Assignment
Machine Learning for Metabolomics
2025-2026

Module Manager: Maria Anastasiadi

Introduction

Apples (Malus × domestica Borkh.) hold a prominent position as the most widely produced and consumed fruit worldwide with the United Kingdom's (UK) apple industry alone valued at approximately £977 million, with a 6.6% growth rate as of 2024. Apples contain a diverse array of bioactive compounds, including vitamins and antioxidants, in addition to essential nutrients such as sugar, fibre, and minerals.

Apple quality is determined by factors such as appearance, firmness, flavour, as well as the absence of physiological and pathological disorders.

Controlled Atmosphere (CA), employing a low O_2, and elevated CO_2 concentration, is a popular post-harvest treatment used in cold storage that decreases ethylene biosynthesis and respiration rate, which are critical biochemical processes accelerating fruit senescence.

A more recent advancement in this field is Dynamic Controlled Atmosphere (DCA), which offers enhanced control by adjusting O_2 levels based on the fruit's metabolic changes. DCA ensures the lowest possible O_2 levels before anaerobic respiration.

In recent decades the UK's domestic apple production has significantly declined and the UK's reliance on imported apples has surged. Therefore, it is crucial for the UK to invest in research and resources to extend the shelf life of locally cultivated apples in alignment with the UK government's objective of self-sufficiency.

This project aimed at assessing the physicochemical changes and quality traits of ‘Royal Gala’ apples stored for a total of 29 weeks under DCA and cold storage over two successive years. In addition, in Year2 the DCA was compared against Controlled Atmosphere (CA), which replicates current industrial storage practices (1 % O_2 and 5 % CO_2).

The apples were received from Kent, England and were divided into six mini pods for the DCA treatment and into three 100 L boxes for CA and placed at 1 °C in the dark. Sampling occurred at timepoints 0, 5, 11, 17, 23 and 29 weeks of storage for both Year 1 and Year2.
At each sampling point the following physicochemical parameters were assessed: **phenolic compounds** in the apple skin (procyanidin B1, catechin, procyanidin B2, epicatechin, chlorogenic acid, ideain, quercetin hyperoside, rutin, quercetin glucopyranoside, quercetin xyloside, quercetin arabinoside, quercetin rhamnoside, and phloridzin, **organic acids** in the skin (SK) and flesh (FL) (glutamic-, quinic-, malic-, ascorbic-, shikimic, and citric-acid) , individual **sugars** in the flesh (fructose, sorbitol, glucose, sucrose), **phytohormones** (ABA, DPA, PA, 7OH-ABA, ABA-GE), **firmness** (N), **total soluble solids** (TSS) (ºBrix).
In this assignment you are provided with the physicochemical parameters measured in Year1 and Year2 in UK "Royal Gala" apples with the aim to build classification models to predict the storage stage of 'Royal Gala' apples - as an indicator of ripening stage- using the available data and identify key differentiation biomarkers.

Dataset

You have been supplied with three separate CSV files: **DCA_Year1.csv**, **DCA_Year2.csv, CA_Year2.csv**. The first three columns in each file correspond to the Sample ID, Storage Week and the Ripening Stage. The storage week denotes the storage week each sample was tested (0, 5, 11, 17, 23, 29), while the storage stage denotes the ripeness stage as "early", "mid", "late". The rest of the columns contain the predictor variables, i.e. the physicochemical parameters measured at each sampling point as described above. The metabolite names have been shortened for ease, and they appear in the same order as above.
Note1: Each parameter and group of metabolites has been measured in different units depending on their concentration in the samples (e.g. sugars, organic acids and phenolic compounds are expressed in mg/g dry weight, while phytohormones are expressed in ng/g dry weight).
Note2: For Year2, week 0 has 18 replicates, while for Year1, week 0 has 6 replicates. All other timepoints have 12 replicates each.

Assignment Objectives:

Objective 1: Use the apple physicochemical data from the DCA trial as predictors to build, optimise and evaluate ***different classification models*** for the prediction of storage time and ripening stage. The algorithms you are required to use to build the predictive models are: **1) Random Forest and 2) Random Forest with Multivariate Longitudinal Predictors.** For each algorithm you are required to build 2 models for different data partitions and to compare the total performance of the two algorithms for the models developed. Provide justification for the method of your choice.

Objective 2: Use the apple physicochemical data from **Year 2** corresponding to DCA and CA treatments to build ***multivariate timeseries classifiers*** for each treatment. Use the test set to evaluate the prediction accuracy of the models. The algorithms you are required to use to build the predictive models are: **1) Random Forest with feature extraction.**
2) An algorithm of your choice suitable for time-series. For each algorithm you are required to build 10 models for different data partitions and to compare the total performance of the two algorithms for the models developed. Provide justification for the method of your choice.

Deliverables:

Analysis Report outlining and discussing your results: [100 marks] (up to 3000 words).

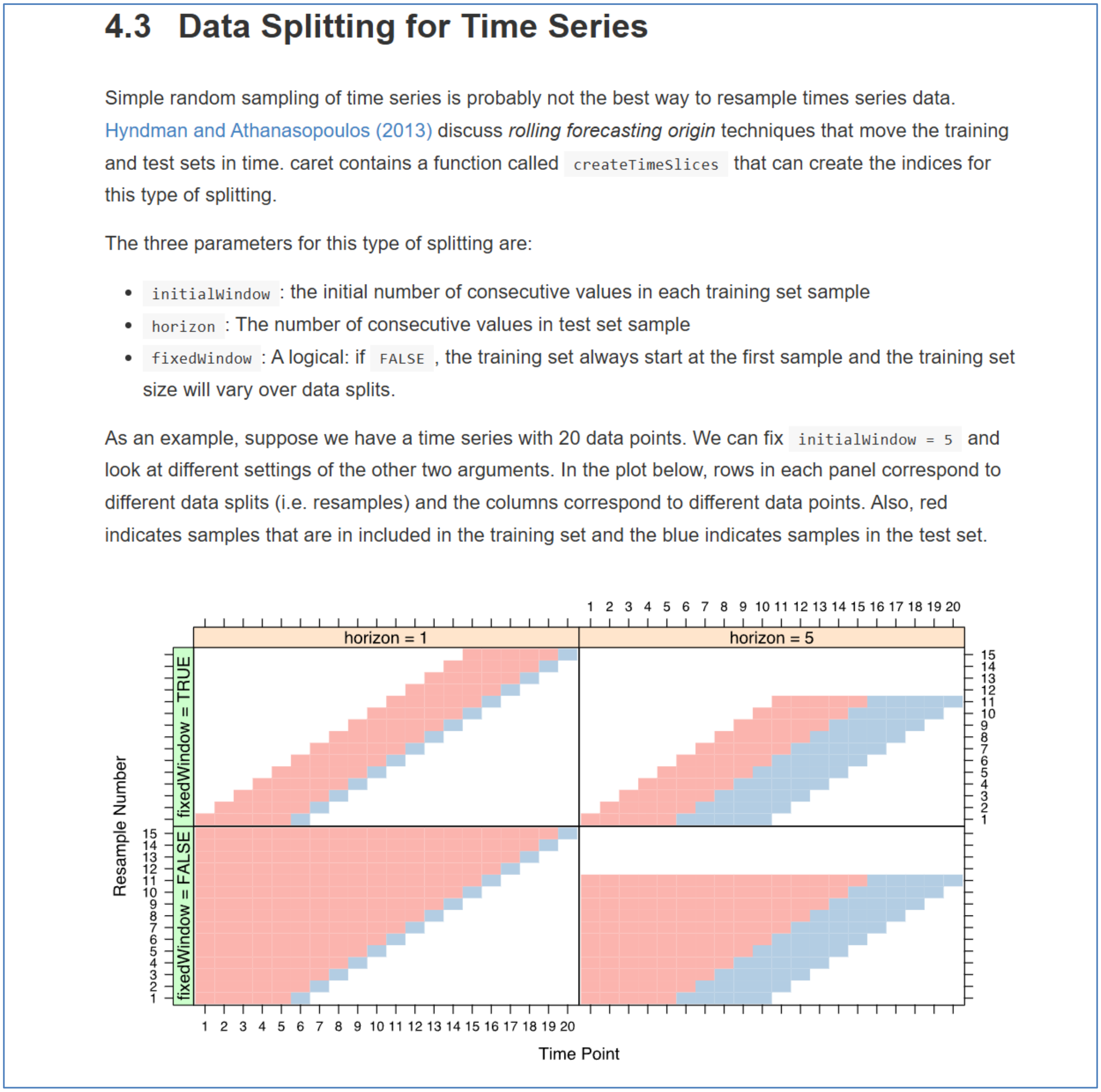
Your analysis report needs to include the following steps:

Data Preparation and Exploratory analysis [15 marks]: Perform QC and data pre-treatment to normalise the data, deal with any outliers and address any batch effect issues. Year1 and Year2 for the DCA treatment can be considered different batches and you may need to apply batch correction prior to the analysis. Use PCA and other suitable visualisation techniques to explore the data and decide if the pre-treatments have any effect. Comment on the clustering of the data in relation to different treatments, batches and storage duration.

Objective 1 [35 marks]: For this task you are required to use the Year1 and Year2 DCA datasets to create:

A) **Multiclass classification models** to predict which *storage week* or *ripening stage* an apple sample belongs to according to their physicochemical profile. You will use the data from Year1 to train and optimise the classification model and the data from Year2 to validate its performance. The first algorithm to use is **Random Forest.** You can use the `createTimeSlices()` function from the caret library (https://topepo.github.io/caret/data-splitting.html) to partition the dataset according to the timepoint they belong to (see screenshot in Figure 1). Then proceed to the training-optimisation-evaluation phase. Apply scaling if required, and grid search for model tuning. Clearly describe the steps you follow for each process. When the model training and optimisation is completed use the test dataset from Year 2 to evaluate it performance. Create confusion matrices and report any other relevant performance metrics. Finally, summarise the most important features (variables) identified by the model and try to provide a biological interpretation of the results. Finally, repeat the process above, but this time use Year2 to train and optimise the model and Year1 as test set.

B) **Random Forest with Multivariate Longitudinal Predictors.** This is a variation of the Random Forest algorithm which allows you to treat the predictor variables (physicochemical parameters) as longitudinal predictors, meaning the model doesn't just look at a snapshot of the profile, but understands the trajectory of changes over the 29-week storage period. The R package **DynForest** (https://cran.r-project.org/web/packages/DynForest/vignettes/overview.html) is specifically designed to address timeseries. Follow the tutorials for this library to perform data partition, training and testing. As previously you are required to use data from Year1 for training and optimising the model and Year2 for testing. The class you can try to predict can be “Storage week” and “Ripeness Stage” and comment on which provides better results. Finally, repeat the process above, but this time use Year2 to train and optimise the model and Year1 as test set.



4.3 Data Splitting for Time Series

Simple random sampling of time series is probably not the best way to resample times series data. Hyndman and Athanasopoulos (2013) discuss *rolling forecasting origin* techniques that move the training and test sets in time. caret contains a function called `createTimeSlices` that can create the indices for this type of splitting.

The three parameters for this type of splitting are:

- `initialWindow` : the initial number of consecutive values in each training set sample
- `horizon` : The number of consecutive values in test set sample
- `fixedWindow` : A logical: if `FALSE` , the training set always start at the first sample and the training set size will vary over data splits.

As an example, suppose we have a time series with 20 data points. We can fix `initialWindow = 5` and look at different settings of the other two arguments. In the plot below, rows in each panel correspond to different data splits (i.e. resamples) and the columns correspond to different data points. Also, red indicates samples that are in included in the training set and the blue indicates samples in the test set.

Figure 1. Example of how to partition data for time series using caret. Snapshot taken from *https://topepo.github.io/caret/data-splitting.html*

Objective 2: [35 marks]: In this objective you are required to create ***multivariate timeseries classifiers*** to be able to differentiate between samples belonging to DCA vs CA treatment in Year2.

A) The first algorithm you are required to employ is RandomForest. Since you have a timeseries, it is recommended to use the **RCatch22 package** (https://cran.r-project.org/web/packages/Rcatch22/index.html) to transform the features (variables) into 22 interpretable feature vectors. Catch22 (CAnonical Time-series Characteristics) is a feature-based transformer that reduces complex time-series data into 22 highly informative, interpretable, and computationally efficient features. It extracts diverse properties—including linear/non-linear autocorrelation, fluctuation scaling, and distribution—to enhance machine learning tasks like classification and clustering. Following feature transformation, follow the typical workflow to train, optimise and validate the classification model. Present the results using a confusion matrix and other performance metrics and comment on the model's performance. Extract importance scores for the original features and attempt biologically meaningful interpretations.

B) In addition to Random Forest you are also required to employ an algorithm of your choice suitable for time-series. For each algorithm you are required to build 10 models for different data partitions and to compare the total performance of the two algorithms for the 20 models developed. Provide justification for the method of your choice. Create confusion matrices and report any other relevant performance metrics. Finally, summarise the most important features (variables), if applicable, and try to provide a biological interpretation of the results.

Assignment report and R code [15]

Produce a comprehensive report (max 3000 words) providing a short background, the aim and objectives of the assignment, the methodology you have applied and the results you acquired from the analysis. Add discussion trying to interpret the results you acquired and relate them to previous findings from the literature. Finally, add a section with the conclusions from your analysis.

In the methodology section specify the R packages you have employed for each analysis type. The results of your analysis need to be clearly documented and accompanied by tables and figures where applicable. Only include the most important results in the main body of the report. Any supplementary figures/tables can be included as an appendix at the end of the report after the references section. Make sure to discuss your results and properly reference any external sources.

R code

The scripts should include the code used for producing the results shown in the report. Make sure it is clearly commented and reproducible. If you decide to create custom made functions include them at the start of the script, or in a separate R file. Feel free to employ RMarkdown for both the analysis and report but make sure when exporting the report as HTML the code is not included.

Submission:

The analysis report and the analysis scripts are to be submitted via the normal route on Canvas. The assignment report must be written in either: MS Office Word (Windows/MAC) or LibreOffice (Ubuntu Linux). Export you report as a pdf and include it in the zip file.

The assignment report, scripts/functions and any accompanying data, image and results (text) files must be archived into one compressed file (e.g. ZIP or TAR) and include your name, student number and course in the filename, e.g. **MAnastasiadi_S30245_BIX.zip**

Deadline:
Full-time students
The archived files must be uploaded on Canvas by: Friday 20th 2026, 12pm

Part-time students
The archived files must be uploaded on Canvas by: March 6th 2026, 12pm