**Machine Learning Model Performance Report**

**1. Introduction**

In this report, we evaluate the performance of various machine learning models after hyperparameter tuning on a given dataset. We assess models using several key evaluation metrics such as **accuracy**, **best score**, and **best parameters** to determine the most effective model for our problem.

**2. Models Evaluated**

The following models were evaluated during the process:

* **Gradient Boosting**
* **Random Forest**
* **Logistic Regression**
* **XGBoost**
* **SVM (Support Vector Machine)**
* **Naive Bayes**
* **Decision Tree**
* **KNN (K-Nearest Neighbors)**

**3. Hyperparameter Tuning**

For each model, hyperparameter tuning was performed to find the optimal set of parameters that maximize performance. The best parameters found for each model are as follows:

* **Gradient Boosting**:  
  {'learning\_rate': 0.1, 'max\_depth': 3, 'min\_samples\_split': 2}
* **Random Forest**:  
  {'max\_depth': 10, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2}
* **Logistic Regression**:  
  {'C': 0.065, 'max\_iter': 300, 'penalty': 'l1'}
* **XGBoost**:  
  {'colsample\_bytree': 0.8, 'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 100}
* **SVM**:  
  {'C': 1, 'class\_weight': None, 'gamma': 'scale'}
* **Naive Bayes**:  
  {'var\_smoothing': 1e-09}
* **Decision Tree**:  
  {'criterion': 'entropy', 'max\_depth': 10, 'min\_samples\_split': 2}
* **KNN**:  
  {'metric': 'manhattan', 'n\_neighbors': 9, 'weights': 'distance'}

**4. Evaluation Metrics**

The models were evaluated based on the following metrics:

* **Accuracy**: The proportion of correct predictions out of total predictions made.
* **Best Score**: The best performance score achieved after hyperparameter tuning.

**5. Performance Results**

Below are the **accuracy** scores and **best scores** for each model:

| **Model** | **Accuracy** | **Best Score** |
| --- | --- | --- |
| **Logistic Regression** | 0.9412 | 0.9373 |
| **SVM** | 0.9408 | 0.9372 |
| **Naive Bayes** | 0.9408 | 0.9372 |
| **Gradient Boosting** | 0.9408 | 0.9397 |
| **Random Forest** | 0.9377 | 0.9374 |
| **XGBoost** | 0.9295 | 0.9373 |
| **Decision Tree** | 0.9075 | 0.9338 |
| **KNN** | 0.7390 | 0.8001 |

**6. Analysis of Results**

* **Best Performing Models**:
  + **Logistic Regression** achieved the highest accuracy at **0.9412**, making it the top-performing model in terms of overall accuracy.
  + **Gradient Boosting** achieved the highest **best score** of **0.9397** after hyperparameter tuning, showcasing its high potential for fine-tuning.
* **Consistently High Performers**:
  + **SVM** and **Naive Bayes** had very similar accuracy scores of **0.9408**, showing solid performance with their respective best scores.
* **Moderate Performers**:
  + **Random Forest** and **XGBoost** both achieved relatively high accuracy but did not outperform **Logistic Regression** or **Gradient Boosting** in this case. These models still provide valuable results, especially in more complex datasets.
* **Lower Performing Models**:
  + **KNN** and **Decision Tree** performed the least effectively with **KNN** achieving an accuracy of **0.7390** and **Decision Tree** with **0.9075** accuracy. KNN especially struggled due to sensitivity to hyperparameters like the number of neighbors and distance metric.

**7. Model Selection**

After careful analysis of the models, the **Logistic Regression** model is selected as the final model for deployment, primarily due to:

* **Highest Accuracy**: Logistic Regression achieved the highest accuracy among all the models evaluated.
* **Interpretability**: Logistic Regression provides a simple, interpretable model, which is advantageous for understanding and explaining predictions.

However, **Gradient Boosting** is a close second due to its superior **best score** and strong performance after hyperparameter tuning. If needed, **Gradient Boosting** can be considered as an alternative model for production, especially for cases where more complex relationships in the data are expected.

**8. Conclusion**

The performance evaluation of multiple machine learning models revealed that **Logistic Regression** is the most effective model for this particular problem, achieving the highest accuracy and stable performance after tuning. While **Gradient Boosting** provides a strong alternative, the simplicity and high performance of Logistic Regression make it the preferred choice for deployment.

Further steps could involve exploring model ensembles or more complex models like **XGBoost** in the future to tackle more challenging datasets.