Thyroid Cancer Risk Detection

Toby Trotta

Updated: February 3, 2025

Foreword

The following dataset was retrieved from Kaggle.com (https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset) on February 3, 2025. In an attempt to remain keen on my machine learning and statistical analysis skills, I plan to continue this series of miscellaneous projects. In this project, we aim to create a machine learning model to predict the diagnosis of a type of thyroid cancer. According to the author of this dataset, the 212,691 records were scraped from the web using an unknown programming language and method, citing 12 unlisted websites as its source. The goal of these independent projects is to showcase my skills in data analysis and machine learning. Enjoy!

Toby

Introduction to Dataset

As mentioned, this dataset was retrieved from Kaggle.com. It contains 212,691 records of simulated real-world thyroid cancer risk factors:

- Gender
- Ethnicity
- Family History
- Radiation Exposure
- Iodine Deficiency
- Smoking
- Obesity
- Diabetes
- TSH, T3, and T4 Levels
- and Nodule Size

Other information included is the patient's Age, Country, Thyroid Cancer Risk, and, of course, the Diagnosis. We will use the next section to investigate the demographic makeup of this dataset.

Demographics:

Country	Count
Russia	21297
Germany	10557

Country	Count
Nigeria	31918
India	42496
UK	10642
South Korea	14965
Brazil	21413
China	31978
Japan	16867
USA	10558

Gender	Diagnosis	Count	Average Age
Male	Benign	65409	51.90292
Male	Malignant	19755	51.95282
Female	Benign	97787	51.92663
Female	Malignant	29740	51.90323