Project D5: Leaf Classification

Team: Kaido Kossas, Sander Miller, Egert Koppel
Repository: https://github.com/Ishua1212/Introduction-to-Data-Science-LTAT.02.002

1. Business understanding

The goal of this project, based on the Kaggle Leaf Classification competition, is to build a machine learning model that can automatically identify plant species from leaf characteristics or straight from images. Identifying plant species by hand takes time, requires expertise and can lead to mistakes. An automated model can speed up the task and help users classify leaves more consistently.

Our main business goal is to develop a reliable classifier that predicts probable species and their likelihood. The value of this goal is that researchers or students could use such a model to quickly identify plant samples without needing specialist knowledge. Additional goals include building a clean and reproducible workflow, demonstrating the practical use of machine learning in biological tasks, and achieving a competitive score on the Kaggle leaderboard. The business success criteria are straightforward: the model should produce accurate predictions (measured via the competition's log-loss metric), the workflow should be understandable and the results should show clear improvements over basic baseline models.

The dataset includes numerical features describing leaf shape, margin and texture, as well as images of leaf silhouettes. We will use Python and common machine learning libraries like scikit-learn, pandas and NumPy.

There are several assumptions and constraints. We assume that the dataset is accurate and representative of the 99 species. We will use only the provided data, but we might use data augmentation. Time is another limitation, as this is a course project. We also assume that our computing resources allow us to train the selected models.

Possible risks include overfitting, because the dataset contains many features but not very many samples. To reduce this risk, we will use regularization and careful tuning of models. Another risk is poor generalization if the model learns patterns that only work for the training set. We plan to avoid this by splitting data into training and validation.

Key terms include "species" (the class label we want to predict), "features" (numerical measurements describing each leaf), "multiclass classification" (predicting one label probability out of all possible labels) and "log-loss" (the evaluation metric used in the competition). In terms of costs and benefits, the main cost is time spent on training and tuning the models, writing the report and designing/presenting the poster. The benefits include gaining experience with machine learning workflows, producing a working classifier, and learning how ML can help support biological research.

The data-mining goals are to train and compare different supervised learning models on the feature dataset, select the best-performing model and optimize it so that the final predictions are as accurate and well-calibrated as possible. In addition, we aim to train convolutional neural networks (CNNs) on the image data. This includes designing preprocessing steps, trying several algorithms, tuning hyperparameters, and generating the final probability outputs required by Kaggle. The data-mining success criteria are to achieve a competitive log-loss score, show that the model performs well and build a pipeline that is clear, reproducible and technically correct.

2. Data understanding

The competition we picked provided a dataset that contains leaf measurements for 99 different plant species and we will use that to train a model to classify species based on their leaf characteristics. The competition dataset includes both numerical features extracted from leaf images and the images themselves. All data is freely available within the competition and can be downloaded directly, so there are no issues with data access or missing information. Our project mainly focuses on the numerical features already extracted from the images, but one model will also be trained on the image dataset. There are 192 features in the extracted dataset that summarize the visual properties in a structured way and are well-suited for classical machine learning algorithms.

The dataset includes 990 training samples (10 samples of each species) and 594 test samples. Each training sample contains an ID, a species label, and 192 numerical features. These features come in three groups: 64 shape features that describe the outline of the leaf, 64 margin features that describe edge patterns and 64 texture features that describe variations in the leaf surface. The labels represent 99 distinct plant species. The images provided are simple black-and-white silhouettes of leaves. They match the numerical features because the features were computed from the same silhouettes. Since the images are clean and free of noise, we plan to use them as additional input sources.

Exploring the dataset reveals several useful patterns. The class distribution across the 99 species is even. This means the model should not be heavily biased toward particular species. Examining the features shows a range of distributions: some look normal, others are skewed. There are also correlations among features, mostly within the same feature group. This means some features capture similar aspects of the leaf, which might later require dimensionality reduction or regularization. Across feature groups, correlations are weaker, suggesting that shape, margin and texture each provide different types of information valuable for classification. Looking at the images confirms that they are simple silhouettes with clear boundaries and little variability in background or lighting. This consistency helps the feature extraction process and reduces noise in the dataset.

Verifying data quality shows that the dataset is clean and reliable. There are no missing values in the numerical features, all IDs are unique and species names are consistent. Because the dataset was prepared for a competition, it appears to have undergone thorough cleaning

beforehand. No obvious outliers or corrupted files were found during exploration. The feature values are within sensible ranges and the images load correctly and contain valid leaf shapes. Still, there are some considerations to keep in mind. The dataset is relatively small, which could increase the risk of overfitting. The high number of features (192) relative to the number of samples also means that models might capture noise if not properly regularized. Another limitation is that the silhouettes do not represent real-world photos, so models trained on this data may not generalize outside the competition. However, for the purpose of building a strong classifier for the given dataset, the data quality is more than adequate.

3. Project plan

Tasks:

- Meetings to discuss and plan our project (everyone 3 h)
- Make a presentation about selected topic (everyone 2 h)
- Write a report about our project (everyone 5 h)
- Check data consistency (everyone 1 h)
- Training CNNs using only images (Sander 10 h)
- Training a neural network on the features data. (Egert 10 h)
- Training simpler models we learned in lectures and practice sessions (RF, KNN, DT) (Kaido 10 h)
- Make ensemble of those models (everyone 1 h)
- Find the best model (highest accuracy) for leaf classification
- Take part of the Kaggle competition
- Create nice graphs and illustrations for poster (everyone 2 h)
- Make a poster (everyone 6 h)

Methods:

- CNNs
- NNs
- RF
- KNN
- DT
- Jupyter Notebook
- Python (scikit-learn, tensorflow, pandas, plotnine, numpy, matplotlib)
- ChatGPT