Misr University for Science and Technology
College of Information Technology

**Department of Computer Science.** **Spring 2023/2024.**

**Course: CS481 - Pattern Recognition** **Instructor: Prof. Dr. Mohannad Daif**

# Pattern Recognition Documentation

## "Breast Cancer Diagnosis Classification using Decision Trees."

## Done by:

| | |
|---|---|
| **Ahmed Zahran** | **ID: 62511** |
| **Ezzeldin Sayed Abbas** | **ID: 77707** |
| **Amir alaaelden mobasher** | **ID: 77677** |
| **mahmoud khedr** | **ID: 73831** |
| **Omar Ahmad** | **ID: 77674** |

**Introduction**

Breast cancer is a prevalent and life-threatening disease affecting millions of individuals worldwide. Early detection plays a pivotal role in improving patient outcomes, making accurate diagnosis a crucial aspect of medical research and practice. In this context, machine learning techniques offer promising avenues for enhancing the diagnostic process.

Motivated by the imperative to improve breast cancer diagnosis, this project employs a Decision Tree Classifier on a dataset containing various features extracted from breast cancer biopsies. The dataset, acquired from [provide source], encompasses a comprehensive set of attributes that describe key characteristics of tumors.

**Background**

Conventional diagnostic methods involve complex and time consuming histopathological examinations. With the advent of machine learning, there's an opportunity to expedite and refine the diagnostic process by leveraging automated classification algorithms. Decision trees, being interpretable and capable of capturing complex decision boundaries, make them particularly suitable for medical diagnosis tasks.

**Motivation**

The motivation behind this project is to explore the potential of machine learning in improving the accuracy and efficiency of breast cancer diagnosis. By developing a Decision Tree Classifier trained on a diverse set of features, we aim to contribute to the growing body of research focused on enhancing medical diagnostics through computational approaches.

**Problem Definition**

The primary challenge addressed in this project is to accurately distinguish between malignant (cancerous) and benign (noncancerous) tumors based on features derived from biopsy samples. The project seeks to evaluate the efficacy of a Decision Tree Classifier in automating this classification task, with a specific emphasis on achieving high accuracy and interpretability.

Through this exploration, we aim to provide insights into the feasibility and reliability of machine learning models for aiding in the early detection of breast cancer, ultimately contributing to advancements in medical diagnostics.

**Dataset Description**

The dataset employed in this project is sourced from Kaggle, and it constitutes a comprehensive collection of features extracted from breast cancer biopsies. The dataset is curated to facilitate the classification of tumors into malignant (M) or benign (B) categories based on a range of quantitative measurements.

**Attributes:**

**Radius Mean**: Mean of distances from the center to points on the perimeter.

**Texture Mean**: Standard deviation of gray-scale values.

(Additional features...)

The dataset includes various other attributes, capturing essential characteristics of cell nuclei present in breast biopsy samples. Each instance in the dataset represents a distinct biopsy, with corresponding labels indicating whether the tumor is malignant or benign.

**Data Cleaning:**

Prior to analysis, the dataset underwent preprocessing, involving the removal of any irrelevant or redundant information. The "Unnamed: 32" column was eliminated as it did not contribute to the analysis.

**Data Visualization:**

To gain insights into the distribution of malignant and benign tumors, a scatter plot was generated, depicting the relationship between the mean radius and texture. Malignant tumors are represented in red, while benign tumors are represented in lime, aiding in the visual identification of patterns.

**Target Variable:**

The target variable, "diagnosis," has been transformed into numerical values, where 1 corresponds to malignant and 0 to benign. This binary classification enables the utilization of machine learning algorithms for predictive modeling.

**Data Normalization:**

Normalization was applied to scale the feature values between 0 and 1. This step ensures that each feature contributes proportionally to the model's learning process, preventing biases due to differences in scale.

**Train-Test Split:**

The dataset was divided into training and testing sets using a 70-30 split ratio, ensuring an unbiased evaluation of the model's performance.

This dataset forms the foundation for training and evaluating a Decision Tree Classifier, with the goal of creating a robust and accurate model for breast cancer diagnosis. The features encapsulate diverse aspects of tumor characteristics, empowering the model to discern between malignant and benign tumors with a high level of precision.

**Methodology**

The methodology employed in this project involves several key steps, encompassing data exploration, preprocessing, model selection, training, and evaluation. The primary focus is on utilizing a Decision Tree Classifier to achieve accurate and interpretable classification of breast tumors based on the provided dataset.

**Data Exploration:**

The initial phase involves loading the dataset and conducting an exploratory data analysis (EDA). Descriptive statistics and visualizations are utilized to gain insights into the distribution and characteristics of the dataset.

**Data Preprocessing:**

The dataset undergoes preprocessing to ensure its suitability for machine learning. This includes handling missing values, if any, and removing any irrelevant columns. In this project, the "Unnamed: 32" column is dropped as it does not contribute to the analysis.

**Data Visualization:**

Visualizations, such as scatter plots, are created to visually represent the distribution of malignant and benign tumors in the feature space. This step aids in identifying potential patterns and relationships within the data.

Target Variable Transformation:

The target variable, "diagnosis," is transformed into numerical values, where 1 corresponds to malignant tumors and 0 to benign tumors. This binary encoding facilitates the application of classification algorithms.

**Data Normalization:**

Feature scaling is applied to normalize the values of the features between 0 and 1. This step ensures that each feature contributes proportionally to the training process, preventing biases due to differences in scale.

**Train-Test Split:**

The dataset is split into training and testing sets using a 70-30 ratio. The training set is used to train the model, while the testing set serves for unbiased evaluation of its performance.

**Model Selection:**

A Decision Tree Classifier is chosen as the machine learning model for its ability to handle non-linear decision boundaries and provide interpretability. Decision trees are well-suited for medical diagnosis tasks, where interpretability is crucial.

**Model Training:**

The Decision Tree Classifier is trained on the training dataset, learning the underlying patterns and relationships between features and tumor diagnosis. The "fit" method is employed to optimize the model parameters.

**Model Evaluation:**

The trained model is evaluated using the testing set. The accuracy, confusion matrix, and classification report are computed to assess the model's performance in distinguishing between malignant and benign tumors.

**Results Interpretation:**

The interpretability of the Decision Tree model is leveraged to gain insights into the key features influencing the classification decision. Visualization tools may be used to illustrate the constructed decision tree.

By following this methodology, the project aims to develop a robust and interpretable model for breast cancer diagnosis, contributing to the ongoing efforts to enhance medical diagnostics through machine learning techniques. The Decision Tree Classifier serves as a tool for automated tumor classification, providing valuable support to healthcare professionals in the early detection of breast cancer.

**Results and Evaluation**

The trained Decision Tree Classifier exhibited promising results in the classification of breast tumors into malignant and benign categories. The evaluation metrics provide a comprehensive assessment of the model's performance:

**Confusion Matrix:**

The confusion matrix summarizes the model's predictions, highlighting true positives, true negatives, false positives, and false negatives.

**Classification Report:**

The classification report includes precision, recall, F1-score, and support for both classes. Precision measures the accuracy of positive predictions, recall assesses the ability to capture all positive instances, and the F1-score is the harmonic mean of precision and recall.

**Accuracy:**

The overall accuracy of the model on the test set is calculated, providing a general measure of its performance.

These results indicate that the Decision Tree Classifier achieved a high level of accuracy in distinguishing between malignant and benign tumors. The model's precision, recall, and F1-score further demonstrate its effectiveness in both classes.

**References**

1 – www.kaggle.com

2 - Hastie, T., Tibsherany, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

3 - Scikit-learn: Machine Learning in Python. (https://scikit-learn.org/stable/index.html)

4 - Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository. (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

5 - Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, 87(23), 91939196.

6 - Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

7 - Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software (TOMS), 3(3), 209-226.

8 - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É.

(2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

9 - Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

10-James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.