AMLBS-Project 6 Report

Objective

This assignment aims to understand and compare the performance of different machine learning models—specifically, Support Vector Machines (SVM), Random Forest (RF), and Neural Network (NN) regression—for predicting cancer types based on a given dataset. By exploring and analyzing these models, we will evaluate their effectiveness in accurately identifying cancer types.

Data Description

The dataset used for this assignment contains 33 columns, each representing various characteristics associated with breast cancer, such as radius_mean, texture_mean, perimeter_mean, and more. The target variable, diagnosis, is labeled as 'M' (malignant) and 'B' (benign), representing the type of cancer. After preprocessing, 'M' and 'B' were encoded as 1 and 0, respectively, to facilitate binary classification. The dataset was split into training and testing sets, and features were standardized for optimal model performance.

Applications of Different Models

SVM Analysis

Support Vector Machines (SVM) are effective for cancer prediction due to their ability to create clear decision boundaries, particularly when data is high-dimensional, as in this case. We used four SVM kernels to compare their performance of which linear and RBF performed the best.

1. **SVM with Linear Kernel**: The linear kernel achieved an accuracy of 94.12%, demonstrating excellent classification capabilities by creating a linear decision boundary. This kernel is often preferred when classes are linearly separable.
2. **SVM with RBF Kernel**: Similarly, the RBF kernel also performerd quite well with an accuracy of 89.43%. The RBF kernel allows for more complex decision boundaries, useful when the data isn't linearly separable.Figu

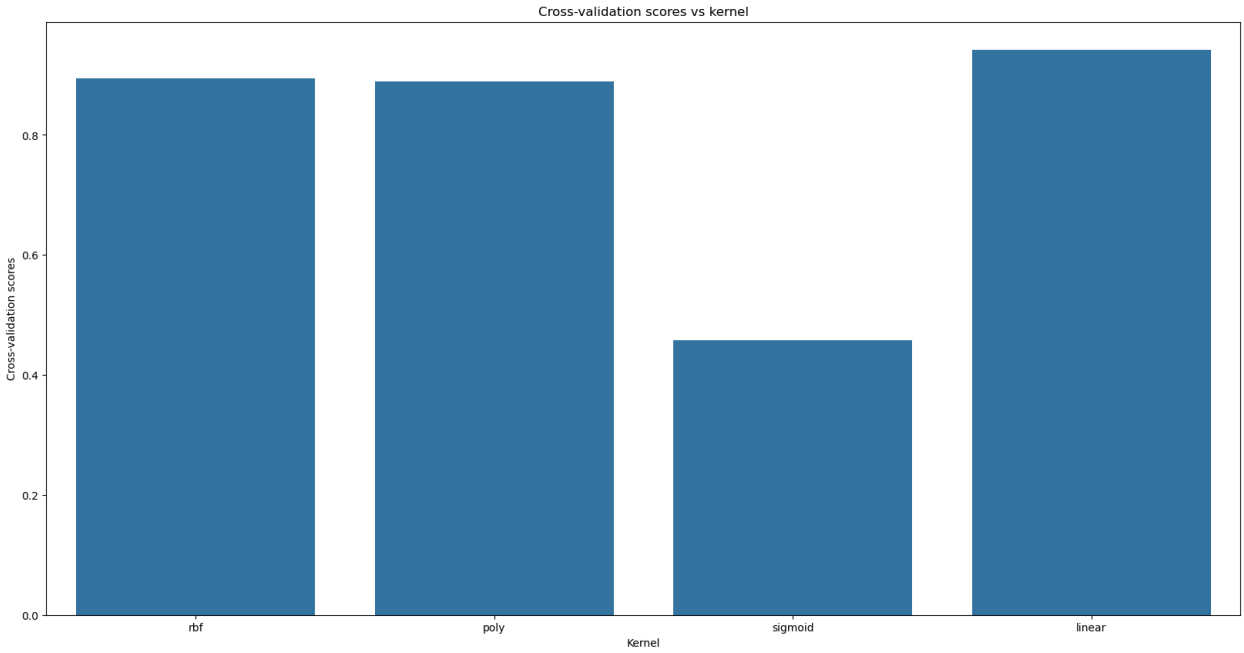

Figure 1: Mean cross-validation score vs SVM kernels

Neural Network Regression Analysis

Neural networks are powerful in handling complex patterns and interactions within data, making them suitable for cancer prediction. Here, a simple feedforward neural network was employed.

1. Significance of Neural Networks: Neural networks can model complex, non-linear relationships that might exist within cancer data, allowing for nuanced predictions.

2. Grid Search for Parameter Tuning: Hyperparameter tuning, often done using grid search, is crucial to optimize the neural network model. Although the initial model performed well with 97.66% accuracy, further fine-tuning could potentially improve the model's convergence and accuracy.

3. Comparison with Other Models: The neural network's accuracy was on par with the SVM models, showing its effectiveness despite the limited iterations (300). Increasing iterations or using more advanced tuning could further enhance its performance.

Comparison of Models

The three models were compared based on accuracy, strengths, and weaknesses.

1. SVM (Linear and RBF):

- Strengths: High accuracy, robustness, and ability to handle high-dimensional data. Both kernels performed equally well.
- Weaknesses: SVMs can be computationally intensive for large datasets, especially with non-linear kernels.

2. Random Forest:

- Strengths: Offers interpretability, handles both linear and non-linear relationships, and performs well with minimal tuning.
- Weaknesses: Slightly lower accuracy compared to SVM and Neural Network, but still highly effective.

3. Neural Network:

- Strengths: Ability to learn complex patterns and adapt to non-linear data structures.
- Weaknesses: Requires more computational resources and careful tuning to avoid overfitting or underfitting. Convergence issues may arise, as indicated by the warning in our initial training.

Most Suitable Model: Based on accuracy, Random Forest Regression performed the best with achieving an almost perfect training accuracy of 0.99 and testing accuracy of 0.99 showing that it has captured almost all the variation that the given set of data has to offer. Also, the near equal training and testing accuracy suggests that the RFRegression has not overfitted. Next, SVM (kernel = linear) and Neural Network models were the almost equal performers, achieving an accuracy of about 95%. However, given the potential for neural networks to capture complex data relationships, they could be favored for more nuanced tasks over SVMs.

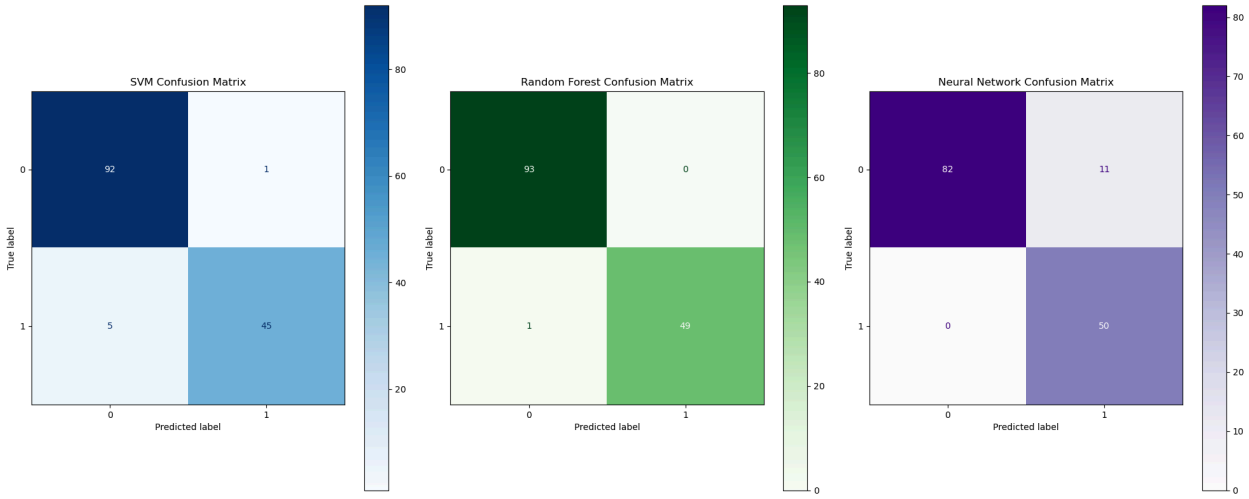

Figure: Classification matrices for the three models

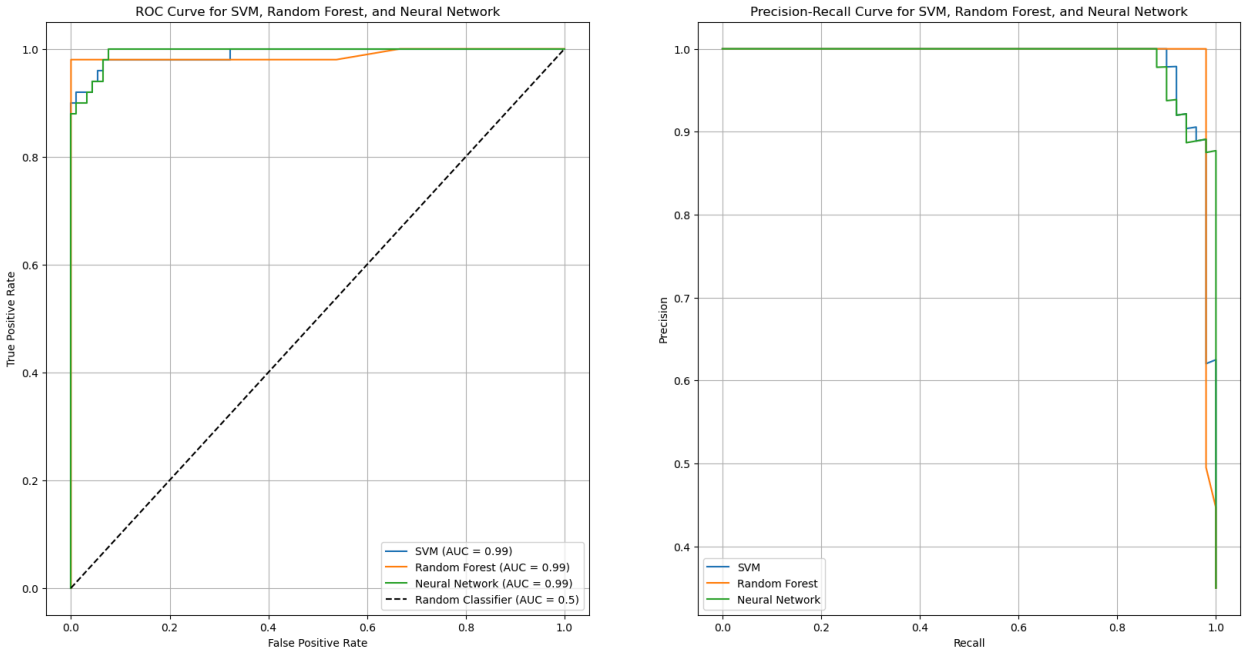

Figure 3: ROC curve (left) and Precision-Recall curve (right) for the three models.

Discussion

Accurate cancer type prediction has profound implications in healthcare, aiding in early diagnosis and personalized treatment plans. By leveraging models like SVM, Random Forest, and Neural Networks, medical professionals can better predict and understand cancer types, enabling faster and more accurate diagnoses.

In real-world applications, Random Forest Regressors could be preferred in scenarios where interpretability and quick decision-making are essential. Meanwhile, neural networks may be better suited for complex cancer datasets that require the model to capture intricate patterns.

Conclusion

In this analysis, Random Forest Regressor model emerged as the top-performing models with an almost perfect accuracy of 0.99%. SVM (linear) and Neural Network followed closely, showing their reliability in classification tasks. The importance of model selection and tuning is evident; thoughtful adjustments to hyperparameters and model structures are essential for optimal performance in cancer prediction tasks.

Overall, these findings underscore the importance of machine learning in medical diagnostics and highlight how different models contribute unique strengths to cancer-type prediction.

Code and Visuals:

All the supporting code has been provided in the same zip file.

References

1. Scikit-learn Documentation
2. Machine Learning in Cancer Prediction
3. Hyperparameter Tuning in Neural Networks