Thyroid Cancer Reoccurrence Prediction Using Machine Learning

Shravani Bande

1 Introduction

Thyroid cancer has seen a rise in diagnosis over the past few decades, particularly among younger women. Although it generally has a favorable prognosis, recurrence can occur years after treatment. Detecting recurrence early through predictive modeling can significantly improve follow-up strategies and outcomes.

The goal of this project is to build a robust machine learning model that predicts the probability of thyroid cancer recurrence using structured clinical data. This involves preprocessing the dataset, selecting suitable algorithms, evaluating performance, and deploying a lightweight interface for predictions.

2 Dataset Description

The dataset used contains 383 records and 17 columns. Each row represents a patient with a history of thyroid cancer. The target variable is **Recurred**, which indicates whether the cancer returned post-treatment.

Features

- **Age, Gender** – Demographic details.
- **Smoking, Hx Smoking, Hx Radiotherapy** – Patient medical history.
- **Thyroid Function, Physical Examination, Adenopathy** – Clinical test outcomes.
- **Pathology, Focality, Risk, T, N, M, Stage, Response** – Cancer classification and treatment details.

3 Data Preprocessing

Before training the model, several preprocessing steps were performed:

- **Missing Values:** No missing values were found in the dataset.
- **Categorical Encoding:** All non-numeric columns were label-encoded to convert text into machine-readable integers.
- **Train-Test Split:** The dataset was split 80:20 using stratified sampling to maintain class distribution across both sets.

4 Model Training and Evaluation

Several models were considered for this classification task. Two were selected and tested:

- **Logistic Regression** — a linear model used as a baseline.
- **Random Forest Classifier** — an ensemble tree-based model known for robustness.

Evaluation Metrics

The models were evaluated using:

- **Accuracy** – overall correctness
- **Precision** – correctness of positive predictions
- **Recall** – ability to identify true recurrences
- **F1-score** – balance between precision and recall

5 Model Performance

The Random Forest model outperformed Logistic Regression in all key areas:

- **Accuracy:** 94.8%
- **Precision:** 95%
- **F1-score:** 90%

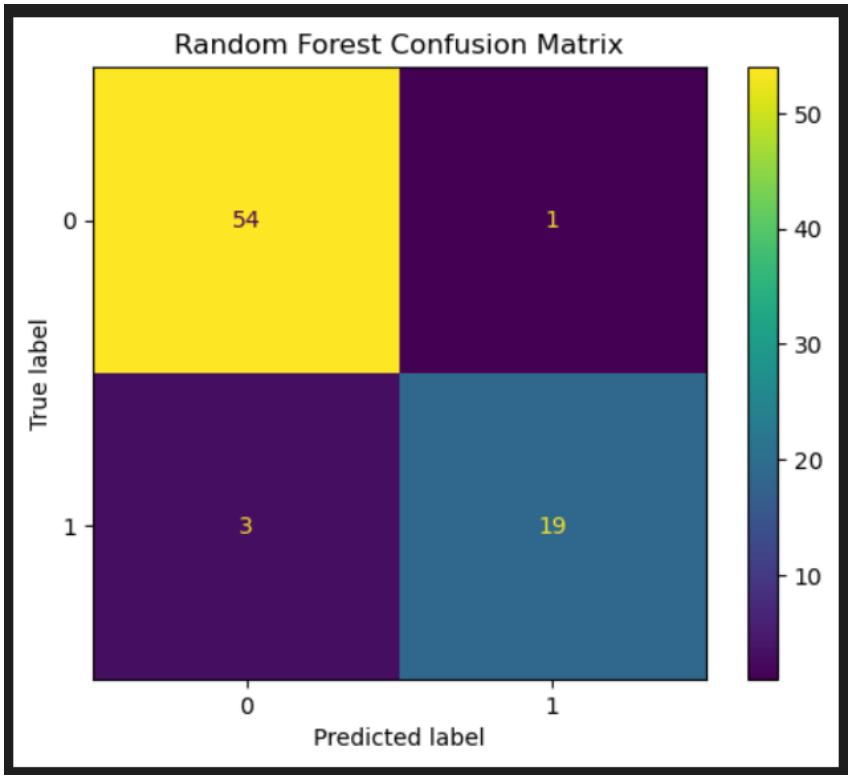

Figure 1: Confusion Matrix of Random Forest Model

6 CLI Prediction Interface

To make the model practical, a Command Line Interface (CLI) tool was developed using Python. It allows users to:

- Manually input patient details
- Randomly select real data samples
- Receive prediction results instantly

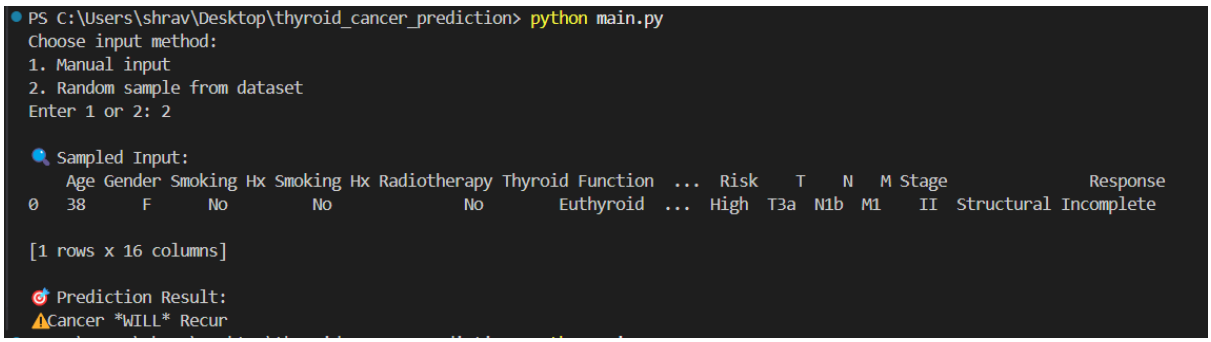

Figure 2: CLI Output – Random Sample Prediction

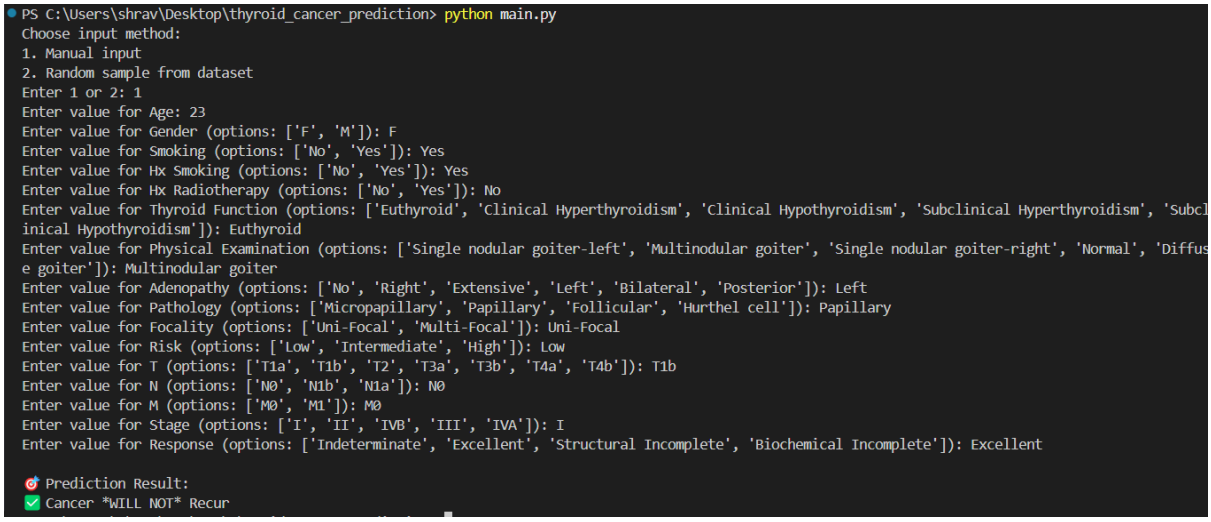

Figure 3: CLI Output – Manual Input Prediction

This interface makes the model usable without needing a complex frontend or server.

7 Conclusion

The Thyroid Cancer Reoccurrence Prediction project successfully demonstrates how machine learning can assist in medical prognosis. With a high-performing Random Forest model and an intuitive CLI, the system provides a ready-to-use solution for analyzing thyroid cancer relapse risk.

In the future, the model can be extended to include time-based features, longitudinal patient data, and even integrated into a web dashboard using Streamlit or Flask.

8 References

- Dataset: Provided internally
- Libraries: pandas, scikit-learn, seaborn, matplotlib, joblib
- GitHub Repo: `https://github.com/savi-08/thyroid-cancer-prediction`
- Scikit-learn Docs: `https://scikit-learn.org/stable/index.html`