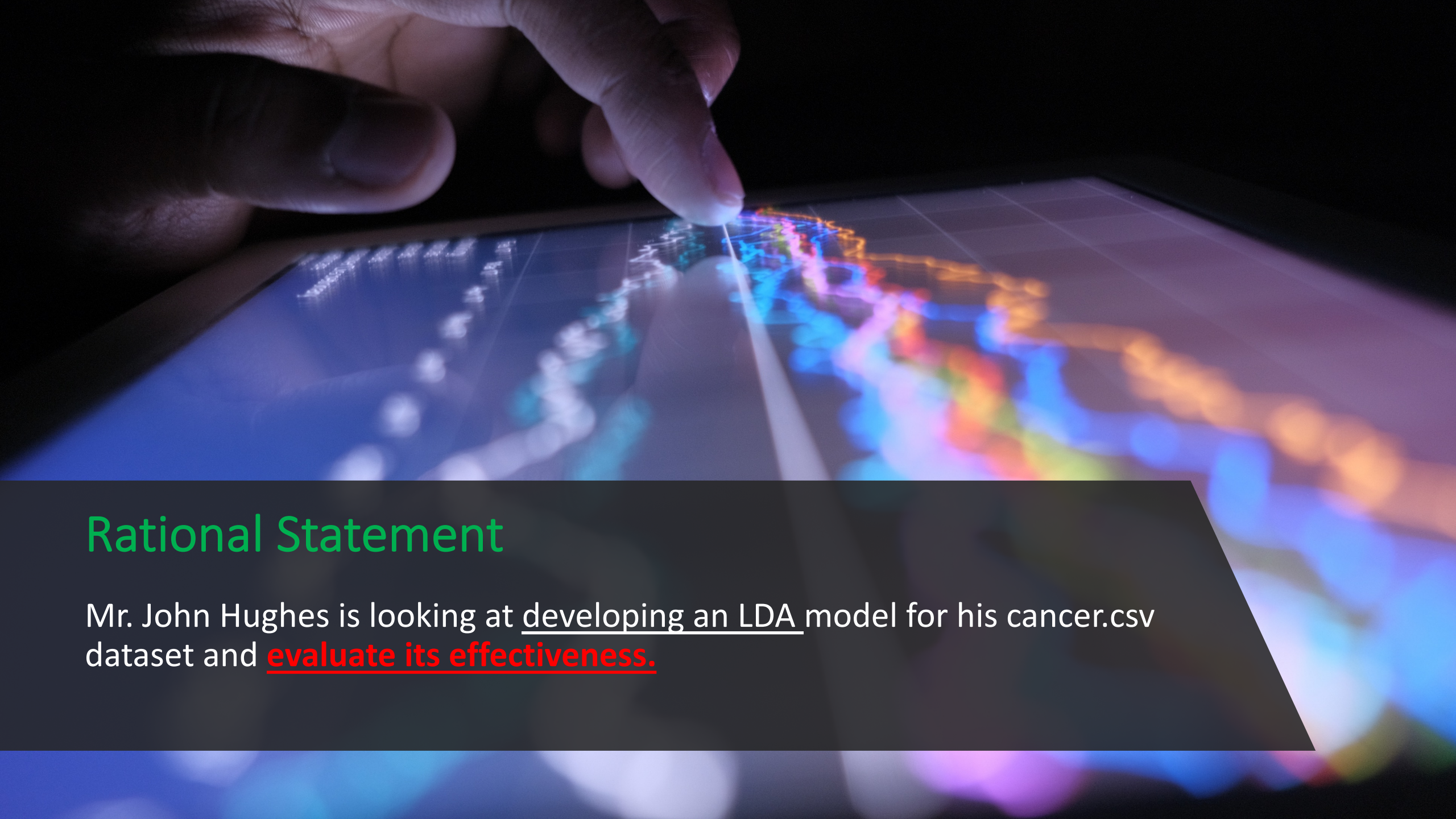
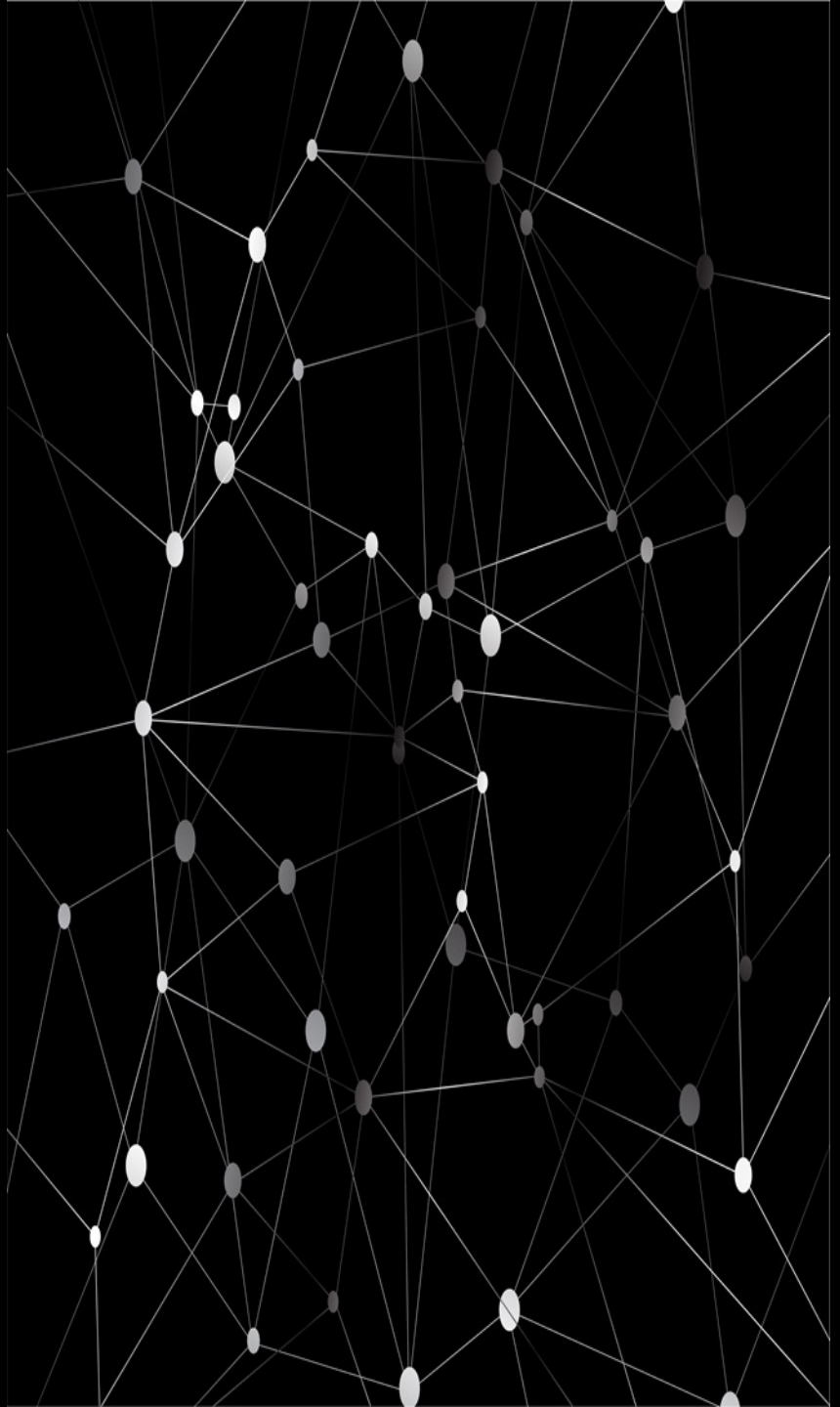
Discriminant Analysis

Prepared By: Rana Khan



Rational Statement

Mr. John Hughes is looking at developing an LDA model for his cancer.csv dataset and **evaluate its effectiveness.**



Independent Variable

- ID - ID number
- Clump Thickness - 1-10
- UofCSize - Uniformity of Cell Size 1-10
- UofShape - Uniformity of Cell Shape 1-10
- Marginal Adhesion - 1-10
- SECSize - Single Epithelial Cell Size 1-10
- Bare Nuclei - 1-10
- Bland Chromatin - 1-10
- Normal Nucleoli - 1-10
- Mitoses - 1-10

Dependent Variable

- Benign (i.e. No Cancer) - 2
- Malignant (i.e. Cancer) - 4

Navigation Synopsis

Copy cancer.csv into Pythondata2204 directory. Ensured the file is called cancer.csv

Launched Jupyter NoteBook

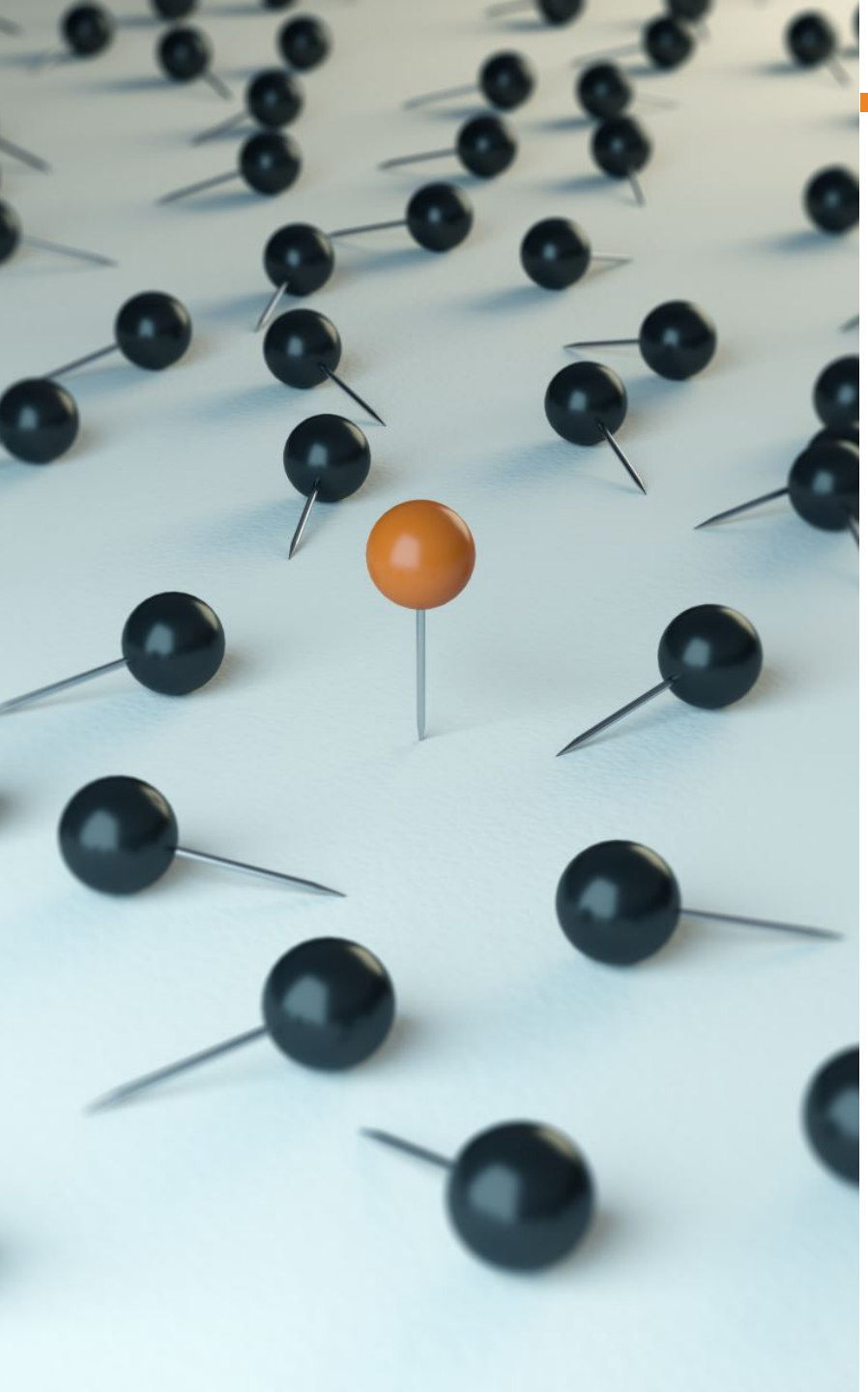
Navigated to Pythondata2204 Directory

Created a new Python NoteBook by clicking on “New Drop Down and Choose “Python3”

Renamed NoteBook by Clicking on File->Rename

Entered “Assignment#3 – Discriminant Analysis” and Clicked OK

Key insights from Panda Profile Report



1.KEY INSIGHTS FROM PANDA PROFILE

Dataset statistics

Number of variables	11
Number of observations	683
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	8
Duplicate rows (%)	1.2%

Variable types

Numeric	10
Categorical	1

1. The fundamental purpose to perform a profile is to determine and investigate different aspects of the dataset while performing EDA.

Using **SMOTE** methodology, we stabilized and balance the class of the dataset while performing Logistic, LDA, and QDA models to support in deciding which model is suitable in the given scenario.

2. KEY INSIGHTS FROM PANDA PROFILE

`Clump Thickness` is highly correlated with `UofCSize` and 6 other fields	High correlation
`UofCSize` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`UofCShape` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Marginal Adhesion` is highly correlated with `UofCSize` and 6 other fields	High correlation
`SECSize` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Bare Nuclei` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Bland Chromatin` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Normal Nucleoli` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Class` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Clump Thickness` is highly correlated with `UofCSize` and 7 other fields	High correlation
`UofCSize` is highly correlated with `Clump Thickness` and 8 other fields	High correlation
`UofCShape` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Marginal Adhesion` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`SECSize` is highly correlated with `Clump Thickness` and 7 other fields	High correlation
`Bare Nuclei` is highly correlated with `Clump Thickness` and 7 other fields	High correlation

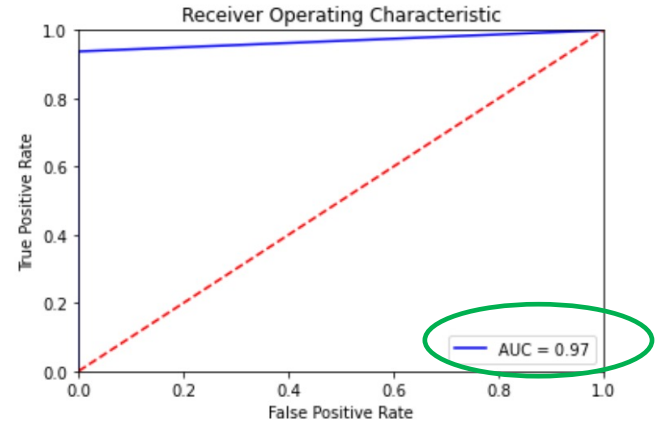
2. Following the snapshot from the Panda Profile, a clear determination can be made that almost all variables are **highly correlated** with each other.

Optimized LDA Classification Report

Key insights from the Optimized LDA Classification Report

```
Optimized Model

Model Name: LinearDiscriminantAnalysis()

Best Parameters: {'clf__solver': 'svd'}

 [[89  0]
 [ 3 45]]

              precision    recall  f1-score   support

     Class 2       0.97      1.00      0.98        89
     Class 4       1.00      0.94      0.97        48

    accuracy                           0.98       137
   macro avg       0.98      0.97      0.98       137
weighted avg       0.98      0.98      0.98       137
```

ROC Curve



1. In reference to the binary classification task, clearly the **higher f1 score the better**, *with 0 being the worst possible* and **1 being the best**. **Our model f1-score is 0.98** which is pretty close to ideal outcomes of higher-end.
2. 2. In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without disease based on test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than **0.9 comes under the category of outstanding**. Since in **our model AUC Curve is 0.97** which indicates model performance is **outstanding.**

3. Given the scenario even though we were in possession of enough data as well as with no hyper-parameter in actual discriminating analysis **LDA Optimized model functionate quite well.**

Comparison between Optimized LDA and Optimized Logistical Regression

Comparison

Optimized LDA Model

```
Optimized Model

Model Name: LinearDiscriminantAnalysis()

Best Parameters: {'clf__solver': 'svd'}

 [[89  0]
 [ 3 45]]
```

	precision	recall	f1-score	support
Class 2	0.97	1.00	0.98	89
Class 4	1.00	0.94	0.97	48
accuracy			0.98	137
macro avg	0.98	0.97	0.98	137
weighted avg	0.98	0.98	0.98	137

1. In optimized model f1-score is 0.98
2. In optimized model accuracy level is 0.98
3. In the optimized model precision is 0.98

Optimized Logistical Regression Model

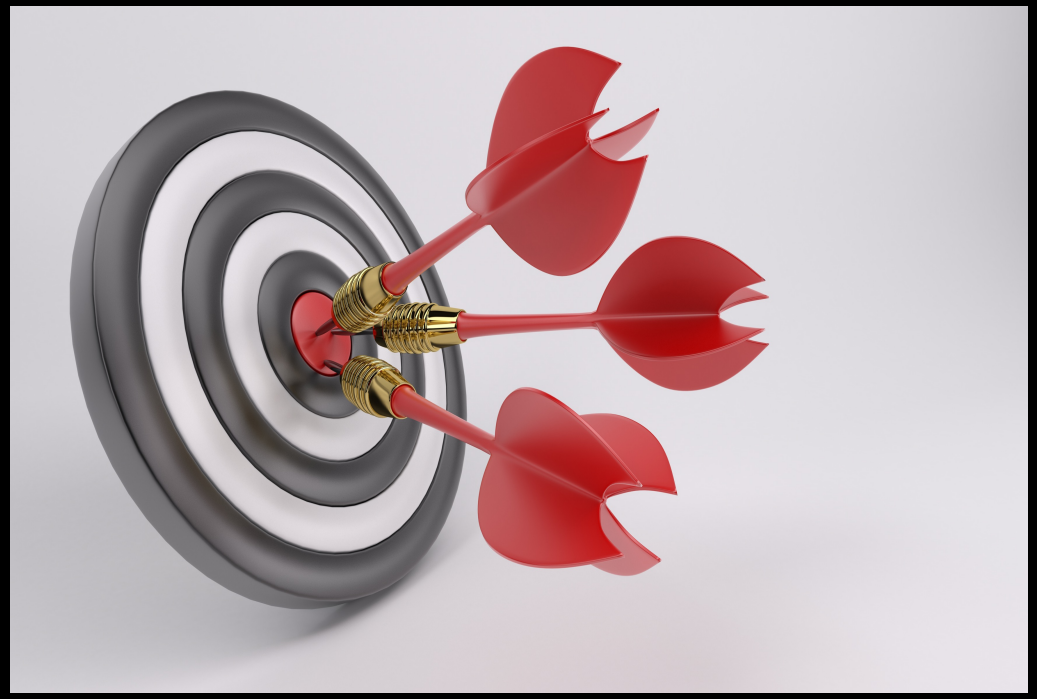
```
Optimized Model

Model Name: LogisticRegression(class_weight='balanced', random_state=100,
                   solver='liblinear')

Best Parameters: {'clf__C': 100, 'clf__penalty': 'l1'}

 [[89  0]
 [ 2 46]]
```

	precision	recall	f1-score	support
Class 2	0.98	1.00	0.99	89
Class 4	1.00	0.96	0.98	48
accuracy			0.99	137
macro avg	0.99	0.98	0.98	137
weighted avg	0.99	0.99	0.99	137

1. In Optimized Logistical Regression model f1-score is 0.99
2. In Optimized Logistical Regression model accuracy level is 0.99
3. In Optimized Logistical Regression model precision is 0.99

Important Note: *The results show that overall, while considering the integral of the model including f1-score, the accuracy of the model, and precision,* ***the Optimized Logistical Regression Model performed better regardless*** *of the distribution of the data is normal or nonnormal.*

Recommendations



Based on the outcome of the model, the followings are the **recommendations** which can be furnished to Mr. John Hughes:

1. In the given problem, while there is a need to identify the effectiveness of the dataset to examine if the patient has No Cancer, Benign, 2 OR Cancer, Malignant, 4 while *both models performed quite well identically to each other.*

2. Taking into consideration trends associated with the f1 score, precision has been observed close in both Optimized LDA and Optimized Logistic Regression. BUT, in common industry practice **accuracy of the model is highly valuable** and it has been found as **99%** for Optimized Logistic Regression Model over 98% for Optimized LDA mode so recommendation goes towards Optimized Logistic Regression Model.

*** *Python HTML file is attached for reference*