

# Report

## Introduction

The aim of this analysis is to build a model that can classify whether breast mass is malignant or benign , using the relevant data. The provided datasets are used to train and test the models then the results are analyzed. Finally the best model is chosen and hyperparameter tuning is used to improve the models performance.

## Data Exploration

The data.csv dataset is loaded and reviewed in a table form to understand the structure and content of the data. While analyzing the dataset its understood which columns are useful for our model building and which columns have missing values and basically the whole plan is formulated. At this stage the variable of interest is identified to be the diagnosis column and the other columns are taken as x-variable (input).

## Data Cleaning

The columns that are irrelevant were dropped.

## Feature Engineering

One hot encoding is used to convert categorical variable values into numerical values.

## Model Building

A logistic regression and Random forest models are built to classify if the breast mass is malignant or benign.

## Model Evaluation

The models are tested and and the results are analyzed then the model with the best results is selected as the most suitable for the classification.

## Model Tuning

Hyperparameter tuning is performed to improve the performance of the model.

## Findings

Upon model evaluation a classification report was used to discover the performance of the model. Accuracy, precision, f1-score and recall were the main metrics that were targeted and used to decide which model was better.

The random forest model was chosen as the most suitable model for classifying if the breast mass is malignant or benign.

## Conclusion

The approaches and steps taken maximized the suitability of this model to classify if a breast mass is malignant or benign.

---