Performance Evaluation of Random Forest Regressor Algorithm Compared to Multiple Linear and Polynomial Regressors for Avocado Sales Volume Forecasting in the USA

Evaluasi Performa Algoritma Regressor Random Forest Dibandingkan dengan Regressor Multiple Linear dan Polinomial untuk Peramalan Volume Penjualan Alpukat di Amerika Serikat

1st Kevin Aditya Hartono
Majoring in Information Systems,
Faculty of
Technology and Information
Universitas Multimedia Nusantara
Tangerang, Indonesia
kevin.aditya@student.umn.ac.id

2nd Nathan Vilbert Kosasih
Majoring in Information Systems,
Faculty of
Technology and Information
Universitas Multimedia Nusantara
Tangerang, Indonesia
nathan.vilbert@student.umn.ac.id

3rd Julius Calvin Saputra
Majoring in Information Systems,
Faculty of
Technology and Information
Universitas Multimedia Nusantara
Tangerang, Indonesia
julius.calvin@student.umn.ac.id

4th Juanito Arvin William
Majoring in Information Systems,
Faculty of
Technology and Information
Universitas Multimedia Nusantara
Tangerang, Indonesia
juanito.arvin@student.umn.ac.id

5th Christopher Abie Diaz
Majoring in Information Systems,
Faculty of
Technology and Information
Universitas Multimedia Nusantara
Tangerang, Indonesia
christopher.abie@student.umn.ac.id

Accepted on May 25, 2024
Approved on May 25, 2024

Abstract— This study employs regression analysis to investigate the impact of independent variables, including average price, PLU codes 4046, 4225, 4770, avocado type, and year, on the total volume of Avocado sales in the USA. Comparing Multiple Linear Regression, Polynomial Regression, and Random Forest methods, the results reveal that the Random Forest algorithm outperforms with an impressive MSE score of 0.04, MAE score of 0.11, and r^2 score of 0,96. This underscores its superior accuracy in predicting avocado sales volume compared to the other methods. The findings affirm the substantial influence of selected independent variables on avocado sales, establishing Random Forest as the optimal regression model for this analysis and emphasizing the significance of these variables in determining total sales volume.

Index Terms— Regression analysis; Random Forest; Multiple Linear Regression; Polynomial Regression; Avocado sales

I. Introduction

Avocado is a popular fruit that has many health benefits and culinary uses. In recent years, the demand for avocado has increased significantly in the USA, creating a need for accurate and reliable forecasting of sales volume. Forecasting can help market planners and stakeholders optimize their production, distribution, and pricing strategies. However, forecasting avocado sales is not a simple task, as it involves multiple factors and complex interactions among them. In the realms of sales and marketing, the utilization of data analytics and visualization tools proves instrumental in providing businesses with supplementary information and valuable insights [1]. Therefore, it is essential to use appropriate statistical methods that can capture the underlying patterns and relationships in the data.

Regression analysis is a common statistical method that can model the dependence of the sales volume (the dependent variable) on various independent variables [2]. It can also provide insights into how each variable affects the sales volume and how they interact with each other. However, different regression methods may have different assumptions, performance, and interpretability, and not all of them may be suitable for forecasting avocado sales. Hence, it is important to compare and evaluate different regression methods and select the best one for the analysis.

This study compares the performance of three regression methods: Random Forest, Multiple Linear Regression, and Polynomial Regression. The researcher uses a dataset of avocado sales from 2015 to 2020, obtained from the Hass Avocado Board 3. The researcher applies each method to the data and measures their accuracy and efficiency using various metrics, such as mean absolute error, root mean square error, and coefficient of determination. The researcher also discusses the advantages and disadvantages of each method and their implications for forecasting avocado sales. The purpose of this study is to evaluate and contrast various regression methods for predicting avocado sales and to determine which method is the most effective for this problem.

II. Literature Review

A. *Hass Avocado Board*

Avocado is a nutritious and versatile fruit that has become a household staple in the United States. The Hass Avocado Board (HAB) is the organization that supports the growth and success of the avocado industry in the U.S. market [4]. HAB represents growers and importers from different countries, such as Mexico, Chile, Peru, Colombia, and the U.S. itself. HAB collects and distributes funds to maintain and expand the demand for avocados in the U.S. HAB also provides the industry with market data, nutrition research, health education, and sustainability practices [4]. HAB works collaboratively with all stakeholders to make avocados America’s most popular fruit.

B. *Big Data*

Big data refers to the large and complex data sets that traditional software cannot process [5]. It comes from various sources, such as the internet, sensors, cameras, logs, etc. It has five characteristics: volume, variety, velocity, veracity, and value. These describe how big, diverse, fast, reliable, and useful the data is.

People can use big data for many purposes, such as business analytics, scientific research, social media, healthcare, etc. Big data can reveal insights and patterns that can enhance decision making, innovation, customer experience, risk management, and more. However, big data also has some challenges, such as data storage, data integration, data analysis, data security, and data privacy [5]. Therefore, big data needs advanced technologies and methods to process and manage it effectively.

C. *Data Mining*

Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [6]. Data mining can use various techniques, such as classification, clustering, association, and others, to help users manage and analyze data [7]. Data mining can help users find useful information from existing data, such as groups of data records, unusual records, and dependencies.

D. *Machine Learning*

Machine learning is a branch of artificial intelligence that uses algorithms to learn from data and make predictions or decisions based on the data analysis [8]. It enables computers to improve their performance over time without explicit programming. In supervised learning, the algorithm is given a dataset with the correct output for each example. The algorithm then learns to predict the output for new examples based on the patterns it has learned from the dataset. In unsupervised learning, the algorithm has to find patterns and relationships in the data without any labels. Reinforcement learning is a method of training an algorithm to make optimal decisions in a situation by rewarding it for good actions [8].

E. *Regression Machine Learning*

Regression machine learning is a technique that uses algorithms to model the relationship between one or more input variables (features) and a continuous output variable (target) [9]. It can be used to predict or estimate the value of the target variable based on the values of the features. Regression machine learning can be divided into different types, such as linear, logistic, polynomial, and random forest, depending on the shape and complexity of the relationship between the variables [9]. Regression machine learning is useful for various applications, such as forecasting, trend analysis, optimization, and data exploration

F. *Random Forest Regressor*

Random Forest Regressor, a machine learning algorithm designed for regression tasks, excels in predicting continuous numerical values [10]. As an ensemble model, it comprises multiple decision trees collaborating to generate predictions, with each tree trained on a distinct subset of the data. This ensemble approach proves beneficial in handling intricate datasets and curbing overfitting by leveraging random subset selection during training [10]. Despite its strengths, challenges exist, including the diminished interpretability of predictions and increased computational costs. Nevertheless, the algorithm finds widespread use in applications prioritizing accuracy, where the interpretability of individual trees takes a back seat. This strategic ensemble methodology effectively addresses the overfitting concerns associated with training a solitary decision tree model on a dataset [10].

G. *Multiple Linear Regressor*

Multiple linear regression is a statistical method that uses an equation to model how a continuous outcome variable depends on two or more predictor variables. It can be used to measure how much each predictor variable contributes to the outcome variable, while holding the other predictor variables constant. Some benefits of multiple linear regression are that it can capture complex and nonlinear patterns in the data, provide coefficients that indicate the direction and strength of the relationships, and test whether the relationships are statistically significant [11]. Some drawbacks are that it can be affected by extreme values, high correlation among predictor variables, unequal variance of errors, large sample size requirements, and low interpretability when there are many predictor variables [11]. The multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

Fig. 1. Multi Linear Regression Equation

Where Y is the outcome variable, X1 to Xp are the predictor variables, b0 is the constant term, b1 to bp are the coefficients, and ϵ is the error term. The coefficients show the effect of each predictor variable on the outcome variable, holding the other predictor variables constant. For example, b1 shows how much Y changes when X1 increases by one unit and the other predictor variables stay the same. Statistical tests can be used to check if the coefficients are significantly different from zero. Multiple linear regression proves beneficial for exploring the influence of multiple factors on an outcome variable [11].

H. *Polynomial Regressor*

Polynomial regression is a statistical technique that utilizes an equation to model the relationship between a continuous dependent variable and one or more independent variables raised to different powers. It is employed to capture nonlinear patterns in the data that a simple linear regression cannot discern [12]. Advantages of polynomial regression include its ability to accommodate more complex data, offer increased accuracy and flexibility, and enable testing the statistical significance of the polynomial terms. However, drawbacks encompass the potential for overfitting, multicollinearity, and high variance. Careful selection of the polynomial degree is necessary, and interpretation and generalization can be challenging. Polynomial regression proves beneficial when modeling data exhibiting curved or wavy trends [12]. The polynomial regression equation is as follows:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2^2 + \ldots + b_k x_k^k$$

Fig. 2. Polynomial Regression Equation

Where Y is the predicted value of the outcome variable, b0 is the intercept, and b1 to bk are the coefficients for each degree of the predictor variable X. The polynomial model is a type of linear regression model with k terms of X raised to different powers from 1 to k. A polynomial model of degree 2 is a quadratic function, a polynomial model of degree 3 is a cubic function, and a polynomial model of degree 4 is a quartic function [12].

III. Methodology

A. *Object of Research*

This research aims to compare three machine learning algorithms for forecasting avocado sales volume in the USA: Random Forest Regressor, Multiple Linear Regressor, and Polynomial Regressor. The objective is to evaluate the effectiveness of these algorithms and identify the best one for predicting avocado sales volume in the USA. To do this, researcher gather data on various factors that influence total sales volume, such as the average price, PLU codes 4046, 4225, 4770, avocado type, and year. Researcher then use different evaluation metrics to assess the performance of the three algorithms and determine which one is the most reliable and accurate for forecasting avocado sales volume in the USA. This research adopts a quantitative approach, as it involves collecting and analyzing numerical data using statistical methods. The data is not obtained through questionnaires, surveys, polls, or interviews, and the sample size is usually larger than in qualitative research. The outcomes of the study can be presented in statistical terms and used to support decision-making and problem-solving.

B. *Methods of Collecting Data*

This research relies on secondary data from other sources, which was accessed through the Kaggle website. Kaggle is a website or platform that offers datasets from different countries, as well as hosting data visualization and data modeling contests [13]. The dataset used in this research is called Avocado Prices (2020), and it contains historical data on avocado prices and sales volume in various US markets. This dataset is an updated version of the one published by Justin Kiggins in 2018, based on the data from the Hass Avocado Board (HAB) [3]. The updated dataset covers the period from 4 January 2015 to 17 May 2020, while the original dataset only had data from 2015 to early 2018. The website for the dataset: (https://www.kaggle.com/datasets/timmate/avocado-prices-2020/data).

The rationale behind choosing this dataset is grounded in its adherence to specific research criteria, boasting a substantial scale of 33,045 rows and 13 columns, far exceeding the minimum threshold of 1,000 rows and 6 columns. The dataset demonstrates a well-balanced distribution of variables, comprising 2 temporal, 9 numeric, and 2 categorical attributes. Significantly, these variables encapsulate crucial and meaningful information that is vital for conducting a thorough analysis of the factors that impact avocado sales volume. The dataset provides an optimal framework, furnishing a meticulous and inclusive viewpoint on the observed population. In essence, it stands as a sturdy foundation for the identification and thorough examination of a myriad of factors linked to the overall sales volume of avocados in the United States.

C. Research Method

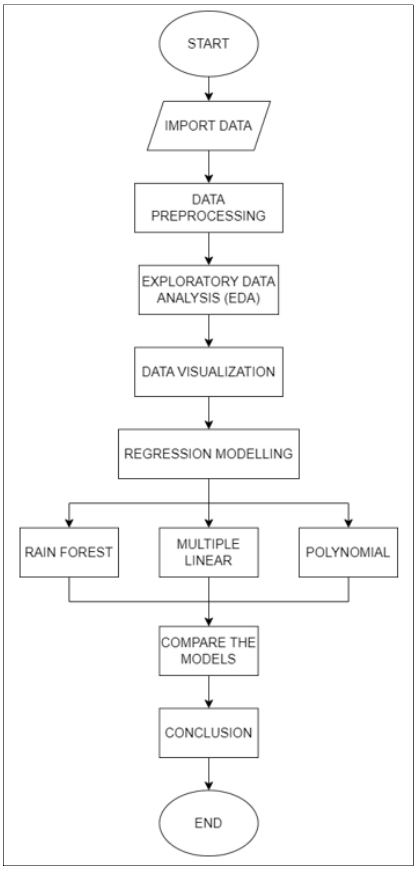


Fig. 3. The Research Framework

Figure 3 illustrates the framework of the research conducted by the team, comprised of Juanito Arvin, Julius Calvin, Nathan Vilbert, Christopher Diaz, and Kevin Aditya (researcher). The team looked for a good dataset on the Kaggle website, and picked the avocado dataset. Then, they put the dataset into the Jupyter notebook and started to analyze the data. They cleaned and changed the data to make it better and easier for the research goals.

After they imported the data, they made some graphs and charts to show the data. The team did this to answer the research question, which was how the different factors (average price, PLU codes 4046, 4225, 4770, avocado type, and year) affected the total amount of avocados sold in the USA.

Next, the team opted for the Regression method, specifically utilizing both Polynomial and Multiple Linear Regression, given the measurable variables directly tied to sales in the avocado dataset. As these methods demonstrated high accuracy, the researchers sought to enhance precision by incorporating and comparing the Random Forest Regressor. In the implementation of this algorithm, the team meticulously compared the models, evaluating correlations with the dataset. The comparison of the three algorithms was based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) score, along with visualizations to enhance interpretability. This comprehensive assessment formed the basis for a definitive conclusion, culminating in the research's intended accurate response to the formulated thesis.

IV. RESULT AND DISCUSSION

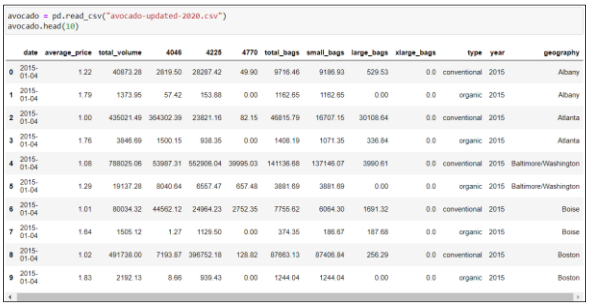
A. Business Understanding

The sales volume of avocados in the United States is influenced by various factors, such as price, PLU code, type, and year. This study aims to find out the most suitable machine learning algorithm for forecasting the sales volume of avocados in the United States. The researcher collected historical data on the factors that affect the sales volume of avocados in the United States. The researcher then applied three machine learning algorithms, namely Random Forest Regressor, Multiple Linear Regressor, and Polynomial Regressor, to create a prediction model for the sales volume of avocados.

This study aims to compare and evaluate the performance of the three machine learning algorithms using different evaluation metrics. This study also wants to identify the most effective, reliable, and accurate machine learning algorithm for forecasting the sales volume of avocados in the United States.

B. Data Understanding

Data understanding is an essential step in the data analysis and machine learning process. It is essential to conduct a thorough examination and understanding of the dataset being utilized [14]. This involves gaining insights into its format, features, and content.

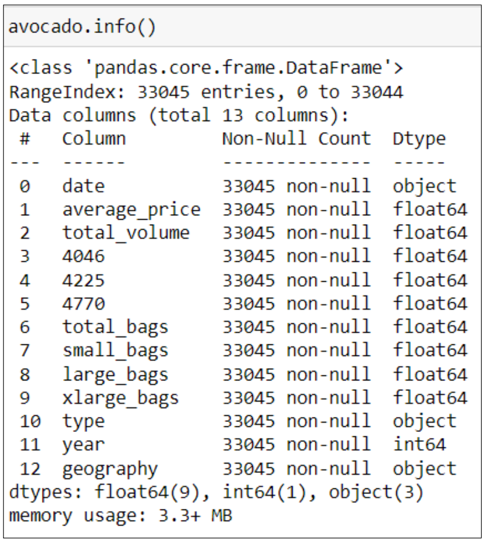


```
avocado = pd.read_csv("avocado-updated-2020.csv")
avocado.head(10)
```

	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
5	2015-01-04	1.29	19137.28	8040.64	6557.47	657.48	3881.69	3881.69	0.00	0.0	organic	2015	Baltimore/Washington
6	2015-01-04	1.01	80034.32	44562.12	24964.23	2752.35	7755.62	6064.30	1691.32	0.0	conventional	2015	Boise
7	2015-01-04	1.64	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.0	organic	2015	Boise
8	2015-01-04	1.02	491738.00	7193.87	396752.18	128.82	87663.13	87406.84	256.29	0.0	conventional	2015	Boston
9	2015-01-04	1.83	2192.13	8.66	939.43	0.00	1244.04	1244.04	0.00	0.0	organic	2015	Boston

Fig. 4. Import Dataset in Python Notebook

The image above shows the output of importing data, which starts with loading the package needed for the analysis. The pandas library in Python is used to convert the “avocado-updated-2020.csv” dataset into a Data Frame. The *pd.read_csv()* function reads the CSV-formatted dataset. The researcher can then examine the dataset’s structure, check the data types, handle missing values, and get a preliminary overview of the variables’ distribution, among other tasks. The next step after loading the data is to show the data using the *.head()* code.

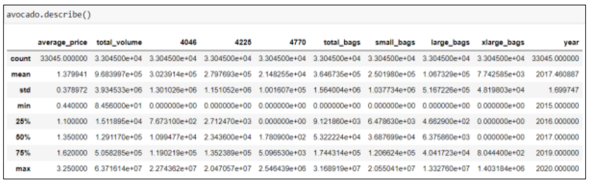


```
avocado.info()
```

```
<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. 5. Data Frame Information

Figure 5 presented showcases the output of the data info, revealing the names of each column along with the corresponding count of data entries. Each column contains 33,045 data points. Furthermore, the image provides insights into the data types of each variable. Notably, there are nine variables with float data types, one with an integer data type, and three with object data types.

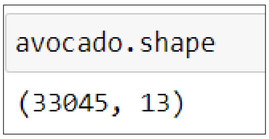


```
avocado.describe()
```

	average_price	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. 6. Describe Data Frame

The image above is data that displays the entire description of the data. The *.describe()* code is a useful tool to get a summary of the statistics from a data frame. It shows the number of entries (count), the mean, the standard deviation (std), the minimum and maximum value (min max), and the quartiles (25%, 50%, and 75%) for each column. This helps to understand the basic characteristics and variability of the data.

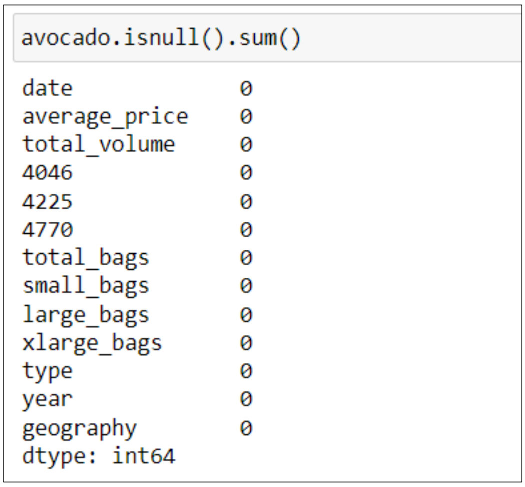


```
avocado.shape
```

```
(33045, 13)
```

Fig. 7. Data Shape

The image displays the data shape, providing fundamental details about the analyzed dataset. It indicates the quantity of rows and columns in the data, revealing that there are 33,045 rows and 13 columns.



```
avocado.isnull().sum()
```

```
date             0
average_price    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
geography        0
dtype: int64
```

Fig. 8. Total of Missing Values from each Columns

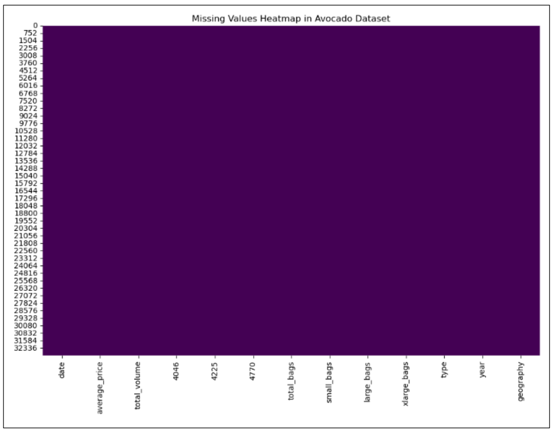
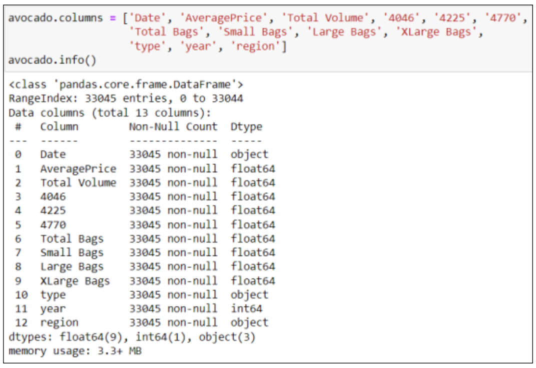


Fig. 9. Visualization of Missing Values from each Columns

The images depicted above illustrate the outcome of a search for missing values in the dataset. This process is crucial for identifying any instances where data may be null or absent, ultimately enhancing the accuracy of the analysis [15]. As indicated in Fig. 6, the total count of missing values, and further confirmed by the visualization in Fig. 7, there are no missing values detected in any columns. Consequently, based on this analysis, we can confidently conclude that among the 33,045 data points, there are no null or empty values present.

C. *Data Preprocessing*

Data preprocessing is the process of getting the dataset ready for processing or analysis [16]. This process involves various activities such as cleaning data, filling in missing values, changing formats, and adjusting variables to fit the analysis or modeling needs.



```
avocado.columns = ['Date', 'AveragePrice', 'Total Volume', '4046', '4225', '4770',
                   'Total Bags', 'Small Bags', 'Large Bags', 'XLarge Bags',
                   'type', 'year', 'region']
avocado.info()
```

```
<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   AveragePrice  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  region        33045 non-null  object 
dtypes: float64(9), int64(1), object(3)
memory usage: 3.3+ MB
```

Fig. 10. Rename Column

Figure 10 displays the columns of the avocado dataset that have undergone a renaming process, resulting in clearer and more user-friendly names. This strategic renaming is intended to enhance the ease of reference and facilitate smoother processing within the domain of machine learning applications. The adoption of intuitive and descriptive column names contributes to improved comprehensibility and accessibility of the dataset, aligning it more effectively with the requirements and conventions of machine learning algorithms and analyses.

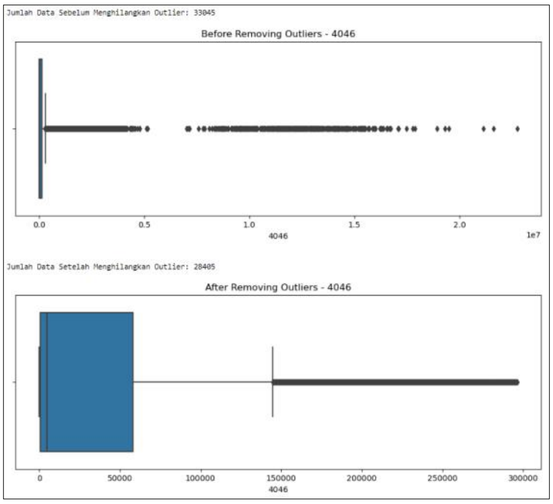


Fig. 11. Handling Outliers for the PLU 4046 Column

The outlier handling process using the Interquartile Range (IQR) method on the PLU 4046 column has had a significant impact on the dataset's total number of records. IQR outlier handling is a method of detecting and removing outliers from a dataset based on the interquartile range (IQR), which is the difference between the 25th and 75th percentiles of the data [17]. Prior to the removal of outliers, the dataset contained a total of 33,045 data points. After implementing the IQR steps to address outliers, the number of data points decreased to 28,405. This action aims to enhance the stability and accuracy of data analysis by mitigating the influence of extreme values that deviate significantly from the quartile values. The objective is to create a more stable dataset, supporting consistent and reliable data analysis and model development.

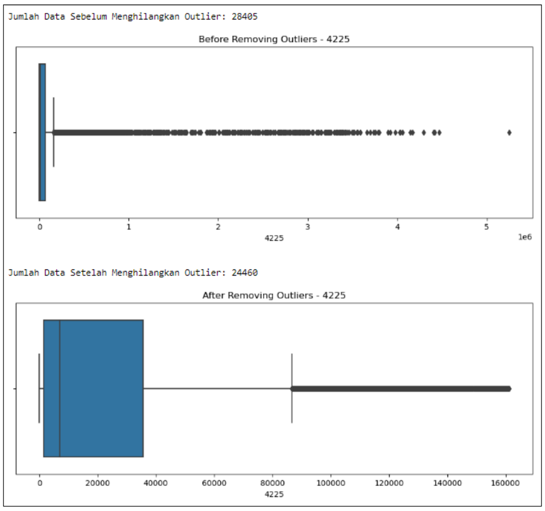


Fig. 12. Handling Outliers for the PLU 4225 Column

The outlier handling process using the Interquartile Range (IQR) method on the PLU 4225 column has a notable impact on the dataset's total number of records. Prior to outlier elimination, the dataset contained a total of 28,405 data points. After implementing the IQR steps to address outliers, the number of data points decreased to 24,460.

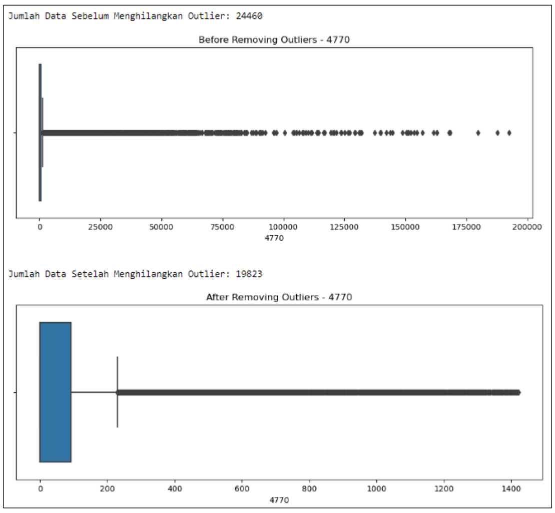
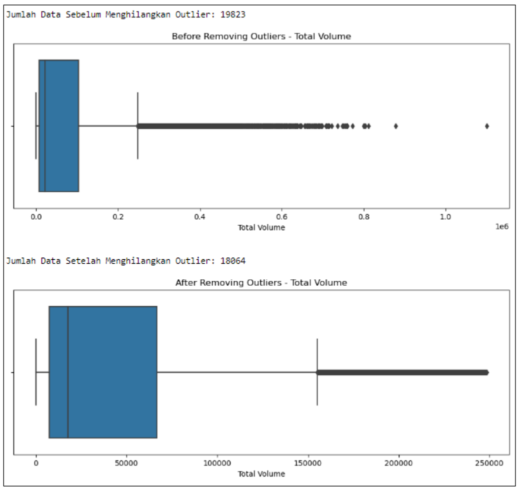


Fig. 13. Handling Outliers for the PLU 4770 Column

The process of outlier handling with the Interquartile Range (IQR) method on the PLU 4770 column has a significant impact on the amount of data in the dataset. Before removing outliers, the total data available was 24,460. After applying the IQR steps to handle outliers, the amount of data decreased to 19,823.

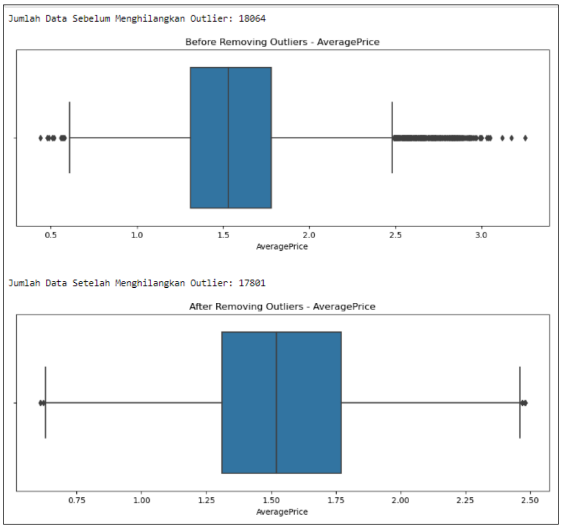


```
Jumlah Data Sebelum Menghilangkan Outlier: 19823
```

```
Jumlah Data Setelah Menghilangkan Outlier: 18064
```

Fig. 14. Handling Outliers for the Total Volume Column

The process of outlier handling using the Interquartile Range (IQR) method on the Total Volume column resulted in a significant change in the amount of data in the dataset. Before removing outliers, the total data available was 19,823. After applying the IQR steps to handle outliers, the amount of data decreased to 18,064.

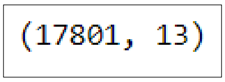


```
Jumlah Data Sebelum Menghilangkan Outlier: 18064
```

```
Jumlah Data Setelah Menghilangkan Outlier: 17801
```

Fig. 15. Handling Outliers for the AveragePrice Column

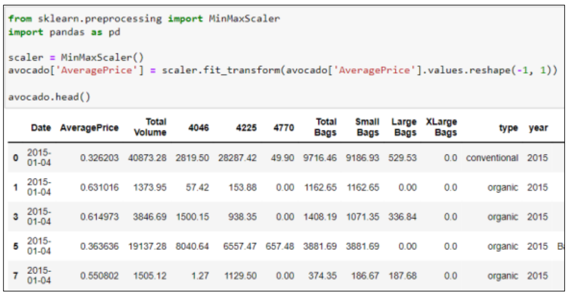
The outlier handling process utilizing the Interquartile Range (IQR) method for the Average Price column has resulted in a discernible impact on the dataset's total number of observations. Prior to outlier removal, the dataset comprised 18,064 data points. Following the implementation of IQR-based outlier handling, the number of data points decreased slightly to 17,801.



```
(17801, 13)
```

Fig. 16. Data Shape After Cleansing

The image above shows the shape of the dataset after using the Interquartile Range (IQR) method to clean the data. The IQR method is a common way to find and remove outliers, which are extreme values that can affect the quality of the data [18]. By using the IQR method, the dataset has fewer rows (17,801) and the same number of columns (13), which means that some outliers have been eliminated. This makes the dataset more reliable and suitable for further analysis.

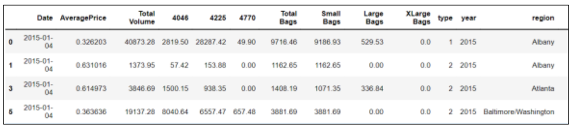


```python
from sklearn.preprocessing import MinMaxScaler
import pandas as pd

scaler = MinMaxScaler()
avocado['AveragePrice'] = scaler.fit_transform(avocado['AveragePrice'].values.reshape(-1, 1))

avocado.head()
```

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

Fig. 17. Normalization on AveragePrice Column

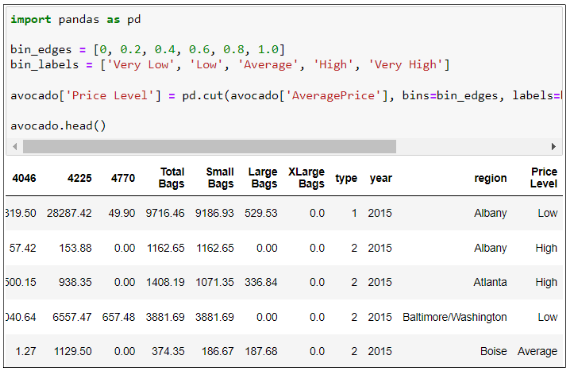
The image above shows the dataset after using Min-Max Scaling, a normalization technique, on the ‘AveragePrice’ column. Min-Max Scaling is a method that changes numerical data [19], such as the ‘AveragePrice,’ to a standard scale, usually from 0 to 1. This normalization makes sure that all values are scaled proportionally, making it easier to compare and understand the ‘AveragePrice’ throughout 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

Fig. 18. Encoding on type Column

Figure 18 shows the dataset after using a technique called encoding on the ‘type’ column. Encoding is a process that changes categorical data, such as ‘conventional’ and ‘organic’, to numerical data, such as 1 and 2 [20]. This makes the data easier to use for some machine learning algorithms that need numerical inputs. The ‘type’ column now has numerical values (1 for ‘conventional’ and 2 for ‘organic’), which helps with further analysis and modeling that depend on numerical data. Encoding also reduces the dimensionality of the data, which means that it reduces the number of possible values that the data can take. This can improve the performance and accuracy of some machine learning algorithms, as well as reduce the computational complexity and memory usage [20]. Encoding is a common and useful technique for transforming categorical data into numerical data for machine learning purposes.



```
import pandas as pd

bin_edges = [0, 0.2, 0.4, 0.6, 0.8, 1.0]
bin_labels = ['Very Low', 'Low', 'Average', 'High', 'Very High']

avocado['Price Level'] = pd.cut(avocado['AveragePrice'], bins=bin_edges, labels=

avocado.head()
```

4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region	Price Level
319.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	1	2015	Albany	Low
57.42	153.88	0.00	1162.65	1162.65	0.00	0.0	2	2015	Albany	High
500.15	938.35	0.00	1408.19	1071.35	336.84	0.0	2	2015	Atlanta	High
040.64	6557.47	657.48	3881.69	3881.69	0.00	0.0	2	2015	Baltimore/Washington	Low
1.27	1129.50	0.00	374.35	186.67	187.68	0.0	2	2015	Boise	Average

Fig. 19. Binning Price Level

The image above explains the binning process on the AveragePrice column of the avocado dataset. Binning is the process of grouping continuous values into several intervals [21]. In this case, the defined intervals are [0, 0.2, 0.4, 0.6, 0.8, 1.0] and each interval is labeled as 'Very Low', 'Low', 'Average', 'High', 'Very High'.

D. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) is a technique that delves into data sets to uncover their key traits, often utilizing graphical representations [22]. Its main purpose is to extract insights from data that go beyond formal modeling or hypothesis testing, thereby enhancing our comprehension of the variables within the data set and the interconnections among them.

Distribution

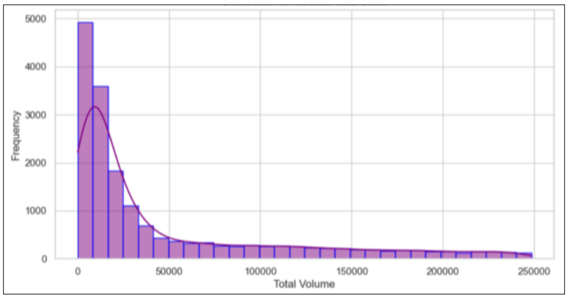


Fig. 20. Total Volume Distribution

The image shown above provides a visual representation of the distribution of the ‘total volume’ variable in our dataset. Histograms are graphical tools that display the frequency of various ranges, or ‘bins’, of values within a dataset [23]. Here, the ‘total volume’ is represented on the horizontal axis and is divided into different bins. The vertical axis, on the other hand, represents the frequency of data points falling within each bin. From the histogram, we can observe that the distribution of ‘total volume’ is skewed to the right, indicating it does not follow a normal distribution.

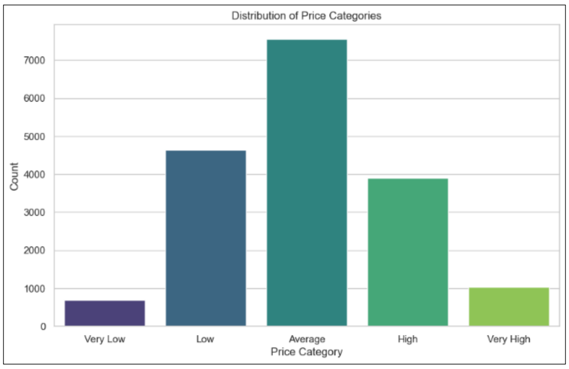


Fig. 21. Price Level Distribution

The Gaussian Distribution, also known as the bell curve because of its unique shape, is a probability distribution that is balanced around its mean [26]. In this scenario, the ‘Price Level’ adheres to a Gaussian or normal distribution. This is due to the uniform distribution of the ‘Price Level’. It can be inferred that the price level of avocados is predominantly average.

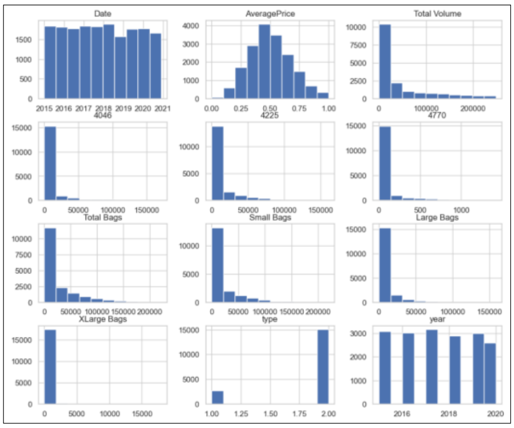


Fig. 22. Price Level Comparison

The histogram displayed above offers an in-depth analysis of the distribution of values throughout the dataset. Notably, the distribution of the ‘AveragePrice’ column bears a striking resemblance to a normal curve. This implies that most of the avocado prices are concentrated around a common value, resulting in a symmetrical distribution with fewer instances of significant deviations.

Composition

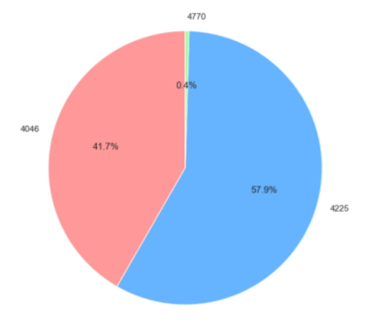


Fig. 23. Composition of Total Volumes between PLU Code

The pie chart shows how much each PLU (Product Lookup) code contributes to the total volume of avocados sold. PLU code ‘4225’ has the largest share with 57.9%, followed by ‘4046’ with 41.7%. PLU code ‘4770’ has the smallest share with only 0.4%. The chart gives a quick overview of the relative importance of each PLU code in the dataset. It also suggests that the demand for different types of avocados is not equal, and some PLU codes have more influence on the sales than others.

Correlation

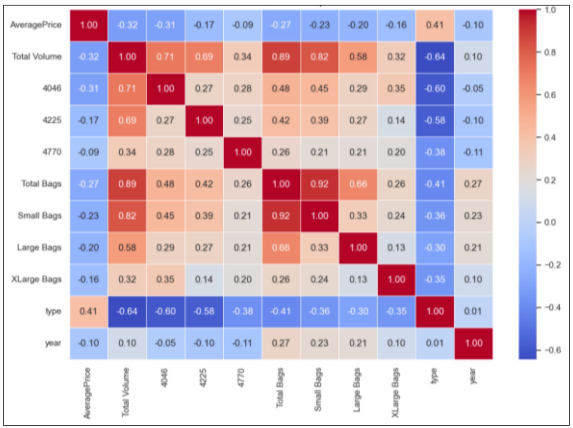


Fig. 24. Correlation between columns

The heatmap output visually illustrates the correlation coefficients between various variables in the dataset [24], with values ranging from -1 to 1. It’s interesting to note the negative correlation between ‘AveragePrice’ and ‘Total Volume’ (-0.3176), suggesting that a rise in total volume typically corresponds to a drop in average prices. Moreover, ‘Total Volume’ is positively correlated 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 generally positive relationship.

The variable ‘type’ has a positive correlation with ‘AveragePrice’ (0.4092) and a negative correlation with ‘Total Volume’ (-0.6445), suggesting that the type of avocado could affect both pricing and volume. The heatmap is an essential tool for comprehending the strength and direction of relationships between different variables, providing vital insights for analysis and decision-making. This heatmap indicates that certain independent variables significantly influence the dependent variable.

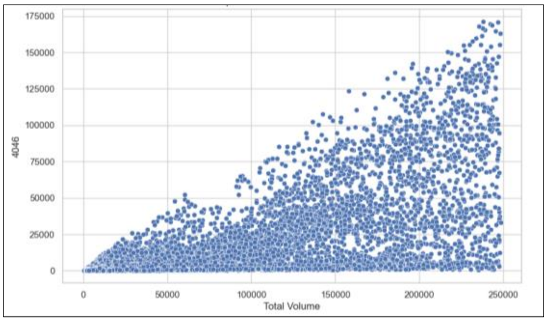


Fig. 25. Correlation Total Volume vs 4046 PLU

The scatterplot visual displayed above depicts the correlation between the two variables: ‘Total Volume’ and ‘4046.’ Each dot on the plot symbolizes a data point in the dataset, with the x-axis denoting the ‘Total Volume’ and the y-axis denoting the ‘4046’ category. The arrangement of dots on the plot uncovers the type of relationship they share. In this specific scatterplot, an increase in ‘4046’ is accompanied by a corresponding increase in ‘Total Volume,’ indicating a positive correlation between these two variables. The scatter plot offers a lucid visual depiction of the data points’ trend and distribution [25], facilitating a swift comprehension of the relationship between ‘Total Volume’ and ‘4046’ PLU code in the dataset. This implies that a specific PLU code can significantly influence the research outcome.

Comparison

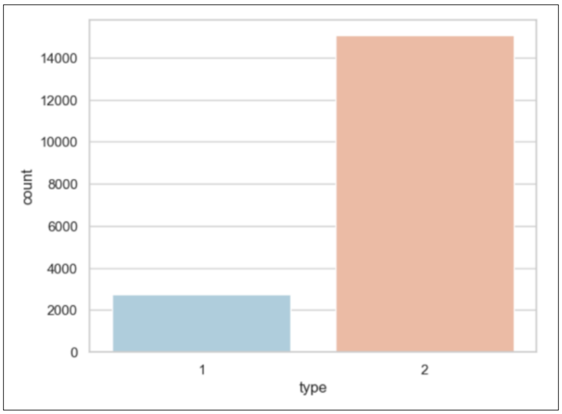


Fig. 26. Total Volume by Type Comparison

The count plot displayed above provides a visual representation of the ‘Total Volume’ distribution across two categories, labeled as Type 1 and Type 2. The frequency of each type is denoted by the height of the bars. In this specific plot, Type 1 (representing conventional avocados) has an approximate count of 2,700, while Type 2 (representing organic avocados) has a count of 15,000.

This count plot facilitates a direct comparison of the ‘Total Volume’ distribution between the two types, highlighting that Type 2 (organic avocados) has a considerably higher count than Type 1 (conventional avocados). It offers an efficient method to comprehend the relative occurrence of each type in the dataset in terms of total volume [27]. This indicates a substantial difference in sales between organic and conventional avocado types, which could significantly influence the research outcomes.

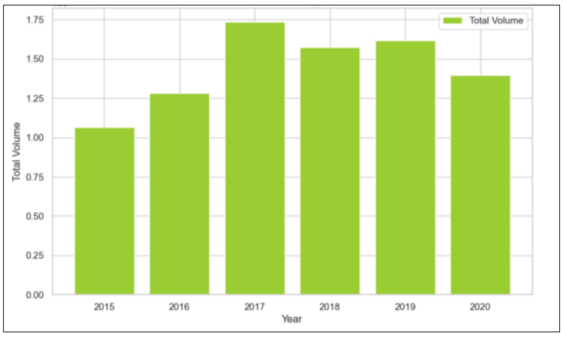


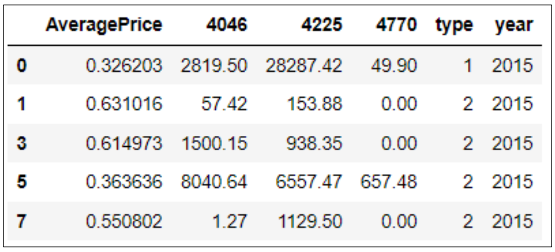
Fig. 27. Total Volume by Year Comparison

The bar plot displayed above provides a visual depiction of the fluctuations in total avocado sales (‘Total Volume’) over different years. Each bar signifies the total volume of avocado sales for a particular year. In this visualization, the peak sales are noted in 2017, while the trough is seen in 2015.

This graphical comparison of total volume per year offers a clear overview of the sales trend during the given time frame. The ascending bar in 2017 signifies a surge in avocado sales for that year, while the descending bar in 2015 indicates a dip in total volume. This kind of bar plot is instrumental in spotting patterns and trends in avocado sales over time [27], facilitating a rapid and intuitive understanding of the data. This visualization underscores the fluctuations in avocado sales over various years, implying that the year could be a contributing factor to avocado sales.

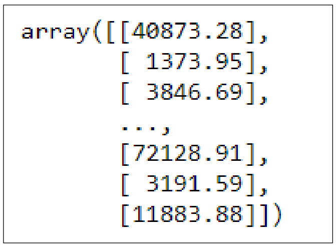
E. *Regression Modelling*

Data modeling in machine learning is the process of creating a representation of data that can be used by a machine learning algorithm to learn from and make predictions [12].



	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

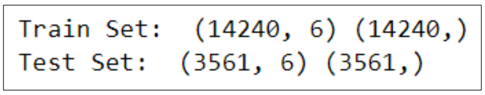
Fig. 28. Independent Variables (data X)



```
array([[40873.28],
       [ 1373.95],
       [ 3846.69],
       ...,
       [72128.91],
       [ 3191.59],
       [11883.88]])
```

Fig. 29. Dependent Variables (data y)

In figure 28, the array of independent variables, labeled as X, is depicted, excluding the 'Total Volume' column. This collection consists of six columns, each signifying a unique variable. The omission of the 'Total Volume' column assigns it the role of the dependent variable or the focal point of interest in the analytical context, as illustrated in figure 29. This refined dataset, comprising solely the independent variables, functions as input for diverse statistical analyses or machine learning models. This setup facilitates the investigation of relationships and patterns free from the influence of the excluded 'Total Volume' variable.



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

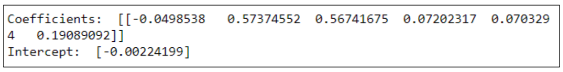
Fig. 30. Train Set and Test Set

The dataset has been subjected to a train-test split for model evaluation, allocating a 20% test size and utilizing a specific random state of 42. This choice of a 20% test size is motivated by the dataset's substantial size, exceeding 1000 columns, ensuring an ample amount of data for both training and testing purposes. The train-test split is a fundamental machine learning technique involving the division of a dataset into a training set, employed for model training, and a test set, employed for evaluating the model's performance on unseen data [28].

Moreover, the data has undergone standardization, ensuring that all variables exhibit a mean of 0 and a standard deviation of 1. Standard Scaler, a preprocessing technique in machine learning, has been applied to scale and standardize the features of the dataset, guaranteeing they possess zero mean and unit variance [29].

Following these procedures, the training set comprises 14,240 samples, each with six features, accompanied by the corresponding target variable. In contrast, the test set encompasses 3,561 samples, featuring the same six attributes along with their respective target variables. This division allows for the training of a predictive model on the training data and the subsequent evaluation of its performance on the distinct test set, offering a robust assessment of the model's generalization capabilities [28].

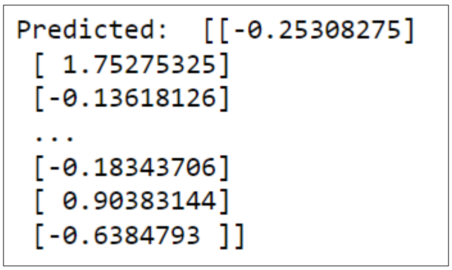
F. *Multiple Linear Regressor*



```
Coefficients:  [[-0.0498538   0.57374552  0.56741675  0.07202317  0.070329
4   0.19089092]]
Intercept:  [-0.00224199]
```

Fig. 31. Multiple Linear Coefficients & Intercept

A Multiple Linear regression model describes how a dependent variable relates to multiple independent variables. The coefficients of the model above show how much the dependent variable changes when one independent variable increases by one unit, holding the others constant. The intercept of the model shows the predicted value of the dependent variable when all independent variables are zero [11]. For example, in this model, the coefficients are -0.0498538, 0.57374552, 0.56741675, 0.07202317, 0.0703294, and 0.19089092 for each independent variable, respectively. The intercept is -0.00224199, which means the expected value of the dependent variable is close to zero when all independent variables are zero.



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

Fig. 32. Multiple Linear y_pred

Model prediction is the process of using a trained machine learning model to estimate or forecast outcomes for new or unseen data based on patterns learned during its training [11]. The values above are predictions for the dependent variable (y) based on a multilinear regression model. Each value in the output corresponds to one value in X_test. For example, the first value [-0.25308275] indicates the prediction for the first test value, the second value [1.75275325] for the second test value, and so on. The meaning of these values will depend on the context of the problem and the dependent variable used in the model.

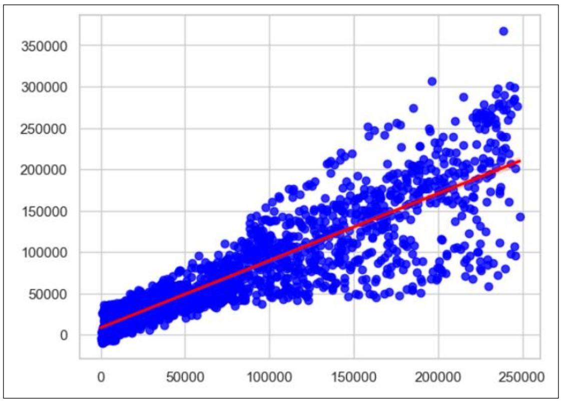
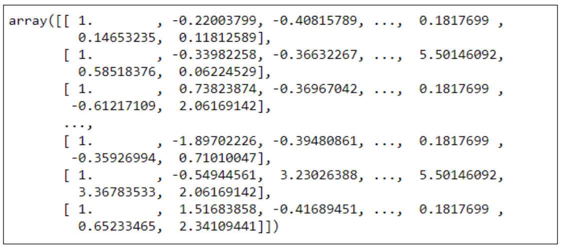


Fig. 33. Scatterplot of the Multiple Linear Regressor Model

The graph above shows how well a multilinear regression model fits the data. The model uses multiple independent variables to predict the total volume of a product. Before applying the model, the data was scaled to have a mean of zero and a standard deviation of one using a scaler object. To interpret the results, the scaled test set (y_test) and the predicted values (y_pred) were transformed back to their original scale using the inverse_transform method of the scaler. The graph plots the actual total volume values (y_test_inverse) on the x-axis and the predicted total volume values (y_pred_inverse) on the y-axis. Each blue point represents a pair of actual and predicted values for a test observation. The closer the points are to the diagonal line, the better the model's performance.

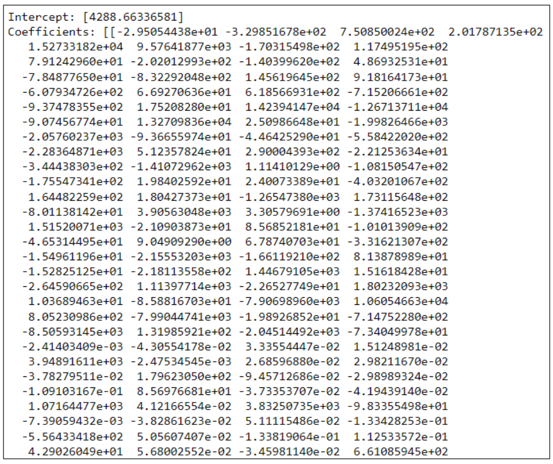
G. *Polynomial Regression*



```
array([[ 1.        , -0.22003799, -0.40815789, ...,  0.1817699 ,
         0.14653235,  0.11812589],
       [ 1.        , -0.33982258, -0.36632267, ...,  5.50146092,
         0.58518376,  0.06224529],
       [ 1.        ,  0.73823874, -0.36967042, ...,  0.1817699 ,
        -0.61217109,  2.06169142],
       ...,
       [ 1.        , -1.89702226, -0.39480861, ...,  0.1817699 ,
        -0.35926994,  0.71010047],
       [ 1.        , -0.54944561,  3.23026388, ...,  5.50146092,
         3.36783533,  2.06169142],
       [ 1.        ,  1.51683858, -0.41689451, ...,  0.1817699 ,
         0.65233465,  2.34109441]])
```

Fig. 34. Polynomial Regression Model X Train

The original features in the training set (X_train) were transformed into polynomial terms using Polynomial Features with a degree of 7 [12]. The image above shows the result of this transformation, which is the polynomial x train array. This array contains not only the original features but also their various combinations up to the seventh power. This creates a larger and more complex set of features.



```
Intercept: [4288.66336581]
Coefficients: [[-2.95054438e+01 -3.29851678e+02  7.50850024e+02  2.01787135e+02
   1.52733182e+04  9.57641877e+03 -1.70315498e+02  1.17495195e+02
   7.91242960e+01 -2.02012993e+02 -1.40399620e+02  4.86932531e+01
  -7.84877650e+01 -8.32292048e+02  1.45619645e+02  9.18164173e+01
  -6.07934726e+02  6.69270636e+01  6.18566931e+02 -7.15206661e+02
  -9.37478355e+02  1.75208280e+01  1.42394147e+04 -1.26713711e+04
  -9.07456774e+01  1.32709836e+04  2.50986648e+01 -1.99826466e+03
  -2.05760237e+03 -9.36655974e+01 -4.46425290e+01 -5.58422020e+02
  -2.28364871e+03  5.12357824e+01  2.90004393e+02 -2.21253634e+01
  -3.44438303e+02 -1.41072962e+03  1.11410129e+00 -1.08150547e+02
  -1.75547341e+02  1.98402592e+01  2.40073389e+01 -4.03201067e+02
   1.64482259e+02  1.80427373e+01 -1.26547380e+03  1.73115648e+02
  -8.01138142e+01  3.90563048e+03  3.30579691e+00 -1.37416523e+03
   1.51520071e+03 -2.10903873e+01  8.56852181e+01 -1.01013909e+02
  -4.65314495e+01  9.04909290e+00  6.78740703e+01 -3.31621307e+02
  -1.54961196e+01 -2.15553203e+03 -1.66119210e+02  8.13878989e+01
  -1.52825125e+01 -2.18113558e+02  1.44679105e+03  1.51618428e+01
  -2.64590665e+02  1.11397714e+03 -2.26527749e+01  1.80232093e+03
   1.03689463e+01 -8.58816703e+01 -7.90698960e+03  1.06054663e+04
   8.05230986e+02 -7.99044741e+03 -1.98926852e+01 -7.14752280e+02
  -8.50593145e+03  1.31985921e+02 -2.04514492e+03 -7.34049978e+01
  -2.41403409e-03 -4.30554178e-02  3.33554447e-02  1.51248981e-02
   3.94891611e+03 -2.47534545e-03  2.68596880e-02  2.98211670e-02
  -3.78279511e-02  1.79623050e+02 -9.45712686e-02 -2.98989324e-02
  -1.09103167e-01  8.56976681e+01 -3.73353707e-02 -4.19439140e-02
   1.07164477e+03  4.12166554e-02  3.83250735e+03 -9.83355498e+01
  -7.39059432e-03 -3.82861623e-02  5.11115486e-02 -1.33428253e-01
  -5.56433418e+02  5.05607407e-02 -1.33819064e-01  1.12533572e-01
   4.29026049e+01  5.68002552e-02 -3.45981140e-02  6.61085945e+02
```

Fig. 35. Polynomial Regression Coefficients & Intercept

When all independent variables (X) are zero, the intercept is a constant value that shows the value of the dependent variable (y). In this case, the intercept is 4288.66336581, which means the value of the dependent variable is around 4288.66 when all independent variables are zero. The coefficients are related to the features in the model. The model has many features, and their coefficients affect the target variable's prediction in a complicated way.

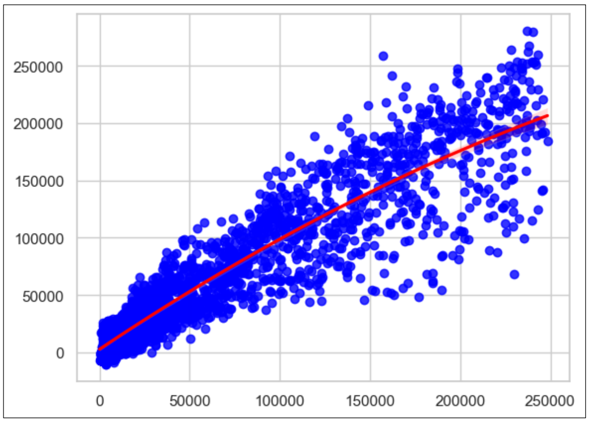
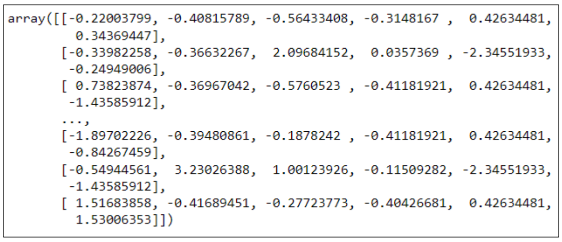


Fig. 36. Scatterplot of the Polynomial Regressor Model

Facilitating the evaluation of the polynomial regression model's data fitting performance, the chart visually represents the relationship between predicted and actual values. The process involves reversing the test set predictions (y_pred_poly) and actual test set values (y_test) to their original scale through an inverse transformation using a scaler. On the graph, the x-axis showcases the actual total volume values (y_test_poly_inverse) after the inverse scaling procedure, while the y-axis exhibits the predicted total volume values (y_pred_poly_inverse) after the inverse scaling. This scatterplot is specifically designed for a polynomial regression model with a degree of 5.

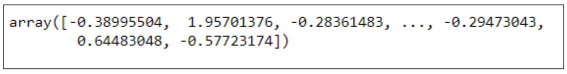
H. *Random Forest Regressor*



```
array([[-0.22003799, -0.40815789, -0.56433408, -0.3148167 ,  0.42634481,
         0.34369447],
       [-0.33982258, -0.36632267,  2.09684152,  0.0357369 , -2.34551933,
        -0.24949006],
       [ 0.73823874, -0.36967042, -0.5760523 , -0.41181921,  0.42634481,
        -1.43585912],
       ...,
       [-1.89702226, -0.39480861, -0.1878242 , -0.41181921,  0.42634481,
        -0.84267459],
       [-0.54944561,  3.23026388,  1.00123926, -0.11509282, -2.34551933,
        -1.43585912],
       [ 1.51683858, -0.41689451, -0.27723773, -0.40426681,  0.42634481,
         1.53006353]])
```

Fig. 37. Random Forest Regression Model X_train

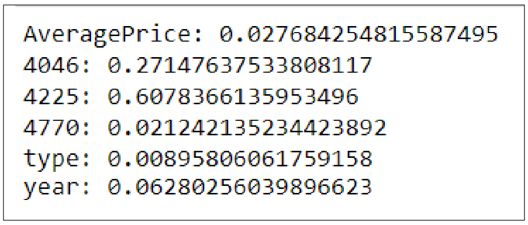
The image above appears to be the feature matrix (X_train) used in training a RandomForestRegressor model. Each row in the matrix corresponds to a specific data point, and each column represents a different feature [10]. The features have been standardized or normalized, as the values vary around zero. The six columns likely correspond to six distinct features used for prediction.



```
array([-0.38995504,  1.95701376, -0.28361483, ..., -0.29473043,
        0.64483048, -0.57723174])
```

Fig. 38. Random Forest Regression Model y_pred

Figure 38 represent the model’s estimates or predictions for the target variable based on the input features. Each value corresponds to the predicted outcome for a specific data point in the dataset [10]. For instance, the first value (-0.38995504) is the predicted result for the first data point, the second value (1.95701376) corresponds to the second data point, and so on. This pattern continues throughout the array, with each value representing a prediction for a specific data point.



```
AveragePrice: 0.027684254815587495
4046: 0.27147637533808117
4225: 0.6078366135953496
4770: 0.021242135234423892
type: 0.008958060617591 58
year: 0.06280256039896623
```

Fig. 39. Random Forest Regressor Model Feature Importance

The total volume of avocado sales is primarily influenced by the number of large Hass avocados sold (4225), which accounts for 60.78% of the variation in sales. Following closely is the total number of small/medium Hass avocados sold (4046), contributing 27.15% to the variation. The year of observation is the third most important feature (feature importance of 0.0628), suggesting a temporal trend or seasonality in sales, possibly influenced by factors like supply, demand, price, weather, or consumer preferences. The average price of avocados (AveragePrice) is the fourth most important feature, explaining 2.77% of the variation, indicating a relatively lower impact compared to other features. The total number of extra-large Hass avocados sold (4770) is the fifth most important feature, with a contribution of 2.12%. The type of avocado (conventional or organic) is the least impactful feature, explaining only 0.90% of the variation, suggesting a negligible influence on total sales compared to other factors.

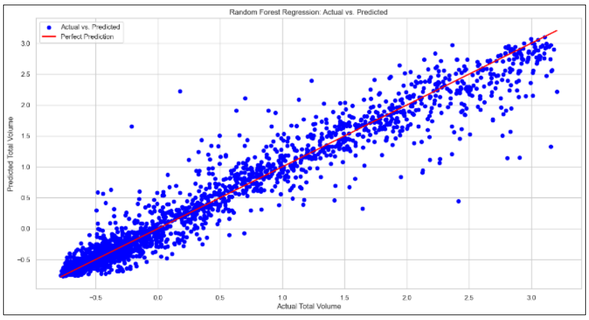


Fig. 40. Scatterplot of the Random Forest Regressor Model

The scatterplot above visualizes the predictions from the Random Forest Regressor model by comparing the actual Total Volume values with the predicted ones. The blue points scattered around the red line, which represents a perfect prediction, signify the model's performance. The red line is a diagonal with a 45-degree slope, indicating the ideal scenario where predicted values perfectly match actual values. The close clustering of the blue points around the red line suggests a high correlation between the model's predictions and the actual values. While there are deviations from the perfect prediction line, the compact distribution and adherence to the red line pattern indicate a high level of accuracy in the model's predictions. In summary, the visualization provides a positive overview of the Random Forest Regressor's predictive quality concerning the actual Total Volume values.

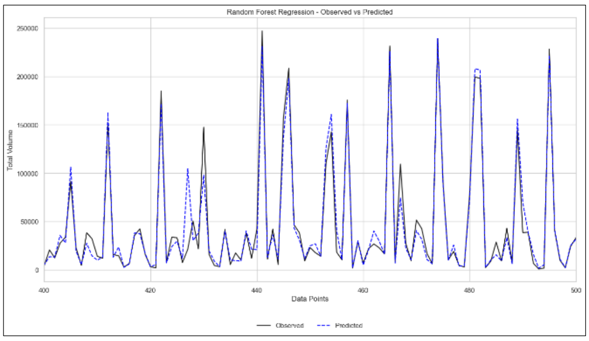
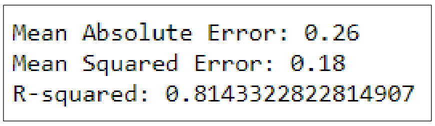


Fig. 41. Line Plot of the Random Forest Regressor Model

The line plot generated from the Random Forest Regressor model indicates a remarkably high level of accuracy, with patterns closely resembling those between observed data and predicted outcomes. This depiction underscores the model's effectiveness in capturing the complexity of relationships within the data. The strong alignment between observations and predictions further substantiates the claim that the model exhibits excellent accuracy in forecasting total sales volume at each level.

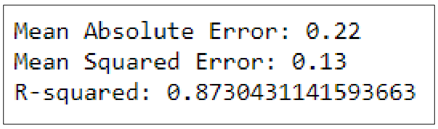
I. Model Evaluation



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

Fig. 42. Multi Linear Regression MAE, MSE and R^2

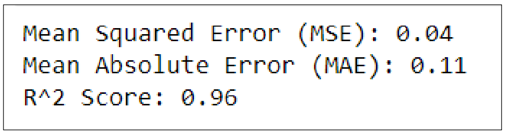
The evaluation of a multi-linear regression model predicting total avocado sales volume reveals promising performance metrics. With a Mean Absolute Error (MAE) of 0.26, the model demonstrates an average absolute difference between predicted and actual values, emphasizing its accuracy. The Mean Squared Error (MSE) of 0.18 underscores the model's precision by measuring the average squared differences, with a lower MSE indicating a higher level of accuracy, particularly in handling larger errors. The commendable R-squared (R²) value of 0.8143 signifies the model's ability to explain approximately 81.43% of the variability in the data, reflecting a strong capacity to capture and elucidate observed variations.



```
Mean Absolute Error: 0.22
Mean Squared Error: 0.13
R-squared: 0.8730431141593663
```

Fig. 43. Polynomial Regression MAE, MSE and R^2

The evaluation results for the polynomial regression model predicting a certain outcome are as follows: The Mean Absolute Error (MAE) of 0.22 indicates that, on average, the absolute difference between the predicted and actual values is approximately 0.22. A lower MAE suggests better accuracy in predicting actual values. The Mean Squared Error (MSE) of 0.13 measures the average squared difference between predicted and actual values. The low MSE implies that the model has a high level of precision in its predictions, with a particular emphasis on minimizing larger errors. The R-squared (R²) value of 0.87 signifies that around 87.30% of the variability in the data can be explained by the polynomial regression model. This high R-squared value indicates a strong ability of the model to capture and elucidate the variations present in prediction total sales volume of avocado.



```
Mean Squared Error (MSE): 0.04
Mean Absolute Error (MAE): 0.11
R^2 Score: 0.96
```

Fig. 44. Random Forest Regression MAE, MSE and R^2

The evaluation results for the random forest regression model are as follows: The Mean Absolute Error (MAE) of 0.04 indicates that, on average, the absolute difference between the predicted and actual values is approximately 0.04. A lower MAE suggests better accuracy in predicting actual values. The Mean Squared Error (MSE) of 0.11 measures the average squared difference between predicted and actual values. The relatively low MSE implies that the model has a high level of precision in its predictions, with a particular emphasis on minimizing larger errors. The R-squared (R²) value of 0.96 is exceptionally high, signifying that around 96% of the variability in the data can be explained by the random forest regression model. This indicates an outstanding ability of the model to capture and elucidate the variations present in the dataset.

V. Conclusion

Multiple machine learning models have been employed to predict the total sales volume of avocados in the USA. This study specifically focuses on three regression models for property price prediction: Multiple Linear Regressor, Polynomial Regressor, and Random Forest Regressor. Among these, the Random Forest Regressor demonstrates the highest accuracy, making it the most suitable choice for forecasting avocado total sales volume due to its utilization of a combination of multiple decision trees.

The Multiple Linear Regression model produced a Mean Absolute Error (MAE) of 0.26, Mean Squared Error (MSE) of 0.18, and an R-squared value of 0.81. Polynomial Regression showed improved performance with a lower MAE of 0.22, MSE of 0.14, and a higher

R-squared value of 0.87. However, the Random Forest Regressor surpass both, achieving a MAE of 0.04, MSE of 0.11, and an impressive R-squared value of 0.96.

The analysis also reveals that several independent variables, including Average Price, PLU Codes (4046, 4225, 4770), Type (conventional or organic), and Year, exert a substantial influence on the total volume of avocado sales in the USA. Particularly noteworthy is the PLU Code 4225, representing the total number of large Hass avocados sold, with a significant impact of 60.78%. To advance future research in this domain, it is advisable to employ a more precise machine learning prediction model, acknowledging the potential limitations of the current study. The researcher welcomes constructive critiques and suggestions to further refine and broaden the coverage of the topic.

VI. Deployment

To apply the findings of this study in a practical context, the Random Forest Regressor model is recommended as the primary tool for predicting avocado sales volume in the USA. This model can be integrated into a data pipeline that continuously collects and processes relevant data, such as average price, PLU codes, avocado type, and yearly data. This pipeline will utilize real-time data sources, ensuring that the predictions remain current and accurate. Additionally, a user-friendly interface can be developed to allow stakeholders, such as marketers and retailers, to input relevant data and receive immediate sales volume forecasts. The implementation will also include regular model retraining and validation cycles to maintain accuracy over time. By leveraging the superior performance of the Random Forest Regressor, businesses can optimize inventory management, pricing strategies, and marketing campaigns, ultimately driving better decision-making and increasing profitability in the avocado market.

Acknowledgment

This research has significantly profited from the expert knowledge and insights generously provided by Dr. Monik, the Data Modelling course instructor in the Information Systems Study Program at Multimedia Nusantara University. The author is also grateful for Dr. Monik’s unwavering support and guidance throughout the entire research process.

References

[1] D. Wu, Z. Xu and J. Li, “Search Data and Geodemographics Determinants of the Avocado Sales in the US Markets,” Journal of Business & Management, vol. 29, no. 1, 2023.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis. John Wiley & Sons, 2021.

[3] T. Mate, “Avocado Prices 2020,” Kaggle, 2020. [Online]. Available: 2. [Accessed: 17-Dec-2023].

[4] Hass Avocado Board, “Hass Avocado Board: 20 Years of Making it Happen,” 2023. [Online]. Available: 2. [Accessed: 17-Dec-2023].

[5] J. Fan, F. Han, and H. Liu, “Challenges of big data analysis,” Natl. Sci. Rev., vol. 1, no. 2, pp. 293-314, 2014, doi: 10.1093/nsr/nwt032.

[6] C. Romero and S. Ventura, “Data mining in education,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 1, pp. 12-27, 2013.

[7] Bariah, S. H., Rahadian, D., & Darmawan, D. (2017). Smart Content Learning Dengan Menggunakan Metode Big Data Analysis Pada Mata Kuliah Media Pembelajaran Ilmu Komputer. Jurnal Teknologi Pendidikan dan Pembelajaran, 1.

[8] B. Mahesh, “Machine learning algorithms - a review - researchgate,” Machine Learning Algorithms -A Review. [Online]. Available: https://www.researchgate.net/profile/BattaMahesh/publication/344717762_Machine_Learning_Algorithms_A_Review/links/5f8b2365299bf1b53e2d243a/MachineLearning-Algorithms-A-Review.pdf?eid=5082902844932096

[9] J. C. Huang, K. M. Ko, M. H. Shu, and B. M. Hsu, “Application and comparison of several machine learning algorithms and their integration models in regression problems,” Neural Comput. Appl., vol. 32, pp. 5461-5469, 2020.

[10] U. Grömping, “Variable importance assessment in regression: Linear regression versus Random Forest,” *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

[11] L. E. Eberly, “Multiple linear regression,” SpringerLink, 01-Jan-1970. [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-59745-530-5_9

[12] E. Ostertagová, “Modelling using polynomial regression,” Procedia Eng., vol. 48, pp. 500-506, 2012.

[13] C. S. Bojer and J. P. Meldgaard, “Kaggle forecasting competitions: An overlooked learning opportunity,” Int. J. Forecast., vol. 37, no. 2, pp. 587-603, Apr.-Jun. 2021, doi: 10.1016/j.ijforecast.2020.08.003.

[14] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed. New York: Wiley, 2002.

[15] W.-C. Lin and C.-F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” Artif. Intell. Rev., vol. 53, pp. 1487-1509, 2020.

[16] S. M. Mefire, “Static regime imaging of certain 3D electromagnetic imperfections from a boundary perturbation formula,” J. Comput. Math., vol. 32, no. 4, pp. 412-441, Jul. 2014, doi: 10.4208/jcm.1401-m4214.

[17] P. J. Rousseeuw and M. Hubert, “Robust statistics for outlier detection,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 1, pp. 73-79, Jan./Feb. 2011, doi: 10.1002/widm.2.

[18] 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.

[19] De Schutter, B., & van den Boom, T. J. (2001, June). Model predictive control for max-min-plus-scaling systems. In Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148) (Vol. 1, pp. 319-324). IEEE.

[20] M. Schuld, R. Sweke, and J. J. Meyer, “Effect of data encoding on the expressive power of variational quantum-machine-learning models,” Phys. Rev. A, vol. 103, no. 3, Mar. 2021, Art. no. 032430, doi: 10.1103/PhysRevA.103.032430.

[21] Padilla, L., Quinan, P. S., Meyer, M., & Creem-Regehr, S. H. (2016). Evaluating the impact of binning 2d scalar fields. IEEE transactions on visualization and computer graphics, 23(1), 431-440.

[22] C. Chatfield, “Exploratory data analysis,” in European Journal of Operational Research, vol. 23, no. 1, pp. 5-13, 1986.

[23] Herho, S. H. S. (2019). Tutorial Visualisasi Data Menggunakan Seaborn.

[24] A. Pryke, S. Mostaghim, and A. Nazemi, “Heatmap visualization of population based multi objective algorithms,” in Proceedings of the 4th International Conference on Evolutionary Multi-Criterion Optimization (EMO 2007), Matsushima, Japan, Mar. 5-8, 2007, pp. 361-375.

[25] 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.

[26] F. Novkaniza, Nico, and R. A. Kafi, “Pemodelan Jumlah Kasus Baru Harian COVID-19 di Indonesia Menggunakan Gaussian Mixture Model,” in Jurnal Riset dan Aplikasi Matematika (JRAM), vol. 7, no. 2, pp. 116-127, Oct. 2023. [Online]. Available: https://doi.org/10.26740/jram.v7n2.p116-127

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

[28] 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*.

[29] 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.

PROJECT ASSIGNMENT WORK ROLE

Kelompok 3:

1. Kevin Aditya Hartono – 00000069875:
   Searching Dataset, Journal Research, Coding, Making PPT, Making Journal

2. Nathan Vilbert Kosasih – 00000069903:
   Searching Dataset, Journal Research, Coding, Making PPT, Making Journal

3. Julius Calvin Saputra – 00000068626:
   Searching Dataset, Journal Research, Coding, Making PPT, Making Journal

4. Juanito Arvin William – 00000069483:
   Searching Dataset, Journal Research, Coding, Making PPT, Making Journal

5. Christopher Abie Diaz Doviano – 00000067692:
   Searching Dataset, Journal Research, Coding, Making PPT, Making Journal