SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

# SUPERMARKET SALES PREDICTION USING MACHINE LEARNING

**Chavali Saathvika Durga Abhinaya [1], Bellamkonda Lahari[2], Chinta Devika Priya[3], Devarapalli Anjali[4], Bathula Sri Navya[5], B. Sai Jyothi[6]**

[1]*B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*
[2]*B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*
[3]*B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*
[4]*B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*
[5]*B. Tech Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*
[6]*Professor Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur*

**Article DOI: https://doi.org/10.36713/epra14814**
**DOI No: 10.36713/epra14814**

**ABSTRACT**

*The huge supermarkets are more data-driven in today's retail world. These businesses tediously analyze sales data for each individual item they provide in order to optimize inventory management and predict managers demand. Using machine learning techniques, anomalies and patterns are being added to the data repository.*

*This data is used to forecast future sales volume, which is critical for merchants like supermarkets. We provide a prediction model, similar to supermarkets, that uses the capabilities of the XGBoost algorithm to forecast a company's sales. Our findings show that our suggested model exceeds existing models in terms of predicted accuracy, illustrating the power of complicated machine learning approaches in optimizing retail operations. This study provides useful information for improving sales forecasting and inventory management.*

**KEY WORDS:** Regression, Sales, Prediction, Data Exploration, Supermarkets, XGBoost.

## 1. INTRODUCTION

Today's board of supermarket, a large grocery chain with locations all over the nation, has issued a challenge to all data scientists to assist them in developing a model that can forecast the sales, per product, for each shop in order to provide accurate findings. Supermarket has gathered sales information from Kaggle for a variety of items across numerous retailers in several cities. The corporation expects that by providing us with this information, we will be able to identify the goods and retailers who are essential to their sales and utilize that knowledge to take the appropriate actions to assure the achievement of their business objective, which is to turn a profit for every supermarket. This is accomplished by selling more products and having a high turnover rate.

Here, jupyter Notebook is utilized as a tool and Python is used as a programming language. This application was created using machine learning components like the Supervised Learning task, There are regression tasks. The major reason for doing this is to forecast future retail sales for a corporation. Many techniques utilized include data collection and Feature engineering, data preprocessing, and model creation Evaluation.

Learning under supervision aids in comprehension of the data flow, understanding of sale pricing, etc. The Regression analysis use variety of techniques to forecast the retail costs. It has tasks like data cleansing, data transformation and visualizing XG Boost algorithms are employed.

In this study, we used the XG Boost approach to create a prediction model and test it on the Supermarket dataset for predicting sales of the product from the particular outlet.

## OBJECTIVES OUR WORK

1. Examine the items' prior sales data
2. Recognizing the elements that influence a product's sales
3. drawing conclusions about those sales

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

4. Computing future sales from the data and making predictions
5. Help for businesses in properly increasing or decreasing product inventories.

## 2. RELATED WORK

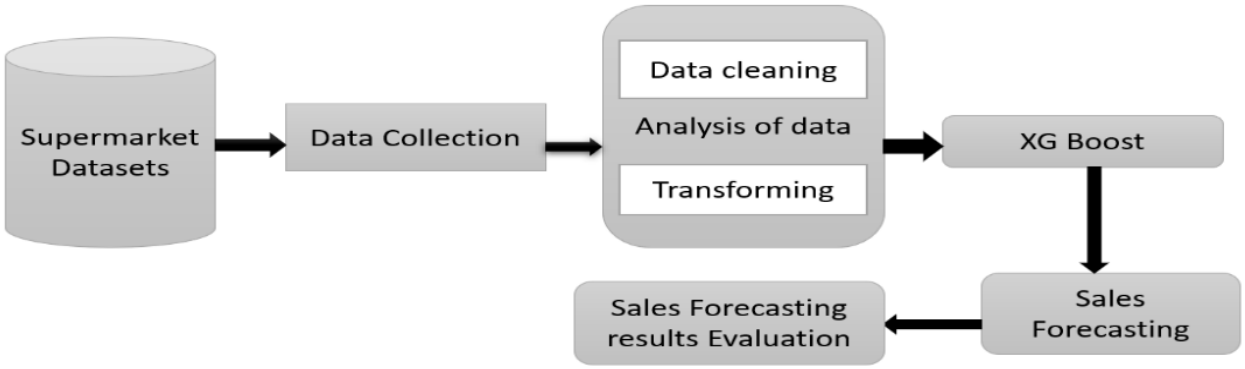
Numerous regression models are used to predict crime, health outcomes, home values, and sales, among other things. cardiovascular risk assessment using XGBoost. To forecast product sales, utilize sales forecasting. Being sold at several Big Mart Company shops. As the items are produced in greater quantity, and increasing regions are greater and more capable of being predicted by hand more challenging. Python is utilized as a programming language here. Jupyter Notebook is used as a tool and a language. In this application, supervised machine learning features Regression and learning functions are also employed. Here is mostly carried out to forecast the company's future revenue store merchandise

The different techniques include data processing, engineering features, model design, and testing. The regression function forecasts using a number of algorithms. prices. This requires labor for data identification, cleaning, and transformation. Profits generated by the business are Accurate sales projections are intimately related to supermarkets want a reliable forecasting method so that the there is no loss to the firm. Experiments confirm this. Our methods result in forecasts that are more accurate. Compared to alternative techniques like decision trees, local gatherings, etc.

## 3. PROPOSED MODEL

Description of the Supermarket. Sales dataset "SUPERMARKET" is the name of the dataset. Every dataset is made up of different properties. Item Outlet Sales is the response variable for these characteristics, while the other features are mostly utilized as predictor factors. This data collection includes diverse items from several cities.

**PROPOSED SYSTEM ARCHITECTURE**




### Advantages of proposed model

- Improved pricing Accuracy which helps supermarkets set competitive prices for their products, maximizing revenue and profit.
- Feature Importance: XGBoost provides insights into the most important features affecting pricing, helping supermarkets make data-driven decisions.
- XGBoost is optimized for performance, making it capable of real-time or near-real-time predictions, vital in a dynamic retail environment.
- XGBoost's flexibility and support for hyperparameter tuning allow for fine-tuning models to best fit the specific needs of a supermarket sales price prediction system.
- XGBoost is robust against outliers and can handle missing data, common challenges in supermarket datasets.
-

### 3.1 The Data

1. Item_Weight: This feature represents the weight of the item being sold. It's typically measured in units like kilograms or pounds.
2. Item_Fat_Content: This feature describes the fat content of the item. It has multiple categories like "Low Fat," "Regular," "LF," "low fat," and "reg." I'll need to preprocess this feature to ensure consistency.
3. Item_Visibility: This feature indicates how prominently the item is displayed in the store. It might be measured as a percentage or another numeric value.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

4. Item_Type: This feature categorizes the item into types such as "Baking Goods," "Dairy," "Frozen Foods," and so on.
5. Item_MRP: This is the Maximum Retail Price (MRP) of the item. It represents the highest price at which the item can be sold.
6. Outlet_Identifier:Each outlet has a unique identifier, and this feature represents that. Different outlets may have distinct characteristics.
7. Outlet_Establishment_Year: This feature represents the year when each outlet was established. It's important for understanding the age of the outlet.
8. Outlet_Size: This feature describes the size of the retail outlet, categorized as "High," "Medium," or "Small."
9. Outlet_Location_Type: It indicates the location of the outlet, such as "Tier 1," "Tier 2," or "Tier 3." These categories might signify different levels of urbanization or geographical areas.
10. Outlet_Type: This feature tells us the type of retail outlet, such as "Grocery Store" or different types of supermarkets.
11. Item_Outlet_Sales: This is the target variable I want to predict. It represents the sales of the item in the outlet and will be used to train and evaluate my predictive model.

The dataset contains a mix of numerical and categorical features, and performed preprocessing steps to handle missing values, one-hot encode categorical variables, and scale numerical variables.

### 3.2 Data Pre Processing

Handling missing values

Load and explore the provided dataset, including both training and testing data. During this investigation, missing values were identified in two key columns: Item_Weight and Outlet_Size. The following steps are taken to resolve this issue:

- Missing values in the "Item_Weight" column are filled with the average value of the column.
- For the "Outlet_Size" column, filled in the missing value with the mode.

This ensures that no values are missing from the records after this process.

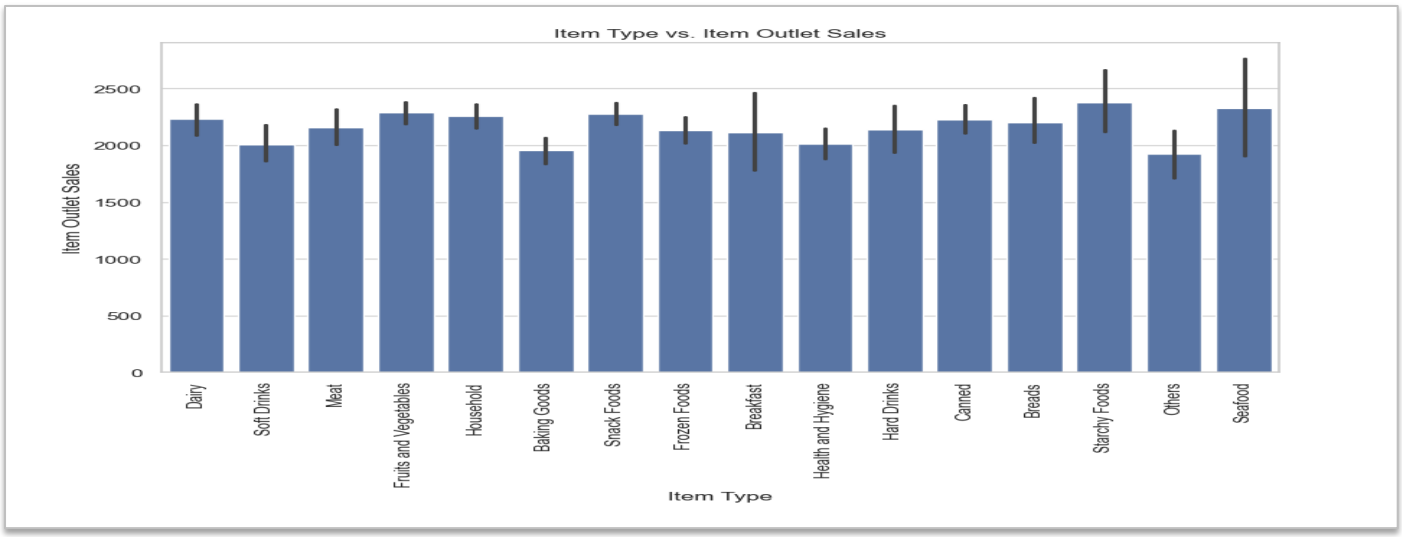
To gain a deeper understanding of the data, we conducted an exploratory data analysis. Here utilized various libraries, including ‘pandas-profiling`, `klib`, and `seaborn`, to:

- Visualize data distributions, correlations, and patterns.
- Uncover insights into the dataset's characteristics.

### 3.3 Feature Engineering

As part of feature engineering, here implemented the following actions:

- Dropped unnecessary columns, including 'Item_Identifier' and 'Outlet_Identifier.'
- Applied label encoding to convert categorical variables into numerical representations.
- Split the data into training and testing sets and standardized the features.




### 3.4 Model Building

For modeling, training of regression models, including XGBoost. And also tuned the hyperparameters for the XGboost model using grid search. To evaluate the models' performance, we employed metrics such as RMSE and R-squared.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

## 4.ALGORITHMS USED

### 4.1 Lasso Regression

The operator that selects the minimum absolute shrinkage rate is called an operator. The typical regression type of linear regression always assumes that there is a linear relationship between input and output variables. A famous linear regression with an L1 penalty is called lasso regression. This reduces the coefficients of input factors that are not useful for prediction. The L1 penalty allows some coefficient values to be zero, essentially removing input variables from the model and allowing automatic feature selection. The mathematical equation for Lasso regression is the degree of shrinkage, expressed as sum of squares + * (sum of absolute values of coefficient magnitudes) Lasso regression. $\lambda=0$ means that all features are considered, similar to linear regression where only sums of squares are considered to create the model. $\lambda = \infty$ means no features are considered. It refers to infinity and excludes other characteristics. As $\lambda$ increases, the deviation also increases. As $\lambda$ decreases, the variance increases. Linear regression refers to a model that assumes a linear relationship between the input variable and the target variable.

### 4.2 Ridge Regression

A common regression technique for estimating the outcome of an equation using any unique solution is ridge regression. This is a common problem in machine learning difficulty of selecting "required" answers.

There is little data. Ridge regression is a well-known and widely used modeling approach that is a variation of linear regression. However, ridge regression stands out because it addresses one of the major problems: multicollinearity.

Traditional linear regression. When there are many independent factors such as seasonal trends or promotions Multicollinearity often occurs in supermarket sales forecasts because area demographics are interrelated. This can lead to irregular and unreliable regression results. Features of ridge regression Managing multicollinearity proves to be a very useful tool in this situation. The custom matrix contains three data sets created from your data. One is the training data, the second is the valid data set, and the third is the test data. The model is trained using the training set you can use the model to provide results. The test data set is ML algorithms.

### 4.3 XGBoost Algorithm

Regardless of the type of prediction task, such as regression or classification, XGboost is one of the most widely used and accurate machine learning algorithms today. This is a competitive implementation of gradient boosting decision trees for machine learning, designed for performance and speed. It is well known that this method produces better results than other machine learning algorithms. Since its inception, it has become a truly "state-of-the-art" machine learning technique for processing well-structured data. A distributed gradient boosting library. This is a software library that you can obtain from the Internet and install and use on your computer.

XGBoost (short for Extreme Gradient Boosting) is a cutting-edge machine learning algorithm that has gained immense popularity and recognition for its superior predictive capabilities. It is known for efficiently processing complex and diverse datasets, making it ideal for supermarket sales forecasting. The goal of this research is to use XGBoost to create a reliable and accurate model for predicting sales in the food industry. Like any other retail industry, supermarkets suffer from various problems that negatively impact sales. Seasonality, geography, marketing, and many other factors come into play. As an ensemble learning method, XGBoost is well suited to address such problems. It is extremely adept at managing both organized and unstructured data, successfully identifying subtle relationships and patterns that contradict traditional linear models. This research attempts to use XGBoost to develop a predictive model that can predict product sales across multiple supermarkets in the future. It could improve retailers' ability to make data-driven decisions, effectively manage inventory, and improve overall performance. The success of this project will not only help retailers, but also serve as an example of the breakthrough potential of cutting-edge machine learning algorithms in tackling difficult real-world problems. We explore the intricacies of XGBoost, its capabilities, and its potential to transform grocery sales forecasts in the process. This project shows how XGBoost can revolutionize retail by enabling data-driven decision-making that supports supermarket performance and sustainability.

## 5.RESULTS

The results of the various models will be presented. The results were obtained by appling various models like lasso regression, Ridge regression, xgboost on supermarket training and testing data.

### 5.1 Performance Metric

Use the mean absolute error (MAE) when evaluating the model. This means that the lower the MAE,A better model.
The choice of performance metrics is based on the fact that the task is a regression task, similar to MAE.
Tested and reliable metrics that provide a good measure of model performance.

#### 5.1.1 Average Absolute Error

The mean absolute error (MAE) is defined as it is the measure of the difference between two continuous variables. Assume

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

X and Y are variables
From the observations, X is the known value and Y is the predicted value of the machine learning model.
The mean absolute error (MAE) is the average vertical distance between each observed and predicted point.
To calculating MAE the below formulae is used
MAE sum i = 1 to n |y_{i} - x_{i}|/n

**5.1.2 RMSE**
Root mean square error, also known as RMSE, is a commonly used statistic to assess accuracy.
Predictive models (such as regression models). You can estimate how well the model's predictions match the observed values.
Analyze the data by quantifying the average size of the error between expected and actual values. Improved model fit
The impact on the data is indicated by the reduced RMSE. The RMSE formula is as follows:
RMSE=(√Σlly(i)-(i)||² N)/N
Where,

- n refers to total number of data points or observations.
- yi represents the actual or observed values in the dataset.
- y^i represents the predicted values generated by the model for the corresponding observations.
- ∑ denotes the summation of the squared differences between actual and predicted values.
- Finally, the entire expression

**5.1.3 R-square method**
R-squared (R2) is a statistical measure of the proportion of variance in a dependent variable.
explained by the independent variables in the regression model. Correlation describes the strength of the relationship between independent and dependent variables;
R-squared describes the extent to which the variance in one variable explains the variance in a second variable. So, if
If the model's R2 is 0.50, approximately half of the observed variation can be explained by the model inputs.
The formula for calculating R-squared is:
R squared = 1-(SSR/SST)

- The R2 value can range from 0 to 1.

The meanings of the various R-squared values are as follows:

- R2=1 The model perfectly explains the variation in the data.
- R2=0 model does not explain variation in the data.
- 0some of the variation in the data, and higher values indicate better fit.

It is important to note that R-square provides information about goodness of fit, but it does not necessarily indicate goodness of fit. The overall quality of the model, or its ability to make accurate predictions.

**Error Measurements &R-Squared:**
In the below table the RMSE and R-squared results are shown respectively.We observe that the XGboost algorithm does best among all three with a R-squared 0.608451. The lasso model has a close R_squared to the ridge but with a much lower RSME

| Algorithms | RSME | R-squared |
|---|---|---|
| XGBoost | 1031.6085175933238 | 0.60845185009 |
| Lasso Regression | 1207.2491022080023 | 0.46377245678 |
| Ridge Regression | 1209.3436327744663 | 0.46190597103 |

## 6. CONCLUSION
This project explains the fundamentals of machine learning, along with the related data processing and modeling methods, and applies them to forecasting sales of various supermarkets products. The many factors taken into account like the location with the highest sales was medium-sized, proposing that other stores should do the same comparable trends to boost sales. Many occurrence parameters and several other elements can be utilized for More successfully and innovatively anticipating the sales.

In prediction systems, accuracy is crucial and can include increased greatly when the parameters employed are increased. Additionally, how the sub-models function might result in increasing the system's productivity

Since the accuracy of the sales estimates directly relates to the profit made, the big stores strive to make accurate predictions to prevent losses for the business.

In this study, we developed a model using the Xgboost method tested with it on lasso regression, ridge regression, and other data. The supermarket sales dataset for estimating the product's sales of a certain outlet. Experiments confirm that our approach i.e is

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

xgboost results in more accurate predictions than compared to alternative methods.

## 7.REFERENCES

1. *Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retails sales forecasting", Int. Journal Production Economics, vol. 86, pp. 217- 231, 2003.*
2. *Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2.*
3. *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 41*
4. *Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.*
5. *https://halobi.com/blog/sales-forecasting-five-uses/.*
6. *Zone-Ching Lin, Wen-Jang Wu, "Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone", IEEE Trans. On Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999.*
7. *O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.*
8. *C. Saunders, A. Gammerman and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", Proc. of Int. Conf. on Machine Learning, pp. 515 – 521, July 1998.IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 7, JULY 2010 3561.*
9. *"Robust Regression and Lasso". Huan Xu, Constantine Caramanis, Member, IEEE, and Shie Mannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration."An improved Adaboost algorithm based on uncertain functions".Shu Xinqing School of Automation Wuhan University of Technology.Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China. 42*
10. *Xinqing Shu, Pan Wang, "An Improved Adaboost Algorithm based on Uncertain Functions", Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.*
11. *A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.*
12. *N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Production Economics 170 (2015) 321-335P*
13. *D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.*
14. *X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science 17 ( 2013) 1055–1062.*
15. *E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: a Case study of the printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.*
16. *P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. GarciaSanchez, A.J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, Knowledge-Based Systems 115 (2017) 133-151.*
17. *R. Majhi, G. Panda and G. Sahoo, "Development and performance evaluation of FLANN based model for forecasting of stock markets". Expert Systems with Applications, vol. 36, issue 3, part 2, pp. 6800-6808, April 2009.*
18. *Pei Chann Chang and Yen-Wen Wang, "Fuzzy Delphi and back propagation model for sales forecasting in PCB industry", Expert systems with applications, vol. 30,pp. 715-726, 2006.*
19. *R. J. Kuo, Tung Lai HU and Zhen Yao Chen "application of radial basis function neural networks for sales forecasting", Proc. Of Int. Asian Conference on Informatics in control, automation, and robotics, pp. 325- 328, 2009.*
20. *R. Majhi, G. Panda, G. Sahoo, and A. Panda, "On the development of Improved Adaptive Models for Efficient Prediction of Stock Indices using Clonal-PSO (CPSO) and PSO Techniques", International Journal of Business Forecasting and Market Intelligence, vol. 1, no. 1, pp.50-67, 2008. 44*
21. *Suresh K and Praveen O, "Extracting of Patterns Using Mining Methods Over Damped Window," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 235-241, DOI: 10.1109/ICIRCA48905.2020.9182893.*
22. *Shobha Rani, N., Kavyashree, S., & Harshitha, R. (2020). Object Detection in Natural Scene Images Using Thresholding Techniques. Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, Iciccs, 509–515.*
23. *https://www.kaggle.com/brijbhushannanda1979/bigmartsalesdata.*