Rusu Andrei Ionut C2 341

Proiect PCLP3 - Partea I

Git Repo

May 25, 2025

1 Setul de Date

Pentru aceasta tema am ales setul de date Plant Communication Dataset (pare sa fie generat sintetic si bazat pe o ipoteza speculativa) de pe platforma Kaggle, pentru o problema de clasificare.

Setul de date initial a fost augmentat prin adaugarea a trei noi coloane:

- **Soil_Nutrient_Level**: O coloana pentru nivelul de nutrienti din sol. Valorile au fost generate dintr-o distributie normala, cu medii si deviatii standard specifice fiecarui tip de mesaj (`Plant_Message_Type`), pentru a reflecta o legatura intre starea plantei si nutrienti. Am folosit deviatii standard mari a.i. plajele de valori sa se intercaleze mult, pentru a nu crea o coloana puternic corelata cu coloana target.
- **Temperature_Stress_Factor**: O coloana categorica ('Low', 'Medium', 'High') derivata din `Ambient_Temperature_C`, indicand nivelul de stres termic.
- **Photosynthetic_Efficiency_Index**: Un indice calculat pe baza factorilor de stres termic, umiditate, nutrienti si expunere la soare, reflectand eficienta fotosintetica.

1.1 Introducerea Zgomotului

S-a adaugat zgomot la coloanele numerice relevante. Acest zgomot a fost generat dintr-o distributie normala cu media 0 si o deviatie standard egala cu 2% din deviatia standard a fiecarei coloane. Coloanele afectate includ cele initiale numerice si cele nou create (`Soil_Nutrient_Level`, `Photosynthetic_Efficiency_Index`). Valorile au fost ulterior limitate pentru a ramane in intervale plauzibile.

1.2 Simularea Valorilor Lipsa (NaN)

S-au introdus valori NaN in mod aleatoriu pentru 5% din instante in urmatoarele coloane: `Pollen_Scent_Complexity`, `Bioluminescence_Intensity_Lux`, `Growth_Rate_mm_day`, `Soil_Nutrient_Level`, si `Soil_Moisture_Level`.

2 Analiza Exploratorie a Datelor (EDA)

2.1 Sumarizare Generala a Setului de Date

Initial, s-a realizat o sumarizare generala a setului de date, care a inclus:

- Dimensiunea setului de date: (1000 de randuri, 14 coloane).
- Tipurile de date pentru fiecare coloana: majoritatea float64, una int64 (`Symbiotic_Fungus_Present`) si doua de tip object (`Plant_Message_Type`, `Temperature_Stress_Factor`) (intial si `Plant_ID`, dar am scos o din setul de date dupa EDA, nefiind relevanta).
- Numarul de valori lipsa per coloana (inainte de imputare): 80 de valori lipsa pentru fiecare din cele 5 coloane mentionate anterior.
- Numarul de randuri duplicate: 0.

2.2 Analiza Variabilelor Numerice

Pentru variabilele numerice, s-au calculat statistici descriptive (medie, deviatie standard, min, max, quartile) si s-au generat boxplot-uri pentru a vizualiza distributiile si a identifica potentiali outlieri. Din boxploturi nu am identficat coloane cu valori aberante, ci mai degraba coloane cu plaje largi de valori. Totusi, am aplicat ulterior un tratament al outlierilor prin metoda IQR capping cu intervalele quantile 0.1/0.9.

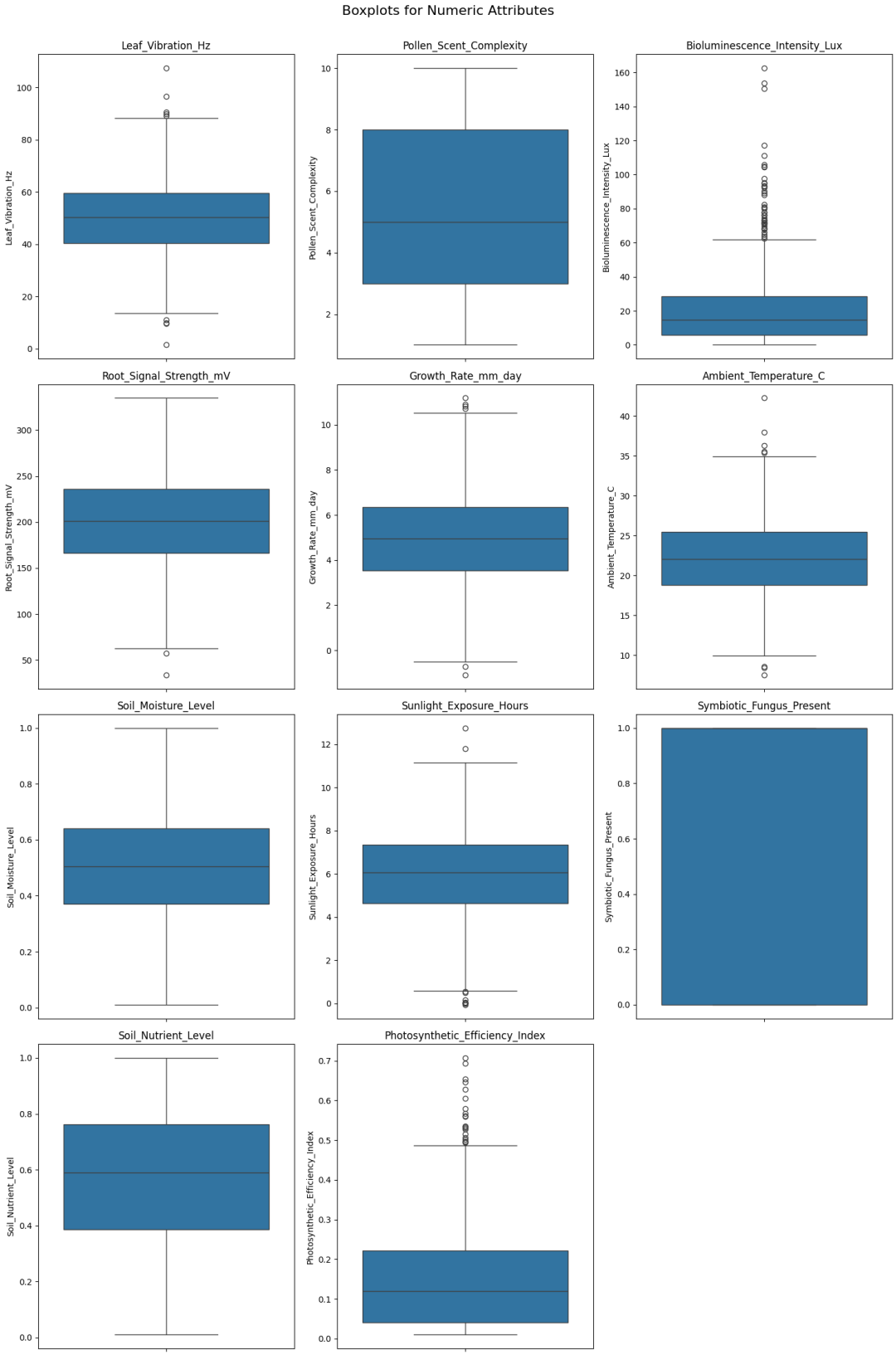

Figure 1: Boxplots date numerice

Observatii:

- `Leaf_Vibration_Hz`: Distributie relativ simetrica in jurul valorii de 50 Hz.
- `Pollen_Scent_Complexity`: Valori intre 1 si 10, cu mediana la 5.

- `Bioluminescence_Intensity_Lux`: O distributie puternic asimetrica la dreapta, cu multe valori mici si cativa posibili outlieri semnificativi cu valori mari.
- `Photosynthetic_Efficiency_Index`: Majoritatea valorilor sunt concentrate in jumatatea inferioara a intervalului [0,1].

2.3 Analiza Variabilelor Categorice

Pentru variabilele categorice (`Plant_Message_Type` si `Temperature_Stress_Factor`), s-au analizat numarul de valori unice si distributia acestora prin countplot-uri.

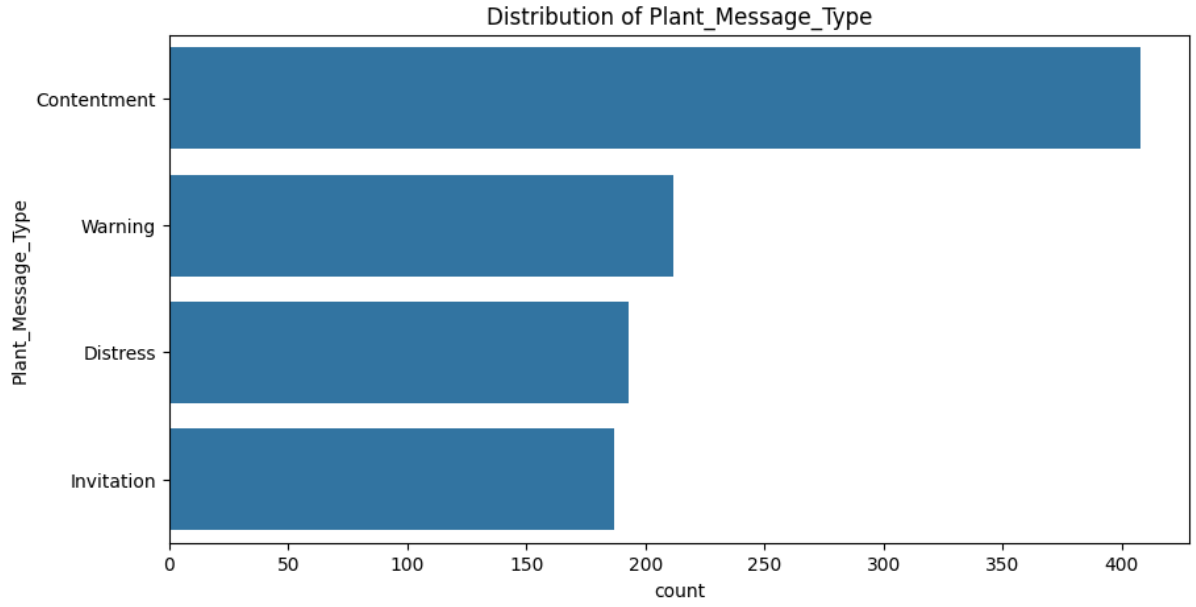

Figure 2: Distributia `Plant_Message_Type`.

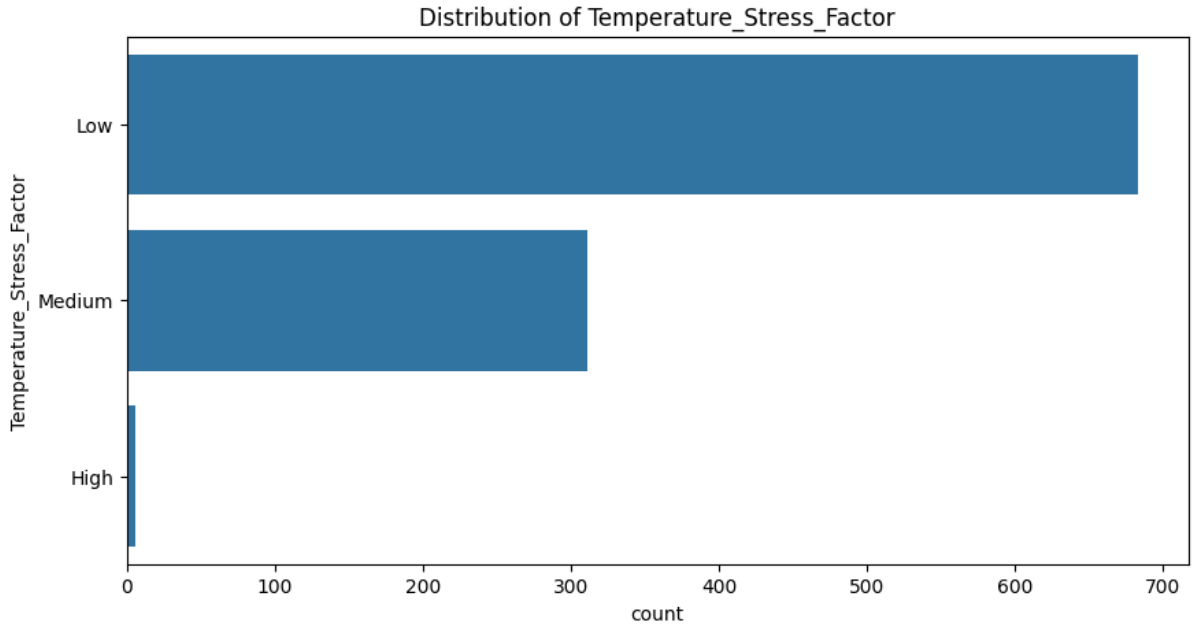

Figure 3: Distributia `Temperature_Stress_Factor`.

Observatii:

- `Plant_Message_Type`: Se observa un usor dezechilibru de clasa, voi folosi class weights pentru antrenarea modelului.

2.4 Analiza Corelatiei intre Atributele Numerice

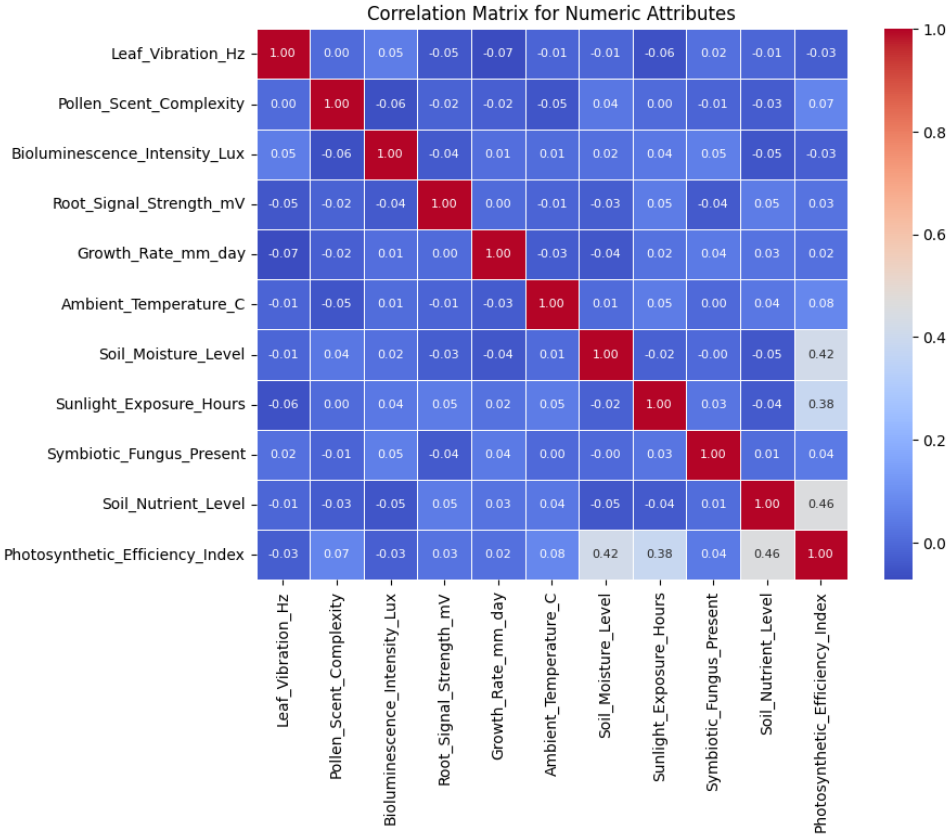

Figure 4: Matricea de Corelatie pentru Atributele Numerice.

Observatii:

- Majoritatea corelatiilor intre variabilele numerice sunt slabe (valori apropiate de 0).
- Corelatiile mai notabile, desi moderate, sunt observate intre `Photosynthetic_Efficiency_Index` si `Soil_Nutrient_Level` (0.52), `Soil_Moisture_Level` (0.43), si `Sunlight_Exposure_Hours` (0.39). Aceste corelatii indica logica de generare a indicelui de eficienta.

3 Preprocesarea Datelor

Inainte de antrenarea modelului setul de date a trecut prin urmatoarele etape de preprocesare:

3.1 Tratarea Valorilor Lipsa (Imputare)

Valorile lipsa identificate in timpul EDA au fost tratate astfel:

- Pentru **coloanele numerice** (`Pollen_Scent_Complexity`, `Bioluminescence_Intensity_Lux`, `Growth_Rate_mm_day`, `Soil_Moisture_Level`, `Soil_Nutrient_Level`), valorile NaN au fost imputate folosind **mediana** fiecarei coloane.
- Pentru **coloanele categorice**, strategia generala ar fi fost imputarea cu **mode**. In acest set de date, dupa augmentare, coloanele categorice ramase in features nu prezentau valori lipsa.

3.2 Tratarea Outlierilor (IQR Capping)

Pentru coloanele numerice, s-a aplicat o metoda de limitare (capping) a outlierilor bazata pe Intervalul Intercuartilic (IQR). S-au utilizat cuantilele $Q1 = 0.1$ si $Q3 = 0.9$ si un factor de $1.5 \times IQR$ pentru a defini limitele. Valorile din afara acestor limite au fost inlocuite cu valoarea limitei corespunzatoare. Efectul acestei operatii a fost vizualizat prin boxplot-uri inainte si dupa capping; de exemplu, pentru `Bioluminescence_Intensity_Lux`, 4 valori superioare au fost limitate.

3.3 Impartirea Setului de Date (Train/Test Split)

Setul de date preprocesat (`df_processed`) a fost impartit intr-un set de antrenare (80%) si un set de testare (20%) folosind `train_test_split` din `scikit-learn`, cu stratificare pe baza variabilei tinta `Plant_Message_Type`.

3.4 Standardizarea Datelor Numerice (Scalare)

Coloanele numerice din `X_train` si `X_test` au fost standardizate folosind `StandardScaler`.

3.5 Encodarea Variabilelor Categorice

- **Variabila Tinta**: `Plant_Message_Type` a fost encodata numeric folosind `LabelEncoder`.
- **Caracteristici (Features)**: Pentru variabilele categorice din `X_train` si `X_test`:
  - `Temperature_Stress_Factor` a fost tratata ca variabila ordinala si encodata cu `OrdinalEncoder` (cu `handle_unknown='use_encoded_value'`, `unknown_value=-1`).
  - Alte coloane categorice (daca ar fi fost prezente si definite ca nominale, ex. `Plant_ID` daca nu ar fi fost eliminat anterior si ar fi fost tratat ca feature) ar fi fost encodate cu `OneHotEncoder`.

4 Modelare si Rezultate

4.1 Modelul Utilizat si Antrenarea

A fost utilizat un clasificator **Random Forest**. Hiperparametrii principali folositi in antrenarea finala au fost:

- `n_estimators=500`
- `criterion='log_loss'`
- `max_depth=None`
- `min_samples_leaf=1`
- `max_features='sqrt'`
- `class_weight='balanced'` (pentru a contracara dezechilibrul claselor)
- `random_state=42`
- `oob_score=True`

4.1.1 Metrici de Clasificare

Principalele metrici de performanta sunt prezentate in tabelul de mai jos.

Table 1: Raport de Clasificare pentru RandomForestClassifier

Clasa	**Precision**	**Recall**	**F1-score**	**Support**
Contentment	0.53	0.84	0.65	82
Distress	0.44	0.62	0.52	39
Invitation	0.00	0.00	0.00	37
Warning	0.25	0.10	0.14	42
Accuracy			**0.48**	200
Macro Avg	0.31	0.39	0.33	200
Weighted Avg	0.36	0.48	0.40	200

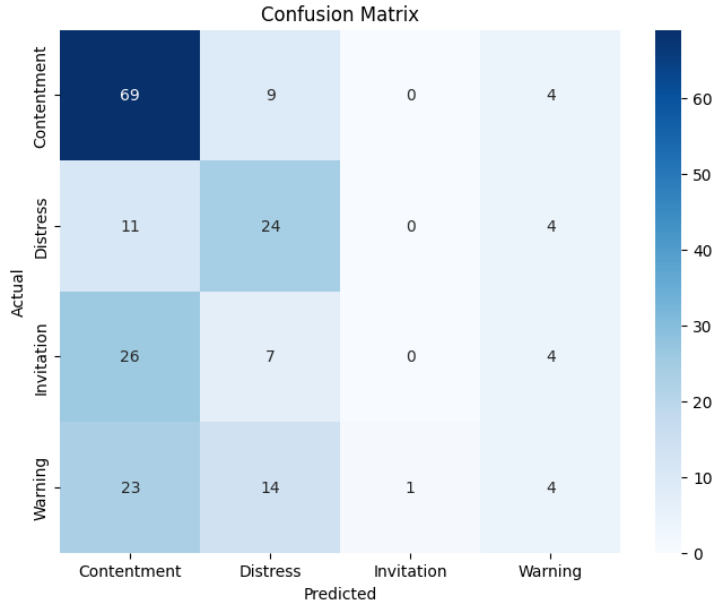

Figure 5: Matricea de Confuzie.

Interpretare: Acuratetea generala a modelului pe setul de test a fost de 48%. Conform Tabelului 1, se observa dificultati majore in clasificarea corecta a clasei 'Invitation', pentru care precision, recall si F1-score sunt 0.00. Acest lucru indica faptul ca toate instantele 'Invitation' din setul de test au fost misclasificate, cel mai des ca 'Contentment', din cauza unei posibile suprapuneri mari a caracteristicilor. Pentru celelalte clase, performanta a fost mai buna, in special Contentment si Distress - scorul recall este de 0.84 respectiv 0.62, indicand o incredere decenta in predictii. Matricea de confuzie (Figura 5) confirma vizual problemele cu clasa 'Invitation', aratand ca toate instantele reale ale acestei clase sunt prezise ca apartinand altei clase. Diagonala principala arata numarul de predictii corecte pentru fiecare clasa. Eu atribui performatele scazute ale modelului naturii sintetice a setului de date.