IRIS FLOWER CLASSIFICATION USING MACHINE LEARNING

DEEP LEARNING AND APPLICATIONS-UEC642

Project

Submitted by : Sameer Pal
Roll no : 102215115

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY,

PATIALA, PUNJAB

IRIS FLOWER CLASSIFICATION USING MACHINE LEARNING

1. Abstract

This project presents the design and implementation of a machine learning based classification system for identifying Iris flower species using supervised learning techniques. The Iris dataset, which is a well-known benchmark dataset in the field of machine learning, is used for this purpose. The system takes four numerical input features related to the physical dimensions of the flower and predicts the corresponding species. Logistic Regression is employed as the classification algorithm due to its simplicity, efficiency, and interpretability. The project demonstrates the complete machine learning workflow including data acquisition, model training, testing, evaluation, and result analysis.

2. Problem Statement

In the field of pattern recognition and data analysis, classification of objects based on measurable attributes is a fundamental problem. The task of this project is to develop a machine learning model that can accurately classify Iris flowers into one of the three species — Setosa, Versicolor, or Virginica — based on four given physical attributes. The challenge lies in learning the decision boundaries between different species using labeled data and ensuring that the model generalizes well to unseen samples.

3. Objectives

The main objectives of this project are: - To understand the concept of supervised machine learning and classification - To study the working of Logistic Regression for multiclass problems - To implement a complete machine learning pipeline using Python - To evaluate the performance of the model using standard metrics - To gain hands-on experience with real-world datasets and ML libraries

4. Dataset Description

The Iris dataset consists of 150 instances of Iris flowers collected from three different species. Each instance is described by four numerical attributes measured in centimeters. The dataset is evenly distributed, with 50 samples per class, which makes it suitable for unbiased model training.

Features: - Sepal Length (cm) - Sepal Width (cm) - Petal Length (cm) - Petal Width (cm)

Target Variable: - Iris Species (Setosa, Versicolor, Virginica)

The dataset does not contain missing values or noise, eliminating the need for complex preprocessing steps. In this project, the dataset is loaded directly from the `scikit-learn` library.

5. Methodology

The methodology followed in this project is systematic and follows standard machine learning practices:

1. Data Loading: The Iris dataset is imported using the scikit-learn library.
2. Data Splitting: The dataset is divided into training and testing sets using an 80:20 ratio.
3. Model Selection: Logistic Regression is chosen for classification.
4. Model Training: The classifier is trained using the training dataset.
5. Prediction: The trained model predicts species labels for the test dataset.
6. Evaluation: Model performance is evaluated using accuracy score and confusion matrix.

This step-by-step approach ensures clarity, reproducibility, and reliable evaluation of the model.

6. Tools & Technologies

- Python 3
- scikit-learn
- pandas
- matplotlib (for visualization)

7. Implementation (Code)

```
# Iris Flower Classification Project

import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
ConfusionMatrixDisplay
```

```
# Load dataset
iris = load_iris()
X = iris.data y
= iris.target

# Train-test split
X_train, X_test, y_train, y_test =
    train_test_split( X, y, test_size=0.2,
    random_state=42
)

# Model
model = LogisticRegression(max_iter=200)
model.fit(X_train, y_train)

# Prediction
y_pred = model.predict(X_test)

# Evaluation
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
display_labels=iris.target_names)
disp.plot()
plt.show()
```

8. Results

The Logistic Regression model achieved high classification accuracy, typically ranging between 96% and 100% depending on the random state used during data splitting. The confusion matrix shows that the Setosa class is classified with 100% accuracy, while minor misclassifications occur between Versicolor and Virginica due to overlapping feature values. Overall, the results indicate that the selected model is well-suited for this dataset.

9. Conclusion

This project successfully demonstrates the application of supervised machine learning for classification problems. By using the Iris dataset and Logistic Regression algorithm, an efficient and accurate classification model was developed. The project highlights the importance of selecting an appropriate algorithm based on dataset characteristics and reinforces key machine learning concepts such as training, testing, and evaluation. The

simplicity of the model also makes it easy to interpret and suitable for educational purposes.

10. Future Scope

- Try other classifiers like KNN or SVM
- Perform feature scaling and hyperparameter tuning
- Deploy the model using a simple web interface

11. References

- scikit-learn documentation
- UCI Machine Learning Repository