2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-7971-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICESC54411.2022.9885315

Optimized Ensemble Model for Wholesale Market Prediction using Machine Learning

Rohit Bajaj
CSE Department
Chandigarh University
Mohali , Punjab
rohit.rick@gmail.com

Gaurav Bathla
CSE Department
Chandigarh University
Mohali , Punjab
gauravbathla86@gmail.com

Abhishek Gupta
CSE Department
Chandigarh University
Mohali , Punjab
abhishekgup232326@gmail.com

Anurag
CSE Department
Chandigarh University
Mohali , Punjab
anuragdharwal7@gmail.com

Lokesh Pawar
CSE Department
Chandigarh University
Mohali , Punjab
lokesh.pawar@gmail.com

***Abstract*—Wholesale is the practice of selling low-cost items to retailers or directly to customers. The wholesale customer dataset primarily provides information on customers' interest in the wholesale market and their annual consumption of a specific product. The goal of this research is to create a preferred machine learning model that can assist wholesale retailers in increasing their production and profit by anticipating the quantity of products necessary in the market. The proposed research study has initially attempted to incorporate typical machine learning models, but it has not resulted in satisfactory results, so this research study has further progressed to propose an ensemble model, and the satisfactory results were obtained for the proposed model. The algorithms utilized in developing this model are SMOreg and Kstar. SMOreg is a programming problem-solving algorithm, which is utilized for training the support vector machine (SVM). The proposed dataset was collected from the UCI Machine Learning repository, which is an open-source platform.**

Keywords—Regression, Ensemble, Retailers, Algorithm, K-star.

I. INTRODUCTION

The like and dislikE of a certain product alter very frequently in the thinking of the end-user (Consumer) in today's fast-paced world. Furthermore, market strategies are always altering in response to client segmentation. Furthermore, Wholesale Retailers [1] may have issues with the bulk manufacture of a product that, despite its quality, fails in the market. To address this issue, researchers have applied advanced tools [2] and machine learning techniques to a dataset called Wholesale Customer Data, which comprises eight attributes as shown in Fig 1.1.

Channels are the linkage or the connections of the brokers through which the products are distributed or sold (hotels, restaurants, Café, and more). These channels are determined by numbers given to every other channel uniquely.

Regions are the places or areas in which the products are being sold, these regions are also determined by numbers given to every other region uniquely.

Fresh, Milk, Grocery, Frozen, Detergents paper, and Delicatessen, are the attributes having the value of annual spending on the respective products mentioned above as attributes. This study uses the eight attributes [3] provided in Fig. 1.1 to predict the best probable quantity and packaging of a certain product in a given region.

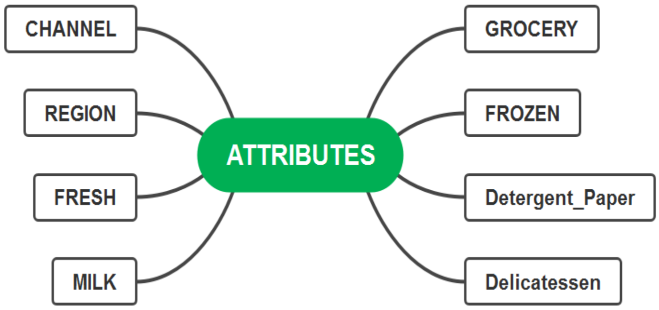

Fig. 1.1 Characteristics Utilized In Predictive Model

We first employed regression [4], in which we used two algorithms, "SMOreg" and "KStar," with the help of stacking, and then we trained an "ensemble model" since we saw a significant improvement above typical machine learning models.

Any Wholesale Retailer [5] may simply increase their client segmentation and market concepts using this strategy [6]. It can also be used to determine how much a wholesale product [7-8] should cost and how much it should make.

II. LITRATURE REVIEW

Lakshmanrao et al. used a clustering algorithm to analyze the wholesale customer data in which they divided the dataset into three clusters by Kmean and Hierarchical clustering algorithm. Time analyzed for used algorithm is not suitable large datasets.

Bangert et al.proposed a Markov-chain model in which he used client segmentation in the wholesale market using data mining with the help of K-mean, a clustering descriptive statistics algorithm. Proposed traditional way to solve the problems that can be accomplished by modern algorithms.

Parvania et al. Used a self-scheduling model for the aggregate wholesale electricity market in which we found that employees and supervisors must become accustomed to their new roles is a big problem here that reduces the productivity of the seller or producer but on the other hand it's also time saving.

Bottani et al. In this paper, they used Artificial-Intelligence based framework to solve the problem of wholesale distribution to reduce the risk of out-of-stock scenarios. This model generally makes the better relationship between the seller and the buyers. High cost of creation and Unemployment are the major drawbacks of this model.

Zarnikau et al. In this he used a statistical model to restructure the wholesale price in the Texas electricity market, here we can also use the analytical model instead of the statistical model. It does not deal with the individual item and result is truly based on average. So, modern techniques can be used to get better outcome.

Nguyen duydat et al. proposed a model in which they explored wholesale customer behavior by outsourcing factories in Vietnam. With the help of the spare GP technique, we can reduce the computational time of data in this model. This technique is very helpful at the time of training data or model to get vector of output values.

Kwok Hung Lau et al. This study basically focuses on balancing the distribution of products efficiently andbalancing is an important aspect that is necessary to be done before the preprocessing. It helps in knowing the responses of customer needs so that a wholesaler to increase its productivity.

Sourav Ray et al. Proposed a model in which they adjusted the cost of price and channel pricing downstream, In this model traditional algorithms can be replaced by modern techniques to get the preferred result. Techniques like Linear regression and Ensembling can bring better outcome.

Joseph O. Chan et al. proposed an enterprise model that comprises of the relationship between the retailers and the consumers. It generally uses the modern algorithms to improve the relationship between the suppliers and the consumers. This model gains competitive advantage and increases control and consistency in the wholesale market.

Lin Chen et al. evaluated latter model that basically lowers the retail price and increase the consumer surplus. This model also shows the profit gained by the retailers from the consumers and their online profit depend upon the customer's loyalty. But this model, risks its productivity on the consumers

YacineRekik et al. made a model which reduce the inaccuracy in the inventory and optimal ordering policy is derived in which demand optimization is done on the basis of inventory record. and it is linked with the well known random yield problem.

FarisOdeh Al Majali et al. implemented a framework in which he manipulated performance factors to the component of proposed framework and this model generally contains the gaining information about the performance and the factors affecting the wholesale markets.

III. Proposed Methodology

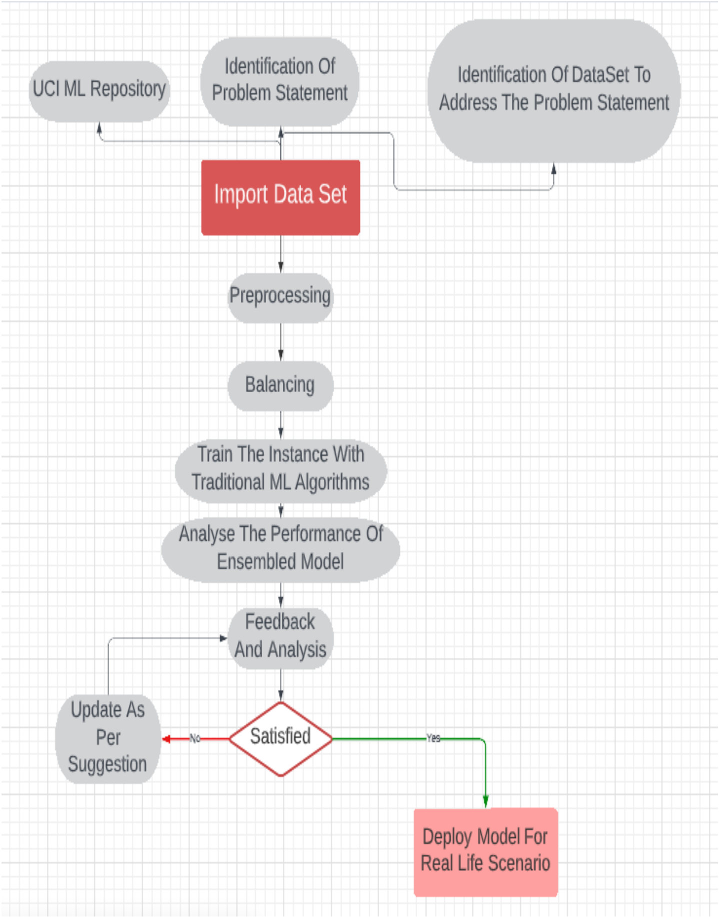
UCI ML Repository
Identification Of Problem Statement
Identification Of DataSet To Address The Problem Statement
Import Data Set
Preprocessing
Balancing
Train The Instance With Traditional ML Algorithms
Analyse The Performance Of Ensembled Model
Feedback And Analysis
Update As Per Suggestion
No
Satisfied
Yes
Deploy Model For Real Life Scenario

Fig.3. 1.Proposed Model

To overcome the problem, we applied several cutting-edge machine learning techniques [9].
To accomplish so, we first gathered data from the UCI Machine Learning Repository, and then used some of the basic methods [10] shown in Figure 3.2. For balancing, the Resampling technique is used, in which variables and compared to find similarity and then tried to balance [10] them. Still if the result are not balanced same process is repeated, to get a balanced. During resampling, various samples in various ratios and attempted to balance the ratios; we repeat this process until the ratios are unbalanced. This balanced data [11-12] will be further used in the predictions. So, after resampling, we receive balanced data, and then we move on to the next step, which is training and testing.

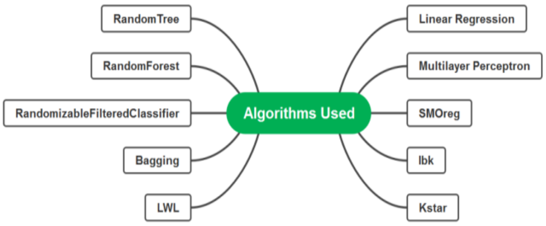

Fig.3. 2.Algorithms Proposed in Wholesale Predictive Model

In this section the unstructured imported data are processed to get it structured by balancing technique. To increase the accuracy now we used two techniques:-
1.Voting
2. Stacking
We were not getting the desired accuracy after training the model with the voting [13] classifier; instead, we were only getting accuracy up to 85%.
When it comes to the stacking classifier [14], we saw a little improvement in accuracy compared to cross-validation [15], up to 3% higher than the prior one.

- The accuracy of Cross-Validation in SMOreg was 90.46 percent.
- We achieve our desired accuracy of 93.49 percent after preprocessing [16] and utilizing the stacking classifier in Ensembled (SMOreg, K-Star).

Refer to Table 1. and Fig 3.3 for a comparison of Cross-validation and Stacking.

Table 1.Comparision of Accuracy for CV and Stacking

Classifier	Accuracy (CV)	Accuracy (STACKING)
Leaner Regression , Random tree	88.01	90.11
SMOreg , Random tree	83.36	90.45
SMOreg , Random Forest	90.46	93.49
SMOreg , Leaner Regression	86.32	90.44
SMOreg , SMOreg	87.22	90.47
SMOreg , Bagging	88.03	90.45
SMOreg , LWL	88.97	90.42
SMOreg , Kstar	85.48	90.41
SMOreg , Kstar , Leaner Regression	88.7	90.45

The feedback and analysis stage is now in progress. In this stage, the data analytics examine the performance of all applied models and provide the feedback and suggestions for updating the model and the data. This must be done up until the point at which we are not pleased with the outcome. Only then can the model be deployed.

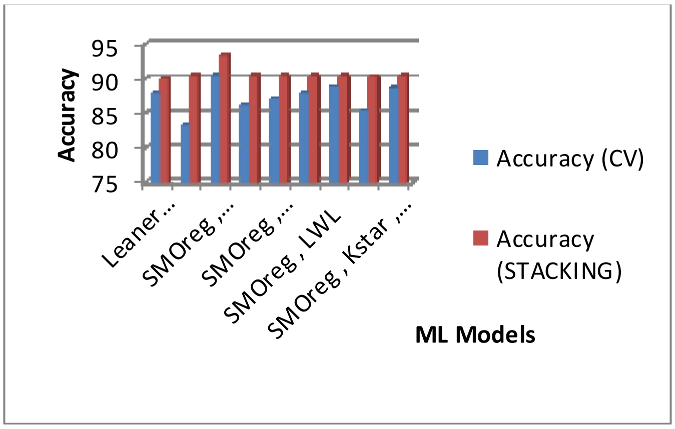

Fig.3. 3.Comparison Of Accuracy Of Cross-Validation And Stacking

IV. Result And Discussion

This section focuses on comparing the performance of frequent machine learning model with the proposed optimized ensemble model [17-18] by simulating the result from weka simulator, and the proposed model's performance is validated as the proposed model's accuracy is increased to 93.49 percent. The performance of the frequent machine learning model and the proposed model are compared in the graph below.

Table 2. Standard Performance Evaluation Parameter

Classification	Correlation Coefficient	Mean-Absolute Error	Root-Mean Squared Error	Accuracy	W-SAW Score
Linear Regression	0.36	11.99	27.66	88.01	22.95
Multilayer Perceptron	0.15	16.64	45.36	83.36	21.89
SMOreg	0.49	9.54	25.34	90.46	23.54
Kstar	0.41	12.78	25.93	87.22	22.82
LWL	0.14	11.97	33.45	88.03	22.94
Random Forest	0.36	11.3	26.57	88.7	23.10

Random Tree	0.16	14.93	40.77	85.07	22.27
Optimized Ensemble	0.62	6.51	23.14	93.49	26.54

A. CORRELATION COEFFICIENT

The correlation coefficient is a measure that can be used to determine the strength of a link between two given attributes.
Correlation Coefficient values typically range from -1 to +1. If the Correlation Coefficient in (1) of the proposed model is close to one, the association between the qualities is strong, and vice versa.

$$Cr = \frac{\sum\left(a_i - \bar{a}\right)\left(b_i - \bar{b}\right)}{\sqrt{\sum\left(a_i - \bar{a}\right)^2 \sum\left(b_i - \bar{b}\right)^2}} \quad (1)$$

Cr = Correlation Coefficient
a_i=Values of the a-variables in a sample
$\bar{a}$=Mean of the values of the a-variable
b_i=Values of the b-variables in a sample
$\bar{b}$=Mean of the values of the b-variable

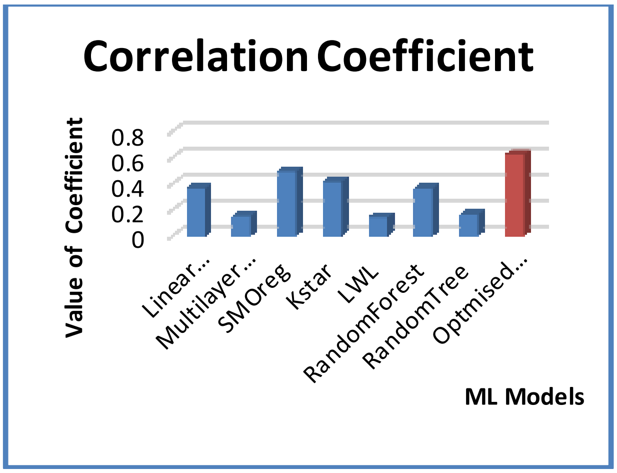

Fig.3. 4.Comparison Of Correlation Coefficient

We can clearly see that the correlation coefficient of our proposed model is high as compared to the frequent machine learning models in the simulated result. Refer to Table 2. and Fig 3.4 for a comparison of correlation coefficients

B. MEAN ABSOLUTE ERROR

MAE is the difference between the true value and the anticipated values from any proposed model, as the name implies. MAE in (2) is commonly used to assess the correctness of a regression problem.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|b_i - \hat{b}\right| \quad (2)$$

MAE = Mean Absolute Error
b_i =Values of the b-variable in a sample
$\hat{b}$ = Predicted value of b-variable
When the MAE value in an ML model is low, it is expected to be deemed the better model when compared to those with larger MAE values.

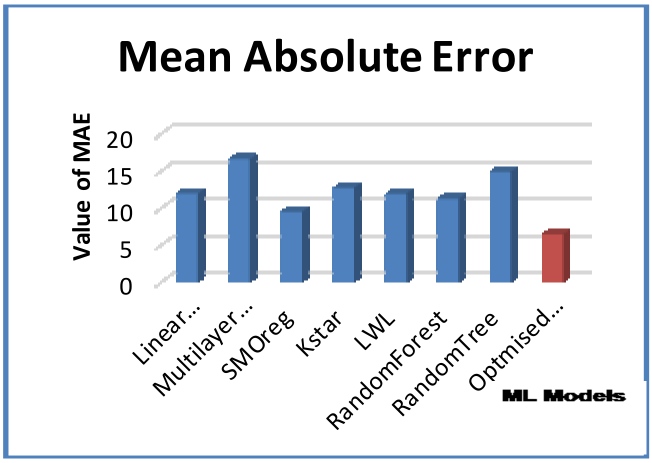

Fig.3. 5.Comparison Of Mean Absolute Error

Moreover the Mean Absolute Error of the proposed model is less as compared to the others in the simulated result. Refer to Table 2. and Fig. 3.5 for a comparison of Mean Absolute Error

C. ROOT MEAN SQUARED ERROR

The Root Mean Square Error can be used to assess the performance of any proposed model during the training and deployment phases. In general, RMSE in (3) is used to determine the accuracy of any forecast based on the input data. It also aids in determining whether the predictions are within the range of correct values measured.

The most common type of learning is supervised learning, and RMSE is often preferred.

The formula for calculating RMSE is as follows.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}\right)^2} \quad (3)$$

$RMSE$ = Root Mean Square Error
MSE = Mean Absolute Error

y_i = Values of the y-variables in a sample

$\hat{y}$ = Predicted value of y-variable

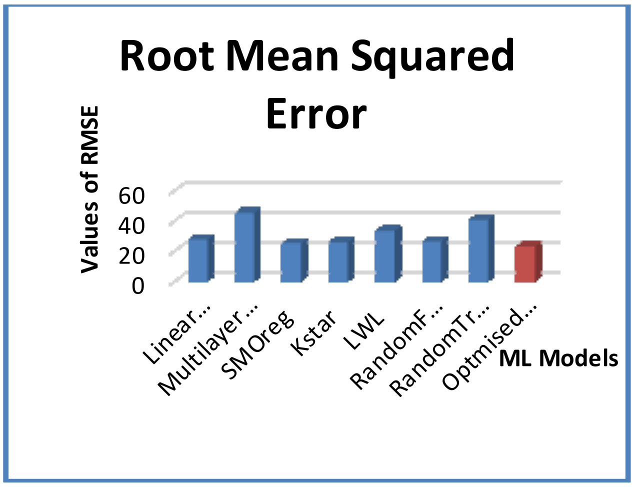

Fig.3. 6.Comparison Of Root Mean Absolute Error

Also the Root Mean Squared Error is less as of the others in the simulated result. Refer to Table 2. and Fig. 3.6 for a comparison Of Root Mean Absolute Error.

D. ACCURACY

The accuracy parameter determines whether the chosen model accurately predicts the true value. Furthermore, anyone can readily verify whether a certain ML model is good at creating and discovering patterns based on the given variables using accuracy.

The difference between MAE and "100" is used to calculate accuracy.

$$\text{Accuracy} = 100 - \text{MAE} \tag{4}$$

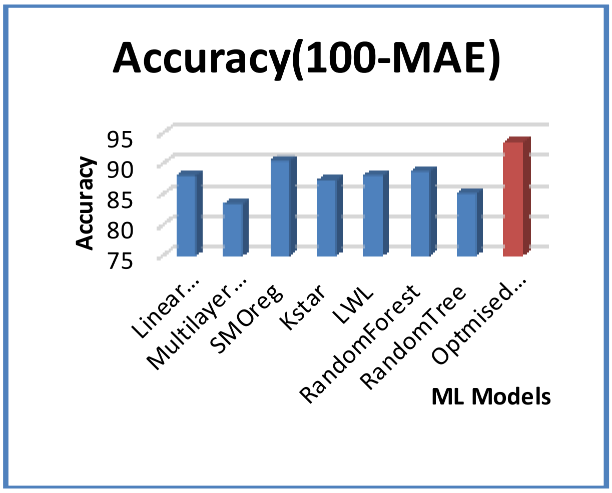

Fig.3. 7.Comparison Of Accuracy

Last but not least the Accuracy in (4) of the proposed model is the highest among the other frequent machine learning models in the simulated result. Refer to Table 2. and Fig. 3.7 for a comparison Of Accuracy.

E. W-SAW SCORE

W-SAW Score generally considers all the performance parameters so as to strengthen the selected model for deployment in the real life scenario. W-Saw Score in (5) of optimized ensemble model is the highest and it will strengthen the proposed model. It also provide novelty to our proposed model as we have used multidisciplinary operational research with computer science machine learning techniques also strengthen our model.

$$\text{W-Saw Score} = \frac{Cr+\sqrt{MAE}+RMSE+ACCURACY}{4} \tag{5}$$

Cr=Correlation Coefficient

MAE = Mean Absolute Error

$RMSE$ = Root Mean Square Error

Refer to fig 3.8 for the comparison of S-Saw Score.

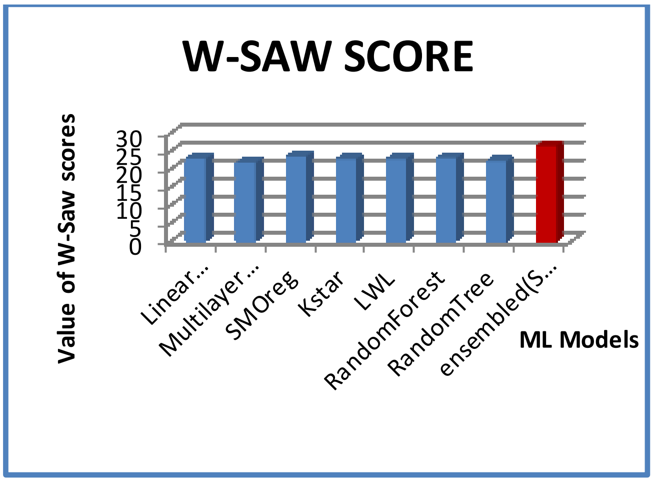

Fig.3. 8.Comparison of W-SAW Score

V. CONCLUDING REMARKS AND FUTURE SCOPE

Frequent machine learning models have room for improvement, thus we proposed optimized ensemble model, whose performance is significantly superior and the results are pretty satisfactory with this precision. Although this work has a satisfactory result but to collaborate with Multidisciplinary domain performance can be enhanced.

Deep learning algorithms can be used to give the strength to the proposed optimized ensemble model. one can use these algorithms to improve the model or develop a better predicting model.

REFERENCES

[1] Lakshmanarao, A., K. Chandra Sekhar, and Vijay Kumar. "Using Machine Learning Clustering Algorithms for Analysing Wholesale Customers Data."

[2] Zarnikau, Jay, Greg Landreth, Ian Hallett, and Subal C. Kumbhakar. "Industrial customer response to wholesale prices in the restructured Texas electricity market." Energy 32, no. 9 (2007): 1715-1723.

[3] Ray, Sourav, Haipeng Chen, Mark E. Bergen, and Daniel Levy. "Asymmetric wholesale pricing: theory and evidence." Marketing Science 25, no. 2 (2006): 131-154.

[4] Parvania, Masood, Mahmud Fotuhi-Firuzabad, and Mohammad Shahidehpour. "Optimal demand response aggregation in wholesale electricity markets." IEEE transactions on smart grid 4, no. 4 (2013): 1957-1965.

[5] Basheda, Gregory N. "Setting stranded costs for retail-turned wholesale customers: why FERC needs to change its approach." Utilities Policy 8, no. 2 (1999): 121-137.

[6] Dat, Nguyen Duy, and Hsing-Kuo Pao. "Exploring wholesale customer behavior via a garment outsourcing factory in Vietnam." In NCS 2019 全國計算機會議, pp. 311-316. 國立金門大學, 2019.

[7] Bangert, Patrick. "Client Segmentation in Wholesale Markets using Data Mining Client Segmentation in Wholesale Markets using Data Mining."

[8] Deepika Sharma at al &quot; deep neuro fuzzy approach for risk and severity prediction using recommendation system in connected health care&quot; pp. 2021/17.

[9] Lokesh pawar et al. &quot; Smart City IOT: smart architectural solution for networking, congestion and heterogeneity&quot; pp. 2019/5/15.

[10] Dinesh Kumar et al. &quot; Feature optimised machine learning framework for unbalanced bioassays&quot; pp. 2021/12/1.

[11] Lokesh pawar et al &quot; elevate primary tumor detection using machine learning&quot; pp. 2021/12/1.

[12] Pankaj rahi et al &quot; Air quality monitoring for smart e-heath system using firefly optimization and support vector machine pp. 2021/10.

[13] Lokesh pawar et al. &quot; Advanced ensemble machine learning model for balanced bioassays&quot; pp. 2021.

[14] Rekik, Yacine. "Inventory inaccuracies in the wholesale supply chain." International Journal of Production Economics 133, no. 1 (2011): 172-181.

[15] Chen, Lin, Guofang Nan, and Minqiang Li. "Wholesale pricing or agency pricing on online retail platforms: The effects of customer loyalty." International Journal of Electronic Commerce 22, no. 4 (2018): 576-608.

[16] Al Majali, Faris Odeh. "A conceptual framework for operational performance measurement in wholesale organisations." International Journal of Productivity and Performance Management (2022).

[17] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

[18] Andi, Hari Krishnan. "An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model." Journal of Soft Computing Paradigm 3, no. 3 (2021): 205-217.