**Data Analyst Project: Iris Dataset Basic Analysis**

**Documentation**

**INTRODUCTION**:

Every machine learning project begins with a thorough understanding of the data and defining the objectives. By applying machine learning algorithms to a dataset, you work through understanding, building, and analyzing the data to achieve the desired results. Here are the steps involved in creating a well-defined ML project:

1. Understand and define the problem
2. Prepare the data
3. Explore and analyze the data
4. Apply the algorithms
5. Reduce the errors
6. Predict the result

To understand various machine learning algorithms, we will use the Iris dataset, one of the most well-known datasets available.

**PROJECT DETAILS:**

* Domain: Data analytics
* Title: Iris Dataset Basic Analysis
* Level: Easy (Basic)

PROJECT OBJECTIVES:

* Utilize the Iris dataset to perform a data science task.
* Conduct a simple Exploratory Data Analysis (EDA) to gain insights into the dataset.

TASKS TO BE COMPLETED:

**1. Data Science Task**

* This task aims to predict the species of flowers based on their unique characteristics.
* Select and implement an appropriate machine learning algorithm (e.g., Decision Trees, Logistic Regression).
* Split the dataset into training and testing sets for model evaluation.
* Train the model on the training set and evaluate its performance on the testing set.
* Use metrics such as accuracy, precision, and recall for model evaluation.

**2. Simple Exploratory Data Analysis (EDA):**

* Perform a basic EDA to understand the structure and characteristics of the Iris dataset.
* Explore the distribution of each feature in the dataset.
* Create visualizations such as histograms, box plots, or scatter plots to highlight relationships between features.

**3. Documentation:**

* Document your approach, methodologies, and any challenges faced during the data science task and EDA.
* Provide clear explanations for the choices made in terms of algorithms, features, and evaluation metrics.
* Include comments in your code to enhance readability.

STEPS:

1. **Data Collection and Loading:**
   * Obtain the Iris dataset from a reliable source (e.g., sci-kit-learn, UCI Machine Learning Repository).
   * Load the dataset into the project environment using suitable libraries (e.g., pandas, sci-kit-learn).
2. **Exploratory Data Analysis (EDA):**
   * Perform preliminary analysis to understand the structure and characteristics of the dataset.
   * Explore basic statistics, such as mean, median, standard deviation, etc., for each feature.
   * Visualize the data using graphs (scatter plots, histograms, pair plots) to identify relationships between features and species.
3. **Data Preprocessing:**
   * Handle missing values, if any.
   * Encode categorical variables (if required).
   * Scale or normalize the features to ensure uniformity in the data.
4. **Model Building:**
   * Split the dataset into training and testing sets.
   * Choose suitable machine learning models for classification (e.g., Decision Trees, Random Forests, Support Vector Machines, K-nearest neighbors).
   * Train the models using the training data.
5. **Model Evaluation:**
   * Evaluate the trained models using appropriate metrics on the test data.
   * Compare the performance of different models to identify the best-performing one.
6. **Model Deployment:**
   * Select the best model based on evaluation results.
   * Deploy the chosen model to make predictions on new or unseen data.

**APPROACH:**

To work on the Iris dataset, we began by loading the dataset into a Jupyter notebook using the read\_csv function. We then proceeded with the project plan.

We implemented various mathematical and statistical operations on the dataset, leveraging Python's built-in libraries. Basic data exploration (EDA) was conducted, including data visualization using the matplotlib and seaborn libraries.

We generated histograms, box plots, and other visualizations using Seaborn to explore the data further. Finally, we applied machine learning algorithms such as Logistic Regression and Decision Trees. These algorithms were evaluated using metrics like accuracy, precision, and recall. Although other machine learning algorithms are suitable for this project, some are more complex to understand.