**Statistical Analysis For Heart Failure Prediction**

**Team Details:**

1. Ananya Samala - 11527196

2. Naveen Vutukuri - 11551117

3. Bindu Gadusu - 11609181

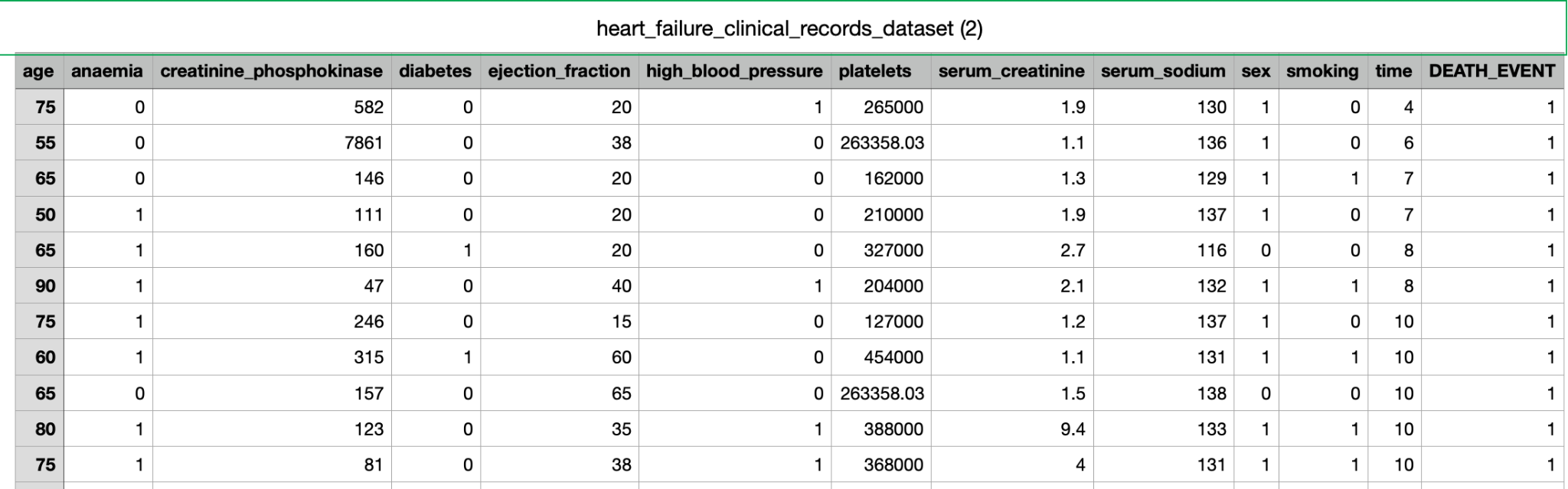
4. Phani Krishna Yadlapati - 11550369

**Abstract:**

* Cardiovascular diseases or most known as CVD’s are one of the primary contributors to mortality. Killing nearly 18 million people a year, CVD’s make up 31% of all fatalities. Population-wide initiatives can be established to prevent these diseases. These initiatives can be targeted at behavioral risk factors like smoking, lack of nutrition and obesity, physical inactivity and excessive alcohol intake.
* Heart failure occurs when the heart fails to pump blood throughout the body. Heart failure affects older individuals more than younger people, and the most prevalent causes of heart failure include excessive blood pressure, diabetes and smoking.
* In this project, we are analyzing several factors that impact cardiac function, and there are 12 features in the dataset we utilized for the analysis as well as a final output variable that represents the death event that occurs depending on the features in the dataset.

**Data Specification:**

* There are 13 features present in the data they are:
* Age: the Age of person
* Anemia: Decrease in red blood cells
* Creatinine phosphokinase: Level of cpk enzyme in blood
* Diabetes: Patient has diabetes or not
* Ejection Fraction: Percentage of blood leaving the heart at each cycle
* High blood pressure: Patient has hypertension or not
* Platelets: Platelets count
* Serum creatinine: serum creatinine level in blood
* Serum sodium: serum sodium level in the blood
* Sex: Gender of a patient
* Smoking: patient smokes or not
* Time: Follow-up period
* Death Event: patient deceased during the follow-up period or not



These are different features that are present in the dataset which are used to build a predictive model to identify heart failures.

* The dataset also included one final variable, which was the Death Event based on all of the features listed above.
* We are excluding younger people because they are less prone to heart failure. In the dataset, the age factor begins at 40 years.
* In the dataset, we have the platelet count, ejection fraction, creatinine phosphokinase, and serum sodium.
* Then there are other features that indicate whether something is positive or negative, such as anemia, diabetes, high blood pressure, sex, and smoking.

**Potential Statistical tests:**

* **Two sample paired T-test**: When the data from two samples are statistically independent, the two-sample t-test is employed. A paired samples t-test is used to compare the means of the two samples where every observation in one sample can be paired with an observation in the other sample.
* **Two sample unpaired T-test**: To compare the means of two independent groups, the unpaired two-samples t-test is employed. For instance, let's say we measured the height of 50 persons, 25 of whom were women (group x) and 25 of whom were men (group y). We're curious to know if women's mean heights (mx) differ noticeably from men's (my).
* **Cross Validation**: To improve overall model performance on unobserved data, we intend to make model adjustments. Performing substantially better on test sets can result from hyperparameter adjustment. A model may perform poorly on data that has not yet been seen if parameters are optimized for the test set because of knowledge leakage. We can use cross validation to make amends for this.

(or)

A statistical technique called cross-validation is used to gauge the effectiveness of machine learning models. It is a technique for determining how well a statistical analysis' findings will translate to a different collection of data.

**Analysis:**

In this project we are doing two different kinds of analysis they are:

1. Univariate Analysis
2. Bivariate Analysis

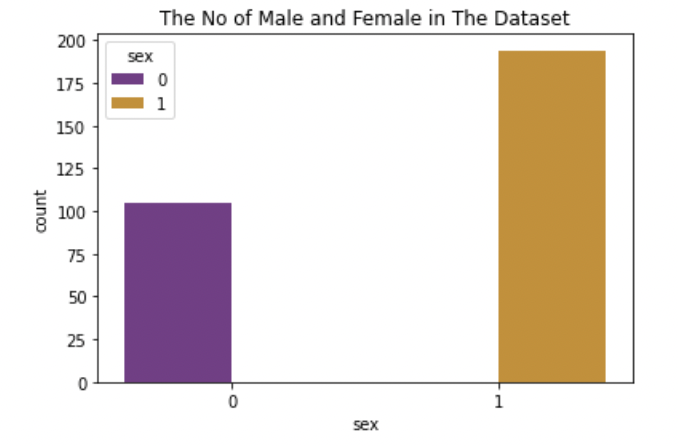
Univariate Analysis: In this univariate analysis we are doing data exploration when we start the analysis and coming to the use of univariate it will show the description of single values that we are interested in doing bivariate analysis.

Bivariate Analysis: In this Bivariate analysis firstly bi means two which means we are finding the relationships between the two data sets which is also called as the two samples data analysis we are

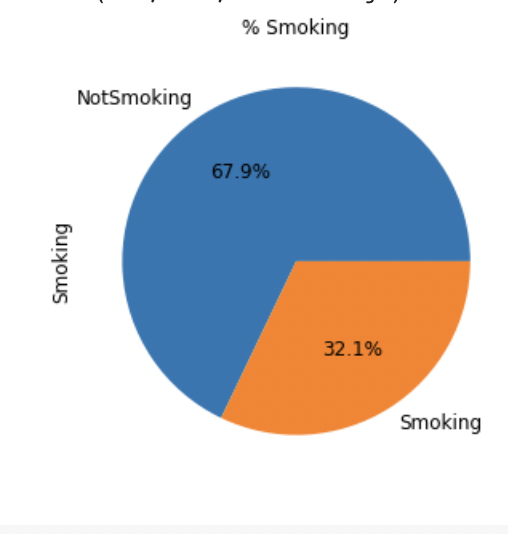
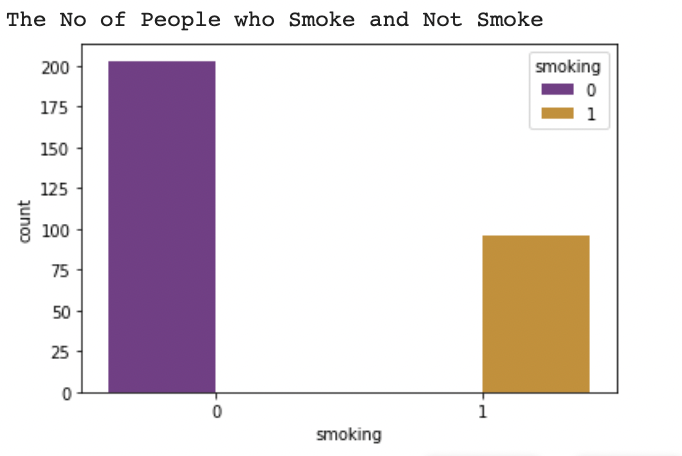
**Data Analysis:**

Univariate Analysis:

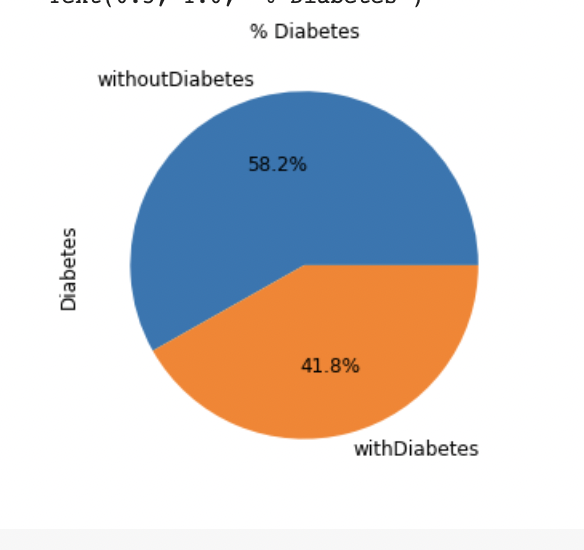
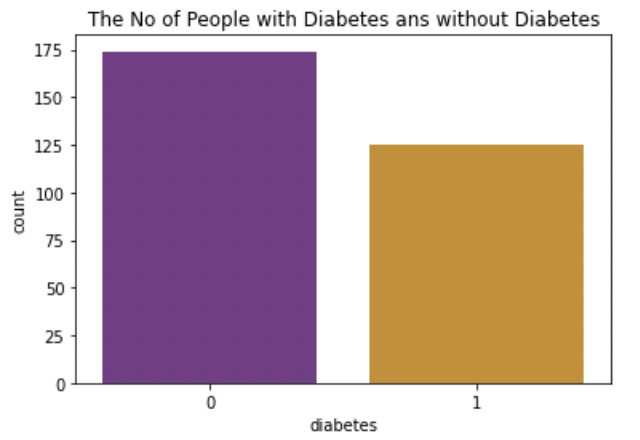
1. Percentage of Male & Female



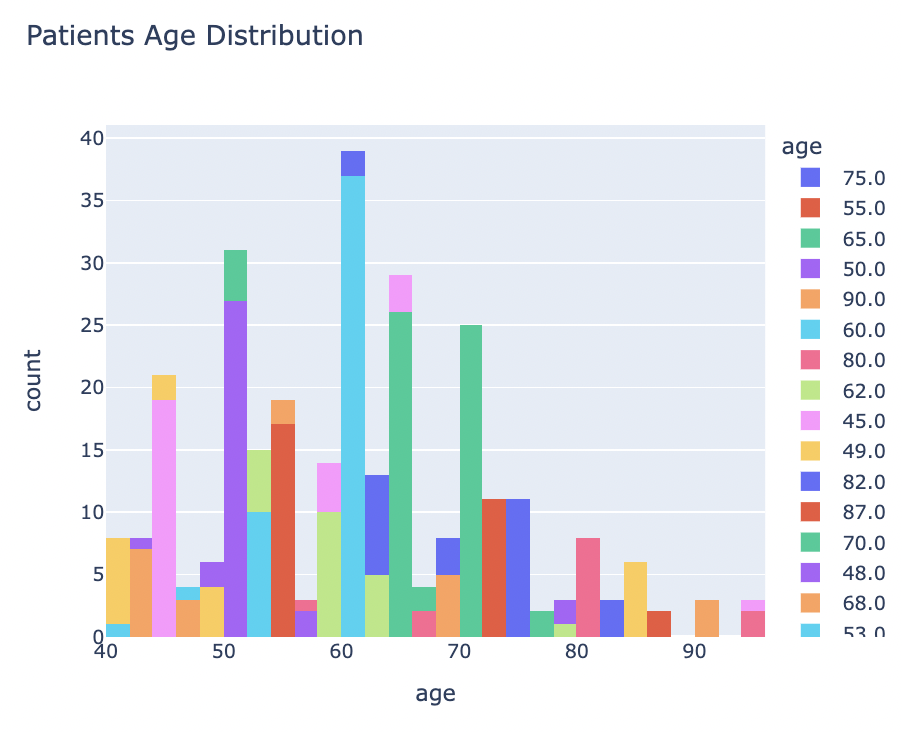
1. Percentage of people smoke & not smoke



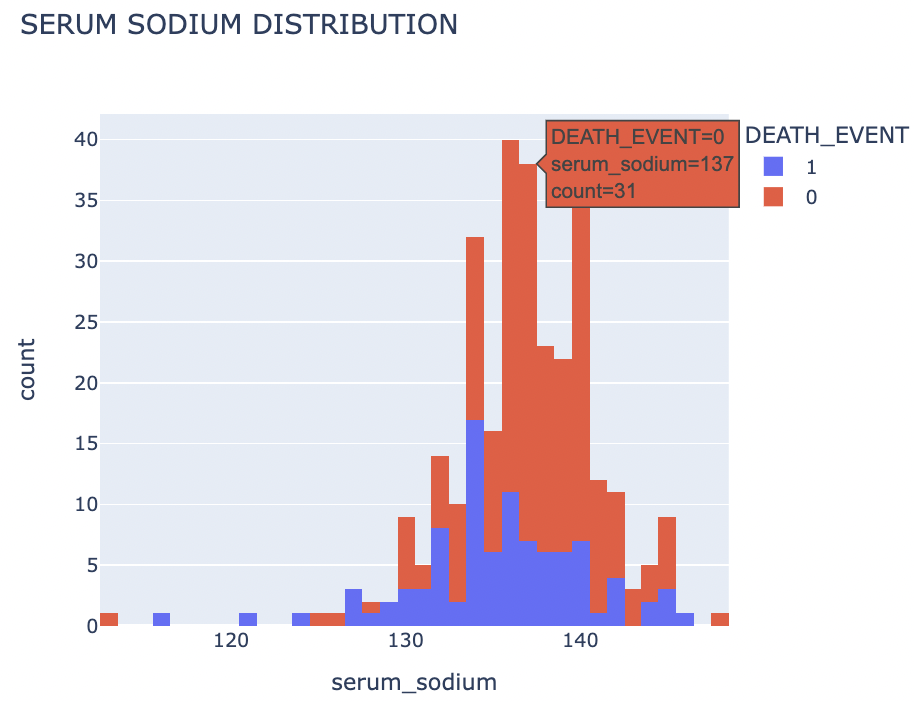
1. Percentage of diabetic and not diabetic



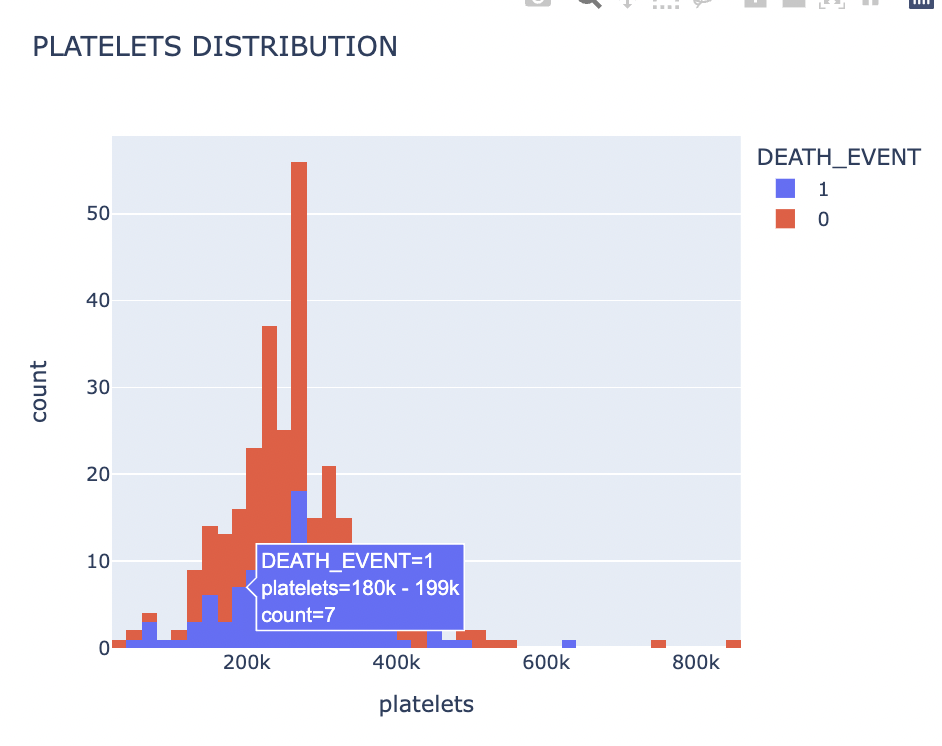
1. Plot for Patient’s Age distribution



1. Plot for serum Sodium distribution

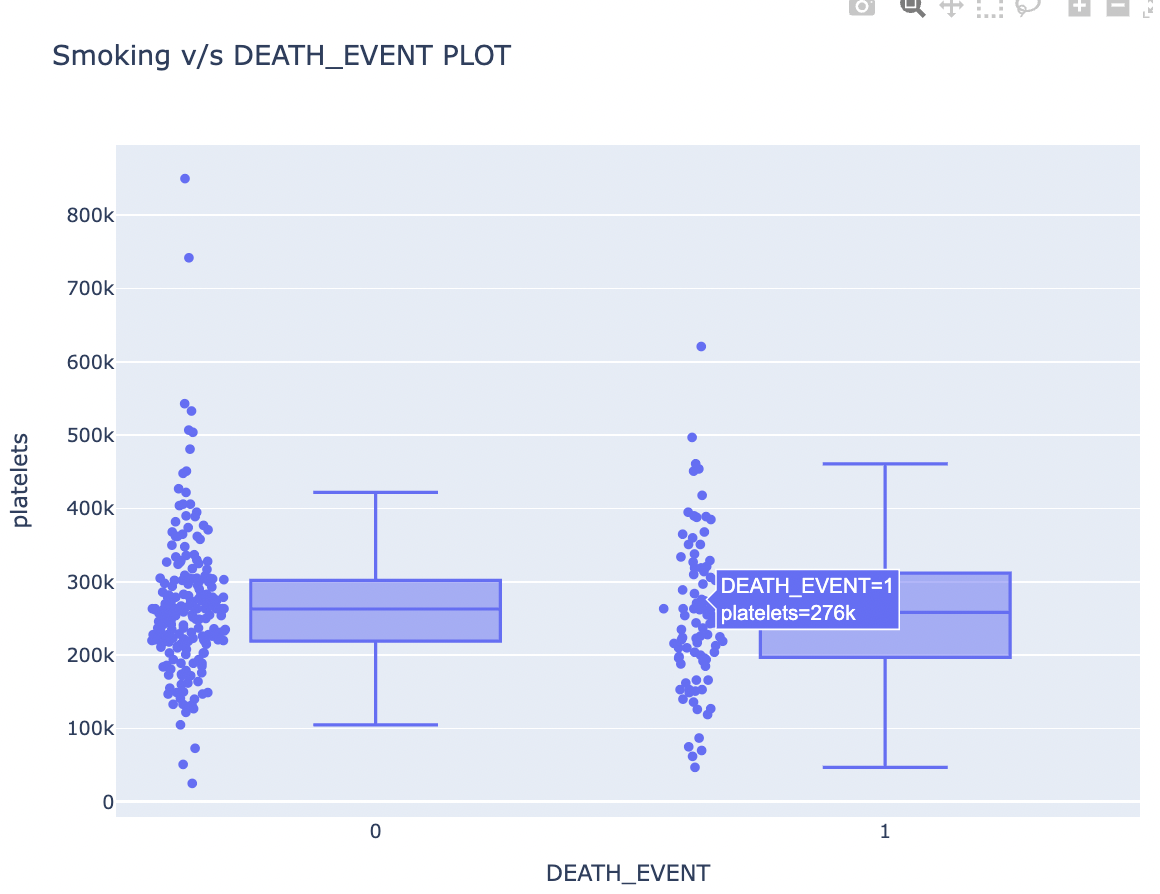


1. Plot for Platelets distribution

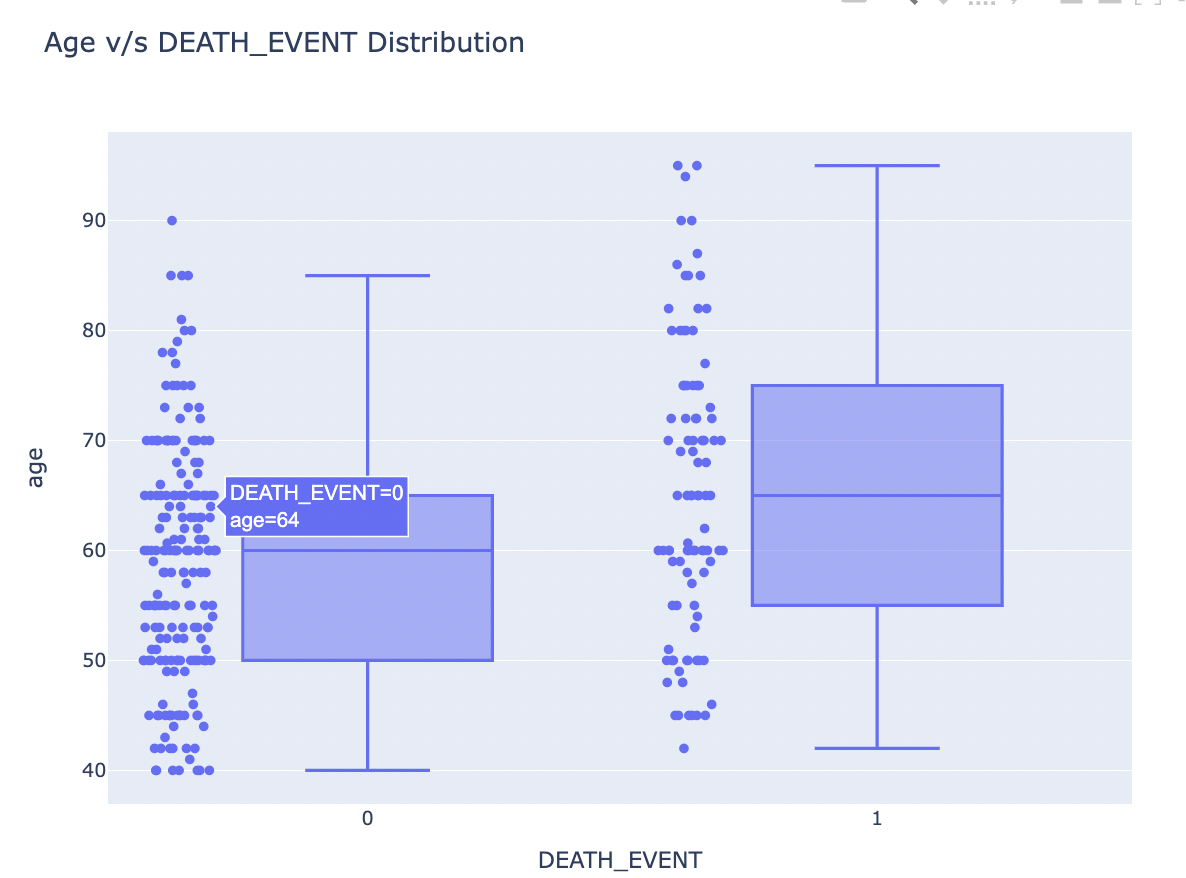


Bivariate Analysis:

1. Boxplot for distribution of Smoking & Death Event



1. Boxplot for distribution of Age & Death Event



**Implementation:**

In the statistical analysis we implement 3 different T-Tests

1. One sample T- test
2. Two Sample Unpaired T- Test
3. Two Sample Paired T- Test

In this project we use machine learning models and compare the accuracy. The two models are

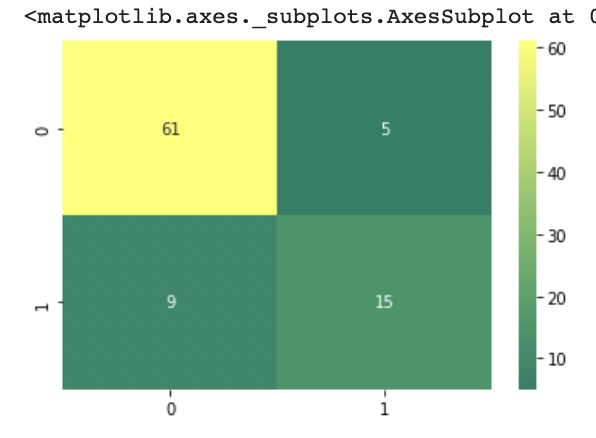
1. Logistic Regression
2. Decision trees

In the models we train the data and then split then test the results. We use cross validation which plays a major role in the project.

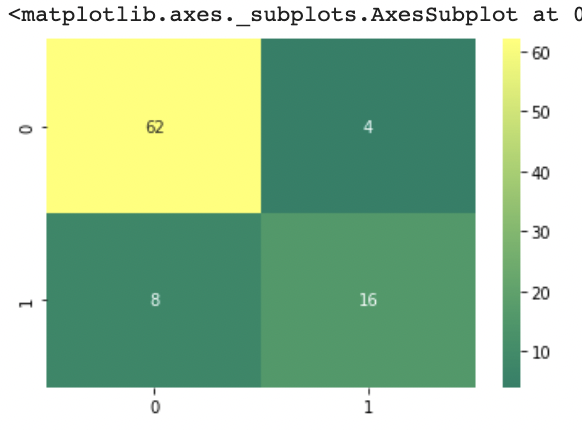
**Preliminary Results:**

* **Model results:**

1.Confusion Matrix for Decision Tree



2.Confusion Matrix for Logistic Regression



* **Statistical Results:**

1. **One sample T- Test**

****

1. **Two Sample Unpaired T-test**

****

1. **Two Sample Paired T-test**

****

**Project Management:**

* **Description:**  For the data we performed data preprocessing, data cleaning & visualization, data modeling , statistical approach
* **Contribution:**
* Ananya Samala (25%)
* Naveen(25%)
* Bindu(25%)
* Phani(25%)

**Implementation Status Report:**

We implemented the machine learning models and the T test ‘s for the dataset .

**Work Completed:**

* Data preprocessing - Phani
* Data cleaning & visualization - Bindu
* Data modeling - Ananya
* Statistical approach - Naveen

**Work to be Completed:**

We need to analysis and compare the results and validate how accurate the the model prediction from the results.

**Resources and Related Projects:**

* **https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical data**
* [**https://www.kaggle.com/code/suprematism/statistics-statistical-tests**](https://www.kaggle.com/code/suprematism/statistics-statistical-tests)
* [**https://www.hindawi.com/journals/cin/2021/8387680/tab1**](https://www.hindawi.com/journals/cin/2021/8387680/tab1)

**Github Link:** [**https://github.com/Naveen4323/Empirical-Project-**](https://github.com/Naveen4323/Empirical-Project-)