Forecasting Sales of Iraqi Dates Using Artificial Intelligence

Hussam Mezher Merdas[1], Ayad Hameed Mousa[2]
[1,2]Department of Computer Science, University of Kerbala, Karbala, Iraq.

Article Info

Article history:

Received 13 October, 2023
Revised 14 November, 2023
Accepted 17 November, 2023

Keywords:

Artificial Intelligence
Gradient Boosting
Random Forest
Bagging Regressor
Decision Tree
Bayesian Ridge

ABSTRACT

Iraq is considered one of the first countries in the world to export Dates of all kinds. This sector at present needs support and serious work to improve sales to provide the country's economy with more revenues. This study proposes building an integrated artificial intelligence model that predicts the quantities of Dates that Iraq will produce in the coming years based on previous data and based on two main points: The first point is to make a comparison between three different food datasets with a different correlation between their features, as the first dataset is of high correlation, the second is of medium correlation, and the third is of weak correlation. The second point is to apply twelve Machine Learning algorithms and evaluate their results to obtain the best three algorithms. The model was applied to predict the quantities of Dates that Iraq will produce for the next five years. The proposed three algorithms were used and gave the following results: (Gradient Boosting: 99.51, Random Forest: 97.05, and Bagging Regressor: 98.54). This study constitutes a starting point for future studies in terms of the process of choosing the datasets, as well as the machine learning technique.

Corresponding Author:

Hussam Mezher Merdas
Department of Computer Science, University of Kerbala, Karbala, Iraq.
Email: hussam.m@s.uokerbala.edu.iq

1. INTRODUCTION

Dates are fast-perishable food items that require multiple and large storage facilities. To reduce losses, countries attempt to forecast the quantity of dates they will produce to provide sufficient storage space. Iraq relies on dates as one of its most important exports after oil. To support this important resource in the country, this model was proposed. Artificial intelligence techniques represented by machine learning algorithms are employed to make an appropriate prediction of the quantities of dates that Iraq will produce. The concern of this study is the prediction algorithms, as there are many of these algorithms and they vary in accuracy according to the type of this algorithm as well as the dataset used. Machine learning (ML) is an important and fundamental branch of AI, which is defined as the ability of a machine to accurately simulate human behavior [1]. AI systems try to perform the functions entrusted to humans by performing the tasks that humans perform and sometimes even tasks that they cannot easily perform.

ML is divided into three main sub-categories: (Supervised, unsupervised, and reinforcement). In the supervised Machine Learning technique, only labeled datasets are used. This allows the models to learn and operate more accurately over time. Supervised ML is divided into two main parts: Classification and Regression. Classification is a technique used when the output variable is categorical, non-continuous, and determinant [2]. For example, yes or no, true or false, sick or not sick, and so on. Regression is a technique that is used when the output variable is continuous and real and is not confined to one or two categories. Where the output variable depends on one or more variables, and these variables depend on each other and affect each other. In unsupervised machine learning, this technique searches for previously unlabeled data and groups them into categories according to the degree of similarity. Reinforcement ML trains algorithms in a way that learns from error and relies on a system of rewards [3]. This technology can train models for games or autonomous driving by learning from errors.

This study relies on supervised ML. Specifically, the proposed model is based on the regression technique. Where the AI algorithms used are employed on real and continuous data to give forecast results for food sales, then the accuracy obtained from these algorithms is evaluated by comparing the results obtained from them with the actual numbers in the datasets. By reviewing and analyzing the relevant studies, found that they differed in terms of choosing the data on which they depend, depending on the companies that provide this data. This data may have a good correlation between its features and may have a weak correlation, which is reflected in the results obtained from the algorithms. On the same aspect, food datasets often need a lot of pre-processing before they can be used for prediction.

After the stage of preparing the datasets, the stage of training the algorithms begins and then testing them to see their accuracy in prediction to use them later, whether in future studies or for the benefit of the food-selling companies' sector. The researchers used several datasets and several artificial intelligence algorithms in their research. Some researchers use one dataset, others use two databases, and so on. Through this study, the researcher found that the way the features correlate with the data was reflected in the accuracy obtained from the algorithms. Also, these algorithms gave different results in different studies. Most researchers used datasets in small numbers (one or two and sometimes three) and also used algorithms in small numbers, often not exceeding four.

As mentioned above, there are two main reasons for the large discrepancy in the results obtained from the proposed models. Therefore, this required the preparation of a comprehensive study and an integrated model that establishes a broad and effective comparison between food datasets on the one hand, and the machine learning algorithms applied to those datasets on the other hand. There are classic algorithms, there are modern algorithms, and others that rely on deep learning. Several practical measures were used to measure the accuracy of the proposed model and the results obtained from it. This study used an actual local dataset of the Dates sales in Iraq for the period (2002-2020) and then predicted the quantities of dates that Iraq will produce for the next five years. This study takes into account the juice of previous studies by designing a model that uses several different datasets and applying the most important and famous AI algorithms. Therefore, this study can be considered a brief and useful summary that is presented to companies without the need to waste time knowing the appropriate food sales forecast models.

2. RELATED WORKS

Many researchers have worked hard over the years to study and develop food sales forecasting models. All of these researches and studies were to serve science and humanity, facilitate his work, and reduce losses resulting from spending and wasting products. The researchers turned to AI algorithms, including DL algorithms, to achieve the best possible results. They used classification and Regression and developed existing algorithms using modern methods. Also, companies have become more flexible by providing the necessary datasets for studies to serve researchers and companies themselves. The most important related works can be summarized in the following paragraphs:

To keep agricultural products fresh and ensure that the products are not spoiled, X. Wang, et al. suggested this model. The suggested model is based on integrating the most important weather factors affecting the sales of perishable agricultural products. Whereby, using the previous sales data for these products, three algorithms Random Forest, Ridge Regression, and Support Vector Machine (SVM), perform the required prediction. The results showed (according to the opinion of the researchers themselves) that this proposed model may significantly improve the prediction results depending on the weather conditions affecting fresh produce. Where the results achieved from this study, using the RMSE "Root Mean Square Error", are 68.90%, 23.66%, and 59.52%. The algorithms gave prediction results using MAE "Mean Absolute Error" as follows: 66.2%, 34.99%, and 61.13%, respectively [4].

V. Prabhakar, et al. proposed in this study a model used to predict the future sales of a competing and leading grocery store. The store provided researchers with a dataset of food sales because they think it deserves study due to its importance and the benefit of the store itself. The store is guaranteed an opportunity to study and manage products that are subject to waste and damage. This study relies on ML algorithms to make an efficient prediction of food sales for this grocery store. The researchers used the R language to generate the statistics needed to predict the use of specific ML algorithms. They used a library called Especially a Caret to perform the prediction and analysis process [5].

Forecasts have always been one of the important things in supply chains, now they have become a necessity, as a result of increasing consumer demand, limited capital, and the importance of dealing with limited time. Any store selling now wants to anticipate the products that consumers need to avoid seasonal shortages, so companies are working to develop forecasting daily. With the aforementioned, Y. F. Akande, et al. suggested using the "*extreme gradient boosting*" (XGBoost) algorithm to predict future sales. Sales data provided by 45 Walmart stores were used to develop the model. The results showed (according to the researcher) that the XGBoost ML algorithm proved effective in terms of predicting future sales [6].

Predicting future sales of groceries increases merchant revenues by avoiding spoilage of surplus products as well as benefits the customer. Likewise, it is important to consider the issue of features affecting sales. For the above, Y. Liu suggested designing a model to make the necessary forecast for sales of large stores. This model is based on the use of two-time series cores and a light gradient boosting machine learning algorithm (LGBM). Where data is processed, analyzed, trained, tested, and then used. The resulting accuracy of this model was measured using the mean squared error, which gave an accuracy of 0.35069 [7].

To facilitate the work of the staff in restaurants to meet the needs of customers, these restaurants need a good sales forecasting program. Using real data from a medium-sized restaurant, A. Schmidt, et al. proposed using several ML algorithms to make the appropriate prediction. Recurrent neural network (RNN) technology was used, as well as three different datasets were used to train and test the models. The results obtained from the samples were compared for one day as well as for one week. The results for the one-day linear models using "*symmetric mean absolute percent error*" sMAPE were only 19.6%. When using RNN with "*long short-term memory*" LSTM and "*Temporal Fusion Transformer*" TFT algorithms, the results are good with errors of less than 20%. When performing the one-week forecasting process with the models without using RNN the result was bad, the results were approximately 20% error. With the use of RNN, the results using the sMAPE scale gave approximately 19.5% of the best result [8].

T. S. Yange, et al. suggested building a model using the (SVM) algorithm to forecast the sales of agricultural products so that managers can determine the quantities needed for the sale of products that may be subject to spoilage. This system used both SVMs and "Fuzzy Theory". The Radial Basis Function (RBF) neural network was taken as a standard for evaluating the result obtained from this model. To train the system, the data provided by "Makurdi University Farm" was used. The system was trained using one part of the data and was tested using another part. The SVM algorithm gave a result of 96.75% and the prediction result of the RBF was 90.55%. From this, the SVM algorithm gave the best results [9].

T. Tanizaki, et al. suggested using POS data as internal data and adopting weather data, events data, and other external data to forecast sales. Where the algorithms of "Bayesian linear regression, decision tree regression, decision forest regression, and stepwise method" were used to perform the prediction process. The results showed closeness to the accuracy obtained from "Bayesian, Decision, and Stepwise", and the results obtained from Boosted were somewhat low. Overall, the expected rate of prediction exceeded 85% [10].

T. Weng, et al. designed a developed model based on "LGBM and LSTM (Long-Short Term Memory)" for sales forecasting. Three databases were used to verify the accuracy and efficiency of the model. According to the researchers in this study, the model works efficiently to forecast supply chain sales. According to the researchers in this study, their model provides a high possibility of predicting long-term sales of the supply chain, which is beneficial for companies. The model is based not only on the efficiency of LSTM but also on the possibility of LGBM, which is efficient in an industrial production environment [11].

I. Vallés-Pérez, et al. developed a prediction system based on three day/store/item alternatives based on deep learning algorithms and their application on "Corporación Favorita" data. According to the results of this study, it provides good prediction accuracy by adopting a simple sequence of geometry sequences to pre-process the data. Also, they used a "training trick" this is to create a model less dependent on time and thus ensure good generalization over time. The model gave results estimated at 0.54 using RMSLE "Root Mean Squared Logarithmic Error" [12].

To predict the future sales of each product accurately, X. Bi, et al. proposed a new model. Relying on "Tensor Methodologies for Context-Aware Recommendation Systems" this model was proposed under the name "Advanced Temporary Latent Factor Approach to Sales Forecasting, or ATLAS." According to the researchers, this model achieves efficient prediction results for each product separately, by building a model of One tensioner worker for stores and products. The proposed model mixes two components of a "tensor framework" (depending on product and stores information), and the second component is the integration of demand mechanics, tensor extrapolation based on the latest (seasonal) statistical data for certain periods as well as using "seasonal autoregressive integrated moving-average models" and RNN models. The main focus of the ATLAS model is the use of eight datasets. The model analyzes data from more than 1,500 grocery stores [13].

Prediction models relied on ML algorithms and deep learning algorithms. However, to achieve higher accuracy and efficiency, other matters must be taken into account, such as epidemics, natural disasters, etc., which is reflected in the purchase of products For the above, N. Kumar, et al. suggested designing a new model that adopts a “multi-modal network” to forecast product sales based on the principle of combining current events with previous data. Moreover, the form also obscures the data collection published by Google Trends. The accuracy obtained from the preliminary results using "symmetric mean absolute percentage error SMAPE" showed approximately 7.37% compared to the results obtained from the current sales forecasting techniques [14].

3. THE MATERIALS AND METHOD

This study aims to design a model to predict the quantities of Iraqi Dates using AI represented by machine learning techniques of the most important types (Figure 1). Several successive steps make up this model: The first step is to feed the model with three different datasets with varying degrees of correlation between their features, which will be discussed in detail in this chapter. The second step is to present the necessary charts and figures that give an initial view of the data used simply and clearly. This step is followed by the step of pre-processing to obtain suitable datasets ready for processing. The fourth step is to measure the degree of correlation between the features of the datasets by using the Heatmap tool from the Seaborn library available in the Python language. In the fifth stage, the three datasets are entered into twelve algorithms used in this model. The sixth stage of this system is the use of several important metrics to measure the accuracy and efficiency of the algorithms used for each dataset separately and represent them graphically using simplified forms. The seventh and final step is to use a realistic dataset for sales of Iraqi Dates for the period (2002-2020), where the correlation of the features of this dataset is measured and according to the degree of correlation, the three appropriate algorithms are used based on the proposed results obtained from the model to give a prediction of the quantities of dates that Iraq will produce in the five years following 2020, to sell them in the local and foreign markets.

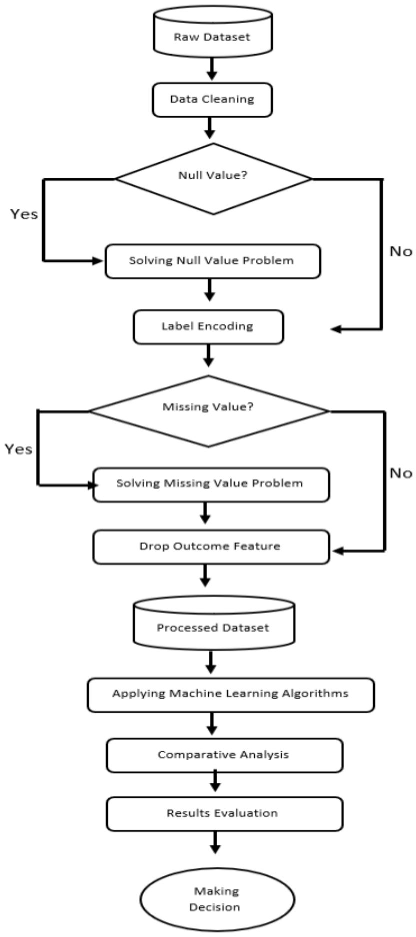
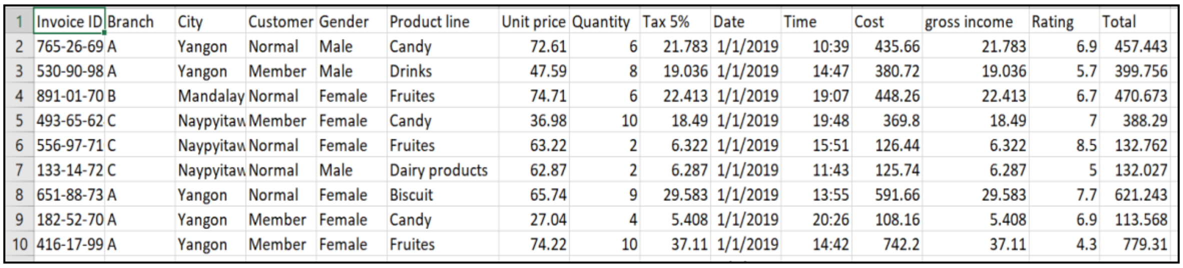


Figure 1. The Proposed Model.

3.1 THE SELECTED DATASETS

This study used four datasets, three of which are entered into the model to be used to clarify the impact of data features correlation on prediction results, as well as to know the best algorithms in terms of accuracy and efficiency in forecasting sales of food products that are applied to each dataset separately. As for the fourth dataset, the model is used to generate a future prediction of the quantities of dates that can be sold according to specific variables. The first dataset is sourced from "*publicly available Alibaba's Tianchi platform data*" and it consists of 1000 rows and 15 columns as shown in Table 1. This data set contains food sales of different categories on the Alibaba platform in several cities in the year 2019.

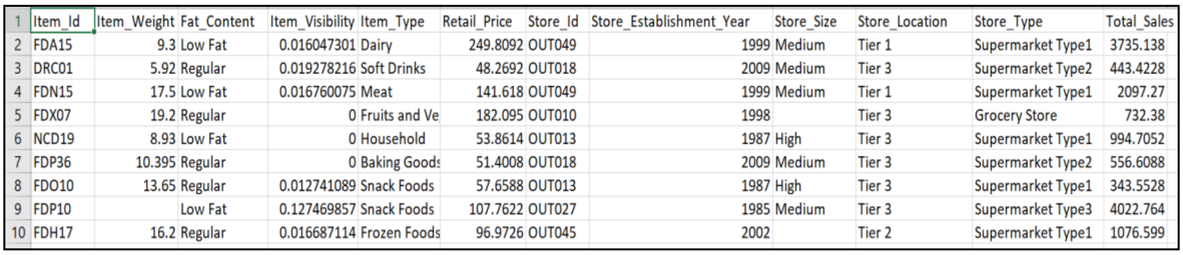
Table 1. A Part of the First Dataset



1	Invoice ID	Branch	City	Customer	Gender	Product line	Unit price	Quantity	Tax 5%	Date	Time	Cost	gross income	Rating	Total
2	765-26-69	A	Yangon	Normal	Male	Candy	72.61	6	21.783	1/1/2019	10:39	435.66	21.783	6.9	457.443
3	530-90-98	A	Yangon	Member	Male	Drinks	47.59	8	19.036	1/1/2019	14:47	380.72	19.036	5.7	399.756
4	891-01-70	B	Mandalay	Normal	Female	Fruites	74.71	6	22.413	1/1/2019	19:07	448.26	22.413	6.7	470.673
5	493-65-62	C	Naypyitaw	Member	Female	Candy	36.98	10	18.49	1/1/2019	19:48	369.8	18.49	7	388.29
6	556-97-71	C	Naypyitaw	Normal	Female	Fruites	63.22	2	6.322	1/1/2019	15:51	126.44	6.322	8.5	132.762
7	133-14-72	C	Naypyitaw	Normal	Male	Dairy products	62.87	2	6.287	1/1/2019	11:43	125.74	6.287	5	132.027
8	651-88-73	A	Yangon	Normal	Female	Biscuit	65.74	9	29.583	1/1/2019	13:55	591.66	29.583	7.7	621.243
9	182-52-70	A	Yangon	Member	Female	Candy	27.04	4	5.408	1/1/2019	20:26	108.16	5.408	6.9	113.568
10	416-17-99	A	Yangon	Member	Female	Fruites	74.22	10	37.11	1/1/2019	14:42	742.2	37.11	4.3	779.31

The second dataset was taken from "*Kaggle"*, and it is related to food sales for ten stores in different places, and it consists of 12 columns and 8523 rows, as shown in Table 2.

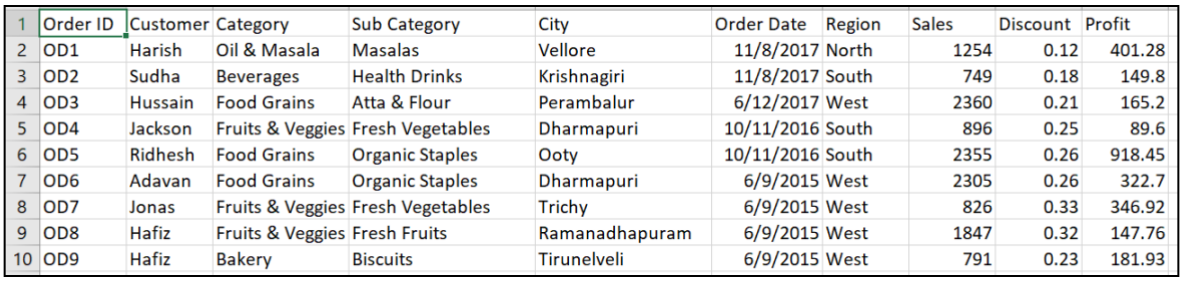
Table 2. A Part of the Second Dataset



1	Item_Id	Item_Weight	Fat_Content	Item_Visibility	Item_Type	Retail_Price	Store_Id	Store_Establishment_Year	Store_Size	Store_Location	Store_Type	Total_Sales
2	FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
3	DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
4	FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
5	FDX07	19.2	Regular	0	Fruits and Ve	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
7	FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
8	FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
9	FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.764
10	FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.599

The third dataset is related to food sales for several main categories and sub-categories in different cities. This dataset is sourced from *Kaggle* and consists of 10 attributes and 9995 objects as shown in Table 3.

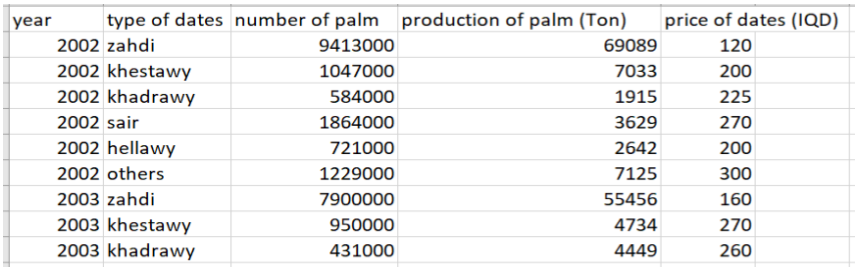
Table 3. A Part of the Third Dataset



1	Order ID	Customer	Category	Sub Category	City	Order Date	Region	Sales	Discount	Profit
2	OD1	Harish	Oil & Masala	Masalas	Vellore	11/8/2017	North	1254	0.12	401.28
3	OD2	Sudha	Beverages	Health Drinks	Krishnagiri	11/8/2017	South	749	0.18	149.8
4	OD3	Hussain	Food Grains	Atta & Flour	Perambalur	6/12/2017	West	2360	0.21	165.2
5	OD4	Jackson	Fruits & Veggies	Fresh Vegetables	Dharmapuri	10/11/2016	South	896	0.25	89.6
6	OD5	Ridhesh	Food Grains	Organic Staples	Ooty	10/11/2016	South	2355	0.26	918.45
7	OD6	Adavan	Food Grains	Organic Staples	Dharmapuri	6/9/2015	West	2305	0.26	322.7
8	OD7	Jonas	Fruits & Veggies	Fresh Vegetables	Trichy	6/9/2015	West	826	0.33	346.92
9	OD8	Hafiz	Fruits & Veggies	Fresh Fruits	Ramanadhapuram	6/9/2015	West	1847	0.32	147.76
10	OD9	Hafiz	Bakery	Biscuits	Tirunelveli	6/9/2015	West	791	0.23	181.93

The fourth and final dataset was used to experiment with the proposed model to generate a prediction of the quantities of food products that could be sold for the next five years under specific variables. This data set is for sales of Dates in the Republic of Iraq for the period from (2002-2020) and includes several features, including the quantity, prices, types of Dates, and the number of palm trees in this period. The source of this dataset is the Iraqi Ministry of Planning / Central Statistical Organization. Table 4 shows this dataset.

Table 4. A Part of the Fourth Dataset



year	type of dates	number of palm	production of palm (Ton)	price of dates (IQD)
2002	zahdi	9413000	69089	120
2002	khestawy	1047000	7033	200
2002	khadrawy	584000	1915	225
2002	sair	1864000	3629	270
2002	hellawy	721000	2642	200
2002	others	1229000	7125	300
2003	zahdi	7900000	55456	160
2003	khestawy	950000	4734	270
2003	khadrawy	431000	4449	260

3.2 PRE-PROCESSING

In this study, three different datasets were used to construct the model. Where a comparison was made between them in the degree of correlation of features. There is also a fourth dataset through which the actual use of the proposed model was made. For the above, the four datasets needed pre-processing before use. Where some of them were suffering from the problem of data loss, where the lost data was compensated by appropriate methods, such as taking the average for the rest of the values. Some data contain errors in the entry, as these errors have been dealt with in appropriate ways. In the end, good and organized datasets were obtained so that they could be dealt with. This study also sought not to cause large changes in the data and contents of these datasets to simulate the use of such datasets in other studies.

3.3 FEATURES CORRELATION

To achieve a comprehensive view of the datasets used, three different datasets were taken into account. One of the most important differences that distinguish one dataset from another is how the features are correlated. To find out how the features of the datasets are correlated, Correlation Matrix Heatmap was used. It is a tool available from the Seaborn library in the Python language. It is a two-dimensional matrix containing colored cells. The colors are graded according to the correlation strength. Whenever the color tends to become darker, this means a stronger correlation, and on the contrary, whenever it tends to become lighter, this means a weaker correlation. The cells of the matrix contain values between (-1 and 1), where whenever this value is directed to the positive, this indicates the strength of the correlation, and whenever it is directed to the negative, this indicates a weak correlation of the features. Using the Correlation Matrix Heatmap tool, the correlation between the features of the three datasets used in this model was measured. Figure 2 below shows the correlation using this tool. What is noticed is that most of the cells tend to be dark in color, and there are many cells close to (1), which indicates that the first dataset is well correlated.

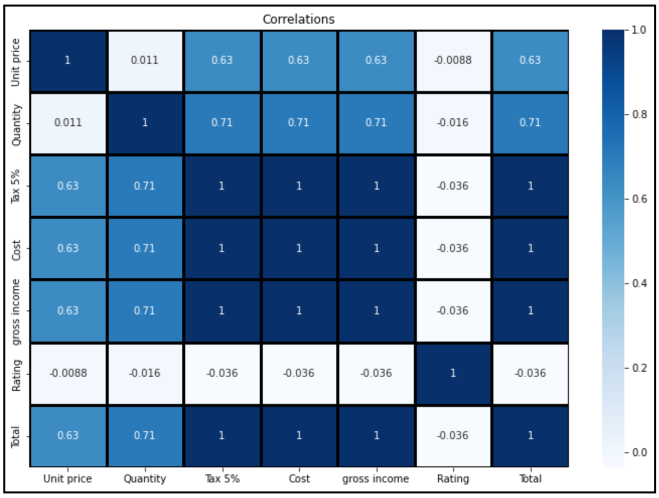


Figure 2. The Correlation between Features of the First Dataset.

In the same way, the correlation of features for the second dataset was measured. Figure 3 below shows that the matrix contains medium values. This indicates that the second dataset has a medium correlation between its features.

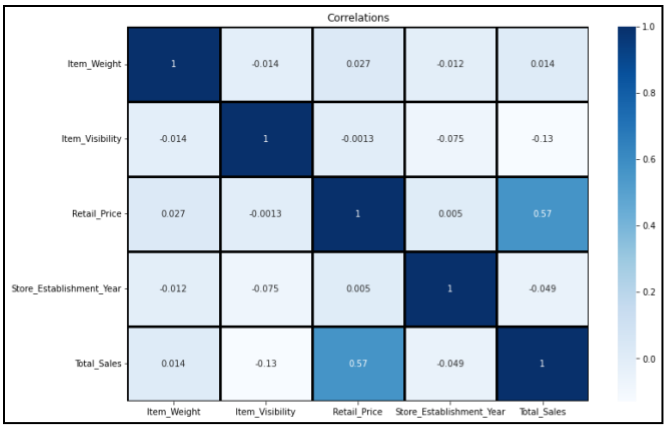


Figure 3. The Correlation between Features of the Second Dataset.

Finally, the correlation between the features of the third dataset was measured. Figure 4 below shows that. It is noted that most cell values tend to be negative, which indicates a weak correlation between the features of this dataset.

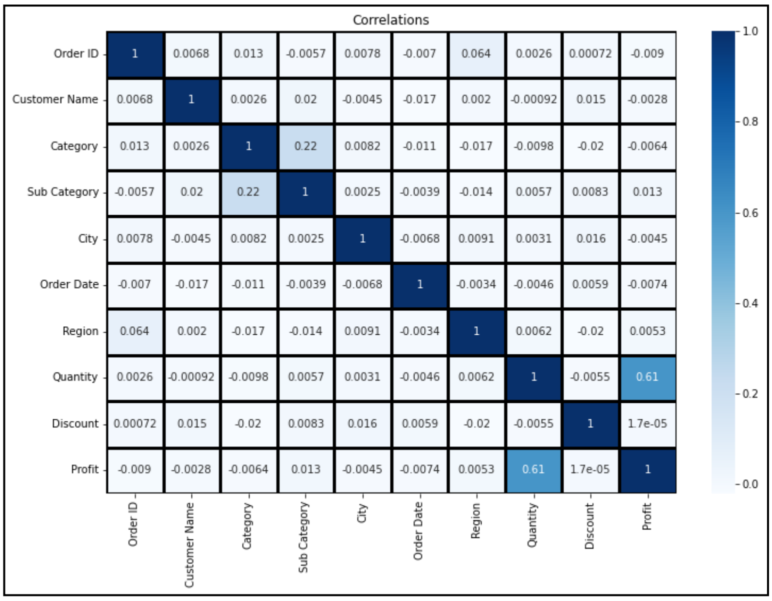


Figure 4. The Correlation between Features of the Third Dataset.

From what was mentioned above, the conclusion was reached that the first dataset is of strong correlation, the second is of medium correlation, and the third is of weak correlation. On this result, this study will be based. It will apply twelve algorithms to each dataset separately to find out the best three algorithms applied to each dataset.

3.4 ALGORITHMS USED

To make a comprehensive and efficient comparison, twelve algorithms are tested including ML algorithms, DL algorithms, as well as models that rely on ensemble techniques (techniques that are based on merging several models instead of using a single model to obtain greater prediction accuracy). These algorithms are the most important and widespread in the field of forecasting food sales. These algorithms are: (Multilayer Perceptron, Support Vector Regressor (SVR), Multiple Linear Regression, Decision Tree, Random Forest Regressor, K-Nearest Neighbors (KNN), Bagging Regressor, LASSO, Gaussian Processes Regressor, RANSAC, Gradient Boosting, Ridge Regressor, Elastic Net, Bayesian Ridge, and Kernel Ridge). Below will discuss the theoretical side of these algorithms in a way that benefits the reader and concerns this study.

3.4.1 MULTILAYER PERCEPTRON

Deep learning is based on the use of multi-layered Artificial Neural Networks to train them, also called Deep Neural Networks. In the fifties of the last century, the Rosenblatt perceptron model was developed, but the topic of neural networks did not receive this important interest until 1986, in this year Dr. Hinton and his colleagues developed the backpropagation technique for training a multi-layered neural network [15]. Now the topic of neural networks has become a popular field for large companies, why it gives good results in prediction.

When the neural network has multiple layers and these layers are fully connected then it is called Multilayer Perceptron (MLP). The simplest MLP networks consist of only three layers, the input layer, and the output layer, and between them, there is one hidden layer. But if there are many hidden layers, then in this case the neural network is called a deep artificial neural network (ANN). MLP is a good example of a feedforward ANN (Figure 5). The number of layers in each neural network must be set. The backpropagation balances the weight. The deeper the neural network, the more accurate it is. However, too many deep layers can lead to problems. The inputs, which are weights, are entered into the activation function. The activation function converts the weights into output values and converts them to the output of the neuron. The activation function is so named because it controls the threshold for activating the neurons and the strength of the output signal [16].

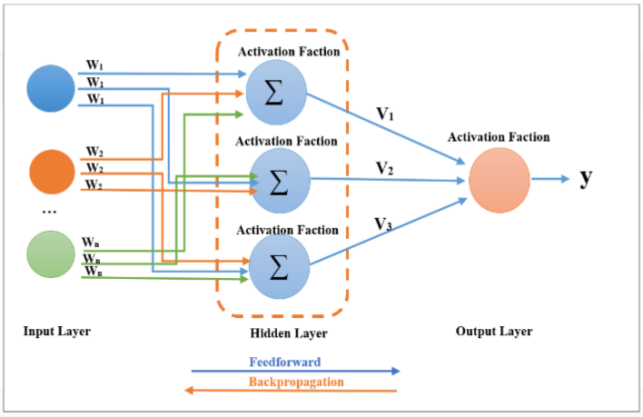


Figure 5. Multilayer Perceptron.

3.4.2 SUPPORT VECTOR REGRESSION

Support Vector Machine (SVM) is widely used as a classification algorithm. However, this algorithm has good use with regression. It relies on non-linearity in building the model and dealing with the data to produce the prediction [17]. Whereas, when SVM is used for regression purposes, it is called Support Vector Regression or SVR. SVR is a supervised algorithm used for prediction purposes. SVM is a classifier, which performs the classification process by predicting discrete labels while SVR acts as a regression technique used to predict continuous variables. The idea of SVM work is based on the principle of forming a hyperplane, the same principle is applied in the algorithm SVR, with the difference in the method of determining the best-fit line which represents the hyperplane that contains the maximum number of points. SVM works to reduce the error rate, but SVR works by fitting the error value within a specified threshold, which results in the best possible value within a certain margin [18].

Using a set threshold, SVR works on the best possible line fit. The threshold is the specified distance between the boundary line and the hyperplane. When the SVR expands so that the samples become large, this causes difficulty in the work of this algorithm. So SVR relies on a subset of the data for training to prevent complicating the algorithm, by having the cost function discard samples that are not useful for prediction and that do not improve accuracy.

3.4.3 MULTIPLE LINEAR REGRESSION

Based on one or more variables, multiple linear regression generates a prediction. It is a type of linear regression. This algorithm is based on the prediction of a variable called the dependent variable and builds its prediction on several other variables called the independent variables. Simple linear regression is also a type of linear regression that uses one independent variable to predict another variable [19]. This type of linear regression attempts to create a straight line to represent these two variables.

Multiple Linear regression depends on a linear relationship between two or more variables. When the dependent and independent variables do not follow a single line, then the relationship in this case is non-linear. Can be written multiple linear regression equation as follows [20]:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 \cdot \cdot \cdot + \beta kXk + \varepsilon \quad (1)$$

- Yi is the dependent value.
- β1 and β2 are parameters.
- Xi is the independent variable.
- ε is the random error.

3.4.4 DECISION TREE

From the shape and structure of trees in nature came the design of the decision tree, which is a well-known supervised learning algorithm that is used in classification and regression. This study focused on regression to form a prediction model. A decision tree consists of three types of nodes. The root node is the main node and the pillar from which the rest of the nodes branch [21]. Dataset features are internal nodes and branches that represent decision rules, and leaf nodes represent the result.

From the root node, the branching process begins, as it continues for several levels according to the data that is processed until it reaches the terminal node that contains the prediction, which represents the final result of the algorithm. Branching or division usually starts from the top, as it produces a new variable at each step that divides better. Each subtree is divided into two branches only. In regression, the tree gives leaf nodes with continuous values (usually real numbers), unlike in classification, where the values are discrete. The regression decision tree divides the data into subgroups and fits the model to each subgroup separately and progresses one level after another until it reaches the best possible prediction [22].

3.4.5 RANDOM FOREST REGRESSION

Random Forest is one of the supervised learning algorithms based mainly on ensemble learning using several decision trees together [23]. This algorithm adopts the bagging technique, so the processing is done in parallel so that each decision tree works separately. This algorithm can be used in classification and regression, in this study the focus will be on regression. The name of the random forest came from the idea of random bagging of data, and based on several decision trees, a prediction is generated like (forest). Overall, this algorithm is an important and powerful algorithm that minimizes the defects of the decision tree. Moreover, random forest is very popular in the field of sales forecasting for its power and accuracy of results.

A random forest is an algorithm that consists of a set of decision trees. These decision trees are generated "randomly" to assemble into a random forest. Decision trees are created by selecting samples from the rows, and at each point, a division is made based on the features of the datasets [24]. Each decision tree produces its own sub-predictor. Then the final result is produced from the average of the results of the sub-trees. The average yields a random forest with efficient results compared to a single decision tree, which results in high accuracy away from overfitting.

3.4.6 K-NEAREST NEIGHBORS (KNN)

The K-Nearest Neighbors (KNN) algorithm is a non-parametric supervised artificial intelligence algorithm that is easy to implement and simple [25]. Evelyn Fix and Joseph Hodges developed KNN in 1951. KNN is used in classification and regression. For classification, this algorithm assigns a label to each class using a majority vote, the point's affiliation to any class is determined by adopting the majority vote of its neighbors. The KNN algorithm works in regression as it does in classification, but the difference is that this algorithm uses to predict discrete values in the classification, while it works to predict continuous values in regression, also the output of the prediction represents the value of the object's property. This value is obtained from the average of the k-nearest neighbors. However, before classification is made, distance or similarity is measured. The best measure of distance is the Euclidean distance [26]. Also, the value of k must be determined based on the entered data. As the value of k affects the prediction results positively or negatively.

3.4.7 BAGGING REGRESSION

A Bagging Regression is a meta-estimator work based on an ensemble technique suggested by Leo Breiman in 1996 [27]. This algorithm works by taking random samples from the original dataset and then fitting them into regressors after which it aggregates its predictions (either by voting or by averaging) to get a final prediction. This method mainly aims to get rid of overfitting problems in regression and reduce variance within a noisy dataset. A Bagging regressor improves the accuracy and performance of machine learning algorithms. The work of this algorithm can be summarized in three main steps [28]:

Bootstrapping: The Bagging relies on bootstrapping sampling technique to generate various samples. This resampling process generates new training samples by randomly selecting and replacing data points. Thus, each time a data point is selected from the training data set.

Parallel training: The resulting samples from the first step are trained independently using either basic or weak learners.

Aggregation: finally, to form the final estimate the average of all the predictions that resulted from the individual predictors is taken; This process is known as soft voting.

3.4.8 LASSO REGRESSION

The word “LASSO” stands for “Least Absolute Shrinkage and Selection Operator”. This model is a type of linear regression and one of the regularization techniques. It is used with regression to obtain a more accurate prediction. LASSO is based on the principle of shrinkage. Shrinking means shrinking data values towards a central point as an average [29]. L1 regularization is the technique used with this algorithm. This technique is used when there are many features because it automatically performs the feature selection process.

Regularization is used to avoid the problem of data overfitting, especially in the case where the trained data and the test data differ a lot. Regularization works by adding a “penalty” term to the best fit produced from the trained data, to obtain less variance with the tested data and also reduce the influence of the predictor variables on the output variable by compressing their coefficients [30]. LASSO fits a model containing all possible predictors, where it selects a variable based on the use of a technique that regularizes coefficient estimates (it is based on the principle of shrinking the coefficients toward zero).

3.4.9 GAUSSIAN PROCESSES REGRESSION

Gaussian Processes Regression (GPR) is a supervised ML algorithm that deals with probabilistic classification and regression problems. This method works well with small datasets as it can provide measures of uncertainty in predictions. GPR differs from the rest of the supervised ML algorithms in that the latter learns the exact values of all parameters in a function, unlike GPR which is based on the principle of probabilities distribution over all possible values [31]. GPR is a non-parametric Bayesian algorithm that works by distributing probabilities over all admissible functions that fit the data and not by calculating the probability of a parameter’s distribution on a custom function as shown in Figure 6 (a and b) [32]. Where the posterior is calculated using the training data, and the predictive posterior distribution is calculated using points of interest.

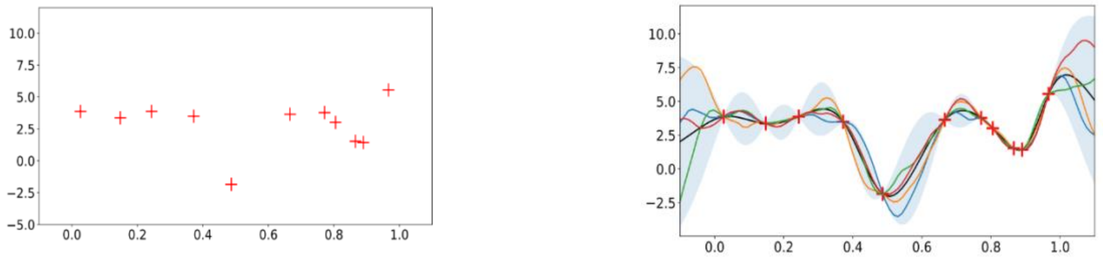


Figure 6. a. The observed data points. b. Five possible functions of GPR.

3.4.10 RANSAC

RANSAC is an iterative ML algorithm, which is an abbreviation for (Random sample consensus) proposed by Fischler and Bolles in 1981 [33]. This algorithm estimates the parameters of a model by taking random samples from the target data. Where the dataset that contains inliers and outliers is used, RANSAC is considered efficient in terms of isolating the required data from the outliers and thus forming an efficient predictive model. RANSAC is used for regression because, as noted, it is good at handling outliers. This algorithm works by randomly selecting a subset of data samples and then using these samples to estimate the model parameters. Then in the next step, RANSAC finds samples that are within the fault tolerance range of the model. These samples are named inliers data and form a consensus set, and the rest data is named outliers data [34]. This algorithm trains the model using the inliers data and iterates these steps several times to produce a more efficient and less error-prone model.

This algorithm does not guarantee to obtain the best parametric model due to its randomness. However, the possibility of obtaining the optimal model remains possible by assigning appropriate values to the algorithm's parameters. One of the most important features of RANSAC is its ability to perform an efficient estimation of model parameters despite the presence of a large number of outliers in the target data set [35]. And the most important defect of this algorithm is that there is no specific time for calculating these parameters. Whereas, when a limited number of iterations is used, the optimal or desirable solution may not be obtained.

3.4.11 GRADIENT BOOSTING REGRESSION

Gradient Boosting Regression (GBR) is a well-known supervised AI algorithm. It works efficiently with data with missing values and outliers, and this model can detect non-linear relationships between features quite well. GBR is one of the most important ensemble methods that depend on obtaining an excellent prediction by collecting the prediction results of several poor methods together. It is an algorithm that reduces the bias error of the model. In regression-related problems, this algorithm uses the mean squared error (MSE) as a cost function. Gradient boosting is a very accurate prediction technique [36].

This algorithm is based on ensemble techniques that rely on grouping several decision trees. The design of decision trees is similar to the design of a natural tree. Where it begins with the root and then the branches until it reaches the leaves, as the terminal leaf is considered the final result or goal. A disadvantage of decision trees is the overfitting of test data if the hierarchy is too deep [37]. To avoid such a problem, the GBR algorithm works to merge several decision trees with a technique somewhat similar to the technique used in the Random Forest. Random Forests generate multiple decision trees by randomly segmenting the data used. It avoids overfitting by obtaining the prediction of all individual decision trees and averaging the results.

3.4.12 ELASTIC NET REGRESSION

Linear regression assumes a linear relationship between the input variables and the target variable. To make linear regression more efficient and less affected by outliers, penalties were added to the loss function to encourage simpler models. After making these developments, two types of linear regression appeared, which are penalized linear regression and regularized linear regression.

An Elastic Net is a well-known type of regular linear regression that relies on two important penalties, namely the penalty functions L1 and L2 [38]. That is, this algorithm is a combination of two techniques: Ridge and Lasso. As mentioned earlier, Lasso uses the L1 penalty and Ridge uses the L2 penalty. The main goal of Elastic Net is to reduce the following loss function [39]:

$$\frac{\sum_{i=1}^{n}\left(y_i-x_i^T\hat{\beta}\right)^2}{2n}+\lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2+\alpha\sum_{j=1}^{m}\left|\hat{\beta}_j\right|\right) \tag{2}$$

As the value of α approaches 0, then Elastic is closer in its work to a Ridge, and vice versa, if the value of α approaches 1, then Elastic is closer in its work to a LASSO.

4. RESULTS AND DISCUSSION

Twelve different algorithms were applied to three different datasets. Three tables will be listed below that represent the results achieved with each dataset separately. Table 5 below shows the results of ML algorithms with the first dataset:

Table 5. The Results of the Algorithms Using the First Dataset.

Algorithm	RMSE	MSE	MAE	Accuracy (R2)
Multilayer Perceptron	1.8082	3.2698	1.3527	58.97
SVR	1.1881	1.4116	0.8377	82.29
Multiple Linear Regression	1.1684	1.3652	0.8384	82.87
Decision Tree	0.5612	0.315	0.2750	96.05
Random Forest Regressor	0.3290	0.1082	0.2085	98.64
K- Nearest Neighbor (KNN)	1.9882	3.9532	1.544	50.40
Bagging Regressor	0.3631	0.1318	0.2254	98.35
LASSO	1.1634	1.3536	0.8247	83.02
Gaussian Processes Regressor	6.2713	39.33	5.6	-393.48
RANSAC	1.1604	1.3466	0.8240	83.10
Gradient Boosting	0.3286	0.1079	0.2354	98.65
Elastic Net	1.1599	1.3454	0.8252	83.12

Table 5 above summarizes the results of the algorithms applied to the first dataset. From these results, the best three algorithms were obtained, which are (Gradient Boosting, Random Forest, and Bagging Regressor) with results (of 98.65, 98.64, and 98.35) respectively. Table 6 below will show the results of the second dataset:

Table 6. The Results of the Algorithms Using the Second Dataset.

Algorithm	RMSE	MSE	MAE	Accuracy (R2)
Multilayer Perceptron	1243.3410	1243.3797	946.9159	47.18
SVR	1238.1194	1532939.8127	925.1295	47.63
Multiple Linear Regression	1203.8843	1449337.6418	916.6681	50.48
Decision Tree	1105.0661	1221171.0931	780.1205	58.28
Random Forest Regressor	1094.0445	1196933.5811	774.0955	59.11
K- Nearest Neighbor (KNN)	1545.5220	2388638.2604	1142.9383	18.39
Bagging Regressor	1212.0442	1469051.3589	838.9364	49.81
LASSO	1203.9653	1449532.5095	916.7961	50.47
Gaussian Processes Regressor	2762.7895	7633005.8353	2173.6021	-160.79
RANSAC	1356.2644	1839453.2686	1044.2302	37.15
Gradient Boosting	1091.6557	1191712.2726	774.0043	59.28
Elastic Net	1237.6056	1531667.8220	935.9901	47.67

Table 6 above lists the results of the algorithms applied to the second dataset with the medium correlation. It is noted that the obtained results are also medium. where the best algorithms are (Gradient Boosting, Random Forest, and Decision Tree) with results (of 59.28, 59.11, and 58.28).

Table 7. The Results of the Algorithms Using the Third Dataset

Algorithm	RMSE	MSE	MAE	Accuracy (R2)
Multilayer Perceptron	463.1929	214547.7335	375.1952	35.58
SVR	458.4476	210174.2116	365.3286	36.89
Multiple Linear Regression	453.9270	206049.7831	373.8584	38.13
Decision Tree	520.1328	270538.2311	408.5757	18.76
Random Forest Regressor	449.2760	201849.0024	368.1400	39.39
K- Nearest Neighbor (KNN)	488.1161	238257.3500	392.3922	28.46
Bagging Regressor	479.9841	230384.8177	381.6961	30.82
LASSO	453.9024	206027.4091	373.8363	38.13
Gaussian Processes Regressor	1498.3207	2244965.0564	1368.0676	- 574.12
RANSAC	522.6554	288187.6189	392.4079	17.97
Gradient Boosting	449.6354	202172.0207	367.6754	39.29
Elastic Net	453.8821	206008.9711	373.8241	38.14

Table 7 above shows the results obtained from the third dataset with a weak correlation between its features. As noted, the results obtained are weak and inaccurate. And the best algorithms in terms of results are (Random Forest, Gradient Boosting, and Elastic Net) with results (of 39.39, 39.29, and 38.14).

For all of the above, it is concluded that the first dataset with a good correlation between its features gave the best results with the three best ML algorithms (Gradient Boosting, Random Forest, and Bagging Regressor). As for the second dataset with medium correlation, it gave average results, as the best three algorithms in terms of accuracy with this dataset were (Gradient Boosting, Random Forest, and Decision Tree). As for the third dataset with weak correlation, it gave modest and weak results, as the three best algorithms were (Random Forest, Gradient Boosting, and Elastic Net). Finally, and based on this study, when using this model to predict the quantities of foodstuffs that can be sold, the degree of correlation of features must be measured, and then based on the degree, the three best-suggested algorithms are used to obtain the best prediction accuracy.

To use the proposed model, a real dataset of Iraqi Dates sales was used for the period (2002-2020). It contains several features, including types of dates, their quantities, prices, number of palm trees, and year of production, as shown in Table 4 above. In the beginning, the dataset was read and pre-processed. Then, the degree of correlation between the features of the dataset was measured using the Correlation Matrix of the Seaborn library, as shown in Figure 7 below:

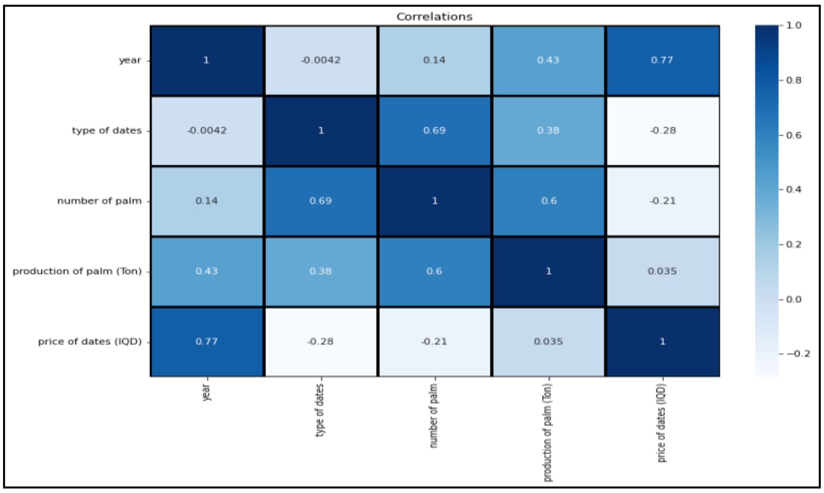


Figure 7. The correlation between Features of the Dataset.

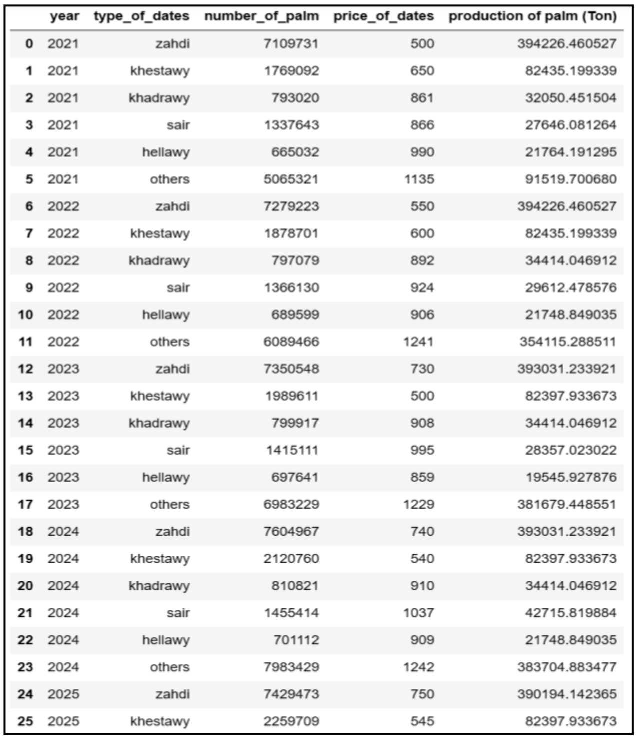
Figure 7 above shows that the correlation of this dataset tends to be good. On this basis, the ML algorithms proposed by the model previously were used, that worked better with datasets with good correlation, which are (Gradient Boosting, Random Forest, and Bagging Regressor). And when used, they gave predictions with excellent accuracy, as shown in Table 8 below:

Table 8. The Results of the Algorithms Using the Iraqi Dates Dataset

Algorithm	Accuracy (R2)
Gradient Boosting	99.51
Random Forest	97.09
Bagging Regressor	98.54

To predict the quantities of Dates that Iraq will sell for the next five years (2021-2025), The best algorithm was used, and gave the results shown in part in Table 9 below:

Table 9. Prediction obtained from Gradient Boosting Algorithm.



	year	type_of_dates	number_of_palm	price_of_dates	production of palm (Ton)
0	2021	zahdi	7109731	500	394226.460527
1	2021	khestawy	1769092	650	82435.199339
2	2021	khadrawy	793020	861	32050.451504
3	2021	sair	1337643	866	27646.081264
4	2021	hellawy	665032	990	21764.191295
5	2021	others	5065321	1135	91519.700680
6	2022	zahdi	7279223	550	394226.460527
7	2022	khestawy	1878701	600	82435.199339
8	2022	khadrawy	797079	892	34414.046912
9	2022	sair	1366130	924	29612.478576
10	2022	hellawy	689599	906	21748.849035
11	2022	others	6089466	1241	354115.288511
12	2023	zahdi	7350548	730	393031.233921
13	2023	khestawy	1989611	500	82397.933673
14	2023	khadrawy	799917	908	34414.046912
15	2023	sair	1415111	995	28357.023022
16	2023	hellawy	697641	859	19545.927876
17	2023	others	6983229	1229	381679.448551
18	2024	zahdi	7604967	740	393031.233921
19	2024	khestawy	2120760	540	82397.933673
20	2024	khadrawy	810821	910	34414.046912
21	2024	sair	1455414	1037	42715.819884
22	2024	hellawy	701112	909	21748.849035
23	2024	others	7983429	1242	383704.883477
24	2025	zahdi	7429473	750	390194.142365
25	2025	khestawy	2259709	545	82397.933673

The above results can be used economically. Where the Gradient Boosting algorithm gave results with an accuracy of (99.51) and for five years (2021_2025). This model can be used with other datasets related to other crops or even sales related to various economic aspects. Where the specific dataset is used and the accuracy of the correlation between its features is measured, the artificial intelligence algorithm is used according to the degree of correlation to obtain the best prediction results.

5. CONCLUSION

A model for predicting Dates sales in Iraq using AI techniques is what this study proposed. The Dates production sector in Iraq needs more technological support, and this is what this study aspires to. This model focused on the correlation of the features of the datasets which had a significant impact on the prediction results obtained from the ML algorithms. The model's prediction results are affected by the type of algorithm used, so determining the appropriate algorithm improves the results obtained. This study laid the foundation for the rest of the subsequent studies. It is possible to build on and expand this study in several aspects, such as increasing the number of used datasets and diversifying them in terms of features and internal data. Also

Include more algorithms and diversify them to include most of the models used to predict sales of food products. And finally using more metrics to measure the accuracy of the algorithms used to make the comparison between them more effective.

REFERENCES

[1] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 2, pp. 31-42, 2023.DOI: https://doi.org/10.52866/ijcsm.2023.02.02.004

[2] A. A. Mukhlif, B. Al-Khateeb, and M. Mohammed, "Classification of breast cancer images using new transfer learning techniques," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 1, pp. 167-180, 2023.DOI: https://doi.org/10.52866/ijcsm.2023.01.01.0014

[3] G. Caminero, M. Lopez-Martin, and B. Carro, "Adversarial environment reinforcement learning algorithm for intrusion detection," Computer Networks, vol. 159, pp. 96-109, 2019. https://doi.org/10.1016/j.comnet.2019.05.013

[4] X. Wang, D. Lin, W. Fan, and T. Wang, "Research on Sales Forecast of Fresh Produce Considering Weather Factors," 2018. https://aisel.aisnet.org/iceb2018/47

[5] V. Prabhakar, D. Sayiner, U. Chakraborty, T. Nguyen, and M. Lanham, "Demand Forecasting for a large grocery chain in Ecuador," Data. Published, 2018. https://varshaprabhakar07.github.io/img/3.pdf

[6] Y. F. Akande, J. Idowu, A. Misra, S. Misra, O. N. Akande, and R. Ahuja, "Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset," in Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2021: Springer, 2022, pp. 147-159. DOI: 10.1007/978-981-19-1111-8_13

[7] Y. Liu, "Grocery Sales Forecasting," in 2022 International Conference on Creative Industry and Knowledge Economy (CIKE 2022), 2022: Atlantis Press, pp. 215-219. DOI:10.2991/aebmr.k.220404.040

[8] A. Schmidt, M. W. U. Kabir, and M. T. Hoque, "Machine Learning Based Restaurant Sales Forecasting," Machine Learning and Knowledge Extraction, vol. 4, no. 1, pp. 105-130, 2022. https://doi.org/10.3390/make4010006

[9] T. S. Yange, C. O. Egbunu, O. Onyekwere, and K. A. Foga, "Prediction of Agro Products Sales Using Regression Algorithm," American Journal of Data Mining and Knowledge Discovery, vol. 5, no. 1, pp. 11-19, 2020.DOI: 10.11648/j.ajdmkd.20200501.12

[10] T. Tanizaki, T. Hoshino, T. Shimmura, and T. Takenaka, "Demand forecasting in restaurants using machine learning and statistical analysis," Procedia CIRP, vol. 79, pp. 679-683, 2019. https://doi.org/10.1016/j.procir.2019.02.042

[11] T. Weng, W. Liu, and J. Xiao, "Supply chain sales forecasting based on lightGBM and LSTM combination model," Industrial Management & Data Systems, vol. 120, no. 2, pp. 265-279, 2020. https://doi.org/10.1108/IMDS-03-2019-0170

[12] I. Vallés-Pérez, E. Soria-Olivas, M. Martínez-Sober, A. J. Serrano-López, J. Gómez-Sanchís, and F. Mateo, "Approaching sales forecasting using recurrent neural networks and transformers," Expert Systems with Applications, vol. 201, p. 116993, 2022. https://doi.org/10.1016/j.eswa.2022.116993

[13] X. Bi, G. Adomavicius, W. Li, and A. Qu, "Improving Sales Forecasting Accuracy: A Tensor Factorization Approach with Demand Awareness," INFORMS Journal on Computing, vol. 34, no. 3, pp. 1644-1660, 2022.https://doi.org/10.1287/ijoc.2021.1147

[14] N. Kumar, K. Dheenadayalan, S. Reddy, and S. Kulkarni, "Multimodal Neural Network For Demand Forecasting," arXiv e-prints, p. arXiv: 2210.11502, 2022. https://doi.org/10.48550/arXiv.2210.11502

[15] S. S. Raoof, M. Jabbar, and S. Tiwari, "Foundations of deep learning and its applications to health informatics," in Deep Learning in Biomedical and Health Informatics: CRC Press, 2021, pp. 1-28.https://www.taylorfrancis.com/chapters/edit/10.1201/9781003161233-1/foundations-deep-learning-applications-health-informatics-syed-saba-raoof-jabbar-sanju-tiwari

[16] I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 1, pp. 1-12, 2019. https://doi.org/10.48550/arXiv.1903.04579

[17] R. Gayathri, S. U. Rani, L. Čepová, M. Rajesh, and K. Kalita, "A Comparative Analysis of Machine Learning Models in Prediction of Mortar Compressive Strength," Processes, vol. 10, no. 7, p. 1387, 2022. https://doi.org/10.3390/pr10071387

[18] J. Manasa, R. Gupta, and N. Narahari, "Machine learning based predicting house prices using regression techniques," in 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA), 2020: IEEE, pp. 624-630. DOI: 10.1109/ICIMIA48430.2020.9074952

[19] M. Sinurat, M. Heikal, A. Simanjuntak, R. Siahaan, and R. N. Ilham, "Product Quality On Consumer Purchase Interest With Customer Satisfaction As A Variable Intervening In Black Online Store High Click Market: Case Study on Customers of the Tebing Tinggi Black Market Online Store," Morfai Journal, vol. 1, no. 1, pp. 13-21, 2021.https://doi.org/10.54443/morfai.v1i1.12

[20] D. Alita, A. D. Putra, and D. Darwis, "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 3, pp. 1-5, 2021. DOI: https://doi.org/10.22146/ijccs.65586

[21] M. Ahmad, N. A. Al-Shayea, X.-W. Tang, A. Jamal, H. M. Al-Ahmadi, and F. Ahmad, "Predicting the pillar stability of underground mines with random trees and C4. 5 decision trees," Applied Sciences, vol. 10, no. 18, p. 6486, 2020.https://doi.org/10.3390/app10186486

[22] C. Peláez-Rodríguez, J. Pérez-Aracil, D. Fister, L. Prieto-Godino, R. Deo, and S. Salcedo-Sanz, "A hierarchical classification/regression algorithm for improving extreme wind speed events prediction," Renewable Energy, vol. 201, pp. 157-178, 2022. https://doi.org/10.1016/j.renene.2022.11.042

[23] R. Sathya, "Ensemble Machine Learning Techniques for Attack Prediction in NIDS Environment," Iraqi Journal For Computer Science and Mathematics, vol. 3, no. 2, pp. 78-82, 2022. DOI: https://doi.org/10.52866/ijcsm.2022.02.01.008

[24] Q. Zhang, "Financial data anomaly detection method based on decision tree and random forest algorithm," Journal of Mathematics, vol. 2022, 2022. https://doi.org/10.1155/2022/9135117

[25] S. Bishnoi and H. Hisar, "k–NEAREST NEIGHBOR (k-NN) ALGORITHM FOR CLASSIFICATION," Dr. Med Ram Verma, p. 109. https://www.bhumipublishing.com/wp-content/uploads/2022/07/Advances-in-Mathematical-and-Statistical-Science.pdf#page=115

[26] B. Wang et al., "A novel weighted KNN algorithm based on RSS similarity and position distance for Wi-Fi fingerprint positioning," IEEE Access, vol. 8, pp. 30591-30602, 2020. DOI: 10.1109/ACCESS.2020.2973212

[27] H. ŞAHİN and İ. Duygu, "Application of Random Forest Algorithm for the Prediction of Online Food Delivery Service Delay," Turkish Journal of Forecasting, vol. 5, no. 1, pp. 1-11. https://doi.org/10.34110/forecasting.842180

[28] T. N. Rincy and R. Gupta, "Ensemble learning techniques and its efficiency in machine learning: A survey," in 2nd International Conference on Data, Engineering and Applications (IDEA), 2020: IEEE, pp. 1-6.DOI: 10.1109/IDEA49133.2020.9170675

[29] A. A. El Sheikh, S. L. Barakat, and S. M. Mohamed, "New aspects on the modified group LASSO using the least angle regression and shrinkage algorithm," Information Sciences Letters, vol. 10, no. 3, pp. 527-536, 2021.doi:10.18576/isl/100317

[30] P. A. Omer, "Improving Prediction Accuracy of Lasso and Ridge Regression as an Alternative to LS Regression to Identify Variable Selection Problems." DOI: https://doi.org/10.31972/ticma2022

[31] D. V. Dao et al., "A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a Monte Carlo simulation," Sustainability, vol. 12, no. 3, p. 830, 2020. https://doi.org/10.3390/su12030830

[32] J. Wang, "An intuitive tutorial to Gaussian processes regression," arXiv preprint arXiv:2009.10862, 2020.https://doi.org/10.48550/arXiv.2009.10862

[33] C. Lara-Alvarez and F. Gonzalez-Herrera, "Testing multiple polynomial models for eye-tracker calibration," Behavior Research Methods, vol. 52, no. 6, pp. 2506-2514, 2020. https://doi.org/10.3758/s13428-020-01371-x

[34] J. M. Martínez-Otzeta, I. Rodríguez-Moreno, I. Mendialdua, and B. Sierra, "RANSAC for Robotic Applications: A Survey," Sensors, vol. 23, no. 1, p. 327, 2022. https://doi.org/10.3390/s23010327

[35] H. K. Sangappa and K. Ramakrishnan, "A probabilistic analysis of a common RANSAC heuristic," Machine Vision and Applications, vol. 30, pp. 71-89, 2019. https://doi.org/10.1007/s00138-018-0973-4

[36] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transportation Research Part C: Emerging Technologies, vol. 58, pp. 308-324, 2015. https://doi.org/10.1016/j.trc.2015.02.019

[37] E. Spiliotis, "Decision trees for time-series forecasting," Foresight, vol. 1, pp. 30-44, 2022.https://econpapers.repec.org/article/forijafaa/y_3a2022_3ai_3a64_3ap_3a30-44.htm

[38] M. Y. Shams, O. M. Elzeki, L. M. Abouelmagd, A. E. Hassanien, M. Abd Elfattah, and H. Salem, "HANA: a healthy artificial nutrition analysis model during COVID-19 pandemic," Computers in Biology and Medicine, vol. 135, p. 104606, 2021. https://doi.org/10.1016/j.compbiomed.2021.104606

[39] B. Yu, C. Chen, X. Wang, Z. Yu, A. Ma, and B. Liu, "Prediction of protein–protein interactions based on elastic net and deep forest," Expert Systems with Applications, vol. 176, p. 114876, 2021. https://doi.org/10.1016/j.eswa.2021.114876