Comparative Analysis of the Supervised Machine Learning Methods for the Prediction of the Breast Cancer

Drashti Mehta

Background

Breast cancer is the second most prominent cancer found in women after the skin cancer. It is a cancer that starts in breast tissue. It happens when cells in the breast change and grow out of control. The cells usually form a tumor.

Causes:
Breast cancer happens when there are changes in the genetic material (DNA). Often, the exact cause of these genetic changes is unknown. But sometimes these genetic changes are inherited, meaning that you are born with them. Breast cancer that is caused by inherited genetic changes is called hereditary breast cancer. Besides genetics, your lifestyle and the environment can affect your risk of breast cancer.

Aim

The aim of this study is to have a comparative analysis for the machine learning methods to check which method fits well for the prediction of Breast cancer.

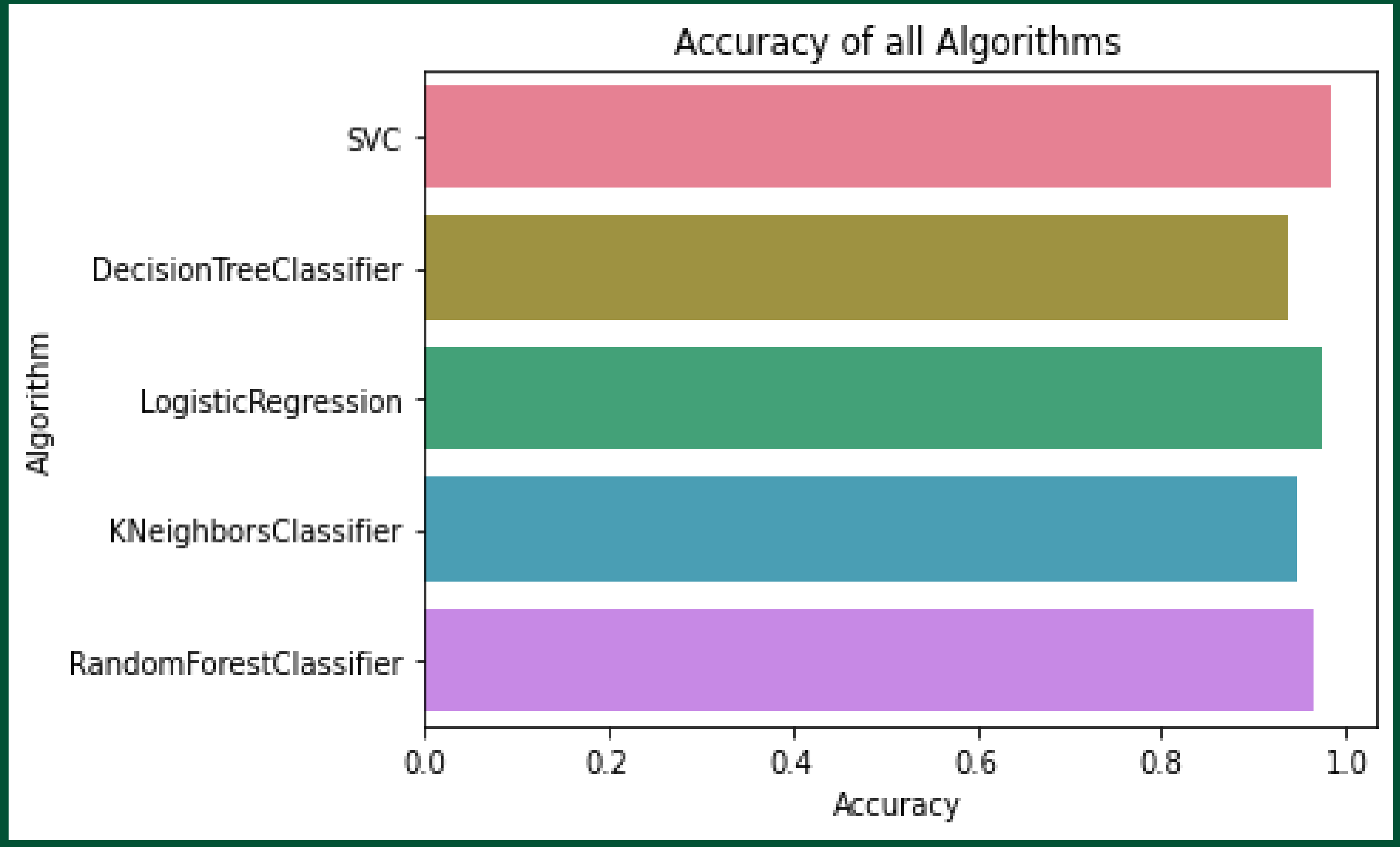
Methodology

The supervised machine learning methods used for the study are SVM, Logistic Regression, Decision Tree, K-Nearest Neighbor and Random Forest with the help of sklearn library.

SVM: Support Vector Machines are models that analyze data for classification and regression for linear and non – linear data by creating a hyper-plane.
Logistic Regression: This estimates the parameters of a model which models the **probability** of an event taking place. Usually used for prediction or classification.
Decision Tree: Machine learning approach for classification or regression used as a predictive model. They **break down** complex data into simpler form, used when evaluation method is needed.
K-Nearest Neighbor: Machine learning method which uses the **proximity** to make classifications or predictions about grouping of an individual data point.
Random Forest: Machine learning methods that generates multiple decision trees and gives output based on those.

UNC CHARLOTTE

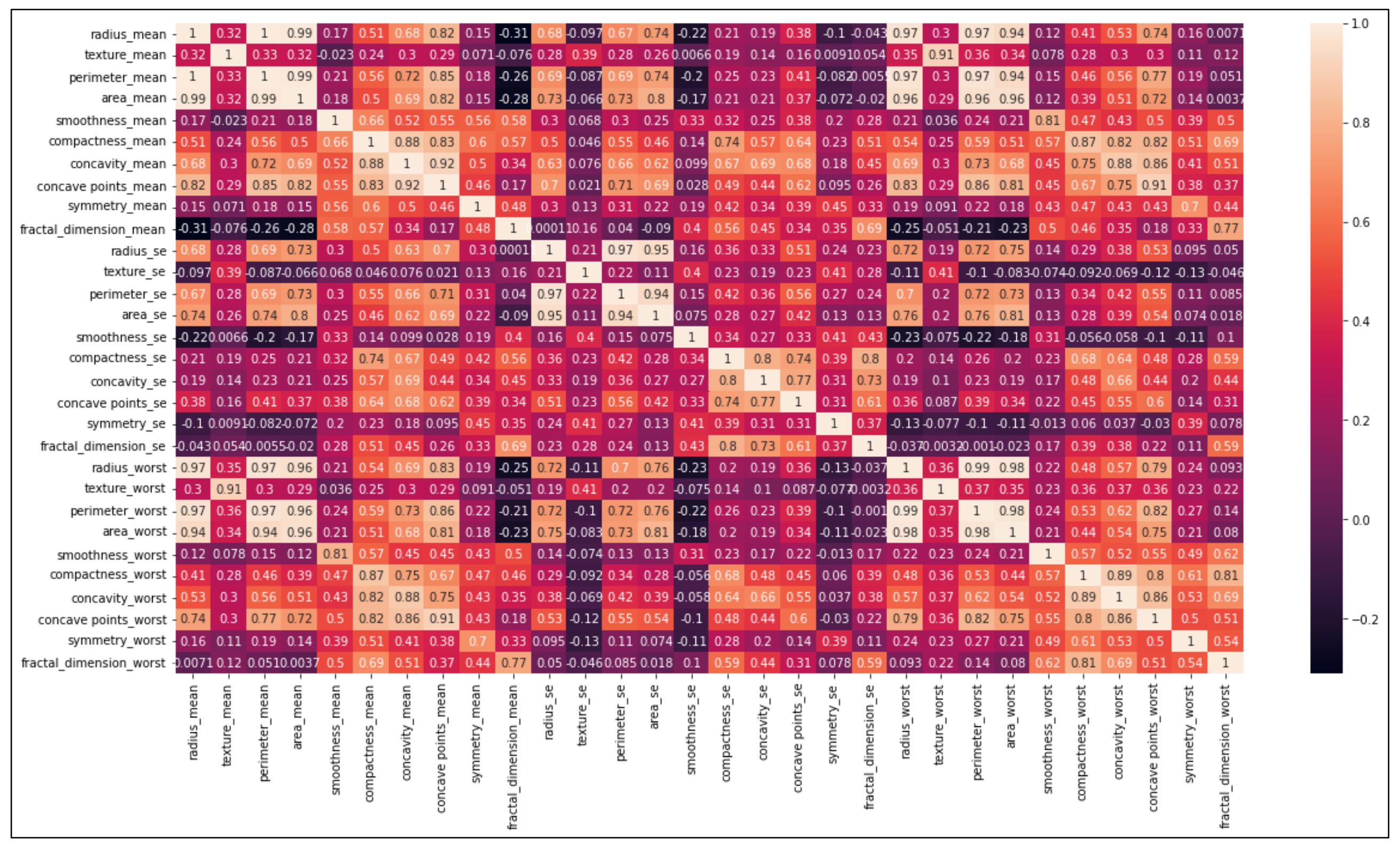
The SVM (Support Vector Machine) proved to be the most accurate method for the prediction of the Breast Cancer (Malignant or Benign) amongst other methods



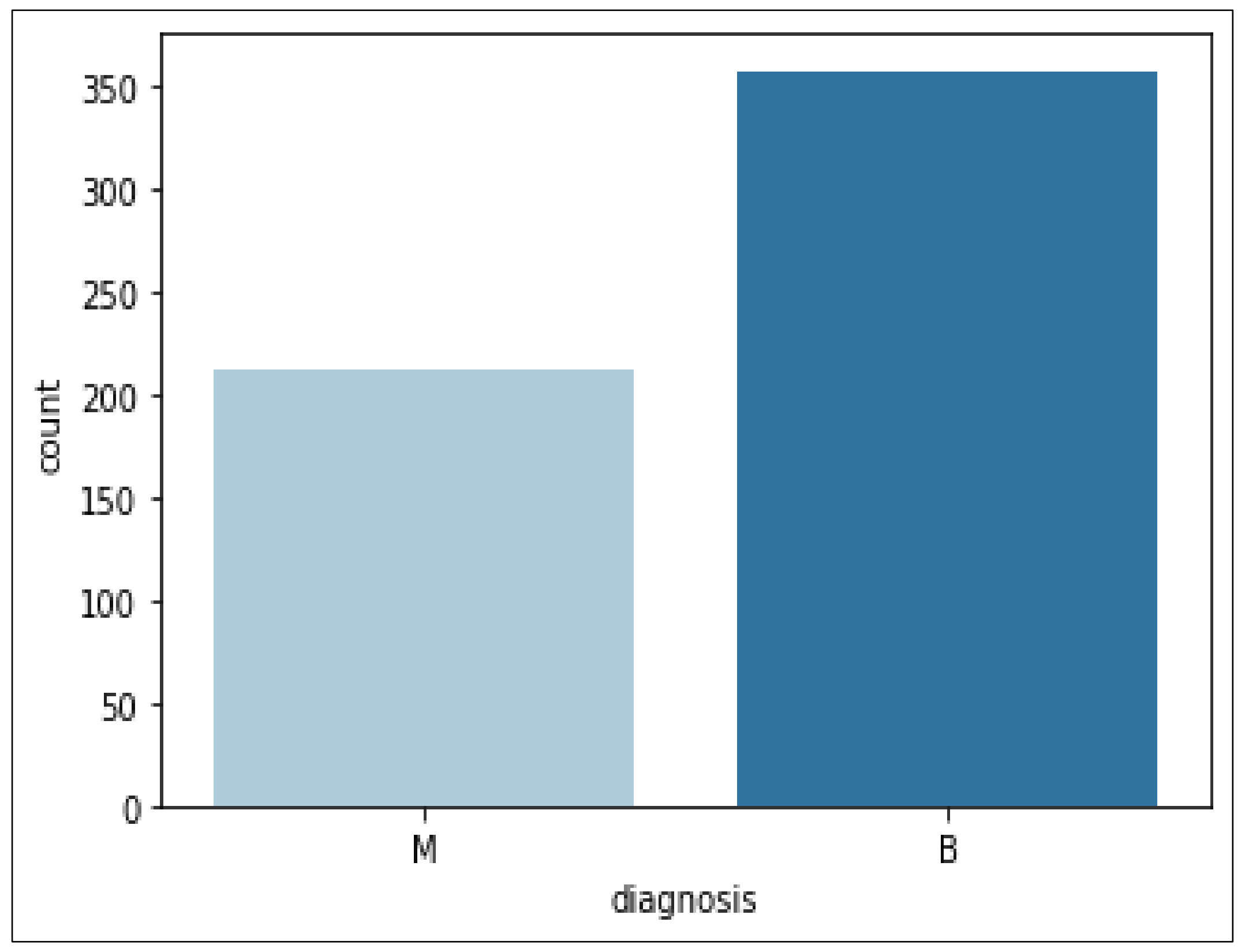
Scan for supplementary reading

Why SVM?

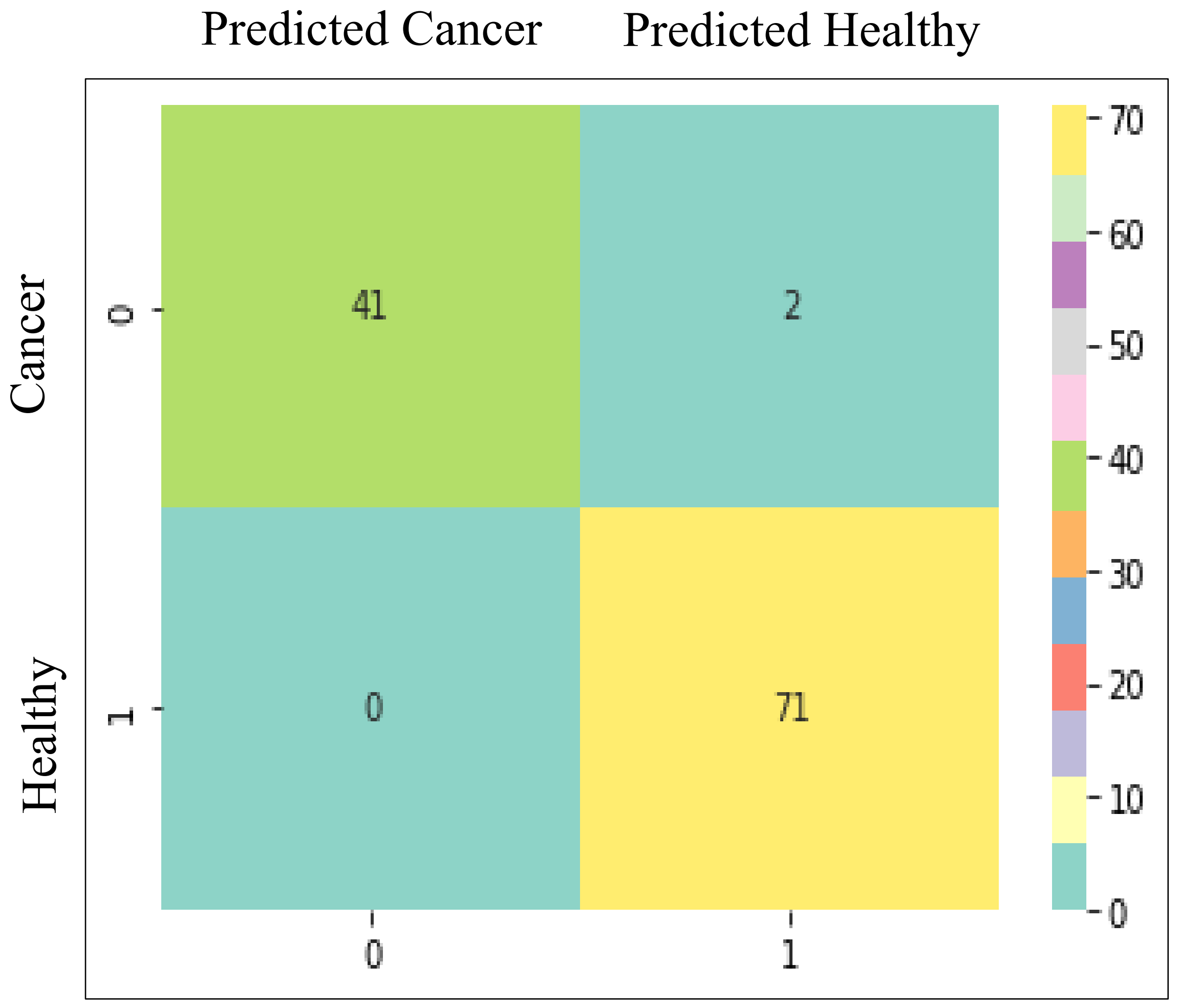
Decision Tree classifier has the lowest accuracy and it causes overfitting of the data. KNN depends on the proximity and scale of the data which does not work with large dataset. As some features of the data are linearly correlated, Random Forest classifier does not depict high accuracy. The result from Logistic Regression is nearly similar to the SVM accuracy, but SVM works better with outliers.



Ratio of Malignant and Benign types



Confusion Matrix (SVC)


Predicted Cancer
Predicted Healthy
Cancer
Healthy