JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

SICAIET
2024

Research Article
Collection: Artificial Intelligence and Emerging Technologies

Market Basket Analysis for Next Basket Item Prediction Using Data Mining and Machine Learning

Mubasher H. Malik[1*], Hamid Ghous[2], Maryem Ismail[2], Sana Jamshaid[3], and Javeria Altaf[3]

[1]Vision, Linguistics and Machine Intelligence Research Lab, Multan, Pakistan.
[2]Department of Computer Science, Institute of Southern Punjab, Multan, Pakistan.
[3]Department of Information Technology, Institute of Southern Punjab, Multan, Pakistan.
[*]Corresponding Author: Mubasher H. Malik. Email: mubasher@isp.edu.pk

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: Data Mining is one of the morst prominent approach used nowadays to identify sales patterns and features from large scale datasets. The primary objective of this research is to develop a model based on advaned Market Basket Analysis (MBA) to increase the sales of any orgnaization. This research focused on adoption of FP-Grwoth algorithm and Machine Learning (ML) algorithms to predict next item basket. Two widely used datasets French Retail Store Dataset (FRSD) and Bread Basket Dataset (BBD) were used for experiments. Experiments showed that FP-Growth algorithms produced most frequent items purchased by customers while ML classifiers such as Logistics Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN) and Decision Tree (DT) showed promising results. Among these ML classifiers RF produced promising accuracy of 0.922% and 0.930% using FRSC and BBD respectively. The proposed model has ability to predict next item basket. This model will help organization to increase their sales.

Keywords: Data Mining; Market Basket Analysis; Association Rules; Machine Learning; Marketing.

1. Introduction

Market Basket Analysis is one of the data mining relevant area that is becoming more widely used in a variety of industries (MBA). Relationships and rules are discovered in MBA by analyzing customer, product, and sales data in a retail store. MBA increases company profits by determining a product's sales relationship with another product and determining the rules of association (Fernandes, 2020). Market basket analysis is centered on the baskets of customers, and it is used to monitor buying patterns and improve customer service in order to improve. Retailers are interested in customer purchasing habits in order to keep up with changing consumer demands and maintain market stability. As a range of products are launched to suit future demands, customers may purchase a diverse collection of items in a single shop visit. Due to the huge number of goods and consumers, the traditional approach takes longer to discover buying behavior. As a result, data mining techniques for market basket analysis are required the data mining methods used to identify common patterns for merchants are association rules. Using a transactional dataset, a market basket analysis was conducted to detect buying behavior. The transactional dataset is made up of transaction IDs and a list of goods bought. The likelihood of one item being associated with another in a transactional collection is increased. Clean and standardized data is examined, and preparation methods are used. Following cleaning, an association rule mining method was used to create common item groups and association rules in order to discover customer purchase patterns (Ilham, 2018).

Data mining extracts ambiguous, previously unknown but valuable information from large amounts of data. Despite the fact that the existence of association rules in a database is apparent from a data mining standpoint, extracting them is a challenging job. It enables for the discovery and summarization of different associations after it has been extracted. Due to the enormous amount of data generated by the retail industry's fast growth, academics are focusing on machine learning techniques to offer novel online and offline shopping experiences. Next basket suggestions, sales forecast, customer segmentation, and churn prediction utilizing physical and online retail information are examples of machine learning applications in MBA. Artificial neural networks were used in machine learning techniques to mimic the organization of the human brain. These neural networks are capable of learning from unlabeled input without the need for human involvement. Layers, a learning algorithm, and an activation function make up a neural network.

The next section showed contribution of different researchers in the domain of MBA predictions.

2. Literature Review

In 2023, a prediction model for next market basket was proposed. Gated Recurrent Unit (GRU) network was used to predict next market basket based on customer purchase habit. The proposed model captures customers purchase behaviors, frequent product co-occurrence in shopping basket, purchase behavior recurrence and dynamic customer purchase taste. The model outperformed on real-life dataset and produced sophisticated prediction results (Van Maasakkers, 2023). In the same year, Augmented Reality (AR) based framework was proposed to predict next market basket. InstaCart dataset was used for experiments. CNN, LSTM, Bi-LSTM and CNN-BiLSTM was used for classification. LSTM produced prominent results and predict next basket successfully (Ghous, 2023). Furthermore, a novel Multi-aspect Neural Recommendation Model was proposed in 2023 by Deng. The proposed model learns from user interests and assign weights to each item from different aspects. The model captures items correlations at low level based on different aspects both customer side and item side. Real time datasets was used to test the proposed model and it produced promising results (Deng, 2023). Another recommendation based model was proposed in 2023. The proposed model recommends bet product to new customers based on historical sales data. MBA algorithm was adopted for recommendation analysis. The proposed model also used FP-Growth, W-apriori and W-Tertius algorithms to produce better results. Experiments shows that MBA algorithm produced confidence value of 67.6% while FP-Growth, W-Apriori and W-TErtius algorithms produced 70.2%, 63.0% and 96.2% respectively (Khadapi, 2023).

In 2022, K-Mean Clustering based model was proposed to predict customer purchase behaviors. The model is based on the analysis of recurrence of purchase behavior, sales frequency of purchases and monetary. The proposed model used cluster of customers to predict purchase behaviors. The proposed model showed promising results (Anitha, 2022). In 2021, proposed a prediction method for next market basket based of customer purchase habits . The proposed method observed customer personnel purchase history as well as purchase history of other customers. Instacart dataset was used to test the proposed method. The results shows that methods can work on small sample size and could not performed well with large record of customers (Maradin, 2021). In 2020, an apriori algorithm was used to identify customers purchase patterns. Frequent items were generated using apriori algorithm using transactional sales data. The experiments showed that the proposed model provides poor results on big dataset while it produces better results on small dataset (Efrat, 2020). In the same year, a data mining based model was proposed to predict customers purchase demands. Real world dataset was taken from Amazon. After preprocessing ML algorithms were used to predict online manufactured products to customers. The proposed model produced promising results (Van Nguyen, 2020). Another product recommendation model was proposed in 2020. The proposed model used Word2Vec encoding technique. XGboost algorithm was used for classification with collaborative filtering technique. The proposed model produced promising results (Shahbazi, 2020). Similarly, a ML based prediction model was proposed using large transactional dataset consist of customers purchase features. Logistic Regress (LR), Extreme Learning Machine (ELM) and Gradient Tree Boosting (GTB) was used for prediction. Experiments showed that Gradient Tree Boosting (GTB) showed promising results (Mart{\'\i}nez, 2020).

In 2019, an Apriori algorithm was used to predict customer purchase patterns. Association rules was also used in this proposed model. The proposed model used Bayesian Algorithm and Decision Tree (DT) algorithm for classification. The model showed that DT produced promising results (Wang L. a., 2019). In 2018, proposed a model based on Temporal Annotated Recurring Sequence (TARS). In his proposed prediction model TARS captures next market basket based on customers decisions factors such as co-occurrence, sub sequentially periodically and recurrency of items purchased. The experimental results showed that TARS could predict customers purchase behaviors (Guidotti, 2018). Similarly, a Neural Network based MBA model was proposed. Feed Forward Neural Network (FFNN) was used to predict customer purchase behavior. The model produced promising results (Gangurde, 2018).

In next section, proposed model along with methods and technique presented in detail. The proposed model takes transactional dataset and after preprocessing FP-Growth method was used for features extractions. Association rules were applied on the extracted features. ML classifiers such as LR, RF, KNN and DT deployed to predict next basket item.

3. Methods and Materials

The purpose is to mention the methodology used for the completion of this thesis and the techniques to satisfy the objectives of this thesis (Biggam, 2018). In today's customer-centric marketplaces, businesses must develop appropriate and low-cost advertising methods that can respond to changes in consumer perceptions and product demands (Gupta, 2021). This was a major problem to be solved which can be done using association rules and machine learning approaches. (Kumar, 2022) From the literature review, several techniques and methodologies came into view and from them, we have adopted our methodology to complete this study. Next Basket Item Prediction Model as shown in Figure 1. The proposed model

3.1. Preprocessing

For improved accuracy findings, preprocessing is an important step in this methodology (Sun, 2022) since it transforms data into a meaningful form by using cleaning, reduction strategies to eliminate missing values, normalizing data form, and reduction techniques to reduce data dimensions.

3.2. Data Mining

Data mining is a predictive modeling technique used to find out hidden patterns or meaningful information(knowledge) from bulky amounts of data. (Wu, 2021).

3.3. FP-Growth Method

A FP-tree is a data structure that represents a collection of tree-shaped records in a compact format. Every transaction is read and sorted according to an FP-tree route (Patel, 2022). This will apply until all transactions have been read out. Because various transactions with similar subsets overlap the routes, the tree stays compact. The primary execution mechanism is the FP-Growth algorithm; (Ahmad, 2023).

1. Scan the database for items equal to hand over the threshold value. (Kim, 2023)
2. Support values for particular goods are shown in descending order of size (Big value to Small value).
3. It then creates a root-only tree

3.4. Association Rules

One of the most significant Data Mining methods used in Market Basket Analysis is the association rule (Sjarif, 2021). The phrase "what goes with what" is linked to the association rule. (Shmueli, 2023) Customers' purchases of goods at a Supermarket are referred to as 'Transactions.' (Bachtiar, 2023) The presence of three factors, namely support, confidence, and lift, may be used to calculate the magnitude of an associative rule (Mishra, 2009).

3.4. Support

Support refers the frequency of buying products within the transactions. This is the percentage of transactions that consist on the list of item set. (Abdulsalam, 2014). Table 1 showed the customer purchase transaction record and Table 2 shows frequency of items purchased.

Table 1. Customer Purchase Transactions Record

Transaction Order	Items
TD1	Apple, Egg, Milk
TD2	Carrot, Milk
TD3	Apple, Egg, Carrot
TD4	Apple, Egg
TD5	Apple, Carrot

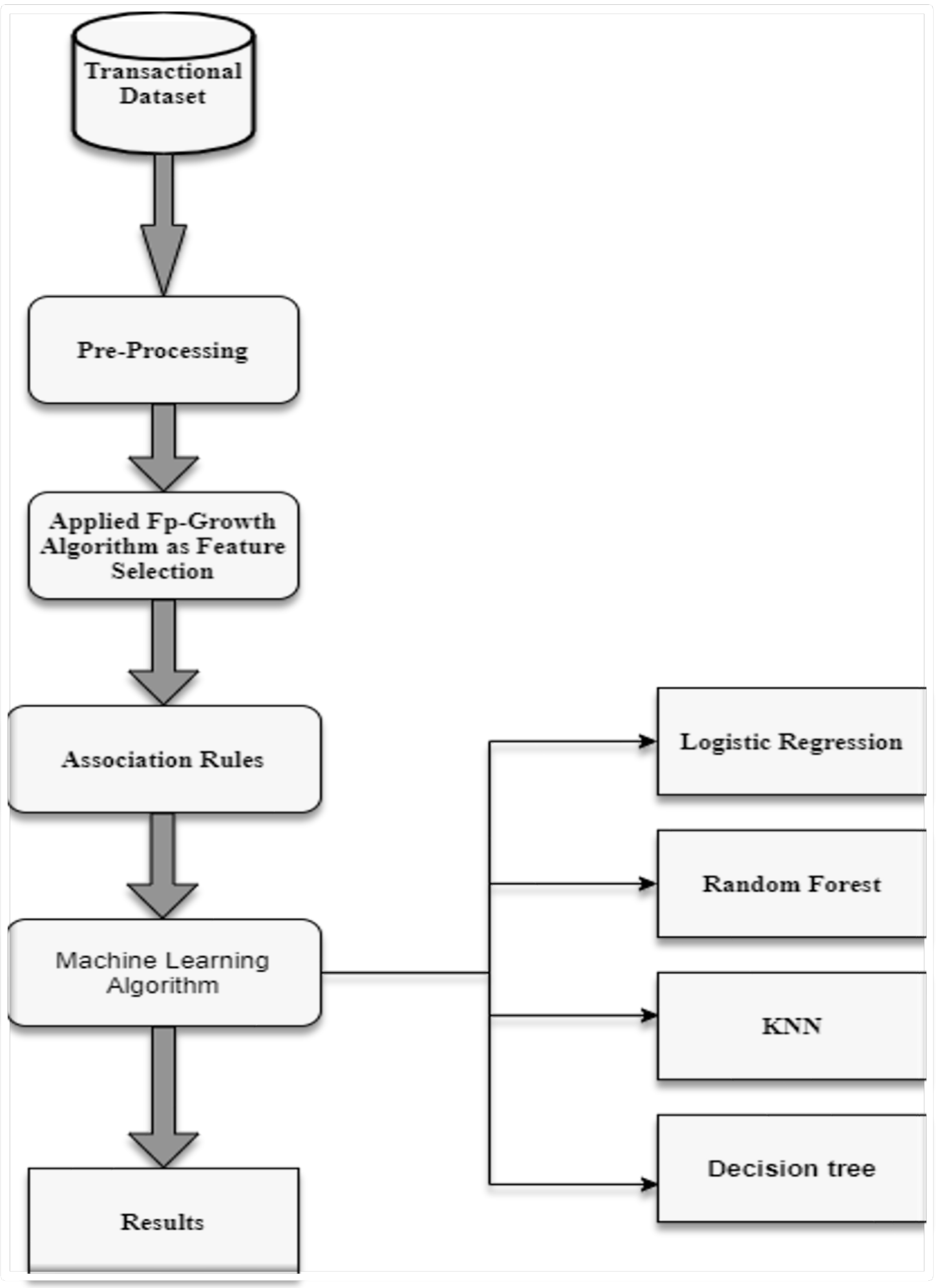

Figure 1. Next Basket Item Prediction Model

Table 2. Purchase Frequency of Items

Product s	Items Frequency
Apple	4
Egg	3
Milk	2
Carrot	2

In transactions two items Carrot and Milk are discarded as these two items do not meet predefined threshold value. In second rotation assemble item set of size 2 so [apple, egg] occurrence count 3 times remain in frequent item set list and further no more item add. Equation 1 depicted how support calculated.

$$\text{Support} = (AB) = \sum(A \cup B)|N \quad (1)$$

Support of[Egg, Apple] = 3/5 or 60%. The minimum support threshold value essential by FP-growth can be set based on information of our domain. (Shabtay, 2021).

3.5. Confidence

Above mentioned two products, B and A, calculate value of confidence by measuring the number of times that product A purchased with product B. Equation 2 and Equation 3 showed Confidence A → B and B → A respectively.

Confidence(A B) = Support → (AUB) / Support(A) (2)

Confidence (B A) = Support → (AUB) / Support (B) (3)

As given the above mentioned example, number of times Apple is purchased so that Egg was purchased calculated the value of

Confidence (Apple Egg) → Support (Apple, Egg) / Support(Apple)

Confidence (Apple (3/5) / → (4/5) = 0.75

By calculating support and confidence value we see that transactions orders contains apple also contains egg.

3.6. Classification

Machine learning is a division of Artificial intelligence that able machine to be trained from data and earlier results. (Tyagi, 2022) Machine learning is study of algorithms to check out improvement after training or knowledge. (Sarker, 2021)This is also called as predictive analytics. Machine learning algorithms used in this research study to perform predictions on selected rules. (Ofori, 2020). ML classifiers such as LR, RF, KNN and DT deployed to predict next basket item.

3.6.1. Logistic Regression Algorithm

Logistic regression is statistical analysis algorithm of machine learning to predicts or classify data consist on prior observations or knowledge. (Celine, 2020) There are three types of logistic regression one is binary logistic regression in which dependent variable would be binary such as 0,1 e.g. spam=1 or not spam=0 (Mrisho, 2021)Third type is Ordinal logistic regression analysis with more than three categories with orders. Equation 4 illustrated LR algorithm. The graphical representation of LR algorithm is shown in Figure

$$p = e^{a+bx/1+e^{a+bx........eq(1)}} \quad (4)$$

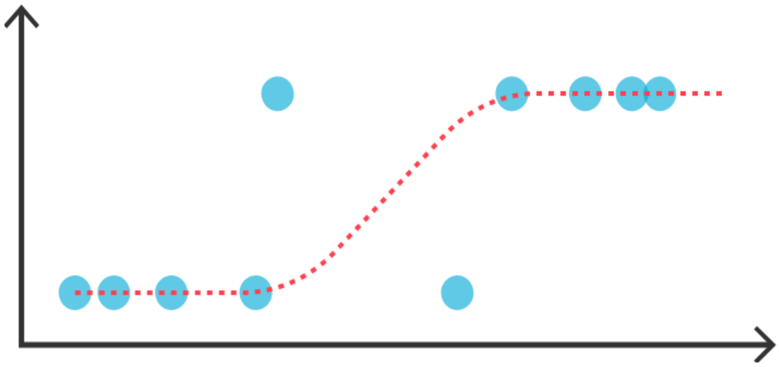

Figure 2. Graphical Representation of Logistic Regression Algorithm

3.6.2.Random Forest Algorithm

Random forest is a supervised machine learning algorithm that used to classify data or solved regression problems (Shah K. a., 2020). Equation 5 illustrated RF algorithm.

$$\text{Regression} = :f_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (5)$$

3.6.3.K-Neareast Neighbor Algorithm

KNN algorithm is a supervised machine learning algorithm used to predict data. It predicts the new dataset or case on the basis of assuming similarities of available dataset categories (Ali, 2021). Equation 6 illustrated RF algorithm.

$$\sqrt{(x_{2}^{-}x_1)} + (y_2 - y_1)^2 \tag{6}$$

3.6.4.Decision Tree Algorithm

Decision tree is a supervised algorithm of machine learning used for classification and regression problems (Charbuty, Classification based on decision tree algorithm for machine learning, 2021). To solve classification problems decision tree generates tree shape structure in keeping with target variable in data (Bertsimas, 2017).

The next section shows the experimental results and graphical representation of results produced by ML algorithms for next basket item prediction.

4. Results

The experiments performed to predict next market basket. Two datasets were taken for experiments. The first dataset is French Retail Store Dataset (FRSD) and second is Bread Basket Dataset (BBD) (Hossain, 2019). Both datasets are available at Kaggle. Firstly, FRSD was used for feature selection. Seven Thousand Five Hundred and One transactions were selected. These transactions contained one hundred twenty one items. The items were reduced up to Fifty to perform experiments. FP-Growth algorithms deployed with support 0.02% threshold value to extract frequent items. After that Eighty Seven frequent items were extracted. Following Table 3 show the list of frequent items with support at 0.02% threshold value.

Table 3. List of Frequent Items from FRSD Dataset

Sr.No.	Support 0.02%	Frequent Items	Item Length
1	0.238368	Mineral water	1
2	0.132116	Green tea	1
3	0.076523	Low fat yogurt	1
4	0.071457	Shrimp	1
5	0.065858	Olive oil	1
6	0.027463	mineral water, cake	2

Following Table 4 showed the accuracy of ML classifiers. 70:30 Train-Test ratio adopted. This Train-Test ratio produced promising results. LR algorithm produced 0.919% accuracy. While Random Forest showed highest accuracy of 0.922%. On the other hands, KNN and DT produced 0.810% and 0.782% accuracy respectively.

Table 4. Evaluation Results of ML classifiers using Accuracy

Sr.No.	Classifier	Accuracy (%)
1	Logistic Regression	0.919
2	Random Forest	0.922
3	K-Nearest Neighbors	0.810
4	Decision Tree	0.782

Figure 3 showed the graphical representation of ML classifiers evaluation results. Experiment performed using BBD. The BBD was used for feature selection. Twenty Thousand Five Hundred Seven transactions were selected. These transactions contained one hundred twenty items. FP-Growth algorithms deployed with support 0.01% threshold value to extract frequent items. After that Ninty Four frequent items were extracted. Following Table 5 show the list of frequent items with support at 0.01% threshold value.

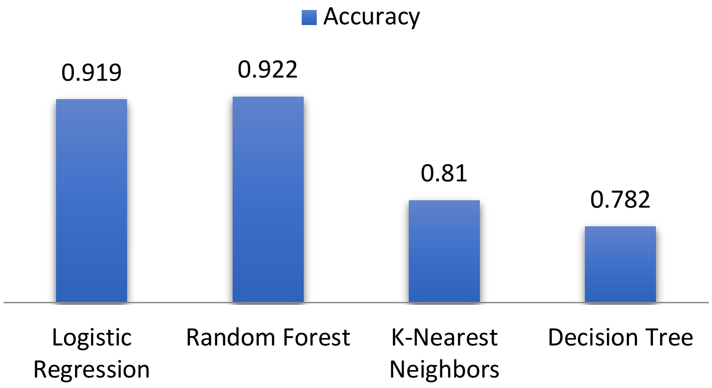

Figure 3. Visual Representation of Training at Ratio 70:30.

Table 5. List of Frequent Items from BBD Dataset

Sr.No	Support 0.01%	Frequent Items	Length
1	0.327205	Bread	1
2	0.029054	Scandinavian	1
3	0.058320	Hot chocolate	1
4	0.054411	Cookies	1
5	0.018067	Scone, Coffee	2
6	0.010882	Spanish Brunch, Coffee	2

Following Table 6 showed the accuracy of ML classifiers. 70:30 Train-Test ratio adopted. This Train-Test ratio produced promising results. LR algorithm produced 0.919% accuracy. While Random Forest showed highest accuracy of 0.930%. On the other hands, KNN and DT produced 0.850% and 0.781% accuracy respectively.

Table 6. Evaluation Results of ML classifiers using Accuracy

Sr.No.	Classifier	Accuracy (%)
1	Logistic Regression	0.919
2	Random Forest	0.930
3	K-Nearest Neighbors	0.850
4	Decision Tree	0.781

Figure 4 showed the graphical representation of ML classifiers evaluation results.

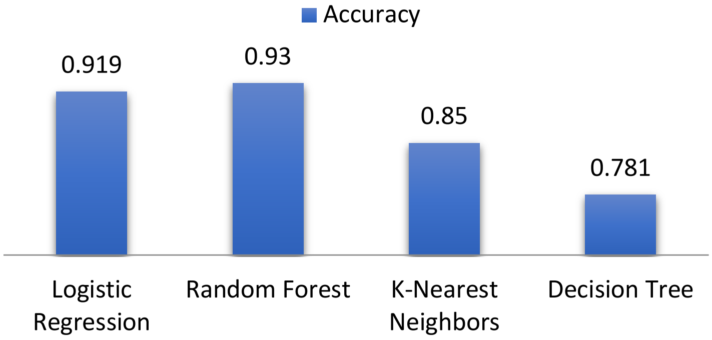

Figure 4. Visual Representation of Training at Ratio 70:30.

The experiments conducted on both dataset named FRSD and BBD. FP-Growth algorithm deployed on both datasets using 0.02% and 0.01% support value. This experiments extracted most frequent items as shown in Table 3 and Table 5. Furthermore, the experiments extended using ML classifiers such as LR, RF, KNN and DT. The ML classifiers deployed on both datasets. RF produced promising accuracy of 0.922% and 0.930% using FRSC and BBD respectively. This model can easily predict next item basket using Data Mining and ML algorithms.

5. Conclusions

Customer purchasing patterns can be observed using MBA approaches. These approaches helps us to identify significant connections with the items putted into shopping basket This may help the decision makers to take decision. This may also increase the sales of any organization. The major focus of this research was to predict next item basket. Two most popular datasets FRSC and BBD were used to perform experiments. FP-Growth algorithm was used to identify most frequent items from the basket while ML classifiers RF, LR, KNN and DT deployed on both datasets. The experimental results showed that proposed approach can easily identify next item basket using Data Mining and ML algorithms. This will help the decision makers to take certain decisions which may increase the sales of any organization. In Future the model will be improved using Deep Learning and Transfer Learning methods to predict next item basket. More popular datasets will be adopted for experiments.

References

1. Abdulsalam, S. a. (2014). Data mining in market basket transaction: An association rule mining approach. International Journal of Applied Information Systems (IJAIS), 7, 15--20.
2. Adadi, A. (2021). A survey on data-efficient algorithms in big data era. Journal of Big Data, 8, 24.
3. Agarwal, M. S. (2020). Tomato leaf disease detection using convolution neural network. Procedia Computer Science. 293-301.
4. Ahmad, S. S. (2023). Hybrid recommender system for mental illness detection in social media using deep learning techniques}. Computational Intelligence and Neuroscience, 2023.
5. Ali, M. M. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. omputers in Biology and Medicine, 136, 104672.
6. Amjad, K. &. ((2021)). Critical review on multi-crops leaves disease detection using artificial intelligence methods. Int J Sci Eng Res, 12, 2.
7. Anaissi, A. a. (2015). SVM-based association rules for knowledge discovery and classification. 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)},.
8. Anitha, P. a. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Information Sciences, 34, 1785--1792.
9. ansal, M. a. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.
10. Azmi, S. S. (2020). An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. Int. Res. J. Eng. Technol, 7.
11. Bachtiar, A. A. (2023). Analysis of the Effect of Usability, Trust, and Risk Perception on Buying Interest in the Purchase Category of Mobile and Tablet Products through the ABC Application. European Journal of Business and Management Research, 8, 40--46.
12. Batmavady, S. &. (2019). Detection of cotton leaf diseases using image processing. Int J Rec Technol Eng (IJRTE), 8.
13. Bertsimas, D. a. (2017). Optimal classification trees. Machine Learning, 106, 1039--1082.
14. Bharadiya, J. P. (2023). A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics. American Journal of Artificial Intelligence}, 7, 24.
15. Biggam, J. (2018). EBOOK: Succeeding with your Master's Dissertation: A Step-by-Step Handbook: Step-by-step Handbook.
16. Celine, S. a. (2020). Logistic regression for employability prediction. International Journal of Innovative Technology and Exploring Engineering, 9, 2471--2478.
17. Charbuty, B. a. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2, 20--28.
18. Charbuty, B. a. (2021). Classification based on decision tree algorithm for machine learning}. Journal of Applied Science and Technology Trends, 2, 20--28.
19. Chen, C. P.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences, 314--347.
20. D.Kokane, M. C. (2016). A Review: To Detect and Identify Cotton leaf disease based on pattern recognition technique. International Journal Of Engineering And Computer Science, 4,, 110–118.
21. De Luna, R. G. ((pp. 1414-1419).). Automated image capturing system for deep learning-based tomato plant leaf disease detection and recognition. In TENCON 2018-2018 IEEE Region 10 Conference,IEEE, (2018, October). .

22. Deng, Z. a. (2023). Multi-view Multi-aspect Neural Networks for Next-basket Recommendation. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, {1283--1292.
23. Despotović, I. G. (2015). MRI segmentation of the human brain: challenges, methods, and applications. Computational and mathematical methods in medicine, 2015.
24. Dhikhi, T. T. (2019). Measuring size of an object using computer vision. International Journal of Innovative Technology and Exploring Engineering, 8(4), 424-426.
25. Durmuş, H. G. (2017). Disease detection on the leaves of the tomato plants by using deep learning. 6th International conference on agro-geoinformatics,IEEE, (pp. 1-5).
26. Efrat, A. a. (2020). Consumer purchase patterns based on market basket analysis using apriori algorithms. IOP Publishing, 012109.
27. Fernandes, J. D. (2020). Research Culture: A survey-based analysis of the academic job market.
28. Ganatra, N. &. (2018). A survey on diseases detection and classification of agriculture products using image processing and machine learning. International Journal of Computer Applications,, 1-13.
29. Gangurde, R. a. (2018). Optimized predictive model using artificial neural network for market basket analysis}. International Journal of Computer Science and Communication, 9, 42--52.
30. Gharghory, S. M. (2020). Performance Analysis of Efficient Pre-trained Networks based on Transfer Learning for Tomato Leaf Diseases Classification. International Journal of Advanced Computer Science and Applications., 11(8).
31. Ghous, H. a. (2023). KIET Journal of Computing and Information Sciences, 6, 14--34.
32. Ghous, H. a. (2023). Deep Learning based Market Basket Analysis using Association Rules. {KIET Journal of Computing and Information Sciences, 6, 14--34.
33. Guidotti, R. a. (2018). Personalized market basket prediction with temporal annotated recurring sequences. IEEE Transactions on Knowledge and Data Engineering, 31(11), 2151--2163.
34. Gupta, S. a. (2021). Emerging market retail: transitioning from a product-centric to a customer-centric approach. Journal of Retailing, 97, 597--620.
35. Hassanien, A. E. (2020). Machine learning in telemetry data mining of space mission: basics, challenging and future directions. Artificial Intelligence Review, 53, 3201--3230.
36. Hossain, M. a. (2019). Market basket analysis using apriori and FP growth algorithm. IEEE, 1--6.
37. Hsu, T.-S. W.-C. (2015). An improvement stereo vision images processing for object distance measurement. International Journal of Automation and Smart Technology, 5(2), 85-90.
38. Ilham, A. G. (2018). Market Basket Analysis Using Apriori and FP-Growth for Analysis Consumer Expenditure Patterns at Berkah Mart in Pekanbaru Riau. Journal of Physics: Conference Series (Vol. 1114, No. 1, p. 012131). IOP Publishing.