Assignment 2 Report

Name: Pola Gnana Shekar

Roll No: 21CS10052

A Brief Description of Code:

Importing and Manipulating Data:

1. Reads training data, test data, and output data from CSV files.
2. Manipulates the data by dropping unnecessary columns, reorganizing columns, and merging the data frames.

Data Encoding:

Encodes categorical data using LabelEncoder.

SVM Models:

1. Trains Support Vector Machine (SVM) models with different kernel functions (linear, rbf, poly, sigmoid).
2. Scales the features using StandardScaler.
3. Evaluates each SVM model using accuracy, precision, recall, F1-score, and confusion matrix.

Random Forest Model:

1. Trains a Random Forest model.
2. Evaluates the Random Forest model using accuracy, precision, recall, F1-score, and confusion matrix.

Neural Network Model with Hyperparameter Tuning:

1. Conducts a grid search over a range of hyperparameters for a Neural Network model using GridSearchCV.
2. Scales the features using StandardScaler.
3. Evaluates the best Neural Network model found during the grid search using accuracy, precision, recall, F1-score, and confusion matrix.

The code uses print statements to display the results of each model and hyperparameter combination. It identifies and prints the best-performing model with its corresponding hyperparameters and evaluation metrics.

SVM’s roles in cancer prediction:

Support Vector Machines (SVMs) play a crucial role in cancer type prediction by providing a powerful classification method that can effectively separate different classes of cancer based on the features (gene expression profiles) of patients. SVMs are particularly useful for cancer type prediction due to their ability to find an optimal hyperplane that maximally separates the data points belonging to different cancer types. Breakdown of SVMs in cancer type prediction:

Linear Kernel SVM: This kernel is suitable when the data is linearly separable, the linear kernel achieved a high accuracy of 91.18%, indicating that the gene expression profiles of patients with different cancer types can be well-separated by a linear boundary.

RBF (Radial Basis Function) Kernel SVM: The RBF kernel is versatile and is often a good choice when you are not sure about the linearity of the data. The RBF kernel SVM performed less effectively with an accuracy of 58.82%, possibly indicating that the data might not be well-suited for an RBF kernel.

Polynomial Kernel SVM: The polynomial kernel can capture non-linear relationships. It achieved an accuracy of 58.82%, which is like the RBF kernel.

Sigmoid Kernel SVM: The sigmoid kernel is suitable when the data has a sigmoid-like shape and can handle non-linear separations. The sigmoid kernel SVM achieved an accuracy of 85.29%, indicating moderate success in capturing the underlying patterns in the data.

Neural Network Analysis:

Neural networks play a significant role in cancer type prediction due to their ability to capture complex patterns and relationships within high-dimensional gene expression data. Breakdown of the significance of neural networks in cancer type prediction:

Complex Pattern Detection: Gene expression data often contains intricate and non-linear relationships between genes and cancer types. Neural networks can learn and represent these complex patterns, allowing them to make accurate predictions.

Feature Extraction: Neural networks can automatically extract relevant features from gene expression data, reducing the need for manual feature engineering. This ability is crucial when dealing with high-dimensional datasets like gene expression profiles.

Scalability: Neural networks can be scaled to handle large and diverse datasets, making them suitable for analyzing extensive collections of patient data.

Flexibility: Neural networks offer flexibility in model architecture, allowing the incorporation of various hidden layers, activation functions, and hyperparameters. This flexibility enables customization for different cancer prediction tasks.

Implementing a neural network regression model and tuning hyperparameters using grid search:

Create a Neural Network Model: Initialize an *MLPClassifier,* which represents a multi-layer perceptron neural network. This model is a common choice for classification tasks.

Define Hyperparameter Grid: Specify a grid of hyperparameters to search through using *GridSearchCV.* The grid includes hyperparameters like *hidden_layer_sizes, alpha, max_iter, solver, and learning_rate_init*. This step allows you to systematically explore different configurations.

Scale the Features: Standardize the input features using *StandardScaler*. Scaling ensures that all features have a similar scale, which can improve the neural network's training convergence.

Perform Grid Search: Fit the grid search object on the training data. Grid search will explore different combinations of hyperparameters using cross-validation and select the best model based on the specified scoring metric.

Retrieve Best Model and Hyperparameters: Retrieve the best model and its corresponding hyperparameters from the grid search results.

Make Predictions and Evaluate: Use the best model to make predictions on the test data and evaluate its performance using various metrics like accuracy, precision, recall, F1 score, and the confusion matrix.

Display Results: Print the evaluation metrics and the best hyperparameters for the tuned neural network model.

The output provided indicates the best hyperparameters found through grid search and the performance metrics of the tuned neural network model. It is crucial to fine-tune neural network hyperparameters to achieve the best possible predictive performance on your specific cancer type prediction task.

Model Comparison:

Performance Metrics:

The performance of the SVM models with linear, RBF, and polynomial kernels varied significantly:

- The SVM model with a linear kernel achieved an accuracy of 0.912 and an F1 score of 0.930, demonstrating strong predictive capabilities.
- The SVM model with an RBF kernel performed poorly with an accuracy of 0.588 and an F1 score of 0.741, indicating issues with precision and recall.
- Similarly, the SVM model with a polynomial kernel also exhibited suboptimal performance with the same accuracy and F1 score of 0.588 and 0.741, respectively.
- The SVM model with a sigmoid kernel demonstrated moderate performance with an accuracy of 0.853 and an F1 score of 0.889.

In contrast, the Random Forest model delivered competitive performance:

- It achieved an accuracy of 0.882 and an F1 score of 0.909, offering a balanced trade-off between precision and recall.

The Neural Network model with hyperparameter tuning emerged as the top-performing model:

- It achieved a perfect accuracy of 1.0 and an F1 score of 1.0, indicating its exceptional ability to predict cancer types accurately.
- The model's precision, recall, and F1 score were all at their highest values, resulting in a perfect confusion matrix.

Support Vector Machines (SVM)

Strengths:

- Effective in high-dimensional spaces.
- Versatile due to different kernel functions.

Weaknesses:

- Sensitive to hyperparameter tuning.
- Limited scalability with large datasets.

Random Forest

Strengths:

- Handles high-dimensional data effectively.
- Robust to overfitting.

Weaknesses:

- Less interpretable compared to simpler models.
- Limited control over individual decision trees.

Neural Network

Strengths:

- Can capture complex relationships in data.
- Highly adaptable with various architectures.
- Suitable for large datasets.

Weaknesses:

- Requires extensive hyperparameter tuning.
- Can overfit without proper regularization.
- Less interpretable.

Conclusion:

In the context of cancer type prediction using gene expression data, the Neural Network model with hyperparameter tuning clearly stands out as the best-performing model. It achieved perfect accuracy and F1 score, indicating its remarkable predictive power.

While SVM models offer versatility with different kernels, they require careful selection and hyperparameter tuning to perform well. The Random Forest model provided a competitive alternative with a balanced trade-off between precision and recall. However, it couldn't match the Neural Network's predictive performance.

For accurate cancer type prediction, the Neural Network model serves as the preferred choice, showcasing the importance of thoughtful model selection and parameter optimization in machine learning tasks.