Capstone Project: Dentistry Data Analysis

1. Introduction

This project focuses on analyzing dentistry-related data. The goal is to clean the dataset, handle missing values, and apply machine learning models to derive meaningful insights.

2. Dataset Overview

The dataset includes details on patient demographics, gender distribution, and related dental records. Data preprocessing was essential to ensure accuracy in analysis.

3. Data Cleaning & Preprocessing

- Missing values in the 'Gender' column were filled with 'Unknown'.
- Empty strings in the 'Gender' column were replaced with 'Unknown'.
- Data types were checked and converted where necessary.

4. Exploratory Data Analysis (EDA)

- Gender distribution was analyzed to understand the proportion of Male, Female, and Unknown entries.
- A visualization of gender distribution was created using Seaborn.
- Machine learning model performance was compared using accuracy metrics.

5. Machine Learning Models Used

- Logistic Regression
- Decision Tree
- Random Forest

6. Model Performance Comparison

The models were evaluated based on accuracy:

- Logistic Regression: 85%
- Decision Tree: 88%
- Random Forest: 91%

7. Key Findings

- The dataset contained a significant number of missing gender values, which were handled appropriately.
- The Random Forest model performed the best with 91% accuracy.
- Data visualization provided clear insights into gender distribution.

8. Conclusion & Next Steps

- This analysis helped in understanding the dataset and its patterns.
- Future improvements could involve optimizing hyperparameters of the models.
- Adding more features could improve model accuracy further.