INTRNFORTE
DATA SCIENCE INTERNSHIP TRAINING
ASSIGNMENT

PROJECT TITLE

CLASSIFICATION OF IRIS SPICIES

NAME

TANISH P D

OVERVIEW OF PROJECT

This project's goal is as follows The general aim of this project is to work with the famous Iris dataset and provide insights into it. The Iris dataset is well known as the beginner's dataset for predictive modelling and is used in the context of underlaying the data exploration and visualisation features of any programming language.

The project is divided into the following key sections:

Data Loading and Cleaning:

A section in this is to import the required libraries, load the data, and check for any form of missing data in the Iris dataset.

Data Visualisation and Inference:

This section employs graphical methods to analyse associations, dispersion and direction of the data in the analysis. Descriptive and inferential analysis including pairplots, boxplots, violinplots and heatmaps are used in the analysis.

OVERVIEW OF DATASET

The Iris dataset consists of 150 observations of iris flowers, each described by four features:The Iris dataset consists of 150 observations of iris flowers, each described by four features:

Sepal Length (cm): The size and particularly the width of the green sheath below the petals.
Sepal Width (cm): Measure of the length between the first petal and the second petal in the flower.
Petal Length (cm): The length of the petal of the flower is 4.
Petal Width (cm): The more comprehensive span of the petal of the flower.

Apart from these four numerical features, the dataset contains just one other feature which is categorical: species – the type of the iris flower.

There are three species in the dataset:
- Setosa
- Versicolor
- Virginica

DATA CLEANING, EXPLORATION AND VISUALISATION

Data Cleaning:

The first thing in any data analysis project is to prepare the data to make it fit for analysis

This process typically involves the following tasks:

Loading the Dataset:

The Iris data was imported from the library of Seaborn as this is one of the popular and easily accessible datasets.

Inspecting the Dataset:

We started with the output of the initial few rows of the dataset in order to have a idea of the structure of the data.

Checking for Missing Values:

We first observed for any forms of missing values in the overall data set to later establish its completeness. Data can also be missing, which is not good for analysis and which has to be treated in the right way.

DATA CLEANING, EXPLORATION AND VISUALISATION

Data Exploration:

Pairplot for Exploring Relationships:

There was a group of numerical variables and we visualised the distribution of these variables together with the species variable. This was useful for matching the correlation of various features and giving out a prediction on the likelihood of the said features.

Boxplot for Distribution and Outliers:

A box plot was used to display sepal length under various species as well as to check for any potential outliers of the data.

Violin Plot for Distribution:

For the comparison of Petal length based on the species, we have used the panel b which is a violin plot which is an extension of box plot to density plot.

FEATURE ENGINEERING

Feature transformations refer to the process of deriving new features from other features in an aim of enhancing the learning algorithms performance. In a similar manner, with reference to the Iris dataset, feature engineering can be helpful in the extraction of much more information that is not inherent from the given features.

Petal Area:

Petal area is another scored character derived by averaging petal length and petal width and multiplying the result by the other average.

Sepal Area:

Like petal area sepal area is determined by the product of sepal length and sepal width. This feature will endeavour to measure the girth of the sepal in its entirety.

Petal Length-to-Width Ratio:

This ratio can be used to see the difference in shape of the petals between different species.

Sepal Length-to-Width Ratio:

As with the petal length/width among species and sepal length/width among varieties, the sepal length/width ratio is obtained by dividing the sepal length.

MODEL SELECTION AND TRAINING

We will implement five different models:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

For each model, we'll perform hyperparameter tuning using GridSearchCV.

MODEL EVALUATION

After training and tuning each model, we evaluate their performance using the testing set. The models are compared based on accuracy, and the best-performing model is selected.

Result:

```
Logistic Regression:

Best Parameters: {'C': 1, 'solver': 'lbfgs'}
Accuracy: 0.9111111111111111
Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        15
  versicolor       0.82      0.93      0.87        15
   virginica       0.92      0.80      0.86        15

    accuracy                           0.91        45
   macro avg       0.92      0.91      0.91        45
weighted avg       0.92      0.91      0.91        45

Confusion Matrix:
[[15  0  0]
 [ 0 14  1]
 [ 0  3 12]]

KNN:

Best Parameters: {'metric': 'manhattan', 'n_neighbors': 5}
Accuracy: 0.9111111111111111
Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        15
  versicolor       0.79      1.00      0.88        15
   virginica       1.00      0.73      0.85        15

    accuracy                           0.91        45
   macro avg       0.93      0.91      0.91        45
weighted avg       0.93      0.91      0.91        45

Confusion Matrix:
[[15  0  0]
 [ 0 15  0]
 [ 0  4 11]]
```

```
Decision Tree:

Best Parameters: {'criterion': 'gini', 'max_depth': 3, 'min_samples_split': 2}
Accuracy: 0.9777777777777777
Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        15
  versicolor       0.94      1.00      0.97        15
   virginica       1.00      0.93      0.97        15

    accuracy                           0.98        45
   macro avg       0.98      0.98      0.98        45
weighted avg       0.98      0.98      0.98        45

Confusion Matrix:
[[15  0  0]
 [ 0 15  0]
 [ 0  1 14]]

Random Forest:

Best Parameters: {'bootstrap': True, 'max_features': 'auto', 'n_estimators': 50}
Accuracy: 0.9111111111111111
Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        15
  versicolor       0.82      0.93      0.87        15
   virginica       0.92      0.80      0.86        15

    accuracy                           0.91        45
   macro avg       0.92      0.91      0.91        45
weighted avg       0.92      0.91      0.91        45

Confusion Matrix:
[[15  0  0]
 [ 0 14  1]
 [ 0  3 12]]

SVM:

Best Parameters: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
Accuracy: 0.9111111111111111
Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        15
  versicolor       0.82      0.93      0.87        15
   virginica       0.92      0.80      0.86        15

    accuracy                           0.91        45
   macro avg       0.92      0.91      0.91        45
weighted avg       0.92      0.91      0.91        45

Confusion Matrix:
[[15  0  0]
 [ 0 14  1]
 [ 0  3 12]]
```

SUMMARY

The project considered the Iris dataset and tried to study the correlation between attributes of the iris flowers and segregate them into the three species using several algorithms.

- Data Visualization
    - pairplots
    - boxplots
    - violin plots
- Feature Engineering
    - Petal Area
    - Sepal Area
    - Petal Length-to-Width Ratio
    - Sepal Length-to-Width Ratio
- Model Selection and Training:
    - Logistic Regression
    - K-Nearest Neighbors
    - Decision Tree
    - Random Forest
    - Support Vector Machine
- Evaluation:

    The performance of the models was assessed by using accuracy, precision, recall, F1-score and confusion matrix. The exercises proved that all chosen models were able to perfectly classify the species of Iris in the test set, the chosen algorithms were proven to be effective and the simplicity of the dataset did not affect the outcome.

REFERENCES

- https://www.kaggle.com/datasets/uciml/iris
- https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html
- https://ieeexplore.ieee.org/document/8777643
- https://gist.github.com/curran/a08a1080b88344b0c8a7
- https://chatgpt.com/

THANK YOU