# Tooth Measurement Dataset

## Project Objective:

Build a predictive model which can take the tooth measurements as input and give the gender of the person as output.

## Data Understanding:

We have our data saved in a CSV file called proj1.csv.The dataset provides the tooth measurement information. It includes 1100 records and 14 fields.

### Input variables:

* Age(numeric)
* Inter Canine Distance Intraoral (Continuous)
* Inter Canine Distance Casts (Continuous)
* Right Canine Width Intraoral (Continuous)
* Right Canine Width Casts (Continuous)
* Left Canine Width Intraoral (Continuous)
* Left Canine Width Casts (Continuous)
* Right Canine Index Intraoral (Continuous)
* Right Canine Index Casts (Continuous)
* Left Canine Index Intraoral (Continuous)
* Left Canine Index Casts (Continuous)
* Sample ID (empty)
* Sl No (numeric)
* Sample ID and Sl No Variables do not contribute to prediction.

### Predict (Target) Variable:

* Gender (Categorical)

Binary Variable: Two Categories (Male and Female)

## Data Exploration:

Data has no missing values or null except for the Sample ID variable which is empty and can be dropped for prediction.

It’s a balanced data as male and female counts match.

Chart, bar chart

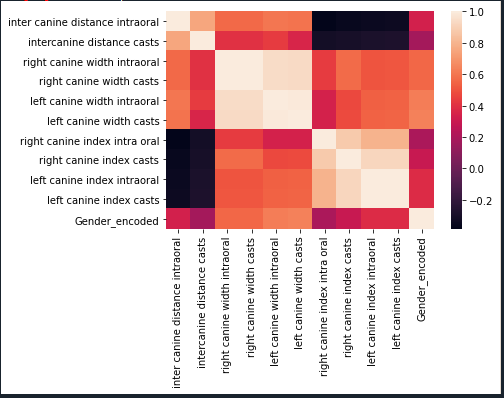
Description automatically generated

### Correlation:

Correlation of every pair of features (and the gender variable) and visualize the correlations using a heatmap.

A screenshot of a computer

Description automatically generated with medium confidence



In the above heatmap, brighter colours indicate more correlation. As we can see from the table and heatmap, right canine width intraoral, right canine width casts, left canine width intraoral and left canine width casts have significant correlation with the gender variable. These four features are highly correlated with each other as well.

Left Canine index intraoral, left canine index casts, right canine index intraoral, right canine index casts are inter-correlated and correlate with Gender variable though not significantly.

Inter canine Distance interval also seem to have impact on Gender variable though less.

### Outliers:

As we can see from below, data has negligible outliers.

Text

Description automatically generated

## Model Building:

Let us build model using Logistic Regression, K-Nearest Neighbours and Stochastic Gradient Descent Machine Learning Algorithms.

## Feature Selection:

For good predictions of the outcome, it is essential to include the good independent variables (features) for fitting the model (e.g., variables that are not highly correlated). If we include all features, there are chances that we may not get all significant predictors in the model.

we visualize how logistic regression model uses the different features and which features have greater effect.

Left canine width intraoral, left canine width casts, Left Canine index intraoral, left canine index Casts and Inter canine Distance Intraoral have significant influence on the model.

Inter canine distance casts, right canine index casts, Right canine index intraoral have negative influence on the prediction.

Chart, bar chart

Description automatically generated

The idea of repeatedly constructing a model and choosing either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted and check P values of the independent features and we filter down on few of the features.

We analyse various performance measures like accuracies, confusion matrix, sensitivity, precision etc on all the below models with different features as input.

## Model Evaluation:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Features** | **Logistic Accuracy** | **KNN Accuracy** | **SGD Accuracy** |
| 1 | * Inter canine distance intraoral * left canine width casts | 79% | 83% | 77% |
| 2 | * inter canine distance intraoral * left canine width casts * left canine index casts | 79% | 83% | 77% |
| 3 | * inter canine distance intraoral * left canine width casts, * left canine index intraoral | 79% | 83% | 77% |
| 4 | * inter canine distance intraoral * left canine width intraoral | 78% | 83% | 77% |
| 5 | * left canine width casts, * right canine width casts * left canine index casts | 80% | 88% | 78% |
| 6 | * left canine width intraoral * right canine width intraoral * left canine index intraoral | 78% | 87 | 77% |

## Logistic Regression Model:

|  |  |
| --- | --- |
| **Model 1** | Chart, histogram  Description automatically generated  **Model 2** |
|  |  |
| Chart, histogram  Description automatically generated  **Model 3** | Chart, histogram  Description automatically generated  **Model 4** |
|  |  |
| Chart, histogram  Description automatically generated  **Model 5** | Chart, histogram  Description automatically generated  **Model 6** |

### Model 6:

|  |  |
| --- | --- |
| Log Loss | 7.11 |
| Confusion Matrix | [[139 26]  [ 42 123]] |
| Sensitivity score | 0.83 |
| Specificity score | 0.77 |
| Precision | 0.75 |
| F1 score | 0.78 |
| Average Accuracy | 0.79 |
| ROC\_AUC score | 0.79 |

Chart, line chart

Description automatically generated

## KNN Model:

Visualizing KNN accuracies of all the above models

Chart, line chart

Description automatically generated

### Model 6:

|  |  |
| --- | --- |
| Confusion Matrix | [[127 11]  [ 27 110]] |
| Sensitivity score | 0.91 |
| Specificity score | 0.82 |
| Precision | 0.8 |
| F1 score | 0.85 |
| Accuracies (n -neighbours(1-4): | [0.87, 0.87, 0.89, 0.86] |
| ROC\_AUC score | 0.86 |

Chart, line chart

Description automatically generated

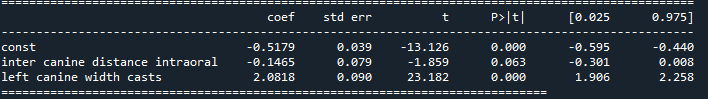
## SGD Model:

### Model 6:

|  |  |
| --- | --- |
| Confusion Matrix | [[125 13]  [ 38 99]] |
| Sensitivity score | 0.88 |
| Specificity score | 0.77 |
| Precision | 0.72 |
| F1 score | 0.8 |
| Average Accuracy | 0.81 |
| ROC\_AUC score | 0.78 |

## Chart, line chart Description automatically generated

## P-Values for above models with selected features:



A screenshot of a computer

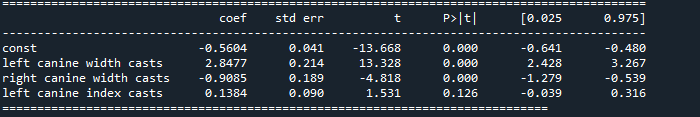
Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence



A screenshot of a computer

Description automatically generated with medium confidence

The *p* values for all independent variables are significant (*p* < 0.05) for model 6 and model 4 and suggests that these variables are highly associated with the outcome.

**Model 6 with features “left canine width intraoral”, “right canine width intraoral”, “left canine index intraoral” seem to be the fittest model with P-value <0.05 and gives better accuracy compared to other models. KNN algorithm gives the good results.**