# CONTENTS OF THE PRESENTATION

## 01

### INTRODUCTION

We'll discuss the types of tumors/cancer and the need to discriminate them.

## 02

### CANCER DATASET

We'll discuss the tumor features cancer data set.

## 03

### OPTIMIZATION

We’ll discuss the optimization model that we're going to use.

## 04

### RESULTS & ACCURACY

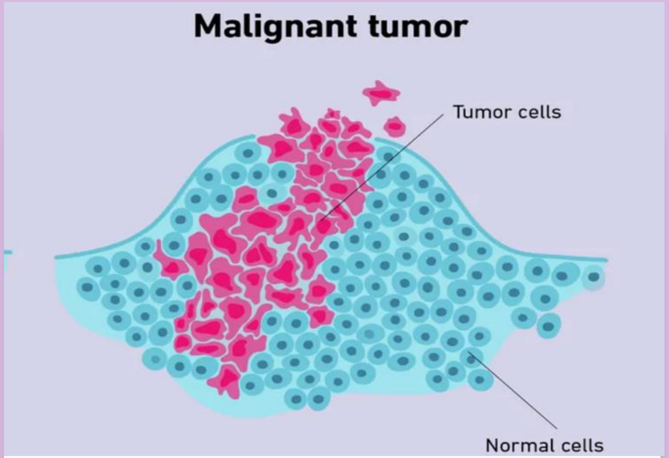
Results of our prediction model and its accuracy.

# INTRODUCTION

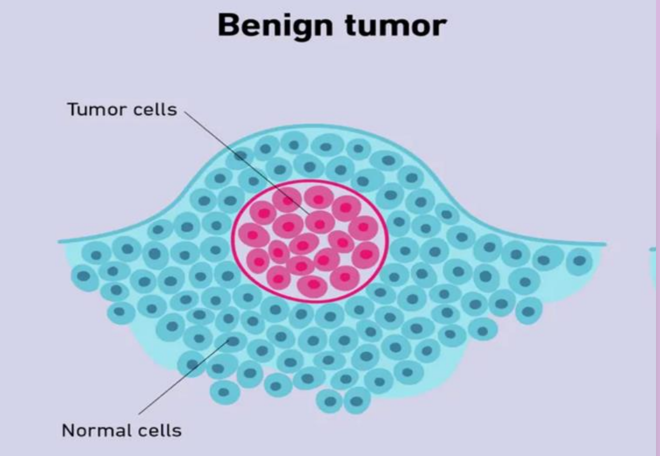
In this presentation, we will look at the two different types of tumors, and how we can use various features to distinguish one type from another. We will delve into how we applied Discriminant Analysis on the characteristic data of tumors such as size, shape, texture, location, and **histopathological** features to accurately classify tumors. Earlier identification of the type of tumor can ultimately lead to better diagnosis and treatment options for patients.

***Histopathological –*** *study of diseased cells and tissues using a microscope.*

# MALIGNANT AND BENIGN TUMORS



| Malignant Tumors: |
|---|
| Cancerous |
| May invade surrounding tissue |
| Most grow rapidly* |
| Irregular shape |
| Needs treatment |



| Benign Tumors: |
|---|
| Not cancerous |
| Doesn’t invade surrounding tissue |
| Most grow slowly* |
| Smooth shape |
| May not need treatment |

# DATASET

SOURCE: KAGGLE

The dataset contains 570 rows of different tumor data containing 30 different features of the tumor and their actual classification(as Benign or Malignant). The following are the attributes in this dataset.

| | |
|---|---|
| 1 | id |
| 2 | radius_mean |
| 3 | texture_mean |
| 4 | perimeter_mean |
| 5 | area_mean |
| 6 | smoothness_mean |
| 7 | compactness_mean |
| 8 | concavity_mean |
| 9 | concave points_mean |
| 10 | symmetry_mean |
| 11 | fractal_dimension_mean |
| 12 | radius_se |
| 13 | texture_se |
| 14 | perimeter_se |
| 15 | area_se |
| 16 | smoothness_se |
| 17 | compactness_se |
| 18 | concavity_se |
| 19 | concave points_se |
| 20 | symmetry_se |
| 21 | fractal_dimension_se |
| 22 | radius_worst |
| 23 | texture_worst |
| 24 | perimeter_worst |
| 25 | area_worst |
| 26 | smoothness_worst |
| 27 | compactness_worst |
| 28 | concavity_worst |
| 29 | concave points_worst |
| 30 | symmetry_worst |
| 31 | fractal_dimension_worst |
| 32 | **Diagnosis** |

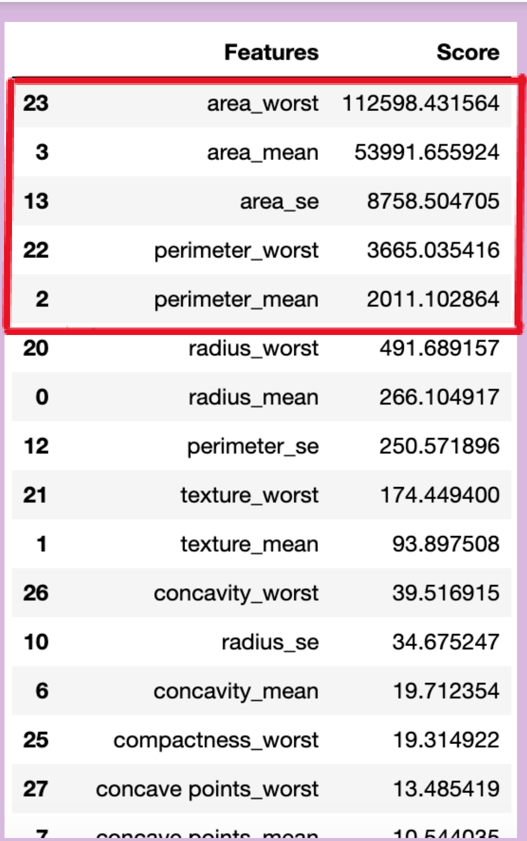
# OPTIMIZATION

**Objective:**
The main goal is to maximize the **Accuracy** of Cancer Prediction using **Discriminant Analysis**.

- Predicting the cancer with a very high accuracy improves the business of a diagnostic centre.

**Best Feature Selection**:

- We have 30 columns(features of tumor) in our dataset but most of them are redundant and doesn't contribute to the determination of the type of tumor. So, we used Genetic Algorithm to find the best 5 high quality features that are relevant to the problem.
- We found this using 'GeneticSelectionCV' module from sklearn-genetic package in Python.
  The best features are shown in the adjacent image.



| | Features | Score |
|---|---|---|
| **23** | area_worst | 112598.431564 |
| **3** | area_mean | 53991.655924 |
| **13** | area_se | 8758.504705 |
| **22** | perimeter_worst | 3665.035416 |
| **2** | perimeter_mean | 2011.102864 |
| **20** | radius_worst | 491.689157 |
| **0** | radius_mean | 266.104917 |
| **12** | perimeter_se | 250.571896 |
| **21** | texture_worst | 174.449400 |
| **1** | texture_mean | 93.897508 |
| **26** | concavity_worst | 39.516915 |
| **10** | radius_se | 34.675247 |
| **6** | concavity_mean | 19.712354 |
| **25** | compactness_worst | 19.314922 |
| **27** | concave points_worst | 13.485419 |
| **7** | concave points_mean | 10.544035 |

# OPTIMIZATION

***Discriminant Analysis* (optimization model that we adopted):**

- It is a statistic tool used in business and marketing by analysts to classify/categorize certain data. If a high % of data is correctly classified, then the Analysis is said to be successful.

**Objective Function**: Accuracy (of prediction) - must be Maximized.

**Features deciding the prediction of tumor:** area_worst, area_mean, area_se, perimeter_worst, perimeter_mean

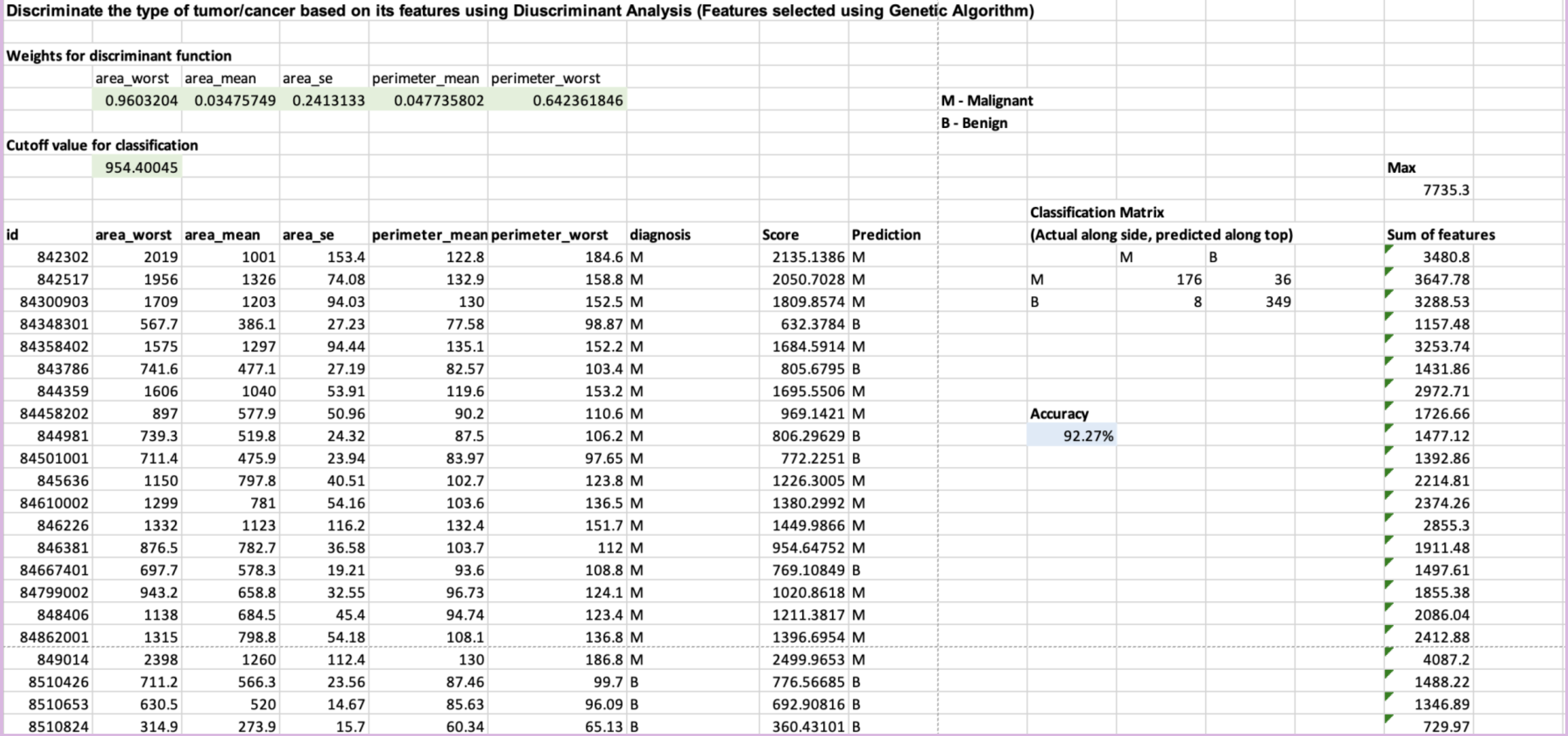
**Constraints for Cutoff value of score**: $-7735.3 < \text{Cutoff} < 7735.3$

*Score = SUMPRODUCT(Features,Weights).*

*Since the Maximum value of the sum of the 5 features is 7735.3, the cut-off value cannot be greater than 7735.3(if weights = 1) and cannot be smaller than –7735.3(if weights = -1)*

**Constraints for Weights of features**: $-1 < \text{Weights} < 1$

# OPTIMIZATION RESULTS & ACCURACY



Discriminate the type of tumor/cancer based on its features using Diuscriminant Analysis (Features selected using Genetic Algorithm)

**Weights for discriminant function**

| area_worst | area_mean | area_se | perimeter_mean | perimeter_worst |
|---|---|---|---|---|
| 0.9603204 | 0.03475749 | 0.2413133 | 0.047735802 | 0.642361846 |

**M - Malignant**

**B - Benign**

**Cutoff value for classification**

954.40045

**Max**

7735.3

**Classification Matrix**

**(Actual along side, predicted along top)**

| | M | B |
|---|---|---|
| M | 176 | 36 |
| B | 8 | 349 |

**Accuracy**

92.27%

| id | area_worst | area_mean | area_se | perimeter_mean | perimeter_worst | diagnosis | Score | Prediction | Sum of features |
|---|---|---|---|---|---|---|---|---|---|
| 842302 | 2019 | 1001 | 153.4 | 122.8 | 184.6 | M | 2135.1386 | M | 3480.8 |
| 842517 | 1956 | 1326 | 74.08 | 132.9 | 158.8 | M | 2050.7028 | M | 3647.78 |
| 84300903 | 1709 | 1203 | 94.03 | 130 | 152.5 | M | 1809.8574 | M | 3288.53 |
| 84348301 | 567.7 | 386.1 | 27.23 | 77.58 | 98.87 | M | 632.3784 | B | 1157.48 |
| 84358402 | 1575 | 1297 | 94.44 | 135.1 | 152.2 | M | 1684.5914 | M | 3253.74 |
| 843786 | 741.6 | 477.1 | 27.19 | 82.57 | 103.4 | M | 805.6795 | B | 1431.86 |
| 844359 | 1606 | 1040 | 53.91 | 119.6 | 153.2 | M | 1695.5506 | M | 2972.71 |
| 84458202 | 897 | 577.9 | 50.96 | 90.2 | 110.6 | M | 969.1421 | M | 1726.66 |
| 844981 | 739.3 | 519.8 | 24.32 | 87.5 | 106.2 | M | 806.29629 | B | 1477.12 |
| 84501001 | 711.4 | 475.9 | 23.94 | 83.97 | 97.65 | M | 772.2251 | B | 1392.86 |
| 845636 | 1150 | 797.8 | 40.51 | 102.7 | 123.8 | M | 1226.3005 | M | 2214.81 |
| 84610002 | 1299 | 781 | 54.16 | 103.6 | 136.5 | M | 1380.2992 | M | 2374.26 |
| 846226 | 1332 | 1123 | 116.2 | 132.4 | 151.7 | M | 1449.9866 | M | 2855.3 |
| 846381 | 876.5 | 782.7 | 36.58 | 103.7 | 112 | M | 954.64752 | M | 1911.48 |
| 84667401 | 697.7 | 578.3 | 19.21 | 93.6 | 108.8 | M | 769.10849 | B | 1497.61 |
| 84799002 | 943.2 | 658.8 | 32.55 | 96.73 | 124.1 | M | 1020.8618 | M | 1855.38 |
| 848406 | 1138 | 684.5 | 45.4 | 94.74 | 123.4 | M | 1211.3817 | M | 2086.04 |
| 84862001 | 1315 | 798.8 | 54.18 | 108.1 | 136.8 | M | 1396.6954 | M | 2412.88 |
| 849014 | 2398 | 1260 | 112.4 | 130 | 186.8 | M | 2499.9653 | M | 4087.2 |
| 8510426 | 711.2 | 566.3 | 23.56 | 87.46 | 99.7 | B | 776.56685 | B | 1488.22 |
| 8510653 | 630.5 | 520 | 14.67 | 85.63 | 96.09 | B | 692.90816 | B | 1346.89 |
| 8510824 | 314.9 | 273.9 | 15.7 | 60.34 | 65.13 | B | 360.43101 | B | 729.97 |

# OPTIMIZATION RESULTS & ACCURACY

| | |
|---|---|
| Total number of tumor samples | 569 |
| Number of samples correctly predicted as Malignant | 176 |
| Number of samples correctly predicted as Benign | 349 |
| Number of Benign tumors wrongly predicted as Malignant | 8 |
| Number of Malignant tumors wrongly predicted as Benign | 36 |
| **Accuracy** (Percentage of tumor samples that are correctly predicted) | = (349+176)/569<br>= 0.9227<br>= **92.27 %** |

**CONCLUSION:**

- We managed to achieve a very good Maximum Accuracy of 92.27% by choosing the best features for Discriminative analysis using Genetic Algorithm. (We previously got a maximum accuracy of only 70% when we used all the features for Discriminative Analysis)
- This prediction can also be achieved by Multiple Linear Regression by training and testing a model. We wish to perform this and compare the results and accuracy of both models.

# THANKS FOR YOUR ATTENTION!!

ANY QUESTIONS!?