**Model Comparison Report**

**Objective:**

This report evaluates the performance of multiple machine learning models used to predict heart disease based on patient data. The objective is to identify the best-performing model for production deployment, balancing accuracy, interpretability, and computational efficiency. Several models were tested, and their performance was assessed using key classification metrics.

**1. Models Evaluated:**

The following models were implemented and tested:

* **Logistic Regression**: A linear model that is widely used for binary classification tasks.
* **Random Forest Classifier**: An ensemble model that aggregates the predictions of multiple decision trees.
* **Support Vector Machine (SVM)**: A model that finds the optimal hyperplane to classify data points.
* **K-Nearest Neighbours (KNN)**: A simple non-parametric model that classifies based on proximity to training data.
* **Decision Tree Classifier**: A tree-based model that makes decisions by splitting data at various nodes.

**2. Performance Metrics:**

The models were compared using the following metrics to ensure a comprehensive evaluation:

* **Accuracy**: The overall percentage of correct predictions.
* **Precision**: The ratio of true positives to the total number of predicted positives.
* **Recall**: The ratio of true positives to the total number of actual positives.
* **F1-Score**: The harmonic mean of Precision and Recall, providing a balance between the two.
* **AUC (Area Under the Curve)**: A measure of the model’s ability to distinguish between classes, irrespective of the threshold.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 85.3% | 0.87 | 0.82 | 0.84 | 0.90 |
| Random Forest | 88.7% | 0.89 | 0.86 | 0.88 | 0.93 |
| Support Vector Machine | 83.5% | 0.85 | 0.80 | 0.82 | 0.88 |
| K-Nearest Neighbours | 81.1% | 0.82 | 0.78 | 0.80 | 0.85 |
| Decision Tree | 80.2% | 0.81 | 0.77 | 0.79 | 0.83 |

**3. Model Selection:**

* **Best Model: Random Forest Classifier**
  + The Random Forest model consistently outperformed other models across all metrics. It achieved the highest accuracy (88.7%), precision (0.89), recall (0.86), F1-score (0.88), and AUC (0.93). This model was particularly effective in capturing complex patterns in the data and was robust against overfitting, making it the best candidate for production deployment.
* **Alternative: Logistic Regression**
  + Logistic Regression achieved a competitive performance with an accuracy of 85.3% and an AUC of 0.90. It is a simpler model with a lower computational footprint and strong interpretability. This makes it a viable option when transparency and speed are important considerations, particularly in healthcare settings where model explainability is critical.

**4. Conclusion:**

The **Random Forest Classifier** is recommended as the best model for production deployment due to its superior performance across all evaluation metrics. It is robust, accurate, and can handle the complexities of the dataset effectively. For applications requiring more interpretability and computational efficiency, **Logistic Regression** provides a good trade-off, offering strong performance while being simpler and easier to implement.