U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

SPRUCE TREE TYPE DETECTION

3LEIC12 | GROUP 120

Daniel Dória | up202108808
Leonor Filipe | up202204354
Luís Alves | up202108727

CONTENTS

SPECIFICATION OF THE MACHINE LEARNING PROBLEM

DATASET

- Spruce tree type detection
- 15.120 observations over different 30m x 30m patches in the forests of Alberta, Canada

CARTOGRAPHIC VARIABLES

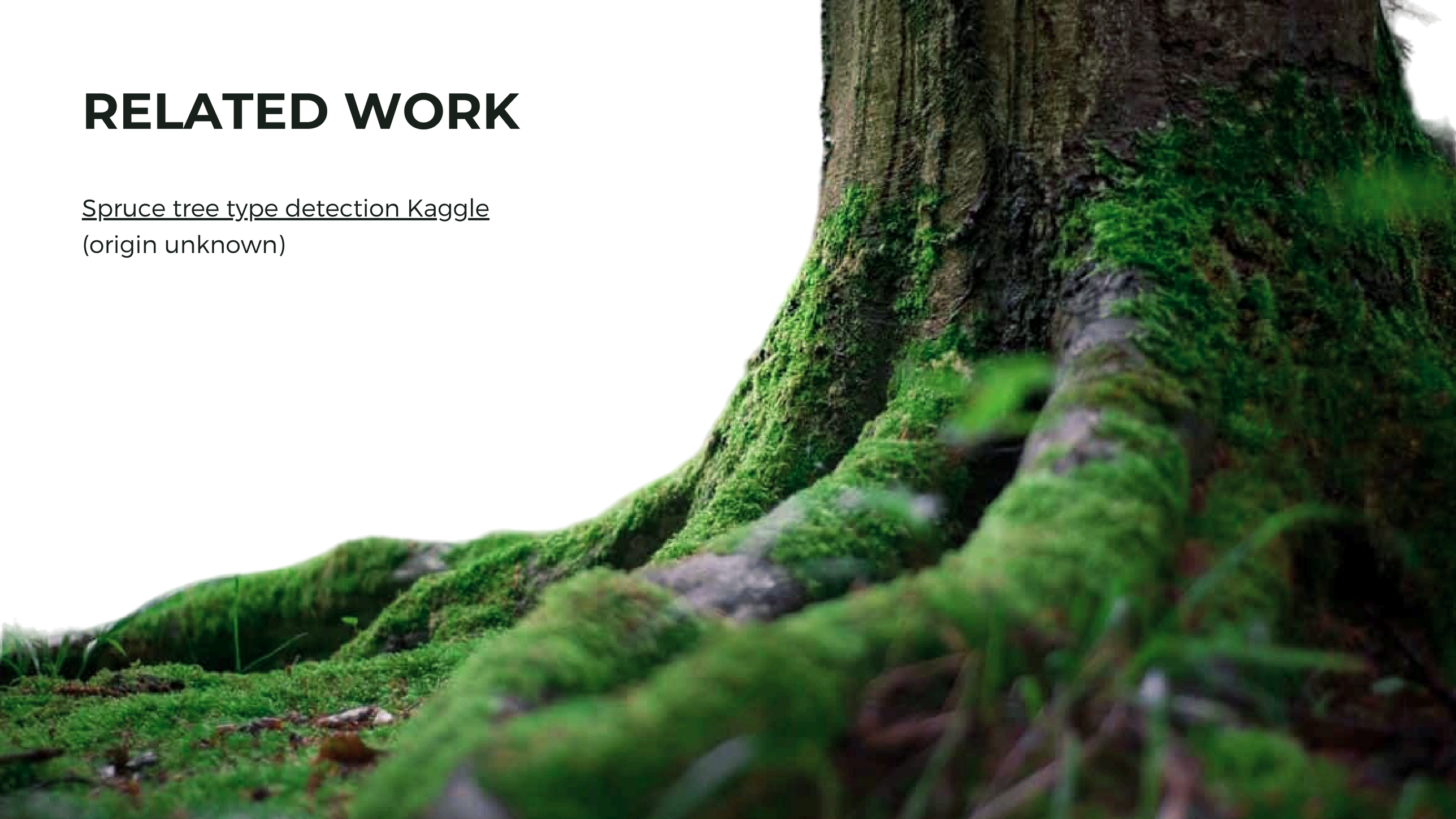
- Spruce tree information (44 integer features): **elevation**, **slope**, distance to **hydrology**, **roadways**, **fire points** and **soil type**

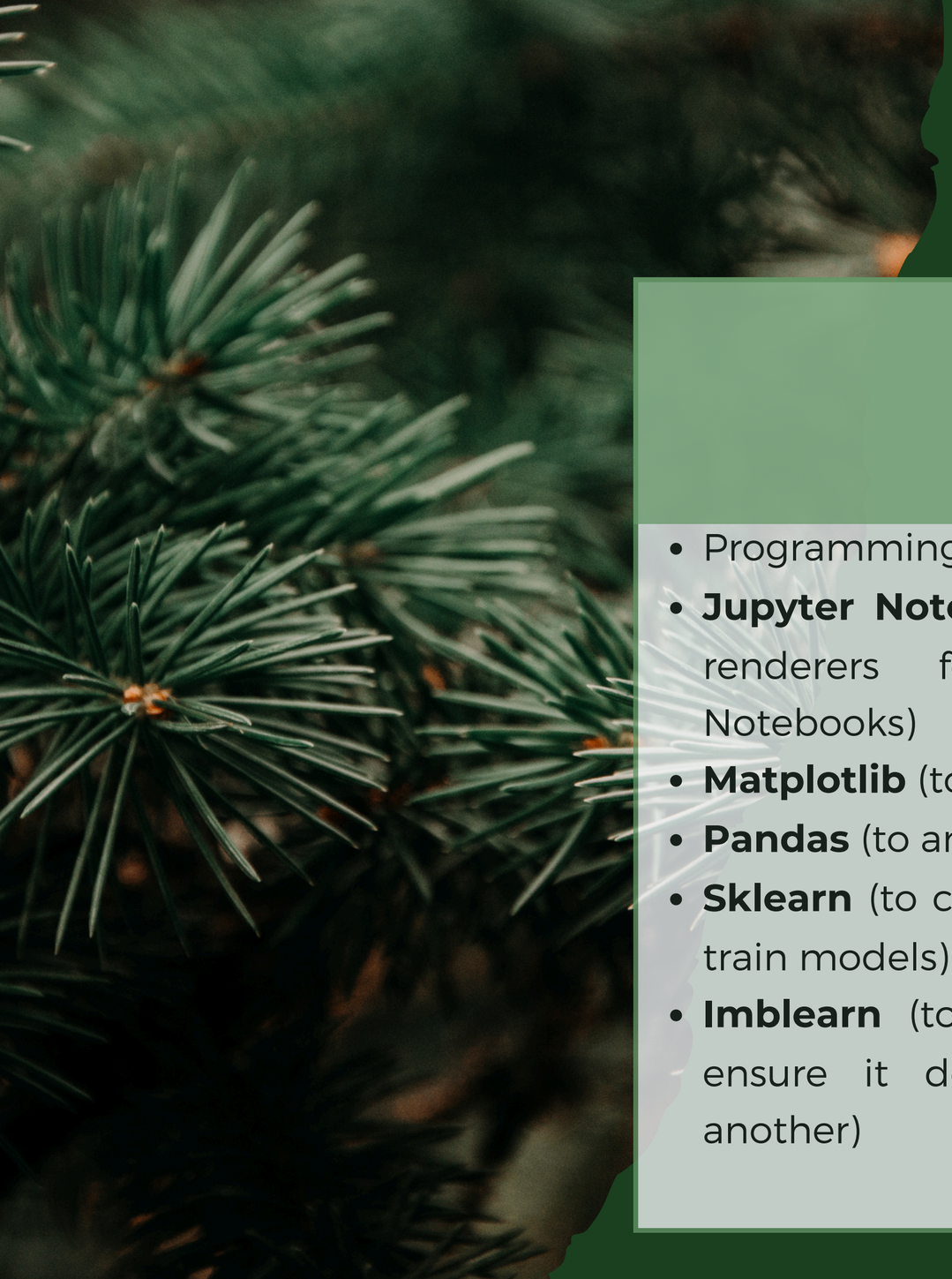
TARGET VARIABLE

- **Tree type**: **Spruce** (meaning Spruce tree was found predominant in the observed patch) or **Other** (meaning trees other than Spruce were found predominant)

RELATED WORK

Spruce tree type detection Kaggle
(origin unknown)





TOOLS & ALGORITHMS

TOOLS

- Programming language: **Python**
- **Jupyter Notebook Renderers** (provides renderers for outputs of Jupyter Notebooks)
- **Matplotlib** (to create charts)
- **Pandas** (to analyse the data)
- **Sklearn** (to create, test performance and train models)
- **Imblearn** (to balance the dataset and ensure it doesn't favor a class over another)

SUPERVISED ALGORITHMS

- Decision Tree
- SVC
- Nearest Neighbor
- Random Forest
- Logistic Regression
- Gaussian Naive Bayes
- Multi-layer Perceptron
- XGBoost
- LightGBM
- CatBoost
- GradientBoosting

DATA PRE-PROCESSING

- **Class Distribution**:
  - Visual representation of Tree_Type distribution.
  - Highlight the need for encoding the target variable.
- **Target Encoding**:
  - Conversion of Tree_Type from string to integer using LabelEncoder.
- **Soil Type Aggregation**:
  - Aggregating 38 soil type columns into a single Soil_Type column.
- **Distance Normalization**:
  - Each feature's values are transformed so that they fall within the range [0, 1].
- **Correlation Analysis**:
  - No highly correlated features were identified, suggesting no need for dimensionality reduction.

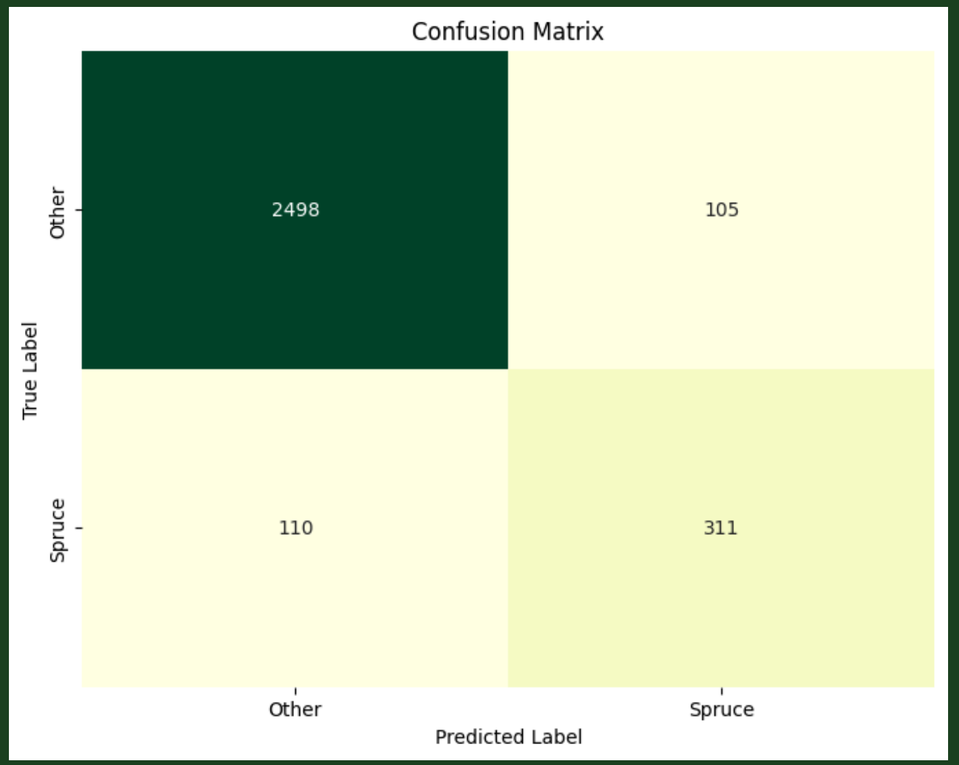
DEVELOPED MODELS

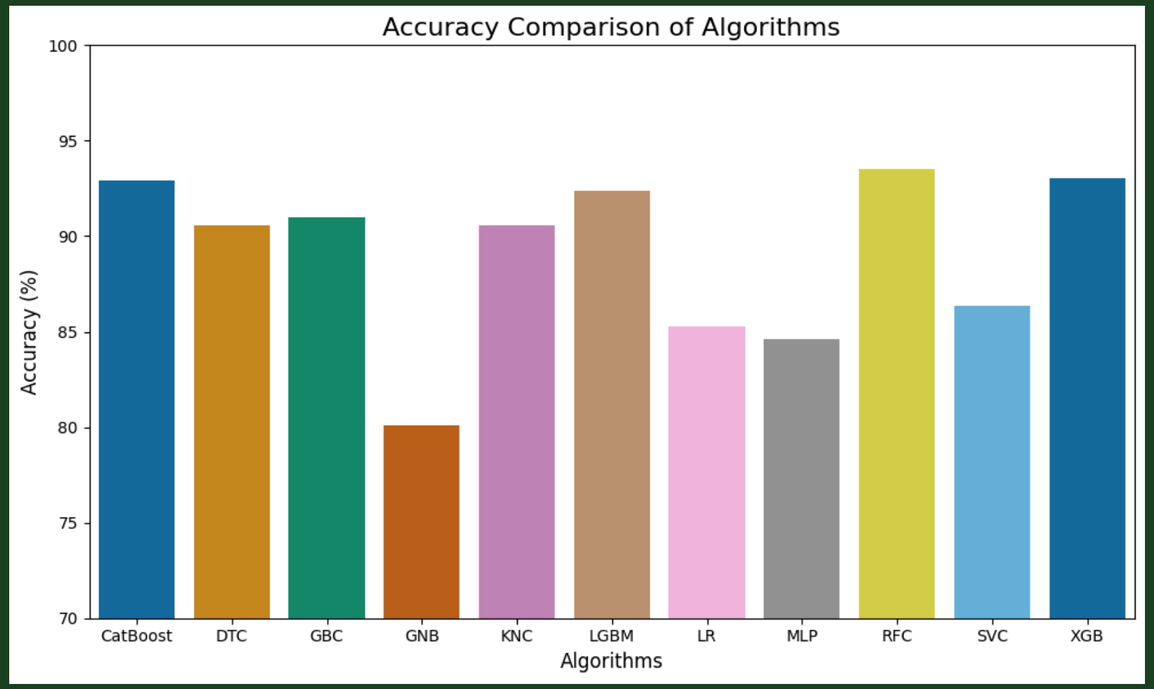
- Address **class imbalance** by **oversampling minority classes**. This function duplicates instances of the minority class to ensure that the model has enough examples from all classes during training.

- **Metrics used**: accuracy, AUC, recall, F1 score, Cohen's kappa, correlation coefficient, precision, and training time.

EVALUATION & COMPARISON

- **Best-performing model**: **Random Forest** (highest accuracy, F1 score, and precision).
- **Importance of features** in the Random Forest model (1st **Elevation**).





- **Accuracy**: 93%, indicating a high level of correctness in its predictions.
- **Precision and Recall for 'Spruce'**: 75% and 74% respectively, highlighting a balanced performance in identifying 'spruce' trees.
- **F1 Score for 'Spruce'**: 74% reflects a balanced trade-off between precision and recall.