Estimation Of Gourami Supplies Using Gradient Boosting Decision Tree Method

By Made Sukarsa

WORD COUNT 3270
TIME SUBMITTED 23-NOV-2020 10:45PM
PAPER ID 65499385

Estimation Of Gourami Supplies Using Gradient Boosting Decision Tree Method

I Made Sukarsa [1], Ngakan Nyoman Pandika Pinata [2], Ni Kadek Dwi Rusjayanthi [3]

Abstract **– The need for food supplies are very crucial in a food business, therefore it is necessary to estimate the right supplies to maximize profit. One of the methods to determine these is by looking for patterns and forecasting transaction data. The purpose of this research is to estimate the gourami supplies using transaction data to forecast using the gradient boosting decision tree method from XGBoost. The transaction data used comes from Restaurant X with a time period from 2016 to 2019. A measurement error rate of the model using MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). This study tried five XGBoost models with different features such as lag, rolling window, mean encoding, and mix. The results of this study indicate that the mixed feature model produces an accuracy of 97.54% with an MAE of 0.63 and a MAPE of 2.64%.**

Keywords **– Forecasting, Time Series, GBDT, XGBoost, Gouramy Inventory.**

1. Introduction

Gourami (Osphronemus goramy) is a type of freshwater fish native to Indonesia that has long been cultivated and consumed by the public because of its delicious meat taste, so it has a high economic value. Therefore, many food businesses sell gourami [1].

Business players in business competition use information technology to be able to analyze transactions carried out by consumers, including forecasting transaction data so that they can provide useful information for the company [2].

Most sales transaction data are only used as archives without being properly utilized, even though these data sets contain information that can be very useful [3]. Cases of data imbalance with the resulting information are usually referred to as the data-rich but information poor phenomenon where organizations have excess data from business processes that are captured and stored but still lack of information [4]. Finding new information, patterns or certain rules in a large amount of data can be done using data mining techniques [5].

Data mining also has many methods and categories, in this study, it focuses on forecasting. Forecasting future data based on analysis of past data is an important way to properly explore data for decision making [6], such as sales in the retail industry. Forecasting sales transactions can have a crucial impact on the success and performance of companies. shortages of inventory and excess inventory that results in losses for the companies usually occurs due to inaccurate forecasts [7].

Researchers made predictions at a restaurant X that sells a special menu of gourami. Restaurant X has a problem in the supplies namely, the supply exceeds demand or is unable to fulfill demand. Uncertainty in sales requires restaurant X to develop a strategy to maximize profits. One of the methods to determine these is by looking for patterns and forecasting data transactions.

Forecasting on sales data is usually performed with a statistical-based methods approach such as research that conducts short-term forecasting of international tourism demand with the F.Y.R. case study Macedonia uses the ARIMA method [8]. Machine learning approaches have also been widely used such as research on forecasting outdoor air temperature and humidity using XGBoost which uses data from Shenzhen, China with excellent model results. [9]. According to [10] doing a sales prediction using several data mining techniques such as linear regression, random forest, and XGBoost, the best result is the XGBoost method. Traffic flow forecasting for one day using the Gradient Boosting Decision Tree (GBDT) method is very effective with a MAPE error rate of 0.097% [11]. Chronic disease classification uses several methods such as logistic regression, random forest, support vector machine (SVM), and gradient boosting. The highest accuracy is obtained using the gradient boosting method of 88.83% [12]. XGBoost also produces the best forecast and fast execution times than several other methods such as ARIMA, SARIMA, and Random Forest in load forecasting [13].

This study applies the decision tree gradient boosting method in forecasting because based on previous research, it is found that the use of the gradient boosting method produces very good accuracy. Forecasting is done on transaction data for gourami and uses the XGBoost library. The results of this study are expected to provide recommendations in purchasing raw gourami or determining the supply of raw gourami.

2. Literature Review

2.1 *XGBoost*

XGBoost is a machine learning ensemble algorithm that is based on gradient boosting. XGBoost has been used in several studies, particularly in data and machine learning competitions. XGBoost has shown great results over the other methods. XGBoost can be used in regression or classification problems, such as prediction of store sales, prediction of customer behavior, prediction of a number of ad clicks, prediction of risk of a case, classification of web text, and classification of malware. [14].

XGBoost requires a number of parameters to choose from. Some of the parameters in XGBoost are as follows [15].

1) **eta**, the learning rate which serves to prevent the model from overfitting.
2) **gamma,** to determine tree pruning when the resulting split experiences a positive reduction in the loss function.
3) **max depth,** maximum depth of tree.
4) **min child weight,** the minimum weight a node should have. If the minimum is not met, splits in a tree will not take place.
5) **subsample**, to use the entire row of data (default 1) while 0.5 means randomly using half of the rows of data.
6) **colsample bytree,** to use the entire column in the data (default 1) while 0.5 means using half of the existing columns.

2.2 *Evaluation Metrics*

A forecast model can be said to be accurate or not, there are two aspects that must be considered. First, it is known that the model can follow the training data pattern, and second, how well the model predicts the test data. One way to evaluate a forecast model is to calculate the Mean Absolute Percentage Error (MAPE). MAPE is one of the most used error metric for measuring regression model [16]. Although MAPE is very sensitive to outliers, MAPE is still very widely used because it is so easy to interpret [17].

MAPE is the average percentage of the absolute result of the difference between the actual value and the predicted value divided by the actual value [18]. MAPE can be found using the formula in Equation (1).

$$MAPE = \frac{100}{n} \Sigma \left| \frac{A_t - y_t}{A_t} \right| \quad (1)$$

The explanation of Equation (1) is as follows:
A_t = actual value on data t
y_t = predicted value on data t
n = the amount of data

A good regression model is a model that has a MAPE smaller than 10% [19] dan juga kriteria nilai MAPE according to [20] is divided into 4 of bad, reasonable, good, and excellent. The range of MAPE values and criteria are shown in Table 1.

Table 1. MAPE score criteria

MAPE	Criterion
<10%	Excellent
10% - 20%	Good
20% - 50%	Reasonable
>50%	Bad

Researchers also use another measure, namely MAE (Mean Absolute Error) which is a simpler value than MAPE. Similar to MAPE, MAE is the absolute value of the difference between the actual value and the predicted value divided by the amount of data expressed in Equation (2).

$$MAE = \Sigma \left| \frac{A_t - y_t}{n} \right| \quad (2)$$

3. Methodology

3.1 *Datasets*

The data used to forecast are the sales transactions of Restaurant X in the period January 2016 to December 2019. A snapshot of the transaction data can be seen in Table 2.

Table 2. Transaction data of Restaurant X

Date	NoTransaksi	Qty	NamaItem
14/01/2016	KS160100002	1	Soup Gurame
14/01/2016	KS160100002	1	Gurame 400 gr
15/01/2016	KS160100006	1	Orange Juice
15/01/2016	KS160100006	1	Soda Gembira
15/01/2016	KS160100006	2	Udang Panggang
15/01/2016	KS160100009	1	Gurame 350 gr

The data is received from the Restaurant X cashier program. A lot of data obtained is around 182,756 data with various food menus, but the focus here is the menu that uses gourami.

3.2 *Research Flow*

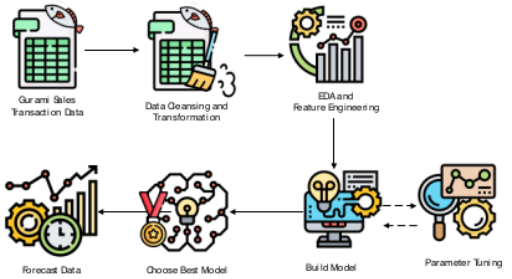

Figure 1. Research Flow

Figure 1. is an overview of this study. The first step is to load gourami sales transaction data in excel file format. Furthermore, data cleaning is carried out, such as deleting unused columns, deleting zero transactions, and transforming data into kilograms. Clean data is used in the EDA (Exploratory Data Analysis) process to see what information the data can provide by visualizing. The next step is performing feature engineering, namely the process of converting data into better features to increase the accuracy of the model.

Model building and parameter configuration are done by trying various models with features that have been made at the feature engineering stage such as lag, rolling window, and mean encoding, then looking for the optimal parameter configuration. All the models that have been built are compared to the MAE and MAPE error values from the predicted results of each model, then a model with a low MAE and MAPE is selected. The last step is if the best model has been found, the model is used to forecast unseen data.

3.3 Data Cleansing

The first step of data cleansing is loading the data then sorting out the data, which ones will be used, and removing those that are not used. The transaction data is then converted into kilograms to make it easier to provide suggestions from the prediction results. The next step is to check the data whether there is a zero sales if it is true that the transaction will be deleted.

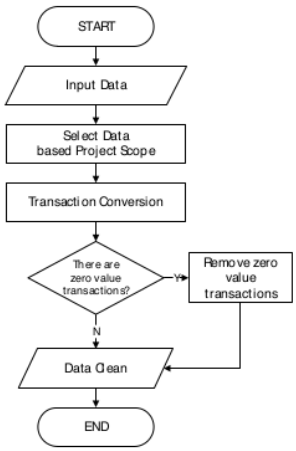

Figure 2. Flowchart Data Cleansing

3.4 Exploratory Data Analysis

Exploratory data analysis (EDA) is a critical first step in analyzing data. EDA is usually carried out by analyzing data descriptively [21]. Exploratory data analysis is used to obtain information from the data, find out the distribution of data, data outliers, the correlation between variables and to obtain the underlying assumptions that occur.

Most EDA techniques are visualizing data. The main reason for visualizing that they make it easy to see data distribution, trends, correlations, and outliers. Visualizations that are often used in EDA are boxplots, standard deviation plots, mean plots, and comparisons between variables in the data [22].

3.5 Build and Testing Model

The first step in model building process is the initiation of the XGBoost model then inputting training data, test data, and parameters such as max_depth, min_child_weight, eta, subsample, colsample_bytree, and gamma. The next process is training the XGBoost model. Details of the model training process can be seen in Figure 3.8. Model evaluation is done by measuring MAE and MAPE. If the error results are still high enough, change the parameters until the best parameter is found. The best model is selected, then visualized to see the trend between the predicted results and the actual data.

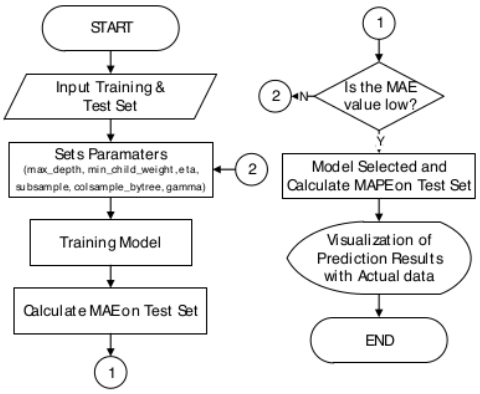

Figure 3. Flowchart Build and Testing Model

4. Implementation and Results

The implementation is carried out in the Kaggle private kernel with the Python version 3. Implementation is done in the Kaggle kernel to speed up computation. The libraries used include the following.

1) Numpy, to do math calculations like matrices and algebra
2) Pandas, for da tabulated
3) Plotly, data visualization
4) Sklearn, calculation of MAE and MAPE error metrics
5) XGBoost, machine learning to build model based on gradient boosting decision tree.

The first implementation is data cleansing, which includes selecting data based on project scope, conversion of transactions, and eliminating zero transactions. Second, EDA and finally building and testing models.

4.1 Select Data based Project Scope

The results of the special data selection process for the gourami menu are as follows.

Date	qty	NoItem	NamaItem
2016-01-16 13:19:22	1.0	F4	Gurame 400 gr
2016-01-16 13:19:22	1.0	F9	Soup Gurame
2016-01-16 13:29:00	3.0	F3	Gurame 350 gr
2016-01-16 13:56:12	1.0	F4	Gurame 400 gr
2016-01-16 14:04:49	1.0	F4	Gurame 400 gr

Figure 4. Gourami Menu Transactions

The menu with raw gourami ingredients is the number item F1, F2, F3, F4, F5, F6, F7, F8, F9. The total data that has the gourami menu is 37,010 data.

4.2 Transformation

There are two stages in data transformation, namely the transformation of transaction units into kilograms and the transformation of transaction times into daily. The results of the transformation of the gourami transaction data are as follows.

Date	qty	kilo
2016-01-16	37.0	12.85
2016-01-17	30.0	10.10
2016-01-18	35.0	11.40
2016-01-19	26.0	8.50
2016-01-20	20.0	6.65

Figure 5. Result of Transformation Data

The result of the transformation of the kilogram unit contained in the kilo column on Figure 5. is obtained by multiplying each row of the qty column value by the kilogram unit owned by the menu. Daily transformations are obtained by daily resampling using the Pandas library.

4.3 Delete Zero Sales

The zero-value transaction write-off stage is carried out to make the model work better. The results of this stage obtained 79 zero transactions. The 79 transactions were later deleted. So the total rows of data after cleaning this data are 1367 rows of data with daily intervals.

4.4 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) stage was carried out by the researcher by visualizing the transaction data with various types of visualization including line charts, bar charts, and box plots. Visualization of the need for gourami according to transactions that occurred from 2016 to 2019.

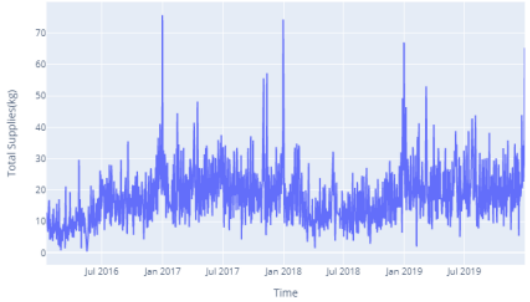

Figure 6. History Transactions

In this visualization, it is known that transactions always experience a spike at the end of each year. The distribution of gourami needs with histogram visualization is as follows.

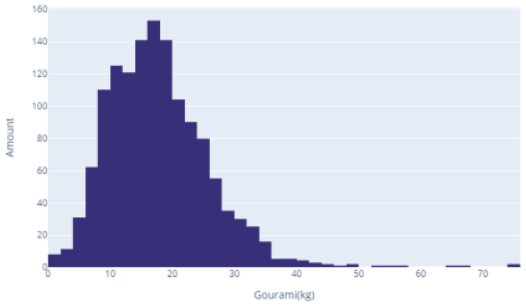

Figure 7. Distribution of Gourami Transaction

Based on the histogram in Figure 7., it is known that there is an abnormal distribution, there are several transactions that are far from the distribution. Transaction distribution top 3 ranges from 14-20kg. The initial assumption of the abnormal distribution is an outlier at the end of the year.

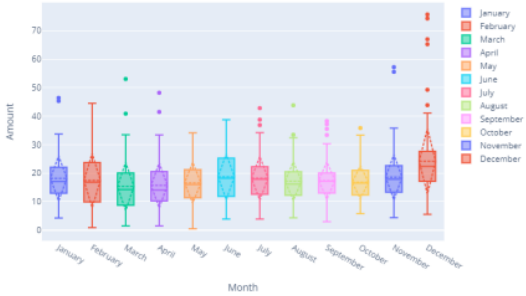

Figure 8. Boxplot of Gourami Transaction Monthly

It turns out that from the boxplot graph the researcher can see data outliers from each month, outliers are data that are far from the distribution of other data, therefore in certain months that have high outliers results in a high mean value, in other words, the mean is influenced by outliers. The initial assumption that the abnormal distribution on the histogram is reinforced by the outliers that occurred in December which can be seen in the boxplot visualization. Some of the other visualizations are as follows.

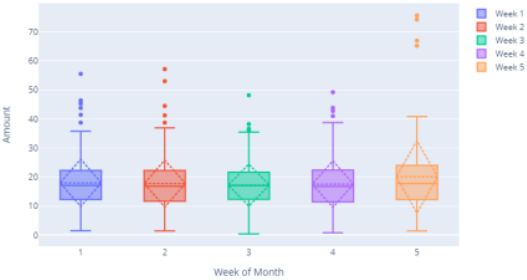

Figure 9. Boxplot of Gourami Transaction Week of Month

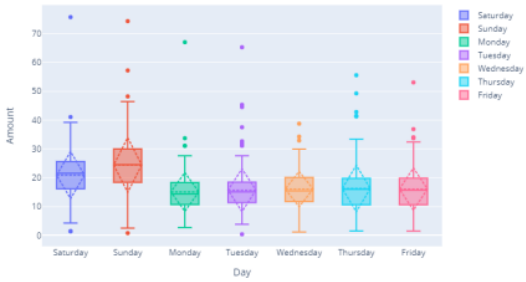

Figure 10. Boxplot of Gourami Transaction Daily

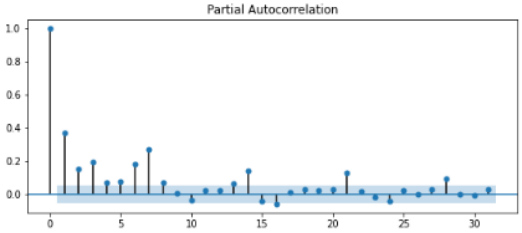

Figure 11. PACF of Gourami Transaction

The information obtained from the exploratory data analysis on the transaction data of gourami at Restaurant X is as follows:

1) The use of gourami supplies in 2018 has decreased.
2) Every year in December the use of gourami supplies is always high.
3) The use of gourami supplies tends to be high at the end of December.
4) Every week in every month tends to be stable.
5) Early to the end of the month, the use of gourami supplies is quite stable, except in December.
6) Weekends of gourami supplies usage are always high.
7) The use of gourami supplies may be influenced by the use of raw materials from yesterday to one week ago.

4.5 Building and Testing Models

The building and testing models are carried out by splitting the data into training sets and test sets. Training set with 1351 data from 16 January 2016 to 15 December 2019 and a test set with 16 data from 16 December 2019 to 31 December 2019.

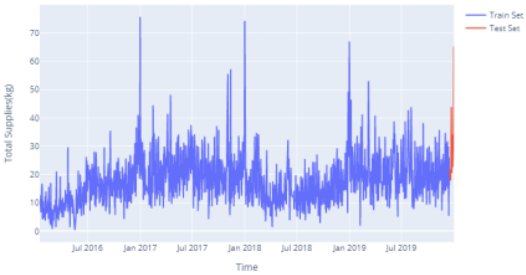

Figure 12. Data Splitting

The models are built with different kinds of features such as date feature, lag feature, mean rolling window feature, mean encoding feature, and combined feature. The description of these features is as follows.

1) Date

Date feature is done by taking components from the time series itself, such as the day component of the year, day of the month, month, and year.

Date	dayofmonth	dayofweek	weekofmonth	month	dayofyear	weekofyear	year
2016-01-16	16	5	3	1	16	2	2016
2016-01-17	17	6	3	1	17	2	2016
2016-01-18	18	0	3	1	18	3	2016
2016-01-19	19	1	3	1	19	3	2016
2016-01-20	20	2	3	1	20	3	2016

Figure 13. Date Features

In Figure 13. January 16, 2016, in the first row data, some components are taken into sections such as *dayofmonth* (day in each month), *dayofweek* (day of the week), *weekofmonth* (week of each month), *month* (month), *dayofyear* (day of the year), *weekofyear* (week of the year), and *year* (year). This date feature can see how the transaction patterns in each time component such as transactions on weekdays and weekends.

2) Lag

Lag Feature is a feature that assigns the value of *t* to the value of *t-1*. The value at time *t* is strongly influenced by the value at time *t-n*. The previous value is known as lag. Thus, *t-1* is lag 1, *t-2* is lag 2.

Date	kilo	lag_1	lag_2	lag_3
2016-01-16	12.85	NaN	NaN	NaN
2016-01-17	10.10	12.85	NaN	NaN
2016-01-18	11.40	10.10	12.85	NaN
2016-01-19	8.50	11.40	10.10	12.85
2016-01-20	6.65	8.50	11.40	10.10

Figure 14. Lag Features

Based on Figure 14., it is known that on January 16, 2016, the needed gourami is 12.85. Lag 1 takes data on the previous day which is 12.85 placed at *t+1*, lag 2 takes data on the previous 2 days is placed at *t+2*.

3) Rolling Window

Rolling Window feature is a feature that is obtained by taking a summary of several previous values with the aggregate. This research uses the *mean* as the aggregate of the rolling windows obtained.

Date	kilo	roll_mean_2	roll_mean_3	roll_mean_4
2016-01-16	12.85	NaN	NaN	NaN
2016-01-17	10.10	11.475	NaN	NaN
2016-01-18	11.40	10.750	11.45	NaN
2016-01-19	8.50	9.950	10.00	10.7125
2016-01-20	6.65	7.575	8.85	9.1625

Figure 15. Rolling Mean Window Feature

Based on Figure 15., the roll_mean_2 column takes the average of the two previous observation data, roll_mean_3 takes the average of the three previous observation data.

4) Mean Encoding

Mean encoding feature is a feature obtained by finding the average of the component time series itself. For example, the average for each month of December, the average for each day of the week, and the average for every 31st.

Date	kilo	weekofyear	weekofyear_enc
2016-01-16	12.85	2	19.202273
2016-01-17	10.10	2	19.202273
2016-01-18	11.40	3	16.161111
2016-01-19	8.50	3	16.161111
2016-01-20	6.65	3	16.161111

Figure 16. Mean Encoding Features

Based on Figure x, the *weekofyear_enc* column takes the average set of values from the week of year time component which has a value of 2.Besides that, this study also uses several time components to be used as a mean encoding feature such as months, days, and days of the year.

5) Combined

Combined features are done by combining pre-made features such as date features, lags, rolling windows and mean encoding.

Model testing is done by searching for the best model of all models with several parameters such as *eta*, *max_depth*, *min_child_weight*, *subsample*, *colsampe_bytree*, and *gamma*. The performance of each model is measured using MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) on the test data. The results of all models found that the best model is with combined features as shown in Table 3. and Figure 17.

Table 3. Result of All Models

Model	MAE	MAPE	Time
Model 1 (Date)	4.03	13,59%	113ms
Model 2 (Lag)	3.39	10,62%	90.6ms
Model 3 (Rol. Window)	2.51	10,34%	231ms
Model 4 (Mean Enc)	3.52	12,56%	139ms
Model 5 (Combined)	**0.63**	**2,46%**	**348ms**

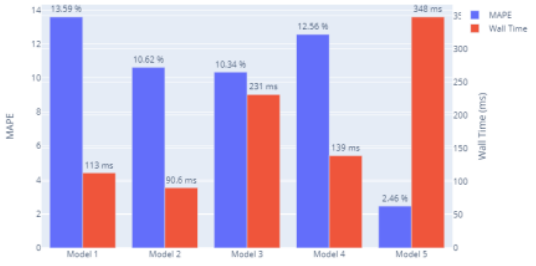

Figure 17. Comparison of the results of all models

4.6 Select Best Model

The best model is chosen which has the smallest error value, namely model 5 which uses combined features. The comparison of 10 actual data with the predicted results of model 5 on the test data can be seen in Figure 18. The accuracy produced by model 5 is 97.54% with 16-time steps from 16 December 2019 to 31 December 2019.

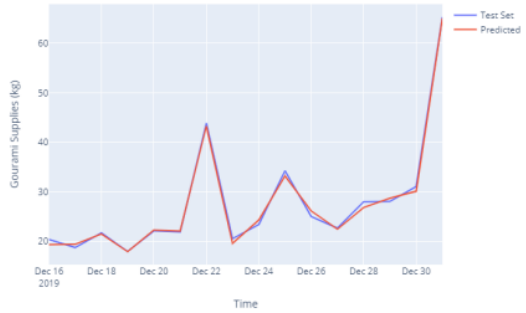

Figure 18. Actual Data vs Predicted

The important features in making model 5 can be seen in Figure 19.

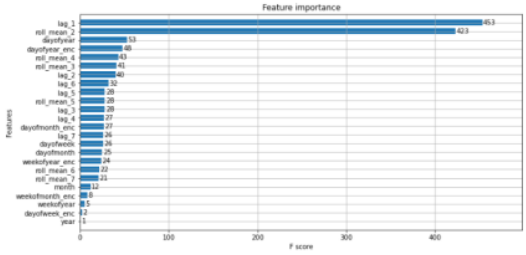

Figure 19. Feature Importance

Based on Figure 19., it is known that two important features in making model 5 are the lag_1 and roll_mean_2 features.

5. Conclusion

Based on the results of testing with 5 XGBoost models, the best XGBoost model is obtained with combined features consisting of components of time, lag, rolling window, and mean encoding. The best parameter is the number of *colsample_bytree* is 1, *eta* 0.175, *min_child_weight* 6, *gamma* 0.6, *max_depth* 11, and *subsample* 0.75 resulting in an accuracy of 97.54% to predict gourami supplies. Suggestions for the further development of this research are to add some external features such as holidays and weather or climate that occurs in the field, besides that, it can also do a stacking model which makes models with various other methods such as linear SVM models, and neural networks.

Estimation Of Gourami Supplies Using Gradient Boosting Decision Tree Method

ORIGINALITY REPORT

5%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.mdpi.com Internet	38 words — 1%
2	worldwidescience.org Internet	27 words — 1%
3	www.analyticsvidhya.com Internet	25 words — 1%
4	upcommons.upc.edu Internet	23 words — 1%
5	www.econstor.eu Internet	12 words — < 1%
6	Vladislav Khramtsov, Alexey Sergeyev, Chiara Spiniello, Crescenzo Tortora et al. "KiDS-SQuaD", Astronomy & Astrophysics, 2019 Crossref	12 words — < 1%
7	www.commerceresources.com Internet	10 words — < 1%
8	www.tandfonline.com Internet	9 words — < 1%
9	www.nature.com Internet	9 words — < 1%
10	docplayer.net Internet	8 words — < 1%

11 Intharathirat, Rotchana, P. Abdul Salam, S. Kumar, and Akarapong Untong. "Forecasting of municipal solid waste quantity in a developing country using multivariate grey models", Waste Management, 2015.
Crossref

6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		