BREAST POWER Awareness - Model on Breast Cancer

By Sana Sharma

Introduction & Abstract

- Significant public health problem
- Most common cancers
- Early diagnosis can improve a chance to survival
- Accurate classification of tumors –
  - Malignant groups
  - Benign groups
- I will use machine learning algorithms: Logistic Regression, K-Nearest Neighbors, Decision Tree and Random Forest

Motivation

- Model selection
- The dataset consists of two set of values, either M (Malign) or B(Benign)
- Use Classification algorithm
- My goal of the research is to assist clinicians in BC screening and detection
- Machine Learning, has proved to play a vital role in predicting diseases such as cancers
- In the medical field, these methods have been used to predict and to make decisions.

Research Question

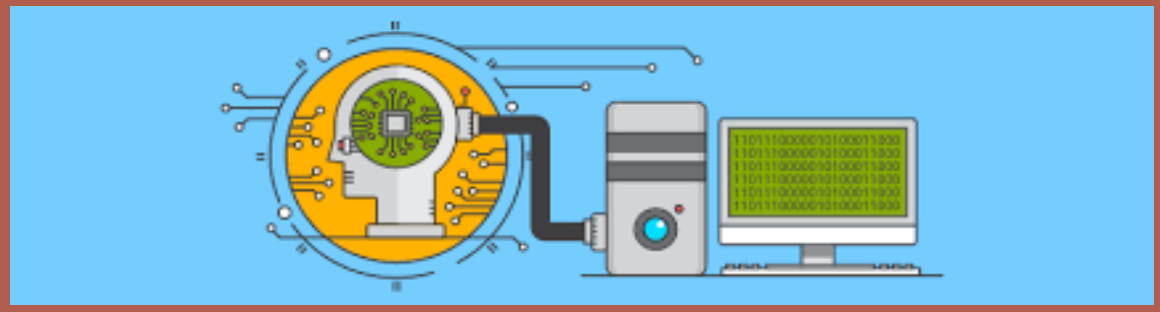
Which ML model from this study best enables the prediction of BC using the classification algorithm of supervised learning by testing and training the data?

Proposed Method

- Exploratory data analysis and machine learning techniques
- SciKit Learn, Classification Algorithm
- Confusion Matrix

Dataset Information

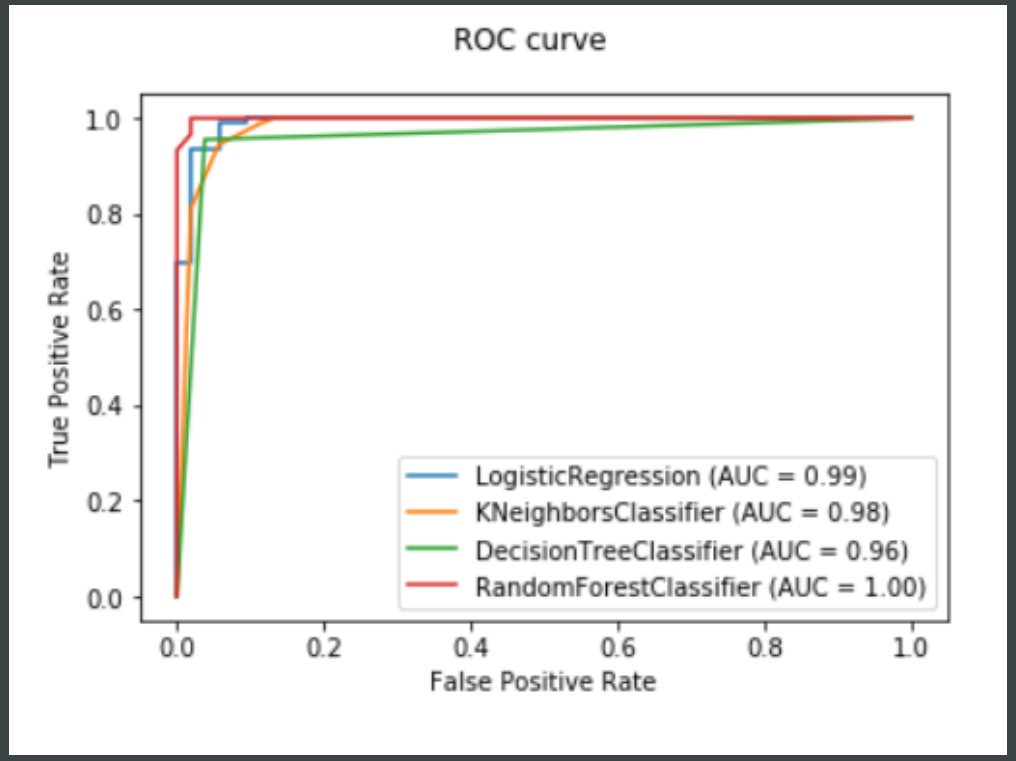
- Breast Cancer Wisconsin dataset, created by Dr. William H. Wolberg
- sklearn.data website, UCI Machine Learning website and Kaggle.
- Attribute Information:
  - It has 569 entries and 30 columns.
  - Diagnosis (M/0 = malignant, B/1 = benign)



Experiments

- Jupyter notebook – python programming language
- Exploratory analysis
  - NaN values
  - Dataframe Info, Attributes
- Proposed Method
  - Logistic Regression
  - KNN
  - Random Forest Regressor
  - Decision Tree Regressor
  - Confusion matrix

Results & Conclusion



- The accuracy of each model while training is predicted as below:
    1. Logistic Regression - 0.96
    2. K-Nearest Neighbors - 0.95
    3. Decision Tree - 0.96
    4. Random Forest - 0.97
- The accuracy of each model while testing is predicted as below:
    1. Logistic Regression - 0.96 (+/- 0.03)
    2. K-Nearest Neighbors - 0.94 (+/- 0.02)
    3. Decision Tree - 0.91 (+/- 0.05)
    4. Random Forest - 0.94 (+/- 0.05)
- Model is over fitting, needs more work

Limitations and later work

- The model is over fitting
- work on keeping the model simple
  - reduce variance
  - use regularization techniques

Hence, work on a better classification model that can prevent overfitting with this dataset.

Thank you!
- Sana Sharma