greatlearning

Oral Cancer Prediction Based on Clinical and Lifestyle Factors

Introduction:

Oral cancer remains one of the most critical public health concerns worldwide, especially in developing countries like India where the prevalence of tobacco, alcohol, and betel quid consumption is high. Early detection of oral cancer significantly increases the chances of successful treatment and long-term survival. However, in many cases, the disease is diagnosed only in its advanced stages due to lack of awareness, poor access to healthcare, and limited use of screening tools.

With the growing availability of healthcare data and advancements in machine learning (ML), there is a significant opportunity to improve early diagnosis through data-driven predictive models. This project aims to leverage supervised learning techniques to predict the likelihood of oral cancer based on individual demographic, lifestyle, and clinical attributes.

The primary objective of this capstone project is to develop a reliable, non-invasive predictive model using historical data. By identifying high-risk individuals early, such models can help support timely clinical intervention, reduce the burden on healthcare systems, and ultimately save lives. This report outlines the interim progress of the project, covering background research, data exploration, and the methodologies applied thus far.

1. Industry Review – Current Practices and Background Research

1.1. Background Research:

- Oral cancer refers to malignant tumors occurring in the oral cavity and oropharynx, including the lips, tongue, cheeks, floor of the mouth, hard and soft palates, sinuses, and pharynx. It is the sixth most common cancer worldwide and a major health concern in countries like India, where tobacco chewing, smoking, and alcohol consumption are prevalent.
- The early stages of oral cancer often go undiagnosed due to the absence of noticeable symptoms, and by the time it is clinically detected, it may have progressed to advanced stages, significantly reducing the survival rate. According to the World Health Organization (WHO), early detection and timely treatment can increase survival rates by 70-90%.

1.2 Current Diagnostic Practices:

- Traditional diagnostic methods for oral cancer typically involve:
  - Clinical examination by specialists (visual and tactile inspection).
  - Biopsy and histopathology, which are invasive and time-consuming.
  - Imaging techniques like CT, MRI, and PET scans for tumour staging.
  - Vital staining and exfoliative cytology, which are useful for identifying suspicious lesions but not always definitive.

- While effective, these methods have limitations:
  - They require trained professionals.
  - They are time- and resource-intensive.
  - Early detection is highly dependent on patient initiative and clinical suspicion.

2. Literature Survey

2.1. General Oral Health

- **Publication:** World Health Organization (WHO) Fact Sheet on Oral Health.
  - Source: https://www.who.int/news-room/fact-sheets/detail/oral-health
- **Application:** Provides foundational information for understanding the global burden of oral diseases, including oral cancer, and emphasizes the importance of oral health in overall well-being.
- **Past and Ongoing Research:** WHO continues to monitor oral health trends globally, conduct surveys, and develop guidelines for oral disease prevention and control. Research focuses on the impact of socioeconomic factors on oral health and effective public health interventions.

2.2. Salivary Extracellular Vesicles as Biomarkers

- **Publication:** "Salivary extracellular vesicles as biomarkers for early oral squamous cell carcinoma detection." Oral Oncology.
  - Source: https://www.sciencedirect.com/science/article/abs/pii/S1368837516302135
- **Application:** The study explores the use of salivary extracellular vesicles as non-invasive biomarkers for early detection of oral squamous cell carcinoma. This could lead to new diagnostic tools.
- **Past and Ongoing Research:** Research in this area involves identifying specific molecular markers in saliva that indicate the presence of oral cancer, validating these markers in larger studies, and developing Point-of-Care diagnostic tests.

2.3 Periodontitis and Oral Cancer Risk

- **Publication:** Research on the relationship between periodontitis and oral cancer risk in the Journal of Periodontology.
  - Source: https://onlinelibrary.wiley.com/doi/abs/10.1111/jop.13157

- **Application:** This research highlights the importance of periodontal health in the context of oral cancer prevention. Dental professionals can use this information to educate patients and identify high-risk individuals.

- **Past and Ongoing Research:** Further research is exploring the biological mechanisms that link periodontitis to oral cancer, including the role of chronic inflammation and specific bacteria. Longitudinal studies are investigating whether treating periodontitis can reduce oral cancer risk.

2.4 Deep Learning for Oral Lesion Detection

- **Publication:** Deep Learning for the Automated Detection and Diagnosis of Oral Lesions and Cancer: A Systematic Review. Cancers.
  - Source: https://www.mdpi.com/2072-6694/16/3/617
- **Application:** This review discusses the application of deep learning techniques to automate the detection and diagnosis of oral lesions and cancer, potentially improving accuracy and efficiency in clinical settings.
- **Past and Ongoing Research:** Ongoing research involves developing more sophisticated deep learning models, using larger and more diverse datasets, and integrating AI tools into clinical workflows. There is also a focus on developing methods that are robust to variations in image quality and lighting conditions.

2.5. Global Burden of Preventable Oral Cancers

- **Publication**: Research on the global burden of potentially preventable oral cavity and pharyngeal cancers. International Journal of Cancer.
  - Source: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.34277
- **Application**: This research informs public health policy and interventions aimed at reducing the incidence of oral cancer by addressing preventable risk factors.
- **Past and Ongoing Research**: Research continues to monitor cancer incidence trends, identify high-risk populations, and evaluate the effectiveness of prevention programs, including tobacco control, alcohol reduction, and HPV vaccination.

2.6 Autofluorescence and Diffuse Reflectance Spectroscopy

- **Publication:** Investigation of autofluorescence and diffuse reflectance spectroscopy for detecting oral cancer and dysplasia. Oral Oncology.
  - Source: https://www.sciencedirect.com/science/article/abs/pii/S1386505621001830
- **Application**: This research explores advanced optical techniques for improving the early detection of oral cancer and precancerous lesions.

- **Past and Ongoing Research**: Current research focuses on refining spectroscopic techniques, developing portable and cost-effective devices, and conducting clinical trials to validate their effectiveness in real-world settings.

2.7. Hyperspectral Image Analysis

- **Publication:** "Hyperspectral image analysis for oral cancer detection."
    - Source: https://ieeexplore.ieee.org/abstract/document/5986713
- **Application:** Development of a non-contact, reflection-mode hyperspectral imaging system to differentiate between healthy and cancerous tissue in the oral cavity.
- **Past and Ongoing Research:** Research in this area includes improving the accuracy and speed of hyperspectral imaging systems, developing more robust image processing algorithms, and integrating this technology into clinical practice for real-time diagnosis.

3. Dataset and Domain

3.1 Data Dictionary

- ID: Unique identifier for each record.
- Country: Country of the patient.
- Age: Age of the patient.
- Gender: Gender of the patient.
- Tobacco Use: Whether the patient has used tobacco (Yes/No).
- Alcohol Consumption: Whether the patient consumes alcohol (Yes/No).
- HPV Infection: Presence of Human Papillomavirus infection (Yes/No).
- Betel Quid Use: Whether the patient has used betel quid (Yes/No).
- Chronic Sun Exposure: Whether the patient has prolonged sun exposure (Yes/No).
- Poor Oral Hygiene: Whether the patient has poor oral hygiene (Yes/No).
- Diet (Fruits & Vegetables Intake): Dietary habits (Good/Poor).

- Family History of Cancer: Whether the patient has a family history of cancer (Yes/No).
- Compromised Immune System: Whether the patient has a weakened immune system (Yes/No).
- Oral Lesions: Presence of oral lesions (Yes/No).
- Unexplained Bleeding: Presence of unexplained bleeding in the mouth (Yes/No).
- Difficulty Swallowing: Whether the patient has trouble swallowing (Yes/No).
- White or Red Patches in Mouth: Presence of abnormal patches in the mouth (Yes/No).
- Tumor Size (cm): Size of the tumor in centimeters.
- Cancer Stage: Stage of oral cancer (0, 1, 2, 3, 4).
- Treatment Type: Type of treatment received (e.g., No Treatment, Surgery).
- Survival Rate (5-Year, %): Estimated 5-year survival rate.
- Cost of Treatment (USD): Cost of treatment in US dollars.
- Economic Burden (Lost Workdays per Year): Number of workdays lost per year.
- Early Diagnosis: Whether the cancer was diagnosed early (Yes/No).
- Oral Cancer (Diagnosis): Final diagnosis of oral cancer (Yes/No).

3.2 Variable categorization (count of numeric and categorical)

- Numeric: ID, Age, Tumor Size (cm), Survival Rate (5-Year, %), Cost of Treatment (USD), Economic Burden (Lost Workdays per Year)
- Categorical: Country, Gender, Tobacco Use, Alcohol Consumption, HPV Infection, Betel Quid Use, Chronic Sun Exposure, Poor Oral Hygiene, Diet (Fruits & Vegetables Intake), Family History of Cancer, Compromised Immune System, Oral Lesions, Unexplained Bleeding, Difficulty Swallowing, White or Red Patches in Mouth, Cancer Stage, Treatment Type, Early Diagnosis, Oral Cancer (Diagnosis)

3.3 Pre Processing Data Analysis

- Missing/Null Values: No missing values were found in the dataset.
- Redundant Columns: There are some redundant columns such as : ID and country columns are dropped before the model building.

3.4 Alternate sources of data that can supplement the core dataset

- National Cancer Institute (NCI) SEER Program: Provides detailed cancer statistics.
- World Health Organization (WHO) Global Cancer Observatory (GLOBOCAN): Offers global cancer incidence and mortality data.

3.5 Project Justification

- **Project Statement:**
  - Analysis of risk factors and prediction of oral cancer diagnosis and outcomes.
- **Complexity involved:**
  - High-dimensional data with mixed variable types.
  - Potential for class imbalance in the target variable.
  - Need to handle categorical variables and potential non-linear relationships.
- **Project Outcome:**
  - Commercial: Development of a tool to assist in early diagnosis.
  - Academic Value: Insights into the key risk factors and progression patterns of oral cancer.
  - Social value: Potential to improve patient outcomes and reduce the burden of oral cancer.

4. Data Exploration (EDA)

Exploratory Data Analysis (EDA) is a critical step in the data science pipeline that involves investigating the dataset to uncover initial patterns, spot anomalies, test hypotheses, and visualizations. The goal is to gain a better understanding of the data and prepare it for modelling.

In this project, EDA was performed to analyse the distribution of variables, identify the presence of outliers, assess correlations between features, and evaluate the significance of variables in relation to the target. This section outlines the major findings and insights derived from the data exploration process.

4.1 Relationship between variables

4.1.1 Multicollinearity

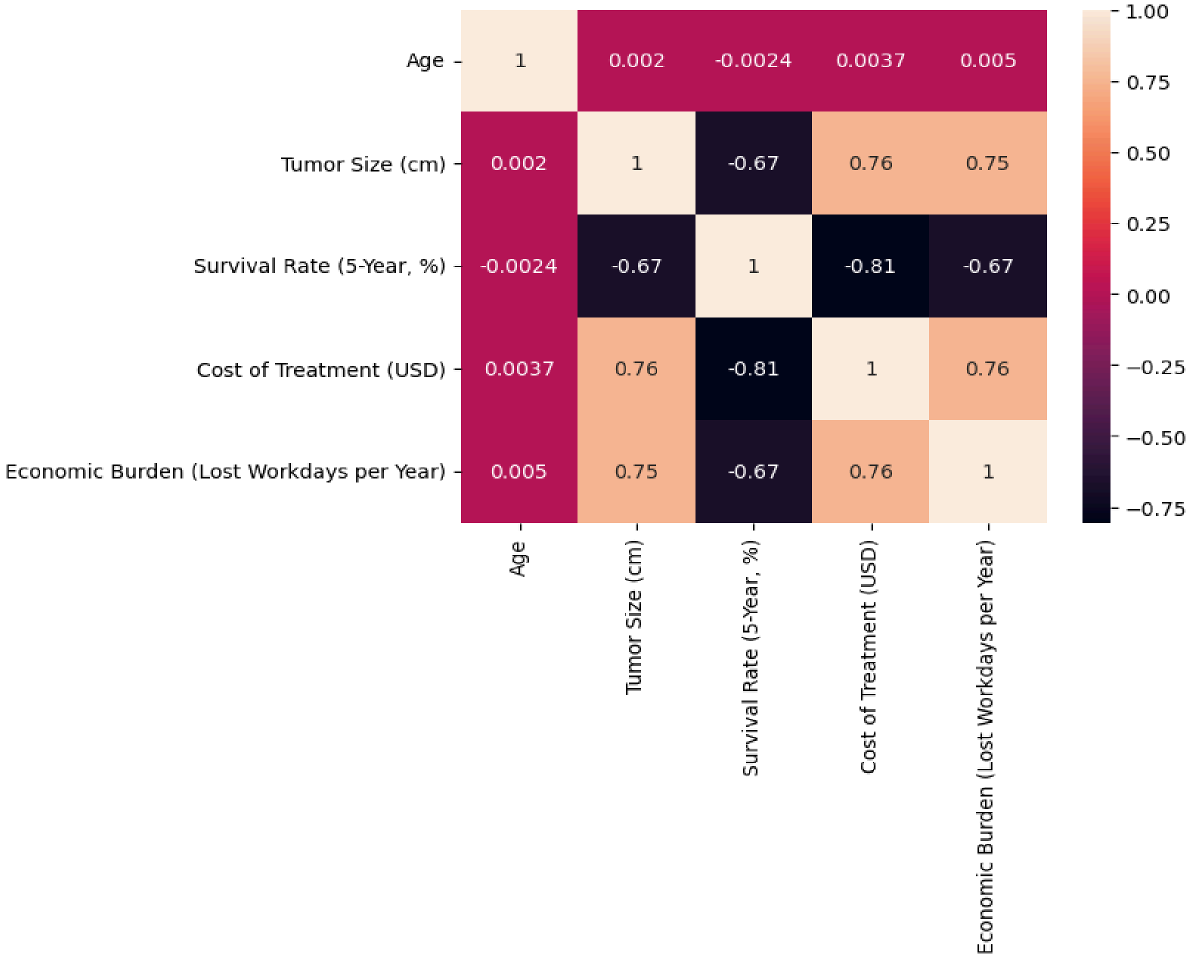

Fig-1 (Correlation between numerical variables)

- **Cost of Treatment & Tumor Size:** Larger tumors are associated with higher treatment costs (0.76 correlation).
- **Economic Burden:** Higher treatment costs and larger tumors lead to a greater economic burden (lost workdays) (0.75 and 0.76 correlations).
- **Survival Rate:** Higher treatment costs, larger tumor size, and greater economic burden are associated with lower 5-year survival rates (-0.81, -0.67, and -0.67 correlations, respectively). Cost of treatment has the strongest negative correlation with survival rate.

4.1.2 Distribution of variables

- **Numerical Features:**

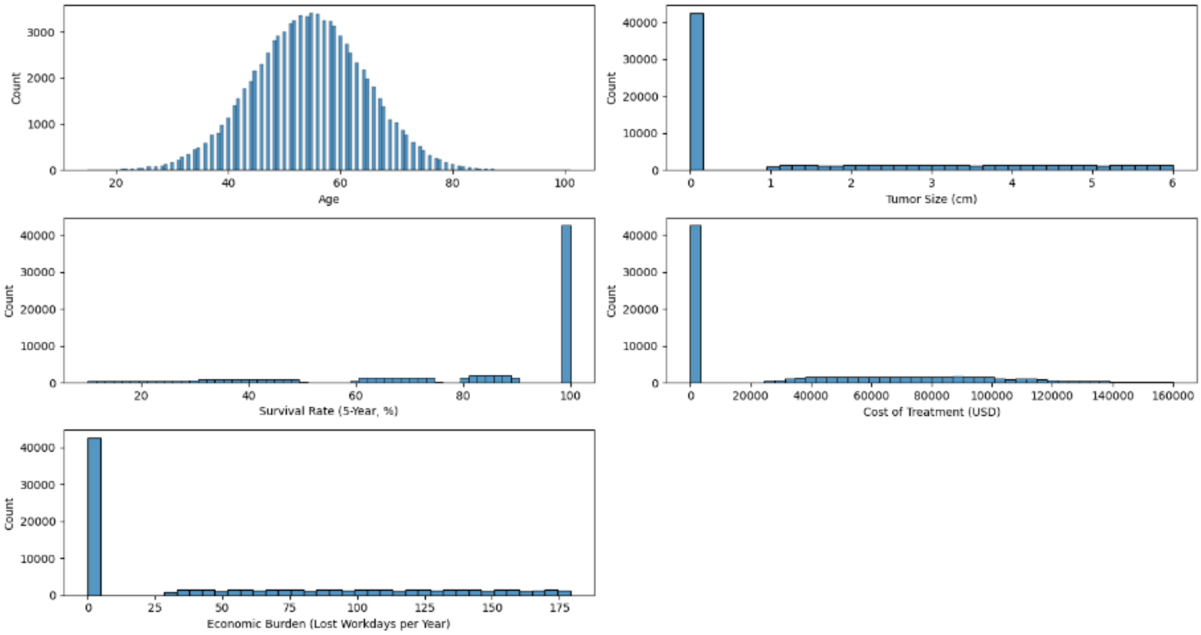

Fig-2 (Distribution of numerical columns)

- **Categorical Features:**

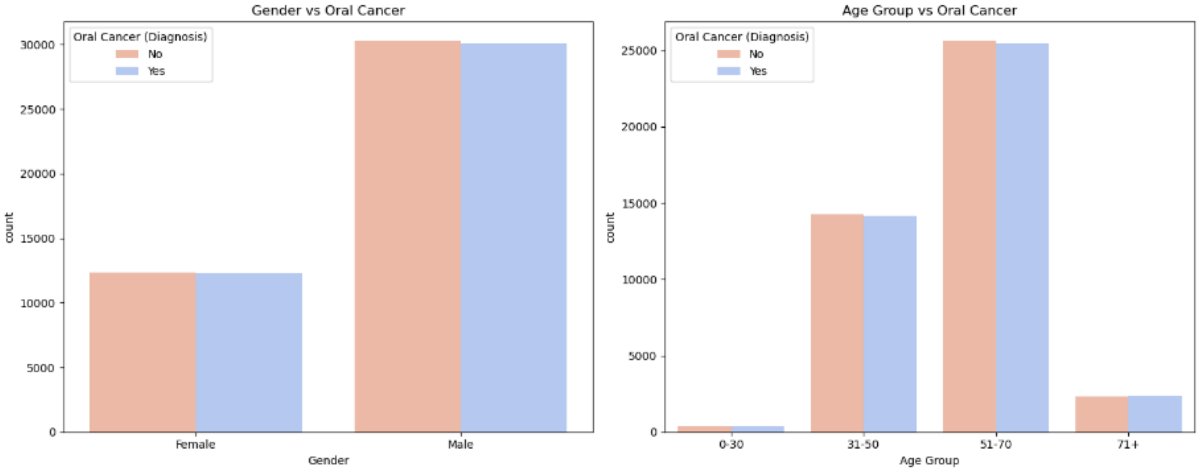

fig-3 (Checking different columns with the target)

- The chart shows that a greater number of males have been diagnosed with oral cancer than females
- The chart demonstrates that the highest incidence of oral cancer is found in the 51-70 age group, followed by the 31-50 age group, indicating that oral cancer is more common in older adults.

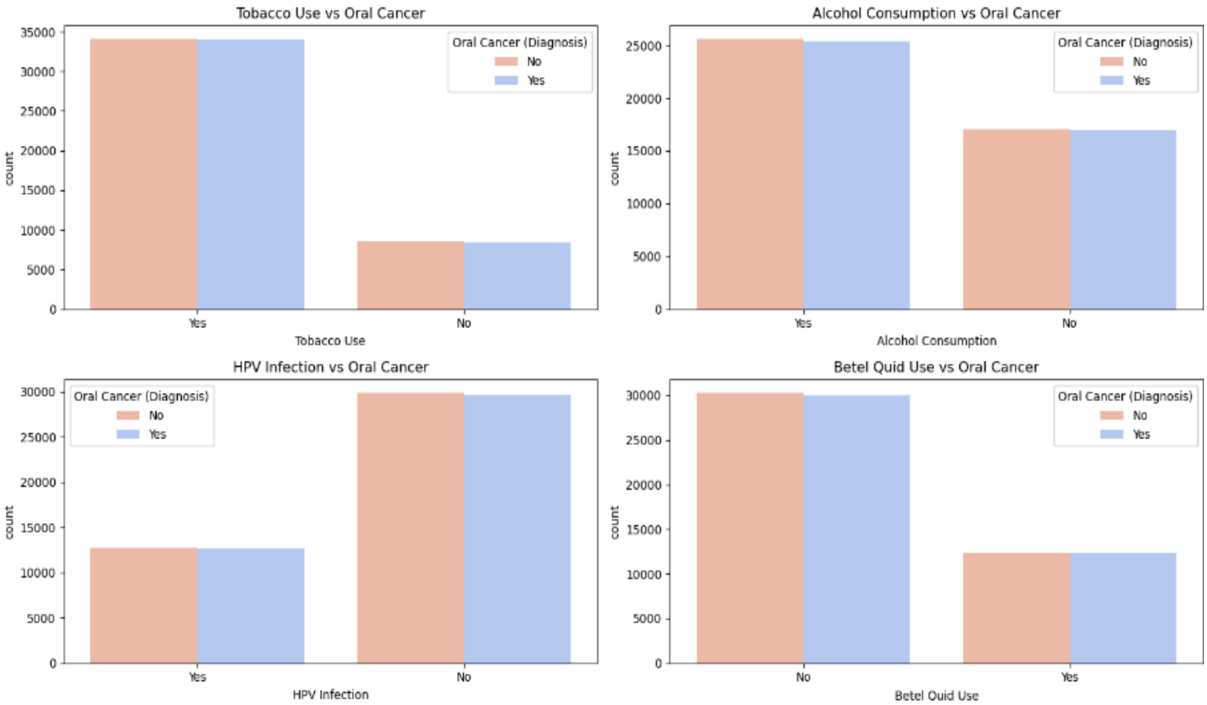

fig-3(a) (Checking different columns with the target)

- Tobacco Use vs Oral Cancer Diagnosis: Individuals who use tobacco show significantly higher instances of oral cancer compared to non-users.
- Alcohol Consumption vs Oral Cancer Diagnosis: Alcohol consumers have a higher prevalence of oral cancer diagnoses than non-consumers.
- HPV Infection vs Oral Cancer Diagnosis: There is a strong correlation between HPV infection and increased rates of oral cancer diagnosis.
- Betel Quid Use vs Oral Cancer Diagnosis: Betel quid users exhibit a markedly higher occurrence of oral cancer than non-users.

4.1.3 Presence of outliers and its treatment

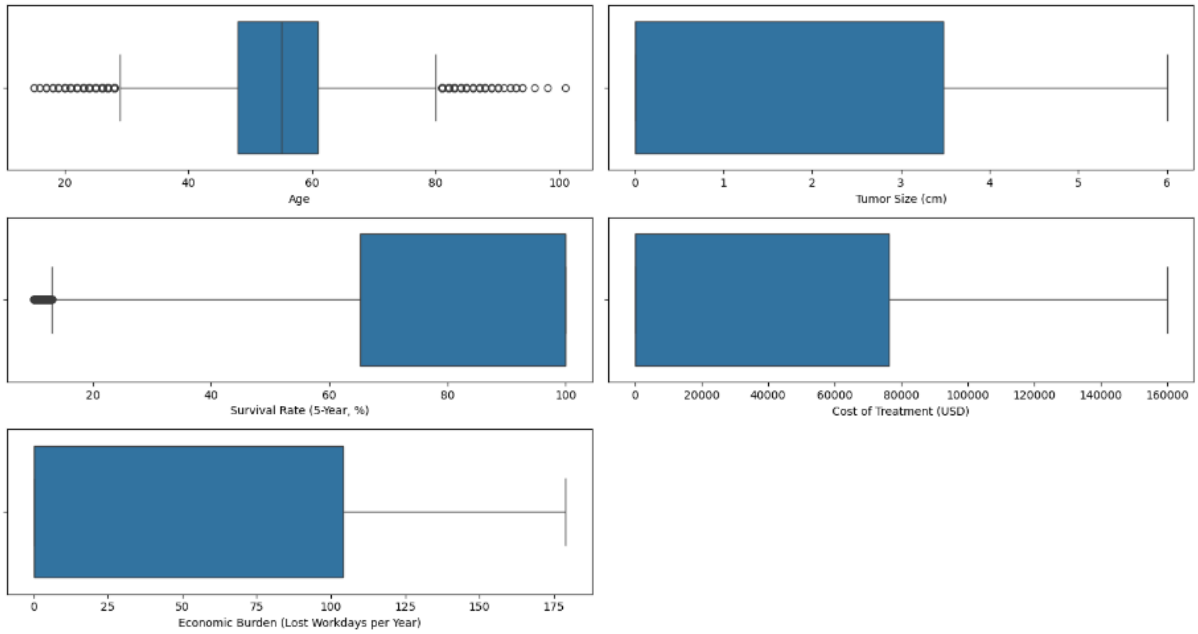

fig-4 (Checking outliers using box plot)

4.1.4 Statistical significance of variables

1. Numeric vs. Target Variable

For numeric columns, we used the independent samples t-test to compare the means of each numeric variable between the two groups of the target variable ('Oral Cancer (Diagnosis)' = 'No' and 'Yes').

- P value of Age is 0.7112023434523236

  There is no statistical significant association between the Age and Oral Cancer (Diagnosis)

- P value of Tumor Size (cm) is 0.0

  There is statistically significant association between the Tumor Size (cm) and Oral Cancer (Diagnosis)

- P value of Survival Rate (5-Year, %) is 0.0

  There is statistically significant association between the Survival Rate (5-Year, %) and Oral Cancer (Diagnosis)

- P value of Cost of Treatment (USD) is 0.0

  There is statistically significant association between the Cost of Treatment (USD) and Oral Cancer (Diagnosis)

- P value of Economic Burden (Lost Workdays per Year) is 0.0

There is statistically significant association between the Economic Burden (Lost Workdays per Year) and Oral Cancer (Diagnosis)

2. Categorical vs. Target Variable

For categorical columns, we used the Chi-squared test for independence. This test assesses the association between two categorical variables by comparing the observed frequencies to the frequencies expected under the assumption of no association.

- P value of Country is 0.35163122093418847
  There is no statistical significant association between the Country and Oral Cancer (Diagnosis)
- P value of Gender is 0.9204976552018163
  There is no statistical significant association between the Gender and Oral Cancer (Diagnosis)
- P value of Tobacco Use is 0.5864637294064725
  There is no statistical significant association between the Tobacco Use and Oral Cancer (Diagnosis)
- P value of Alcohol Consumption is 0.6457066556474733
  There is no statistical significant association between the Alcohol Consumption and Oral Cancer (Diagnosis)
- P value of HPV Infection is 0.9104274259235184
  There is no statistical significant association between the HPV Infection and Oral Cancer (Diagnosis)
- P value of Betel Quid Use is 0.6448383589549511
  There is no statistical significant association between the Betel Quid Use and Oral Cancer (Diagnosis)
- P value of Chronic Sun Exposure is 0.7863289778674327
  There is no statistical significant association between the Chronic Sun Exposure and Oral Cancer (Diagnosis)
- P value of Poor Oral Hygiene is 0.15653614324609877
  There is no statistical significant association between the Poor Oral Hygiene and Oral Cancer (Diagnosis)
- P value of Diet (Fruits & Vegetables Intake) is 0.6137104582742303
  There is no statistical significant association between the Diet (Fruits & Vegetables Intake) and Oral Cancer (Diagnosis)
- P value of Family History of Cancer is 0.7373912327789546
  There is no statistical significant association between the Family History of Cancer and Oral Cancer (Diagnosis)
- P value of Compromised Immune System is 0.11155068735152987
  There is no statistical significant association between the Compromised Immune System and Oral Cancer (Diagnosis)
- P value of Oral Lesions is 0.794868712800618

There is no statistical significant association between the Oral Lesions and Oral Cancer (Diagnosis)

- P value of Unexplained Bleeding is 0.951636963391214
  There is no statistical significant association between the Unexplained Bleeding and Oral Cancer (Diagnosis)
- P value of Difficulty Swallowing is 1.0
  There is no statistical significant association between the Difficulty Swallowing and Oral Cancer (Diagnosis)
- P value of White or Red Patches in Mouth is 0.523898932645378
  There is no statistical significant association between the White or Red Patches in Mouth and Oral Cancer (Diagnosis)
- P value of Cancer Stage is 0.0
  There is statistically significant association between the Cancer Stage and Oral Cancer (Diagnosis)
- P value of Treatment Type is 0.0
  There is statistically significant association between the Treatment Type and Oral Cancer (Diagnosis)
- P value of Early Diagnosis is 0.8378899394669141
  There is no statistical significant association between the Early Diagnosis and Oral Cancer (Diagnosis)

4.1.5 Class imbalance and its treatment

The distribution of the target variable 'Oral Cancer (Diagnosis)' is as follows: 49.8% 'No' and 50.1% 'Yes'. This balanced distribution indicates the absence of class imbalance.

5.Feature Engineering:

Feature engineering plays a vital role in enhancing model performance by optimizing the quality and relevance of the input features. In this project, the dataset was well-structured and mostly composed of categorical variables with clean labels. Therefore, extensive transformations or scaling were not required.

5.1. Transformations

The dataset primarily consists of binary and nominal categorical variables (e.g., Tobacco Use, Alcohol Consumption, Oral Lesions) with standardized "Yes" or "No" values. These were directly converted to numerical form using label encoding without the need for any complex transformations.

No logarithmic, square root, or polynomial transformations were necessary, as the features were already in interpretable and usable form.

5.2 Scaling

Feature scaling (e.g., Min-Max Scaling, Standardization) is typically required for distance-based algorithms such as SVMs or KNN. However, in this case, the majority of models used—Decision

Trees, Random Forests, and other ensemble methods—are not sensitive to feature magnitudes, making scaling unnecessary.

Even for numerical columns like Tumor Size (cm) and Cost of Treatment, scaling was skipped since tree-based models handle variable scales internally without affecting model performance.

5.3 Feature Selection

All available features were retained for model training. No feature selection methods (such as recursive elimination, filter methods, or model-based selection) were used. This decision was made because:

- The dataset had a manageable number of features (25 columns).
- Each feature was clinically or contextually relevant to oral cancer prediction.
- Tree-based models inherently manage feature relevance during training, reducing the need for pre-selection.

6. Assumptions

Logistic Regression:

- Assumes a linear relationship between the independent variables and the log-odds of the dependent variable.
- Requires independence of observations.
- Assumes little to no multicollinearity among the independent variables.

Decision Tree:

- Does not assume a specific distribution of data.
- Works well with both categorical and numerical data.
- Assumes that the features are relevant to the target variable.