TaxaTherapy: A Data-driven Lens on Medicinal Plants

CodeOp DST19

Collaborative Project

24.07.2025

Eliana, Leila, Roza

What to Expect?

Introduction

What is a medicinal plant? Why do we need to predict them?

Methods

How do we predict medicinal property?

Analysis

Correlation and classification of medicinal property

Conclusion

Could we classify medicinal property?

What is a medicinal plant?

Medicinal plants are species whose harvested parts—leaves, roots, bark, seeds, or flowers—contain **bioactive compounds** used to:

What is a medicinal plant?

Medicinal plants are species whose harvested parts—leaves, roots, bark, seeds, or flowers—contain **bioactive compounds** used to:

- Prevent

What is a medicinal plant?

Medicinal plants are species whose harvested parts—leaves, roots, bark, seeds, or flowers—contain **bioactive compounds** used to:

- Prevent
- Alleviate

What is a medicinal plant?

Medicinal plants are species whose harvested parts—leaves, roots, bark, seeds, or flowers—contain **bioactive compounds** used to:

- Prevent
- Alleviate
- Or cure human and animal diseases.

What is a medicinal plant?

Medicinal plants are species whose harvested parts—leaves, roots, bark, seeds, or flowers—contain **bioactive compounds** used to:

- Prevent
- Alleviate
- Or cure human and animal diseases.

Across traditional healing systems and modern pharmacology, these plants serve as natural reservoirs of chemicals many of which have become the **basis for today's prescription drugs**.

30%

of all plants carry a medicinal property, while only 5% are used in pharmaceuticals.*

*according to WHO (World Health Organization)

Aim: Can we classify medicinal property based on…

Edibility

Context and variability for consumption

Family

Taxonomical family (not phylogenetic)

Native range

Location to which the plant is endogenous

Methods

How did we analyze data?

Plants for a Future Dataset

Digestive (26%)

Compounds acting on GI tack

Neurological (19.2%)

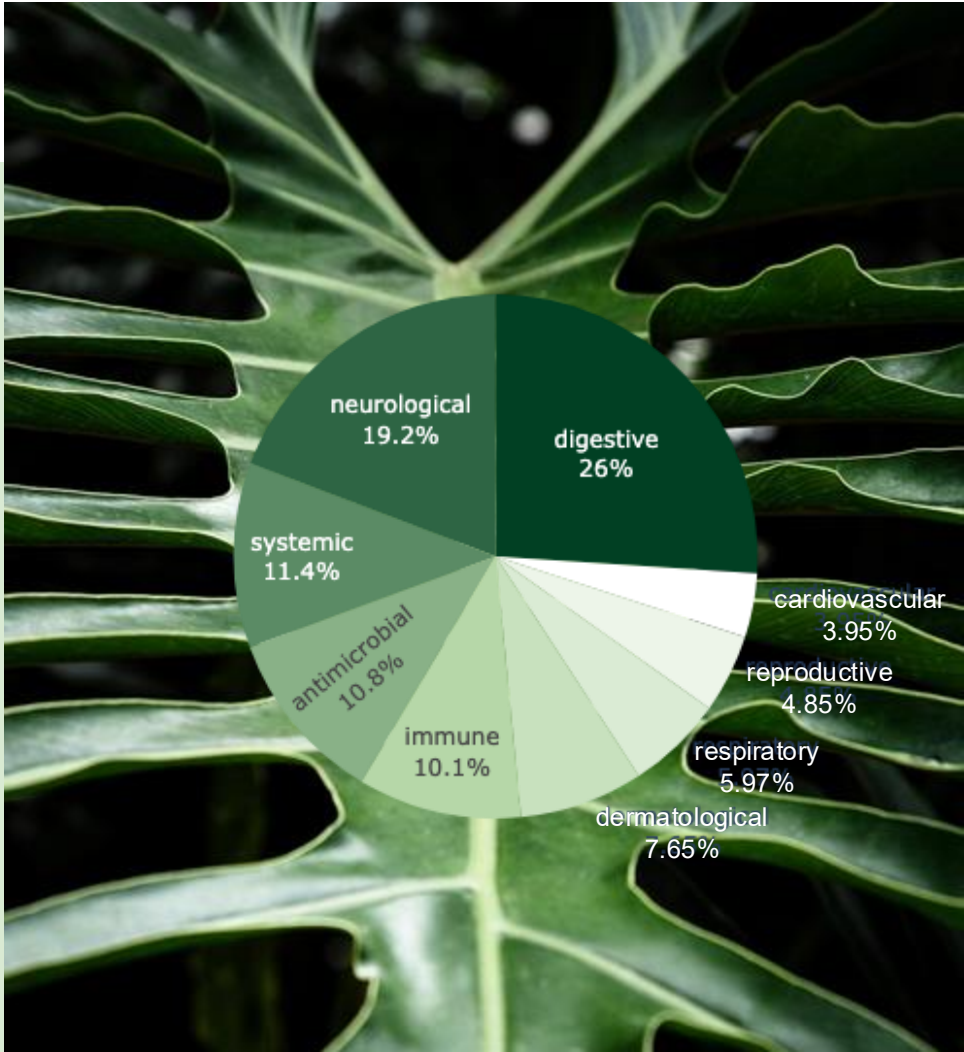
Includes hallucinogenics, “uppers”, ”downers” etc.

Systemic (11.4%)

Mostly metabolic agents e.g. hypoglycaemic

Kaggle dataset from PFAF organization‘s database:

- 17950 rows x 27 columns
- 4131 distinct plants



Features

Native Range

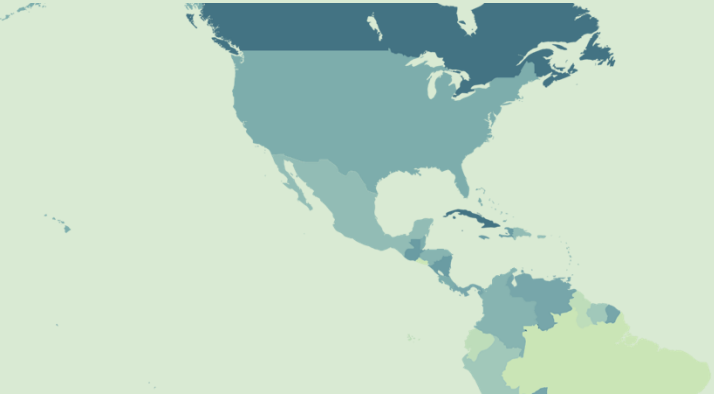
Location where the plant is known to be native

Family

Taxonomic classification of the plant

Edibility

Both the context and the variability (score) of consumption



Features

Native Range

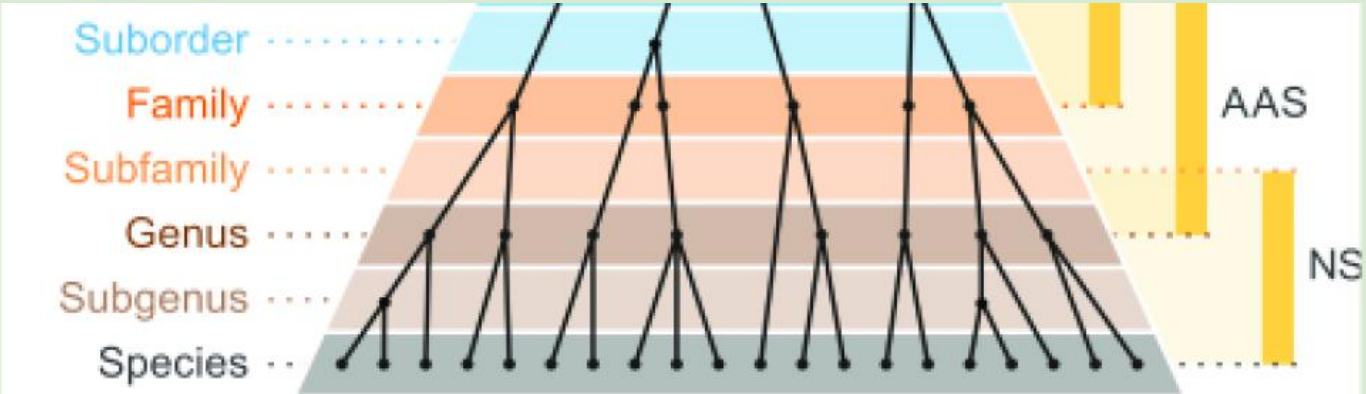
Location where the plant is known to be native

Family

Taxonomic classification of the plant

Edibility

Both the context and the variability (score) of consumption



Features

Native Range

Location where the plant is known to be native

Family

Taxonomic classification of the plant

Edibility

Both the context and the variability (score) of consumption

Natural Language Processing (NLP)

NLP originated from computational linguistics and has evolved into an engineering discipline within computer science. It's used for both Natural Language Understanding (NLU), and Natural Language Generation (NLG).

Natural Language Processing (NLP)

NLP originated from computational linguistics and has evolved into an engineering discipline within computer science. It's used for both Natural Language Understanding (NLU), and Natural Language Generation (NLG).

Data pre-processing

- Tokenization
- Stop word removal
- Lemmatization
- Text normalization

Techniques

- Part-of-speech (POS) tagging
- Named-entity recognition (NER)

Library

- spaCy

Natural Language Processing (NLP)

NLP originated from computational linguistics and has evolved into an engineering discipline within computer science. It's used for both Natural Language Understanding (NLU), and Natural Language Generation (NLG).

Data pre-processing

- **Tokenization**: splits the text into tokens *(words or phrases)*
- **Stop word removal**: removes words with little meaning *(e.g. articles, conjunctions, prepositions, pronouns, auxiliary verbs)*
- **Lemmatization**: uses context and grammar to give the root of the words *(e.g. sowing = sow)*
- **Text normalization**: standardize text *(e.g. removing punctuation, lowercasing)*

Techniques

- **Part-of-speech (POS) tagging**
- **Named-entity recognition (NER)**

Library

- **spaCy**

Natural Language Processing (NLP)

NLP originated from computational linguistics and has evolved into an engineering discipline within computer science. It's used for both Natural Language Understanding (NLU), and Natural Language Generation (NLG).

Data pre-processing

- **Tokenization**: splits the text into tokens *(words or phrases)*
- **Stop word removal**: removes words with little meaning *(e.g. articles, conjunctions, prepositions, pronouns, auxiliary verbs)*
- **Lemmatization**: uses context and grammar to give the root of the words *(e.g. sowing = sow)*
- **Text normalization**: standardize text *(e.g. removing punctuation, lowercasing)*

Techniques

- **Part-of-speech (POS) tagging**: identifies the grammatical role of each word *(e.g. noun, verb, adjective, ...)*
- **Named-entity recognition (NER)**: recognizes and classifies named entities in the text *(e.g. names of people, locations, organizations, ...)*

Library

- **spaCy**

Natural Language Processing (NLP)

NLP originated from computational linguistics and has evolved into an engineering discipline within computer science. It's used for both Natural Language Understanding (NLU), and Natural Language Generation (NLG).

Data pre-processing

- **Tokenization**: splits the text into tokens *(words or phrases)*
- **Stop word removal**: removes words with little meaning *(e.g. articles, conjunctions, prepositions, pronouns, auxiliary verbs)*
- **Lemmatization**: uses context and grammar to give the root of the words *(e.g. sowing = sow)*
- **Text normalization**: standardize text *(e.g. removing punctuation, lowercasing)*

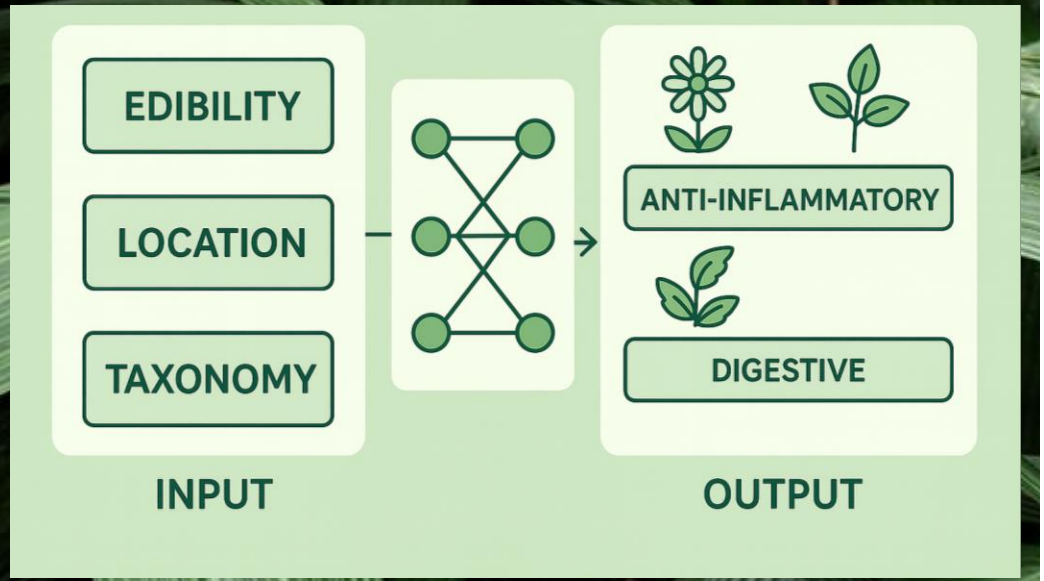
Techniques

- **Part-of-speech (POS) tagging**: identifies the grammatical role of each word *(e.g. noun, verb, adjective, ...)*
- **Named-entity recognition (NER)**: recognizes and classifies named entities in the text *(e.g. names of people, locations, organizations, ...)*

Library

- **spaCy**: is an open-source NLP library that supports over 66 languages. It's used to apply core NLP techniques *(e.g. tokenization, POS tagging, lemmatization, NER)*

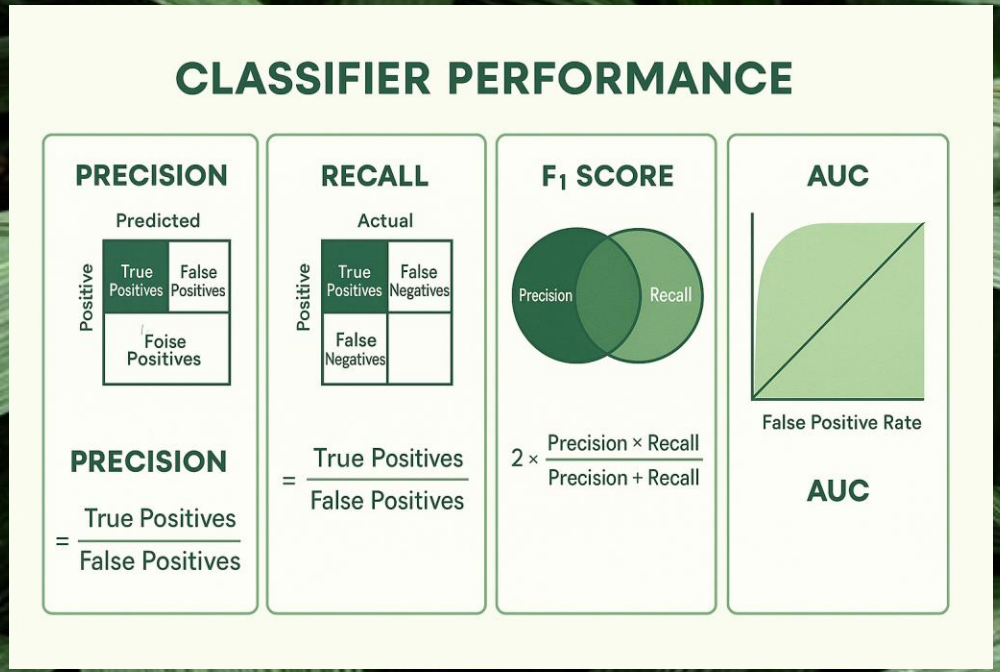
ML: Classifying leads to predicting



ML model that can sort into groups based on features:

- e.g. group plants in medicinal properties based on edibility, location, taxonomy

Classifying leads to predicting



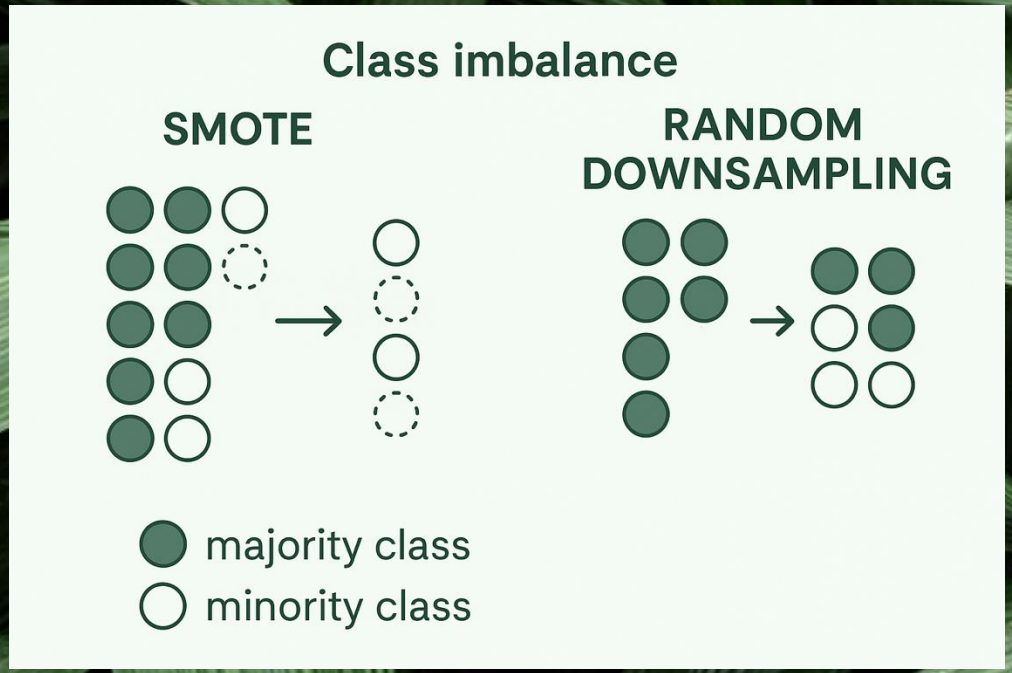
ML model that can sort into groups based on features:

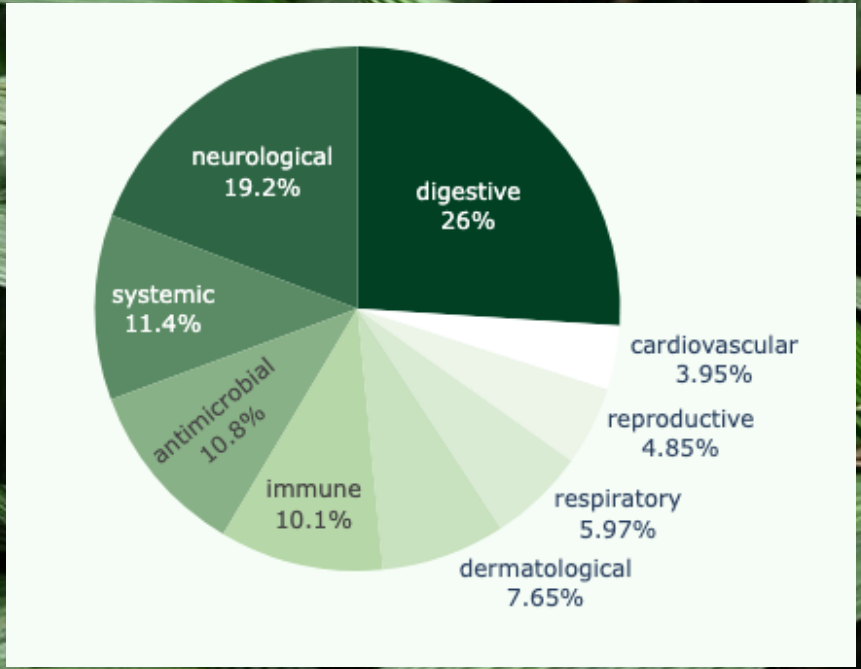
- e.g. group plants in medicinal properties based on edibility, location, taxonomy

To confirm an ML can be used for predicting we test for its performance. These metrics show:

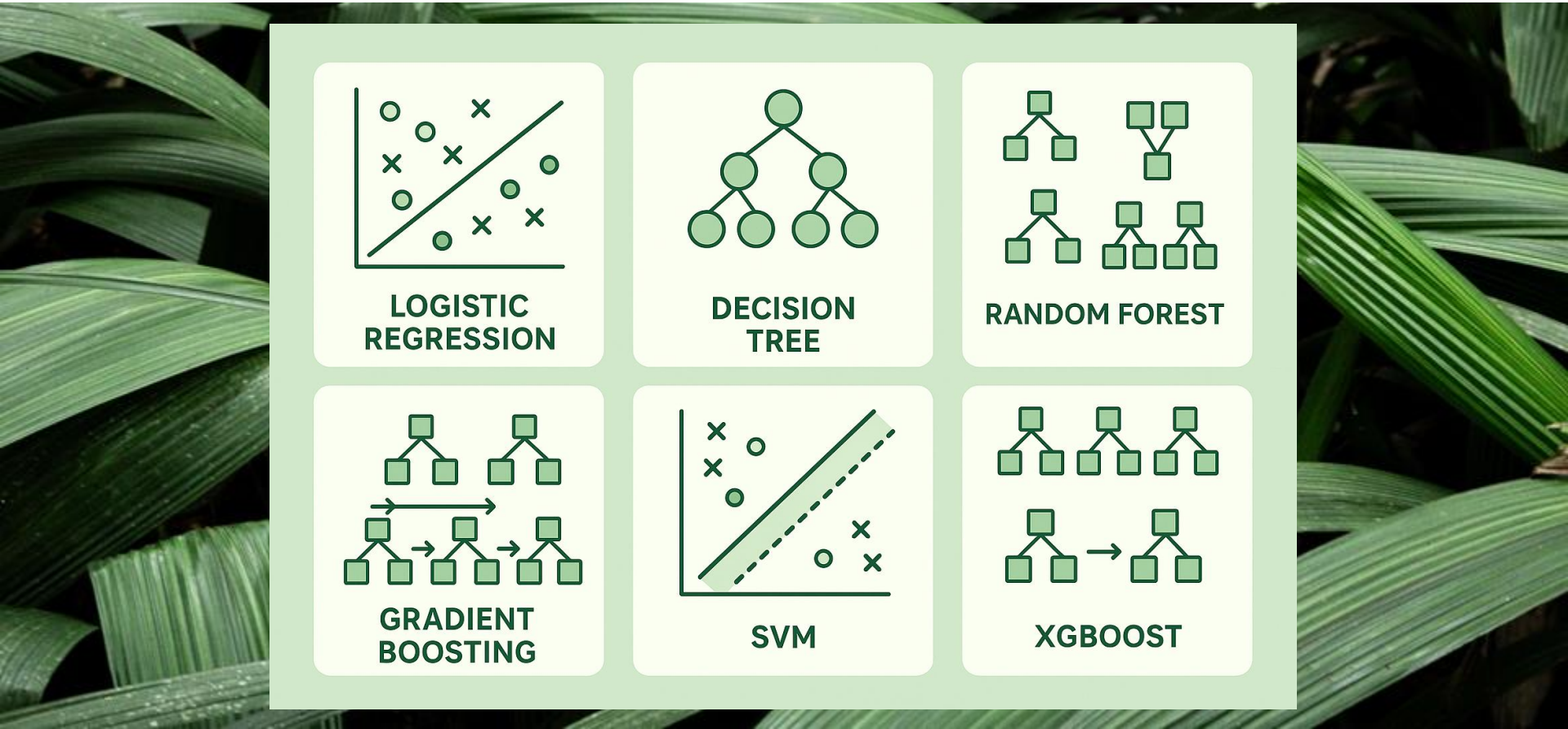
- The rate of true positives vs false positives

Addressing class imbalance

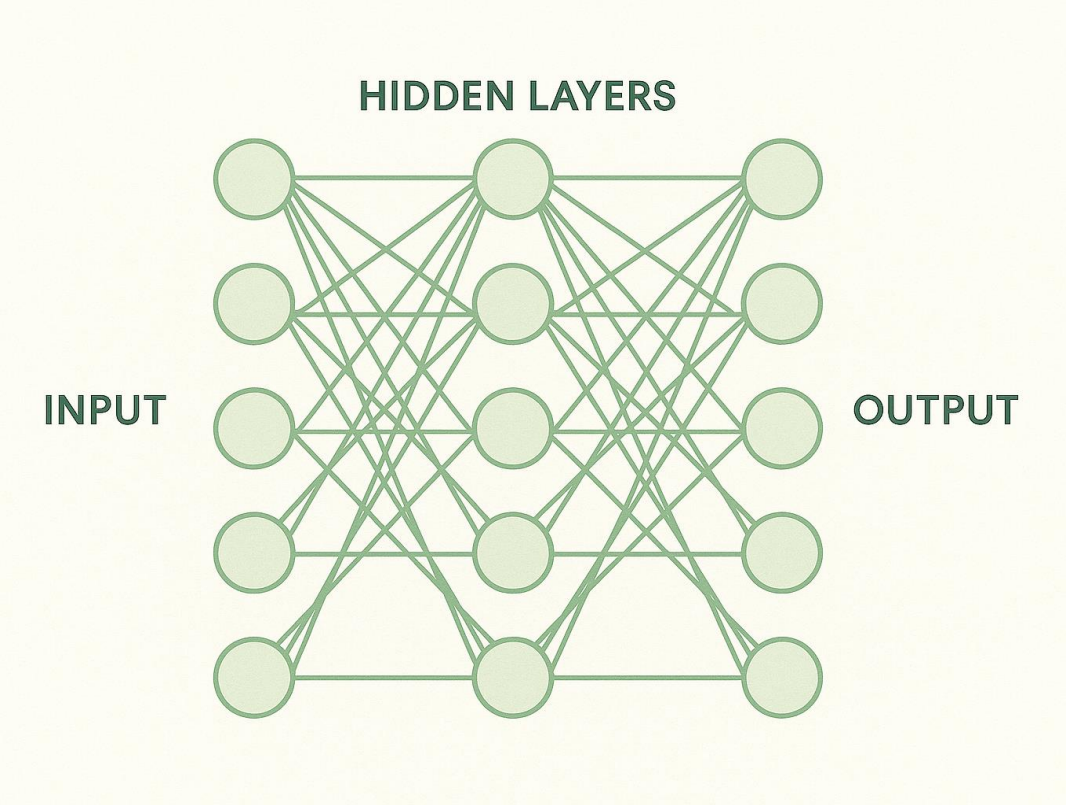




Classifier Selection



Neural Network Models: more suitable to real data



Feature Engineering

No predefined features

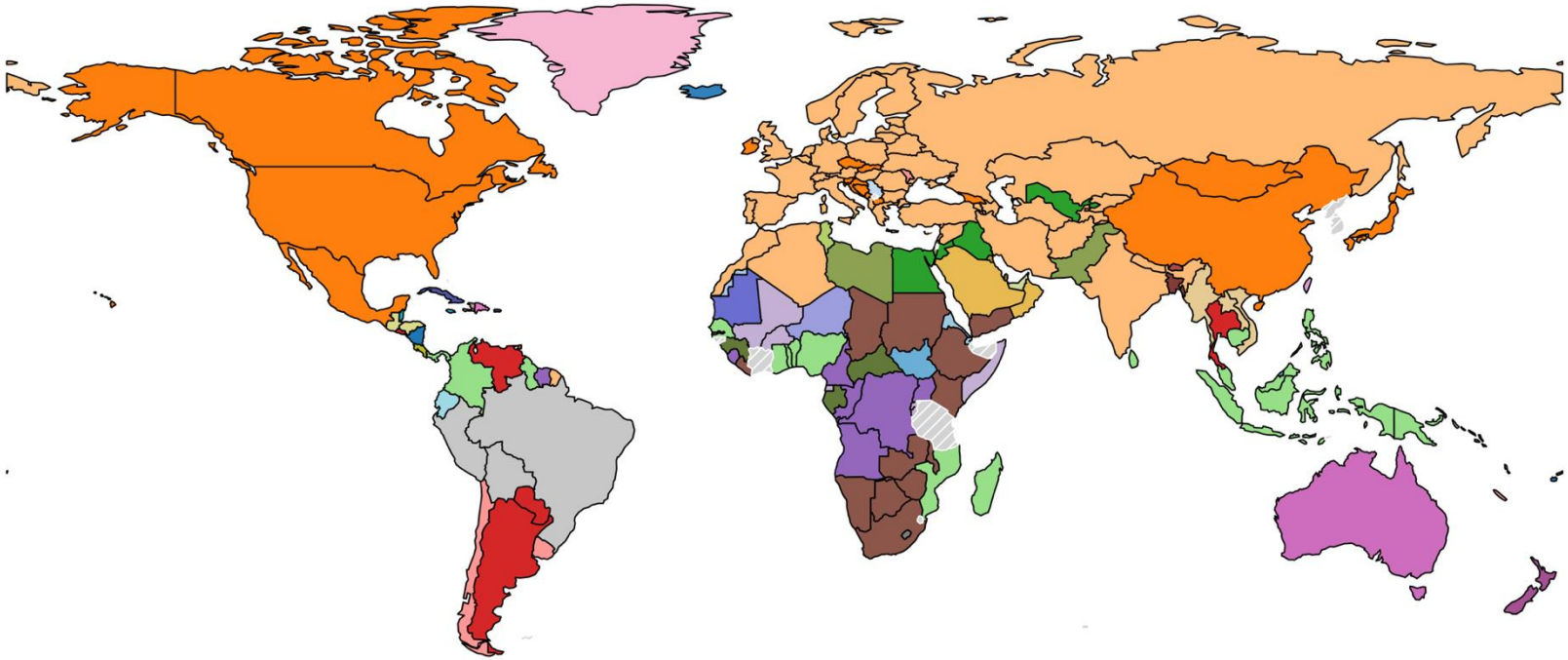
Complexity & Flexibility

No mathematical assumptions, can learn nonlinear patterns

Results

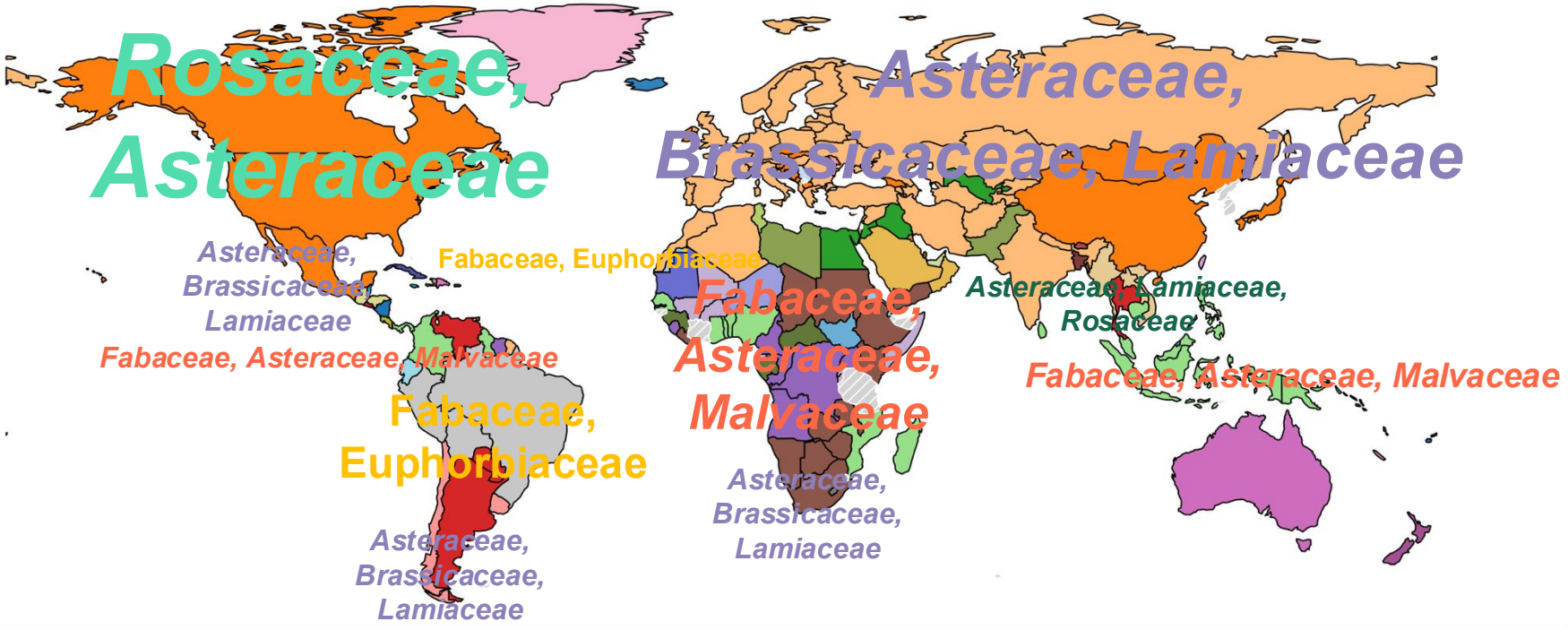
What did we see
in data?

Geographical Origins of the Top Medicinal Plant Families



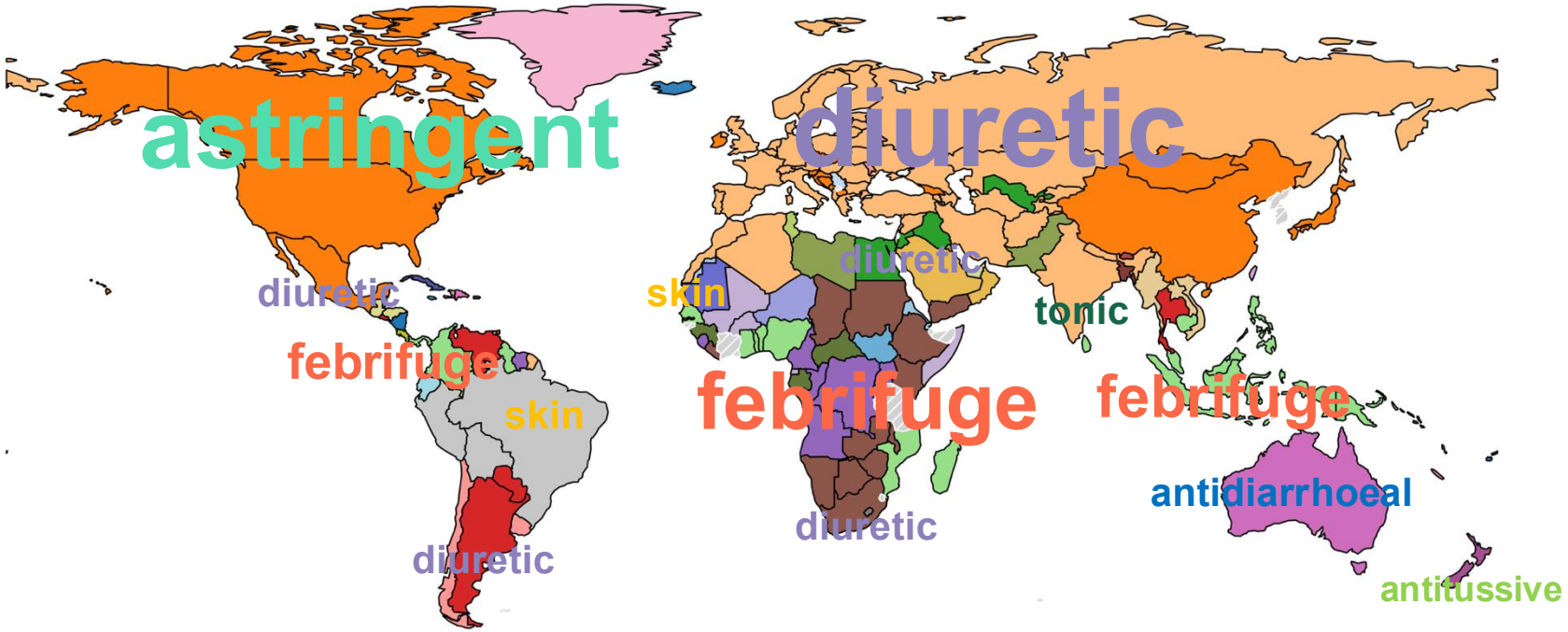
- Malvaceae | Fabaceae
- Cyperaceae
- Asteraceae or Compositae | Rosaceae
- Asteraceae or Compositae | Lamiaceae or Labiatae
- Asteraceae or Compositae | Brassicaceae or Cruciferae
- Fabaceae | Malvaceae
- Fabaceae | Asteraceae or Compositae
- Asteraceae or Compositae | Apiaceae or Umbelliferae
- Fabaceae | Meliaceae
- Fabaceae | Cyperaceae
- Fabaceae | Poaceae or Gramineae
- Solanaceae | Verbenaceae
- Meliaceae | Verbenaceae
- Ericaceae | Lycopodiaceae
- Polypodiaceae | Poaceae or Gramineae
- Fabaceae | Euphorbiaceae
- Malvaceae | Meliaceae
- Asteraceae or Compositae | Fabaceae
- Malvaceae | Solanaceae
- Asteraceae or Compositae | Solanaceae
- Myrtaceae | Malpighiaceae
- Malvaceae | Verbenaceae
- Fabaceae | Zygophyllaceae
- Apocynaceae | Cucurbitaceae
- Fabaceae | Apocynaceae
- Asteraceae or Compositae | Fabaceae or Leguminosae
- Lamiaceae or Labiatae | Asteraceae or Compositae
- Brassicaceae or Cruciferae | Fabaceae
- Capparidaceae | Brassicaceae or Cruciferae
- Zygophyllaceae | Brassicaceae or Cruciferae
- Fabaceae | Plantaginaceae
- Fabaceae | Fabaceae or Leguminosae
- Fabaceae | Rubiaceae
- Rosaceae | Fabaceae or Leguminosae
- Lemnaceae | Rosaceae
- Fabaceae | Celastraceae
- Cucurbitaceae | Fabaceae
- Asteraceae or Compositae | Myrtaceae
- Myrtaceae | Fabaceae
- Asteraceae or Compositae | Rutaceae
- Rosaceae | Ericaceae
- Fabaceae | Musaceae
- Cyperaceae | Fabaceae
- Lamiaceae or Labiatae | Fabaceae or Leguminosae
- Meliaceae | Urticaceae
- No data

Geographical Origins of the Top Medicinal Plant Families



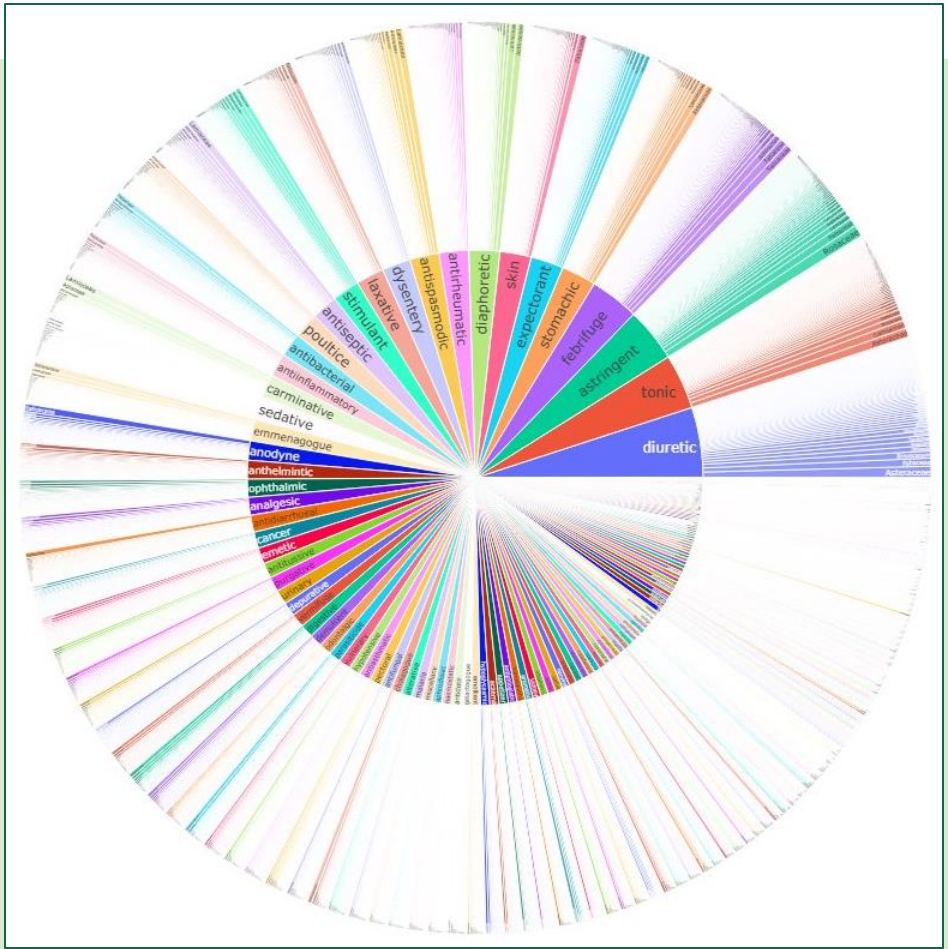
- Malvaceae | Fabaceae
- Cyperaceae
- Asteraceae or Compositae | Rosaceae
- Asteraceae or Compositae | Lamiaceae or Labiatae
- Asteraceae or Compositae | Brassicaceae or Cruciferae
- Fabaceae | Malvaceae
- Fabaceae | Asteraceae or Compositae
- Asteraceae or Compositae | Apiaceae or Umbelliferae
- Fabaceae | Meliaceae
- Fabaceae | Cyperaceae
- Fabaceae | Poaceae or Gramineae
- Solanaceae | Verbenaceae
- Meliaceae | Verbenaceae
- Ericaceae | Lycopodiaceae
- Polypodiaceae | Poaceae or Gramineae
- Fabaceae | Euphorbiaceae
- Malvaceae | Meliaceae
- Asteraceae or Compositae | Fabaceae
- Malvaceae | Solanaceae
- Asteraceae or Compositae | Solanaceae
- Myrtaceae | Malpighiaceae
- Malvaceae | Verbenaceae
- Fabaceae | Zygophyllaceae
- Apocynaceae | Cucurbitaceae
- Fabaceae | Apocynaceae
- Asteraceae or Compositae | Fabaceae or Leguminosae
- Lamiaceae or Labiatae | Asteraceae or Compositae
- Brassicaceae or Cruciferae | Fabaceae
- Capparidaceae | Brassicaceae or Cruciferae
- Zygophyllaceae | Brassicaceae or Cruciferae
- Fabaceae | Plantaginaceae
- Fabaceae | Fabaceae or Leguminosae
- Fabaceae | Rubiaceae
- Rosaceae | Fabaceae or Leguminosae
- Lemnaceae | Rosaceae
- Fabaceae | Celastraceae
- Cucurbitaceae | Fabaceae
- Asteraceae or Compositae | Myrtaceae
- Myrtaceae | Fabaceae
- Asteraceae or Compositae | Rutaceae
- Rosaceae | Ericaceae
- Fabaceae | Musaceae
- Cyperaceae | Fabaceae
- Lamiaceae or Labiatae | Fabaceae or Leguminosae
- Meliaceae | Urticaceae
- No data

Medicinal Properties Pattern around the Globe



- Malvaceae | Fabaceae
- Cyperaceae
- Asteraceae or Compositae | Rosaceae
- Asteraceae or Compositae | Lamiaceae or Labiatae
- Asteraceae or Compositae | Brassicaceae or Cruciferae
- Fabaceae | Malvaceae
- Fabaceae | Asteraceae or Compositae
- Asteraceae or Compositae | Apiaceae or Umbelliferae
- Fabaceae | Meliaceae
- Fabaceae | Cyperaceae
- Fabaceae | Poaceae or Gramineae
- Solanaceae | Verbenaceae
- Meliaceae | Verbenaceae
- Ericaceae | Lycopodiaceae
- Polypodiaceae | Poaceae or Gramineae
- Fabaceae | Euphorbiaceae
- Malvaceae | Meliaceae
- Asteraceae or Compositae | Fabaceae
- Malvaceae | Solanaceae
- Asteraceae or Compositae | Solanaceae
- Myrtaceae | Malpighiaceae
- Malvaceae | Verbenaceae
- Fabaceae | Zygophyllaceae
- Apocynaceae | Cucurbitaceae
- Fabaceae | Apocynaceae
- Asteraceae or Compositae | Fabaceae or Leguminosae
- Lamiaceae or Labiatae | Asteraceae or Compositae
- Brassicaceae or Cruciferae | Fabaceae
- Capparidaceae | Brassicaceae or Cruciferae
- Zygophyllaceae | Brassicaceae or Cruciferae
- Fabaceae | Plantaginaceae
- Fabaceae | Fabaceae or Leguminosae
- Fabaceae | Rubiaceae
- Rosaceae | Fabaceae or Leguminosae
- Lemnaceae | Rosaceae
- Fabaceae | Celastraceae
- Cucurbitaceae | Fabaceae
- Asteraceae or Compositae | Myrtaceae
- Myrtaceae | Fabaceae
- Asteraceae or Compositae | Rutaceae
- Rosaceae | Ericaceae
- Fabaceae | Musaceae
- Cyperaceae | Fabaceae
- Lamiaceae or Labiatae | Fabaceae or Leguminosae
- Meliaceae | Urticaceae
- No data

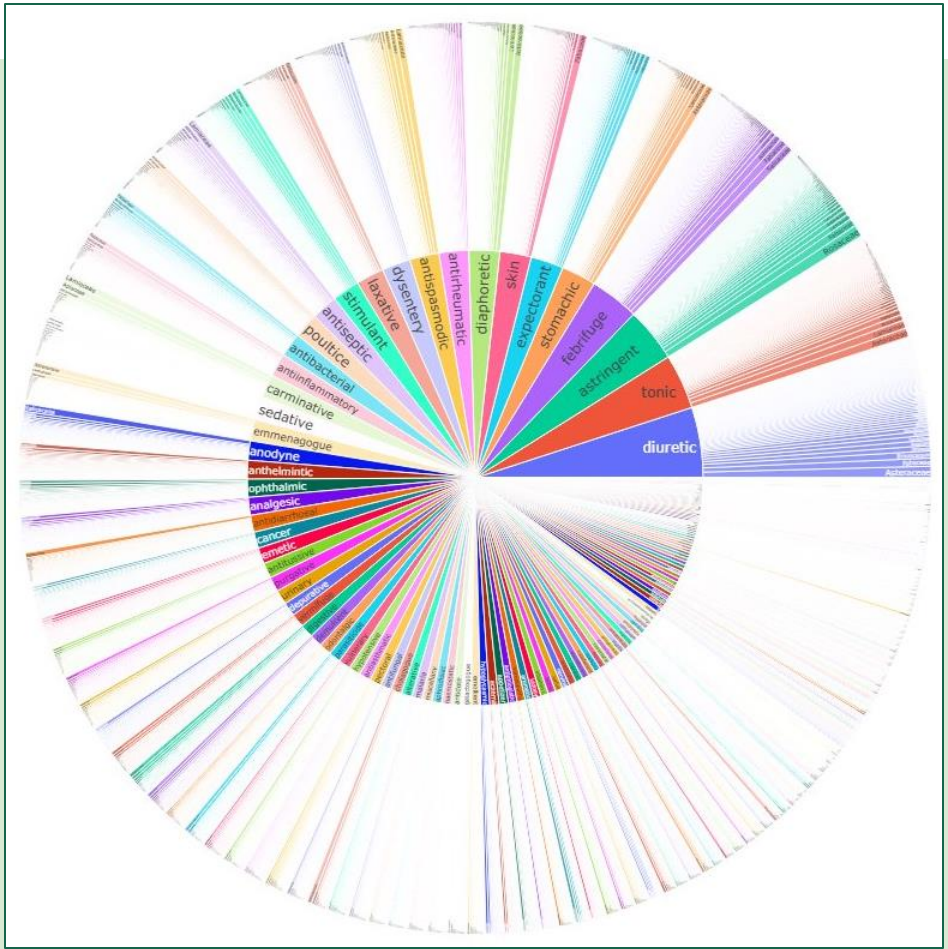
Most Common Medicinal Properties of Plant Families



Most common medicinal properties of the plant families used in different countries:

- **Diuretic** *(Asteraceae, Brassicaceae, Lamiaceae)* = most of Eurasia, Mexico, Argentina, South Africa, Middle East

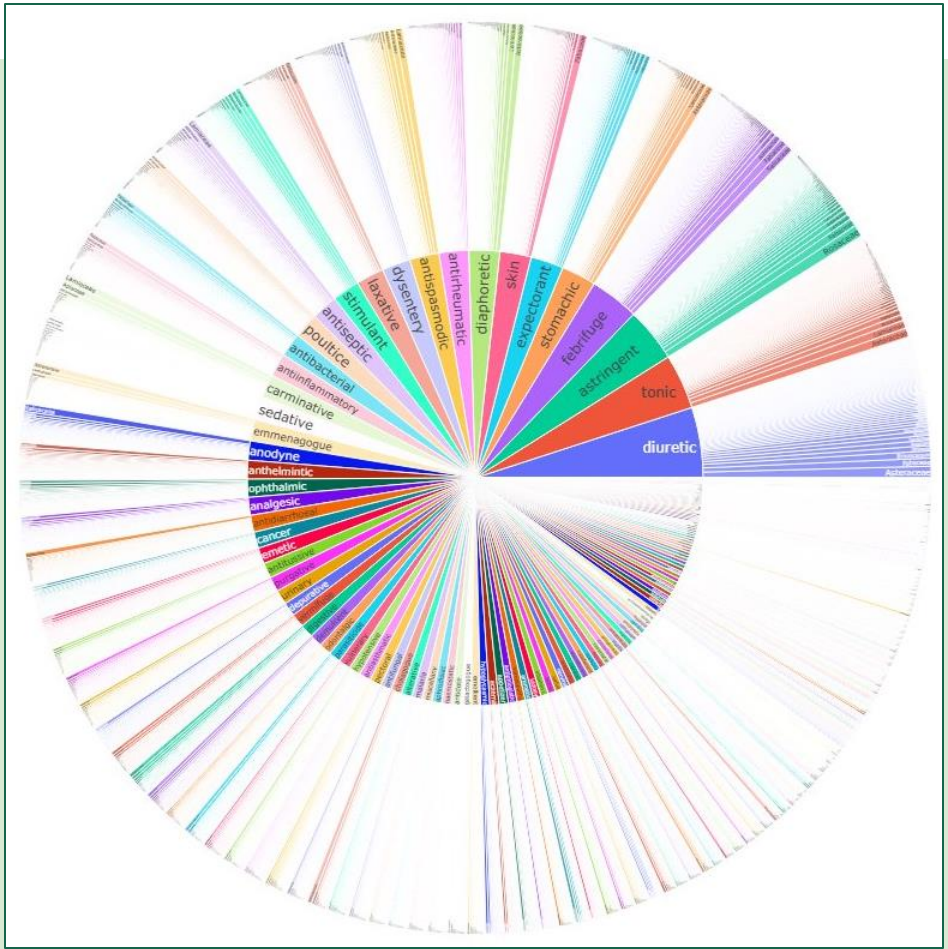
Most Common Medicinal Properties of Plant Families



Most common medicinal properties of the plant families used in different countries:

- **Diuretic** *(Asteraceae, Brassicaceae, Lamiaceae)* = most of Eurasia, Mexico, Argentina, South Africa, Middle East
- **Tonic** (*Asteraceae, Lamiaceae, Rosaceae)* = India, Southern Eurasia

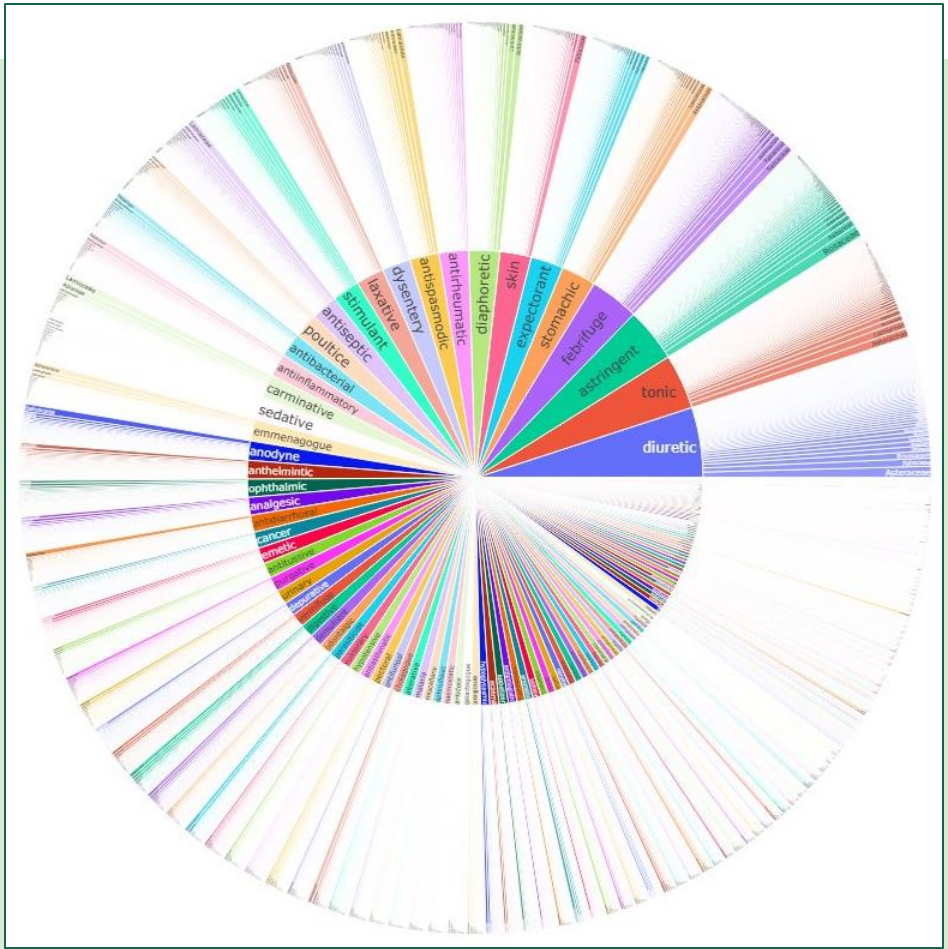
Most Common Medicinal Properties of Plant Families



Most common medicinal properties of the plant families used in different countries:

- **Diuretic** *(Asteraceae, Brassicaceae, Lamiaceae)* = most of Eurasia, Mexico, Argentina, South Africa, Middle East
- **Tonic** (*Asteraceae, Lamiaceae, Rosaceae)* = India, Southern Eurasia
- **Astringent** (*Rosaceae, Asteraceae*) = North America, Greenland, Alaska

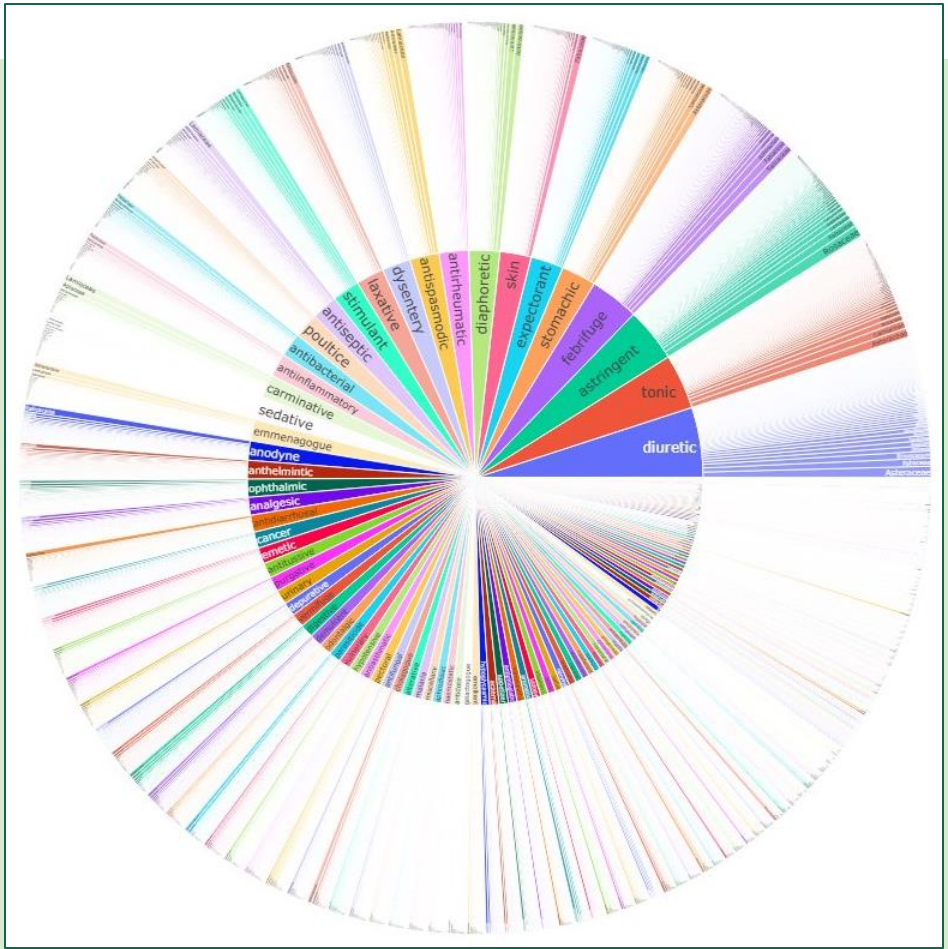
Most Common Medicinal Properties of Plant Families



Most common medicinal properties of the plant families used in different countries:

- **Diuretic** *(Asteraceae, Brassicaceae, Lamiaceae)* = most of Eurasia, Mexico, Argentina, South Africa, Middle East
- **Tonic** (*Asteraceae, Lamiaceae, Rosaceae)* = India, Southern Eurasia
- **Astringent** (*Rosaceae, Asteraceae*) = North America, Greenland, Alaska
- **Febrifuge** *(Fabaceae, Asteraceae, Malvaceae)* = South Asia, Central Africa, Northern South America

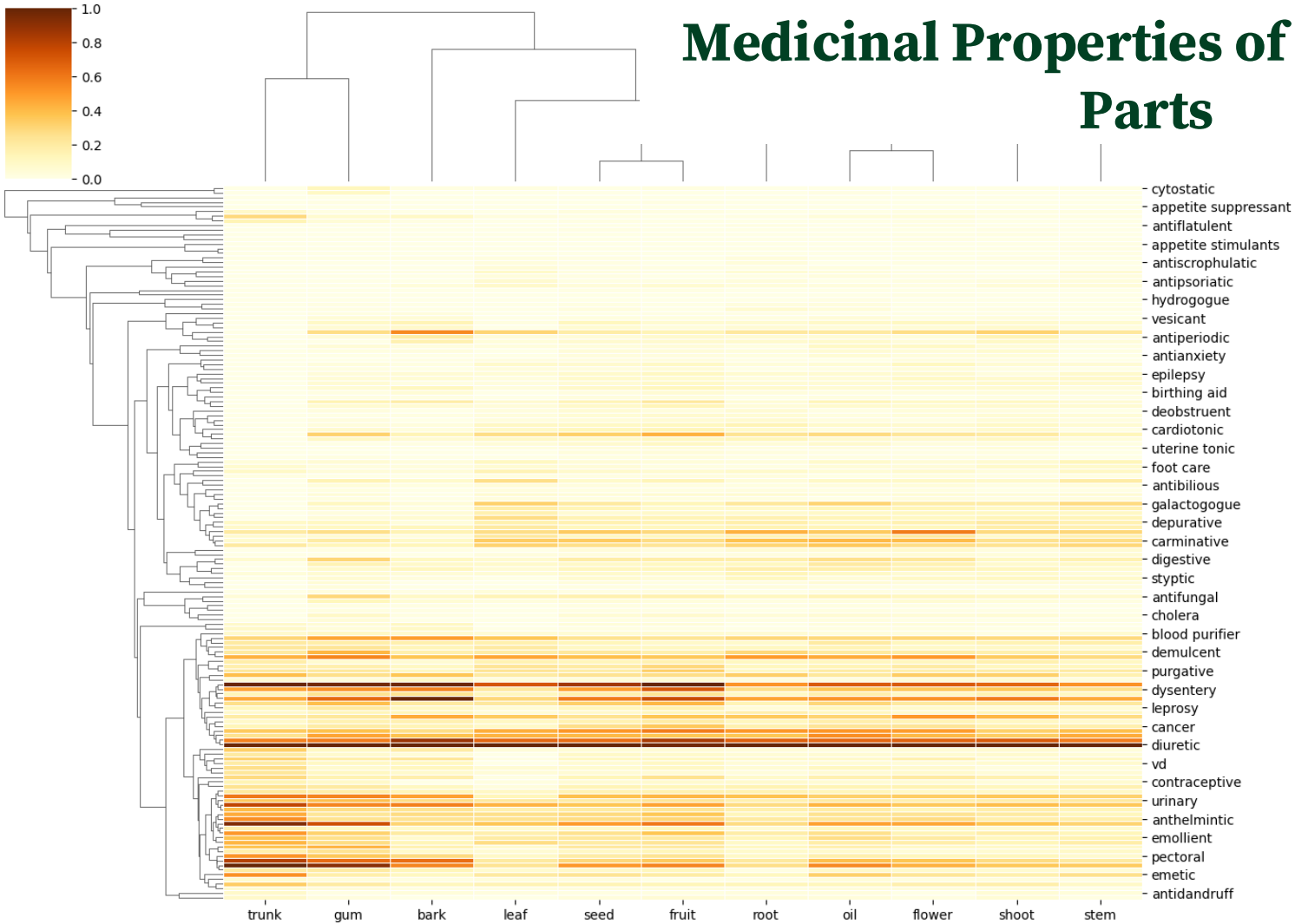
Most Common Medicinal Properties of Plant Families



Most common medicinal properties of the plant families used in different countries:

- **Diuretic** *(Asteraceae, Brassicaceae, Lamiaceae)* = most of Eurasia, Mexico, Argentina, South Africa, Middle East
- **Tonic** (*Asteraceae, Lamiaceae, Rosaceae)* = India, Southern Eurasia
- **Astringent** (*Rosaceae, Asteraceae*) = North America, Greenland, Alaska
- **Febrifuge** *(Fabaceae, Asteraceae, Malvaceae)* = South Asia, Central Africa, Northern South America
- **Skin** *(Fabaceae, Euphorbiaceae)* = Brasil, Northwest Africa
- ...

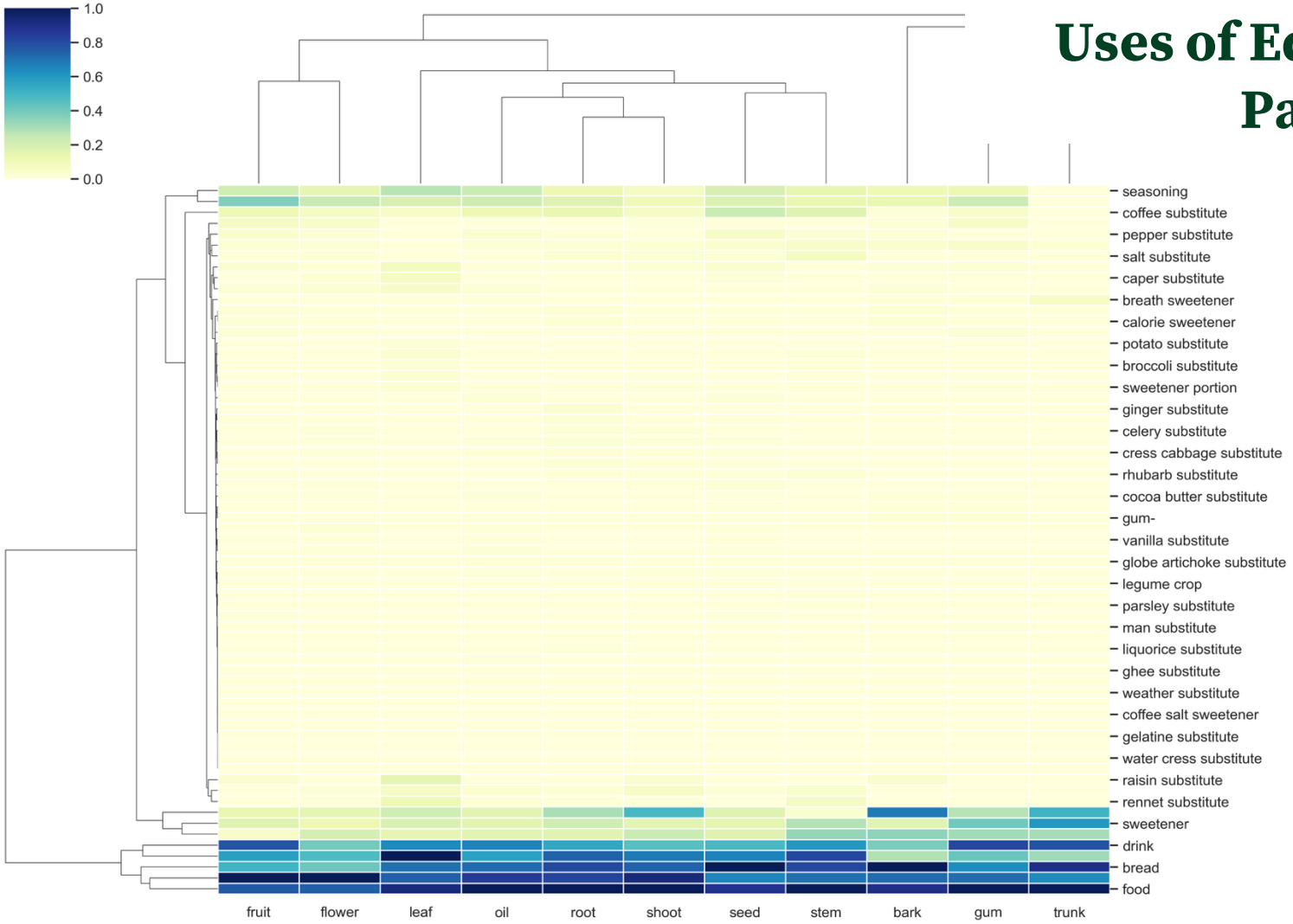
Medicinal Properties of Edible Plant Parts



Most common medicinal properties of edible plant parts used in different countries:

- **Diuretic** (all plant parts)
- **Dysentery** (trunk, gum, bark, fruit)
- **Anthelmintic** (trunk)
- **Emetic** (trunk, gum)
- ...

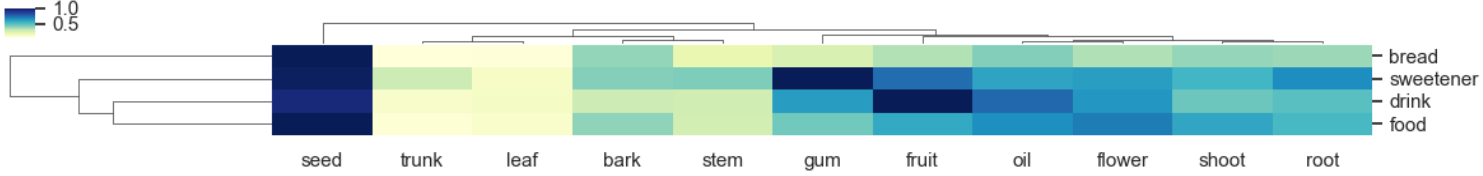
Uses of Edible Plant Parts



The populations use most frequently:

- **Food** and **bread** (all plant parts)
- **Drink** (trunk, gum, fruit)
- **Sweetener** (trunk, bark)
- ...

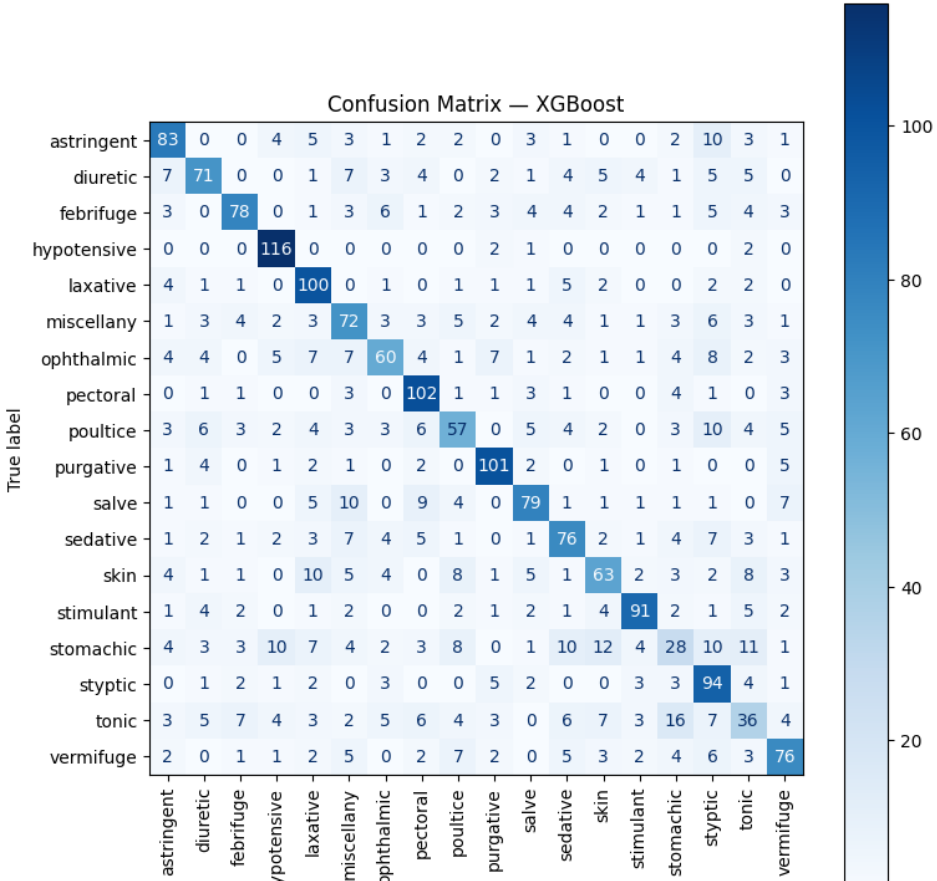
Uses of Edible Plant Parts



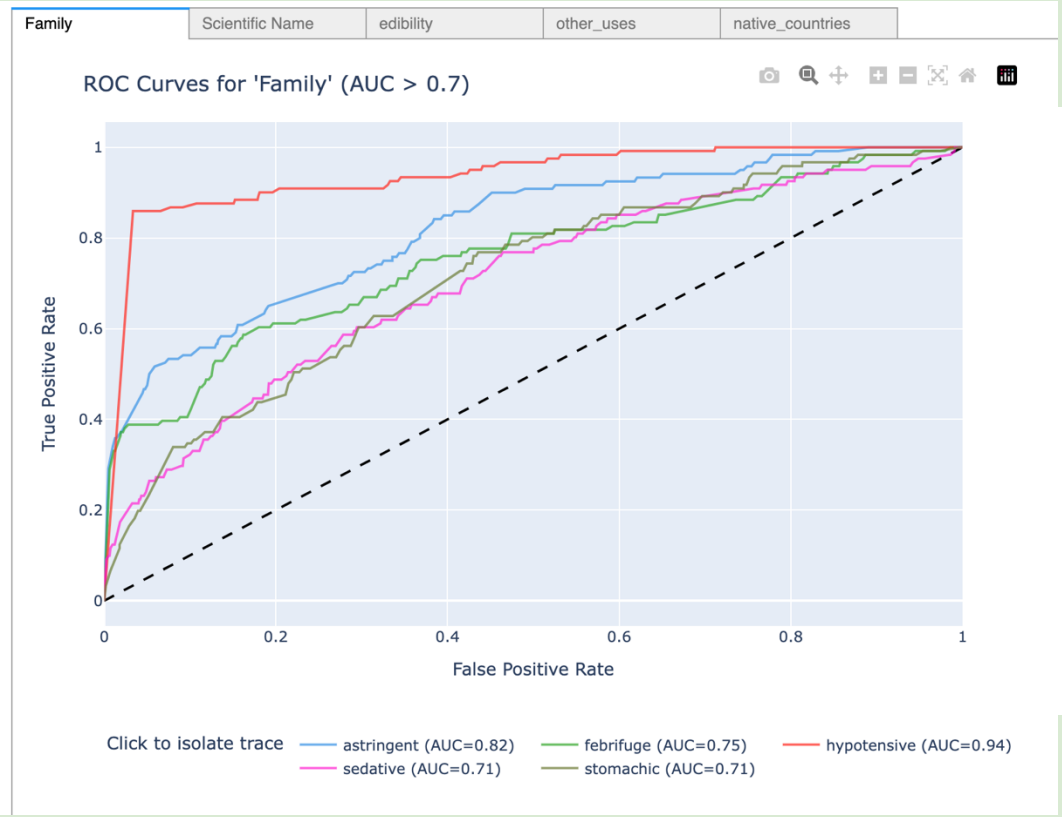
The plants are **used mostly for nutritional purposes** (food, drink, bread), and **considerebly less for other uses**:

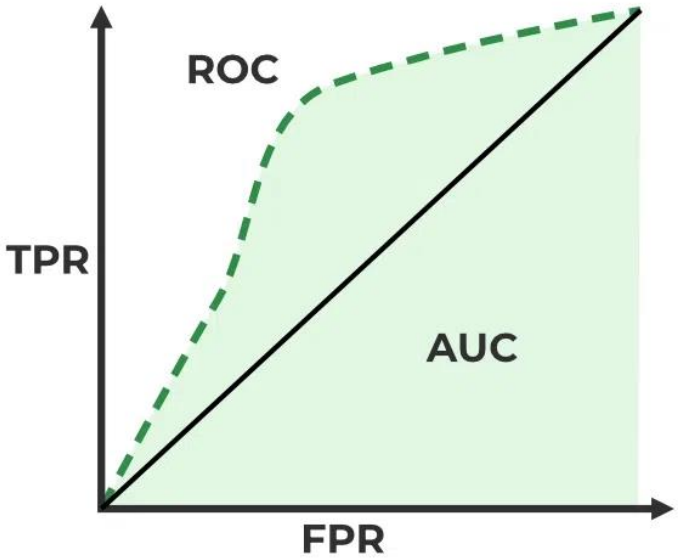
- **Sugar substitute** (flower, shoot, stem, leaf, gum)
- **Rhubarb substitute** (root, fruit, stem)
- **Meat / cheese substitute** (seed, oil, flower)
- ...

Picking the best ML for our dataset

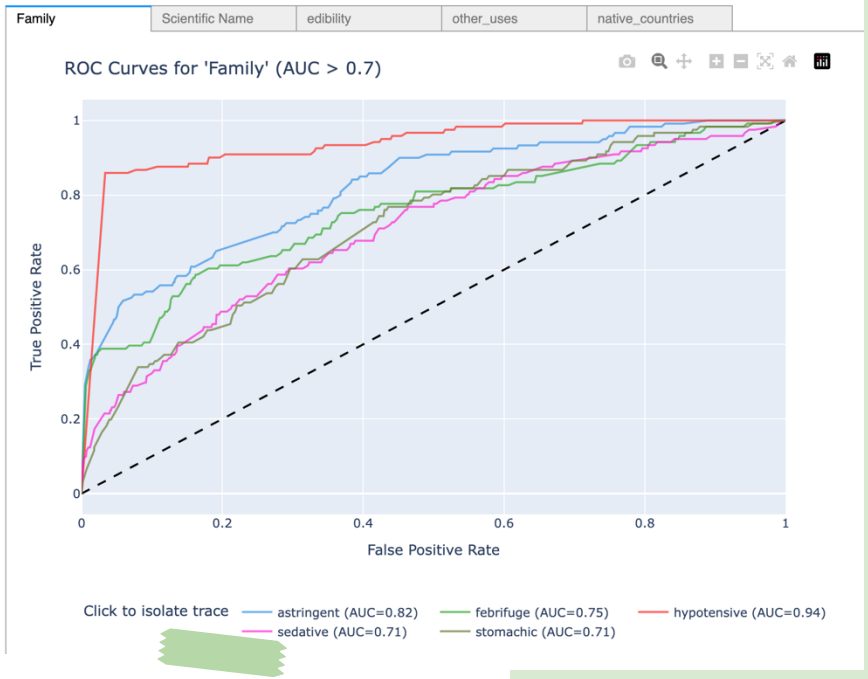


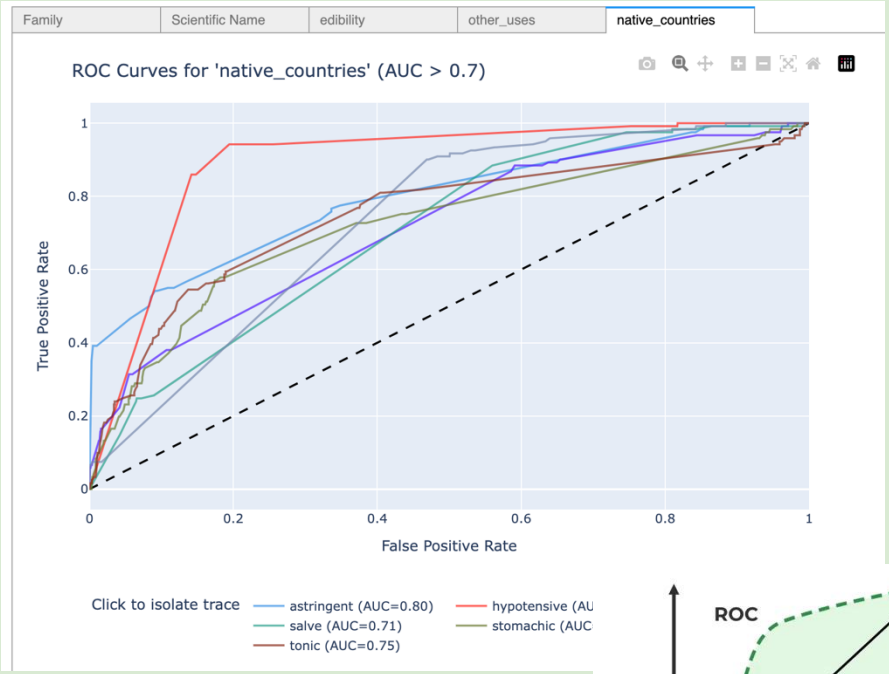
Classifying medicinal properties based on correlated features



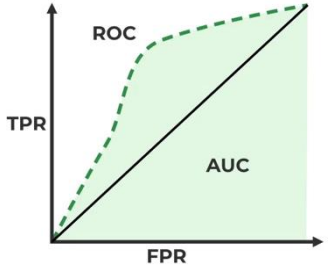


Family and Location perform the best

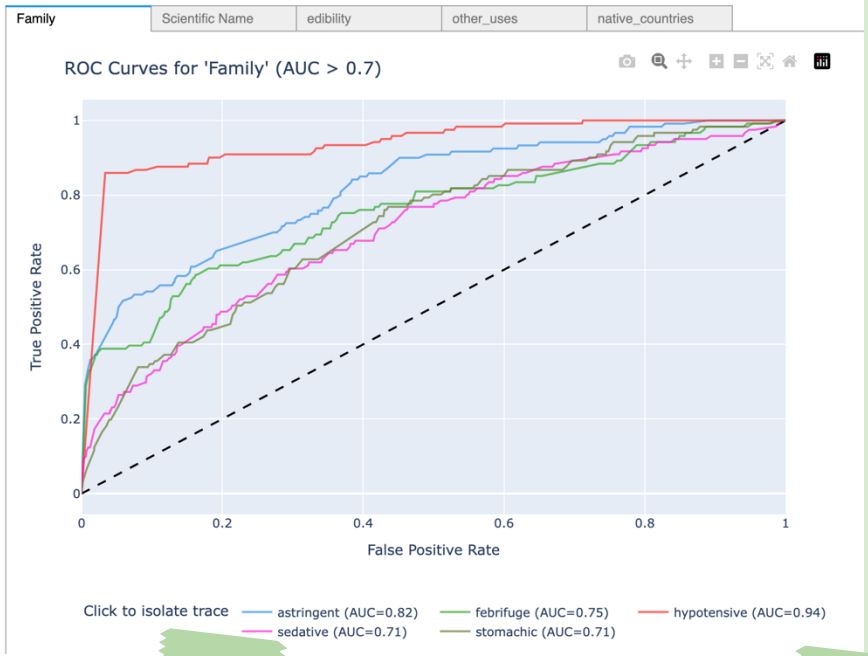


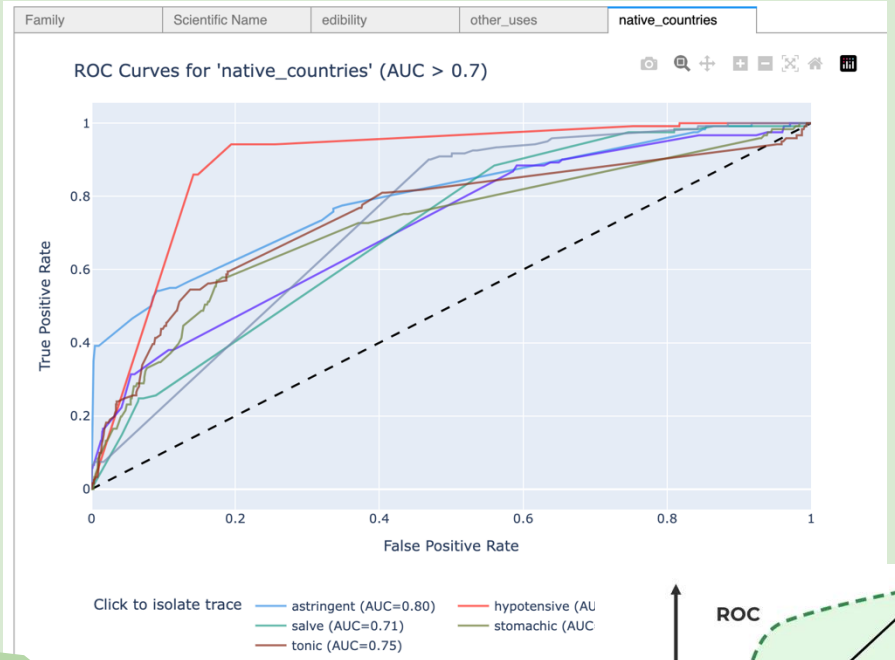


Family can classify hypotensive plants with 94% accuracy

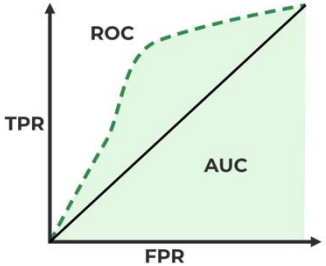


Family and Location perform the best

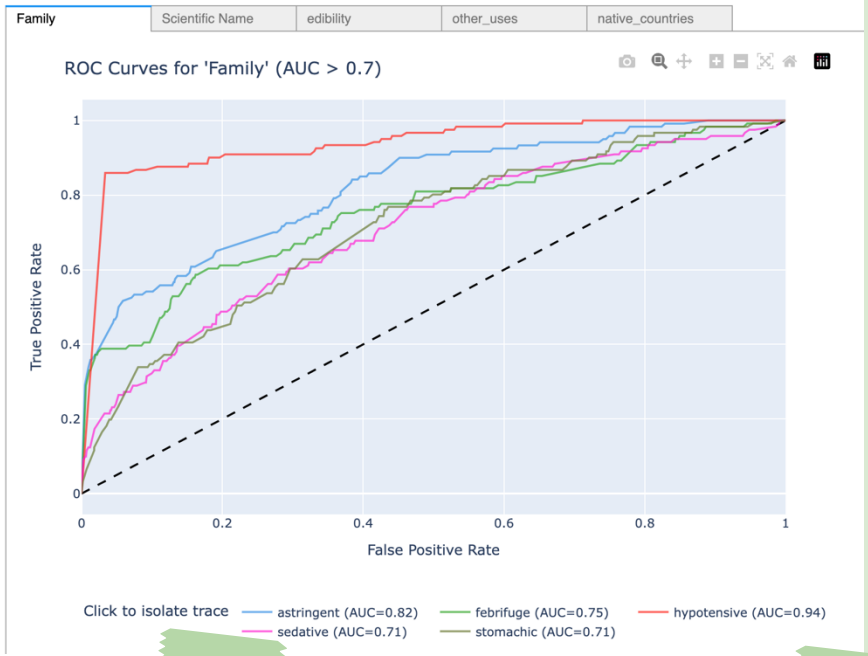


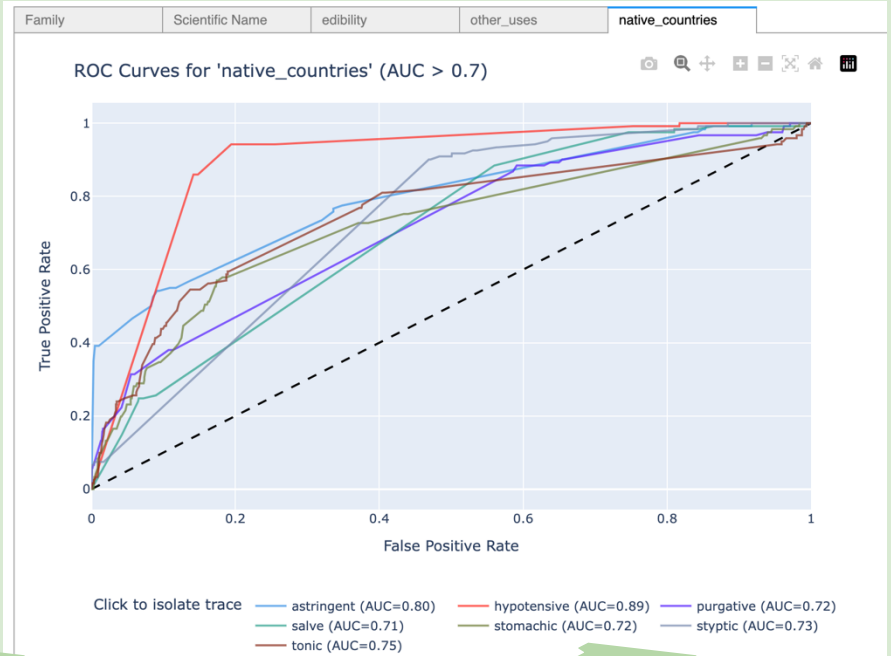


- Family can classify hypotensive plants with 94% accuracy
- 'Hypotensive' is the best classified property across all



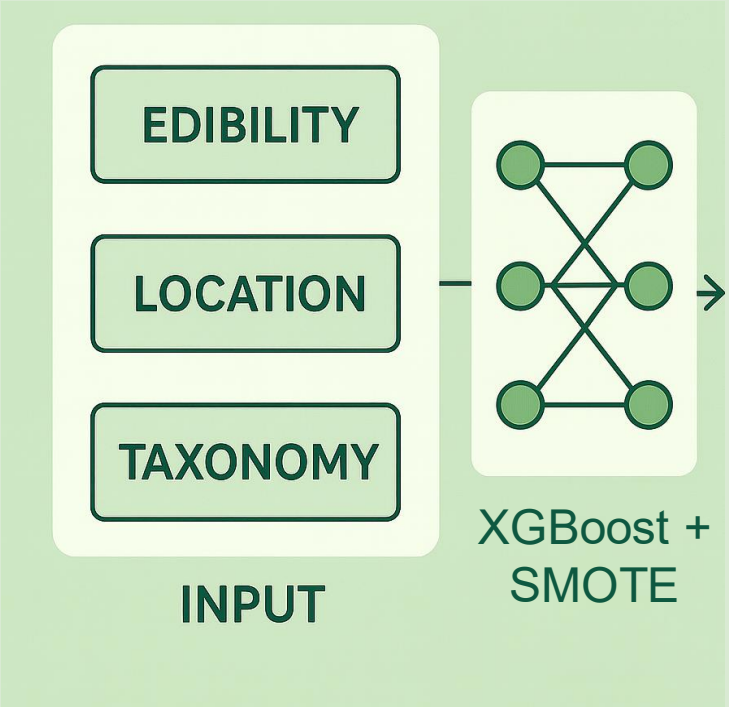
Family and Location perform the best

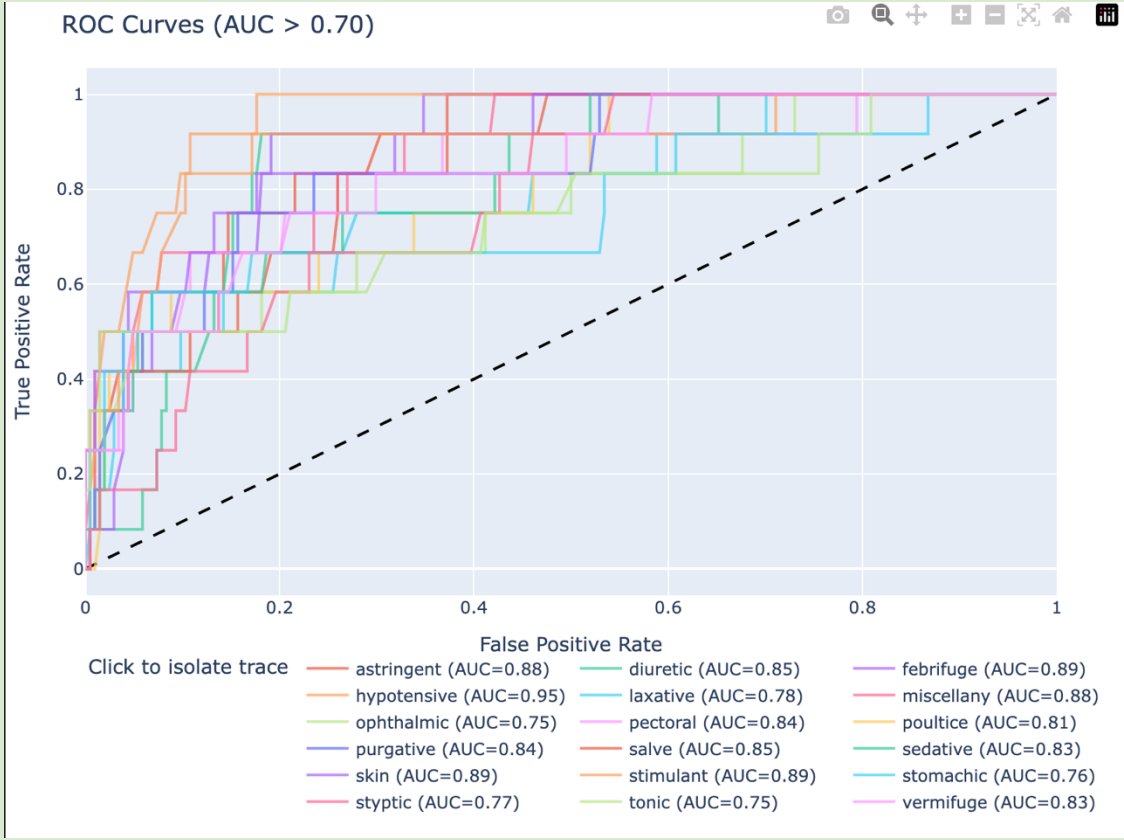




- Family can classify hypotensive plants with 94% accuracy
- 'Hypotensive' is the best classified property across all
- Location performs close to taxonomy

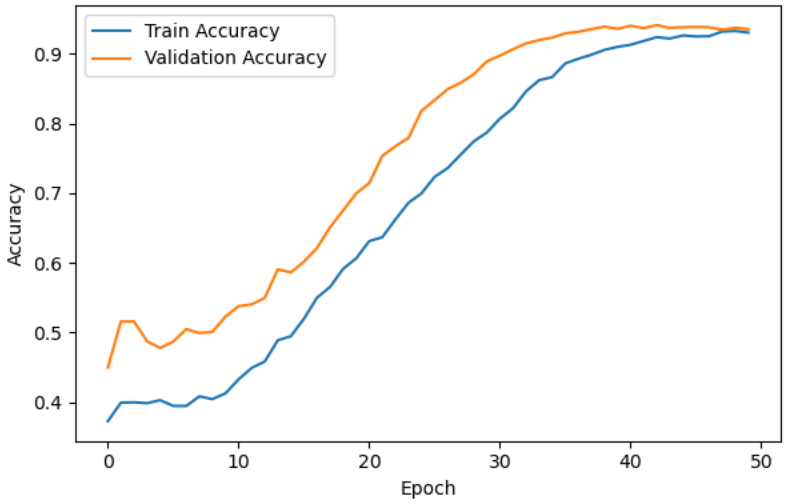
Performance of an “all features” classifier

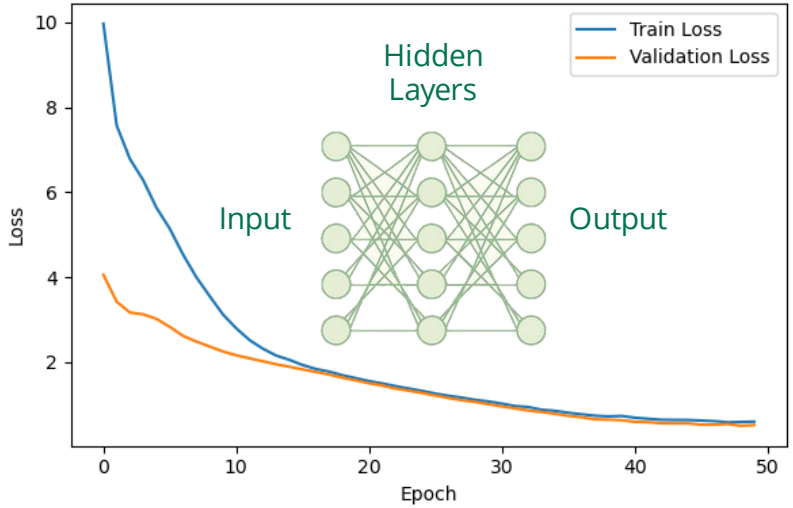




Could a NN model perform better?

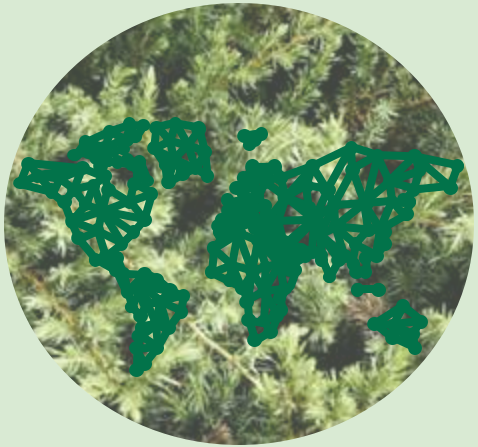
- **Fast initial learning:** Training loss plummets ~80% in the first 10 epochs, and accuracy jumps from ~40% to ~60%.
- **Steady convergence:** Both training and validation curves rise smoothly from epoch 10–50, ending near 95% accuracy.
- **Low overfitting:** Validation metrics track the training curves closely, with no widening gap.
- **Diminishing returns:** After ~epoch 35, gains flatten out—model performance plateaus.
- **Good generalization:** Validation accuracy briefly exceeds training early on and stays on par throughout.





So what did we learn?

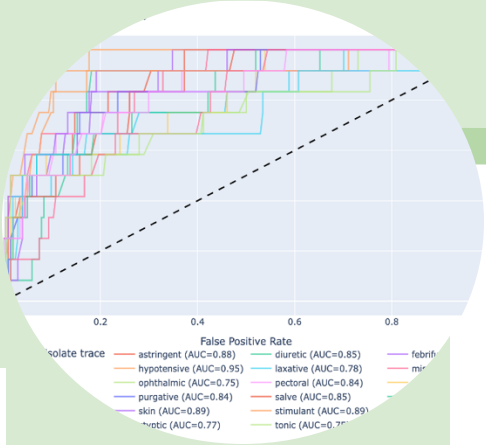
Conclusions



Location and family correlates with medicinal property

- Plants with certain medicinal trait correlate with specific climate conditions

Conclusions



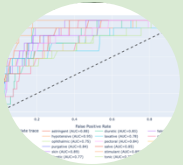
Location & Family

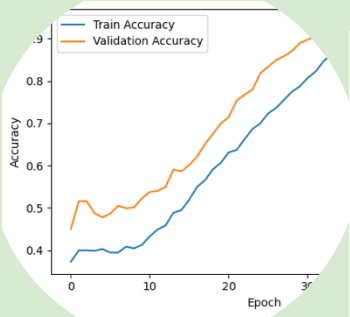
relates to
medicinal property

Mathematical models resulted in overfitting

- When all features are combined, the best model indicates an poor generalization
- A bigger data set is needed for further training.

Conclusions





Location & Family

relates to
medicinal property

XGBoost

overfits bigger
data set needed

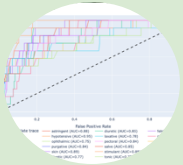
NN model can identify nonlinear patterns in property subgroups

As expected real-life/categorical data sets have complex patterns which require a more adaptable classifier

Conclusions

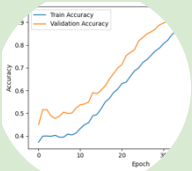
Location & Family

relates to
medicinal property



XGBoost

overfits bigger
data set needed



NN model

captures complex
patterns

OUTLOOK

- **Feature engineering**: identify NN layers that better classify
- **Data collection:** a larger more balanced data set

Thank you for listening!

A huge thank you to:
Filipa Peleja,
Oumaima,
DSPT19,
CodeOp,
& Plants for a Future!

contact us @ Slack!