Thyroid Cancer Detection - Project Report

# 1. Introduction

Thyroid cancer is a type of cancer that forms in the thyroid gland. Even after successful treatment, there is a possibility of recurrence in some patients. This project aims to build a predictive system that can determine the likelihood of thyroid cancer reoccurring based on a patient's clinical history and diagnosis details.

# 2. Problem Statement

The objective is to build a machine learning model to predict whether a thyroid cancer survivor is at risk of relapse. The system is designed to provide clinicians with insights based on multiple diagnostic features.

# 3. Dataset Overview

Key features include:

* - Age, Gender  
  - Smoking, Smoking History, Radiotherapy History  
  - Thyroid Function, Physical Exam, Adenopathy  
  - Pathology, Focality, Risk Category  
  - Tumor (T), Node (N), Metastasis (M) classifications  
  - Stage, Response to Treatment, and Recurrence (Target)

# 4. Data Preprocessing

* The dataset underwent several preprocessing steps:  
  - Missing values were handled using appropriate imputation  
  - Categorical features were encoded for model compatibility  
  - Feature scaling was applied for numeric columns  
  - The target column 'Recurred' was encoded to binary classes

# 5. Exploratory Data Analysis (EDA)

The following visualizations helped analyze feature distributions and recurrence patterns:

Figure 1: Recurrence by Risk Category

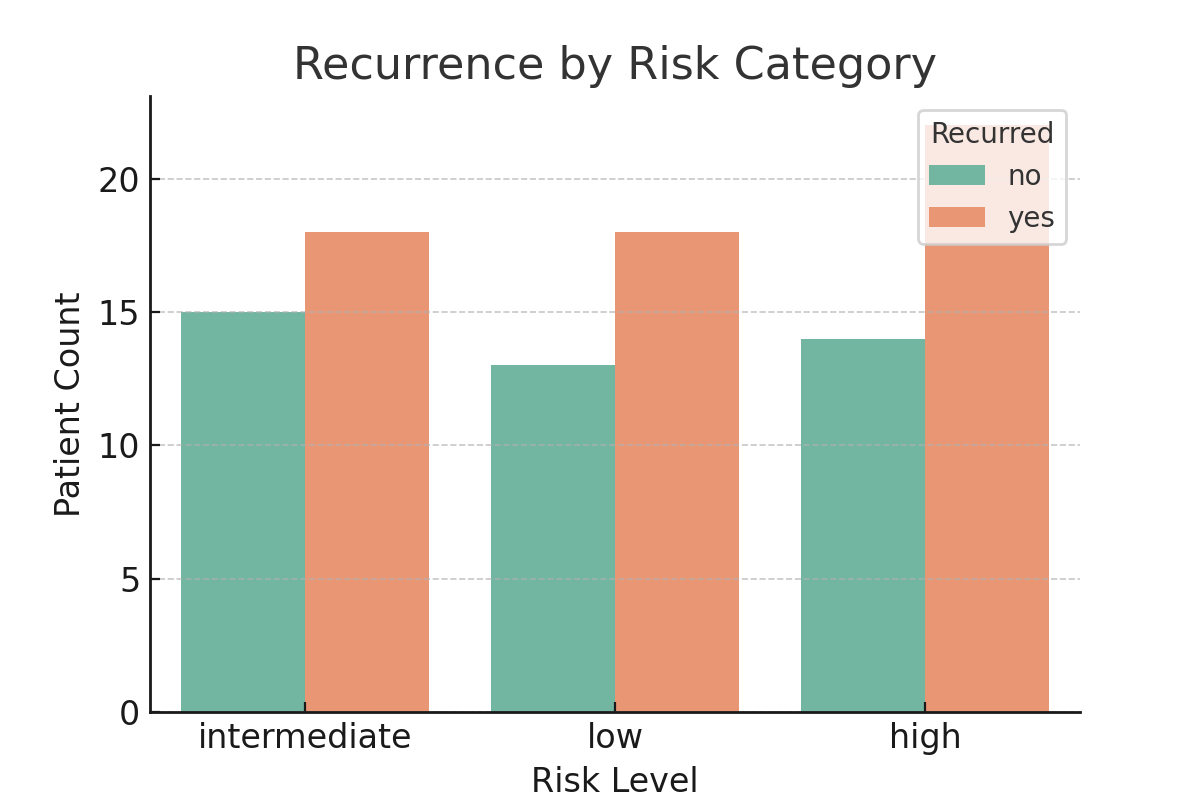
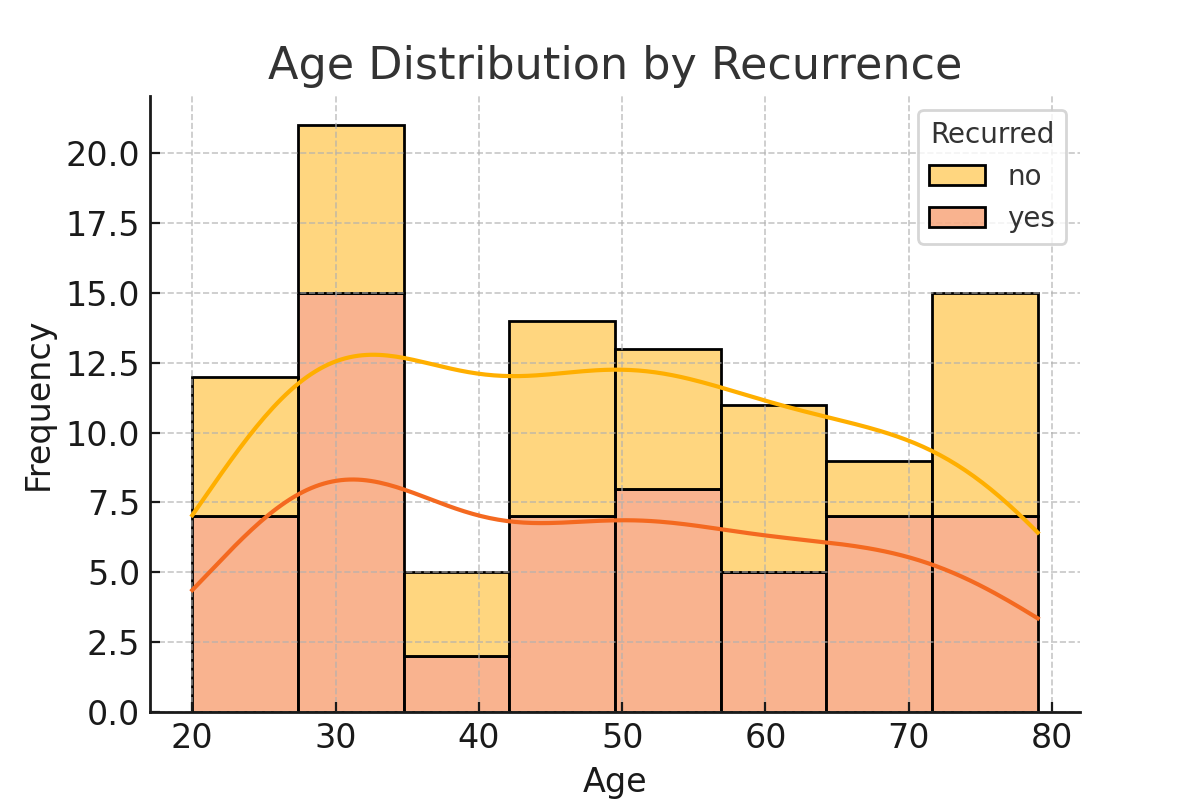


Figure 2: Age Distribution by Recurrence



# 6. Model Building and Evaluation

A Random Forest classifier was trained to predict thyroid cancer recurrence. The data was split into training and testing sets using an 80/20 ratio. The model's performance was evaluated using accuracy, precision, recall, and F1-score. Random Forest delivered strong results and was able to capture complex patterns in patient data.

# 7. Conclusion

The developed model effectively identifies thyroid cancer patients at risk of recurrence. Risk level, age, and staging features were among the most predictive. This model can support oncologists in follow-up treatment planning.

# 8. Future Work

* - Include additional longitudinal health records for better prediction  
  - Explore survival analysis for time-to-recurrence modeling  
  - Improve interpretability through SHAP or LIME methods  
  - Deploy the model as a clinical decision-support tool