

Thyroid Cancer Risk Prediction Using Machine Learning

Capstone Project Report

Author: S. Jenifer

Introduction

Thyroid cancer is increasing globally. Early detection of thyroid cancer risk is crucial. With the growing availability of medical datasets, machine learning provides a reliable approach to automate diagnosis and support clinical decision-making. This project focuses on building a **Thyroid Cancer Risk Prediction** System using a structured data science and machine learning pipeline. The goal is to identify the most important diagnostic factors, build a predictive model, optimize it, evaluate its performance, and finally deploy the model for real-world use.

Problem Statement

A requirement from the Hospital Management asked us to create a predictive model which will predict thyroid nodules as benign or malignant.

This project aims to build a model that predicts whether a patient is Benign (0) or Malignant (1) using clinical features such as TSH, T3, T4, Nodule Size, and Cancer Risk score.

1.Objective of the Project

- ✓ To analyze the Thyroid Cancer Risk Dataset
 - ✓ To preprocess the data and conduct EDA
 - ✓ To Apply feature selection
 - ✓ To train multiple classifiers and compare performance
 - ✓ To optimize the final model using Grid Search
 - ✓ To deploy the best model using Pickle
-

2.Dataset Description

Dataset Name: Thyroid Cancer Risk Dataset

Total Rows: 2,12,691

Columns include:

- Age
- Gender
- TSH, T3, T4 levels
- Nodule Size
- Family history
- Cancer risk levels
- Diagnosis (Benign/Malignant)

```
dataset.columns
```

```
Index(['Age', 'Gender', 'Country', 'Family_History', 'Radiation_Exposure',  
      'Iodine_Deficiency', 'Smoking', 'Obesity', 'Diabetes', 'TSH_Level',  
      'T3_Level', 'T4_Level', 'Nodule_Size', 'Thyroid_Cancer_Risk',  
      'Diagnosis'],  
      dtype='object')
```

3.Data Preprocessing

The dataset was reviewed and transformed to prepare it for machine learning. The following steps were completed:

✓ Verified Dataset Completeness

The dataset had no missing values, so no filling or imputation techniques were required.

✓ Checked for Outliers

All numerical columns were analyzed, and no major outliers were detected. The distribution was acceptable, so no outlier removal was applied.

✓ Label Encoding Applied

Some categorical features (for example: Thyroid_cancer_risk) was converted into numerical values using Label Encoding.

This step helps machine learning models understand categorical information.

✓ One-Hot Encoding Applied

Categorical variables such as Country (Gender,Family history,etc.,) were transformed using One-Hot Encoding to avoid ordinal relationships.

✓ Scaled Numerical Features

Numerical features such as:

- T3_Level
- TSH_Level
- T4_Level
- Nodule_Size
- Thyroid_Cancer_Risk

were scaled using StandardScaler to ensure uniform value ranges.

✓ Encoded Target Column

The “Diagnosis” column was converted into numeric form:

- Benign → 0
- Malignant → 1

✓ Final Dataset Prepared

After encoding and scaling, the cleaned dataset was ready for modeling.

```
dataset.columns
```

```
Index(['Age', 'TSH_Level', 'T3_Level', 'T4_Level', 'Nodule_Size',  
      'Thyroid_Cancer_Risk', 'Gender_Male', 'Country_China',  
      'Country_Germany', 'Country_India', 'Country_Japan', 'Country_Nigeria',  
      'Country_Russia', 'Country_South Korea', 'Country_UK', 'Country_USA',  
      'Family_History_Yes', 'Radiation_Exposure_Yes', 'Iodine_Deficiency_Yes',  
      'Smoking_Yes', 'Obesity_Yes', 'Diabetes_Yes', 'Diagnosis_Malignant'],  
      dtype='object')
```

Train-Test Split

To evaluate model performance, the dataset was split into:

- Training Set: 75%
 - Testing Set: 25%
-

4.Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution of the clinical features, identify patterns in hormone levels, detect imbalance in diagnosis classes, and study relationships between predictors.

4.1 Data Overview

- ❖ Verified dataset structure, row count, and column types
- ❖ Confirmed there were no missing values
- ❖ Ensured data types (numerical/categorical) matched expected medical features

Observation: The dataset was clean with consistent formats for all numerical hormone features.

4.2 Univariate Analysis

Univariate analysis was used to study each feature individually.

Numerical Features (TSH, T3, T4, Nodule_Size)

Techniques used:

- Histograms

Insights:

- Hormone levels showed natural variability across patients.
- Nodule size evenly spread across patients.

Categorical Features

- Diagnosis (Benign vs Malignant)
- Thyroid_Cancer_Risk (Low, Medium, High)

Insights:

- Diagnosis was slightly imbalanced, with more benign cases.
 - Higher risk categories had more malignant outcomes.
-

4.3. Bivariate Analysis

Bivariate analysis compared two features to identify relationships.

- No Significant Difference in T4 Level Between Benign and Malignant Cases
 - No Significant Relationship Between TSH Level and Diagnosis
 - Correlation Between Numeric Features is Very Low
 - Nodule Size Shows No Linear Relationship With Age
 - The Average hormone levels for both Malignant and benign are nearly same.
-

4.4. Outlier Detection

Outliers were inspected using:

- IQR method

Finding:

No severe outliers requiring removal. Hormone-level variations were medically expected.

5.Feature Selection

Feature selection was performed to identify the most important predictors contributing to thyroid cancer classification. This step helps improve model performance, reduce overfitting, and simplify the final model.

5.1 Method 1: SelectKBest (Statistical Feature Selection)

The **SelectKBest** technique was used to evaluate each numerical feature using statistical scoring functions.

- Features were ranked based on their relevance to the target variable (Diagnosis).
- The best top 6 and top 5 features were tested.

Insights

- With 6 features, Random Forest and Decision Tree achieved the highest accuracy (0.97).
 - With 5 features, performance remained strong, confirming these features are highly informative.
-

5.2 Method 2: Feature Importance

Used four tree-based models:

- Random Forest
- Extra Trees
- Gradient Boosting
- Decision Tree

Top 5 features consistently identified as top contributors across both methods.

Thyroid_Cancer_Risk , TSH_Level, T4_Level, Nodule_Size, T3_Level

5.3 Final Selected Features

After comparing both statistical and model-based feature selection methods, the final set of **5 most impactful features** was chosen:

- Thyroid Cancer Risk
- TSH Level
- T4 Level
- Nodule Size
- T3 Level

These features showed the strongest relationship with the target and delivered the best model performance.

5.4 Conclusion

Feature selection confirmed that thyroid hormone levels and nodule characteristics play a crucial role in predicting malignant outcomes.

By reducing the dataset to the top features, the model became:

- ✓ Faster
- ✓ Less complex
- ✓ More generalizable
- ✓ More interpretable

These optimized features used for final model training and hyperparameter tuning.

6. Model Building

This stage involved training machine learning models using the selected top 5 features to predict whether a thyroid nodule is Benign or Malignant. The primary goal was to build an accurate, reliable, and generalizable classification model.

6.1 Train-Test Split

The cleaned dataset was split into:

- Training Data: **75%**
- Testing Data: **25%**

This helps evaluate how well the model performs on unseen data and reduces the risk of overfitting.

6.2 Model Used: Random Forest Classifier

The Random Forest Classifier was chosen based on consistent accuracy and interpretability. The Model achieved:

- Accuracy: 0.8249
- AUC Score: 0.694

Although performance was decent, Grid Search was performed to further optimize the model.

6.3 Model Optimization Using Grid Search

A GridSearchCV pipeline was used to tune key Random Forest hyperparameters such as:

- Number of trees (n_estimators)
- Tree depth (max_depth)
- Minimum samples at leaf (min_samples_leaf)

- Splitting criteria (gini/entropy)
 - ✓ Accuracy: 0.8271
 - ✓ AUC Score: 0.6985
-

7. Model Evaluation

After training and optimizing the Random Forest model, the final step involved evaluating its performance using the test dataset. Multiple evaluation metrics were used to ensure the model is accurate, reliable, and performs well in distinguishing between Benign and Malignant thyroid cases.

Baseline vs Tuned Model Comparison

Model Version	Accuracy	AUC Score
Baseline Random Forest	0.82	0.69
Optimized Random Forest (Grid Search)	0.8271	0.6985

Best Parameters:

```
{ 'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 100 }
```

The tuned model is more stable, less prone to overfitting, and provides better malignant-case detection, which is crucial for medical prediction.

8. Deployment

The final model was saved as a .pkl file.

Final Model: Random Forest Classifier

Accuracy -> 0.8271

Python script created for user-driven predictions

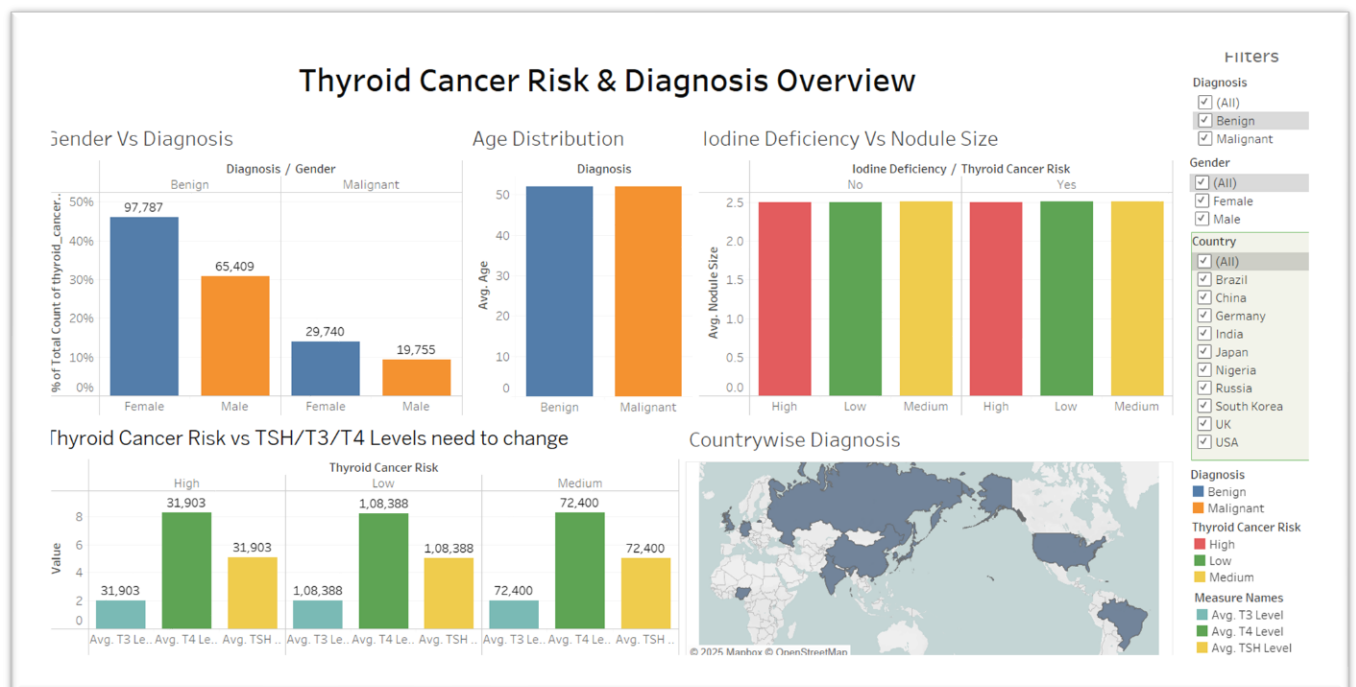
Output returns **Benign** or **Malignant** diagnosis

9. Tableau Dashboard

An interactive Tableau dashboard was built to visually explore the thyroid cancer risk dataset. The dashboard includes:

- **Diagnosis Distribution:** Benign vs. Malignant comparison
- **Gender-wise & Age-wise Analysis**
- **Hormone Level Insights:** TSH, T3, T4 levels across risk categories
- **Iodine Deficiency vs. Nodule Size Relationship**
- **Country-wise Diagnosis Map**
- **Interactive Filters:** Gender, Country, Thyroid Cancer Risk, Diagnosis

This dashboard provides a clear, intuitive visual summary of patient characteristics and helps understand the key patterns in the dataset.



10.Conclusion

The analysis revealed that although hormone indicators such as TSH and T4 did not show statistically significant differences between benign and malignant groups, key predictors—including **Thyroid Cancer Risk score, Nodule Size, T3 Level, and T4 Level**—played a crucial role in classification. The absence of strong linear correlations among features further supported the suitability of non-linear machine learning algorithms, justifying the final selection of the Random Forest model.

This capstone project successfully developed an end-to-end machine learning solution for predicting thyroid cancer risk using a structured clinical dataset. Through detailed preprocessing, EDA, bivariate analysis, feature selection, model comparison, and hyperparameter tuning, the **Random Forest classifier** emerged as the best-performing model. With an accuracy of **82.71%**, the model demonstrates a reliable ability to distinguish between benign and malignant thyroid conditions.
