Predicting Medical Symptom Categories of Plants Using Supervised Machine Learning

A comparative Study of KNN and Logistic Regression for Medicinal Plant Symptom Classification

The classification of medicinal plants according to their therapeutic properties is a challenging task due to the diversity and complexity of symptom descriptions found in real-world botanical datasets. This project addresses a supervised multiclass classification problem aimed at predicting medicinal symptom categories of plants based on structured numerical and categorical attributes. A real-world dataset containing plant characteristics such as medicinal rating, edibility rating, weed potential, and climatic tolerance was used. Due to the high granularity of the original symptom labels, these were grouped into a smaller number of broader classes to make the learning task tractable. Extensive preprocessing was performed, including exploratory data analysis, missing value handling, categorical encoding, feature scaling, and class imbalance analysis. Two supervised machine learning models were implemented and compared: Logistic Regression as a baseline linear classifier and K-Nearest Neighbors (KNN) as a non-parametric, distance-based model. Hyperparameter optimization using GridSearchCV was applied to both models. The experimental results show that KNN significantly outperforms Logistic Regression, highlighting the non-linear structure of the data and the importance of local similarity in symptom classification. These findings emphasize the need to select machine learning models that align with the underlying characteristics of real-world datasets.

Table des matières

Introduction

Medicinal plants are widely used in traditional and modern medicine to treat a broad range of health conditions. With the growing availability of digital plant databases, there is increasing interest in applying machine learning techniques to better organize and analyze medicinal plant information. However, data related to medicinal plants is often complex and difficult to exploit automatically, as it typically includes heterogeneous attributes, missing values, and a large variety of symptom descriptions.

At the beginning of this project, the initial objective was to work within the healthcare domain by building a predictive model capable of recommending medicinal plants corresponding to specific symptoms. During an extensive search for suitable datasets, several limitations were encountered. Many available datasets were incomplete, still under development, poorly structured, or accessible only through paid services. As a result, a publicly available dataset derived from the Plants For A Future (PFAF) database was selected from Kaggle. This dataset provides structured information describing plant characteristics along with their associated medicinal uses.

Although the selected dataset does not directly allow the prediction of plants from symptoms, it enables the formulation of a related and meaningful machine learning problem. The task was therefore reformulated as a supervised multiclass classification problem, where the objective is to predict medicinal symptom categories based on plant characteristics such as medicinal rating, edibility rating, weed potential, and climatic tolerance. The original symptom labels were highly granular and unevenly distributed, making direct classification impractical. To address this issue, the symptom labels were grouped into a smaller number of broader categories, resulting in a more tractable learning problem.

This problem is of particular interest because it reflects realistic constraints commonly encountered in applied machine learning projects. The dataset is imperfect, contains both numerical and categorical features, and exhibits class imbalance and non-linear relationships. Addressing these challenges requires careful data preprocessing, appropriate model selection, and robust evaluation methods. In this project, the problem was addressed by implementing a complete supervised machine learning pipeline and comparing two classification approaches: Logistic Regression as a baseline linear model and K-Nearest Neighbors as a non-parametric, distance-based model.

Aim

The aim of this project is to develop and evaluate supervised machine learning models capable of predicting medicinal symptom categories of plants based on their characteristics, and to identify the most suitable model for this classification task.

Objectives

The main objectives of this project are:

1. To explore and understand a real-world medicinal plant dataset using exploratory data analysis.
2. To preprocess the dataset by handling missing values and preparing numerical and categorical features for machine learning.
3. To reduce the complexity of the original symptom labels by grouping them into a smaller number of meaningful classes.
4. To implement and evaluate two supervised classification models: Logistic Regression and K-Nearest Neighbors.
5. To apply hyperparameter tuning and appropriate evaluation metrics to compare model performance.
6. To analyze the results and determine which model is best suited to the problem.

Scope of the work

This study focuses exclusively on supervised classification techniques applied to a single medicinal plant dataset. The scope is limited to predicting predefined symptom categories and does not aim to provide medical recommendations or discover new medicinal properties. Only classical machine learning models available in the scikit-learn library are considered, and deep learning approaches are outside the scope of this work. The analysis is based on structured tabular data and does not include text mining or image-based features.

Related works

The use of computational approaches for analysing medicinal plant data has increased in recent years, mainly due to the availability of ethnobotanical databases and open-access biological datasets. Previous work related to this project can be grouped into three main categories: ethnobotanical databases documenting medicinal plant uses, exploratory and descriptive analyses of medicinal plant datasets, and the application of classical supervised machine learning methods to structured biological data.

Several well-established ethnobotanical databases aim to document and preserve traditional medicinal knowledge. The Prelude Medicinal Plants Database, developed by the Belgian Biodiversity Platform and available through GBIF, aggregates information from a large number of scientific publications and focuses on traditional human and veterinary medicinal uses of plants, particularly in Sub-Saharan Africa. Similarly, Dr. Duke's Phytochemical and Ethnobotanical Databases, maintained by the United States Department of Agriculture, provide extensive information on plant species, phytochemical compounds, biological activities, and

traditional uses. These resources are widely used in ethnobotanical and pharmacological research.

Although these databases contain rich and valuable information, they are primarily designed for documentation and exploration rather than prediction. Medicinal uses are typically represented as textual descriptions or categorical annotations, and no machine learning models are applied to predict medicinal properties or symptom categories. As a result, these databases serve as important data sources but do not address supervised classification or model evaluation.

Other related studies focus on organizing, cleaning, and visualizing ethnobotanical data. For example, the Ewé database presented by Souza et al. provides a web-based platform for storing and visualizing medicinal plant data from Brazil. The emphasis of this work is on data integration, visualization, and hypothesis generation rather than on predictive modeling. While such studies demonstrate the importance of structured data and data preprocessing, they do not evaluate supervised machine learning models or compare classification performance.

More closely related to this project are exploratory analyses conducted on the Pfaf medicinal plants dataset. Several publicly available studies and notebooks, particularly on Kaggle, have explored this dataset through descriptive statistics and visualizations, examining relationships between medicinal ratings, edibility ratings, and other plant attributes. Some works apply regression models to predict numerical medicinal ratings, showing that meaningful relationships exist between plant characteristics. However, these studies generally focus on regression or exploratory analysis rather than multiclass classification of medicinal symptom categories.

In the broader machine learning literature, Logistic Regression and K-Nearest Neighbors are commonly used for classification tasks involving structured tabular data. Logistic Regression is often employed as a baseline model due to its simplicity and interpretability, but it is known to perform poorly when the underlying relationships between features and target variables are non-linear. In contrast, K-Nearest Neighbors is a non-parametric, distance-based method that can capture local and non-linear patterns, provided that appropriate feature scaling and parameter tuning are applied. Previous studies emphasize the importance of preprocessing steps such as encoding categorical variables, scaling numerical features, and addressing class imbalance to ensure reliable model performance.

Table 1 summarizes the main online resources and datasets related to medicinal plant data that are relevant to this study. The table provides a brief description of each resource along with its corresponding URL, highlighting their scope and purpose.

This summary helps position the present project within the existing landscape of ethnobotanical data sources. While these resources offer rich documentation and exploratory access to medicinal plant information, they do not apply supervised machine learning techniques to predict symptom categories. In contrast, the current project builds upon such data sources by framing the problem as a supervised multiclass classification task and by evaluating predictive models on structured plant attributes.

Source/Database	Description	URL
Pfaf Medical Plants Use Dataset	Open-access dataset containing medicinal and edible plant information, including ratings, symptom keywords, climatic tolerance, and weed potential. Used as the primary dataset in this project.	https://www.kaggle.com/datasets/edwardgaibor/pfaf-medical-plants-use-dataset
Prelude Medicinal Plants Database	Ethnobotanical database summarizing medicinal plant uses from scientific literature, with a focus on traditional human and veterinary medicine in Africa.	https://www.gbif.org/dataset/49c5b4ac-e3bf-401b-94b1-c94a2ad5c8d6
Dr.Duke's Phytochemical and Ethnobotanical Databases	Public database providing information on plant species, phytochemicals, biological activities, and traditional medicinal uses.	https://phytochem.nal.usda.gov

Ewé Ethnobotanical Database	Web-based ethnobotanical database designed for storing, visualizing, and analyzing traditional medicinal plant use data.	https://www.ewedb.com

Methods

Dataset Overview

This project uses a real-world dataset related to medicinal plants collected from an open botanical database. The dataset contains numerous plant records described by a combination of numerical and categorical attributes. Numerical features include ratings related to medicinal usefulness and edibility, while categorical features describe properties such as weed potential and climatic tolerance.

The target variable corresponds to medicinal symptom information associated with each plant. Due to the high granularity of the original symptom labels, these were grouped into a smaller number of broader classes, allowing the problem to be formulated as a multiclass classification task.

Machine learning approaches

Two supervised learning algorithms were selected to address the classification problem.

Logistic Regression was chosen as a baseline classifier. As a linear and interpretable model, it provides a reference point for assessing whether the dataset can be effectively modeled using linear decision boundaries.

K-Nearest Neighbors (KNN) was selected as a non-parametric, distance-based algorithm. Unlike Logistic Regression, KNN makes no assumptions about the underlying data distribution and classifies observations based on similarity to neighboring samples. This makes it well-suited for capturing non-linear relationships and local patterns within the data.

The use of these two contrasting models enables a meaningful comparison between a simple linear approach and a more flexible distance-based method.

Experimental Framework and Tools

All experiments were conducted using the Python programming language. Data preprocessing, model training, and evaluation were performed using the scikit-learn library, with pandas and NumPy used for data manipulation. Visualization tools were employed during exploratory data analysis.

The dataset was split into training and testing subsets to evaluate model generalization. Model training and hyperparameter tuning were performed exclusively on the training data, while the test set was reserved for final evaluation. Hyperparameter optimization was conducted using cross-validation to ensure robust model selection.

Evaluation Metrics

Model performance was evaluated using standard classification metrics. Accuracy was used to measure overall predictive performance, but due to the multiclass setting and class imbalance, additional metrics were required.

Precision, recall, and F1-score were used to assess performance across all classes. Particular emphasis was placed on the macro-averaged F1-score, which assigns equal importance to each class regardless of frequency. Confusion matrices were also used to support the interpretation of classification errors.

Methodological Summary

In summary, this project follows a structured supervised learning methodology involving dataset exploration, preprocessing, model comparison, hyperparameter tuning, and performance evaluation. The chosen methods and metrics provide a coherent and practical framework for assessing the suitability of different machine learning approaches for multiclass classification of medicinal plant data.

DATASET PREPARATION

Dataset Description and Selection

The raw dataset contains **17,950 observations and 27 attributes**, describing medicinal plants and their associated properties. It is a real-world dataset combining heterogeneous data types,

including numerical ratings, categorical attributes, and textual descriptions, making it suitable for demonstrating practical machine learning techniques on imperfect data.

The dataset includes descriptive text fields (such as summaries, medicinal properties, and care requirements), identification-related information (Latin names, URLs, and image references), numerical rating values (including edibility and medicinal ratings), and practical attributes such as **USDA hardiness** and **weed potential**. A central therapeutic keyword (*use keyword*) is provided for each plant and represents the medicinal symptom category associated with the plant.

According to the info() output, **22 out of the 27 variables are of type object**, corresponding to textual or categorical data, while **five variables are numerical ratings**. Due to the presence of long textual descriptions, duplicated attributes, and identifier fields, not all variables are suitable for supervised learning. A feature selection step was therefore required to retain only the most relevant attributes for the classification task.

Dataset Variables Overview

The dataset contains 27 variables describing medicinal plants. These variables can be grouped according to their data type and relevance for the machine learning task.

Target variable

- **use_keyword**
  Type: categorical (text)
  Description: therapeutic use or medicinal symptom associated with the plant
  Usage: **used as target variable** (after grouping into broader classes)

Numerical variables

- **edibility_rating_search**
  Type: integer (0–5)
  Description: edibility score of the plant
  Usage: **used** (informative numerical feature)
- **medicinal_rating_search**
  Type: integer (0–5)
  Description: medicinal usefulness score
  Usage: **used** (strongly related to therapeutic properties)

- **Edibility Rating**
  Type: float
  Description: alternative edibility score
  Usage: **not used** (duplicate information)

- **Medicinal Rating**
  Type: float
  Description: alternative medicinal score
  Usage: **not used** (duplicate of rating_search variable)

- **Other Uses Rating**
  Type: float
  Description: score related to non-medicinal uses
  Usage: **not used** (not directly relevant to symptom prediction)

Categorical variables used for modeling

- **USDA hardiness**
  Type: categorical (text)
  Description: climatic tolerance of the plant
  Usage: **used** (captures ecological constraints)

- **Weed Potential**
  Type: categorical (text: Yes / No)
  Description: invasive or weed-like behavior
  Usage: **used** (may correlate with toxicity or potency)

Textual and identifier variables (not used)

- **latin_name_search**, **Scientific Name**
  Type: text
  Description: scientific plant identifiers
  Usage: **not used** (identifiers only)

- **common_name_search**, **Common Name**, **Common Names**
  Type: text
  Description: plant names
  Usage: **not used** (high cardinality, low predictive value)

- **plant_url**, **Image URLs**
  Type: text
  Description: external references
  Usage: **not used**

- **Care Requirements**, **Cultivation Details**, **Propagation**
  Type: text
  Description: cultivation instructions
  Usage: **not used** (free text)

- **Medicinal Properties**, **Edible Uses**, **Other Uses**, **Special Uses**, **Summary**
  Type: text
  Description: long descriptive fields
  Usage: **not used** (unstructured text, outside scope)

- **Native Range**, **Range**, **Family**
  Type: categorical / text
  Description: botanical and geographic information
  Usage: **not used** (too granular, not directly linked to target)

Target Variable Construction

The original target variable (use_keyword) contains approximately 120 distinct classes, each corresponding to a specific medicinal symptom or therapeutic use. Treating each symptom as an independent target class would result in a highly fragmented classification problem, with many categories being severely underrepresented.

To make the classification task tractable, the symptom labels were grouped into three broader categories according to their general level of medicinal severity. This transformation reduces target sparsity while preserving meaningful distinctions between symptom types. The impact of this transformation and the resulting class distribution are further examined in the exploratory data analysis section.

Exploratory Data Analysis

Exploratory Data Analysis was performed to understand the structure of the dataset and identify potential challenges for supervised classification.

The original target variable (use_keyword) contains around **120 distinct symptom categories**, with a highly imbalanced distribution dominated by a few frequent classes.

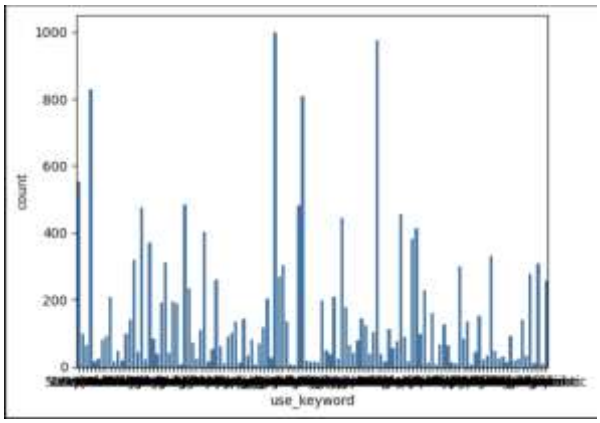

Fig. 1. Distribution of original medicinal symptom categories

This fragmentation made direct classification ineffective. After grouping symptoms into **three broader classes** based on their general medicinal severity, the target distribution became more manageable, although some imbalance remains.

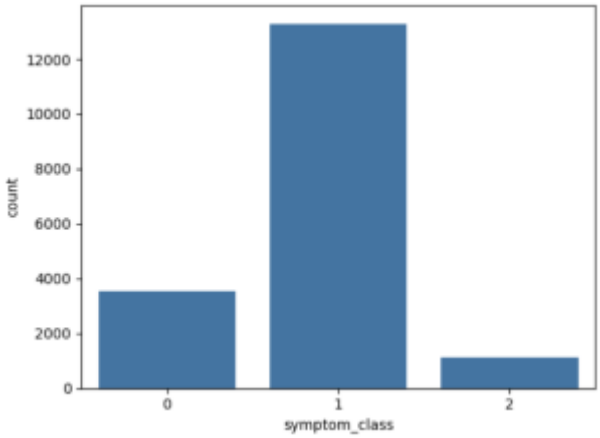

Fig. 2. Distribution of target classes after mapping

The numerical features *medicinal rating* and *edibility rating* are bounded between **0 and 5** and show **skewed distributions**, with most values concentrated at low to medium levels.

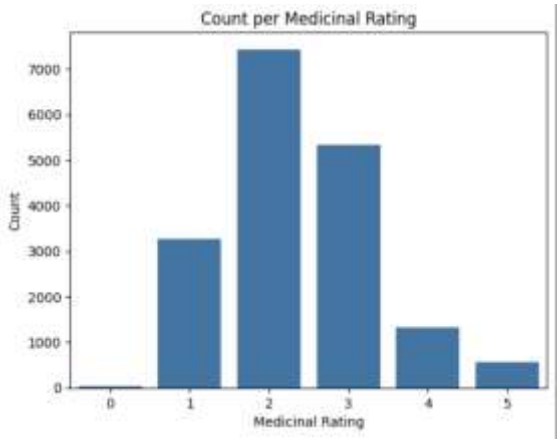

Fig. 3. Distribution of medicinal rating values

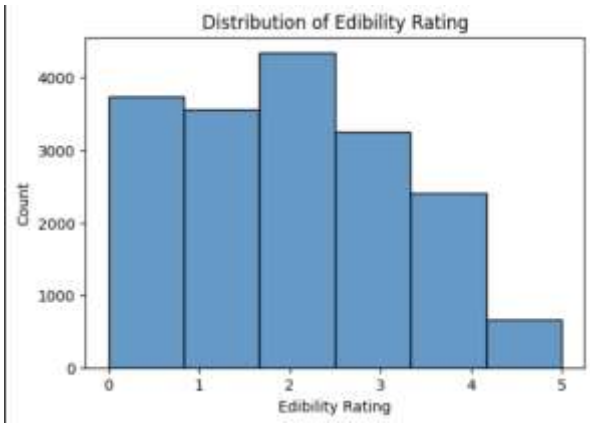

Fig. 4. Distribution of edibility rating values

Categorical features also exhibit imbalance. *Weed Potential* is dominated by the **"No"** category, while *USDA Hardiness* shows a long-tailed distribution with a few dominant climatic zones.

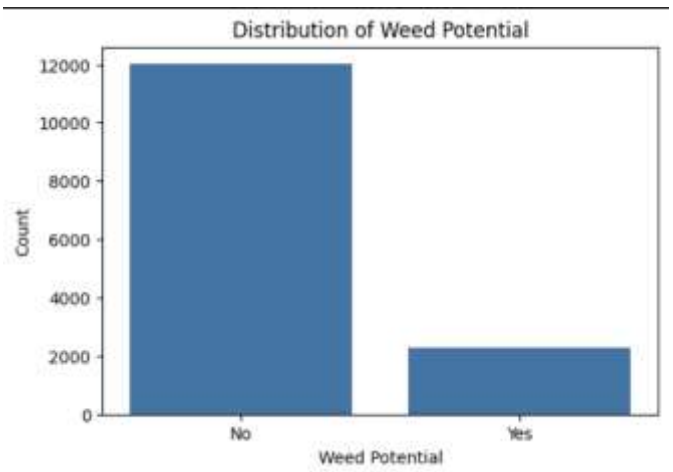

Fig. 5. Distribution of Weed Potential values

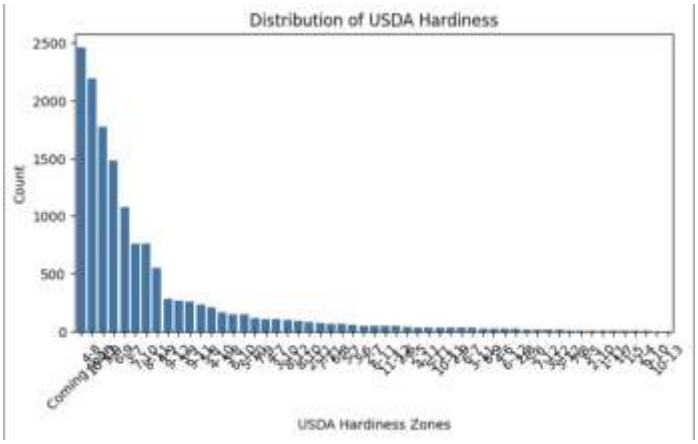

Fig. 6. Distribution of USDA Hardiness zones

Overall, the dataset is heterogeneous, imbalanced, and non-linear in nature. These observations motivated the use of target mapping, feature scaling, appropriate encoding strategies, and balanced evaluation metrics in the subsequent stages of the project.

Data Cleaning and Formatting

Data cleaning was performed to prepare the dataset for supervised learning by improving consistency and ensuring that only relevant information was retained. The dataset was first reduced to a subset of meaningful features, including the target variable, numerical ratings, and selected categorical attributes, while identifiers and textual descriptions not suitable for modeling were excluded.

Missing and ambiguous values in categorical variables were handled by introducing a unified "unknown" category. This approach preserves all observations while ensuring consistent representation of missing information. In addition, categorical text values were normalized by applying lowercase conversion and trimming whitespace, preventing the creation of artificial categories caused by formatting inconsistencies.

Column names were then standardized to improve readability and maintain consistency throughout the analysis. Final validation checks were conducted to confirm the absence of missing values, verify data types, and ensure that categorical values and target labels were properly formatted before proceeding to feature encoding and dataset splitting.

Train-Test Split

After data cleaning and formatting, the dataset was divided into training and testing subsets. A stratified split was applied in order to preserve the distribution of the target classes across both subsets. The training set was used exclusively for model training and preprocessing, while the test set was reserved for final performance evaluation.

Feature Preparation and Encoding

After the train–test split, the input features were prepared for model training. Categorical variables were converted into a numerical format using one-hot encoding, while numerical features were kept in their original form. To address the class imbalance observed in the target variable, random oversampling was applied only to the training data. Feature scaling was then performed using standardization to ensure that distance-based models, such as K-Nearest Neighbors, were not influenced by differences in feature magnitude. All preprocessing

operations were fitted exclusively on the training data and subsequently applied to the test data to prevent data leakage.

MODEL IMPLEMENTATION

Overview of Implement Models

To address the multiclass classification problem, two supervised machine learning models were implemented and compared. Logistic Regression was used as a baseline linear classifier, while K-Nearest Neighbors (KNN) was selected as a non-parametric, distance-based model. These two approaches were chosen because they rely on fundamentally different learning assumptions, allowing an evaluation of how model complexity and non-linearity affect predictive performance on structured tabular data.

Logistic Regression Model

The Logistic Regression model was first implemented as a baseline classifier. Despite class weighting and hyperparameter tuning, the model achieved limited performance, with an accuracy of approximately 31% and a macro-averaged F1-score around 0.26.

The model performs reasonably on the majority class but struggles to correctly classify minority classes, particularly the most severe symptom category. This behaviour indicates that the linear decision boundaries imposed by Logistic Regression are insufficient to capture the complex and non-linear relationships present in the data.

Hyperparameter tuning using GridSearchCV resulted in only marginal improvements, suggesting that the observed limitations are inherent to the model rather than caused by suboptimal parameter selection. These results motivate the use of a more flexible, non-linear model in the next stage of the analysis.

K-Nearest Neighbors Model (KNN)

The K-Nearest Neighbors (KNN) classifier was implemented as a non-parametric, distance-based model. Since KNN relies on distance computations, feature scaling was applied prior to training to ensure that all input variables contribute equally to the distance measure.

An initial KNN model trained with default hyperparameters already showed a clear improvement over Logistic Regression, achieving an accuracy close to 65% on the test set, compared to approximately 30% for the linear model. This substantial performance gain

confirms that the dataset exhibits strong non-linear and local patterns that cannot be captured by a linear decision boundary. However, class-level performance remained uneven, with the minority class being particularly difficult to predict.

To further improve performance, hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation on the training set. The optimization explored different numbers of neighbors, distance metrics, and weighting strategies, using the macro-averaged F1-score as the selection criterion to explicitly account for class imbalance.

The optimized KNN model achieved a macro-averaged F1-score of approximately 0.32, compared to around 0.26 for Logistic Regression, while maintaining an overall accuracy of about 64%. The dominant class was predicted with high recall (above 80%), whereas the rarest class remained challenging to classify. These results confirm the suitability of KNN for this dataset while also highlighting the limitations imposed by highly imbalanced multiclass data.

Hyperparameter Optimization

Hyperparameter optimization was applied to improve model performance and ensure a fair comparison between the selected algorithms. GridSearchCV was used to systematically explore combinations of hyperparameters using 5-fold cross-validation on the training data only, in order to prevent any information leakage from the test set.

For the K-Nearest Neighbors model, the optimization focused on the number of neighbors, the distance metric, and the weighting strategy. The best-performing configuration corresponded to a KNN model using three neighbors, a uniform weighting scheme, and the Euclidean distance metric. This configuration achieved a cross-validated macro-averaged F1-score of approximately 0.33, indicating improved balance across classes compared to the baseline model.

For Logistic Regression, hyperparameter tuning primarily targeted the regularization strength. Although slight improvements were observed, overall performance remained limited, confirming that the linear nature of the model is not well suited to the non-linear structure of the dataset.

The use of cross-validation during hyperparameter tuning ensures that the selected configurations generalize beyond a single train–test split and reduces the risk of overfitting

Model Outputs

Both models generate predicted class labels for unseen observations, which are subsequently used for performance evaluation. Logistic Regression also produces class probability estimates, providing additional insight into prediction confidence. In contrast, K-Nearest Neighbors predictions are based on a voting mechanism among the nearest observations in the feature space, reflecting local similarity patterns in the data.

After training and hyperparameter optimization, the final versions of both models were applied to the test dataset to produce predictions. These outputs serve as the basis for the quantitative evaluation and comparative analysis presented in the following section.

MODEL VALIDATION

After implementing and optimizing the selected models, their performance was evaluated using the held-out test dataset in order to assess generalization to unseen data and identify the most suitable approach for the classification task.

Logistic Regression achieved limited performance, with an accuracy of approximately 30% and a macro-averaged F1-score around 0.26. These results indicate that the model struggles to correctly separate the symptom classes, particularly the minority class, despite the use of class weighting. This suggests that linear decision boundaries are insufficient to capture the complexity of the dataset.

In contrast, the K-Nearest Neighbors model demonstrated substantially better performance. After hyperparameter optimization, KNN achieved an accuracy of approximately 64% and a macro-averaged F1-score close to 0.32. The dominant class was predicted with high recall (above 80%), confirming that KNN effectively captures local and non-linear patterns in the feature space. However, performance on the rarest class remained weak, reflecting the strong class imbalance present in the dataset.

Based on this comparative evaluation, K-Nearest Neighbors was selected as the most suitable model for this classification task. Its superior performance across multiple evaluation metrics demonstrates its ability to better model the structure and characteristics of the data, while Logistic Regression serves as a useful baseline for comparison.

ANALYSIS & RECOMMENDATIONS

Analysis of Expermimental Results

The experimental results highlight clear differences between the two implemented models, consistent with their theoretical characteristics and the structure of the dataset.

Logistic Regression exhibits underwhelming performance due to its reliance on linear decision boundaries and global relationships between features and the target variable. Given the heterogeneous nature of the data and the presence of complex interactions between numerical and categorical features, these assumptions are not satisfied, leading to underfitting and poor discrimination between symptom classes.

In contrast, the K-Nearest Neighbors model demonstrates stronger performance by leveraging local similarities between plants. Observations sharing similar medicinal ratings, edibility scores, and categorical characteristics tend to belong to the same symptom class, which favors a distance-based classification approach. Feature scaling played a crucial role in this performance by ensuring that distance computations were not dominated by features with larger numeric ranges.

Expected Outcomes and Observed Anomalies

The superior performance of KNN over Logistic Regression was expected due to the non-linear and fragmented nature of the dataset. However, an important limitation was observed in the classification of the rarest symptom class, which remained difficult to predict even after hyperparameter optimization.

This outcome highlights the strong impact of class imbalance and overlapping feature distributions. It also illustrates that higher overall accuracy does not necessarily imply better performance across all classes. In particular, Logistic Regression occasionally achieved moderate accuracy while performing poorly on minority classes, reinforcing the importance of balanced evaluation metrics such as the macro-averaged F1-score in multiclass classification problems.

Comparison with Related Work

The findings of this project are consistent with results reported in related studies on medicinal plant classification. Previous research indicates that linear models often struggle with complex, non-linear datasets containing mixed data types, whereas distance-based or non-parametric models tend to perform better in such settings. The improved performance of KNN observed in this study aligns with these findings and supports the use of flexible models for heterogeneous biological datasets.

Unlike approaches relying on highly specialized features or domain-specific transformations, this project demonstrates that competitive performance can be achieved using a limited set of structured features combined with appropriate preprocessing and model selection, reinforcing the practical relevance of the proposed methodology.

Recommendations

Based on the experimental results, several recommendations can be made. For datasets exhibiting non-linear structures and heterogeneous feature types, non-parametric models such as KNN should be preferred over simple linear classifiers. Careful preprocessing, including feature scaling and explicit handling of class imbalance, is essential to ensure reliable performance. Additionally, systematic hyperparameter tuning is critical, as model performance is highly sensitive to parameter selection.

Future work could explore ensemble-based methods or alternative imbalance-handling strategies to further improve robustness. Incorporating additional structured features or domain-specific knowledge may also help improve the classification of rare symptom categories.

Conclusion

This project explored the use of supervised machine learning to classify medicinal plant symptom categories based on plant characteristics. A real-world dataset containing numerical and categorical attributes was used, requiring extensive preprocessing due to missing values, duplicated fields, and class imbalance.

Two models were implemented and compared. Logistic Regression was used as a baseline model but showed limited performance, indicating that linear decision boundaries were not sufficient for this task. In contrast, the K-Nearest Neighbors model achieved significantly better results after feature scaling and hyperparameter tuning, demonstrating its ability to capture non-linear and local patterns in the data.

The results highlight the importance of selecting models that are well suited to the structure of the dataset and of using appropriate evaluation metrics in imbalanced multiclass problems. Although KNN outperformed Logistic Regression, the classification of minority classes remained challenging, showing the limitations of the current approach.

Overall, this project provided valuable experience in building a complete machine learning pipeline on real-world data, from preprocessing to model evaluation. Future work could explore more advanced models or alternative techniques to better handle class imbalance and further improve classification performance.