# BCUR Poster Section Drafts

## Modelling Techniques

The results of the data to be analysed is received from other teams tracing the whistles in a spectrogram and providing the traced data to tools such as ROCCA and ARTwarp. ROCCA is a MATLAB-based tool designed for real time species identification of delphinid whistles during shipboard surveys, and ARTwarp is an automated method for categorizing bioacoustic signals, particularly focusing on dolphin whistles and killer whale calls. Both these tools are used to categorise the whistles into different groups in a variable called Category with an output of NUMBER variables for each whistle analysed.

The issue with all the variables output by ROCCA and ARTwarp from each whistle, is that many of these are collinear, meaning that they provide no tangible benefit to be included in the model, instead complicating the process. Therefore, all the collinear variables were removed reducing the number of variables to 6, being ‘DURATION’, ‘FREQABSSLOPEMEAN’, ‘FREQBEG’, ‘FREQMAX’, ‘FREQQUARTER1’ and ‘FREQPOSSLOPEMEAN’.

To determine the number of data points (whistles) to use for the modelling, a completeness graph was used. From this graph, we can see the discovery curve begins flattening around the mark of 1000 whistles. This means that using more whistles will have negligible additions to the number of categories, therefore the number of 1000 whistles was categories.

* Data is received from teams tracing whistles in a spectrogram and analysed using ROCCA and ARTwarp.
* ROCCA identifies delphinid whistles in real-time, while ARTwarp categorizes dolphin and killer whale calls.
* These tools classify whistles into categories, outputting multiple variables per whistle.
* Many output variables are collinear, so only six key variables were retained: DURATION, FREQABSSLOPEMEAN, FREQBEG, FREQMAX, FREQQUARTER1, and FREQPOSSLOPEMEAN.
* A completeness graph was used to determine the ideal number of whistles for modelling.
* The discovery curve flattens at 1000 whistles, meaning additional data offers minimal benefit.

A graph of a number of whistles

AI-generated content may be incorrect.

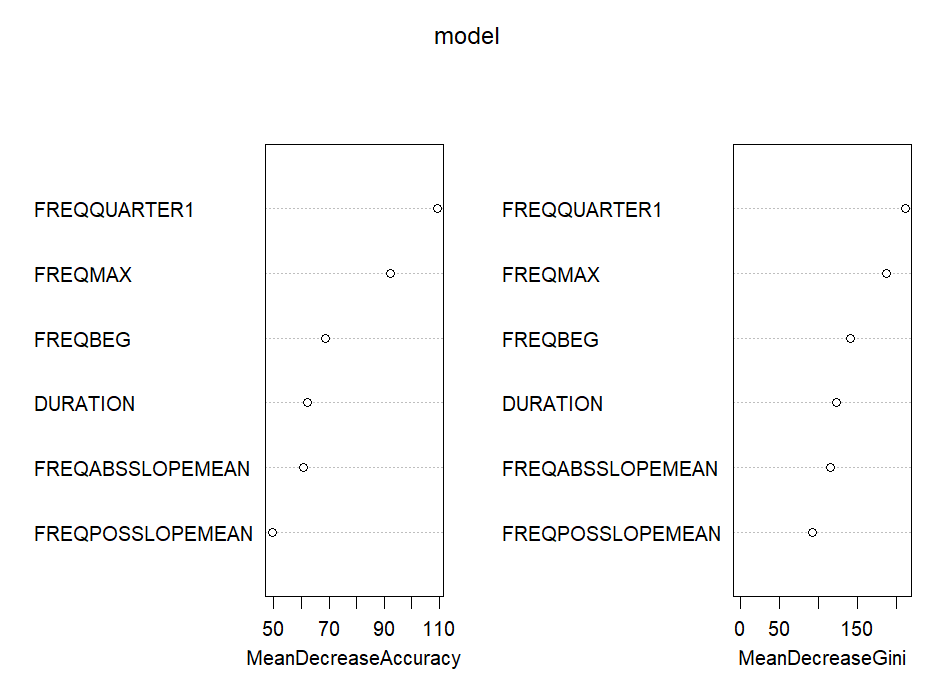
## What is Multinomial GLM?

A Multinomial Generalized Linear Model (GLM) is a type of statistical model used when the response variable is categorical with more than two possible outcomes (i.e., a multinomial outcome). It is an extension of logistic regression that can handle multiple categories rather than just binary classification. In this instance, the categories would be considered as multiple possible outcomes, derived from the machine learning grouping techniques used to provide the data to analyse.

* A Multinomial Generalized Linear Model (GLM) is used for categorical response variables with more than two outcomes.
* It extends logistic regression to handle multiple categories instead of just binary classification.
* In this case, categories represent outcomes from machine learning grouping techniques used for data analysis.

## Random Forest Results

* Using a set seed of ‘123’ the accuracy of the Random Forest model fitted with this data was found to be 74.32 %
* When another set seed of ‘456’ was used, the accuracy of the Random Forest model fitted with this data was found to be 63.39 %
* After averaging the results of 100 trials with random seeds, the accuracy of this model was found to be 67.68%



## Sources

Deecke VB, Janik VM. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. The Journal of the Acoustical Society of America. 2006 Jan 1;119(1):645-53.

Oswald JN, Rankin S, Barlow J, Lammers MO. A tool for real-time acoustic species identification of delphinid whistles. The Journal of the Acoustical Society of America. 2007 Jul 1;122(1):587-95.