Breast Cancer Prediction: A Comparative Analysis of Support Vector Machine and Tree-Based Machine Learning Models

BY

Seyi Falope

Introduction:

The aim of this report project is to gain a deep and comprehensive understanding of Support Vector Machine models, Decision Tree Models, and Random Forest Models. The objective is to apply these models in breast cancer prediction and conduct a comparative analysis based on feature importance selection and the performance level of each model.

Understanding the Models:

Support Vector Machine: SVM is a robust model that is applicable to both regression and classification tasks. Its primary objective is to determine the hyperplane that effectively separates data points into different classes while maximizing the margin between the hyperplane and datapoints. Additionally, SVM has the capability through its use of kernel functions to handle non-linear tasks

Decision Tree: This model also functions in both classification and regression tasks. It employs an if-else sequence tree method to partition the data based on input values, aiming to maximize the information gain at each split. While Decision Trees are easily interpretable and can handle non-linear tasks, they are susceptible to overfitting due to the absence of randomness in the model.

Random Forest: This model combines multiple decision trees to create a robust ensemble model. It utilizes the bootstrap method to generate subsets of features for individual trees and then aggregates the results by averaging the results for regression tasks or using majority voting for classification problems. This approach effectively addresses the issue of overfitting and also offers insights into the estimation of important features in the data for prediction.

Data Preprocessing:

The following preprocessing steps were implemented in the study based on insights from the Exploratory Data Analysis to ensure that the data is well-suited for optimal model performance.

Label Encoding the Target Class: The target features were represented as 2 for benign and 4 for malignant classes. These were encoded as 0 and 1, respectively, to ensure compatibility with the models and evaluation metrics.

Features Selection Based on Multicollinearity: The uniformity of cell size and the uniformity of cell shape exhibited a high correlation of 0.91, indicating the presence of multicollinearity. To address this, the uniformity of cell size was removed, and the remaining 8 features were used in the models.

Splitting the Data into Train and Test: A train and test splitting ratio of 80:20 was employed. This is crucial to assess how well the models generalize to unseen data.

Outlier Removal Using Z-score: Z score outlier removal technique was applied to the "Mitoses" column to ensure that extreme values which could skew our analysis are excluded

Handling Imbalance Data with Sampling: The benign class and malignant class constituted approximately 65% and 35% of the data, respectively, which could potentially introduce bias in the

model's results. To mitigate this issue, the Synthetic Minority Over-sampling Technique sampling technique was utilized, as illustrated in the image below.

Fig1:

Feature Importance Selection Using RFE (Recursive Feature Elimination)

The RFE technique was applied to each model to evaluate the three most crucial features influencing their predictions. It's worth noting that the predictive decision-making process for SVM differs from that of the tree-based model as illustrated in the table below:

Table 1:

3 MOST IMPORTANT FEATURES DRIVING EACH MODEL PREDICTION		
Support Vector Machine	**Decision Tree**	**Random Forest**
Clump_thickness Uniformity_of_cell_shape Bare_nuclei	Uniformity_of_cell_shape Single_epithelial_cell_size Bare_nuclei	Uniformity_of_cell_shape Single_epithelial_cell_size Bare_nuclei

Evaluating Model Performance Using Cross Validation And Roc-Auc :

Cross Validation techinique was applied to get a robust estimates performance. ROC AUC metric was employed to assess the models' effectiveness in distinguishing between benign and malignant cases, considering the trade-off between True Positive Rate and False Positive Rate . The table and the graph below illustrates the models performance against each others

Mean ROC-AUC Cross Validation Score		Standard Deviation	Performance on Test Data
SVM:	0.983	0.013	0.973
Decision Tree	0.945	0.019	0.969
Random Forest	0.992	0.006	0.983

Fig 2

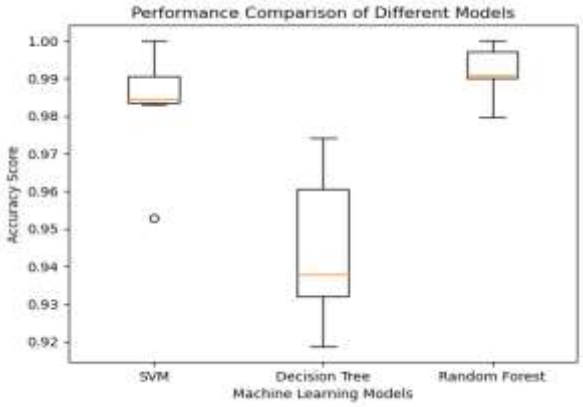
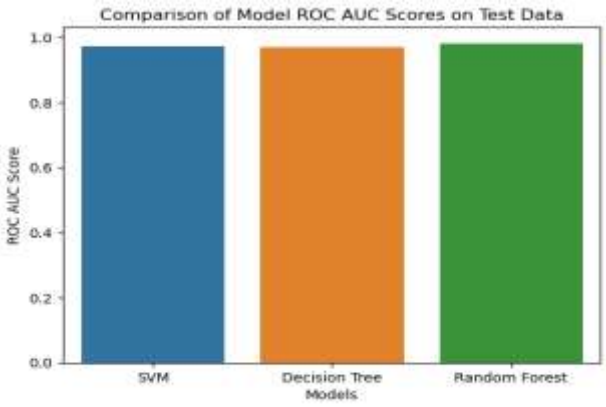


Fig 3



Conclusion

Upon rigorous evaluation, it is observed that through the feature importance selection using RFE method, that the predictive decision-making process for SVM differs from that of the Tree-Based

Models. Allthough, the 3 models performed well, Random Forest model consistently surpassed both SVM and Decision Tree models in cross-validation and test result in predicting breast cancer.

Refrences

1. Evans, J.D.P. (2023). Data Mining and Discovery Lecture Notes, University of Herfordshire, Uk,
2. Cortes, C., & Vapnik, V. (1995). Support-vector networks**.** *Machine Learning, 20*(3), 273-297.
3. Breiman, L. (2001). Random forests**.** *Machine Learning, 45*(1), 5-32.
4. YouTube Channel: StatQuest with Josh Starmer URL: https://www.youtube.com/user/joshstarmer
5. Kaur, R., & Kaur, A. (2019). A survey of breast cancer classification: from traditional to recent emerging techniques. *Expert Systems with Applications, 129*, 71-94.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830. URL: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf