Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Coffee Quality Prediction with Light Gradient Boosting Machine Algorithm Through Data Science Approach

Adya Zizwan Putra[1)], Mawaddah Harahap[2*)], Achmad Nurhadi[3)], Andro Eriel Tambun[4)], Syahmir Defha[5)]
[1,2,3,4,5)]Universitas Prima Indonesia, Indonesia
[1)]adyazizwanputra@unprimdn.ac.id, [2)]mawaddah@unprimdn.ac.id, [3)]achmad.nurhadi@ymail.com,
[4)]androeriel8@gmail.com, [5)]syahmirsyahmir18@gmail.com

Submitted : Jan 27, 2023 | **Accepted** : Feb 2, 2023 | **Published** : Feb 4, 2023

Abstract: In increasing sales by increasing consumer satisfaction with the quality of coffee sold. A way is needed to make it easier to predict the determination of quality coffee so as to increase the efficiency of the coffee sorting process which does not take a long time and can increase the productivity of companies that have competitiveness. Several developments have been made to improve the performance of the algorithm which has the potential to produce good quality predictions. Import Copy Data into a format that can be processed to a later stage or with a Machine Learning algorithm. Copy data that can be processed is then modified in such a way as to ensure that the data is suitable for use in Data Science or Machine Learning processes. By using coffee data specifications from the plantation to the coffee beans produced, it is expected that coffee quality can be predicted quickly without the need for manual calculations or analysis by humans. The working procedures for selecting the quality of coffee beans are coffee import data, coffee data processing, split test-train coffee data, light gradient enhancement machine, yield prediction, and Performance Prediction Evaluation. The amount of data used is 1,339 data. The dependent variable in this data is Coffee Quality while the rest will be cleaned and processed to serve as an independent variable. The accuracy rate of the algorithm in predicting coffee quality is 72%.

Keywords: Coffe quality prediction, light gradient boosting, Data Science Approach

INTRODUCTION

Judging coffee quality was traditionally done by weighing the beans individually into a small cup, grinding the beans, then pouring boiling water over them, and after 5 minutes sipping the coffee to the palate with a large spoon. Although coffee may have a good and general visual appearance, as indicated by its color, bean uniformity, and lack of defective beans, it can have a poor taste due to contamination during processing, storage, and transport from the coffee farm to the roaster's warehouse. This process certainly takes a long time and requires a large amount of money to employ experts. On the other hand, technological assistance can make the coffee quality assessment process more automatic, efficient, and relatively affordable compared to hiring experts.

However, research and system development is needed to achieve technology that can solve these problems. Research related to coffee quality has been carried out by several other studies. These include smart farming and intelligent imaging to predict defects in coffee leaves(Chemura, Mutanga, Sibanda & Chidoko, 2018), genomic classification of Arabica coffee(Sousa, Nascimento, Silva, Nascimento, Cruz, Silva, Almeida, Pestana, Azevedo, Zambolim & Caixeta, 2020) and prediction of coffee quality based on coffee specifications(Christiana & Darmawana, 2020). Coffee specifications are proven to be used to determine the quality of coffee. In previous research, the Neural Network algorithm was successfully designed to assess coffee quality.

Apart from Neural Networks, there are also other similar algorithms, namely Light Gradient Boosting. What distinguishes Light Gradient Boosting from Neural Networks is that Light Gradient Boosting uses a decision tree method that has been developed(Akbilgic, Butler, Karabayir, Chang, Kitzman, Alonso, Chen & Soliman, (2021). Light Gradient Boosting (LGB) extends the Gradient Boosting algorithm by adding a type of automatic object selection, and focuses on boosting examples with large gradients. This can lead to significant learning acceleration and better predictive performance. Thus, the LGB algorithm has become the flagship algorithm for

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

machine learning competitions when working with tabular data for regression and classification predictive modeling problems(Ustuner & Sanli, 2019). Several developments have been made to improve the performance of the algorithm which has the potential to produce good quality predictions.

LITERATURE REVIEW

Some of the obstacles that have occurred in selecting each coffee bean directly are considered less effective. It is difficult to determine the quality of coffee without clear provisions. So far, choosing the quality of coffee based on guesswork. Customers feel disappointed with the quality of the coffee being sold. Everyone has a different opinion on the quality of coffee, so it is difficult to determine the average quality for everyone.

Prediction of specialty coffee flavors based on near-infrared spectra using machine- and deep -learning methods, predicting the taste quality of specialty coffee using the near infrared spectrum, based on 7 categories with 266 samples. Where is coffee grounds as input by training Machine Learning (ML) and Deep Learning (DL) models with an accuracy of 70–73% and 75–77% respectively(Chang, Hsueh, Hung, Lu, Peng & Chen, 2021). Light Gradient Boosting Machine-Based Link Quality Prediction for Wireless Sensor Networks. Effective link quality prediction can select high-quality links for communication and improve the reliability of data transmission. To improve the accuracy of the link quality prediction model and reduce the model complexity, the link quality prediction model is based on the light gradient enhancement engine (LightGBM-LQP). Specifically, agglomerative hierarchical clustering and manual splitting were combined to assess link quality and derive sample labels. Then, the light gradient boosting machine (LightGBM) and Focal Loss classification algorithms are used to estimate the link quality value(Niu, Zhang & Shu, 2022).

Evaluation of Light Gradient Boosted Machine Learning Technique in Large Scale Land Use and Land Cover Classification. Comparison of three machine learning techniques: Random Forest, Support Vector Machines, and Light Gradient Boosted Machine, using a training evaluation model of 70%. And 30% testing. Evaluate the accuracy of the Light Gradient Boosted Machine model against the more classical and reliable Random Forest and Support Vector Machines in terms of classifying land use and land cover over a wide geographic area. It was found that the Light Gradient Booted model was slightly more accurate with an increase in overall accuracy of 0.01 and 0.059 respectively compared to the Support Vector and Random Forests, but also performed about 25% faster on average(McCarty, Woo & Lee, 2020). Forward Chaining Method in an Expert System for Assessment of Web-Based Coffee Bean Quality(Raharjo &Agustini, 2020). The quality of each coffee bean is assessed based on: color, size, impurities and the level of coffee physical defects. Each sorting of coffee beans produces a grade/quality of coffee beans ranging from the best quality (grade 1) to the worst quality (grade 6)(Raharjo & Agustini, 2020).

METHOD

The coffee data used in this study was obtained from the Coffee Quality Institute. This data has also been studied by other researchers to predict coffee quality using a Neural Network algorithm or an Artificial Neural Network(Christiana & Darmawana, (2020). This data has 44 columns or variables with a total of 1339 data. The dependent variables predicted in this study sequentially from best to worst quality include the quality of Specialty Grade, Premium, Exchange, and Below Standard coffee. The expectation in this study is to be successful in using independent variable data to predict the quality of the coffee. In this research Machine Learning will be used to predict the quality of coffee which consists of four outputs, namely, Specialty Grade, Premium Grade, Exchange Grade, and finally Below Standard. It is hoped that the algorithm can be used to facilitate the assessment of coffee, especially in an effort to increase the efficiency of coffee producing companies(Traore, Wilson & Fields, 2018).

Using coffee specification data from the plantation to the coffee beans produced, it is hoped that coffee quality can be predicted quickly without the need for manual calculations or analysis by humans. The work procedure for selecting the quality of coffee beans is shown in Fig. 2 below:

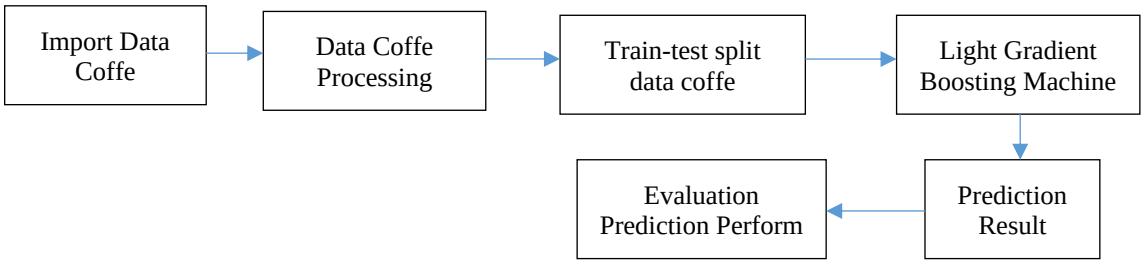

Fig. 1 Work Procedure

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data coffe import into a format that can be processed to the next stage with a Machine Learning algorithm. Imported data has data type .CSV to read each line in the file using commas as delimiters. The results of import data in this study are shown in Fig. 2.

```
data = pd.read_csv("../merged_data_cleaned.csv")
data.head()
```

	Unnamed: 0	Species	Owner	Country.of.Origin	Farm.Name	Lot.Number	Mill	ICO.Number	Company	Altitude	...	Color	Category.Two.Defects	Expiration	Certification.Body	
0	0	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	...	Green	0	April 3rd, 2016	METAD Agricultural Development plc	309fcf77415a
1	1	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	...	Green	1	April 3rd, 2016	METAD Agricultural Development plc	309fcf77415a
2	2	Arabica	grounds for health admin	Guatemala	san marcos barrancas "san cristobal cuch	NaN	NaN	NaN	NaN	1600 - 1800 m	...	NaN	0	May 31st, 2011	Specialty Coffee Association	36d0d00a3724
3	3	Arabica	yidnekachew dabessa	Ethiopia	yidnekachew dabessa coffee plantation	NaN	wolensu	NaN	yidnekachew debessa coffee plantation	1800-2200	...	Green	2	March 25th, 2016	METAD Agricultural Development plc	309fcf77415a
4	4	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	...	Green	2	April 3rd, 2016	METAD Agricultural Development plc	309fcf77415a

5 rows × 44 columns

Fig. 2 Layout import data coffe

Previous studies used pre-processing techniques before processing data with algorithms. In this study, there are several things that will be considered to improve the quality of the Light Gradient Boosting Machine algorithm in this study, including data voids, inappropriate data types, and data imbalances. Data emptiness is a problem that occurs because there is missing or empty data in part to the whole of a data variables(Madley-Dowd, Hughes, Tilling & Heron, 2019). Encoding labels can handle data types that do not match or cannot be processed. The data type before being processed has an order from bad to very good(Yang, Hou, Zhou, Wang & Yan, 2021). In addition, One-hot Encoding is different from Label Encoding, which converts data types that have not been processed into numbers 1 or 0 so that the algorithm can understanded(Li, Si, Xu & Jiang, 2018).

Coffe data that can be processed is then modified in such a way as to ensure that the data is fit for use for Data Science or Machine Learning processes. Observing data gaps is important to prevent confusion in Machine Learning algorithms in training to predict coffee quality. The results of observing empty data are shown in Figure 3.2 where the black color indicates filled data, while the white color indicates empty data. As shown in Figure 3.2, data voids can be found in several independent variables, such as lots. Number, Mill, Company, and so on. Overcome the problem of data emptiness by deleting empty data. Then select most of the data to be studied. The independent variables used are as many as 8 variables, including "Country.of.Origin", "Harvest.Year", "Variety", "Processing.Method", "Category.One.Defects", "Category.Two.Defects" ,"Quakers","altitude_mean_meters" and "Total.Cup.Points". While the dependent variable used is "Cupping.Grade". Next is to observe the data type, this is done by observing the data after the data void problem has been resolved as shown in Fig. 3.

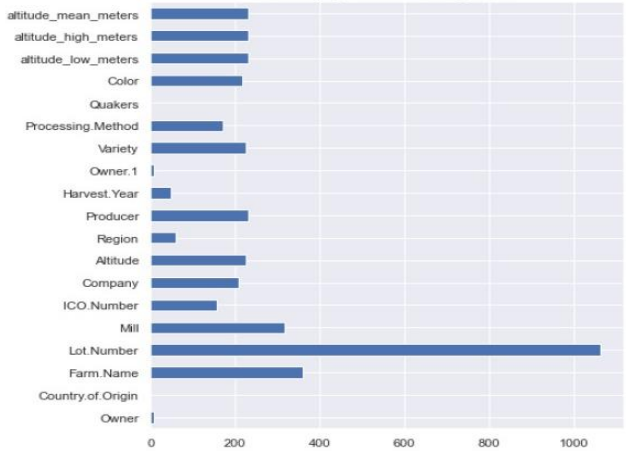

Fig. 3 Observation results of data emptiness

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Such as data blanks on owner, quakers, and owner 1. Data blanks will always occur if the accuracy of the test is not too high.

```
df.head()
```

	Country.of.Origin	Harvest.Year	Variety	Processing.Method	Category.One.Defects	Category.Two.Defects	Quakers	altitude_mean_meters	Total.Cup.Points
0	Ethiopia	2014	Other	Washed / Wet	0	1	0.0	2075.0	89.92
1	Ethiopia	2014	Other	Washed / Wet	0	2	0.0	2075.0	88.83
2	Ethiopia	2014	Other	Natural / Dry	0	4	0.0	1822.5	88.25
3	United States	2014	Other	Washed / Wet	0	0	0.0	1872.0	87.92
4	United States	2014	Other	Washed / Wet	0	0	0.0	1943.0	87.92

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 975 entries, 0 to 974
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country.of.Origin     975 non-null    object
 1   Harvest.Year          975 non-null    object
 2   Variety               975 non-null    object
 3   Processing.Method     975 non-null    object
 4   Category.One.Defects  975 non-null    int64
 5   Category.Two.Defects  975 non-null    int64
 6   Quakers               975 non-null    float64
 7   altitude_mean_meters  975 non-null    float64
 8   Total.Cup.Points      975 non-null    float64
dtypes: float64(3), int64(2), object(4)
memory usage: 68.7+ KB
```

Fig. 4 Observation results of data after cleaning from empty data

The next step is to overcome inappropriate data types as shown in Table 1. To reduce variations in country data that are too diverse, the authors group data from countries of origin into 'Others' for countries that have data below 10. Then, the results of the research are presented in accordance with the proposed problem-solving stages as previously described, which include:

Table 1. Data Processing

Variabel	Jenis Pemrosesan Data
Country of Origin	Mengelompokkan data minoritas menjadi ‘other’
Harvest Year	Label encoding dari tipe data string menjadi integer
Variety	Mengelompokkan data minoritas menjadi ‘other’
Processing	Label encoding dari tipe data string menjadi integer
Category	-
Quakers	-
Altitude	Menghapus data yang terlalu besar atau rendah
Total Cup Points	Transformasi data menjadi tiga kualitas kopi (1=Specialty Quality, 2 = Premium Quality, 3 = Usually Good Quality)

After the data is processed, it will produce something like in Figure 3.5. It can be seen that the data type has become a classification. However, in the figure it can be seen that there is an imbalance in the data. This can be seen from the Premium Quality data which is too much compared to other data. To overcome this kind of data imbalance, Random Over Sampling can be done.

Fig. 5 Observation results of the dependent variable are not balanced

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The next step is to observe again whether the Random Over Sampling works properly. The results of these observations are shown in Fig.6. It can be seen that the distance between Premium coffee data and the others is not too far.

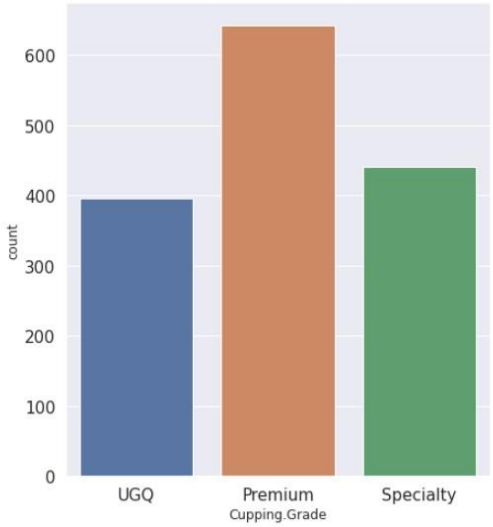

Fig. 6 Results of observations after the data has been balanced

Train-Test Split Coffee Data After Processing: To be able to assess the performance of the algorithm in this study, it is necessary to divide the data into two types, namely Train Data or training data and Test Data or testing data. This process is useful for estimating the performance of Machine Learning algorithms that can be applied to prediction-based algorithms. This method is a quick and easy procedure to perform so that we can compare the results of our own machine learning models with those of the machines. In general, the test data is divided into 30% of the actual data and the training data is divided into 70% of the actual data. In this research, it is necessary to divide the data into data for train and test in order to evaluate how well the machine learning model performs. Train Data is used to be able to train Machine Learning models. The second data is called Test Data, this data is only used for predictions. After performing the Split Data process as shown in Fig 7, the resulting Train Data is 1,035 data and the Test Data is 444 data.

```
# Proses Data Split untuk membagi data menjadi dua macam data, Test Data dan Train Data.
from sklearn.model_selection import train_test_split
X = encode_df.drop("Cupping.Grade", axis = 1)
Y = encode_df["Cupping.Grade"]

x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state = 1)

print('Training Features Shape:', x_train.shape)
print('Training Labels Shape:', y_train.shape)
print('Testing Features Shape:', x_test.shape)
print('Testing Labels Shape:', y_test.shape)
```

```
Training Features Shape: (1035, 8)
Training Labels Shape: (1035,)
Testing Features Shape: (444, 8)
Testing Labels Shape: (444,)
```

Fig. 7. Split data result

Light Gradient Boosting algorithm training: Train data or data prepared for the training process will then be used to train the Light Gradient Boosting algorithm so that the algorithm can practice predicting coffee quality until it achieves the expected performance. Prediction of coffee quality with the Light Gradient Boosting algorithm: After the training process, the trained Light Gradient Boosting algorithm will then be used to predict data that has never been studied before. This is important because if the algorithm predicts previously recognized data, it is not certain that the algorithm can predict new data. Based on Fig. 8, using the Jupyter Notebook application is processing the prediction of the quality of a coffee.

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

```
# Pelatihan menggunakan Light Gradient Boosting
import lightgbm as lgb
lgbm = lgb.LGBMClassifier()
lgbm.fit(x_train, y_train)

# Prediksi menggunakan Light Gradient Boosting
y_pred=lgbm.predict(x_test)

print("\nHasil prediksi:")
display(y_pred)
print("\nKeterangan: # 0 = Specialty Quality, 1 = Premium Quality, 2 = Usually Good Quality")
```

```
Hasil prediksi:
array([1, 0, 0, 1, 0, 0, 2, 0, 1, 1, 0, 2, 1, 1, 2, 0, 0, 0, 0, 2, 0,
       0, 2, 0, 1, 0, 2, 2, 1, 0, 0, 0, 0, 0, 0, 2, 2, 1, 1, 1, 1, 1,
       1, 2, 1, 2, 0, 2, 1, 0, 0, 0, 0, 2, 0, 0, 1, 2, 2, 1, 2, 2, 1, 0,
       0, 0, 0, 2, 2, 1, 2, 1, 0, 0, 2, 2, 2, 1, 0, 1, 0, 1, 2, 0, 1, 0,
       1, 1, 0, 1, 0, 2, 2, 1, 2, 2, 2, 1, 2, 1, 0, 2, 2, 0, 2, 0, 1, 0,
       1, 2, 0, 1, 2, 0, 1, 0, 0, 1, 1, 2, 2, 0, 0, 1, 0, 2, 2, 1, 0, 1,
       0, 2, 2, 1, 0, 2, 2, 1, 2, 0, 0, 0, 2, 0, 2, 2, 1, 0, 2, 0, 1, 1,
       0, 1, 2, 0, 2, 1, 0, 2, 2, 1, 1, 0, 1, 2, 1, 0, 1, 1, 1, 0, 2, 1,
       0, 0, 1, 1, 2, 1, 0, 0, 0, 0, 1, 2, 1, 2, 1, 2, 0, 2, 1, 0, 2, 1,
       0, 2, 1, 0, 1, 0, 1, 0, 0, 2, 1, 2, 0, 1, 1, 2, 1, 2, 2, 2, 0, 1, 1,
       0, 1, 1, 0, 1, 1, 2, 1, 0, 0, 2, 2, 1, 1, 1, 2, 0, 1, 0, 0, 1,
       1, 1, 1, 0, 2, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 2, 0, 0, 2, 1, 1, 1,
       2, 0, 0, 2, 1, 1, 1, 1, 0, 1, 1, 0, 0, 2, 0, 0, 0, 2, 0, 1, 0, 0,
       0, 1, 1, 2, 1, 0, 1, 0, 2, 1, 2, 0, 2, 2, 1, 0, 2, 1, 2, 2, 0, 2,
       0, 1, 1, 1, 0, 2, 0, 0, 1, 2, 2, 0, 1, 1, 2, 1, 2, 0, 0, 0, 1, 0,
       0, 2, 0, 2, 0, 0, 0, 1, 1, 2, 2, 0, 2, 0, 2, 0, 1, 2, 1, 0, 2, 2,
       0, 2, 2, 2, 0, 2, 1, 2, 2, 0, 2, 1, 0, 0, 1, 0, 1, 0, 0, 2, 2, 1,
       1, 2, 0, 1, 0, 1, 2, 0, 2, 1, 0, 1, 1, 2, 0, 1, 2, 1, 0, 0, 2, 0,
       2, 0, 2, 1, 2, 2, 1, 1, 2, 0, 2, 2, 0, 2, 0, 2, 0, 2, 0, 0, 0, 1,
       0, 2, 1, 2, 1, 1, 2, 1, 2, 1, 0, 1, 1, 0, 2, 2, 0, 1, 0, 2, 0, 0,
       1, 0, 1, 2])

Keterangan: # 0 = Specialty Quality, 1 = Premium Quality, 2 = Usually Good Quality
```

Fig. 8 Coffee quality prediction process

According to the information, number 0 is Specialty Quality, number 1 is Premium Quality and number 2 is Good Quality. Once predicted, it will proceed to the stage of comparing the predicted and actual data results according to Fig. 8.

```
# Menggabungkan data hasil prediksi
hasil_prediksi = pd.DataFrame(y_pred, columns = ['Hasil Prediksi'])
hasil_prediksi = hasil_prediksi.join(y_test.reset_index().drop('index', axis=1))
hasil_prediksi.rename(columns = {'Cupping.Grade':'Hasil Aktual'}, inplace = True)

display(hasil_prediksi)
```

	Hasil Prediksi	Hasil Aktual
0	1	1
1	0	0
2	0	0
3	1	2
4	0	1
...	...	...
439	0	0
440	1	1
441	0	0
442	1	1
443	2	1

444 rows × 2 columns

Fig. 9 Comparison of predicted and actual data results

It can be seen that there are 2 columns containing predicted data and actual data and numbers are quality data. It can be seen after predicting that the results are different. Like experiment 3, the predicted result is number 1 which means premium quality but if the actual data results are number 2 which means good quality. This means that sometimes the predicted results may differ from the actual results. Fig. 9 shows a comparison of the predicted and actual data generated by combining the predicted data from the Light Gradient Boosting algorithm with actual coffee quality data. There are 444 data as a whole generated in Test Data or test data. Furthermore, the result of a correct prediction is the result of the prediction which agrees with the actual result. For example, if the actual results are 1, 2, and 1 respectively, then the correct prediction results are 1, 2, and 1.

Evaluation of the ability of the Light Gradient Boosting algorithm: The last step is to evaluate the performance of the algorithm in predicting the test data. Good ability is characterized by high accuracy. The more correct predictions, the higher the resulting accuracy value. Thus the correct prediction results in this study are shown in Figure 3.10. The circled areas show the actual results which match the predicted results. This means that the algorithm successfully predicts coffee quality according to the actual coffee quality. The more accurate prediction results, as shown in Fig 10, the better the accuracy of the algorithm in predicting coffee quality. However, other predictions may not go well and there are several predictions that differ from the actual results.

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

```
# Menggabungkan data hasil prediksi
hasil_prediksi = pd.DataFrame(y_pred, columns = ['Hasil Prediksi'])
hasil_prediksi = hasil_prediksi.join(y_test.reset_index().drop('index', axis=1))
hasil_prediksi.rename(columns = {'Cupping.Grade':'Hasil Aktual'}, inplace = True)

display(hasil_prediksi)
```

	Hasil Prediksi	Hasil Aktual
0	1	1
1	0	0
2	0	0
3	1	2
4	0	1
...	...	...
439	0	0
440	1	1
441	0	0
442	1	1
443	2	1

444 rows × 2 columns

Fig. 10 Correct prediction

RESULT

The amount of data used is 1,339 data. The dependent variable in this data is Coffee Quality while the rest will be cleaned and processed to be used as an independent variable. The data was obtained from the Coffee Quality Institute which has also been studied by previous researchers using Deep Learning techniques (Yang, Hou, Zhou, Wang, & Yan, 2021).

In addition, the data type in the harvest year variable also needs to be converted to an integer. An example is the previous text or string value "2017/2018" was changed to 2018 (integer). This needs to be done because Machine Learning algorithms do not understand text data types. So that integer type data is prioritized in order to facilitate assessment. Changes to Onehot encoding form (for example 2018 has a value of 1, and 2017 has a value of 0) are not made because the annual data has a hierarchy between bigger and higher.

Furthermore, the Country of Origin data or country of origin of coffee production also needs to be simplified. This is done by changing the data which amounts to only 1 data so that it is changed to 'others'. This needs to be done because the amount of information about the data is too small so that it cannot provide significant learning to the Machine Learning algorithm. However, if it is categorized as 'other', there will be a lot of data that has similar values. For example, there are 10 countries that only have 1 copy of data, if. the data is changed to 'other' then there will be 10 data that have similar values. Then delete the altitude data or the altitude of the plantation area where coffee is planted that is too high or too low, namely above 2,000 meters and below 200 meters. Make improvements to the data type of variables in the dataset.

Then the last is to make the dependent variable the grading of coffee beans based on data on defects, quakers, and Total Cupping Point. Defects are coffee defects, and quakers are coffees that do not abnormally brown when roasted. Both are used to judge low quality coffee. Meanwhile, a high Total Cupping Point is used to indicate high quality coffee. Cupping Grade is the dependent variable used in this study with four levels sequentially from highest to lowest, including Specialty, Premium, Exchange, and Below Standard.

This technique can be effective for Machine Learning algorithms that are affected by data imbalances and where multiple duplicate instances for a particular class can affect predictive performance. This can include algorithms that learn coefficients iteratively, such as artificial neural networks that use stochastic gradient descent. This can also affect models that seek good data separation, such as Supervised Vector Machines and Decision Trees. The process includes training and prediction of coffee quality with the Light Gradient Boosting algorithm which will produce predictive results that will be evaluated in this study. shows the resulting prediction results. The resulting data type is an array with a collection of numbers. This figure is the dependent variable predicted in this study. In accordance with the previous design, the numbers 0, 1, and 2 respectively symbolize Specialty, Premium, and Usually Good Quality.

Things to consider when conducting training and predicting coffee quality are the amount of data used and the quality of the data that has been processed. The amount of data used is important because it will determine the performance of the prediction results. Little data, for example 50-100 data can produce less than optimal performance while sufficient data, for example 500-1000 data can produce optimal performance. Then the quality of the data also needs to be considered so that the data can be processed and understood by the Machine Learning algorithm.

By showing a comparison of predicted and actual results. This is done by combining the resulting predicted data with actual data that has previously been known. Of course the algorithm doesn't know the actual data so the results can be different. The more precise the prediction results are in guessing the actual results, the higher the accuracy value of the algorithm will be.

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The final step is to evaluate the Light Gradient Boosting algorithm. This stage is carried out with the Classification Report function which belongs to the SKLearn library and is shown in Fig. 12. This is useful for automating the evaluation of the Light Gradient Boosting algorithm in predicting coffee quality in this study. So that each algorithm has finished predicting the data prepared, then a report on the performance of the algorithm will be displayed.

First of all, the classification_report function needs to be imported into Jupyter Notebook so that it can be used. Then the coffee quality prediction results in the form of the DataFrame data type are entered into the y_pred variable and the actual y_test coffee quality results are entered into the function along with the three predicted coffee quality levels, including Premium, Specialty, and UGQ. The three coffee quality levels are entered into variable a which is an array data type with the three coffee quality levels with string data type.

The accuracy value in Fig. 12 is a benchmark for the performance of the Light Gradient Boosting algorithm which will be discussed further in the next chapter. This value is the accuracy performance generated by predicting Test Data or test data. The higher the accuracy value, the better the algorithm under study is in predicting coffee quality. The lowest value is 0 and the highest value is 1 (100%).

	precision	recall	f1-score	support
0	0.74	0.62	0.68	192
1	0.69	0.82	0.75	125
2	0.72	0.76	0.74	127
accuracy			0.72	444
macro avg	0.72	0.73	0.72	444
weighted avg	0.72	0.72	0.71	444

Fig. 12 Evaluation result performance Light Gradient Boosting Machine

In addition, the Confusion Matrix technique from the SKLearn library is also used to produce a more detailed explanation of the predicted results between actual coffee quality data and predicted coffee quality data. By using the Confusion Matrix technique which initially includes data on the three levels of coffee and the prediction results as well as the results of the actual coffee quality data, Table 1 can be produced as a truth table for coffee quality predictions. The higher the same value between the actual data and the predicted coffee quality data, the higher the accuracy value will be. Conversely, if the actual and predicted coffee data values are different, the accuracy of the Light Gradient Boosting algorithm in predicting coffee quality will be lower.

Table 2. Evaluation of the coffee quality prediction truth table

Actual Data Coffe Quality	Premium	Specialty	UGQ
Premium	120	41	31
Specialty	17	102	6
UGQ	26	5	96

Then the next is to observe feature importance. It is necessary to. The results of the feature importance observations are shown in Fig.13. It can be seen that the variables or features that most influence the prediction of coffee quality are 'Category.Two.Defects', 'Country.of.Origin', 'Variety', and 'Harvest.Year'. The four data sequentially represent coffee bean defects, country of origin, type of variety, and year of harvest.

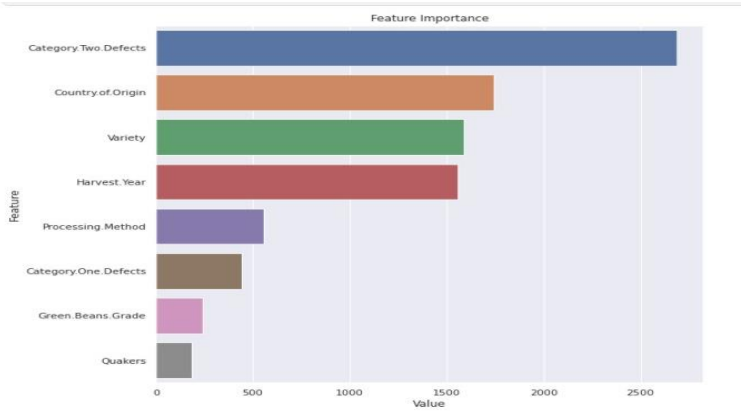

Fig. 13 Feature Importance

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

Because there are three possible outputs on the dependent variable, namely the three types of coffee quality to be predicted, there will be 9 possible prediction results. The prediction results in this study are divided into two types, true and false. The correct result is the predicted value that corresponds to the actual value. An example of a correct prediction is that the actual value of the quality of a coffee is Premium, and the Light Gradient Boosting algorithm successfully predicts correctly that the quality of the coffee is Premium, then. So as Likewise for other coli quality values, including Specialty and UGQ.

Table. 2 Correct prediction of coffee quality

Actual Data Coffe Quality	Premium	Specialty	UGQ
Premium	120	-	-
Specialty	-	102	-
UGQ	-	-	96

Table 2 displays the results of the number of correct prediction results, including: 120 actual results of Premium predicted as Premium, 102 Specialty actual results predicted as Specialty, 96 actual results of UGQ predicted as UGQ. On the other hand, the incorrect prediction value is indicated by the difference between the actual coffee quality value and that predicted by the Light Gradient Boosting algorithm. For example, if the actual quality level of a coffee is Specialty, but the algorithm predicts the quality of the coffee as Premium. Of course this can cause losses to the company because the predicted quality of the coffee is different from the actual one. Therefore it is necessary to improve the algorithm to achieve the highest possible accuracy value so as to reduce the error rate of the algorithm.

Table 3. Incorrect coffee quality prediction results

Actual Data Coffe Quality	Premium	Specialty	UGQ
Premium	-	41	31
Specialty	17	-	6
UGQ	26	5	-

Table 3. shows the number of incorrect prediction results, including: 41 actual results of Premium predicted as Specialty, 31 actual results of Premium predicted as UGQ, 17 Specialties actual results predicted as Premium, 6 Specialty actual results predicted as UGQ, 25 UGQ actual results predicted as Premium, 5 actual results of UGQ predicted as a Specialty

By using these results, the accuracy value is obtained by 72%, it can be said that the Light Gradient Boosting algorithm in this study succeeded in predicting the quality of the coffee, but there is a 28% possibility that the predicted results do not match the actual results. Further research is needed to improve the accuracy of Light Gradient Boosting in predicting coffee quality even better. An accuracy rate close to 100% would be the expected achievement for the algorithm to be reliable in predicting coffee quality. Collecting data from the perspective of coffee producing companies can also increase the accuracy of predictions. An example is by adding data on other coffee specifications, for example the brightness level of the coffee color, the shape and size of the coffee, and so on. From the researcher's point of view, the Light Gradient Boosting parameter can also be adjusted in order to achieve a higher level of accuracy.

After the research results are completed, it can be concluded that the things that contribute to the Light Gradient Boosting model which successfully predicts coffee quality in this study include: Selection of independent variables that ensure no empty data, Coffee data processing which makes the data usable by the Light Gradient Boosting algorithm to predict coffee quality, A sufficient amount of data to ensure the algorithm can practice until it reaches the expected performance.

While the low value of accuracy in this study can be caused by: Lack of independent variables used. This study only uses 8 of the 43 variables available. So there are still variables that have the potential to increase accuracy; Not setting the available Light Gradient Boosting parameters to improve accuracy.

*mawaddah@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12169
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Variety	Processing Method	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points	Moisture	Category.One.Defects	Quakers	Color	Category.Two.Defects
	Washed / Wet	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	90.58	0.12	0	0	Green	0
Other	Washed / Wet	8.75	8.67	8.5	8.58	8.4	8.42	10	10	10	8.58	89.92	0.12	0	0	Green	1
Bourbon		8.42	8.5	8.42	8.42	8.3	8.42	10	10	10	9.25	89.75	0	0	0		0
	Natural / Dry	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	89	0.11	0	0	Green	2
Other	Washed / Wet	8.25	8.5	8.25	8.5	8.4	8.33	10	10	10	8.58	88.83	0.12	0	0	Green	2
	Natural / Dry	8.58	8.42	8.42	8.5	8.3	8.33	10	10	10	8.33	88.83	0.11	0	0	Bluish-Green	1
Other	Washed / Wet	8.42	8.5	8.33	8.5	8.3	8.25	10	10	10	8.5	88.75	0.11	0	0	Bluish-Green	0
		8.25	8.33	8.5	8.42	8.3	8.5	10	10	9.33	9	88.67	0.03	0	0		0
		8.67	8.67	8.58	8.42	8.3	8.42	9.33	10	9.33	8.67	88.42	0.03	0	0		0
Other	Natural / Dry	8.08	8.58	8.5	8.5	7.7	8.42	10	10	10	8.5	88.25	0.1	0	0	Green	4
	Natural / Dry	8.17	8.67	8.25	8.5	7.8	8.17	10	10	10	8.58	88.08	0.1	0	0		1
Other	Washed / Wet	8.25	8.42	8.17	8.33	8.1	8.17	10	10	10	8.5	87.92	0	0	0		0
Other	Washed / Wet	8.08	8.67	8.33	8.42	8	8.08	10	10	10	8.33	87.92	0	0	0		0
		8.33	8.42	8.08	8.25	8.3	8	10	10	10	8.58	87.92	0	0	0		2
		8.25	8.33	8.5	8.25	8.6	8.75	9.33	10	9.33	8.5	87.83	0.05	0	0		2
Other	Washed / Wet	8	8.5	8.58	8.17	8.2	8	10	10	10	8.17	87.58	0	0	0		0
		8.33	8.25	7.83	7.75	8.5	8.42	10	10	10	8.33	87.42	0.03	0	0		0
		8.17	8.33	8.25	8.33	8.4	8.33	9.33	10	9.33	8.83	87.33	0.05	0	0		2
Catimor	Washed / Wet	8.42	8.25	8.08	8.17	7.9	8	10	10	10	8.42	87.25	0.1	0	0	Green	0
Ethiopian Yirgacheffe	Natural / Dry	8.17	8.17	8	8.17	8.1	8.33	10	10	10	8.33	87.25	0	0	0		8
Other	Washed / Wet	8	8.25	8.08	8.5	8.3	8	10	10	10	8.17	87.25	0	0	0	None	0
Caturra	Washed / Wet	8.08	8.25	8	8.17	8	8.33	10	10	10	8.33	87.17	0.11	0	0	Green	2
Other	Washed / Wet	8.17	8.25	8.17	8	7.8	8.17	10	10	10	8.58	87.17	0.13	0	0	Green	0
Other	Washed / Wet	8.25	8.33	8.17	8.17	7.8	8.17	10	10	10	8.17	87.08	0	0	0		0
	Natural / Dry	8.42	8.17	7.92	8.17	8.3	8	10	10	10	8.08	87.08	0.11	0	0	Bluish-Green	1
Bourbon	Natural / Dry	8.5	8.5	8	8	8	8	10	10	10	7.92	86.92	0.12	0	0	Green	2
		7.83	8.25	8.08	8.17	8.2	8.17	10	10	10	8.25	86.92	0.05	0	0		2
SL14	Washed / Wet	8.42	8.17	8.17	8.17	7.8	7.92	10	10	10	8.17	86.83	0.12	0	0	Green	1
Caturra		8.17	8.08	8.08	8	8.1	8	10	10	10	8.25	86.67	0.1	0	0	Green	3
Sumatra	Pulped natural / honey	8	8	8	8.25	8	8.17	10	10	10	8.17	86.58	0	0	0	Green	0
		7.92	8.25	8	8.33	8	8.08	10	10	10	8	86.58	0.08	0	0		2
	Natural / Dry	8.42	8.17	8.17	8	7.6	8	10	10	10	8.17	86.5	0.11	0	0	Bluish-Green	1
Bourbon	Natural / Dry	8.5	8.17	8	7.75	8	8	10	10	10	8	86.42	0.12	0	0	Green	2
Caturra	Washed / Wet	8.17	7.83	8	8.08	7.8	8	10	10	10	8.42	86.33	0	0	0	Blue-Green	0
	Natural / Dry	8	8.08	7.92	8	8.1	8.08	10	10	10	8.08	86.25	0.1	0	0	Green	3
SL34	Washed / Wet	8.08	8	8	8.25	7.9	7.92	10	10	10	8.08	86.25	0.12	0	0	Blue-Green	1

Fig. 14 Overview of coffee yield data that has been researched

In accordance with Fig. 14, is an image of the results of the data that has been studied. Those studied were varieties, research processing methods, aromas, tastes, acidity levels, defects, and many others. Thus the research results, the more data, the stronger the evidence of the research results. As well as accuracy also helps to get a higher percentage of truth and avoid mistakes.

Species	Owner	Country.of.Origin	Farm.Name
Arabica	metad plc	Ethiopia	metad plc
Arabica	metad plc	Ethiopia	metad plc
Arabica	grounds for health admin	Guatemala	san marcos barrancas "san cristobal cuch
Arabica	yidnekachew dabessa	Ethiopia	yidnekachew dabessa coffee plantation
Arabica	metad plc	Ethiopia	metad plc
Arabica	ji-ae ahn	Brazil	
Arabica	hugo valdivia	Peru	
Arabica	ethiopia commodity exchange	Ethiopia	aolme
Arabica	ethiopia commodity exchange	Ethiopia	aolme
Arabica	diamond enterprise plc	Ethiopia	tulla coffee farm
Arabica	mohammed lalo	Ethiopia	fahem coffee plantation
Arabica	cqi q coffee sample representative	United States	el filo
Arabica	cqi q coffee sample representative	United States	los cedros
Arabica	grounds for health admin	United States (Hawaii)	arianna farms
Arabica	ethiopia commodity exchange	Ethiopia	aolme
Arabica	cqi q coffee sample representative	United States	el Ã¡guila
Arabica	grounds for health admin	Indonesia	toarco jaya
Arabica	ethiopia commodity exchange	Ethiopia	
Arabica	yunnan coffee exchange	China	echo coffee
Arabica	essencecoffee	Ethiopia	drima zede
Arabica	cqi q coffee sample representative	United States	el rodeo
Arabica	the coffee source inc.	Costa Rica	several
Arabica	roberto licona franco	Mexico	la herradura

Fig. 15 At a Glance Data on coffee owner, country of origin, and farm name

In accordance with Fig. 15, is the data owner, country of origin, and farm name. Where everything is recorded in detail so that the research process can take place precisely and in more detail because there is the owner's name, country of origin and the name of the coffee plantation. And don't forget the species of coffee which is very important in coffee research. It can be seen from Table 4, there is an owner of a coffee with the owner and the origin of the plant.

Table 4. Brief Data Coffe

Owner	Species	Country of Origin
Metad plc	Arabica	Ethiopia
Ground for helath admin	Arabica	Guatemala
Mohammed Lalo	Arabica	Arabica

CONCLUSION

After the research results are completed, it can be concluded that the things that contribute to the Light Gradient Boosting model which successfully predicts coffee quality in this study include: selection of independent variables that ensure no empty data; Coffee data processing which makes the data usable by the Light Gradient Boosting algorithm to predict coffee quality; A sufficient amount of data to ensure the algorithm can practice until it reaches the expected performance. Based on the results of the research that has been done, several conclusions can be drawn: Predict coffee quality with 1,339 coffee specification data using the Light

*mawaddah@unprimdn.ac.id

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Gradient Boosting Machine algorithm; The level of accuracy of the algorithm in predicting coffee quality is 72%.

REFERENCES

Akbilgic, O., Butler, L., Karabayir, I., Chang, P., Kitzman, D., Alonso, A., Chen, L., and Soliman, E. (2021). Artificial intelligence applied to ECG improves heart failure prediction accuracy. J Am Coll Cardiol, vol. 77(18).

Chang, Y., Hsueh, M., Hung, S., Lu, J., Peng, J., and Chen, S. (2021). Prediction of specialty coffee flavors based on near-infrared spectra using machine-and deeplearning methods," Journal of the Science of Food and Agriculture. 101(11). 4705–4714

Chemura, A., Mutanga, O., Sibanda, M., & Chidoko, P. (2018). Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. Tropical Plant Pathology, 43(2). 117–127.

Christiana, Y., and Darmawana, I. D. M. B. A. (2020). Specialty Coffee Cupping Score Prediction with General Regression Neural Network (GRNN)," Jurnal Elektronik Ilmu Komputer Udayana p-ISSN. 9(2). 185-190.

Li, J., Si, Y., Xu, T., and Jiang, S., (2018). Deep convolutional neural network based ECG classification system using information fusion and one-hot encoding techniques. Math Probl Eng.

Liu, L., Niu, M., Zhang, C., and Shu, J. (2022). Light Gradient Boosting Machine-Based Link Quality Prediction for Wireless Sensor Networks. Wireless Communications and Mobile Computing. Article ID 8278087. 13 pages. https://doi.org/10.1155/2022/8278087 2

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J., (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 110. 63–73.

McCarty, D. A., Kim, H. W., and Lee, H, K. (2020). Evaluation of Light Gradient Boosted Machine Learning Technique in Large Scale Land Use and Land Cover Classification, Environments. 7(10). doi:10.3390/environments7100084.

Raharjo, B., and Agustini, F. (2020). Metode Forward Chaining pada Sistem Pakar Penilaian Kualitas Biji Kopi Berbasis Web. International Journal of Natural Sciences and Engineering. 4(2). 73-82.

Sousa, I. C., Nascimento, M., Silva, G. N., Nascimento, A. C. C., Cruz, C. D., Silva, F. F., Almeida, D. P., Pestana, K. N., Azevedo, C. F., Zambolim, L., & Caixeta, E. T. (2020). Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola. 78.

Traore, T. M., Wilson, N. L. W., and Fields, D. (2018). What explains specialty coffee quality scores and prices: A case study from the cup of excellence program. Journal of Agricultural and Applied Economics. 50(3). 349–368

Ustuner, M., and Sanli, F. B. (2019). Polarimetric target decompositions and light gradient boosting machine for crop classification: A comparative evaluation. ISPRS International Journal of Geo-Information. 8(2).

Yang, X., Hou, L., Zhou, Y., Wang, W., and Yan, J., (2021). Dense label encoding for boundary discontinuity free rotation detection. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.15819–15829.

*mawaddah@unprimdn.ac.id