# **Gender Prediction Using Dental Metric**

## **1. Introduction**

**Problem Statement:** The goal of this project is to analyse the data and predict, based on a combination of dental features that describes the Gender of the person

**Dataset:** The dataset contains multiple dental attributes such as inter canine distance intraoral, right canine width intraoral, left canine width intraoral. The target variable is Gender, categorized as Male or Female.

## **2. Data Preprocessing**

### **Steps Taken:**

#### **Data Cleaning:**

* Removed unnecessary columns such as Sample ID and Sl No.
* Checked for missing values and handled them appropriately.

#### **Encoding Categorical Data:**

* Used Binary Encoding for the Gender column:

#### **Feature Normalization:**

* Applied Normalizer from sklearn.preprocessing to scale the dataset, ensuring all features fall within a comparable range.

#### **Feature Selection:**

* Highly correlated features were identified using a heatmap and removed to reduce multicollinearity.

## **3. Exploratory Data Analysis (EDA)**

* Used Seaborn’s heatmap to visualize correlations between features.
* Identified which features had the strongest relationships with the target variable.

## **4. Model Building**

### **Algorithms Used:**

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* XGBoost Classifier

### **Train-Test Split:**

* 80% training data, 20% testing data used for model evaluation.

## **5. Model Evaluation & Results**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Best Performing |
| Logistic Regression | 64.54% | No |
| Decision Tree | 86.36% | No |
| Random Forest | 88.63% | Yes |
| XGBoost | 89.54% | Yes (Best Overall) |

### **Confusion Matrices:**

#### **Random Forest Confusion Matrix:**

* True Positives: 92
* False Positives: 15
* True Negatives: 103
* False Negatives: 10

#### **XGBoost Confusion Matrix:**

* True Positives: 95
* False Positives: 12
* True Negatives: 102
* False Negatives: 11

### **ROC Curve & AUC Score:**

* Random Forest AUC Score: 0.86 (High, indicating strong model performance.)
* XGBoost AUC Score: 0.96 (High, indicating strong model performance.)

## **6. Conclusion**

* XGBoost was the best-performing model with an accuracy of 89.54% and an AUC Score of 0.96.
* Feature selection played a crucial role in improving model accuracy by removing correlated features.