**Title**: Machine Learning Classification Project - Model Comparison and Insights

**Introduction**

The objective of this project was to classify data effectively using machine learning techniques. We applied rigorous preprocessing, feature selection, model training, and evaluation techniques to derive actionable insights. The findings, approaches, and results are documented below.

**1. Data Preprocessing**

Data preprocessing is a critical step to ensure clean and structured input for machine learning models. The following steps were performed:

* **Data Cleaning**: Handled missing values by imputing mean/median for numerical features and the most frequent value for categorical features.
* **Normalization**: Scaled numerical data using MinMaxScaler to ensure all features are on the same scale.
* **Encoding**: Converted categorical variables to numerical formats using one-hot encoding where applicable.

**2. Feature Selection**

To improve the model's performance and reduce computational complexity, we applied feature selection techniques:

* **Correlation Analysis**: Removed highly correlated features to mitigate multicollinearity.
* **Feature Importance**: Extracted important features using tree-based methods like Random Forest.

**3. Model Selection & Training**

Three machine learning models were trained for classification:

* **Random Forest Classifier**: Utilized an ensemble of decision trees to achieve high accuracy.
* **Support Vector Machine (SVM)**: Optimized with a radial basis function (RBF) kernel for non-linear classification.
* **XGBoost Classifier**: Trained using gradient boosting for robust performance on complex datasets.

Hyperparameters for all models were tuned using GridSearchCV to maximize performance.

**4. Evaluation Metrics**

To evaluate and compare model performances, the following metrics were calculated:

* **Accuracy**: The ratio of correctly predicted instances to total instances.
* **Precision**: The proportion of true positives among predicted positives.
* **Recall**: The proportion of true positives among actual positives.
* **F1-Score**: Harmonic mean of precision and recall.
* **AUC-ROC**: Evaluated model’s ability to distinguish between classes.
* **Cohen Kappa Score**: Measured inter-annotator agreement.
* **Matthews Correlation Coefficient (MCC)**: Considered true/false positives and negatives for a balanced evaluation.

**5. Visualizations**

Insights were visualized to facilitate understanding:

* **Confusion Matrix**: Illustrated the classification performance for each model.
* **Feature Importance Plot**: Highlighted the most influential features.
* **ROC Curve**: Showed trade-offs between sensitivity and specificity.
* **Precision-Recall Curve**: Focused on precision-recall trade-offs for imbalanced datasets.

**6. Results and Insights**

* **Model Comparison**: XGBoost outperformed other models with the highest F1-score and AUC-ROC. Random Forest also delivered competitive results, while SVM lagged due to its sensitivity to parameter settings.
* **Feature Importance**: The top 5 features contributed over 70% of the predictive power.
* **Confusion Matrix Analysis**: Most errors were concentrated in borderline cases.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| XGBoost | 0.95 | 0.96 | 0.94 | 0.95 | 0.98 |
| Random Forest | 0.93 | 0.92 | 0.91 | 0.91 | 0.96 |
| SVM | 0.89 | 0.90 | 0.85 | 0.87 | 0.91 |

**Conclusion**

This project demonstrated the power of machine learning in classification tasks. By employing preprocessing, feature selection, and evaluation, XGBoost was identified as the most effective model. The insights and visualizations provide actionable knowledge for future decision-making processes.

**References**

1. Scikit-learn Documentation
2. XGBoost Documentation
3. SHAP Interpretability Resources

Report prepared by Vemulapelli.Saakshitha . For any inquiries, contact [saakshitha2004@gmail.com](mailto:saakshitha2004@gmail.com) | +91 7396618668 .