Comparative Performance Evaluation of Gradient Boosting Regressor Against Multiple Linear Regression and Polynomial Regression in Forecasting Avocado Sales Volume in the USA

Evaluasi Kinerja Komparatif Gradient Boosting Regressor Terhadap Regresi Multi Linier dan Regresi Polinomial dalam Peramalan Volume Penjualan Alpukat di Amerika Serikat

Nathan Vilbert Kosasih
Department of Information System, Multimedia Nusantara University, Indonesia
nathan.vilbert@student.umn.ac.id

Accepted on December 21, 2023
Approved on December 21, 2023

***Abstract* – This study employs regression analysis, including multi-linear regression, polynomial regression, and gradient boosting regressor, to predict avocado sales volume in the USA. Investigating the influence of factors such as Average Price, PLU Codes (4046, 4225, 4770), Type (conventional or organic), and Year on Total Volume, the research uncovers significant correlations within the dataset, encompassing date, region, and bags distribution. These findings provide valuable insights for stakeholders in the avocado industry, emphasizing the versatility of regression methods in capturing nuanced patterns within complex datasets. The results contribute to a comprehensive understanding of the diverse factors shaping avocado sales in the USA, concluding that Average Price, PLU Codes, Type, and Year are pivotal in determining sales volume. Further research is recommended to deepen insights into these factors and explore their intricate relationships.**

Keywords - Avocado; Regression; Total Sales

I. INTRODUCTION

The avocado market in the United States has not only witnessed substantial growth but has also become increasingly dynamic, necessitating advanced forecasting methods to comprehend and predict sales volumes accurately. With an annual crop volume exceeding 200,000 metric tons, the U.S. now stands as the world's second-largest producer of avocados. A significant portion of this production, nearly 90%, hails from California, with the 'Hass' avocado variety dominating the landscape, constituting 82% of the state's annual crop [1].

The allure of 'Hass' avocados lies in their widespread acceptance and consistent availability throughout the year, contributing to market dominance. The industry's success is underscored by the remarkable surge in avocado consumption, surpassing the 900,000-tonne mark in 2015 and more than doubling since 2008 [2]. This growth has prompted strategic shifts in marketing approaches, including the development of year-round programs to sustain a continuous presence in key regional and national markets.

However, this success story is not devoid of challenges. In the early 20th century, the market grappled with fragmentation and a myriad of seasonal avocado varieties, each presenting unique obstacles such as alternate bearing patterns [12]. A pivotal moment in addressing these challenges came in the 1960s with the establishment of the California Avocado Advisory Board, which later evolved into the California Avocado Commission [2]. These entities played a crucial role in fostering a unified approach to marketing, steering the industry toward maturity.

The prominence of 'Fuerte' and, notably, 'Hass' avocados in the 1970s and 80s laid the foundation for a more organized U.S. market [1]. Despite persistent challenges, such as alternate bearing, the industry's resilience became evident in the 1980s with concerted efforts to stabilize supply. Today, the avocado market reflects a nationwide presence, a testament to the industry's adaptability and ability to overcome hurdles.

This paper aims to delve into the methodology behind each regression technique, emphasizing their unique characteristics and applications. The ensuing sections will present a detailed analysis of the models' predictive accuracy, emphasizing metrics such as mean squared error, R-squared, and others. The ultimate goal is to discern which regression method, among the three considered, demonstrates superior performance in forecasting avocado sales volume, thereby providing a practical guide for industry professionals and researchers. This comparative study contributes to the ongoing discourse on the application of advanced regression techniques in agriculture and market forecasting, with specific relevance to the burgeoning avocado industry in the United States.

II. Literature

A. *Data Statistics*

Statistics, with its long history in human civilization, has evolved from simple data collection to becoming a fundamental pillar in decision-making and forecasting. Statistics involves the systematic collection, analysis, and interpretation of data to uncover patterns and insights. It encompasses both descriptive and inferential techniques, facilitating informed decision-making across various fields [3].

With the discovery of probability theory and decision-making principles, statistics became instrumental in achieving efficiency across various fields. In the 1950s, it entered the realm of decision-making through generalization and forecasting, accounting for risk and uncertainty factors [3]. The understanding of statistics has progressed from being a tool for data presentation to serving as the foundation for intelligent decision-making, illustrating its essential role in the evolution of modern society.

B. *Variables Concept*

The concept of the relationship between independent and dependent variables is a fundamental aspect of statistical analysis. An independent variable is one that explains or influences another variable, often considered as a presumed cause or antecedent [4]. Conversely, a dependent variable is the one that is described or affected by the independent variable, typically seen as a presumed effect or consequent [4]. This relationship allows for the identification and measurement of the impact of one variable on another, providing a basis for causal analysis in research.

C. *Data Modelling*

Data modeling is the systematic process of structuring and organizing data to represent relationships within a system [5]. It involves creating an abstract representation of real-world entities and their interactions to understand, interpret, and visualize information.

The two main types are conceptual, focusing on high-level entities and relationships, and physical, translating the conceptual model into a specific database system [4]. This process includes identifying entities, defining attributes, and establishing relationships, providing a blueprint for efficient database systems in various applications such as database design, software development, and business analysis.

D. *Machine Learning*

Machine learning, a subset of artificial intelligence, involves creating algorithms that enable computers to learn from data and make predictions or decisions without explicit programming. There are three main types: supervised learning (using labeled data for predictions), unsupervised learning (finding patterns in unlabeled data), and reinforcement learning (learning through interactions to maximize rewards) [5].

Applications include image and speech recognition, natural language processing, and recommendation systems. Success depends on quality data and algorithm selection. Machine learning continues to drive advancements across various fields, shaping the future of technology and problem-solving.

E. *Regression*

Regression is a statistical technique employed for two primary purposes. Firstly, it is commonly utilized for forecasting and prediction, with significant overlaps in its application with machine learning. Secondly, regression analysis can be employed in certain cases to establish causal relations between independent and dependent variables [4]. It's crucial to note that regressions, on their own, reveal relationships between a dependent variable and a fixed dataset collection of various variables [6].

F. *Multiple Linear Regression*

Multiple Linear Regression (MLR) is a statistical technique used to predict the outcome of a response variable by considering multiple explanatory variables [6]. The objective of MLR is to model the linear relationship between the independent variables (represented by 'x') and the dependent variable (represented by 'y'), which is the variable under analysis [5].

This method extends the basic concept of linear regression to accommodate situations where more than one independent variable influences the dependent variable, allowing for a more comprehensive understanding of the relationships in the data. The fundamental model for Multiple Linear Regression (MLR) is represented as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

The formula to compute the parameter vector (β) in Multiple Linear Regression is expressed as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

G. Polynomial Regression

Polynomial regression is a regression method that models the relationship between an independent variable (x) and a dependent variable (y) using a polynomial function. In polynomial regression, the degree of the polynomial (the highest power of the independent variable) can be adjusted to better fit the data than simple linear regression [6]. The general equation for polynomial regression is expressed as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_h x^h + \varepsilon$$

This equation captures the relationship between the variables and allows for a more flexible modeling of non-linear patterns in the data

H. Gradient Boosting Regression

Gradient Boosting Regressor is a machine learning algorithm for regression tasks within ensemble learning. It builds a predictive model using a series of weak learners, typically decision trees, to correct errors iteratively. The algorithm minimizes residuals from each step by training new weak learners on the errors of the current model [7].

This sequential process refines predictions, and the term "gradient" refers to using gradient descent to optimize the loss function. Gradient Boosting Regressor is effective for modeling complex relationships, resistant to overfitting, and finds applications in predicting numerical values like house or stock prices.

III. Methodology

A. Object of Research

This research aims to explore and compare the factors influencing avocado sales volume in the USA using regression methods, including Gradient Boosting Regressor, Multiple Linear Regression, and Polynomial Regression. The study focuses on parameters such as Average Price, PLU Codes (4046, 4225, 4770), Type, and Year to understand their impact on the dependent variable, Total Volume, within the dynamic U.S. avocado market. The dataset, comprising entries for various regions and years, provides a comprehensive overview. This study focuses on using predictive modeling to find important patterns and connections in the changing avocado industry. The aim is to give stakeholders helpful insights for making informed decisions.

B. Methods of Collecting Data

The data for this research has been sourced from Kaggle, a reputable platform for hosting datasets, and is specifically derived from the "Avocado Prices 2020" dataset created by Timofei Kornev. This dataset, last updated on February 26, 2021, provides detailed information on weekly avocado sales across the United States for the year 2020 [Dataset Link: https://www.kaggle.com/datasets/timmate/avocado-prices-2020].

Kaggle serves as a collaborative platform for data science, offering datasets from various domains. The choice of this dataset aligns with the research objective of analyzing factors influencing avocado sales volume in the USA. As a secondary data source, it leverages real sales data, ensuring the reliability and authenticity of the information. The dataset covers a range of parameters, including date, average price, avocado type, year, region, and various volume metrics, providing a comprehensive foundation for the research analysis.

C. Methods of Research

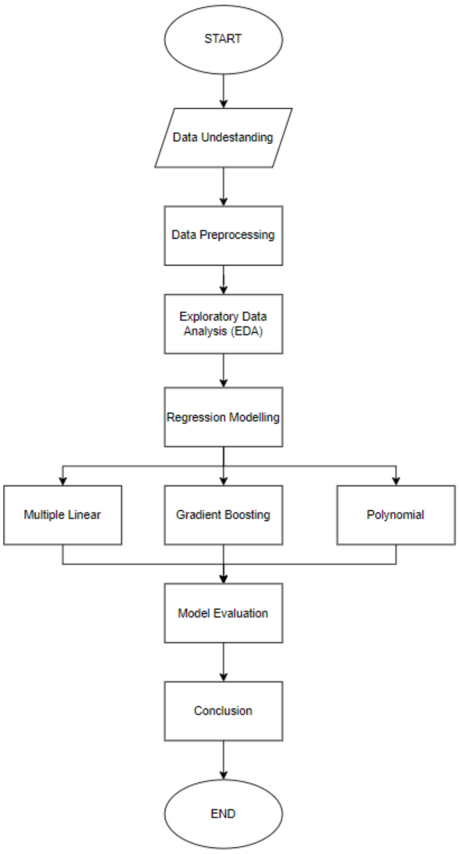


Fig. 1. The Research Framework

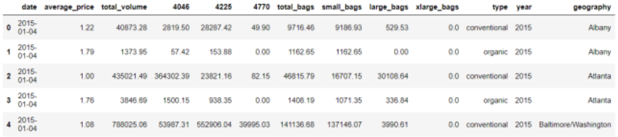
The research team followed a structured approach to analyze avocado sales data in the USA. In the initial phase, the dataset was selected and imported from Kaggle, specifically opting for the "Avocado Prices 2020" dataset created by Timofei Kornev, due to its alignment with the research focus. Once the dataset was imported, the collective effort focused on cleaning and organizing the data to eliminate any irrelevant or messy information, ensuring the dataset's suitability for analysis.

Moving forward, the team explored the dataset to comprehend its characteristics. After this initial phase, each team member individually wrote code to visualize the data, facilitating the identification of trends, patterns, and relationships within the dataset. As regression models emerged as the primary analytical tool, each team member selected a different regression algorithm, including Multiple Linear Regression, Gradient Boosting Regression, and Polynomial Regression.

Following the implementation of individual regression models, the team compared the results, considering the accuracy, predictive capabilities, and overall fit of each algorithm to the dataset. The final step involved drawing conclusions based on the outcomes of the regression models, aiming to provide insights into the factors influencing avocado sales volume in the USA. Each algorithm contributed unique perspectives to the findings.

IV. Result and Discussion

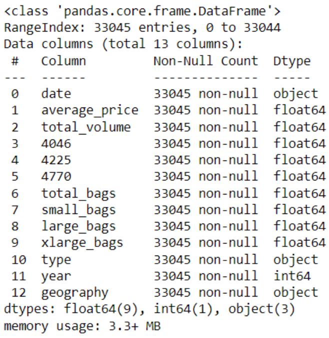
A. Data Understanding



	date	average_price	total_volume	4046	4225	4770	total_bags	small_bags	large_bags	xlarge_bags	type	year	geography
0	2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	conventional	2015	Albany
1	2015-01-04	1.79	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.0	organic	2015	Albany
2	2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.0	conventional	2015	Atlanta
3	2015-01-04	1.76	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.0	organic	2015	Atlanta
4	2015-01-04	1.08	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.0	conventional	2015	Baltimore/Washington

Fig. 2. Import Data

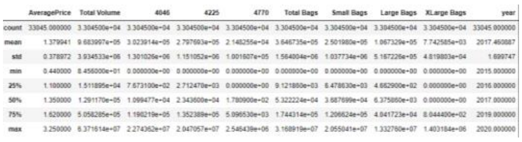
The result above is an output of importing and reading data CSV from “avocado-updated-2020.csv”. Using the Pandas library in Python, the pd.read_csv() function is employed to read the CSV file and create a DataFrame called "avocado" to store the dataset. The head(5) method is then applied to display the first five rows of the dataset, providing a quick overview of its structure and content.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33045 entries, 0 to 33044
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   date           33045 non-null  object 
 1   average_price  33045 non-null  float64
 2   total_volume   33045 non-null  float64
 3   4046           33045 non-null  float64
 4   4225           33045 non-null  float64
 5   4770           33045 non-null  float64
 6   total_bags     33045 non-null  float64
 7   small_bags     33045 non-null  float64
 8   large_bags     33045 non-null  float64
 9   xlarge_bags    33045 non-null  float64
 10  type           33045 non-null  object 
 11  year           33045 non-null  int64  
 12  geography      33045 non-null  object 
dtypes: float64(9), int64(1), object(3)
memory usage: 3.3+ MB
```

Fig. 3. Data Info

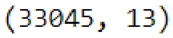
The result presented here is a summary of a dataset using Pandas in Python. It provides information about the structure and content of the dataset. The dataset has 33,045 entries (rows) and 13 columns. Each column represents a different aspect of the data, such as date, average price, total volume, specific volume categories (e.g., 4046, 4225, 4770), total bags, bag sizes (small, large, xlarge), type, year, and geography. The "Non-Null Count" column indicates that there are no missing values in any of the columns.



	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	year
count	33045.000000	3.304500e+04	3.304500e+04	3.304500e+04	3.304500e+04	3.304500e+04	3.304500e+04	3.304500e+04	3.304500e+04	33045.000000
mean	1.379941	9.683997e+05	3.023914e+05	2.797693e+05	2.148255e+04	3.646735e+05	2.501980e+05	1.067329e+05	7.742585e+03	2017.460887
std	0.378972	3.934533e+06	1.301026e+06	1.151052e+06	1.001607e+05	1.564004e+06	1.037734e+06	5.167226e+05	4.819803e+04	1.699747
min	0.440000	8.456000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2015.000000
25%	1.100000	1.511895e+04	7.673100e+02	2.712470e+03	0.000000e+00	9.121860e+03	6.478630e+03	4.662900e+02	0.000000e+00	2016.000000
50%	1.350000	1.291170e+05	1.099477e+04	2.343600e+04	1.780900e+02	5.322224e+04	3.687699e+04	6.375860e+03	0.000000e+00	2017.000000
75%	1.620000	5.058285e+05	1.190219e+05	1.352389e+05	5.096530e+03	1.744314e+05	1.206624e+05	4.041723e+04	8.044400e+02	2019.000000
max	3.250000	6.371614e+07	2.274362e+07	2.047057e+07	2.546439e+06	3.168919e+07	2.055041e+07	1.332760e+07	1.403184e+06	2020.000000

Fig. 4. Data Describe

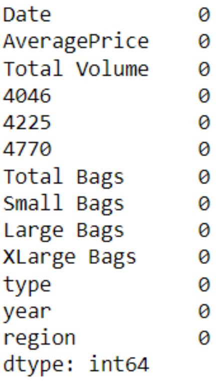
The output above provides statistical summaries for various columns in the dataset after the column name changed. It includes information such as the count of non-null entries, the mean (average), standard deviation (variability), minimum and maximum values, and percentiles (25th, 50th, and 75th).



```
(33045, 13)
```

Fig. 5. Data Shape

The visual representation shown above is the result of analyzing the data structure, providing information about the dataset under examination. This output indicates that the analyzed data comprises 33,045 rows and 13 columns.



```
Date            0
AveragePrice    0
Total Volume    0
4046            0
4225            0
4770            0
Total Bags      0
Small Bags      0
Large Bags      0
XLarge Bags     0
type            0
year            0
region          0
dtype: int64
```

Fig. 6. Find missing value

The visual presented above displays the outcome of a search for missing values in the dataset. This process is employed to identify whether there are any null values in the data, aiming to enhance the accuracy of the analysis. Consequently, based on the displayed result, it can be inferred that there are no null or empty values in the dataset comprising 33,045 rows.

B. *Data Preprocessing*

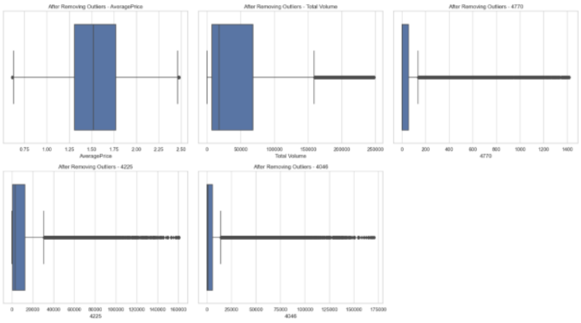
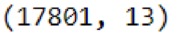


Fig. 7. Data Cleansing

The image presented illustrates a bar plot that visualizes specific aspects of the dataset post-cleansing. The focus is on key variables, namely AveragePrice, 4046, 4770, 4225, and Total Volume. A bar plot is a graphical representation that provides a clear comparison of these variables [8]. This visual representation aids in understanding the relative magnitudes or patterns of the specified variables, offering valuable insights into their post-cleansing characteristics.

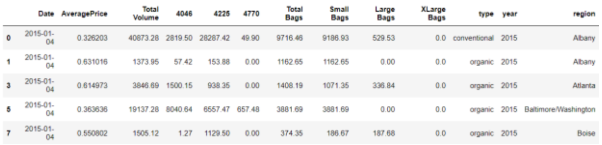


```
(17801, 13)
```

Fig. 8. Data Shape After Cleansing

The output above illustrates the dimensions of the dataset following the application of the Interquartile Range (IQR) method for data cleansing. The result indicates that, after employing the IQR method, the dataset now consists of 17,801 rows and 13 columns.

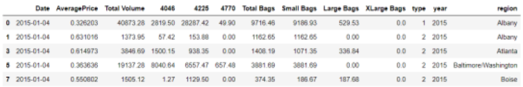
The IQR method is a statistical technique commonly used for outlier detection and removal, aiming to enhance the quality of the dataset by addressing extreme values [9]. This reduction in the number of rows suggests that outliers may have been identified and filtered out during the cleansing process, contributing to a more refined and potentially more reliable dataset for further analysis.



	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-01-04	0.326203	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	conventional	2015	Albany
1	2015-01-04	0.631016	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.0	organic	2015	Albany
3	2015-01-04	0.614973	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.0	organic	2015	Atlanta
5	2015-01-04	0.363636	19137.28	8040.64	6557.47	657.48	3881.69	3881.69	0.00	0.0	organic	2015	Baltimore/Washington
7	2015-01-04	0.550802	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.0	organic	2015	Boise

Fig. 9. Data After Normalization

The output depicts the dataset after applying a normalization technique known as Min-Max Scaling specifically to the 'AveragePrice' column. Min-Max Scaling is a process that transforms numerical data [10], in this case, the 'AveragePrice,' to a common scale, typically between 0 and 1. This normalization ensures that all values are proportionally adjusted, allowing for a more consistent comparison and interpretation of the 'AveragePrice' across the dataset.



	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-01-04	0.326203	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	1	2015	Albany
1	2015-01-04	0.631016	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.0	2	2015	Albany
3	2015-01-04	0.614973	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.0	2	2015	Atlanta
5	2015-01-04	0.363636	19137.28	8040.64	6557.47	657.48	3881.69	3881.69	0.00	0.0	2	2015	Baltimore/Washington
7	2015-01-04	0.550802	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.0	2	2015	Boise

Fig. 10. Data After Encoding

The displayed result represents the dataset after applying a process called encoding, specifically for the 'type' column. In this encoding, the original values 'conventional' have been replaced with the numerical value 1, and 'organic' has been replaced with the numerical value 2. This transformation simplifies the representation of categorical data, making it more suitable for certain machine learning algorithms that require numerical inputs [11]. Consequently, the 'type' column now holds numerical values (1 for 'conventional' and 2 for 'organic), facilitating further analysis and modeling tasks that rely on numerical data.

C. *Exploratory Data Analysis (EDA)*

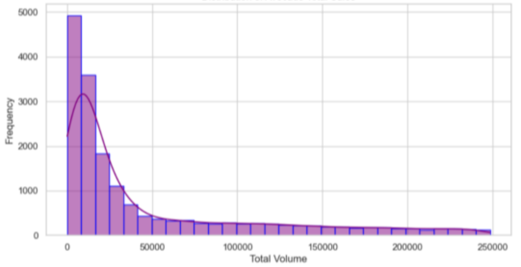


Fig. 11. Total Volume Histogram

The histogram visual above depicts the distribution of the total volume variable in the dataset. A histogram is a graphical representation that illustrates the frequency or occurrence of different ranges or bins of values within a dataset [13]. In this case, the horizontal axis represents the total volume, divided into intervals or bins, while the vertical axis shows the frequency of occurrences within each bin. The histogram indicates that the distribution is right-skewed in nature and not a normal distribution.

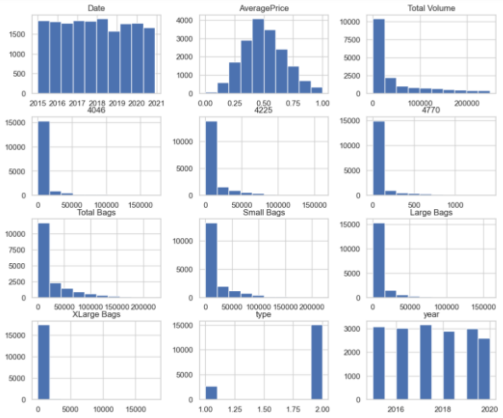


Fig. 12. Data Histogram

The histogram visual above provides a comprehensive overview of the distribution of values across the entire dataset. Remarkably, the 'AveragePrice' column displays a distribution that closely mirrors a normal curve. This suggests that the majority of avocado prices tend to center around a typical value, forming a balanced distribution with fewer prices deviating significantly.

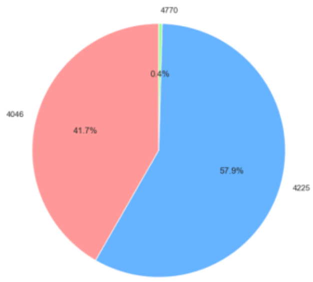


Fig. 13. Composition of Total Volumes between PLU Code

The pie chart output illustrates the composition of total volumes categorized by PLU (Product Lookup) codes. The percentages represent the proportion of each PLU code in the total volume. Specifically, PLU code '4046' constitutes 41.7% of the total volume, '4770' makes up 0.4%, and '4225' represents the majority with 57.9%. This visual representation provides a clear snapshot of how different PLU codes contribute to the overall distribution of total volumes, highlighting the predominant role of PLU code '4225' in the dataset. This visualization indicates that the sales of Avocado PLU Codes vary, and not all of them have a significant impact on the dependent variable.

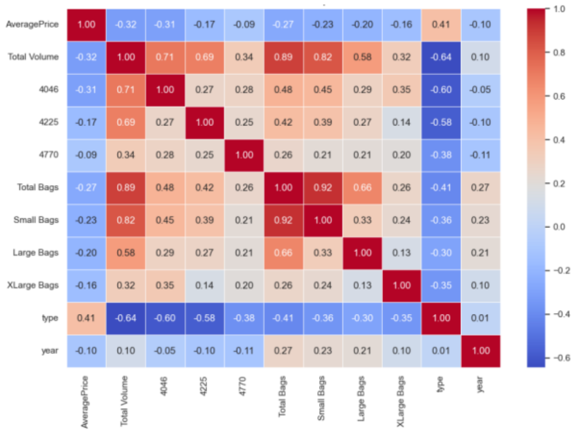


Fig. 14. Correlation between columns

The heatmap output visually represents the correlation coefficients between various variables in the dataset, with scores ranging from -1 to 1. Notable correlations include a negative correlation between 'AveragePrice' and 'Total Volume' (-0.3176), suggesting that as average prices decrease, total volume tends to increase. Additionally, 'Total Volume' exhibits positive correlations with several subcategories such as '4046' (0.7135), '4225' (0.6888), '4770' (0.3353), 'Total Bags' (0.8892), and 'Small Bags' (0.8248), indicating a general positive relationship.

The 'type' variable shows a positive correlation with 'AveragePrice' (0.4092) and a negative correlation with 'Total Volume' (-0.6445), indicating that avocado type may influence pricing and volume. The heatmap serves as a valuable tool for understanding the strength and direction of relationships between different variables, providing insights crucial for analysis and decision-making [14]. This heatmap shows that some of the independent variables has impact on the dependent variable.

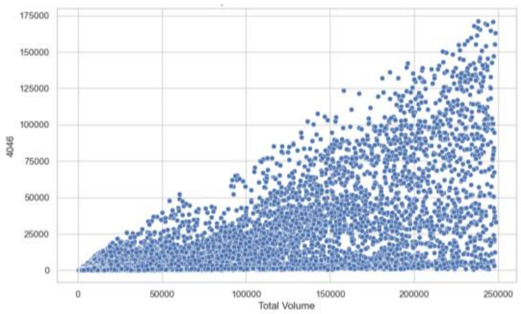


Fig. 15. Scatterplot Total Volume vs 4046 PLU

The scatterplot visual above illustrates the relationship between two variables: 'Total Volume' and '4046.' Each point on the plot represents an observation in the dataset, with the x-axis representing the 'Total Volume' and the y-axis representing the '4046' category. The pattern of points on the plot reveals the nature of their association.

In this specific scatterplot, as ‘4046’ increases, the corresponding values of ‘Total Volume’ also tend to increase, indicating a positive correlation between these two variables. The scatter plot provides a clear visual representation of the trend and distribution of data points [14], allowing quick understanding of the relationship between 'Total Volume' and '4046' PLU code in the dataset. This suggests that a particular PLU code can have a substantial impact on the research.

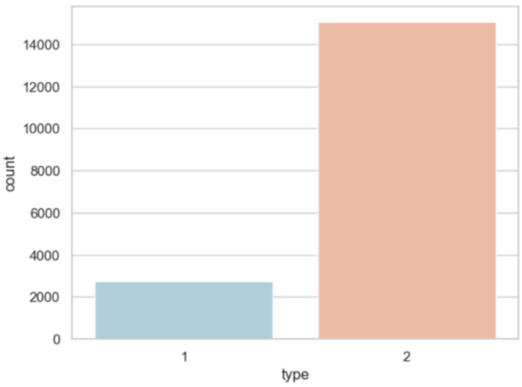


Fig. 16. Count plot of Total Volume by Type

The count plot above visualizes the distribution of 'Total Volume' across two types, denoted as Type 1 and Type 2. The count of occurrences for each type is represented by the height of the bars. In this plot, Type 1 (conventional avocado) has an estimated count of 2,700, while Type 2 (organic avocado) has a count of 15,000.

This count plot offers a straightforward comparison of the 'Total Volume' distribution between the two types, indicating that Type 2 which is organic avocado has a significantly higher count than Type 1 which is conventional avocado. It provides a quick and effective way to grasp the relative prevalence of each type in the dataset in terms of total volume [15]. This suggests a significant difference in sales between organic and conventional types of avocados, which can impact the research.

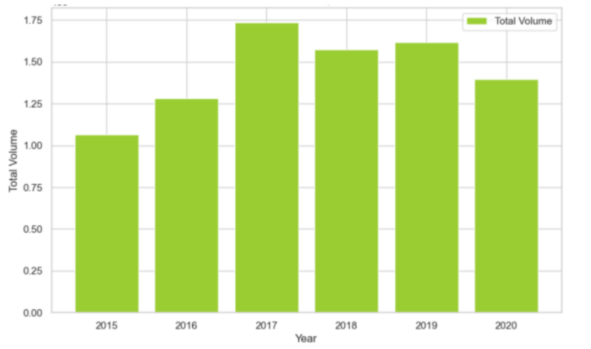
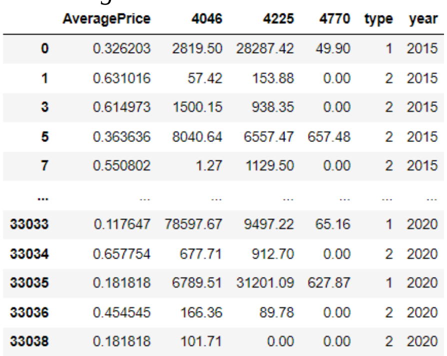


Fig. 17. Bar plot of Total Volume by Year

The bar plot above illustrates the variation in total avocado sales ('Total Volume') across different years. Each bar represents the total volume of avocado sales for a specific year. In this representation, the highest sales are observed in 2017, while the lowest sales occurred in 2015.

This visual comparison of total volume by year provides a clear snapshot of the sales trend over the specified time period. The rising bar in 2017 indicates a peak in avocado sales during that year, while the shorter bar in 2015 reflects a lower total volume. This type of bar plot is valuable for identifying patterns and trends in avocado sales over the years, allowing for a quick and intuitive interpretation of the data [16]. This visualization illustrates the variation in avocado sales across different years, suggesting that the year may be a factor influencing avocado sales.

D. Modelling

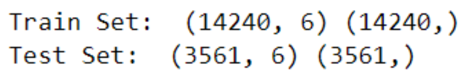


	AveragePrice	4046	4225	4770	type	year
0	0.326203	2819.50	28287.42	49.90	1	2015
1	0.631016	57.42	153.88	0.00	2	2015
3	0.614973	1500.15	938.35	0.00	2	2015
5	0.363636	8040.64	6557.47	657.48	2	2015
7	0.550802	1.27	1129.50	0.00	2	2015
...	...	...	...	...	...	...
33033	0.117647	78597.67	9497.22	65.16	1	2020
33034	0.657754	677.71	912.70	0.00	2	2020
33035	0.181818	6789.51	31201.09	627.87	1	2020
33036	0.454545	166.36	89.78	0.00	2	2020
33038	0.181818	101.71	0.00	0.00	2	2020

Fig. 18. Data X (Independent Variables Column)

The output of the data reveals the set of independent variables, denoted as X, where the 'Total Volume' column has been excluded. This set comprises six columns, each representing a distinct variable.

The removal of the 'Total Volume' column designates it as the dependent variable or the outcome of interest in the context of the analysis. This prepared dataset with the independent variables serves as input for various statistical analyses or machine learning models, allowing for the exploration of relationships and patterns without the influence of the excluded 'Total Volume' variable.



```
Train Set:  (14240, 6) (14240,)
Test Set:  (3561, 6) (3561,)
```

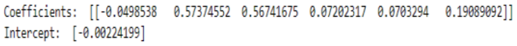
Fig. 19. Data Train Test Split Shape

The shape information above shows that the dataset has undergone a train-test split for model evaluation, with a test size of 20% and a specified random state of 42. Train-test split is a method in machine learning that involves dividing a dataset into a training set, used for model training, and a test set, used to assess the model's performance on unseen data [17]. Additionally, the data has been standardized, ensuring that all variables have a mean of 0 and a standard deviation of 1. Standard Scaler is a preprocessing technique in machine learning that scales and standardizes the features of a dataset, ensuring they have zero mean and unit variance [18].

After this process, the training set comprises 14,240 samples with six features each, along with a corresponding target variable.

On the other hand, the test set consists of 3,561 samples with the same six features and their respective target variables. This split allows for training a predictive model on the training data and evaluating its performance on the separate test set, providing a robust assessment of the model's generalization capabilities [17].

E. Multiple Linear Regression

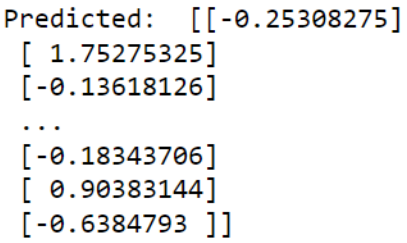


```
Coefficients:  [[-0.0498538   0.57374552  0.56741675  0.07202317  0.0703294   0.19089092]]
Intercept:  [-0.00224199]
```

Fig. 20. Multi Linear Coefficients & Intercept

The results above show the value of coefficient and intercept from the multi linear regression. The output for the coefficients of the multilinear regression model indicates the weights assigned to each independent variable. The intercept is the predicted value of the dependent variable when all independent variables are zero [20]. Together, they define the linear relationship between variables in the model.

In this case, the coefficients are as follows: -0.0498538, 0.57374552, 0.56741675, 0.07202317, 0.0703294, and 0.19089092 for the respective independent variables. The intercept, represented by -0.00224199, signifies the expected value of the dependent variable when all independent variables are zero.



```
Predicted:  [[-0.25308275]
 [ 1.75275325]
 [-0.13618126]
 ...
 [-0.18343706]
 [ 0.90383144]
 [-0.6384793 ]]
```

Fig. 21. y Predict Multi Linear

The results above show an output of the predicted model from y. Model prediction is the process where a trained machine learning model estimates or forecasts outcomes for new or unseen data based on patterns learned during its training [19].

In the provided result, the predicted values are printed, showcasing an array of numerical values, such as [-0.25308275, 1.75275325, -0.13618126, ..., -0.18343706, 0.90383144, -0.6384793]. Each value corresponds to the model's prediction for a specific instance in the test set.

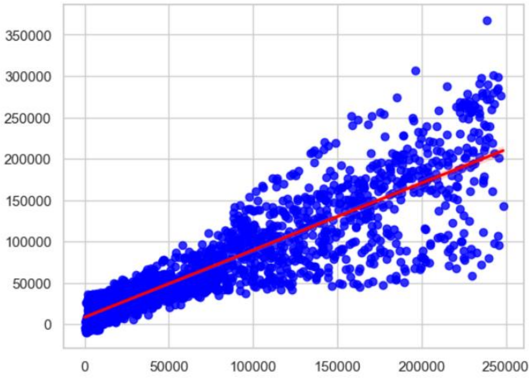
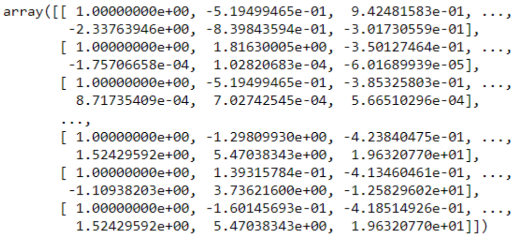


Fig. 22. Scatterplot Multilinear Regression

The output above is a visualization for a multilinear regression model. It first transforms the scaled test set (y_test) and predicted values (y_pred) back to their original scale using the inverse_transform method of the scaler. Then, it creates a scatter plot with blue points representing the actual versus predicted values. The x-axis represents the actual total volume values (y_test_inverse) after inverse scaling, and the y-axis represents the predicted total volume values (y_pred_inverse) after inverse scaling.

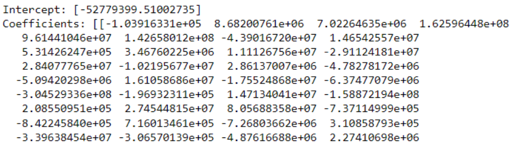
F. Polynomial Regression



```
array([[ 1.00000000e+00, -5.19499465e-01,  9.42481583e-01, ...,
        -2.33763946e+00, -8.39843594e-01, -3.01730559e-01],
       [ 1.00000000e+00,  1.81630005e+00, -3.50127464e-01, ...,
        -1.75706658e-04,  1.02820683e-04, -6.01689939e-05],
       [ 1.00000000e+00, -5.19499465e-01, -3.85325803e-01, ...,
         8.71735409e-04,  7.02742545e-04,  5.66510296e-04],
       ...,
       [ 1.00000000e+00, -1.29809930e+00, -4.23840475e-01, ...,
         1.52429592e+00,  5.47038343e+00,  1.96320770e+01],
       [ 1.00000000e+00,  1.39315784e-01, -4.13460461e-01, ...,
        -1.10938203e+00,  3.73621600e+00, -1.25829602e+01],
       [ 1.00000000e+00, -1.60145693e-01, -4.18514926e-01, ...,
         1.52429592e+00,  5.47038343e+00,  1.96320770e+01]])
```

Fig. 23. Polynomial X Train

The image above is the output of polynomial x train array after it utilizes Polynomial Features with a degree of 7 to transform the original features in the training set (X_train) into polynomial terms. This shows the expanded set of features, including not only the original features but also their various polynomial combinations up to the seventh degree.



```
Intercept: [-52779399.51002735]
Coefficients: [[-1.03916331e+05  8.68200761e+06  7.02264635e+06  1.62596448e+08
   9.61441046e+07  1.42658012e+08 -4.39016720e+07  1.46542557e+07
   5.31426247e+05  3.46760225e+06  1.11126756e+07 -2.91124181e+07
   2.84077765e+07 -1.02195677e+07  2.86137007e+06 -4.78278172e+06
  -5.09420298e+06  1.61058686e+07 -1.75524868e+07 -6.37477079e+06
  -3.04529336e+08 -1.96932311e+05  1.47134041e+07 -1.58872194e+08
   2.08550951e+05  2.74544815e+07  8.05688358e+07 -7.37114999e+05
  -8.42245840e+05  7.16013461e+05 -7.26803662e+06  3.10858793e+05
  -3.39638454e+07 -3.06570139e+05 -4.87616688e+06  2.27410698e+06
```

Fig. 24. Polynomial Coefficient & Intercept

The output displays the coefficients and intercept of a linear regression model. The intercept is approximately -52,779,399.51, and the coefficients correspond to the features in the model. The model appears to have a large number of features, and their coefficients influence the prediction of the target variable in a complex manner.

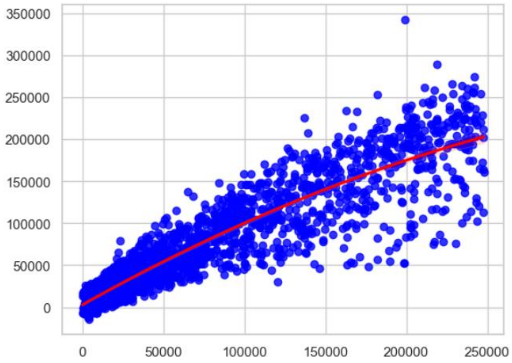
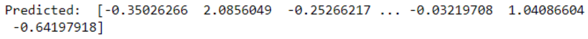


Fig. 25. Scatterplot Polynomial Regression

The plot visualizes the relationship between the predicted and actual values, allowing for an assessment of how well the polynomial regression model fits the data. It transforms the test set predictions (y_pred_poly) and actual test set values (y_test) back to their original scale using inverse transformation with a scaler.

The x-axis represents the actual total volume values (y_test_poly_inverse) after inverse scaling, and the y-axis represents the predicted total volume values (y_pred_poly_inverse) after inverse scaling. The scatterplot is generated for a polynomial regression model with a degree of 2.

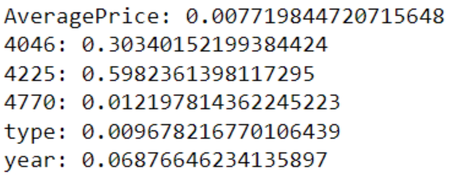
G. Gradient Boosting Regressor



```
Predicted:  [-0.35026266  2.0856049  -0.25266217 ... -0.03219708  1.04086604
 -0.64197918]
```

Fig. 26. y Predict Gradient Boosting Regressor

The image above is the output of using KNN, which has an accuracy of 81%. From this, it can be said that this data has a good value for processing the data.



```
AveragePrice: 0.007719844720715648
4046: 0.30340152199384424
4225: 0.5982361398117295
4770: 0.012197814362245223
type: 0.009678216770106439
year: 0.06876646234135897
```

Fig. 27. Gradient Boosting Regressor Feature Importance

The image above reveals the feature importances of a Gradient Boosting Regressor model applied to avocado sales data. Feature importances are valuable metrics that convey the relative contribution of each input feature to the model's predictions [21].

Among the features, "PLU 4225" (representing a specific avocado type) is the most influential, contributing around 59.82%, followed by "PLU 4046" at approximately 30.34%. The "year" feature also plays a notable role, contributing 6.88%. Conversely, "AveragePrice," "PLU 4770," and "type" exhibit minor importance, each contributing less than 1%.

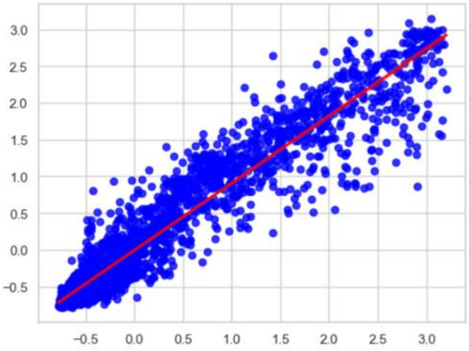
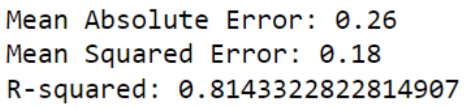


Fig. 28. Scatterplot Gradient Boosting Regressor

The scatterplot visualizes the performance of a Gradient Boosting Regressor (GBR) in predicting total avocado volume based on various features, including average price, specific avocado types (4046, 4225, 4770), and the year. The GBR model is configured with parameters such as n_estimators=100, indicating the use of 100 decision trees in the boosting process, learning_rate=0.1 to control the step size, max_depth=4 to limit tree complexity, and random_state=42 for reproducibility.

The x-axis represents the actual total avocado volume values (y_test) after inverse scaling, while the y-axis represents the predicted total avocado volume values (y_pred_gbr) after inverse scaling.

H. Model Evaluation

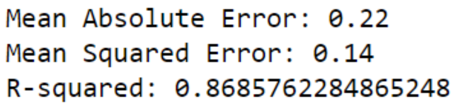


```
Mean Absolute Error: 0.26
Mean Squared Error: 0.18
R-squared: 0.8143322822814907
```

Fig. 29. Multi Linear Regression Evaluation

The picture above shows the evaluation metrics for the multi-linear regression model. The Mean Absolute Error (MAE), which measures the average absolute differences between predicted and actual values [21], stands at 0.26. This implies that, on average, the model's predictions deviate by 0.26 units from the actual values.

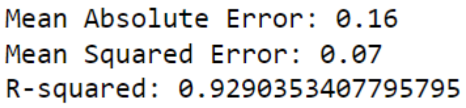
Moving on to the Mean Squared Error (MSE), which averages the squared differences between predictions and actual values [21], it is reported as 0.18. A lower MSE indicates that the model is adept at minimizing larger errors. Lastly, the R-squared value, a measure of how well the model explains the variance in the target variable [21], is notably high at 0.81. This indicates that approximately 81% of the variability in the dependent variable is accounted for by the model.



```
Mean Absolute Error: 0.22
Mean Squared Error: 0.14
R-squared: 0.8685762284865248
```

Fig. 30. Polynomial Regression Evaluation

The picture above shows the evaluation of polynomial regression. The output from the polynomial regression evaluation shows that the model has a Mean Absolute Error (MAE) of 0.22, a Mean Squared Error (MSE) of 0.14, and an R-squared value of 0.87. These metrics collectively indicate the accuracy and performance of the polynomial regression model, with lower MAE and MSE values suggesting less prediction error and a higher R-squared indicating better explanatory power of the model [21].



```
Mean Absolute Error: 0.16
Mean Squared Error: 0.07
R-squared: 0.9290353407795795
```

Fig. 31. Gradient Boosting Regressor Evaluation

The picture above shows the evaluation of gradient boosting regressor. The output from the gradient boosting regressor evaluation reveals that the model has a Mean Absolute Error (MAE) of 0.16, a Mean Squared Error (MSE) of 0.07, and an R-squared value of 0.93. These metrics provide insights into the performance of the gradient boosting regression model. A lower MAE and MSE signify less prediction error, while a higher R-squared indicates a greater proportion of variance in the target variable explained by the model [21].

V. CONCLUSION

In conclusion, the evaluation of three regression models, namely Multi Linear Regression, Polynomial Regression, and Gradient Boosting Regressor, provides valuable insights into their performance in predicting the total volume of avocados based on various features. The Multi Linear Regression model yielded a Mean Absolute Error (MAE) of 0.26, Mean Squared Error (MSE) of 0.18, and an R-squared value of 0.81. The Polynomial Regression demonstrated improved performance with a lower MAE of 0.22, MSE of 0.14, and a higher R-squared value of 0.87. The Gradient Boosting Regressor outperformed both with an even lower MAE of 0.16, MSE of 0.07, and an impressive R-squared value of 0.93. These results suggest that the Gradient Boosting Regressor model excels in capturing the complex relationships within the avocado dataset, providing the most accurate predictions. The Polynomial Regression model also outperformed the Multi Linear Regression, indicating that introducing polynomial features enhances predictive capabilities. Researchers can confidently consider the Gradient Boosting Regressor as the preferred model for predicting total avocado sales, leveraging its superior performance for more reliable and precise forecasting.

ACKNOWLEDGMENT

The author expresses gratitude to Dr. Irmawati, the instructor for the Data Analysis course in the Information Systems Study Program at Multimedia Nusantara University. Dr. Irmawati has generously shared valuable knowledge and insights that have greatly contributed to the progress of this research. Additionally, the author acknowledges Dr. Irmawati's willingness to offer input and guidance from the initial planning stages of this research.

REFERENCES

[1] Rincon-Patino, J., Lasso, E., & Corrales, J. C. (2018). Estimating avocado sales using machine learning algorithms and weather data. *Sustainability*, *10*(10), 3498.

[2] Cavaletto, G. (2015). The avocado market in the United States. *In VIII Congreso Mundial de la Palta*.

[3] Arifin, M. H. (2014). Konsep-konsep Dasar statistika. Jakarta: *Universitas Terbuka*.

[4] Liana, L. (2009). Penggunaan MRA dengan SPSS untuk menguji pengaruh variabel moderating terhadap hubungan antara variabel independen dan variabel dependen. *Dinamik*, 14(2).

[5] Muhamad, I. M., Wardana, S. A., Wanto, A., & Windarto, A. P. (2022). Algoritma Machine Learning untuk penentuan Model Prediksi Produksi Telur Ayam Petelur di Sumatera. *Journal of Informatics, Electrical and Electronics Engineering*, 1(4), 126-134.

[6] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4), 140-147.

[7] Zhan, X., Zhang, S., Szeto, W. Y., & Chen, X. (2020). Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree. *Journal of Intelligent Transportation Systems*, 24(2), 125-141.

[8] Midway, S. R. (2020). Principles of effective data visualization. *Patterns*, 1(9).

[9] Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14, 1-13.

[10] Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. *Int. J. Health Sci*, 6, 10981-10996.

[11] Guedrez, R., Dugeon, O., Lahoud, S., & Texier, G. (2016, October). Label encoding algorithm for MPLS segment routing. In 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA) (pp. 113-117). IEEE.

[12] Affleck, M. (1992). The United States avocado market. In Proc. 2nd World Avocado Congr (Vol. 2, pp. 643-645).

[13] Herho, S. H. S. (2019). Tutorial Visualisasi Data

Menggunakan Seaborn.

[14] Guntara, R. G. (2023). Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab. *ULIL ALBAB: Jurnal Ilmiah Multidisiplin*, 2(6), 2091-2100.

[15] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, 1, 1-6.

[16] Larson–Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics 1. *The Modern Language Journal*, 101(1), 244-270.

[17] Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). A critical look at the current train/test split in machine learning. *arXiv preprint arXiv:2106.04525*.

[18] Shivani, C., Anusha, B., Druvitha, B., & Swamy, K. K. (2022). RNN-LSTM Model Based Forecasting of Cryptocurrency Prices Using Standard Scaler Transform. *J. Crit. Rev*, *10*, 144-158.

[19] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., ... & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.

[20] Bouwmeester, W., Twisk, J. W., Kappen, T. H., van Klei, W. A., Moons, K. G., & Vergouwe, Y. (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC medical research methodology, 13, 1-10.

[21] Mohan, A., Chen, Z., & Weinberger, K. (2011, January). Web-search ranking with initialized gradient boosted regression trees. *In Proceedings of the learning to rank challenge* (pp. 77-89). PMLR.

[22] Priya Varshini, A. G., & Anitha Kumari, K. (2020). Predictive analytics approaches for software effort estimation: A review. *Indian J. Sci. Technol*, 13, 2094-2103.