Heart Failure Prediction

BDV3ILV: Big Data Analytics und Interactive Visualization

Łukasz Kołtun

Dataset Information

1

The dataset

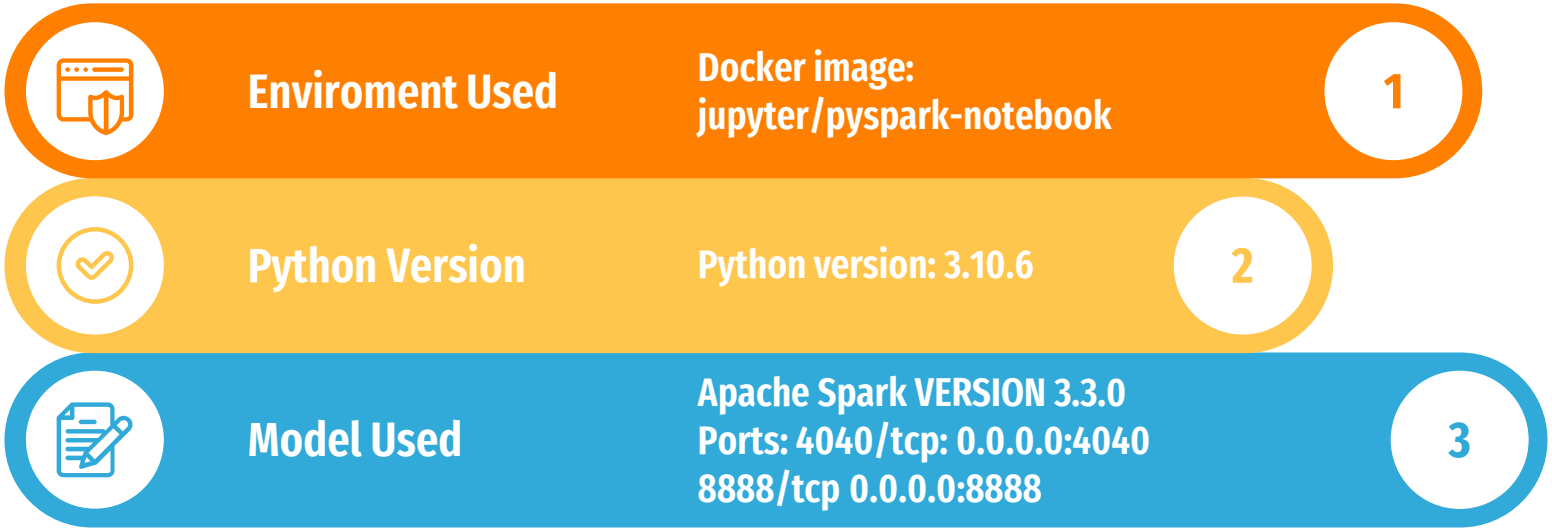
Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

2

Aim of the project

Create a model for predicting mortality caused by Heart Failure.

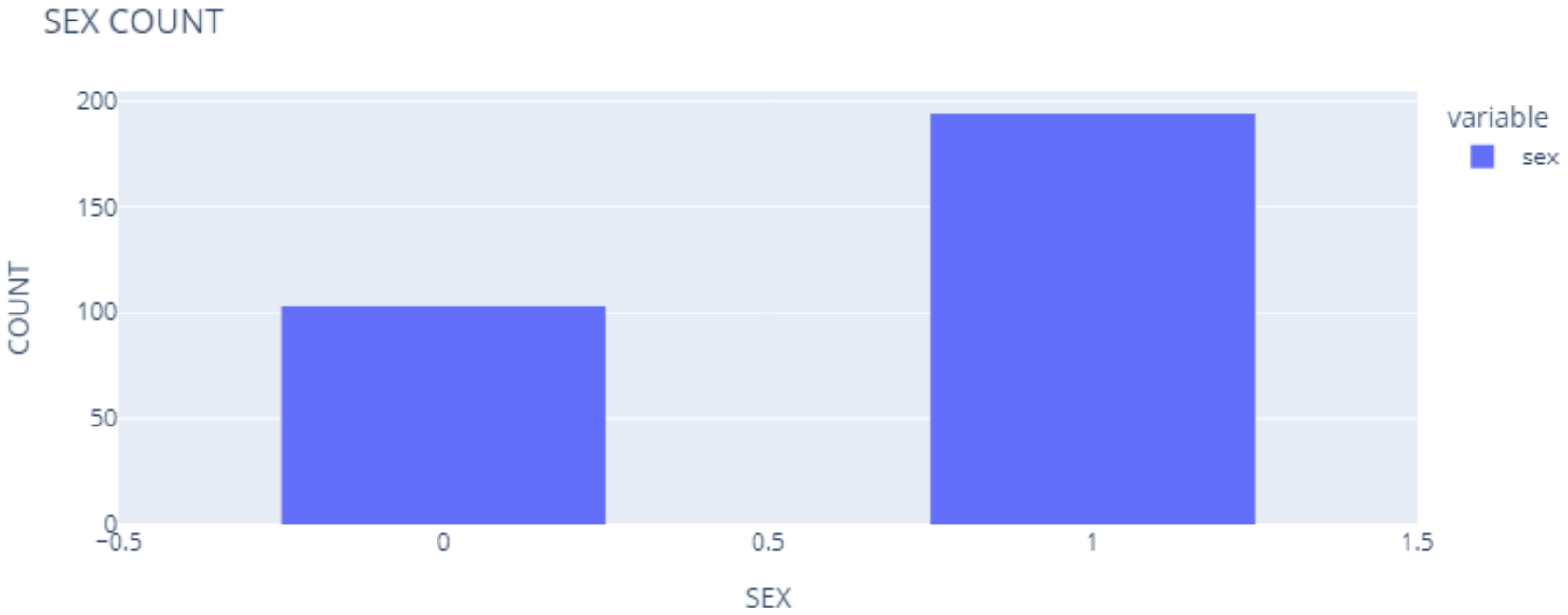
Environment



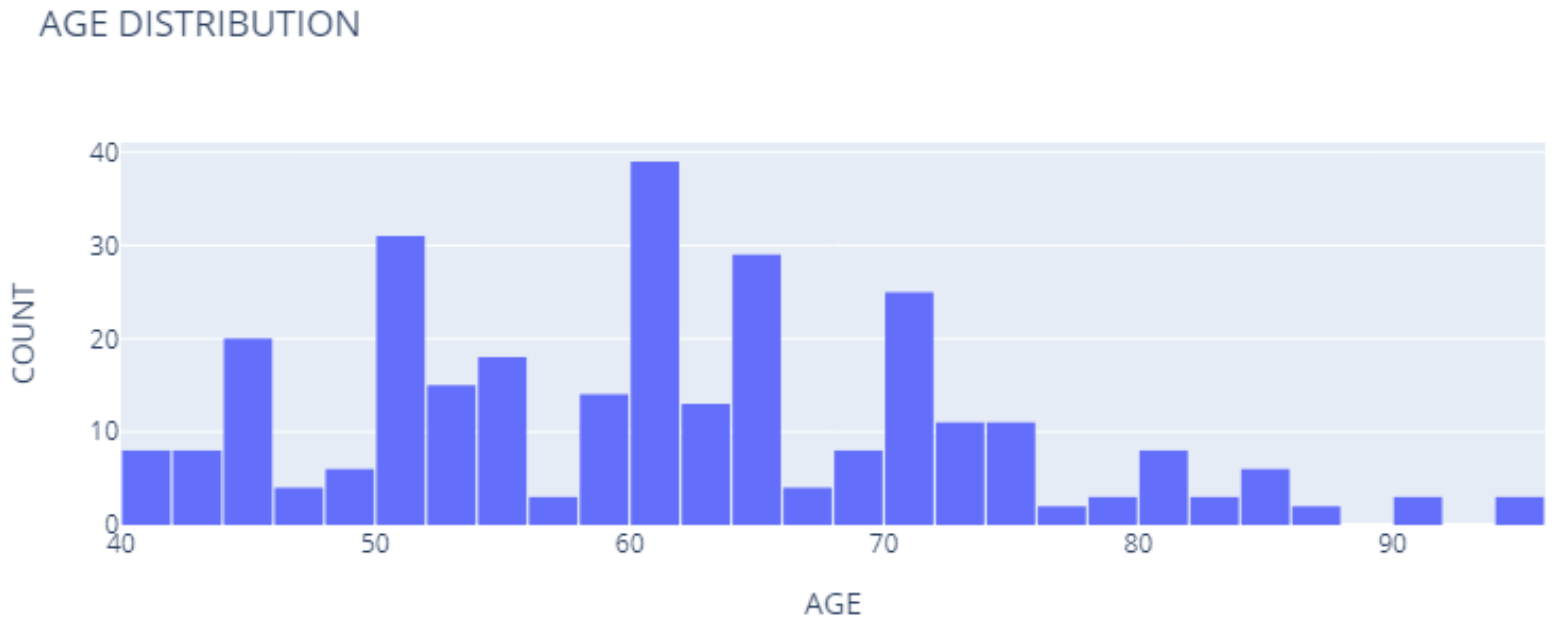
Research Questions

1. Is age and sex an indicator for death event?
2. How many smoking persons survived?
3. Has the smoking an influence on the ejection fraction?

Is age and sex an indicator for death event?

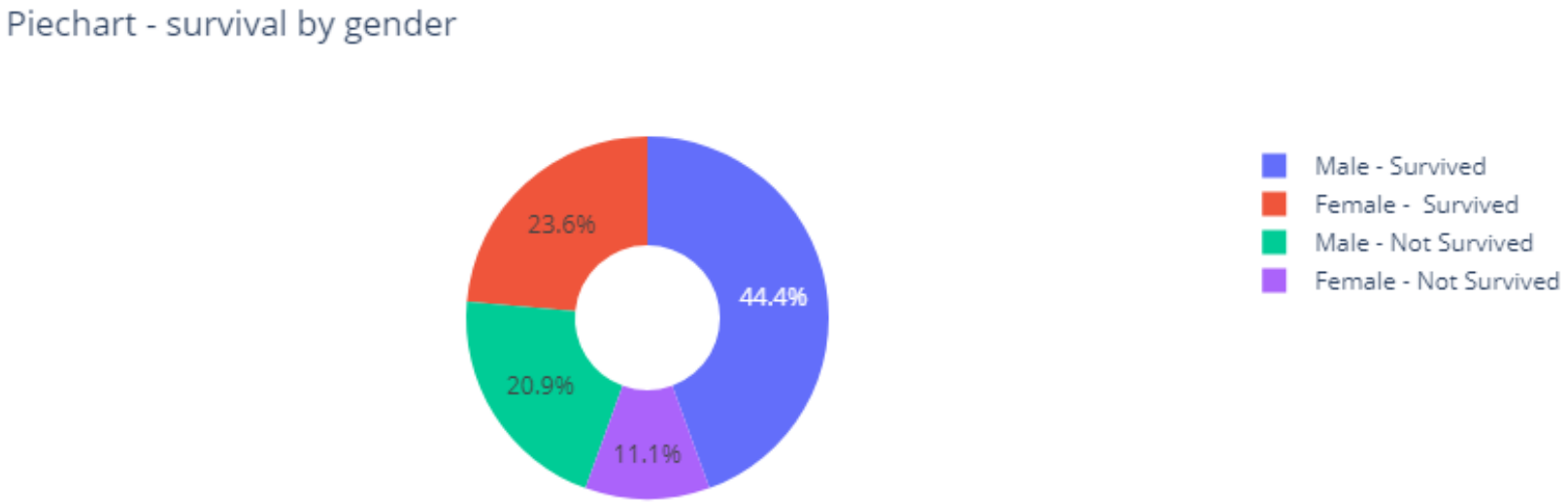


Is age and sex an indicator for death event?



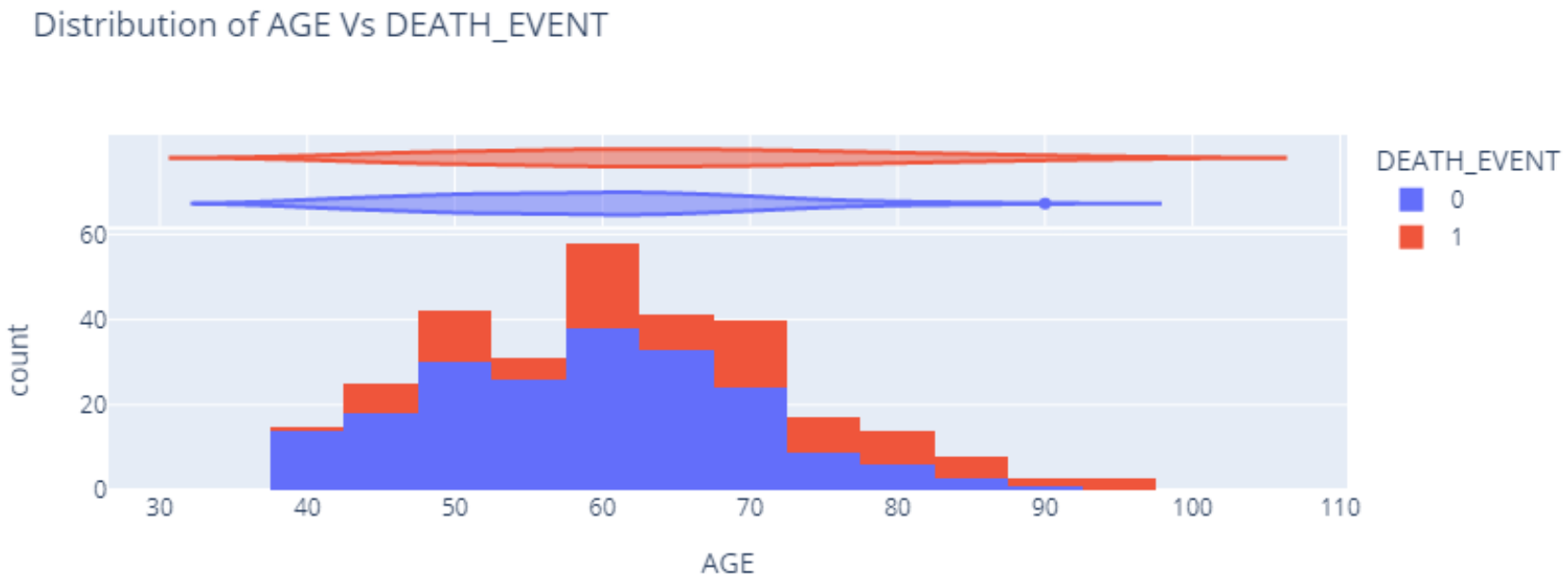
We don't have any people under 40 and above 95 years. Age higher than 80 are very low.

Is age and sex an indicator for death event?



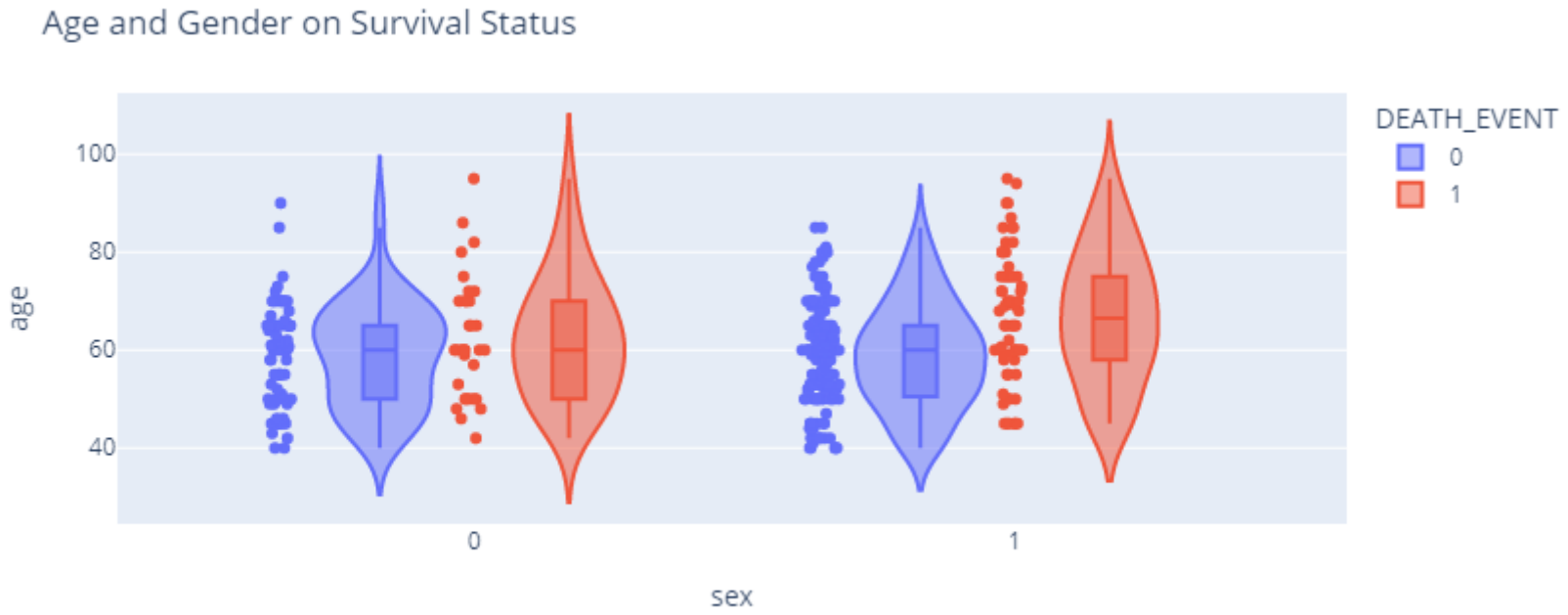
Survival (by percent): Female 71/105 = 67,6 % Male 132/194 = 68,0 %

Is age and sex an indicator for death event?



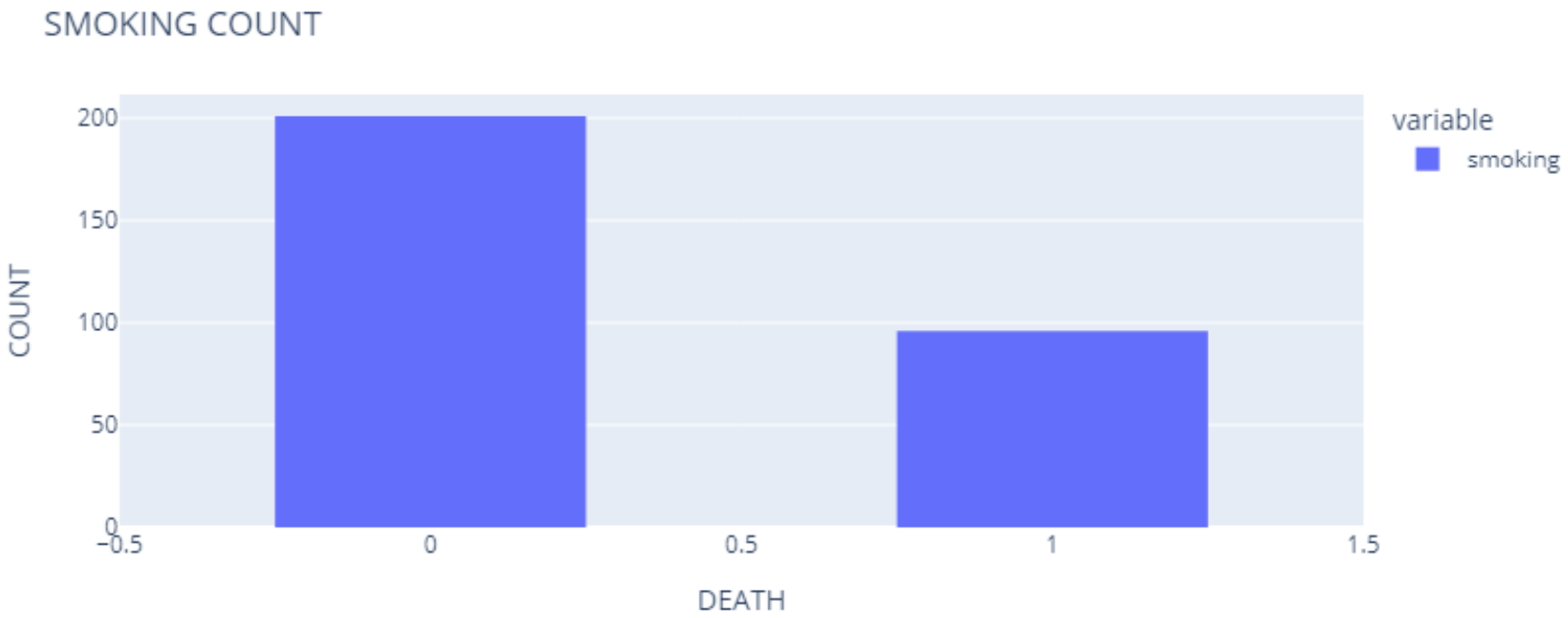
Survival is high on 40-70.

Is age and sex an indicator for death event?



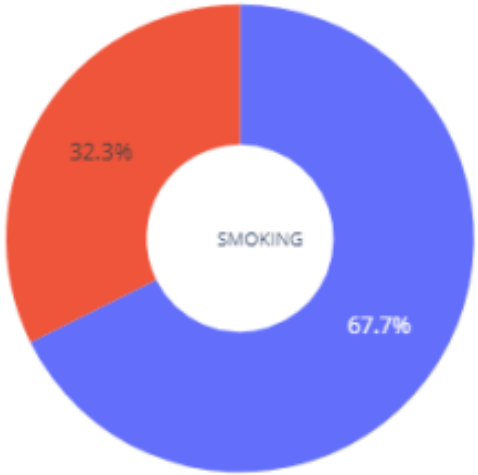
The Survival is the highest for male between 50 to 60 and female's between 60 to 70.

How many smoking persons survived?

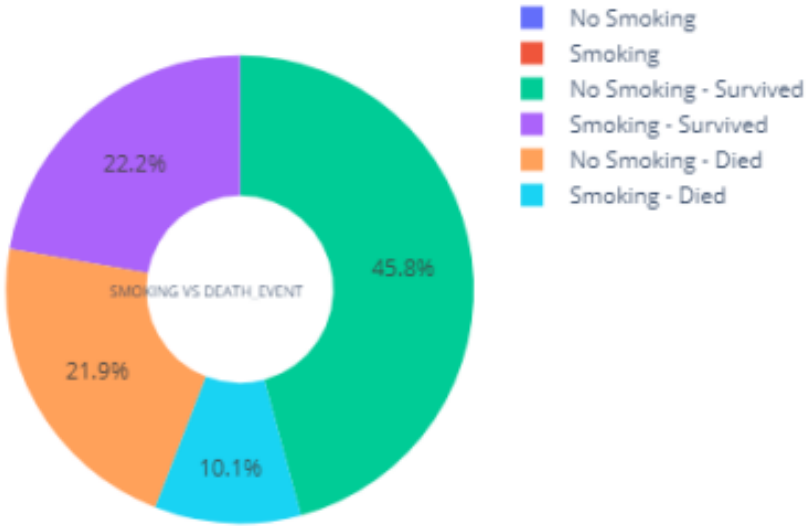


How many smoking persons survived?

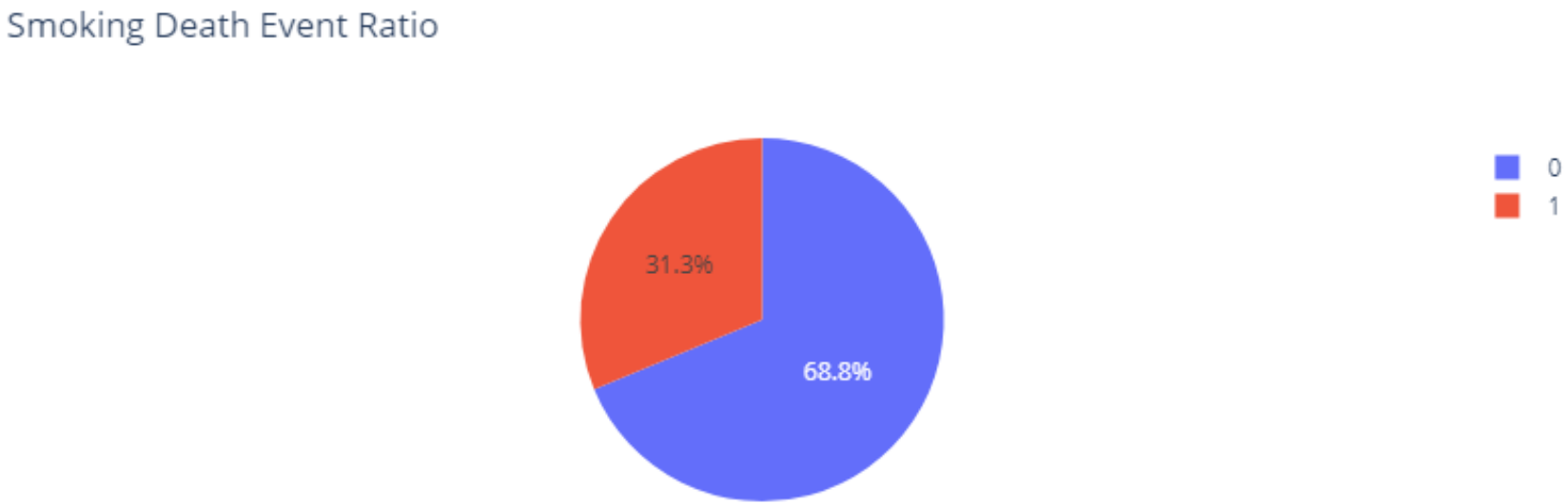
SMOKING DISTRIBUTION IN THE DATASET



SMOKING VS DEATH_EVENT

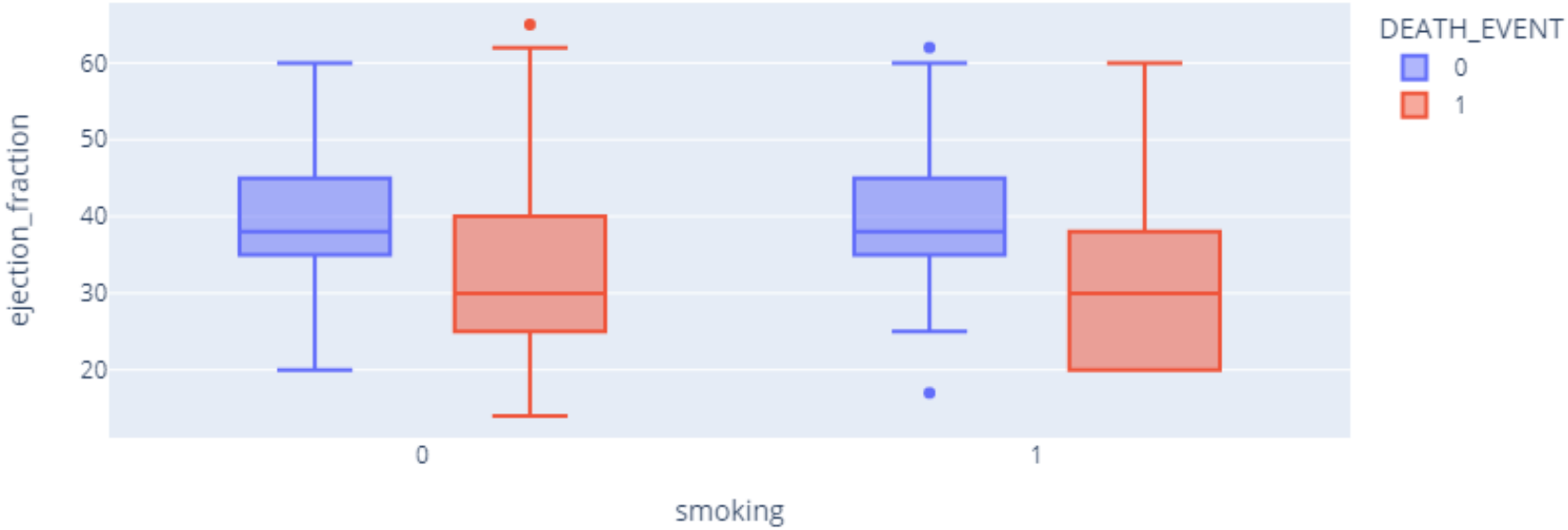


How many smoking persons survived?



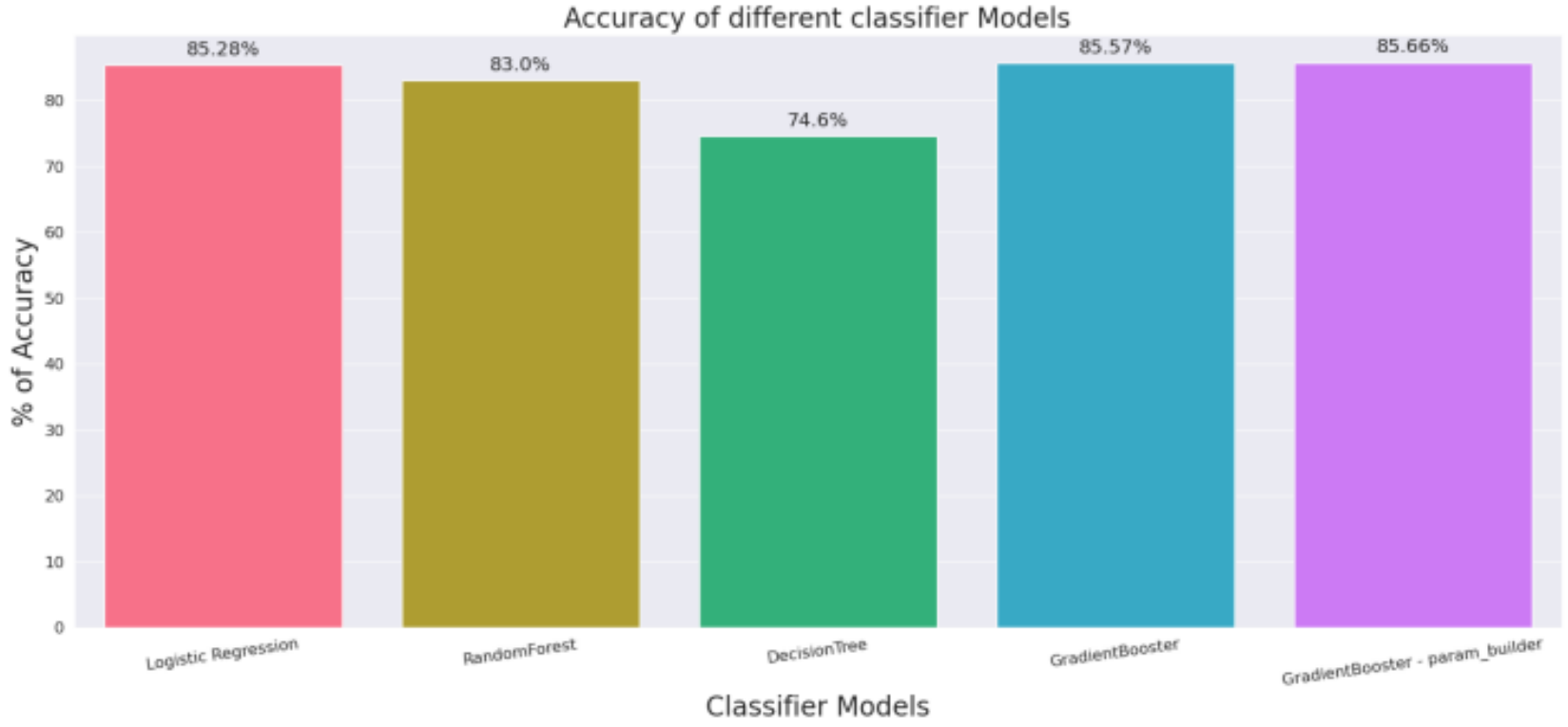
We can observe that in our dataset from 96 smoking persons 66 survived (30 not), so is giving us a survival rate of 68,75% smoking persons.

Has the smoking an influence on the ejection fraction?



Smoking causes slight decrease in ejection fraction

Data models accuracy comparison:



Summary

The best accuracy was achieved with Gradient Booster classifier, but during experiments sometimes even higher accuracies wad achieved for Logistic Regression also. But overall Gradient Booster and Logistic Regression had the highest accuracies.