***Predicting recurrence of Thyroid Cancer***

**A PROJECT REPORT**

**Submitted by**

***Pushan T – 1RVU23CSE360***

***Samarth Swarup K – 1RVU23CSE402***

***Sagar N Rao – 1RVU23CSE393***

***Rashi Jajodia* – 1RVU23CSE371**

*in partial fulfillment for the award of the degree of*

**B.Tech (Hons) – Computer Science & Engineering**

*in* **Data Analysis with Python**

**to Dr. Sahana Prasad**



**School of Computer Science and Engineering**

**RV University**

**RV Vidyaniketan, 8th Mile, Mysuru Road, Bengaluru, Karnataka, India – 562112.**

**December 2023**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** |  |
| **1.0** | **INTRODUCTION** | **3** |
|  | 1.1 Data Description | 3 |
|  |  |  |
| **2.0** | **METHODOLOGY** | **5** |
|  |  |  |
| **3.0** | **INFERENCE** | **6** |
|  |  |  |
| **4.0** | **RESULT AND DISCUSSION** | **7** |
|  |  |  |
| **5.0** | **CONCLUSION** | **8** |

1. **INTRODUCTION**

Cancer has been a very concerning factor in the survival of the human race, mainly because it has no cure. Thyroid cancer is a prevalent malignancy in the thyroid gland and it requires an analysis of its factors and patterns in several patients who have suffered it.

In this project, we perform Exploratory Data Analysis (EDA) and visualization to gain useful insights and understand patterns seen in thyroid cancer patients.

# Data description

## Brief explanation of each column:

* **Age**: The age of the patient in years
* **Gender**: The gender of the patient (F for female, M for male)
* **Smoking**: Whether the patient has a history of smoking (Yes or No)
* **Hx Smoking**: Whether the patient has a history of smoking or radiotherapy (Yes or No)
* **Thyroid Function**: This column represents the thyroid function status of the patient. The thyroid gland produces hormones that regulate the body’s metabolism. The three possible values in this column are:
  + **Euthyroid**: The thyroid gland is functioning normally.
  + **Clinical Hyperthyroidism**: The thyroid gland is overactive and producing too much thyroid hormone. Symptoms include weight loss, agitation and increased heart rate.
  + **Clinical Hypothyroidism**: The thyroid gland is underactive and not producing enough thyroid hormone.
* **Physical Examination**: The type of thyroid enlargement or goiter observed in the patient (Single nodular, Multinodular, or Diffuse)
* **Adenopathy**: Whether the patient has swollen glands or lymph nodes (Yes, No, or Bilateral)
* **Pathology**: The type of thyroid cancer diagnosed in the patient (Papillary, Follicular, Medullary, Anaplastic, Hurthel cell, or Micropapillary)
* **Focality**: The number of tumors in the thyroid gland (Uni-Focal, Multi-Focal, or Extensive)
* **Risk**: The risk level of the patient based on the tumor size and other factors (Low, Intermediate, or High)
* **T**: This column represents the tumor stage based on the size and extent of the primary tumor. The possible values in this column are:
  + **T1a**: The tumor is 2 cm or smaller and is limited to the thyroid gland.
  + **T1b**: The tumor is larger than 2 cm but not larger than 4 cm and is limited to the thyroid gland.
  + **T2**: The tumor is larger than 4 cm but not larger than 5 cm and is limited to the thyroid gland, or the tumor is any size and has grown outside the thyroid gland and invaded nearby tissues.
  + **T3**: The tumor is larger than 5 cm and is limited to the thyroid gland, or the tumor has grown outside the thyroid gland and invaded nearby tissues.
  + **T4a/b**: The tumor has grown outside the thyroid gland and invaded nearby tissues, such as the trachea, esophagus, or larynx.
* **N**: This column represents the node stage based on the presence and number of regional lymph node metastases (spread of cancer).
  + **N0**: No regional lymph node metastases are present.
  + **N1a**: Metastases are present in the same side of the neck as primary tumor.
  + **N1b**: Metastases are present in both sides of the neck.
* **M**: This column represents the metastasis stage based on the presence of distant metastases. The possible values in this column are:
  + **M0**: No distant metastases are present.
  + **M1**: Cancer has spread to other organs.
* **Stage**: The overall stage of the cancer based on the TNM system (I, II, III, IVA, IVB, or IVC)
* **Response**: This column represents the response of the patient to the treatment. The possible values in this column are:
  + **Excellent**: The patient has responded well to the treatment and shows no signs of cancer.
  + **Biochemical Incomplete**: The patient’s blood tests show no signs of cancer, but imaging tests show that the cancer is still present.
  + **Structural Incomplete**: Imaging tests show that the cancer is still present and has not responded to the treatment.
  + **Indeterminate**: The response to the treatment is unclear.
* **Recurred**: Whether the cancer recurred after the treatment.

**2. METHODOLOGY**

Python is the primary programming language used, we utilize the Pandas, Matplotlib, and Seaborn libraries to perform Exploratory Data Analysis.

* We use various statistical measures such as finding the average and standard deviation of the ages.
* There is no need for data cleaning as the dataset has no missing values or duplicates.
* The standard deviation appears to be very high, therefore we stick to median and mode to find the central tendency of patient ages.
* We use a boxplot to visualize the distribution of patients who had a recurrence based on their ages.
* Among those with a recurrence of thyroid cancer, we sort the number of patients by the state of their thyroid function to check if issues with their thyroid function have any effect on their chances of cancer recurrence.
* We have plotted multiple heatmaps on the following data fields to check for a correlation between them and the likelihood of a cancer recurrence:
  + Stage of initial cancer vs a recurrence
  + Tumor size vs response to previous cancer treatment
  + Recurrence vs type of cancer pathology
  + Nodal Stage of initial cancer vs spread of cancer to other organs
* We also use a bar plot to visualize the number of recurrences based on the Nodal stage and the spread of cancer to other organs.
* We utilize a parallel categories plot to visualize to correlation between the patient's smoking history, radiotherapy history and the presence of swollen lymph nodes.

**3. INFERENCE**

We find that the average age of patients based on the recurrence of cancer:

* Average age of all patients: 41 years
* Average age of those who suffer a reoccurrence of cancer: 47

We find the standard deviation to be very high (18.2) meaning that the mean would be a bad measure of central tendency. Instead, we use mode: The most common age of patients who have suffered thyroid cancer (including those without recurrence) is 31 years and the median age is 37 years.

The boxplot visualizes the distribution of patients based on age by dividing the data into quartiles:

* The youngest age in the dataset is 15 years.
* The oldest person in the dataset is 82 years old.
* 50% of the patients are between the ages of 32-62 years.

We also see that abnormal thyroid function has little to no effect on the chances of cancer recurrence as most of the patients under this category have normal thyroid functions.

Based on the Initial Cancer Stage heatmap, the most important thing that can be noted is that those whose initial cancer was at stage 3 or farther, always have a cancer recurrence.

The heatmap of the tumor stage against the treatment response of initial cancer shows that those with a completely successful cancer treatment have a very low chance of having a recurrence of cancer whereas those with a structurally incomplete treatment response are highly likely to have a cancer recurrence.

From the heatmap of the cancer pathology type, we can note that Papillary cancer is the most common type of cancer. However, those who suffer from follicular type have the highest probability of cancer recurrence.

The heatmap of the nodal stage against the spread of cancer to other organs shows that recurrence is most likely among those where the cancer has stayed within the thyroid.

Furthermore, the bar plot indicates that cancer is most likely to re-occur among those where metastasis is present on both sides of the neck.

The parallel categories plot shows us that having no history of smoking and radiotherapy treatment is an indicator that the patient is likely not going to have a cancer recurrence in the future. However, having swollen lymph nodes puts the patient at a high risk of having a recurrence.

**4. RESULT AND DISCUSSION**

The findings on the relationship between cancer stage and recurrence underscore the critical role of early detection and intervention in mitigating recurrence risks, offering actionable insights for clinicians and healthcare providers.

The exploration of age as a predictor revealed a significant correlation, emphasizing the importance of age-dependent risk assessments in thyroid cancer. We observe that most patients fall under the age range of 32-62 years warranting additional post-treatment monitoring for individuals who fall into this age group.

The unexpected inverse relationship between a history of metastasis and the likelihood of recurrence adds a layer of complexity to our understanding, warranting further investigation into the complex interplay between metastases and recurrence dynamics. Additionally, the impact of papillary and follicular cancer on recurrence rates highlights the need for more personalized post-treatment surveillance and therapeutic strategies for individuals with this specific cancer subtype. Finally, the elevated recurrence risk associated with tumor class T3a is a crucial indicator that guides the identification of individuals requiring intensified monitoring.

The insights gained from this analysis serve as valuable contributions to the ongoing efforts to combat this formidable malignancy as cancer has been a serious concern for people. As we reflect on this report, it becomes clear that this project has not only deepened our comprehension of thyroid cancer but also provides actionable recommendations for clinicians and researchers to enhance healthcare options for thyroid cancer patients and improve their outcomes.

**5. CONCLUSION**

The insights gained from this analysis serve as valuable contributions to the ongoing efforts to combat this formidable malignancy as cancer has been a serious concern for people. As we reflect on this report, it becomes clear that this project has not only deepened our comprehension of thyroid cancer but also provides actionable recommendations for clinicians and researchers to enhance healthcare options for thyroid cancer patients and improve their outcomes.

### Sources:

1. Thyroid gland: Anatomy, functions and hormones | Kenhub. <https://www.kenhub.com/en/library/anatomy/thyroid-gland>.
2. Thyroid function tests - British Thyroid Foundation. <https://www.btf-thyroid.org/thyroid-function-tests>.
3. Overview of Thyroid Function - Endocrine and Metabolic Disorders <https://www.msdmanuals.com/en-in/professional/endocrine-and-metabolic-disorders/thyroid-disorders/overview-of-thyroid-function>.
4. Thyroid Function Tests | American Thyroid Association. <https://www.thyroid.org/thyroid-function-tests/>.
5. What is Metastasis.  
   <https://www.cancer.net/navigating-cancer-care/cancer-basics/what-metastasis>
6. Dataset | Kaggle: [Differentiated Thyroid Cancer Recurrence (kaggle.com)](https://www.kaggle.com/datasets/zsinghrahulk/differentiated-thyroid-cancer-recurrence)