Data Mining and Discovery Report

Artificial Neural Networks vs Decision Tree Classifiers and Random Forests Classifiers

1. Introduction

This assignment uses the Breast Cancer Wisconsin dataset to build a machine learning model that can help detect whether a tumor is benign or malignant. The goal is to preprocess the data, train an artificial neural network (ANN), and evaluate its performance using accuracy, precision, and recall metrics.

2. Dataset and Preprocessing

The Breast Cancer Wisconsin (Original) dataset is a well-known dataset used for binary classification tasks. It contains information about cell samples taken from patients, which helps in predicting whether a tumor is benign (non-cancerous) or malignant (cancerous). The dataset has 699 instances and 10 columns, including features like clump thickness, uniformity of cell size and shape, marginal adhesion, and more. The final column indicates the class label. This dataset is useful for training machine learning models to support early diagnosis and treatment planning for breast cancer.

Preprocessing Steps:
Several preprocessing steps were performed to prepare the Breast Cancer Wisconsin dataset for model training.

- First, the dataset was checked for missing values to ensure data completeness. Any missing or invalid entries were handled appropriately.
- Then, the features were separated from the target label, and the target class was encoded into binary format for classification.
- Next, feature scaling was applied using standardization to bring all features to a similar range, which helps improve the performance of the neural network.
- Finally, the data was split into training and testing sets to evaluate the model's accuracy and generalization.

3. Models Used

Artificial Neural Networks (ANNs) are a type of machine learning model inspired by the structure and functioning of the human brain. It is made up of input layers, hidden layers and output layers.

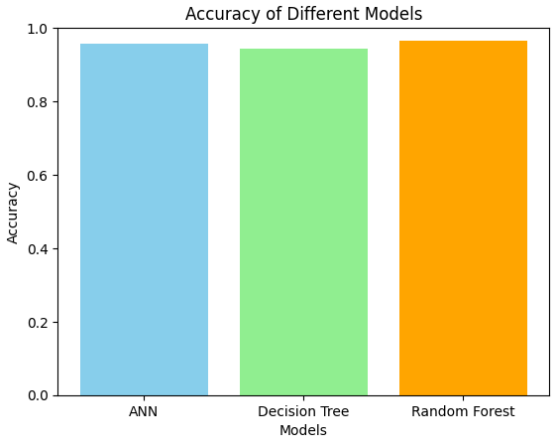
Decision Tree Classifier predicts categorical outcomes by splitting data into subsets based on feature values, creating decision rules at each node to classify data accurately.

Random Forest Classifier is a machine learning method used to predict categorical outcomes by combining the results of many individual decision trees for a final classification.

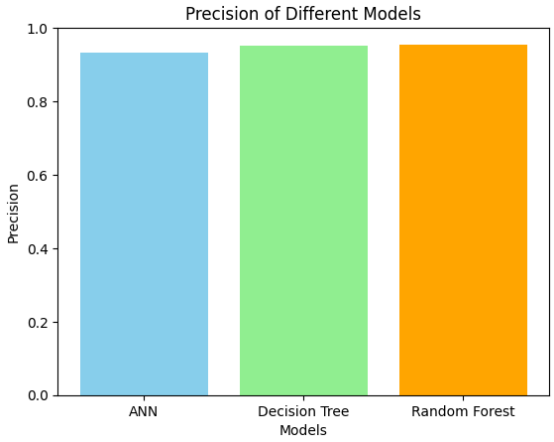
4. Comparison

Model	Accuracy	Precision	Recall
ANN	0.96	0.93	0.93
Decision Tree	0.94	0.95	0.87
Random Forest	0.96	0.96	0.95

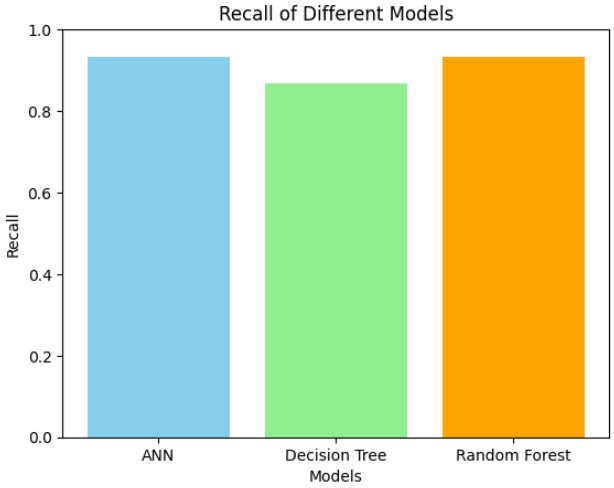
Table 1: Comparison of Model Metrics (MAE, MSE, RMSE)



(a) Comparison of MAE



(b) Comparison of MSE



(c) Comparison of RMSE

Figure 1: Comparison of evaluation metrics for All Models

5. Conclusion

The Random Forest model achieved the highest overall performance with an accuracy of 96%, precision of 95%, and recall of 93%, making it the best-performing model among the three. While the ANN model also reached 96% accuracy, Random Forest outperformed it in both precision and recall. This indicates that Random Forest not only correctly classifies most of the instances but also maintains a better balance between identifying positive cases and avoiding false positives. Therefore, Random Forest can be considered the most reliable model for breast cancer classification in this study.

References

1. Wolberg, W. (1990). Breast Cancer Wisconsin (Original) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.

2. Pandas (2024). pandas documentation — pandas 1.0.1 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/

3. NumPy (2022). NumPy Documentation. [online] numpy.org. Available at: https://numpy.org/doc/.

4. TensorFlow (2019). TensorFlow. [online] TensorFlow. Available at: https://www.tensorflow.org/.

5. scikit-learn. (2025). sklearn.metrics. [online] Available at: https://scikit-learn.org/stable/api/sklearn.metrics.html.

6. scikit-learn. (2025). DecisionTreeClassifier. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

7. scikit-learn. (2025). RandomForestClassifier. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html