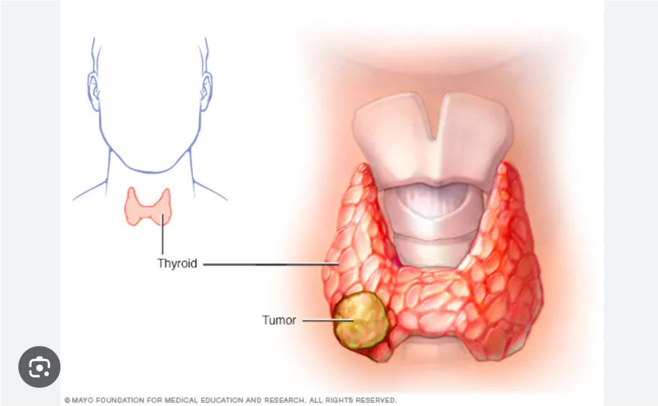
# Capstone 3 Final Presentation

Ashley Kim

# Background:

- Thyroid cancer is the most common endocrine cancer
- Symptoms are very hard to detect, therefore it is hard to identify the cancer in its early stages
- There are several types of thyroid cancer
- Prior thyroid cancer patients have a 20% risk of recurrence



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

# What this project will be doing:

- With data analysis, thyroid cancer recurrence can potentially be prevented
- Identifying common trends in clinicopathological features such as:
    - Smoking
    - Radiotherapy
    - Thyroid function
    - Physical examination
    - Adenopathy
    - Pathology
    - Focality
    - And a few more
- Identifying these trends can help in creating models that can predict recurrence in past thyroid patients

# Steps taken for this project:

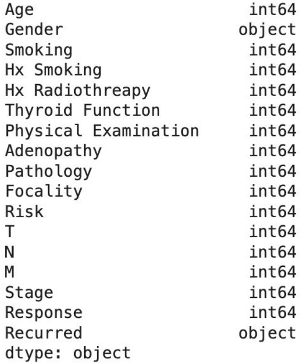
- Data wrangling
  - Identify and eliminate any missing values
  - Change any column data types
  - Replace categorical values with numerical values
- EDA
  - Observe trends using box plots, linear plot graphs, and bar graphs
  - Observe correlation between features using heat maps
- Pre-processing
  - Create a training and testing set to use in the model
- Model
  - Use Random Forest Classification and confusion matrix

# Dataset description:

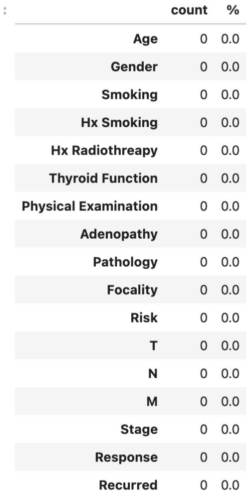
- Dataset includes the following columns:
  - Age: age of patient
  - Gender: gender of patient
  - Smoking: currently smoking
  - Hx Smoking: history of smoking
  - Hx radiotherapy: history of receiving radiotherapy
  - Risk: level of risk for recurrence
  - T: size of tumor
  - N: spread of cancer to nearby lymph nodes
  - M: metastasis (spread to other parts of body)
  - Stage: stage of cancer
  - Response: efficacy of therapy
  - Recurred: recurred thyroid cancer

# Data wrangling:

- No null values found
- Unique values per column
  - Age showed a wide variation of values
  - Other categories were all categorical
- Changed all categorical values into numerical values to use for the pre-processing stage

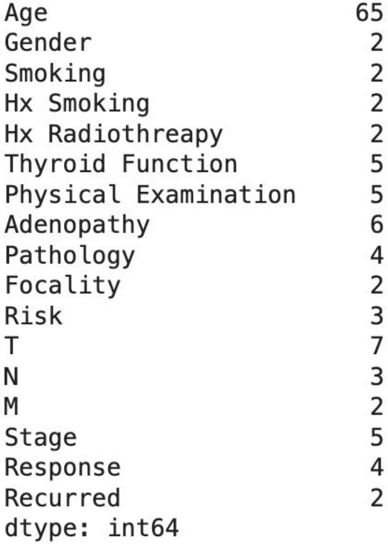


```
Age                     int64
Gender                 object
Smoking                 int64
Hx Smoking              int64
Hx Radiothreapy         int64
Thyroid Function        int64
Physical Examination    int64
Adenopathy              int64
Pathology               int64
Focality                int64
Risk                    int64
T                       int64
N                       int64
M                       int64
Stage                   int64
Response                int64
Recurred               object
dtype: object
```



| | count | % |
|---|---|---|
| Age | 0 | 0.0 |
| Gender | 0 | 0.0 |
| Smoking | 0 | 0.0 |
| Hx Smoking | 0 | 0.0 |
| Hx Radiothreapy | 0 | 0.0 |
| Thyroid Function | 0 | 0.0 |
| Physical Examination | 0 | 0.0 |
| Adenopathy | 0 | 0.0 |
| Pathology | 0 | 0.0 |
| Focality | 0 | 0.0 |
| Risk | 0 | 0.0 |
| T | 0 | 0.0 |
| N | 0 | 0.0 |
| M | 0 | 0.0 |
| Stage | 0 | 0.0 |
| Response | 0 | 0.0 |
| Recurred | 0 | 0.0 |

Null values

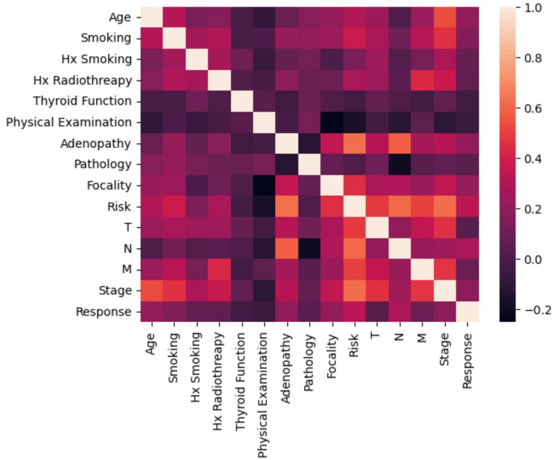


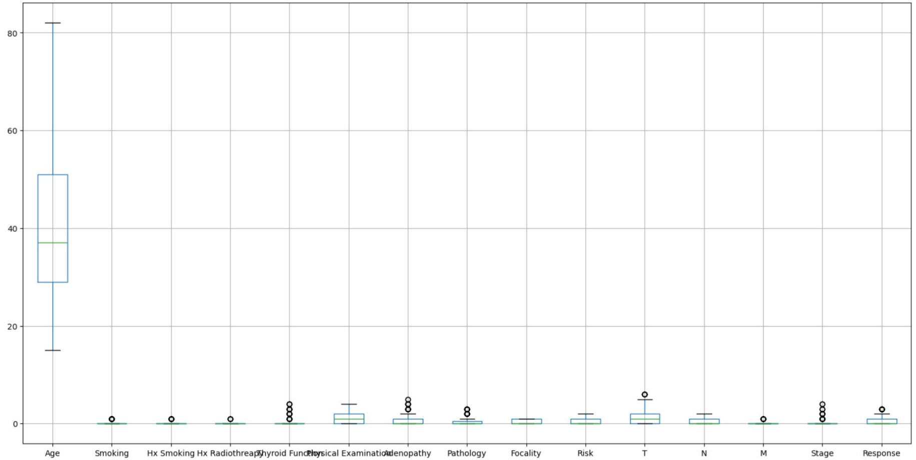
```
Age                     65
Gender                   2
Smoking                  2
Hx Smoking               2
Hx Radiothreapy          2
Thyroid Function         5
Physical Examination     5
Adenopathy               6
Pathology                4
Focality                 2
Risk                     3
T                        7
N                        3
M                        2
Stage                    5
Response                 4
Recurred                 2
dtype: int64
```

Unique values

# EDA:

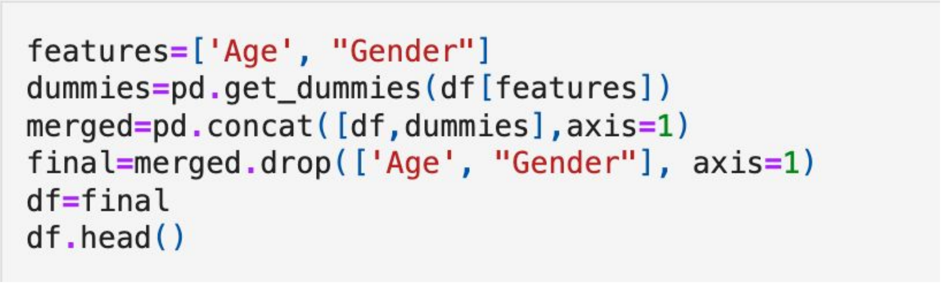
- Heatmap shows some correlation between categories
  - Stage and Age
  - Risk and Stage
  - N and Risk
- Boxplot shows the variability in age
- Other features have low variability because they are categorical data



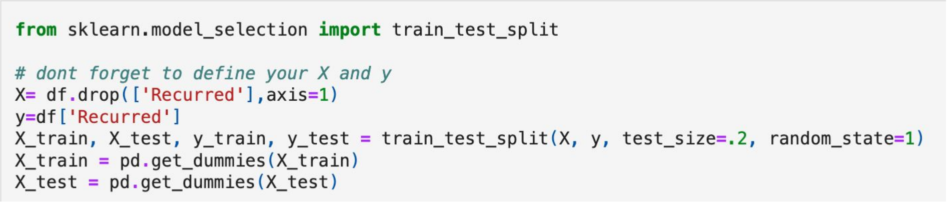


# Pre-processing data:

- Features used to create dummy variables are “Age” and “Gender”
- Did a test train split to create data to use for the model



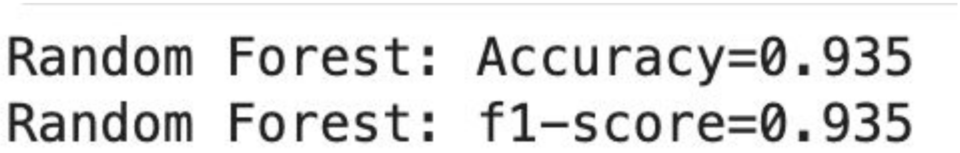
```
features=['Age', "Gender"]
dummies=pd.get_dummies(df[features])
merged=pd.concat([df,dummies],axis=1)
final=merged.drop(['Age', "Gender"], axis=1)
df=final
df.head()
```



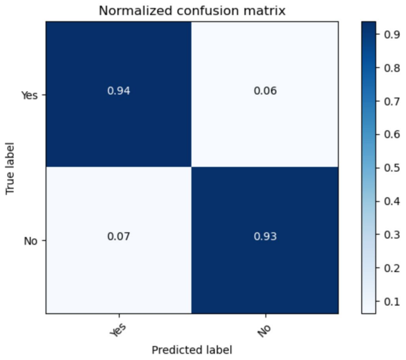
```
from sklearn.model_selection import train_test_split

# dont forget to define your X and y
X= df.drop(['Recurred'],axis=1)
y=df['Recurred']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.2, random_state=1)
X_train = pd.get_dummies(X_train)
X_test = pd.get_dummies(X_test)
```

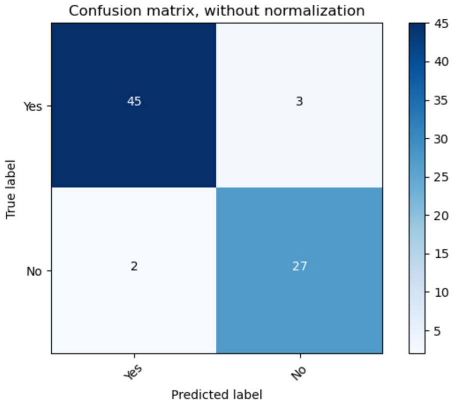
# Model:

- Used the Random Forest Classifier
    - Accuracy = 0.935
    - F1-score = 0.935
- Used the normalized confusion matrix to identify where the misclassification occurred
    - 6% misclassified for the Yes true label
    - 7% misclassified for the No true label



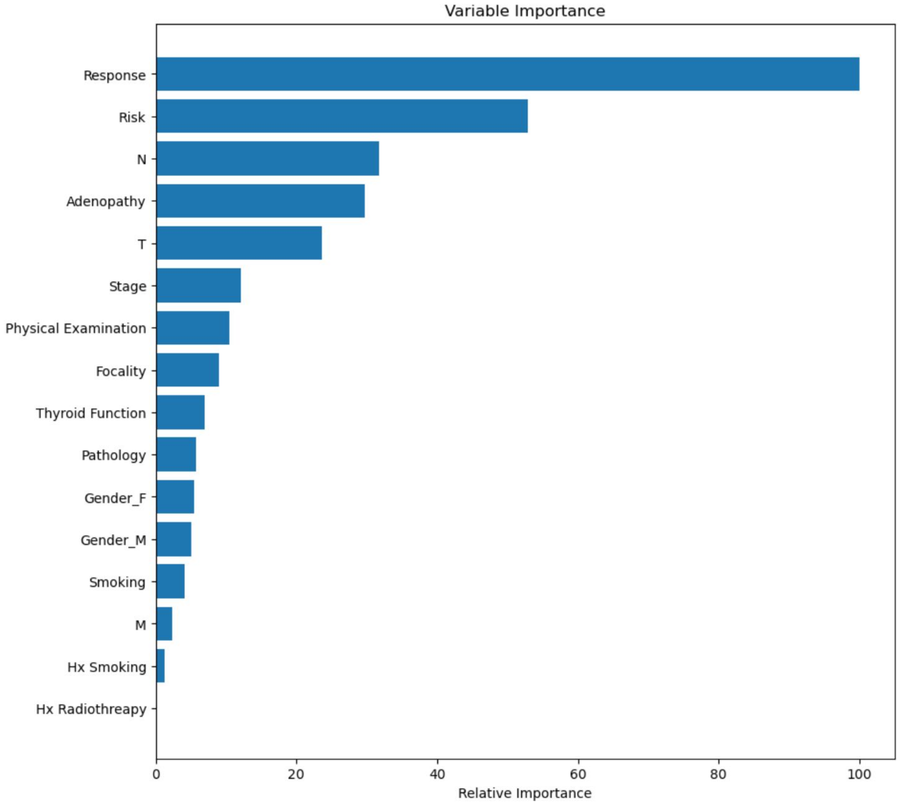
```
Random Forest: Accuracy=0.935
Random Forest: f1-score=0.935
```





# Model continued:

- Created a bar graph to identify which clinicopathological features were of most importance when determining recurrence
  - Response had almost 100% relative importance
  - Risk had almost 60% relative importance
  - Hx Smoking and Hx radiotherapy seemed to have the lowest relative importance



# Future Research:

- Test out additional models
    - Ex: linear regression model
- Research and collect additional characteristics that have a high variability in numerical value
    - The given dataset only had categorical data, which could limit the prediction ability