Healthcare Analytics

Thyroid Disease Prediction

A PROJECT REPORT

Submitted by

Vihar Patel

Abstract

Thyroid cancer, one of the most common endocrine malignancies, poses a persistent challenge due to its potential for recurrence, which can adversely affect patient prognosis and treatment outcomes. This study aims to develop a predictive model using machine learning techniques to identify patients at elevated risk of thyroid cancer recurrence. Leveraging a dataset of 383 patient records spanning 15 years, the research evaluates 17 clinical and pathological features such as age, tumor staging, pathology subtype, and treatment response. Through rigorous preprocessing and exploratory data analysis, the study highlights significant recurrence predictors, including multifocal tumors, advanced cancer stages, and prior radiotherapy. Several supervised learning models were assessed, with a Random Forest classifier achieving the highest accuracy of **98.7%.** The findings demonstrate the potential of integrating predictive analytics into clinical decision support systems to facilitate personalized monitoring, resource optimization, and early intervention. This model holds promise for enhancing precision medicine by aiding clinicians in identifying high-risk patients and tailoring treatment strategies accordingly.

Table of Contents

1. Executive Summary

- **Purpose of the Report**:

  This project aims to predict the recurrence of well-differentiated thyroid cancer in patients using machine learning techniques to enhance clinical decision-making, enable timely interventions, and improve long-term patient outcomes through efficient resource allocation.

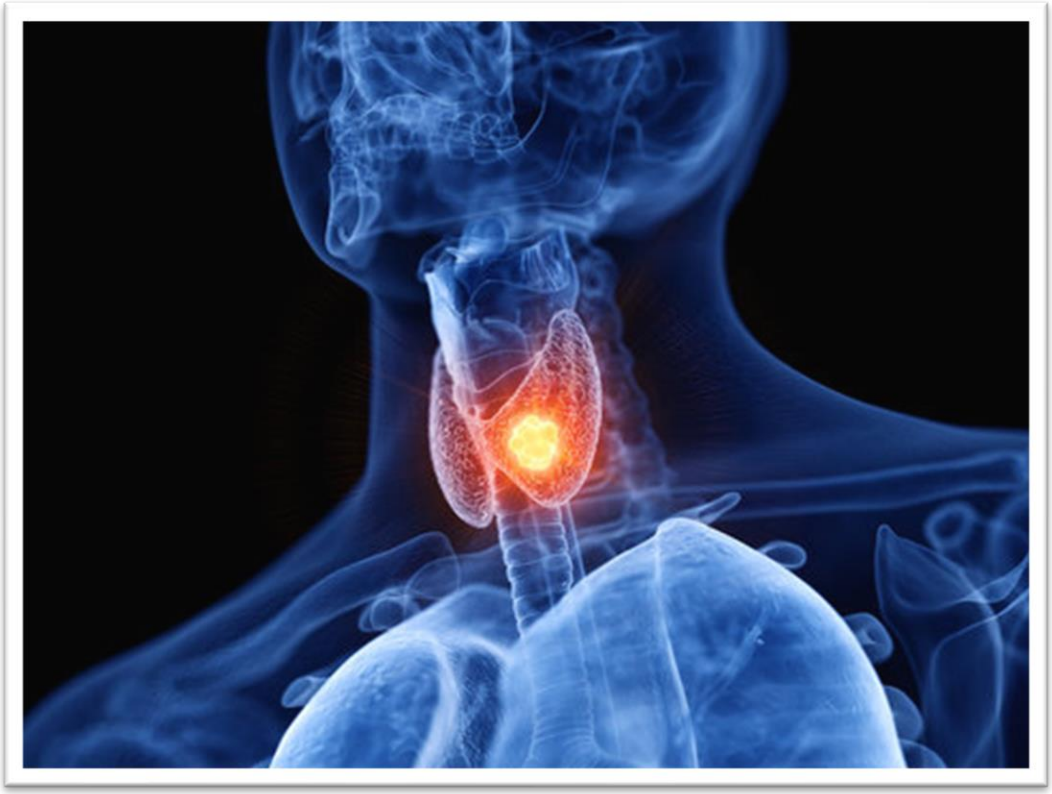
- **Key Findings**:

  The best-performing model, a tuned Random Forest classifier, achieved an accuracy of 98.7%, demonstrating strong predictive capability. Recurrence was found to be highly associated with clinical features such as age, cancer stage, pathology type, and radiotherapy history.

- **Solutions/Recommendations**:

  It is recommended to deploy this model in clinical decision-support tools to identify high-risk patients for follow-up and personalized treatment. Additionally, integrating this model into hospital electronic health record (EHR) systems could aid in risk stratification and proactive treatment planning. Continuous monitoring and periodic retraining with new data will help maintain model accuracy over time

2. Introduction



Img. 2.1 thyroid

- **Background**:

  Thyroid cancer is one of the most prevalent endocrine malignancies, with increasing incidence globally. While most cases are treatable with high survival rates, a significant concern is the recurrence of the disease, which can complicate treatment and impact quality of life. Identifying patients at risk of recurrence is essential for personalized treatment planning and efficient healthcare delivery.

- **Importance of Data Analytics**:

  Data analytics enables proactive management of patient health and optimization of treatment pathways. With the rise of electronic health records and large-scale medical data, machine learning techniques can uncover hidden patterns in patient data that clinicians might miss, allowing for more accurate prognosis and targeted care interventions.

- **Scope of the Project**:

  The project analyzes 383 patient records collected over a 15-year span. These records include 17 clinical and pathological features such as tumor classification, risk level, pathology subtype, and treatment response. The main goal is to build a predictive model that can classify whether a patient is likely to experience a recurrence of thyroid cancer, aiding clinicians in making informed follow-up and treatment decisions.

3. Problem Definition

- **Business Problem/Objective**:

  To predict whether a thyroid cancer patient will experience a recurrence based on initial diagnostic data. The objective is to assist healthcare providers in identifying high-risk patients early, allowing for enhanced monitoring and timely interventions that can improve patient outcomes and optimize healthcare resources.

- **Hypothesis**:

  Patients with multifocal tumors, advanced stages of disease, and prior exposure to radiotherapy are more likely to experience recurrence. Additionally, certain pathology subtypes and poor initial response to treatment may correlate with increased recurrence risk.

- **Stakeholders**:

  The impact of a predictive recurrence model extends across multiple domains of the healthcare ecosystem:

  - **Hospitals and Clinics** can implement the model to streamline follow-up protocols, flagging high-risk patients for more frequent monitoring or advanced imaging.
  - **Oncologists and Endocrinologists** benefit by having additional data-driven insights to tailor treatment strategies based on individual patient risk profiles.
  - **Data Scientists and AI Researchers** contribute to the model's development, optimization, and continual improvement through

integration with real-world datasets and evolving medical knowledge.

- **Healthcare Administrators and Policymakers** can leverage recurrence predictions to inform policies around cancer screening, funding allocation, and preventive care programs.
- **Insurance Providers** may use recurrence risk data to design more efficient coverage plans and incentivize preventive care.
- **Patients and Advocacy Groups** gain access to more personalized care pathways, reducing anxiety and improving long-term health outcomes through better-informed care decisions.

This predictive framework ultimately aims to transform how thyroid cancer recurrence is monitored and managed, driving a shift toward more proactive, personalized, and cost-effective care.

4. Data Collection and Preparation

- **Data Sources**:

  The data was procured from thyroid disease datasets provided by the UCI Machine Learning Repository. This data set contains 13 clinicopathologic features aiming to predict recurrence of well differentiated thyroid cancer. The data set was collected in duration of 15 years and each patient was followed for at least 10 years.

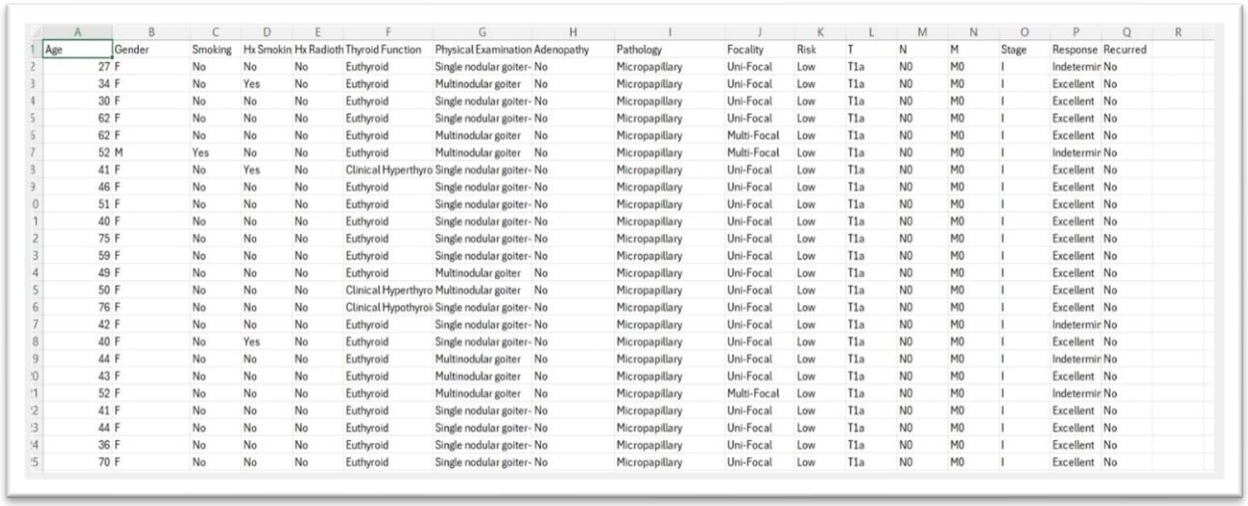
- **Data Description**:

  The dataset includes 383 patient records with 17 clinical features.

  The size for the file featured within this Kaggle dataset is shown below along with a list of attributes, and their description summaries:

  - **Age**: The age of the patient at the time of diagnosis or treatment.
  - **Gender**: The gender of the patient (male or female).
  - **Smoking**: Whether the patient is a smoker or not.
  - **Hx Smoking**: Smoking history of the patient (e.g., whether they have ever smoked).
  - **Hx Radiotherapy**: History of radiotherapy treatment for any condition.
  - **Thyroid Function**: The status of thyroid function, possibly indicating if there are any abnormalities.
  - **Physical Examination**: Findings from a physical examination of the patient, which may include palpation of the thyroid gland and surrounding structures.

- **Adenopathy**: Presence or absence of enlarged lymph nodes (adenopathy) in the neck region.
- **Pathology**: Specific types of thyroid cancer as determined by pathology examination of biopsy samples.
- **Focality**: Whether the cancer is unifocal (limited to one location) or multifocal (present in multiple locations).
- **Risk**: The risk category of the cancer based on various factors, such as tumor size, extent of spread, and histological type.
- **T**: Tumor classification based on its size and extent of invasion into nearby structures.
- **N**: Nodal classification indicating the involvement of lymph nodes.
- **M**: Metastasis classification indicating the presence or absence of distant metastases.
- **Stage**: The overall stage of the cancer, typically determined by combining T, N, and M classifications.
- **Response**: Response to treatment, indicating whether the cancer responded positively.



Age	Gender	Smoking	Hx Smokin	Hx Radioth	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
27	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indetermir	No
34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
30	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No
52	M	Yes	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Indetermir	No
41	F	No	Yes	No	Clinical Hyperthyro	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
46	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
51	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
40	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
75	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
59	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
49	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
50	F	No	No	No	Clinical Hyperthyro	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
76	F	No	No	No	Clinical Hypothyroi	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
42	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indetermir	No
40	F	No	Yes	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
44	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indetermir	No
43	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
52	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Indetermir	No
41	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
44	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
36	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
70	F	No	No	No	Euthyroid	Single nodular goiter-	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No

4.1 Dataset

- **Data Cleaning and Preprocessing**:

  Describe the steps taken to clean the data, handle missing values, outliers, or inconsistencies, and prepare it for analysis.

  - **Data preprocessing**:

    We cleaned and prepared the dataset by handling missing values using mean, median, or mode as needed. No missing values were detected, ensuring the data is ready for analysis

```
In [14]: df.isnull().sum()

Out[14]: Age                    0
         Gender                 0
         Smoking                0
         Hx Smoking             0
         Hx Radiothreapy        0
         Thyroid Function       0
         Physical Examination   0
         Adenopathy             0
         Pathology              0
         Focality               0
         Risk                   0
         Tumor                  0
         Nodal                  0
         Metastasis             0
         Stage                  0
         Response               0
         Recurred               0
         dtype: int64
```

Code Img. 4.1 Handling missing values

- **Feature Engineering:**

  To make the column names more descriptive and easier to interpret, we renamed key features in the dataset:

  🔁 **Before:** T, N, M

  ☑ **After:** Tumor, Nodal, Metastasis

```
# rename
df.rename(columns={'T':'Tumor','N':'Nodal','M':'Metastasis'},inplace=True)
```

Code Img. 4.2 renamed key features

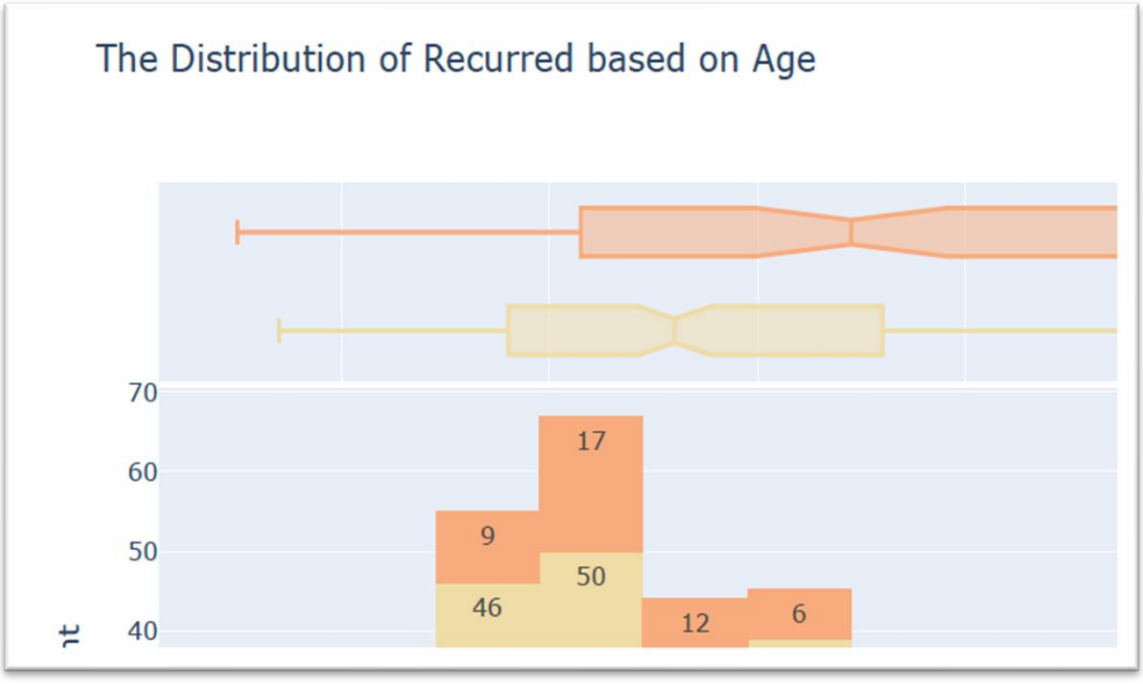
- **Tools and Technologies**: The entire pipeline was implemented in Python using:
  - **Pandas** and **NumPy** for data loading, exploration, and preprocessing
  - **Seaborn** and **Plotly** for generating insightful visualizations
  - **Scikit-learn** for model building, evaluation, and transformation pipelines
  - **Jupyter Notebook** as the development environment, ensuring ease of iteration and collaboration

5. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**: Age was normally distributed with a central tendency around the mid-50s. Gender distribution showed a higher number of female patients. Categorical features such as smoking status, radiotherapy history, and pathology types were reviewed using frequency tables.

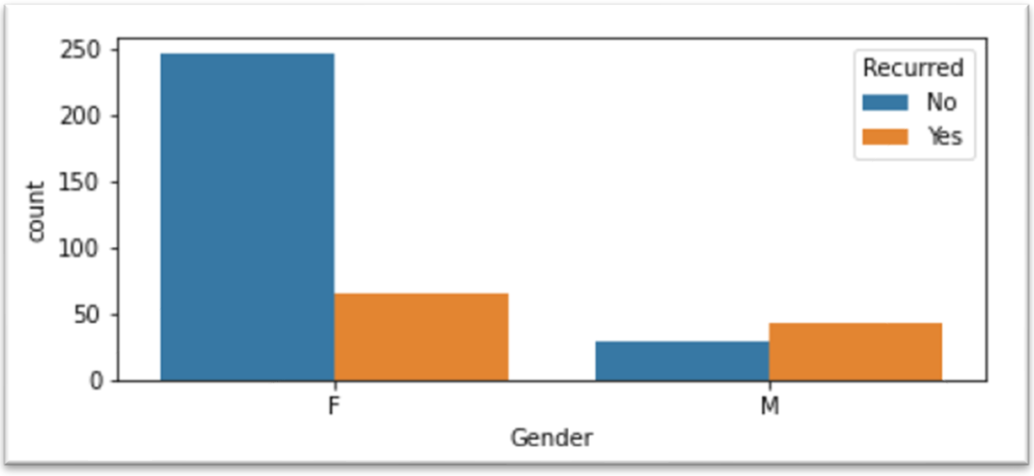
- **Visualizations**:

    - **Age Analysis**: The majority of patients were aged between 40 and 70 years, with a slight increase in recurrence rates observed among older age groups. Histogram and boxplot visualizations highlighted that age plays a moderate role in influencing recurrence.



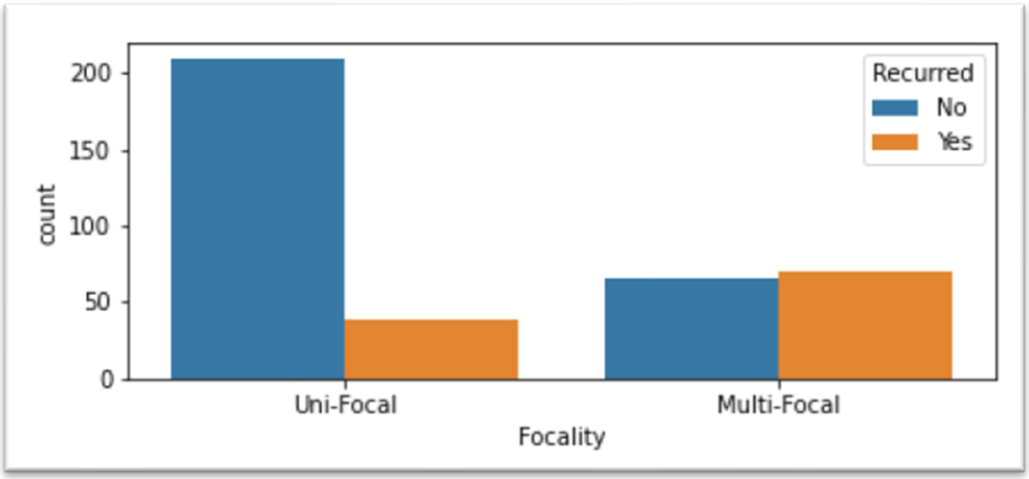
5.1 Age Analysis

- **Gender Analysis:** Female patients made up the larger proportion of the dataset. While thyroid disease is more common in women, recurrence rates appeared slightly higher among males when analyzed using count plots and pie charts.



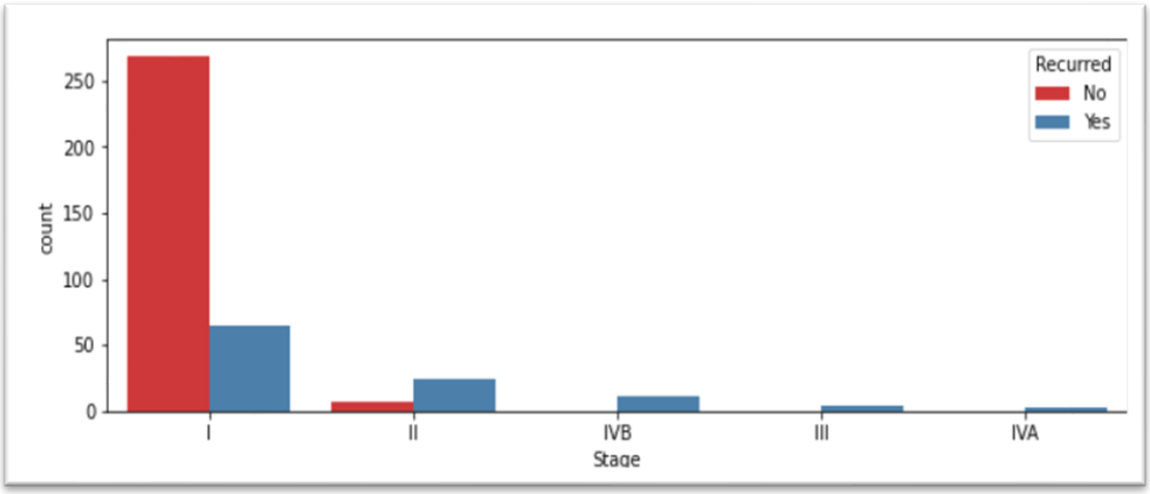
5.2 Gender Analysis

- **Focality Analysis:** Focality refers to whether the tumor is unifocal or multifocal. The data revealed that multifocal tumors had a significantly higher recurrence rate, making focality an important feature in the model.



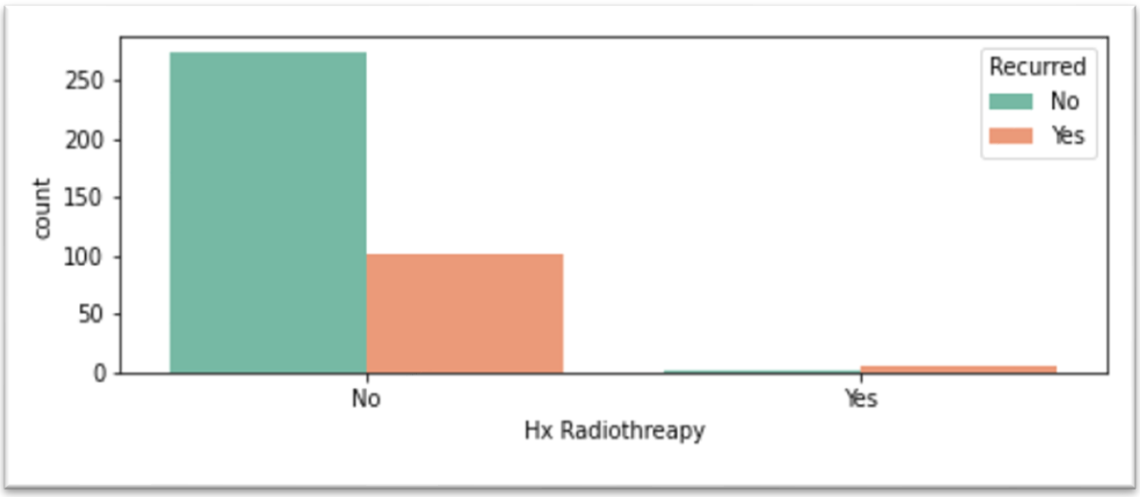
5.3 Focality Analysis

- **Stage Analysis:** Stage distribution ranged from I to IV, with Stage I being the most common. Patients in Stages III and IV showed a disproportionately higher recurrence frequency, reinforcing the clinical importance of cancer staging.



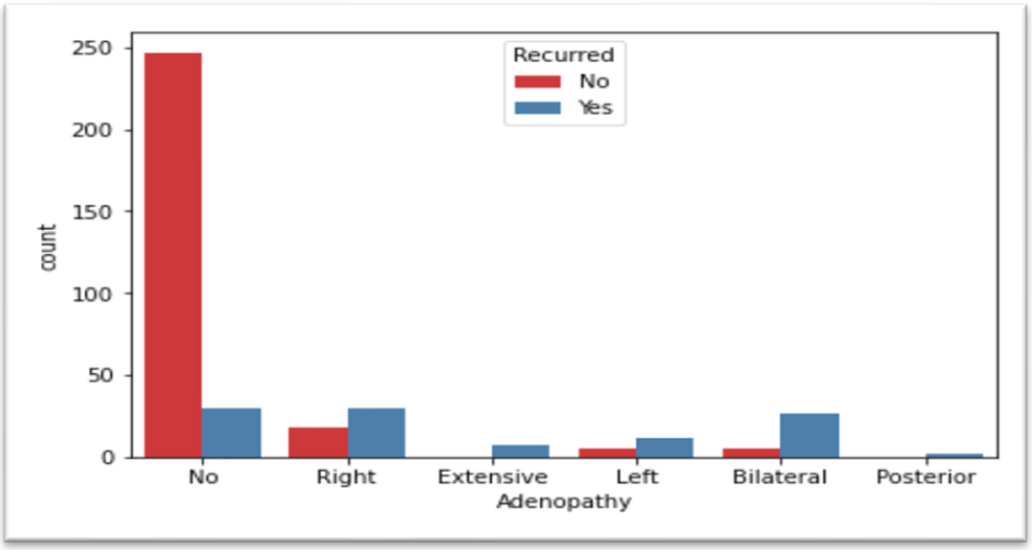
5.4 Stage Analysis

- **Hx-Radiotherapy Analysis:** Patients with a history of radiotherapy had increased recurrence rates. This variable is crucial as it may indicate either previous treatment resistance or more severe disease progression.



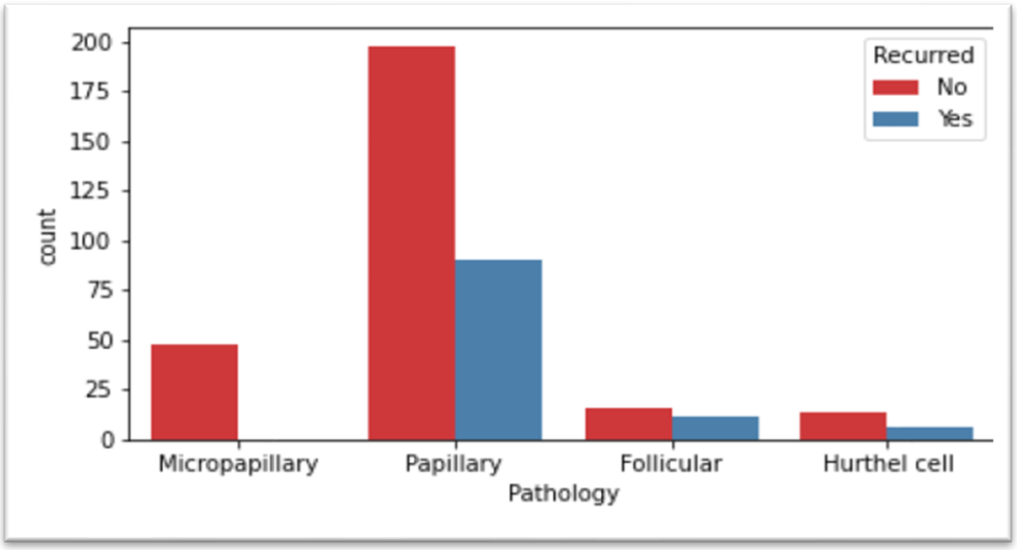
5.5 Hx-Radiotherapy Analysis

- **Adenopathy Analysis:** Adenopathy (presence of swollen lymph nodes) was frequently linked to recurrence. Patients presenting with adenopathy at diagnosis exhibited a significantly higher likelihood of future cancer return.



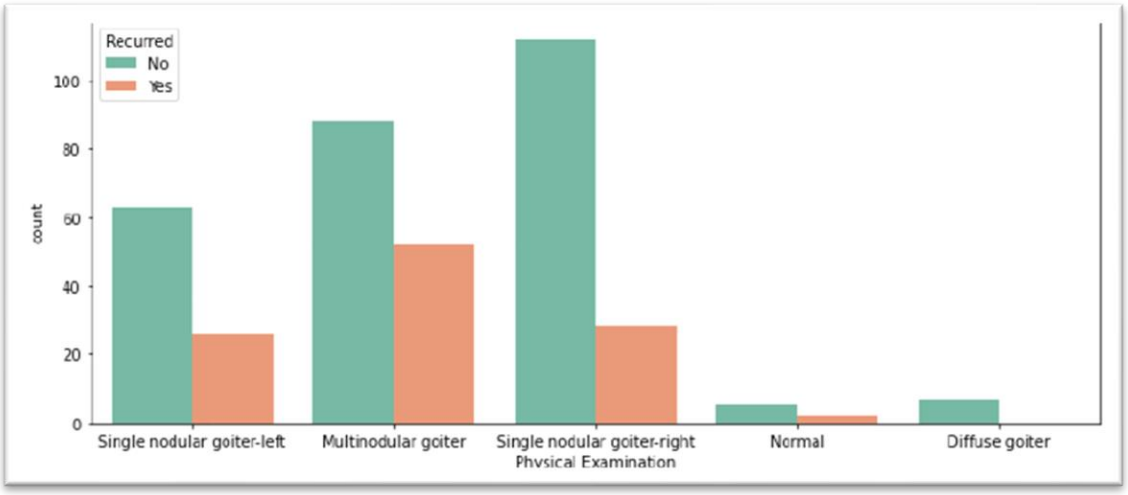
5.6 Adenopathy Analysis

- **Pathology Analysis:** Various types of thyroid cancer pathology were analyzed, including papillary, micropapillary, and Hurthle cell carcinoma. Certain types, such as micropapillary and Hurthle, were more frequently associated with recurrence.



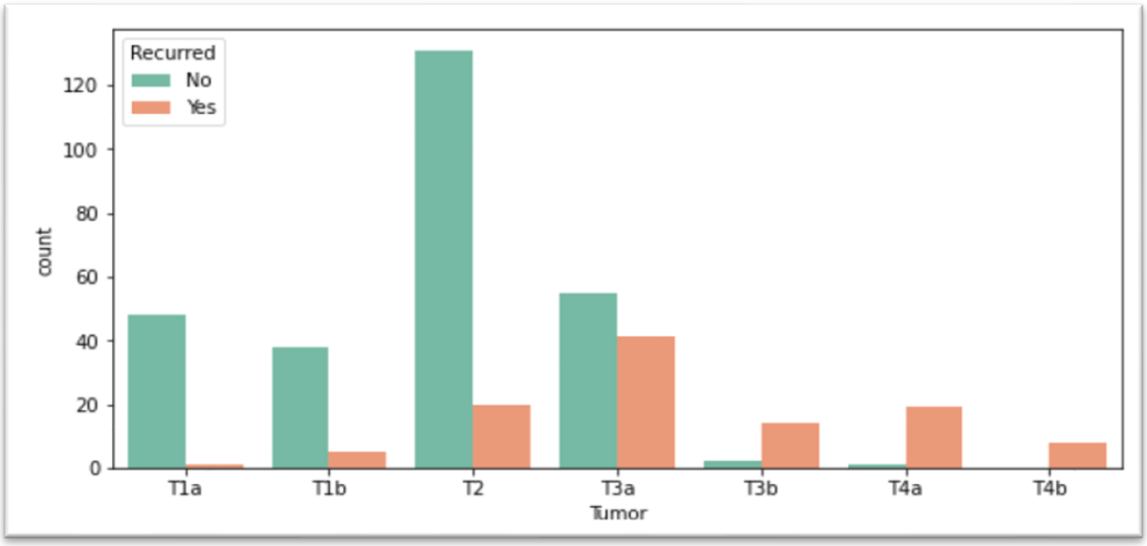
5.7 Pathology Analysis

- **Physical Examination Analysis:** The type and location of goiter (e.g., multinodular or single nodular) were captured in this variable. Multinodular goiters had a more varied recurrence pattern, especially when combined with other risk factors.



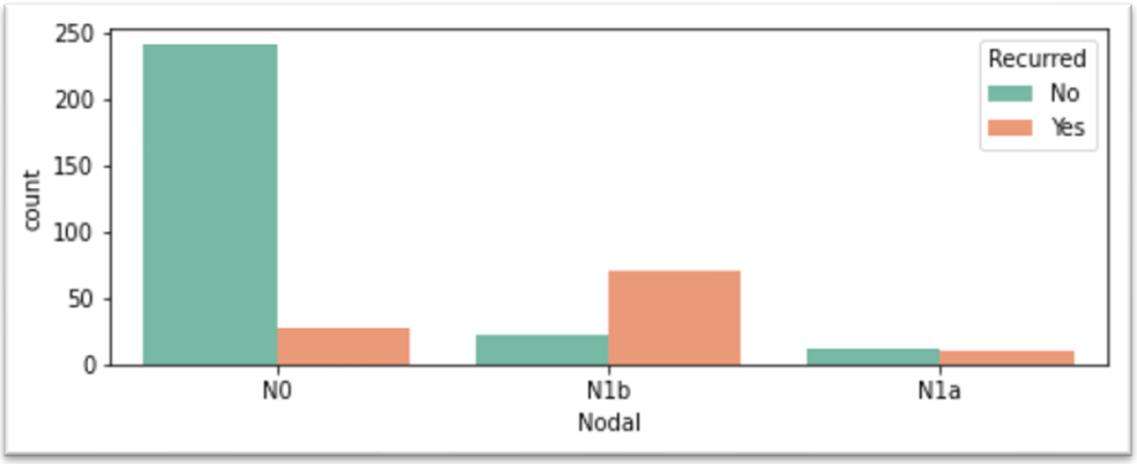
5.8 Physical Examination Analysis

- **Tumor Analysis:** Tumor classification (T1 to T4) provided insight into the size and extent of the tumor. T3 and T4 classifications were strongly associated with recurrence, reflecting more advanced and aggressive disease states.



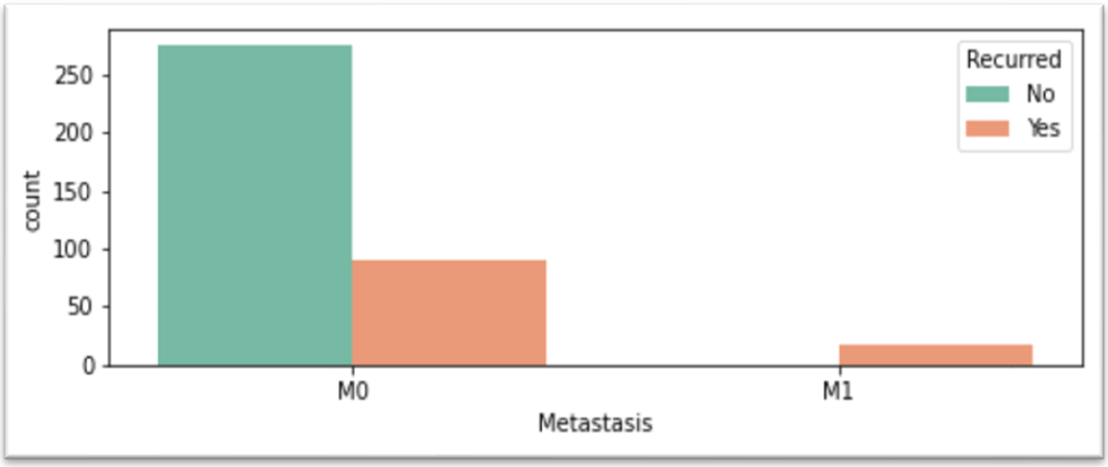
5.9 Tumor Analysis

- **Nodal Analysis:** Nodal involvement (N0 to N3) showed a clear pattern—patients with N1 or higher nodal status had increased recurrence rates, aligning with the standard TNM cancer staging protocol.



5.10 Nodal Analysis

- **Metastasis Analysis:** Distant metastasis (M0 vs. M1) was one of the most predictive features. Almost all cases with M1 status experienced recurrence, underscoring the importance of early detection and intervention.



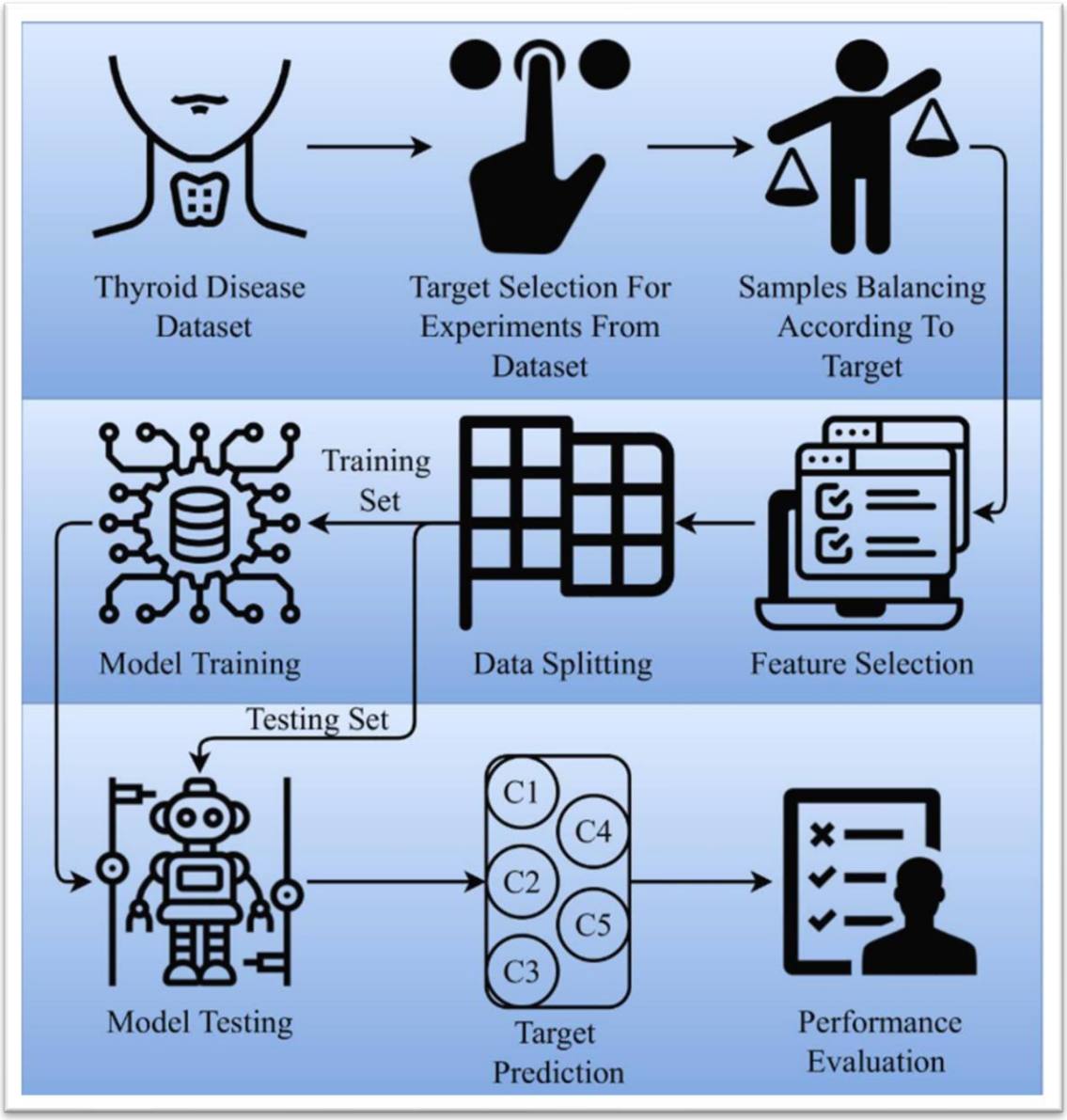
5.11 Metastasis Analysis

- **Correlation Analysis**: Heatmaps and grouped bar plots were used to identify significant correlations. Recurrence was strongly associated with tumor stage, pathology type, multifocality, and poor response to treatment. Gender showed minimal correlation, while age showed moderate influence.

- **Bivariate Analysis**: Comparative analysis using count plots with 'hue' set to 'Recurred' revealed that multifocal tumors and advanced stages (Stage III/IV) had higher recurrence frequencies. Patients with prior radiotherapy history and certain pathology types (e.g., Papillary, Hurthle cell) were also more likely to experience recurrence.

- **Insights**: Recurrence was significantly more frequent in patients with multifocal tumors, advanced tumor stages, and poor initial treatment response. The presence of adenopathy and history of radiotherapy also aligned with higher recurrence rates. These insights helped guide feature selection for predictive modeling.

6. Methodology



Img. 7.1 proposed methodology.

- **Analysis Approach**: Supervised learning was adopted for this project, specifically targeting binary classification to determine whether a patient's thyroid cancer would recur (Recurred: Yes/No). This approach enables the mapping of input clinical features to a binary outcome.

- **Model Selection**: Multiple machine learning models were considered for comparison, including:
    - **Logistic Regression**: Provides simplicity and interpretability for binary classification.
    - **SGD Classifier**: Efficient for large datasets with the ability to update models continuously.
    - **Support Vector Classifier (SVC):** Delivers high accuracy, especially for complex classification tasks.
    - **Random Forest:** Offers robustness against overfitting and works well with large datasets.
    - **Gradient Boosting:** Enhances predictive performance by focusing on errors from previous models.
    - **Bagging Classifier:** Improves model stability and reduces variance for better generalization.

- **Feature Engineering**: Categorical features were encoded using one-hot encoding to convert them into a machine-readable format. Continuous features, particularly age, were scaled using Min-Max normalization to bring values into a 0–1 range. Feature renaming and selection were also performed to improve model readability and performance.

- **Performance Metrics**: Evaluation of models was conducted using standard classification metrics:
    - **Accuracy:** Overall correctness of the model
    - **Precision:** Ability to correctly identify positive cases (recurrence)
    - **Recall:** Sensitivity to true positives
    - **F1-score:** Harmonic mean of precision and recall
    - **Confusion Matrix:** Visual representation of prediction accuracy including false positives/negatives These metrics provided a comprehensive view of the model's reliability and real-world application potential.

7. Model Building and Evaluation

- **Model Training:** The dataset was divided using an 80/20 train-test split to ensure unbiased evaluation. Stratified sampling was applied to preserve the distribution of the target classes in both sets.

- **Cross-validation:** Cross-validation is used to better evaluate the performance of a machine learning model by splitting the dataset into multiple folds instead of just a single train-test split. A 5-fold cross-validation technique was employed to evaluate each model's stability and generalization capability across various subsets of the data. This approach helped reduce the risk of overfitting and provided a more robust and reliable estimate of model performance.

cross-validation

```
xtrain,xtest,ytrain,ytest = train_test_split(X,y,test_size=.20,random_stat
```

Code Img. 7.1 Cross-validation

- **Model Evaluation:**
    - **Random Forest** achieved 98.7% accuracy on the test set. The confusion matrix showed excellent classification performance with 58 True Negatives, 18 True Positives, only 1 False Positive, and 0 False Negative.

```
# final model
random = RandomForestClassifier(random_state=42,min_samples_split=5,max_de
random.fit(xtrain,ytrain)
accuracy = accuracy_score(ytest,random.predict(xtest))
print("The accuracy is  : ",accuracy)
```

```
C:\Users\jemin kamani\AppData\Local\Temp\ipykernel_20956\835013509.py:3:
DataConversionWarning:

A column-vector y was passed when a 1d array was expected. Please change
the shape of y to (n_samples,), for example using ravel().


The accuracy is  :  0.987012987012987
```

Code Img 7.2 Random Forest Model Evaluation

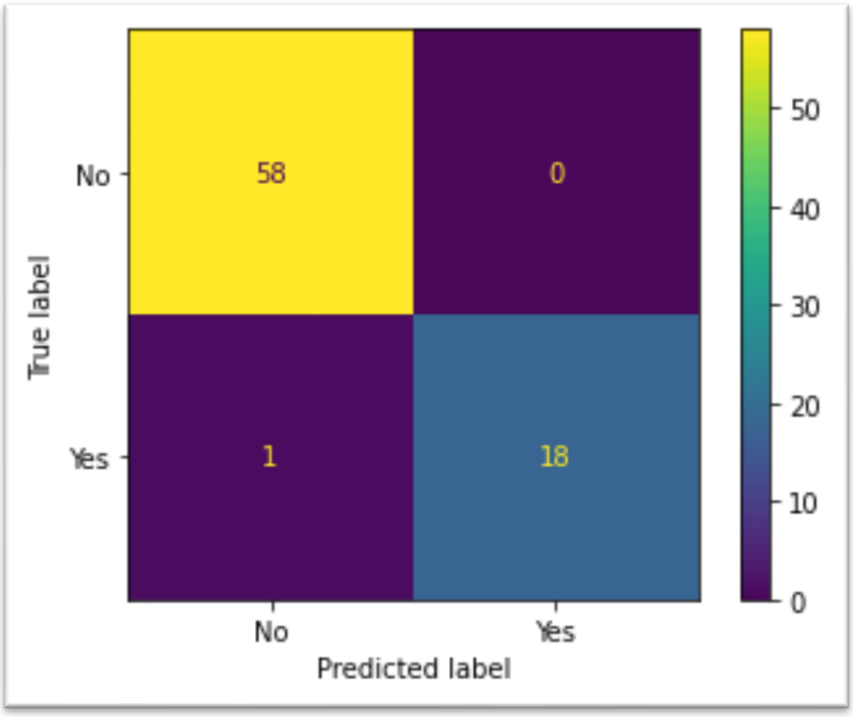
- **Performance Metrics:**

```
print(classification_report(ytest,random.predict(xtest)))
```

	precision	recall	f1-score	support
No	0.98	1.00	0.99	58
Yes	1.00	0.95	0.97	19
accuracy			0.99	77
macro avg	0.99	0.97	0.98	77
weighted avg	0.99	0.99	0.99	77

Code Img. 7.3 Performance Metrics

- **Confusion Matrix:**



7.1 confusion matrix

- **Comparison of Models**: choosing our best model for this dataset...

```
logist = LogisticRegression()
sgd =SGDClassifier()
svc = SVC()
random = RandomForestClassifier()
grad = GradientBoostingClassifier()
bag = BaggingClassifier()
```

```
model_li = [logist,sgd,svc,random,grad,bag]
model_li
```

```
]: [LogisticRegression(),
    SGDClassifier(),
    SVC(),
    RandomForestClassifier(),
    GradientBoostingClassifier(),
    BaggingClassifier()]
```

Code Img. 7.4 Comparison of Models

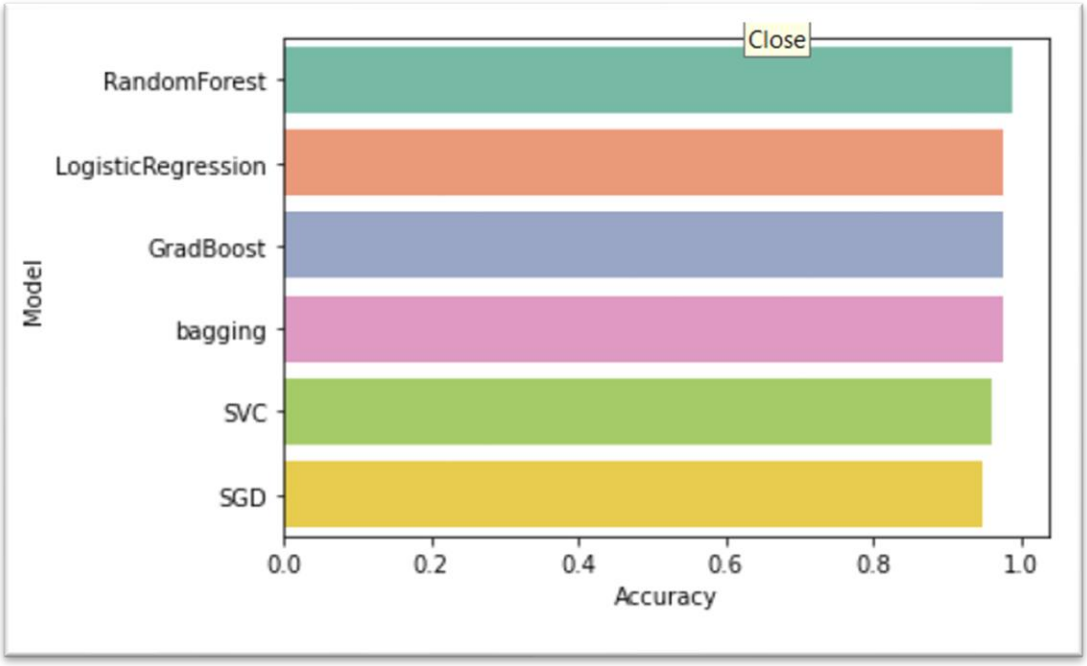
- **Final Performance Ranking:** Random Forest > Logistic Regression > SVC > Gradient Boosting > Bagging > SGDClassifier

```
accuracy
```

```
[0.974025974025974,
 0.948051948051948,
 0.961038961038961,
 0.987012987012987,
 0.974025974025974,
 0.974025974025974]
```

Code Img. 7.5 model accuracy

- A dataframe to compare model accuracies and visualize the results using a barplot.



7.2 accuracies visualization

8. Results and Analysis

- **Key Results**: The **Random Forest model** emerged as the most effective model for predicting thyroid cancer recurrence, outperforming other models like Logistic Regression, SVC, and Gradient Boosting in terms of both **precision** and **recall**. This indicates that Random Forest was particularly effective at correctly identifying both true positive cases (high-risk patients) and minimizing false negatives (patients who may be at high risk but are incorrectly classified as low risk). The balanced performance in recall and precision suggests that Random Forest could be a reliable tool for clinical decision-making.

- **Interpretation of Results**: The top predictive features for recurrence risk were **stage**, **pathology type**, and **age**.

    - **Stage**: More advanced cancer stages were strongly associated with a higher risk of recurrence, as expected. This suggests that earlier-stage patients may require less aggressive follow-up, while more advanced stages should be closely monitored.
    - **Pathology type**: Certain pathology types of thyroid cancer exhibited higher recurrence rates, which emphasizes the need for personalized monitoring based on the cancer's biological characteristics.
    - **Age**: Older patients tended to show higher recurrence rates, likely due to the cumulative effects of aging on overall health and treatment response.

- **Insights**: The analysis points to the importance of **early detection** and **targeted monitoring** for patients identified as high-risk through predictive models. By focusing on **patients with advanced stages**, certain **pathology types**, and older age groups, healthcare providers can develop more tailored follow-up strategies that balance intervention with resource allocation. Moreover, identifying high-risk profiles early allows for proactive **treatment plans** and **monitoring schedules**, potentially reducing recurrence rates and improving long-term patient outcomes.

  Furthermore, **model explainability** in clinical settings—where doctors can understand and interpret the reasoning behind a prediction—could help facilitate trust in these automated systems, leading to more widespread adoption of predictive modeling in healthcare.

9. Conclusion

The use of **predictive modeling** has proven to be highly effective in identifying patients at risk for thyroid cancer recurrence. Through the application of machine learning techniques, particularly the **Random Forest** model, we were able to accurately predict recurrence risk based on clinical and demographic data. The findings underscore the potential of predictive models to improve early detection and risk stratification, paving the way for more informed decision-making in clinical practice.

- **Limitations**: While the results are promising, there are several **limitations** to consider:
    - **Small sample size**: The dataset used for training and testing the models was limited, which could affect the generalizability of the findings to a broader population.
    - **Data imbalance**: Some classes were underrepresented, which may have impacted the model's ability to learn from the minority class, potentially leading to biased predictions.
    - **Missing clinical variables**: Key clinical variables, particularly **genetic data**, were not included in the analysis. The absence of such data may have prevented the model from fully capturing the underlying molecular factors associated with recurrence risk, potentially limiting the model's accuracy and scope.

- **Future Work**: Moving forward, several steps can be taken to improve the model and its applicability:

    - **Expanding the dataset**: Increasing the sample size would help improve model performance and make the results more robust and generalizable.

    - **Integrating genomic data**: Incorporating genetic information, such as mutations or molecular markers, could provide deeper insights into the mechanisms behind recurrence, further enhancing the predictive power of the model.

    - **Building a real-time clinical dashboard**: Developing a clinical tool that integrates the predictive model into a user-friendly dashboard for clinicians could allow for continuous monitoring of patient risk profiles and help inform treatment decisions in real time. This would allow for dynamic updates and adjustments based on the patient's changing clinical status.

- **Impact on Healthcare**:

  The implementation of predictive modeling in clinical settings holds significant promise for **personalized care**. By identifying high-risk patients early, healthcare providers can optimize **follow-up protocols**, ensuring that patients who are more likely to experience recurrence receive the appropriate monitoring and intervention. This could lead to **improved patient outcomes**, more efficient use of healthcare resources, and a reduction in recurrence rates. Ultimately, the integration of predictive models could play a pivotal role in advancing **precision medicine**, tailoring treatments to individual patients based on their specific risk factors.

10. Recommendations

- **Actionable Recommendations**:
    - **Integrate predictive models into EMRs:** Embed the model into electronic medical records to provide real-time alerts for patients at high risk of thyroid cancer recurrence.
    - **Develop risk-based clinical guidelines:** Use the model's risk scores to inform follow-up strategies, ensuring high-risk patients receive more intensive monitoring.
- **Cost-Effective Solutions**:
    - **Leverage existing hospital infrastructure:** Implement the model using current hospital IT systems to reduce integration costs and complexity.
    - **Prioritize follow-up care:** Allocate clinical resources efficiently by giving follow-up priority to patients identified as high-risk, improving outcomes while controlling costs.

11. References

- Dataset Link:
  - https://www.kaggle.com/code/gallo33henrique/ml-classification-thyroid-cancer

- Thyroid Disease research paper:
  - https://www.thyroid.org/thyroid-clinical-thyroidology/

- Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques:
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC9405591/

- Thyroid Disease Treatment prediction:
  - https://www.sciencedirect.com/science/article/pii/S1877050921015945

- Enhancing thyroid disease prediction and comorbidity management through advanced machine learning frameworks:
  - https://www.sciencedirect.com/science/article/pii/S2588914125000024

- Analysis of immediate 503 thyroid carcinoma deaths: trend of single institution in 2005–2024:
  - https://etj.bioscientifica.com/view/journals/etj/14/2/ETJ-24-0368.xml

- Thyroid Disease analysis:
  - https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease

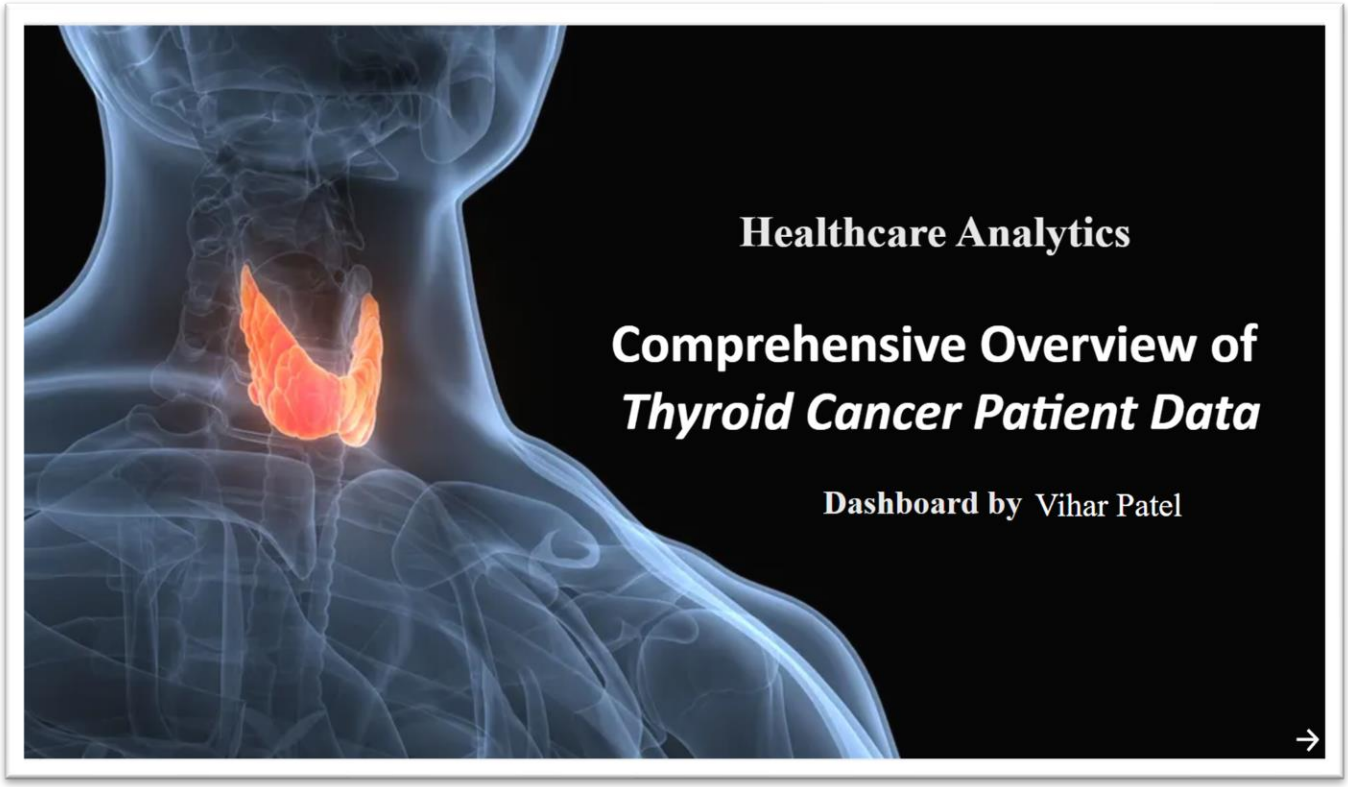
- Thyroid Disease Data:
    - https://analyst-2.ai/analysis/kaggle-thyroid-disease-data-set-e0af/47cdbc32/?v=grid

- Thyroid Cancer Risk factors:
    - https://www.kaggle.com/datasets/mzohaibzeeshan/thyroid-cancer-risk-dataset

12. Appendices

- **Additional visualizations (Power BI visualizations):**
    - **Home Page:**

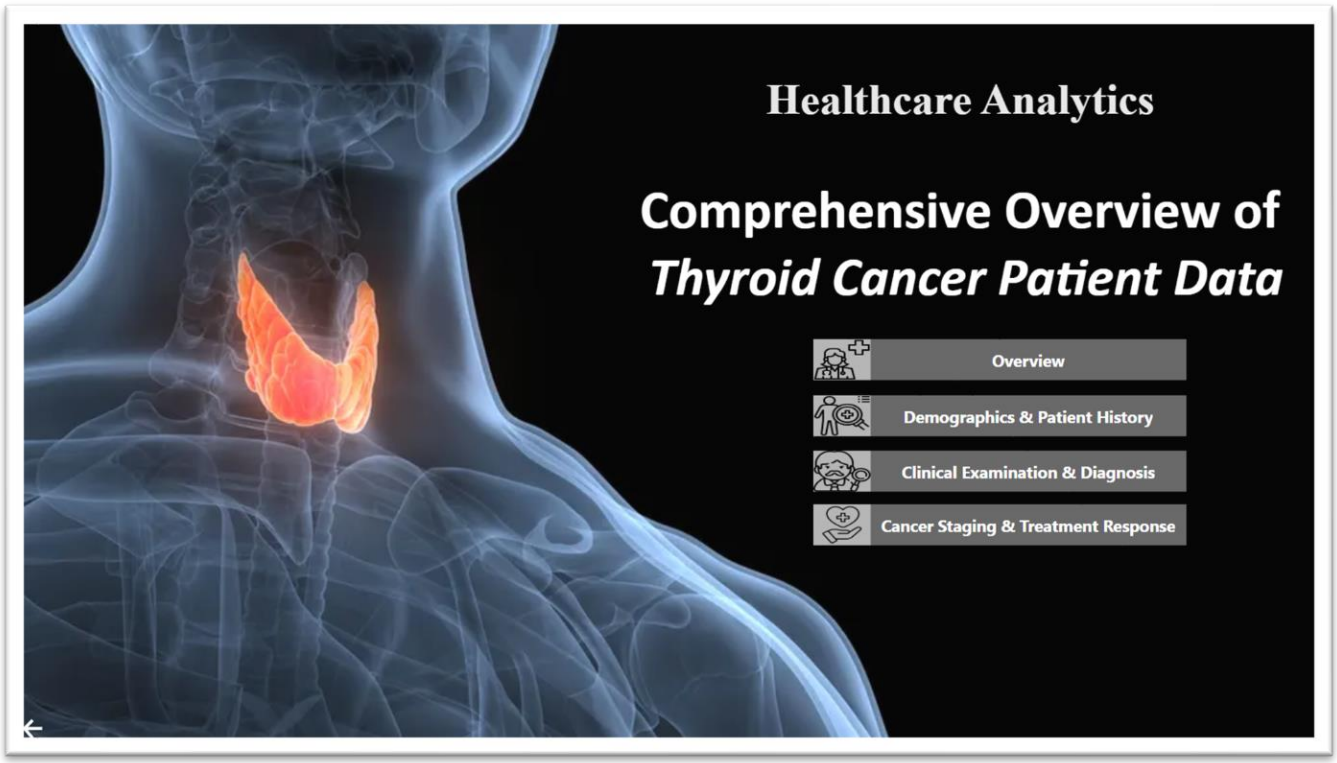
      This dashboard offers a comprehensive analysis of thyroid cancer patient data, highlighting key trends and insights in healthcare analytics. It visualizes critical metrics to support data-driven decisions in oncology care.



12.1 Home Page

- **Selection Panel:**

  The dashboard about "Comprehensive Overview of Thyroid Cancer Patient Data" under the Healthcare Analytics theme. It visually explores patterns in demographics, clinical features, cancer staging, and treatment response. The aim is to support data-driven insights for better diagnosis and care strategies.



12.2 Selection Panel

- **Overview:**

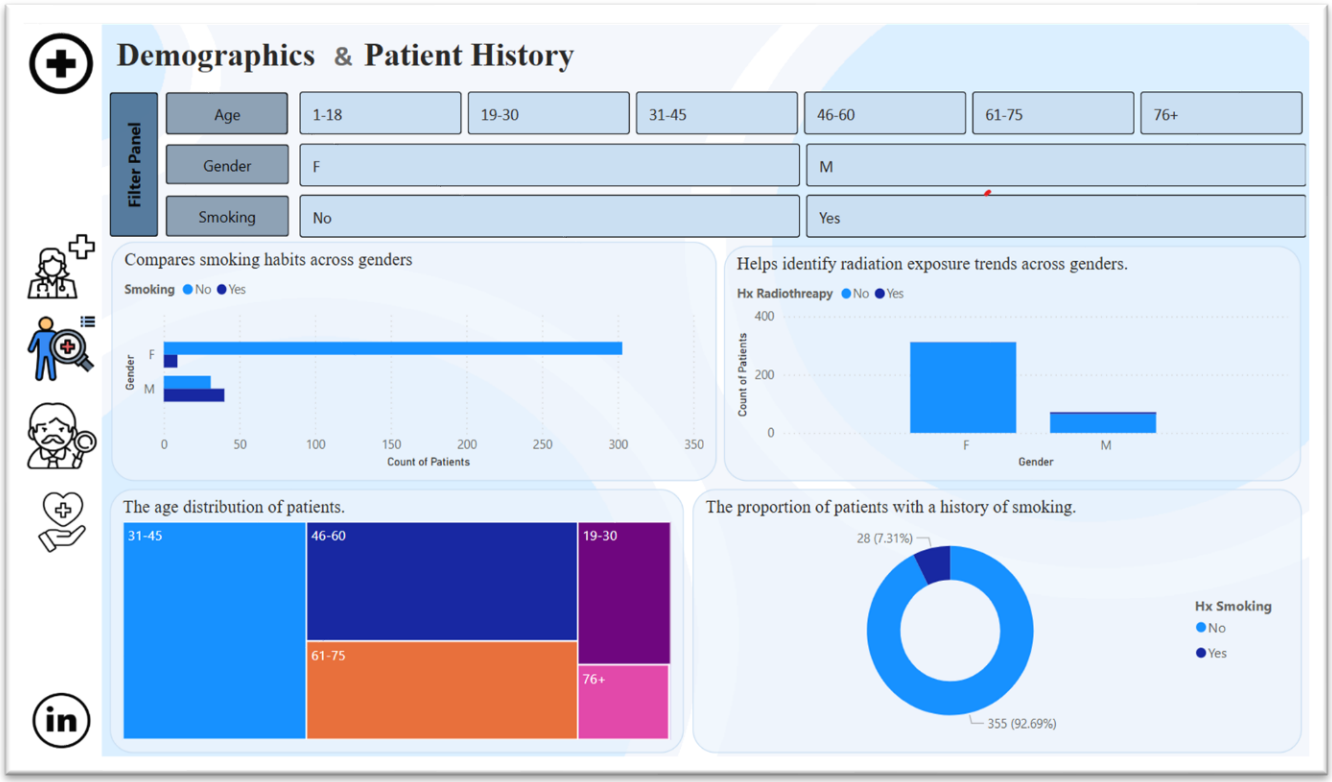
  The dashboard is structured into three core areas:

  - Demographics & Patient History
  - Clinical Examination & Diagnosis
  - Cancer Staging & Treatment Response

  Each section includes interactive slicers (filters) such as age group, gender, smoking history, recurrence status, and more—allowing users to dynamically explore and drill down into the data.

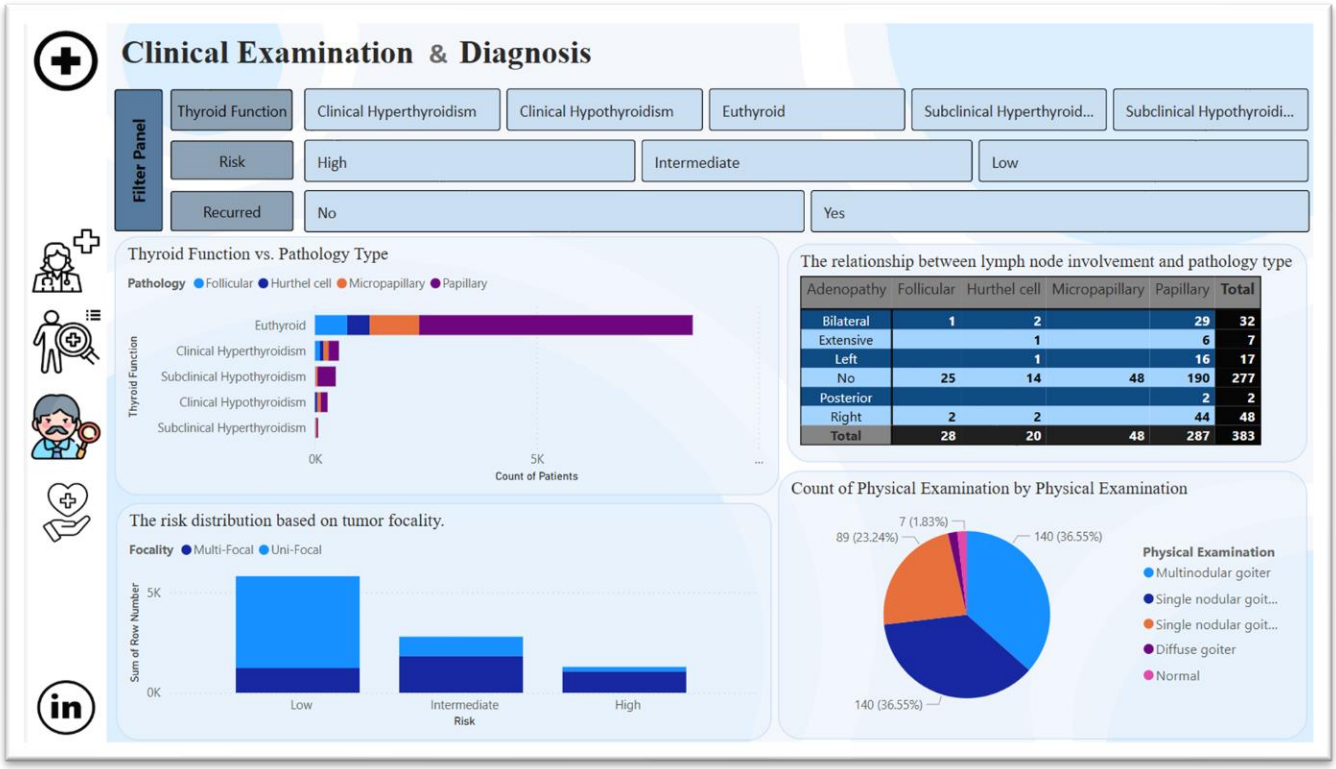
12.3 Overview

- **Demographics & Patient History:**
  Most thyroid cancer patients are female, and the 31–60 age range is the most affected. A large majority have no smoking history or prior radiation exposure. Gender-based differences in risk factors like smoking and radiotherapy are also observed. This section provides baseline population insights, with slicers enabling customized demographic views.



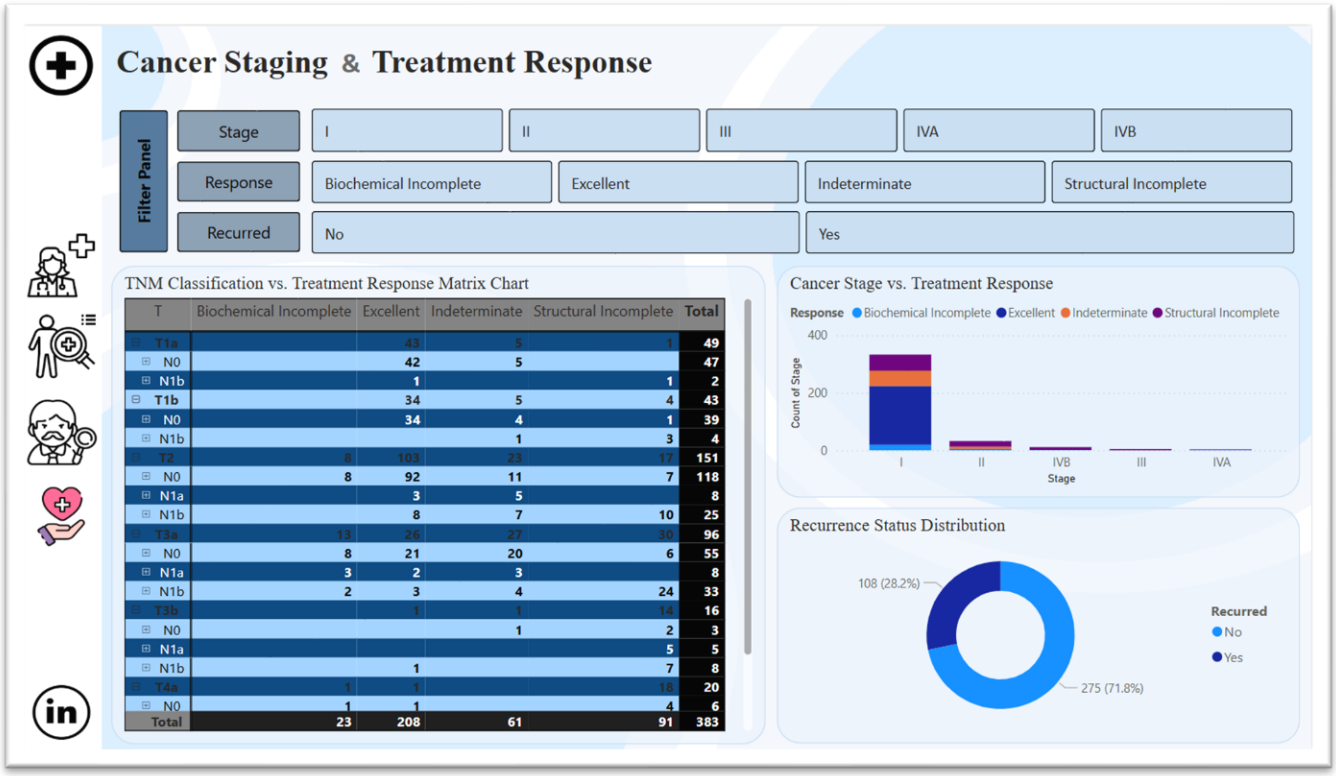
12.4 Demographics & Patient History

- **Clinical Examination & Diagnosis:**
  Papillary carcinoma is the most frequent pathology, with most patients showing normal thyroid function (euthyroid). Uni-focal tumors are more common than multi-focal ones. Lymph node involvement (adenopathy) shows a strong link with recurrence. Goiters are the most common finding on physical exams. Slicers help isolate trends by pathology type, thyroid function, and tumor focality.



12.5 Clinical Examination & Diagnosis

- **Cancer Staging & Treatment Response:**
  Early-stage cancers (Stage I and II) dominate, showing higher rates of excellent treatment response. Tumors with higher T and N classifications tend to have worse outcomes. The recurrence rate stands at 28.2%. The TNM classification matrix helps visualize treatment effectiveness across tumor spread levels. All visuals are slicer-enabled for deeper stage-by-stage analysis.



12.6 Cancer Staging & Treatment Response

- **Feature importance plot (Random Forest)**
  - The single_prediction function is designed to predict the risk of thyroid cancer recurrence for a single new patient based on their input data. It performs preprocessing, model inference, and outputs the prediction along with its confidence score.

  - If a doctor inputs a new patient's data, this function:
    - Predicts whether that patient is at risk of thyroid cancer recurrence.

    - Tells how confident the model is about that prediction.

```
new_input ={
    'Age':30,
    'Gender':'F',
    'Smoking':'No',
    'Hx Smoking':'No',
    'Hx Radiothreapy':'No',
    'Thyroid Function':'Euthyroid',
    'Physical Examination':'Multinodular goiter',
    'Adenopathy':'No',
    'Pathology':'Micropapillary',
    'Focality':'Uni-Focal',
    'Risk':'Low',
    'Tumor':'T1a',
    'Nodal':'N0',
    'Metastasis':'M0',
    'Stage':'I',
    'Response':'Excellent'
}
```

```
new_df = pd.DataFrame([new_input])
new_df
```

]:

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	
0	30	F	No	No	No	Euthyroid	Multinodular goiter	No	Mi

Code Img. 12.1 create a single input and predict stepwise

create single predict function

```
def single_prediction(new_input):
    new_df = pd.DataFrame([new_input])
    new_df[encoder_cols] =encoder.transform(new_df[categorical_cols])
    new_df[['Age']] = scaler.fit_transform(new_df[['Age']])
    new_df = pd.concat([new_df['Age'],new_df[encoder_cols]],axis=1)
    pred = random.predict(new_df)[0]
    proba = random.predict_proba(new_df)[0][0]
    print('Single value prediction is : ',pred)
    print('Single value predic probability is : ',proba)
```

```
single_prediction(new_input)
```

```
Single value prediction is :  No
Single value predic probability is :  0.97875
```

Code Img. 12.2 create single predict function