Tumor Analysis

Krishany Alageshwaran

The Rationale Statement

- ✓ In our data set we will be examining tumor analysis related dataset consisting of 569 observations. The data set contains 30 attributes that identifies tumor measurements such as texture, perimeter, smoothness & etc.

- ✓ With these attributes tumors are categorized as two types which are named as Malignant (M) & Benign (B). Benign tumors stay in their primary location without invading other sites of the body. Malignant tumors are cancerous and have cells that grow uncontrollably and spread locally and to distant sites

- ✓ In this case Mr. John Hughes wants to determine the best suited algorithm which can accurately predict whether the tumor is a Malignant type tumor or Benign type tumor based on the attributes of tumor measurements

- ✓ In addition to that, he expects us to provide recommendations to improve the effectiveness of the selected model

- ✓ The given dependent variables is categorical data and we should approach models which can handle classification problems well

The methodology used in tackling the problem

When looking at the dependent variable - diagnosis, which is categorical and suggests a classification problem

We have selected below 3 algorithms which works well on addressing classification problems

1. **Support Vector Machines** – It is considered to be a classification approach, it creates a hyper plane (decision boundary) which separates data into classes. It uses the kernel trick to find the best line separator (decision boundary that has same distance from the boundary point of both classes). It is also a clear and more powerful way of learning complex non linear functions

2. **Logistic Regression** – It is a classification algorithm which is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (Malignant) or 0 (Benign)

3. **Decision Tree** - Classification tree type decision tree predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs to.

Initially, we will be creating x, y variables, then the training & test dataset will be created and scaled before executing the models

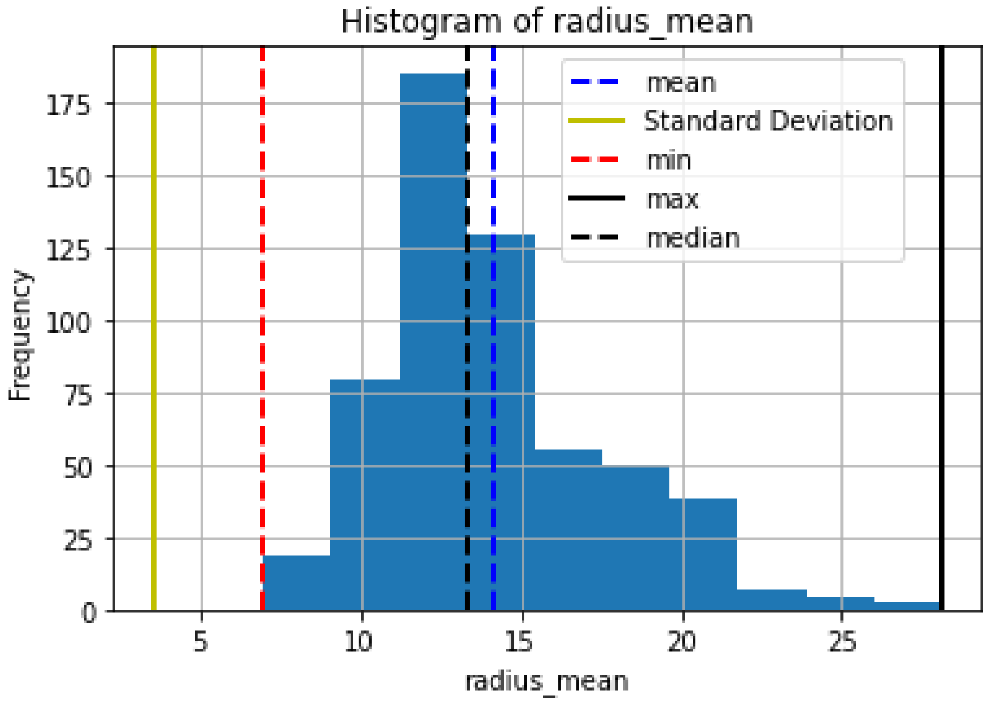
We will be using the above mentioned algorithms to compare results of confusion matrix and classification report to recommend the best model

Key insights from the basic statistics of the dataset

1) radius_mean

count	mean	std	min	25%	50%	75%	max
569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000

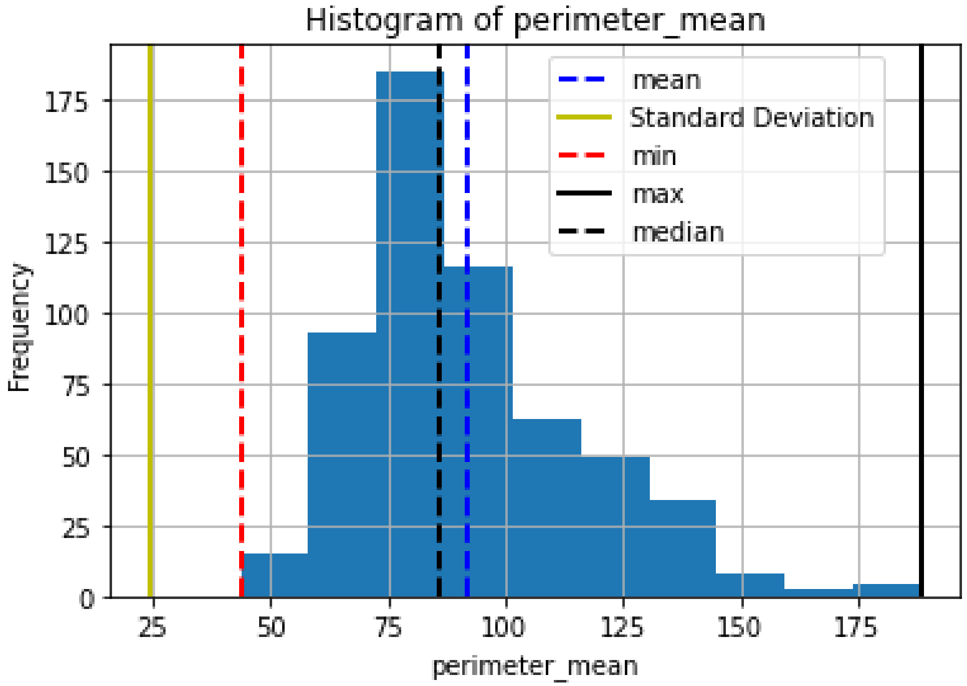
- ➢ The observations are based on 569 tumors
- ➢ On average, the distance from the center point of a tumor is 14.127292mm
- ➢ The standard deviation of 3.524049mm show the data scattered from the average 14.127292mm
- ➢ The smallest average distance from the center point was 6.981mm
- ➢ The highest average radius reported among the tumors is 28.11mm
- ➢ 75% of the tumors had a probability of having 15.78mm radius or lesser, while 25% of the tumors had 11.7 mm or lesser average radius
- ➢ Half of the tumors had at least 13.37mm of average radius



2) perimeter_mean

count	mean	std	min	25%	50%	75%	max
569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000

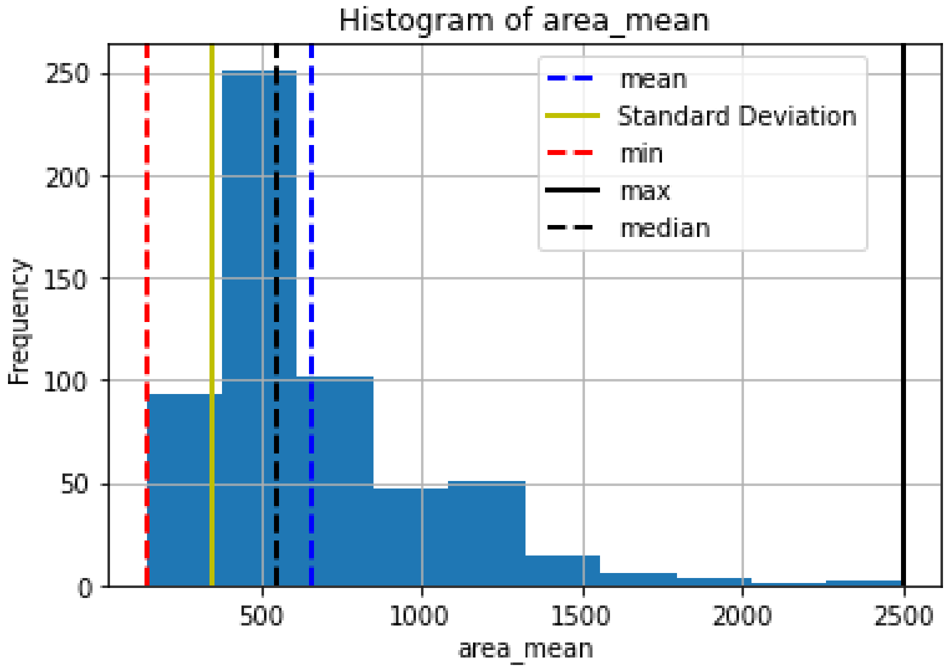
- The observations are based on 569 tumors
- On average, the distance around a tumor is 91.969033mm
- The high standard deviation of 24.298981mm show the data scattered from the average 14.127292mm
- The minimum average distance around a tumor was 43.79mm
- The maximum average perimeter of tumor reported among others is 188.5mm
- 75% of the tumors had at least a perimeter of 104.1mm or lesser, while 25% of the tumors had 75.17mm or lesser average perimeter
- Half of the tumors had at least 86.24mm of average distance around



3) area_mean

count	mean	std	min	25%	50%	75%	max
569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000

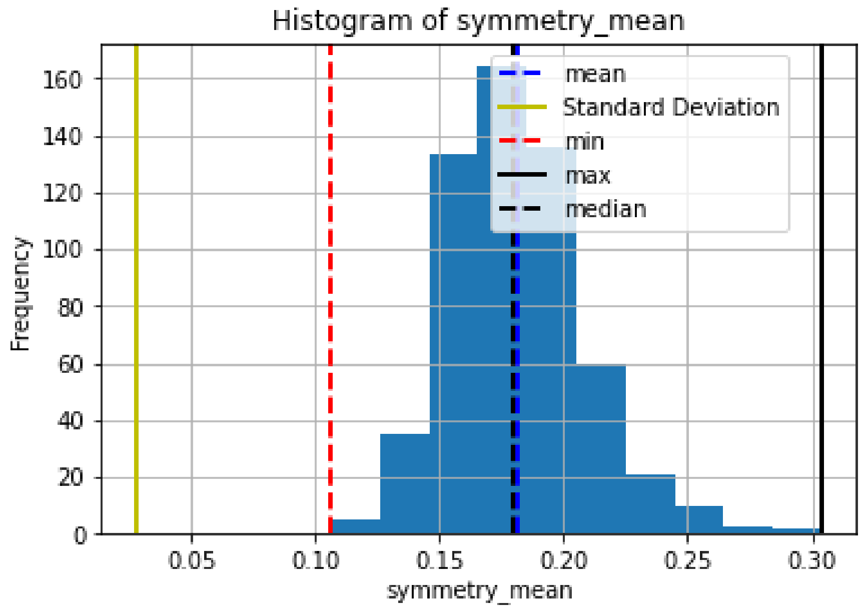
- The observations are based on 569 tumors
- On average, the surface area of a tumor is 654.889104mm^2
- The high standard deviation of 351.914129mm^2 shows the data scattered from the average 654.889104mm^2
- The smallest average surface area around a tumor was 143.5mm^2
- The largest average surface area around a tumor reported is 2501mm^2
- 75% of the tumors had at least a surface area of 782.7mm^2 or lesser, while 25% of the tumors had 420.3mm^2 or lesser average surface area around them
- Half of the tumors had at least 551.1mm^2 of average surface area around



4) symmetry_mean

count	mean	std	min	25%	50%	75%	max
569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400

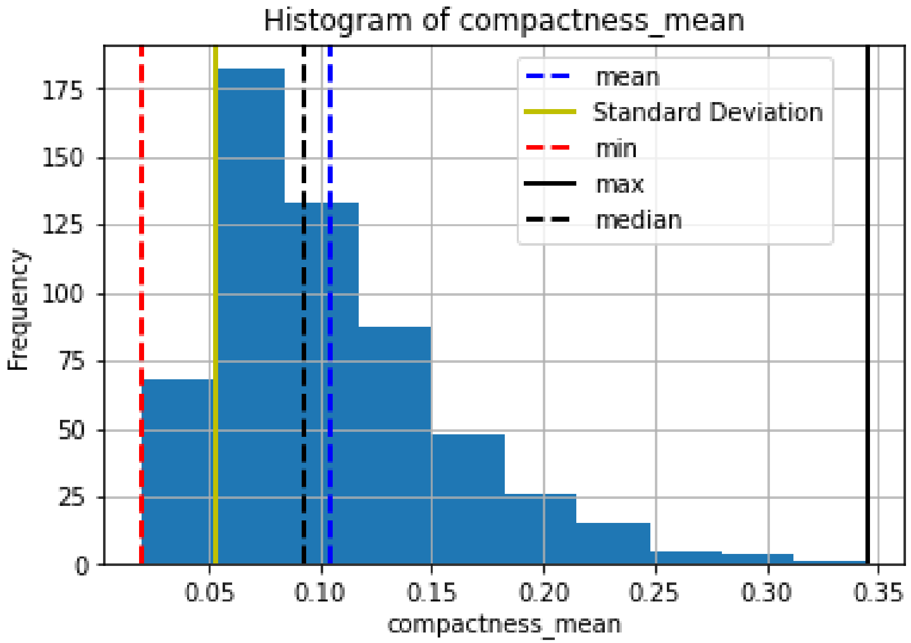
- The observations are based on 569 tumors
- On average, the uniformity between a tumors is 0.181162
- The standard deviation of 0.027414 shows the data scattered from the average 0.181162
- The minimum average uniformity of tumors was 0.106
- The highest average uniformity of tumors reported is 0.304
- 75% of the tumors had at least 0.1957 of uniformity or lesser, while 25% of the tumors had 0.161900 of uniformity or lesser
- Half of the tumors had at least 0.179200 of uniformity in-between tumors



5) compactness_mean

count	mean	std	min	25%	50%	75%	max
569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540

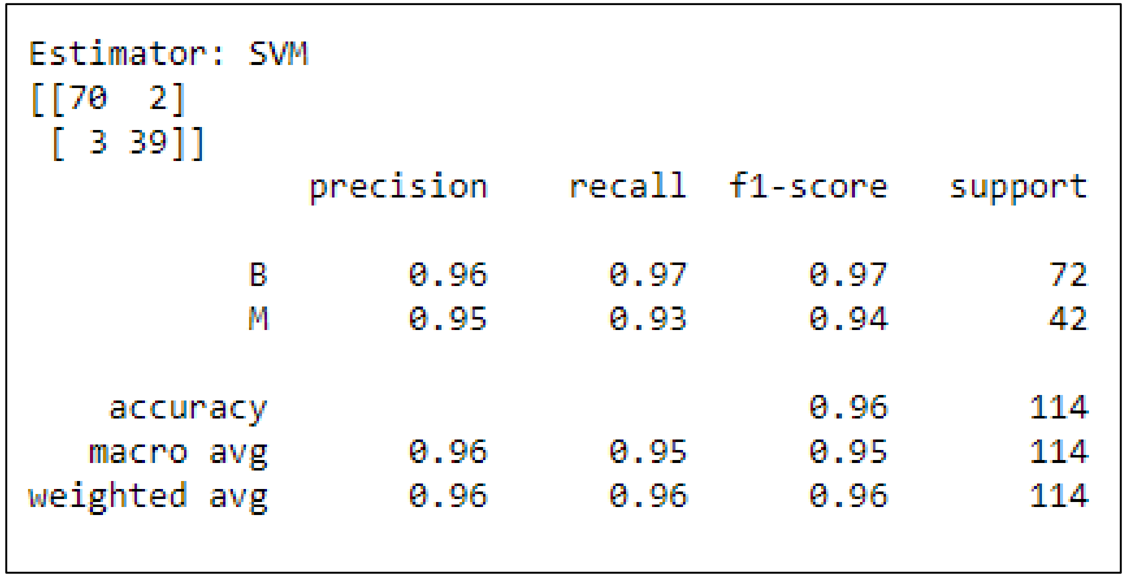
- The observations are based on 569 tumors
- Tumor compactness is defined as the ratio of the volume and the surface area
- Average compactness of tumors is 0.104341
- The standard deviation of 0.052813 shows the data scattered from the average 0.104341
- The minimum average compactness of tumors was 0.019380
- The maximum average compactness of tumor reported is 0.34540
- 75% of the tumors had a compactness of at least 0.1304 or lesser, while 25% of the tumors had a compactness of 0.064920 or lesser
- 50% of the tumors had a compactness of at least 0.092630 or higher



Support Vector Machine Algorithm

Confusion Matrix

- The first row (B): The model has classified 70 Benign Tumors correctly and 2 incorrectly
- The Second row (M): The model has classified 39 Malignant tumors correctly and 3 incorrectly



```
Estimator: SVM
[[70  2]
 [ 3 39]]
              precision    recall  f1-score   support

           B       0.96      0.97      0.97        72
           M       0.95      0.93      0.94        42

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Classification Report

Precision – 0.96

✓ Precision is about being precise and how accurate the model is. In other words, when a model makes a prediction, how often it is correct

✓ SVM has 96% correctly predicted the diagnosis(Tumors)

Recall – 0.96

✓ Recall is the ratio of correctly predicted positive observations to the all observations in actual class. When the model predicts positive, how often is it correct

✓ SVM Model can identify them 96% of the time

F1 – 0.96

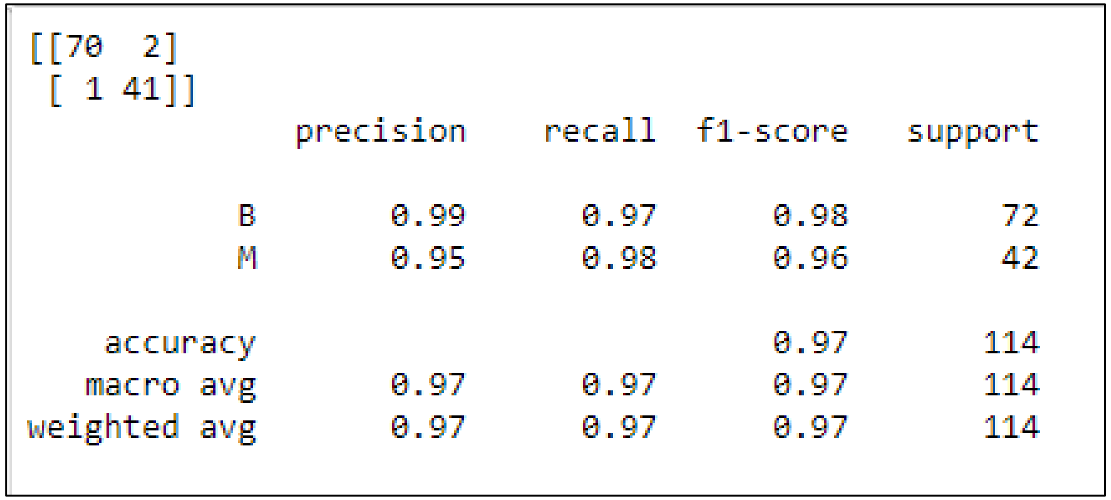
✓ F1 Score is the weighted average of Precision and Recall

✓ I this model we have received a F1 score of 96%

Logistic Algorithm

Confusion Matrix

- The first row (B): The model has classified 70 Benign Tumors correctly and 2 incorrectly
- The Second row (M): The model has classified 41 Malignant tumors correctly and 1 incorrectly



```
[[70  2]
 [ 1 41]]
              precision    recall  f1-score   support

           B       0.99      0.97      0.98        72
           M       0.95      0.98      0.96        42

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

Classification Report

Precision – 0.97

✓ Precision is about being precise and how accurate the model is. In other words, when a model makes a prediction, how often it is correct

✓ SVM has 97% correctly predicted the diagnosis(Tumors)

Recall – 0.97

✓ Recall is the ratio of correctly predicted positive observations to the all observations in actual class. When the model predicts positive, how often is it correct

✓ SVM Model can identify them 97% of the time

F1 – 0.97

✓ F1 Score is the weighted average of Precision and Recall

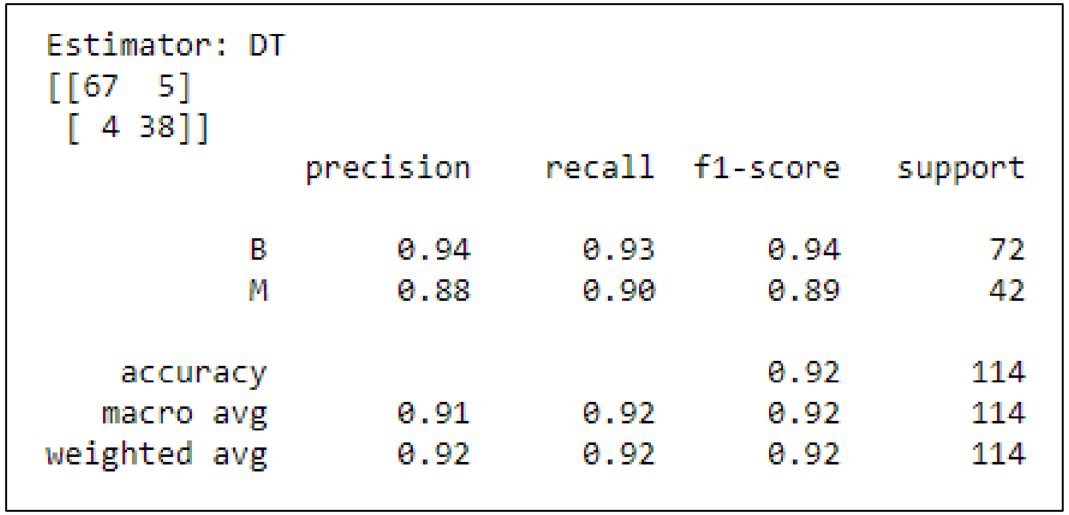
✓ I this model we have received a F1 score of 97%

Decision Tree Algorithm

Confusion Matrix

- The first row (B): The model has classified 67 Benign Tumors correctly and 5 incorrectly
- The Second row (M): The model has classified 38 Malignant tumors correctly and 4 incorrectly

Classification Report



```
Estimator: DT
[[67  5]
 [ 4 38]]
              precision    recall  f1-score   support

           B       0.94      0.93      0.94        72
           M       0.88      0.90      0.89        42

    accuracy                           0.92       114
   macro avg       0.91      0.92      0.92       114
weighted avg       0.92      0.92      0.92       114
```

Precision – 0.92

- ✓ Precision is about being precise and how accurate the model is. In other words, when a model makes a prediction, how often it is correct
- ✓ SVM has 92% correctly predicted the diagnosis(Tumors)

Recall – 0.92

- ✓ Recall is the ratio of correctly predicted positive observations to the all observations in actual class. When the model predicts positive, how often is it correct
- ✓ SVM Model can identify them 92% of the time

F1 – 0.96

- ✓ F1 Score is the weighted average of Precision and Recall
- ✓ I this model we have received a F1 score of 92%

Recommending a model to be utilized by Mr. John Hughes

Recommendation: I recommend Logistic algorithm to Mr. John Hughes which will help him predict tumor type classifications better

Justification:

✓ When considering the F1 values of all 3 Algorithms for the given dataset,

SVM – 96%
Logistic – 97%
Decision Trees – 92%

✓ the logistic model has 97% predicted accurately and has performed well than the SVM and Decision Tree Algorithm

2 possible improvements that can be made to increase the effectiveness of the model chosen

- ✓ Using more data to train and test the model could result in better and accurate outcomes. For example, the data set provided contains 569 records and within that 80% of data is used to train the data. Adding more data would allow more information and helps the model to improve its accuracy

- ✓ Features that has been used such as fractal_dimension_worst might not have a close relationship to the Diagnosis type trying to predict. Using better new features might improve the performance of the models