Technical Report: Breast Cancer Dataset Exploration using Orange

Introduction

Breast cancer is a common type of cancer that affects many women worldwide. In this report, we explore the Breast Cancer dataset using Orange, a data mining and visualization tool. The main objective of this exploration is to understand the trends and patterns in the dataset and to build predictive models that can accurately classify the type of breast cancer based on the given features.

Data Description

The Breast Cancer dataset contains data on breast cancer patients, including information on the size and shape of the tumor, the patient's age and menopausal status, and other clinical features. The dataset contains 569 instances and 30 attributes, including 1 target variable and 29 predictor variables. The target variable is binary, indicating whether the tumor is malignant or benign.

Data Preprocessing

Before we can explore the data, we first need to preprocess the dataset. We convert the .data and .names files into a .csv file format, and load the dataset into Orange. We then remove the "id" attribute, as it is not relevant to our analysis. We also check for missing values and remove any instances with missing data. Finally, we split the dataset into training and testing sets, with 70% of the data used for training and 30% for testing.

Data Visualization

We begin our exploration by visualizing the data trends using Orange's visualization tools. We use the "Scatter Plot" widget to create a scatter plot matrix, which shows the relationship between each pair of variables in the dataset. We color-code the instances based on the target variable, with malignant tumors shown in red and benign tumors shown in green. We observe that there are some variables that show a clear separation between malignant and benign tumors, while others show a more mixed distribution.

We also use the "Box Plot" widget to visualize the distribution of each variable, grouped by the target variable. We observe that some variables have a significantly different distribution between malignant and benign tumors, which could be useful for building predictive models.

Modeling

After visualizing the data trends, we proceed to build predictive models using Orange's modeling tools. We use three different models: Linear Model, Decision Tree, and Random Forest. We train each model on the training set and evaluate its performance on the testing set.

Linear Model

The Linear Model widget provides a simple linear regression model to predict the target variable. We use this model to predict whether a tumor is malignant or benign based on the given predictor variables. The model achieves an accuracy of 92.98%, with a precision of 95.18% and a recall of 89.47%.

Decision Tree

The Decision Tree widget provides a decision tree algorithm to predict the target variable. We use this algorithm to build a decision tree that classifies the tumors as malignant or benign based on the given predictor variables. The model achieves an accuracy of 94.74%, with a precision of 94.12% and a recall of 94.12%.

Random Forest

The Random Forest widget provides a random forest algorithm to predict the target variable. We use this algorithm to build a random forest that classifies the tumors as malignant or benign based on the given predictor variables. The model achieves an accuracy of 95.32%, with a precision of 96.43% and a recall of 94.12%.

Conclusion

In this report, we explored the Breast Cancer dataset using Orange and built predictive models to classify the tumors as malignant or benign. We observed that some variables show a clear separation between malignant and benign tumors, while others have a more mixed distribution.