distinguish between different types of fruits

**Machine Learning**

**Final Version 1.0**

**Tshepo Makgoba**

# Revision and Approval

**Revision History**

The table 1 below explains the changes between approved revisions of this document.

Table 1: Revision History

| Ver. | Date | Author | Comment/Summary of Updates |
| --- | --- | --- | --- |
| 0.1 | 07 March 2021 | Tshepo Makgoba | Initial Draft. |
| 0.2 | 09 March 2021 | Tshepo Makgoba | Changes based on initial review |
| 1.0 | 10 March 2021 | Tshepo Makgoba | Introduction and relevant information |
| 1.1 | - | Tshepo Makgoba | Changes based on feedback from FNB interview panel |
| 1.2 |  |  |  |

# Focused Reviews

| **No.** | **Location** | **Reviewers** | **Sections Discussed** |
| --- | --- | --- | --- |
| 1 | TEAMS | FNB Panel |  |
| 2 |  |  |  |
| 3 |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Document Distribution

|  |  |
| --- | --- |
| **Full Name** | **Business Area** |
| FNB Panel |  |
|  |  |
|  |  |

# 

# Reference Documents

| No. | Document Name | Version / Date | Attachment |
| --- | --- | --- | --- |
| 1 | Baseline model code | 10 March 2021 |  |
| 2 | Exploratory Data Analysis code | 10 March 2021 |  |
| 3 | Final model | 10 March 2021 |  |
| 4 | Dataset | 10 March 2021 |  |

# Abbreviations

The Table below details the terms, entities, acronyms and abbreviations relevant to this document.

| **Terms** | **Definition** |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Table of Contents**

[Revision and Approval 2](#_Toc66270264)

[Focused Reviews 3](#_Toc66270265)

[Document Distribution 3](#_Toc66270266)

[Reference Documents 3](#_Toc66270267)

[Abbreviations 4](#_Toc66270268)

[1. Introduction 6](#_Toc66270269)

[2 Roadmap 7](#_Toc66270270)

[2.1 Stating the question and determining the required data 7](#_Toc66270271)

[2.2 Acquire the data in an accessible format 7](#_Toc66270272)

[2.3 Identify and correct missing data points/anomalies as required 7](#_Toc66270273)

[2.4 Prepare the data for the machine learning model 9](#_Toc66270274)

[2.5 Establish a baseline model to exceed 10](#_Toc66270275)

[2.6 Train the model on the training data 10](#_Toc66270276)

[2.7 Make predictions on the test data 10](#_Toc66270277)

[2.8 Compare predictions to the known test set targets and calculate performance metrics 10](#_Toc66270278)

[2.9 If performance is not satisfactory, adjust the model, acquire more data, or try a different modeling technique 11](#_Toc66270279)

# Introduction

This is a business requirement document (BRD) for the classification of different types of fruits dated 10th of March 2021. This document is titled, DISTINGUISH BETWEEN DIFFERENT TYPES OF FRUITS, available in the [Reference Documents](#_Reference_Documents) section above.

This document serves as the requirements elicitation to the fruits classification project.

The objective of this project is to identify the machine learning algorithm that is best-suited to distinguish different types of fruits thus, we want to compare different algorithms, selecting the best-performing one.

The fruits dataset was created by Dr. Iain Murray from University of Edinburgh. He bought a few dozen oranges, lemons and apples of different varieties, and recorded their measurements in a table. And then the professors at University of Michigan formatted the fruits data slightly and it can be downloaded at the link available in the [Reference Documents](#_Reference_Documents) section above.

This is a supervised classification machine learning problem. It is supervised because there is labelled features and the target that will be used for prediction. During training the model is fed the features and target to learn how to map the data to a prediction. Moreover, this is a classification task because the target value is discrete classes (as opposed to continuous values in regression)

# Roadmap

Before any programming there must be a guide/overview/architecture to keep things on track

This roadmap pertains to the machine learning workflow once there is a problem and a model in mind. The ten steps to a successful end to end model build.

## Stating the question and determining the required data

Can we distinguish different types of fruits ? Although the question has already been checked off in the introduction section above.

## Acquire the data in an accessible format

The data is accessed through a link available in the reference section above. It is a text document that can be parsed in languages such as Python and R

Figure 1 below depicts a data frame view of the data

Table

Description automatically generated

Figure 1: Data view

## Identify and correct missing data points/anomalies as required

Look at the dimensions of the data below, there are only 59 rows and 7 columns. Looking through the data, notice that there are no missing values, which is a great. Missing data can impact an analysis as can incorrect data or outliers.

Figure 2 below depicts a view and calculation of missing values

Shape

Description automatically generated

Figure 2: Missing values

If there were missing values:

1. Missing categorical values will be imputed by a categorical base imputer
2. Missing numerical data will be imputed by the mean or median imputer based on the distribution of the data.
3. The third option will be to drop the column based on the 5% rule

Figure 3 below depicts an overview information of the dataset.

Table

Description automatically generated

Figure 3: Overview

Figure 3: Tranche 2 - Capabilities and Features Overview

## Prepare the data for the machine learning model

The figure below depicts the target variable is distributed and we can see that all the values are balanced except mandarin

Chart, pie chart

Description automatically generatedFigure 3: Tranche 3 – Balanced & imbalanced target

The figures below depict the distribution of the features and shows how mass and width are correlatedChart

Description automatically generated

A picture containing chart

Description automatically generated

## Establish a baseline model to exceed

A picture containing text, receipt, screenshot

Description automatically generated

## Train the model on the training data

This step is available in the code document attached in the reference section above

## Make predictions on the test data

This step is available in the code document attached in the reference section above

## Compare predictions to the known test set targets and calculate performance metrics

This step is available in the code document attached in the reference section above

## If performance is not satisfactory, adjust the model, acquire more data, or try a different modeling technique

This step is available in the code document attached in the reference section above