Final Project Report
Introduction to Data Analytics

Project Title:
Breast Cancer Prediction

Prepared by:
Utsav Jitendrabhai Patel (N01516259)

ITE 5201 – Summer 2022
Humber College

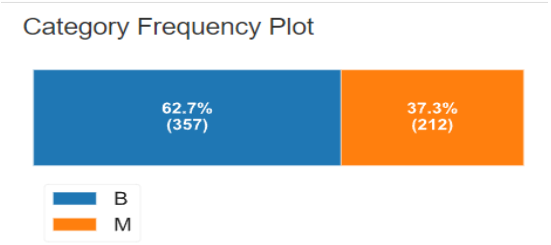
1. Problem Statement: To classify that the specific cancer node is malignant or benign by observing parameters.

2. Dataset Description: The used dataset carries the information regarding the cancer nodes observed in various patients. These parameters include Diagnosis (Diagnosis of breast tissue<M, B>), radius_mean (mean of the distances from center to points on the parameter), perimeter_mean (The mean size of the core tumor), texture_mean (standard deviation of gray-scale values) and so on.

Detail information of dataset: It is showing detail of columns and null values, datatypes of columns regarding this dataset

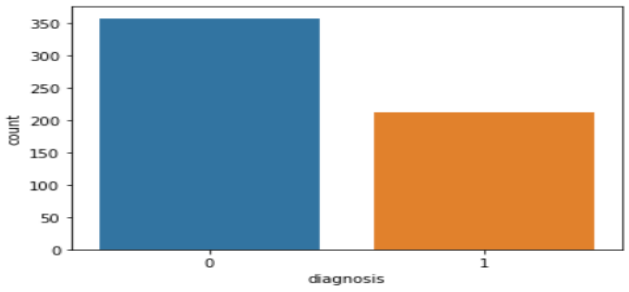
```
RangeIndex: 569 entries, 0 to 568
Data columns (total 12 columns):
 #    Column                  Non-Null Count  Dtype
---   ------                  --------------  -----
 0    id                      569 non-null    int64
 1    diagnosis               569 non-null    object
 2    radius_mean             569 non-null    float64
 3    texture_mean            569 non-null    float64
 4    perimeter_mean          569 non-null    float64
 5    area_mean               569 non-null    float64
 6    smoothness_mean         569 non-null    float64
 7    compactness_mean        569 non-null    float64
 8    concavity_mean          569 non-null    float64
 9    concave points_mean     569 non-null    float64
 10   symmetry_mean           569 non-null    float64
 11   fractal_dimension_mean  569 non-null    float64
dtypes: float64(10), int64(1), object(1)
memory usage: 53.5+ KB
```

Diagnosis (Category-Frequency Plot): 62.7% data benign and 37.3% malignant



3. Dataset Analysis and Observations: In this dataset Id has unique values. Diagnosis is highly correlated with radius_mean, area_mean is highly correlated with radius_mean, Perimeter_mean is also highly correlated with the radius_mean etc.

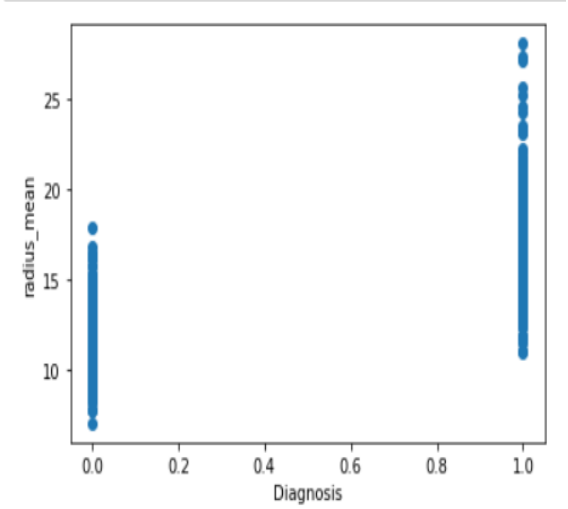
Univariate: Diagnosis plot: Diagnosis have categorical data (0,1)



Radius_mean observations: Distinct values are 456, missing 0 after cleaning, Mean 14.12729174

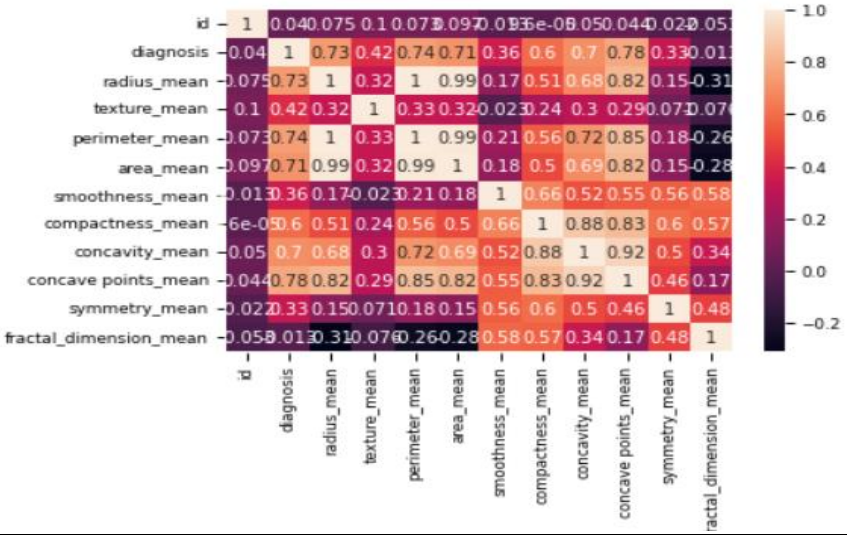
Distinct	456
Distinct (%)	80.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	14.12729174

Bivariate: Scatter plot of Diagnosis and radius_mean



This scatter plot predicts the range of radius mean based on Diagnosis

Heatmap for visualization



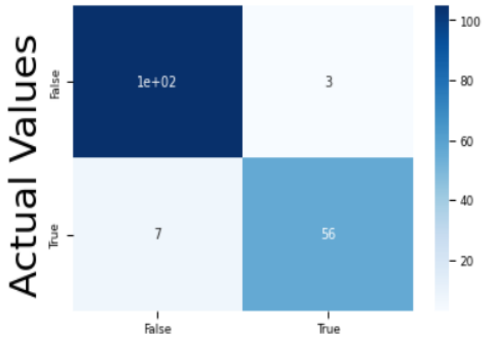
Heatmap denotes the relationship between two variable's data in R * R=1 strong positive relationship * R=0 Not linearly correlated * R=-1 strong negative relationship. From this heatmap, I observed that radius_mean, texture_mean, perimeter_mean, area_mean, concavity_mean, concavepoint_mean means are highly correlated to diagnosis. And diagnosis, radius_mean, area_mean are highly important part in classifying Breast cancer type.

4. Proposed Analytical Model: I have categorical output. I have used logistic regression, Random Forest and kneighborClassifier. Also, I designed confusion matrix and classification report and based on that I designed confusion matrix diagram and evaluation of predicted data. I wanted to choose these three models because for categorical these models gives accurate prediction.

5. Results and Discussions: I created three models and in diagram showing accuracy, confusion matrix, classification report.

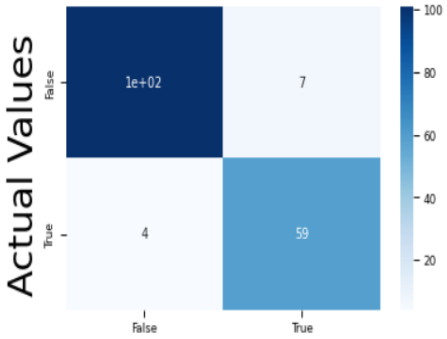
Logistic Regression:

```
ACCURACY OF THE Logistic Regression:  0.9415204678362573
```



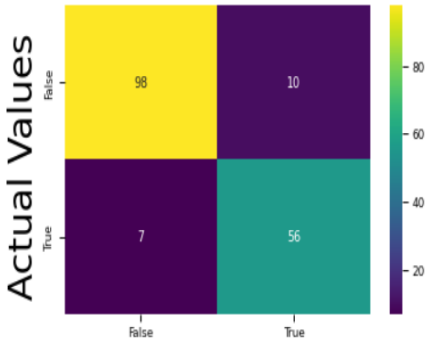
Random Forest:

```
ACCURACY OF THE Random Forest:  0.935672514619883
```



KNN:

```
Accuracy of knn model:  0.9005847953216374
```



True positive (TP): We predicted positive, but value is positive. **False positive (FP):** We predicted positive, but the actual value is negative.

True negative (TN): We predicted negative, but it is negative. **False negative (FN):** We predicted negative, but value is negative.

	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Logistic Regression	56	3	105	7
Random Forest	59	7	101	4
KNN	56	10	98	7

Classification Report of Logistic Regression:

	precision	recall	f1-score	support
0	0.94	0.97	0.95	108
1	0.95	0.89	0.92	63
accuracy			0.94	171
macro avg	0.94	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171

Classification Report of Random Forest:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	108
1	0.89	0.94	0.91	63
accuracy			0.94	171
macro avg	0.93	0.94	0.93	171
weighted avg	0.94	0.94	0.94	171

Classification Report of KNN:

	precision	recall	f1-score	support
0	0.93	0.91	0.92	108
1	0.85	0.89	0.87	63
accuracy			0.90	171
macro avg	0.89	0.90	0.89	171
weighted avg	0.90	0.90	0.90	171

Precision	Out of all the model's prediction data 95% were correct.	Out of all the model's prediction data 89% were correct.	Out of all the model's prediction data 85% were correct.
Recall	The model predicted Data 89% correctly.	The model predicted Data 94% correctly.	The model predicted Data 89% correctly.
F1-Score	0.92 score is near to 1 so model did a good job.	0.91 score is near to 1 so model did a good job.	0.87 score is near to 1 so model did a good job.
Support	We can see that among the prediction of M and B in the test dataset,108 datasets did not predict properly and 63 did.	We can see that among the prediction of M and B in the test dataset,108 datasets did not predict properly and 63 did.	We can see that among the prediction of M and B in the test dataset, 108 datasets did not predict properly and 63 did.

Conclusion: Hammering the last nail, the logistic regression is the best for modeling for breast cancer dataset. Because the accuracy score for logistic regression is high. And I believe that for categorical data logistic regression gives the best accuracy score.