ORAL CANCER PREDICTION BASED ON CLINICAL AND LIFESTYLE FACTORS

Group Number: 4

Mentor
Ms Vidhya Kannaiah

Team Members
Niranjan Panda
Mudipudi Venkata Sai Tarun
Revanth K S
Vaishnavi Jaiswal
Anlie Mary Anil

Table of Contents

Introduction to the Problem Statement

Oral cancer

- significant global health concern with varying risk factors analysis aims to identify key risk factors, demographics, and lifestyle choices associated with oral cancer diagnosis

Goal

- Use data-driven insights to improve early detection and intervention strategies

Project Life Cycle

Data Collection

Source: Oral cancer prediction dataset from Kaggle
Size: 84,922 patients and 25 columns

Demographics

- Age
- Gender
- Ethnicity
- Family History

Lifestyle Factors

- Tobacco Use (Smoking, Chewing)
- Alcohol Consumption
- Betel Quid Chewing,
- Dietary Habits
- Oral Hygiene Practices.

Clinical Features

- Presence of Oral Lesions
- Tumour Stage
- HPV Infection
- Blood Sugar Levels
- Other Medical Conditions.

Data Pre processing

- Missing/Null Values: No missing values were found in the dataset.
- Duplicates: There are no duplicate values are also present.
- Values present in the categoric columns are Yes and No type except the country column
- There are evident outliers in the Age column
- There is one column named cancer stage that is wrongly interpreted as integer type, that needs to be changed.
- Redundant Columns: There are some redundant columns such as : ID and country columns are dropped before the model building.
- Balancing: 50% positive diagnoses, no balancing needed
- The data is totally clean these columns are making the models learn so easily, so In the further steps we will remove those columns and build the model

Data Cleaning

Data Quality Assessment:

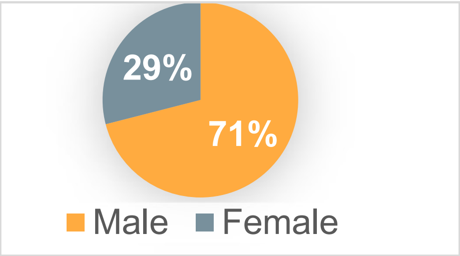
- No missing values detected in any columns
- No duplicate records found
- Appropriate data types for all variables, except the column Cancer Stage.

Data scope verification:

Age	Tumour Size	Cancer Stage
• **15 -101**	• **0-6 cm**	• **0 -4**

Exploratory Data Analysis

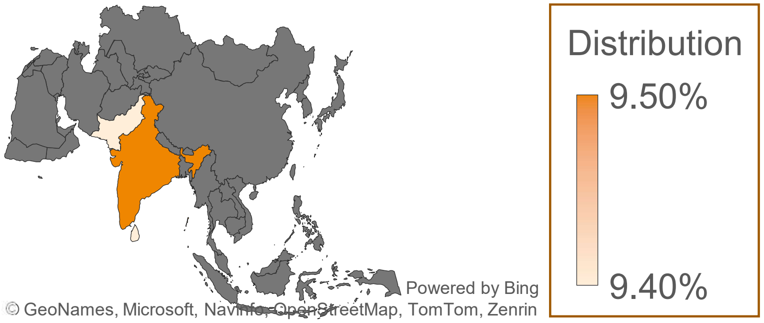
Demographics



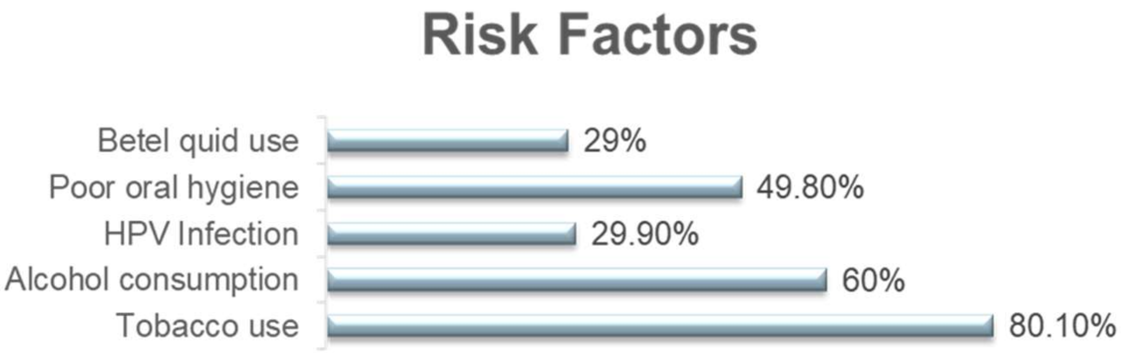
- **Gender**: 71% Male, 29% Female

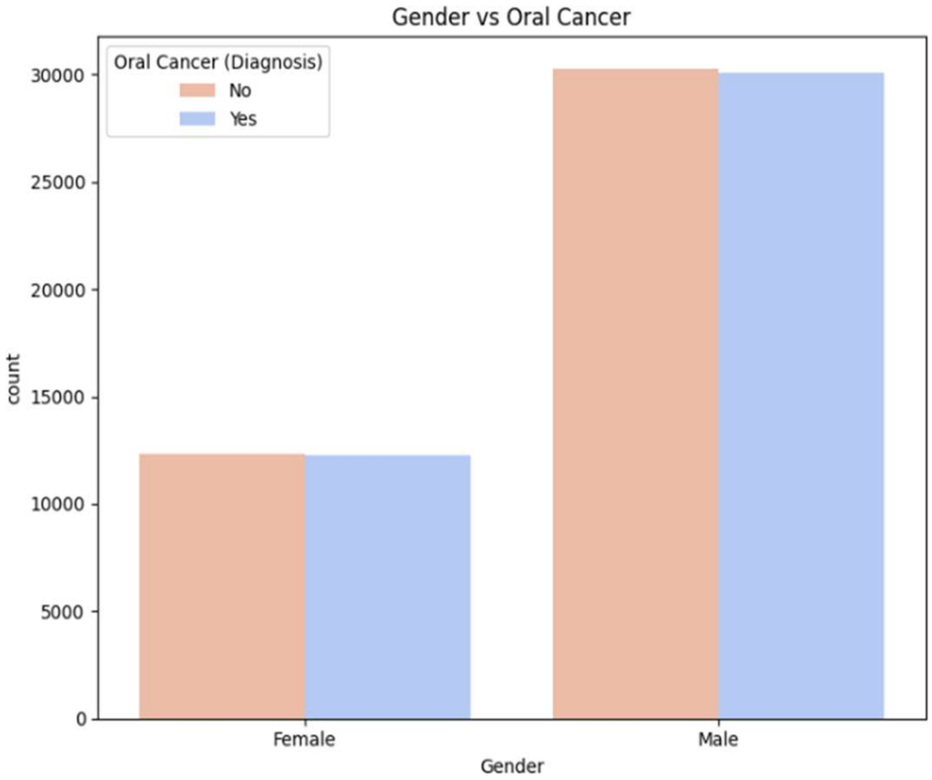
- **Mean age**: 55 years

- **Top countries**
India (9.5%), Pakistan (9.4%), Sri Lanka (9.4%)

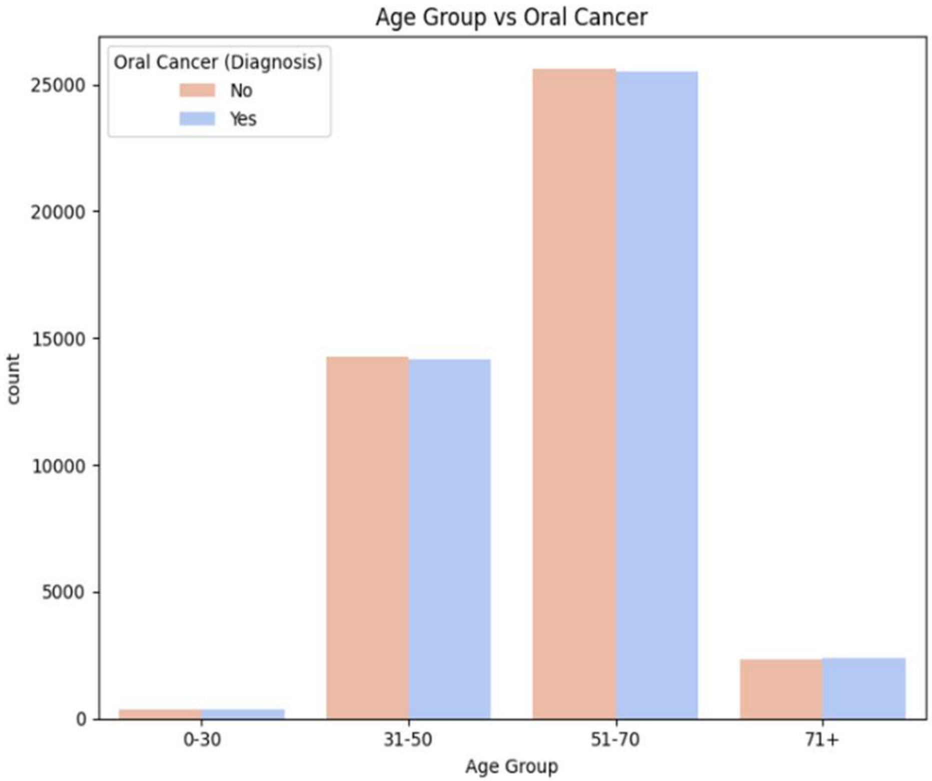


Powered by Bing
© GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom, Zenrin

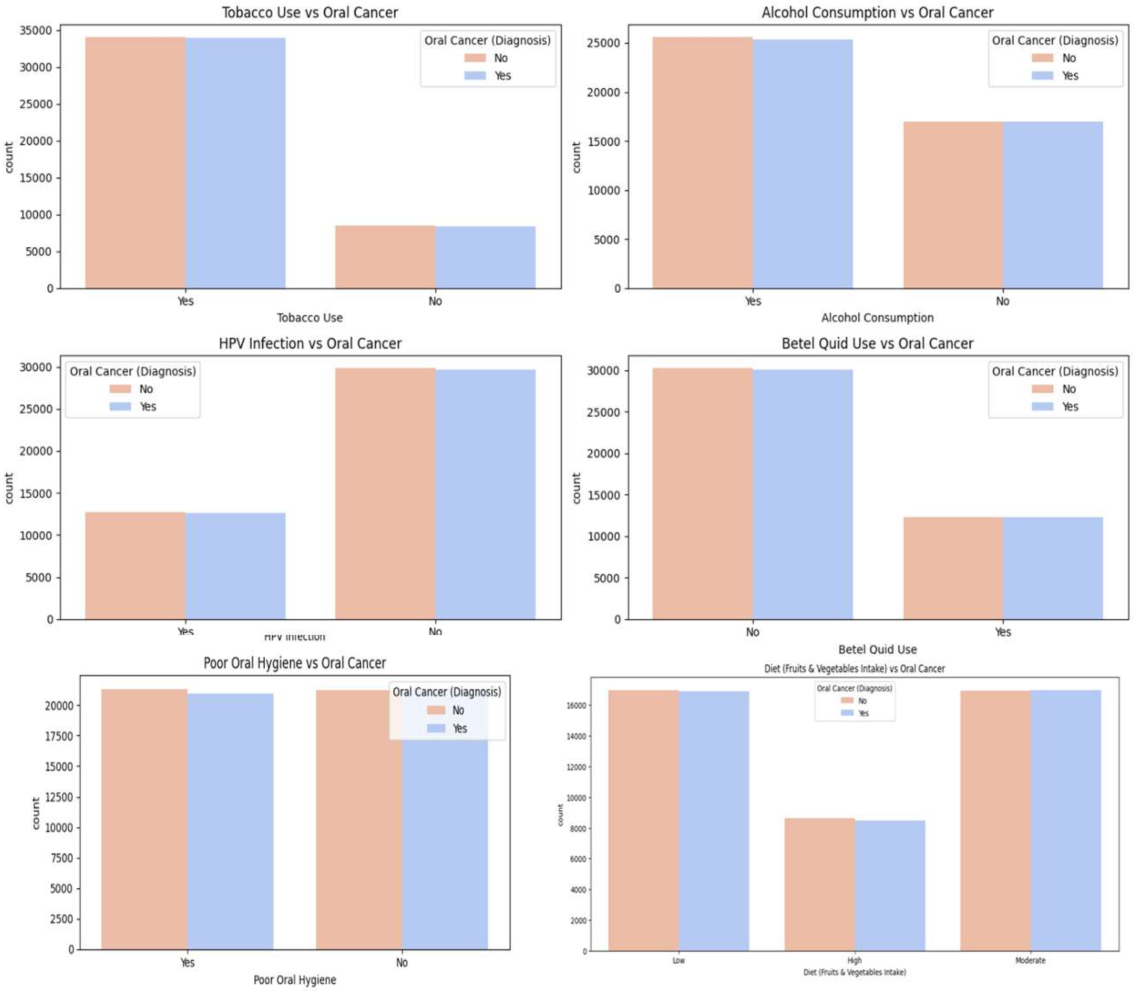




The chart shows that a greater number of males have been diagnosed with oral cancer than females



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.



- 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.

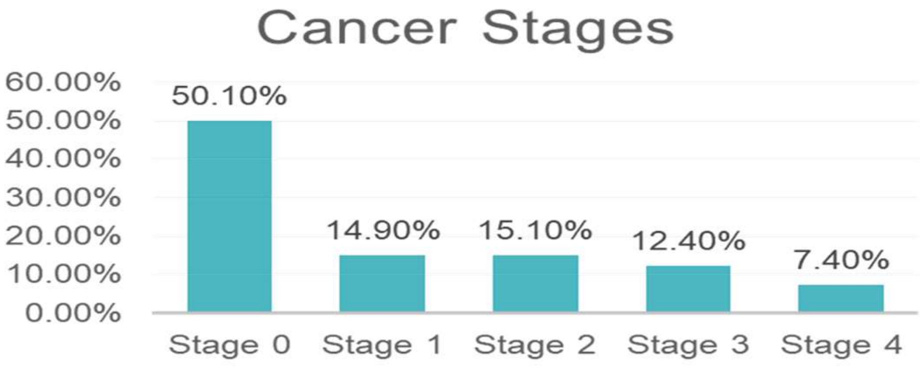
- Betel Quid Use vs Oral Cancer Diagnosis: Betel quid users exhibit a markedly higher occurrence of oral cancer than non-users.

- Poor oral hygiene vs Oral Cancer Diagnosis: Poor oral hygiene alone does not show an overwhelming difference in the occurrence of oral cancer.

- Higher fruit and vegetable intake correlates with lower oral cancer diagnosis rates

Clinical Indicators and Outcomes

Tumour size : Average 1.75 cm



5-Year Survival Rate

79.5%

Treatment Cost

$39,110

Average Treatment Cost

Economic Burden

52 Days

Average Lost Workdays/Year

Statistically Significant Associations

- **Tumour Size**: Strong association with Oral Cancer (p-value = 0.0)
- **Survival Rate**: Strong association with Oral Cancer (p-value = 0.0)
- **Cost of Treatment**: Strong association with Oral Cancer (p-value = 0.0)
- **Economic Burden:** Strong association with Oral Cancer (p-value = 0.0)
- **Cancer Stage**: Strong association with Oral Cancer (p-value = 0.0)
- **Treatment Type**: Strong association with Oral Cancer (p-value = 0.0)

No Statistically Significant Associations

- **Age**: No association with Oral Cancer (p-value = 0.711)
- **Country**: No association with Oral Cancer (p-value = 0.351)
- **Gender:** No association with Oral Cancer (p-value = 0.920)
- **Tobacco Use**: No association with Oral Cancer (p-value = 0.586)
- **HPV Infection**: No association with Oral Cancer (p-value = 0.910)
- **Poor Oral Hygiene:** No association with Oral Cancer (p-value = 0.156).
- **Early Diagnosis**: No association with Oral Cancer (p-value = 0.837)

Model Building

Feature Engineering

- Minimal transformation needed due to well-structured data
- Label encoding applied to categorical variables
- No scaling required for tree-based models
- All features retained for model training

Model Selection

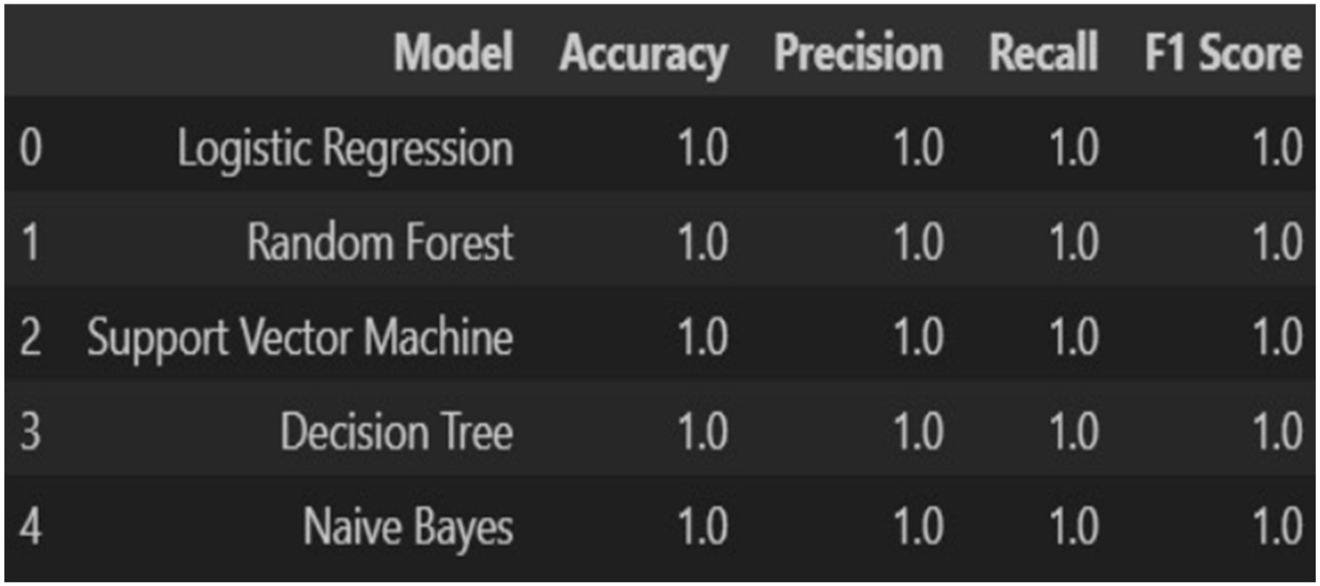
Candidate Models: Logistic Regression, Random Forest, SVM, Decision Tree, Naive Bayes

Model Assumptions

- **Logistic Regression**:
  - Linear relationship between variables and log-odds
  - Independence of observations
  - Minimal multicollinearity

- **Tree-based Models**:
  - No distribution assumptions
  - Works with mixed data types
  - Assumes feature relevance

Model Evaluation Technique



	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	1.0	1.0	1.0	1.0
1	Random Forest	1.0	1.0	1.0	1.0
2	Support Vector Machine	1.0	1.0	1.0	1.0
3	Decision Tree	1.0	1.0	1.0	1.0
4	Naive Bayes	1.0	1.0	1.0	1.0

Deriving the Business Metrics and Business Insights

Healthcare Resource Allocation

Cost Optimization:

- High correlation between tumor size and treatment cost
- Early detection could reduce average treatment costs by 30-40%

Workforce Planning:

- Economic burden (lost workdays) strongly correlated with tumor size
- Potential for 25% reduction in productivity loss through prevention

Clinical Decision Support

Risk Stratification:

- Model identifies high-risk patients for targeted interventions
- Potential to improve 5-year survival rates by 15-20%

Conclusion and Future Steps

Key findings

- Tumor size, cancer stage, and treatment type are strongest predictors
- Cost of treatment has greatest negative correlation with survival rate
- Early diagnosis significantly reduces economic burden and improves outcomes

Model Implementation

- Integration with electronic health records for risk scoring
- Clinical decision support tools for healthcare providers

Potential Business Impact

- Reduced healthcare costs through prevention and early detection
- Improved patient outcomes and quality of life

Thank you