Title: CSIRO - Image2Biomass Prediction
Team Members: Anneli Oro, Robert Koor

# 1. Business understanding

## 1.1. Identifying your business goals

### 1.1.1. Background

Pasture biomass estimation is crucial for grazing management. Farmers must know whether there is enough feed available for livestock and when paddocks need rest. Traditional biomass measurement techniques such as “clip and weigh” are accurate but slow and labour-intensive, while automated tools like plate meters often fail under variable conditions. The competition offers annotated pasture images, NDVI values, and ground-truth biomass measurements to build a machine learning model that predicts biomass from visual data.

### 1.1.2. Business goals

Although this is not a commercial business project, industry stakeholders such as farmers, agricultural researchers and agribusiness agencies do directly benefit from the results. Our project aims to:

- develop a predictive model that estimates pasture biomass from images;
- improve accuracy compared to manual or traditional measurement tools;
- contribute to scalable, automated solutions for grazing management and sustainable agriculture.

### 1.1.3. Business success criteria

The project will be considered successful if we manage to train and create a model that is capable of accurately (with a weighted coefficient of determination value $R^2 > 0.5$) predicting pasture biomass.

## 1.2. Assessing your situation

### 1.2.1. Inventory of resources

Data used in this project originates from a Kaggle competition and consists of

- high-resolution pasture images (357 images, 2000x1000px each);
- ground-truth measurements of biomass

Tools that are used in this project include, but are not limited to: Jupyter Notebook, different Python libraries and pre-trained models.

#### 1.2.2. Requirements, assumptions, and constraints

Based on Kaggle competition rules:

- submissions must be made through Notebooks and follow Kaggle competition format;
- Notebook run-time must be under 9 hours;
- Notebooks are not allowed to have internet access

For this competition we can assume that images and csv files contain meaningful data to estimate biomass. Main constraints are strict submission rules and limited dataset size.

#### 1.2.3. Risks and contingencies

Because of limited data there is a risk of overfitting the models.

#### 1.2.4. Terminology

Biomass - Amount of feed available, measured in grams
State - Australian state where sample was collected
NDVI - Normalized Difference Vegetation Index, measures the "greenness" of vegetation and indicates vegetation health
GDM - Green dry matter
Weighted $R^2$ - Single globally weighted coefficient of determination using weights: Dry_Green_g: 0.1, Dry_Dead_g: 0.1, Dry_Clover_g: 0.1, GDM_g: 0.2 and Dry_Total_g: 0.5

#### 1.2.5. Costs and benefits

As it is a research competition we were given the data for free and tools we plan to use are also completely free, therefore the cost of this project is just the time and energy of team members. Main benefits to the team members are potential prize money and an opportunity to improve their knowledge in machine learning. The results of this project also contribute to sustainable agriculture.

### 1.3. Defining your data-mining goals

#### 1.3.1. Data-mining goals

The main goal of this project is to build and improve a model that predicts five biomass measurements (Dry_Green_g, Dry_Dead_g, Dry_Clover_g, GDM_g and Dry_Total_g) from each image. In order to get the best results, we should experiment with different approaches.

#### 1.3.2. Data-mining success criteria

This project can be considered successful when we manage to achieve a high $R^2$ score.

# 2. Data understanding

## 2.1. Gathering data

### 2.1.1. Outline data requirements

We require data that allows prediction of biomass (multiple components such as Dry_Clover_g, Dry_Total_g, GDM_g, etc) from visual and environmental features.
That includes:

- RGB images of pasture lawns
- Metadata describing environment and vegetation
- Ground truth biomass target values for supervised learning

The data must:

- Link each image to its tabular features via sample_id
- Contain the 5 biomass target variables for training samples
- Include same feature structure for test samples

### 2.1.2. Verify data availability

Available data sources:

- “train/” folder - pasture images for training
- train.csv - tabular data with metadata and target columns
- “test/” folder - images for prediction
- test.csv - tabular data without targets
- sample_submission.csv - expected output format

All required fields for metadata and targets exist in training data.

### 2.1.3. Define selection criteria

To include a sample in training, it must:

1. Have a valid sample_id
2. Have a corresponding image existing in the train/ directory
3. Contain valid numeric target values

## 2.2. Describing data

Training dataset consists of:

- Number of unique images: 357
- Number of tabular columns: 9
- 5 prediction targets
- Mixed feature types:
    - Categorical: State, Species
    - Numerical Pre_GSHH_NVDI, Height_Ave_cm, target

    - Date: Sampling_Date

Target names:

1. Dry_Clover_g
2. Dry_Dead_g
3. Dry_Green_g
4. Dry_Total_g,
5. GDM_g

Data storage format:

- Images are stored as JPG, with fixed resolution of 1200x1000px
- Metadata is stored in CSV file

### 2.3. Exploring data

Initial checks performed:

- Categorical Features:
    - The dataset contains several categorical features, including target_name, State, and Species. Frequency analysis shows that data was collected from 4 Australian states: Tasmania, New South Wales, Western Australia, and Victoria. After filtering, 15 distinct pasture species remain. target_name is repeated per image, representing multiple biomass measures.

- Numeric Features:
    - Features such as Height_Ave_cm and Pre_GSHH_NDVI vary across samples.

- Correlations:
    - Positive correlations were observed between plant height, vegetation indices and biomass targets. This suggests taller plants and greener vegetation are associated with higher biomass.

- Patterns and Trends:
    - Visualization of sample images and numeric features revealed expected trends: greener and taller plants tend to have higher total biomass (GDM_g). While categorical distributions are relatively uniform, numeric features show meaningful variation useful for modeling.

### 2.4. Verifying data quality

We systematically verified the quality of the dataset.

- Missing values: None of the columns contained missing values
- Duplicates: No duplicate rows were found in either training or test datasets.
- Target validity: All biomass values were numeric and within plausible ranges for pasture measurements

- Categorical consistency: Categorical features contained consistent labels with no unexpected labels and with no unexpected values.
- Image files: all referenced image files existed in their respective paths
- Target validity: All biomass values were numeric, non-negative and within plausible ranges for pasture measurements.

All these checks confirm that the dataset is clean and suitable for modelling, so no further preprocessing or cleaning steps were required.

# 3. Planning your project

## 3.1. List of tasks

- Data exploration
    - Understand the data and the task
    - Make sure there is no missing data
    - Clean the data if necessary
    - Team contribution: Anneli 1h, Robert 1h
- Get familiar with the task, data and Kaggle
    - Create a test notebook and submit answers to understand how to use the environment
    - Team contribution: Anneli 1h, Robert 1h
- Train a simple baseline model
    - Use LightGBM (Light Gradient-Boosting Machine), extract simple image features
    - Get positive result
    - Team contribution: Anneli 4h
- Read and learn about neural networks
    - As it is not really covered by the materials of this course we need to do some extra work to really understand how use neural networks
    - Team contribution: Anneli 5h, Robert 5h
- Team meetings to discuss progress / next steps
    - Team contribution: Anneli 1h, Robert 1h
- Train neural network
    - Build a CNN using PyTorch
    - Test different backbones
    - Team contribution: Robert 10h
- Data augmentation
    - Rotate, flip or modify existing images to increase dataset size
    - Experiment with different regularisations / parameters
    - Avoid overfitting
    - Team contribution: Anneli 3h, Robert 1h
- Get a model that predicts at least 0.5 accuracy
    - Test basic models
    - Hyperparameter tuning
    - Team contribution: Anneli 10h, Robert 10h
- Optimize for maximum accuracy
    - Try different approaches, experiment with different architectures or features, tune hyperparameters, combine models when needed etc to get a model thats as accurate as possible
    - Team contribution: Anneli 5h, Robert 5h

### 3.2. Methods and tools

- LightGBM – A gradient boosting framework applied to tabular features, serving as a baseline and comparison model.

- Convolutional Neural Networks (CNNs) – Deep learning models applied to image data:

    - ResNet18 – Small CNN suitable for training from scratch on limited data.
    - EfficientNetB0 – Baseline CNN architecture for comparison.
    - Testing others as well, the grind ain't over yet

- Fusion Model – Combines image embeddings from CNNs with tabular features using a small multi-layer perceptron (MLP) for multi-output prediction.

- Data Augmentation – Aggressive transformations applied to images to increase the effective dataset size and reduce overfitting. Includes: random resized crops, horizontal flips, rotations, color jitter, grayscale conversion, and random erasing.

- Training Strategies – Techniques applied to improve model performance on a small dataset:

    - Early stopping based on validation loss to prevent overfitting.
    - Dropout and Batch Normalization in the fusion head for regularization.
    - Careful learning rate selection and optimization using AdamW

Tools

- Python – Programming language used for data processing, modeling, and analysis.

- Kaggle Notebooks – Environment for development and execution of experiments.

- PyTorch – Deep learning framework for building and training CNNs and fusion models.

- LightGBM – Gradient boosting framework for tabular data modeling.

- pandas & numpy – Libraries for data preprocessing, manipulation, and numerical operations.

- scikit-learn – Provides evaluation metrics ($R^2$, MAE), preprocessing functions, and train/test splitting.

- torchvision.transforms – Provides image augmentation utilities such as cropping, flipping, rotation, and color adjustments.

## Additional Notes

- Dataset Size – The image dataset contains only 357 images, so all deep learning approaches incorporate strategies to mitigate overfitting.

- Evaluation Metrics – Models are evaluated using $R^2$ and Mean Absolute Error (MAE) to measure prediction accuracy.

- Fusion Justification – Combining tabular features with image embeddings allows the model to leverage complementary information from both modalities for multi-output prediction.