Predicting Whether a Thyroid Nodule is Cancerous using Machine Learning Algorithms

Group 8
Chandralekha Venkatesh Perumal
Ruchika Jha
Sneha Krishnan Akavalapil
Vanisri Kirubakaran

What Exactly happens ?

Thyroid cancer is the growth of abnormal cells in the thyroid glands located at the base of the necks.

“It is currently the sixth most frequent cancer in women and the thirteenth most frequent overall in all populations worldwide.”

— Cari M. Kitahara and Arthur B. Schneider, “Epidemiology of thyroid cancer,” Cancer Epidemiology, Biomarkers & Prevention, vol. 31, no. 7, pp. 1284–1297, 2022.

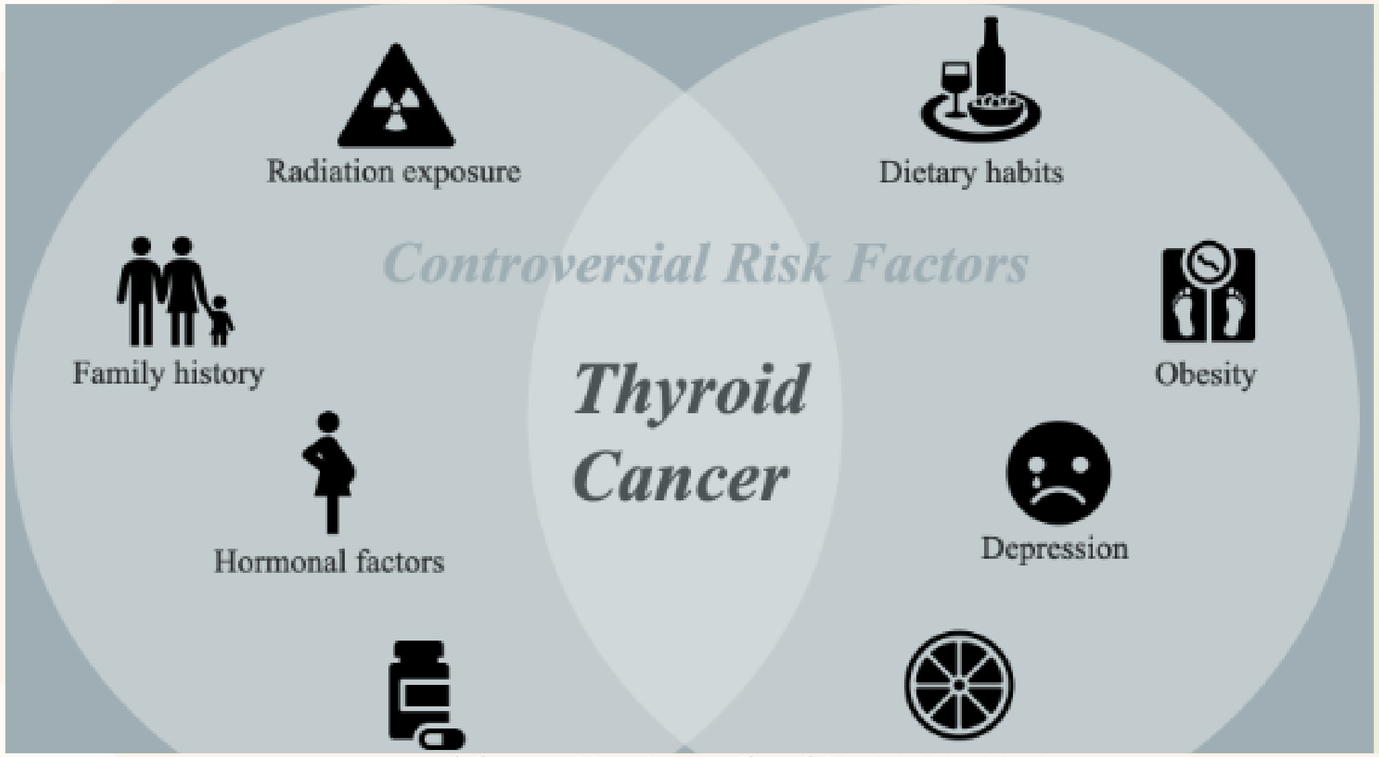


Image ref: https://arxiv.org/html/2401.03722v1

Research Question

How well can machine learning models predict if a thyroid nodule is cancerous or not, based on patient data ?

"Thyroid nodules are common, but only 5–15% of them are malignant. Accurate classification is critical to avoid unnecessary biopsies or surgeries."

— Haugen BR, et al. (2016). 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid, 26(1), 1–133.

Dataset

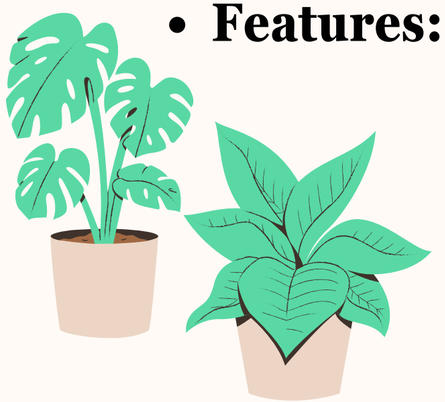
- **URL:**https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset

- **Description:** The dataset represents real-world thyroid cancer risk factors such as Age, Thyroid Levels, Radiation Exposure, Iodine Deficiency, Family History and so on.

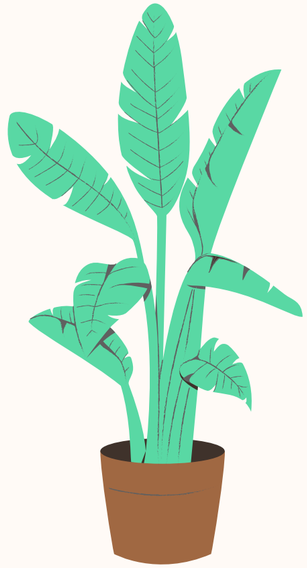
- **No. of records:** 212,691 rows and 23 attributes.

- **Target Variable:** Diagnosis (Benign and Malignant)

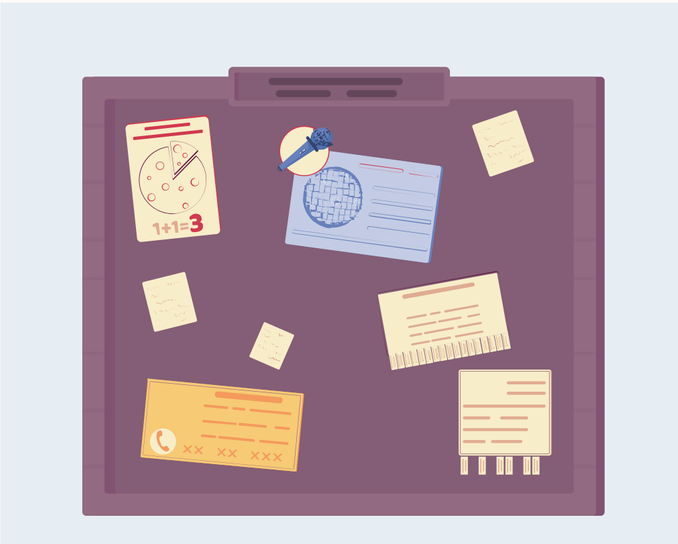
- **Features:** Age, Gender, Country, Thyroid Test Levels.



Source : Kaggle



Literature Review



Literature Review

Ranking	Authors	Paper Name	Summary
Q1	Carolyn Dacey Seib and Julie Ann Sosa	Evolving understanding of the epidemiology of thyroid cancer	The paper analyzes the evolution of understanding using data-driven methods, identifying key factors and proposing strategies to enhance comprehensive decision-making.
Q1	Cari M. Kitahara and Arthur B. Schneider	Epidemiology of Thyroid Cancer	The paper talks about risk factors, and trends over time. It emphasizes the role of environmental, genetic, and lifestyle factors in influencing disease prevalence and outcomes.
Q1	Nan Miles Xi, Lin Wang and Chuanjia Yang	Improving the diagnosis of thyroid cancer by machine learning and clinical data	The paper discusses how machine learning models, combined with clinical data, improve the accuracy and efficiency of thyroid cancer diagnosis. It evaluates different algorithms and their predictive performance, highlighting AI's potential in medical diagnostics.
Q1	Jianhua Gu, Rongli Xie, Yanna Zhao, Zhifeng Zhao, Dan Xu, Min Ding, Tingyu Lin, Wenjuan Xu, Zihuai Nie, Enjun Miao et al.	A machine learning-based approach to predicting the malignant and metastasis of thyroid cancer	The study used machine learning to predict how dangerous thyroid cancer is and whether it might spread (metastasize). The researchers tested different models like random forest, logistic regression, and gradient boosting using patient data. The models worked well, helping doctors make better and more personalized decisions.

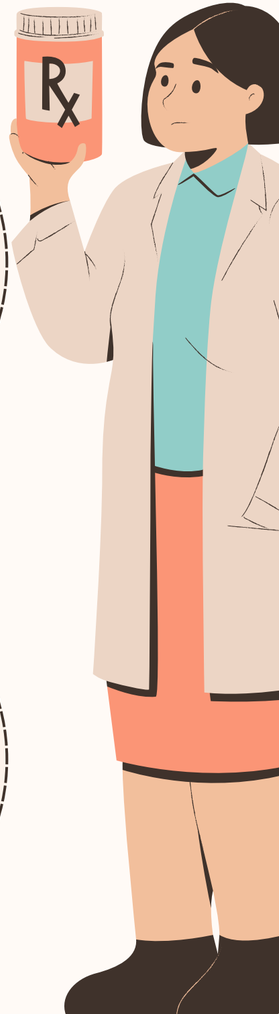
Steps Followed

Data Preprocessing

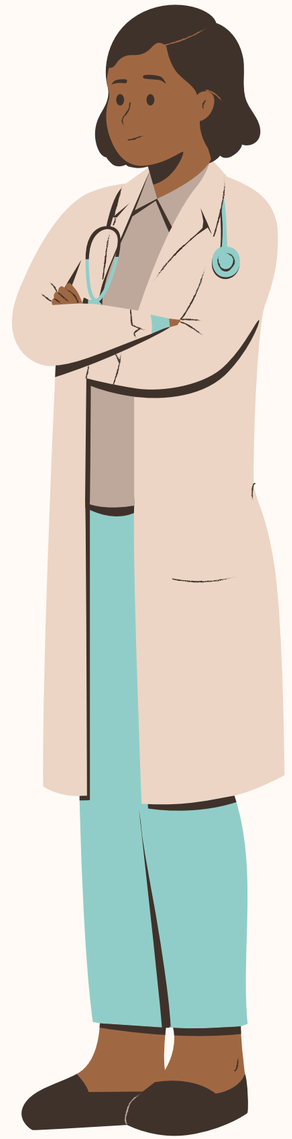
Statistical Analysis

Implementation

User Interface



Data Preprocessing

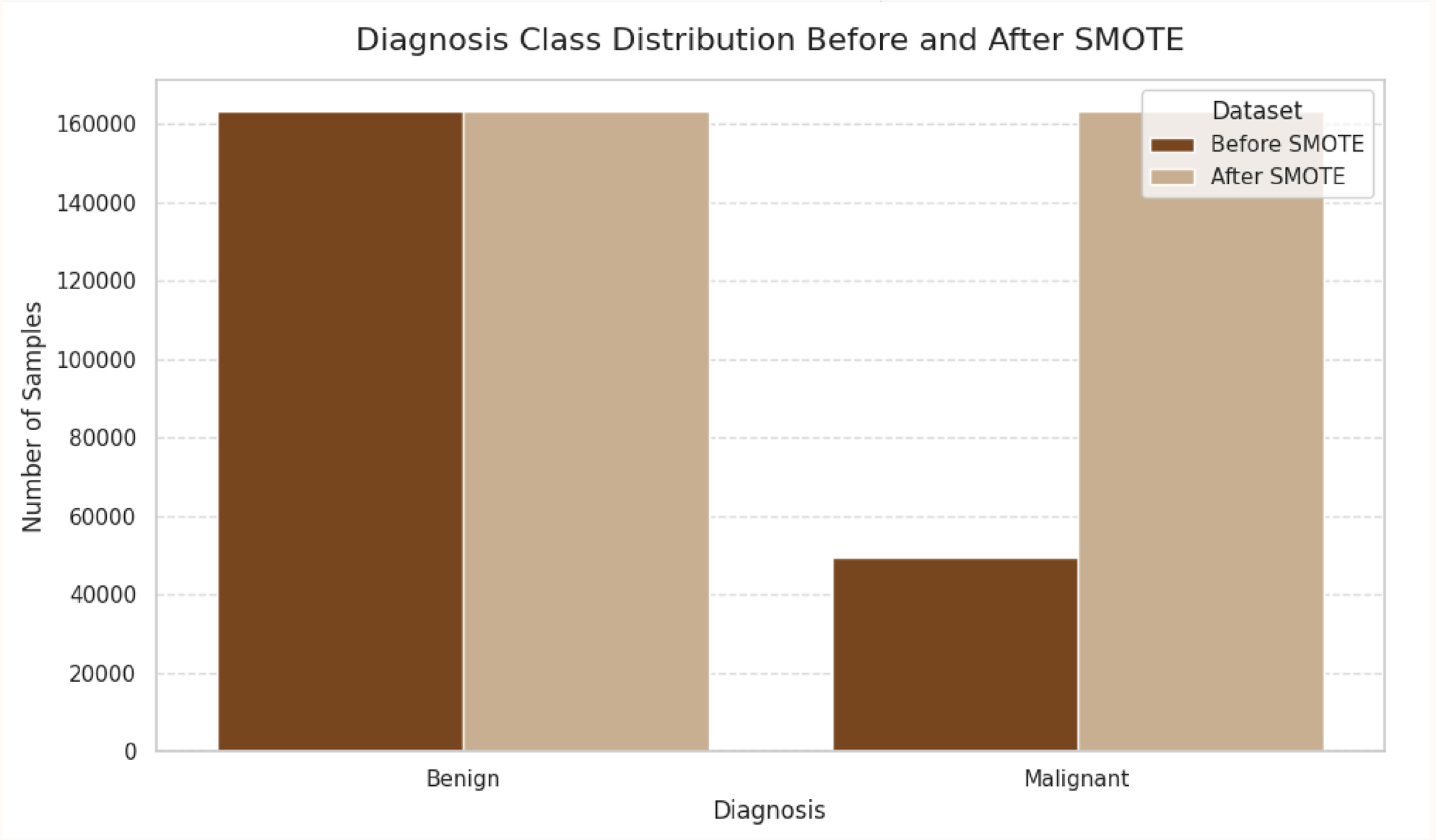


Balancing data

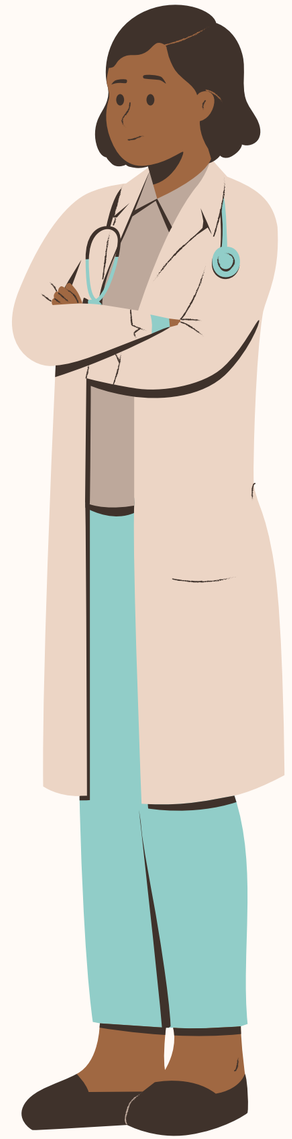
Applied Encoding steps to transform categorical columns into numeric form and classes were balanced using following techniques:

SMOTE **Random Under Sampling** **Random Over Sampling**

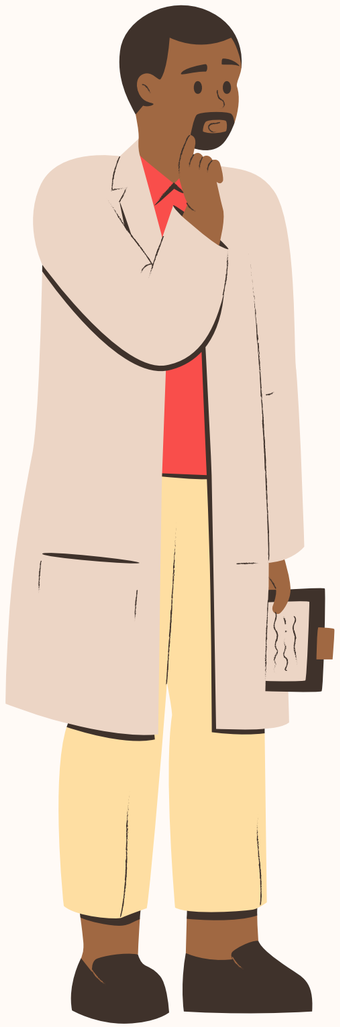
TECHNIQUE	BENIGN	MALIGNANT
SMOTE	1,63,196	1,63,196
RUS	39,571	39,571
ROS	1,63,196	1,63,196


Diagnosis Class Distribution Before and After SMOTE
Dataset
Before SMOTE
After SMOTE
Number of Samples
160000
140000
120000
100000
80000
60000
40000
20000
0
Benign
Malignant
Diagnosis

Statistical Analysis



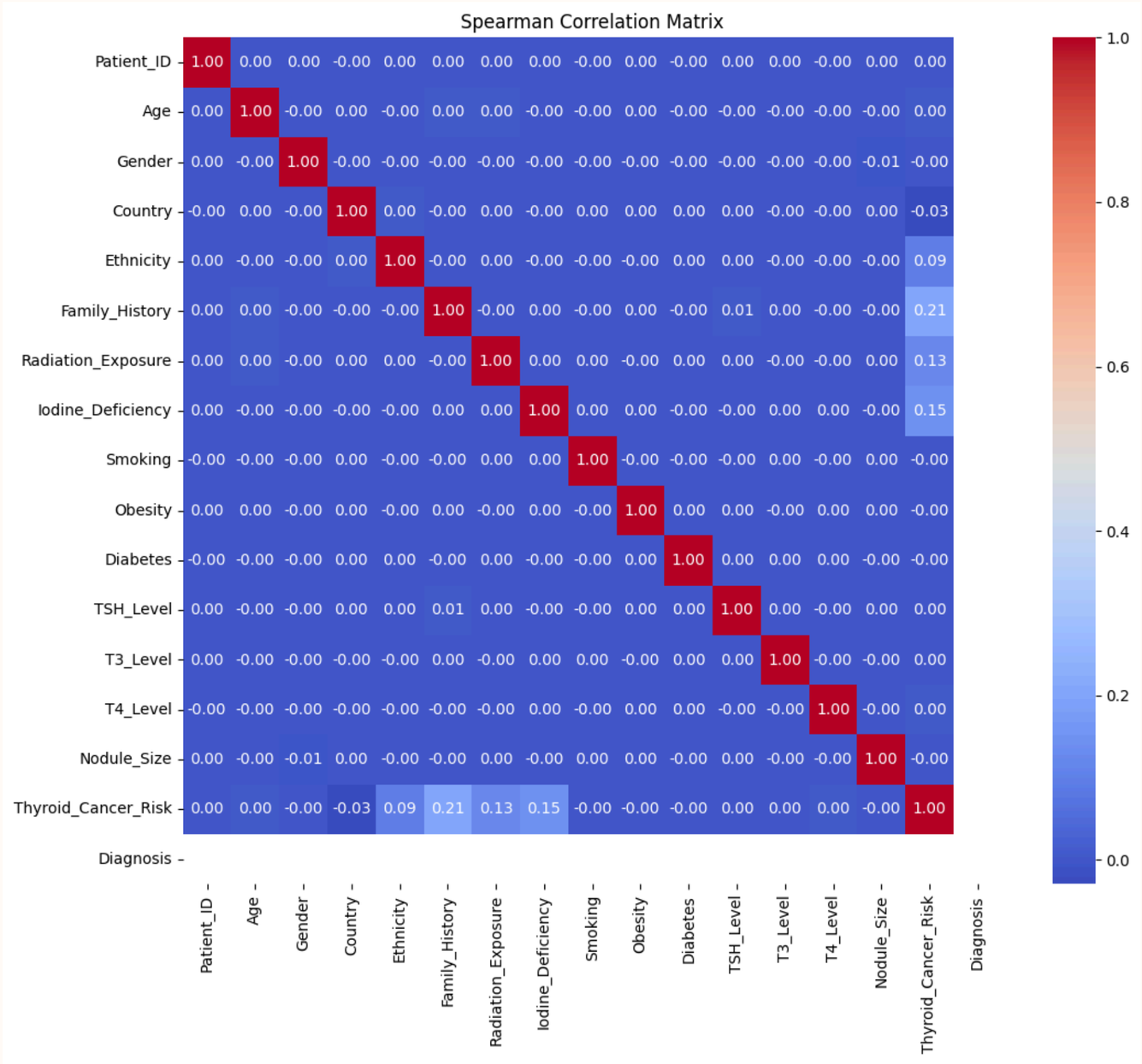
Correlation



Feature	VIF
Age	1.000108
TSH_Level	1.000112
T3_Level	1.000066
T4_Level	1.000081
Nodule_Size	1.000090

No Correlation Found Among Variables

Spearman’s Correlation

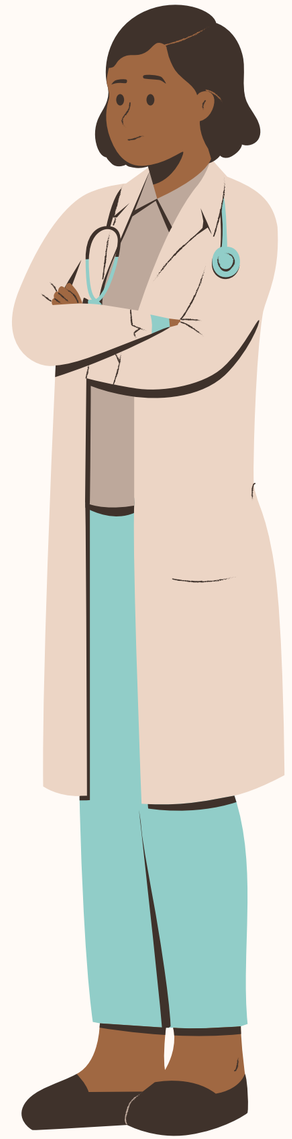


Statistical Tests

Action	Methods	Insights
Normalization	Shapiro-Wilks Kolmogorov-Smirnov Anderson-Darling	Data Not Normally Distributed
Tried to Normalize Data	Log Transformation Box-Cox Transformatiom	Data Remains Unnormalized
Decided to Work on unnormalized data, Identify top features that impact the target	Mann-Whitney U Test Chi Squared Test	Significant: Thyroid Cancer Risk, Diabetes, Smoking, Obesity, Country

H0 Accepted - Found significant difference between benign and malignant cases.

Implementation



- Feature Selection
- 80 : 20 Dataset
- 7 Machine Learning Algorithms
  - Categorical Output
  - No Linear Relationship

Model Training

Model	Recall	Precision	Accuracy	F1 Score
XGBoost	0.74	0.74	0.74	0.74
Random Forest	0.84	0.85	0.84	0.84
Decision Tree	0.84	0.84	0.84	0.84
Gradient Boost	0.72	0.73	0.72	0.72
Light GBM	0.73	0.73	0.73	0.73
Ada Boost	0.72	0.73	0.72	0.72
Stochastic GB	0.72	0.73	0.72	0.72

Model Testing

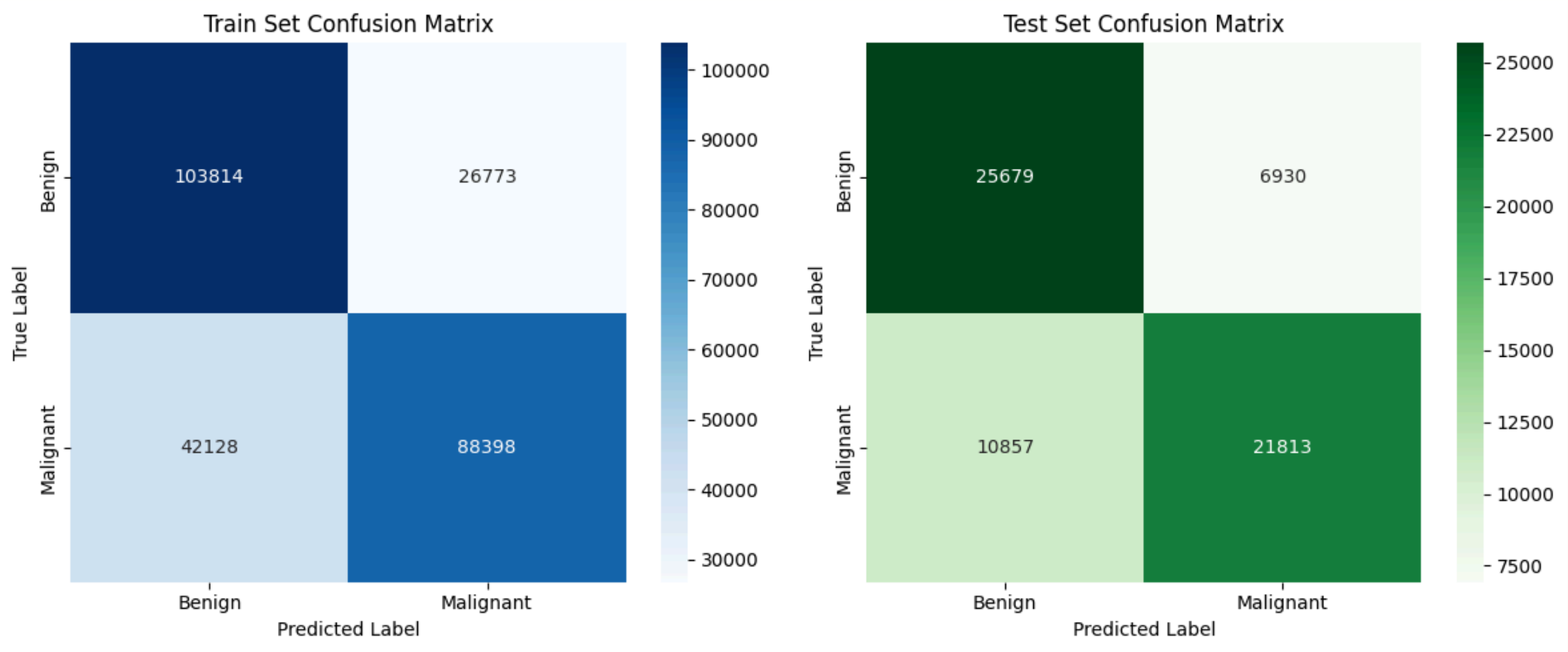
Model	Recall	Precision	Accuracy	F1 Score
XGBoost	0.73	0.73	0.73	0.73
Random Forest	0.72	0.72	0.72	0.72
Decision Tree	0.70	0.70	0.70	0.70
Gradient Boost	0.72	0.73	0.72	0.72
Light GBM	0.72	0.73	0.72	0.72
Ada Boost	0.72	0.73	0.72	0.72
Stochastic GB	0.72	0.73	0.72	0.72

Evaluation

Recall Values for Train & Test Data for each Model

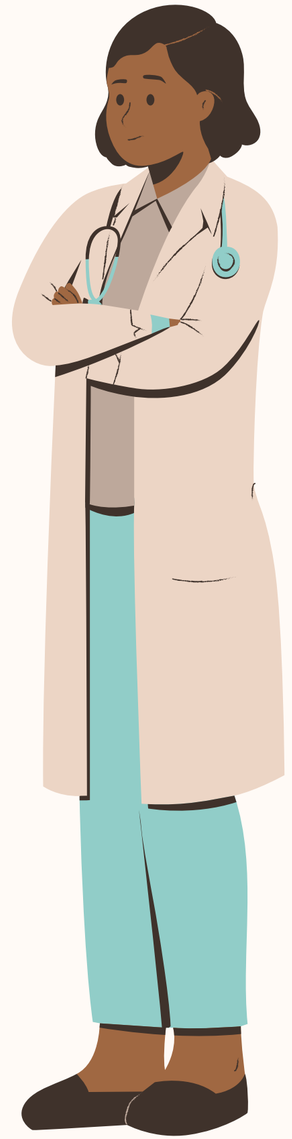
- Olli Rainio, Juha Teuho, and Rami Kl´ en, “Evaluation metrics and statistical tests for machine learning,” Scientific Reports, vol. 14, Art. no. 6086, 2024.

Evaluation



Confusion Matrix for XGBoost Model

User Interface



output

| Family_History - No |
| Radiation_Exposure - No |
| Iodine_Deficiency - No |
| Smoking - No |
| Obesity - No |
| Diabetes - Yes |
| TSH_Level - 0.99 |
| T3_Level - 3.45 |
| T4_Level - 6.16 |
| Nodule_Size - 4.49 |
| Thyroid_Cancer_Risk - Medium |
| Gender - Female |
| Country - Brazil |
| Ethnicity - African |
| Age - 25.0 |

Actual Diagnosis: Benign
Predicted Diagnosis: Benign
Prediction Status: Correct

Clear

Generate

References

[1] Cari M. Kitahara and Arthur B. Schneider, “Epidemiology of thyroid cancer,” Cancer Epidemiology, Biomarkers & Prevention, vol. 31, no. 7, pp. 1284–1297, 2022.

[2] Naykky Singh Ospina, Nicole M. I˜ niguez-Ariza, and M. Regina Castro, “Thyroid nodules: diagnostic evaluation based on thyroid cancer risk assessment,” BMJ, vol. 368, pp. 1–20, 2020.

[3] Jianhua Gu, Rongli Xie, Yanna Zhao, Zhifeng Zhao, Dan Xu, Min Ding, Tingyu Lin, Wenjuan Xu, Zihuai Nie, Enjun Miao, Dan Tan, Sibo Zhu, Dongjie Shen, and Jian Fei, “A machine learning-based approach to predicting the malignant and metastasis of thyroid cancer,” Frontiers in Oncology, vol. 12, pp. 1–13, 2022.

[4] Nan Miles Xi, Lin Wang, and Chuanjia Yang, “Improving the diagnosis of thyroid cancer by machine learning and clinical data,” Scientific Reports, vol. 12, pp. 1–11, 2022.

[5] Jina Kim, Jessica E. Gosnell, and Sanziana A. Roman, “Geographic influences in the global rise of thyroid cancer,” Nature Reviews En docrinology, vol. 16, pp. 17–29, 2020.

[6] Quang T. Nguyen, Eun Joo Lee, Melinda Gingman Huang, Young In Park, Aashish Khullar, and Raymond A. Plodkowski, “Diagnosis and treatment of patients with thyroid cancer,” American Health & Drug Benefits, vol. 8, no. 1, pp. 30–40, 2015.

[7] Luca Giovanella, Alfredo Campenn`ı, Murat Tuncel, and Petra Petranovi´ c Ov˘ cari˘ cek, “Integrated diagnostics of thyroid nodules,” Cancers, vol. 16, no. 2, Art. no. 311, 2024.

[8] Carolyn Dacey Seib and Julie Ann Sosa, “Evolving understanding of the epidemiology of thyroid cancer,” Endocrinology and Metabolism Clinics, vol. 48, no. 1, pp. 23–35, 2019.

References

[9] Yuan-yuan Guo, Zhi-jie Li, Chao Du, Jun Gong, Pu Liao, Jia-xing Zhang, and Cong Shao, “Machine learning for identifying benign and malignant thyroid tumors: A retrospective study of 2,423 patients,” Frontiers in Public Health, vol. 10, 2022.

[10] Umut Percem Orhan Soylemez and Nesrin Gunduz, “Diagnostic accu racy of five different classification systems for thyroid nodules,” Journal of Ultrasound in Medicine, vol. 41, no. 5, pp. 1125–1136, 2021.

[11] Laura Boucai, Mark Zafereo, and Maria E. Cabanillas, “Thyroid cancer: a review,” Clinical Review & Education, vol. 331, no. 5, pp. 425–435, 2024.

[12] Hongxi Wang, Chao Zhang, Qianrui Li, Tian Tian, Rui Huang, Jiajun Qiu, and Rong Tian, “Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches,” BMC Cancer, vol. 24, Art. no. 427, 2024.

[13] C. Schr¨oer, F. Kruse, and J. Marx G´omez, “A systematic literature review on applying CRISP-DM process model,” Procedia Computer Science, vol. 181, pp. 526–534, 2021.

[14] C. Yang, M. D. Naser, S. L. Happy, and T. R. Mahesh, “Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data,” Journal of Big Data, vol. 11, no. 1, p. 7, 2024.

[15] Ankush Panday, “Thyroid Cancer Risk Prediction Dataset,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset

[16] Vanisri Kirubakaran,Chandralekha Venkatesh Perumal, Sneha Krishnan Akavalapil, Ruchika Jha, “Predicting a Cancerous Thyroid Nod ule by Applying Data Balancing Techniques and Machine Learn ing Algorithms,” Github, 2025. [Online]. Available: https://github.com/SnehaKrishnan96/Thyroid-cancer-prediction.

Thank you

Dataset Snippet

Patient_ID	Age	Gender	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
1	66	Male	Russia	Caucasian	No	Yes	No	No	No	No	9.37	1.67	6.16	1.08	Low	Benign
2	29	Male	Germany	Hispanic	No	Yes	No	No	No	No	1.83	1.73	10.54	4.05	Low	Benign
3	86	Male	Nigeria	Caucasian	No	No	No	No	No	No	6.26	2.59	10.57	4.61	Low	Benign
4	75	Female	India	Asian	No	No	No	No	No	No	4.1	2.62	11.04	2.46	Medium	Benign
5	35	Female	Germany	African	Yes	Yes	No	No	No	No	9.1	2.11	10.71	2.11	High	Benign
6	89	Male	UK	African	No	No	No	Yes	Yes	No	4	0.98	5.52	0.02	Medium	Benign
7	89	Female	South Korea	Asian	Yes	Yes	No	No	Yes	No	4.7	0.62	11.73	0.01	High	Malignant
8	38	Female	India	African	No	No	No	No	No	No	5.54	3.49	9.47	4.3	Medium	Benign
9	17	Female	Russia	African	No	Yes	No	No	No	Yes	2.3	2.6	11.89	0.81	High	Malignant
10	36	Male	Germany	Asian	No	No	No	No	Yes	No	1.34	0.56	4.51	1.44	Low	Benign
11	67	Male	Nigeria	African	No	Yes	No	No	No	No	9.65	1.82	8.17	0.35	High	Malignant
12	16	Female	Nigeria	Asian	No	No	No	Yes	No	No	0.53	1.13	9.56	3.87	Medium	Benign
13	44	Male	South Korea	Asian	Yes	No	No	No	No	Yes	6.77	1.37	6.13	4.15	High	Malignant
14	52	Male	Brazil	Asian	No	No	No	No	No	No	4.91	0.95	6	0.38	Low	Benign
15	16	Female	China	Asian	No	No	No	No	No	No	6.84	0.62	6.8	1.68	Medium	Benign
16	78	Female	Nigeria	Caucasian	Yes	No	Yes	Yes	No	No	7.32	1.9	11.82	2.86	Low	Benign
17	74	Female	India	African	Yes	No	No	No	No	No	9.6	2.86	11.5	0.25	Low	Benign
18	35	Male	Japan	Hispanic	No	No	No	No	No	No	3.59	1.83	4.95	4.93	Medium	Benign
19	47	Female	USA	Caucasian	No	No	No	No	No	Yes	6.43	3.39	5.66	1.63	Medium	Benign
20	72	Female	Japan	Caucasian	No	No	No	No	No	No	5.96	1.26	7.89	2.27	Low	Benign
21	36	Male	Russia	Hispanic	No	No	Yes	No	Yes	No	4.17	2.92	10.24	2.41	Low	Benign
22	63	Female	Nigeria	Asian	No	No	No	No	Yes	Yes	6.97	3.48	7.67	0.46	Low	Malignant

thyroid_cancer_risk_data

Ranking	Authors	Paper Name	Summary
Q1	Carolyn Dacey Seib and Julie Ann Sosa	Evolving understanding of the epidemiology of thyroid cancer	The paper analyzes the evolution of understanding using data-driven methods, identifying key factors and proposing strategies to enhance comprehension and decision-making.
Q1	Cari M. Kitahara and Arthur B. Schneider	Epidemiology of Thyroid Cancer	The paper talks about risk factors, and trends over time. It emphasizes the role of environmental, genetic, and lifestyle factors in influencing disease prevalence and outcomes.
Q1	Albano et al.	A data-driven approach to refine predictions of differentiated thyroid cancer outcomes: A prospective multicenter study	This study used a data-driven (machine learning) approach to better predict the outcomes of patients with differentiated thyroid cancer. By analyzing real patient data from multiple hospitals, the model helped doctors make more accurate treatment decisions.
Q1	Hongxi Wang, Chao Zhang, Qianrui Li, Tian Tian, Rui Huang, Jiajun Qiu and Rong Tian	Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches	The paper focuses on developing and validating machine learning-based prediction models for structural recurrence in papillary thyroid cancer. It evaluates various ML approaches to enhance early detection and improve patient outcomes.

Ranking	Authors	Paper Name	Summary
Q1	Nan Miles Xi, Lin Wang and Chuanjia Yang	Improving the diagnosis of thyroid cancer by machine learning and clinical data	The paper discusses how machine learning models, combined with clinical data, improve the accuracy and efficiency of thyroid cancer diagnosis. It evaluates different algorithms and their predictive performance, highlighting AI's potential in medical diagnostics.
Q1	Xie et al.	A machine learning-based approach to predicting the malignant and metastasis of thyroid cancer	The study used machine learning to predict how dangerous thyroid cancer is and whether it might spread (metastasize). The researchers tested different models like random forest, logistic regression, and gradient boosting using patient data. The models worked well, helping doctors make better and more personalized decisions.
Q1	Luca Giovanella, Alfredo Campennì, Murat Tuncel and Petra Petranovi ć Ov ˇ cari ˇ cek	Integrated diagnostics of thyroid nodules	This paper discusses how combining different diagnostic tools—like ultrasound, blood tests, and molecular testing—can improve the accuracy of thyroid nodule diagnosis. It shows that using an integrated approach helps avoid unnecessary surgeries and better detect cancer
Q2	Chan Kwon Jung,Andrey Bychkov and Kennichi Kakudo	Update from the 2022 World Health Organization Classification of Thyroid Tumors: A standardized diagnostic approach	This paper explains the updates made in the 2022 WHO classification of thyroid tumors, offering a more standardized way to diagnose them. It helps doctors better identify different types of thyroid cancer based on clear guidelines.

Predicting Thyroid Cancer using XGBoost

Randomly selects a record from the dataset and checks if our trained XGBoost model predicts the diagnosis correctly.

output

Clear

Generate

Predicting Thyroid Cancer using XGBoost

Randomly selects a record from the dataset and checks if our trained XGBoost model predicts the diagnosis correctly.

output

Feature - Value
Family_History - No
Radiation_Exposure - No
Iodine_Deficiency - Yes
Smoking - No
Obesity - No
Diabetes - No
TSH_Level - 4.878228388519075
T3_Level - 1.4040268944756988
T4_Level - 11.591365207953618
Nodule_Size - 1.28974813244578
Thyroid_Cancer_Risk - High
Gender - Female
Country - India
Ethnicity - Middle Eastern
Age - 21.0

Actual Diagnosis: Malignant
Predicted Diagnosis: Malignant
Prediction Status: Correct

Clear

Generate