¹st Proceeding of International Conference on Science and Technology UISU (ICST)
"Contribution of Engineering Disciplines to Achieving Sustainable Development Goals Research Realizing the Vision of Golden Indonesia 2045"
ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

An Analysis Of Cash On Delivery (COD) Purchase Classification Using the C4.5 and ID3 Algorithms

Asrianda[1)], Herman Mawengkang[2], Poltak Sihombing[3)], Mahyuddin K. M. Nasution[4)]

[1,2,3,4)] Program Studi Doktor Ilmu Komputer, Fakultas Ilmu Komputer dan Informasi Teknologi,
Universitas street No. 9A, Medan, 20155, Universitas Sumatera Utara, Indonesia
[1] Corresponding author: asrianda@students.usu.ac.id
[2)] mawengkang@usu.ac.id
[3)] poltak@usu.ac.id
[4)] mahyuddin@usu.ac.id

Abstract - This study investigates the prevalence of COD (Cash on Delivery) payment methods across different price categories of goods and the impact of normalization on classification accuracy using the C4.5 and ID3 algorithms. The data reveals that the majority of COD payments occur for items priced below 1,000 PKR, with a decreasing trend as the price increases. conversely, higher-priced items see more non-COD transactions. Monthly analysis shows the highest number of COD transactions in November. Among various product categories, Men's Fashion, Soghaat, and Beauty & Grooming dominate COD payments. The implementation of min-max normalization improves the accuracy of both C4.5 and ID3 algorithms, with C4.5 showing a notable improvement in precision and recall metrics.

Keywords: Normalization, C4.5 and ID3, Classification, and COD.

1. INTRODUCTION

Online purchasing using the Cash on Delivery (COD) payment method provides convenience and security for consumers, consumers are not comfortable making payments in advance. COD has become an option in several developing countries due to limited access to banking services and low trust in online transactions[1]. In the e-commerce industry, understanding and predicting COD purchasing patterns is important for improving marketing strategies and customer service. Allows consumers to pay when goods are received, providing more confidence to customers worried about fraud or product quality. This incident is very relevant to customers who have low trust in electronic payments and credit cards[2]. Studies show the COD method increases consumer confidence and consumer convenience. In turn increasing sales on e-commerce platforms[1].

Offering solutions for unbanked consumers, enabling them to participate in the digital economy. Overcoming trust issues in online transactions, because you pay after receiving and inspecting the goods. Provides a sense of security and comfort for consumers, worrying about online fraud. The risk of losing money because the goods do not match the description, and the goods can never be minimized. Consumers are sure to shop online, they can check the goods first before making payment.

COD has many challenges, high operational costs, fraud risks and increased logistics burdens faced by sellers and delivery service providers. The trend of increasing consumer interest in switching to electronic payment services. The Covid-19 pandemic has encouraged the adoption of new technology and reduced physical contact[3]. Consumer trust and ease of transaction factors play an important role in the adoption of COD in the Shopee market[4]. Problems are faced that influence COD purchasing decisions, by determining the variables that influence consumer decisions.

Decision trees are an effective classification technique for solving problems, the ID3 algorithm uses entropy and information gain to build models. Works by selecting attributes and providing the highest information gain, to divide data at tree nodes[5]. Has several weaknesses, cannot handle continuous attributes and missing values. Tends to overfitting because it does not have a pruning mechanism[6]. Overcoming the limitations of ID3, Quinlan introduced C4.5 in handling continuous attributes. Divides data at optimal split points based on gain ratio[7]. Handle missing values by estimating the distribution of missing values, reducing overfitting. By cutting off tree branches that do not contribute significantly to model accuracy[8]. Studies show C4.5 is often superior to ID3 in a variety of applications, medical, financial and marketing classifications[9].

Use. C4.5 and ID3 are proven to be effective, the main problem in data analysis is different feature scales, which can affect the performance of classification models. C4.5 has higher computational complexity than ID3. When dealing with large datasets and many attributes, it causes bias in the resulting model. The trees produced by C4.5 become dense and complex, especially on large datasets and reduce interrepresentation. Min-max normalization is used by scaling the features in the range [0,1], to test the performance of the ID3 and C4.5 algorithms.

There have been many studies using the C4.5 and ID3 algorithms for data classification in various contexts. There is a gap in studies that specifically examine the effects of data normalization. In particular, Min-Max normalization of the performance of the two algorithms in the context of COD purchases. Previous research often focuses only on direct comparisons between algorithms without considering the impact of the data pre-processing techniques used. There has not been much research specifically exploring how to normalize Min-Max. In influencing classification accuracy in e-commerce scenarios with COD payment methods. This research seeks to fill the gap, by examining how min-max normalization can affect the accuracy of COD purchase data classification using the C4.5 and ID3 algorithms.

2. LITERATURE REVIEW

2.1 Cash on Delivery (COD) purchases.

Cash on Delivery (COD) is a payment method made by buyers when the goods are received. This method has been used for a long time, especially in retail and postal commerce, before becoming popular in e-commerce. COD offers confidence to consumers who are reluctant to pay up front, thereby facilitating transactions in markets that are less developed in terms of digital financial infrastructure. Cash on Delivery plays an important role in increasing consumer confidence in e-commerce. COD can increase the hedonic value of online purchases without affecting the utilitarian value or risk perceived by consumers[10].

The presence of COD makes the shopping experience more enjoyable for consumers, which ultimately increases their purchasing intentions. COD can increase consumer confidence, many consumers feel safer using COD. They don't need to pay for the goods before receiving the goods. Payment is made upon delivery, the risk of fraud can be minimized. Consumers can check the goods before making payment[11]. Can attract consumers, previously hesitant to shop online, due to distrust of other payment methods.

The COD problem involves higher operational costs for the seller. Including the cost of returning goods, if the consumer refuses to accept the goods sent. The risk of consumers refusing to accept goods upon delivery. Resulting in additional costs for sellers in the form of returning goods and managing inventory. The COD process extends the time for delivery and receipt of payment, payment is made after the goods arrive in the hands of the consumer. The presence of COD makes the shopping experience more enjoyable for consumers, ultimately increasing their purchasing intentions.

2.2 ID3 algorithm

The ID3 algorithm (Iterative Dichotomiser 3) is a machine learning algorithm used to build decision trees based on the concepts of entropy and information gain. The ID3 algorithm has the advantages of simplicity and computational speed, easy to implement and fast in processing data. ID3 has some limitations, such as its inability to handle attributes with continuous values. Tends to produce overfitting trees[5], if the training data contains noise. ID3 does not have a mechanism to handle missing data, which can affect model accuracy[6]. ID3 calculates the entropy of the entire dataset, selecting the attribute that provides the greatest information value to become the root of the tree. The process is repeated recursively for each branch until all attributes are exhausted or the tree reaches a certain depth.

The ID3 algorithm uses entropy to measure uncertainty in a data set. Entropy(E) for a data set S with n categories is defined as:

$$E(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

Pi is the proportion of items in category i, the information gain (IG) for attribute A is used to divide data S into smaller subsets.

$$IG(S,A) = E(S) - \sum_{v\in \text{Values}(A)} \frac{|S_v|}{|S|} E(S_v)$$

Siis a subset S that has a value v in attribute A, the highest gain information attribute is selected as the root of the decision tree.

2.3 Algorithm C4.5

The C4.5 algorithm is superior to ID3, CART on student datasets in terms of accuracy[12]. Presenting analysis in the academic environment, predicting the academic performance of undergraduate and postgraduate students. Shows that decision tree accuracy is consistently 3-12% more accurate than Bayesian Network. The C4.5 algorithm solves problems using certain instructions to produce output[13]. The C4.5 algorithm produces an accuracy of 54.17% compared to ID3. CART produces an accuracy of 55.83%, compared to the C4.5 and ID3 algorithms[12]. ID3, can only perform categorical features (nominal/ordinal, while numeric types cannot be used[13]. C4.5 as an algorithm uses gain ratio as the selection of split attributes, the formula is as follows:

$$GainRatio(A) = \frac{\text{Gain}(A)}{\text{SplitEntropy}(A)}$$

To calculate split entropy use the following formula:

$$SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|S_i|}{|S|} * log_2 \frac{|S_i|}{|S|}$$

2.4 Min-Max Normalization

The ID3 and C4.5 algorithms are the best known and most widely used methods. These two algorithms are used to build a decision tree model from a data set, which can then be used to make predictions on new data. However, to improve model performance, data pre-processing stages such as normalization are often required. One normalization technique that is often used is the min-max method.

Min-max normalization provides a linear transformation to normalize a dataset, scaling the input data into a range of -1 to 1 or 0 to 1[14]. Helps in resolving the issue of varying scales between attributes, and can impact performance. The formula is as follows:

$$\mathbf{X}_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

3. METHOD

The research methodology includes several stages:

1. Data collection: e-commerce transaction data obtained from the Pakistan Largest Ecommerce dataset Pakistan Largest Ecommerce Dataset.csv[15]. Includes several features influencing COD purchasing decisions.
2. Data preprocessing has been carried out on the dataset, missing values have been removed, with a total of 10 columns and 97,228 rows. Table as below:

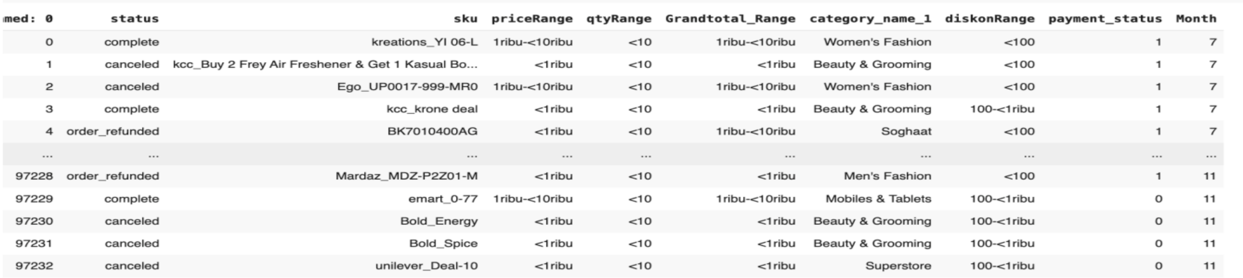
Table 1. Column name dataset ecommerce.xlsx

Column name	Missing value	type
status	0	Objects
sku	0	Objects
priceRange	0	Objects
qtyRange	0	Objects
Grandtotal_Range	**0**	**Objects**
category_name_1	0	Objects
discountRange	0	Objects
payment_status	0	int64
Month	0	int64

3. Min-max normalization is applied to scale the feature range [0,1, to improve the algorithm performance.
4. The model was built using the ID3 and C4.5 algorithms, and the model was evaluated using accuracy, precision, recall and ROC curve metrics.

4. RESULTS AND DISCUSSION

There are 97,228 e-commerce data, dividing training data into 70% and testing data into 30%, the training data is 77,782 and testing data is 19,446. Pakistan ecommerce dataset with label payment_status, as shown below:



med: 0	status	sku	priceRange	qtyRange	Grandtotal_Range	category_name_1	diskonRange	payment_status	Month
0	complete	kreations_YI 06-L	1ribu-<10ribu	<10	1ribu-<10ribu	Women's Fashion	<100	1	7
1	canceled	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	<1ribu	<10	<1ribu	Beauty & Grooming	<100	1	7
2	canceled	Ego_UP0017-999-MR0	1ribu-<10ribu	<10	1ribu-<10ribu	Women's Fashion	<100	1	7
3	complete	kcc_krone deal	<1ribu	<10	<1ribu	Beauty & Grooming	100-<1ribu	1	7
4	order_refunded	BK7010400AG	<1ribu	<10	1ribu-<10ribu	Soghaat	<100	1	7
...	...	...	...	...	...	...	...	...	...
97228	order_refunded	Mardaz_MDZ-P2Z01-M	<1ribu	<10	<1ribu	Men's Fashion	<100	1	11
97229	complete	emart_0-77	1ribu-<10ribu	<10	1ribu-<10ribu	Mobiles & Tablets	100-<1ribu	0	11
97230	canceled	Bold_Energy	<1ribu	<10	<1ribu	Beauty & Grooming	100-<1ribu	0	11
97231	canceled	Bold_Spice	<1ribu	<10	<1ribu	Beauty & Grooming	100-<1ribu	0	11
97232	canceled	unilever_Deal-10	<1ribu	<10	<1ribu	Superstore	100-<1ribu	0	11

Figure 1. Pakistan ecommerce dataset

The payment_status label has data for the true category of 44,752 records and the false category of 33,030 records. Below is a table of Pakistani e-commerce shopping categories:

Table 2. Average e-commerce COD price category (data processing)

price	Cod	code no
<1 thousand	41165	24143
1 thousand -<10 thousand	10486	10954
10 thousand-<50 thousand	4040	5245
50 thousand-<100 thousand	334	741
100 thousand-<500 thousand	1	121

From table 2, 41,165 people made the most cod payments for items under 1 thousand and 24,143 did not use a cod. Goods above 1 thousand to under 10 thousand make a COD payment of 10,485 and do not make a COD payment of 10,954. Prices of goods above 10 thousand to below 50 thousand make a payment with a COD of 4,040 and do not make a payment with a COD of 5,245. Prices above 50 thousand to below 100 thousand make 334 cod payments, and do not make 741 cod payments. Above 100 thousand to below 500 thousand make 1 cod payment, and do not make 121 cod payments. From the table it can be seen that many cod payments are the price. items under 1 thousand. Goods are becoming more expensive, more consumers make payments without using COD than consumers using COD.

Table 3. Average e-commerce COD qtyRange category (data processing)

qty	cod	code no
<10	55907	41118
10-<50	106	69
50--<100	9	9
100-<500	4	6

Table 3 shows that the average number of goods purchased using COD <10 was 55907 and not using COD was 4118. Second, those who ordered goods using COD 10-<50 were 106 and did not use COD 69. The average number of items ordered was 50-< 100 people use COD and 9 don't use COD. On average, orders are 100-<500 items that use COD 4 and don't use COD 6.

Table 4. Monthly shopping categories (data processing)

month	cod	code no
7	7310	1527
8	10518	1016
9	8572	6860
10	9498	3623
11	20128	28176

The highest monthly shopping category was in month 11 with 20128 COD payments, and 28176 not using COD. The second highest was in month 8 which used 10518 COD payments, and 1016 not using COD. The third most used COD in month 10 was 9498, not using COD was 3623. In month 9, 8572 COD payments were used, 6860 were not used. Finally, in month 7, 7310 COD payments were used, and 1527 COD payments were not made.

Table 5**.** Categories of purchasing goods (data processing)

Category name	**cod**	**code no**
Men's Fashion	13658	3818
Mobile & Tablets	6167	9495
Soghaat	9029	2993
Beauty & Grooming	7817	3460
Appliances	2297	4402
Women's Fashion	4290	2319
Superstore	1142	2815
\N	2519	2794
Computing	921	2282
Home & Living	2262	2268

Entertainment	459	2192
Kids & Babies	1974	456
Health & Sports	1927	746
Others	517	1005
School & Education	733	114
Books	314	43

There are 16 categories of goods sold in e-commerce Pakistan, Men's Fashion made the most payments with cod 13658 and did not use cod 3818. The two categories of goods Soghaat made cod payments 9029, did not make cod payments 2993. Beauty & Grooming made cod payments 7817, did not make cod payments as many as 3460. Mobile & Tablets made cod payments as many as 6167, and did not make cod payments as many as 9495. The fifth category made the most cods, namely the Women's Fashion category, as many as 4290, did not make cod payments as many as 2319. The last category made cod payments, namely Books as many as 314, and did not make cod payments 43. Finally, the two Entertainment categories made cod payments as many as 459, and did not make cod payments as many as 2192.

The research results show that min-max normalization succeeded in increasing the accuracy of the classification model in both the C4.5 and ID3 algorithms. Tables 6 and 7 show the model performance before and after normalization. The ID3 algorithm shows an increase in accuracy of 0.03% after applying min-max normalization. C4.5 shows an increase of 0.57% after applying min-max normalization. Precision in the ID3 algorithm shows a decrease of 0.02% after min-max normalization. C4.5 experienced an increase of 0.01% after normalization. Recall shows an increase in the ID3 algorithm of 0.1% after normalization. The C4.5 algorithm shows an improvement of 1.45% after carrying out min-max normalization.

Table 6. Model performance before normalization

Algorithm	**Accuracy**	**Precision**	**Recall**
ID3	90.33%	91.89%	84.64%
C4.5	89.75%	91.86%	83.20%

Table 7**.** Model performance after normalization

Algorithm	**Accuracy**	**Precision**	**Recall**
ID3	90.36%	91.87%	84.74%
C4.5	90.32%	91.87%	84.65%

5. CONCLUSION

The analysis indicates that consumers in pakistan predominantly use cod for low-priced items, while higher-priced goods are more likely to be purchased without cod. Monthly trends highlight a peak in cod usage during november. Among product categories, men's fashion leads in cod transactions, followed by soghaat and beauty & grooming. The study demonstrates that min-max normalization effectively improves the accuracy of classification algorithms, with c4.5 benefiting more significantly than id3. These findings suggest that normalization can be a valuable step in enhancing model performance for classification tasks.

ACKNOWLEDGMENTS

We would like to extend our deepest gratitude to all parties who have provided support and assistance in the completion of this research. First and foremost, we would like to thank the institution and the faculty of computer science and information technology at university of north sumatra for providing the facilities and resources necessary for this research.

REFERENCES

[1]. M. Halaweh, "Cash on delivery (COD) as an alternative payment method for e-commerce transactions: Analysis and implications," *Int. J. Sociotechnology Knowl. Dev.*, vol. 10, no. 4, pp. 1–12, 2018, doi: 10.4018/IJSKD.2018100101.

[2]. DI Hajati, "The Effect of Cash on Delivery, Online Consumer Ratings and Reviews on the Online Product Purchase Decisions," *Bus. Innov. Entrep. J.*, vol. 4, no. 1, pp. 18–26, 2022, doi: 10.35899/biej.v4i1.348.

[3]. B. Purwandari, SA Suriazdin, AN Hidayanto, S. Setiawan, K. Phusavat, and M. Maulida, "Factors Affecting Switching Intention from Cash on Delivery to E-Payment Services in C2C E-Commerce Transactions: COVID-19, Transaction, and Technology Perspectives," *Emergency. Sci. J.*, vol. 6, no. Special Issue, pp. 136–150, 2022, doi: 10.28991/esj-2022-SPER-010.

[4]. M. Alfarizi and RK Sari, "Adoption Of Cash on Delivery (COD) Payment System in Shopee Marketplace Transaction," *Procedia Comput. Sci.*, vol. 227, pp. 110–118, 2023, doi: 10.1016/j.procs.2023.10.508.

[5]. JR Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/bf00116251.

[6]. HMS B, C. Lei, and D. Neagu, *Computational Complexity Analysis of Decision Tree Algorithms*. Springer International Publishing, 2018. doi: 10.1007/978-3-030-04191-5.

[7]. JR Quinlan, "Improved use of continuous attributes in C4.5," *J. Artif. Intel. Res.*, vol. 4, no. 1996, pp. 77–90, 1996, doi: 10.1613/jair.279.

[8]. S. STEVEN L., "Book Review : C4 . 5 : Programs for Machine Learning," *Mach. Learn.*, vol. 240, pp. 235–240, 1994.

[9]. HO Salami, RS Ibrahim, and MO Yahaya, "Detecting Anomalies in Students' Results Using Decision Trees,"*Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 7, pp. 31–40, 2016, doi: 10.5815/ijmecs.2016.07.04.

[10]. S. Hamed and S. El-Deeb, "Cash on Delivery as a Determinant of E-Commerce Growth in Emerging Markets," *J. Glob. Mark.*, vol. 33, no. 4, pp. 242–265, 2020, doi: 10.1080/08911762.2020.1738002.

[11]. NA Hamdani and GAF Maulani, "The influence of E-WOM on purchase intentions in the local culinary business sector," *Int. J.Eng. Technol.*, vol. 7, no. 2, pp. 246–250, 2018, doi: 10.14419/ijet.v7i2.29.13325.

[12]. TM Lakshmi, A. Martin, RM Begum, and VP Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," *Int. J. Mod. Educ. Comput. Sci.*, vol. 5, no. 5, pp. 18–27, 2013, doi: 10.5815/ijmecs.2013.05.03.

[13]. HA Prihanditya, "The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4. 5 Algorithms in Diagnosing Chronic Kidney Disease," *J. Soft Comput. Explore.*, vol. 5, no. 1, pp. 63–69, 2020, doi: https://doi.org/10.52465/joscex.v1i1.8.

[14]. D. Rajeswari and K. Thangavel, "the Performance of Data Normalization Techniques on Heart Disease Datasets," *Int. J. Adv. Res. Eng. Technol.*, vol. 11, no. 12, pp. 2350–2357, 2020, doi: 10.34218/IJARET.11.12.2020.222.

[15]. OpenDataPakistan, "No Title," *Pakistan Largest Ecommerce*, 2021. https://opendata.com.pk/dataset/pakistan-largest-ecommerce-dataset/resource/7395e1d0-c02b-4e1d-abb8-84ae52681ffb