Thyroid Cancer Risk Prediction Project Report

Project Objective:

To build a machine learning model that predicts the risk of thyroid cancer based on various health and demographic features. The goal is to support early screening and risk assessment.

---

1. Data Understanding & Preprocessing

Dataset:

- **Dataset Used:** thyroid_cancer_risk_data.csv
- **Rows & Columns:** ~ 212691 rows × 15 columns
- **Features:**
  - Age
  - Gender
  - Country
  - Ethnicity
  - Family_History
  - Radiation_Exposure
  - Iodine_Deficiency
  - Smoking
  - Obesity
  - Diabetes
  - TSH_Level
  - T3_Level
  - T4_Level
  - Nodule_Size
- **Target Variable:** Diagnosis

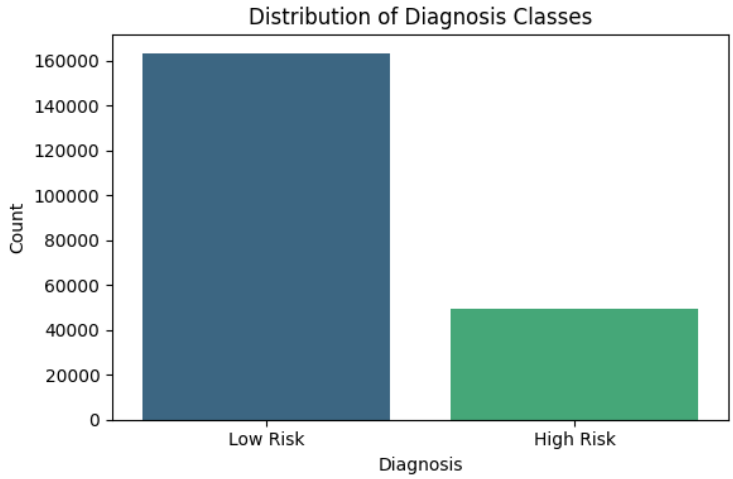
Preprocessing Steps:

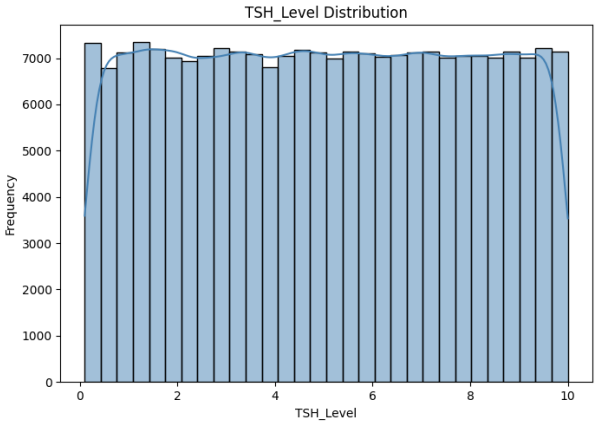
- Removed ID and target leakage columns
- Handled missing and duplicate values
- Encoded:
  - Binary categorical features using LabelEncoder
  - Multiclass categorical features like Country, Ethnicity using OneHotEncoder

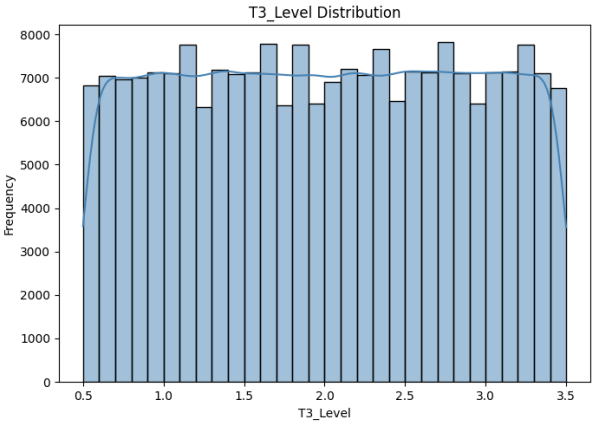
---

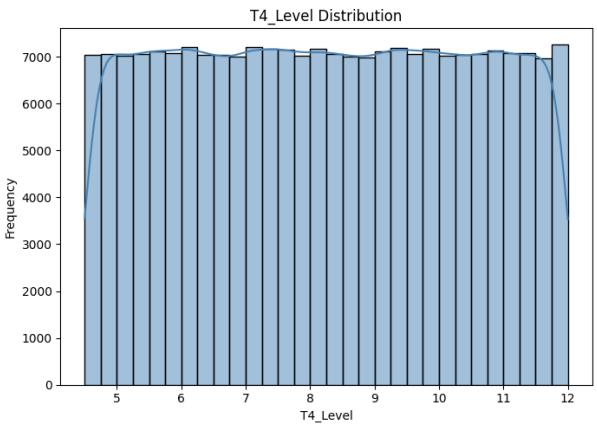
2. Exploratory Data Analysis (EDA)

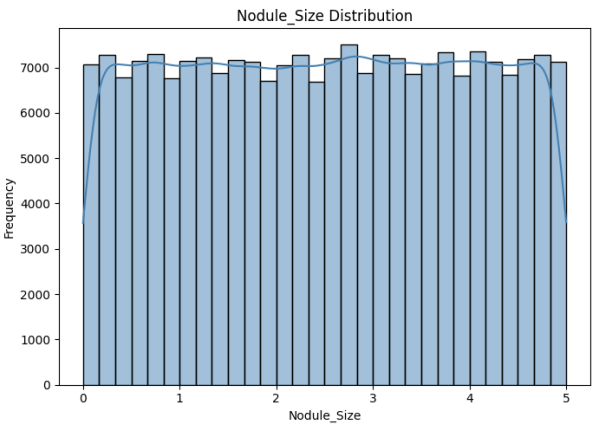
- Value counts and distributions were inspected for each feature
- Highlighted imbalance in target labels
- Diagnosis class was found to be skewed, prompting sampling strategies

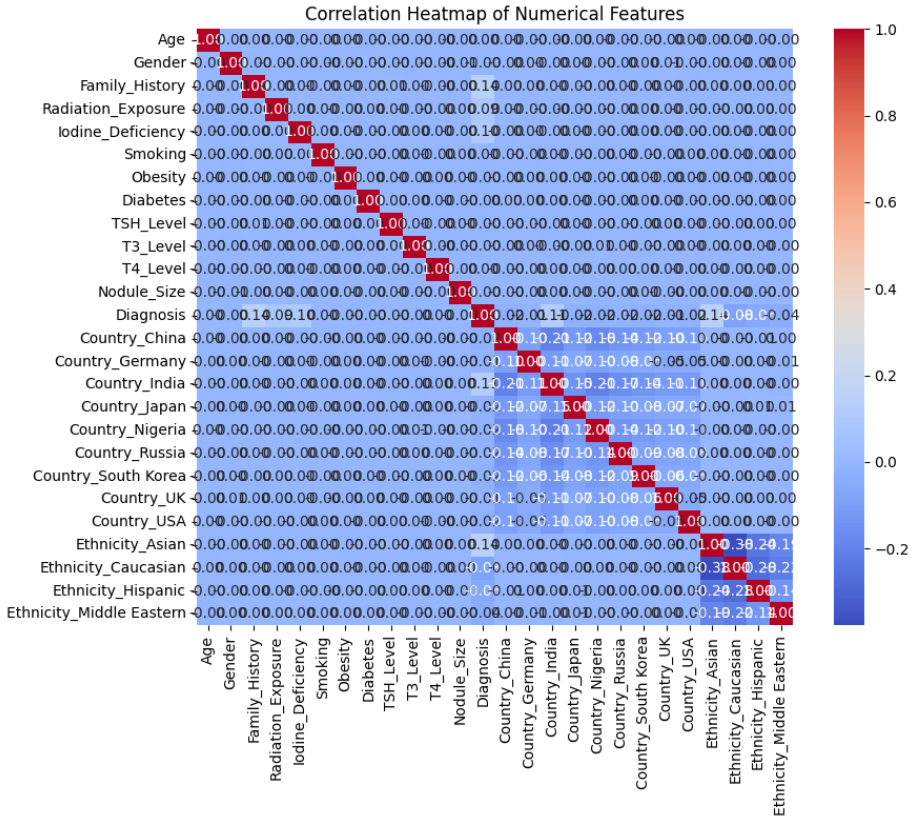












---

3. Handling Class Imbalance

- Applied:
  - **Undersampling**: To balance the dataset by reducing majority class
  - **Oversampling (SMOTE)**: To synthetically increase minority class samples

---

4. Feature Scaling

- Applied MinMaxScaler to normalize all numerical features to the [0, 1] range.
- Essential for SVM, KNN models to perform fairly across features.

---

5. Model Training & Evaluation

Algorithms Used:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

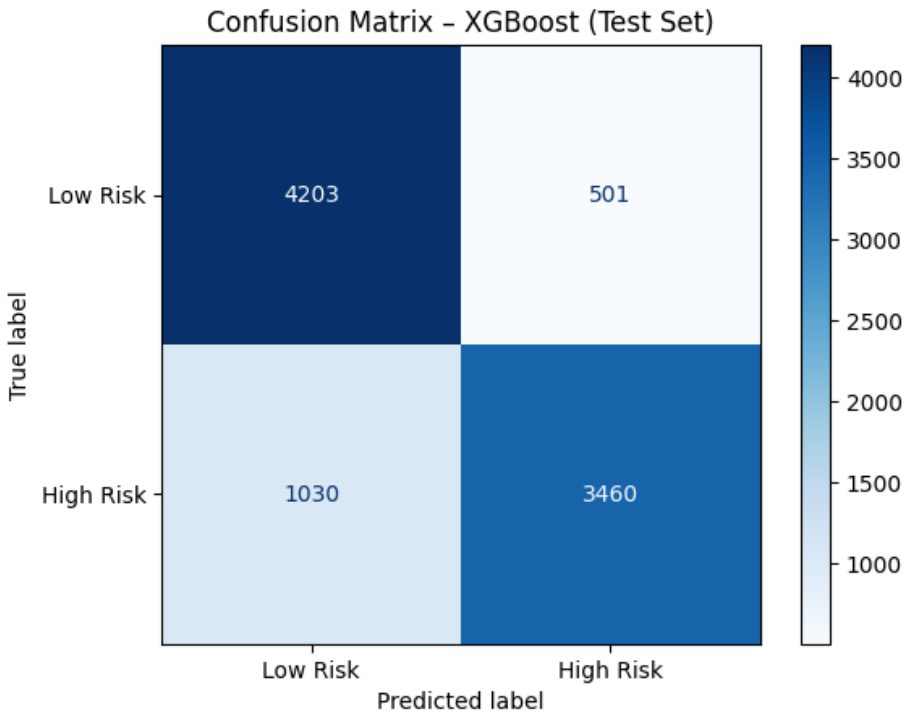
Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Findings:

- XGBoost and Random Forest consistently showed better performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	76.2%	80%	68%	73.6%
Decision Tree	78.1%	83%	70%	76.0%
Random Forest	80.0%	88%	69%	77.4%
XGBoost	**80.4%**	**91%**	**67%**	**77.0%**



6. Final Model

Model Selected: XGBoostClassifier

- Trained on balanced data (after applying under/oversampling)
- Parameters used:
    - n_estimators=200
    - learning_rate=0.1
    - max_depth=10
    - gamma=1
    - subsample=0.8
    - colsample_bytree=0.8
    - use_label_encoder=False
    - eval_metric='logloss'
- Delivered strong performance on both train and test datasets
- Chosen for its robustness, scalability, and interpretability

---

7. Model Deployment

Final Model:

- XGBoostClassifier with customized parameters
- Model and preprocessors saved together using pickle as model_and_encoders.pkl
- Includes:
    - The trained model
    - LabelEncoders for binary features
    - OneHotEncoders for multiclass features
    - MinMaxScaler for feature scaling

Streamlit App (thyroid_app.py**):**

- Interactive web interface for thyroid cancer risk prediction
- Inputs taken via sliders and dropdowns:
    - Age, Gender, Country, Ethnicity, Family History, Radiation Exposure, Iodine Deficiency, Smoking, Obesity, Diabetes, TSH, T3, T4, Nodule Size

- Uses pre-loaded encoders and scaler for consistent preprocessing
- Predicts: High Risk or Low Risk of Thyroid Cancer
