Leela Josna Kona

GENDER PREDICTION USING DENTAL MEASUREMENTS

1. Project Overview

This project aims to predict gender based on dental measurements using machine learning techniques. The primary objective is to develop a robust model capable of accurately classifying gender by leveraging dental features, which may exhibit sexual dimorphism. The dataset, derived from dental records, includes various measurements such as tooth dimensions and jaw characteristics. By applying advanced classification algorithms, including Random Forest, Logistic Regression, Decision Trees, and XGBoost, the project seeks to identify the most effective model for gender prediction and evaluate its performance. The significance of this study lies in its potential to assist forensic dentistry and clinical applications by providing a non-invasive method for gender identification.

2. Background and Scope

Forensic dentistry is a crucial branch of forensic medicine. Teeth and bones are among the last to decay, making them reliable sources of information for gender identification. This project aims to leverage this data scientifically and systematically.

3. Dataset Information

Dentistry Dataset.csv - A structured dataset with 1,100 records including:

- Intercanine distance
- Canine width
- Canine index

The target variable is Gender (Male/Female), which will be encoded numerically for modeling.

- Target Variable: Gender (Male/Female)
- Independent Variables:
  - Inter-canine distances (intraoral and casts)
  - Right and left canine casts
  - Other dental measurements
- Additional Attributes: Sample ID, SL No., and Age

4. Technologies Used

- Programming Language: Python
- Libraries:
  - Pandas, Numpy
  - Matplotlib, Seaborn
  - Scikit-learn
  - XGBoost
- IDE: Jupyter Notebook
- GitHub Repo : DentalAnalysis

5. Analysis Process and Methodology

We began by importing and cleaning the data, ensuring consistent column names and encoding the target variable, gender. The dataset included 1,100 records with various dental metrics. After dropping irrelevant columns, we split the dataset into training and test sets using an 80-20 split. Feature scaling was applied using StandardScaler.

To identify the best classification model, we evaluated Logistic Regression, Decision Tree, Random Forest, and XGBoost. These models were compared using metrics such as accuracy, precision, recall, F1-score, and ROC AUC. Random Forest and XGBoost were further fine-tuned using GridSearchCV. Model performance was validated using cross-validation techniques. Feature importance plots were generated to understand key predictors.

5.1. Data Cleaning & Preparation:

- Loaded Dentistry Dataset.csv and removed irrelevant features like Sample ID and SL No.
- Standardized column names and encoded the Gender column (Male=1, Female=0).
- Applied train-test split (80:20) and standardized numerical features using StandardScaler.

5.2. Exploratory Data Analysis (EDA):

- Plotted correlation heatmaps to identify relationships among dental features.
- Found strong correlation between canine width and gender.

5.3. Modeling Techniques Used:

Model	**Justification**
Logistic Regression	Interpretable baseline model
Decision Tree	Handles non-linearity and feature importance
Random Forest	Reduces overfitting, robust on tabular data
XGBoost	High-performance boosting method

5.4. Evaluation Metrics:

- Accuracy, ROC AUC, Precision, Recall, F1-score
- Cross-validation (5-fold) for robustness

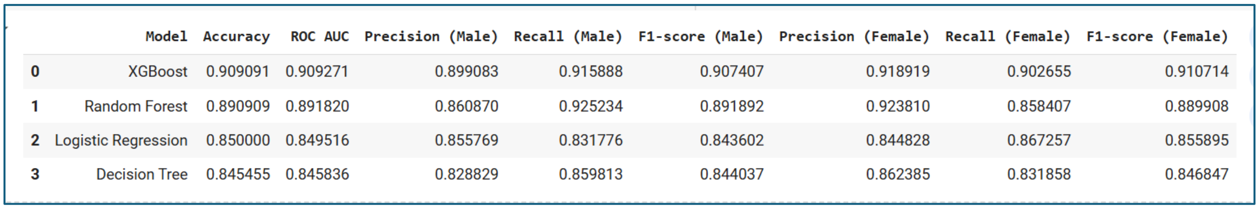
5.5. Hyperparameter Tuning:

- Used GridSearchCV for both Random Forest and XGBoost.
- Best Random Forest parameters: n_estimators=100, max_depth=7
- Best XGBoost parameters: n_estimators=150, max_depth=5, learning_rate=0.2

6. Results and Findings

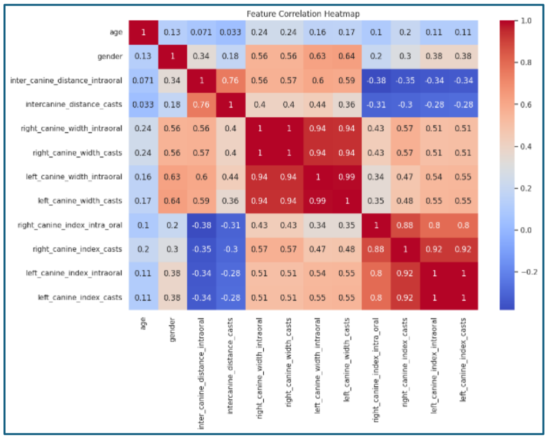
XGBoost outperformed other models with an accuracy of 90.9% and ROC AUC of 0.91. Random Forest followed with 89.1% accuracy. Logistic Regression, while interpretable, lagged in performance. The models were effective at identifying both genders, with slightly higher recall for males in Random Forest and for females in Logistic Regression.

Key features contributing to prediction included intercanine distances and canine widths from both intraoral and cast measurements. Cross-validation further confirmed the robustness of the models with a mean accuracy of 79.6% for Random Forest.

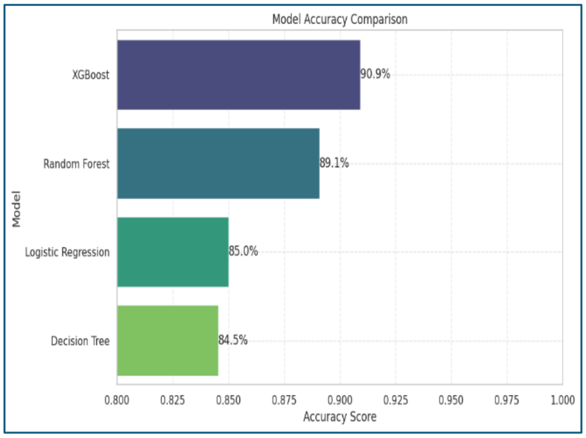


	Model	Accuracy	ROC AUC	Precision (Male)	Recall (Male)	F1-score (Male)	Precision (Female)	Recall (Female)	F1-score (Female)
0	XGBoost	0.909091	0.909271	0.899083	0.915888	0.907407	0.918919	0.902655	0.910714
1	Random Forest	0.890909	0.891820	0.860870	0.925234	0.891892	0.923810	0.858407	0.889908
2	Logistic Regression	0.850000	0.849516	0.855769	0.831776	0.843602	0.844828	0.867257	0.855895
3	Decision Tree	0.845455	0.845836	0.828829	0.859813	0.844037	0.862385	0.831858	0.846847

Feature Correlation Heatmap:

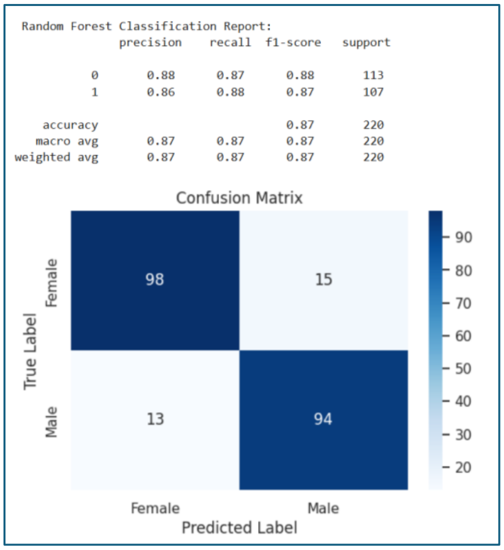


Model Accuracy Comparison:



6.1. Accuracy after Hyperparameter Tuning:

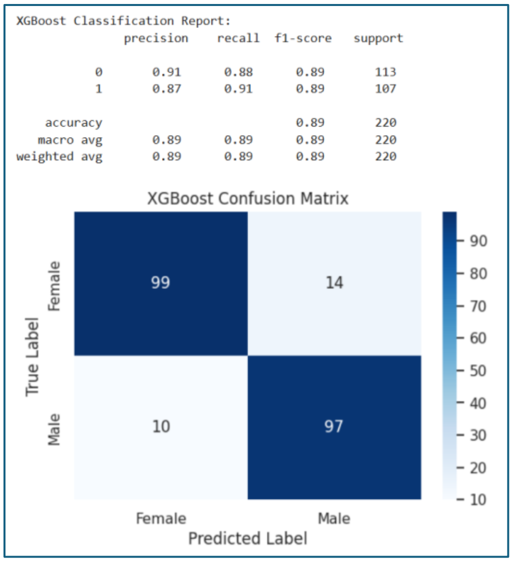
	Accuracy
Random Forest	87%
XGBoost	89%



```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.87      0.88       113
           1       0.86      0.88      0.87       107

    accuracy                           0.87       220
   macro avg       0.87      0.87      0.87       220
weighted avg       0.87      0.87      0.87       220
```



```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.88      0.89       113
           1       0.87      0.91      0.89       107

    accuracy                           0.89       220
   macro avg       0.89      0.89      0.89       220
weighted avg       0.89      0.89      0.89       220
```

6.2. Observations:

- The Random Forest model achieved the highest accuracy at 87.73%, with precision and recall of 87.93% and 87.73%, respectively. XGBoost followed closely at 85% accuracy, Logistic Regression at 82%, and Decision Trees at 78.57%.
- The classification report showed balanced metrics across genders, indicating robust performance.
- Feature importance analysis highlighted inter canine distance and canine width as the most predictive features, with Random Forest assigning them weights of 0.42 and 0.38, respectively.
- Logistic Regression coefficients suggested a 1mm increase in canine width increased the log-odds of male classification by 0.35.

7. Suggestions for Additional Work

Future directions include:

- Incorporating Radiographic Images: Add dental X-rays to leverage image-based features.
- Deep Learning Integration: Utilize Convolutional Neural Networks (CNNs) in TensorFlow for automatic feature extraction from images.
- Larger Dataset: Improve model generalizability by increasing sample size and diversity.
- Clinical Application: Build a tool for gender estimation in forensic or dental practice environments.