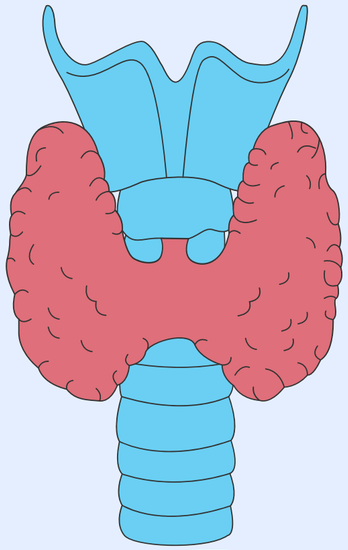
Predicting Thyroid Malignancy Using Machine Learning

Presented By: Group 02
Tishani Wijekoon,
Chami Sewwandi,
W.K.Hiruni Hasara,
S.Luxan

What Is the Thyroid Gland?

Butterfly-shaped gland in the neck



A healthy thyroid keeps your body running smoothly

Produces hormones that control metabolism, energy use, and growth

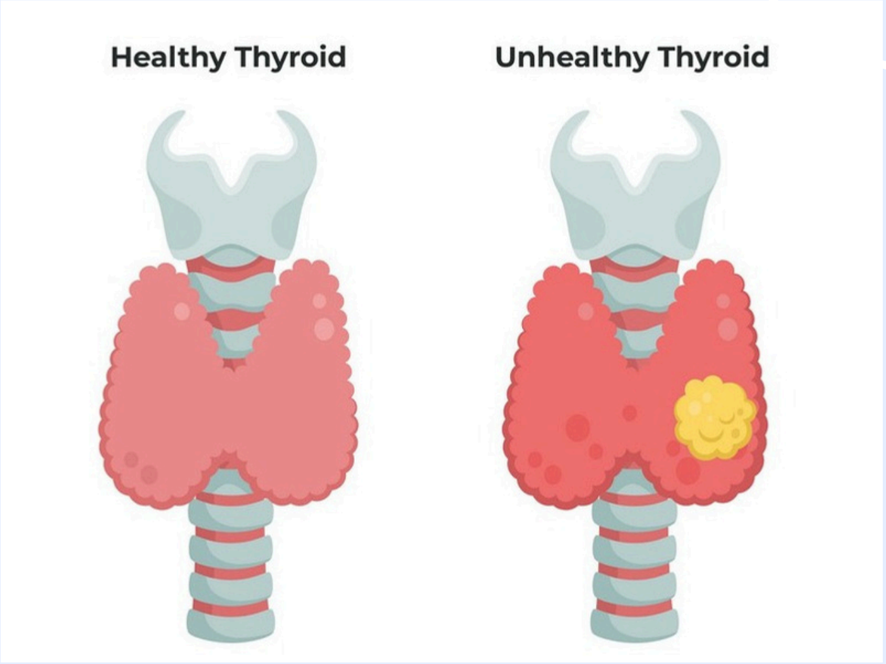
What Is Thyroid Cancer?

Abnormal cells grow in the thyroid gland

Usually grows slowly

Most common types: papillary and follicular

Highly treatable if diagnosed early

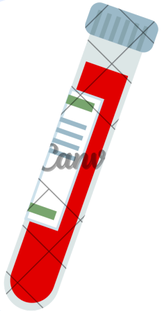


Benign vs. Malignant

Benign (Non-Cancerous)	Malignant (Cancerous)
Harmless growth	Dangerous tumor
Doesn’t spread	Can invade nearby tissue
May not need treatment	Needs treatment early

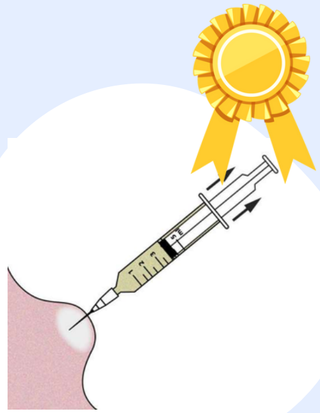
How Is Thyroid Cancer Diagnosed?

Physical exam



Blood tests (TSH, T3, T4)

Ultrasound imaging

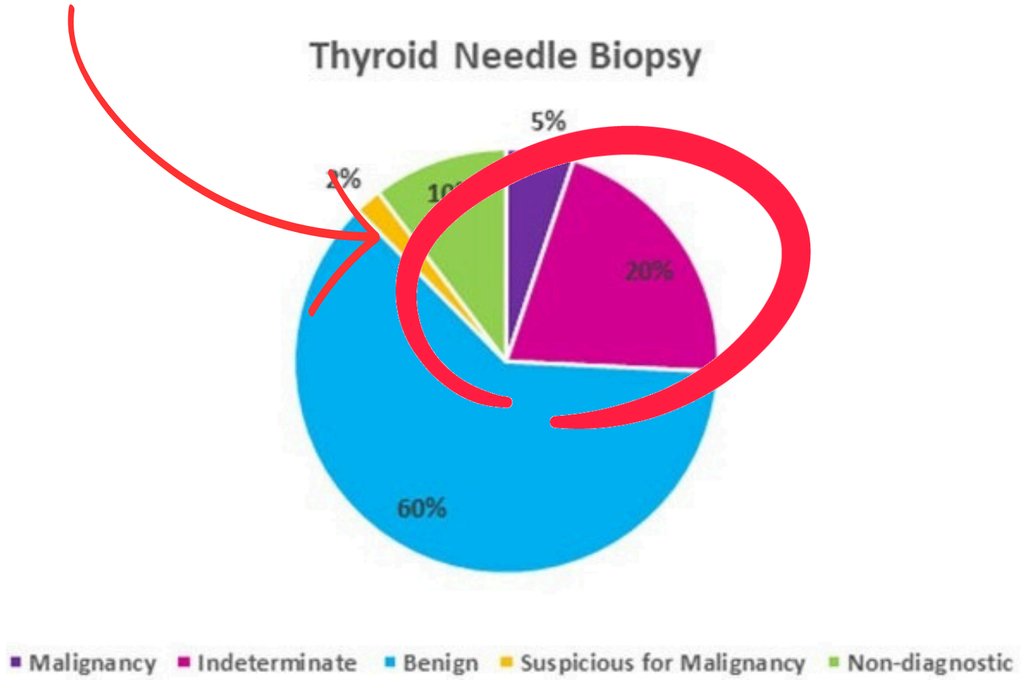


FNAB- gold standard

Limitations of FNAB

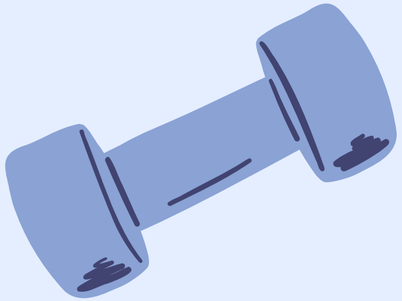
Why FNAB Isn't Always Enough?

- Invasive and costly
- Results often fall into “indeterminate” zone (Gray zone)
- May lead to unnecessary surgery
- Sometimes fails to detect cancer



Our Research Question

“Can we help doctors make that decision earlier — using just basic data? ”



About the Dataset

210,000+ patient records (212,691)

- 16 variables
- Features: ID, Age, Gender, Country, Ethnicity, Family_History, Radiation_Exposure, Iodine_Deficiency, Smoking, Obesity ,Diabetes, TSH_Level, T3_Level, T4_Level, Nodule_Size, Diagnosis
- Target: Benign or Malignant diagnosis
- No missing
- No duplicate values
- No outliers

Feature Selection & Evaluation

Feature Selection Approach:
Lasso Regularization

Statistical Evaluation:

- Mann-Whitney U Test for continuous features (e.g., TSH, T3, T4, nodule size)
- Chi-Squared Test for categorical features (e.g., gender, family history, exposure)

Country

Ethnicity

Family History

Radiation Exposure

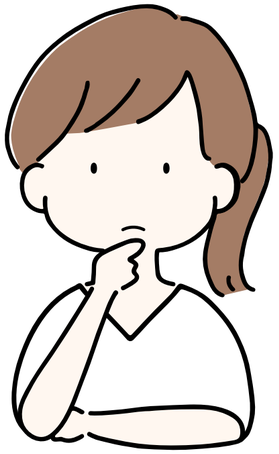
Iodine Deficiency

Insights

Understanding the Unexpected!!!

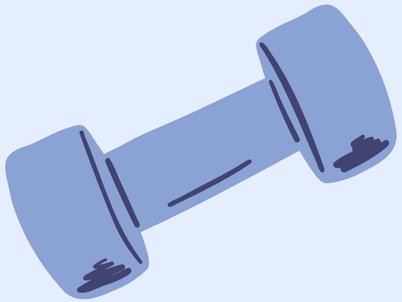
- **Age - Most common around 30-60 years** [American Cancer Society]
- **Gender - Occurs more often in women** [MAYO Clinic]
- **Family History - Close relatives with cancer increases risk** [Cancer Research UK]
- **Radiation Exposure** [Cancer Research UK]
- **Obesity** [Cancer Research UK]
- **Nodule Size** [Clayman Thyroid Center]
- **TSH Level** [Alaraifi et al. (2023)]

WHY DOESN'T OUR DATA REFLECT THESE FACTS?



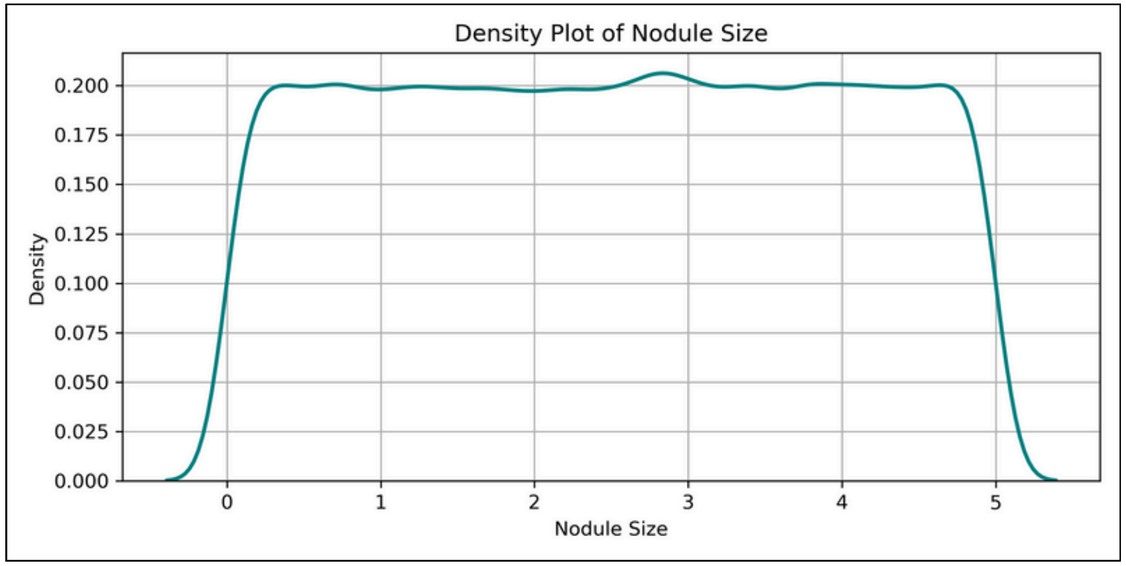
Our New Research Question

“Can we support diagnosis in patients with indeterminate biopsy results?”

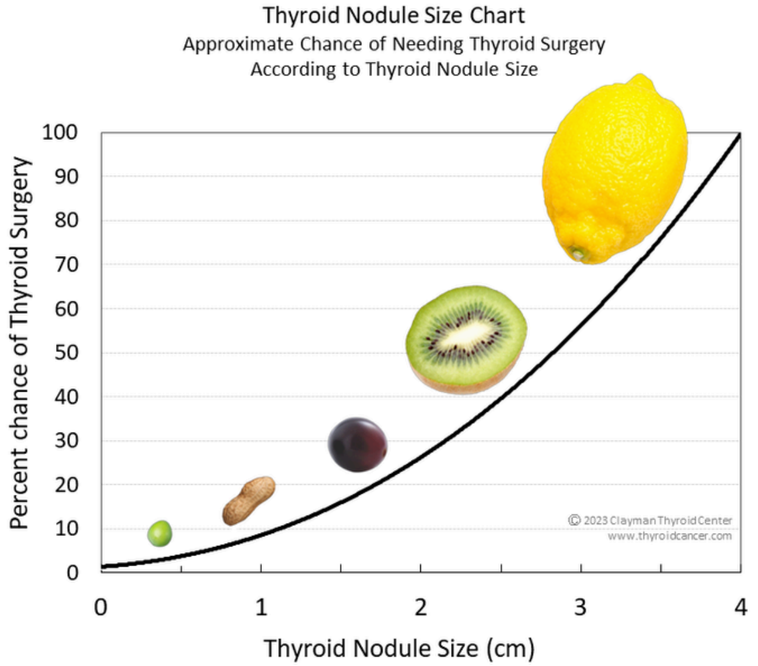


Data Exploration & Insights

Nodule Size



Majority of our dataset need a Fine Needle Aspiration Biopsy Result



Clayman Thyroid Center

Percentage of observations with nodule size > 1 cm is almost 80%

Nodules of size greater than 1 cm usually need an FNAB

-Clayman Thyroid Center -

-American Thyroid Association (ATA) Guidelines (2015)-

What happens if Results are not Clear?

Bethesda Categories III to VI

Majority of our dataset have suspicious or ambiguous presentations.

Analyzing Cases with Diagnostic Uncertainty

Bethesda Categories III and IV
-Gray Zone-

References

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.

Cibas & Ali, The Bethesda System for Reporting Thyroid Cytopathology (2017)

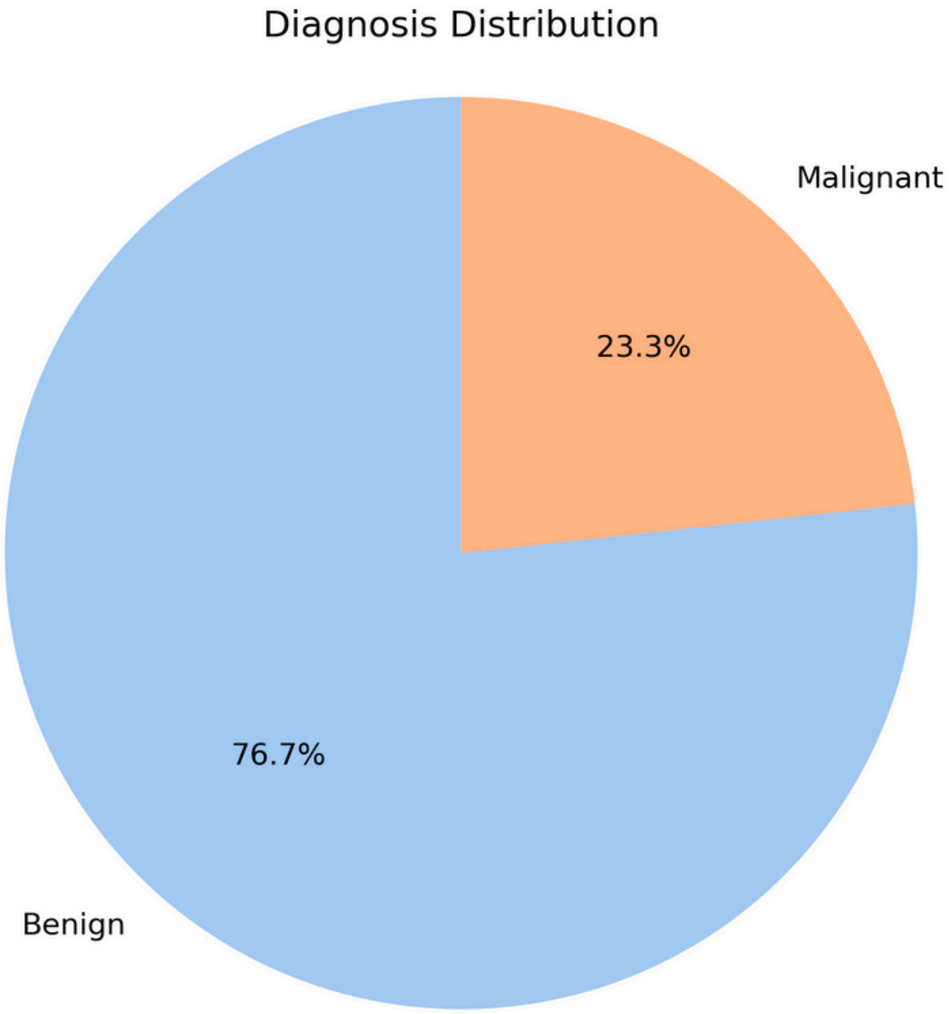
Response Variable–Diagnosis

Presence of Class Imbalance

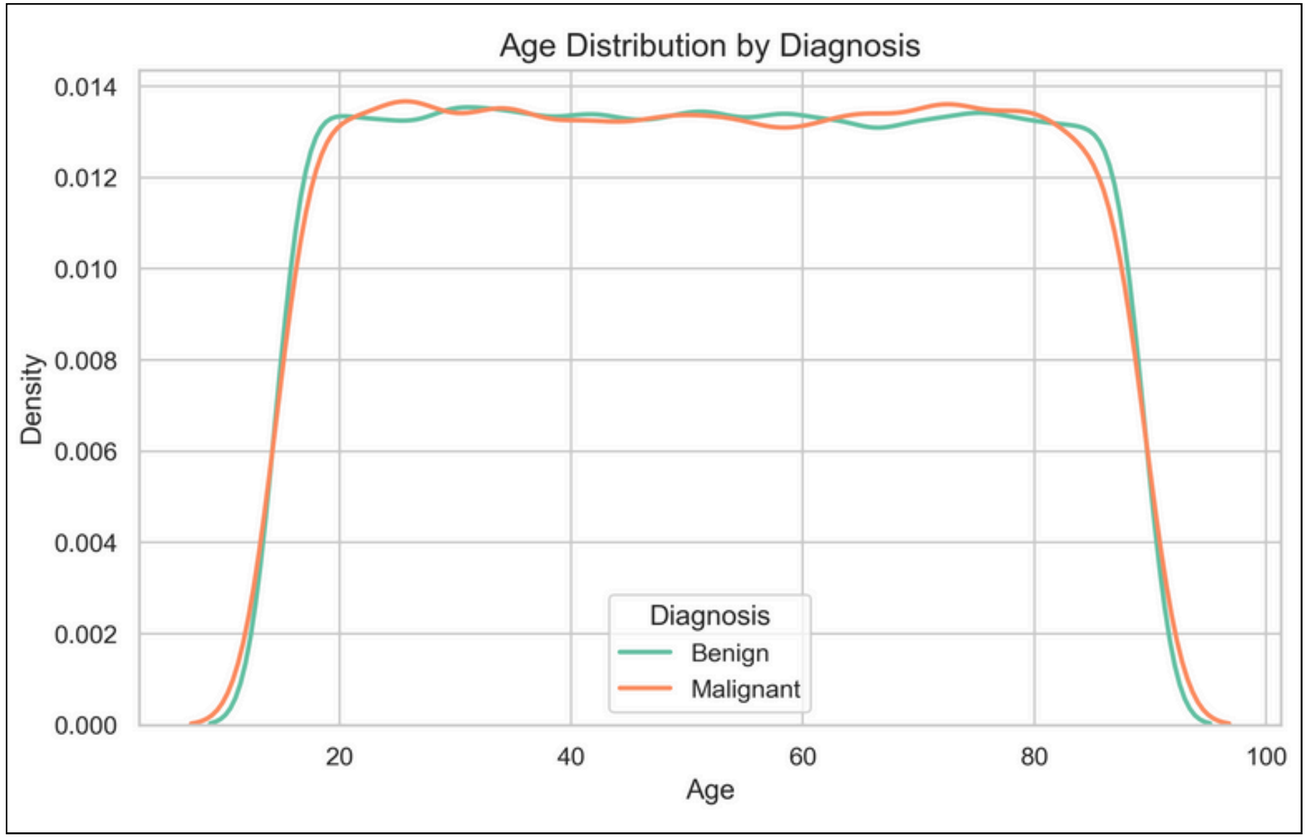
- Benign Tumors 76.7%
- Malignant Tumors 23.3%

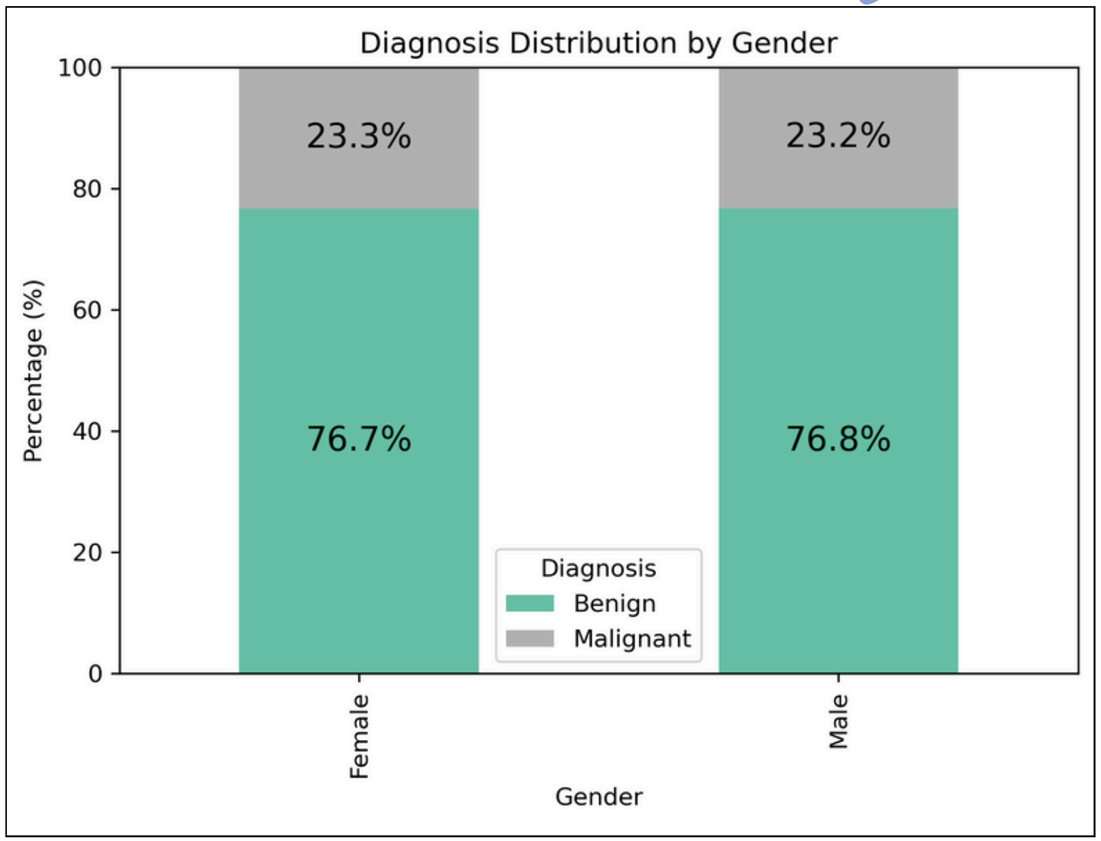
Solution: Resampling Techniques

→ SMOTE



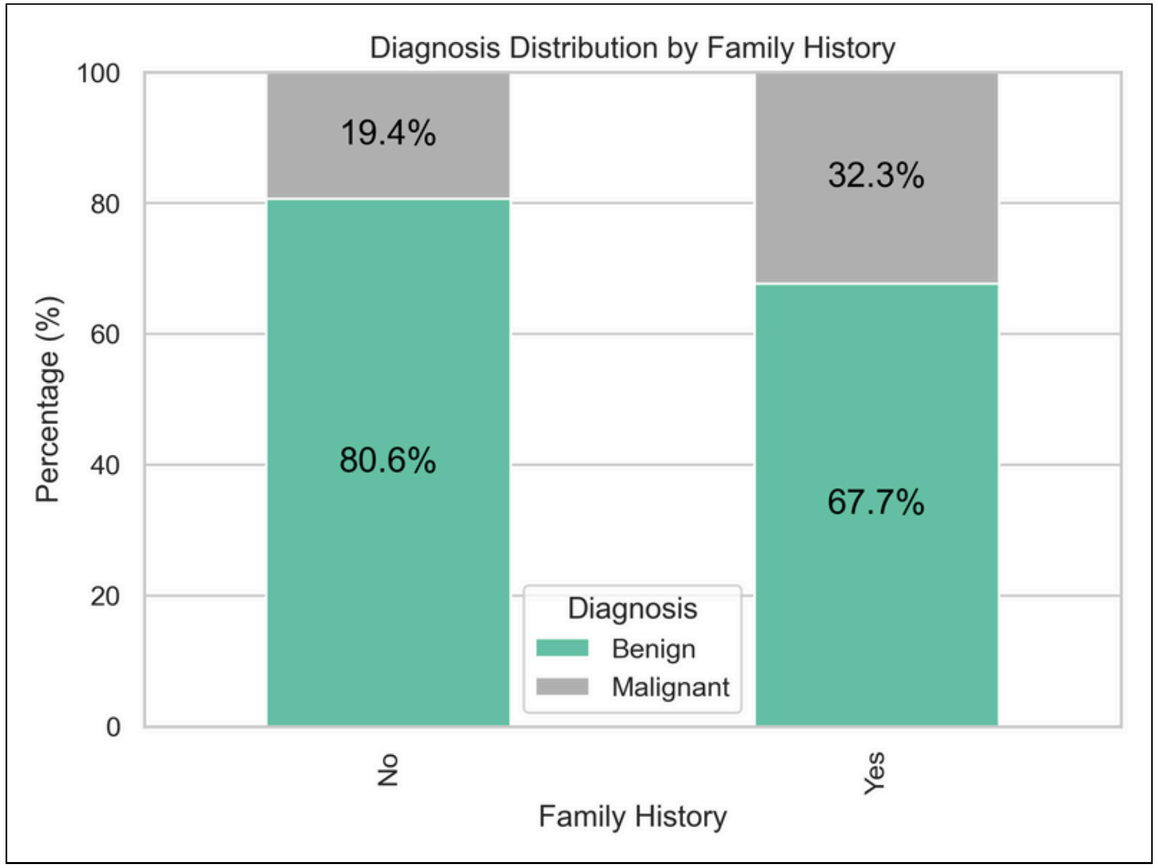
Age and Gender

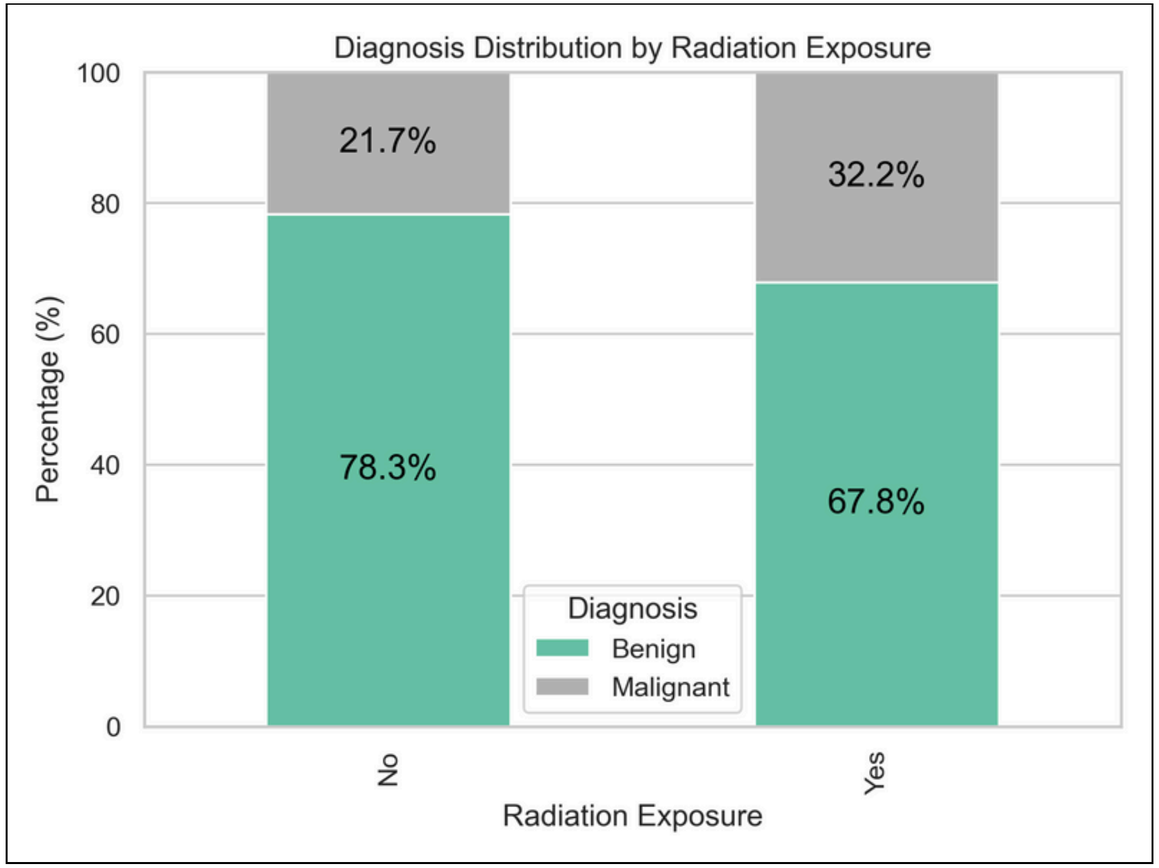




- Similar patterns observed for both benign and malignant tumors.
- No relationship can be observed between age or gender with diagnosis.

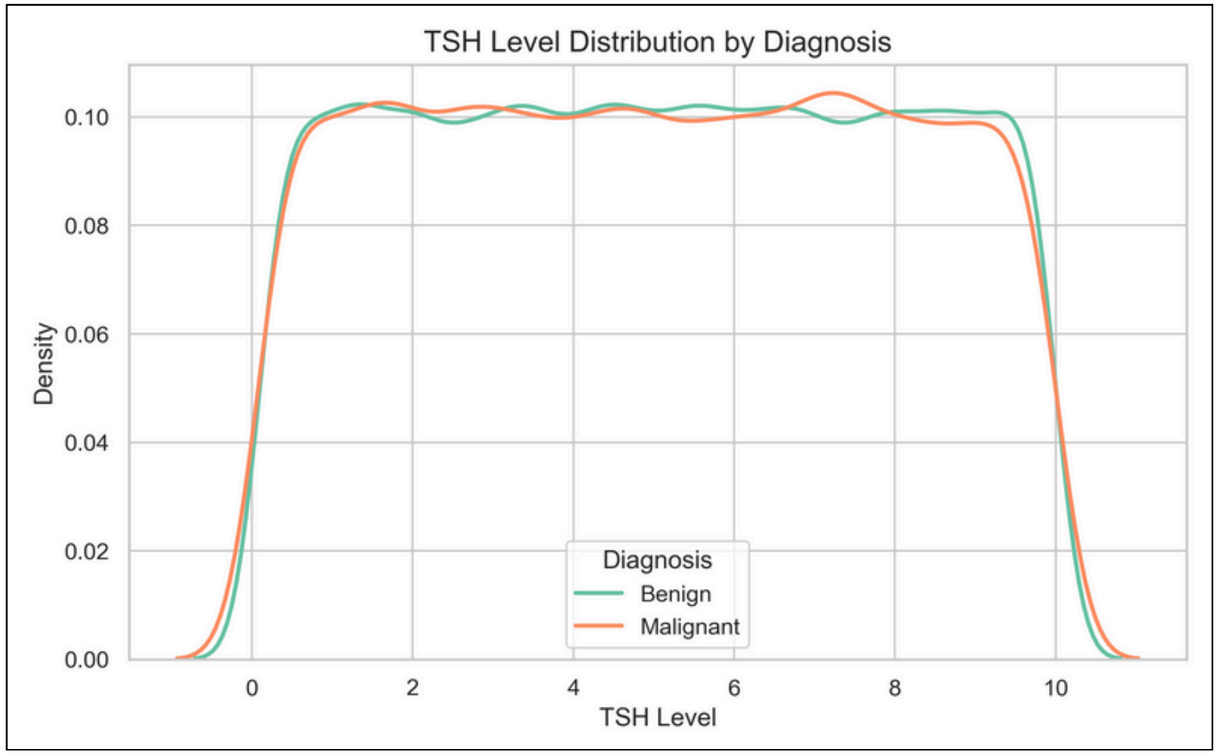
Family History & Radiation Exposure

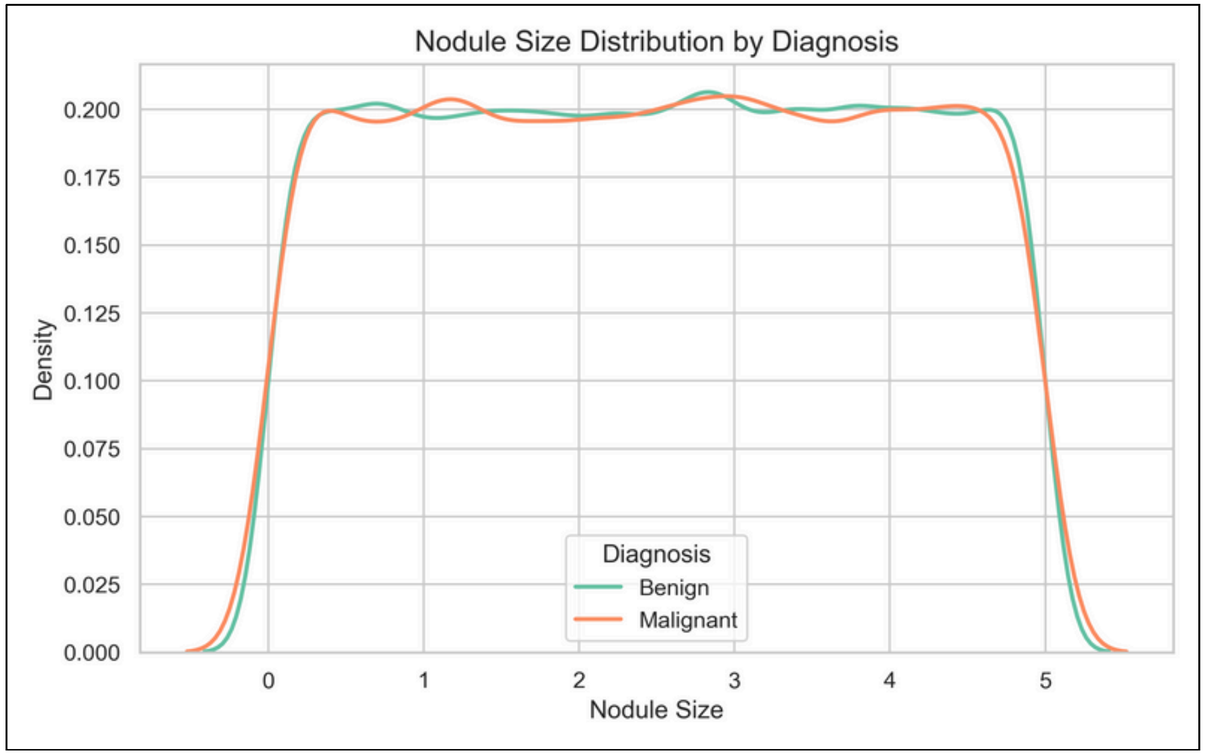




A relationship between both Family History and Radiation Exposure with Diagnosis can be observed.

TSH Level and Nodule Size





- Similar patterns observed for both benign and malignant tumors.
- No relationship can be observed between TSH level or Nodule Size with diagnosis.

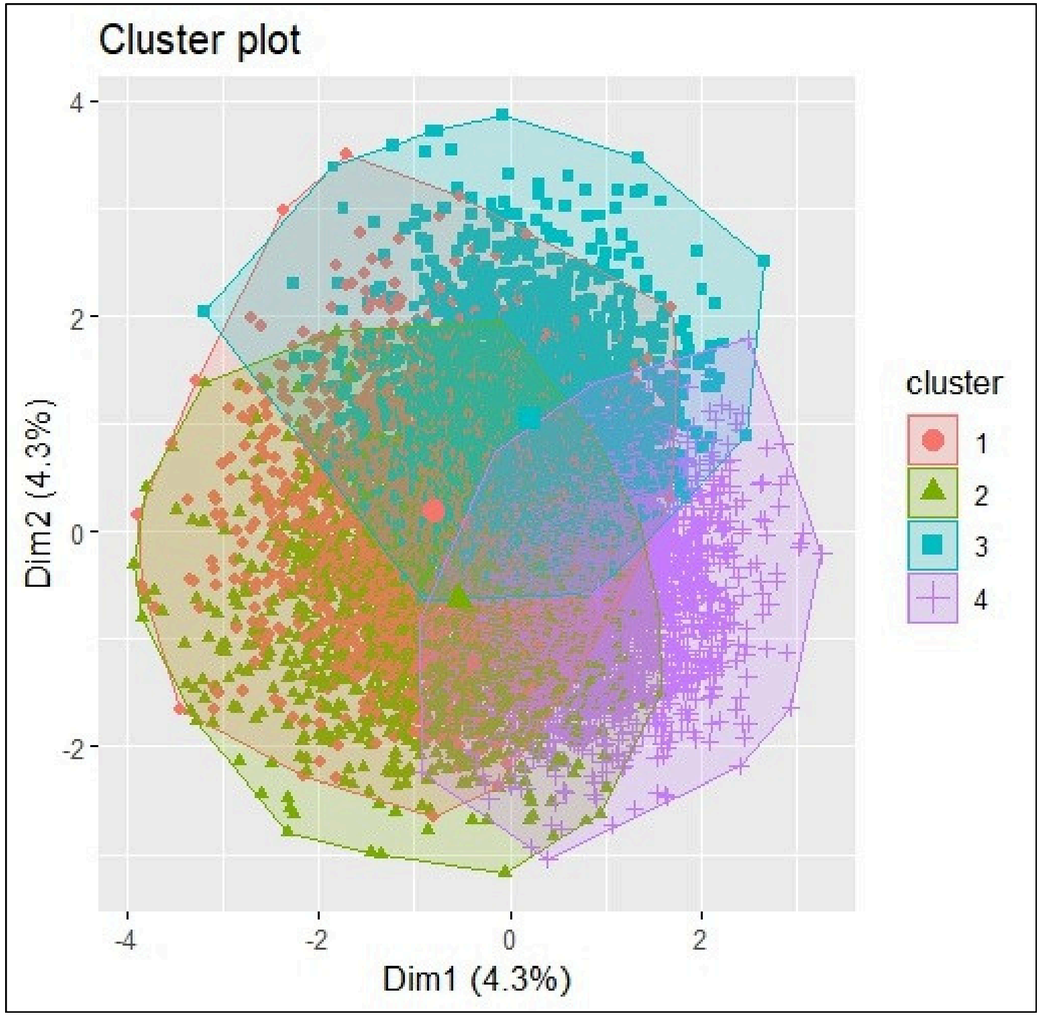
Multivariate Analysis

Factor Analysis for Mixed Data

Top 5 FAMD components explained 21.22% of total variance.

Component 1 - 4.39%
Component 2 - 4.28%
Component 3 - 4.26%
Component 4 - 4.24%
Component 5 - 4.19%

Clusters??



Advanced Analysis

PREDICTIVE MODELS: STRATEGY & SETUP

- Applied SMOTE to balance the class distribution before model training.

Implemented multiple models:

- Logistic Regression
- Decision Tree
- SVM
- Random Forest
- XGBoost

Performed hyperparameter tuning using GridSearchCV.

Evaluated models using:

- Precision
- Recall
- F1-score
- Accuracy

MODEL COMPARISON

	TRAIN ACCURACY	TEST ACCURACY	PRECISION	RECALL	F1_SCORE
LOGISTIC	71%	67 %	33%	41%	36%
SVM	71%	67%	33%	40%	36%
DECISION TREE	99%	66%	34%	47%	39%
RANDOM FOREST	99%	78%	54%	44%	48%
XG BOOST	80%	**80%**	**56%**	**47%**	**51%**

WINNER–XG BOOST
WHY?

Highest Test Accuracy 80%

Best F1 Score 51%

No Overfitting

Strong Precision & Recall

WHY NOT OTHER MODELS?

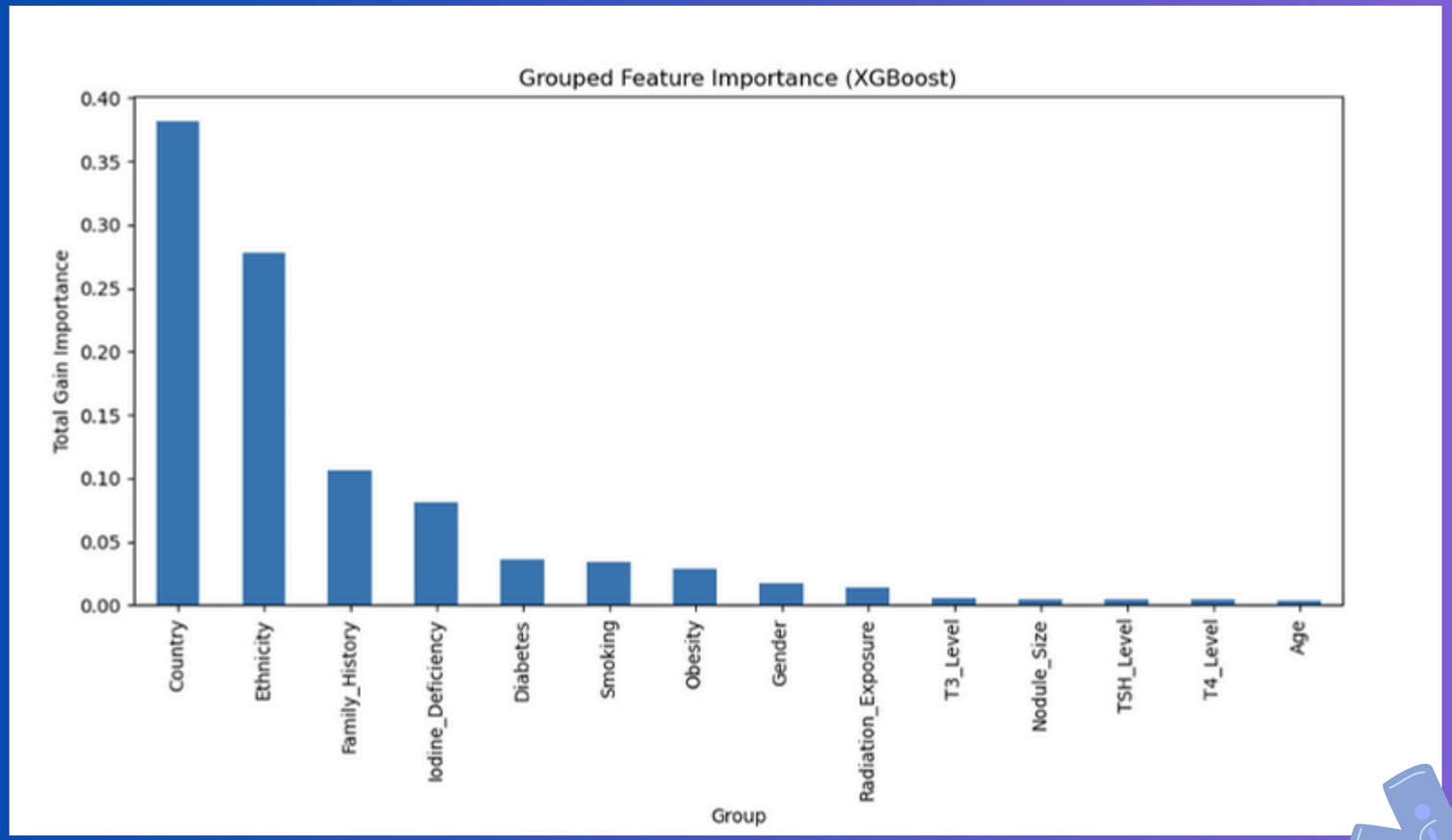
SVM and Logistic -

- F1 Score is low.Not suitable for real-world prediction.

Random Forest and Decission Tree-

- Overfitted

FEATURE IMPORTANCE PLOT XG BOOST



1. Country-38%
2. Ethnicity-28%
3. Family History-10%
4. Iodine Deficiency-8%

Model Interpretation





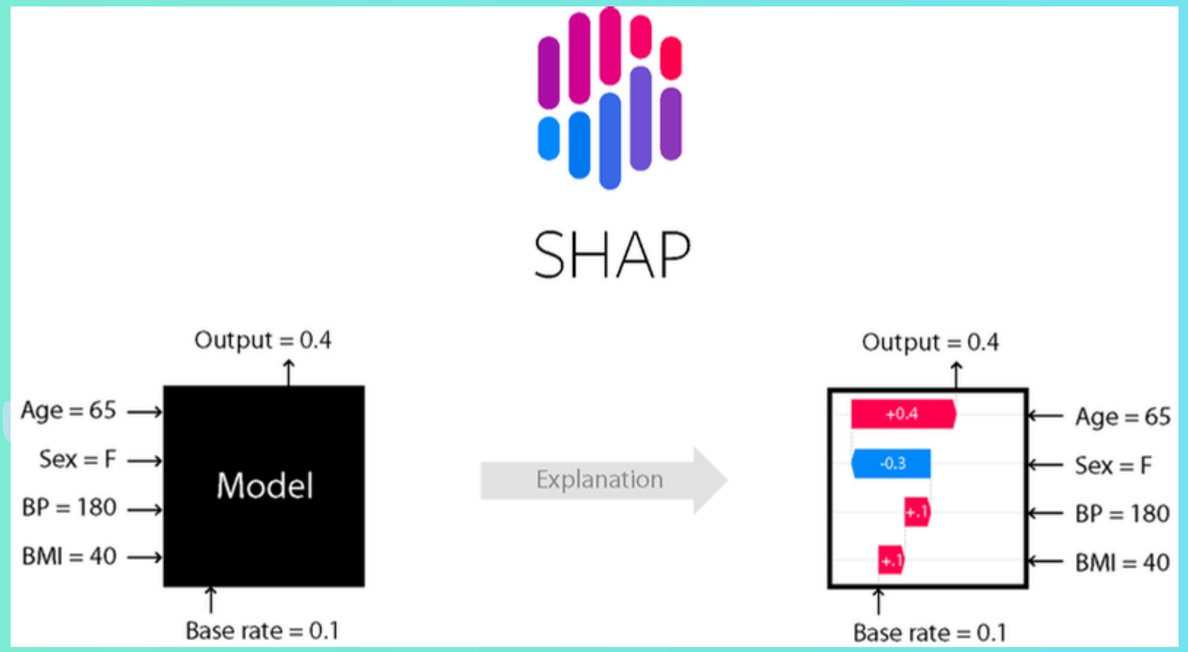

Explaining Model Predictions with SHAP

Overview

- Stands for ‘SHapley Additive exPlanations’.
- Based on SHapley values from Game Theory.
- Measures each feature's marginal contribution to a prediction.

Why SHAP?

- Traditional methods (eg: PDP) assume continuous variables.
- Handles both numerical and continuous variables.
- See which features are driving predictions.



A Graphical representation of how SHAP works

How the Model Thinks...

🎯 Dot position = influence on prediction

➡️ Positive SHAP → pushes toward Malignant

⬅️ Negative SHAP → pushes toward Benign

🔴 = High value 🔵 = Low value

Most impactful : Country, Ethnicity

Pushing Towards Malignancy:
Country, Ethnicity

Pushing Towards Benign:
Obesity, Diabetes, Smoking

Real-World Insights from Model Interpretation

Impact of Country of Residence on Thyroid Cancer Risk

Higher SDI Countries

- Increased incidence is observed.
- Possible reasons include:

1. Widespread use of advanced diagnostic techniques.
2. Socioeconomic development and improved health awareness.
3. Higher rates of overdiagnosis, especially of small, slow-growing tumors.

Lower SDI Countries

- Some studies report higher mortality rates.
- Potential contributing factors:

1. Limited access to quality healthcare.
2. Lack of advanced medical services and accurate laboratory investigations.

Impact of Ethnic Background on Thyroid Cancer Risk

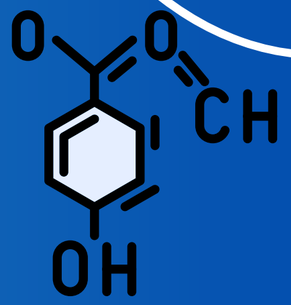
- Ethnic disparities may influence thyroid cancer risk, often in combination with other factors.
- A study in Detroit (Peterson et al., 2011) found that:
  - Arab and non-Hispanic White women had different medical practices, dietary habits, and lifestyle factors, potentially affecting risk levels.
- Sanabria et al. (2018) identified ethnicity as a contributing factor, linked to:
  - Geographic location of ethnic groups
  - Higher education, urban living, and socioeconomic status, which often result in better healthcare access and, in some cases, higher detection rates.

- Limited studies have explored ethnicity as an independent risk factor; most findings suggest it interacts with social and environmental determinants.

Impact of Family History of Thyroid Cancer on Thyroid Cancer Risk

- According to the American Cancer Society, having a first-degree relative (parent, sibling, or child) with thyroid cancer increases the risk of developing the disease.
- However, most people diagnosed with thyroid cancer do not have a family history of the condition.
- A 2018 review in the World Cancer Research Journal reported:
  - A clear association between family history and thyroid cancer incidence.
  - Recommended regular screenings for individuals with a family history to enable early detection and prevent disease progression.

Impact of Iodine Deficiency on Thyroid Cancer Risk



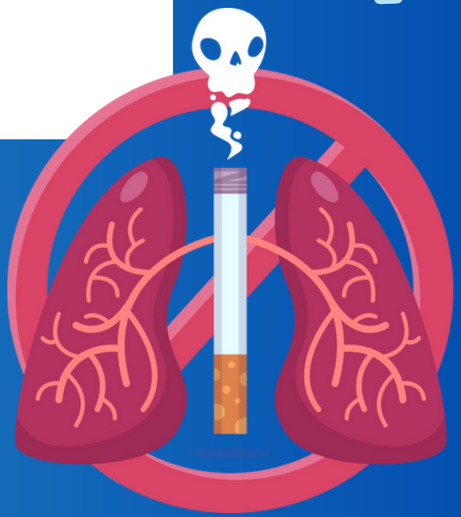
- The American Cancer Society notes:
  - Follicular thyroid cancer is more common in regions with low dietary iodine.
  - Conversely, a high-iodine diet may increase the risk of papillary thyroid cancer.
- Numerous studies identify iodine deficiency as a significant risk factor for thyroid cancer.
- Iodized salt programs have been linked to:
  - Reduced iodine deficiency in populations.
  - A decline in thyroid cancer mortality in many countries.

Impact of Smoking on Thyroid Cancer Risk

- While smoking is a known risk factor for many types of cancer,studies suggest it may be linked to a reduced risk of developing thyroid cancer.
- This risk reduction has been observed across different population groups.

- Possible biological explanations include:
  - Anti-estrogenic effects of tobacco smoke
  - Lower levels of Thyroid Stimulating Hormone (TSH) in smokers

Note: This does not imply smoking is beneficial — it increases the risk of many serious diseases.

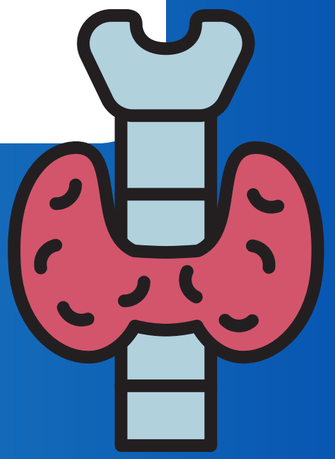


Impact of Obesity on Thyroid Cancer Risk

- The American Cancer Society reports that individuals with excess weight have a higher risk of developing thyroid cancer compared to those at a healthy weight.
- Multiple studies support this finding, identifying obesity as a significant and severe risk factor.
- The increased risk is believed to be linked to:
  - Hormonal imbalances
  - Chronic inflammation
  - Metabolic changes associated with excess body fat

Impact of Gender on Thyroid Cancer Risk

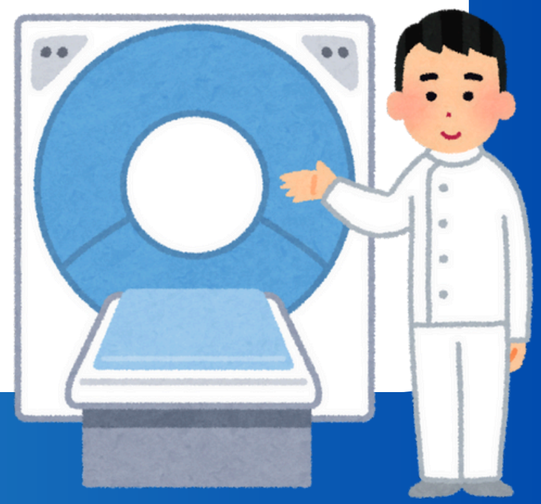
- According to the American Cancer Society, thyroid cancer occurs nearly 3 times more often in women than in men — though the exact reason remains unclear.
- Possible contributing factors include:
  - Hormonal influences, especially abnormal estrogen activity
  - Unhealthy dietary patterns
  - Higher rates of medical visits among women, particularly during pregnancy and gynecological care, may lead to more frequent detection



Impact of History of Radiation Exposure on Thyroid Cancer Risk

- According to the American Cancer Society, radiation exposure, especially during childhood, significantly increases the risk of thyroid cancer. The risk depends on the type, dose, and age at exposure.
- Types of Radiation Exposure:

1. Medical Treatments
    - Radiation therapy to the head or neck during childhood poses the highest risk.
    - Risk increases with higher doses and younger age at exposure.
2. Imaging Tests
    - X-rays and CT scans also emit radiation but at much lower levels.
    - These carry a very low risk, especially in children.
3. Radiation Fallout
    - Exposure to radioactive fallout from nuclear accidents (e.g., Chernobyl) increases risk.



📌 Supported by multiple studies affirming these findings.

Impact of Some other Factors on Thyroid Cancer Risk

TSH Level

- **Numerous studies have shown that higher TSH levels—even within the normal range—are strongly associated with an increased risk of thyroid cancer, especially in individuals with thyroid nodules.**

- **T3 (Triiodothyronine)** is biologically active, but current evidence does not consistently support it as an independent risk factor for thyroid cancer.
- **Thyroxine (T4) levels** are still under investigation, with some evidence suggesting they may contribute to the overall risk profile for thyroid cancer.
- While **larger thyroid nodules** may be associated with **a higher risk of malignancy** in some studies, findings are not consistent; clinical assessment typically considers **nodule size alongside other diagnostic features** for accurate risk evaluation.
- Thyroid cancer can occur at **any age**, but is **most common between ages 30–60**, with **earlier peak risk in women** than men for reasons not yet fully understood.
- **Diabetes** has been linked to thyroid cancer in some studies, but **current evidence is inconclusive**; more research is needed to determine its role as a risk factor.

ThyroPredict

ThyroPredict | Thyroid Tumor

127.0.0.1:5000/predict#predict

coursera courses | css assignment

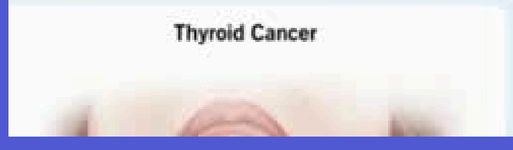
ThyroPredict

Home About Predict Treatments Contact

Empowering Informed Decisions: Thyroid Tumor Risk Prediction

Gain clarity on your thyroid health. Our predictive tool helps assess the nature of your tumor, supporting early and effective decision-making.

Start Assessment



Summary

- Majority of cases fell under **Bethesda Categories III & IV** — the "**Gray Zone**" with ambiguous presentations.
- **SMOTE** resampling was applied to address class imbalance in the dataset.
- **Exploratory Data Analysis (EDA)** revealed:
  - **No strong relationship** between diagnosis and age, gender, TSH level, or nodule size.
  - **Demographic factors** showed a notable association with malignancy.
- Multiple models were tested; **XGBoost** outperformed others with 80% test accuracy.
- **SHAP interpretation** identified **country and ethnicity** as top predictors pushing towards malignancy, while **diabetes, obesity, and smoking** leaned toward benign outcomes.
- Real-world insights supported and explained the model's behavior through literature review.
- **ThyroPredict** — a user-friendly web app — was developed to predict thyroid malignancy in **Gray Zone** cases.

References

- Cancer Research UK - Risks and Causes of Thyroid Cancer
  - https://www.cancerresearchuk.org/about-cancer/thyroid-cancer/causes-risks

- MAYO Clinic - Thyroid Cancer
  - https://www.mayoclinic.org/diseases-conditions/thyroid-cancer/symptoms-causes/syc-20354161

- American Cancer Society - Causes, Risk Factors, and Prevention
  - https://www.cancer.org/cancer/types/thyroid-cancer/causes-risks-prevention

- Clark, E., Price, S., Lucena, T., Haberlein, B., Wahbeh, A., & Seetan, R. (2024). Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach. Knowledge, 4, 557–570. https://doi.org/10.3390/knowledge4040029.

- Salman, K., & Sonuç, E. (2021). Comparative Analysis of Machine Learning Models for Thyroid Cancer Recurrence Prediction. J. Phys.: Conf. Ser., 1963, 012140

- Iqbal & Shahzad (2024). Comparative Analysis of ML Models for Thyroid Cancer Recurrence. PJNMed

- Bongiovanni et al. (2012). The Bethesda System for Reporting Thyroid Cytopathology: A meta-analysis. Acta Cytologica.
- Lee et al. (2020). Risk stratification of indeterminate thyroid nodules using non-invasive data. Journal of Endocrinological Investigation

- Feature Selection - Mutual Information
  - https://medium.com/@miramnair/feature-selection-mutual-information-a0def943e1ed
- Canadian Cancer Society - Risk Factors for Thyroid Cancer

- Cibas ES, Ali SZ. The Bethesda System for Reporting Thyroid Cytopathology. Springer (2017)
- Haugen et al., 2015 ATA Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer.

- Scikit-learn Documentation – GridSearchCV - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- Cibas, E. S., & Ali, S. Z. (2017)-The Bethesda System for Reporting Thyroid Cytopathology (2nd ed.). Springer
- https://www.thyroid.org/patient-thyroid-information/ct-for-patients/april-2024/vol-17-issue-4-p-11-12/
- https://www.sciencedirect.com/science/article/pii/S0720048X25001354
- https://www.researchgate.net/publication/385996218_Predictive_Analytics_for_Thyroid_Cancer_Recurrence_A_Machine_Learning_Approach

References

- Bao WQ, Zi H, Yuan QQ, Li LY, Deng T. (2021). Global burden of thyroid cancer and its attributable risk factors in 204 countries and territories from 1990 to 2019. Thorac Cancer, 12(18), 2494–2503. https://doi.org/10.1111/1759-7714.14099
- Aschebrook-Kilfoy B, Kaplan E, Chiu BC, et al. (2015). The association between race/ethnicity and the prevalence of thyroid cancer. Cancer Causes Control, 26(3), 383–386. https://doi.org/10.1007/s10552-015-0525-4
- Guth S, Theune U, Aberle J, Galach A, Bamberger CM. (2015). Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Thyroid Res, 8, 3. https://doi.org/10.1186/s13044-015-0020-8
- Ranasinghe I, Ranasinghe N. (2018). Incidence, mortality and risk factors of thyroid cancer in the world: a review. WCRJ, 5:e1093. https://www.wcrj.net/wp-content/uploads/sites/5/2018/06/e1093-Incidence-mortality-and-risk-factors-of-thyroid-cancer-in-the-world-a-review.pdf
- Rogel Cancer Center. Thyroid cancer risk factors and symptoms. https://www.rogelcancercenter.org/thyroid-cancer/learn-about-thyroid-cancer/risk-factors-and-symptoms
- Stojanovic M, Milosevic Z, Milosevic V, et al. (2020). Thyroid cancer incidence and overdiagnosis. Genes, 11(10), 1039. https://doi.org/10.3390/genes11101039
- Canadian Cancer Society. Thyroid Cancer Risks. https://cancer.ca/en/cancer-information/cancer-types/thyroid/risks
- Chen J, Wang C, Shao B. (2023). Global, regional, and national thyroid cancer age-period-cohort modeling and Bayesian predictive modeling studies. Heliyon, 9(11), e22490. https://doi.org/10.1016/j.heliyon.2023.e22490
- Peterson L, Soliman A, Ruterbusch JJ, Smith N, Schwartz K. (2011). Comparison of exposures among Arab American and non-Hispanic White female thyroid cancer cases in metropolitan Detroit. J Immigr Minor Health, 13(6), 1033–1040. https://doi.org/10.1007/s10903-011-9485-2
- Sanabria A, Kowalski LP, Shah JP, et al. (2018). Growing incidence of thyroid carcinoma in recent years: Factors underlying overdiagnosis. Head Neck, 40(4), 855–866. https://doi.org/10.1002/hed.25029
- American Cancer Society. Thyroid Cancer – Risk Factors. https://www.cancer.org/cancer/types/thyroid-cancer/causes-risks-prevention/risk-factors.html
- Zhang Q, Li H, Li L, et al. (2021). Role of estrogen in thyroid cancer: Molecular mechanisms and potential therapeutic targets. Front Endocrinol (Lausanne), 12:738213. https://doi.org/10.3389/fendo.2021.738213
- Delange F, Lecomte P, et al. (2014). Iodine deficiency as a cause of thyroid cancer: A review. https://thyroidresearchjournal.biomedcentral.com/articles/10.1186/s13044-015-0020-8
- Liu Y, Su L, et al. (2015). Thyroid nodules and risk factors for malignancy in patients with thyroid nodules. Oncol Lett, 10(3), 1479–1483. https://doi.org/10.3892/ol.2015.3417
- Qian D, Chen J, et al. (2021). Risk factors of thyroid cancer: insights from the Global Burden of Disease Study 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC8527095/
- Alqahtani S, Almalki N, et al. (2023). The prevalence and risk factors of thyroid nodules in the general population: A review. https://pmc.ncbi.nlm.nih.gov/articles/PMC10311569/

References

- Thyroid Cancer. Top 4 Reasons to Worry About Thyroid Nodules. https://www.thyroidcancer.com/blog/top-4-reasons-to-worry-about-thyroid-nodules#:~:text=%233%20Reason%20to%20Worry%20About,evaluation%20of%20your%20thyroid%20nodule
- Thyroid Cancer. Thyroid Nodule Size. https://www.thyroidcancer.com/blog/thyroid-nodule-size
- Columbia Surgery. Thyroid Cancer Diagnosis. https://columbiasurgery.org/thyroid/thyroid-cancer-diagnosis#:~:text=Most%20patients%20with%20thyroid%20cancer,seen%20in%20medullary%20thyroid%20cancer
- American Cancer Society. Thyroid Cancer Diagnosis and Staging. https://www.cancer.org/cancer/types/thyroid-cancer/detection-diagnosis-staging/how-diagnosed.html#:~:text=T3%20and%20T4%20are%20the,in%20people%20with%20thyroid%20cancer
- WebMD. What Is the T3 Test? https://www.webmd.com/a-to-z-guides/what-is-t3-test
- Nair, M. (2020). Feature Selection: Mutual Information. https://medium.com/@miramnair/feature-selection-mutual-information-a0def943e1ed
- IBM. Feature Engineering. https://www.ibm.com/think/topics/feature-engineering
- Murphy, P. J. (2019). Running UMAP for Data Visualization in R. https://rpubs.com/pjmurphy/758265
- Chelaru, N. (2020). FAMD in R for Data Analysis. https://rpubs.com/nchelaru/famd
- PubMed Central. Study on UMAP for Data Visualization. https://pmc.ncbi.nlm.nih.gov/articles/PMC9806162/
- R-Bloggers. Running UMAP for Data Visualization in R. https://www.r-bloggers.com/2019/06/running-umap-for-data-visualisation-in-r/

Thank You!