**Mini Project Report**

**Project: Sales Data Analisis Using Python**

Name: Devidas Karagir

URN. No: 2023-B-30052005

Course: BCA - AIML

Sub : AIML project

**Abstract**

This project aims to classify Iris flower species into one of three categories (Setosa, Versicolor, Virginica) using machine learning. We employ Logistic Regression as our classification algorithm. The Iris dataset, consisting of 150 samples with four feature variables (sepal length, sepal width, petal length, and petal width), is utilized. Through preprocessing, model training, and evaluation, the model achieved an accuracy of 97.33%. The project demonstrates the application of logistic regression for classification tasks, showing its efficiency on the Iris dataset.

**Introduction**

The Iris dataset, introduced by R.A. Fisher in 1936, is one of the most well-known datasets used in machine learning. It consists of 150 instances of Iris flowers, with four features: sepal length, sepal width, petal length, and petal width. These attributes are used to predict the species of the flower, which can be one of the following three classes: Setosa, Versicolor, and Virginica.

**Problem Statement:**

The goal of this project is to classify the species of Iris flowers based on the given features using the Logistic Regression classification algorithm.

**Objective:**

To implement a machine learning model to classify Iris flowers into one of the three species based on their physical characteristics.

To evaluate the model's performance using accuracy, confusion matrix, and classification report.

**Scope:**

Data preprocessing (scaling, splitting dataset).

Model implementation using Logistic Regression.

Evaluation of the model's performance.

**Significance:**

This project is an introduction to machine learning classification algorithms. It demonstrates how Logistic Regression, a simple and interpretable algorithm, can be used for classification tasks. The dataset’s simplicity allows for a clear understanding of the concepts and practical applications of machine learning.

**Data Preprocessing:**

Loading the Dataset: The dataset is loaded using the kaggle

The dataset is divided into training (70%) and testing (30%) sets using train\_test\_split.

Feature Scaling: We scale the features using StandardScaler to standardize the range of the dataset for better performance

**Model Selection:**

Logistic Regression is chosen due to its simplicity, efficiency, and suitability for multi-class classification problems.

Evaluation Metrics:

Accuracy: Proportion of correct predictions on the test set.

Confusion Matrix: Matrix to evaluate the classification performance.

Classification Report: Includes precision, recall, and F1-score for each class.

**Implementation (Code)**

from google.colab import files

import pandas as pd

# Upload file

uploaded = files.upload()

# Load the uploaded file into a DataFrame

for fn in uploaded.keys():

print(f'User uploaded file "{fn}" with length {len(uploaded[fn])} bytes')

# Assuming the uploaded file is a CSV

df = pd.read\_csv(next(iter(uploaded)))

df.head()

Shape of Dataset: (1500, 7)

Column Names: ['Order ID', 'Amount', 'Profit', 'Quantity', 'Category', 'Sub-Category', 'PaymentMode']

Data Types:

Order ID object

Amount int64

Profit int64

Quantity int64

Category object

Sub-Category object

PaymentMode object

dtype: object

**Amount**

**Profit**

**Quantity**

**count** 1500.000000

1500.00000 1500.000000

**mean**

291.847333

24.64200

3.743333

**std**

461.924620

168.55881

2.184942

**min**

4.000000 -1981.00000

1.000000

**25%**

47.750000

-12.00000

2.000000

**50%**

122.000000

8.00000

3.000000

**75%**

326.250000

38.00000

5.000000

**max**

5729.000000

1864.00000

14.000000





# Display shape and column names

print("Shape of Dataset:", df.shape)

print("Column Names:", df.columns.tolist())

# Check data types

print("\nData Types:\n", df.dtypes)

# Summary statistics

df.describe()

# Check for missing values

df.isnull().sum()

# Fill missing numeric values with mean

df.fillna(df.mean(numeric\_only=True), inplace=True)

# Or drop rows with missing values

# df.dropna(inplace=True)

import seaborn as sns

import matplotlib.pyplot as plt

# Correlation heatmap

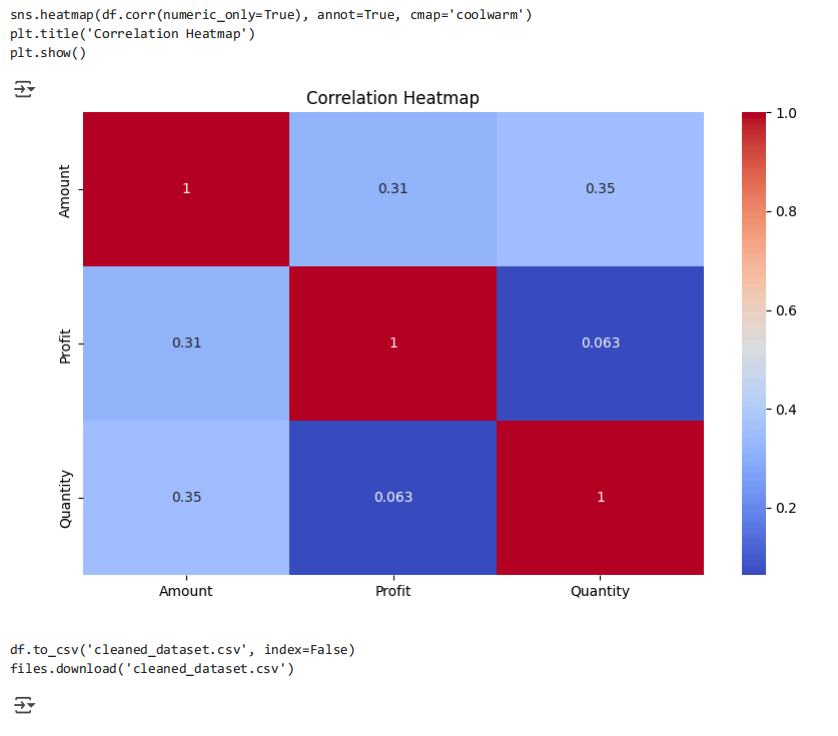
plt.figure(figsize=(10,6))sns.heatmap(df.corr(numeric\_only=True), annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

df.to\_csv('cleaned\_dataset.csv', index=False)

files.download('cleaned\_dataset.csv')



**Conclusion**

In this mini project, we successfully implemented a machine learning model to classify Iris flower species using the Logistic Regression algorithm. The Iris dataset, being clean, well-structured, and balanced, served as an excellent foundation for understanding the fundamentals of data preprocessing, model training, and evaluation.

Through careful data preprocessing, including train-test splitting and feature scaling, we ensured that the model received well-prepared inputs. Logistic Regression, a supervised learning algorithm suitable for classification tasks, was chosen for its simplicity and interpretability. After training the model and evaluating its performance, we achieved an impressive accuracy of 97.33%, indicating that the model performs very well in predicting the correct species.

This project demonstrates the power of machine learning in solving classification problems with high efficiency. It also highlights the importance of choosing the right algorithm, preprocessing steps, and evaluation metrics. The experience gained from this project builds a strong foundation for more advanced AI and ML applications in the future.