**Task 1: Data Collection & Exploration**

**1. Dataset Loading**

* The dataset was loaded from an Excel file

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

* The dataset was examined for missing values (none found).
* Statistical summaries were generated to understand feature distributions.
* **Visualizations:**
  + **Pair plots** and **scatter plots** were used to analyze feature relationships.

**Task 2: Preprocessing & Feature Engineering**

**1. Data Cleaning**

* No missing values were found, so no imputation was needed.

**2. Feature Scaling & Splitting**

* The dataset was split into **training (80%)** and **testing (20%)** sets using train\_test\_split.

**Task 3: Model Training & Evaluation**

We trained five different classification models:

1. **Logistic Regression**
2. **Decision Tree**
3. **Random Forest**
4. **Support Vector Machine (SVM)**
5. **K-Nearest Neighbors (KNN)**

**Initial Model Performance**

| **Model** | **Accuracy** |
| --- | --- |
| Logistic Regression | 0.93 |
| Decision Tree | 0.93 |
| Random Forest | 0.93 |
| SVM | 0.96 |
| KNN | 0.93 |

**Observations:**

* **SVM performed the best (96% accuracy)**, followed by Logistic Regression and KNN.
* Random Forest had a slightly lower performance compared to other models.

**Task 4: Model Optimization & Hyperparameter Tuning**

**1.Hyperparameter Tuning (Using GridSearchCV)**

We optimized each model using GridSearchCV to find the best hyperparameters.

| **Model** | **Best Parameters** | **Cross-Validation Accuracy** |
| --- | --- | --- |
| Logistic Regression | C=10 | 0.9667 |
| Decision Tree | max\_depth=None, min\_samples\_split=2 | 0.9417 |
| Random Forest | n\_estimators=10, max\_depth=5, min\_samples\_split=2 | 0.9667 |
| SVM | C=0.1, kernel='linear', gamma='scale' | 0.9750 |
| KNN | n\_neighbors=9, metric='euclidean' | 0.9583 |

**2.Final Model Performance After Tuning**

| **Model** | **Final Accuracy** |
| --- | --- |
| Logistic Regression | **1.00** |
| Decision Tree | 0.90 |
| Random Forest | 0.90 |
| SVM | **0.93** |
| KNN | **0.97** |

**3.Cross-Validation for Model Robustness**

To ensure model reliability, we implemented **5-fold cross-validation** using cross\_val\_score.

| **Model** | **Cross-Validation Accuracy** |
| --- | --- |
| Logistic Regression | 0.9667 |
| Decision Tree | 0.9417 |
| Random Forest | 0.9583 |
| SVM | 0.9750 |
| KNN | 0.9583 |

**Final Observations:**

* **SVM (97.5%) and Logistic Regression (96.67%) performed the best** in cross-validation.
* **KNN and Random Forest followed closely behind.**
* **Decision Tree had the lowest robustness.**

**Conclusion:**

**1.Support Vector Machine (SVM) emerged as the best model**, showing the highest accuracy (97.5%) in cross-validation.  
**2.Logistic Regression also performed excellently**, achieving 100% accuracy on the test set.