**CS 519 Applied Machine Learning I**

**HW1: Basic Python Programming**

**Indronil Bhattacharjee**

**Task 1: Read data from the Iris dataset**

import pandas as pd

import matplotlib.pyplot as plt

# Task 1

def read\_dataset():

iris\_data = pd.read\_csv("iris.data", header=None)

print("Data reading completed")

return iris\_data

**Task 2: Counting number of rows and columns**

# Task 2

def row\_column(dataset):

rows, columns = dataset.shape

print(f"Number of rows: {rows}")

print(f"Number of columns: {columns}")

**Task 3: Get the distinct values of the last column**

# Task 3

def distinct\_values(dataset):

distinct\_values = dataset.iloc[:, -1].unique()

print(f"Distinct values of the last column: {distinct\_values}")

**Task 4: Count, Average, Minimum, Maximum**

1

# Task 4

def analyze\_setosa\_data(dataset):

setosa\_data = dataset[dataset.iloc[:, -1] == "Iris-setosa"] num\_rows = setosa\_data.shape[0]

avg\_first\_col = setosa\_data.iloc[:, 0].mean()

max\_second\_col = setosa\_data.iloc[:, 1].max()

min\_third\_col = setosa\_data.iloc[:, 2].min()

print(f"Number of rows with 'Iris-setosa': {num\_rows}")

print(f"Average value of the first column: {avg\_first\_col}")

print(f"Maximum value of the second column: {max\_second\_col}")

print(f"Minimum value of the third column: {min\_third\_col}")

**Task 5: Visualization of data**

# Task 5

def plot\_scatter\_plot(dataset):

colors = {'Iris-setosa': 'red', 'Iris-versicolor': 'blue', 'Iris-virginica': 'green'}

shapes = {'Iris-setosa': 's', 'Iris-versicolor': 'o', 'Iris-virginica': '^'}

for species, group in dataset.groupby(dataset.iloc[:, -1]):

plt.scatter(group.iloc[:, 0], group.iloc[:, 1], color=colors[species],

marker=shapes[species], label=species)

plt.xlabel("First Column")

plt.ylabel("Second Column")

plt.legend()

plt.show()

**Task 6: Readme file**

* Task 6
* Please read instructions in the readme.txt file

**Execution of the task functions**

# Main program

if \_\_name\_\_ == "\_\_main\_\_":

print("Task 1")

iris\_dataset = read\_dataset()

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print("Task 2")

row\_column(iris\_dataset)

2

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print("Task 3")

distinct\_values(iris\_dataset)

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print("Task 4")

analyze\_setosa\_data(iris\_dataset)

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print("Task 5")

plot\_scatter\_plot(iris\_dataset)

Task 1

Data reading completed.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Task 2

Number of rows: 150

Number of columns: 5

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Task 3

Distinct values of the last column: ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Task 4

Number of rows with 'Iris-setosa': 50

Average value of the first column: 5.006

Maximum value of the second column: 4.4

Minimum value of the third column: 1.0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Task 5

