A Comparative Study of Artificial Neural Network, Decision Tree, and Random Forest Classifiers for Breast Cancer Prediction

Name: Friday Odeh Ovbiroro
Student No: 22034665

Introduction:

The objective of this project is to have an understanding of Artificial Neural Networks and Tree-based Machine Learning models, specifically focusing on their application in the prediction of breast cancer. The study aims to apply these models to a classification problem, using them to predict the likelihood of breast cancer in patients.

Understanding the Models:

Artificial Neural Network: An Artificial Neural Network (ANN) is a computational model that draws inspiration from the neural structure of the human brain. It consists of interconnected nodes or neurons that work together to learn from data patterns. ANNs are highly acclaimed for their ability to handle complex tasks such as classification and prediction, making them extremely useful in fields like image recognition and natural language processing.

Decision Tree: This model handles classification and regression tasks, it uses a structured if-else decision-making tree to segment data according to various input variables. Its primary aim at each split is to optimize the information gain. This model stands out for their straightforward interpretability and proficiency in managing non-linear tasks. However, they have a propensity to overfit, a consequence stemming from the model's deterministic nature and lack of randomness

Random Forest: Random Forest is an ensemble model that integrates numerous decision trees to enhance overall performance and robustness. It employs the bootstrap aggregation method, generating different subsets of features for each tree in the ensemble. This methodology effectively counters the overfitting problem often seen in single decision trees. Additionally, Random Forest provides valuable insights into feature importance, aiding in the identification of key predictors within the dataset.

Data Preprocessing:

Based on the findings from the Exploratory Data Analysis, several preprocessing measures were adopted in this study to optimize the data for effective model performance

Encoding the Target Class: The target features were initially labeled as 2 for benign and 4 for malignant classes. To ensure these values were compatible with the modeling and evaluation processes, they were re-encoded: benign (2) was changed to 0, and malignant (4) was changed to 1.

Features Selection Based on Multicollinearity: The uniformity of cell size and the uniformity of cell shape showed a significant correlation of 0.91, revealing multicollinearity in the dataset. To tackle this, the uniformity of cell shape was excluded, leaving 8 features for use in the modeling process.

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: To address the data imbalance where the benign class made up roughly 65% and the malignant class 35%, potentially biasing the model's performance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This approach artificially generates new

examples in the minority class to achieve a more balanced class distribution, as depicted in the subsequent image.

Fig1:

Normalization: Normalization of the data was achieved using the Standard Scaler. This technique adjusted the dataset so that each feature had a mean of 0 and a standard deviation of 1. This standardization is crucial as it brings all the variables to a uniform scale,

Assessment of Model Performance Using Cross Validation and F1 Score:

The Cross Validation method was implemented to ensure reliable and comprehensive performance evaluations. The F1 score, a critical metric combining precision and recall, was utilized to measure the models' proficiency in distinguishing between benign and malignant cases. This balanced approach, as depicted in the subsequent table and graph, highlights the comparative effectiveness of the models, with an emphasis on their consistent performance, free from overfitting.

Table 1

Mean F1 Cross Validation Score		Standard Deviation	Performance on Test Data
ANN:	0.963	0.011	0.952
Decision Tree	0.944	0.020	0.934
Random Forest	0.976	0.012	0.965

Fig 2

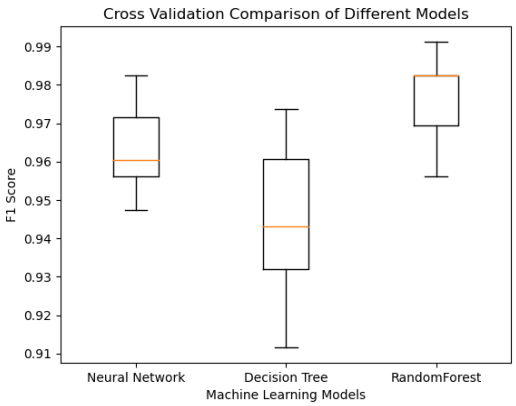
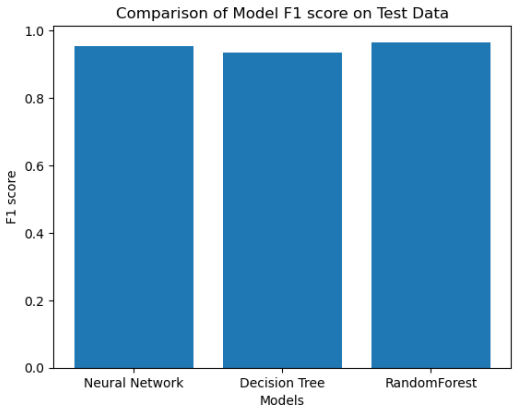


Fig 3



Conclusion

In conclusion, the detailed analysis of the performance scores clearly highlights the Random Forest model's superior capability in predicting breast cancer. It leads the group with the highest F1 scores in both cross-validation and test data, showcasing an optimal balance between bias and variance. This performance distinctly overshadows that of the Artificial Neural Network (ANN) and Decision Tree models, which, despite showing respectable capabilities, fall short of the Random Forest's consistenand accuracy. Therefore, when it comes to reliably and effectively predicting breast cancer, the Random Forest model stands out as the most advantageous choice.

Refrences

1. Evans, J.D.P. (2023). Data Mining and Discovery Lecture Notes, University of Herfordshire, Uk,
2. Cruz, J.A., & Wishart, D.S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2, 59-77.*
3. Dua, S., Acharya, U.R., & Dua, P. (2013). Machine learning for the prediction of breast cancer using a decision tree model. *Journal of Medical Systems, 37(4), 9954.*
4. Karabatak, M., & Ince, M.C. (2015). An expert system for detection of breast cancer based on association rules and neural network. *Journal of Medical Systems, 39(3), 31.*
5. Chen, H.L., Yang, B., Liu, J., & Liu, D.Y. (2017). A support vector machine classifier reduces interscanner variation in the HRCT classification of regional disease pattern in diffuse lung disease: Comparison to a convolutional neural network classifier. *Journal of Digital Imaging, 30(4), 476-482.*
6. Breiman, L. (2001). Random forests. *Machine Learning, 45(1), 5-32.*