (11)$$

3.2.5. R^2

R^2 , also known as the coefficient of determination, measures the proportion of variance in the dependent variable that can be explained by the independent variable, as represented by Equation (12).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

4. Results

4.1. Model Metrics

The accuracy, precision, recall, and F1-score metrics provide valuable insights into the performance of the machine learning models used to assess the influence of exogenous variables on soybean road freight prices. Accuracy measures the overall correctness of the model's predictions, while precision reflects the proportion of true positive predictions among all positive predictions, indicating how reliable the model is when it predicts a positive outcome. On the other hand, recall captures the model's ability to identify all relevant positive instances, while the F1-score balances precision and recall, providing a comprehensive view of the model's performance, especially in cases of imbalanced data.

Analyzing Table 3, it is evident that LightGBM outperformed the other models across all metrics, demonstrating its superiority for the classification task. Additionally, the close

values of the metrics underscore the model's stability and consistent performance. However, XGBoost delivered comparable results, further emphasizing the strong performance of tree-based algorithms. Both models excel in handling large datasets and effectively extracting the most relevant features. Moreover, they offer a broad set of parameters, enabling extensive hyperparameter tuning to further optimize performance.

It can be observed that the obtained standard deviations are small values, demonstrating that the results do not experience significant variations between the folds.

Table 3. Classification results for Decision Tree, Extra Trees, KNN, LightGBM, Logistic Regression, Passive-Aggressive, Random Forest, and XGBoost evaluated using accuracy, precision, recall, and F1-score. Bold font is used to indicate the best result based on the F1-score.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.752 ± 0.007	0.748 ± 0.007	0.752 ± 0.007	0.748 ± 0.006
Extra Trees	0.757 ± 0.011	0.760 ± 0.012	0.757 ± 0.011	0.744 ± 0.012
KNN	0.737 ± 0.007	0.732 ± 0.007	0.737 ± 0.007	0.732 ± 0.007
LightGBM	0.791 ± 0.007	0.788 ± 0.008	0.791 ± 0.007	0.788 ± 0.008
Logistic Regression	0.686 ± 0.008	0.681 ± 0.009	0.686 ± 0.008	0.682 ± 0.009
Passive Aggressive	0.671 ± 0.008	0.665 ± 0.009	0.671 ± 0.008	0.661 ± 0.010
Random Forest	0.710 ± 0.002	0.737 ± 0.003	0.710 ± 0.002	0.675 ± 0.003
XGBoost	0.786 ± 0.009	0.783 ± 0.009	0.786 ± 0.009	0.782 ± 0.009

For the regression task, analyzing Table 4, the XGBoost model shows an advantage for the MSE, RMSE, and MAE metrics. Despite this, its performance is similar to that of LightGBM, which has a slight advantage in the MdAE metric. Regarding the R² metric, the Passive-Aggressive model performs better, which was expected because the R² metric measures the linearity of the data and the Passive-Aggressive model is linear, allowing it to minimize this metric more efficiently compared to the other models.

Table 4. Regression results for Decision Tree, Extra Trees, KNN, LightGBM, Passive-Aggressive, Random Forest, and XGBoost evaluated using MSE, RMSE, MAE, MdAE, and R². Bold font is used to indicate the best result for each metric (column).

Model	MSE	RMSE	MAE	MdAE	R ²
Decision Tree	861.969 ± 46.459	29.351 ± 0.790	19.620 ± 0.400	11.733 ± 0.218	0.686 ± 0.018
Extra Trees	754.412 ± 34.273	27.461 ± 0.620	18.240 ± 0.363	10.629 ± 0.116	0.725 ± 0.011
KNN	914.181 ± 31.191	30.232 ± 0.514	21.389 ± 0.419	14.665 ± 0.282	0.667 ± 0.011
LightGBM	706.650 ± 24.156	26.580 ± 0.455	17.062 ± 0.362	9.442 ± 0.269	0.742 ± 0.010
Passive Aggressive	1325.033 ± 51.281	36.396 ± 0.702	25.804 ± 0.296	17.330 ± 0.376	0.517 ± 0.014
Random Forest	719.125 ± 35.900	26.810 ± 0.669	17.240 ± 0.394	9.640 ± 0.156	0.738 ± 0.013
XGBoost	697.468 ± 23.423	26.407 ± 0.443	17.022 ± 0.285	9.454 ± 0.175	0.746 ± 0.009

Thus, it can be observed that the LightGBM and XGBoost models are extremely competitive with each other, and achieve better results than the other models for both tasks. This is due to their robust tree structure and the presence of various hyperparameters, allowing for better optimization. However, LightGBM generally exhibits better computational efficiency compared to XGBoost, an aspect that deserves attention.

4.2. Exogenous Variables Influences

Analysis of the predictive variables in the regression and classification models using the unstratified dataset (i.e., the complete dataset \mathcal{D}) revealed the predominance of the “Distance” variable as the most influential factor in determining the price of soybean road

freight. This finding aligns with the expectation of direct influence in this process, as there is a clear relationship with operational costs such as fuel expenses, travel time, tire and parts wear, maintenance, and labor. “Average Monthly Exchange Rate” and “Yearly Export Tonnage by Origin State” emerged as the second and third most relevant variables, respectively, highlighting the importance of the macroeconomic context of Brazilian soybean exports. The exchange rate directly influences the competitiveness of the product in the international market, affecting demand, and consequently the export volume. In turn, the export volume impacts the demand for road freight to transport the production to export ports. Furthermore, for the evaluated models, we observed the presence of similar variables in both scenarios: “Year”, “Crushing Capacity of Industry by Origin State”, “Destination: *Paranaguá*”, “Average Price of Ethanol”, and “Destination State: *São Paulo*” (Figures 1 and 2).

The “Year” variable can capture macroeconomic trends, structural changes in the domestic market, and regulatory adjustments, such as the Truckers’ Law—*Lei dos Caminhoneiros* (Law No. 13,103/2015), which regulated aspects such as working hours, rest periods, waiting times, and the establishment of a minimum floor for freight rates. Additionally, this variable may reflect specific events, such as the 2018 truckers’ strike, which halted cargo transportation nationwide and particularly impacted the agricultural sector [50].

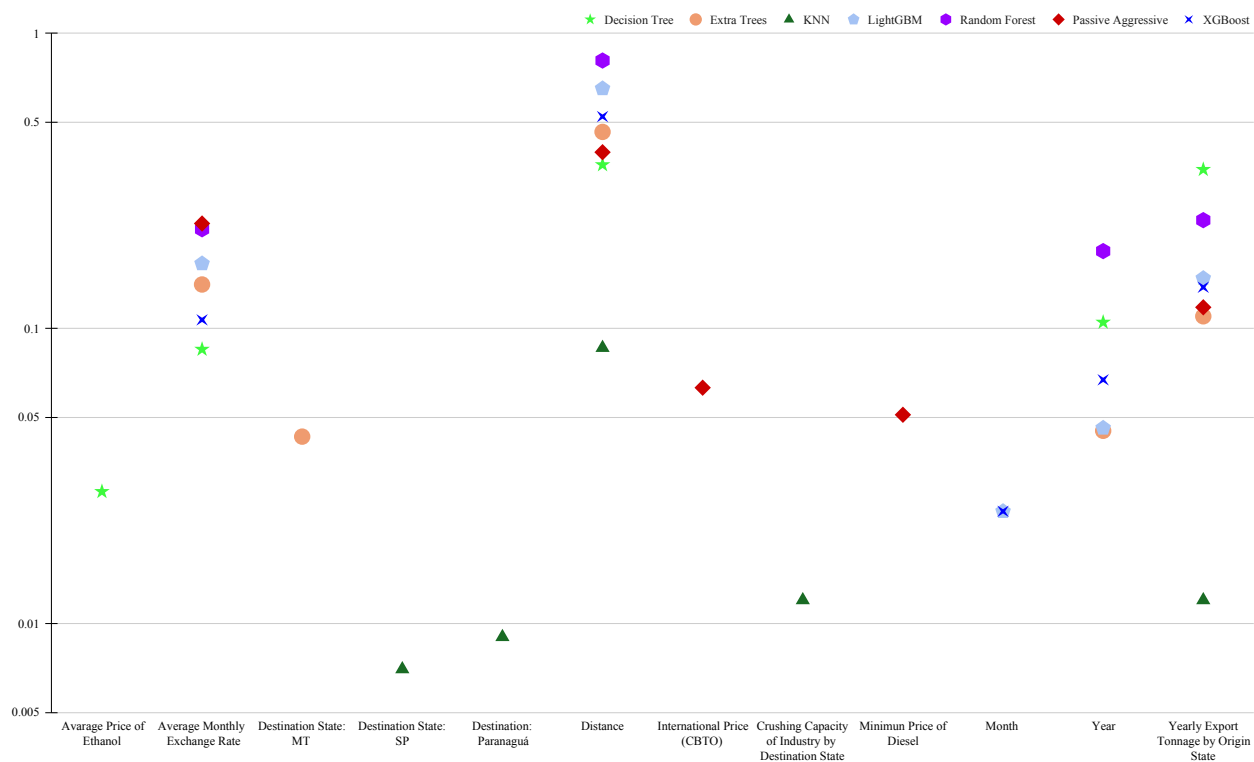


Figure 1. Key variables identified by the regression models along with their importance values determined through permutation importance.

Also pointed out as an influential variable is “Crushing Capacity of Industry”, which is a differentiating factor in soybean marketing, as greater milling capacity leads to stronger bargaining power [51]. In contrast, soybeans that do not undergo industrial processing are largely destined for the external market. In this context, the Port of *Paranaguá* stands out as one of the main hubs for exporting Brazilian soybeans to the international market, especially to China. Its relevance on the national stage is comparable to that of other major export ports, such as the Port of *Santos* located in *São Paulo* [52]. The strategic location of the

Port of *Santos* provides insights into the importance of the "Destination State: São Paulo" variable in the analyzed models.

The data stratification reveals the influence of additional variables in determining the freight price for soybeans. The analysis of high freight rates (Figure 3) also highlights the influence of the "Yearly Export Tonnage by Origin State" variable, especially for the Decision Tree, ExtraTrees, and XGBoost models. Distance is particularly influential for the LightGBM and Random Forest models. "Volume of Exports by State of Origin per Year" is especially relevant in the XGBoost and Decision Tree models. Variables related to the final destination suggest a correlation with the location of the main producing states, as *Mato Grosso* and *Goiás*, both large soybean producers, used the Port of *Santos* as the main export route for most of the period under study. In contrast, *Paraná* and *Mato Grosso do Sul*, both of which are also significant producers, directed their output to the Port of *Paranaguá* [53].

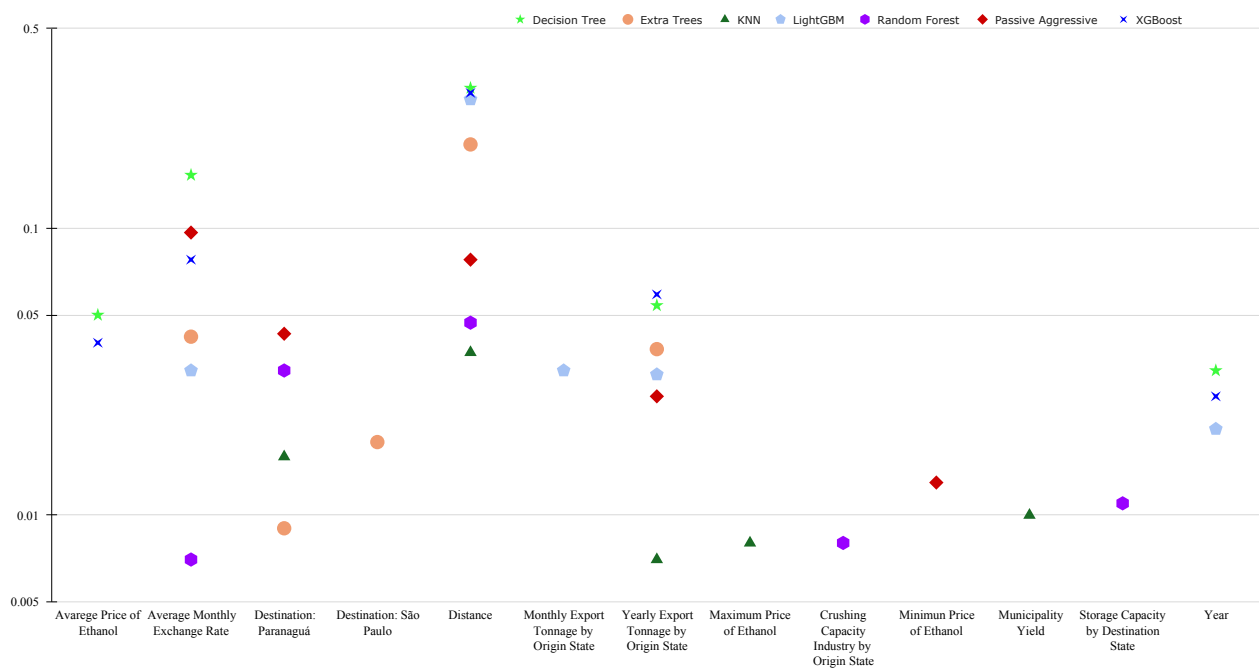


Figure 2. Key variables identified by the classification models along with their importance values determined through permutation importance.

Another variable that appears in high-value freight is "Storage Capacity by Destination State". Storage has a significant impact on prices of agricultural commodities, as it allows for better marketing strategies [54]. In the Brazilian scenario, the storage network does not keep pace with the dynamism of the sector, resulting in a storage deficit. This leads to the so-called sales rush, i.e., when there is a peak in the harvest with a large supply of the product. At this time, soybean prices are low due to the abundant supply, while freight prices are high. Storage is essential for the transfer of production to processing and export centers, as Brazil's main soybean-producing areas are located far from the export ports [55]. Innovative strategies such as rural storage cooperatives are being adopted by producers to reduce costs and improve logistical efficiency [56].

The "Soybean Price (Parity)" variable, which refers to the relationship between domestic soybean prices and international prices, is relevant in the KNN and LightGBM models. Price parity directly impacts producers' marketing strategies, influencing their bargaining power with companies.

Similar to high freight rates, the average freight rate (Figure 4) highlights the importance of the final destination, such as the Ports of *Paranaguá* and *São Luis*, which unlike

the Port of *Santos* are located closer to the main soybean-producing areas. The Port of *São Luis*, in particular, has become increasingly relevant for soybean exports from Mato Grosso, with international destinations such as Hamburg, Germany and Shanghai, China. The increased use of the *Arco Norte* routes, including *São Luis*, presents an efficient alternative that can reduce logistical costs and ease the burden on southern ports, improving Brazil's competitiveness in the global soybean market [52].

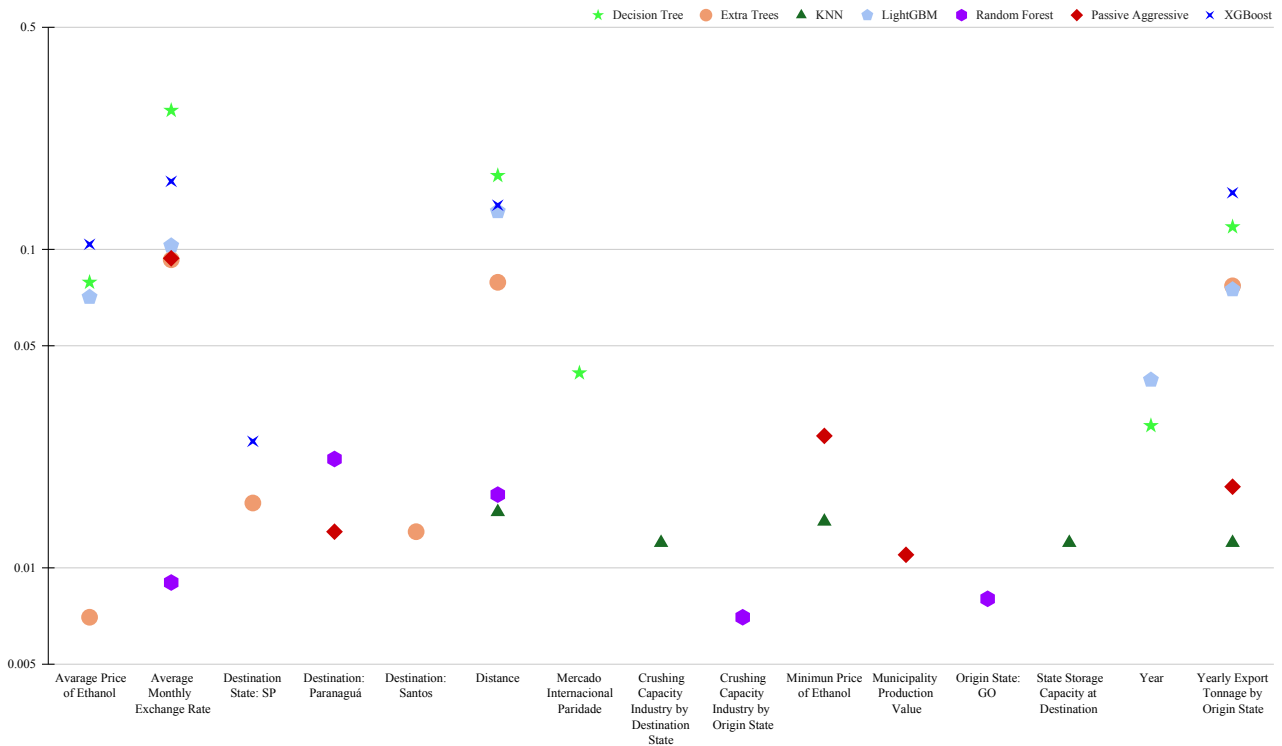


Figure 3. Key variables identified by the classification models for high freight rates along with their importance values determined through permutation importance.

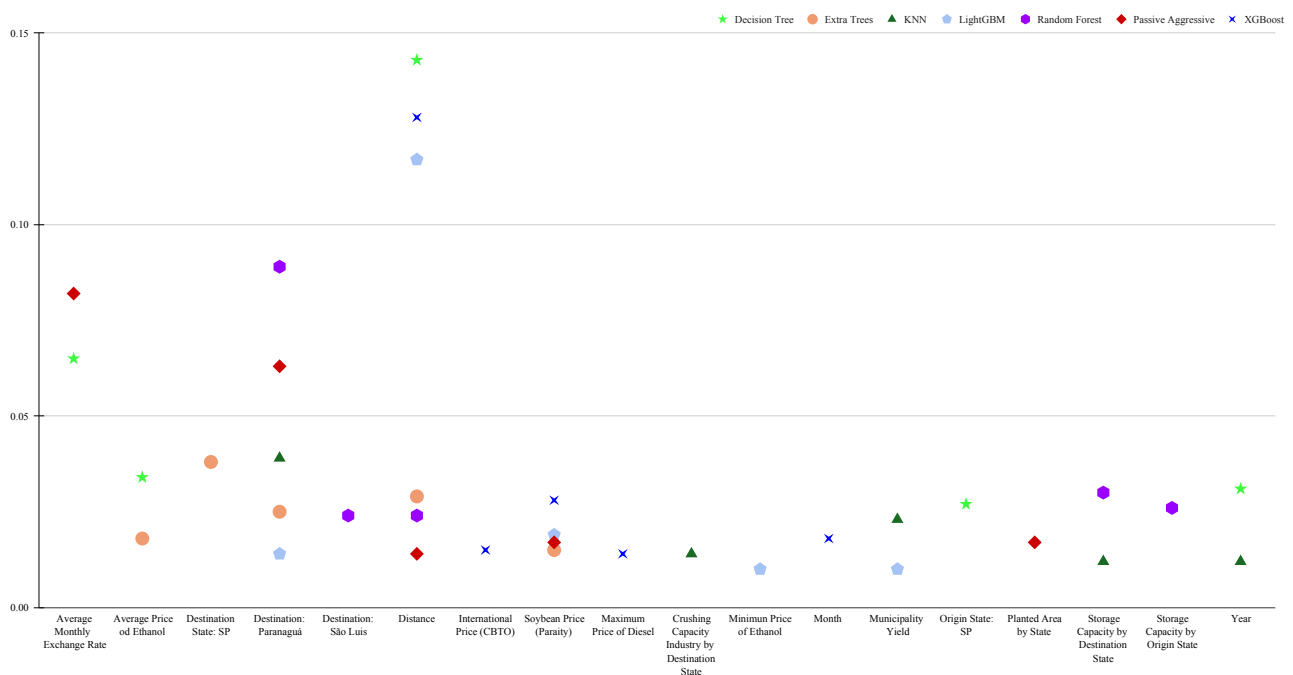


Figure 4. Key variables identified by the classification models for medium freight rates along with their importance values determined through permutation importance.

In short-distance freight (Figure 5), we observe a different scenario from that of medium- and long-distance freight. While long-distance freight clearly indicates the transportation of soybeans destined for export, short-distance freight is more associated with domestic supply. The four cities that appear as short-distance destinations are *Barreiras*, *Maringá*, *Osvaldo Cruz*, and *Uberlândia*. All of these cities have soybean crushing facilities [57] dedicated to the production of oil and meal products, which are widely used in animal feed and biofuel production in the domestic market.

The stratification of the database using the same models highlights the influence of freight rate division on the relevance of predictive variables. While the general scenario analysis identified thirteen variables with significant influence, segmentation by price ranges revealed a gradual increase in this number: fifteen variables in the high freight price scenario, eighteen in the medium freight price scenario, and nineteen in the low freight price scenario. The variables “Year”, “Average Monthly Exchange Rate”, “Distance”, and “Minimum Price of Ethanol” consistently proved to be influential across all scenarios, suggesting their fundamental importance in determining freight costs regardless of the price range. This variation in the number of relevant variables in each stratum demonstrates how stratification can allow for a more granular analysis that captures the influence of different variables in each price range and reaffirms the dynamics of the freight market.

In Brazil, transportation of grains such as soybeans generally begins with road transport, which connects farm production to final destinations such as industries or export ports. In some cases, cargo is initially sent to warehouses or transshipment terminals, from where it proceeds to the final destination via other transport modes such as railways or waterways. This intermodal system aims to reduce costs and optimize the logistics of production flow. Although distance is a determining factor in the road freight price for agricultural products, the negotiation process between market agents has a fundamental impact. The grain transport market is highly competitive, and is marked by an imbalance of power between demand (represented by agricultural trading companies) and supply (composed of small transport companies and independent truck drivers). The trading companies, mostly transnational corporations, use their large cargo volumes to negotiate better freight conditions, taking advantage of fragmentation among transport providers.

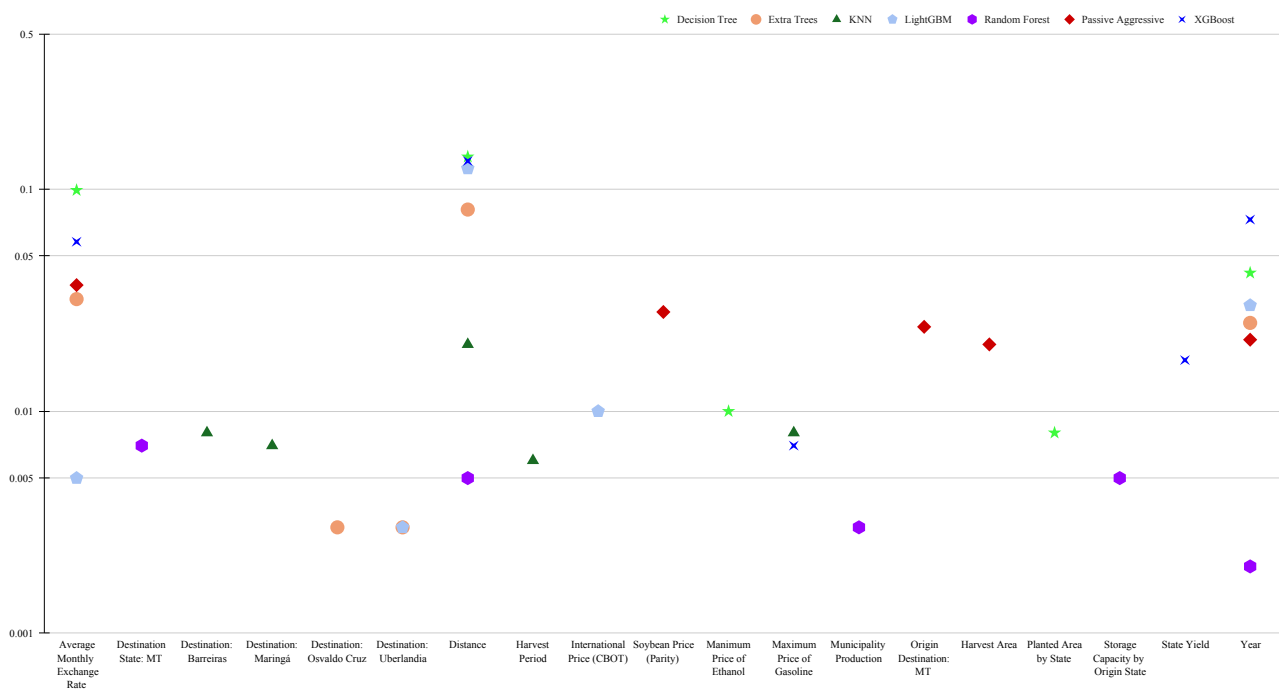


Figure 5. Key variables identified by the classification models for low freight rates along with their importance values determined through permutation importance.

Additionally, the behavior of commodity markets such as the soybean market is complex, and can be influenced by a variety of factors over time. Different elements can determine the prices of these commodities during certain periods, for instance the availability of soybeans during the off-season [58]. Many forecasting models use only historical prices of commodities [59]; however, large price fluctuations can impact not only production and consumption costs but also government regulatory policies [60].

5. Discussion

The results of this study demonstrate the effectiveness of incorporating a wide range of variables when predicting freight costs, such as economic indicators, regional productivity, and logistical infrastructure.

Distance is recognized as the primary factor in determining freight prices due to its direct and significant impact on transportation operational costs [36]. The link between distance and freight cost is both intuitive and supported by empirical evidence, as longer journeys typically demand higher fuel consumption, increased labor hours, and elevated vehicle maintenance expenses [61]. These elements collectively lead to greater costs for logistics providers, which are subsequently transferred to customers.

Furthermore, the findings in this paper highlight the complexity of this relationship within specific logistical frameworks such as agricultural supply chains. In the context of soybean logistics, distance not only quantifies the physical gap between origin and destination but also encapsulates challenges related to infrastructure quality, road conditions, and connectivity, all of which can escalate costs. In areas with inadequate infrastructure, the financial repercussions of distance are even more significant, as poor road conditions prolong travel time, increase fuel usage, and accelerate vehicle deterioration.

Additionally, distance frequently influences the selection of transportation modes, thereby shaping pricing dynamics. Shorter distances may favor road transport due to its flexibility, whereas longer distances might necessitate rail or waterways to achieve cost efficiency. As revealed in the study, this segmentation illustrates how distance interacts

with other variables such as macroeconomic factors (e.g., fuel prices and exchange rates) to affect freight prices across various scenarios.

By quantifying the influence of distance and integrating it into predictive models, this research offers a nuanced perspective on how this variable governs freight pricing, especially in intricate and variable settings such as agricultural logistics. This understanding can enable stakeholders to better anticipate cost fluctuations and strategically optimize their transportation operations.

From a financial and management point of view, our findings highlight the importance of systematically monitoring and updating road freight costs. This periodic systematization can serve as a valuable managerial tool, empowering agricultural traders and grain producers to negotiate transportation service contracts more effectively. In addition, access to accurate and regularly updated data on road transportation costs enhances stakeholders' ability to secure favorable terms and optimize logistics expenditures, ultimately contributing to more efficient supply chain management.

Based on the findings of this research, several public policy recommendations can be proposed to improve freight cost management, logistical efficiency, and sustainability in agricultural supply chains.

First, investment in transportation infrastructure, particularly road networks, is essential for fostering economic growth, enhancing logistical efficiency, and ensuring the competitiveness of agricultural supply chains. Roads serve as the primary mode of transportation in many regions, particularly in developing countries. Importantly, they connect production hubs, processing facilities, and export terminals. Improved road infrastructure reduces transportation costs, minimizes delays, and enhances the reliability of freight movement, directly benefiting sectors such as agribusiness that are heavily reliant on efficient logistics. Furthermore, investments in road infrastructure contribute to regional development by bridging urban and rural areas, facilitating market access for small-scale producers, and fostering socioeconomic inclusion.

Second, investment in multimodal transport infrastructure should be prioritized. This includes improving railways, waterways, and port facilities in order to reduce dependency on road transportation and diversify logistical options. Additionally, addressing regional disparities in infrastructure investment is essential for ensuring equitable access to efficient transportation systems, particularly for agricultural producers in remote or underserved areas.

Third, investment in data infrastructure for freight cost monitoring is crucial. Governments should establish public systems to collect, organize, and disseminate freight cost data while integrating metrics such as economic indicators, regional productivity, logistical infrastructure, and sustainability measures. Public–private partnerships can play a vital role in ensuring that these systems are comprehensive, up-to-date, and accessible to all stakeholders.

Finally, to ensure fairness in pricing, regulatory frameworks should be developed to prevent exploitation of disparities between regions or transportation modes. For example, price caps could protect smaller producers from inflated road freight rates. Transparency in pricing across all transportation modes should also be encouraged to facilitate fair competition and better decision-making by logistics agents.

A direct comparison with other studies is challenging due to variations in the datasets, particularly as our study relies on a uniquely created dataset. Instead, we evaluate other factors, such as the variables involved, the specific characteristics of the datasets, and the modeling approaches employed. Due to these variations, solely relying on evaluation metrics such as accuracy or F1-score would not result in a reliable comparison, as they would measure entirely different contexts and datasets.

Compared to the work of [26], which focused primarily on a limited set of features, our approach provides a more comprehensive understanding of the factors influencing transportation costs. While [26] achieved reasonable accuracy using ensemble models such as LightGBM and XGBoost, their narrower scope potentially overlooked critical variables that contribute to freight dynamics, particularly in contexts with high variability such as agricultural logistics. By including additional explanatory variables and leveraging permutation importance to assess their impact for different models, our findings underscore the importance of a holistic approach to freight cost prediction, offering deeper insights and greater applicability to complex logistical scenarios.

Furthermore, our work is focused on analyzing the impacts of exogenous variables on soybean freight costs in Brazil, offering distinctive contributions compared to the study of [34]. While the latter emphasized operational variables such as vehicle type and cargo weight to determine transportation costs, our study highlights the role of macroeconomic factors, including exchange rates and export volumes, thereby providing a broader and more tailored analysis for the agribusiness context. Additionally, by segmenting our results across short, medium, and long transportation distances and employing explainability techniques, our approach ensures greater interpretability and practical applicability. This supports the international competitiveness of Brazilian soybeans within complex logistics chains. Consequently, our work expands the understanding of freight pricing by integrating economic and geographic dimensions beyond traditional operational factors.

6. Ethical Implications

The application of machine learning models in predicting freight prices has significant ethical implications that must be carefully considered. One key concern is the potential for bias in the data used for training these models. Historical data may reflect systemic inequalities, such as regional disparities in infrastructure investment or socioeconomic imbalances, which could inadvertently be perpetuated by predictive models. Thus, ensuring fairness in predictions is crucial in order to avoid reinforcing existing inequities in logistics and transportation planning [62].

Another concern involves the environmental and logistical implications of optimizing transportation logistics. While machine learning can lead to increased efficiency and reduced costs, these optimizations might unintentionally encourage overuse of certain routes or resources, exacerbating environmental degradation and creating new logistical challenges. For instance, the concentration of freight traffic on specific routes could result in congestion, increased wear on infrastructure, and delays in transportation. Thus, a balanced approach is needed in order to incorporate sustainability metrics such as carbon footprint and resource consumption into predictive models while also considering strategies for distributing traffic evenly and preventing overloading on critical routes.

Finally, the privacy of data sources, such as sensitive commercial or logistical information, must be safeguarded; adopting robust data protection measures is essential to prevent unauthorized access or misuse and comply with legal frameworks such as the General Data Protection Regulation (GDPR) in applicable jurisdictions [63].

Addressing these ethical considerations is fundamental to ensuring that the deployment of machine learning in freight prediction contributes positively to society and aligns with broader ethical standards.

7. Conclusions

Transportation is a crucial component in the final cost of soybeans with a complex and nonlinear relationship. The different variables associated with each price range of soybean freight emphasize the nonlinearity of this behavior across the spectrum. When evaluating

variable classification, the LightGBM model proved to be the most accurate, while the XGBoost, Passive-Aggressive, and LightGBM models stood out in our regression analysis.

Distance is the most significant variable in determining freight costs, aligning with operational expenses such as fuel and labor. This study advances the theoretical understanding of this variable, demonstrating that while distance remains the primary determinant of freight prices, macroeconomic factors such as exchange rates and export volumes significantly impact price variations across different logistic scenarios.

Our findings also contribute to the theory of supply chain management. First, this study suggests that machine learning models for efficiently forecasting freight costs can assist economic agents in agricultural supply chains in mitigating the financial impact of transportation. Second, our identification of the stratified impact of different variables depending on the distance of transport (short, medium, and long) adds a more detailed perspective to transportation cost theories, suggesting that different models should be applied according to the scope of the supply chain. Short-distance freight is more related to domestic supply, such as transportation to soybean crushing plants for the production of oil and meal; in contrast, medium- and long-distance freight are predominantly tied to export logistics. The segmentation of our dataset also highlights an increase in the number of relevant variables for each price range, underscoring the importance of a more granular analysis to capture freight dynamics.

The practical contributions of this study provide a robust foundation for improving predictive models of freight costs. By incorporating a wide range of variables, including economic indicators, regional productivity, and logistical infrastructure, this research offers insights for building efficient models to anticipate market fluctuations. This approach highlights the importance of considering factors beyond traditional metrics, offering a more detailed understanding of cost dynamics in the transportation sector.

From a managerial perspective, our results provide actionable insights for logistics managers and transportation planners. Understanding that distance, exchange rates, and export volumes are key determinants of freight costs allows managers to better anticipate price fluctuations and strategically plan logistics operations. For instance, managers could optimize routing and scheduling decisions by considering macroeconomic indicators in addition to their operational expenses. This predictive capability can support decision-making processes such as choosing optimal transport routes and timing shipments to minimize costs.

Moreover, the different variables can interact with each other, meaning that the impact of one variable partially depends on the values of others. For instance, the influence of distance on freight costs can be amplified or reduced by factors such as road infrastructure and weather conditions. The inclusion of additional variables such as these, along with public policies, can further enhance the analysis, providing a more detailed and accurate view of freight price formation.

One promising direction is to model the transportation problem as a graph in which the nodes represent municipalities, production centers, and ports and the edges correspond to transportation routes with attributes such as distance, cost, and capacity. Leveraging graph-based learning techniques such as Graph Neural Networks (GNNs) or Graph Convolutional Networks (GCNs) could provide a deeper understanding of the structural and relational factors influencing freight costs. These methods are particularly well suited to capturing complex interactions between variables, and could improve prediction accuracy while offering new insights into logistical optimization.

Another potential avenue for future research is expanding the analysis to include data from other countries, particularly in South America. Understanding the dynamics of soybean freight costs across different regions could provide a more comprehensive

view of the global market and help to identify regional factors that influence pricing. This geographical expansion would contribute to a better understanding of logistical challenges and improve the generalizability of the model.

Furthermore, incorporating variables related to sustainability, such as the carbon footprint of transportation and greener agricultural practices, could provide insights into not only financial costs but also the environmental impact of the soybean freight process. This would allow for a more holistic approach that addresses both economic and ecological concerns within the logistics sector.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/su17031067/s1>, Table S1.

Author Contributions: Conceptualization: K.B.M. and A.L.R.d.O.; methodology: P.M. and M.Y.R.U.; validation: K.B.M., M.Y.R.U. and T.C.O.d.C.; writing: K.B.M., A.L.R.d.O., P.M., M.Y.R.U. and T.C.O.d.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Scientific and Technological Development (CNPq), grant number n° 144566/2019-2, and by São Paulo Research Foundation (FAPESP), grant number n° 2018/19571-1.

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- da Silva e Souza, G.; Gomes, E.G.; de Andrade Alves, E.R.; Gasques, J.G. Technological progress in the Brazilian agriculture. *Socio-Econ. Plan. Sci.* **2020**, *72*, 100879. [\[CrossRef\]](#)
- Clott, C.; Hartman, B.C.; Ogard, E.; Gatto, A. Container repositioning and agricultural commodities: Shipping soybeans by container from US hinterland to overseas markets. *Res. Transp. Bus. Manag.* **2015**, *14*, 56–65. [\[CrossRef\]](#)
- Morais, G.R.; Calil, Y.C.D.; de Oliveira, G.F.; Saldanha, R.R.; Maia, C.A. A Sustainable Location Model of Transshipment Terminals Applied to the Expansion Strategies of the Soybean Intermodal Transport Network in the State of Mato Grosso, Brazil. *Sustainability* **2023**, *15*, 63. [\[CrossRef\]](#)
- Kamrud, G.; Wilson, W.W.; Bullock, D.W. Logistics competition between the U.S. and Brazil for soybean shipments to China: An optimized Monte Carlo simulation approach. *J. Commod. Mark.* **2023**, *31*, 290. [\[CrossRef\]](#)
- Wanke, P.; Fleury, P.F. Transporte de Cargas no Brasil: Estudo exploratório das principais variáveis relacionadas aos diferentes modais e às suas estruturas de custos. In *Estrutura e Dinâmica do Setor de Serviços no Brasil*; IPEA: Brasília, Brasil, 2006; pp. 409–464.
- Filassi, M.; Marsola, K.B.; Oliveira, A.L.R. Armazenagem de grãos no Brasil: Um gargalo logístico a ser superado. In Proceedings of the 58° Congresso SOBER—Sociedade Brasileira de Economia, Administração e Sociologia Rural, Foz do Iguaçu, Brazil, 26–28 October 2020.
- Isler, C.A.; Asaff, Y.; Marinov, M. Designing a Geo-Strategic Railway Freight Network in Brazil Using GIS. *Sustainability* **2021**, *13*, 85. [\[CrossRef\]](#)
- Friend, D.J.; da Lima, R.S. Impact of transportation policies on competitiveness of Brazilian and U.S. soybeans: From field to port. *Transp. Res. Rec.* **2011**, *2238*, 61–67. [\[CrossRef\]](#)
- Eurostat. Eurostat Statistics Explained, Road Freight Transport by Journey Characteristics. 2024. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road_freight_transport_by_journey_characteristics (accessed on 20 August 2024).
- CIA (Central Intelligence Agency). The World Factbook. 2021. Available online: <https://www.cia.gov/the-world-factbook/> (accessed on 20 August 2024).
- OECD. The Organization for Economic Co-operation and Development. 2024. Available online: <https://www.oecd.org/en/data/indicators/freight-transport.html> (accessed on 5 August 2024).
- CNT. Boletim Unificado, Setembro 2024. 2024. Available online: <https://www.cnt.org.br/boletins> (accessed on 20 September 2024).
- Brasil. Observatório Nacional de Transportes e Logística. 2024. Available online: <https://ontl.infrasa.gov.br/> (accessed on 2 August 2024).

14. Savić, B.; Petrović, M.; Vasiljević, Z. The impact of transportation costs on economic performances in crop production. *Econ. Agric.* **2020**, *67*, 683–697. [\[CrossRef\]](#)
15. Adisa, O.M.; Botai, J.O.; Adeola, A.M.; Hassen, A.; Botai, C.M.; Darkey, D.; Tesfamariam, E. Application of artificial neural network for predicting maize production in South Africa. *Sustainability* **2019**, *11*, 1145. [\[CrossRef\]](#)
16. Benos, L.; Tagarakis, A.C.; Dolias, G.; Berruto, R.; Kateris, D.; Bochtis, D. Machine learning in agriculture: A comprehensive updated review. *Sensors* **2021**, *21*, 3758. [\[CrossRef\]](#)
17. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM neural network based forecasting model for wheat production in Pakistan. *Agronomy* **2019**, *9*, 72. [\[CrossRef\]](#)
18. Mahesh, P.; Soundrapandiyan, R. Yield prediction for crops by gradient-based algorithms. *PLoS ONE* **2024**, *19*, 0291928. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Sun, C.; Pei, M.; Cao, B.; Chang, S.; Si, H. A Study on Agricultural Commodity Price Prediction Model Based on Secondary Decomposition and Long Short-Term Memory Network. *Agriculture* **2024**, *14*, 60. [\[CrossRef\]](#)
20. Ghutake, I.; Verma, R.; Chaudhari, R.; Amarsinh, V. An intelligent Crop Price Prediction using suitable Machine Learning Algorithm. *ITM Web Conf.* **2021**, *40*, 03040. [\[CrossRef\]](#)
21. Kurumatani, K. Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. *SN Appl. Sci.* **2020**, *2*, 1434. [\[CrossRef\]](#)
22. Araújo, S.O.; Peres, R.S.; Ramalho, J.C.; Lidon, F.; Barata, J. Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy* **2023**, *13*, 2976. [\[CrossRef\]](#)
23. Macarrigue, A.M.J.S.; de Oliveira, A.L.R.; Dias, C.T.D.S.; Marsola, K.B. Multidimensionality of agricultural grain road freight price: A multiple linear regression model approach by variable selection. *Cienc. Rural* **2024**, *54*, 335. [\[CrossRef\]](#)
24. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [\[CrossRef\]](#)
25. Mohanty, M.K.; Thakurta, P.K.G.; Kar, S. Agricultural commodity price prediction model: A machine learning framework. *Neural Comput. Appl.* **2023**, *35*, 15109–15128. [\[CrossRef\]](#)
26. Kulkarni, P.; Gala, I.; Nargundkar, A. Freight Cost Prediction Using Machine Learning Algorithms. In *Proceedings of the Intelligent Systems and Applications*; Kulkarni, A.J., Mirjalili, S., Udgata, S.K., Eds.; Springer: Singapore, 2023; pp. 507–515.
27. Tsolaki, K.; Vafeiadis, T.; Nizamis, A.; Ioannidis, D.; Tzovaras, D. Utilizing machine learning on freight transportation and logistics applications: A review. *ICT Express* **2023**, *9*, 284–295. [\[CrossRef\]](#)
28. Silva, R.F.; Paula, A.F.M.; Mostaço, G.M.; Costa, A.H.R.; Cugnasca, C.E., Soybean Price Trend Forecast Using Deep Learning Techniques Based on Prices and Text Sentiments. In *Information and Communication Technologies for Agriculture—Theme II: Data*; Bochtis, D.D., Moshou, D.E., Vasileiadis, G., Balafoutis, A., Pardalos, P.M., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 235–266. [\[CrossRef\]](#)
29. Das, N.; Sadhukhan, B.; Chatterjee, R.; Chakrabarti, S. Integrating sentiment analysis with graph neural networks for enhanced stock prediction: A comprehensive survey. *Decis. Anal. J.* **2024**, *10*, 100417. [\[CrossRef\]](#)
30. Wu, S.; Liu, Y.; Zou, Z.; Weng, T.H. S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. *Connect. Sci.* **2022**, *34*, 44–62. [\[CrossRef\]](#)
31. Fan, K.; Hu, Y.; Liu, H.; Liu, Q. Soybean futures price prediction with dual-stage attention-based long short-term memory: A decomposition and extension approach. *J. Intell. Fuzzy Syst.* **2023**, *45*, 10579–10602. [\[CrossRef\]](#)
32. Wang, C.; Gao, Q. High and Low Prices Prediction of Soybean Futures with LSTM Neural Network. In *Proceedings of the 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 23–25 November 2018; pp. 140–143. [\[CrossRef\]](#)
33. Osinga, S.A.; Paudel, D.; Mouzakitis, S.A.; Athanasiadis, I.N. Big data in agriculture: Between opportunity and solution. *Agric. Syst.* **2022**, *195*, 103298. [\[CrossRef\]](#)
34. Jang, H.S.; Chang, T.W.; Kim, S.H. Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning. *Sustainability* **2023**, *15*, 1122. [\[CrossRef\]](#)
35. Liachovičius, E.; Šabanovič, E.; Skrickij, V. Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport* **2023**, *38*, 231–242. [\[CrossRef\]](#)
36. Archetti, C.; Peirano, L.; Speranza, M.G. Optimization in multimodal freight transportation problems: A Survey. *Eur. J. Oper. Res.* **2022**, *299*, 1–20. [\[CrossRef\]](#)
37. Moreira, C.E.S.; de Oliveira, A.L.R.; de Medeiros Oliveira, S.R.; Yamakami, A. Identification of freight patterns via association rules: The case of agricultural grains. *Bulg. J. Agric. Sci.* **2017**, *23*, 887–893.
38. Oliveira, A.L.R.; Filassi, M.; Lopes, B.F.R.; Marsola, K.B. Logistical transportation routes optimization for Brazilian soybean : An application of the origin-destination matrix. *Ciênc. Rural* **2021**, *51*, 20190786. [\[CrossRef\]](#)
39. Kengpol, A.; Tuammee, S.; Tuominen, M. The development of a framework for route selection in multimodal transportation. *Int. J. Logist. Manag.* **2014**, *25*, 581–610. [\[CrossRef\]](#)

40. Márquez, L.; Cantillo, V. Evaluating strategic freight transport corridors including external costs. *Transp. Plan. Technol.* **2013**, *36*, 529–546. [CrossRef]
41. Péra, T.G.; Bartholomeu, D.B.; Su, C.T.; Filho, J.V.C. Evaluation of green transport corridors of Brazilian soybean exports to China. *Braz. J. Oper. Prod. Manag.* **2019**, *16*, 398–412. [CrossRef]
42. Melo, I.C.; Junior, P.N.A.; Perico, A.E.; Guzman, M.G.S.; do Nascimento Rebelatto, D.A. Benchmarking freight transportation corridors and routes with data envelopment analysis (DEA). *Benchmarking* **2018**, *25*, 713–742. [CrossRef]
43. de Oliveira Melo Cicolin, L.; de Oliveira, A.L.R. Avaliação de desempenho do processo logístico de exportação do milho brasileiro: Uma aplicação da análise envoltória de dados—DEA. *J. Transp. Lit.* **2016**, *10*, 30–34. [CrossRef]
44. Filippi, A.C.G.; Guarnieri, P. Novas formas de organização rural: Os Condomínios de Armazéns Rurais. *Rev. Econ. Sociol. Rural* **2019**, *57*, 270–287. [CrossRef]
45. Teixeira, M.M.A.; Losekann, L.D.; Rodrigues, N. Mercado de frete rodoviário e transmissão assimétrica de preço do diesel no Brasil. *Rev. Bras. Energ.* **2020**, *26*, 29–38. [CrossRef]
46. Wetzstein, B.; Florax, R.; Foster, K.; Binkley, J. Transportation costs: Mississippi River barge rates. *J. Commod. Mark.* **2021**, *21*, 100123. [CrossRef]
47. Asai, G.; Piacenti, C.A.; Gurgel, A.C. Impactos no Comportamento do Frete: Uma Aplicação de Equilíbrio Geral Computável para os Produtos Agropecuários do Brasil. *Internext* **2020**, *15*, 17–33. [CrossRef]
48. Sonaglio, C.M.; Zamberlam, C.O.; Filho, R.B. Variações cambiais e os efeitos sobre exportações brasileiras de soja e carnes. *Rev. Política Agrícola* **2011**, *20*, 5–23.
49. Watanabe, S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv* **2023**, arXiv:2304.11127. [CrossRef]
50. Kreter, A.C.; de Castro Junior, J.S.R.; Associado, J.S.P. Impactos Iniciais da Greve dos Caminhoneiros no Setor Agropecuário, 2018. Available online: <https://www.ipea.gov.br/cartadeconjuntura/index.php/2018/06/impactos-iniciais-da-greve-dos-caminhoneiros-no-setor-agropecuario/> (accessed on 10 July 2024).
51. Martins, R.S.; Lobo, D.S.; Araújo, P. Sazonalidade nos fretes e preferências dos embarcadores no mercado de transporte de grãos agrícolas. *Rev. Econ. Adm.* **2005**, *4*, 68–96. [CrossRef]
52. USDA. Oilseeds: World Markets and Trade, 2024. Available online: <https://apps.fas.usda.gov/psdonline/app/index.html#/app/home> (accessed on 3 August 2024).
53. COMEXSTAT. COMEXSTAT, 2024. Available online: <https://comexstat.mdic.gov.br/pt/home> (accessed on 16 June 2022).
54. Delai, A.P.D.; Araujo, J.B.; Reis, J.G.M.; da Silva, L.F. Armazenagem e ganhos logísticos: Uma análise comparativa para comercialização da soja em Mato Grosso do Sul. *Rev. Agronegócio Meio Ambiente* **2015**, *10*, 395–414. [CrossRef]
55. de Lima, D.P.; Fiorioli, J.C.; Padula, A.D.; Pumi, G. The impact of Chinese imports of soybean on port infrastructure in Brazil: A study based on the concept of the “Bullwhip Effect”. *J. Commod. Mark.* **2017**, *9*, 55–79. [CrossRef]
56. Filippi, A.C.G.; Figueiredo, R.S. Associação da relação entre os preços de fretes de soja e óleo diesel no período de 2015 a 2018. *Rev. Eniac Pesquisa* **2019**, *8*, 254–269. [CrossRef]
57. ABIOVE. Estatísticas Cadeia da Soja, 2024, 2024. Available online: <https://abiove.org.br/estatisticas-cadeia-da-soja/> (accessed on 5 September 2024).
58. Ghalayini, L. Modeling and forecasting spot oil price. *Eurasian Bus. Rev.* **2017**, *7*, 355–373. [CrossRef]
59. Drachal, K. Some Novel Bayesian Model Combination Schemes : An Application to Commodities Prices. *Sustainability* **2018**, *10*, 1–27. [CrossRef]
60. Guo, Y.; Tang, D.; Tang, W.; Yang, S.; Tang, Q.; Feng, Y.; Zhang, F. Agricultural Price Prediction Based on Combined Forecasting Model under Spatial-Temporal Influencing Factors. *Sustainability* **2022**, *14*, 2–18. [CrossRef]
61. Huertas, J.I.; Serrano-Guevara, O.; Díaz-Ramírez, J.; Prato, D.; Tabares, L. Real vehicle fuel consumption in logistic corridors. *Appl. Energy* **2022**, *314*, 118921. [CrossRef]
62. Mehrabi, N.; Morstatter, F.; Saxena, N.A.; Lerman, K.; Galstyan, A.G. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv. (CSUR)* **2019**, *54*, 1–35. [CrossRef]
63. Voigt, P.; von dem Bussche, A. *The EU General Data Protection Regulation (GDPR)*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.