JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
Published by: Indonesian Informatics Association
e-ISSN:2829-310X

Comparative Analysis of Muslim Clothing Sales Predictions Using the C4.5 Method and Linear Regression

Alpa Gustiana
Informatics Engineering, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia

alpagustiana2001@gmail.com

Abstract. This study aims to develop a sales data prediction model using the machine learning method. Sales are an important indicator in the business world because they can provide information about company performance, market trends, and support better decision-making. However, accurate and reliable prediction of sales data is often a complex challenge. In this study, the researchers collected historical sales data from Farhan Stores that included information about time, product, category, and price. This study also aims to apply data mining techniques to predict sales of Muslim clothes at Farhan stores using the C4.5 algorithm and the linear regression algorithm. The prediction method is used in this study and the calculations are performed using Google Collab. The results of the research that was conducted to predict sales of robes and shirts at Farhan Stores show that the best-selling item during the sales period from January to July 2022 was Sabiyan robes, which were the most sold item or can be said to be the Best Seller item at Farhan Stores. In this study, the parameters MAE (Mean Absolute Error), MSE (Mean Squared Error), and the R2 score are used to evaluate prediction performance. In the linear regression algorithm, the MAE value is 43,633.21, the MSE value is 4,005,924,352.66, and the R2 score is 0.94. Whereas in the C4.5 algorithm, the MAE value was 44,823.96, the MSE value was 50,233,775.14, and the R2 score was 0.94.

Keywords: *prediction, data mining, algorithm C 4.5, linear regression algorithm*

Cite as: A. Gustiana, "Comparative Analysis of Muslim Clothing Sales Predictions Using the C4.5 Method and Linear Regression," Journal of Systems Engineering and Information Technology, vol. 3, no. 1, pp. 30–36, Mar. 2024.
DOI: 10.29207/joseit.v3i1.5678

Received by the Editor: 2024-01-17
Final Revision: 2024-03-05
Published: 2024-03-07

1. Introduction

The Farhan store is a shop that specializes in the sale of Muslim clothing. In its business, this store is often faced with the problem of determining the stock of Muslim clothing and cannot overcome sales predictions in the next period. To be able to make the right decisions for predicting the sale of Muslim clothing that meets the needs of the owner and helps in the sales field, the researchers used a data mining technique using the C4.5 method and linear regression to suit sales needs and can overcome the accumulation of Muslim clothing stocks [1], [2]. In the previous study entitled "Application of Data Mining to Predict Wallpaper Sales Using the C4.5 Algorithm", data mining with the C4.5 algorithm can be implemented to determine the availability of e-Commerce goods with two categories, namely, goods are not available or not yet available to order goods for customers / backorders, and goods are available to order goods for customers. The results of determining the availability of e-commerce goods from this research application can help the company determine the status of goods that are ready to be traded.

This can be a recommendation for decision making in accepting orders so that the goods are maintained in stock and do not experience backorders. This is important so that customers maintain their trust in the conduct of goods transactions [3]. The large number of visitors who come to the Farhan Store often run out of Muslim clothing stocks, so when a customer buys or orders, the order is not available. This results in disappointment for the customer. Therefore, this research helps Farhan Store predict the stock of Muslim clothing for the next period, so that there are no shortages or stockouts and Muslim clothing on the menu is always available. In a previous study entitled Prediction of Jersey Sales Turnover Using the Linear Regression Method concluded that to maintain business stability and make plans for the following months [4]. Furthermore, in this research, the method used is linear regression and the data used are turnover history data for the past year. From the tests that have been carried out, almost every test produces the largest MAPE when used to predict turnover in November

2021, the MAPE produced is very large because there is a significant decrease in turnover from October to November.

However, in general, the resulting average is good because the MAPE is only 10.23201. That means Linear Regression is a fairly good method to use to predict turnover, especially in businesses whose turnover tends to be stable [4]. This study will use the C4.5 algorithm and linear regression, where the C4.5 algorithm is an algorithm developed by Ross. Quinlan, which is used to process data mining by forming a decision tree and the linear regression algorithm is composed based on patterns of data relationships relevant to the past [1], [5]. The independent variable X and the dependent variable Y are expressed by the regression of Y in X1 [6], [7].

In this study implemented using the C4.5 algorithm and simple linear regression to predict sales of Muslim clothing in the future period using sales data from the previous period. The choice of C4.5 and linear regression methods as prediction methods in this study is based on their advantages in estimating simple model parameters and data based on time series. In addition, this method can perform analysis using several independent variables (X) so that the prediction results can be more accurate [8]. Algorithm C4.5 is an algorithm used to form a decision tree. The C4.5 algorithm is a development of the ID3 algorithm; therefore, the C4.5 algorithm has the same working concept. The C4.5 algorithm includes a data mining method, which is the process of finding patterns by sorting large amounts of data using pattern recognition technology [9], [10].

Determine the decision tree in the C4.5 algorithm, there are 4 formulas as follows:

$$Entropy\,(S) += -\sum pi * log\ 2\ pi\ n\ i = 1 \quad (1)$$

Information:
S = Number of cases
n = Total set of S
pi = Proportion to i
Calculating the Gain value by doing the equation:

$$Gain\,(S, A) = (S) - \sum \frac{|Si|}{|s|} * Entropy(Si) n\ i = 1 \quad (2)$$

Information:
S = Number of cases
A = Character
n = Total A
Si = Number of cases on partition to i
s = Number of cases in S

Furthermore, the value of Split Info is calculated by the equation:

$$SplitInfo(S, A) = -\sum \frac{si}{s} log\ log\ 2 \frac{si}{s} ni = 1$$

Information:
S = Number of cases
A = Character
Si = Number of character samples i

What determines an attribute can be used as the root or branch of a decision tree is obtained from the Gain Ratio value with the equation:

$$Gain\ Ratio\ (S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

Information:
$Gain(S, A)$ = Gain info on character A
$SplitInfo(S, A)$ = Split info on characters A

Based on this background, the author will try to examine two problems, namely how to predict sales of Muslim clothing using the C 4.5 algorithm at Farhan's Store and how to predict Muslim clothing sales to predict the availability of Muslim clothing stocks using the Linear Regression algorithm at Farhan's Store.

2. Methods

The type of research used is quantitative in nature[11], where the research is based on the amount of data and will be examined in a structured and systematic manner. This research will also use a mathematical method so

that it is quantitative. Evaluation is needed to analyze and measure the precision of the results that have been achieved. In research, there are steps, steps so that what is set is achieved. Here are the steps that will be used.

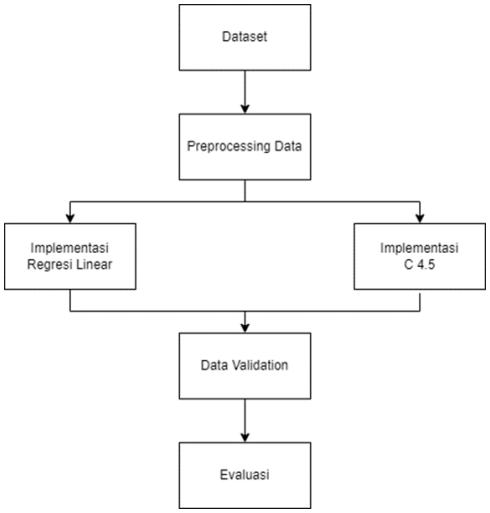

Figure 1. Research stages

The data that has been obtained needs to be pre-processed. Data preprocessing is something that must be done in the data mining process, because not all data or data attributes in the data are used in the data mining process. This process is carried out so that the data to be used are in accordance with the needs [12]. There are several types of preprocessing that will be used in this study, including. Linear regression is a regression method that is a statistical method that makes predictions using the development of mathematical relationships between variables, namely the dependent variable (Y) and the independent variable (X). The dependent variable is the effect variable or the affected variable, while the independent variable is the cause variable or the influencing variable. Predictions of the value of the dependent variable can be made if the independent variables are known.

3. Results and Discussion

In this study, the data collection method used in this study is the Farhan Store dataset. The period used in this data set is January - July 2022 in the form of excel. The amount of data used is 1501 data and has 9 attributes consisting of No, Date, Item Name, Category Name, Size, Price, Item Sold, Gross Sales, and Item Refunded. The attributes contained in the data can be presented in Table 1. The data set in Farhan's shop is as follows:

Table 1. Dataset at Farhan's Store

Attribute	Type	Information
No	*Int*	The number on the dataset
Date	*Datetime*	Purchase date
Item Name	*Object*	Muslim clothing name
Category Name	*Object*	Types of Muslim Clothing
Size	*Object*	Muslim Clothing Size
Price	*Int*	Prices of Muslim Clothing
Item Sold	*Int*	Muslim clothing sold
Gross Sales	*Int*	Total revenue from the results of items sold
Item Refunded	*Int*	Canceled transactions

3.1 Pre-Processing

In the data preprocessing stage, data cleaning will be performed on duplicated data and will remove columns 'No', 'Date', 'Item Name', 'Category Name', 'Size', and 'Item refunded' which is not used for the preprocessing stage.

```
data_stok_olah = data_stok_olah.drop(["No","Date", "Item Name", "Category Name", "Size", "Item Refunded"], axis=1)

data_stok_olah.columns
```

```
Index(['Price', 'Item Sold', 'Gross Sales'], dtype='object')
```

Then it will calculate total sales per month based on item sold, then it will sort from lowest to highest, and calculate total sales per month based on item sold and category name in the period January to July 2022.

```
data_stok_olah_totalPenjualanPerBulan = data_stok_olah.groupby(data_stok_olah["Date"].dt.strftime("%B")
                                                               )["Item Sold"].sum().sort_values()
data_stok_olah_totalPenjualanPerBulan
```

```
Date
February    170
June        267
March       349
May         416
January     454
July        534
April       591
Name: Item Sold, dtype: int64
```

From the results of the output above, it can be seen that sales in April were the highest sales compared to sales in other months, which reached 591 items sold. Meanwhile, in February, sales were the lowest among other months, namely only 170 items were sold.

3.2 Data Visualization

After pre-processing the data, this visualization data will see the sales trend of Muslim clothes along with the gross sales trend during the period January - July 2022. Next, compare the total number related to the menu categories sold at Farhan Stores during the period January - July 2022. After making comparisons, researchers will take into account the frequency of data in several ranges by looking at the distribution of data in items sold. Then look at the shape of the data distribution between the 2 variables (Independent Variable 'X' and Variable The dependent 'Y') is determined by the item type.

```
# Melihat trend penjualan baju muslim beserta trend penjualan kotor selama periode bulan Januari - Juli 2021
plt.figure(figsize=(30, 25))

plt.subplot(4, 3, 1)
sns.lineplot(x="Date", y="Item Sold", data=data_stok)
plt.title("Grafik Trend Terhadap Penjualan Baju Muslim \n yang ada di Toko Farhan Selama Bulan Januari - Juli 2021",
          fontsize=14, pad=14)
plt.xlabel("Periode Penjualan")
plt.ylabel("Jumlah Menu Terjual")

plt.subplot(4, 3, 2)
sns.lineplot(x="Date", y="Gross Sales", data=data_stok)
plt.title("Grafik Trend Terhadap Penjualan Kotor \n yang ada di Toko Farhan Selama Bulan Januari - Juli 2021",
          fontsize=14, pad=14)
plt.xlabel("Periode Penjualan")
plt.ylabel("Jumlah Penjualan Kotor")
plt.show()
```

The code above is a code to graph the trend of selling Muslim clothing and the trend of gross sales during the period from January to July 2022. Using the code above, a figure consisting of two subplots will be produced. The first subplot shows a graph of the trend of sales of Muslim clothes, while the second subplot shows a graph of the trend of gross sales. Each graph is equipped with a title, axis labels, and the data used comes from the data_stok variable.

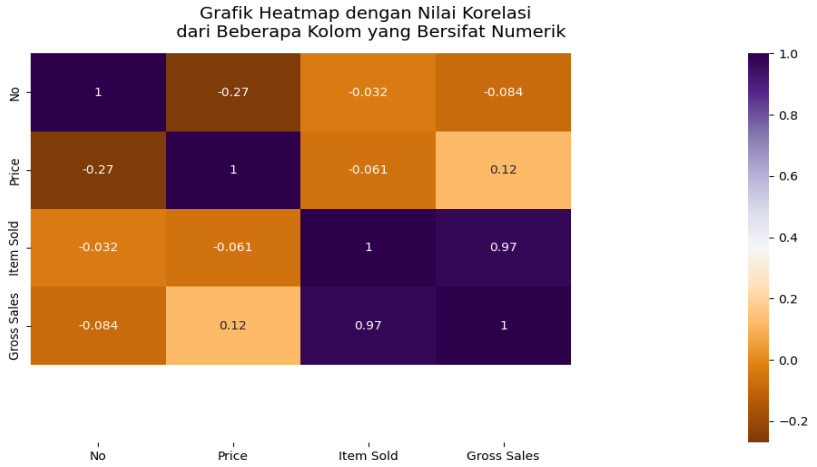

Figure 2. Heatmap

The above is a heatmap graphic that displays the mutual correlation between each column. The heatmap graph shows the correlation value or the connectedness value that has a reciprocal relationship between each column with the correlation value or the level of connectedness. Correlation values can only be observed in numerical data, whereas categorical data cannot display correlation values. Therefore, in the graph, focus only on the numeric data to see the correlation values in each column. An explanation of how to read and analyze the correlation values between columns or variables contained on the Y axis and the X axis in a heatmap is as follows:

- On the Y-axis the variable No and the X-axis on the variable Price have a low correlation value. Likewise, on the Y axis the variable Price and on the X axis the variable No have a correlation value of -0.27.
- On the Y-axis, the variable No and on the X-axis the variable No have a high correlation value. Likewise, on the Y axis, the variable 'Price' and on the X axis the variable 'Price' has a correlation value of 1.

In this study, an experiment was carried out that involved dividing the data into training data by 65% and data testing by 35%. To test performance, the parameters MAE (Mean Absolute Error), MSE (Mean Squared Error), and the R2 score are used. The following are the results of the tests using training data and testing data using linear regression and C 4.5 algorithms.

3.3 Linear Regression Algorithm

Based on the results of the graph below, it can be seen from the testing data and training data that it is positive between the X-axis and Y-axis. An increase in the value on the X-axis results in an increase in the value on the Y-axis. Both graphs are linear because the lines show an upward trend. It can also be seen that the two graphs have differences, although not so significant, between the two graph lines. The two graph lines show almost the same increase between the testing data and the training data, even though the graphic lines are not so perfect but the same straight up.

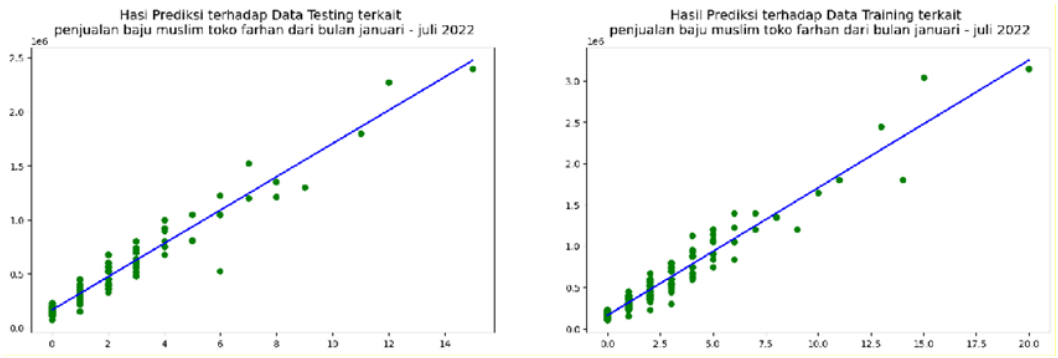

Figure 2. Linear Regression Algorithm

There is a difference between the two graph lines, although not so significant between the two graph lines. In data testing, there are several data points that are still clustered at certain points, although there are also several data points that are spread out, even though they are very small. On the training data, there are some data points that tend to be more clustered, although there are also some data points that are spread out. The resulting comparison shows that the data points in the training data are slightly spread out compared to the data points in the testing data. However, both still have linear and positive properties.

3.4 Results of Linear Regression Algorithm Analysis and Algorithm Analysis C4.5

The following are the results of the MAE, MSE, and R2 scores in the Linear Regression algorithm:

Table 2. Calculation results for MAE, MSE, R2 Score Accuracy using the Linear Regression Algorithm

	Data Testing	**Data Training**
MAE	43633.21	40805.98
MSE	4005924352.66	3830831325.48
R2Score	0.94	0.95

Based on the results of the MAE, MSE, and R2 scores of the linear regression Algorithm accuracy score, it can be seen that the results of the MAE, MSE, and R2 score testing and training. The results of the calculation of the MAE evaluation on the data testing are 43633.21, MSE on the data testing is 400592435.66, and R2Score on

the data testing is 0.94. Although the results of the calculation of the MAE on the training data is 40805.98, the MSE on the training data is 3830831325.48, and the R2Score on the training data is 0.95.

The following are the results of the MAE, MSE, and R2 scores in the C4.5 algorithm:

Table 3. Calculation results for MAE, MSE, R2Score Accuracy using the C4.5 Algorithm

	Data Testing	Data Training
MAE	44823.96	39148.97
MSE	5023377540.14	3030592966.30
R2Score	0.94	0.95

Based on the results of the MAE, MSE, and R2 scores of the linear regression Algorithm accuracy score, it can be seen that the results of the MAE, MSE, and R2 score testing and training. The results of the calculation of the MAE evaluation on the test data are 44823.96, the MSE on the test data is 5023377540.14, and the R2Score on the data test is 0.93. While the MAE calculation results on the training data are 39148.97, MSE on the training data are 3030592966.30, and the R2Score on the training data is 0.96.

4. Conclusions

This study tested the model by comparing two methods, namely the linear regression algorithm and the C4.5 algorithm, on Muslim clothing sales data at Farhan Stores. The results of this study were also tested using linear regression and C4.5 algorithms using MAE, MSE, and R2Score parameters and obtained standard results. In the linear regression algorithm, the MAE value is 43,633.21, the MSE value is 4,005,924,352.66, and the R2Score value is 0.94. While in the C4.5 algorithm, the MAE value is 44,823.96, the MSE value is 50,233,775.14, and the R2Score value is 0.94. Based on the results of the research, the best-selling item in the Farhan Store during the sales period from January to July 2022 was Sabiyan robes, with a total of 592 items sold. Therefore, it is important for the Farhan Store to maximize the stock of Sabiyan robe supplies. When paying attention to the number of items that are most in demand, Farhan Shop can manage inventory more efficiently. Meanwhile, the fewest items sold were Gliter robes, Polka Dot robes, Cadar robes, and Satin robes, with only one item sold each during the period January - July 2022.

From the two algorithms, namely the Linear Regression Algorithm and Decision Tree, Algorithm Linear Regression is considered superior to Decision Tree. Linear regression is a very effective and complex algorithm to find the relationship between independent and dependent variables. Using this method, we can predict future values. Suggestions for future researchers are to make predictions for robes at the Farhan Shop. This analysis can be performed during the period from August to December, based on the last update of the test script file, or within a full year, namely 2022, from January to December. The results of this analysis can be used as material for reports, discussions, or year-end meetings to create a new strategy to sell each new item in the coming 2023.

References

[1] M. A. Barata, E. Noersasongko, Purwanto, and M. A. Soeleman, "Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 226–235, 2023, doi: 10.29207/resti.v7i2.4687.

[2] I. Romli, F. Kharida, and C. Naya, "Determination of Customer Satisfaction of Tax Service Office Services Using C4.5 and PSO," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 2, pp. 296–302, 2020, doi: 10.29207/resti.v4i2.1718.

[3] G. Lukhayu Pritalia, "Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce," *Indonesian Journal of Information Systems*, vol. 1, no. 1, pp. 47–56, Aug. 2018, doi: 10.24002/ijis.v1i1.1727.

[4] R. G. F. Junior, N. Hidayat, and A. A. Soebroto, "Prediksi Omzet Penjualan Jersey menggunakan Metode Regresi Linier (Studi Kasus CV. Quattro Project Bululawang)," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 6, no. 10, 2022.

[5] D. M. Tarigan, D. P. Rini, and Samsuryadi, "Feature Selection in Classification of Blood Sugar Disease Using Particle Swarm Optimization (PSO) on C4.5 Algorithm," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 3, pp. 569–575, 2020, doi: 10.29207/resti.v4i3.1881.

[6] V. L. D. Pasaribu, F. Septiani, S. Rahayu, L. Lismiatun, and ..., *Forecast Analysis of Gross Regional Domestic Product based on the Linear Regression Algorithm Technique*. ceeol.com, 2021. [Online]. Available: https://www.ceeol.com/search/article-detail?id=955135

[7] D. Suryani, M. Fadhilla, and A. Labellapansa, "Indonesian Crude Oil Price (ICP) Prediction Using Multiple Linear Regression Algorithm," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 6, pp. 1057–1063, 2022, doi: 10.29207/resti.v6i6.4590.

[8] O. J. Ababil, S. A. Wibowo, and H. Zulfia Zahro', "PENERAPAN METODE REGRESI LINIER DALAM PREDIKSI PENJUALAN LIQUID VAPE DI TOKO VAPOR PANDAAN BERBASIS WEBSITE," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 1, pp. 186–195, Mar. 2022, doi: 10.36040/jati.v6i1.4537.

[9] T. H. Sinaga, A. Wanto, I. Gunawan, S. Sumarno, and Z. M. Nasution, "Implementation of Data Mining Using C4.5 Algorithm on Customer Satisfaction in Tirta Lihou PDAM," *Journal of Computer Networks, Architecture, and High-Performance Computing*, vol. 3, no. 1, 2021, doi: 10.47709/cnahpc.v3i1.923.

[10] Muhasshanah, M. Tohir, D. A. Ningsih, N. Y. Susanti, A. Umiyah, and L. Fitria, "Comparison of the performance results of c4.5 and random forest algorithm in data mining to predict childbirth process," *CommIT Journal*, vol. 17, no. 1, 2023, doi: 10.21512/commit.v17i1.8236.

[11] J. W. Creswell, *Research design : qualitative, quantitative, and mixed methods approaches*. California: SAGE Publications, 2014.

[12] I. Djamaludin and A. Nursikuwagus, "ANALISIS POLA PEMBELIAN KONSUMEN PADA TRANSAKSI PENJUALAN MENGGUNAKAN ALGORITMA APRIORI," *Simetris : Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 8, no. 2, p. 671, Nov. 2017, doi: 10.24176/simet.v8i2.1566.