Project 8: Decision Trees and Neural Networks on Breast Cancer Prediction

Blake Barber and Sayali Sali

GitHub Repository: Project 8 **Presentation:** Project 8

Introduction:

This project aims to use machine learning techniques, decision trees and neural networks for classification. The objective is to evaluate the performance of these algorithms in predicting the diagnosis of breast cancer based on various features extracted from cell nuclei. Breast cancer is one of the most powerful forms of cancer among women, and early detection plays a crucial role in improving treatment outcomes and survival rates. By leveraging machine learning techniques, we can help medical professionals in making informed decisions, and leading to earlier detection

Dataset:

In this project, we will be focusing on a dataset that involves the classification of cell nuclei as either malignant (M) or benign (B). The dataset comprises 30 real-valued features extracted from cell nuclei, including attributes like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset was very clean and required no additional cleaning by us. Each feature provides valuable insights into the characteristics of the cell nucleus.

Decision Tree

Analysis Technique:

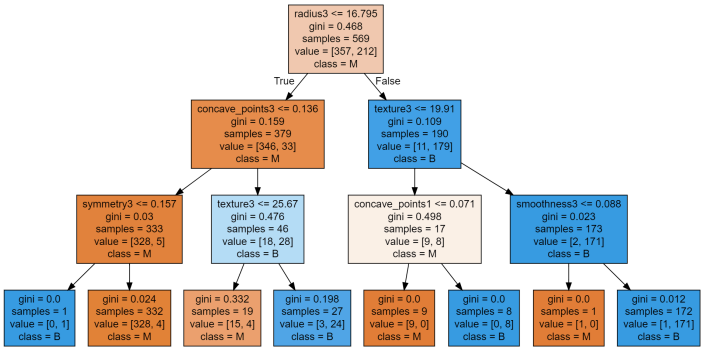
The decision tree classifier is trained to predict whether a breast tumor is malignant or benign based on features extracted from cell nuclei. By visualizing the decision tree, we can understand which features are most important in distinguishing between malignant and benign tumors. Each node in the decision tree represents a feature, and each edge represents a decision based on that feature.

Results:

Two decision trees with different maximum depths (3 and 5) have been generated and visualized.

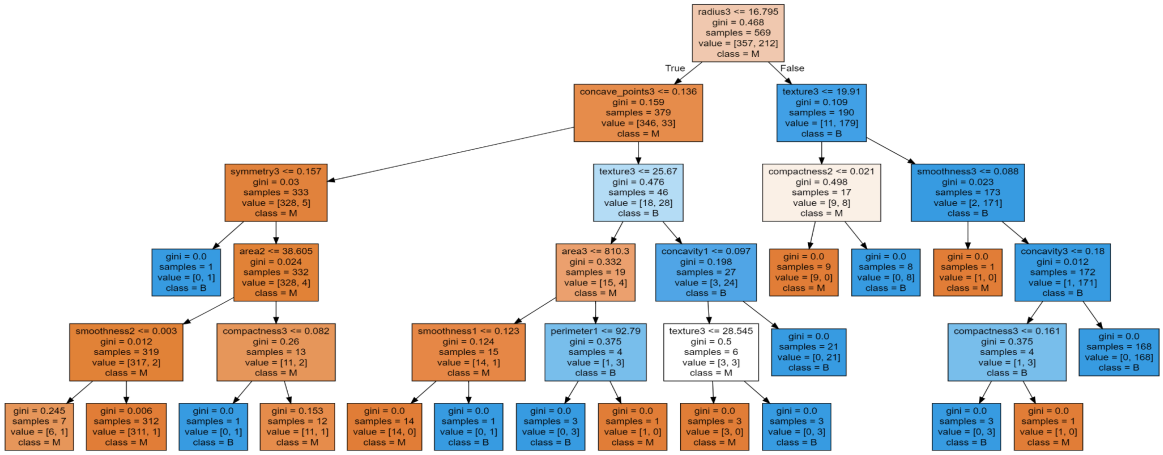
In the decision tree with a maximum depth of 3, the decision making process is relatively simple, with only a few splits. Features like radius, concave points, symmetry,

area, and smoothness play key roles in this simplified model. It focuses on features that appear to have higher importance in distinguishing between malignant and benign tumors.



With a lower depth, the model has high bias as it may oversimplify the relationships in the data, potentially missing important variation. Lower variance compared to deeper trees. Model is less likely to overfit since it's less complex.

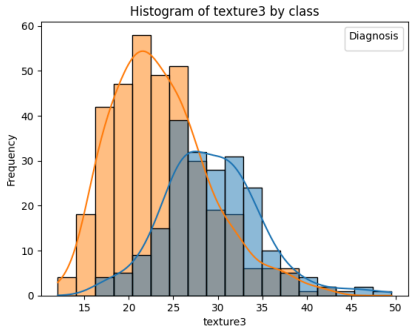
In the decision tree with a maximum depth of 5, the model becomes more complex, with more splits and nodes. It captures smaller details of the dataset, including more intricate patterns and relationships.

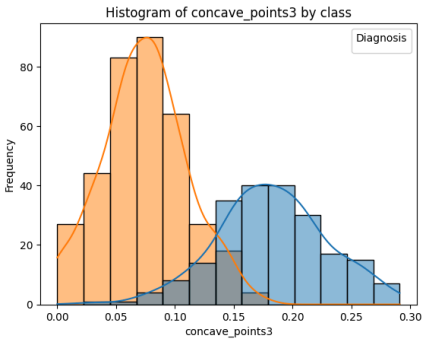


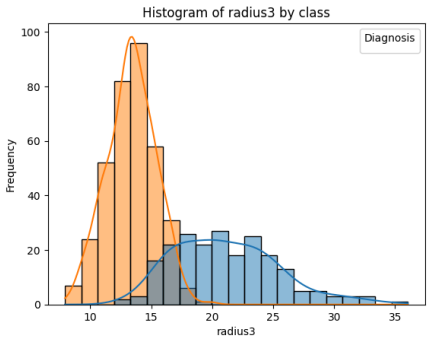
This mode has lower bias compared to the tree with depth 3. Models can capture more complex relationships in the data, reducing bias. Higher variance compared to the shallower tree. It becomes more prone to overfitting as it captures noise in the data.

The feature importance scores calculated from the decision tree indicate the relative importance of each feature in predicting the diagnosis of breast cancer. Features such as 'radius3' 0.710391, 'texture3' 0.082792, 'compactness2' 0.032519 , and 'concave_points3' 0.109375 are identified as the most important features for classification, while other features have negligible importance.

In the below histogram we can see different mean points for both classes and it is easily distinguishable that is the reason this feature is considered to split in the decision tree.







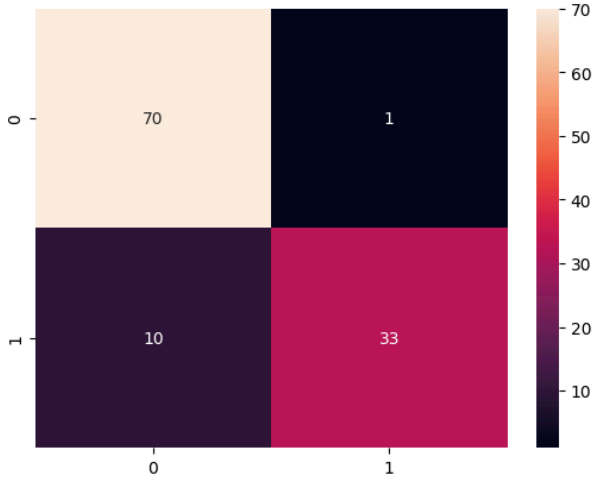
Neural Networks

Analysis Technique:

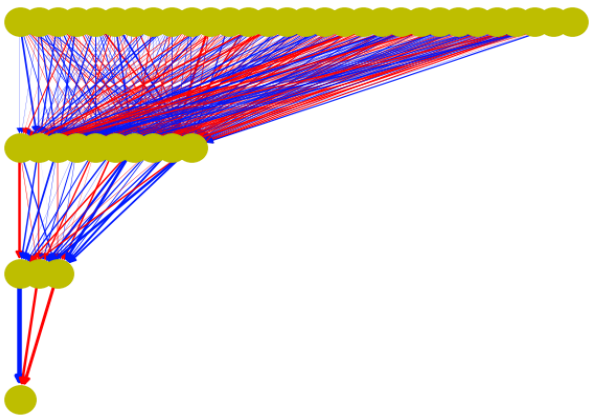
In addition to decision trees, we used neural networks, which are a powerful analysis technique known for their adaptability and proficiency in handling complex pattern recognition and classification tasks. Neural networks have the ability to learn and model non-linear and complex relationships within data, making them suited for medical datasets where such complexities often exist. We experimented with different architectures, varying numbers of hidden layers and nodes, so that we could observe the impact of network complexity on the model's performance. This approach allowed for an improved understanding of how each of the different neural network configurations could affect classification accuracy, which helped to provide a comprehensive view of the dataset's underlying patterns.

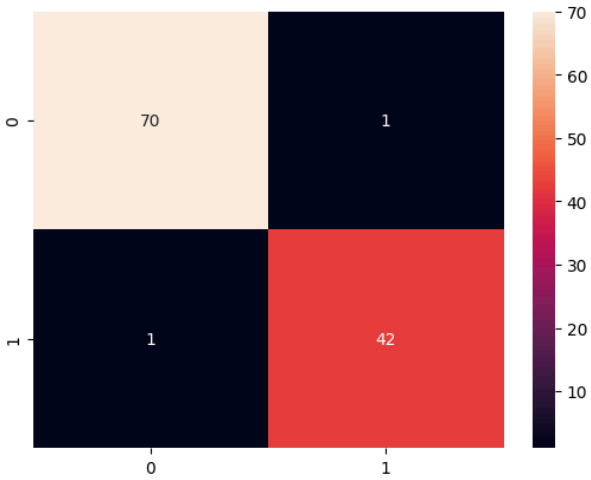
Results:

The neural networks employed for this project provided interesting insights into how the complexity of a network could yield varying results. The first neural network that was created contained a single hidden layer with three nodes. This network surprisingly performed decently, though we believe this is due to allowing for 500 iterations. The single layer neural network obtained an F1 score of 0.93 in predicting the benign class and a score of 0.86 predicting the malignant class. Overall for such a simple network we were impressed with the results. The figure to the left is the confusion matrix for this single layer network.

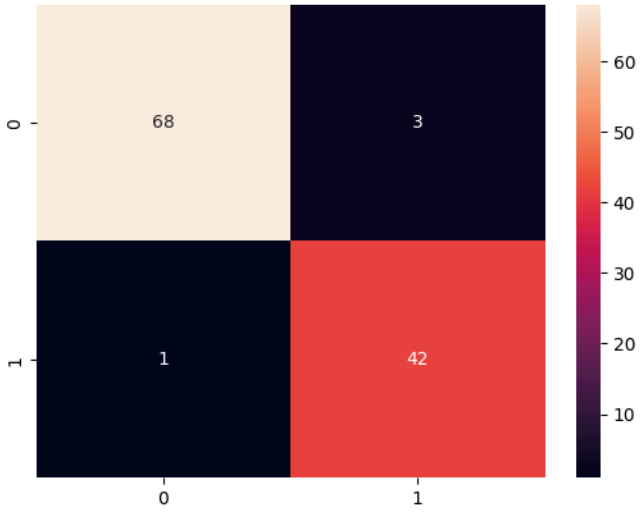


The second structure we tried was a multilayer perceptron network containing two hidden layers with ten nodes and 3 nodes respectively. This network performed the best on our dataset and achieved an F1 score of 0.99 in predicting benign and 0.98 in predicting cancerous. We believe this network successfully captured the intricacies of the dataset while also not overfitting the dataset. Below is both the confusion matrix and the structure for this neural network.





The final network we tried contained three layers with 15 nodes, 7 nodes, and finally 3 nodes respectively. While this network still performed better than the single layer network, we believe that it began to overfit the data. This network achieved an F1 score of 0.97 in predicting a benign tumor and a score of 0.95 in predicting a cancerous tumor. To the left is the confusion matrix for this neural network.



Exploring the Breast Cancer Wisconsin dataset further could involve us employing ensemble methods such as Random Forests or Gradient Boosting, which could further enhance our models predictive accuracy by combining multiple decision trees. Additionally, deep learning techniques, particularly Convolutional Neural Networks (CNNs), could be applied if image data of cellular structures are available, offering potentially superior performance in image-based classification tasks.