2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)
Karnataka, India. Dec 29-31, 2023

2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) | 979-8-3503-1912-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/SMARTGENCON60755.2023.10442766

Demand Prediction of Consumer Intention to Buy Edible Items using Machine Learning Techniques

Hitesh Goyal
Chitkara University Institute of Engineering & Technology,
Chitkara University,
Punjab, India
hitesh6020.be21@chitkara.edu.in

Ashna Sukhija
Chitkara University Institute of Engineering & Technology,
Chitkara University,
Punjab, India
ashna6045.be21@chitkara.edu.in

Kavish Bhatia
Chitkara University Institute of Engineering & Technology,
Chitkara University,
Punjab, India
kavish6051.be21@chitkara.edu.in

Kalpna Guleria*
Chitkara University Institute of Engineering & Technology,
Chitkara University,
Punjab, India
guleria.kalpna@gmail.com*

Shagun Sharma
Chitkara University Institute of Engineering & Technology,
Chitkara University,
Punjab, India
shagunsharma7098@gmail.com

***Abstract*— Agriculture is a process of simplifying the food chain in nature and redistributing the resources for animals and plant production. It is the process of nurturing livestock and growing crops. It consists of preparing animal and plant items for human consumption and distributing them to marketplaces. A maximum number of textiles and food items in the world are produced by agriculture. In India, the primary source of income is agriculture. Various companies buy crops from the farmers and directly sell them to the consumers. As per a study by "The India Express", there are 90-150 million farmers who rely on agriculture to run the expenses of their daily lives. Each day there is a huge amount of vegetable waste as their demand doesn't get predicted due to which a volume of vegetables gets rotten. So to deal with this issue, there is a need for the development of an automated model which can accurately predict the demand for vegetables for the next day based on the previous data. In this work, a framework has been developed for predicting the demand for okra and tomato for the next day, so that the farmers won't pluck too many vegetables and also keep track of daily sales. We have used four machine learning models namely, linear regression, decision tree, random forest, and logistic regression for identifying the demand for vegetables. In this experiment, two parameters %error and accuracy have been taken which shows that the decision tree model performs best with the highest accuracy of 99.62% and logistic regression has shown the least accuracy at 85.62%. This work can help various edible-based eCommerce companies for predicting the demand for the items for the next day based on the previous data.**

Keywords—Agriculture, machine learning, linear regression, decision tree, random forest, logistic regression.

I. INTRODUCTION

Agriculture is a primary source of livelihood for the Indian economy [1]. Half of the population in India is employed in the agriculture sector which has a great contribution to the country's GDP and almost 70% of the rural population depends on agriculture for living their daily life. Agriculture is the industry that farmers are a part of and this industry produces a variety of food products for human and animal consumption [2]. Farmers' main goal is to feed the population and make a living expense by producing good crops. They work hard for the whole year and produce good quality food for a satisfactory life. This food is purchased from the farmers at a reasonable price and stored in warehouses or cold storage until it gets a good market price. From the warehouses, food is transported to the wholesaler then from the wholesale market it is sold to local shops and vendors (retailers) which is then finally purchased by consumers in the market. Over the past four decades, India's agricultural productivity has increased rapidly [2][3]. According to the Food and Agriculture Organization (FAO), India comes in second place as the largest producer of fruits and vegetables in the world. However, the availability of food remains a major concern due to the issues of poor packaging, absence of sheltered storage, inefficient traders and many more [4]. The fragmented nature of the food industry leads to inefficiency and losses in the supply chain. To overcome these challenges, there are various agri-tech startups and companies which aim at connecting farmers directly to the consumer through online channels and providing businesses with fresh produce in the most efficient manner. Fresh fruits and vegetables are directly collected from the farmers [3]. Then, in the warehouse quality is checked and the best ones are packed and transported to the consumers in a limited period. The company buys the required amount of fruits and vegetables as per the demands of the consumers. Keeping all these factors in mind created a Machine Learning(ML) model to predict the demand for fruits and vegetables using historical data [4].

ML is a branch of Artificial Intelligence (AI) which enables systems to make successful predictions using past experiences [4][5][6]. The system is given the ability to think like humans, with the help of various algorithms and techniques, on its own [7][8]. In today's modern time, only agriculture is a field which lacks technological advancements. ML can play a big role in making agriculture more efficient and effective and help farmers in getting better MSP for their crops. ML algorithms can easily and accurately predict the weather and soil conditions, demand for the crop and correct price by comparing it with historical data. When we train an algorithm, based on the historical dataset and apply new data to it, the output which comes is called prediction. Demand prediction is a technique that gives the output a requirement of a particular product as per the demand of the consumer. And in agriculture, it gives the output of a particular crop required as per the demand of the consumers. It helps the farmers to select and grow the appropriate crop to satisfy the demand. This reduces the gap or mismatch between the demand and supply of the crop and also reduces the wastage of crops.

This paper is structured as follows: Section II introduces various state-of-the-art models for the demand prediction of edible items. Section III summarizes the methodology used for predicting the demand for okra and tomato. Lastly, the prediction results of the proposed model and conclusion have been mentioned in sections IV and V, respectively.

II. Related Work

An ML model has been developed in [9] for the prediction of nitrogen in rice crops. The research contained three different datasets for the prediction namely, multi-temporal, multi-source, and multi-scale and applied XGBoost, and various other ML models for the prediction which resulted in the XGBoost model performing best among other algorithms with the highest accuracy. An ML model has been presented in [10] to work on the problem of farmers like price fluctuations of farmer tools in the market and supply-demand of the tools required. The research done used various ML models- decision trees, deep neural networks and others, for price prediction and supply and demand for farmer tools. It was concluded that the most accurate and effective ML model for sharing and renting the farmer's equipment is the decision tree. A review of the price prediction of crops has been done in [4]. The research done contains a prediction of attributes like soil condition, weather conditions and demand of the crop in the marketplace to predict the price using past scenarios. The research has used many ML models for different purposes- Linear Regression, Decision Tree and Random Forest. Various ML algorithms have been identified in [2] for the identification of crop yield. This survey has identified that artificial neural networks (ANN) and SVM models perform best for the prediction of crop yield. A survey on emerging trends in ML has been done in [3] to predict crop yield and its influence factors. The motive of the research is to explore various ML techniques used for predicting crop yield and solving the food problem in the world. The various techniques surveyed are ANN, LSTM, RNN, Fuzzy Technique and various other ML models. The best results were obtained through the Neural Network technique, ANN and ANFIS.

TABLE I. Comparison of various existing works in terms of algorithm and domain of the research work

Ref.	Technique	Domain
[1]	Decision tree	Crop price prediction
[2]	ANN and SVM	Crop yield prediction
[3]	RNN, LSTM, ANN and fuzzy technique	Crop yield prediction
[4]	Linear regression, random forest and decision tree	Crop price prediction
[9]	XGBoost	Rice crop prediction
[10]	Decision tree and Deep neural networks	Farmer tools price prediction
[11]	KNN and Naive bayes	Crop yield prediction
[12]	ANN	Food demand prediction

Authors in [11] have done a semantic analysis of the opinions of experts on crop productivity through ML. The study has shown a descriptive analysis to increase the crop yield and provided the requirement of awareness in the current agriculture system using ML. The ML algorithms used are KNN and Naive Bayes. It concluded that Naive Bayes performs better on a text dataset of agriculture experts' opinions. An ML model has been developed in [1] for the price prediction of crops using supervised ML algorithms. The developed model predicted the crop price and forecasted how much the price will be there in the upcoming 12 months. The model has been used to predict the crop price using decision tree regression. An analysis of neural networks and regression models has been performed in [12] to Forecast the demand for food. The research done is implementing multiple linear regression and an ANN model for predicting the food crops demand which is mostly used in daily life. Table. I show various ML approaches used in various different sectors of prediction such as rice, crop, and farmer tools price prediction.

III. Materials and Method

This section of the paper provides an overview of the dataset along with the methodology applied to that for achieving efficient prediction results.

A. Dataset

The dataset is collected from a new startup company which directly deals with the demand prediction problem of fruits and vegetable packets. The dataset contains a total number of 330 data points including four attributes: AVG. PRICE, PROD_NAME, PACK_SOLD, ORDER_OF_PRODUCT. In the proposed model, the AVG. PRICE, PROD_NAME, and PACK_SOLD have been taken as the inputs whereas the model predicts the output in the form of ORDER_OF_PRODUCT.

B. Methodology

This subsection represents the proposed methodology used for the prediction of the demand for okra and tomato.

In this work, the preprocessing step has been performed on the raw data to make it efficient and useful. In this phase of the proposed model, the merging and removing of attributes have been done manually according to the need of the output and a single .csv file has been formed. In the second step modification of the data type of an attribute, PROD_NAME is done by assigning it a numerical value instead of categories. The third phase includes the replacement of null or empty values in the dataset by the mean/median values of the attributes and then randomizing data to evacuate the impact of ordered data gathered. Perform further analysis such as visualizing data using histograms to understand the data and trends more glaringly. The train and test ratio of the data has been taken as 80:20. After cleaning data into two parts, the first part is utilized to build the model and the latter is to probe the model, check the accuracy of the model or examine the error in predicted values. For further evaluation, the CROSS VALIDATION technique will be used. In cross-validation, the k-fold cross-validation has been used where the value of k has been taken as 9 as shown in Fig. 1.

In this technique, the train set is further divided into 10 sets. Out of 10 sets, 1 part is utilized to examine the model and the left 9 are utilized to compose it. Repeatedly, the

cross-validation technique has been done on alternative sets. Finally, the %error of the values has been identified for predicting the performance of the models. There are various ML models which have been used to identify the prediction of the demand for okra and tomato. These algorithms are used to train the model to forecast the precise/best prediction outcome. The proposed methodology has been shown in Fig. 2.

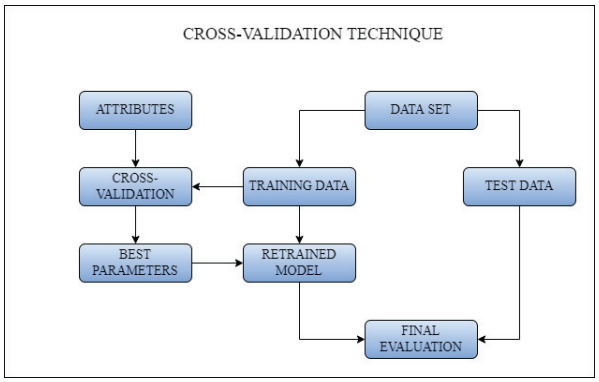

Fig. 1. Cross-validation technique

The algorithms used in the proposed work are as follows:

Linear regression: Linear regression is a statistical way used for predictive analysis [13][14]. It could be explained as a single-layer perceptron neural layout involving input and output variables. It exhibits the linear relationship, which signifies how the values of dependent variables are altered according to the values of independent variables.

Decision tree regressor: The decision tree regressor algorithm is a supervised learning algorithm whose primary motive is to forecast the aimed variables by framing a training model[15]. It is a technique based on trees in which each path originating from the root is represented by a data-isolated sequence until a boolean value is achieved at the leaf node[15][16]. The decision tree algorithm is also optimized if the data set carries the minimum number of nodes and is properly classified.

Random forest regressor: It regressor is a well-admired algorithm that pertains to the supervised learning technique [17]. It can help solve both classification and regression problems in ML. It mainly revolves around the concept of ensemble learning, which is used to solve complex problems by combining multiple classifiers for improving the performance of the model. In short, a random forest is a classifier in which numerous decision trees are carried out on various subparts of a given dataset and takes the mean to improve the predictive accuracy of the model. Simply, a random forest makes the prediction from each tree and based on the maximum votes prediction, it will predict the final output [18][19].

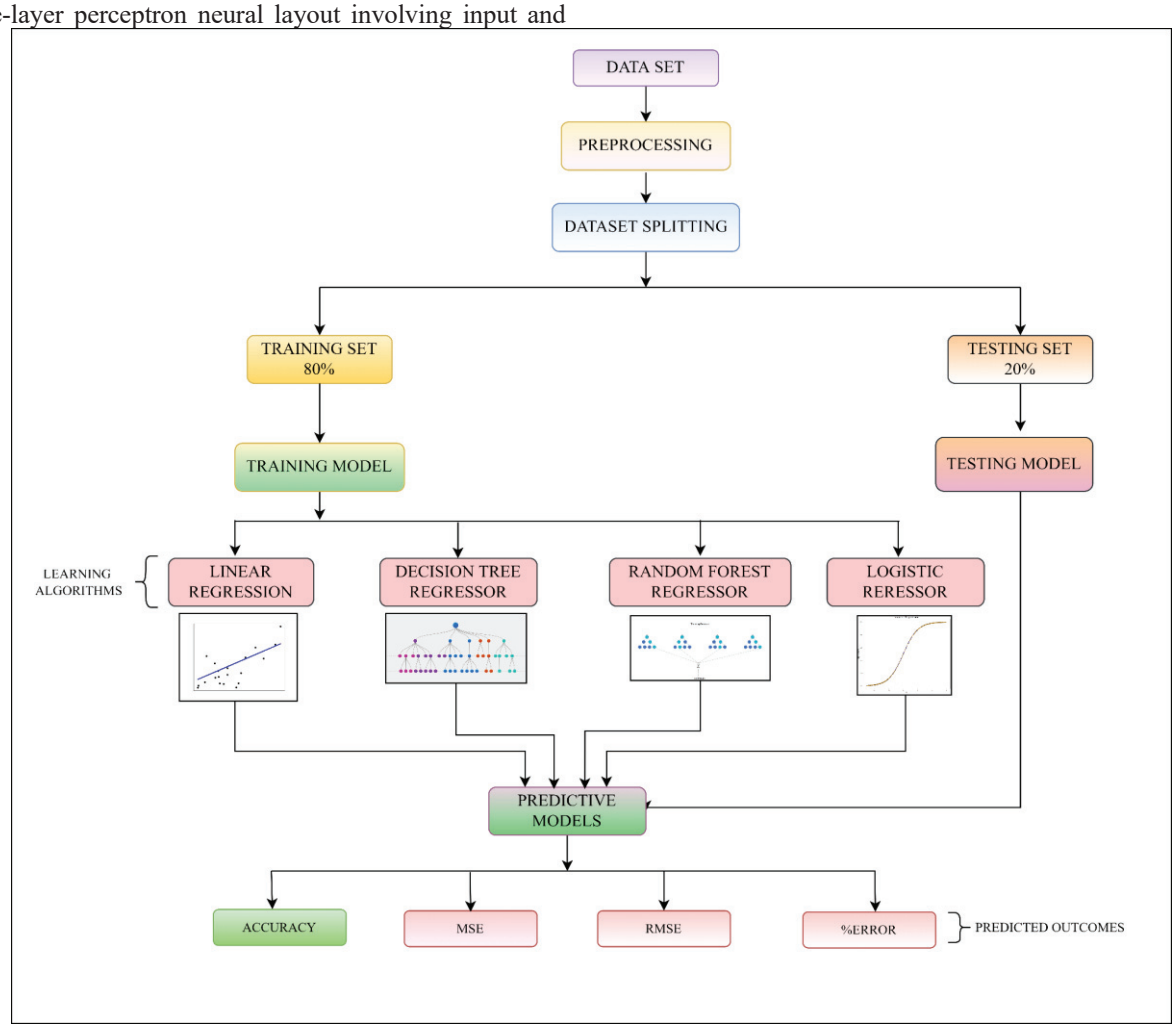

Fig. 2. Proposed methodology

Logistic regression: It is also associated with supervised learning techniques which are used to predict categorical dependent variables from the given set of independent variables [20][21]. It is used in various predictive models such as healthcare, stock movement, crop yield and price prediction [22][23][24]. The output of this model must be categorical or distinct values which are in the boolean form either 0 or 1, true or false, Yes or No. It gives the values in probability form between 0 and 1 instead of showing the values as 0 and 1. This type of model is similar to linear regression which is best known for solving and predicting regression-based challenges while logistic regression works for solving the problems of classification.

IV. Implementation Results

This section identifies the prediction performance of various ML algorithms used in this paper which are shown with the help of illustrations in the graphs below.

In the evaluation phase, various algorithms have been used to forecast the demand for fruits and vegetables and gained different results from each algorithm. The proposed model is based on a regression problem hence there are various regression-based parameters which have been utilized to identify the performance results of the proposed work which are: accuracy and %error.

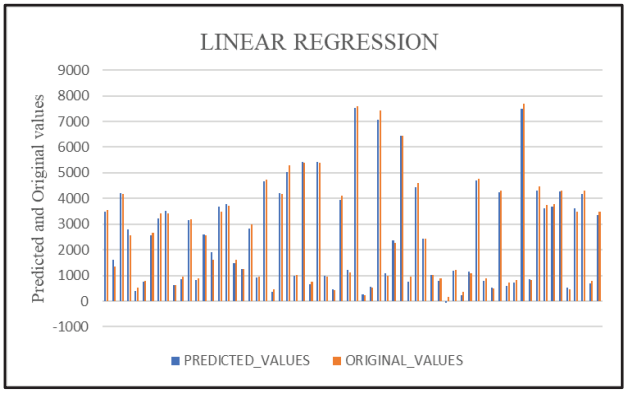

Fig. 3. Predicted vs original values using linear regression

Fig. 3 shows the relationship between the original values and predicted outputs using the linear regression algorithm. This graph shows that the model has some variations between the prediction results and the actual values.

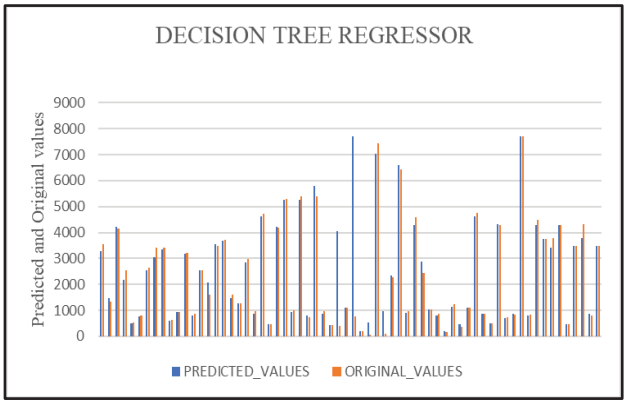

Fig. 4. Predicted vs original values using decision tree

Fig. 4 describes the results of the decision tree algorithm which performs quite well on the dataset and predicts the precise values. The comparison between the values of prediction and actual attribute shows that the execution of the model is good and the predicted values are so close to the original values.

Fig. 5 depicts the relationship between the actual value and predicted values of the demands of okra and tomato. The graph shows that the random forest has performed quite good but some predicted values are too large which is directly affecting the accuracy of the model.

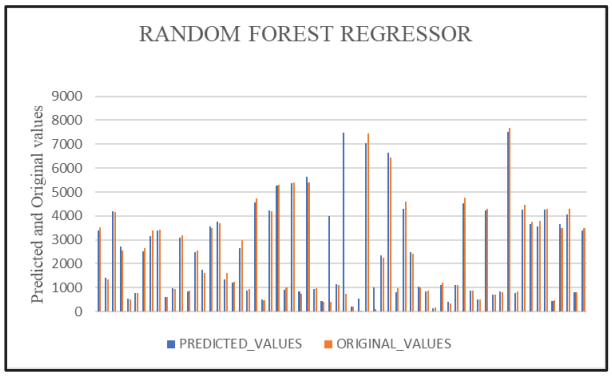

Fig. 5. Predicted vs original values using random forest

Fig. 6 shows the prediction performance of the logistic regression model in the case of predicting the demand for okra and tomato. The graph identifies that the logistic regression model has a very huge difference between the actual demands and the predicted demand which results that this model is not suitable for prediction with such kind of dataset.

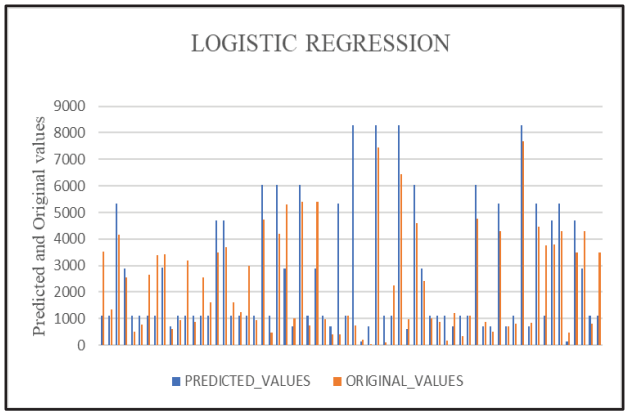

Fig. 6. Predicted vs original values using logistic regression

Fig. 7 shows the comparative analysis of various ML algorithms in the form of accuracy. The comparison results show that the decision tree algorithm performs best with the highest accuracy of 99.62% among all the algorithms for predicting the demand for okra and tomato. The random forest model also performs better after the decision tree with an accuracy rate of 98.61%. Furthermore, the accuracy of the linear and logistic regression have been identified as 95.90% and 85.62%, respectively.

Fig. 8 shows the comparative analysis of various ML models based on %error. This result identifies that the decision shows the least %error as 0.37 whereas random forest, linear regression and logistic regression perform after that with the %error of 1.38, 4.09 and 14.377, respectively.

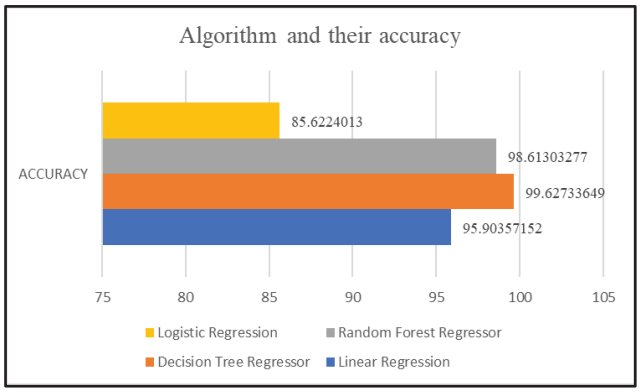

Fig. 7. Accuracy comparison of various machine learning algorithms

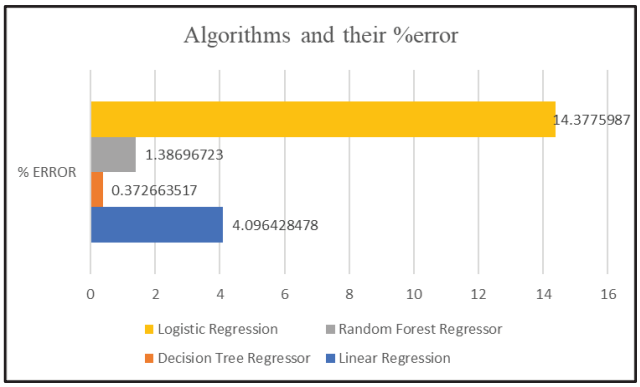

Fig. 8. %error comparison of various machine learning algorithms

The performance results of logistic regression, linear regression, random forest and decision tree have been tabulated in Table II. as %error and accuracy.

TABLE II. PERFORMANCE COMPARISON OF VARIOUS TECHNIQUES PROPOSED IN THE WORK WITH EXISTING WORK

Technique	**Accuracy**	**%error**
Linear Regression	95.9%	4.09%
Decision Tree	99.62%	0.37%
Random Forest	98.61%	1.38%
Logistic Regression	85.62%	14.37%
Naive Bayes[11]	87%	13%
KNN[11]	85%	15%

Table II. presents the performance comparison of the proposed work and existing work. This comparison shows that the proposed decision tree model as an outperforming model having the highest accuracy rate of 99.62% while other proposed models namely linear regression and random forest also outperform the existing models. Furthermore, in comparison with logistic regression, it performs quite better than KNN[11] but underperformed when compared to Naive Bayes[11].

V. CONCLUSION

The agriculture sector is extremely important to the country's economy in the nation. In the past few years, several advancements in the realm of agriculture have been put into practice. With the advent of AI and ML, various issues which are faced by farmers have been resolved. Similarly, there is another issue causing a big loss to the farmers which is the demand for crops and vegetables required the next day. Various companies buy crops along with vegetables from farmers and sell them directly to consumers. But the problem with this is unpredicted demand due to a huge amount of rotten vegetable and fruit waste each day. These companies buy only limited stock. However, it causes problems for the farmers as the remaining vegetable and fruit get rotten and wasteful. The objective of this paper is to describe modern ML techniques in the commercial agricultural sector. To reduce the wastage of fruits and vegetables by predicting the optimum demand of fruits and vegetables required. In this paper, we are discovering the solution to a real-time problem in which we predicted the demand for a particular product, utilizing historical data. ML plays a vital role in the prediction of the demand for vegetables and fruit. The visualising techniques of ML are also used for a better understanding of data, such as graphs and histograms that provide ease in determining the performance of the algorithm. Four different ML approaches namely random forest, decision tree, logistic regression and linear regression have been used for the prediction based on the previous data. In this paper, the decision tree has been identified as the best method with the highest accuracy value of 99.62% for predicting the demand for crops in the upcoming day. While the performance of random forest, linear regression and logistic regression have been identified with the accuracy value of 98.61%, 95.90% and 85.62%, respectively. This article can help various eCommerce companies for predicting the vegetable and fruit demand each day. Furthermore, this work is also helpful for academia to understand the applications of various algorithms used in the paper.

REFERENCES

[1] R. Dhanapal, A. AjanRaj, S. Balavinayagapragathish, and J. Balaji, “Crop price prediction using supervised machine learning algorithms,” J. Phys. Conf. Ser., vol. 1916, no. 1, p. 012042, 2021.

[2] V. Nathgosavi, “A survey on crop yield prediction using machine learning,” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 13, pp. 2343–2347, 2021.

[3] N. Bali and A. Singla, “Emerging trends in machine learning to predict crop yield and study its influential factors: A survey,” Arch. Comput. Methods Eng., vol. 29, no. 1, pp. 95–112, 2022.

[4] M. Rakhra et al., “Crop price prediction using random forest and decision tree regression:-A review,” Mater. Today, 2021.

[5] S. Sharma and K. Guleria, “Deep learning models for image classification: Comparison and applications,” in 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022.

[6] R. Sharma and V. Kukreja, “Mustard Downy Mildew Disease Severity Detection using Deep Learning Model,” in 2021 International Conference on Decision Aid Sciences and Application (DASA), pp. 466-470, 2021.

[7] S. Sharma and K. Guleria, “A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images,” Multimed. Tools Appl., 2023.

[8] R. Sharma and V. Kukreja, “Amalgamated convolutional long term network (CLTN) model for Lemon Citrus Canker Disease Multi-classification,” in 2022 International Conference on Decision Aid Sciences and Applications (DASA), 2022.

[9] M. Iatrou et al., “Topdressing nitrogen demand prediction in rice crop using machine learning systems,” Agriculture, vol. 11, no. 4, p. 312, 2021.

[10] M. Rakhra, A. Bhargava, D. Bhargava, R. Singh, A. Bhanot, and A. W. Rahmani, “Implementing machine learning for supply-demand shifts and price impacts in farmer market for tool and equipment sharing,” Journal of Food Quality, 2022.

[11] M. Rehman et al., “Semantics analysis of agricultural experts’ opinions for crop productivity through machine learning,” Appl. Artif. Intell., vol. 36, no. 1, pp. 1–16, 2022.

[12] B. P. Bv and M. Dakshayini, “Computational Performance Analysis of Neural Network and Regression Models in Forecasting the Societal Demand for Agricultural Food Harvests,” in Research Anthology on Artificial Neural Network Applications, IGI Global, 2022, pp. 1287–1300.

[13] W. Wei and X. Yang, “Comparison of diagnosis accuracy between a backpropagation artificial neural network model and linear regression in digestive disease patients: An empirical research,” Comput. Math. Methods Med., vol. 2021, p. 6662779, 2021.

[14] T. K. Saha, S. Pal, and R. Sarkar, “Prediction of wetland area and depth using linear regression model and artificial neural network based cellular automata,” Ecol. Inform., vol. 62, no. 101272, p. 101272, 2021.

[15] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021.

[16] C. S. Lee and P. Y. S. Cheang, “Predictive analysis in business analytics: Application of decision tree in business decision making,” Adv. Decis. Sci., vol. 26, no. 1, pp. 1–29, 2021.

[17] M. A. Khan et al., “Application of random forest for modelling of surface water salinity,” Ain Shams Engineering Journal, vol. 13, no. 4, 2022.

[18] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, “Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model,” Big Data Min. Anal., vol. 4, no. 2, pp. 116–123, 2021.

[19] K. Guleria, S. Sharma, S. Kumar, and S. Tiwari, “Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning,” Measurement: Sensors, vol. 24, no. 100482, p. 100482, 2022.

[20] P. Schober and T. R. Vetter, “Logistic regression in medical research,” Anesth. Analg., vol. 132, no. 2, pp. 365–366, 2021.

[21] X. Song, X. Liu, F. Liu, and C. Wang, “Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis,” Int. J. Med. Inform., vol. 151, no. 104484, p. 104484, 2021.

[22] P. K. Sarangi, K. Guleria, D. Prasad, and D. K. Verma, “Stock movement prediction using neuro genetic hybrid approach and impact on growth trend due to COVID-19,” Int. j. netw. virtual organ., vol. 25, no. 3/4, p. 333, 2021.

[23] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, “A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans,” Measurement: Sensors, vol. 24, no. 100506, p. 100506, 2022.

[24] S. Srivastav, K. Guleria and S. Sharma, "Tea Leaf Disease Detection Using Deep Learning-based Convolutional Neural Networks," 2023 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2023, pp. 569-574, doi: 10.1109/AIC57670.2023.10263835.