Journal of Management Matters

Journal of Management Matters
10(1) (2023) 19-34
@Faculty of Management Studies
Rajarata University of Sri Lanka
https://fms.rjt.ac.lk/mgt-matters-about/

Article

Crop Price Prediction Using Machine Learning Approaches: Reference to the Sri Lankan Vegetable Market

H.M.B.P. Ranaweera[1, *], R.M.G.H.N. Rathnayake[2], and A.S.G.J.K. Ananda[3]

1, 2. Department of Information Systems, Faculty of Management Studies, Rajarata University of Sri Lanka

3. Digiratina Technology Solutions (Pvt.) Ltd., 399/2/1, Pepiliyana, Colombo.

* buddhika@mgt.rjt.ac.lk

Article History
Received Date: 03.04.2023
Revised Date : 03.08.2023
Published date: 30.09.2023

Abstract

Every sector of this digital world is experiencing a noticeable change due to the development of information and communication technology, and the agriculture sector can be introduced as one sector that has undergone a remarkable revolution. Sri Lanka is a developing country in a tropical region and agriculture is the backbone of the nation, and plays a major role in the country's economy. The price that farmers receive for their harvest is critical to farmer satisfaction as well as the future survival of agriculture and primarily it depends on several factors such as demand, seasonal trends, and price offers from multiple suppliers. In recent years, crop prices in Sri Lanka have fluctuated drastically due to unpredictable climate change, natural calamities and many other circumstances. As the farmers were unaware of these uncertainties, they suffered huge losses in their harvest and became disillusioned and most of them intended to give up farming. Therefore, crop price forecasting seems to be a crucial factor in considering the future of agricultural production. Since the properties of crop prices are highly non-linear and combined with significant noise, forecasting crop prices is not an easy problem. Recently, many researchers have proposed various approaches for crop price predicting, among which data mining can be identified as an emerging approach that plays an important role in decision-making related to agricultural product price forecasting. However, in the context of Sri Lanka, there is no evidence of extensive studies on the use of data mining approaches for predicting crop prices, particularly for vegetables. The main objective of this research is to fill the gaps in the literature by assessing the predictability of vegetable prices in the context of Sri Lanka using data mining techniques. The variation in crop price was analyzed based on four factors namely rainfall, temperature, fuel price and crop production and experiments were conducted on four systematically selected vegetables covering up-country and low-country. Analysis was performed using five widely-used machine learning algorithms on similar phenomena and performance was evaluated using common evaluation metrics such as mean absolute error and root-mean-square error. Experimental results revealed that tree-based models are superior among the classifiers considered in forecasting vegetable prices in Sri Lanka.

Keywords: agriculture; crop price; data mining; machine learning; prediction; Sri Lanka, vegetable.

1. Introduction

Sri Lanka has more than 2500 years of agricultural history and today agriculture has become one of the main pillars of our economy. Paddy cultivation has been an economic activity on the island for centuries and it reflects the social, cultural, religious and economic lifestyle of the countrymen (Dilmah, 2020). Over time, agriculture was transformed into a consumption-based trade, which was further taken into the international market. In the post-colonial era and the open economy, the agriculture, industrial and service sectors diversified and faced different challenges and experiences (Gunawardana, 2018).

On average, 40 percent of the agricultural sector goes to rice and 22 percent to other crops like fruits, vegetables and field crops (Kumara, 2017). Although vegetables are not the staple food of Sri Lankans, the vegetable industry has become an important means of increasing farmers' income (Ye et al., 2016). Looking at the progress of vegetable crop cultivation in Sri Lanka in 2018, the progress of up-country vegetable cultivation was 5,091 hectares and the progress of low-country vegetable cultivation was 14,373 hectares (Socio-Economics & Planning Centre, 2018). Today, agriculture as a whole is facing many problems such as low mechanization, low crop productivity and low-value addition to primary products. In addition, the unexpected fluctuation in crop prices is a major problem (Kumara, 2017).

Predicting the price of their yield is very important for any farmer and it is very important to know in advance how much they will earn for their crop (Kaur, 2014). Vegetable prices have fluctuated dramatically in recent years due to unpredictable climate change, natural disasters, economic crises and many other issues. There is no proper mechanism to compensate farmers when prices of vegetables unexpectedly fall, and consumers have to spend more on food in case of rising prices. Therefore, an unexpected increase or decrease in vegetable prices affects the stable life of every family. Accurate forecasting of vegetable prices can effectively guide farmers to make reasonable production decisions, and reduce economic losses for consumers (Ye et al., 2016).

Vegetable prices fluctuate due to natural climate, production, market supply, consumer demand, government regulations and many other factors (Ye et al., 2016). The properties of vegetable prices are highly non-linear and combined with significant noise, and thus vegetable prices are difficult to predict (Nasira & Hemageetha, 2012). According to the daily price report of the Central Bank of Sri Lanka, price inconsistencies in most crops have increased the risk faced by farmers over the past few years (Central Bank of Sri Lanka, 2020). As fluctuations in crop prices adversely affect farmers and consumers, finding a systematic and fair mechanism for forecasting crop prices becomes a more important and high-priority task. Accurate forecasting of vegetable prices will enable farmers to make correct crop decisions and reduce economic losses to consumers (Ye et al., 2016).

The main objective of the forecasting model is to enable farmers to make informed decisions and manage price risk. Moreover, it affects the understanding and prediction of crop performance under different environmental conditions and can increase farm productivity. Price forecasting can be modelled as a regression function that allows researchers to determine how much the predictors of crop prices affect the target variable. In the past few years, many researchers have proposed various approaches such as time series analysis and data mining for crop price prediction. Currently, there is increasing interest in applying data mining and machine learning approaches to crop price forecasting (Kaur, 2014; Rachana et al., 2019; Varun et al., 2019). However, there is a dearth of extensive research on crop price forecasting using machine learning techniques in the context of Sri Lanka. It is important to conduct research on the applicability of machine learning for predicting crop market prices and it would be

beneficial for future researchers to extend the same direction. Considering all the pieces of evidence revealed, the main objective of the study is to identify the suitability of machine learning approaches to forecast crop prices more efficiently in the context of the Sri Lankan vegetable market.

The remainder of this article is organized as follows. The next section contains a review of the literature. Then, the research methodology adopted is discussed. Afterwards, data analysis and results are presented. The results are discussed in the subsequent section. The final section summarizes and concludes the study by presenting implications and recommendations for future studies.

2. Literature Review

The vegetable sub-sector in the agricultural sector occupies a major place and vegetables are grown all over the country and a large number of farmers are engaged in it. The majority of the upland farmers earn their primary income from vegetable farming, while in urban areas some paddy fields grow vegetables and this trend is increasing (Rupasena, 1999). Substandard seeds, high input costs, lack of finance and scarcity of water resources are some of the problems faced by vegetable farmers while price fluctuation is a major problem faced by vegetable farmers (Rupasena, 1999, Henegedara, 2016). As in many developing countries, price volatility and declining trade in small-scale farming have become critical factors in household agriculture in Sri Lanka today, greatly affecting the livelihoods of small farmers as well as local food consumption, leading to macroeconomic imbalances (Henegedara, 2016). On the other hand, nearly 40 percent of the monthly income of Sri Lankans is spent on food and thus any change in the price of food crops will adversely affect the income and lifestyle of the people of the country (Henegedara & Abeykoon, 2015).

Vegetable prices are analyzed both seasonally and annually. As vegetable consumption does not change in the short term, prices are mainly determined by seasonal supply. Seasonal changes are mainly due to the seasonality of production and its relatively high variability, while the change in annual prices reflects the balance of supply and demand over time (Rupasena, 1999). Although the theory provides sufficient coverage to explain the output of vegetable products, despite the technological innovations initiated by policymakers, empirical evidence on price and supply changes in vegetable cultivation in Sri Lanka shows that there are convergent and constant fluctuations in vegetable cultivation. Hence price volatility has adversely affected declining farm income and sustainable farming (Henegedara & Abeykoon, 2015).

Factors affecting vegetable price variability were mainly associated with physical and climatic factors. Vegetable crops are very sensitive to climate change and sudden rises in temperature as well as erratic rainfall can affect any stage of crop growth (Abewoy, 2018). Rainfall and temperature are the main weather factors that affect the fluctuation of vegetable prices. Among all standard climatic parameters, rainfall is the most variable parameter in time and space and is extremely important to the physical and cultural landscape of any region. Rainfall intensity varies considerably across the island and several agro-climatic zones such as the wet zone, intermediate zone and dry zone are identified based on rainfall. The amount of rainfall that changes over time in an area is an important feature of that area's climate. Changes in magnitude, intensity and frequency affect the environment and society. If the frequency and intensity of rainfall are different, the climate can be very different. Therefore, rainfall and temperature affect the production and prices of different crops in different regions (Mathanraj & Kaleel, 2016).

The literature on price behaviour and supply response to food crop agriculture has shown that the prices of vegetables are determined by variations in supply patterns, which have occurred due to seasonality and the production gap for vegetables (Henegedara & Abeykoon, 2015). Its impact is critical in determining food supply, food security, employment and farm income.

Fuel prices directly affect vegetable prices as many farmers and middlemen in Sri Lanka use heavy vehicles such as Lorries and trucks to transport their products to market (Kumarage, 2000). The fuel used for most heavy vehicles is diesel and the cost of fuel has to be borne by farmers and middlemen.

Although the theory provides adequate coverage for price and supply changes in vegetable cultivation in Sri Lanka despite technological innovations initiated by policymakers, it shows that there are convergent and persistent fluctuations in vegetable cultivation. Hence price volatility has adversely affected declining farm income and sustainable farming (Henegedara & Abeykoon, 2015).

Nowadays, devices handled by people, built-in sensors and various organizations generate large amounts of data but the knowledge hidden in these databases does not seem to be effectively used in the decision-making process, especially in developing countries. Data mining tools can be used to uncover hidden information in these large amounts of data, which attempt to identify patterns in data that are difficult to identify with automated patterns and traditional statistical methods (Ashoori et al., 2015). Data mining is the extraction of knowledge from large databases or sets and some of these tasks are searching for concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis and similarity analysis. Moreover, data mining is a creative process that requires various skills and knowledge, and data mining is still expected to be a push-button technology in the marketplace (Wirth & Hipp, 2000).

Farmers not only harvest vegetables and crops but also harvest huge amounts of data. Data mining along with machine learning provides a methodology for transforming this data into useful information for decision making (Nasira & Hemageetha, 2012). The application of data mining and machine learning in agriculture is a novel direction of research. The vegetable market can be considered as a specific application of data mining and can be used to create innovative models for market agricultural forecasting of yields and market prices of related commodities and is also useful for farmers to plan their cropping operations. There is a growing trend of applying data mining approaches, especially machine learning, to forecasting crop traits such as yield and price, and many researchers invest their time every day in coming up with strategies that can improve the accuracy of crop projection models (Crane-Droesch, 2018; Cai et al., 2017).

Recently, data mining techniques have been used in vegetable market price forecasting and have been successfully demonstrated to generate high predictive accuracy of crop price movement (Jothi et al., 2017). Rachana et al. (2019) used the Naïve Bayes algorithm in crop price forecasting considering yield, rainfall, minimum support price and maximum trade as primary determinants of crop price and observed that Naïve Bayes has high applicability in crop price prediction. In evaluating the models, the researchers used WEKA, a collection of machine learning algorithms for data mining tasks developed at the University of Waikato in New Zealand. Nasira and Hemageetha (2012) developed a forecasting model using a neural network to predict tomato prices and observed that a neural network is one acceptable method of forecasting vegetable market prices with non-linear time series. In the study of Varun et al. (2019), researchers applied linear regression considering yield, maximum trade, minimum

trade, rainfall, wind speed, humidity, temperature, cloud cover, pesticides, yield, fertilizer, soil and sunlight as determinants of crop prices.

During their study, the researchers revealed that because the target variable is always numerical, price forecasting can be formulated as a regression function, and linear regression is well-applied for crop price forecasting in Indian markets. Deepalakshmi et al. (2019) found a new algorithm for dynamic price forecasting of vegetable commodities using statistical price forecasting techniques combining data mining methodologies. In that study, vegetable pricing methodology on agricultural price forecasting, various surveys on agricultural data mining methods for process price forecasting and chronological analysis of estimated prices have been discussed. Guo et al. (2022) proposed a price forecasting model based on Back Propagation Neural Network models. There, experiments were conducted considering corn prices in Sichuan Province, China, and it was shown that the proposed mechanism not only predicts the price in steady data changes, but also provides accurate predictions in periods of large price changes. For Boro, Jute and Potato price forecasting in Bangladesh, Shakoor et al. (2017) have studied the applicability of Decision Tree and K-Nearest Neighbors Regression algorithms.

There, the researchers observed that Decision Tree Learning was slightly better than K-Nearest Neighbor Regression in predicting crop prices, although both algorithms offered better accuracy. The study by Jothi et al. (2017), introduced data analysis techniques for crop price estimation using Linear Regression, Decision Trees, XGBoost, and Neural Networks classifiers with the help of the WEKA data mining tool. Subhasree and Priya (2016) developed a model for crop price forecasting which was derived with high accuracy using Backpropagation Neural Network, Genetic algorithm, and Radial Basis function provided in WEKA. Paul et al. (2022) applied Generalized Neural Network, Support Vector Regression, Random Forest and Gradient Boost to forecast the wholesale price of brinjal in India. In their study, they used accuracy measures such as Mean Error, Root Mean Square Error, Mean Absolute Error, and Mean Absolute Prediction Error and found that the Generalized Neural Network outperformed other considered algorithms.

3. Methodology

The objective of this study was to fill the gap in assessing the applicability of machine learning in vegetable price forecasting in the Sri Lankan context while improving crop price forecasting through data mining and thereby improving understanding of the field.

The proposed work mainly focused on forecasting vegetable prices in Sri Lanka based on the best machine learning algorithm identified among logically selected classifiers. The inputs of the system were crop prices, rainfall data, temperature figures, diesel prices and crop production for the last four years ranging from 2018 to 2021. All inputs considered were numeric values. Since the expected output crop price is also a numerical value, the price forecast can be formulated as a regression function.

The collected data were pre-processed as the raw data consisted of many errors and inconsistencies. Then, a machine learning algorithm followed by 10-fold cross-validation was applied to each vegetable individually to identify the best-performing algorithm among the classifiers considered. Finally, the predictive ability of the best algorithm was verified based on a known validation set. The working flow of the proposed setup is shown in Figure 1.

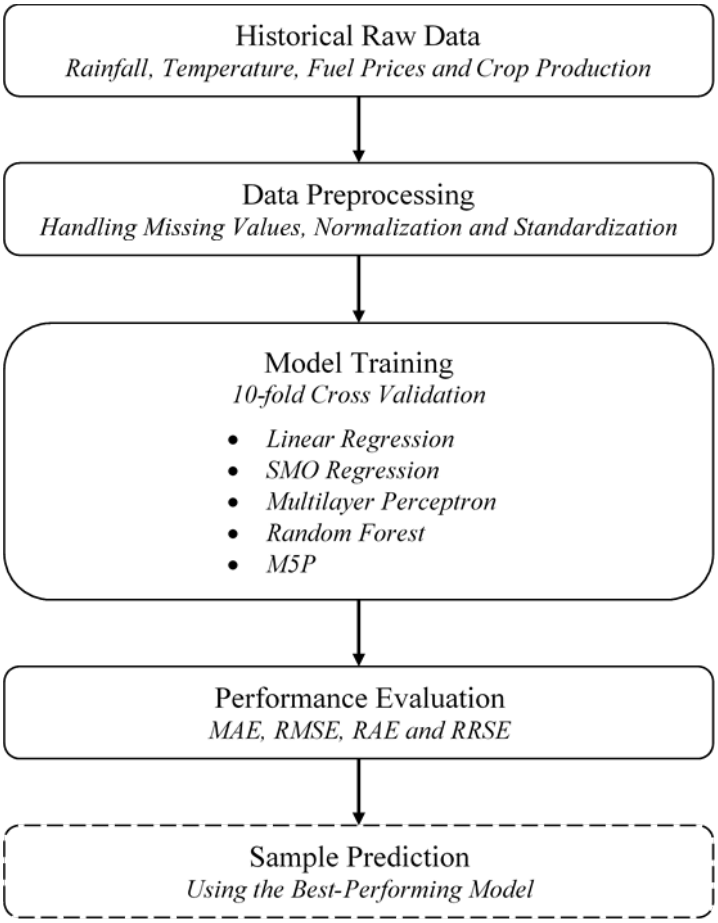

Figure 1. General Framework of the Proposed Arrangement

According to available statistics, more than 16 vegetables are grown in Sri Lanka. Considering the priority given by the farmers to the crops, four crops with the highest priority were selected for the experiments covering two crops from the upland area and another two from the lowland area. There are twenty-six dedicated economic centres in Sri Lanka and crop price and daily supply data were collected from five conveniently selected Dambulla, Thabuttegama, Nuwara Eliya, Narahenpita, and Embilipitiya to represent the Sri Lankan vegetable market. Prices of selected vegetables collected from daily price reports of the Central Bank of Sri Lanka were tallied with prices obtained from economic centres. Weather data covering rainfall and temperature were collected from the Meteorological Department. Meanwhile, the daily supply of vegetables is obtained from the Department of Agriculture and compared with the data obtained from economic centres. Diesel prices were obtained from the website of the Ceylon Petroleum Corporation. The data from the last four years has been taken into consideration in this study.

The subsequent task of historical crop data collection was data preprocessing. Data preprocessing is an essential sub-task of data mining, which involves converting the raw data into a more compact format as the raw data is usually inconsistent or incomplete and usually contains many errors (Jothi et al., 2017). Data were preprocessed by scanning for missing values, finding rank values, and finally comparing them to common environments by applying a feature scale to limit the range of variability. Especially, some of the production data was generated by interpolation, since the Department of Agriculture maintains production data on a time frame limited to specific cropping periods. In addition, data cleaning, data integration, data transformation and data reduction were performed during this phase.

After data preprocessing, models were trained using logically selected machine learning algorithms and the performance of each algorithm was observed. A brief description of the machine learning algorithm applied in this study is as follows.

3.1. Linear Regression

Linear regression is a basic and widely used predictive analysis that can be easily used to classify domains with numerical properties. The simplest form of regression equation has one dependent and one independent variable and is defined by the following formula.

$$y = c + bx$$

Where,

y - Estimated dependent variable score

c - Constant

b - Regression coefficient

x - Score on the independent variable

Regression analysis is a predictive modelling technique that examines the relationship between dependent and independent variables. This technique is used to forecast, create time series models and find causal relationships between variables (Lewis-Beck, 2015). Regression analysis is an important tool for data formatting and analysis. Here, the researcher fits the curve to the data points so that the difference between the distances between the data points on the curve or line is minimized.

3.2. SMO Regression

Sequential Minimum Optimization (SMO) uses Support Vector Machine (SVM) algorithm to enable regression. SMO Regression implements an SMO algorithm for training support vector regression using polynomial or Radial Basis Function (RBF) kernels. This activation restores all missing values and converts nominal attributes to binary ones. It normalizes all attributes by default, so the output coefficients are based on the standardized data and not the original data (Shevade et al., 2000).

3.3. Multilayer Perceptron

A multilayer perceptron is a class of fully connected feed-forward artificial neural networks that uses back-propagation to learn a multi-layer sense to classify instances. Here, the network can be built manually or configured using simple heuristics, and network parameters can also be monitored and modified during training. The nodes in this network are all sigmoid and the output nodes become unbounded linear units, except when the class is numeric.

The most widely used Artificial Neural Network (ANN) is the Multilayer Perceptron, in which neurons are organized into layers. These neural networks have self-learning capabilities that can produce better results when more data is available. Input layer neurons receive the input signal and feed it to the network and these neurons perform no other function. The neurons in the output layers are activated and the output provided by them is considered as the output provided by the network. There may be several hidden layers between the input and

output layers and each neuron receives an input signal from the neurons in the previous layer and transmits its output to the neurons in the successive layer (Nasira & Hemageetha, 2012).

3.4. Random Forest

Random Forest is an ensemble learning algorithm that can be successfully used for classification and regression tasks. This algorithm is a collection of tree predictors and each tree depends on the values of an independently sampled random vector with the same distribution for all trees in the forest (Breiman, 2001). This approach has shown excellent performance in settings where the number of variables is larger than the number of observations, as it combines multiple random decision trees and aggregates their predictions by averaging (Biau, 2016). In addition, this algorithm can be used for complex problems and because it is simple to modify, it can be successfully applied to various ad-hoc learning tasks. The Random Forest algorithm builds decision trees using a greedy algorithm that selects the best split point at each step of the tree-building process. In some cases, this adversely affects the performance of the algorithm, and the resulting trees look too similar, which reduces the variance of the predictions from all the bags and thus damages the robustness of the predictions made.

3.5. M5P

M5P is a refactoring of Quinlan's M5 algorithm for creating trees in regression forms. M5P combines a traditional critical tree with the ability to enable linear regression at 32 nodes. In addition, M5P can work effectively and efficiently with calculated properties and missing values. Here, initially, a decision-tree inductive algorithm is used to create a tree, and a split criterion is used, which minimizes the internal sub-variability of the class values below each branch, rather than maximizing the information from each internal node. Second, the tree is pruned at every leaf, and when pruned, an internal node becomes a leaf with a reflective plane. Finally, to avoid sharp discontinuities between subspecies, a smoothing procedure is adopted, combining the leaf model prediction with each node along the path back to the root, and smoothing it by combining it with the predicted value of each of these nodes. A linear model for smoothing those nodes significantly increases the prediction accuracy (Moraru et al., 2010).

In this study, descriptive statistics were used to justify the quantitative nature of data collection. Descriptive statistics provide simple summaries of samples and measurements and are used to describe the basic characteristics of a study's data (Williams, 2007). Detailed analysis in this study was performed with the help of WEKA version 3.8.6 and many evaluation measures are described in terms of classification scenarios rather than numerical predictive scenarios. In evaluating the performance of machine learning algorithms, several evaluation metrics were used, a brief explanation of which is given below.

3.6. Performance Evaluation

Mean Absolute Error (MAE)

MAE is measured regardless of the direction of error in a prediction group. It measures accuracy for continuous variables. In other words, MAE is more common than the verification sample of absolute values of the difference between forecast and corresponding observation. MAE is a linear score, which means that all individual variations are equally weighted.

$$\boldsymbol{MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}}$$

Where, p_i represents predicted values and a_i represents actual values.

Root Mean Square Error (RMSE)

RMSE is a square marking rule that measures the average size of an error. The equation for RMSE is given in both references. The difference between the observed, predicted, and correspondingly observed values in the formula are categorized and then averaged over the sample. Finally, the square root of the average is taken. Because errors are classified as normal, RMSE carries more weight than larger errors. This means that RMSE is especially useful when large errors are inappropriate.

MAE and RMSE can be used to detect the variability of errors in a prediction group. RMSE will always be greater than or equal to MAE; The greater the difference between them, the greater the variability of the individual errors in the sample. If RMSE = MAE, all errors are the same size.

$$\boldsymbol{RMSE} = \sqrt{\frac{(\boldsymbol{p_1} - \boldsymbol{a_1})^2 + \cdots + (\boldsymbol{p_n} - \boldsymbol{a_n})^2}{\boldsymbol{n}}}$$

Where, p_i represents predicted values and a_i represents actual values. Both MAE and RMSE can range from 0. They are negatively biased: lower values are better.

Relative Absolute Error (RAE)

It is a relative absolute error with the same generalization. In relative error measurements, errors are normalized by a simple prediction error that predicts average values.

$$\boldsymbol{RAE} = \frac{|\boldsymbol{p_1} - \boldsymbol{a_1}| + \cdots + |\boldsymbol{p_n} - \boldsymbol{a_n}|}{|\boldsymbol{a_1} - \overline{\boldsymbol{a}}| + \cdots |\boldsymbol{a_n} - \overline{\boldsymbol{a}}|}$$

Where,

$$\overline{\boldsymbol{a}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} \boldsymbol{a_i}$$

Root-Relative Square Error (RRSE)

The relative square error is normalized by taking the total square error and dividing it by the total square error of the default predictor.

$$\boldsymbol{RAE} = \frac{|\boldsymbol{p_1} - \boldsymbol{a_1}| + \cdots + |\boldsymbol{p_n} - \boldsymbol{a_n}|}{|\boldsymbol{a_1} - \overline{\boldsymbol{a}}| + \cdots |\boldsymbol{a_n} - \overline{\boldsymbol{a}}|}$$

Where,

$$\overline{\boldsymbol{a}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} \boldsymbol{a_i}$$

4. Results

In compiling the data, the researchers considered daily data on beans, eggplant, carrots and pumpkins over the past four years ranging from 2018 to 2021. The statistical values for rainfall, temperature, fuel prices, and crop production with bean cultivation are given in Table 1.

Table 1: Statistical Values of Beans

Attribute	Min	Max	Mean	Std. Dev.	Distinct	Unique
Rainfall (mm)	0.00	53.40	4.688	8.330	363	240 (27%)
Temperature (°C)	20.40	29.65	26.169	1.819	406	279 (32%)
Fuel Price (LKR)	95.00	123.00	102.735	6.827	9	0 (0%)
Production (t)	23.79	560.54	298.457	146.67	138	108(12%)

Here, distinct is the number of different values that contain data for an attribute, and unique is the percentage of data that has a value for this attribute that does not exist in any other instance. According to Table 1, the average rainfall is 4.688 mm, while the average temperature and average fuel price are 26.169 Celsius and LKR 102.735, respectively. Meanwhile, the average bean production in the last four years was 298.457 Metric Tons with high variation. Moreover, there appears to be high variability in rainfall and fuel prices, while the country's temperature appears to be fairly stable throughout the four years considered.

In the case of beans, the distribution of data on rainfall, temperature, fuel prices and production is shown in Figure 2. In Figure 2, the x-axis of each graph represents the range of attribute values, and the y-axis represents the frequency of occurrence. This figure further explains the meaning of the values given in Table 1. The performance of the considered machine learning algorithms for beans, eggplant, carrot and pumpkin are presented in Table 2, Table 3, Table 4 and Table 5 respectively.

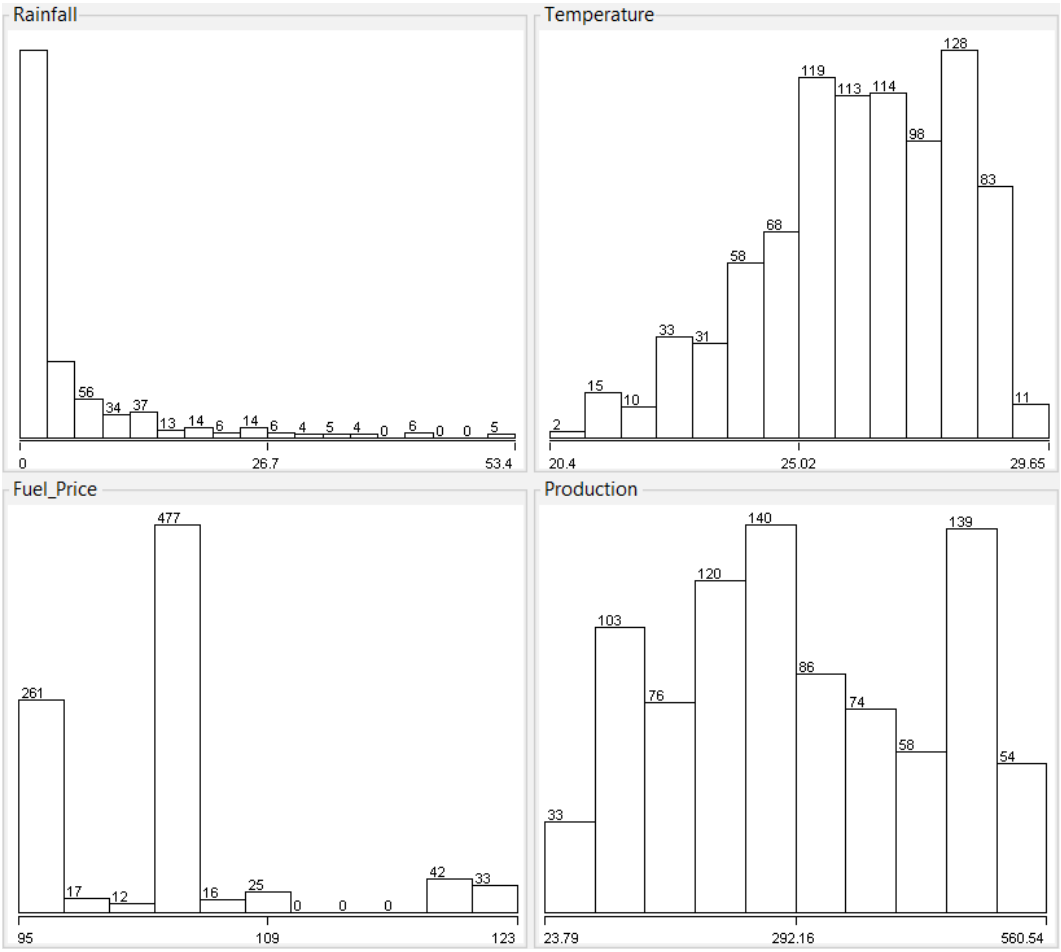

Figure 2: Statistical Values of Beans

Table 2: Performance of Algorithms for Beans

Evaluation Metrics	Linear Regression	SMO Regression	Multilayer Perceptron	Random Forest	M5P
MAE	46.7569	46.0079	50.0942	27.6227	29.9279
RMSE	58.4234	59.7346	62.2429	36.6394	37.8036
RAE	99.2503%	97.6603%	106.3342%	58.6344%	63.5275%
RRSE	99.3738%	101.6039%	105.8704%	62.3208%	64.3009%

Table 3: Performance of Algorithms for Eggplant

Evaluation Metrics	Linear Regression	SMO Regression	Multilayer Perceptron	Random Forest	M5P
MAE	24.7692	24.8500	27.2838	19.1100	21.7530
RMSE	34.6392	35.0344	37.4102	29.0625	31.6049
RAE	98.5547%	98.8763%	108.5602%	76.0371%	86.5536%
RRSE	99.1943%	100.3259%	107.1293%	83.2246%	90.5051%

Table 4: Performance of Algorithms for Carrots

Evaluation Metrics	Linear Regression	SMO Regression	Multilayer Perceptron	Random Forest	M5P
MAE	41.2798	37.9900	47.4884	20.0318	22.4018
RMSE	55.3191	58.3035	62.4240	29.1959	32.5995
RAE	97.1065%	89.3675%	111.7115%	47.1228%	52.6979%
RRSE	95.5017%	100.654%	107.7676%	50.4032%	56.2791%

Table 5: Performance of Algorithms for Pumpkins

Evaluation Metrics	Linear Regression	SMO Regression	Multilayer Perceptron	Random Forest	M5P
MAE	22.0790	21.8453	25.6043	10.5792	12.5173
RMSE	28.4544	28.7126	31.6965	15.4747	17.4574
RAE	93.1705%	92.1843%	108.0469%	44.6429%	52.8215%
RRSE	95.4439%	96.3102%	106.319%	51.9065%	58.5568%

According to Table 2, Table 3, Table 4 and Table 5, it can be observed that Random Forest presented better accuracy in predicting the price of the considered vegetables, followed by M5P and SMO Regression. In addition, the results revealed that all machine learning algorithms considered could more accurately predict the price of pumpkins. Figure 3 illustrates this observation further.

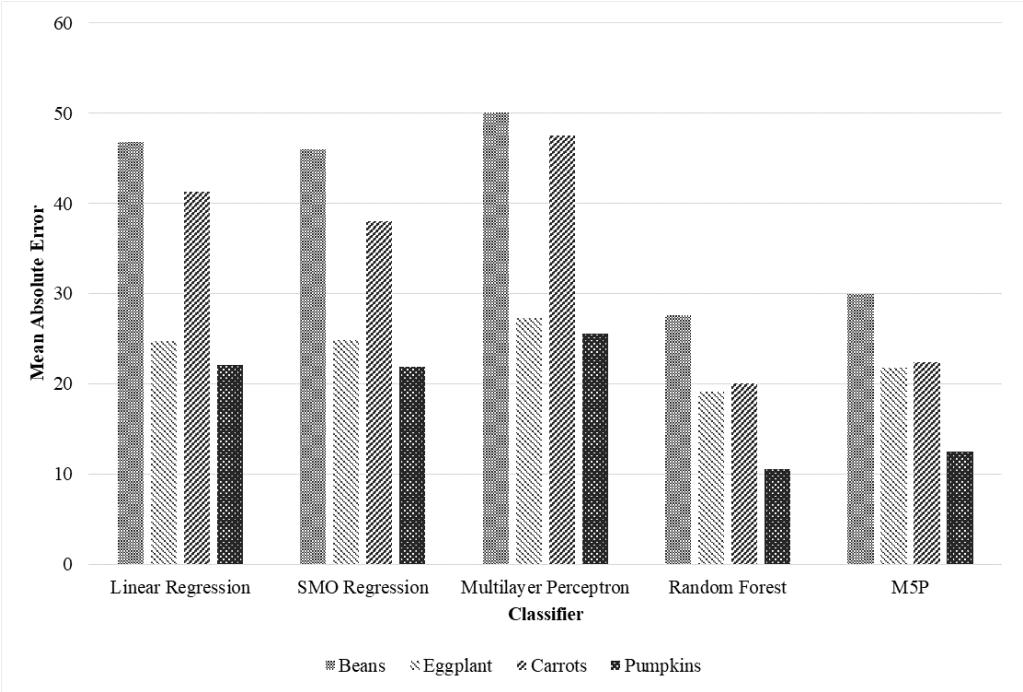

Figure 3: Comparison of Mean Absolute Errors

According to Figure 3, it can be observed that Random Forest outperforms other considered machine learning algorithms in terms of MAE. Moreover, among the crops considered, pumpkins recorded the lowest MAE for all the classifiers considered. Therefore, the ability to predict prices using a model constructed from the Random Forest was assessed with ten known records of pumpkins and the result is shown in Table 6.

Table 6: Pumpkin Price Prediction Using the Model Created from Random Forest

Rainfall	Temperature	Fuel Price	Production	Actual Price	Predicted Price	Error	Error %
3.53	28.83	104	230.53	20	20.54	0.54	2.70
1.29	28.67	104	443.32	195	140.07	(54.93)	28.17
18.50	22.70	106	86.82	19	18.53	(0.47)	2.49
7.35	24.90	123	230.53	23	21.60	(1.40)	6.10
0.50	26.55	123	348.20	18	16.69	(1.31)	7.30
2.95	24.05	118	869.22	19	24.26	5.26	27.69
0.35	24.65	118	9.54	38	35.23	(2.77)	7.29
2.85	23.25	109	117.66	28	33.03	5.03	17.96
0.00	24.95	95	153.03	43	46.81	3.81	8.87
0.00	28.80	95	271.60	53	55.66	2.66	5.01
Mean Error Percentage							11.36

Table 6 further strengthens the observations with experiments, illustrating that the developed model using Random Forest is capable of predicting pumpkin prices with a high accuracy of over 85 percent.

5. Discussion

This section discusses the findings of the present study in comparison with previous research findings in the existing literature. The literature survey revealed that no published work has been done on crop price forecasting in the context of the Sri Lankan vegetable market. Therefore, the findings of the present study have to be discussed by comparing the studies conducted with foreign countries.

Varun et al. (2019), revealed that linear regression is a better approach to vegetable price prediction. Although linear regression was not the best-performing algorithm in this study, the present study observed that linear regression has a good ability to predict vegetable prices in Sri Lanka. In the study by Nasira and Hemageetha (2012), researchers found that neural networks can predict tomato prices in India with high accuracy. Contrary to Nasira and Hemageetha's findings, in the present study, Multilayer Perceptron has shown a high error rate in predicting vegetable prices in Sri Lanka. However, only tomato price was considered as input to the algorithm in the study of Nasira and Hemageetha while in the present study, several factors like rainfall, temperature, fuel price and crop production along with the crop price were considered as input to the algorithm. This may account for the conflicting findings in the two studies. In the study of Shakoor et al. (2017), the researchers observed that Decision Tree based algorithms performed better than other machine learning algorithms considered in their study for crop price forecasting problems. In harmony with the findings of Shakoor et al. (2017), the present study also revealed that Random Forest which is a tree-based algorithm best performed in crop price prediction in the context of the Sri Lankan vegetable market. Paul et al. (2022) showed that Generalized Neural Network outperformed Support Vector Regression, Random

Forest and Gradient Boost in forecasting brinjal price in India. Furthermore, the researchers noted that the Random Forest performed similarly to the Generalized Neural Network in four markets. Although the neural network-based algorithm did not perform better in the present study, the tree-based algorithms employed in this study showed the best ability to predict crop prices. Therefore, it can be said that Paul's findings partially support the findings of the present study.

6. Conclusions

Based on the literature survey and the findings of the present study, several implications can be presented. Mainly during the last decade, the prices of vegetables fluctuated sporadically and from place to place but there is no fixed vegetable price for the entire country. The present study as well as the literature has shown that there is a non-linear relationship between crop prices and influencing factors. Hence, predicting crop prices is a difficult task as price movements move randomly and change over time. The specialty of this model is that in addition to the price, if necessary, other factors that affect the price can also be predicted. However, in this study, only price was predicted. Moreover, it can be mentioned that tree-based algorithms have a high ability to predict crop prices in the Sri Lankan vegetable market. This model will be useful for individual farmers in planning their cultivation and harvesting activities while the agricultural sector can apply this model for macro-level planning of crop production. Furthermore, the outcomes of this study will open avenues for the government to direct the country toward development, for traders to make trade decisions, for the entire population of the country to make daily consumption decisions, and for future researchers to plan future research.

This field of research spans a wide range. Future researchers can focus on developing a better model for crop price forecasting in Sri Lanka by parameter-tuning the best-performing algorithms identified in this study. Moreover, going beyond the current study, future researchers could experiment with weather data specific to specific crop-growing regions. It will affect the prediction accuracy of the developed models. Different machine learning algorithms have a high potential to perform better based on different markets. Therefore, future studies can be narrowed down to the selected vegetable market in Sri Lanka. Moreover, comparing the performance of different algorithms among different vegetable markets is also a direction for future research.

Acknowledgements:

This research was supported by the RJT/R&PC/2021/R/FMS/03 grant received from the Rajarata University of Sri Lanka, Mihintale. In addition, the Department of Agriculture, Department of Meteorology, and Economic Centers in Sri Lanka assisted in data acquisition.

References

Abewoy, D. (2018). Review on impacts of climate change on vegetable production and its management practices. *Advances in Crop Science and Technology*, *6*(1), 1-7.

Ashoori, M., Alizade, S., Hosseiny Eivary, H. S., Rastad, S., & Hossieny Eivary, S. S. (2015). A model to predict the sequential behaviour of healthy blood donors using data mining. *Journal of Research and Health*, 5(2), 141-148.

Biau, G., & Scornet, E. (2016). A random forest-guided tour. *Test*, 25, 197-227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., & Semret, N. (2017). Crop yield predictions-high resolution statistical model for intra-season forecasts applied to corn in the US. In *2017 Fall Meeting*. Gro Intelligence Inc.

Central Bank of Sri Lanka. (2020). *Daily Price Report, Central Bank of Sri Lanka*. https://www.cbsl.gov.lk/en/statistics/economic-indicators/price-report.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003.

Deepalakshmi, R., Devi, S. P., Revathy, J. S., & Shalini, T. G. (2019). Scheming a new algorithm for dynamic price prediction of vegetable commodities using statistical price prediction technique. *International Journal of Computational Complexity and Intelligent Algorithms*, *1*(2), 117-128.

Dilmah Ceylon Tea Company PLC. (2020). *Traditional Agriculture in Sri Lanka, Agriculture Practices of Sri Lanka*. https://www.dilmahconservation.org/arboretum/traditional-agriculture.html

Socio Economics & Planning Centre. (2018). *Crop Forecast*. Department of Peradeniya, 1–10.

Gunawardana, A. (2018). Agriculture sector performance in the Sri Lankan Economy: A systematic review and a Meta data analysis from year 2012 to 2016. *Experiment Findings*, 17.

Guo, Y., Tang, D., Tang, W., Yang, S., Tang, Q., Feng, Y., & Zhang, F. (2022). Agricultural Price Prediction Based on Combined Forecasting Model under Spatial-Temporal Influencing Factors. *Sustainability*, *14*(17), 10483.

Henegedara, G. M., & Abeykoon, A. M. N. J. (2016). Price Volatility of Vegetable Farming in Sri Lanka: With special reference in up country Vegetable Farming in Nuwara Eliya District in Sri Lanka. Sri Lanka Forum of University Economists (SLFUE), Department of Economics, Faculty of Social Sciences, University of Kelaniya.

Henegedara. (2016). Price instability and change of terms of trade in small farming sector in Sri Lanka (with special reference to cultivation of paddy and vegetables). *International Journal of Development Research*, *6*(7), 757–876.

Jothi, V. L., Lavanya, C., Sri, A., Kalaivani, M., & Kiruthikadevi, T. Price and Demand Forecasting for Agricultural Commodity using Data mining Techniques.

Kaur, M., Gulati, H., & Kundra, H. (2014). Data mining in Agriculture on crop price prediction: Techniques and Applications. *International Journal of Computer Applications*, *99*(12), 1-3.

Kumara, S. (2017). *Sri Lankan Agriculture: Goals, Challenges & E-solutions*. 1–31.

Kumarage, A. S. (2000). Traffic and Transportation Plan for the Shifting of the Vegetable Wholesale Trading Activities from Manning Market to Orugodawatte, *Colombo Final Report*. November.

Lewis-Beck, C., & Lewis-Beck, M. (2015). *Applied regression: An introduction, 22*. Sage publications.

Lu, Y. E., Yuping, L. I., Weihong, L., Qidao, S. O. N. G., Yanqun, L. I. U., & Xiaoli, Q. I. N. (2015). Vegetable price prediction based on pso-bp neural network. *8th international conference on intelligent computation technology and automation (ICICTA)* 1093-1096, IEEE.

Mathanraj, S., & Kaleel, M. I. M. (2016). The influence of rainfall variability on paddy production: a case study in Batticalloa district. *World Scientific News*, *52*, 265-275.

Moraru, A., Pesko, M., Porcius, M., Fortuna, C., & Mladenic, D. (2010). Using machine learning on sensor data. *Journal of computing and information technology*, *18*(4), 341-347.

Nasira, G. M., & Hemageetha, N. (2012, March). Vegetable price prediction using data mining classification technique. In *International conference on pattern recognition, informatics and medical engineering (PRIME-2012),* 99-102. IEEE.

Paul, R. K., Yeasin, M., Kumar, P., Kumar, P., Balasubramanian, M., Roy, H. S., & Gupta, A. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *Plos one*, *17*(7), e0270553.

Rachana, P. S., Rashmi, G., Shravani, D., Shruthi, N., & Kousar, R. S. (2019). Crop price forecasting system using supervised machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET)*, *6*, 4805-4807.

Rupasena, L. P., & Hector Kobbekaduwa. (1999). *Production and marketing of vegetables*. Hector Kobbekaduwa Agrarian Research and Training Institute.

Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). Agricultural production output prediction using supervised machine learning techniques. *1st international conference on next generation computing applications (NextComp)*, 182-187, IEEE.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE transactions on neural networks*, *11*(5), 1188-1193.

Subhasree, M., & Priya, C. A. (2016). Forecasting vegetable price using time series data. *International Journal of Advanced Research*, *3*, 535-641.

Varun, R., Neema, N., Sahana, H. P., Sathvik, A., & Muddasir, M. (2019). Agriculture commodity price forecasting using ML techniques. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2S), 729.

Williams, C. (2007). Research methods. *Journal of Business & Economics Research (JBER)*, *5*(3).

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 1*, 29-39.

Ye, L., Li, Y., Liang, W., Song, Q., Liu, Y., & Qin, X. (2016). *Vegetable Price Prediction Based on PSO-BP Neural Network*. Proceedings - 8th International Conference on Intelligent Computation Technology and Automation, ICICTA 2015, 1093–1096. https://doi.org/10.1109/ICICTA.2015.274