Oral Cancer Prediction Using Machine Learning

This project explores machine learning methods to predict oral cancer early, aiding timely diagnosis and treatment.

Presented by Sahib Chouhan and Gargi Sharma under Mr. Ashwin RamKishor Pal.

Introduction to Oral Cancer

What is Oral Cancer?

Cancer impacting mouth, lips, tongue, and gums; often detected late.

Importance of Early Prediction

Early detection boosts survival rates; Machine learning aids early diagnosis using patient data.

Problem and Machine Learning Solution

Problem

Delayed diagnosis reduces treatment success. Need early risk prediction from patient features.

Why Machine Learning?

- Handles large, complex data
- Detects subtle patterns missed by traditional methods
- Random Forest excels with non-linear data

Dataset Overview

Data Source

Collected from trusted medical databases.

Size

85,000 samples, 25 features.

Key Features

- Age, Gender
- Tobacco, Alcohol Use
- HPV, Betel Quid Use
- Tumor Size, Cancer Stage

Target Variable

Oral Cancer presence (1 = Yes, 0 = No)

Data Preprocessing Steps

Handling Missing Data

Missing values removed or imputed for consistency.

Label Encoding

Categorical variables converted to numeric (e.g. Gender: 1=Male, 0=Female).

Feature Scaling

Numerical data scaled using StandardScaler for uniformity.

Model Selection: Random Forest Classifier

Robust & Accurate

Ensemble of trees reduces overfitting and enhances prediction accuracy.

Non-linearity

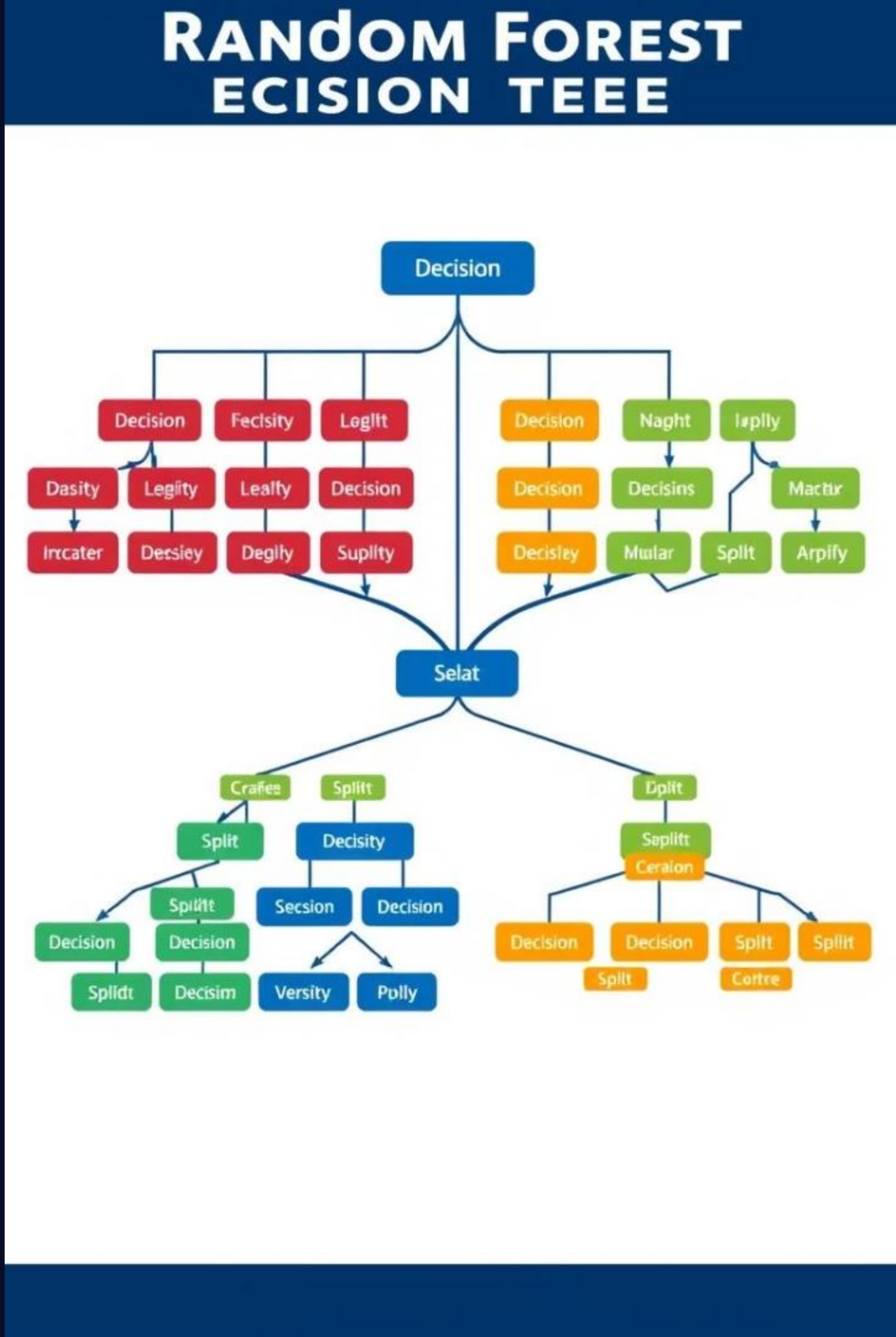
Captures complex, non-linear relationships beyond linear models like logistic regression.

Feature Importance

Highlights key predictors that influence the risk of oral cancer.

Interpretability

Provides decision rules enabling explainable and transparent results.



Model Training Methodology

Train-Test Split

Data divided (commonly 70/30) to train and evaluate model performance.

Cross-Validation

Multiple training tests on data subsets prevent overfitting and improve generalization.

Evaluation Metrics

Accuracy

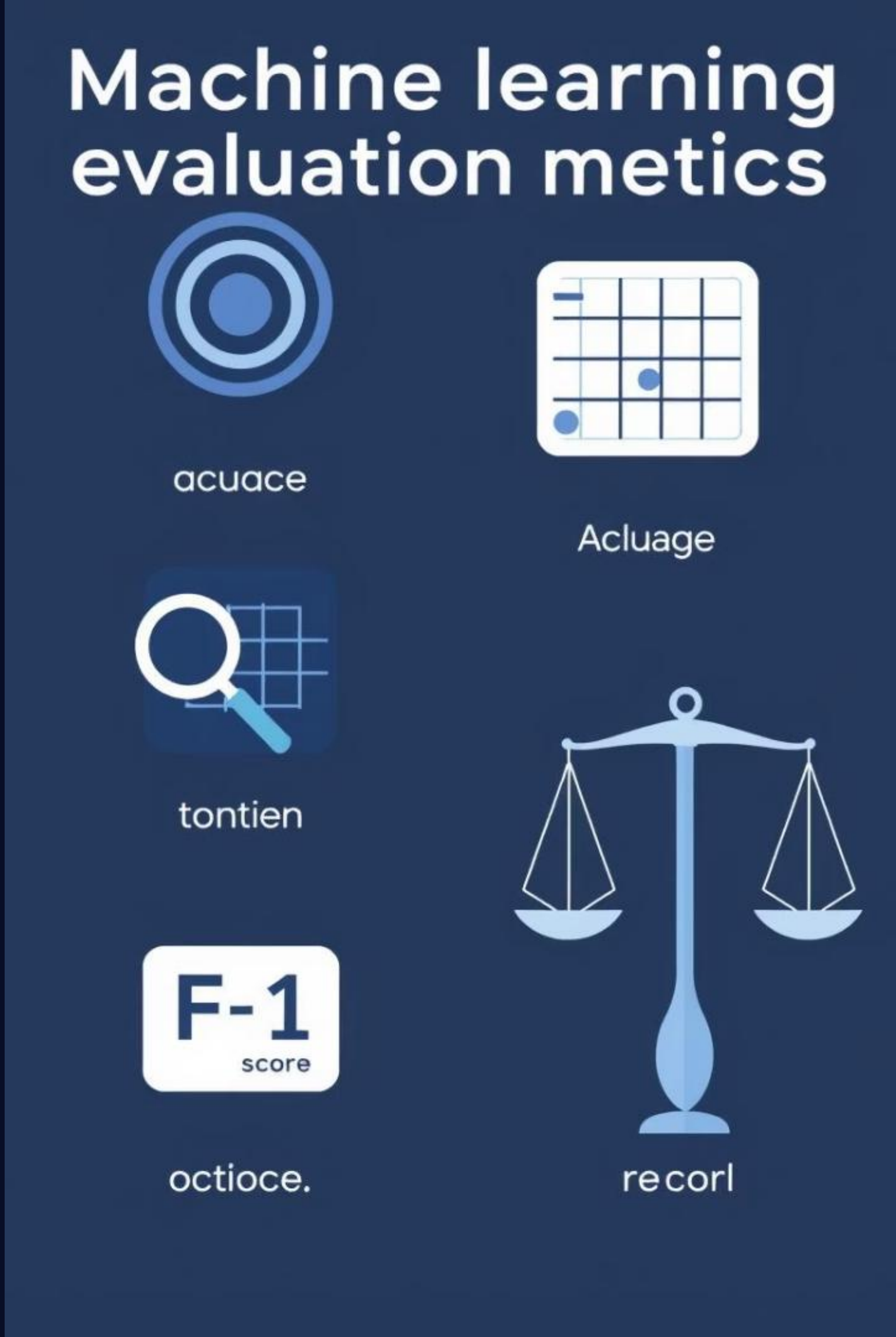
Proportion of correct predictions out of total cases.

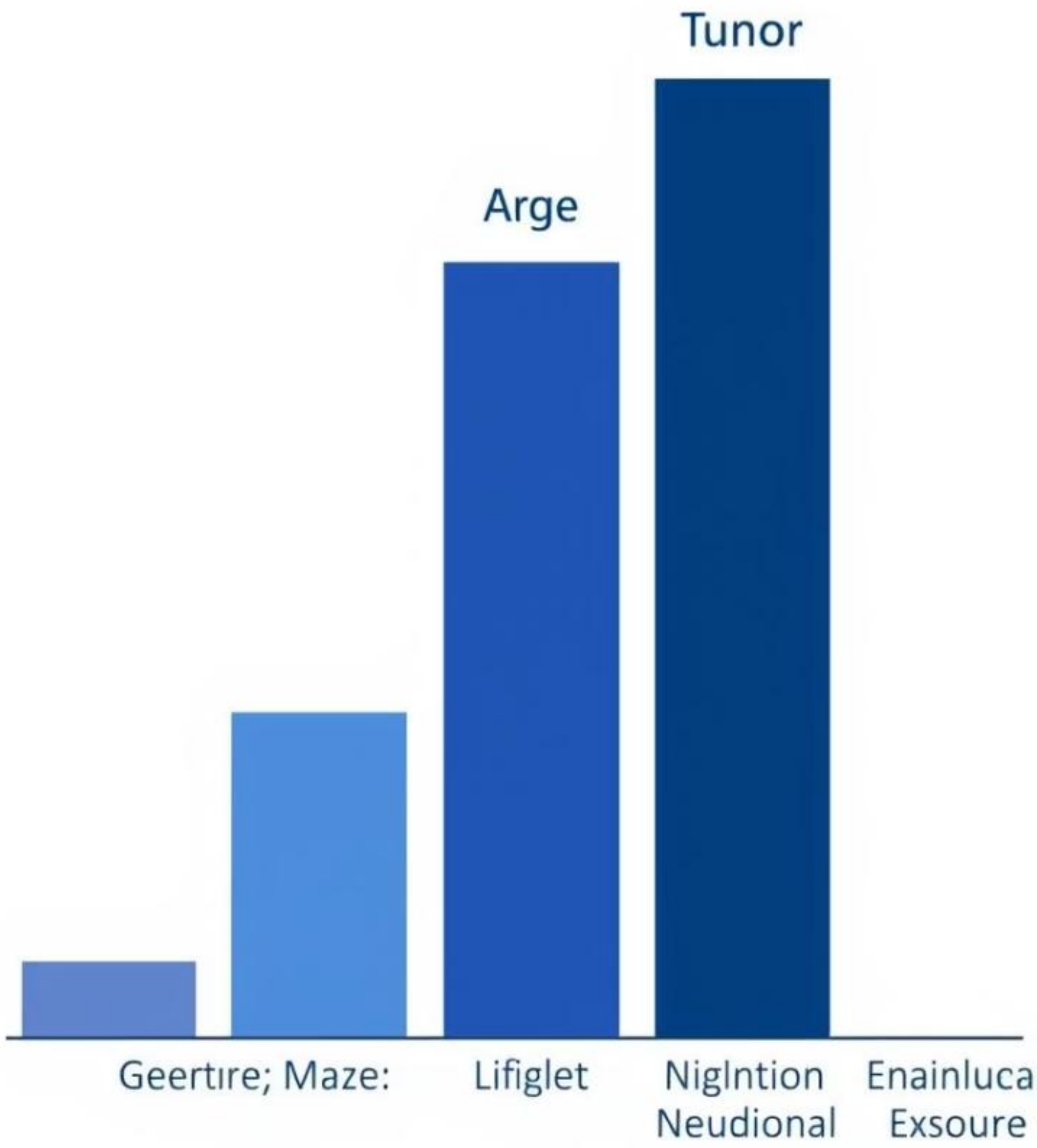
Confusion Matrix

Details true/false positives and negatives for comprehensive performance insight.

Precision, Recall, F1-Score

- Precision: Correct positive predictions
- Recall: Correctly identified actual positives
- F1-Score: Balance of precision & recall





Results and Insights

Test Accuracy

Model achieved 100% accuracy on test data.

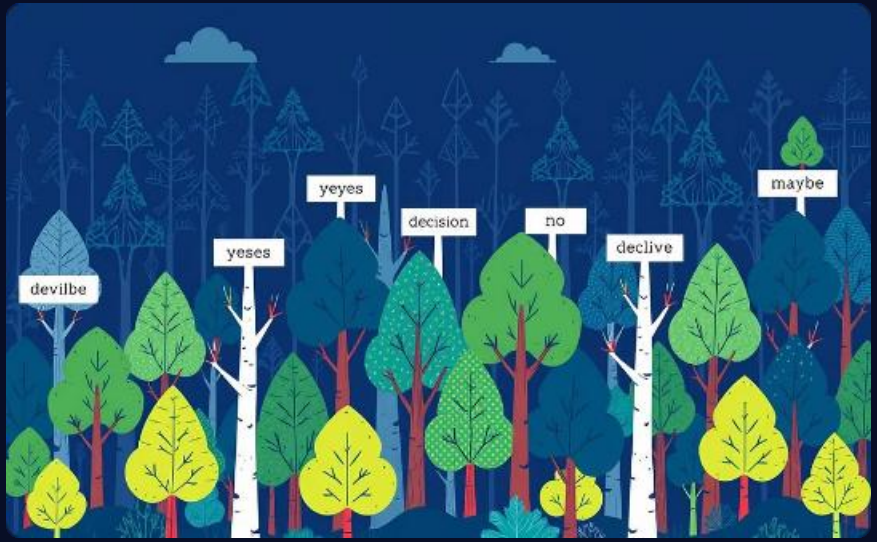
Key Features

Tumor Size, Age, and Tobacco Use ranked highest in influence.

Sample Predictions

- Patient A → Cancer with 81% confidence
- Patient B → No Cancer with 100% confidence

Conclusion and Future Directions



Achievements

Developed a robust Random Forest model to detect oral cancer risk early.

Impact

Enables clinicians to make quick and accurate assessments, enhancing patient outcomes.

Future Work

Plan to expand data and explore XGBoost and Deep Learning to improve accuracy.

GITHUB PROFILE LINK

https://github.com/Sahib2306/Oral_Cancer