Report

Assignment-6

Application of Machine Learning in Biological Systems

Name: Jatin Mahawar

Roll.NO: 22CS30032

Objective

This assignment explores and compares machine learning models for breast cancer type prediction using the Breast Cancer Wisconsin (Diagnostic) dataset. The models analyzed include Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). The goal is to identify the model that best balances accuracy and computational efficiency for this classification task.

Code Overview

The notebook is organized into sections that cover:

1. **Data Import and Preprocessing**: Loading the dataset, encoding the target variable, and preparing features.
2. **Model Training**: Implementing SVM with multiple kernels, Random Forest, and Neural Network models.
3. **Results and Analysis**: Evaluating and comparing model performances with respect to accuracy.

Data Description

The dataset used in this assignment contains measurements from digitized images of fine needle aspirate (FNA) of breast masses. The target variable (`diagnosis`) classifies tumors as malignant (`M`) or benign (`B`), which is mapped to binary values (`1` for malignant and `0` for benign). The preprocessing includes dropping irrelevant columns (`id`), encoding the `diagnosis`, and scaling features for improved model performance.

```
data['diagnosis'] = data['diagnosis'].map({'M': 1, 'B': 0})
X = data.drop(['id', 'diagnosis'], axis=1)
y = data['diagnosis']
```

Data Insights

The dataset includes 569 instances with 32 columns (30 feature columns, one target, and one `id` column). The feature columns represent various attributes of cell nuclei, such as radius, texture, perimeter, and area.

Model Analyses

SVM Analysis

Support Vector Machine (SVM) was used with different kernels to assess the impact of kernel choice on model accuracy:

- **Linear Kernel**: With an accuracy of **95.6%**, the linear kernel performed well but showed limitations in handling non-linear relationships.

```
svm_linear = SVC(kernel='linear')
svm_linear.fit(X_train, y_train)
y_pred_svm_linear = svm_linear.predict(X_test)
print("SVM Linear Kernel Accuracy:", accuracy_score(y_test, y_pred_svm_linear))
```

- **RBF (Radial Basis Function) Kernel**: Achieved the highest accuracy of **98.2%**, demonstrating its suitability for capturing complex patterns in the data.

```
svm_rbf = SVC(kernel='rbf')
svm_rbf.fit(X_train, y_train)
y_pred_svm_rbf = svm_rbf.predict(X_test)
print("SVM RBF Kernel Accuracy:", accuracy_score(y_test, y_pred_svm_rbf))
```

Random Forest Analysis

The Random Forest model was implemented for its capability to handle high-dimensional data and its effectiveness in reducing overfitting. The Random Forest achieved an accuracy of **96.5%**.

```
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
```

Neural Network Regression

A neural network model was trained using a grid search for parameter tuning, exploring various configurations to improve model performance. With optimized parameters, the neural network model achieved an accuracy of **96.5%**.

```
nn = MLPClassifier(max_iter=1000, random_state=42)
param_grid = {
    'hidden_layer_sizes': [(50,), (100,), (50, 50)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant', 'adaptive'],
}

grid_search = GridSearchCV(nn, param_grid, n_jobs=-1, cv=3)
grid_search.fit(X_train, y_train)
best_nn = grid_search.best_estimator_
y_pred_nn = best_nn.predict(X_test)
print("Neural Network Best Model Accuracy:", accuracy_score(y_test, y_pred_nn))
```

Comparison of Models

The following summarizes each model's performance:

- **SVM**: The RBF kernel achieved the highest accuracy at **98.2%**, making it the most accurate model for this dataset. The linear kernel was also effective but less accurate at **95.6%**.
- **Random Forest**: Achieved a high accuracy of **96.5%** and showed resilience to overfitting with its ensemble method.
- **Neural Network**: With optimized parameters, the neural network model reached an accuracy of **96.5%**. This model benefits from tuning flexibility but requires more computational resources for training.

Performance Summary Table

Model	Kernel/Method	Accuracy (%)
SVM	Linear	95.6
SVM	RBF	98.2
Random Forest	N/A	96.5

Neural Network	Tuned	96.5

Discussion

The high accuracy achieved by the RBF SVM model (98.2%) highlights its ability to capture non-linear relationships, making it highly effective for cancer-type prediction. Neural networks also performed well (96.5%) but required careful parameter tuning. Random Forest balanced accuracy and training efficiency, which is practical when interpretability and speed are priorities.

Accurate cancer-type prediction models have significant implications in clinical diagnosis, where early and precise detection can save lives. These models could aid radiologists by providing preliminary classification, reducing diagnostic time, and enhancing treatment planning.

Conclusion

The RBF kernel SVM model emerged as the best performer in terms of accuracy (98.2%) for predicting cancer types. This assignment underscores the importance of selecting a suitable model and tuning its parameters to meet task requirements. The analysis demonstrated that SVM with an RBF kernel is well-suited for this data, while neural networks offer flexibility with proper tuning.

References

- Scikit-learn documentation for SVM, Random Forest, and MLPClassifier.
- Dataset source: Breast Cancer Wisconsin (Diagnostic) Data Set.