Contents lists available at IdNNS

Journal of Intelligent Systems for Community Development

Volume 1 ; Issue 1

journal homepage: https://journal.idnns.org/index.php/jiscode

Food and Beverage Sales Prediction Using Linear Regression and Random Forest Regression at Ayam Serayu Restaurant Bekasi

Nabila Ramadhani Sari[a,*], Herlawati[a] and prima dina atika[a]

[a]*Program Study; Institution, Institution address, Bekasi, West Java, Indonesia*

ARTICLE INFO

Keywords:
Sales Prediction
Linear Regression
Random Forest Regression
Data Mining

ABSTRACT

The large number of sales transactions at Ayam Serayu Restaurant Bekasi creates challenges in managing and analyzing sales data. Manual processes make it difficult to predict future sales, affecting inventory management and decision-making. Therefore, an accurate prediction method is needed. This study applies Linear Regression and Random Forest Regression to predict sales based on historical data. The research stages include data collection, preprocessing, modeling, and evaluation using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results show that Random Forest Regression provides better accuracy than Linear Regression. The resulting model is expected to improve operational efficiency and support decision-making at Ayam Serayu Restaurant Bekasi.

1. Introduction

The development of information technology has had a major impact on various business sectors, including the culinary industry. Daily sales transaction data can be utilized to support decision-making processes, particularly for predicting food and beverage sales. Statistical learning methods provide a foundation for building predictive models from historical observations and for evaluating their predictive performance (James et al., 2021).

Ayam Serayu Restaurant Bekasi has a relatively high number of sales transactions with fluctuating sales patterns. Manual sales-data management makes it difficult to estimate sales in subsequent periods. This condition can affect raw-material inventory management and operational efficiency.

Machine Learning is one method that can be used to perform predictions based on historical data. This study uses Linear Regression and Random Forest Regression. Linear Regression is used to identify linear relationships among variables, whereas Random Forest Regression can handle more complex and non-linear data patterns (James et al., 2021; Breiman, 2001).

Previous studies have indicated that Random Forest Regression can perform well for sales prediction because it can achieve high accuracy compared with conventional methods. However, research on food and beverage sales prediction in restaurants still needs to be further developed using more complex transaction data.

Based on these problems, this study aims to apply and compare Linear Regression and Random Forest Regression for predicting food and beverage sales at Ayam Serayu Restaurant Bekasi. The results are expected to assist raw-material inventory management and support data-driven decision-making.

2. Research Method

This study uses the Cross Industry Standard Process for Data Mining (CRISP-DM) as the research framework. CRISP-DM provides a structured process for data-mining projects and comprises six main stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000; Shearer, 2000).

The dataset consists of historical sales transactions from Ayam Serayu Restaurant Bekasi over a certain period, containing approximately 26,610 transaction records. The data include transaction date, product name, product category, price, number of items sold, and total sales.

The research framework describes the process from problem identification to the resulting food and beverage sales predictions. The study begins by identifying the problem of manual sales prediction, which is considered less optimal for supporting inventory management and business decision-making. Data are then collected through observation, interviews, and literature study.

The collected data undergo preprocessing, including data cleaning, data transformation, and format adjustment according to model requirements. The next stage is modeling

*Corresponding author: xxxx@xxxx.xxx
ORCID(s):

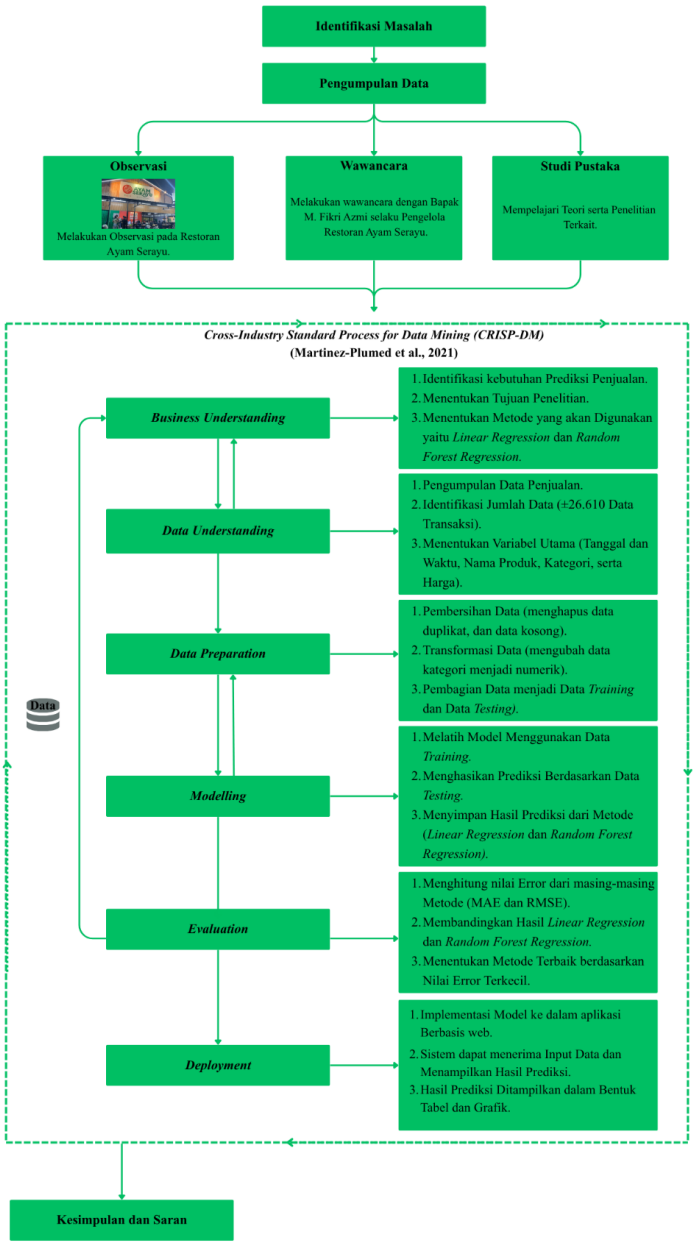

Figure 1: Research framework.

using Linear Regression and Random Forest Regression. Linear Regression is used to identify linear relationships among variables, whereas Random Forest Regression is used to handle more complex and non-linear patterns.

After the models are built, evaluation is conducted using MAE and RMSE to measure prediction error. The evaluation results are analyzed to compare the performance of the two algorithms and determine the best method. The final stage is drawing conclusions based on the analysis results.

3. Results and Discussion

This study aims to develop a sales prediction model for Ayam Serayu Restaurant Bekasi by applying Linear Regression and Random Forest Regression. The research process

follows the CRISP-DM stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

3.1. Business Understanding

The Business Understanding stage was conducted to understand the problems at Ayam Serayu Restaurant Bekasi related to sales-data management. Based on observation, the large volume of transaction data had not been optimally utilized for sales analysis and prediction. This condition makes it difficult for the restaurant to estimate sales and consequently affects raw-material inventory management and production planning.

Therefore, this study aims to build a sales prediction model using Linear Regression and Random Forest Regression based on historical transaction data. The resulting model is expected to help the restaurant improve inventory-management effectiveness and support data-driven decision-making.

3.2. Data Understanding

The Data Understanding stage was conducted to understand the structure and characteristics of the sales-transaction dataset. The activities included reading the dataset, checking its dimensions, analyzing data types, identifying missing values, and conducting descriptive statistical analysis.

3.2.1. Reading Sales Transaction Data

At the initial stage, the sales-transaction dataset was read using the pandas library. The dataset was loaded from a CSV file and stored in the variable `df_raw`. pandas is a widely used Python library for practical data manipulation and analysis (McKinney, 2010).

3.2.2. Displaying a Data Sample

After the dataset was successfully loaded, an initial review was performed by displaying the first five rows using the `head()` function.

The first five rows allow the researcher to understand the dataset content and identify possible data-format errors before preprocessing.

3.2.3. Data-Type and Missing-Value Analysis

The next stage analyzed the data type of each column and identified missing values.

This stage is important because missing data can affect the results of the Machine Learning prediction model.

The output presents information about data types and the number of missing values in each dataset column, including integer, float, and object attributes.

3.2.4. Descriptive Statistical Analysis

The final activity in Data Understanding was descriptive statistical analysis of numerical data.

The descriptive statistics help identify general sales patterns and possible extreme values or outliers that may influence prediction-model performance.

3.3. Data Preparation

Data Preparation involves preparing the data before entering the modeling stage.

3.3.1. Data Cleaning

The data-cleaning stage includes cleaning numerical formats, checking invalid data, and normalizing product categories so that the data become more consistent.

The purpose of this process is to ensure that the system can correctly process time-related data. The resulting cleaned data have consistent numerical formats and are ready for data transformation and modeling.

3.3.2. Data Transformation

Data Transformation and Feature Engineering were conducted to convert raw data into more informative features so that the Machine Learning model can better recognize sales patterns.

The sales distribution by transaction hour indicates the busiest operating periods. This exploration provides an overview of the sales pattern in the restaurant.

3.3.3. Feature Engineering

Feature Engineering was conducted by creating additional features so that the model can understand historical sales patterns.

The heatmap shows the level of correlation among features and the sales target. A higher correlation value indicates a stronger relationship between the feature and the prediction target.

3.3.4. Train–Test Split

The Split Data stage divides the dataset into training and testing data.

3.4. Modeling

Modeling is the process of constructing prediction models using the prepared data.

3.4.1. Selecting the Methods

Two Machine Learning algorithms were used in this study: Linear Regression and Random Forest Regression, both for predicting the number of sales. Linear Regression is a standard statistical learning approach for modeling the relationship between a response variable and one or more predictors, while Random Forests construct an ensemble of tree-based predictors using randomized feature selection (James et al., 2021; Breiman, 2001).

The two methods are subsequently compared to determine the model with the best performance.

3.4.2. Model Training

Training was conducted so that the models could learn sales patterns from the training data.

The training process is used to train the Linear Regression model with the training data so that it can learn the relationship between sales features and the sales target.

3.4.3. Model Testing

Testing was conducted to evaluate the ability of the models to predict new data that had not previously been observed.

The testing data are used to evaluate model performance. The predictions are subsequently compared with actual data to calculate the model error.

3.5. Evaluation

The Evaluation stage tests the Machine Learning models to determine their prediction accuracy.

3.5.1. Calculating Predictions

This stage calculates model predictions and compares them with actual sales. MAE and RMSE are used to measure prediction errors for Linear Regression and Random Forest Regression.

Lower MAE and RMSE values indicate better prediction performance. In this study, RMSE is used as the primary criterion for selecting the final model, while MAE is retained as a complementary measure (Chai and Draxler, 2014).

3.5.2. Calculating MAE and RMSE Error

This stage compares the prediction errors of Linear Regression and Random Forest using a bar-chart visualization. The chart makes it easier to identify the model with the smallest error. MAE and RMSE are commonly used complementary measures for assessing prediction error, and their interpretation depends on the characteristics of the error distribution (Chai and Draxler, 2014).

The graph shows that the model with the smallest error has the best prediction performance.

3.5.3. Analyzing Evaluation Results and Comparing Model Performance

The best model is determined based on the evaluation results. The model is selected according to the smallest RMSE because it is considered to have the lowest prediction error.

The selected model is subsequently used during deployment to predict restaurant product stock requirements.

3.6. Deployment

During Deployment, the Machine Learning model that has completed training and evaluation is implemented in a web-based system.

3.6.1. System Dashboard

The dashboard is the main page displayed when users open the web application. It provides a summary of information related to the sales prediction system.

The dashboard presents important information such as the total number of transaction records, the number of food categories, the number of beverage categories, and the Machine Learning methods used, namely Linear Regression and Random Forest Regression. This information helps users understand the general characteristics of the data used by the system.

3.6.2. Data Input Menu

The data-input menu is used to add new sales transaction data to the system. The page is designed to allow users to manage sales data directly through the web application.

The form contains several fields, including transaction date, product category, product name, quantity sold, and product price.

3.6.3. Prediction Process Menu

On this page, users can configure model parameters before running the prediction process.

The left side of the page provides parameters that can be adjusted by the user, such as test-data size (test size) and the number of decision trees ($n_{estimators}$) in the Random Forest Regression method. The implementation uses standard Python machine-learning tooling for supervised learning and model evaluation (Pedregosa et al., 2011).

3.6.4. Stock Prediction Results

The stock-prediction results menu displays sales predictions and product-stock recommendations based on the selected Machine Learning model.

The page displays important information such as product name, latest sales quantity, average sales over the previous seven days, predictions using Linear Regression, predictions using Random Forest Regression, predicted sales for the following day, and additional stock recommendations. The system automatically selects the best model based on the smallest RMSE. The prediction from the selected model is then used to determine the recommended stock quantity for the following day.

```
from google.colab import files

print("📂 Silakan upload file CSV transaksi Pawoon...")
uploaded = files.upload()

# Ambil nama file yang baru diupload
CSV_FILE = list(uploaded.keys())[0]
print(f"\n✅ File berhasil diupload: {CSV_FILE}")
```

Figure 2: Reading sales transaction data.

```
print("📄 5 Baris pertama data mentah:")
df_raw.head()
```

Figure 3: Displaying a data sample.

```
print("📊 Informasi tipe data & missing values:\n")
info_df = pd.DataFrame({
    'Tipe Data': df_raw.dtypes,
    'Non-Null': df_raw.count(),
    'Null Count': df_raw.isnull().sum(),
    'Null %': (df_raw.isnull().sum() / len(df_raw) * 100).round(2)
})
print(info_df.to_string())
```

Figure 4: Analyzing data types and missing values.

```
📊 Informasi tipe data & missing values:

                    Tipe Data  Non-Null  Null Count  Null %
Tanggal & Waktu        object     26610           0    0.00
ID Struk               object     26610           0    0.00
Status Pembayaran      object     26610           0    0.00
ID / Kode Outlet       object     26610           0    0.00
Outlet                 object     26610           0    0.00
Tipe Penjualan         object     26610           0    0.00
Kasir                  object     26610           0    0.00
No. Hp Pelanggan      float64         0       26610  100.00
Nama Pelanggan        float64         0       26610  100.00
SKU                   float64         0       26610  100.00
Nama Produk            object     26610           0    0.00
Kategori               object     26610           0    0.00
Jumlah Produk           int64     26610           0    0.00
Harga Produk           object     26610           0    0.00
Penjualan Kotor        object     26610           0    0.00
Diskon Produk           int64     26610           0    0.00
Subtotal               object      5447       21163   79.53
Diskon Transaksi       object      5447       21163   79.53
Service Charge        float64      5447       21163   79.53
```

Figure 5: Results of data-type and missing-value analysis.

```
print(df_raw.describe().to_string())
```

Figure 6: Initial descriptive statistical analysis.

```
def clean_number(val):
```

Figure 7: Number-format cleaning.

```
df['tanggal']   = df['_tgl'].dt.date
df['jam']       = df['_tgl'].dt.hour
df['hari']      = df['_tgl'].dt.day_name()
df['bulan_nm'] = df['_tgl'].dt.month_name()
```

Figure 8: Extracting transaction date and time information.

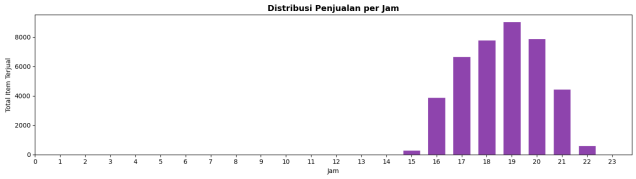

Figure 9: Sales distribution by transaction hour.

```
FEATURES_ITEM = [
```

Figure 10: Defining the feature list.

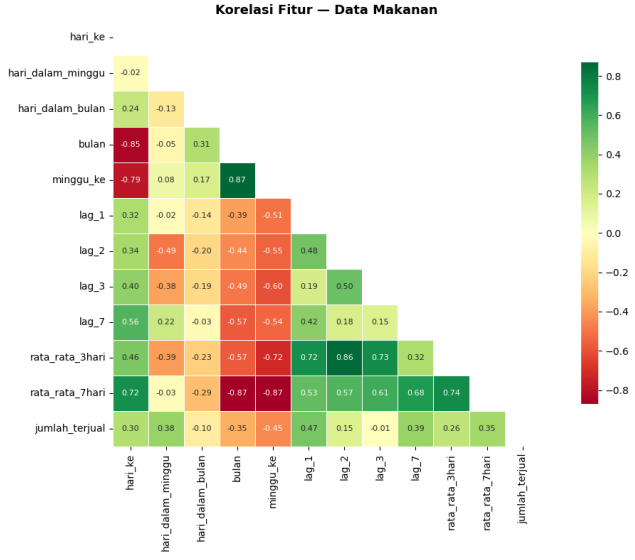

Figure 11: Correlation among features.

```
train = feat_df.iloc[:n_train].copy()
test  = feat_df.iloc[n_train:].copy()
```

Figure 12: Training and testing data split.

```
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
```

Figure 13: Selecting the Machine Learning methods.

```
lr = LinearRegression()
lr.fit(X_train, y_train)
```

Figure 14: Model training process.

```
pred_lr = lr.predict(X_test)
```

Figure 15: Model testing process.

```
mae  = mean_absolute_error(y_true, y_pred)
rmse = np.sqrt(mean_squared_error(y_true, y_pred))
```

Figure 16: Calculating prediction results.

```
bars1 = ax.bar(x - width/2, [data_plot['Makanan']['Linear Regression'],
                             data_plot['Minuman']['Linear Regression']],
               width, label='Linear Regression', color='#3498DB', edgecolor='white')
bars2 = ax.bar(x + width/2, [data_plot['Makanan']['Random Forest'],
                             data_plot['Minuman']['Random Forest']],
               width, label='Random Forest', color='#E74C3C', edgecolor='white')
```

Figure 17: Calculation and visualization of MAE and RMSE.

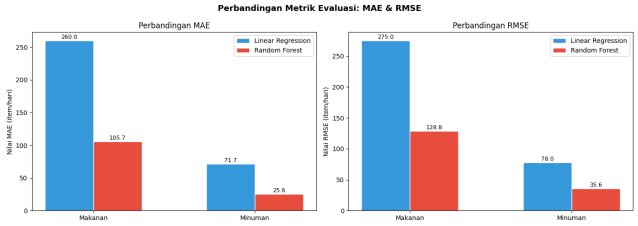

Figure 18: Evaluation metric results.

```
best    = 'Random Forest' if rf_row['RMSE'] <= lr_row['RMSE'] else 'Linear Regression'
```

Figure 19: Determining the best model.

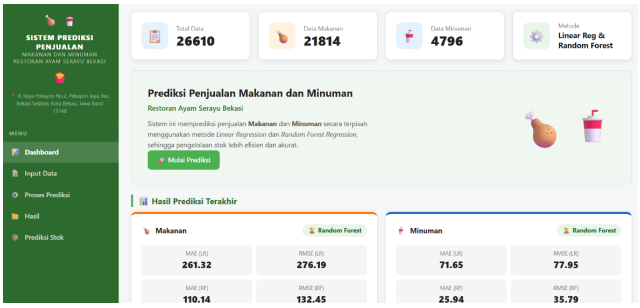

Figure 20: System dashboard.

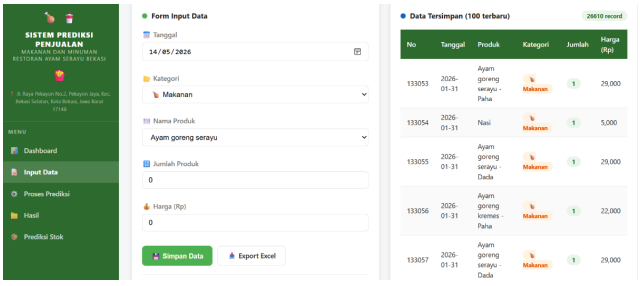

Figure 21: Data input page.

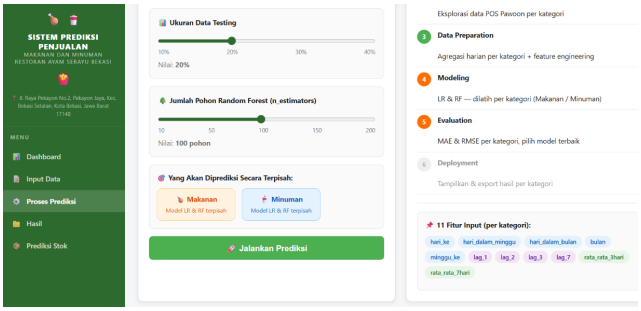

Figure 22: Prediction process page.

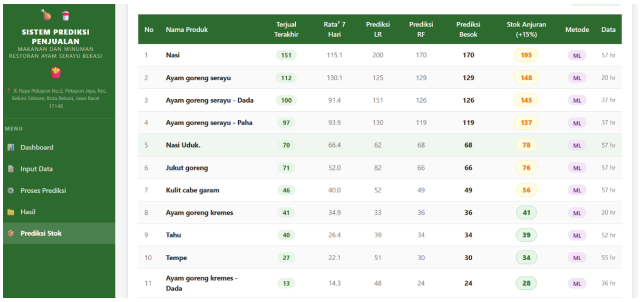

Figure 23: Stock prediction results page.

4. Conclusion

Based on the research results, Linear Regression and Random Forest Regression were successfully applied to predict food and beverage sales at Ayam Serayu Restaurant Bekasi using historical transaction data. The results indicate that Random Forest Regression performs better than Linear Regression because it produces lower prediction errors and predictions that are closer to the actual data. In addition, the prediction model was successfully implemented in a web-based system, which can help the restaurant manage inventory, plan sales, and make decisions more effectively.

Acknowledgments

The source manuscript does not provide an acknowledgments section. No additional acknowledgment content has therefore been introduced in this conversion.

References

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Chai, T., Draxler, R.R., 2014. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. Geoscientific Model Development 7, 1247–1250. doi:10.5194/gmd-7-1247-2014.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0: Step-by-Step Data Mining Guide. SPSS, Chicago, IL, USA.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An Introduction to Statistical Learning: With Applications in R. 2 ed., Springer, New York, NY, USA. doi:10.1007/978-1-0716-1418-1.

McKinney, W., 2010. Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, pp. 51–56. doi:10.25080/Majora-92bf1922-00a.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830.

Shearer, C., 2000. The crisp-dm model: The new blueprint for data mining. Journal of Data Warehousing 5, 13–22.