Comparing Machine Learning Models for Classification of Breast Cancer

Introduction

Breast cancer is a disease that affects many women globally and is associated with a high fatality rate. The second-deadliest malignant tumour worldwide, behind lung cancer, is breast cancer. The duct region of a woman is where breast cancer is most frequently seen. Using a variety of techniques and instruments, including magnetic resonance imaging (MRI) and biopsy, doctors can detect breast cancer. For breast cancer, early detection is critical. Early detection of cancer threats may increase the patient's chances of survival. Therefore, research is being conducted to precisely classify individuals into malignant and benign groups. There are two types of breast cancer: benign tumours and malignant tumours.

Machine learning is the scientific study of computational algorithms and statistical models that seek to outperform human intelligence. Machine learning trains the system to predict future event outcomes using historical data. In order to quickly predict the risk of breast cancer, we want to create a machine learning system that incorporates a lot of data. Different forms of breast cancer can be categorised using a variety of ML techniques. Among Logistic Regression, Random Forest classification, Support Vector Machine, and k-nearest neighbours, the most accurate method for classifying breast cancer is chosen for comparison. All of these strategies use supervised machine learning models.

As a result, the Wisconsin Breast Cancer dataset, which was also included in the research, was employed in this machine learning study. The dataset was created by Dr. William H. Wolberg, a medical professional at the University of Wisconsin Hospital in Madison, Wisconsin. The collection includes 569 data pieces spanning 33 features including 357 benign (non-cancerous) and 212 malignant (cancerous) data pieces.

The dataset was first preprocessed to accommodate for zero values and missing values. The "train-test-split" approach is then used to divide this filtered dataset into training and testing analyses. Using the machine learning methods selected for the categorization of breast cancer, these subgroups are trained and tested. The effectiveness of these techniques is assessed using the following metrics: accuracy, precision, recall, specificity, and sensitivity after applying a variety of machine learning techniques to the dataset, including Logistic Regression, Random Forest classification, Support Vector Machine, and k-nearest neighbours.

Executive Summary:

This project compares two or more machine learning algorithm types on a dataset to determine which is the best. Breast Cancer Wisconsin is the dataset that was found for this research (Diagnostic). The algorithms that are suggested are Random Forest, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors for the purposes of this project.

Project Background:

Breast cancer is one of the most common cancers in women today, and data indicates that it is the second biggest killer of women after lung cancer. The disorders are separated into cancerous and benign tumors (non-cancerous). Malignant tumors are those that occur when breast tissue cells proliferate and grow without the normal controls on cell death and division. A benign tumor does not penetrate nearby tissue or spread to other parts of the body, but it might become harmful if it puts pressure on vital structures like blood vessels or nerves.
This is accomplished by classifying tumors as benign or malignant using the well-known dataset known as Breast Cancer Wisconsin which is available at Kaggle.com and served as the source of the data for this study. The performance of the suggested machine learning algorithm will be evaluated using this dataset. These statistics were gathered between 1989 and 1991 at the University of Wisconsin Madison Hospitals by Dr. William H. Wolberg.
With 32 attributes, including ID, diagnosis (M = malignant, B = benign), and 30 real-valued input features, there are 569 cases of Breast Cancer in Wisconsin(Diagnostic). Each cell nucleus is given a computed set of real-valued attributes, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave spots, symmetry, and fractal dimension. There are no empty attribute values in the dataset, which has 357 benign and 212 malignant samples, respectively, based on the class distribution. These attributes, which were calculated from a digitized image of a fine needle aspirate (FNA) of a breast tumor, will be used to explain the characteristics of the cell nuclei included in the image. It will be easier to distinguish benign from malignant tumors utilising this statistical information since the patients can be sorted into benign or malignant groups using the finest machine learning method.

Objectives:

i) To perform Data processing and exploratory Data Analysis on the dataset.

ii) On dataset, Apply the Feature Engineering technique.

iii) Using the dataset apply the 4 ML models

iv) To train the models using the product dataset

v) Evaluation and comparison of each models performance on the data

vi) Testing the dataset using ML models

vii) Identify the most efficient algorithm by comparing the performance of these models by accuracy.

Scope:

1. To help Medical or Health Related personalities for the prediction of breast cancer to be benign or malignant for the women.
2. This project will use Feature Engineering Algorithm that can classify the type of breast cancer based on the attributes into several groups that share a similarity in different ways that are relevant
3. Through analysis and the visualization of the result, this will help the doctors or the field related personality to diagnose the type of breast cancer for the patient more effectively using the feature selection method.

Project Significant:

Our main goal is to find the traits that are most helpful in determining whether a malignancy is benign or malignant. This classification will help with early detection of breast cancer, which can significantly increase the prognosis and likelihood of survival. It is possible for doctors to save patients from unnecessary operations by correctly diagnosing benign tumours.

The methodology used in the machine learning project is a supervised machine learning algorithm. The supervised machine learning method, which is trained on labelled data, is one of the most basic types of machine learning in this scenario. Although the data must be correctly classified for it to function properly, when used in the appropriate circumstances, it is quite valuable.

Supervised Machine Learning Algorithms

- **Random Forest Classification**

  Random forest classification is the most extensively used method in the supervised learning sector. This method's foundation is ensemble learning, which is further divided into a variety of classifiers and then merged to get accurate predictions. To solve complex issues and improve the performance of the model, a variety of classifiers are combined. By averaging the subsets in a random forest, which is made up of different Decision trees that are subsets of the supplied dataset, the accuracy of the dataset is increased. Rugged Forest Decision trees created during the training phase are used in this ensemble learning approach to carry out classification, regression, and other tasks. The class that most trees select as their output from a classification challenge.

- **Support Vector Machine**

  The employment of support vector machines can be used to tackle regression and classification problems. But the majority of the time, classification problems are addressed. In the Support Vector Machine algorithm, the value of each feature is represented by the value of a specific coordinate, and each piece of data is displayed as a point in n-dimensional space (where n is the number of characteristics you have). Then, classification is performed by locating the hyper-plane that most successfully distinguishes the two classes. The coordinates of each observation serve as the basis for support vectors. The SVM classifier is a frontier that separates the two classes the most effectively (hyper-plane and line).

- **Logistic Regression**

  One of the most widely used machine learning algorithms in the supervised learning approach is logistic regression. It is used to predict a categorical dependent variable from a set of independent variables. Logistic regression predicts a dependent categorical variable's output. Because it can generate probabilities and categorise new data using both continuous and discrete datasets, logistic regression is an important machine learning technique. Logistic regression can be used to categorise observations using a range of data formats and can quickly identify the most useful categorization criteria.

- **K-nearest neighbours**

  For classification and regression problems based on the feature similarity approach, the Supervised Learning technique uses one of the most well-known Machine Learning algorithms, k-nearest neighbours. Based on how similar the new data points are to the previously stored data points, KNN classifies them. The best k value, which can range from 1 to 100, is chosen from the testing results on the trained model with the highest accuracy. The closest neighbours are represented by the number K. The KNN algorithm does not have a training phase. Predictions are made based on the Euclidean distance to the k-nearest neighbours. This method is applied to the dataset for the prediction of breast cancer because it already contains labels for benign and malignant tumours. Based on whatever class label is the closest neighbour to that of the label's neighbours, the label is categorised.

Methodology

1. Data Collection

   The dataset was first collected from any kind of resource repository. The Wisconsin Breast Cancer dataset (WBCD) repository, which is available at Kaggle.com, served as the source of the data for this study. The dataset can be viewed using any tool, including MS-Excel or notepad, because it is supplied in.csv format. The dataset is mostly numerical, and the categories M for malignancy and B for benignity are used to identify the different forms of cancer. The dataset was produced using 569 samples and 32 attributes. The 32 features of the dataset are listed below. Identification, diagnosis, radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension mean; radius, texture, perimeter, area, smoothness, concave points, symmetry, and fractal dimension se

   Due to their exclusion from biological datasets, the first two features—id and diagnosis—out of a total of 32 features are not taken into consideration for the experimental analysis in this work. Based on the mean, standard error (se), and worst characteristics, each of which comprises ten attributes, the datasets are split into three groups.

2. Data Cleansing

Data cleaning is the next phase, when we check to determine if the dataset has any null or missing values. Label encoding is finished at this point. Words or numbers are used to indicate these labels. To make training data easier for individuals to grasp and interpret, labels have been added to the data. Label encoding is the process of translating labels from numerical to computer-readable representations. After that, ML algorithms will decide how to use those labels in the most efficient way. In supervised learning systems, label encoding is a crucial stage in the data pre-processing process. In order to translate labels into numbers between 0 and 1, the class Label Encoder is helpful. The minor drawback of label encoding is that it transforms data into a machine-readable format. Since each class in a dataset is given a distinct number starting at zero, priority difficulties occur when training datasets. A label with a high value, for instance, might be given preference over one with a low value.

3. Train Test Split

The train-test-split method can be used to evaluate machine learning algorithms for classification and regression applications. The technique divides the provided dataset into two categories: The training dataset is used to train the machine learning algorithm and to fit models. The algorithms make predictions using the input element from the training data in the test dataset. There are 20% of test data and 80% of training data in this dataset.

4. Feature Selection

Using the feature selection approach, the significant features with the strongest correlation to the target label will be selected. When irrelevant features are included, the model's performance degrades. Prior to creating a model, features should be decided. The benefits of feature selection include decreased over-fitting, better precision, and reduced training times. In this thesis, the key features were determined using the SelectKBest feature selection and the feature's chi-square score. It will be used to select the top five features. prediction based on all or just certain features.

Performance Metrics:

1. Confusion Matrix:

The confusion matrix is a performance metric that is widely used in classification issues with two or more output class labels. In this matrix, there are four different sets of expected and actual values. A confusion matrix is used to calculate accuracy, precision, f1-score, and recall. The following definitions must be considered in order to comprehend the confusion matrix.

2. Accuracy:

A measurement of misclassifications is the amount of projected classes that were erroneously classified based on the actual classes.

3. Precision:

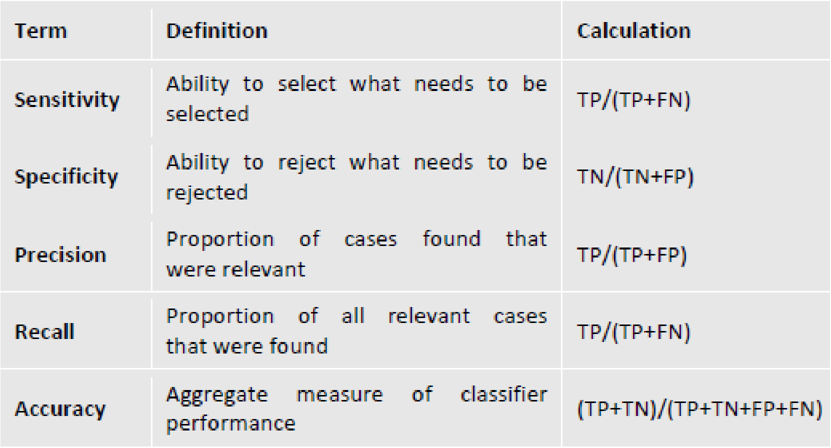
Precision, often known as confidence, is the proportion of genuinely good and genuinely negative events that are absolutely positive. This displays the classifier's ability to deal with positive findings while expressing little opinion about negative ones.

4. Recall and Sensitivity:

The frequency with which favourable forecasts are really projected to be positive is known as recall, also known as sensitivity. This metric is appealing because to the accuracy with which the data are reviewed, especially in the clinical situation. It is more important to correctly detect a dangerous cancer during this examination than it is to incorrectly identify a benign one.

5. Specificity

Because it is not specific, the rate at which negative projections are discovered will have a greater False Positive Value.



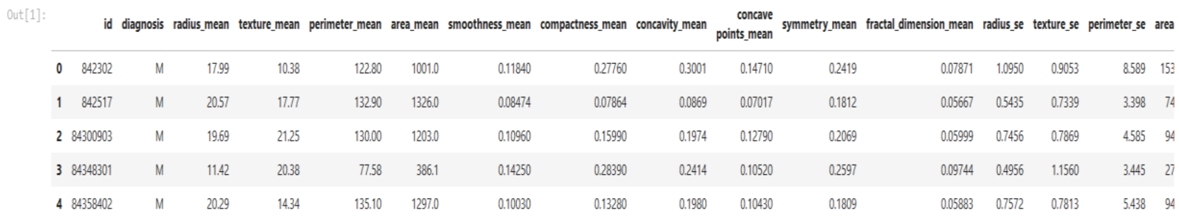
Term	Definition	Calculation
Sensitivity	Ability to select what needs to be selected	TP/(TP+FN)
Specificity	Ability to reject what needs to be rejected	TN/(TN+FP)
Precision	Proportion of cases found that were relevant	TP/(TP+FP)
Recall	Proportion of all relevant cases that were found	TP/(TP+FN)
Accuracy	Aggregate measure of classifier performance	(TP+TN)/(TP+TN+FP+FN)

6. Comparing the prediction results.

The best model to classify the type of breast cancer will next be determined by comparing the models' accuracy.

Results

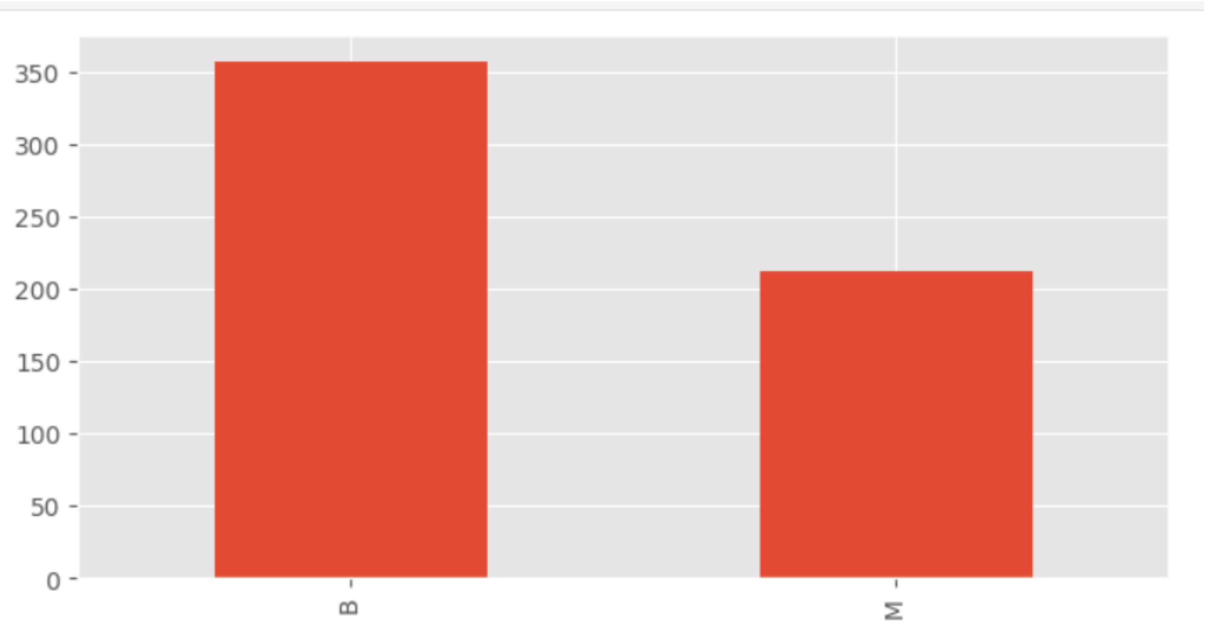
1. Importing Dataset



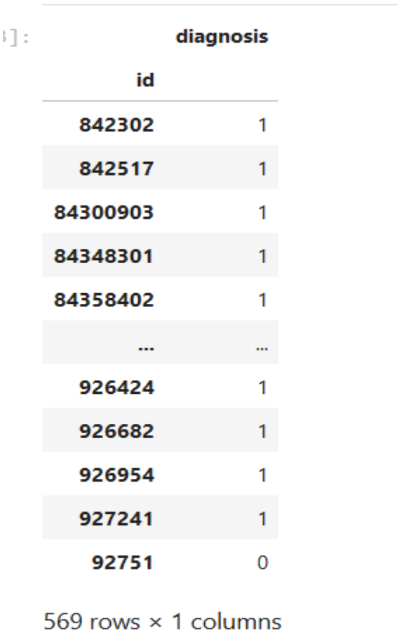
Out[1]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950	0.9053	8.589	153
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	0.7456	0.7869	4.585	94
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	0.4956	1.1560	3.445	27
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	0.7572	0.7813	5.438	94

2. Proportion of Benign and Malignant



3. Binarisation- Feature Transformation

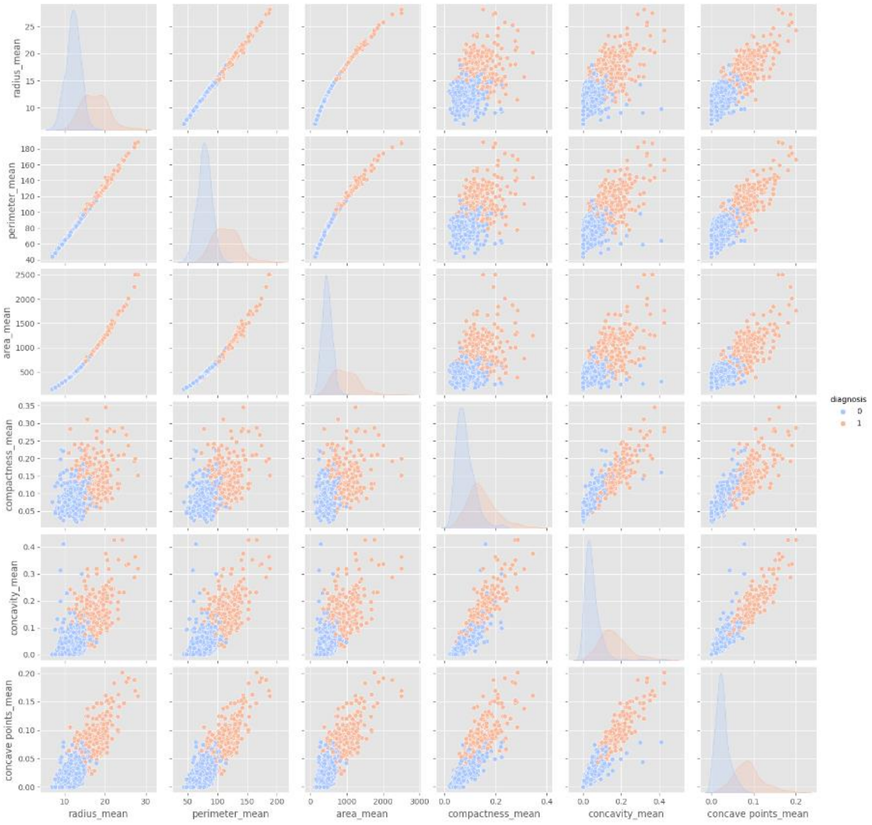


id	diagnosis
842302	1
842517	1
84300903	1
84348301	1
84358402	1
...	...
926424	1
926682	1
926954	1
927241	1
92751	0

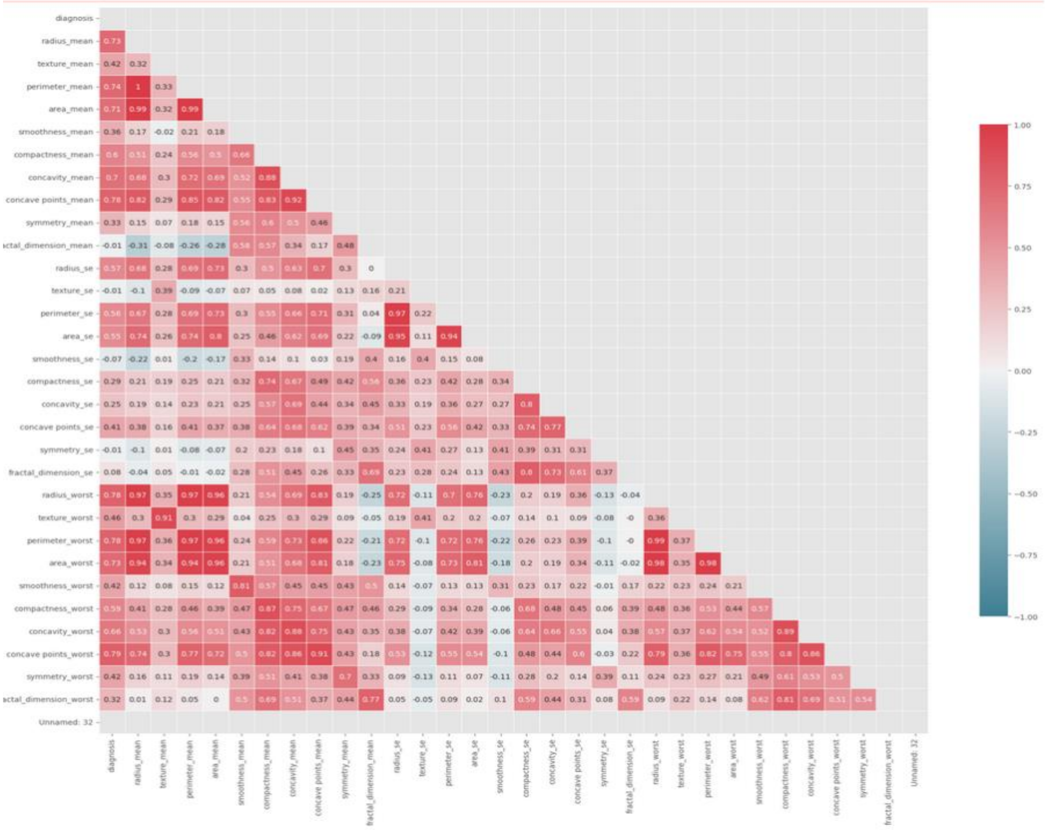
569 rows × 1 columns

4. Pairplot of the features

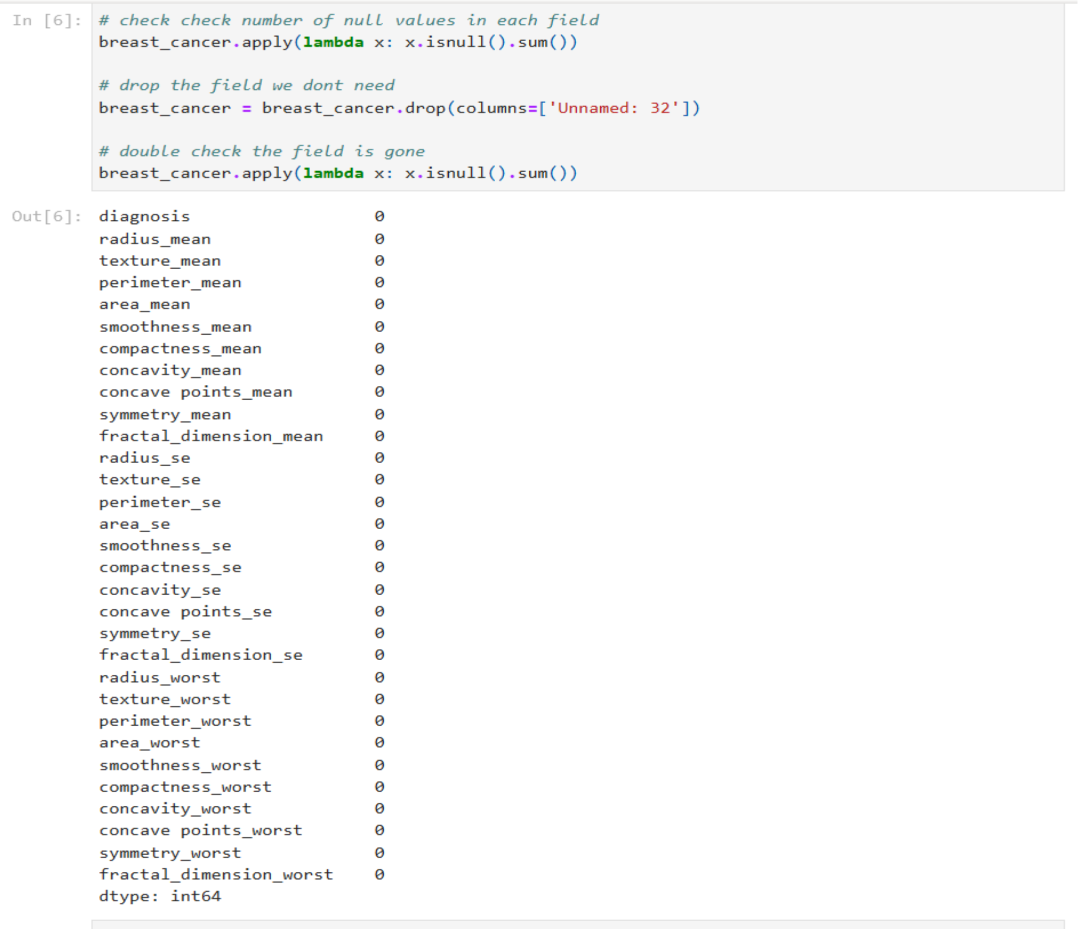
```
<seaborn.axisgrid.PairGrid at 0x2646d5f1330>
```



5. Finding the correlation through heatmap



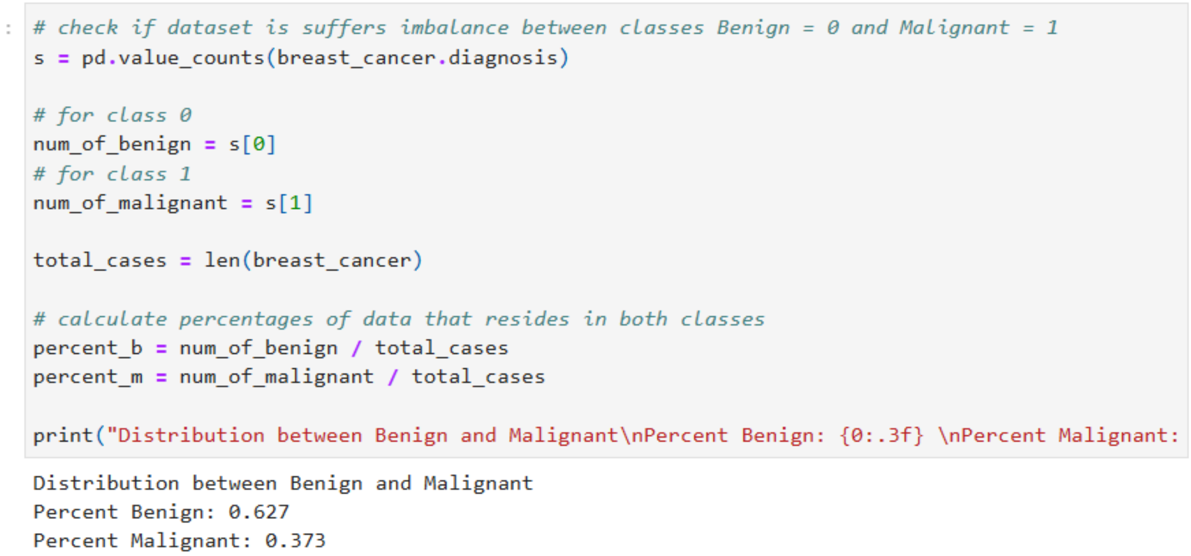
6. Checking if there is any missing values and dropping the unnecessary columns



```
In [6]: # check check number of null values in each field
        breast_cancer.apply(lambda x: x.isnull().sum())

        # drop the field we dont need
        breast_cancer = breast_cancer.drop(columns=['Unnamed: 32'])

        # double check the field is gone
        breast_cancer.apply(lambda x: x.isnull().sum())
```

```
Out[6]: diagnosis                  0
        radius_mean                0
        texture_mean               0
        perimeter_mean             0
        area_mean                  0
        smoothness_mean            0
        compactness_mean           0
        concavity_mean             0
        concave points_mean        0
        symmetry_mean              0
        fractal_dimension_mean     0
        radius_se                  0
        texture_se                 0
        perimeter_se               0
        area_se                    0
        smoothness_se              0
        compactness_se             0
        concavity_se               0
        concave points_se          0
        symmetry_se                0
        fractal_dimension_se       0
        radius_worst               0
        texture_worst              0
        perimeter_worst            0
        area_worst                 0
        smoothness_worst           0
        compactness_worst          0
        concavity_worst            0
        concave points_worst       0
        symmetry_worst             0
        fractal_dimension_worst    0
        dtype: int64
```

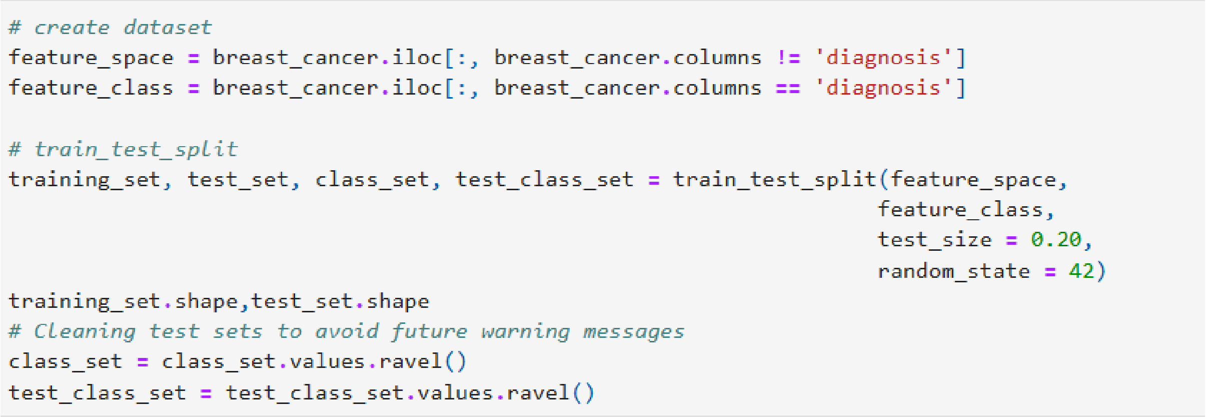
7. Checks the dataset if they are imbalance of benign and malignant



```
: # check if dataset is suffers imbalance between classes Benign = 0 and Malignant = 1
  s = pd.value_counts(breast_cancer.diagnosis)

  # for class 0
  num_of_benign = s[0]
  # for class 1
  num_of_malignant = s[1]

  total_cases = len(breast_cancer)

  # calculate percentages of data that resides in both classes
  percent_b = num_of_benign / total_cases
  percent_m = num_of_malignant / total_cases

  print("Distribution between Benign and Malignant\nPercent Benign: {0:.3f} \nPercent Malignant:
```

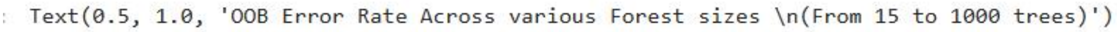
```
  Distribution between Benign and Malignant
  Percent Benign: 0.627
  Percent Malignant: 0.373
```

8. Train-Test-Split

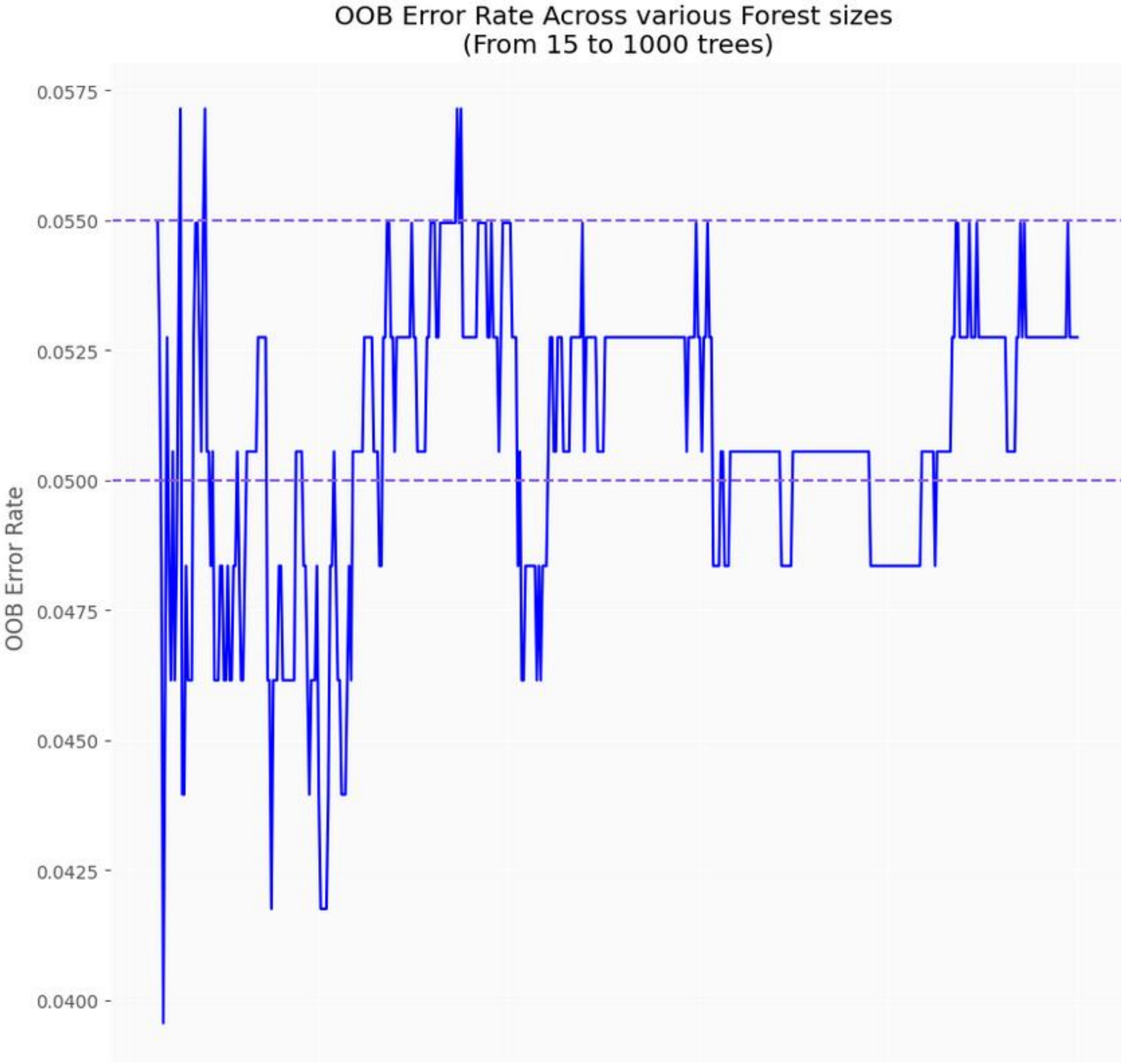


```
# create dataset
feature_space = breast_cancer.iloc[:, breast_cancer.columns != 'diagnosis']
feature_class = breast_cancer.iloc[:, breast_cancer.columns == 'diagnosis']

# train_test_split
training_set, test_set, class_set, test_class_set = train_test_split(feature_space,
                                                                      feature_class,
                                                                      test_size = 0.20,
                                                                      random_state = 42)
training_set.shape,test_set.shape
# Cleaning test sets to avoid future warning messages
class_set = class_set.values.ravel()
test_class_set = test_class_set.values.ravel()
```

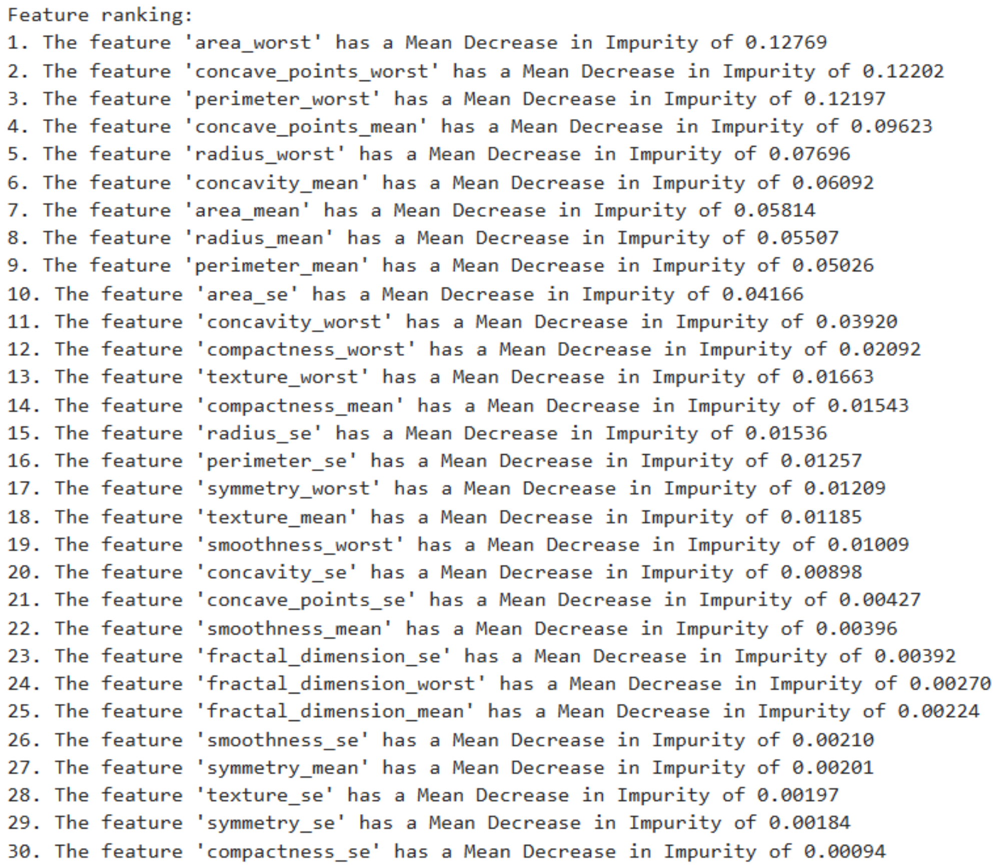
9. Random Forest Error Rate



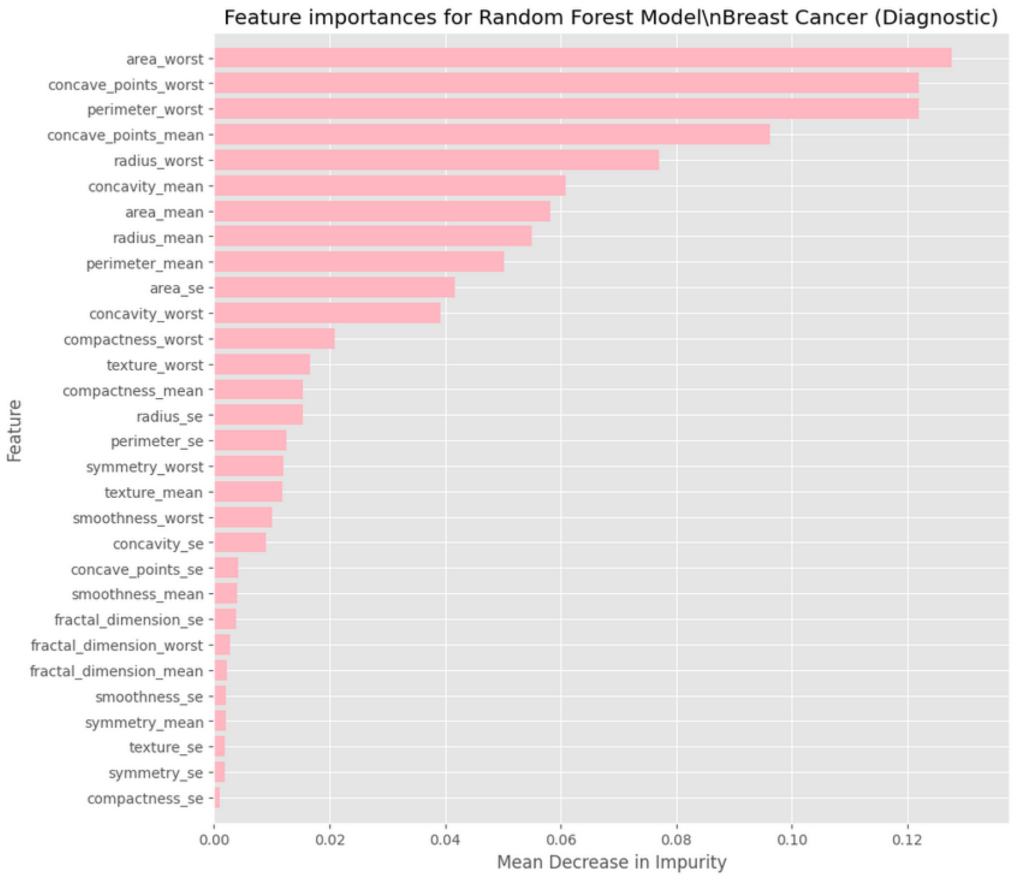
```
:  Text(0.5, 1.0, 'OOB Error Rate Across various Forest sizes \n(From 15 to 1000 trees)')
```



10.Random Forest Importance for Random Forest Model

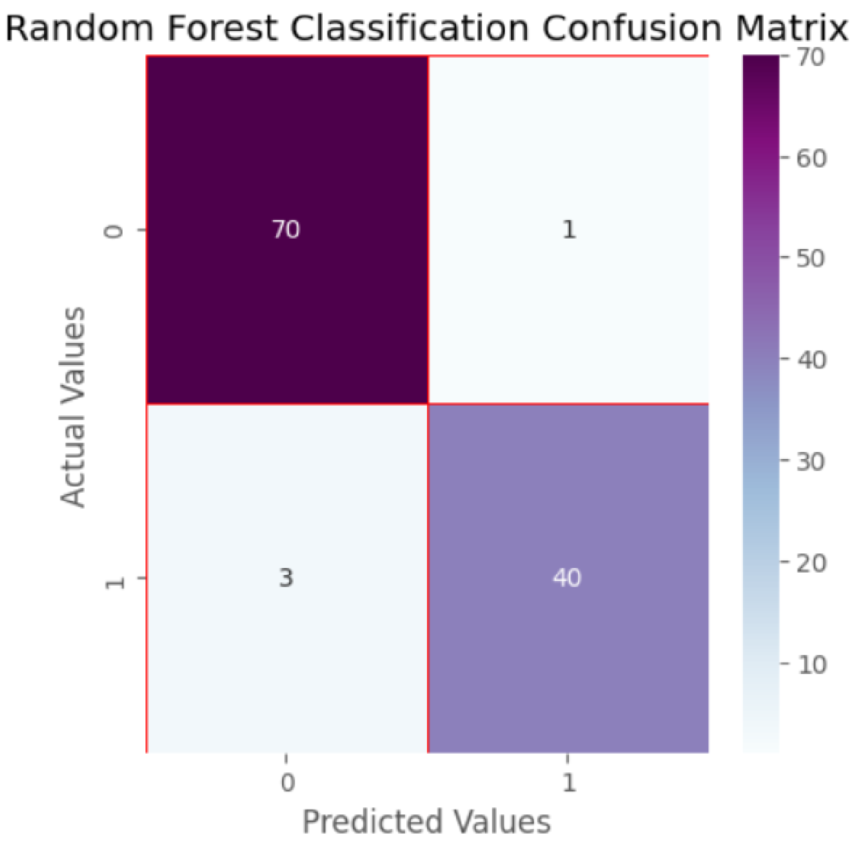


```
Feature ranking:
1. The feature 'area_worst' has a Mean Decrease in Impurity of 0.12769
2. The feature 'concave_points_worst' has a Mean Decrease in Impurity of 0.12202
3. The feature 'perimeter_worst' has a Mean Decrease in Impurity of 0.12197
4. The feature 'concave_points_mean' has a Mean Decrease in Impurity of 0.09623
5. The feature 'radius_worst' has a Mean Decrease in Impurity of 0.07696
6. The feature 'concavity_mean' has a Mean Decrease in Impurity of 0.06092
7. The feature 'area_mean' has a Mean Decrease in Impurity of 0.05814
8. The feature 'radius_mean' has a Mean Decrease in Impurity of 0.05507
9. The feature 'perimeter_mean' has a Mean Decrease in Impurity of 0.05026
10. The feature 'area_se' has a Mean Decrease in Impurity of 0.04166
11. The feature 'concavity_worst' has a Mean Decrease in Impurity of 0.03920
12. The feature 'compactness_worst' has a Mean Decrease in Impurity of 0.02092
13. The feature 'texture_worst' has a Mean Decrease in Impurity of 0.01663
14. The feature 'compactness_mean' has a Mean Decrease in Impurity of 0.01543
15. The feature 'radius_se' has a Mean Decrease in Impurity of 0.01536
16. The feature 'perimeter_se' has a Mean Decrease in Impurity of 0.01257
17. The feature 'symmetry_worst' has a Mean Decrease in Impurity of 0.01209
18. The feature 'texture_mean' has a Mean Decrease in Impurity of 0.01185
19. The feature 'smoothness_worst' has a Mean Decrease in Impurity of 0.01009
20. The feature 'concavity_se' has a Mean Decrease in Impurity of 0.00898
21. The feature 'concave_points_se' has a Mean Decrease in Impurity of 0.00427
22. The feature 'smoothness_mean' has a Mean Decrease in Impurity of 0.00396
23. The feature 'fractal_dimension_se' has a Mean Decrease in Impurity of 0.00392
24. The feature 'fractal_dimension_worst' has a Mean Decrease in Impurity of 0.00270
25. The feature 'fractal_dimension_mean' has a Mean Decrease in Impurity of 0.00224
26. The feature 'smoothness_se' has a Mean Decrease in Impurity of 0.00210
27. The feature 'symmetry_mean' has a Mean Decrease in Impurity of 0.00201
28. The feature 'texture_se' has a Mean Decrease in Impurity of 0.00197
29. The feature 'symmetry_se' has a Mean Decrease in Impurity of 0.00184
30. The feature 'compactness_se' has a Mean Decrease in Impurity of 0.00094
```



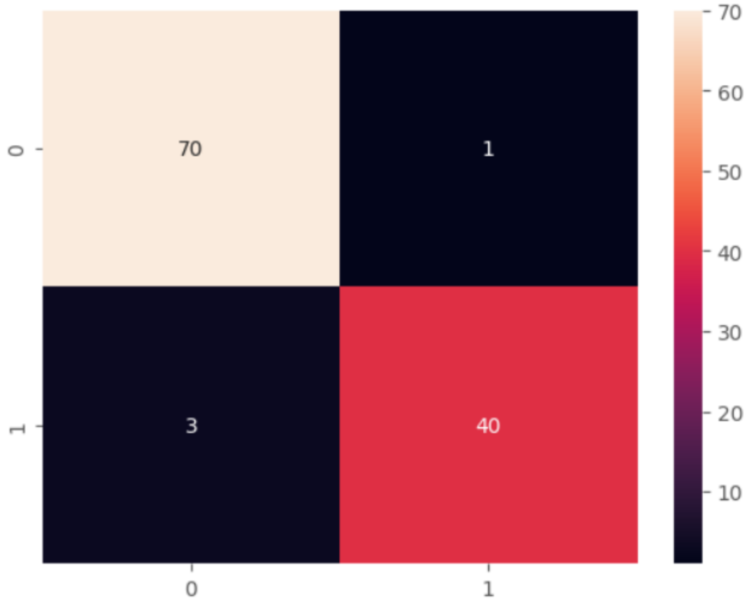
Results of each Models:

1. Confusion Matrix of Random Forest Model

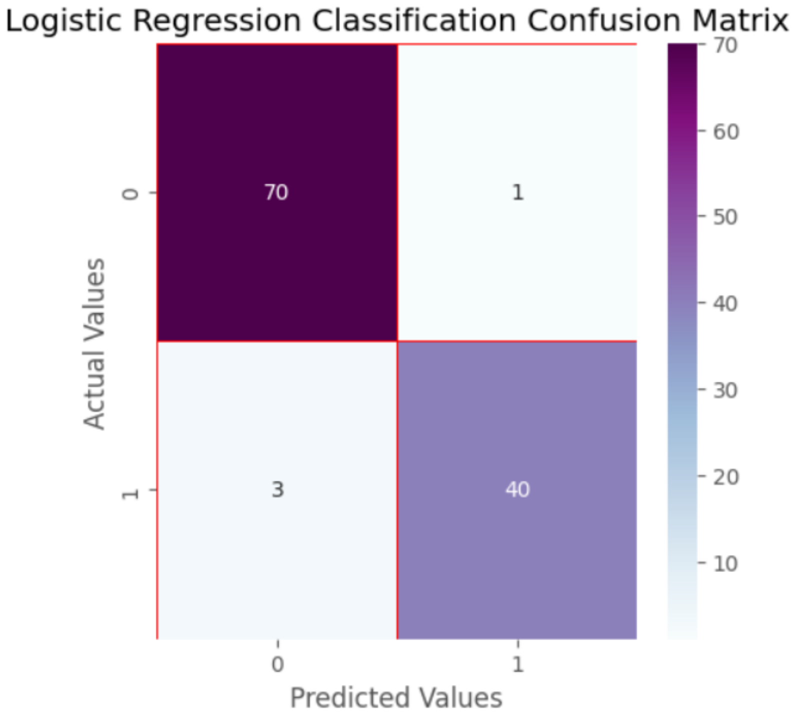


Confusion Matrix of Random Forest after feature selection of selectKBest

```
       Accuracy is:  0.9649122807017544
[40]:  <AxesSubplot: >
```

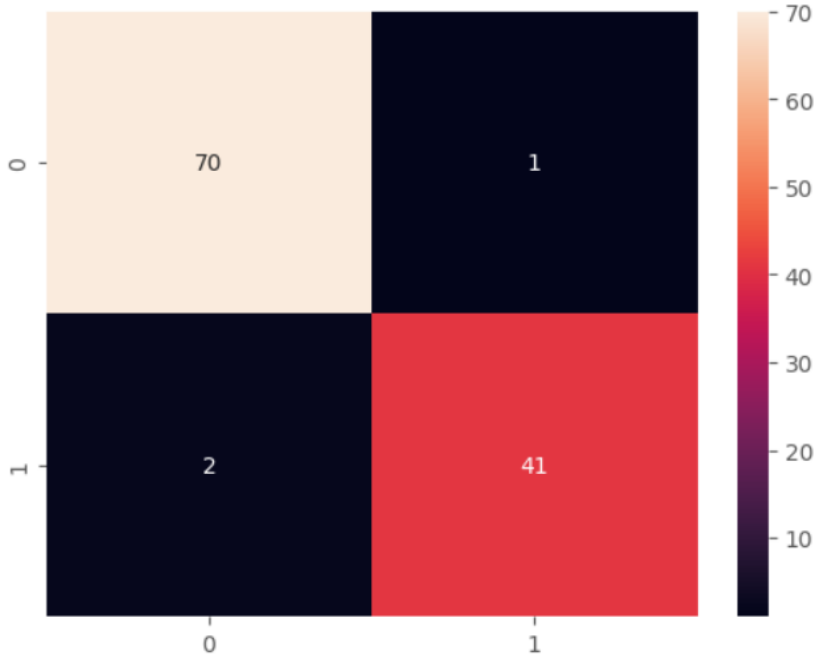


2. Confusion Matrix of Logistic Regression Model

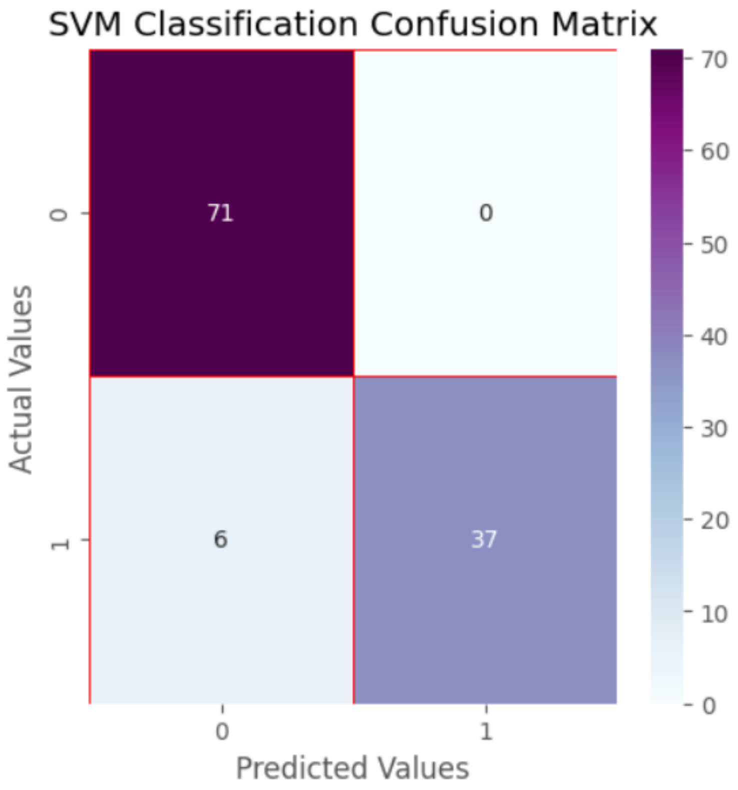


Confusion Matrix of Logistic Regression after feature selection of selectKBest

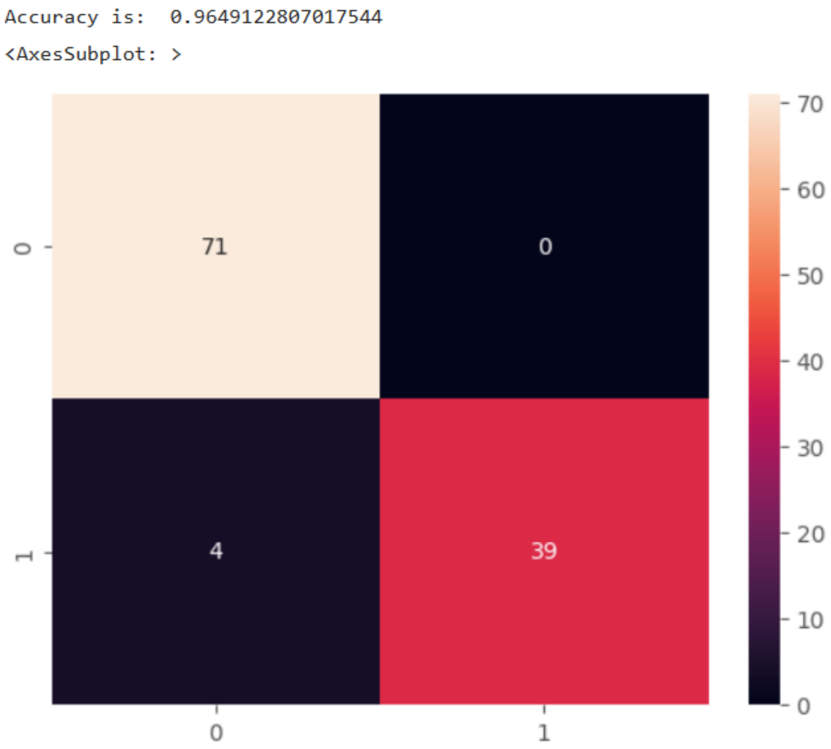
```
Accuracy is:  0.9736842105263158
<AxesSubplot: >
```



3. Confusion Matrix of Support Vector Machine Model

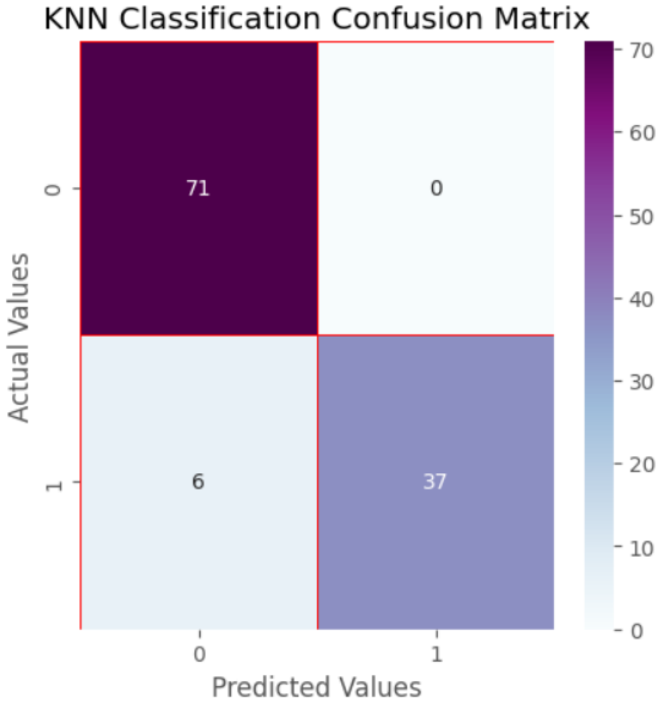


Confusion Matrix of Support Vector Machine Model after feature selection of selectKBest

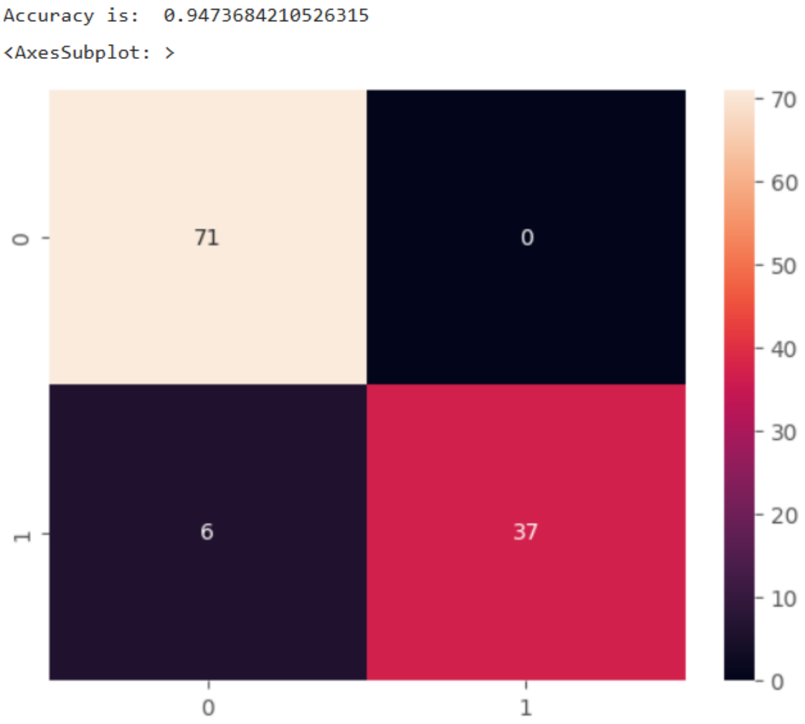


```
Accuracy is:  0.9649122807017544
<AxesSubplot: >
```

4. Confusion Matrix of K-nearest neighbors Model



Confusion Matrix of K-nearest neighbors Model after feature selection of selectKBest

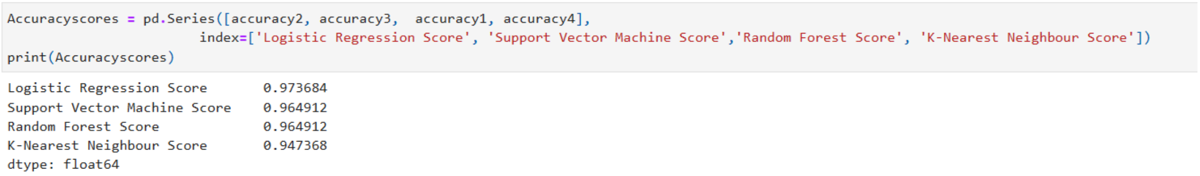


```
Accuracy is:  0.9473684210526315
<AxesSubplot: >
```

Comparative Analysis

Algorithm	Sensitivity	Specificity	Precision	Recall	Accuracy
Random Forest	0.98591549	0.930232558	0.98591549	0.636363636	0.96491228
Logistic Regression	0.98591549	0.95348837	0.98591549	0.63063063	0.97368421
Support Vector Machine	1.0	0.90697674	1.0	0.645454545	0.96491228
K-Nearest Neighbors	1.0	0.860465116	1.0	0.657407407	0.94736842

Comparing the accuracy of each model:-



```
Accuracyscores = pd.Series([accuracy2, accuracy3,  accuracy1, accuracy4],
                        index=['Logistic Regression Score', 'Support Vector Machine Score','Random Forest Score', 'K-Nearest Neighbour Score'])
print(Accuracyscores)
```

```
Logistic Regression Score       0.973684
Support Vector Machine Score    0.964912
Random Forest Score             0.964912
K-Nearest Neighbour Score       0.947368
dtype: float64
```

Conclusion:

Using the chosen Machine Learning approaches, such as Random Forest classification, Logistic Regression, Support Vector Machine, and K-nearest Neighbors, the dataset is utilized to evaluate and train the study's findings. After model creation, the dataset is used to evaluate the performance metrics accuracy, specificity, sensitivity, recall, and precision to determine how well the models classify the type of breast cancer. A tabulation of the performance measures was then used to assess the efficiency of each technique.

Only precise predictions are considered when measuring True Positives and True Negatives. When compared to other algorithms, Logistic Regression has the highest accuracy. When comparing the top performers of the two algorithms, Logistic Regression offers superior performance metrics. Thus, the Logistic Regression, a supervised learning method, produced the best training and testing outcomes when compared to Random Forest, Support Vector Machine, and k-nearest neighbors. The strength of Logistic Regression is the outputs have a good probabilistic interpretation, and the algorithm can be regularized to avoid overfitting Weaknesses of Logistic regression is that tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships.

Tools and Algorithms:

For this project, we will be using Python Programming Language and several needed libraries as matplotlib, pandas, and many more. Algorithms that will be used in this project are Logistic Regression, Random Forest, Support Vector Machine and K-Nearest Neighbours.

Reference:

1. *BITI 2223 machine learning mini project- Feature engineering*. (2023, January 27). YouTube. https://youtu.be/L7jShgs50GI
2. *Breast cancer Wisconsin (Diagnostic) data set*. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
3. Goswami, S. (2020, November 13). *Using the CHI-squared test for feature selection with implementation*. Medium. https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-b15a4dad93f1
4. Kumar, S. (2021, September 4). *Feature selection using logistic regression model*. Medium. https://towardsdatascience.com/feature-selection-using-logistic-regression-model-efc949569f58
5. *Support vector machines (SVM) algorithm explained*. (2017, June 22). MonkeyLearn Blog. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
6. Dubey, A. (2018, December 15). *Feature selection using random forest*. Medium. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f