**Capstone Project Documentation**

**Project Title: Gender Classification using Dental Measurements**

**1. Introduction**

This project aims to classify gender based on various dental measurements using machine learning techniques. The dataset consists of multiple numerical features related to dental measurements and a categorical target variable (Gender).

**2. Dataset**

**Source:** Dentistry Dataset.csv

**Key Columns:**

* **Sl No**: Serial number (Dropped during preprocessing)
* **Sample ID**: Unique sample identifier (Dropped during preprocessing)
* **Gender**: Target variable (Male/Female)
* **Various Dental Measurements**: Independent variables used for classification

**3. Data Preprocessing**

* Removed unnecessary columns (**Sl No, Sample ID**).
* Encoded the categorical variable (**Gender**) using Label Encoding.
* Normalized the feature values using **Normalizer** to standardize the range.
* Split the dataset into **80% training** and **20% testing** using stratified sampling.

**4. Exploratory Data Analysis (EDA)**

* Generated a **correlation heatmap** to analyze feature relationships.
* Identified highly correlated features that could impact model performance.

**5. Model Selection & Training**

Four machine learning models were trained and evaluated:

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

**6. Model Evaluation**

Each model was assessed based on:

* **Accuracy**
* **AUC (Area Under Curve)**
* **Confusion Matrix**
* **Classification Report (Precision, Recall, F1-Score)**

**Results**

| **Model** | **Accuracy** | **AUC** |
| --- | --- | --- |
| Logistic Regression | 68.2% | 68.0% |
| Decision Tree | 70.5% | 70.3% |
| Random Forest | 76.4% | 76.2% |
| XGBoost | 78.2% | 78.0% |

**Best Performing Model: XGBoost**

**7. Feature Importance**

For tree-based models (Decision Tree, Random Forest, XGBoost), feature importance was analyzed to identify the most impactful dental measurements in predicting gender.

**8. Conclusion**

* **XGBoost** achieved the highest accuracy and AUC, making it the best model for gender classification.
* Some dental measurements were highly correlated, indicating potential for feature selection or dimensionality reduction.
* Future work could involve hyperparameter tuning, additional feature engineering, or using deep learning models for improved accuracy.

**9. References**

* Machine Learning in Dentistry Research
* Scikit-learn and XGBoost documentation for model implementation