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

Goals & Objective:

The main aim of the project is to predict the heart failure in the people of different ages. We use machine learning and statistical approach to analyse the data mainly we use T-test in this project. We implement the models for better accuracy.

Data Specification:

- There are 13 features present in the data they are:

heart_failure_clinical_records_dataset (2)

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1

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.

Analysis:

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

1. Univariate Analysis
2. Bivariate Analysis

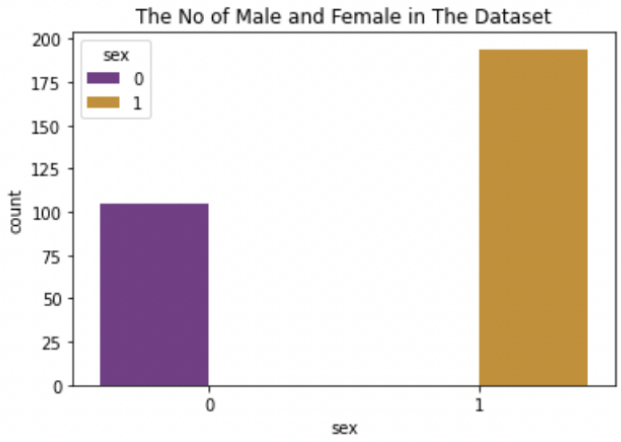
Univariate Analysis: In this univariate analysis we are doing data exploration when we start the analysis and coming to the use of univariate it will show the description of single values that we are interested in doing bivariate analysis.

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

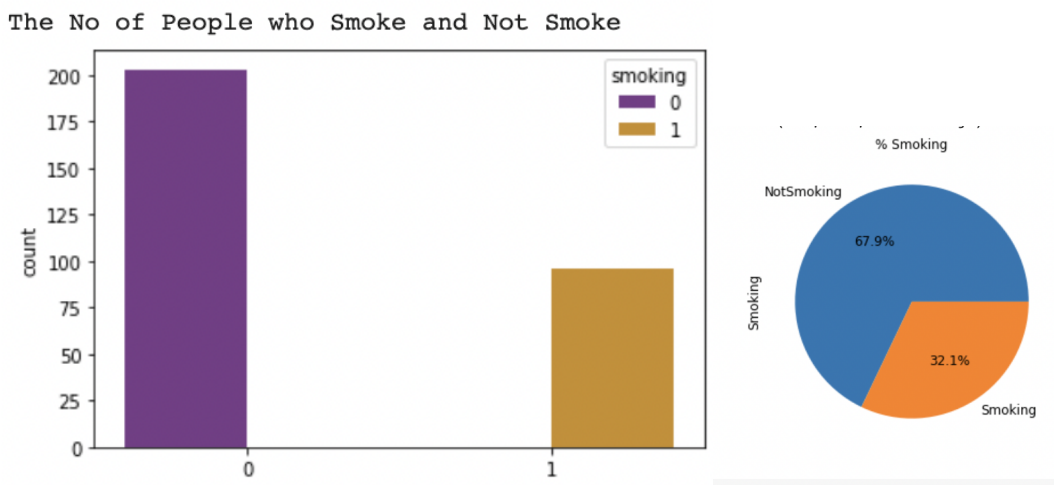
Data Analysis:

Univariate Analysis:

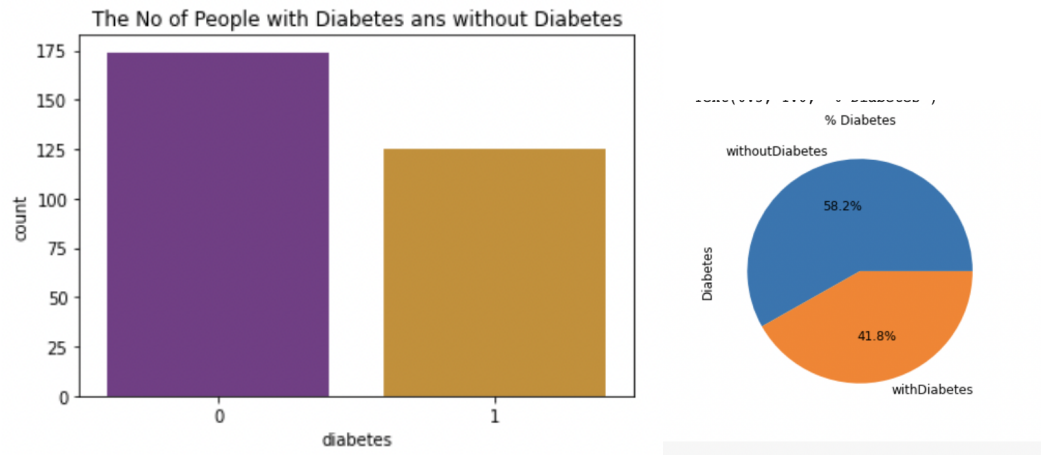
1. Percentage of Male & Female



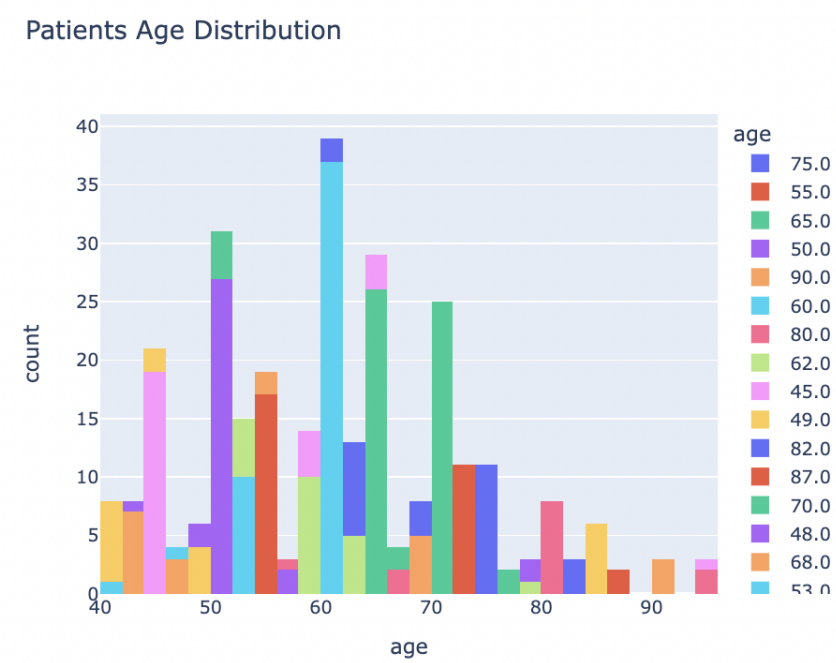
2. Percentage of people smoke & not smoke



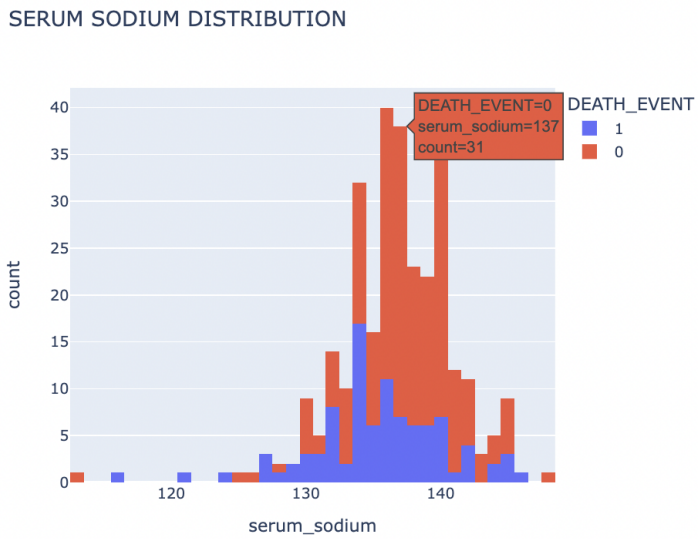
3. Percentage of diabetic and not diabetic



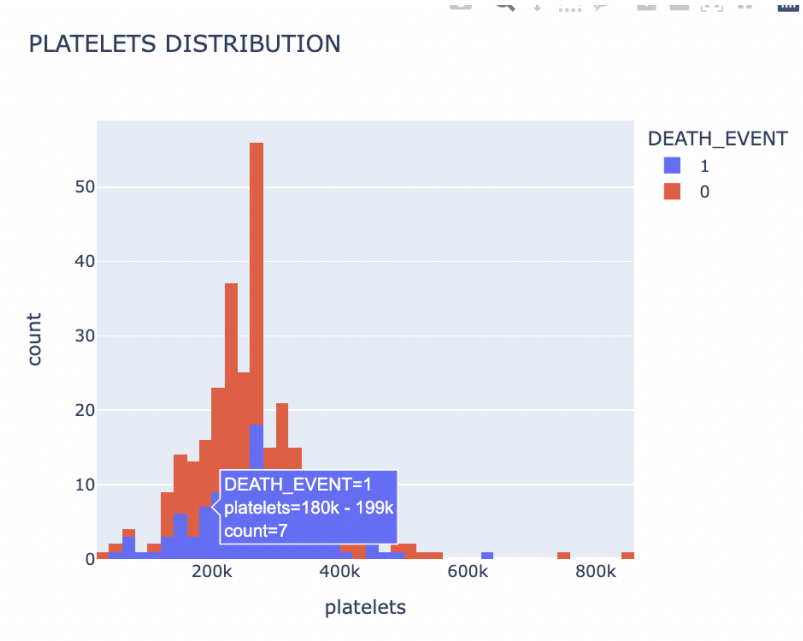
4. Plot for Patient's Age distribution



5. Plot for serum Sodium distribution



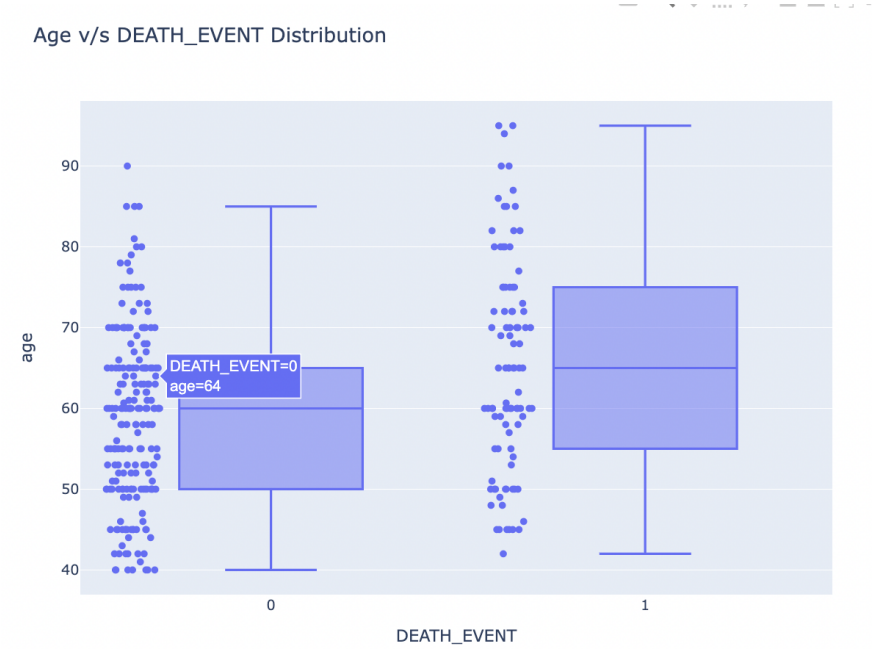
6. Plot for Platelets distribution



Bivariate Analysis:

1. Boxplot for distribution of Smoking & Death Event

2. 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

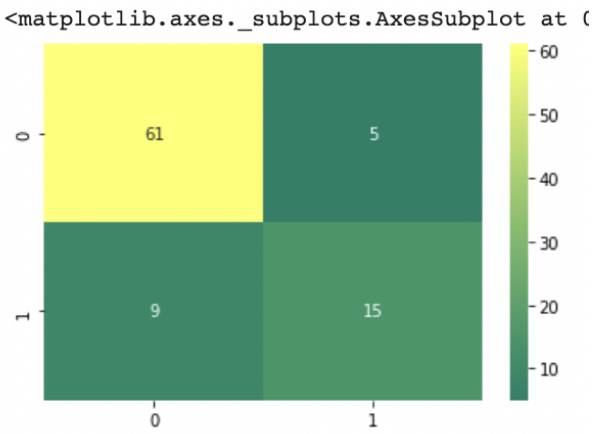
1. Logistic Regression
2. Decision trees

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:

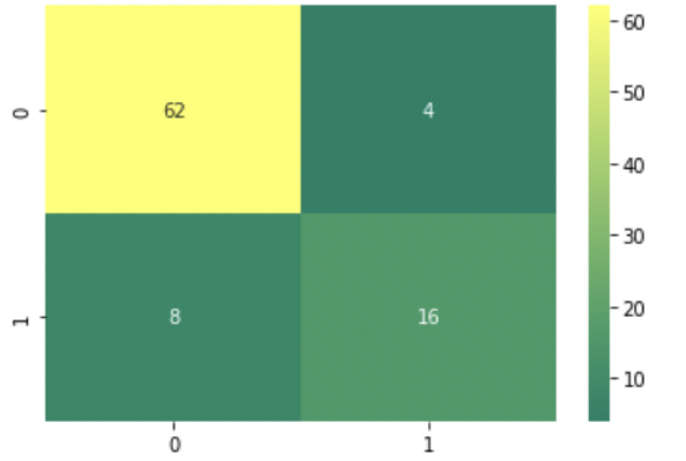
❖ Model results:

1. Confusion Matrix for Decision Tree



2. Confusion Matrix for Logistic Regression

<matplotlib.axes._subplots.AxesSubplot at 0



❖ Statistical Results:

1. One sample T- Test

```
[67] male_age=data.loc[data.sex==1,'age']

[68] male_age.mean()

     61.405499999999996

[69] scipy.stats.ttest_1samp(male_age,popmean=60)

     Ttest_1sampResult(statistic=1.6014139978952975, pvalue=0.11092074578307773)

[70] female_age=data.loc[data.sex==0,'age']
     female_age.mean()

     59.777780952380944

[71] scipy.stats.ttest_1samp(female_age,popmean=60)

     Ttest_1sampResult(statistic=-0.20256952029853986, pvalue=0.8398672748505357)
```

2. Two Sample Unpaired T-test

```
[73] male=data.loc[data.sex==1,'platelets' ]

[74] female=data.loc[data.sex==0,'platelets']

[75] scipy.stats.ttest_ind(male,female)

     Ttest_indResult(statistic=-2.1733666381904304, pvalue=0.03054272778909673)

[76] male_ejectionfraction=data.loc[data.sex==1,'ejection_fraction']

[77] female_ejectionfraction=data.loc[data.sex==0,'ejection_fraction']

[78] scipy.stats.ttest_ind(male_ejectionfraction,female_ejectionfraction)

     Ttest_indResult(statistic=-2.585864165814152, pvalue=0.010189690578840916)
```

3. Two Sample Paired T-test

```
male=data.loc[data.sex==1,'age' ]

[85] male1=male[:97]
     male1.shape

     (97,)

[86] male2=male[97:]
     male2.shape

     (97,)

[87] scipy.stats.ttest_rel(male1,male2)

     Ttest_relResult(statistic=3.129543569066593, pvalue=0.0023192690887055417)

[88] female=data.loc[data.sex==0,'age' ]
     female.shape

     (105,)

[89] female1=female[:52]
     female2=female[52:104]

[90] scipy.stats.ttest_rel(female1,female2)

     Ttest_relResult(statistic=1.1424610281926113, pvalue=0.2585976712210964)
```

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 analyze and compare the results and validate how accurate the model prediction is 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.hindawi.com/journals/cin/2021/8387680/tab1**

Github Link: **https://github.com/Naveen4323/Empirical-Project-**