CYBR 486 - Lab #1: Introduction to Datasets and Simple Machine Learning Models

Overview

This repository contains the solution for Lab #1 from the CYBR 486 course. The lab serves as an introduction to working with datasets and implementing simple machine learning models using Python. In this lab, the Iris dataset is used to explore fundamental concepts in data manipulation, exploration, and basic machine learning.

Dataset

The Iris dataset contains data about three species of iris flowers, with four features (sepal length, sepal width, petal length, and petal width) measured for each flower. The goal is to predict the species based on these features.

Objectives

1. Loading the Dataset: Learn how to load the Iris dataset from a CSV file as well as directly from scikit-learn.
2. Exploring the Data: Use pandas to inspect and explore the dataset, including selecting specific columns and filtering data.
3. Statistical Analysis: Calculate various statistics (mean, standard deviation, min, max) for different features in the dataset.
4. Machine Learning Preparation: Understand how to prepare the data by separating features and labels for training machine learning models.

Getting Started

Prerequisites

Before running the code, you need to install the following Python packages:

* pandas
* numpy
* scikit-learn

You can install them using pip:

bash

Copy code

pip install pandas numpy scikit-learn

Files

* iris.csv: The dataset file containing flower data.
* iris\_lab1\_solution.ipynb: The Jupyter notebook containing the code and solutions for the lab.

Running the Notebook

1. Clone this repository to your local machine:

bash

Copy code

git clone https://github.com/your-username/cybr-486-lab1.git

1. Navigate to the directory where the notebook is located:

bash

Copy code

cd cybr-486-lab1

1. Open the Jupyter notebook:

bash

Copy code

jupyter notebook iris\_lab1\_solution.ipynb

1. Follow the instructions in the notebook to run each section and observe the results.

Lab Tasks

1. Loading and Exploring the Dataset

* Load the dataset from a CSV file into a pandas DataFrame.
* Use the head() method to view the first few entries.
* Describe the dataset and analyze the types of data present.
* Print specific columns, such as petal.length and petal.width.

2. Statistical Analysis

* Use pandas' describe() method to generate statistical data for the dataset.
* Calculate key statistics for different features like the standard deviation, mean, min, and max values.

3. Importing Dataset from Scikit-Learn

* Load the Iris dataset directly from scikit-learn and separate it into features (X) and labels (y).
* Calculate and print the standard deviation for the dataset features.
* Print the data types of the dataset features.

Example Code

python

Copy code

import pandas as pd

from sklearn import datasets

# Load the dataset from a CSV file

input\_file = "iris.csv"

iris\_data\_frame = pd.read\_csv(input\_file)

# Print the first 5 rows of the dataset

print(iris\_data\_frame.head())

# Calculate the mean of the sepal length

print(iris\_data\_frame['sepal.length'].mean())

# Load the dataset directly from scikit-learn

X, y = datasets.load\_iris(return\_X\_y=True, as\_frame=True)

# Print the standard deviation of the features

print(X.std())

# Print the data types of the features

print(X.dtypes)

Conclusion

This lab provides an introductory exploration of data manipulation using pandas and prepares you for machine learning by separating features and labels. By completing this lab, you'll have a better understanding of how to work with datasets and perform basic statistical analysis, which are essential skills for data science and machine learning.