Malaysian Journal of Computing, 9 (2): 1905-1915, 2024

eISSN: 2600-8238

DECISION TREE AND RULE-BASED CLASSIFICATION FOR PREDICTING ONLINE PURCHASE BEHAVIOR IN MALAYSIA

Maslina Abdul Aziz [1*], Nurul Ain Mustakim[2] and Shuzlina Abdul Rahman[3]

[1,2,3] School of Computing Sciences, College of Computing, Informatics, and Mathemathics, Universiti Teknologi MARA, 40450 Shah Alam*

[1*]maslina_aziz@uitm.edu.my , [2]ainmustakim@uitm.edu.my, [3]shuzlina@uitm.edu.my

ABSTRACT

In Malaysia, fast growth in e-commerce speeds a business need to understand and predict consumer online behavior in order to be more competitive. While the whole world is embracing big data analytics, many businesses in Malaysia, particularly those in the e-commerce sector, find it hard to harness these technologies to their benefit. The absence of specific predictive models and the complexity of socio-cultural diversity further complicate the efforts toward understanding consumer preferences. Therefore, this research tries to fill in some of the gaps by applying decision tree and rule-based algorithms to classify online purchasing behavior amongst Malaysian consumers. The study looks into the data from an online survey comprising 560 respondents with a view to demographic, factors influences, and purchasing behaviour. The performance of six machine learning models comprising J48, Random Tree, REPTree representing decision trees and JRip, PART, and OneR as rule-based algorithms was assessed. Feature selection, pre-processing, and SMOTE were applied in order to balance class inequalities of the dataset. The result indicated that the highest accuracy of 89.34% was achieved by the Random Tree algorithm, while the rule-based algorithm PART reached an accuracy of 87.56%. Results of these models open up the possibility of providing very important insights from a business perspective into consumer behaviour and thus offer actionable data which allows them to complete their job of fine-tuning marketing strategies and engaging customers. The current study contributes to the literature by highlighting decision tree and rule-based classification models as very useful in the Malaysian e-commerce context. These developed predictive models can serve as building blocks where businesses might know more about consumer behavior, personalize marketing, and reach operationally efficient levels. Future research may involve integrating other influencing variables and applying them across industries.

Keywords: *Decision Trees, E-Commerce, Malaysian Consumers, Predictive Models, Rule-Based Algorithm.*

Received for review: 11-07-2024; Accepted: 28-09-2024; Published: 01-10-2024
DOI: 10.24191/mjoc.v9i2.27130

1. Introduction

The ability to predict consumer behavior has indeed been highlighted as one of the major challenges for any company eager to seek an advantage. Against the dynamics of change in the digital ecosystem, proper understanding and anticipation of the demand, preference, or buying habits of customers can be realized through data analytics, AI, and proper market

research (Isa et al., 2020; Loheswar, 2021; Rahayu et al., 2020; Ru et al., 2021; Wong et al., 2023). Previous research from the MCMC (2018) showed that from 2016 to 2018, online purchasing amongst Malaysians saw an increase of 48.8 percent. According to Statista, the number of users in Malaysia's e-commerce sector is expected to add 2.9 million users from 2023 to 2027. Both sources cite that Malaysians are buying more online.

When companies worldwide started embracing data-driven strategies, big data started to be implemented in companies in Malaysia to uncover rich insights about their tech-savvy and diversified customers. This trend shift therefore means that e-commerce companies in Malaysia can tailor their services to the needs of their customers through better data usage for targeted marketing, personalized shopping experiences, and supply chain control. This is especially true in Malaysia, where differences along national, regional, and socio-economic lines make it difficult for people to agree on what they want. The use of big data has increased customer satisfaction and involvement thus enabling the local enterprises to fight both at home and abroad by putting the customer first. The use of big data to provide services to the local customers still has its problems. Big data allow for personalized marketing, an improvement of shopping experiences, and the acceleration of supply chain management. However, multiple companies are still anticipated to make wild guesses about consumers' needs in Malaysia's diversified market. As such, businesses are under obligation to address these challenges so they will be able to gain the most value from smart data and remain competitive in the digital marketplace.

While the progress with e-commerce and big data analytics is impressive across countries, not all nations of the world are in a good position regarding the integration of these technologies; Malaysia is not an exception. This also agreed that "big data analytics adoption in Malaysia, particularly for MSMEs, is considered to be at its infancy by the researcher such as (Loh & Teoh, 2021) and (Vachkova et al., 2023) Such technologies might play a better role in contributing to economic development and enabling cross-border business ventures, but efficient exploitation of technological resources becomes key. In this context, both studies recommend further research to learn about how big data analytics can be better adopted in the e-commerce industry within Malaysia. Among the major issues, designing a robust technology framework and preparing the labor force at an advanced level of data analytics require huge investment. Besides, consumer data handling raises several problems concerning privacy and security that have to balance innovation against regulatory compliance.

The present research attempts to apply some of the established techniques of classification, such as decision trees and rule-based algorithms, to predict online consumer behavior in Malaysia. These are powerful techniques for processing complicated and nonlinear patterns of data in a clear and interpretable way, thus being appropriate for business decision-making purposes (Khairudin et al., 2020). The Malaysian online market presents a unique combination of rich cultural diversity, rapid technology adoption, and a demographic mix with technologically savvy users. This creates certain opportunities and challenges for businesses. This study fills the gap in literature by analyzing the effective use of these techniques in Malaysia's online market, which has combined factors such as attitude, motivation, perceived risk, belief, and demographics into a single comprehensive predictive model. The rest of the paper is organized in a way that several studies reviewing the classification and supervised learning algorithms for the consumer prediction in consumer behavior are presented in Section 2 (Related Work). That section shortly describes methodology techniques employed within the study. Further, experimental findings and comparative analysis concerning this study are presented in Section 4 (Results and Discussion). Finally, Section 5 stands for the Conclusion, where the results of the Research are presented.

2. Related Work

Machine learning finds wide applications in domains such as e-commerce, finance, employability prediction, e-learning, and mobile security. Some recent references can be found in (Musleh et al., 2024; Ying, 2023). Applications in e-commerce prove the potential

importance of predictive models in understanding consumer behavior with a view to enhancing decision-making. The previous works evidence that machine learning models have been widely used in e-commerce and other related sectors to predict customer purchase behavior.

However, the major challenge in model selection is balancing accuracy and interpretability. While rule-based models like JRip and PART are transparent-a very important feature in areas such as finance and education-complex models like Neural Networks and Random Forest yield more accurate results in general Chaubey et al. (2022)**.** This may imply a trade-off that needs to be done judiciously, based on the application for which the model is being chosen. It has, for instance, emerged as one of the best models in its use across many applications in healthcare, including in the prediction of heart failure, as it made predictions with high sensitivity and accuracy, beating other models. (Huang et al., 2021).

Chaubey et al.(2022) explored various ML models, namely Decision Trees, Random Forest, SVM, and Neural Networks, and found that the fusion of classifiers achieved an accuracy of 92.42%. Another related work on machine learning models for predicting customer purchasing decisions in food delivery over the online platform was conducted by Madani & Alshraideh (2021). In their work, it was shown that the C4.5 decision tree achieved the highest accuracy of 91.67%. It is also illustrated that C5.0 and the decision tree model perform effectively in other domains, including divorce prediction, with 77.96% accuracy of divorce classification among Malaysian women (Aimran et al., 2022). Given the use case, different algorithms will perform better. For example, REPTree outperformed in mobile adware detection (Ndagi & Alhassan, 2019), while stochastic gradient boosting worked for customer churn prediction (Prabadevi et al., 2023). FZ Besides, quality and availability of data are also one of the most important factors that determine the model's effectiveness. Large and clean datasets tend to benefit complex models such as Random Forest, while rule-based models may be preferred for smaller or more interpretable datasets.

Beyond e-commerce, rule-based algorithms like JRip, PART, and REPTree have made great achievements in employability prediction, credit allocation, and e-learning. Outside e-commerce, the rule-based algorithms like JRip and PART were applied to predict employability among IT graduates by Gatbonton & Aguinaldo (2018); Ayhan & Uçar (2022) used such models for credit allocation. Beside these Ndagi & Alhassan (2019) presented the performance of these classifiers in mobile adware detection and showed outstanding accuracy by REPTree, readable rules generated by JRip and PART. Besides this, Siebra et al. (2020) showed JRip and PART useful in the prediction of e-learning dropout rates where interpretability and rule generation plays a prime role in educational settings. With the increasing demand for transparency within machine learning, this area of research also became important, which is known as Explainable AI. XAI tries to show how decisions are made by black-box models like Neural Networks and Random Forest, which is much needed in areas where decisions have to be crystal clear.

Another research was done by Zuo et al. (2017) , where machine learning methods for predicting customer purchase behaviors in grocery stores were applied. The Bayes classifier and the Support Vector Machine are two representatives of machine learning methods used in this research, analyzed in their capacity to forecast purchasing decisions. Other studies have pointed out that improvement of the customers' churn needs advanced models. Prabadevi et al. (2023) conducted studies to improve the techniques of machine learning in early stage customer churn prediction. The algorithms considered in the study included stochastic gradient boosting, Random Forest, logistic regression, and k-nearest neighbors. The results of the various algorithms showed varying levels of accuracy, with the highest being from stochastic gradient boosting at 83.9%, followed by logistic regression at 82.9%, Random Forest at 82.6%, and k-nearest neighbors at 78.1%.

Similarly, Ying (2023) conducted research into the construction of predictive models for e-commerce customer churn within Malaysia's online marketplace. The motivation for conducting this research was due to the peculiar socio-cultural factors, diverse consumer preferences, and regulatory frameworks underpinning the manner of online purchasing in

Malaysia. In order to handle such complexities, the study applied six machine learning techniques surveyed through the extensive review of literature. These include Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and eXtreme Gradient Boosting. The purpose of the paper was to describe how those machine learning algorithms are effective in predicting customer churn within the e-commerce industry in Malaysia. By harnessing such techniques, the study purposed to furnish valuable insights into customers' behavioral traits and improve retention strategies based on evidence of unique market dynamics in the Malaysian online marketplace. Although machine learning models have been successful in many sectors, the deployment of these models in real-world settings seems to be challenging on a number of issues, including but not limited to scalability, computational cost, and the need for clean and high-quality data.

All these are some of the challenges that any business enterprise has to address if it intends to use machine learning for better prediction of consumer behavior and thereby optimizing business strategy. The next section summarizes ongoing research on machine learning applications across different domains, discussing the promises and challenges of predictive models. Such a focus on culturally diverse markets, such as Malaysia, in the present work will add much depth to the understanding of the practical deployment of machine learning algorithms for predicting consumer behavior and optimizing business strategy. By resolving trade-offs between accuracy and interpretability, this research contributes to an emerging knowledge base on real-world applications of machine learning models.

3. Methodology

This section consists of the methodology of this study, including data collection in Section 3.1, data preparation and pre-processing in Section 3.2, data modelling in section 3.3 and performance measurement in section 3.4.

3.1 Data Collection and Features

The data collection in this research was done by distributing the online survey to Malaysian consumers aged 18 years and above who had made at least one online purchase. The preliminary dataset included 560 instances and 85 features categorized into demographic, factors, and online consumer behavior features. Each category entailed several sub-features that indicate a variety of influencing factors on consumer behaviors.

- Demographic features accounted for about 10 features, like gender, age, level of education, and annual income.

- The factors influencing online purchase behavior have been further sub-divided into many features for specific categories. For example, Attitude was further divided into BA1 to BA7; Perceived Risk into BB1 to BB6; Trust & Security into BC1 to BC6; Psychological factors into BD1 to BD5; and Hedonic Motivation into BE1 to BE5. These influencing factors give a deeper understanding of attitude, perceived risks, and motivations that govern consumers while making online purchases.

- The target or output features were Online Purchasing Behaviour. This consisted of 10 sub-features ranging from C1 to C10 that map various purchasing decisions which the respondents would have performed.

This dataset originally had a count of 85 features before doing any preprocessing. These features gave a wide view into what affects consumer purchasing decisions, as was shown in Table 1. The dataset captured a wide range of information on personal details, shopping habits, attitudes, and concerns, therefore giving a good basis for the analysis of consumer behavior.

Table 1. Data Description before pre-processing

No	Features	Description	Number of features
1	Gender	Consumer's gender	Demographic 10
2	Age	Consumer's age	
3	Level of Education	Consumer's level of education	
4	Ethnicity	Consumer's ethnicity	
5	Annual income	Consumer's annual income	
6	Employment status	Consumer's employment status	
7	Current residential	Consumer's current residential	
8	Often online shopping	How frequently the consumer visits online shopping platforms	
9	Average time	The average time the consumer spends online shopping	
10	Type of product	Products that are usually purchased by the consumer	
11	Attitude (BA1-BA7)	Attitude factors influence online purchasing behavior	7
12	Perceived Risk (BB1- BB6)	Perceived risk factors influence online purchasing behavior	6
13	Trust and Security (BC1 - BC6)	Trust and security factors influence online purchasing behavior.	6
14	Psychological (BD1- BD5)	Psychological factors influence online purchasing behavior	5
15	Hedonic Motivation (BE1- BE5)	Hedonic motivation factors influence online purchasing behavior	5
16	Promotion (BF1- BF6)	Promotion factors influence online purchasing behavior	6
17	Product Price (BG1- BG6)	Product price factors influence online purchasing behavior	6
18	Privacy (BH1- BH6)	Privacy factors influence online purchasing behavior	6
19	Emotional (BI1- BI6)	Emotional factors influence online purchasing behavior	6
20	Perceived Benefits (BJ1- BJ6)	Perceived benefit factors influence online purchasing behavior	6
21	Accessibility (BK1- BK6)	Accessible factors influence online purchasing behavior	6
22	Online Purchasing Behavior (C1- C10)	Will the consumer purchase? (Output)	10

3.2 Data Preparation and Pre-Processing

After the collection of raw data, pre-processing cleaning and reduction were performed. The technique involved in this may include feature selection, whereby the most relevant features would remain and the redundant or less important ones discarded. In such a way, the dataset was reduced to a total of 30 features, pinpointing mainly the most influencing features on consumer purchasing behavior. This step involves the handling of missing values and normalization of data in variables that required this treatment to improve model performance.

Additionally, this could involve upsizing the dataset to 1126 instances, perhaps by resampling methods like SMOTE to handle class imbalance issues for better generalization of the model performance. The final dataset was based on five important class labels, reflecting a spectrum from "definitely would not buy" to "probably would not buy", "might buy", "probably would buy", and finally "definitely buy." This ordinal classification allows for even more subtlety in the analysis of consumer behavior, essential to making predictions about purchase decisions under a variety of influences.

3.3 Dataset Description

After the pre-processing stage, the refined dataset was divided into three disjoint subsets, each aimed at exploring different facets of consumer behavior and purchase intention. The splitting of these datasets was conducted as follows:

1. **Dataset 1:** This is the base dataset that consists of features from A, which is Demographics; B, Factors influencing consumer purchasing behavior; and C, Online Purchasing Behavior. This is the complete dataset across all aspects of the consumer behavior model and encompasses a full view of the features that affect purchasing decisions.
2. **Dataset 2:** The data selected in this will be features from A (Demographics) and C (Online Purchasing Behavior) without the psychographic and behavioral factors in B. This way, a direct correlation of demographic features to online purchasing behavior could be put in focus.
3. **Dataset 3:** This covers all features from A to C but does not include the types of products purchased. It would thus provide insight into the wider emotional and psychological issues of consumers without concern for product-specific preferences.

3.4 Data Modelling

At this stage, applied data mining approaches on the data include Decision Tree (DT) and Rules. In this context, the Decision Tree classifier encompasses three different methods: J48, Random Tree, and REPTree. On the other hand, the classifier Rules utilises JRip, OneR, and PART.

3.4.1 Decision Tree (DT)

Decision trees are a flexible and interpretable machine learning algorithm used for classification and regression. In WEKA, three key decision tree algorithms are J48, Random Tree, and REPTree. J48, an implementation of the C4.5 algorithm, splits data based on information gain and includes pruning to prevent overfitting, making it easy to interpret. Random Tree builds multiple unpruned trees using random data subsets, enhancing robustness through ensemble learning. REPTree constructs trees using information gain or variance reduction and prunes them with reduced error pruning, creating efficient and generalizable models. These algorithms provide a balance of simplicity, robustness, and efficiency for various data analysis tasks.

3.4.2 Rules

There were three rule-based algorithms used. These were PART, JRip, and OneR. JRip is a version of the RIPPER algorithm that creates and removes rules to make the algorithm more general and better at dealing with noisy data. OneR makes one rule for each characteristic and chooses the one with the lowest mistake rate. This makes the model easy to understand. PART builds partial decision trees and pulls out the best rules, which makes sure that the results are accurate and easy to understand. These algorithms have easy to understand rules for classifying things, which makes them useful for making models that are clear.

3.5 Performance Measurement

This section presents the performance measurement that compares both models-Decision Tree and Rule-Based-based on accuracy, precision, recall, and F1 measure. Each model evaluation is, respectively, conducted three times through cross-validation and split percentage methods. Experiment 1 applies both algorithms to the original dataset. Experiment 2 analyzes the models by using a dataset with demographic and online purchase behavior features without some specific influencing factors. Because of that, Experiment 3 is based on a much rich data set: containing demographic, influencing factors, and online purchase behavior variables but without

product types.

Accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions being made. The formula is shown in the equation below. The calculation is performed by dividing the number of correct predictions by the total number of predictions made. Precision is a metric that measures the proportion of positive class predictions that correctly belong to the positive class. A higher precision score indicates that the model exhibits greater accuracy and reliability in correctly identifying positive samples.

Recall is the measure of performance in a classification task that shows how well a model outperforms other models in terms of identifying all the positive instances from among all the actually positive ones in a dataset. It basically refers to the proportion of positive instances that were among the truly positive ones, which have been correctly predicted as positive by a model. The F-measure or F1-score condenses precision and recall into one metric and provides a better sense of the real performance of a classification model. The F-measure will be particularly useful in cases where the classes in data sets are not regularly distributed, and accuracy cannot provide a completely realistic measure of performance. The formula for the F1-score calculation is as follows: F1 score = 2 * (precision * recall) / (precision + recall).

4. Result and Discussion

This section show the comparative experiments trough three dataset using two algorithms decision tree and rule base. This study employed the cross validation approach for model evaluation. In this study this research evaluated and compared the classification models using four types of evaluation metrics: accuracy, precision, recall and F1.

Table 2, performance measures of six algorithms on Dataset 1 evaluated in terms of accuracy, precision, recall, and F-Measure. In the regard, the J48 algorithm had an accuracy rate of 86.14%, reflecting balanced precision, recall, and F-Measure values of 0.859, 0.857, and 0.858 respectively. The performance of the best model, the Random Tree, was 89.16%, with precision, recall, and F-Measure at 0.888, 0.890, and 0.889, respectively. REPTree, on the other hand, has a very low accuracy of 83.83% at an individual 0.830, 0.836, and 0.833. JRip had done comparatively, coming up with an accuracy rate of 84.90% while receiving 0.845, 0.848, and 0.846 scores. The worst performance was done by OneR, which came up with an accuracy of 60.92%, while it came with scores of 0.558, 0.592, and 0.574. PART had a solid accuracy rate of 87.48% while reaping in impressive scores of 0.873, 0.871, and 0.872, placing it amongst the top performing models.

Table 3 depicts the performances of the six algorithms on Dataset 2 for accuracy, precision, recall, and F-Measure. J48 showed a promising performance by having an accuracy of 81.26% and scores of 0.805, 0.809, and 0.807. Random Tree performed best among all the trees with an accuracy of 87.56% along with scores of 0.870, 0.872, and 0.871. REPTree had an accuracy of 76.55% with scores of 0.755, 0.759, and 0.757. JRip had an accuracy of 79.13% while sustaining scores of 0.784 throughout. OneR had an accuracy of 61.98% with lower scores of 0.572, 0.605, and 0.588. PART was also very strong in performance, having an accuracy of 79.57% and scores of 0.793, 0.789, and 0.791.

Table 4 presents the performance of the six algorithms for Dataset 3. J48 yields an accuracy of 86.05%, precision, recall, and F-Measure values of 0.857, 0.856, and 0.857, respectively. Thus, it is pretty consistent and reliable. Random Tree performed the best, yielding the highest accuracy of 89.34% and precision, recall, and F-Measure scores of 0.889, 0.890, and 0.890, respectively, proving to be quite effective. For REPTree, the accuracy was 82.86%, with a precision of 0.820, recall of 0.825, and F-Measure of 0.823, hence satisfactory. JRip delivered the best performance among the considered algorithms, with an accuracy of 86.32% and precision, recall, and F-Measure values of 0.859, 0.861, and 0.861, respectively, hence reliable. At the bottom, OneR had an accuracy of 60.92% while recording the lowest precision, recall, and F-Measure values of 0.563, 0.590, and 0.576, respectively, hence of limited effectiveness. PART

managed an accuracy of 87.56%, with precision, recall and F-Measure values of 0.872, 0.873, and 0.873, respectively, making it one of the top performers.

Based on the findings, it will be correct to establish that the topmost models that can be used for any classification task are those of Random Tree and PART algorithms because both have outstanding and consistent performances among the tested datasets. J48 and JRip algorithms are follow-through good models; their metrics are good and reliable. REPTree can be considered a middle performance algorithm and works, but not as well as some of the other options. The OneR model is the worst of all the models that will be compared, having the lowest measures of performance in general.

Table 2. Precision, Recall, and F-Measure for Dataset 1

Dataset 1	Accuracy	Precision	Recall	F-Measure
J48	86.14%	0.859	0.857	0.858
Random Tree	89.16%	0.888	0.890	0.889
REPTree	83.83%	0.830	0.836	0.833
JRip	84.90%	0.845	0.848	0.846
OneR	60.92%	0.558	0.592	0.574
PART	87.48%	0.873	0.871	0.872

Table 3. Precision, Recall, and F-Measure for Dataset 2

Dataset 2	Accuracy	Precision	Recall	F-Measure
J48	81.26%	0.805	0.809	0.807
Random Tree	87.56%	0.870	0.872	0.871
REPTree	76.55%	0.755	0.759	0.757
JRip	79.13%	0.784	0.783	0.784
OneR	61.98%	0.572	0.605	0.588
PART	79.57%	0.793	0.789	0.791

Table 4. Precisiion, Recall, and F-Measure for Dataset 3

Dataset 3	Accuracy	Precision	Recall	F-Measure
J48	86.05%	0.857	0.856	0.857
Random Tree	89.34%	0.889	0.890	0.890
REPTree	82.86%	0.820	0.825	0.823
JRip	86.32%	0.859	0.861	0.861
OneR	60.92%	0.563	0.590	0.576
PART	87.56%	0.872	0.873	0.873

5. Conclusion and Recommendations

In all, six prediction models were used and explored by their performance and accuracy of the prediction, where the original 560 instances of the dataset have turned into 1126 instances after the use of the SMOTE technique, in order to resolve problems of class imbalance. The three different datasets were divided, each having a collection of features. The various methodologies that were followed in developing the predictive models were the decision tree approach, which included J48, Random Tree, and REPTree, while rule-based approaches included JRip, OneR, and PART. Three rounds of experiments with these studies reported that among the decision tree classifiers, Random Tree was the best performer, while PART was the most accurate algorithm among the rule-based categories. These findings indicate that the Random Tree and PART algorithms demonstrate promise in the prediction of consumer buying behavior in e-commerce. However, there are a few limitations that must be taken into account when interpreting these findings. First, since the dataset reflects Malaysian consumers alone, it may be difficult to generalize such findings to other geographic regions characterized by different

consumer buying behaviors. Besides, the application of SMOTE may have altered the natural distribution within the data; hence, some sort of bias in generalizing the real world capability of the model is invited. Another limitation relates to the fixed set of features. Adding other relevant features will give more substantial predictive power for the models. Lastly, the analysis was done using only the WEKA toolset, and looking into other machine-learning platforms or techniques may give different outcomes or improvements in model performance. Other directions for future research could be the use of different and more diverse e-commerce datasets, together with experimenting with other classification techniques using WEKA or other data-mining tools. Other work may involve the performance comparison of decision tree algorithms by weight allocation, and ranking methods to achieve higher accuracy. Classifying e-commerce data using any other data mining techniques may further enhance the outcomes of predictive models.

Acknowledgements

The researchers would like to thank UiTM and MARA for their support of academic research and funding possibilities to share research findings locally.

Funding

The author(s) received no specific funding for this work.

Author Contribution

Author 1 prepared the literature review, wrote the content, developed the research methodology, and conducted the results analysis and discussion. Author 2 and Author 3 reviewed the literature review and provided constructive feedback on the manuscript. All authors have reviewed and approved the final version of the manuscript.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Aimran, N., Rambli, A., Afthanorhan, A., Mahmud, A., Sapri, A., & Aireen, A. (2022). Prediction of Malaysian Women Divorce Using Machine Learning Techniques. *Malaysian Journal of Computing*, *7*(2), 1067–1081. https://doi.org/10.24191/mjoc.v7i2.17077

Ayhan, T., & Uçar, T. (2022). Determining customer limits by data mining methods in credit allocation process. *International Journal of Electrical and Computer Engineering*, *12*(2), 1910–1915. https://doi.org/10.11591/ijece.v12i2.pp1910-1915

Chaubey, G., Gavhane, P. R., Bisen, D., & Arjaria, S. K. (2022). Customer purchasing behavior prediction using machine learning classification techniques. *Journal of Ambient Intelligence and Humanized Computing*, *April*. https://doi.org/10.1007/s12652-022-03837-6

Gatbonton, T. M. C., & Aguinaldo, B. E. (2018). Employability predictive model evaluator using part and JRIP classifier. *ACM International Conference Proceeding Series*, 307–310. https://doi.org/10.1145/3301551.3301569

Huang, N. S. M., Ibrahim, Z., & Diah, N. M. (2021). Machine Learning Techniques for Early Heart Failure Prediction. *Mjoc.Uitm.Edu.My*, *6*(2), 872–884. https://mjoc.uitm.edu.my/main/images/journal/vol6-2-2021/7-Huang-et-al-Vol-62.pdf

Isa, K., Shah, J. M., Palpanadan, S. T., & Isa, F. (2020). Malaysians' popular online shopping websites during movement control order (Mco). *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2154–2158. https://doi.org/10.30534/ijatcse/2020/190922020

Khairudin, Z., Abdul Razak, N. A., Abd Rahman, H. A., Kamaruddin, N., & Abd Aziz, N. A. (2020). Prediction of Diabetic Retinopathy Among Type Ii Diabetic Patients Using Data Mining Techniques. *Malaysian Journal of Computing*, *5*(2), 572. https://doi.org/10.24191/mjoc.v5i2.10554

Loh, C.-H., & Teoh, A.-P. (2021). The Adoption of Big Data Analytics Among Manufacturing Small and Medium Enterprises During Covid-19 Crisis in Malaysia. *Proceedings of the Ninth International Conference on Entrepreneurship and Business Management (ICEBM 2020)*, *174*(Icebm 2020), 95–100. https://doi.org/10.2991/aebmr.k.210507.015

Loheswar, R. (2021, June 8). Ipsos poll: Shopee is Malaysia's preferred e-shopping platform during pandemic times, used by four in five buyers. *Malay Mail*. https://www.malaymail.com/news/malaysia/2021/06/08/ipsos-poll-shopee-is-malaysias-preferred-e-shopping-platform-during-pandemi/1980472

Madani, B., & Alshraideh, H. (2021). *Predicting Consumer Purchasing Decisions in the Online Food Delivery Industry*. 103–117. https://doi.org/10.5121/csit.2021.111510

MCMC. (2018). Internet users survey 2018: Statistical brief number twenty-three. *Internet Users Survey 2018*, 1–39. https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Internet-Users-Survey-2018.pdf

Musleh, D., Alkhwaja, A., Alkhwaja, I., Alghamdi, M., Abahussain, H., Albugami, M., Alfawaz, F., El-Ashker, S., & Al-Hariri, M. (2024). Machine Learning Approaches for Predicting Risk of Cardiometabolic Disease among University Students. *Big Data and Cognitive Computing*, *8*(3). https://doi.org/10.3390/bdcc8030031

Ndagi, J. Y., & Alhassan, J. K. (2019). Machine learning classification algorithms for adware in android devices: A comparative evaluation and analysis. *2019 15th International Conference on Electronics, Computer and Computation, ICECCO 2019*, *Icecco*, 1–6. https://doi.org/10.1109/ICECCO48375.2019.9043288

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, *4*(June), 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

Rahayu, E., Fauzan, F., Wijaya, H., & Gunadi, W. (2020). The Effect of Trust and Satisfaction on Customer Loyalty in Online Shop: Case of C2C E-Commerce in Indonesia. *International Journal of Academic Research in Business and Social Sciences*, *10*(8), 699–709. https://doi.org/10.6007/ijarbss/v10-i8/7619

Ru, L. J., Kowang, T. O., Long, C. S., Fun, F. S., & Fei, G. C. (2021). Factors Influencing Online Purchase Intention of Shopee's Consumers in Malaysia. *International Journal of Academic Research in Business and Social Sciences*, *11*(1). https://doi.org/10.6007/ijarbss/v11-i1/8577

Siebra, C. A., Santos, R. N., & Lino, N. C. Q. (2020). A self-adjusting approach for temporal dropout prediction of E-learning students. *International Journal of Distance Education Technologies*, *18*(2), 19–33. https://doi.org/10.4018/IJDET.2020040102

Vachkova, M., Ghouri, A., Ashour, H., Isa, N. B. M., & Barnes, G. (2023). Big data and predictive analytics and Malaysian micro-, small and medium businesses. *SN Business & Economics*, *3*(8), 1–28. https://doi.org/10.1007/s43546-023-00528-y

Wong, K. X., Wang, Y., Wang, R., Wang, M., Oh, Z. J., Lok, Y. H., Khan, N., & Khan, F. (2023). Shopee: How Does E-commerce Platforms Affect Consumer Behavior during the COVID-19 Pandemic in Malaysia? *International Journal of Accounting & Finance in Asia Pasific*, *6*(1), 38–52. https://doi.org/10.32535/ijafap.v6i1.1934

Ying, T. H. (2023). *E-Commerce Customer Churn Prediction for the Marketplace in Malaysia*. *11*(2), 58–66.

Zuo, Y., Yada, K., & Ali, A. B. M. S. (2017). Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques. *Proceedings - Asia-Pacific World Congress on Computer Science and Engineering 2016 and Asia-Pacific World Congress on Engineering 2016, APWC on CSE/APWCE 2016*, 18–25. https://doi.org/10.1109/APWC-on-CSE.2016.015