Project Report

Thyroid Cancer Recurrence Prediction

1. Project Overview

This project aims to **predict whether a thyroid cancer survivor's cancer will recur** using patient clinical, pathological, and treatment history data. The final deliverable is a **trained machine learning model**, deployed as a **FastAPI service**, that accepts patient information and returns a **Yes/No** recurrence prediction with a probability score.

The **core workflow** has three main stages:

1	2	3
EDA (Exploratory Data Analysis) eda.ipynb	Model Training & Evaluation train.ipynb	Model Deployment main.py (FastAPI API)

2. Dataset Summary

Source: dataset.csv

Target variable: Recurred (Yes / No)

Features: 16 clinical & diagnostic attributes:

Demographics: Age, Gender

Lifestyle & History:

Smoking, Hx Smoking, Hx Radiotherapy

Thyroid & Clinical Status:

Thyroid Function, Physical Examination, Adenopathy

Cancer Characteristics:

Pathology, Focality, Risk, T, N, M, Stage

Treatment Feedback:

Response

3. EDA (Exploratory Data Analysis) – eda.ipynb

Purpose: To deeply understand the data, detect patterns, and guide model design.

A. Data Quality Checks

Missing Values: Identified gaps in several categorical fields (e.g., Smoking history, Radiotherapy status).
Duplicates: Checked for duplicate patient records.
Data Types: Ensured numerical features (e.g., Age) are integers, and categorical features are treated as strings.

B. Univariate Analysis

- **Categorical Features:**
  - Count plots for Gender, Smoking, Pathology, etc.

Found that most patients **do not** experience recurrence.

- **Numerical Features:**
  - Distribution of Age showed a concentration in middle-aged to older adults.
  - Potential slight right-skew in age distribution.

C. Bivariate Analysis

Age vs Recurred: Recurrence slightly higher in younger age brackets in dataset.

Risk Category vs Recurred: High-risk patients have significantly higher recurrence rates.
Adenopathy vs Recurred: Presence of adenopathy correlates strongly with recurrence.

D. Correlation & Feature Relationships

- Used heatmaps to see correlations between encoded features.
- Discovered T, N, M, and Stage are strongly correlated — relevant for model selection to avoid redundancy.

E. Class Imbalance Check

- Dataset is imbalanced: “No recurrence” cases dominate.

Implies **Recall** and **F1-score** are more important than pure Accuracy.

4. Model Training – train.ipynb

Purpose: Build, evaluate, and save a robust machine learning model.

1

A. Preprocessing Pipeline

Encoding: OneHotEncoding for categorical variables.

Scaling: StandardScaler for Age.

Missing Values: None

2

B. Model Selection & Experiments

Baseline: Logistic Regression

- **Other Models Tried:**
  - Random Forest – better accuracy, more complex.
  - SVM – less suitable for larger datasets but tested for completeness.

3

C. Model Evaluation Metrics

Used **train-test split** and **cross-validation**:

- Recall (critical for identifying actual recurrence cases)
- F1-score
- ROC-AUC for probability-based evaluation

4

D. Results

- Logistic Regression gave strong baseline with interpretability.
- Tree-based models slightly better in recall but were harder to explain.

Final choice: **Logistic Regression pipeline** (due to preference for transparency).

5

E. Model Persistence

- Best model saved as artifacts/model.joblib for deployment.

5. Deployment

Backend

FastAPI (main.py)

API Endpoint

/predict – takes JSON patient data, outputs prediction + probability.

Server

uvicorn in Docker container.

UI

Static HTML form served from static/ directory.

6. Tech Stack

Layer	Tools / Libraries
Data Handling	pandas, numpy
Visualization	matplotlib, seaborn, plotly
EDA Automation	ydata-profiling
Machine Learning	scikit-learn
Explainability	shap
Model Serving	FastAPI, Pydantic
Persistence	joblib
Deployment	Docker, uvicorn
Experiment Tracking	mlflow

7. Key Insights from EDA & Training

- High-risk, adenopathy-positive, and certain pathological types are strong recurrence indicators.
- Younger patients showed slightly higher recurrence probability in this dataset.
- Logistic Regression with preprocessing pipeline provided both solid predictive power and interpretability.
- Class imbalance required paying attention to recall rather than raw accuracy.