# **Data Science Internship Report**

TASK 1 :

Project Title: **Iris Flower Classification Using Machine Learning**

Intern: **REEMA - M**

Company: **Oasis Infobyte**

Duration: **5th October – 5th November**

College: **Nehru Institute Of Engineering And Technology**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Oasis Infobyte for providing me with the opportunity to work as a Data Science intern. I am thankful to my mentor for their guidance and support throughout this project.

**Abstract**

The goal of this project was to classify Iris flowers into three species — Setosa, Versicolor, and Virginica — using machine learning algorithms. The dataset contains 150 samples with four features: sepal length, sepal width, petal length, and petal width. Multiple models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest, were trained and evaluated. The Random Forest classifier achieved the highest accuracy, demonstrating its effectiveness in this classification task.

**Introduction**

Iris flower classification is a supervised machine learning problem. The task is to predict the species of a flower based on its measurements. Accurate classification of Iris species has applications in botany, agriculture, and biological research.Machine learning allows us to train algorithms that can learn patterns from the dataset and predict the species of unseen flowers automatically.

**Dataset Description**

The dataset used is the Iris dataset, containing 150 samples and 5 columns:

**Column Name:**  **Description:**

Id Unique identifier for each sample

SepalLengthCm. Length of sepal (cm)

SepalWidthCm. Width of sepal (cm)

PetalLengthCm. Length of petal (cm)

PetalWidthCm Width of petal (cm)

Species Target variable: Setosa, Versicolor, Virginica

**Class Distribution:**

Species Count

Iris-setosa 50

Iris-versicolor 50

Iris-virginica 50

**Exploratory Data Analysis (EDA)**

**Summary Statistics and Visualization:**

Import pandas as pd

Import seaborn as sns

Import matplotlib.pyplot as plt

# Load dataset

Df = pd.read\_csv(“Iris.csv”)

# Summary statistics

Print(df.describe())

# Pairplot to see relationships

Sns.pairplot(df, hue=’Species’)

Plt.show()

# Correlation heatmap

Sns.heatmap(df.drop(‘Id’, axis=1).corr(), annot=True, cmap=’coolwarm’)

Plt.show()

**Observations:**

* Petal length and petal width are highly correlated and crucial for classification.
* Sepal measurements show overlap between Versicolor and Virginica species.

**Methodology**

**Step 1: Prepare Data**

From sklearn.model\_selection import train\_test\_split

X = df.drop([‘Id’,’Species’], axis=1)

Y = df[‘Species’]

# Split dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, test\_size=0.2, random\_state=42

)

**Results and Discussion**

Model. Accuracy

Logistic Regression 0.9667

K-Nearest Neighbors 0.9667

Support Vector Machine. 0.9667

Decision Tree. 0.9667

Random Forest 1.0

**Observations:**

* Random Forest achieved the highest accuracy of 100%.
* All models performed well due to the clear separability of the Iris dataset.

**Saving the Best Model**

Import joblib

Best\_model = RandomForestClassifier(random\_state=42)

Best\_model.fit(X\_train, y\_train)

Joblib.dump(best\_model, “iris\_rf\_model.pkl”)

Print(“Random Forest model saved as ‘iris\_rf\_model.pkl’”)

**Conclusion**

* The project successfully classified Iris flowers into three species.
* Random Forest was the best-performing model.
* Future work can include hyperparameter tuning and deployment as a web application.

**References**

* Scikit-learn documentation: <https://scikit-learn.org/>
* UCI Machine Learning Repository: Iris Dataset
* Python libraries: pandas, seaborn, matplotlib