# AppleCheck: Logistic Regression for Qualitative Assessment

Elissa Mohd Azhari

Data Science Analytics Bootcamp
2024 Cohort 2
The Future Talent
PEOPLElogy Sdn Bhd, Malaysia
*elissa.azhri@gmail.com*

**Abstract.** In this capstone project, the quality of food is studied, whereby it focuses mainly on apples. A logistic regression model is built to predict the quality of apples, based on pre-existing data. The model trained achieved an accuracy of 75.5%.

## 1 Introduction

In 2012, the United Nations introduced the world to Sustainable Development Goals (SDGs), whereby it listed over 17 agendas. The second SDG focuses on Zero Hunger which promotes to create a world free of hunger and an increase of food security by 2030. The theme of *AppleCheck* aligns to this goal and intends to have a continuous improvement in the future. The main objective of the capstone project is to build a predictive model that identifies the quality of apples by learning the pre-existing data. Prediction model will help to sustain the quality of apples sold in the supermarket and to ensure food insecurity. Building an API and plug-ins to help store mass and assorted data in a cloud database. Assists user to manage mass data.

## 2 Problem Statement

**Food Insecurity.** Inconsistency of fruit quality sold in supermarkets which contributes to food insecurity.

## 3 Methodology

Following Cross-industry Standard Process for Data Mining (CRISP-DM), a process model that serves as the base for data science process (Hotz, 2023). The outline of the work process is defined in a flowchart.

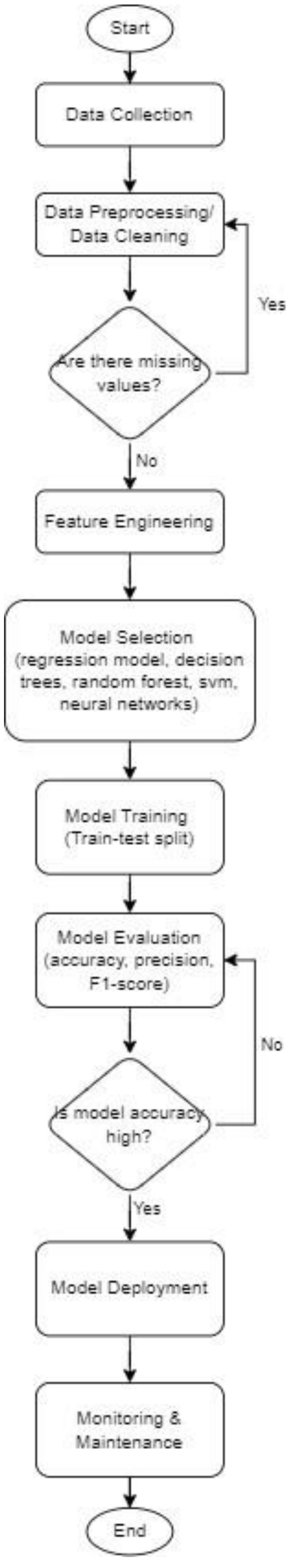

Start
Data Collection
Data Preprocessing/ Data Cleaning
Are there missing values?
Yes
No
Feature Engineering
Model Selection (regression model, decision trees, random forest, svm, neural networks)
Model Training (Train-test split)
Model Evaluation (accuracy, precision, F1-score)
Is model accuracy high?
No
Yes
Model Deployment
Monitoring & Maintenance
End

**Figure 1:** Flowchart shows the outline of the project that follows CRISP-DM

Algorithms are realized in Phyton.

### 3.1 Data Preparation

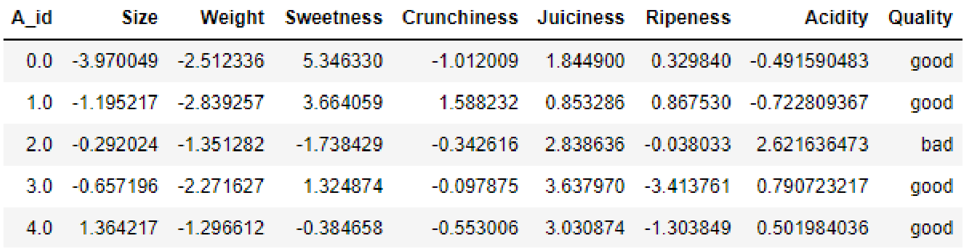
The data is extracted from *Kaggle.com*. Dataset contains a lot of attributes that help to understand the characteristics of the fruit. Dataset contains 4000 rows, representing 4000 different apples with unique identities.

Each column of the dataset is self-explanatory except for A_id,

A_id: Unique ID of each apple

The independent variables of the dataset is *Size, Weight, Sweetness, Crunchiness, Juiciness, Ripeness* and *Acidity*. While the dependent variable of the dataset is *Quality*.

**Table 1**: The first five rows of the data set from <*apple_quality.csv*>



| A_id | Size | Weight | Sweetness | Crunchiness | Juiciness | Ripeness | Acidity | Quality |
|---|---|---|---|---|---|---|---|---|
| 0.0 | -3.970049 | -2.512336 | 5.346330 | -1.012009 | 1.844900 | 0.329840 | -0.491590483 | good |
| 1.0 | -1.195217 | -2.839257 | 3.664059 | 1.588232 | 0.853286 | 0.867530 | -0.722809367 | good |
| 2.0 | -0.292024 | -1.351282 | -1.738429 | -0.342616 | 2.838636 | -0.038033 | 2.621636473 | bad |
| 3.0 | -0.657196 | -2.271627 | 1.324874 | -0.097875 | 3.637970 | -3.413761 | 0.790723217 | good |
| 4.0 | 1.364217 | -1.296612 | -0.384658 | -0.553006 | 3.030874 | -1.303849 | 0.501984036 | good |

### 3.2 Data Pre-processing

Firstly, focusing on our target column, *Quality*. Currently, the column classifies the apples into two groups: good and bad. To ease the classifications for the next steps like modelling, the two classified groups are reassigned with numbers, whereby:

- 0 indicates bad quality apples.
- 1 indicates good quality apples.

The total number of good and bad apples are sum respectively, to ensure that the dataset is not bias. From the calculation, it is found that:

- 2004 data points are classified as good quality apples.
- 1996 data points are classified as bad quality apples.

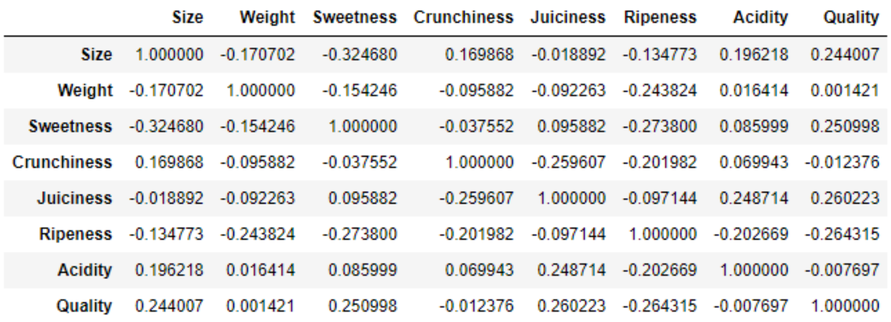
Since the classification is separated to around 50% for each group, the dataset is considered not biased. Moving on, we check whether there are null values in each row and column. All null values will be replaced by the mean value of the columns. However, in this case, there are no null values present. The unique identifier of each apple is dropped from the dataset as it serves no purpose in the investigation.

Data is pre-processed and cleaned.

### 3.3 Correlation

Once the dataset is pre-processed, a correlation check is done to further understand the relationship of each independent variable.

**Table 2**: The correlation values of the independent variables



| | Size | Weight | Sweetness | Crunchiness | Juiciness | Ripeness | Acidity | Quality |
|---|---|---|---|---|---|---|---|---|
| **Size** | 1.000000 | -0.170702 | -0.324680 | 0.169868 | -0.018892 | -0.134773 | 0.196218 | 0.244007 |
| **Weight** | -0.170702 | 1.000000 | -0.154246 | -0.095882 | -0.092263 | -0.243824 | 0.016414 | 0.001421 |
| **Sweetness** | -0.324680 | -0.154246 | 1.000000 | -0.037552 | 0.095882 | -0.273800 | 0.085999 | 0.250998 |
| **Crunchiness** | 0.169868 | -0.095882 | -0.037552 | 1.000000 | -0.259607 | -0.201982 | 0.069943 | -0.012376 |
| **Juiciness** | -0.018892 | -0.092263 | 0.095882 | -0.259607 | 1.000000 | -0.097144 | 0.248714 | 0.260223 |
| **Ripeness** | -0.134773 | -0.243824 | -0.273800 | -0.201982 | -0.097144 | 1.000000 | -0.202669 | -0.264315 |
| **Acidity** | 0.196218 | 0.016414 | 0.085999 | 0.069943 | 0.248714 | -0.202669 | 1.000000 | -0.007697 |
| **Quality** | 0.244007 | 0.001421 | 0.250998 | -0.012376 | 0.260223 | -0.264315 | -0.007697 | 1.000000 |

Using the information gathered from Table 2, seaborn correlation heatmap is generated. From the heatmap, we could conclude that each independent variable has very low correlation to each other.

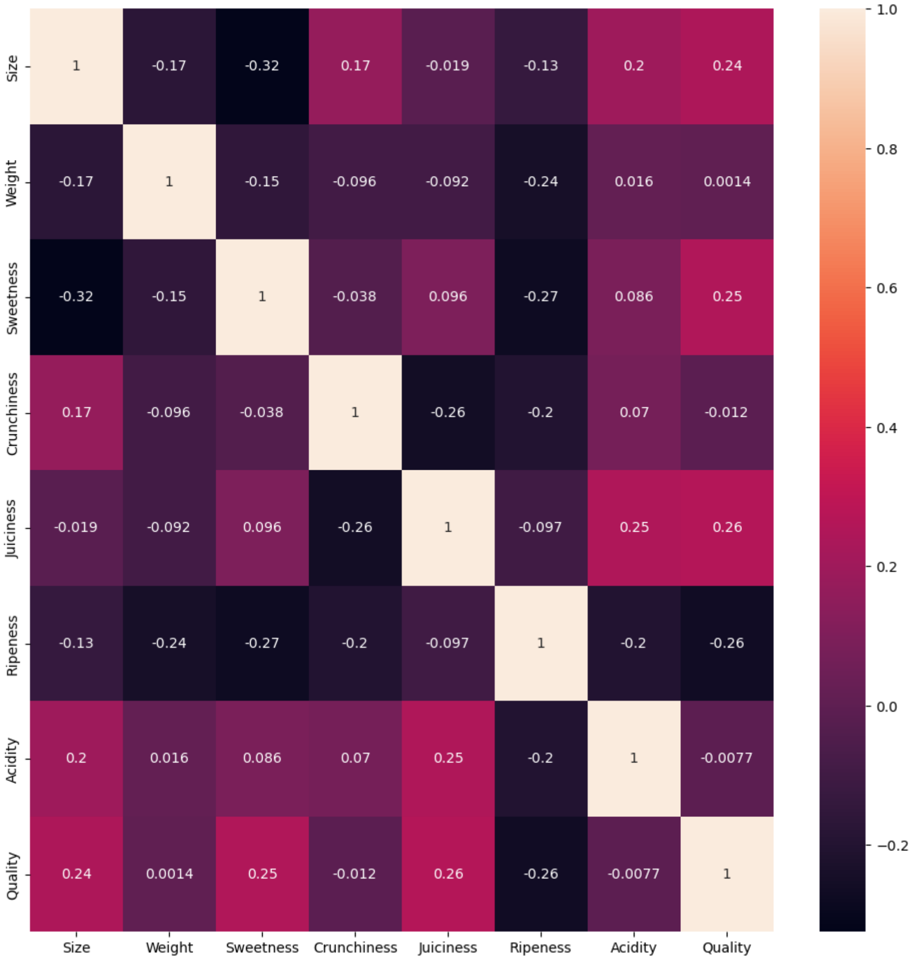


**Figure 2:** The correlation heatmap of the independent variable

### 3.4 Outliers: Boxplot

The next steps to further understand the data points, outliers for sizes and weights are observed. While outliers may distort the analysis or modelling of the data, it is decided that the outliers for both sizes and weights are not removed from the data. Keeping these outliers will enhance model robustness and help the model to handle unexpected observations in real-world situations.

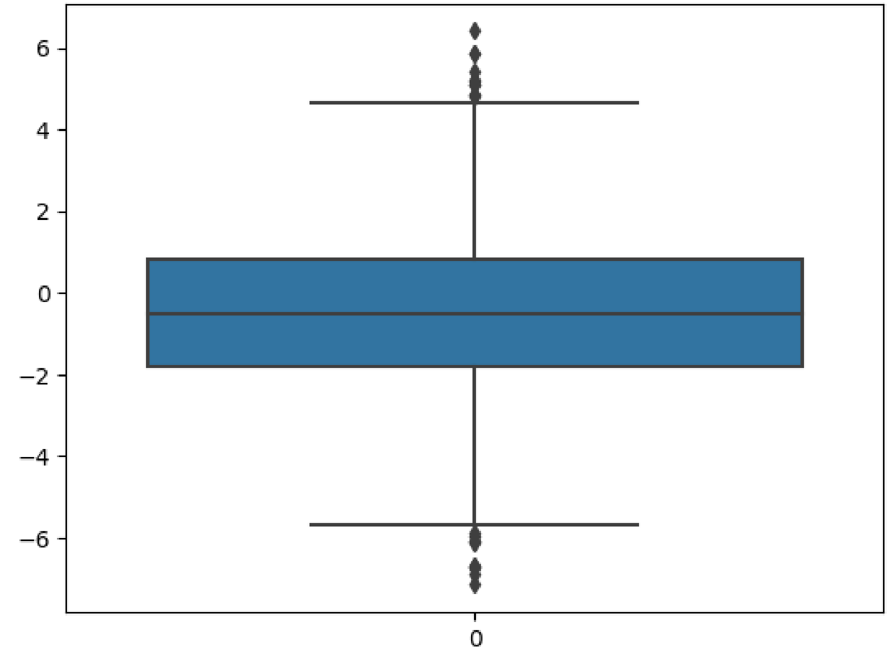


**Figure 3:** The boxplot showing the outliers for apple sizes

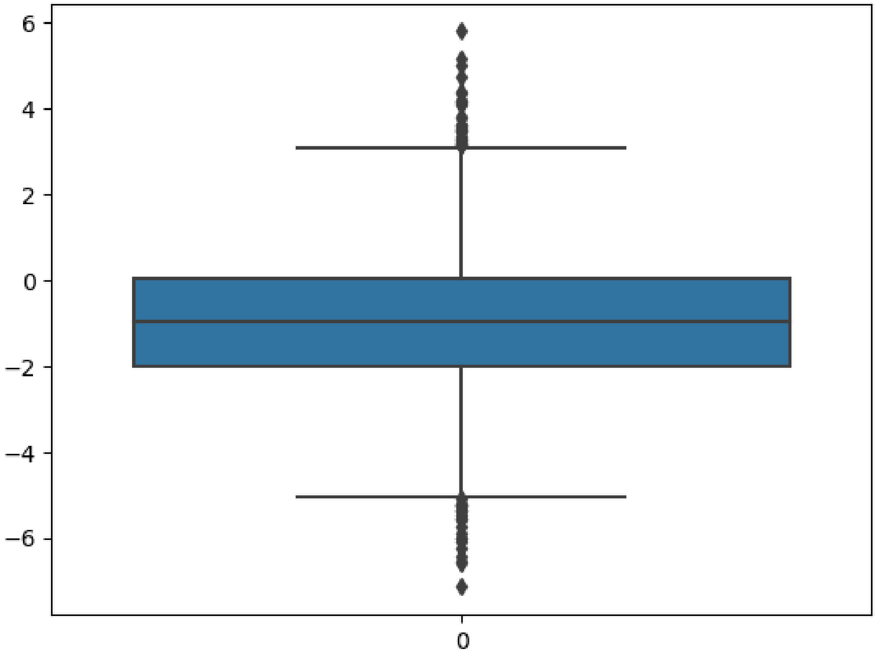


**Figure 4:** The boxplot showing the outliers for apple weights

### 3.4 EDA: Pairplot

Another powerful visualization tool, scatterplot matrix is generated from the dataset. Pairwise relationships between the variables are observed. The blue scatterplots represent bad quality apples (*Quality = 0)* and the orange scatterplots represent good quality apples (*Quality = 1).*

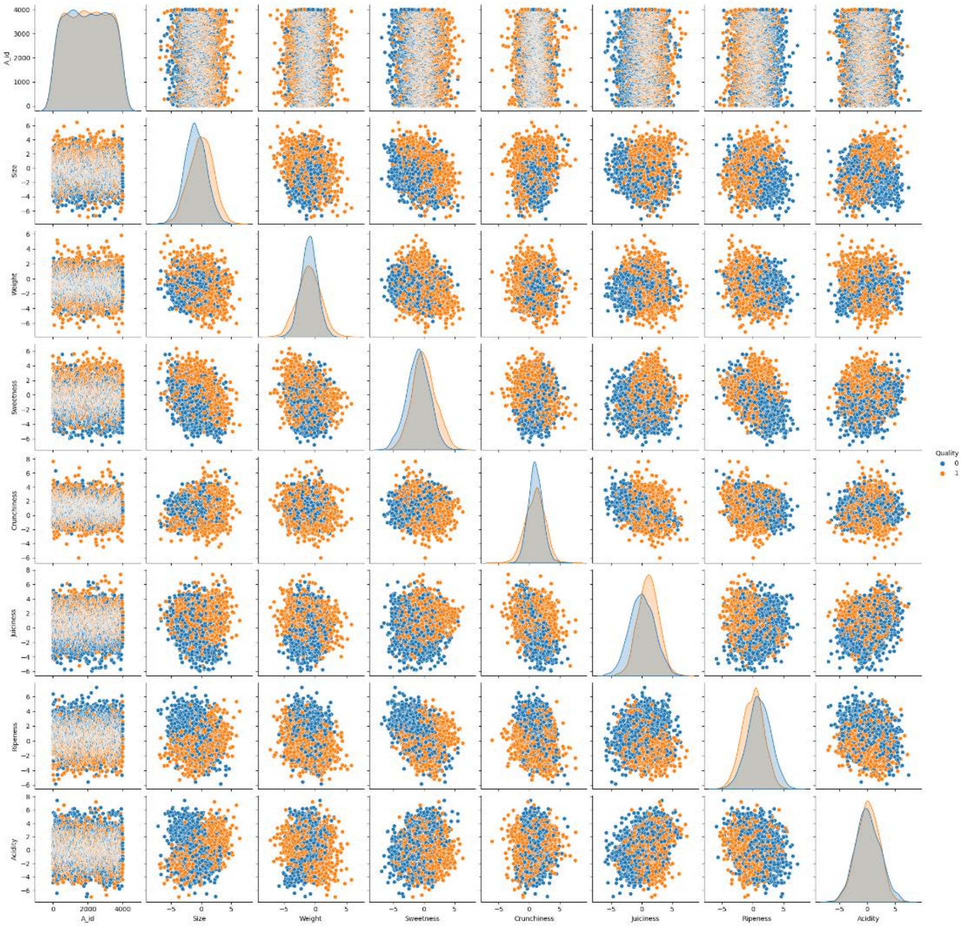


**Figure 5:** The boxplot showing the outliers for apple weights

## 4 Modelling: Binomial Logistic Regression

Algorithms are realized in Phyton.

### 4.1 Defining Parameters

Logistic regression is a supervised machine learning algorithm that is widely used for classification class. Inputs are taken as independent variables and produce a probability value of 0 and 1.

For this dataset,

- Good quality apples = 1
- Bad quality apples = 0

After the necessary libraries are imported, the independent and dependent variables are defined. Whereby, X is an independent variable and Y is dependent variable.

### 4.2 Train-Test Split

The dataset is split into two separate datasets, which are training dataset and testing dataset. 25% of the dataset is reserved for testing. A constant random state of 42 is also defined to ensure that the train-test split is not random.

The model is then instantiated and the training dataset fitted into a binomial linear regression model.

## 5 Results and Discussion

Algorithms are realized in Phyton.

### 5.1 Confusion Metrics

A prediction is made using the remaining 25% of the dataset, also known as the testing dataset. A confusion metric is generated to evaluate the performance of the classification models by comparing the predicted labels against the actual labels.

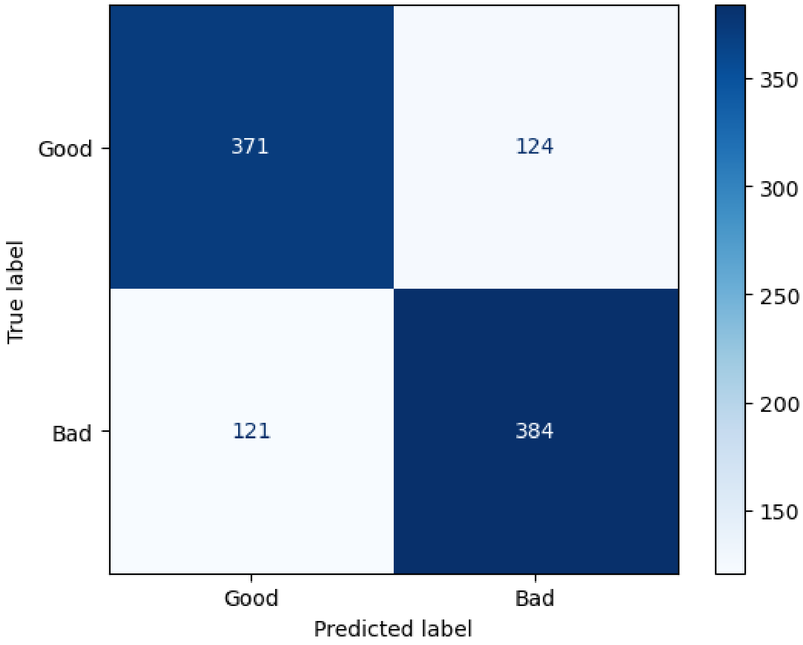


**Figure 6:** The confusion metric, comparing the predicted labels and actual labels

From the confusion metric, we can observe that 371 predicted good quality apples are indeed good quality apples. In contrast, 121 predicted good quality apples are falsely predicted, as it turns out to be bad quality apples. On the other hand, 384 predicted bad quality apples are indeed bad quality apples and 124 predicted bad quality apples are actually good quality apples.

## 5.2 Model Accuracy

The accuracy of the binomial logistic regression can be calculated from the confusion metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Whereby,

| | |
|---|---|
| *TP* : | True Positive |
| *TN* : | True Negative |
| *FP* : | False Positive |
| *FN* : | False Negative |

Using the formula above, the accuracy of the binomial logistic regression model is 75.5%. The accuracy is considered good and acceptable, without overfitting or underfitting of dataset.

## 5.3 F1-score

The f1-score of the model can also be calculated using the following formula.

$$f1 - score = 2 \times \frac{Precision \times Recall}{Presicion + Recall} \qquad (2)$$

Whereby,

*Precision*: The proportion of true positive predictions, among all instances predicted as positive

*Recall*: The proportion of true positive predictions, among all actual positive instances

Using the formula above, the f1-score of the testing dataset is 0.75 and the f1-score of the prediction is 0.76. The closer the f1-score to 1, the better the model. Hence, we can conclude the f1-score for both testing dataset and prediction is good.

## 5.4 ROC and ROC AUC

To further evaluate our binomial logistic regression, a Receiver Operating Characteristics (ROC) Curve is generated.

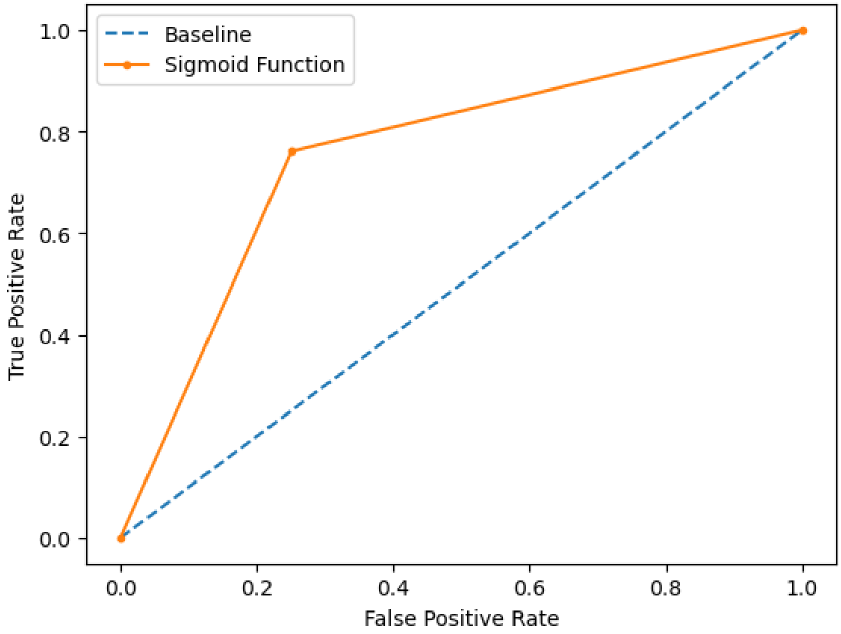


**Figure 7:** The ROC Curve of the logistic regression model

From the ROC Curve, the baseline represents the random classifier. The further the sigmoid curve to the baseline, indicates that the classifier is good. A perfect classifier would be 1. The ROC Area Under Curve (AUC) is 0.821 which concludes that the binomial logistic regression is good.

## 6 Model Deployment

The next step is to integrate the model in an existing production environment. Model deployment highlights taking an input and returning an output.

### 6.1 MLOPS Lifecycle

A structured set of stages are defined in deploying the model into an existing production environment. Since the dataset in real-time scenarios are mass and assorted, it would be costly to have an inhouse database. Therefore, our solution is to store the database in cloud. Users can access to the cloud database at any point. A real-time performance of *AppleCheck* can be observed by users during quality-checking of apples in production. Aside from that, a summary report is sent to users bi-weekly. Moving forward, *AppleCheck* will undergo maintenance, software updates and model improvements when necessary.

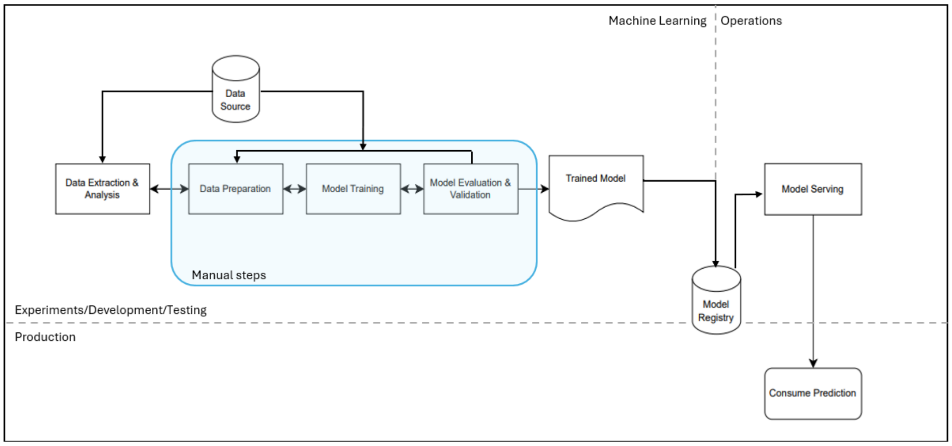


**Figure 8:** Mlops lifecyle of *AppleCheck: Logistic Regression for Quality Assessment*

## 7 Findings and Lessons Learned

From the capstone project, a lot of lessons could be taken such as realistically planning the work outline according to the timeframe given. Aside from that, the accuracy of the model could also be improved, given that more time could be reserved for model research as to deepen the knowledge and to have a better comprehension of different data analytics tool. Moving forward, it would be great if this project could be realized in real-life scenarios.

## 8 References

1. *THE 17 GOALS | Sustainable Development*. (n.d.). https://sdgs.un.org/goals

2. Courtney, R., & Courtney, R. (2021, September 28). *Hot start continues for Cosmic Crisp - Good Fruit Grower*. Good Fruit Grower -. https://www.goodfruit.com/hot-start-continues-for-cosmic-crisp/

3. Shi, X., Chai, X., Yang, C., Xia, X., & Sun, T. (2022, April). Vision-based apple quality grading with multi-view spatial network. *Computers and Electronics in Agriculture*, *195*, 106793. https://doi.org/10.1016/j.compag.2022.106793

4. *Apple Quality*. (2024, January 11). Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality

5. Bodor, A., Hnida, M., & Daoudi, N. (2023, December 12). Machine Learning Models Monitoring in MLOps Context: Metrics and Tools. *International Journal of Interactive Mobile Technologies (IJIM)*, *17*(23), 125–139. https://doi.org/10.3991/ijim.v17i23.43479

6. D. (2022, December 15). *Picking & storing apples to enjoy all year - Plantura*. Plantura. https://plantura.garden/uk/fruits/apple-tree/picking-and-storing-apples

7. Hotz, N. (2023, January 19). *What is CRISP DM? - Data Science Process Alliance*. Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/