MINI-PROJECT

ARTIFICIAL INTELLIGENCE(CO401)

TOPIC: CANCER PREDICTION USING SUPERVISED LEARNING

SUBMITTED TO:

SHOBHANJANA KALITA

SUBMITTED BY:

AKSHAT GUPTA (CSB20038)

SHIVANI TIWARI (CSB20044)

TANUSHREE DAS (CSB20092)

B.TECH 7TH SEMESTER

DEPT:- COMPUTER SCIENCE & ENGG

Introduction

Cancer has identified a diverse condition of several various subtypes. The timely screening and course of treatment of a cancer form is now a requirement in early cancer research because it supports the medical treatment of patients. Many research teams studied the application of ML and Deep Learning methods in the field of biomedicine and bioinformatics in the classification of people with cancer across high- or low risk categories. These techniques have therefore been used as a model for the development and treatment of cancer. As, it is important that ML instruments are capable of detecting key features from complex datasets. Many of these methods are widely used for the development of predictive models for predicating a cure for cancer, some of the methods are artificial neural networks (ANNs), support vector machine (SVMs) and decision trees (DTs).

Objective:

Breast Cancer is one of the most dreadful diseases and is a potential cause of death in women. Late prediction of Breast Cancer may greatly reduce survival chances, and as a solution to that the automatic disease detection system aids the medical field to diagnose and analyse, which offers rapid response, reliability, effectiveness as well as decrease the risk of death. In this paper, we explain how breast cancer can be predicted using a Machine Learning Technique named Random Forest Classifiers.

Here's a roadmap for cancer prediction using supervised learning.

1. **Data Collection**:
   Gather a comprehensive dataset that contains relevant features(variables) and labelled outcomes (cancer or no cancer).
2. **Data analysis:**
   Data visualization is an extremely necessary skill in machine learning. It is responsible for providing a qualitative understanding of the given data, visualizations we use in this project are a heat map. Heat maps are very convenient to use when understanding complex data sets. Heat map is a 2D representation of data in which values can be represented by different color schemes.
3. **Data Preprocessing:**
   Data preprocessing is the conversion of unstructured data into structured data. The steps in data preprocessing are
   - Reading the dataset
   - Check for missing values and fill with required data
   - Splitting data into dependent and independent data.
   - Label encoding
   - One hot encoding (Binarization)
   - Splitting the independent and dependent values into train and test set.
4. **Proposed Algorithm- Random Forest Classifier:**
   Random forest classifier initially selects a random subset of data and generates numerous decision trees. It then summarizes the votes from different decision trees and then takes the decision to decide the final classification of the test object.

Parameters for Random Forest Classifier:

- n_estimators: Specifies the number of trees we use in the algorithm.
- criterion: "Gini" and "Entropy" are the two criterions that we can use.
- min_samples_split: Refers to the minimum number of working set size at node that is required to split. Default value is taken to be 2.

5. **Evaluation Metrics:** Evaluation metrics are responsible for explaining the performance of a model. It is extremely important to check the accuracy of the model before computing the predicted values. Different types of metrics can be considered while evaluating the models. We can evaluate using Confusion Matrix.

Conclusion:

The Random Forest classification algorithm is implemented to train and test the model. The prediction can be 1 or 0. If the prediction obtained is 1, the result displayed on the user interface is ‘Person is at risk of being diagnosed with breast cancer in the future’. If the prediction obtained is 0, the result displayed on the user interface is ‘Person is not at risk of being diagnosed with breast cancer in the future’. Thus, classification takes place.