# Differentiated Thyroid Cancer Prediction Model

*SC1015 Mini Project*

**Group 8: Claire Chia Wan Ni (U2430719H), Narmandakh Nomt (U2430263A)**

# WHAT IS IT?

**Most Common** Cause of Thyroid Cancer

↓

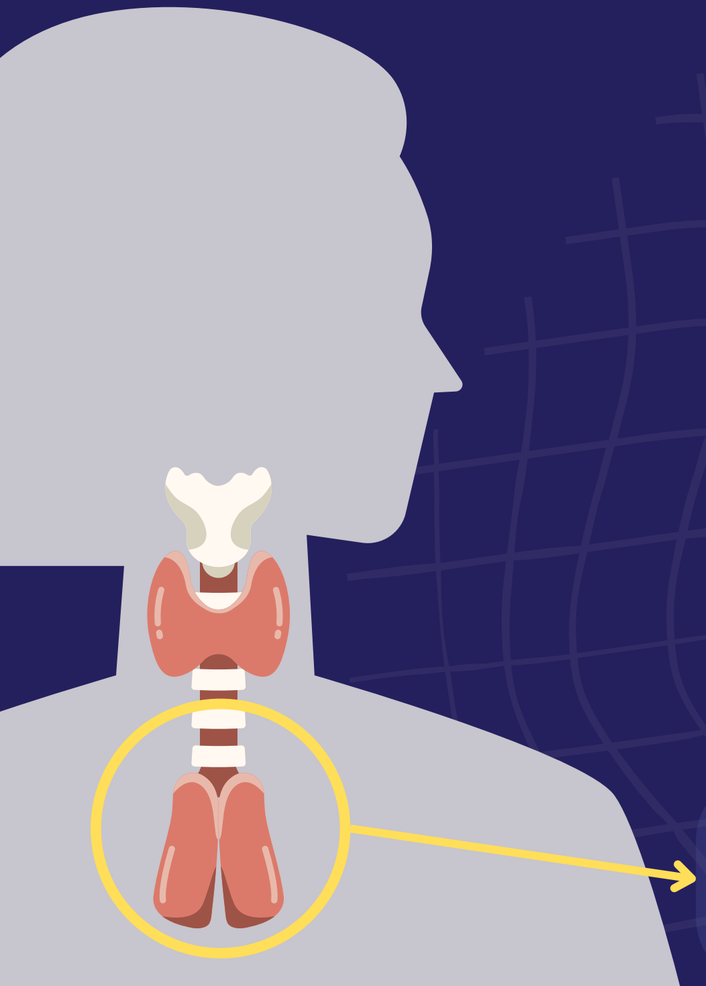
Accounts for > 90% of Thyroid Cancer cases

**CHALLENGE:**
**UNPREDICTABILITY**

- 30% of patients experience recurrence
- Causes Clinical Complications
- Emotional Distress and Financial Burden

**Why is it unpredictable?**

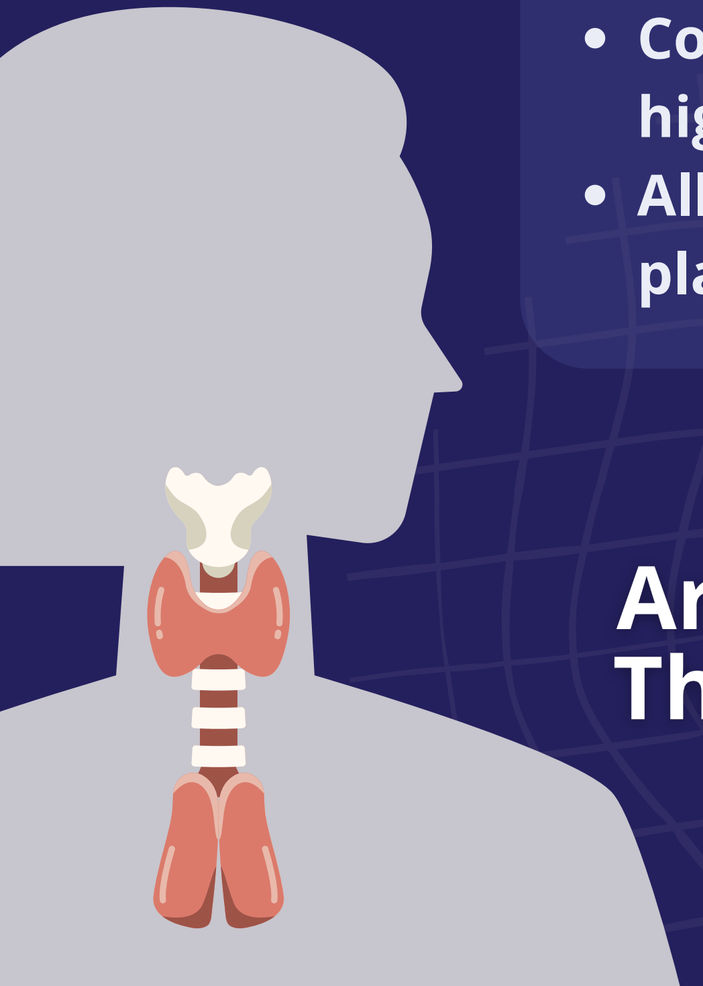
- Assessed by **Physicians' Experience** and **Generalized Clinical Guidelines**
- Overlook nuanced, **personalised factors** that lead to **recurrence**



Originates from the Follicular Cells

# Our Objective

- Identify hidden patterns across clinical attributes
- ~~Replace Clinical Decision-Making~~
- Complement Clinical Decision Making in flagging high-risk patients earlier
- Allows for more tailored monitoring and treatment plans



## Problem Statement:

**Are we able to Predict Recurrence of Thyroid Cancer based on Individuals' Physical Attributes?**


# EXPLORATORY DATA ANALYSIS

# Exploratory Data Analysis

UCI Machine Learning Repository

Thyroid_Diff.csv

- Tracks 17 Clinicopathologic Features
- 383 Thyroid Cancer Patients
- Long-term, curated dataset

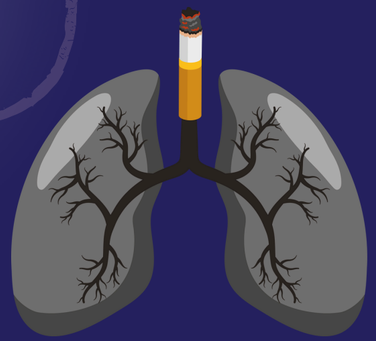
# Statistical Profile

```
thyroiddata.describe()
```

✓ 0.0s

| | Age |
|---|---|
| count | 383.000000 |
| mean | 40.866841 |
| std | 15.134494 |
| min | 15.000000 |
| 25% | 29.000000 |
| 50% | 37.000000 |
| 75% | 51.000000 |
| max | 82.000000 |

Thyroid_Diff.csv

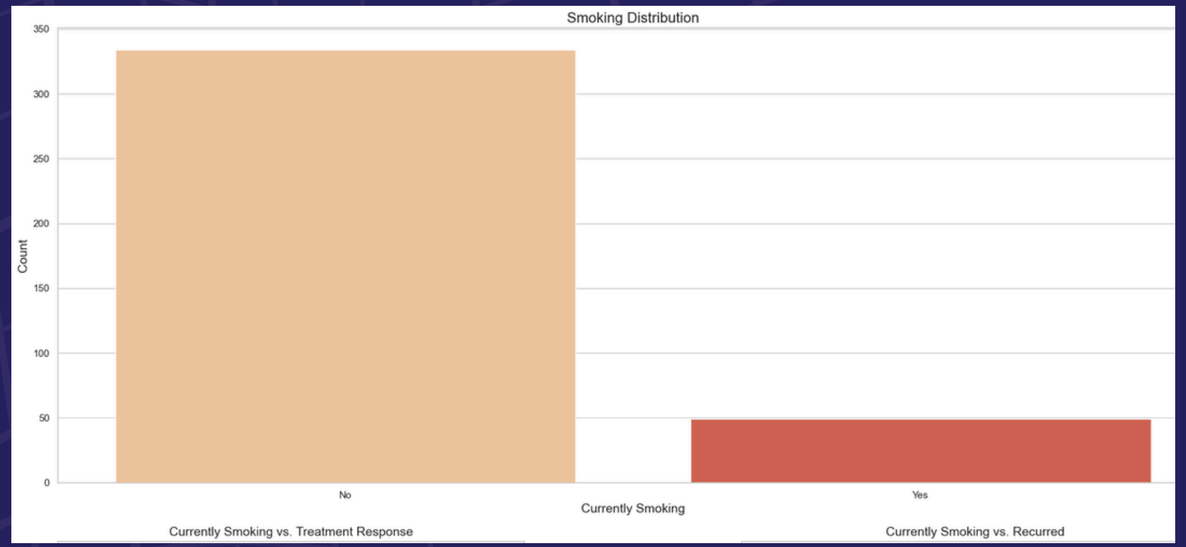


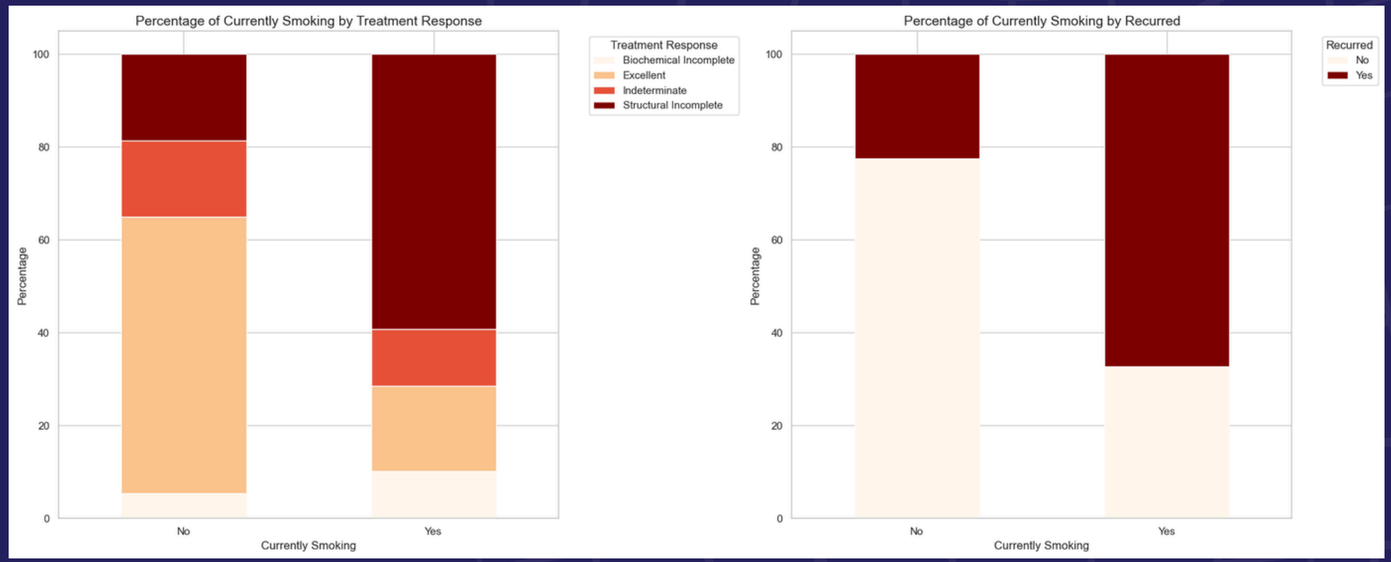
# Data Visualisation

## 1. Categorical Visualisation of Smoking

### Count Plots

- Visualise distribution of smokers based on given data



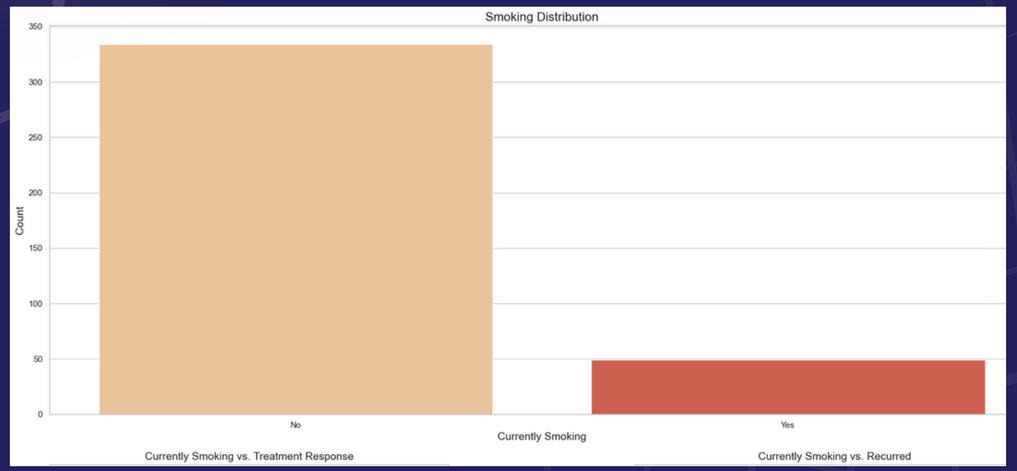

### Bar Graphs

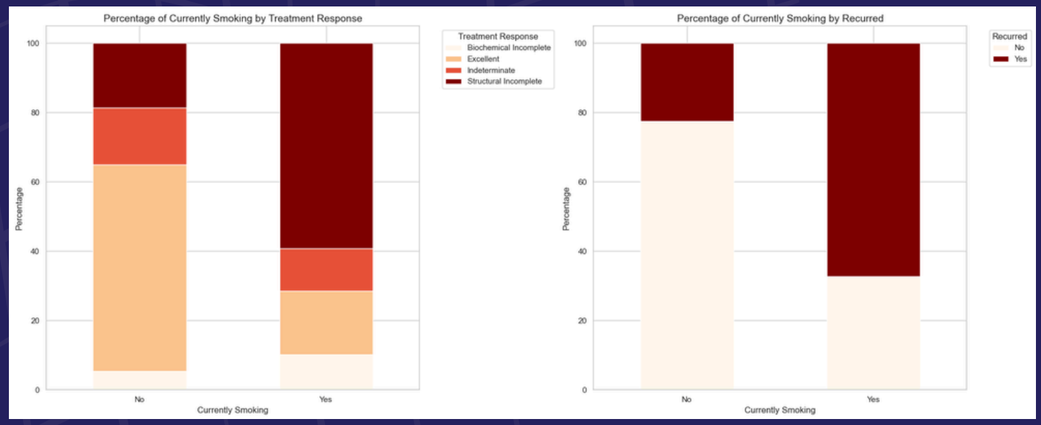
- Visualise statistics of smokers based on our calculated percentages

*Exploratory Data Analysis*

# Data Visualisation

## 1. Categorical Visualisation of Smoking




**1. Majority of Thyroid Cancer patients do not smoke.**

**2. Smoking reduces the effectiveness of the initial thyroid cancer treatment.**

-17% of patients who are currently smoking experiences excellence in initial treatment response, with about 60% experiencing structural incomplete.

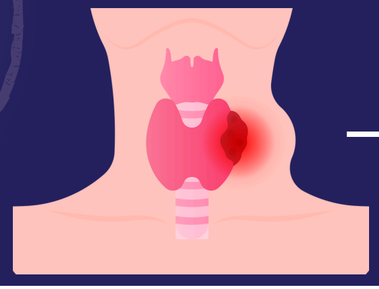
**3. Smoking increases the likelihood of Thyroid Cancer Recurrence.**

- About 67% experiences a recurrence of the thyroid cancer, which is significantly higher than the < 25% of patients who aren't smoking .

**While most patients were non-smokers, smokers consistently showed poorer recovery and higher recurrence rates, marking smoking as a strong risk factor for our model.**

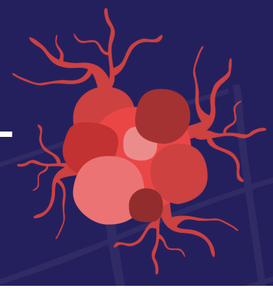
*More explained in the notebook

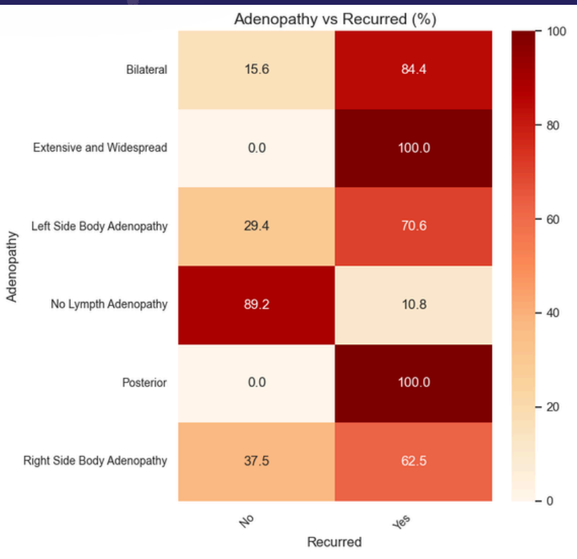
*Exploratory Data Analysis*

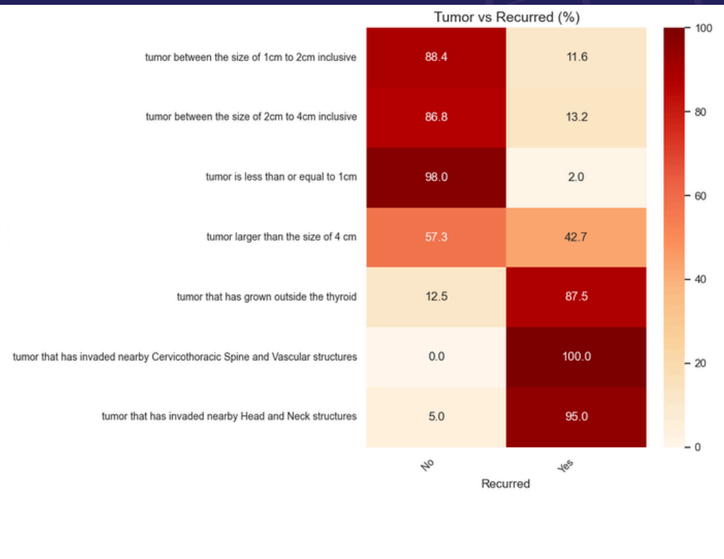


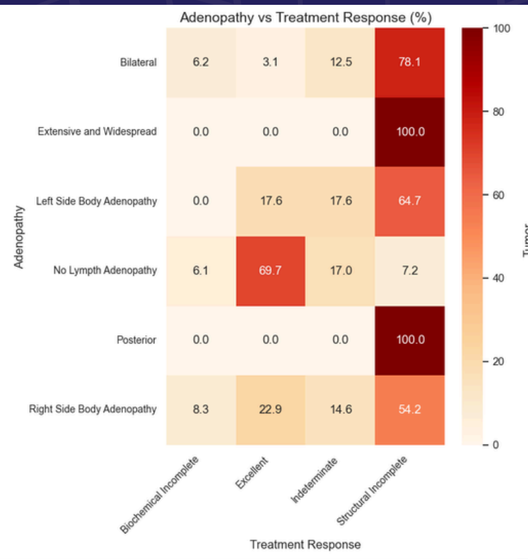
# Data Visualisation

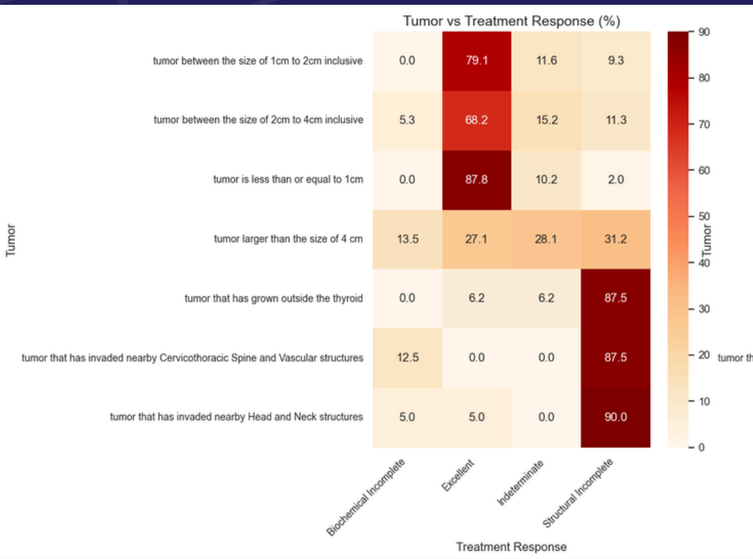
## 2. Bivariate Heatmap of Adenopathy and Tumor (Type)






## Heatmaps

- Strong **Intensity Gradient** and **correlation**, easier for visualisation

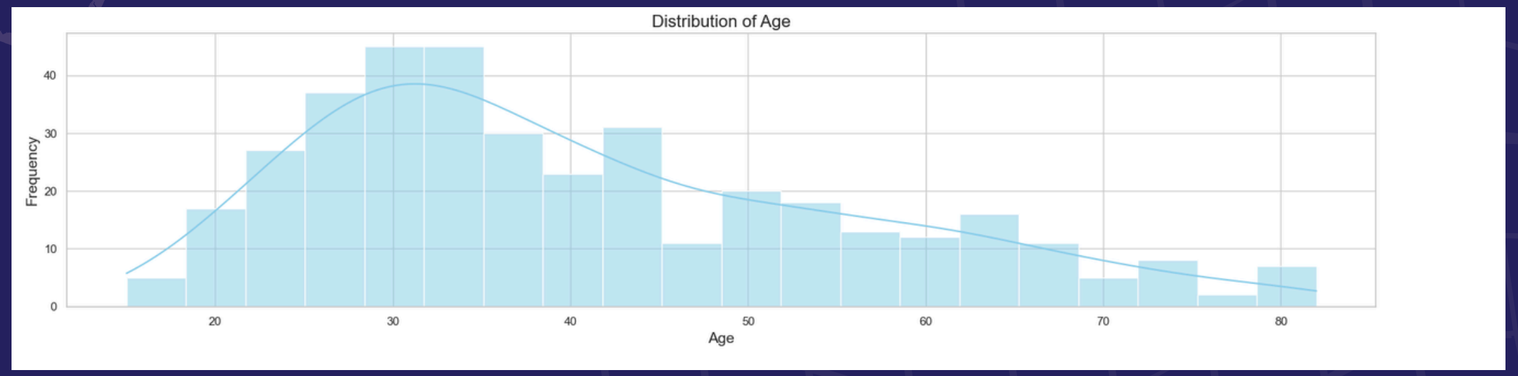
**Patients with Adenopathy or Tumors extending beyond the thyroid had a much higher likelihood of recurrence and poorer treatment outcome, suggesting these conditions are major clinical indicators.**

*More explained in the notebook

*Exploratory Data Analysis*

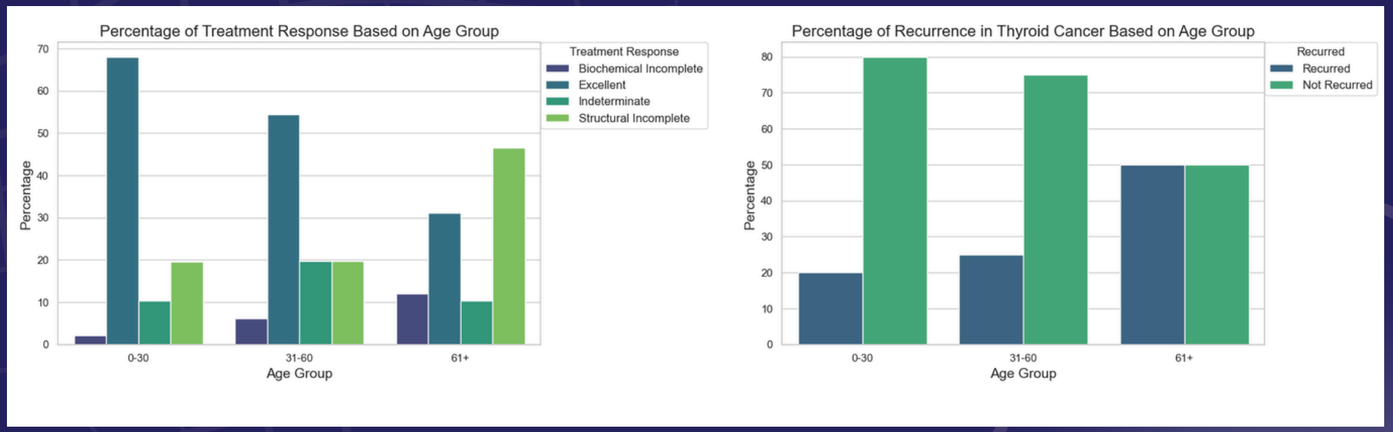
# Data Visualisation

## 3. Univariate and Distribution Visualisation of Age



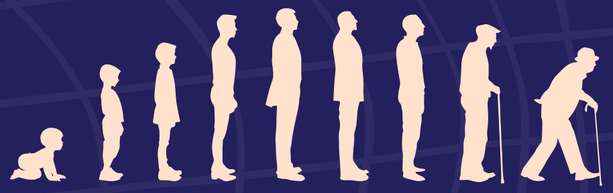

### Histogram

- Shows the frequency distribution of age, helping us see which age ranges are most common in the dataset.




### Bar Graph

- Visualise statistics of responses and recurrence by age based on our calculated percentages

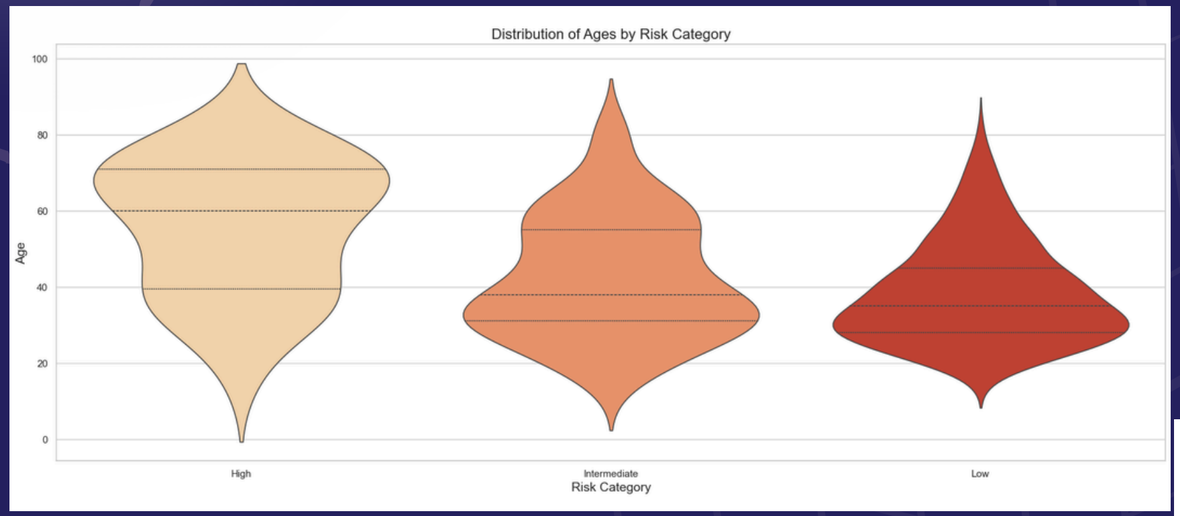


*More explained in the notebook

*Exploratory Data Analysis*

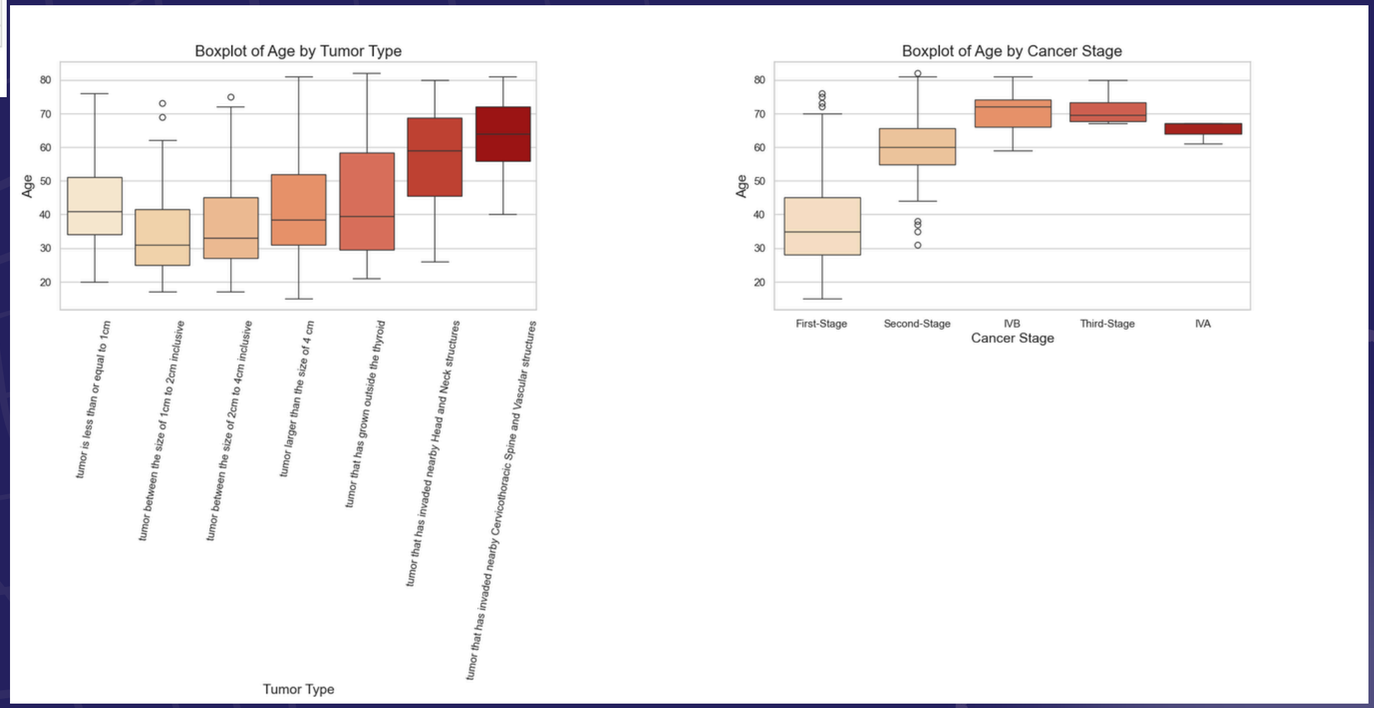
# Data Visualisation

## 3. Univariate and Distribution Visualisation of Age




### Violin Plot

- Shows distribution and density of age data across recurrence outcomes
- Illustrates how common certain age ranges are



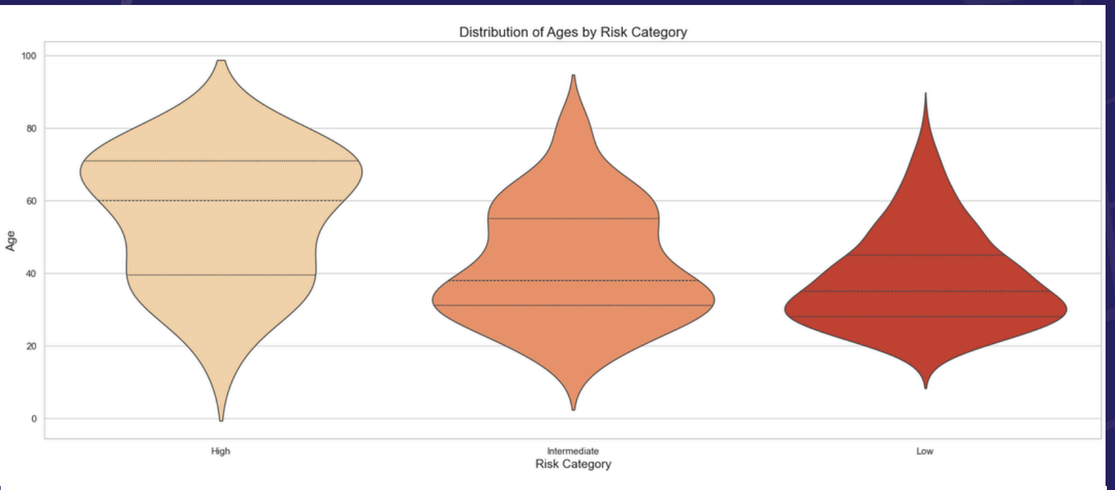

### Box Plot

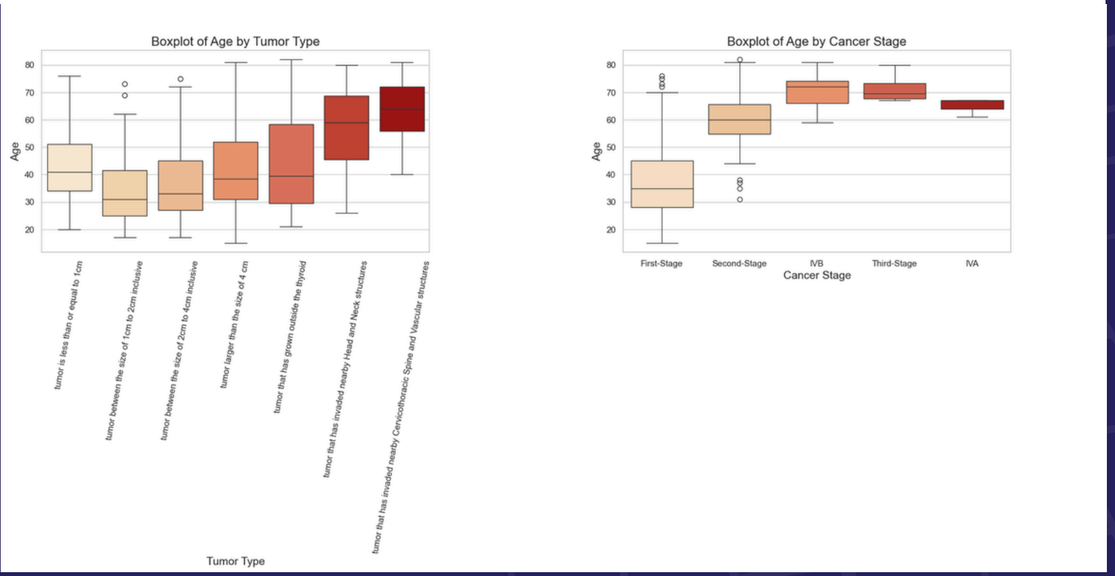
- To identify the median, spread, and outliers in age data across different recurrence outcomes.

*More explained in the notebook

# Data Visualisation

## 3. Univariate and Distribution Visualisation of Age







**Older patients have a higher probability of recurrence and poorer treatment response.**

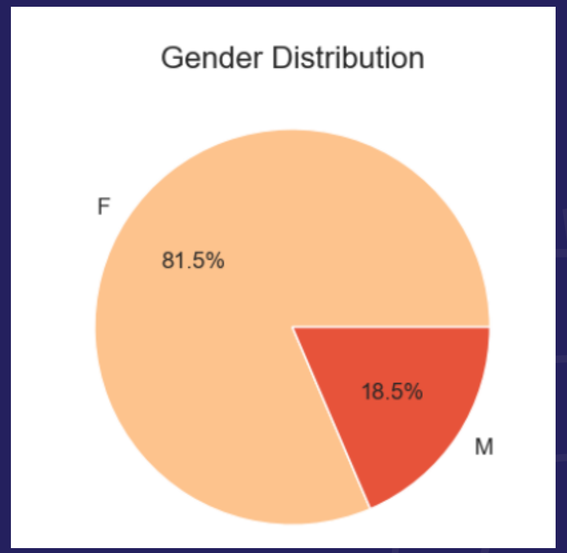
**Recurrence risk increases steadily with age.**

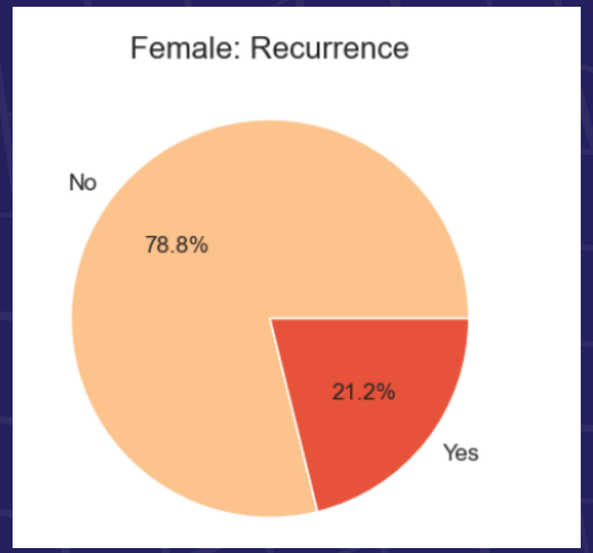
*More explained in the notebook

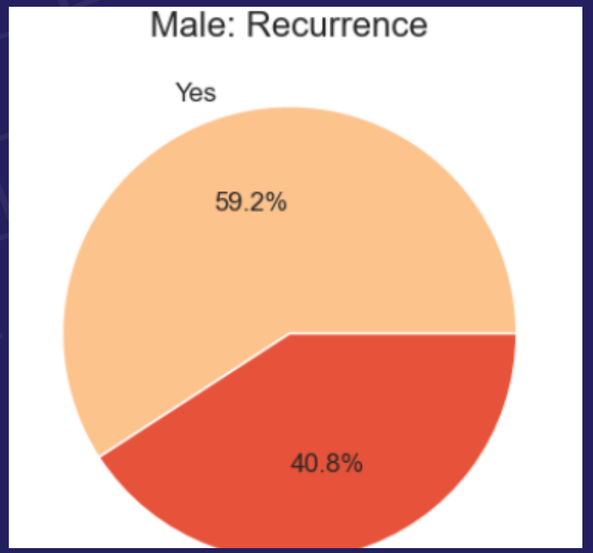
*Exploratory Data Analysis*

# Data Visualisation

## 4. Categorical Visualisation of Gender




### Pie Chart

- Illustrates proportional distribution of male and female patients

**While more females are likely to be diagosed, Male patients are more likely to experience severe complications that increases likelihood of recurrence and poorer treatment response.**

*More explained in the notebook

*Exploratory Data Analysis*

# Data Preparation

# 1. Exploratory data analysis

Medical termonoligies / abbreviations

| Risk | T | N | M | Stage |
|---|---|---|---|---|
| Low | T1a | N0 | M0 | I |
| Low | T1a | N0 | M0 | I |
| Low | T1a | N0 | M0 | I |
| Low | T1a | N0 | M0 | I |
| Low | T1a | N0 | M0 | I |

Understandable phrases

| Tumor | Lymph Nodes | Cancer Metastasis | Stage |
|---|---|---|---|
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |

# 2. Machine learning

- Encoding categorical variables into numerical ones

| Tumor | Lymph Nodes | Cancer Metastasis | Stage |
|---|---|---|---|
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |
| tumor is less than or equal to 1cm | no evidence of regional lymph node metastasis | no evidence of distant metastasis | First-Stage |

| Tumor | Lymph Nodes | Cancer Metastasis | Stage |
|---|---|---|---|
| 2 | 1 | 0 | 0 |
| 3 | 1 | 1 | 1 |
| 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 |

# 2. Machine learning

- Initializing train and test datasets

**75% - 25%**

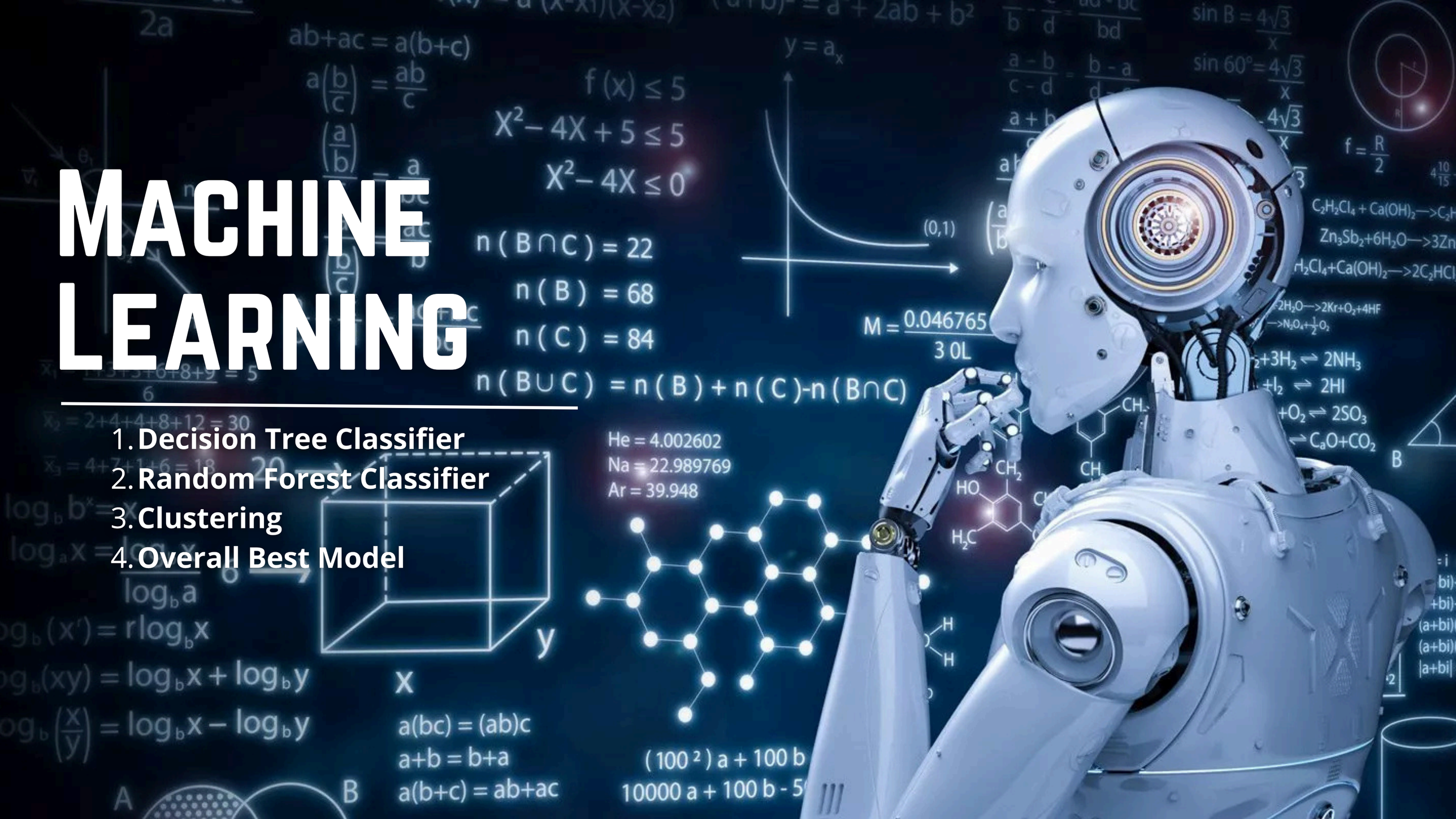
**The Target Variable**: "Recurrence”

**The Predictor Variables**: **Varies*

- **X:** "Age", "Gender", "Smoking History"
- **X2**: "Age", "Gender", "Smoking History", "Currently Smoking", "Adenopathy"
- **X3**: "Age", "Gender", "Currently Smoking", "Smoking History", "Adenopathy", "Risk", "Treatment Response"
- **X4**: All given variables

# Machine Learning

1. **Decision Tree Classifier**
2. **Random Forest Classifier**
3. **Clustering**
4. **Overall Best Model**



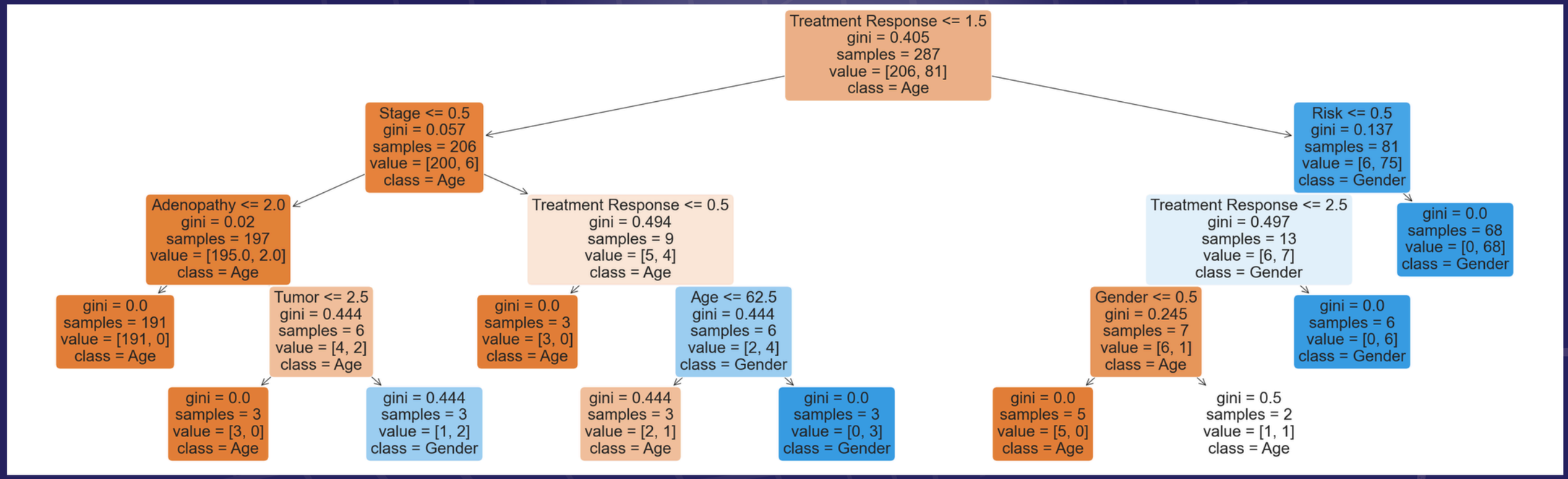
# 1. Decision tree Classifier

## Accuracy

**Train dataset:** 81.5% - 98.9%

**Test dataset:** 63.5% - **93.7%**

**X3_test**



# 1. Decision tree Classifier

X3_test

**Train Dataset**

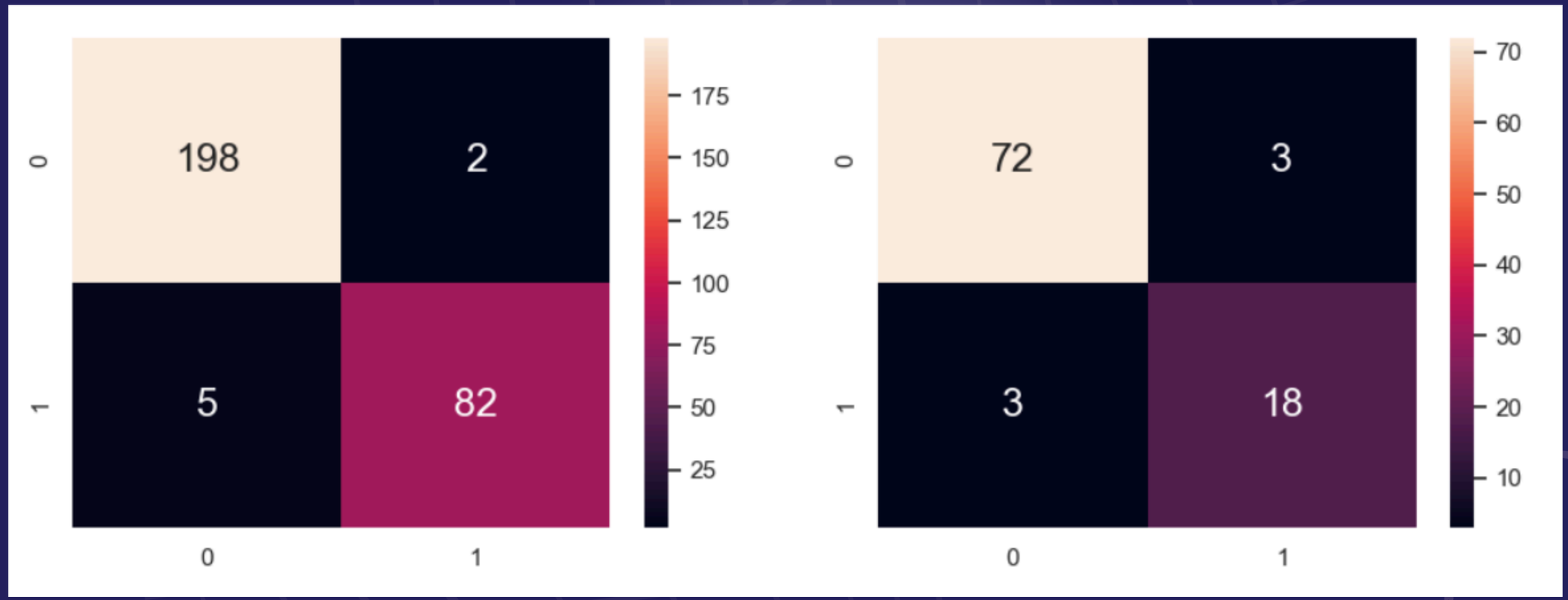
*tpr:* 0.9754 *fpr:* 0.02381

*fnr:* 0.0246 *tnr:* 0.9762

**Test Dataset**

*tpr:* 0.960 *fpr:* 0.14286

*fnr:* 0.04 *tnr:* 0.8571



# 2. Random Forest Tree

**Accuracy Range:** 0.6145 - 0.947

## X4_test

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| No | 0.96 | 0.97 | 0.96 | 69 |
| Yes | 0.92 | 0.89 | 0.91 | 27 |
| accuracy | | | 0.95 | 96 |
| macro avg | 0.94 | 0.93 | 0.93 | 96 |
| weighted avg | 0.95 | 0.95 | 0.95 | 96 |

| | 0 |
|---|---|
| Age | 0.050564 |
| Gender | 0.012299 |
| Currently Smoking | 0.006237 |
| Smoking History | 0.002227 |
| Radiotherapy History | 0.000190 |
| Thyroid Function | 0.005358 |
| Physical Examination | 0.016643 |
| Adenopathy | 0.117673 |
| Types of Thyroid Cancer (Pathology) | 0.009604 |
| Focality | 0.016834 |
| Risk | 0.184580 |
| Tumor | 0.061071 |
| Lymph Nodes | 0.062083 |
| Cancer Metastasis | 0.007518 |
| Stage | 0.031882 |
| Treatment Response | 0.415237 |

# 2. Random Forest Tree (with hyperparameters)

**Accuracy Range:** 0.66 - 0.96


## X3_test

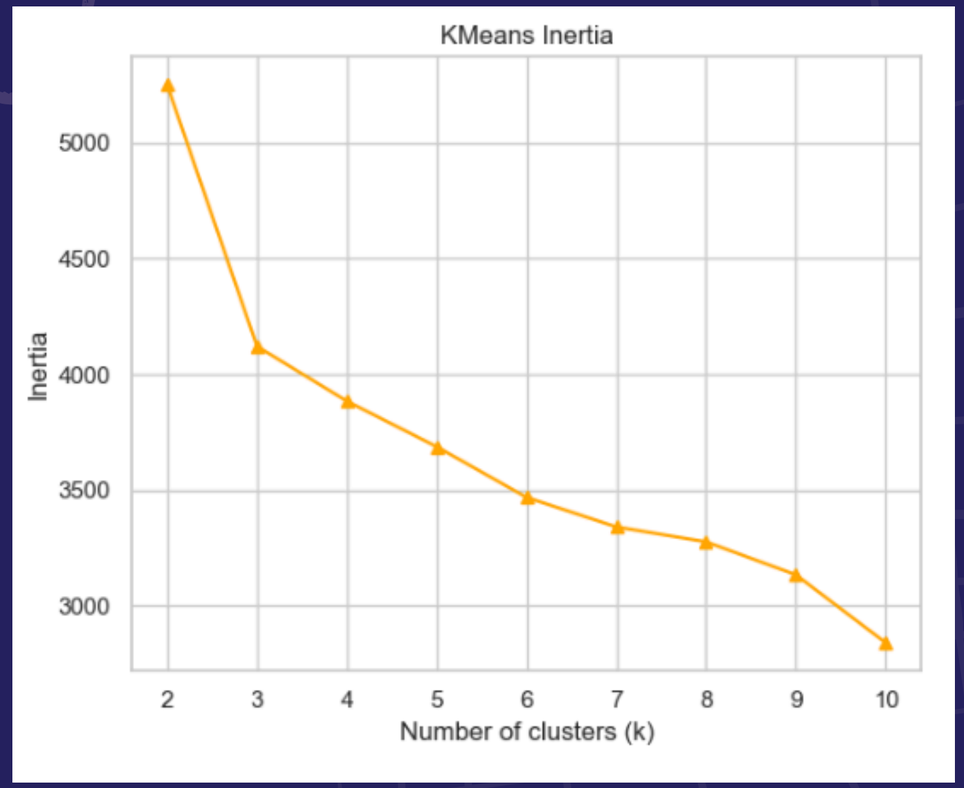
```
y3_pred = rf2.predict(X3_test)
print(classification_report(y3_test, y3_pred))
```

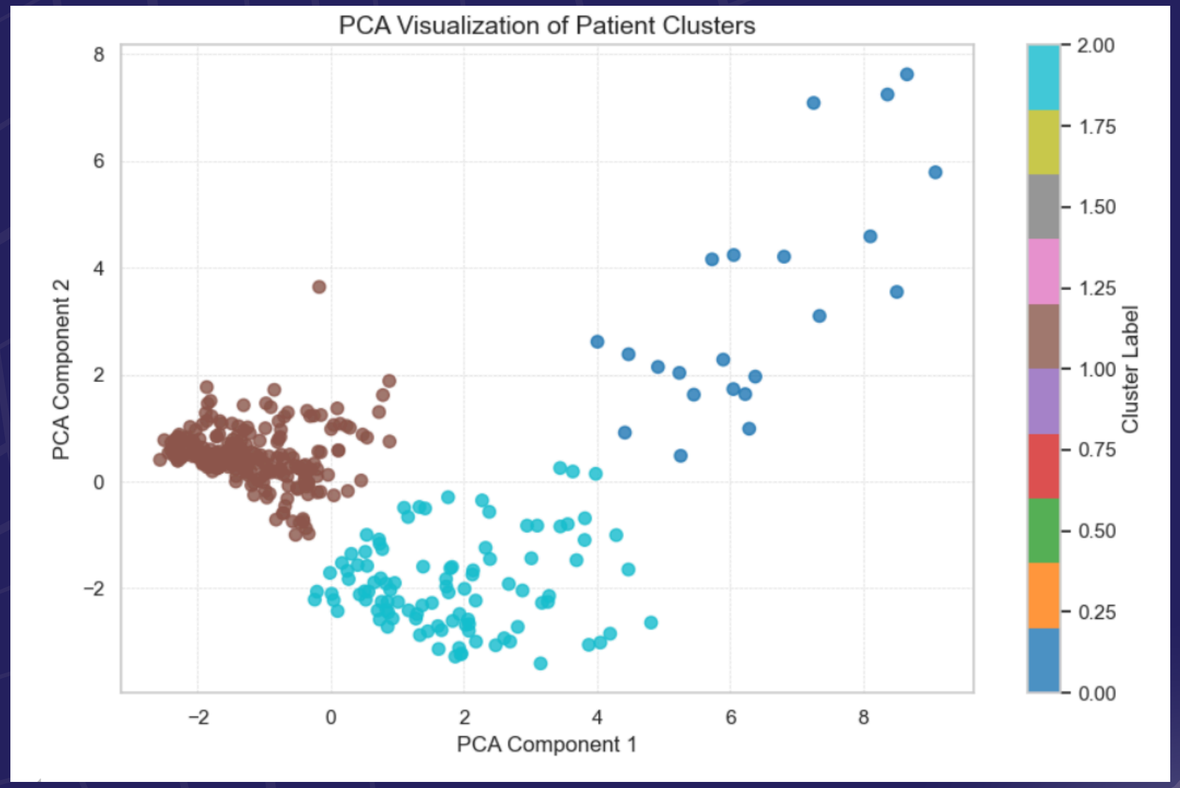
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| No | 0.96 | 1.00 | 0.98 | 75 |
| Yes | 1.00 | 0.86 | 0.92 | 21 |
| accuracy | | | 0.97 | 96 |
| macro avg | 0.98 | 0.93 | 0.95 | 96 |
| weighted avg | 0.97 | 0.97 | 0.97 | 96 |

### 2.1 Introducing hyperparameter tuning to increase accuracy

```
rf2 = RandomForestClassifier(n_estimators = 1000,
                             criterion = "entropy",
                             min_samples_split = 10,
                             max_depth = 14,
                             random_state = 42
)
```

# 3. K-Clustering




**KMeans Inertia**

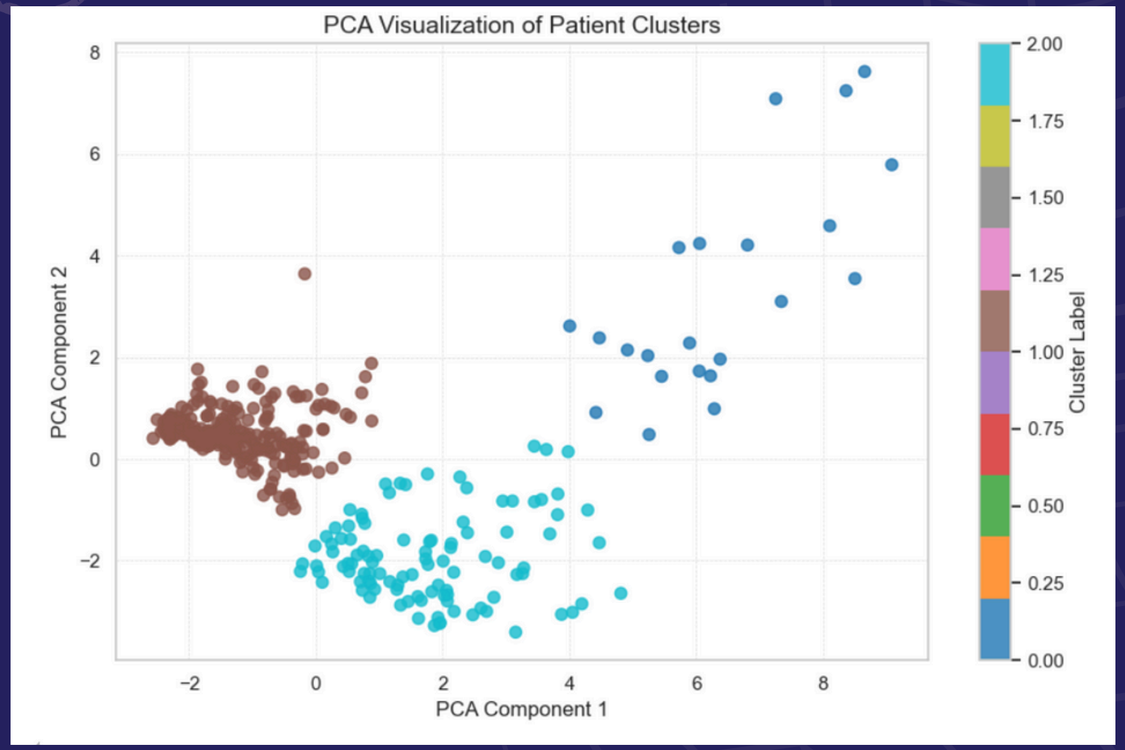
Optimal n of clusters = 3

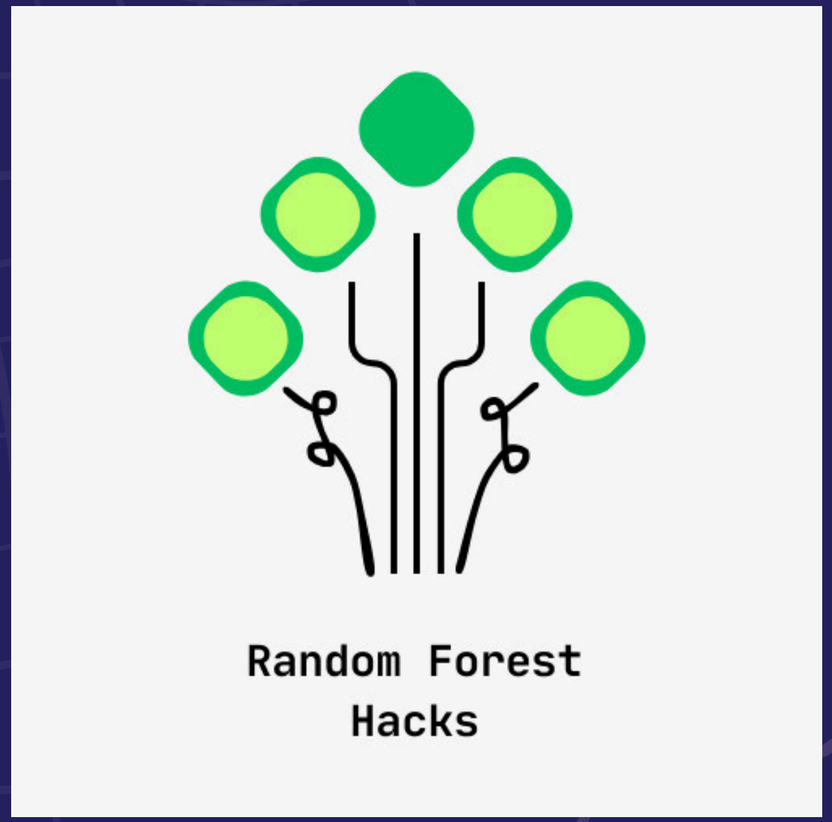
**Key advantages**

Min overlapping & density

# 4. Overall Best Model (Integration)

**Accuracy: 98%**




Machine Learning


# Insights & recommendations

# Insights


Through our analysis, we found strong relation of recurrence with the following:

**Age -** The older the patient is, the more likely to suffer from cancer recurrence

**Adenopathy / Tumor type** - Patient with lymphatic spread should be considered high-risk even before recurrence signs

**Smoking Habit** - It is directly related to worse treatment response

**Gender** - Male patients have higher recurrence rate

**Response to initial cancer treatment -** Patients with bad medical history should have regular follow-up.

# RECOMMENDATIONS

Improve data granularity and balance

Explore ensemble or time-based models


Embed model outputs into clinical settings

# Thank you!