Iris Data Set

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Tècnica Superior d'Enginyeria
Industrial de Barcelona

Names: Victoria Vallejo, Paloma Rodríguez Larreta and Carmela Martinez Casas

Date: 11/06/2023

Course: 240AR064 - SCIENTIFIC PYTHON PER A ENGINYERS

Profesores: Marc Dalmasso and Alexandre Perera

Introduction

The project is focused on classifying different types of iris flowers based on four variables using logistic regression. The variables are: Sepal length, Sepal width, Petal length, Petal width and Species. The objective is to develop a classification model using logistic regression that can accurately differentiate between these flower species.

Dataset Overview

The project utilizes the Iris dataset, which contains measurements of four features (sepal length, sepal width, petal length, and petal width) for three different iris species: Setosa, Versicolor, and Virginica. The dataset can be accessed at [https://archive.ics.uci.edu/dataset/53/iris].

Project Objectives

The main objective of this project is to develop a logistic regression model that can accurately predict the species of an iris flower based on its four measured features. By achieving this objective, we can gain insights into the distinguishing characteristics of each species and improve our understanding of iris flower classification.

Data Exploration and Preprocessing

The first step in the project involved exploring the dataset and performing necessary preprocessing steps. This included checking for missing values, handling outliers, and assessing the distribution of the variables. Data visualization is performed using the Matplotlib and Seaborn libraries to create pair plots and bar plots for analyzing the relationships and distributions of the variables.

Logistic Regression Model

Logistic regression is a widely used classification algorithm that models the relationship between the independent variables and the probability of a certain outcome. In this project, logistic regression will be applied to classify the iris flower species based on the four measured features.

Model Training and Evaluation

The dataset will be divided into a training set and a test set. The logistic regression model will be trained using the training set and evaluated using the test set. The evaluation metrics used to assess the model's performance will include accuracy, precision, recall, and F1 score.

Results and Analysis

The performance of the logistic regression model will be analyzed, and the model's accuracy and overall performance will be assessed. Additionally, feature importance analysis can be conducted to determine the most influential features in differentiating between the iris flower species.

Limitations and Future Work

While logistic regression is a powerful classification algorithm, it may have limitations in handling complex relationships between features. Future work can explore other classification algorithms or ensemble methods to further improve the classification accuracy. Additionally, the dataset can be expanded or augmented with additional features to enhance the model's performance.

Application

Based on the project's findings, the developed logistic regression model can be used by botanists, researchers, or enthusiasts to classify iris flower species accurately. The insights gained from this project can contribute to the conservation, research, and cultivation efforts of iris flowers by understanding their unique characteristics and requirements.

Conclusion

In conclusion, this project aims to develop a logistic regression model for differentiating between iris flowers based on four measured features. By achieving accurate classification, we can gain insights into the distinguishing characteristics of each iris species. The results will provide valuable information for iris flower classification and contribute to our understanding of these flower species.