<u>Evaluation and Prediction of Recurrence of Thyroid Cancer</u>

A project work done in partial fulfilment of the "**Certificate course on Data Analytics & Business Intelligence**"

Submitted by:

KRITIKA MITTAL

&

KRITI KHURANA

Certificate Course on Data Analytics & Business Intelligence Batch-10

Shaheed Sukhdev College of Business Studies May,

2024

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our supervisor, **Dr. Rishi Ranjan Sahay**, for his invaluable guidance and support throughout this project. This project helped us to explore and undertake research work and gaining practical knowledge and expertise to perform tasks. We also extend our thanks to the **Shaheed Sukhdev College of Business Studies, University of Delhi** for providing the necessary resources and facilities.

Thanking You,
Kriti Khurana
Kritika Mittal

DECLARATION

We, **Kritika Mittal** and **Kriti Khurana**, hereby declare that the project entitled "**Evaluation and Prediction of Recurrence of Thyroid Cancer**" is a result of our original research work carried out under the guidance and supervision of **Dr. Rishi Ranjan Sahay, Assistant Professor at Shaheed Sukhdev College of Business Studies, University of Delhi**.

This project work is undertaken as part of our **certificate course in Data Analytics and Business Intelligence** and is submitted in partial fulfilment of the requirements for **the award of the certificate of Data Analytics and Business Intelligence at Shaheed Sukhdev College of Business Studies, University of Delhi**.

We affirm that the research and findings presented in this project are genuine. All sources of information and data have been acknowledged appropriately.

We also declare that any help received in carrying out this project and preparing the report has been duly acknowledged.

ABSTRACT

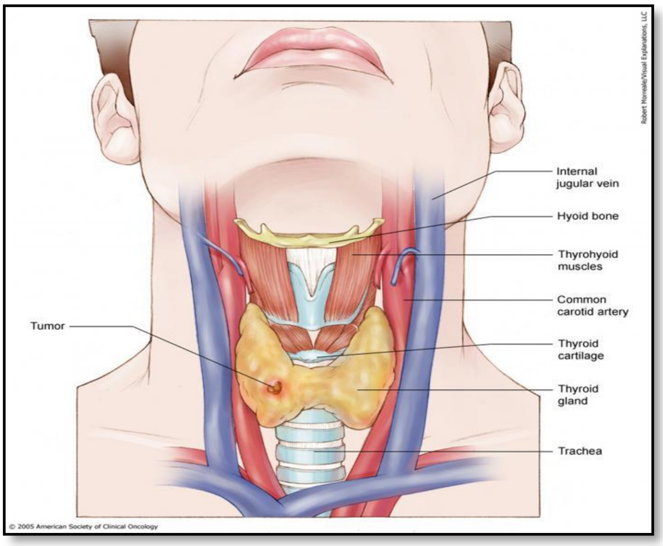
Thyroid cancer, despite its generally favourable prognosis, presents a significant challenge due to the potential for recurrence. This project, "Evaluation and Prediction of Recurrence of Thyroid Cancer," focuses on identifying factors associated with thyroid cancer recurrence and developing predictive models to assess recurrence risk using clinical data.

The dataset used for this study includes patient demographics (age, gender), lifestyle factors (smoking history), clinical and pathological data (thyroid function, physical examination findings, pathology type, focality, TNM classification, and stage), and treatment response. The primary objective is to evaluate these factors to identify key predictors of recurrence.

INTRODUCTION

WHAT IS THYROID CANCER?

Thyroid cancer is the most common malignancy of the endocrine system, characterized by the abnormal growth of cells within the thyroid gland. Located at the base of the neck, the thyroid gland plays a crucial role in regulating metabolism through the production of hormones such as thyroxine (T4) and triiodothyronine (T3). Despite the overall good prognosis for many patients, thyroid cancer can be a significant health concern due to its potential for recurrence and the necessity for lifelong management.



TNM STAGING SYSTEM

One tool that doctors use to describe the stage is the TNM system. Doctors use the results from diagnostic tests and scans to answer these questions:

- **Tumor (T):** How large is the primary tumor? Where is it located?
- **Node (N):** Has the tumor spread to the lymph nodes? If so, where and how many?
- **Metastasis (M):** Has the cancer spread to other parts of the body? If so, where and how much?

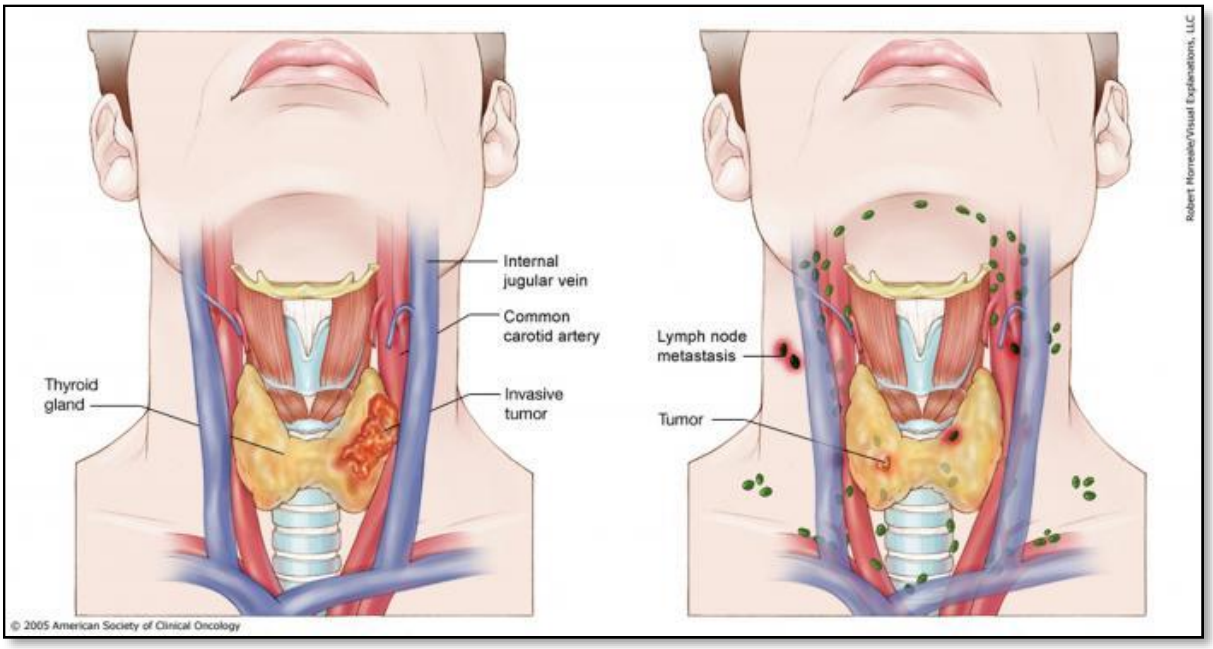
The results are combined to determine the stage of cancer for each person. For thyroid cancer, there are 5 stages: stage 0 (zero) and stages I through IV (1 through 4).

TUMOR (T)

Using the TNM system, the letter “T” plus a letter or number (0 to 4) is used to describe the size and location of the tumor. Tumor size is measured in centimeters (cm).

Stages may also be divided into smaller groups that help describe the tumor in even more detail. When describing “T” in thyroid cancer, doctors may subdivide the general categories by adding the letter “s” to indicate a solitary (single) tumor or “m” to indicate multifocal (more than 1) tumors. Specific tumor stage information is listed below.

- **TX:** The primary tumor cannot be evaluated.
- **T0 (T zero):** There is no evidence of a tumor.
- **T1:** The tumor is 2 centimeters (cm) or smaller and limited to the thyroid.
- **T1a:** The tumor is 1 cm or smaller.
- **T1b:** The tumor is larger than 1 cm but less than 2 cm.
- **T2:** The tumor is larger than 2 cm but smaller than 4 cm and is limited to the thyroid.
- **T3:** The tumor is larger than 4 cm, but the tumor does not extend beyond the thyroid gland.
- **T4:** The tumor is any size and has extended beyond the thyroid.
- **T4a:** The tumor has spread beyond the thyroid to nearby soft tissues, the larynx, trachea, esophagus, or recurrent laryngeal nerve.
- **T4b:** The tumor has spread beyond the regions in T4a (above).



Robert Morreale/Visual Explanations, LLC

© 2005 American Society of Clinical Oncology

NODE (N)

The “N” in the TNM staging system stands for lymph nodes. Careful evaluation of lymph nodes is an important part of staging thyroid cancer. There are many regional lymph nodes located in the head and neck area. Lymph nodes in other parts of the body are called distant lymph nodes.

- **NX:** The regional lymph nodes cannot be evaluated.
- **N0 (N zero):** There is no evidence of cancer in the regional lymph nodes.
- **N1:** Cancer has spread to the lymph nodes.
- **N1a:** Cancer has spread to the lymph nodes around the thyroid (called the central compartment; the pretracheal, paratracheal, and prelaryngeal lymph nodes).
- **N1b:** Cancer has spread beyond the central compartment, including unilateral cervical (lymph nodes on 1 side of the neck), bilateral cervical (lymph nodes on both sides of the neck), contralateral cervical (the opposite side of the tumor), or mediastinal (the chest) lyph nodes.

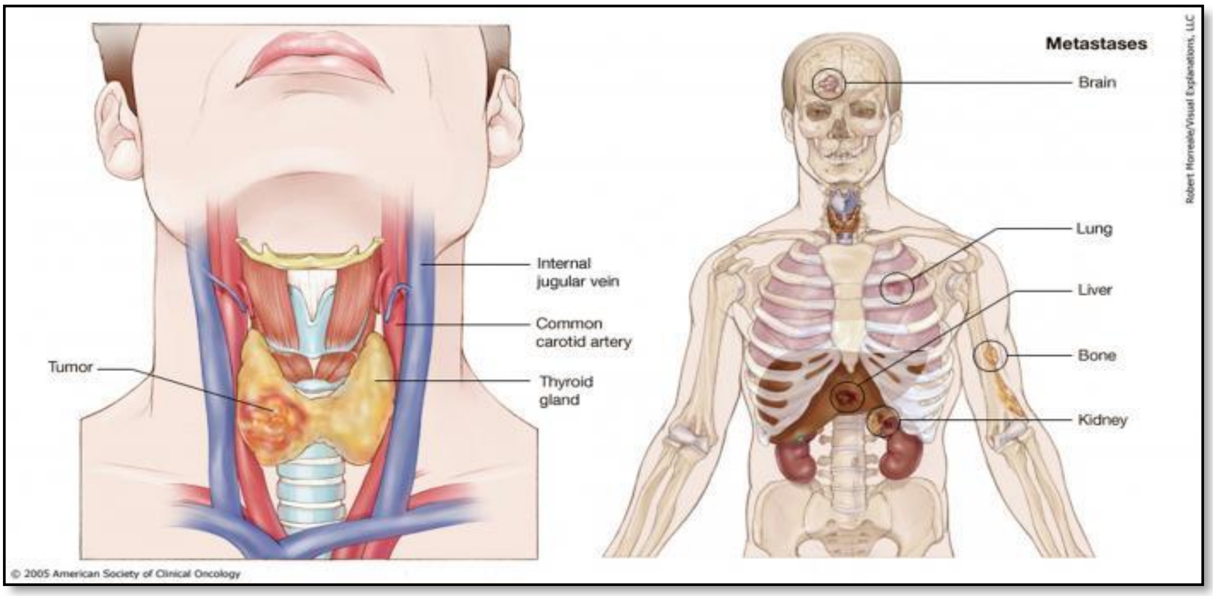
METASTASIS (M)

The “M” in the TNM system describes whether cancer has spread to other parts of the body, called metastasis.

MX: Distant metastasis cannot be evaluated.

M0 (M zero): Cancer has not spread to other parts of the body.

M1: Cancer has spread to other parts of the body.



Robert Morreale/Visual Explanations, LLC

© 2005 American Society of Clinical Oncology

Management of thyroid cancer is tailored to the type and stage of the disease, as well as patient-specific factors. Treatment options include surgical resection (thyroidectomy), radioactive iodine therapy, thyroid hormone suppression therapy, and in some cases, external beam radiation or targeted therapies. While many patients achieve remission, the risk of recurrence necessitates regular follow-up and monitoring.

Despite advancements in diagnostic and therapeutic strategies, predicting the recurrence of thyroid cancer remains a clinical challenge. Recurrence can occur locally within the thyroid bed or regionally in lymph nodes, and less commonly, it can present as distant metastases. Understanding the factors that contribute to recurrence and developing robust predictive models are essential for improving patient outcomes and tailoring individualized treatment plans.

This project aims to evaluate and predict the recurrence of thyroid cancer by analyzing clinical and pathological data. By identifying significant prognostic factors and developing predictive models, we strive to enhance the understanding of recurrence patterns and provide tools for early identification of high-risk patients. This, in turn, can lead to more personalized and effective management strategies, ultimately improving the quality of life and prognosis for thyroid cancer patients.

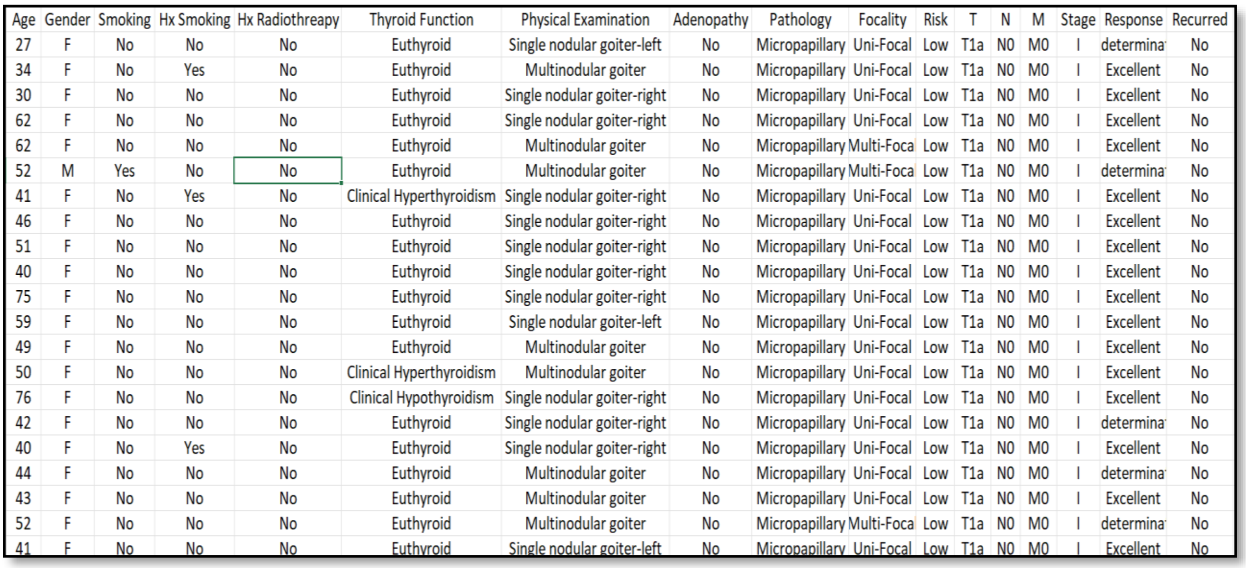
RESEARCH OBJECTIVE

The primary objective of this project is to evaluate the performance of KNN, Logistic Regression, and SVM algorithms in predicting thyroid cancer recurrence based on clinical, demographic, and pathological features. By comparing the accuracy, precision, F1 score, and confusion matrices of these classification models, we aim to identify the most effective algorithm for predicting thyroid cancer recurrence and provide valuable insights for clinical decision-making and patient management.

METHODOLOGY

DATASET

We took this dataset from 'Kaggel.com' in a csv file format & this was uploaded just four months ago. Thus, the data is still new to make a project. It has 383 rows and 17 columns. Following is a picture of the same:



Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	determina	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-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Foca	Low	T1a	N0	M0	I	Excellent	No
52	M	Yes	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Foca	Low	T1a	N0	M0	I	determina	No
41	F	No	Yes	No	Clinical Hyperthyroidism	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
46	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
51	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
40	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
75	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
59	F	No	No	No	Euthyroid	Single nodular goiter-left	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 Hyperthyroidism	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
76	F	No	No	No	Clinical Hypothyroidism	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
42	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	determina	No
40	F	No	Yes	No	Euthyroid	Single nodular goiter-right	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	determina	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-Foca	Low	T1a	N0	M0	I	determina	No
41	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No

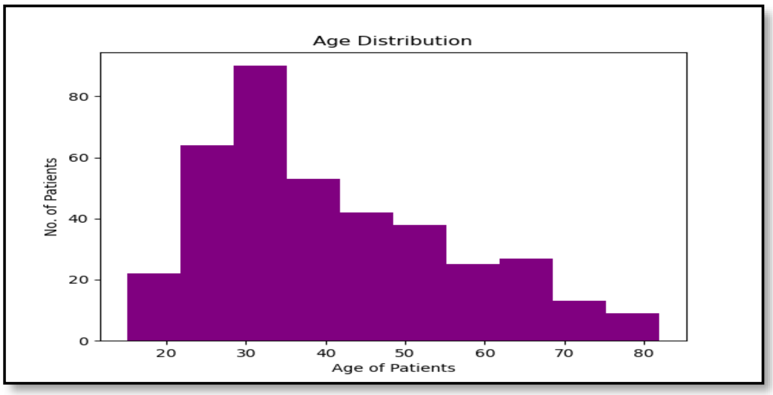
DETAILED INFORMATION OF COLUMNS OF THE DATASET

Column Name	Feature	Type
AGE	Represents the age of the patient.	Integer
GENDER	Represents the gender of the patient.	Categorical
SMOKING	Indicates whether the patient is currently smoking.	Categorical
Hx SMOKING	Indicates whether the patient has smoked in the past.	Categorical
Hx RADIOTHERAPY	Indicates whether the patient has undergone radiotherapy in the past.	Categorical
THYROID FUNCTION	The functional status of thyroid.	Categorical
PHYSICAL EXAMINATION	Results of the physical examination.	Categorical

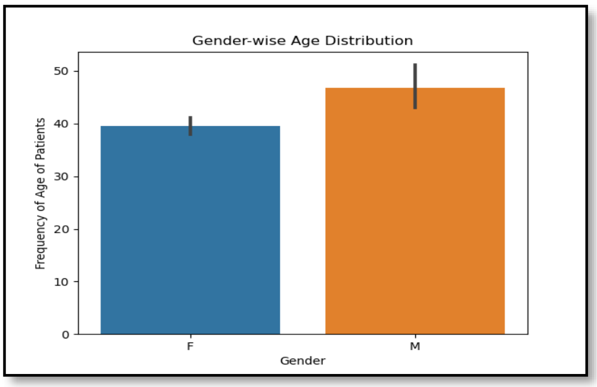
ADENOPATHY	Presence of swollen or abnormal lymph nodes.	Categorical
PATHOLOGY	Details of the pathology report.	Categorical
FOCALITY	Whether the cancer is focal (localized) or multifocal.	Categorical
RISK	Checking the risk of patient.	Categorical
T	T classification in the TNM staging system (size/extent of the primary tumor).	Categorical
N	N classification in the TNM staging system (spread to nearby lymph nodes).	Categorical
M	M classification in the TNM staging system (metastastis).	Categorical
STAGE	Overall stage of the cancer	Categorical
RESPONSE	Response to the treatment	Categorical
RECURRED	Indicates whether the cancer has recurred or not.	Categorical

EXPLORATORY DATA ANALYSIS (EDA)

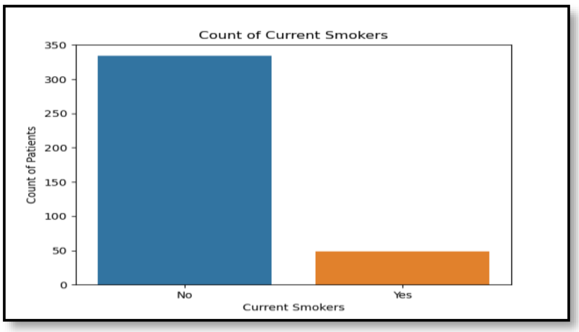
1. **Age Distribution Graph-** This is a histogram of age of patients and their frequencies. This graph shows that the maximum frequency of patients belongs to the age group of 30-35 years in this dataset.



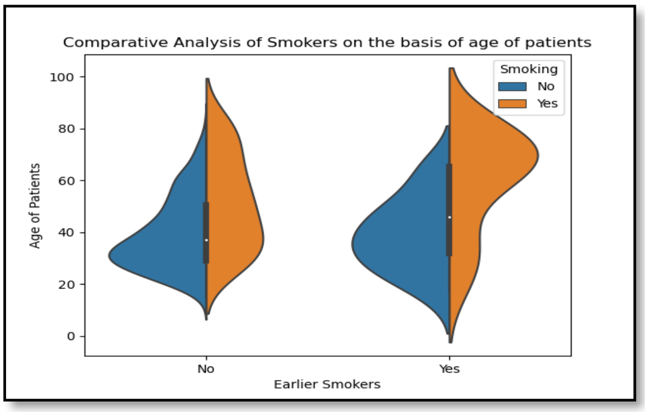
2. **Gender-wise Age Distribution-** This is a bar plot of frequencies of age of the patients on the basis of their gender. This graph shows that there are a more male patients than female in this dataset.



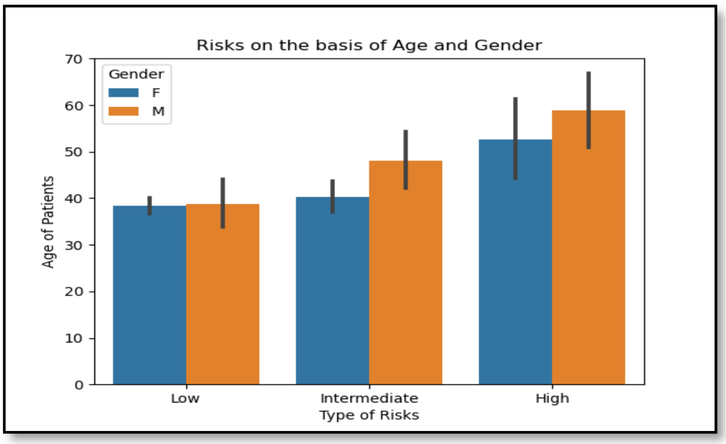
3. **Count of Current Smokers-** This is a count plot which counts the no. smokers who are currently smoking. This graph shows that there are more no. of non-smokers than smokers at present in our dataset.



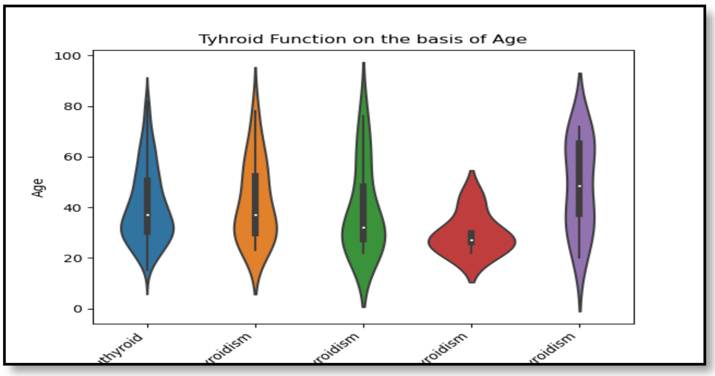
4. **Comparative Analysis of Smokers on the basis of age of patients-** This is a violin plot showing a comparative analysis of current smokers and earlier smokers on the basis of the age of patients. This graph shows a pattern among smokers and non-smokers.



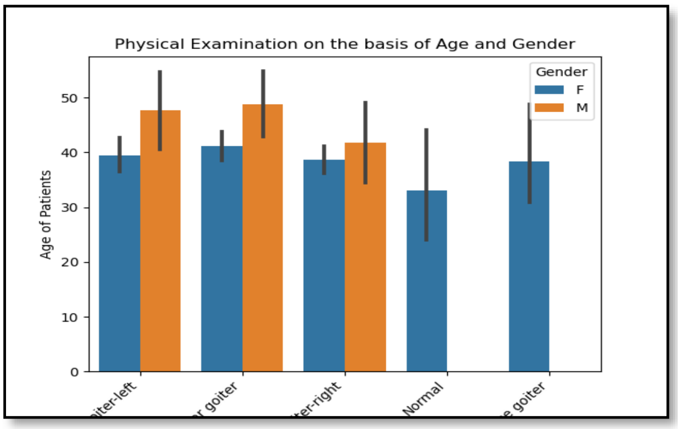
5. **Risk on the basis of Age and Gender-** This is a bar plot which shows the risk of patients on the basis of their age and their gender. This graph shows that the people of age of 50+ are at higher risk and people below the age of 40 are at low risk.



6. **Thyroid Function on the basis of Age-** This is a violin plot which shows the distribution of patients among the thyroid function on the basis of their respective age.



7. **Physical Examination on the basis of Age and Gender-** This is a bar plot which shows the distribution of patients according to their respective physical examination and on the basis of their age and gender. This graph shows that more physical examinations are involved for females than males.



CLASSIFICATION ALGORITHMS

Thyroid cancer is one of the most common endocrine malignancies globally, and its recurrence poses a significant challenge in clinical management. Predicting the likelihood of thyroid cancer recurrence is crucial for tailoring treatment strategies and optimizing patient outcomes. In this data analysis project, we explore the efficacy of three classification algorithms—K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine (SVM)—in evaluating and predicting thyroid cancer recurrence.

K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors is a simple yet powerful algorithm used for classification tasks. It operates on the principle of proximity, where the class of a data point is determined by the classes of its nearest neighbors in the feature space. KNN does not make any assumptions about the underlying data distribution, making it particularly suitable for non-linear and complex datasets. The algorithm's performance heavily relies on the choice of the number of neighbors (K) and the distance metric used for measuring similarity between data points.

LOGISTIC REGRESSION

Logistic Regression is a widely used statistical method for binary classification tasks. Despite its name, logistic regression is a linear model that predicts the probability of an instance belonging to a particular class. It employs the logistic function (sigmoid function) to map the output of a linear combination of input features to a probability score between 0 and 1. Logistic regression provides interpretable results, as it allows for estimating the effect of each predictor variable on the probability of the outcome.

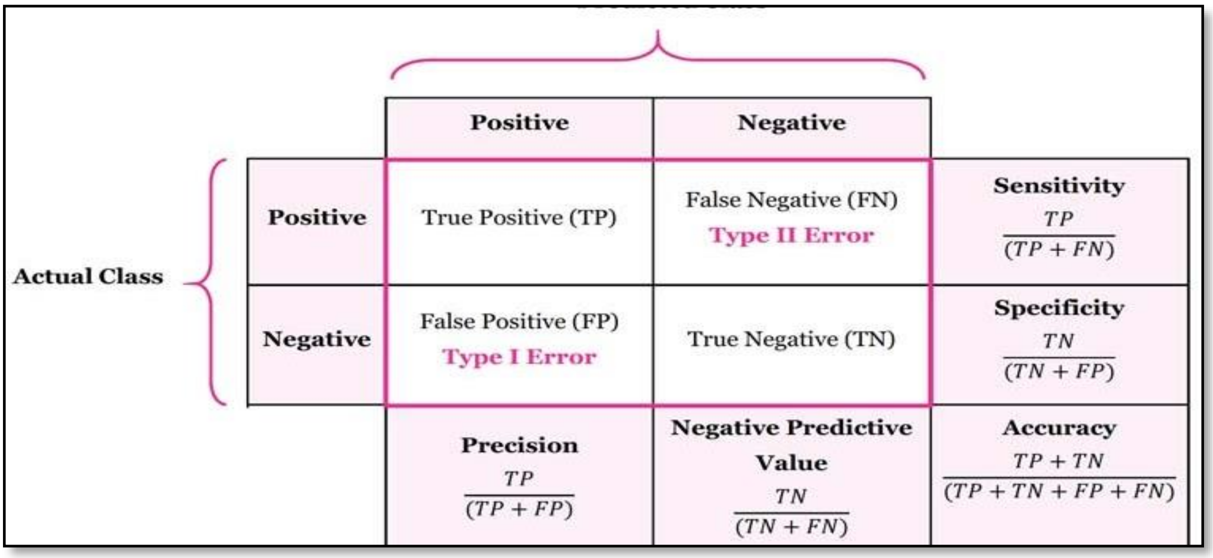
SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a powerful supervised learning algorithm for both classification and regression tasks. SVM aims to find the optimal hyperplane that separates different classes in the feature space while maximizing the margin between the classes. By transforming the input features into a higher-dimensional space using kernel functions, SVM can efficiently handle non-linear decision boundaries. SVM offers flexibility in choosing different kernel functions such as linear, polynomial, and radial basis function (RBF), allowing for versatile classification tasks.

CONFUSION MATRIX

CONFUSION MATRIX

A confusion matrix is a performance evaluation tool used in classification tasks to visualize the performance of a machine learning model. It presents a summary of the predictions made by the model compared to the actual ground truth across different classes.



		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

COMPONENTS OF CONFUSION MATRIX

1) **True Positive (TP):** The instances that were correctly predicted as positive (or belonging to the target class) by the model.
2) **True Negative (TN):** The instances that were correctly predicted as negative (or not belonging to the target class) by the model.
3) **False Positive (FP):** The instances that were incorrectly predicted as positive by the model when they actually belong to the negative class. Also known as Type I error.
4) **False Negative (FN):** The instances that were incorrectly predicted as negative by the model when they actually belong to the positive class. Also known as Type II error.

INTERPRETATION

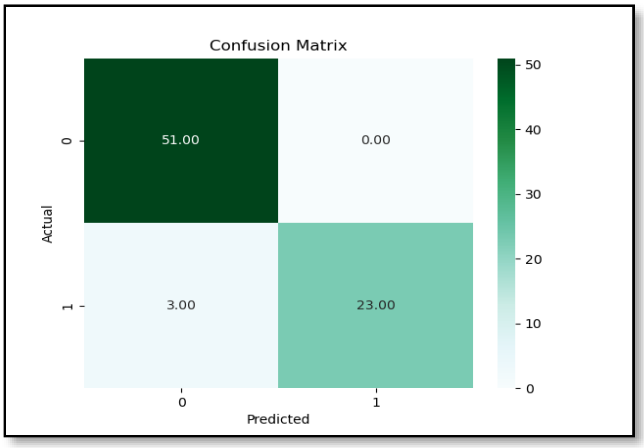
- **Accuracy:** The overall correctness of the model, calculated as (TP + TN) / (TP + TN + FP + FN).
- **Precision:** The proportion of true positive predictions out of all positive predictions made by the model, calculated as TP / (TP + FP).
- **Recall (Sensitivity):** The proportion of true positive predictions out of all actual positive instances, calculated as TP / (TP + FN).
- **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two metrics, calculated as 2 * (Precision * Recall) / (Precision + Recall).

MODEL DEVELOPMENT

- Before starting with the algorithms, we have to convert the categorical data into dummy variable i.e. (True/False).
- After that we will convert this data into integer format i.e. (0 to 1). After that observe the independent and dependent variable in our dataset.
- We then split the data using Sklearn library into Training data (80% of dataset) and Test data (20% of dataset).

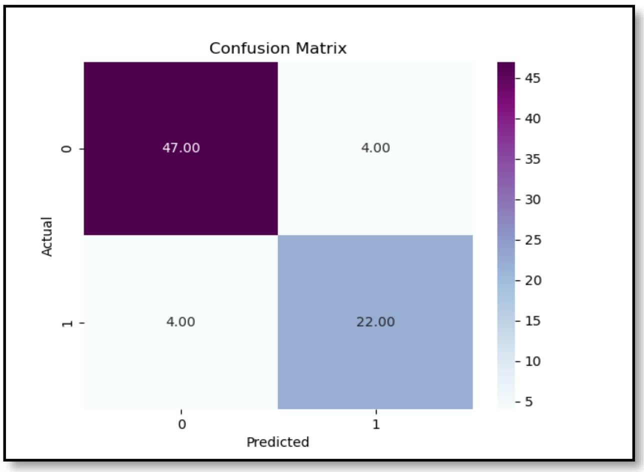
RESULTS OF CLASSIFICATION ALGORITHMS

LOGISTIC REGRESSION



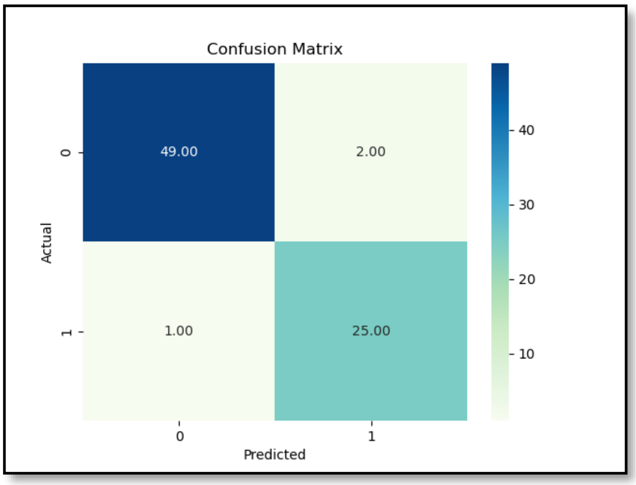
i. Accuracy score is 0.961038961038961
ii. Precision score is 1.0
iii. Recall score is 0.8846153846153846
iv. F1 score is 0.9387755102040816

K-NEAREST NEIGHBORS (KNN)



i. Accuracy score is 0.8961038961038961
ii. Precision score is 0.8461538461538461
iii. Recall score is 0.8461538461538461
iv. F1 score is 0.8461538461538461

SUPPORT VECTOR MACHINE (SVM)



i. Accuracy score is 0.961038961038961
ii. Precision score is 0.9259259259259259
iii. Recall score is 0.9615384615384616
iv. F1 score is 0.9433962264150944

CONCLUSION

The analysis of the dataset pertaining to thyroid cancer recurrence reveals several insightful patterns and trends. The age distribution graph highlights a peak frequency among patients aged 30-35 years, suggesting a potential age-related factor in thyroid cancer. Gender-wise analysis indicates a higher prevalence of male patients. However, risk assessment based on age and gender underscores higher risk among individuals aged 50+ and lower risk among those below 40. Moreover, physical examination distribution showcases gender disparities, with more examinations conducted for females. These findings collectively provide valuable insights into the demographics and characteristics associated with thyroid cancer recurrence, paving the way for targeted interventions.

Based on the evaluation metrics provided, the Logistic Regression and Support Vector Machine (SVM) models outperform the K-Nearest Neighbors (KNN) model for predicting thyroid cancer recurrence. Both Logistic Regression and SVM exhibit high accuracy scores of approximately 96.10%, indicating their robustness in making correct predictions. However, considering precision, recall, and F1 score, SVM demonstrates slightly better performance, with higher precision and recall values.

Therefore, the Support Vector Machine (SVM) emerges as the best-performing model for our dataset. Its ability to accurately classify instances and effectively balance precision and recall makes it the preferred choice for predicting thyroid cancer recurrence.

BIBLIOGRAPHY

1) https://www.cancer.net/cancer-types/thyroid-cancer/stages
2) https://www.cancercenter.com/cancer-types/thyroid-cancer/stages#:~:text=A%20thyroid%20cancer%20recurrence%20typically,in%20the%20bones%20and%20lungs.
3) https://www.kaggle.com/datasets/abuchionwuegbusi/thyroid-cancer-recurrence-prediction
4) https://www.cancer.net/sites/cancer.net/files/styles/medical_illustration_web/public/thyroid_l45_stagei_print.jpg
5) https://www.cancer.net/sites/cancer.net/files/styles/medical_illustration_web/public/thyroid_l45_stageii_print.jpg
6) https://www.cancer.net/sites/cancer.net/files/styles/medical_illustration_web/public/thyroid_g45_stageiii_print.jpg
7) https://www.geeksforgeeks.org/understanding-logistic-regression/
8) https://www.natasshaselvaraj.com/logistic-regression-explained-in-7-minutes/
9) https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEiqxBWDxr82HWBNDD-XI22ht1rEZdhCUXilxAMtKFQugD4WxvEUUfqBv6fZfPwQ7MZZn_AM-xXGjhl2yeaHfWq47U2yKuigKcqy3wi5LQx7ExuJEnkdIJfyZHjpzRCkbYXil8z69kLf4a2_/w1200-h630-p-k-no-nu/confusionMatrxiUpdated.jpg