Problem Statement:

In a packing factory, they want to pack quality stock for wholesalers and retailers, however, they have large quantities and want an efficient method to classify the apples. Now given three kinds of apples namely Golden Delicious (GD), Granny Smith (GS), and Royal Gala (RG). Our main purpose is to use infrared spectrum data to classify between bruised (B) sound (S) samples using Machine Learning (ML). We used Logistic Regression (LR) model to determine the baseline and models such as Random Forest to precisely predict the condition of the sample.

Aim:

Aim:

The aim of the project is to build a model that can precisely predict if an apple is bruised or sound using machine learning.

Objective 1: Change the best features used for Selection Feature

Change the best features to evaluate improvement on the model's precision score.

Objective 2: Change methods used for feature engineering

Consecutive columns contain very similar wavelength data meaning it is to some degree redundant, to reduce the amount of data one could get the average for a certain window of data which will produce a smaller equivalent dataframe to use for the model. The objective is to use different methods to group the data such as variance, skewness and the sum of the mean and variance to find that method that will produce the best precision for a particular data set. Finally we will produce a plot of the precision scores of the different methods against the baseline.

Objective 3: Evaluating different models

Apply different models to the data namely Decision Tree (DC), XGBoost model (XGB) and Support vector machine (SVM) for classification and compare these new models to the baseline and originally used Random forest model (RF).

DATA PRESENTATION:

Given 3 datasets of infrared spectrum for different kinds of apples with sizes about 500 rows x 2000 columns, where the rows contain the samples and the columns being the wavenumber of the spectrum.

Data Pre-processing:

Feature Engineering

Creating new features that extract important information needed for classification. This serves the purpose of cutting down redundant information. To do this we will calculate the mean, variance and skewness of groups of a number of consecutive columns in a dataframe hence then creating a new smaller dataframe. (Explain the graphs when changing the methods)

Feature Selection

Feature selection is the method of reducing the input variable to your model by using only relevant data and getting rid of redundant data. This finds the best number of features for a given N from the dataset. Selecting N=10, gives us the best 10 features which are relevant for our Machine learning model. We will use Sequential Feature Selection (SFS) , an optimization tool from sklearn, which removes features based on a cross-validation score. To obtain this score,we specify a Machine Learning Model which will run for the different features. To explore the impact of feature selection we then change this number of features from 10 to 5,20,40.

RESULTS:

Changing feature number:

Values are randomly spread.
Feature number =10 is above line, but random.
Can see from scale that this is not significantly above the other precision scores.
Conclude it is not worthwhile optimising the number of features.

For the rest, we stick with number of features=10

<u>Changing the methods:</u>

GS data:

For the GS data the mean method aggregator had a precision score above the baseline in the 50 roll bin and a relatively close precision score to the other methods in the other bins except variance which had the lowest precision score across all the roll bins.

GD data:

Mean-variance sum had the highest precision score in the 30 and 50 roll bins with close scores to the other aggregators for the rest of the bins.

RG data:

For the RG data the skewness aggregator performed best for the 30, 50 and 100 bin roll.

Based on the above most of the methods were still randomly placed and you can't conclude on who performed the best. The variance aggregator constantly performed the worst for all the bins.

Machine Learning models
We tried a few machine learning techniques besides Random Forest, Decision Tree, XGBoost and Support Vector Machine.

A brief summary is shown on the slide.

By keeping the number of features to a fixed value of 10, we have applied the different techniques to the 3 data sets. We can see that XBoost has been consistent in the 50 roll bin where it outperforms the other techniques. Overall we cannot conclude on which is the best method. Only XGB and RF went over the baseline in the 50 roll bin for GS data.
CONCLUSIONS:

Sometimes we believe that machine learning can solve a lot of problems but ultimately, sticking to the basic method works best since we could not fine a technique that outperforms the baseline.

In conclusion we can see that no best aggregator could be chosen since none of them really revealed scores that out performed the others in all roll bins but the variance aggregator showed the worst performance for all bins consistently.

TEAMWORK:
Teamwork makes dream work. We helped one