Thyroid Cancer Prediction

Decision Tree Classification & Patient Phenotype Discovery

Data Mining Project

Chergui Mohamed Bahae Eddine - Remmache Hibaterrhman

The Challenge

➔ Key Facts:

Thyroid cancer is one of the **fastest-growing cancers** worldwide. Where its incidence in **Algeria** has **increased significantly** between 1993 and 2013.

It accounts for **88% of all endocrine cancers**, making it the most common endocrine malignancy. Like in most countries, thyroid cancer in Algeria affects **women far more than men (94.7%)**.

Regional Studies (Oran, Tlemcen, Béjaïa, Tizi Ouzou) reported Increasing diagnosis of **papillary thyroid carcinoma** in the last **10 years.**

- Reference: Thyroid Cancer in Western Algeria: Histopathological and Epidemiological Study

The Challenge

Early and **accurate detection** is crucial for better patient outcomes and saving lives.

Resources:

213k

Patients in Dataset

15

Clinical Features

50/50

Balanced Classes

Our Solution: A Two-Part System

➔ Part 1: Supervised Learning

- **Decision Tree Classifier**

Predicts if thyroid is:
- Benign (non-cancerous)
- Malignant (cancerous)

Provides confidence score

➔ Part 2: Unsupervised Learning

- **K-Means Clustering**

Groups patients into 6 types:
- Young Healthy
- Elderly Low-Hormone
- High-Risk Nodular
- And 3 more patterns

Helps understand patient profile

How It Works: Simple 3-Step Process

1

Collect Patient Data

Age, hormones (TSH, T3, T4), nodule size, medical history

2

Model Analyzes the Data

Decision tree examines patterns to predict cancer risk

3

Get Comprehensive Results

Diagnosis + Patient type + Risk zones for each feature

The 6 Patient Types We Discovered

Each patient is automatically assigned to one of these groups

Type 0

Young Healthy

34 yrs

Small nodules, balanced hormones

Type 1

Elderly Low-Hormone

70 yrs

Reduced thyroid function

Type 2

Young High-Risk

40 yrs

Large nodules - needs attention!

Type 3

Hypothyroid Nodular

43 yrs

High TSH, large nodules

Type 4

Elderly Hyperthyroid

69 yrs

Elevated hormone levels

Type 5

Mature T4-Dominant

56 yrs

Very high T4, low T3

What Data Do We Use?

➔ Lab Results (Numerical)

- Age: Patient's age
- TSH Level: Thyroid stimulating hormone
- T3 & T4: Thyroid hormones
- Nodule Size: Size in cm

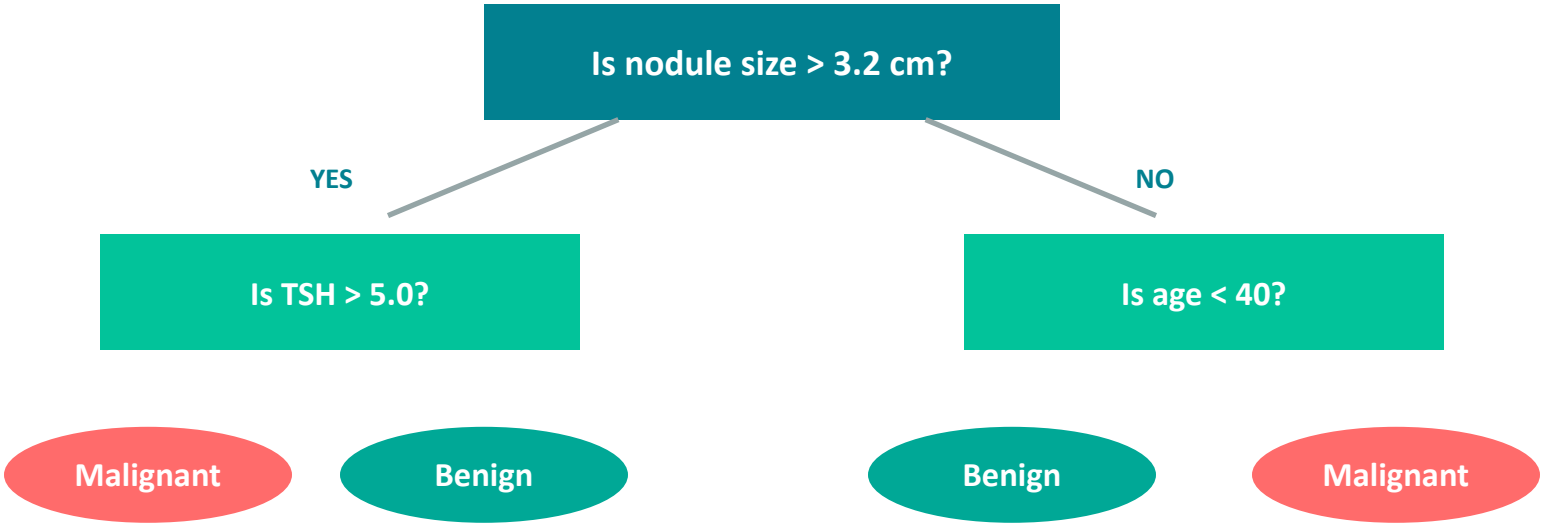
➔ Patient Information (Categorical)

- Gender: Male/Female
- Family History: Thyroid disease
- Radiation: Past exposure
- Risk Factors: Smoking, obesity

Total: 15 Clinical Features

Decision Tree: How We Make Predictions

Think of it like a flowchart that asks questions:



The tree keeps asking questions until it reaches a prediction

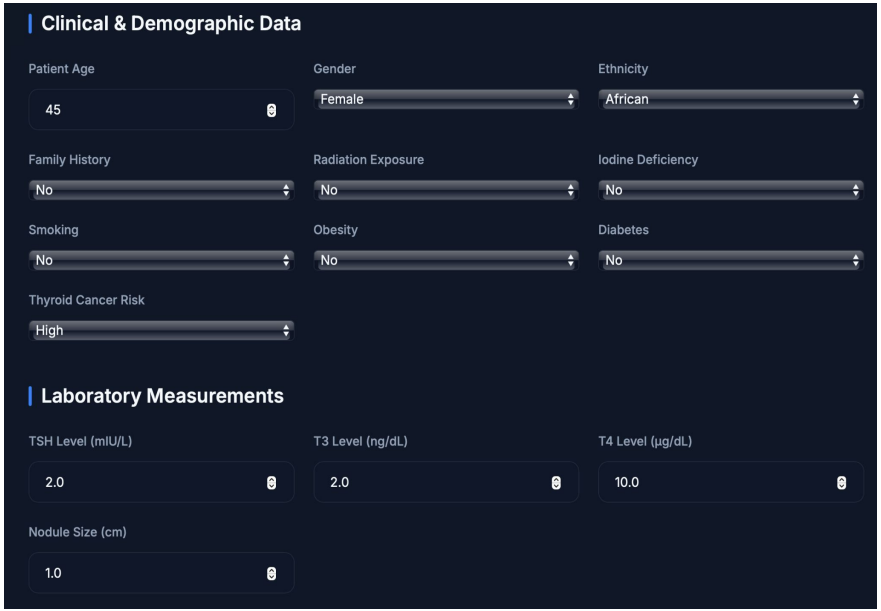
Danger Zones: Understanding Your Risk

We identify specific risk thresholds for each measurement

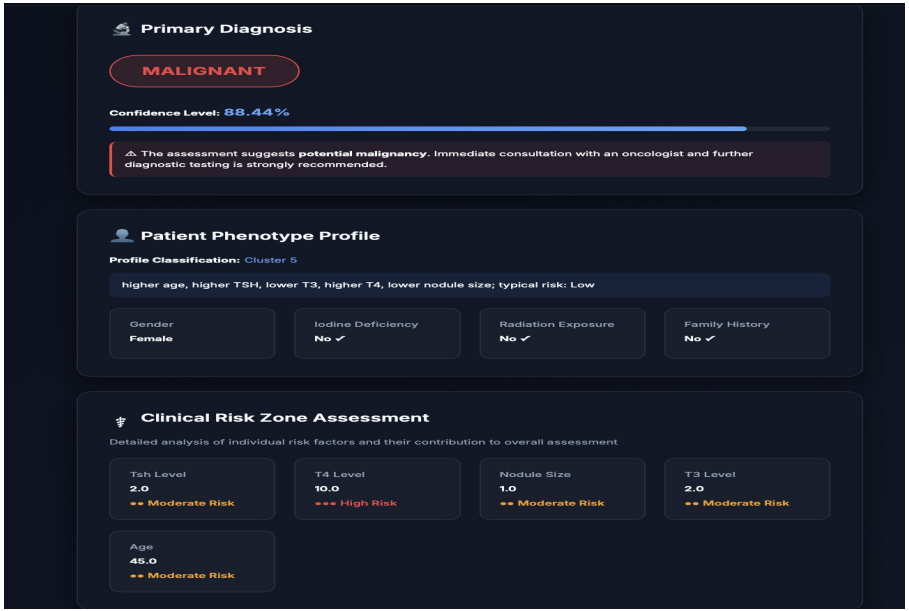
TSH Level	**Low Risk** < 0.48	**Moderate** 0.48 – 9.97	**High Risk** ≥ 9.97	*16 zones*
Nodule Size	**Low Risk** < 0.27 cm	**Moderate** 0.27 – 4.58 cm	**High Risk** ≥ 4.58 cm	*13 zones*
Age	**Low Risk** < 23 yrs	**Moderate** 23 - 86.5 yrs	**High Risk** > 86.5 yrs	*13 zones*
T4 Level	**Low Risk** < 4.51	**Moderate** 4.51 - 11.76	**High Risk** > 11.76	*19 zones*
T3 Level	**Low Risk** < 4.51	**Moderate** 4.51 - 11.76	**High Risk** > 11.76	*18 zones*

Example of Real-Time Prediction from Our Website

➔ Patient Information :



➔ Prediction Result :



What We Achieved

Interpretable Model

Decision tree with clear rules doctors can understand

Risk Thresholds

79 data-driven cutoff values across 5 key features

6 Patient Types

Discovered distinct patterns helping personalize treatment

Comprehensive System

Combines diagnosis, phenotyping, and risk assessment

Why This Matters for Healthcare

→ For Doctors

- See WHY the AI predicted
- Get specific risk markers
- Understand patient profile
- Make informed decisions

→ For Patients

- Earlier cancer detection
- Personalized treatment
- Better risk understanding
- More confidence

→ For Healthcare

- Reduce unnecessary procedures
- Improve screening efficiency
- Data-driven guidelines
- Better resource use

Technical Approach

➔ Dataset

- 98,990 patient records
- Balanced 50/50
- 15 clinical features
- 80/20 split

➔ Model

- Decision Tree (depth=7)
- StandardScaler + One-Hot
- K-Means (k=6)
- Sklearn pipeline

➔ Features

- **Numerical**: Age, TSH, T3, T4, Nodule
- **Categorical**: Gender, History, Factors
- No manual selection

➔ Validation

- Stratified train-test
- Threshold extraction
- Unsupervised discovery

Current Model Limitations

Balanced Dataset assumption

The model was trained on a balanced benign/malignant dataset which may not reflect real-world class distributions and could affect performance in practice .

Recall-Focused Tradeoff

Emphasizing high recall increases detection of malignant cases but may lead to more false positives, reducing overall precision .

Limited Model Complexity

The Decision Tree depth (7) was restricted to maintain interpretability, which may limit the model's ability to capture more complex patterns .

Next Steps & Future Improvements

Short Term

- External dataset testing
- Radiologist comparison
- User interface

Medium Term

- Add ultrasound images
- Temporal patient data
- Mobile app

Long Term

- EHR integration
- Multi-center trials
- FDA approval

Key Takeaways

- Built interpretable AI for thyroid cancer prediction
- Discovered 6 distinct patient phenotypes automatically
- Extracted 79 data-driven risk thresholds
- Created comprehensive tool: diagnosis + phenotype + risk
- Ready for validation and clinical testing

Thank You!

Questions?