# Apple Classification

using Machine Learning

# Team Presentation :BITDN 3

Justine Crook-Mansour

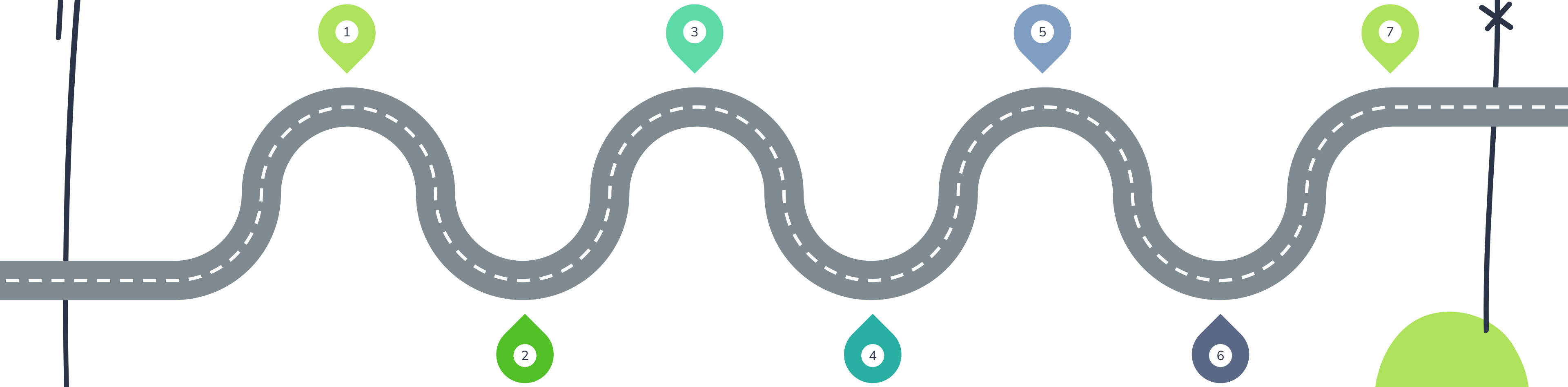
Nhlakanipho Kunene

Ansil Nefdt

Adivhaho Sithari

Venu Prayag

# Roadmap



1. Background, Aim & Data
2. Number of Features
3. Feature Engineering Methods
4. Machine Learning Comparison
5. Results
6. Challenges/ Improvements
7. Teamwork & Conclusion

# BACKGROUND

- Workers in apple factories need to pack quality stock for wholesalers and retailers.
- Some apples are sound (S) and some are bruised (B).
- They need an efficient method to classify the apples, and throw out the bruised ones.

- In this project, we were given infrared spectrum data for apples.
- We used this to predict whether the apples are B or S, using Machine Learning (ML).
- A Logistic Regression (LR) model was used to determine the baseline, and Random Forest (RF) was used to predict the condition of the sample.

# AIM

- Our aim is to experiment with the various parameters in the ML process to examine its effect on the precision score.

# DATA

- ✗ 3 datasets showing data for different types of apples:
  - Golden Delicious (GD)
  - Granny Smith (GS)
  - Royal Gala (RG)
- ✗ Sizes: about  500 rows x 2000 columns
- ✗ Data
  - Each row is a sample.
  - The columns are the condition, age, source, and absorbance at various wavenumbers.

# Precision scores

- Precision score was used as a measure of the efficiency of the ML process.

1

# Number of Features

Varying the number of best features.

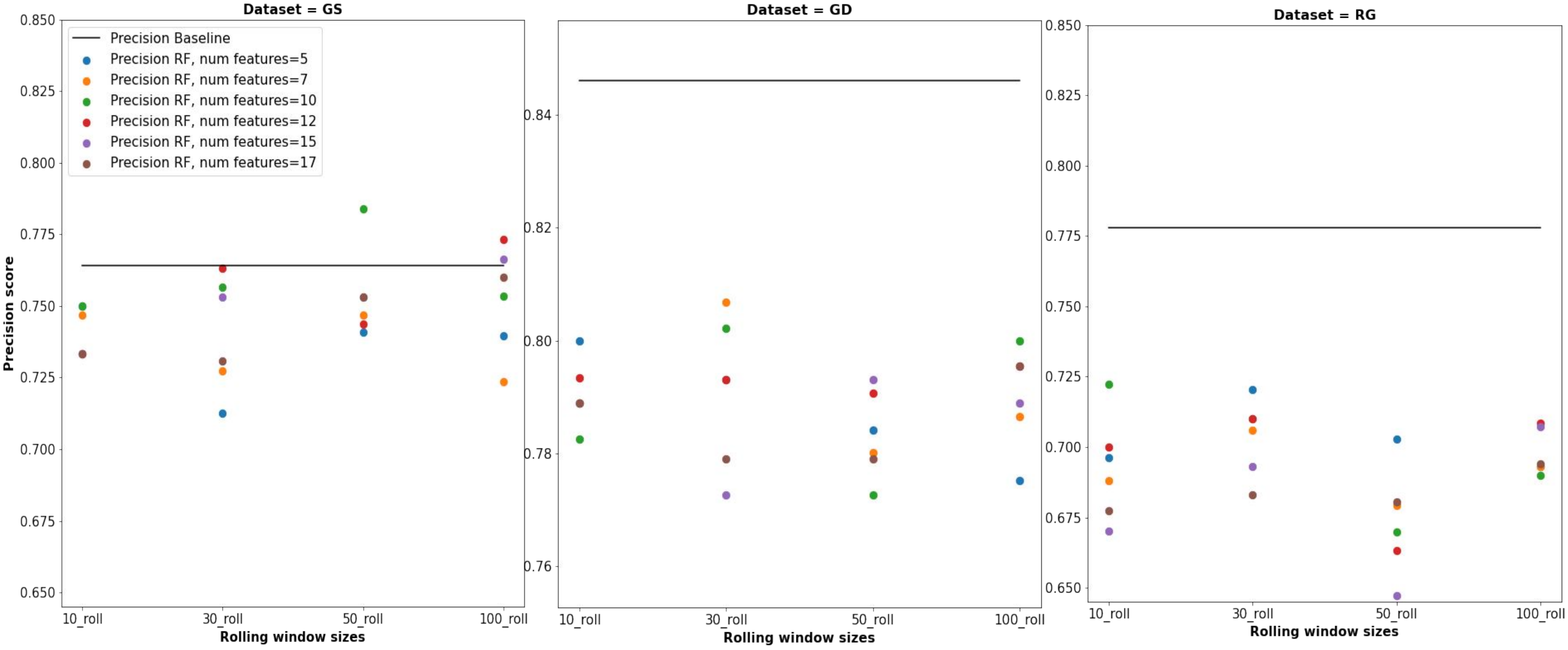
# Number of Features

- ✗ 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 ML model which will run for the different features.

- ✗ To explore the impact of feature selection we changed the number of features to 5, 7, 10, 12, 15, 17.

# Number of Feature Results



# RESULTS

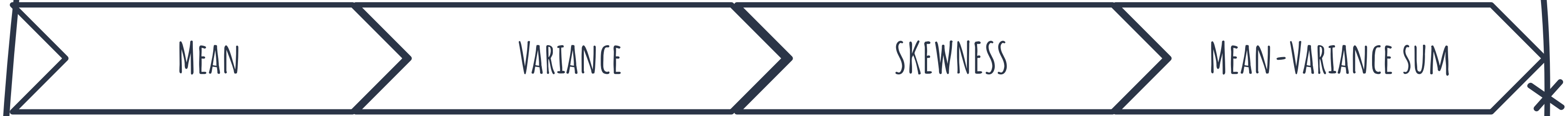
- ✗ Results are randomly spread.
- ✗ The results when feature dataset= GS, number =10, window size= 50 is above baseline, but this is likely random as the higher result is not reflected in other datasets.
- ✗ Also, we can see from the graph scale that this outlier is not significantly above the other precision scores.
- ✗ We conclude it is not worthwhile optimising the number of features, as the effect is negligible.
- ✗ For the rest of this project, we set number of features=10

2

# FEATURE ENGINEERING METHOD

Changing the method used.

# FEATURE ENGINEERING METHODS



## MEAN

Sum of all the data points divided by the total amount of data points.

## VARIANCE

Variance is the square of the mean deviation of the data from the mean value.

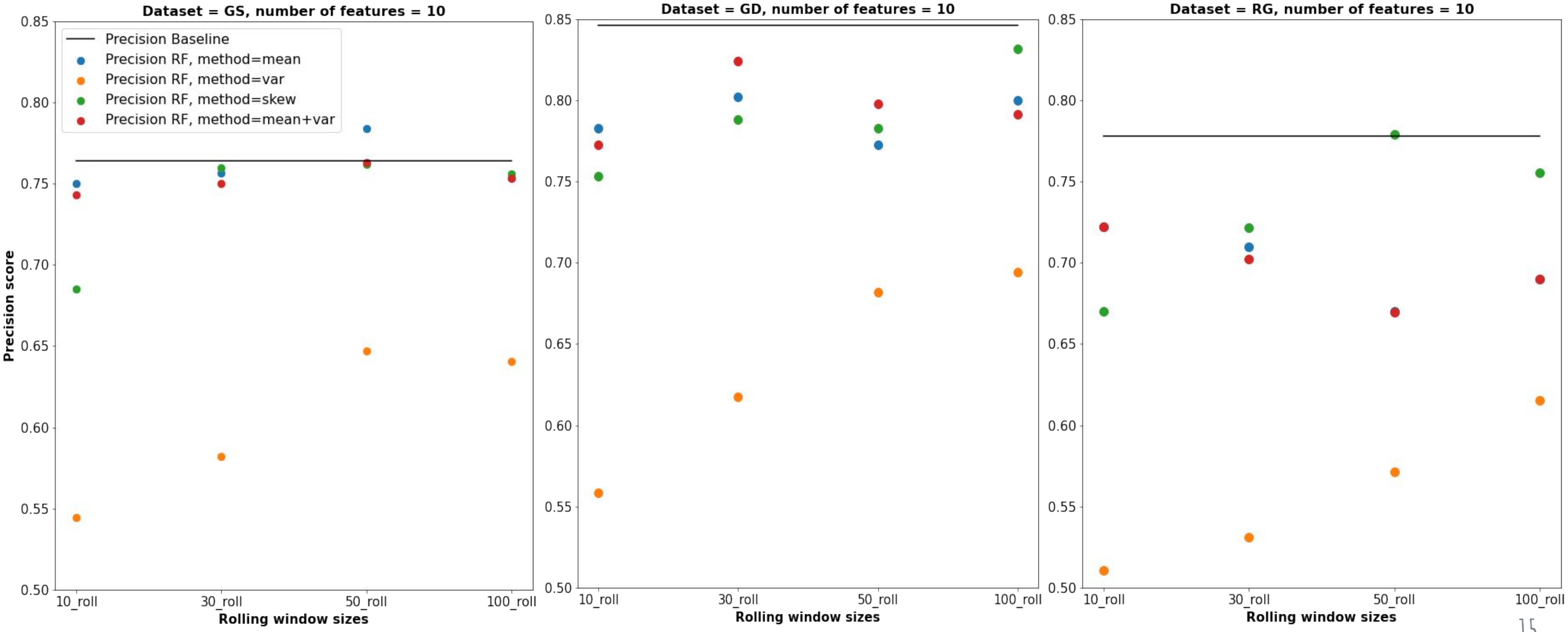
## SKEWNESS

Skewness is a measure of the asymmetry of the distribution.

## MEAN-VARIANCE SUM

The sum of the mean and variance of the roll window.

# Feature Engineering Results



# Results

- ✗ Method = variance performs consistently lower, independent of the dataset.
- ✗ The other method results are randomly spread.
- ✗ For the rest of this project, we use method = mean.

3

# Machine learning comparison

# Machine learning Techniques

## Random Forest (RF)

A way of making a decision based on votes from many decision trees which are created by using a subset of the attributes.

## Decision tree (DT)

A type of flow chart. Basically we have a choice. From each choice we have "branches" that lead to the outcomes of those choices which may, themselves be (or lead to) more choices which have more branches etc.

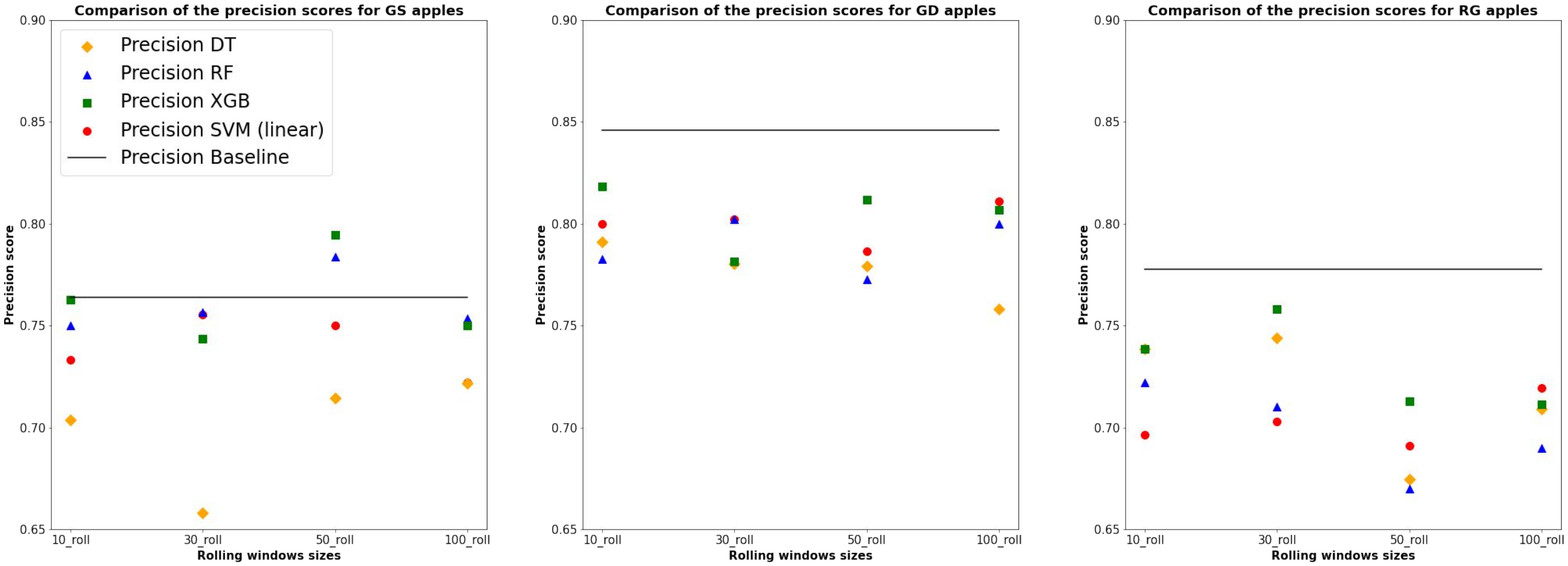
## XGBoost (XGM)

A gradient boosting algorithm that uses decision trees as its “weak” predictors. Beyond that, its implementation was specifically engineered for optimal performance and speed.

## Support vector machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

# Machine learning technique Results



Number of features = 10

# Conclusion

- ✗ From our results, we see that changing the number of features and aggregation method had a negligible effect on the precision scores.
- ✗ No machine learning technique was the clear winner.
- ✗ Machine learning can not solve every problem.

# Problem/improvements

- ✗ We tried implementing cross validation techniques to include errors on our plots.
- ✗ Try more Machine Learning techniques.
- ✗ Use a different splits between training data and testing data.
- ✗ Use other measures for efficiency of the algorithm like accuracy.
- ✗ More data!

# Teamwork

- Teamwork makes the dreamwork.
- Different overlapping tasks were allocated and help was given wherever required.
- We made use of Google workspace to share our results and work.
- Different level of skills in our team, which complemented each other.
- No-one was a bad apple :).

# Thanks!

Any questions?