Breast Cancer Prediction
Using Python & Machine Learning Techniques

Jesmeen Fatema

May 2020

Business Problem

- Detection of breast cancer is extremely important to save women lives
  - Breast cancer is a common cancer for women around the world.
  - Early detection of breast cancer can greatly improve the prognosis and survival chances of the patients by implementing proper course of treatments.
- Create a model that can accurately predict / classify whether a patient has breast cancer.

Data Description

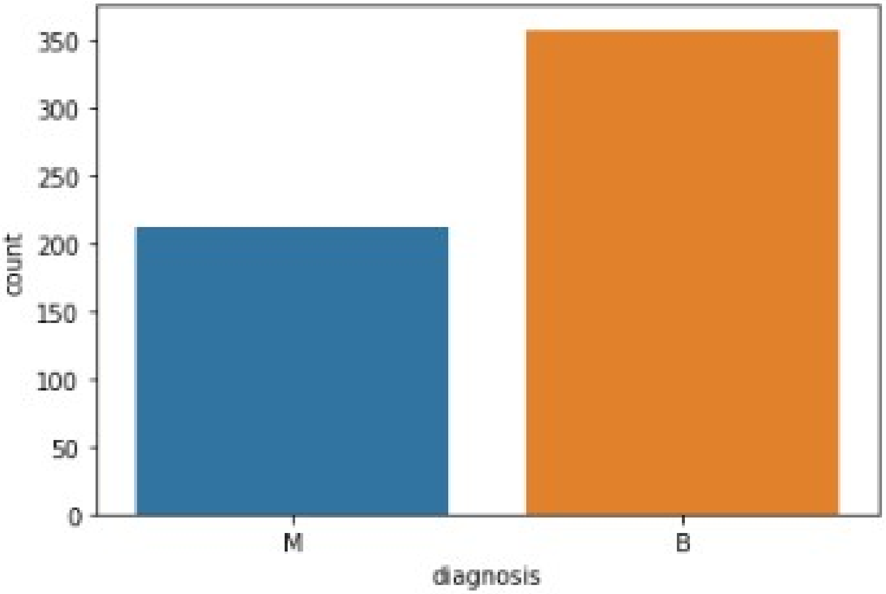
- Sources of Data
  - The data set that will be used in this analysis will come from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data).

Methodology / Data Analysis

- Downloaded (csv file) the data from Kaggle.
- Imported relevant libraries that would be used throughout the program.
- Analyzed the data.
- There are 569 patients and 33 data points on each patient
- None of the columns contain any empty values except the column named ‘*Unnamed: 32*’ , which contains 569 empty values (the same number of rows in the data set, this column is useless for the analysis).

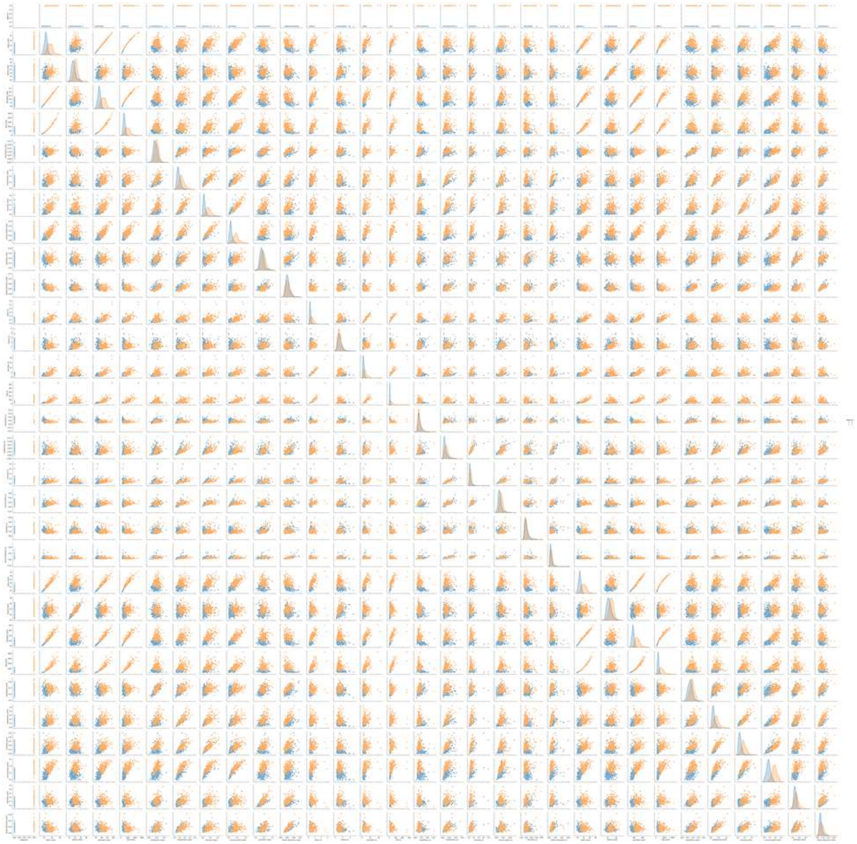
Methodology / Data Analysis (Contd.)

- Based on the data set 212 patients have breast cancer (Malignant -> M) and 357 do not have breast cancer (Benign -> B).



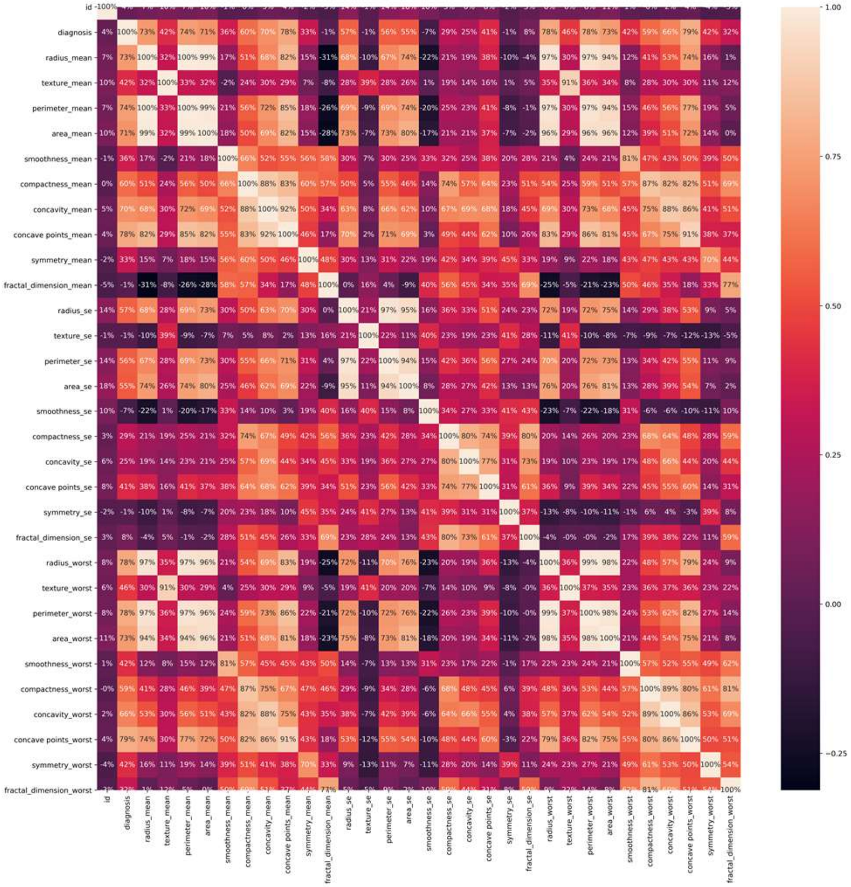
Methodology / Data Analysis (Contd.)

- **Pair plot** to explain a relationship between two variables. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our dataset.



Methodology / Data Analysis (Contd.)

- Heatmap to show the correlations of the columns



Data Processing & Cleaning

- Dropped unnecessary columns (i.e. Unnamed: 32)
- Covert the categorical data (‘diagnosis’ column) to numerical data types by ***Label Encoding***
- Split data into independent / feature (X) and dependent (Y) variables
- Standardizing / scaling the features (X)
- Split the data into training (75%) and testing (25%) data sets

Create Machine Learning Models

- Created a function for different machine learning models to identify which model performs the best :
  - Logistic Regression
  - KNN Classifier
  - Support Vector Classifier
  - Gaussian NB
  - Decision Tree Classifier, &
  - Random Forest Classifier

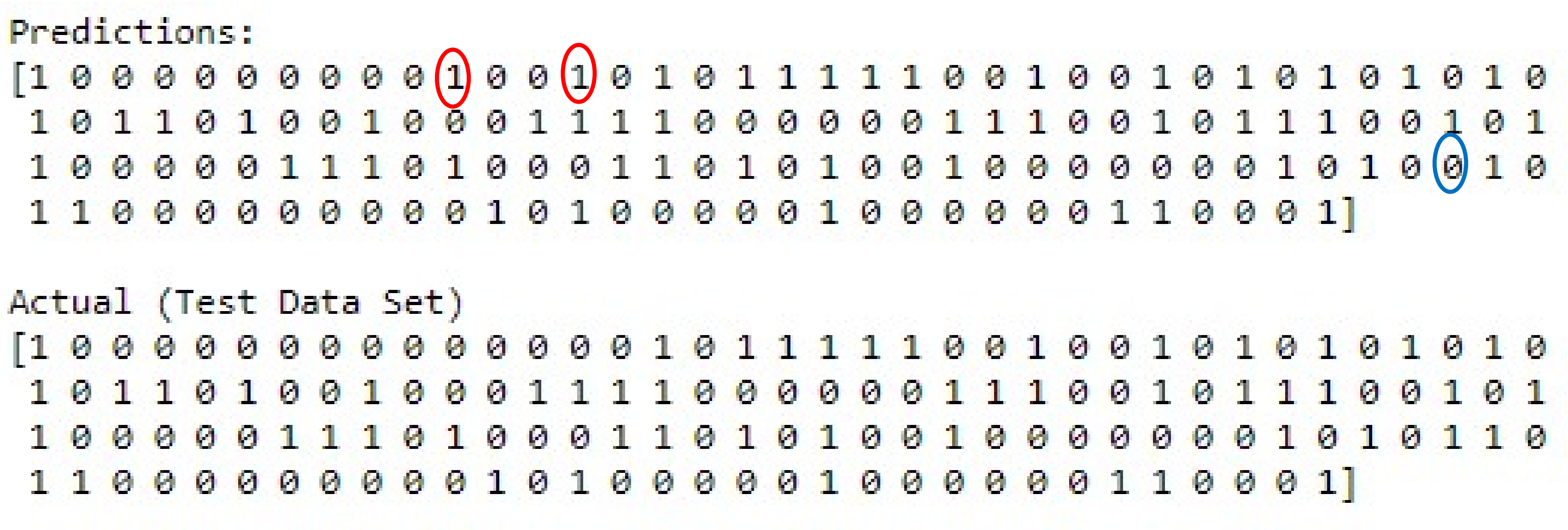
Evaluate Machine Learning Models

- Precisions, recalls, f-scores, and accuracies of some of models are as follows:

M/L Models	Precision	Recall	f1-score	Accuracy (Test Set)	Accuracy (Training Set)
Log. Regression	0.98	0.96	0.97	0.96	0.99
SVC	0.98	0.94	0.96	0.95	0.98
Random Forest	0.99	0.98	0.98	0.98	0.99
Decision Tree	0.97	0.94	0.96	0.94	1.0

- As expected the Decision Tree Classifier has the best training accuracies (overfitting tendency)
- However, the accuracy with the testing data set for the Decision Tree is not the best.
- Considering all the info in the above table, the Random Forest model performs the best.

Results Comparison – Actual vs. Prediction



```
Predictions:
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

Actual (Test Data Set)
[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]
```

○ False Positive

○ False Negative

Conclusions and Recommendations

- The Random Forest model misdiagnosed a few patients as having cancer when they didn’t and it misdiagnosed patients that did have cancer as not having cancer.
- Although this model is good, when dealing with the people’ lives the accuracy of the model should as close to 100% as possible or at least as good as if not better than doctors.
- So a little more tuning of each of the models is necessary.