Breast Cancer Prediction

Machine learning Project Report

1- Introduction

Breast cancer can occur in women and rarely in men. Symptoms of breast cancer include a lump in the breast, bloody discharge from the nipple and changes in the shape or texture of the nipple or breast. Treatment depends on the stage of cancer. It may consist of chemotherapy, radiation and surgery.

2- Problem Definition and Algorithms

2.1 Task Definition

The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy.

2.2 Algorithm Definition

Four algorithm XGBoost, Random Forest, LightGBM and Decision Tree which predict the breast cancer outcome have been compared in the notebook using only one dataset.

3. Experimental Evaluation

3.1 Methodology

Phase 1- Pre-Processing Data

The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format.

Phase 2- Data Preparation

Data Preparation, where we load our data into a suitable place and prepare it for use in our machine learning training. We'll first put all our data together, and then randomize the ordering.

Phase 3- Features Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction.

Phase 4- Feature Projection

Feature projection is transformation of high-dimensional space data to a lower dimensional space (with few attributes). Both linear and nonlinear reduction techniques can be used in accordance with the type of relationships among the features in the dataset.

Phase 5- Feature Scaling

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Phase 6- Model Selection

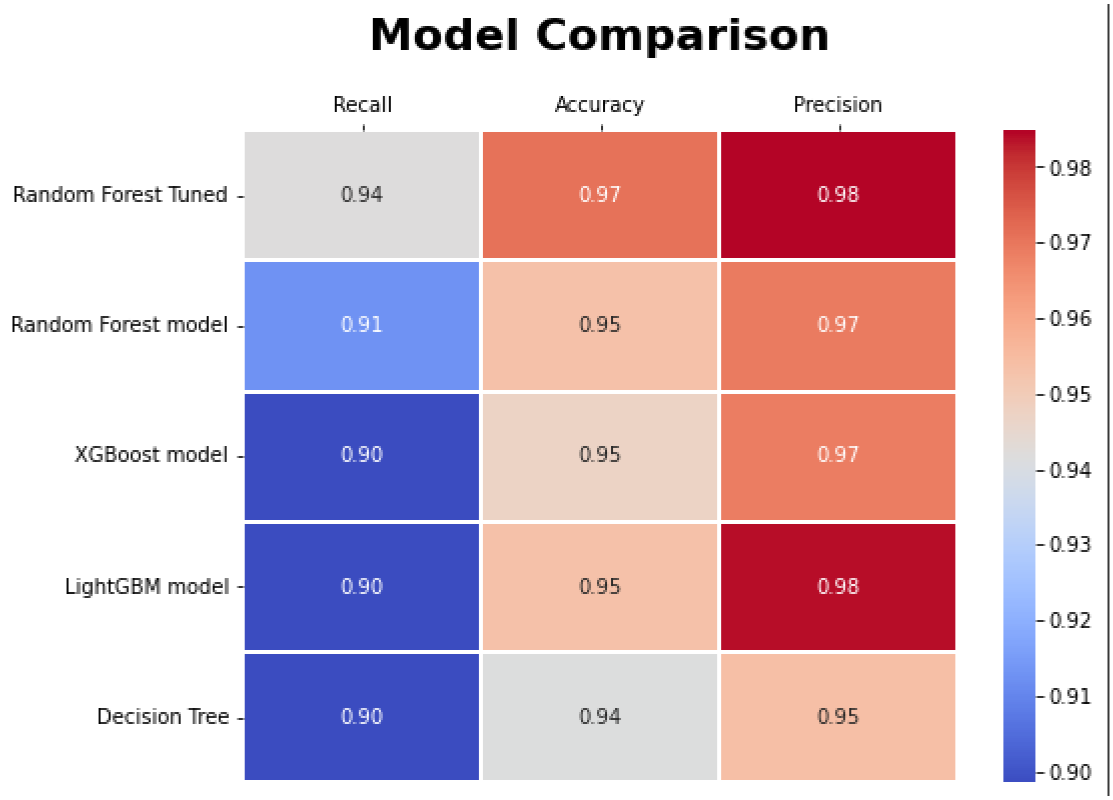
Supervised learning is the method in which the machine is trained on the data which the input and output are well labelled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques. A regression problem is when the result is a real or continuous value, such as "salary" or "weight". A classification problem is when the result is a category like filtering emails spam" or "not spam".

Phase 7- Prediction

Machine learning is using data to answer questions. So, Prediction, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is real.

3.2 Results

After phase 7 we did a model comparison to pick the best model based on accuracy and recall.


Model Comparison
Recall
Accuracy
Precision
Random Forest Tuned
Random Forest model
XGBoost model
LightGBM model
Decision Tree