Artificial Intelligence - ITE2010

J-Component

Slot – C2+TC2

Breast Cancer Prediction Using Machine Learning Techniques

Members:

Mohit Kewalramani	15BIT0048
Sai Srivastav Rajput	15BIT0181
Patel Riddhi Parimalkumar	15BIT0258

Objective:

- Predicting the chances of mammograms to be benign or malignant based on the data available about it.
- Building a Keras Deep Learning neural network model for the Breast Cancer data and predict the same based on the trained model.
- Using other Data Mining algorithms like Decision Tree, Naive Bayes, Random Forest, KNN, SVM, and Logistic Regression.

Methodology:

1. Problem Statement:

Mammography is the most effective method for breast cancer screening available today. So we will be analyzing the patterns and predicting the chances of a mammogram to be benign or malignant.

2. Datasets and Inputs:

We'll be using the "mammographic masses" public dataset from the UCI repository (Source: https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass)

This data contains 961 instances of masses detected in mammograms, and contains the following attributes:

- **BI-RADS assessment:** 1 to 5 (ordinal)
- **Age:** patient's age in years (integer)
- **Shape:** mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
- **Margin:** mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
- **Density:** mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
- **Severity:** benign=0 or malignant=1 (binominal)

3. Solution Statement:

The Goal is to train a model with the capability to minimize the risk by making predictions based on the dataset. The model predicts whether mammogram is benign or malignant leading to cancer.

We will apply several different supervised machine learning techniques to this data set, and see which one yields the highest accuracy as measured with K-Fold cross validation (K=10). Algorithms to be used:

- Decision tree
- Random forest

- KNN
- Naive Bayes
- SVM
- Logistic Regression
- a neural network using Keras.

4. Evaluation Matrices:

   To validate the model and determine its performance, the actual information from the data will be compared with the predictions, and the Recall and precision will be analyzed. Those two are calculated by looking at four things:

   - **True positive:** accounts got defaulted and model predict them as defaulted
   - **False positive:** accounts did not defaulted and model predict them as defaulted
   - **True negative:** accounts did not defaulted and model predict them as not defaulted
   - **False negative:** accounts got defaulted and model predict them as not defaulted
   - **Cross Validation score (cv_score) :** In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality.

Design:

1. **Data Preprocessing:** Will check the data and prepare it by removing data input errors, missing values. Also checking for outliers and removing 5% - 10% or the leading and tailing data narrow the outliers

2. **Data Processing:** Will visualize the data and check the correlation and try to find pattern first visually. After that one of the supervised learning Algorithm will be use. It is yet uncertain whether supervised learning will be employed.

3. **Training and Testing:** will split the data and do 75% training and 25% for testing and to ensure the model is robust and we getting a result we can trust a cross validation will be applied.