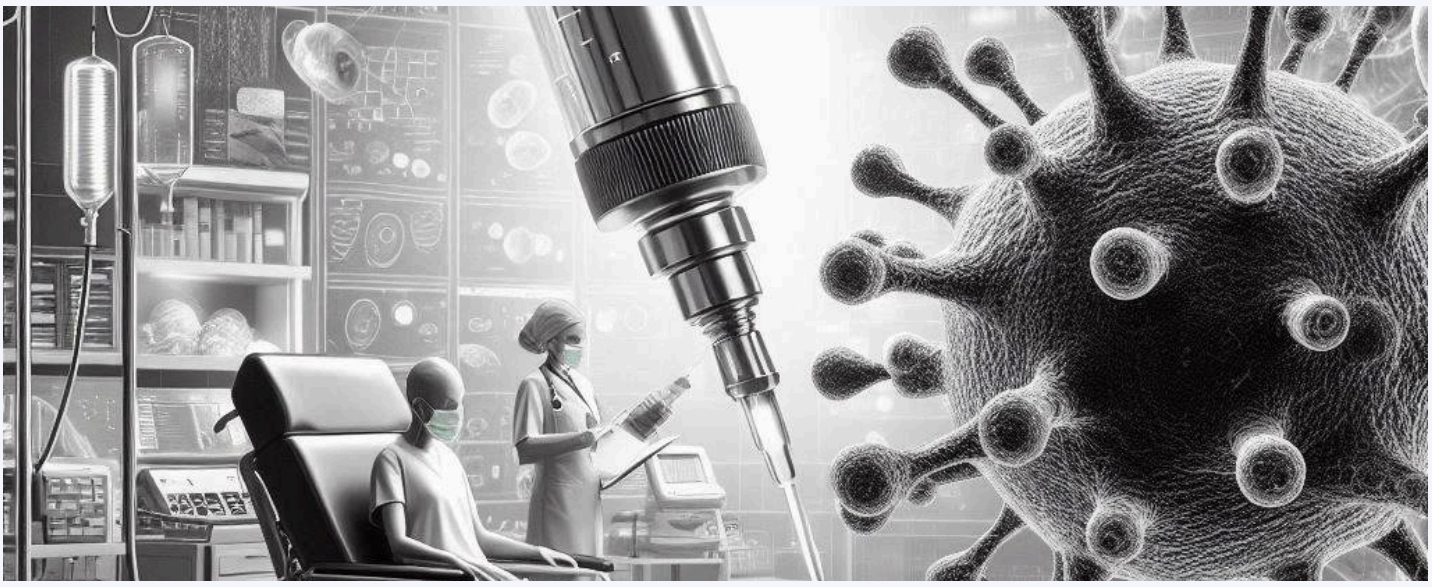
Machine Learning in Health Care: Predicting Recurrence of (thyroid) Cancer -

94.78%

MODEL ACCURACY



PREPARED BY: ISAIAH OLATERU

ISAIAH.OLATERU@100XPONENTIAL.COM

Executive Summary

☑ Business ☐ Policy ☑ Healthcare

Business Problem: Patients diagnosed with thyroid cancer are known to have a high survival rate; however, thyroid cancer is notorious for its high recurrence rate. According to the leading medical newsletter, Medical News Today (M.N.T), 1 in 5 thyroid cancer patients experiences a recurrence.
This significant rate of recurrence inspired a 15-year study of 383 patients diagnosed with thyroid cancer. **Can we predict who is likely to experience recurrence to improve health outcomes and reduce the burden on the healthcare system?** The data-set used for this project was provided by the University of California Irvine (UCI) team (cited in the Appendix).

Project Goal(s): Using October 2023 UCI patient data, build a machine learning (ML) model that can predict thyroid cancer re-occurrence with over 90% accuracy

Key Results:

Out of Three ML models explored, a **Decision Tree (DT) Classifier model (with bagging)** achieved the best predictive performance.
The top 3 factors that best predicts recurrence are i. Response to initial treatment, ii. Age, iii. Pathology_Hurtle Cell (a specific type of thyroid cancer).

Recommended Model (DT): Performance Metric

78.5%	94.7%	89.3%
RECALL	ACCURACY	ROC - AUC

This model correctly predicts 7.8 out of 10 patience that will experience a recurrence. This is based on the recall metric which is the most important metric in this use case.

Insights and Performance Metrics of three ML Models

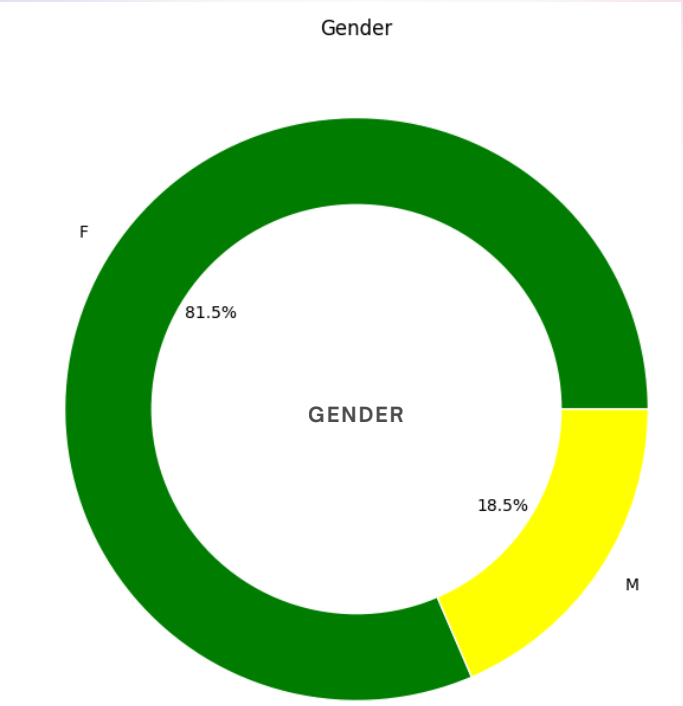
The primary goal of this project was to enhance the original model published by UCI researchers using the same data-set. The aim was to determine if a different type of classification model could yield better results. If yes, I then sought to understand whether there were differences in the features that had the most impact in prediction recurrence. In this project, we explored three different classifier models: a Decision Tree base model (with bagging), a logistic regression model, and a K-Nearest Neighbor (KNN) classifier. The original UCI research published results using the KNN and Support Vector Machine (SVM) models. The K.N.N model served as a comparative baseline. Here are the results of the classifier models I trained and their performance evaluations on a holdout set. As reported on the prior slide, the Decision Tree (DT) model performed best. Here are the results across the three models we trained.

	Decision Tree Classifier (with Bagging)	**Logistic Classifier**	**K-Nearest Neighbours**
Accuracy	0.9478	0.8956	0.8956
Precision	1.0000	0.8636	0.8333
Recall	0.7857	0.6786	0.7142
F1 - Score	0.8800	0.7600	0.7692
ROC (AUC)	0.8929	0.8220	0.8341

You can see from the table above that the **Decision Tree Classifier with Bagging** performed best across all metrics compared to the other two models we trained to predict cancer recurrence.

About the data-set. Patients Studied

- 383 Patients (n)
- 81.5% were female
- 18.5% were male
- 40.9 years - average age
- 15 years - youngest
- 82 years - oldest



	Age
count	383.000000
mean	40.866841
std	15.134494
min	15.000000
25%	29.000000
50%	37.000000

Appendices

1. Feature description of data set and citations

2. Statistical summary description of data features

3. Table of most predictive features for the recommended model

Path to my GitHub repository to view code related to this summary: **https://github.com/YNWA-Algo/T-Cancer-Reoccurrence-Predictor-Model/**blob/3a74564638e03a5079b12d49cd6d12d0000a7a49/TCancer_R_Prediction.ipynb

1

Feature description and citation

1. Age: The age of the patient at the time of diagnosis or treatment.
2. Gender: The gender of the patient (male or female).
3. Smoking: Whether the patient is a smoker or not.
4. Hx Smoking: Smoking history of the patient (e.g., whether they have ever smoked).
5. Hx Radiotherapy: History of radiotherapy treatment for any condition.
6. Thyroid Function: The status of thyroid function, possibly indicating if there are any abnormalities.
7. Physical Examination: Findings from a physical examination of the patient, which may include palpation of the thyroid gland and surrounding structures.
8. Adenopathy: Presence or absence of enlarged lymph nodes (adenopathy) in the neck region.
9. Pathology: Specific types of thyroid cancer as determined by pathology examination of biopsy samples.
10. Focality: Whether the cancer is unifocal (limited to one location) or multifocal (present in multiple locations).
11. Risk: The risk category of cancer is based on several factors, such as tumor size, extent of spread, and histological type.
12. T: Tumor classification is based on its size and extent of invasion into nearby structures.
13. N: Nodal classification indicating the involvement of lymph nodes.
14. M: Metastasis classification indicating the presence or absence of distant metastases.
15. Stage: The overall stage of the cancer, typically determined by combining T, N, and M classifications.
16. Response: Response to treatment, indicating whether the cancer responded positively, negatively, or remained stable after treatment.
17. Recurred: Indicates whether the cancer has recurred after initial treatment.

Data-set Citation: Borzooei,Shiva and Tarokhian,Aidin. (2023). Differentiated Thyroid Cancer Recurrence. UCI Machine Learning Repository. https://doi.org/10.24432/C5632J.

2 Statistical summary description of data-set

```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Age                    383 non-null    int64
 1   Gender                 383 non-null    object
 2   Smoking                383 non-null    object
 3   Hx Smoking             383 non-null    object
 4   Hx Radiothreapy        383 non-null    object
 5   Thyroid Function       383 non-null    object
 6   Physical Examination   383 non-null    object
 7   Adenopathy             383 non-null    object
 8   Pathology              383 non-null    object
 9   Focality               383 non-null    object
 10  Risk                   383 non-null    object
 11  T                      383 non-null    object
 12  N                      383 non-null    object
 13  M                      383 non-null    object
 14  Stage                  383 non-null    object
 15  Response               383 non-null    object
 16  Recurred               383 non-null    object
dtypes: int64(1), object(16)
```

The data summary above reflects the original data and data types at the start of this project as pulled from the UCI repository. Object type variables were then sorted into different types of categorical variables and transformed via encoding prior to training the model

3

Feature importance

We identified the features that contributed most to the recommended model's predictive performance. Reference the bar chart below

0	Response	0.353731
1	Age	0.007388
2	Pathology_Hurthel cell	0.004104
3	Risk	0.003806
4	Adenopathy_Left	0.003433
5	Thyroid Function_Clinical Hypothyroidism	0.003209
6	T	0.002239
7	Focality	0.001791
8	Gender	0.000970
9	Adenopathy_Posterior	0.000000
10	Smoking	0.000000
11	Hx Radiotherapy	0.000000

How well a patient responds to treatment the first time is by far the best predictor of recurrence. Effective response to treatment means much less likelihood of recurrence. The next top 3 features that predict recurrence are Age, Hurthel Cell (a type of thyroid cancer cell), the predefined risk category of the cancer. Nine factors registered some impact on the model. Smoking seemed to have no predictive impact