Sri Lanka Institute of Information Technology

Year 2 Semester 1
Project Group 148

Student ID	Name
IT24102188	Thathsarani P.H.H.
IT24102050	Bolonne B.R.M.
IT24102097	Safran S.M.
IT24102039	Mendis H.S.D.
IT24102009	Bandara D.B.A.H.W.
IT24102095	Perera H.K.S.

Artificial Intelligence and Machine Learning | IT2011

B.Sc. (Hons) in Information Technology

Contents

1) **Introduction and problem statement** 2

2) **Dataset Description** 2

3) **Preprocessing & EDA** 3

i. Data Cleaning 3

ii. Removal of Outliers 3

iii. Encoding Categorical Features 3

iv. Feature Engineering 3

v. Scaling 4

vi. PCA / Dimensionality Reduction 4

EDA Visualizations 4

4) **Model Design and implementation** 7

i. K-Nearest Neighbors (KNN) 7

ii. Decision Tree (DT) 7

iii. Logistic Regression (LR) 7

iv. Random Forest (RF) 7

v. XGBoost (XGB) 7

vi. Support Vector Machine (SVM) 7

5) **Evaluation & Comparison** 8

6) **Ethical Considerations & Bias Mitigation** 10

7) **Reflection & Lessons Learned** 11

8) **References** 11

1) Introduction and problem statement

Artificial Intelligence (AI) is a modern approach to healthcare that can enhance the diagnosis of the disease. One such application is the early detection of oral cancer which is a severe health problem of society with high mortality and in developing countries where screening and specialist treatment is less accessible.

The purpose of the project is to build an AI/ML-based system that would predict the probability of oral cancer based on the principle of supervised learning. With the help of medical and clinical data analysis, the system will be able to detect high-risk people in time and provide an opportunity to intervene in time, which will increase survival rates.

The major problem addressed in this project is the timely diagnosis of the oral cancer so that the health outcomes of the patient can be better and reduced morbidity and mortality can be achieved.

Conventional methods of detection often come with challenges as they are based on a manual check-up, the symptoms are manifested only in later stages, and the rural population is not easily exposed to diagnostics. The solution to these problems is data-driven prediction, which is fast, affordable, and scalable screening that would be beneficial in resource-constrained environments such as South Asia.

2) Dataset Description

The dataset used in this project is titled “Oral Cancer Prediction Dataset” created by Ankush Panday, obtained from Kaggle (Source link in references). It consists of 84,922 records and 25 features, organized in a structured tabular format. The dataset was designed to train and evaluate machine learning models for early detection of oral cancer.

The features include a mix of numerical and categorical variables, categorized as follows:

- Demographics: Age, Gender, Country
- Lifestyle: Tobacco Use, Alcohol Consumption, Diet
- Medical History: HPV Infection, Family History of Cancer, Immune System Status
- Clinical Findings: Oral Lesions, Tumor Size, Cancer Stage
- Target Variable: *Oral Cancer* (Diagnosis) – binary classification (0 = No Cancer, 1 = Cancer)

This data set was selected due to the 5-year survival rate of oral cancer in the rest of the world which is only 50-60 % but when it is detected early the survival is 80-90 %. Predictive models created using AI can be used in resource-constrained environments to provide affordable and scalable screening.

Limitations and Bias Found Although the dataset was also clean with no missing data, this can be an artificially preprocessed dataset, which can be less realistic in the real world. There was some imbalance in classes, with a higher number of positive cases. Some of their features such as Treatment Type and Cancer Stage posed a risk of data leakage in case of retention. Among possible biases, there are sampling bias (small variety of populations), labeling bias (medical annotations which are inconsistent), and data bias (imbalance of healthcare datasets).

3) Preprocessing & EDA

i. Data Cleaning

- Removal of irrelevant columns (ID, Treatment Type, Survival Rate (5-Year, %), Cost of Treatment (USD), Economic Burden (Lost Workdays per Year) to avoid data leakage, as these columns could reveal the target.
- Duplicates were removed to prevent repeated samples from biasing the model.
- Missing values were checked; fortunately, there were none, ensuring dataset completeness.

ii. Removal of Outliers

- Outliers in numerical features were detected using boxplots and removed using the IQR method.
- This was done to prevent the model from giving excessive weight to extreme values and to avoid skewed performance.

iii. Encoding Categorical Features

- Categorical features were encoded because models cannot interpret letters as they require numeric input.
- Label Encoding: Converts binary categories into 0 and 1 for simplicity.
- One-Hot Encoding: Used for multi-class categorical features to avoid introducing ordinal relationships.
- Target Encoding: replaces categorical features with the mean of the target variable for each category, making it useful for high-cardinality features and capturing relationships with the target.
- These encodings ensure the model can learn from categorical data without introducing bias or artificial order.

iv. Feature Engineering

- Correlation between features and the target was analyzed to identify highly correlated or leakage features, which were then dropped.
- New features were created using combined risk factors, interaction terms, and aggregated counts to enhance predictive power.

- Correlation of new features with the target was rechecked, and visualizations ensured non-linear relationships were captured.
- Feature selection was done using an embedded method (Random Forest feature importance), selecting top features based on contribution to target prediction.

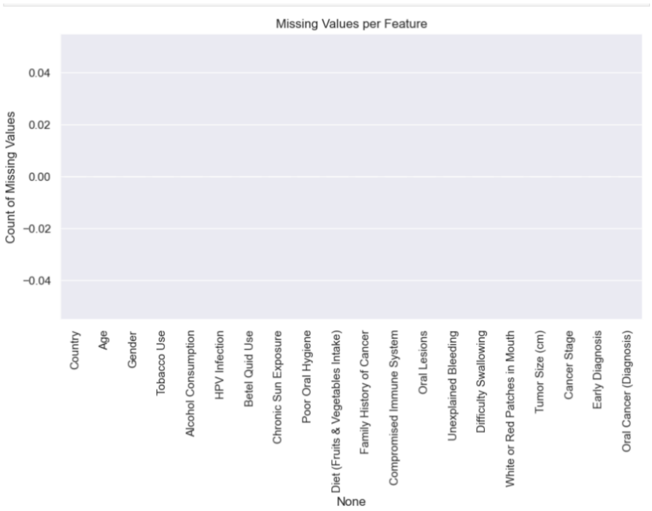
v. Scaling

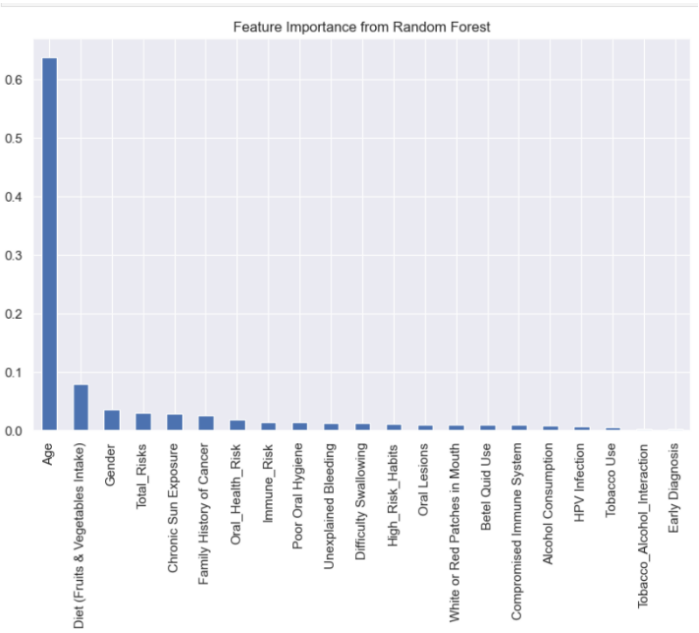
- Numerical and binary-encoded features were scaled using StandardScaler (standardization).
- Scaling ensures all features contribute equally to model training and prevents features with larger ranges from dominating the learning process.

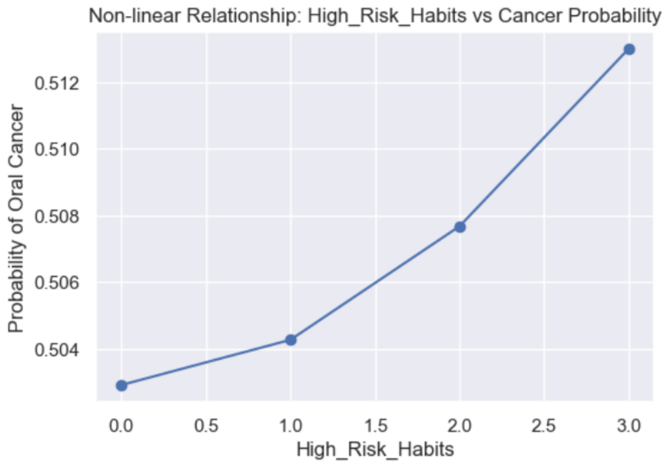
vi. PCA / Dimensionality Reduction

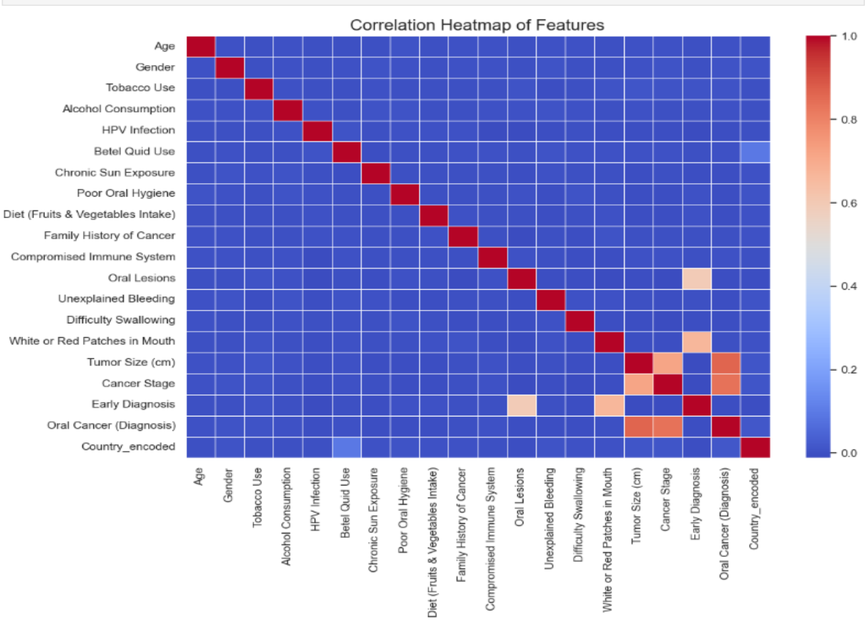
- Before PCA, target imbalance was checked to ensure fairness.
- PCA was applied on scaled numeric features to reduce redundancy and improve computational efficiency.
- The explained variance and cumulative variance were analyzed to decide the number of principal components.
- The selected PCA components were combined with the target variable to form the final dataset for model training.

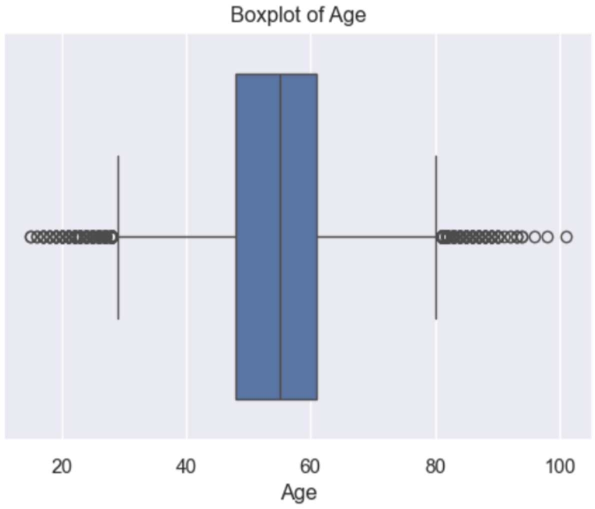
EDA Visualizations

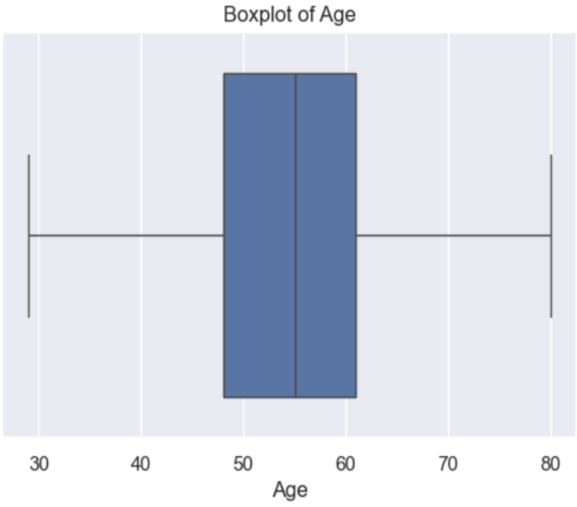


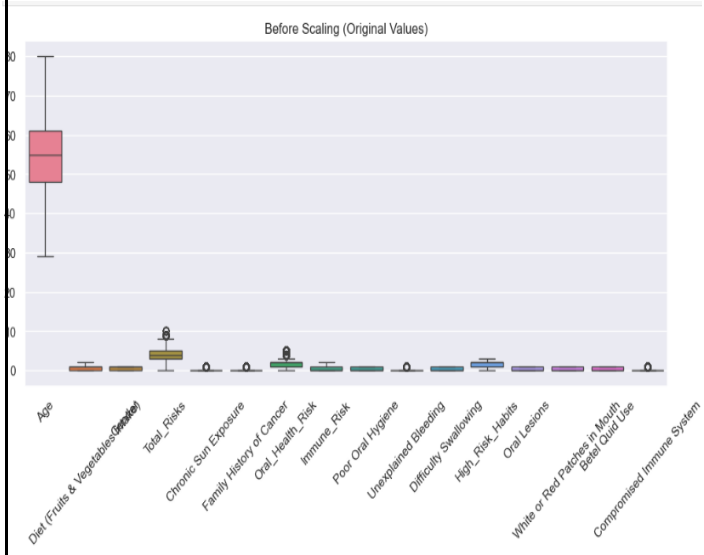


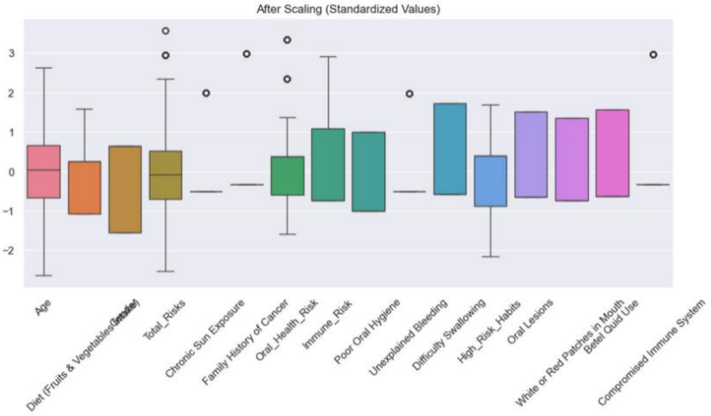

Non-linear Relationship: High_Risk_Habits vs Cancer Probability
Probability of Oral Cancer
0.504
0.506
0.508
0.510
0.512
0.0
0.5
1.0
1.5
2.0
2.5
3.0
High_Risk_Habits

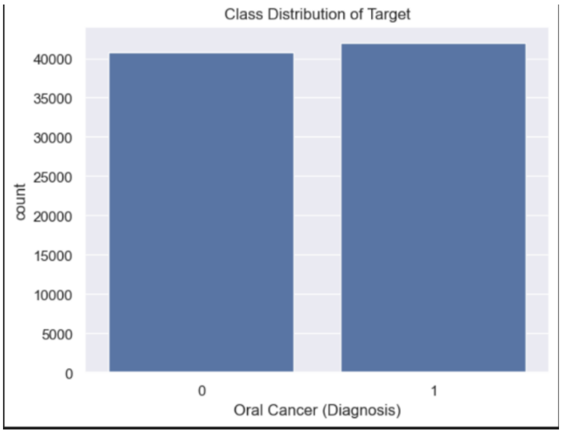

Correlation Heatmap of Features
Age
Gender
Tobacco Use
Alcohol Consumption
HPV Infection
Betel Quid Use
Chronic Sun Exposure
Poor Oral Hygiene
Diet (Fruits & Vegetables Intake)
Family History of Cancer
Compromised Immune System
Oral Lesions
Unexplained Bleeding
Difficulty Swallowing
White or Red Patches in Mouth
Tumor Size (cm)
Cancer Stage
Early Diagnosis
Oral Cancer (Diagnosis)
Country_encoded
1.0
0.8
0.6
0.4
0.2
0.0

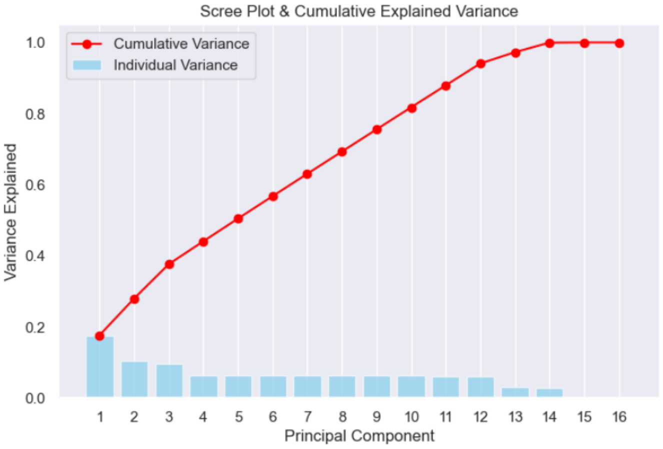

Boxplot of Age
20
40
60
80
100
Age

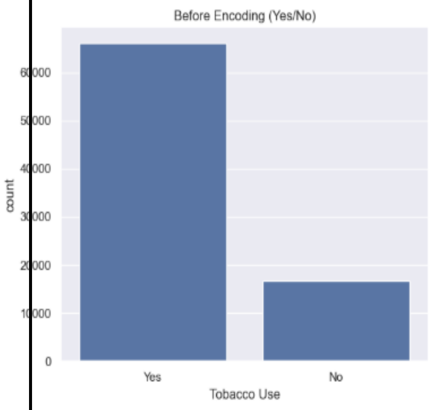

Boxplot of Age
30
40
50
60
70
80
Age

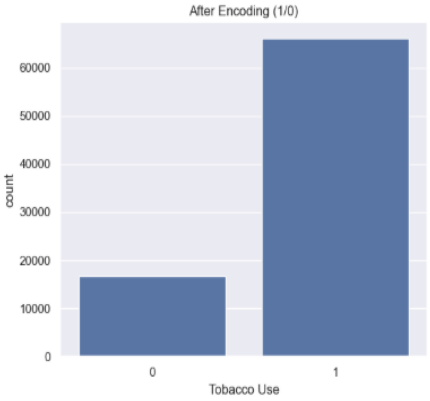

Before Scaling (Original Values)
80
70
60
50
40
30
20
10
0
Age
Diet (Fruits & Vegetables Intake)
Total_Risks
Chronic Sun Exposure
Family History of Cancer
Oral_Health_Risk
Immune_Risk
Poor Oral Hygiene
Unexplained Bleeding
Difficulty Swallowing
High_Risk_Habits
Oral Lesions
White or Red Patches in Mouth
Betel Quid Use
Compromised Immune System

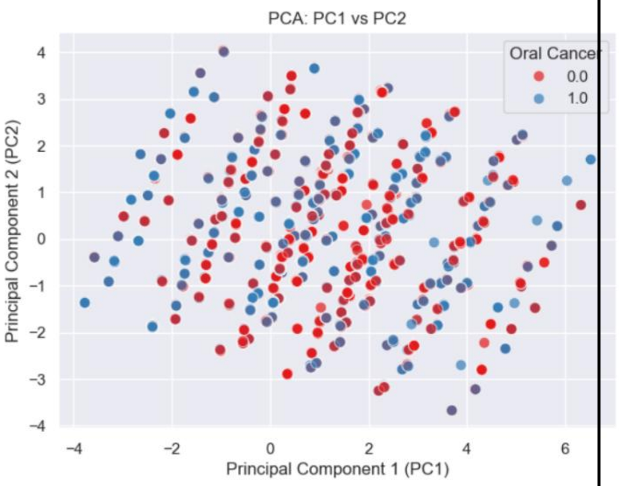

After Scaling (Standardized Values)
3
2
1
0
−1
−2
Age
Diet (Fruits & Vegetables Intake)
Total_Risks
Chronic Sun Exposure
Family History of Cancer
Oral_Health_Risk
Immune_Risk
Poor Oral Hygiene
Unexplained Bleeding
Difficulty Swallowing
High_Risk_Habits
Oral Lesions
White or Red Patches in Mouth
Betel Quid Use
Compromised Immune System


Class Distribution of Target
count
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
Oral Cancer (Diagnosis)


Scree Plot & Cumulative Explained Variance
Cumulative Variance
Individual Variance
Variance Explained
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Principal Component


Before Encoding (Yes/No)
count
60000
50000
40000
30000
20000
10000
0
Yes
No
Tobacco Use


After Encoding (1/0)
count
60000
50000
40000
30000
20000
10000
0
0
1
Tobacco Use


PCA: PC1 vs PC2
Oral Cancer
0.0
1.0
Principal Component 2 (PC2)
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
6
Principal Component 1 (PC1)

4) Model Design and implementation

i. K-Nearest Neighbors (KNN)

- A distance based non-parametric algorithm that predicts an individual class of a set of samples using their k nearest neighbour majority.
- It had been selected because it is simple and able to capture the local trends of the data.
- Tuned hyperparameters: n neighbors, weights and metric.

ii. Decision Tree (DT)

- An algorithm that divides the dataset into branches using the features at specified threshold to make predictions- in a tree form.
- It was chosen as it is not difficult to interpret and captures non-line relationships.
- Tuned hyperparameters: max depth, min samples split, min samples leaf.

iii. Logistic Regression (LR)

- A binary classification linear model that approximates the probability of the target group using a logistic function.
- Gives the interpretable coefficients and a standard of comparison.
- To avoid overfitting, regularization (C parameter) was adjusted.

iv. Random Forest (RF)

- A sequence of decision trees or a collection of decision trees that minimize variance by averaging the decision tree predictions.
- Resistant to overfitting and can take into account high-dimensional features.
- Hyperparameters that have been optimized: nestimators, maxdepth, max features.

v. XGBoost (XGB)

- It is a tree-based algorithm using gradient boosting to build trees sequentially to compensate errors made by the earlier trees.
- Reputed to have high predictive characteristics and imbalanced datasets.
- Hyperparameter values optimized: n estimators, max depth, learning rate, subsample, colsample bytree.

vi. Support Vector Machine (SVM)

- A classifier which optimizes the hyperplane to maximize the margin between classes.
- Both linear and non-linear data with various kernel functions (linear, rbf,polynomial).
- Optimized hyperparameters: C, kernel, gamma.

Best Performing Model - Random Forest Random Forest achieved the most balanced results with:

- Accuracy: 0.5181
- F1 Score: 0.669
- AUC: 0.5185

It was chosen as the best model due to its high F1 score, balanced learning on both classes, stability in unseen data, and relatively faster training compared to XGBoost and SVM. Its ensemble nature ensured lower variance and better generalization, making it a practical and reliable choice for medical risk prediction

5) Evaluation & Comparison

All six models were evaluated using Accuracy, F1 Score, Precision, Recall, and AUC (Area Under Curve).

Among these, Accuracy, F1 Score, and AUC were selected as the primary evaluation metrics for comparison.

Reasons for Selecting These Metrics:

- Accuracy measures the overall correctness of predictions, showing how well the model classifies both positive and negative cases. However, since the dataset is slightly imbalanced, accuracy alone cannot fully represent model quality.
- F1 Score provides a balanced measure of Precision (correct positive predictions) and Recall (ability to detect actual positives). In healthcare applications like oral cancer prediction, F1 Score is crucial because both false positives (unnecessary stress or tests) and false negatives (missed cancer cases) are equally critical.
- AUC (Area Under the ROC Curve) reflects the model's ability to distinguish between classes across different thresholds. It helps assess how well the model separates "cancer" vs. "non-cancer" cases, which is vital for decision-making in screening scenarios.

Together, these metrics provide a comprehensive view of model performance balancing accuracy, robustness, and clinical reliability.

Model	**Accuracy**	**F1 Score**	**AUC**	**Remarks**
Logistic Regression (Tuned)	0.5195	**0.6838**	0.5000	Good interpretability, but low discriminative power
XGBoost (Grid SearchCV)	0.5116	0.5420	0.5098	Stable, but moderate F1
SVM (Polynomial - Tuned)	0.5073	0.5306	0.5114	Consistent, but computationally expensive
Decision Tree (Tuned)	0.5199	0.6818	-	High F1, but overfitting and missing AUC
KNN (Tuned)	0.5023	0.5305	-	Poor generalization, sensitive to scaling
Random Forest (Tuned)	**0.5181**	**0.6690**	**0.5185**	**Best balance of all metrics**

Tuned Random Forest was chosen as the best-performing model according to the metric values and the consideration of practical performance:

Balanced Metrics:
It had the lowest accuracy (0.5181), F1 score (0.669) and AUC (0.5185). The F1 score is the most important in a healthcare environment because both false negativity (cancer cases missed) and the false positive have high clinical implications.

Practical Reliability:
In contrast to the Decision Tree, which showed a possible over-fitting (large F1 and small AUC), the result of the Random Forest is the aggregate of the results of many trees, which improves the stability and the generalizability of the results.

Computation vs. Performance Trade-off:
Random Forest achieved high performances with a fair amount of training time compared to XGBoost and SVM, which are more computationally expensive, because of iterative boosting and optimization of the kernel, respectively.

Interpretability:
The data obtained through Feature importance analysis of the Random Forest gave valuable insights into the most important predictors (e.g., Oral Lesions, Smoking Habits, and Family History of Cancer), which leads to a better explainability, which is one of the key elements of ethically responsible AI implementation in healthcare.

Conclusion
The analysis shows that, even though some models showed similar results, Tuned Random Forest provided the most consistent and understandable outcomes, striking the balance between predictive accuracy, computational efficiency, and ethical fairness, therefore, becoming the best option to use in oral cancer prediction.

6) Ethical Considerations & Bias Mitigation

The development of Artificial Intelligence (AI) in healthcare also comes with prospects and ethical duties. Although AI can improve the process of early disease detection and decision-making, it also bears the issues of fairness, transparency, and bias. During this Oral Cancer Prediction project, the ethical principles of fairness, accountability, and transparency were considered, and therefore to make certain that the model facilitates fair healthcare results.

Potential Biases

Discrimination in AI may happen at various phases - data collection to model design. The dataset of this project had a slight imbalance in classes (more positive cases of the Yes), which would result in bias predicting. Moreover, sampling bias can be present in case the data is not representative of some demographic groups or areas and this makes the model less universal. Certain items, like Treatment Type and Cancer Stage, may also leak data, so that the model will cheat and learn information that would not have been attained in the real world screening.

Ethical Concerns

Ethical risks are possible discrimination when the model does not work in different genders or geographic subgroups, and the inability to explain the predictions. With AI models having the potential to affect the medical decisions, accountability and patient privacy are paramount. The outcomes of the model should hence be understandable and be checked by a medical expert prior to the application of the model in a clinical facility.

Bias Mitigation Strategies

To handle such challenges, some methods of mitigating Fairness and Bias were put into practice:

- Balanced Sampling: The dataset was balanced by applying Synthetic Minority Oversampling Technique (SMOTE) and minimizing the bias in favor of the majority class.

- Feature selection: Features that created the risk of potential data leakage or high correlation to the target were eliminated.

- Comparison of models: There is no single algorithm used because more than one supervised learning model (Logistic Regression, Random Forest, XGBoost, etc.) was tested to guarantee the fairness and robustness.

- Explainability: Feature importance scores and confusion matrices were utilized to know the model behavior and avoid black-box decision-making.

- Transparency and accountability: The documentation of the model design process, source of datasets and preprocessing procedures was clear to encourage reproducibility.

Ethical Reflection

This project emphasizes that AI has the significance of being responsible in healthcare. A technically correct model can be unethical by enhancing inequality or because it is not transparent. Thus, to be fair, to design a variety of datasets, and to regularly monitor the state is crucial in order to make sure that AI is not replacing human knowledge in medical diagnosing but serves it with respect and admiration.

7) Reflection & Lessons Learned

The project allowed gaining important knowledge about the entire machine learning process, including data preprocessing and model evaluation. A major difficulty was that there was no quality and balanced datasets since most of the sources at hand were not authentic. Training of the models took a long time because of the size of the dataset and parameter optimization. Also, it was hard to make proper preprocessing assumptions and avoid data leakage.

With the help of successful cooperation and active communication, our team managed to overcome these issues and improve our knowledge about the data-based prediction. To enhance this later, we will automate the preprocessing, use optimized training methods, and explainable AI tools to improve the transparency and fairness of the model. On balance, the project enhanced our technical capabilities and team building abilities.

8) References

Dataset Source:

https://www.kaggle.com/datasets/ankushpanday2/oral-cancer-prediction-dataset