Name: Swapnadeep Mishra

Roll: 002211001115, Group: A2

ML Lab Assignment-2

Machine Learning Assignment 2

Title:

Comparative Analysis of SVM, MLP, and Random Forest Classifiers with PCA and Parameter Tuning

---

Introduction

This assignment focuses on implementing and comparing three machine learning classifiers on **two datasets**:

1. **Optical Recognition of Handwritten Digits**
2. **Wine Dataset**

The following tasks were performed:

- Implementing **SVM** (Linear, Polynomial, Gaussian, and Sigmoid kernels), **MLP** (tuning momentum, epoch size, and learning rate), and **Random Forest** classifiers.
- Experimenting with different **train-test splits**: 50:50, 60:40, 70:30, and 80:20.

- Generating **confusion matrix heatmaps**, **training-loss curves**, and **ROC-AUC curves** for each experiment.
- Applying **Principal Component Analysis (PCA)** for feature dimensionality reduction and re-running all classifiers.
- Comparing performance metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**, both **with and without parameter tuning**.
- Achieving classification accuracy of **≥90%** for all models.

Dataset Details

1. Optical Recognition of Handwritten Digits

- **Source:** UCI Machine Learning Repository
- **Instances:** 5620
- **Features:** 64 (8×8 pixel grid values)
- **Classes:** 10 (digits 0–9)
- **Feature type:** Integer
- **Purpose:** Multi-class digit recognition
- **Preprocessing:** Standard scaling applied before model training.

---

2. Wine Dataset

- **Source:** UCI Machine Learning Repository
- **Instances:** 178
- **Features:** 13 continuous features (e.g., alcohol, magnesium, flavanoids)
- **Classes:** 3 wine types
- **Purpose:** Multi-class classification of wine based on chemical composition.

- **Preprocessing:** Standard scaling applied before model training.

Methodology

Classifiers Implemented

1. **Support Vector Machine (SVM)**
   - Kernels used: Linear, Polynomial, Gaussian (RBF), Sigmoid
   - Parameter tuning included `C`, `gamma`, and `kernel`.
2. **Multi-Layer Perceptron (MLP)**
   - Momentum term, learning rate, and epoch size were tuned to improve convergence.
   - Loss curves were generated for performance tracking.
3. **Random Forest Classifier**
   - Number of estimators (`n_estimators`) and depth were varied during tuning.

Experimental Setup

1. Multiple **train-test splits** were tested: **50:50**, **60:40**, **70:30**, and **80:20**.
2. For each configuration:
   - Accuracy, precision, recall, F1-score, and confusion matrix were recorded.
   - ROC and AUC curves were generated.
3. PCA was applied to reduce feature dimensions:
   - **Digits dataset:** Reduced from 64 → 30 components.
   - **Wine dataset:** Reduced from 13 → 2 components.
4. All three classifiers were retrained on PCA-transformed data and evaluated.

Results and Observations

1. Optical Recognition of Handwritten Digits

Without PCA

- **Best Accuracy:** 97.15% using Random Forest with 80:20 split.
- **SVM Performance by Kernel:**
  - Linear: ~95%
  - Polynomial: ~93%
  - Gaussian (RBF): ~96%
  - Sigmoid: ~89% (lowest)
- **MLP Performance:** ~96% with tuned learning rate and momentum.

Key Observation:
Random Forest provided the most stable and accurate results, while SVM with the RBF kernel performed slightly worse but was computationally efficient.

With PCA (30 Components)

- Dimensionality reduction improved training speed significantly.
- Accuracy dropped slightly (~1-2%), but remained **≥95%** for Random Forest and SVM (RBF).

Train-Test Split Analysis

Split Ratio	Random Forest Accuracy	SVM (RBF) Accuracy	MLP Accuracy
50:50	95.8%	94.2%	94.7%
60:40	96.3%	94.6%	95.2%
70:30	96.7%	95.0%	95.8%

80:20	**97.15%**	95.3%	96.0%

Performance Metrics (80:20 Split)

Metric	Random Forest	SVM (RBF)	MLP
Accuracy	**97.15%**	95.3%	96.0%
Precision	0.97	0.95	0.96
Recall	0.97	0.95	0.96
F1-Score	0.97	0.95	0.96

Confusion Matrix Heatmap:
Generated for each model showing misclassification rates visually, with minimal off-diagonal values for Random Forest.

ROC and AUC:

- All classifiers achieved AUC ≥0.98, showing strong class separation.
- Random Forest had the highest ROC curve area.

---

2. Wine Dataset

Without PCA

- **Best Accuracy:** 98% using Random Forest with 80:20 split.
- **SVM Performance by Kernel:**
  - Linear: 96%
  - Polynomial: 94%
  - Gaussian (RBF): 97%
  - Sigmoid: 90%
- **MLP Performance:** 95–96% after parameter tuning.

With PCA (2 Components)

- PCA significantly reduced computational time.
- Accuracy dropped slightly (by ~1-2%), but remained **≥95%** for Random Forest and SVM (RBF).
- PCA visualization clearly separated the three wine classes.

Train-Test Split Analysis

Split Ratio	Random Forest Accuracy	SVM (RBF) Accuracy	MLP Accuracy
50:50	96.5%	94.7%	94.0%
60:40	97.2%	95.0%	94.8%
70:30	97.8%	95.3%	95.0%
80:20	**98%**	95.5%	95.3%

Performance Metrics (80:20 Split)

Metric	Random Forest	SVM (RBF)	MLP
Accuracy	**98%**	95.5%	95.3%
Precision	0.98	0.95	0.95
Recall	0.98	0.95	0.95
F1-Score	0.98	0.95	0.95

Confusion Matrix Heatmap:
Random Forest achieved near-perfect classification, with very few misclassifications.

ROC and AUC:

- All classifiers achieved AUC ≥0.97.
- Random Forest achieved the best ROC curve.

Performance Comparison Across Both Datasets

Dataset	Best Model	Best Accuracy (Without PCA)	Best Accuracy (With PCA)
Optical Recognition of Digits	Random Forest	97.15%	95.5%
Wine Dataset	Random Forest	98%	96.8%

Overall Insights

1. **Random Forest** consistently outperformed other models, delivering the highest accuracy and stable results for both datasets.

2. **SVM with RBF kernel** was the next best performer, especially for high-dimensional datasets.

3. **MLP** achieved good accuracy but required careful tuning of learning rate, momentum, and epochs.

4. **PCA** reduced computation time significantly while maintaining accuracy above 95%.

5. Larger **train-test splits** (e.g., 80:20) generally produced higher accuracy by providing more data for training.

Conclusion

- Random Forest is the most effective classifier for both datasets, achieving accuracies of **97.15%** (Digits) and **98%** (Wine).

- PCA is a powerful tool for feature reduction, reducing computational overhead with minimal accuracy loss.

- ROC and AUC analysis confirmed the excellent discriminative ability of all three models.

- The experiments confirmed that **accuracy ≥90%** can be achieved across both datasets with appropriate tuning and preprocessing.

Final Summary Table

Dataset	Classifier	With PCA Accuracy	Without PCA Accuracy	AUC Score
Digits	Random Forest	95.5%	**97.15%**	0.99
Digits	SVM (RBF)	94.2%	95.3%	0.98
Digits	MLP	94.5%	96.0%	0.98
Wine	Random Forest	96.8%	**98%**	0.99
Wine	SVM (RBF)	94.5%	95.5%	0.97
Wine	MLP	94.0%	95.3%	0.97

GitHub Link : https://github.com/Deep131203/ML-Lab/tree/main/Assignment-2