Leela Josna Kona

MIDTERM REPORT – GENDER PREDICTION USING DENTAL MEASUREMENTS

Objective

The goal of this project is to use dental measurements to predict if a person is male or female. By applying **predictive modeling** techniques likes Logistic, Random Forest, we will try to find patterns in tooth data that can help in gender identification. These prediction methods can support in areas like forensic science and anthropology.

Datasets

We will integrate data from two main sources:

1. Forensic Anthropology Data Bank (FDB) / FORDISC
   Source: https://fac.utk.edu/background/
   This dataset contains detailed information such as cranial and postcranial measurements, dental traits, trauma indicators, and demographic data (e.g., height, weight, medical history). It provides rich dental observations useful for gender analysis.
2. 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.

Proposed Methods

- Data Preparation – Clean the data, manage missing values, encode gender, and normalize measurements.
- Exploratory Analysis – Use charts to understand the data and identify the right features.
- Model Training – Try different models like:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest
  - KNN
- Model Testing – Check accuracy and performance using test data.

  - Final Selection – Pick the best model and identify key dental features used for prediction.

Through this project, we hope to show that using simple biological data like teeth measurements can help in identifying gender using modern technology.

Tasks Completed

- Problem Definition and Objective

- Data Collection: Selected two datasets:
  1. *Dentistry Dataset.csv* (1,100 records with dental measurements)
  2. *Forensic Anthropology Data Bank (FDB)* – includes dental and demographic data.

- Data Preparation:
  - Cleaned data and handled missing values
    - Dropped Sl No, Sample ID
  - Encoded gender (Male = 1, Female = 0)
  - Scaled numeric features
  - Merged selected features from both datasets
  - Split the dataset into 80:20

- Exploratory Analysis:
  - Identified feature patterns using histograms and correlations
  - Found intercanine distance and canine index to show strong gender-based trends

- Initial Modeling:
  - Trained Logistic Regression, Decision Tree, Random Forest, and KNN
  - Used 80/20 split and cross-validation for evaluation

Preliminary Results

Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	70.75%	71.05%	70.75%
Decision Tree	78.57%	77.25%	76.20%
Random Forest	87.73%	87.93%	87.73%
KNN	79.71%	79.71%	79.71%

Observations:

- Random Forest shows the best performance so far.
- Dental metrics like intercanine distance are highly predictive.
- Adding FDB features improves accuracy slightly.

Tasks Pending

- Hyperparameter Tuning: Use GridSearchCV to optimize Random Forest, KNN, and others.
- Feature Importance
- Final Model Selection: Choose the best model based on test performance and interpretability.
- Consumer Matrix and Feature Importance Charts and visuals

Issues and Challenges

- Dataset Integration: Needed transformation to align FDB and Dentistry data formats.
- Imbalanced Classes: Used class weights and SMOTE to balance Male/Female ratios.
- Outliers and Multicollinearity

Conclusion

The project is progressing well, and early results are promising. With further tuning and validation, the final model is expected to deliver reliable gender predictions using simple dental measurements.