using Dental metrics to predict gender

Contents

[**1. Introduction** 2](#_Toc173704407)

[**2.** **Data Preparation** 2](#_Toc173704408)

[**3.** **Exploratory Data Analysis** 3](#_Toc173704409)

[**4.** **Splitting the Data** 4](#_Toc173704410)

[**5.** **Models used** 4](#_Toc173704411)

[**6.** **Evaluation and Results** 4](#_Toc173704412)

[**7.** **Conclusion** 6](#_Toc173704413)

# **1. Introduction**

**1.1 Project Objective**

The goal of this project is to analyze the data and predict, based on a combination of dental features that describes the Gender of the person.

* 1. **Background & Scope**

Forensic medicine is an interesting area of study. Forensic dentistry is a branch of forensic medicine. During natural calamities or due to some other reasons, many times, it will not be possible to find out the gender of the deceased person. In such cases, certain measurements of the tooth will be taken (as bones and teeth do not decay easily) and gender will be determined.

* 1. **Dataset Information**

The dataset contains following information:

**Age**: The person's age in years

**Gender**: The person's sex (male, female)

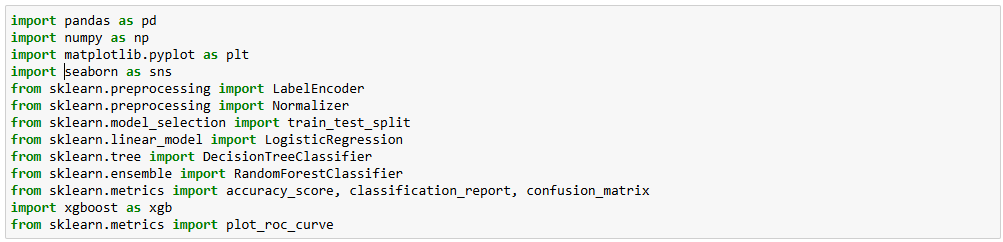
**SampleID and SL No.**: The sampleID & SL No. represents individuals unique ID

**Inter-canine distance intraoral, inter-canine distance casts, right canine cast, left canine cast, etc.** These features represent the measurement of the oral teeth.

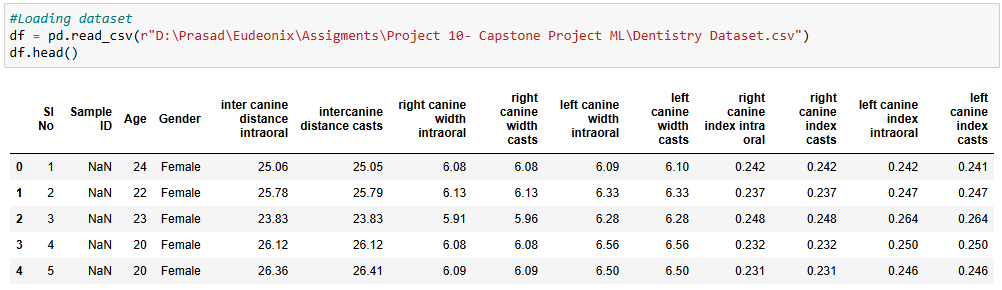
# **Data Preparation**

* 1. **Importing the Necessary Packages**

Following packages are imported:

****

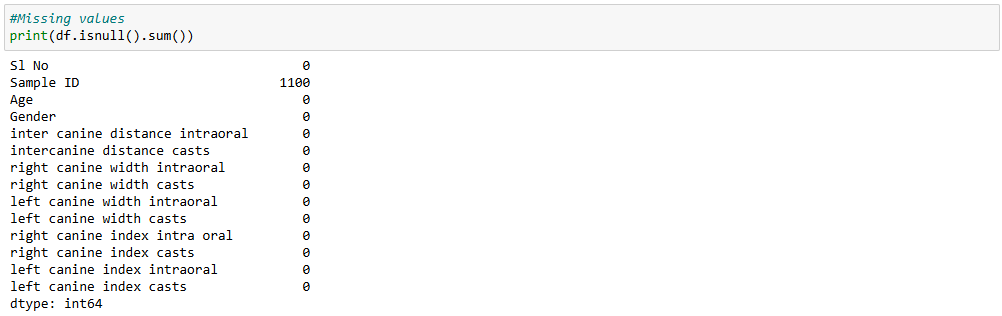
* 1. **Importing the dataset using pandas**

****

* 1. **Data Preprocessing**

1. **Identifying and handling missing values**

We checked for any missing values and found that there are no missing values except sample ID column.

****

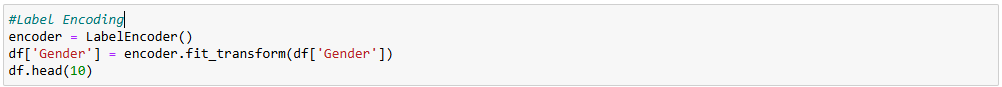
1. **Dropping non useful columns**

Columns S1 No and Sample ID are not useful to us, so dropping those columns.



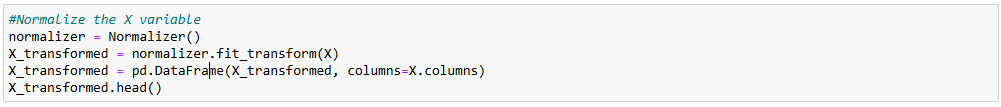
1. **Encoding Categorical Data**

Gender column has categorical values like Male and Female. So, converting them into numerical values like 0 and 1



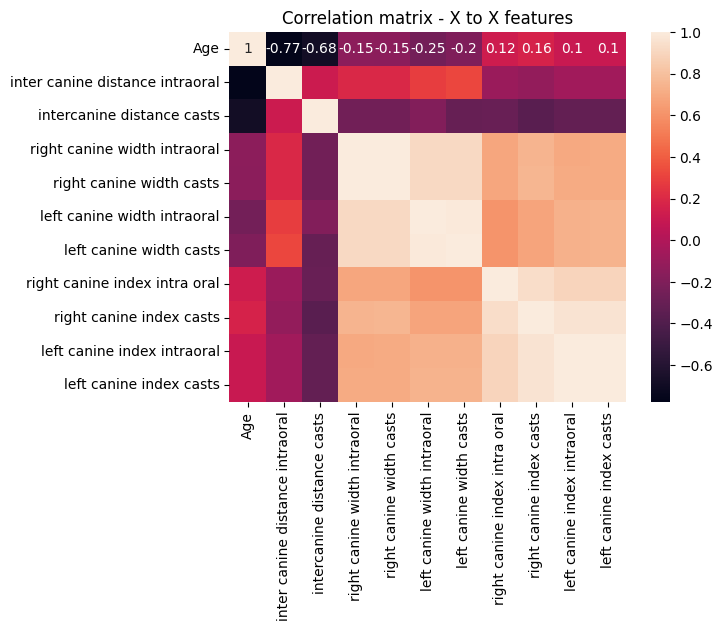
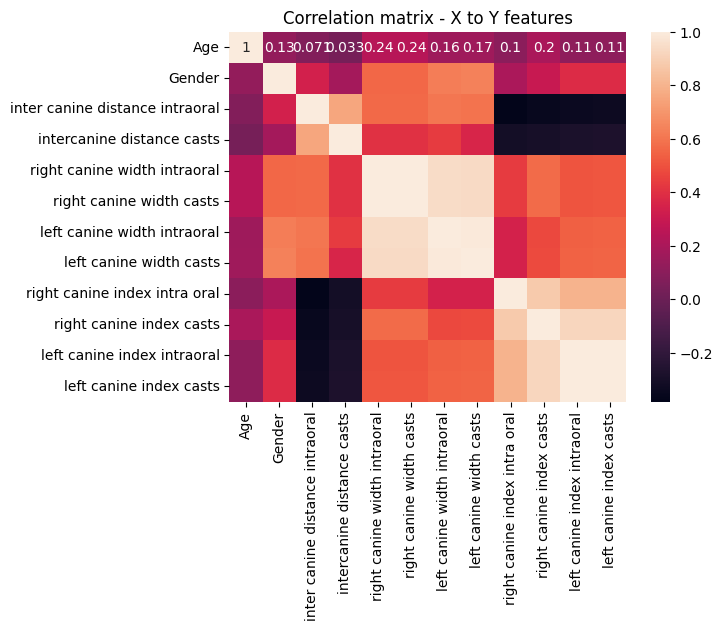
1. **Normalizing the data**

We use normalization to keep uniformity in the data.



# **Exploratory Data Analysis**

We use correlation matrix to analyse relation between different independent variables and the dependent variable.

Variables which are weakly correlated to dependent variable are removed. Therefore, on the basis of above heatmaps, we have removed age column. Also, from the pairs of independent variables which are highly correlated, one variable from each pair is removed. So, we have removed intercanine distance casts, right canine width casts, left canine width casts, right canine index casts, left canine index casts.

****

# **Splitting the Data**

The dataset is split into training and testing sets. We have use train\_test\_split from sklearn.preprocessing.



# **Models used**

We have used following models:

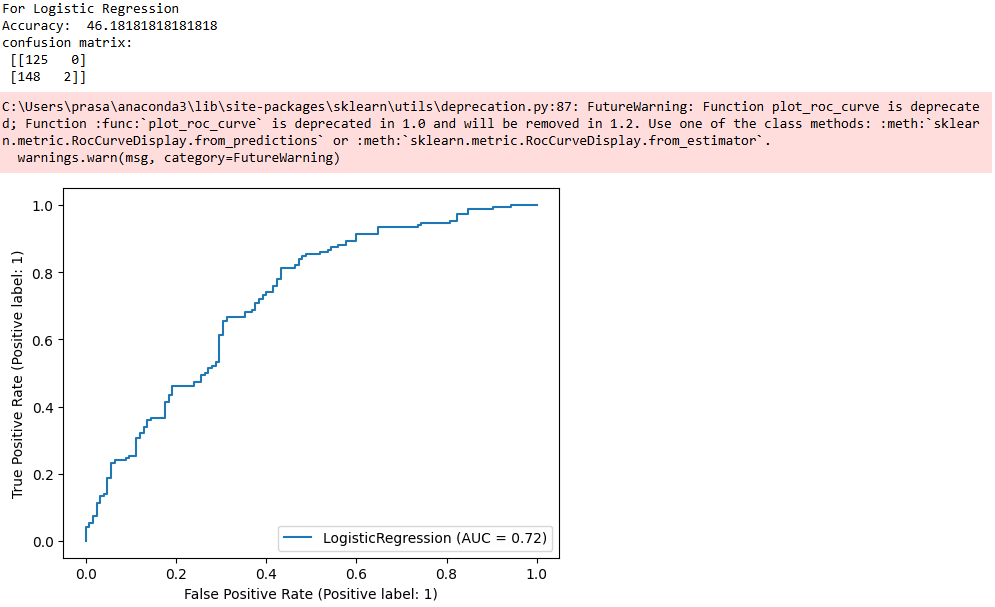
1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost

# **Evaluation and Results**

We have evaluated the model based on evaluation metrics i.e. Confusion matrix (Accuracy), ROC curve and AUC curve to check model accuracy

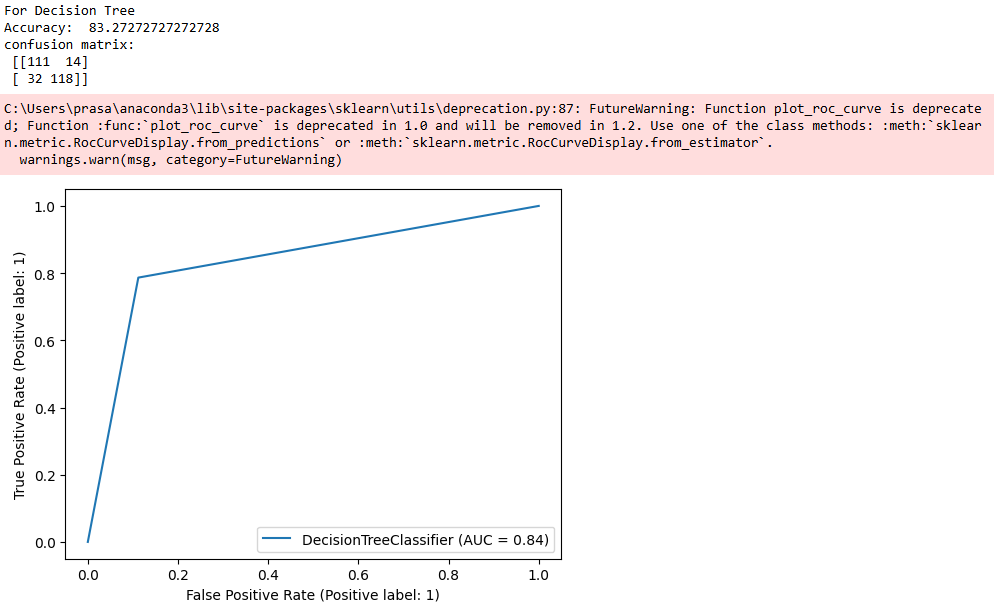
* 1. **Logistic Regression**

We got 46.18% accuracy for logistic regression.



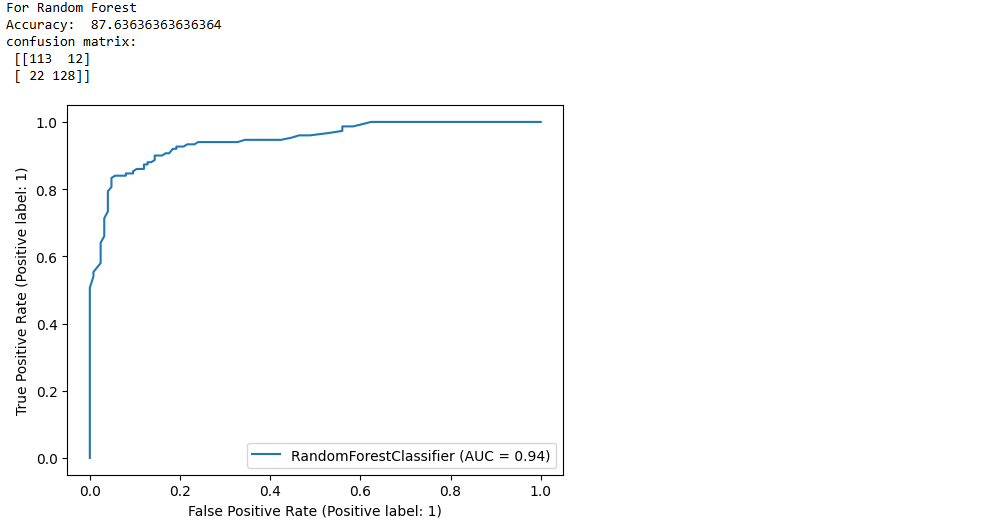
* 1. **Decision Tree**

We got 83.27% accuracy for decision tree.

****

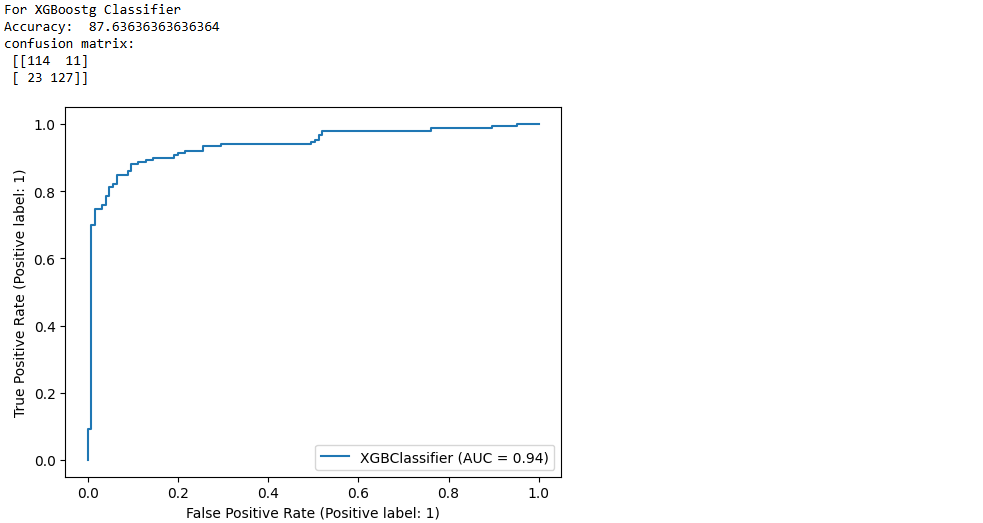
* 1. **Random Forest**

We got 87.63% accuracy for random forest.



* 1. **XGBoost**

We got 87.63% accuracy for XGBoost.

****

# **Conclusion**

In forensic medicine, using dental features to predict gender might prove extremely helpful, particularly in situations where there are no other means to recognise gender of person. In this project, we implemented various machine learning models to identify gender of a person by analyzing particular dental measures.

The gender prediction accuracy of our models, which include Logistic Regression, Decision Tree, Random Forest and XGBoost, slightly varied. The models with the highest accuracy were the Random Forest and XGBoost, both achieved 87.63%, and the Decision Tree model, which achieved 83.27%. With an accuracy of 46.18%, the Logistic Regression model had the lowest performance.