DATA MINING AND DISCOVERY

REPORT ASSIGNMENT

NAME – Mukesh Avudaiappan

STUDENT ID – 22024161

STUDY GROUP – 1

REPORT TOPIC – 1

I. Bayesian Network
II. Decision Tree Classifiers and Random Forests

References -

Sahni, A., Singh, S. and Srivastava, G., 2021, July. Exploring Data Science for Highlighting Breast Cancer Prediction Using Python. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC).

Ak, M.F., 2020, April. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In Healthcare (Vol. 8, No. 2, p. 111). MDPI.

Naji, M.A., El Filali, S., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A. and Debauche, O., 2021. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, pp.487-492.

INTRODUCTION:

The aim of this report is to investigate and implement three distinct algorithms—Bayesian Network, Decision Tree, and Random Forest. The analysis is conducted using the breast cancer dataset selected from the available datasets. Bayesian Networks utilize directed acyclic graphs to model probabilistic relationships among variables, which is particularly beneficial for handling uncertainty (Sahni et al., 2021). Random Forests, as an ensemble of decision trees, enhance predictive accuracy through the introduction of randomness in data sampling and feature selection, thereby mitigating overfitting. Decision Trees, while simple and interpretable, may face overfitting challenges without employing techniques like pruning or ensemble methods. The chosen dataset includes various features related to cell characteristics, providing valuable information for predicting the presence of malignancy (1) or benignity (0).

DATA PREPROCESSING: The clean and organised dataset produced by these preprocessing techniques is prepared for additional analysis with Random Forests, Decision Trees, and Bayesian Networks.

1. Data Import and Inspection: A Pandas Data Frame is used to import the breast cancer dataset. To comprehend the data structure, the first few rows should be examined during the initial inspection.
2. Handling Missing Values: Missing values are represented by '?' entries, which are recognised. To guarantee a full dataset, they are substituted with NaN and rows with NaN values are then removed.
3. Target Variable Transformation: To ease the classification work, the target variable 'Class' is converted to binary values (1 representing malignancy and 0 representing benignity).

BAYESIAN NETWORK:

The creation and application of a Bayesian Network for breast cancer analysis is described in depth. After data preprocessing, which includes managing missing variables and visualising patient counts, a directed acyclic graph is used to build the Bayesian network. **Utilising a BDeu prior and Bayesian Estimation**, the model was trained on a dataset that was divided into training and testing sets. For each randomly selected occurrence in the test set, predictions regarding malignancy or benignity are produced. The interpretability and efficacy of the Bayesian Network in capturing probabilistic correlations among variables that are critical for the diagnosis of breast cancer are demonstrated. Reiterating its dependability, the model predicts the class of cases with accuracy. This method not only helps forecast the course of the disease but also offers insightful information about the complex interactions between many characteristics in breast cancer patients.

The vital role of the Bayesian Network in comprehending the intricacies of breast cancer features and presents it as a potent tool for medical decision support.

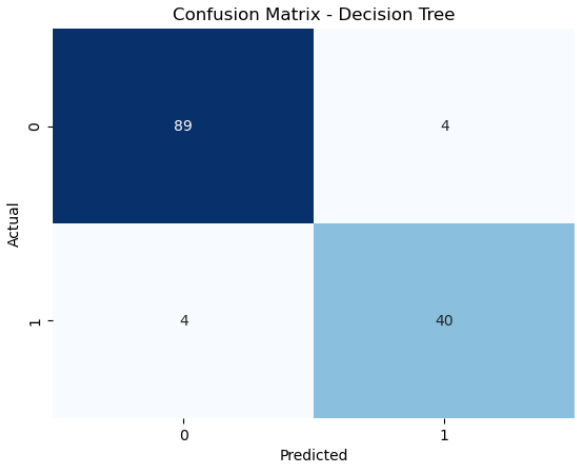
Bayesian Network: Assessing Model Accuracy

Benefits of Using Bayesian Network:

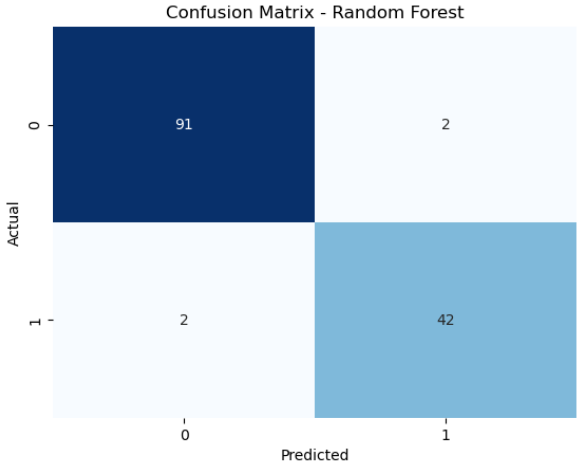
- Accurate prediction is made possible by this probabilistic graphical model, which allows inference about the likelihood of breast cancer based on observable data.

DECISION TREE CLASSIFIERS AND RANDOM FOREST:

The dataset is first divided into feature variables (X) and the target variable (y) for breast cancer classification. Feature scaling is applied using **MinMaxScaler** to normalize the data, and then a train-test split is performed for model evaluation. The **Decision Tree Classifier,** initialized with a random state for reproducibility, **achieves an accuracy of 94%.** The resulting Confusion Matrix illustrates the classifier's ability to predict benign and malignant cases. Subsequently, a **Random Forest** Classifier, also configured with a random state, attains **an even higher accuracy of 97%.** The accompanying Confusion Matrix showcases the robust performance of the Random Forest model. Heatmaps for both classifiers provide a visual representation of their predictive capabilities, aiding in the interpretation and assessment of their diagnostic potential for breast cancer classification.



Decision Tree- Confusion Matrix Heatmap



Random Forest- Confusion Matrix Heatmap

Benefits of Using Decision Tree Classifiers and Random Forest:

- Decision trees are interpretable, making it possible to identify crucial characteristics impacting breast cancer categorization. The decision-making process is made easier to grasp by the tree structure.
- To create a more accurate breast cancer categorization model, it makes use of the collective wisdom of trees.

CONCLUSION:

In summary, this report explores the application of Bayesian Network, Decision Tree, and Random Forest algorithms in breast cancer classification using a well-prepared dataset. Bayesian Networks excel in capturing intricate probabilistic relationships among variables, providing valuable insights into breast cancer features. Decision Trees offer interpretability, aiding in the identification of key characteristics impacting classification, while Random Forest, as an ensemble, enhances accuracy and mitigates overfitting. The preprocessing ensures a clean dataset for effective model implementation. These models exhibit commendable accuracy, with Bayesian Network offering probabilistic predictions and Decision Tree and Random Forest excelling in precise and interpretable classification. Collectively, these algorithms present a holistic approach to breast cancer diagnosis, balancing accuracy, and interpretability. The study underscores the significance of diverse machine learning methodologies in advancing medical decision support systems, contributing to improved breast cancer prognosis, and reinforcing the integration of sophisticated algorithms in healthcare.