**Supervised Data Mining Project: Binary Classification**

**Introduction**

This project aims to compare the performance of three supervised classification algorithms—Random Forest, LSTM (Long Short-Term Memory), and SVM (Support Vector Machine)—for a binary classification task. The evaluation metrics include True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), False Positive Rate (FPR), False Negative Rate (FNR), True Skill Statistics (TSS), Heidke Skill Score (HSS), and ROC-AUC. The dataset chosen for this project is the Breast Cancer dataset from Scikit-learn. The primary goal is to determine the best-performing algorithm based on these metrics, evaluated using 10-fold cross-validation.

**Dataset**

The dataset used is the **Breast Cancer Dataset**, a built-in dataset in Scikit-learn. It contains:

* **Features**: 30 numerical attributes derived from digitized images of fine needle aspirates of breast masses.
* **Target Variable**: A binary classification:
  + 1: Malignant tumor (positive class).
  + 0: Benign tumor (negative class).

The dataset is standardized using StandardScaler to normalize the features, ensuring that all attributes have the same scale and range. This step is crucial for models like SVM and LSTM, which are sensitive to feature scaling.

**Methodology**

**Algorithms Implemented**

1. **Random Forest**:
   * An ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions for improved accuracy.
   * Handles non-linearity well and is robust against overfitting.
2. **Support Vector Machine (SVM)**:
   * A kernel-based algorithm that constructs hyperplanes to maximize the separation margin between two classes.
   * A linear kernel is used in this project for simplicity.
3. **LSTM (Long Short-Term Memory)**:
   * A type of recurrent neural network (RNN) that can model sequential or time-dependent data.
   * Though primarily used for sequence prediction tasks, it is adapted here for binary classification by flattening the data into sequential input.

**Cross-Validation**

* **10-fold cross-validation** was used to divide the dataset into training and testing subsets. Each fold is evaluated individually, and results are averaged to ensure robust evaluation.

**Metrics Calculation**

The evaluation metrics include:

1. **Confusion Matrix Components**:
   * **TP** (True Positives): Correctly predicted positive cases.
   * **TN** (True Negatives): Correctly predicted negative cases.
   * **FP** (False Positives): Negative cases incorrectly predicted as positive.
   * **FN** (False Negatives): Positive cases incorrectly predicted as negative.
2. **Derived Metrics**:
   * **FPR (False Positive Rate)**: FPR=FPFP+TNFPR=FP+TNFP​
   * **FNR (False Negative Rate)**: FNR=FNFN+TPFNR=FN+TPFN​
   * **TSS (True Skill Statistics)**: TSS=TPTP+FN−FPFP+TNTSS=TP+FNTP​−FP+TNFP​
   * **HSS (Heidke Skill Score)**: HSS=2(TP×TN−FP×FN)(TP+FN)(FN+TN)+(TP+FP)(FP+TN)HSS=(TP+FN)(FN+TN)+(TP+FP)(FP+TN)2(TP×TN−FP×FN)​
3. **ROC-AUC**: Measures the area under the Receiver Operating Characteristic curve to evaluate classification performance.

**Results**

The following table summarizes the performance metrics for each algorithm:

| **Metric** | **Random Forest** | **SVM** | **LSTM** |
| --- | --- | --- | --- |
| **TP** | 35.0 | 34.9 | 34.9 |
| **TN** | 19.8 | 20.3 | 19.3 |
| **FP** | 1.4 | 0.9 | 1.9 |
| **FN** | 0.7 | 0.8 | 0.8 |
| **FPR** | 0.064 | 0.040 | 0.088 |
| **FNR** | 0.019 | 0.022 | 0.021 |
| **TSS** | 0.917 | 0.938 | 0.890 |
| **HSS** | 0.919 | 0.935 | 0.895 |
| **ROC-AUC** | 0.993 | 0.994 | 0.988 |

**Discussion**

**Random Forest**:

* Demonstrated robust performance with balanced results across all metrics.
* Achieved high TSS (0.917) and HSS (0.919), making it a reliable choice for this binary classification task.
* The algorithm is resilient to noise and provides consistent results across folds.

**SVM**:

* Delivered the best performance overall, achieving the highest TSS (0.938) and HSS (0.935).
* The False Positive Rate (FPR = 0.040) was the lowest among all models, making SVM a strong choice for applications where minimizing false positives is critical.

**LSTM**:

* While competitive, LSTM underperformed slightly compared to SVM and Random Forest in TSS (0.890) and HSS (0.895).
* The higher False Positive Rate (FPR = 0.088) suggests potential overfitting or insufficient optimization for this dataset.

**Conclusion**

Based on the evaluation:

* **SVM** is the best-performing algorithm for the Breast Cancer dataset, achieving the highest TSS, HSS, and ROC-AUC. It is particularly effective in scenarios where false positives must be minimized.
* **Random Forest** is a close second, providing balanced performance and robustness across all metrics.
* **LSTM**, while showing good performance, is less suited for this task due to its slightly higher false positive rate and lower TSS and HSS scores.

For future work, optimizing hyperparameters for LSTM or using a more complex architecture could improve its performance. Additionally, experimenting with other datasets or larger feature sets could further validate these findings.