# 

Assignment#1: Machine Learning – Iris Dataset – Basic Classification Model.

Swapnali Dashrath

Oakland University

CSI6160 - Machine Learning

Prof. Mohammad Wardat.

September 26, 2024.

**Iris Dataset Modeling and Evaluation Report**

This report presents the implementation and evaluation of two classifiers – Decision Tree and Logistic Regression on Iris Dataset using existing libraries (numpy, pandas, seaborn, matplotlib etc). Implementation is using Python (code file submitted along with report) and Evaluation is based on two evaluation metrics from Accuracy and Confusion Matrix.

**Problem Statement:**

We are trying to use attributes of flowers to predict the species of the flower, specifically, we are trying to use Sepal length, Sepal width, Petal length, Petal width to predict if an Iris flower is of species setosa' 'versicolor' 'virginica

**Approach:**

1. Analyzing Data – Iris Dataset
2. Data Cleanup and formatting using Data Frames.
3. Visualize the data using Data Plots.
4. Defining Relationship of data features with target.
5. Exploratory Data Analysis (EDA) – Pairplots.
6. Training Data Split.
7. Data prep for modeling
8. Simple Manual Modeling – Decision Tree Classifier.
9. Modeling using Logistic Regression Classifier.
10. Evaluation based on Accuracy and Confusion Matrix.
11. Using Cross validation for Evaluation.

**Iris Dataset:**

The dataset consists of 150 samples of iris flowers, divided into three species: Setosa, Versicolor, and Virginica, with four features: sepal length, sepal width, petal length, and petal width.

**Classifiers Implemented**

1. Decision Tree Classifier
2. Logistic Regression Classifier

**Data Splitting**

The dataset was split into training (70%) and testing (30%) sets. This allows us to train the models on the training data and evaluate them on unseen test data.

**Modeling**

Our base line model is just randomly guessing the species of flower, or guessing a single species for every data point with certain accuracy.

**Evaluation Metrics**

We used the following metrics to evaluate the models:

* Accuracy: Proportion of correctly classified instances.
* Confusion Matrix: Summarizes the performance by showing true positives, false positives, true negatives, and false negatives.

**Results**

**1**. **Decision Tree Classifier**

* **Accuracy**: 94.68% (accuracy on the test set), For different test data set model provided 100% accuracy.The Decision Tree achieved accuracy, indicating that the model has fully captured the patterns in the Iris dataset.

2**. Logistic Regression Classifier**

* **Accuracy**: 0.96 (96% accuracy on the test set)
* Confusion Matrix:

[[10 0 0]

[ 0 11 0]

[ 0 1 6]]

Although Logistic Regression performed well with an accuracy of 96%, it slightly misclassified one instance of the Versicolor class as Virginica.

**Interpretations**

* Decision Tree Classifier: The model perfectly classified all samples in the test set. While this performance is impressive, decision trees are prone to over fitting, especially on small datasets like Iris, where the tree may become too specific.
* Logistic Regression Classifier: Logistic regression also performed exceptionally well, but it made one misclassification. Logistic regression assumes a linear relationship between the features and the target class, which may not fully capture the complexity of the dataset. However, its performance is robust and generalizes well.

**Conclusion**

Both models performed exceptionally well, with the Decision Tree achieving perfect accuracy and Logistic Regression following closely behind. However, due to its interpretability and lower tendency to overfit, Logistic Regression could be considered a better option for generalization in real-world scenarios.

**References**

Andreas C. Müller, Sarah Guido. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists O'Reilly Media, Inc., October 2016.

Indently. (2023), [How To Use Jupyter NoteBook For Data Analysis (Beginner Tutorial) (youtube.com)](https://www.youtube.com/watch?v=IMrxB8Mq5KU) [(125) How To Use Jupyter NoteBook For Data Analysis (Beginner Tutorial) - YouTube](https://www.youtube.com/watch?v=IMrxB8Mq5KU)

# Project Data Science (2020).Intro to Machine Learning with Python 1: Welcome and Project Setup [Intro to Machine Learning with Python 1: Welcome and Project Setup (youtube.com)](https://www.youtube.com/watch?v=rdaG53khzv0&list=PLMAyPTgGwv2DUV6DZib9eMetsTTX87JNr)