CSCE 5310 - Methods in Empirical Analysis

Increment - 2

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

Introduction:

Cardiovascular diseases are one of the most common factors that can cause the mortality which nearly kills more than the 18 million people in the world. They are several initiatives taken to prevent these risks and mainly these can occur due to behavioral habits like smoking, lack of nutrition and obesity and lack of physical work and also excessive alcohol consumption can cause the heart diseases.

Heart failure occurs only when it is having an issue with pumping blood to heart and coming to issue it will cause more due to the excessive blood pressure and diabetes and smoking.

In this project we are analyze causes that can lead to heart failure we are have 12 features those we are used for analysis and there is having one feature that occur death event which is dependent to other features. In this we are doing different kinds of tests which is used to take the quantitative decisions using the given models.

Background:

Nowadays we observe there are many cases of heart diseases in order to prevent them we perform statistics analysis and predict by previous observations. In statistics we use some of the models and methods by which our analysis is accurate.

Model: In this project we are using two machine learning models and compares their accuracy. They are

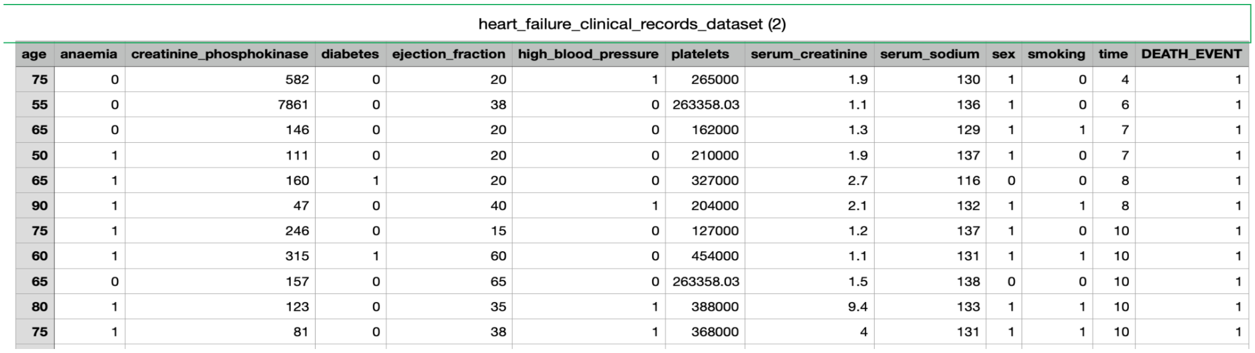
1. Logistic Regression
2. Decision Tree

1. **Logistic Regression:** Logistic regression estimates the probability of event occur based on the previous data values. the output is in between 0 and 1 as the outcome is probability event.

2. **Decision tree:** Decision tree algorithm is a supervised learning algorithm which is preforms both tasks classification and regression tasks. And it's a tree structure which has root node, internal nodes, and leaf nodes

Dataset:

In the data set there are totally 13 values present in the whole dataset. In those features they are several features which are used to build a predictive model to identify the heart failures and also the data set has 13 features in that one feature can decide the final event as it describes the death event which is dependent by 12 features and below the detail description of features.



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

The above are the features of our dataset and coming to the detailed description

- In the dataset, we have the platelet count, ejection fraction, creatinine phosphokinase, and serum sodium
- we are have several features which are the Boolean values like 0 or 1 they are anemia, diabetes, blood pressure, sex, smocking and also death event.
- Death event is the one which is final variable which can decide based on the 12 features it is also a Boolean value

Data Analysis:

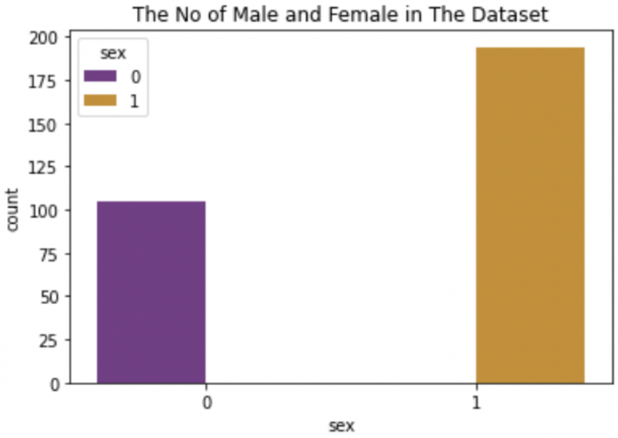
In the data analysis part we perform the data preprocessing in the data preprocessing first we count the features and differentiate the features and also find the null values and clearly understand the features by the count of the features and later we are perform different types of analysis on the data features they are

1. Univariate analysis
2. Bivariate analysis

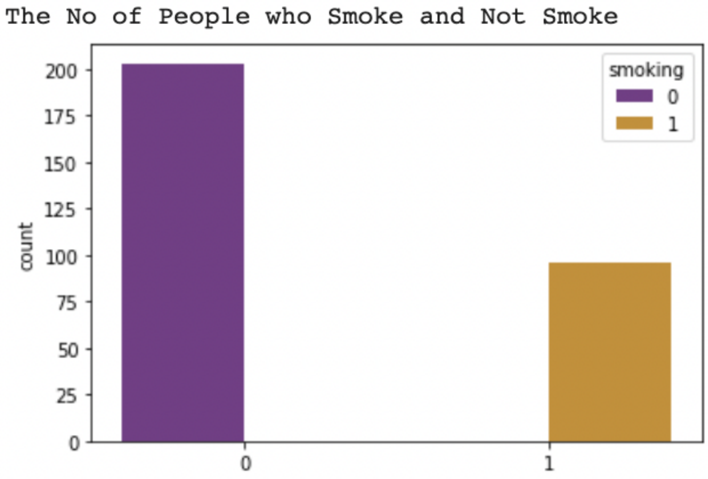
1. Univariate analysis:

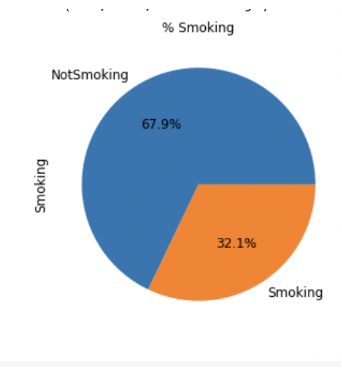
In the univariate analysis it will show us the description of values features that are present in the dataset that we are using for bivariate analysis

1. Male & Female ratios

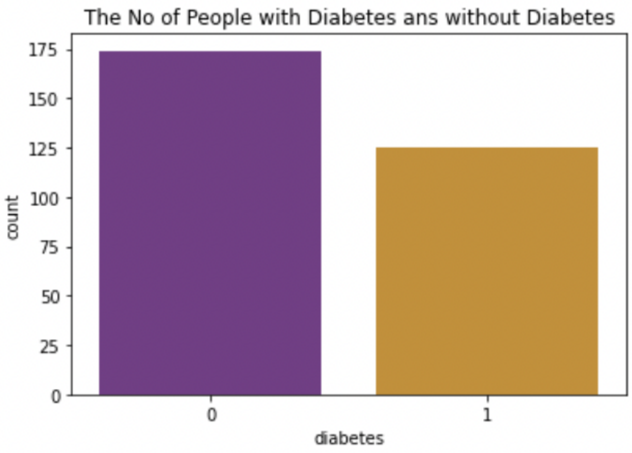


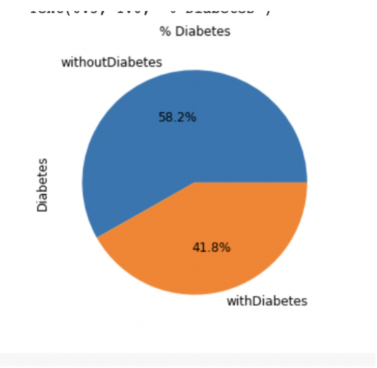
2. Smoke & not smoke ratios



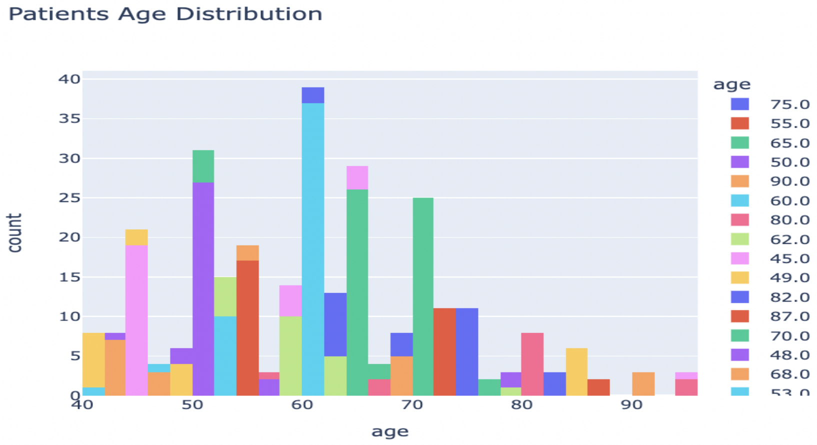


3. Diabetic and not diabetic ratios

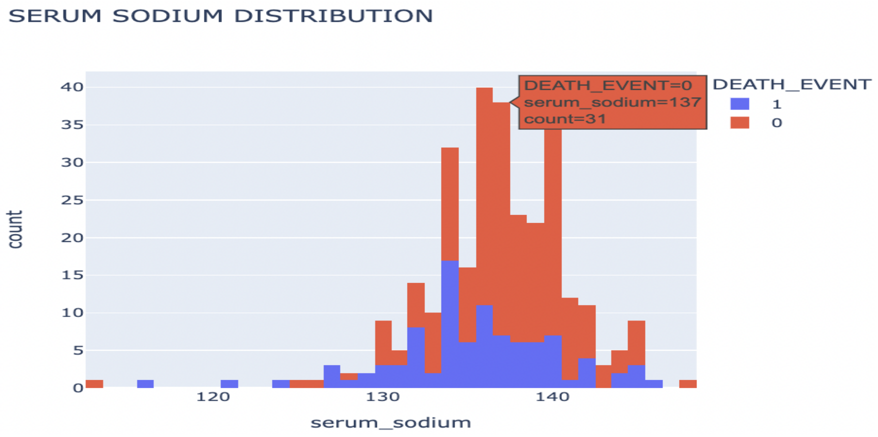




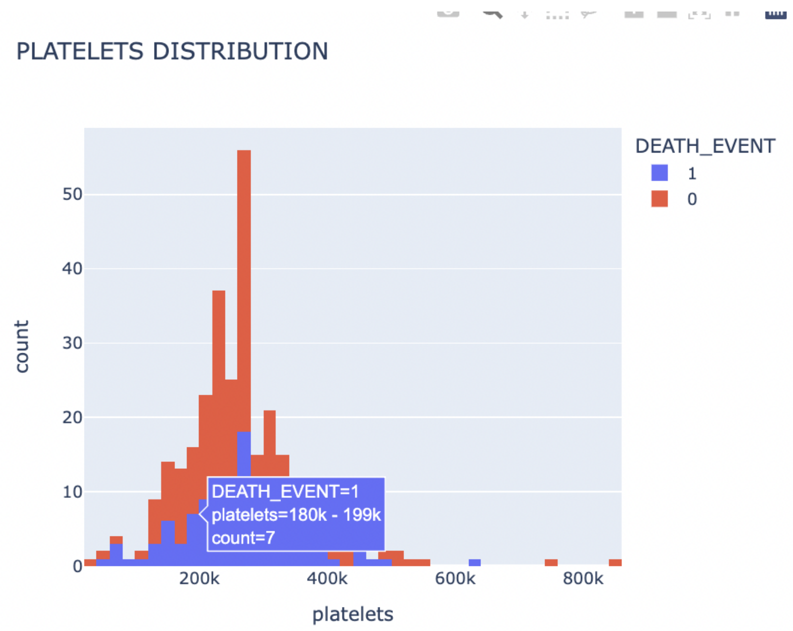
4. Patient’s Age distribution



5. Serum Sodium distribution plot



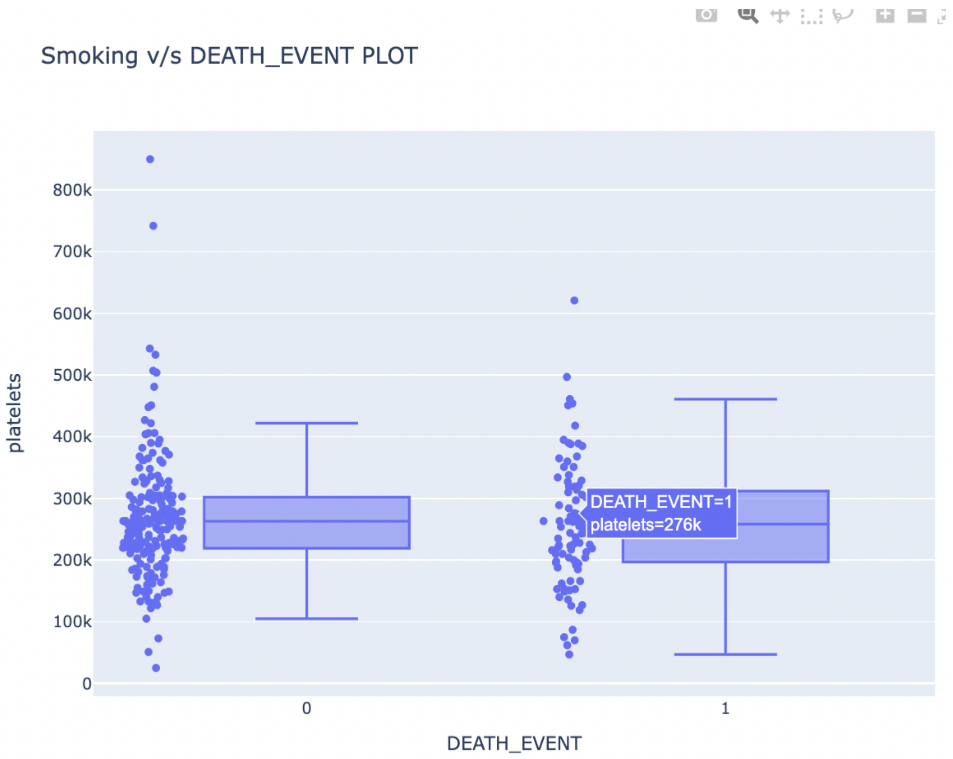
6. Plot for Platelets distribution



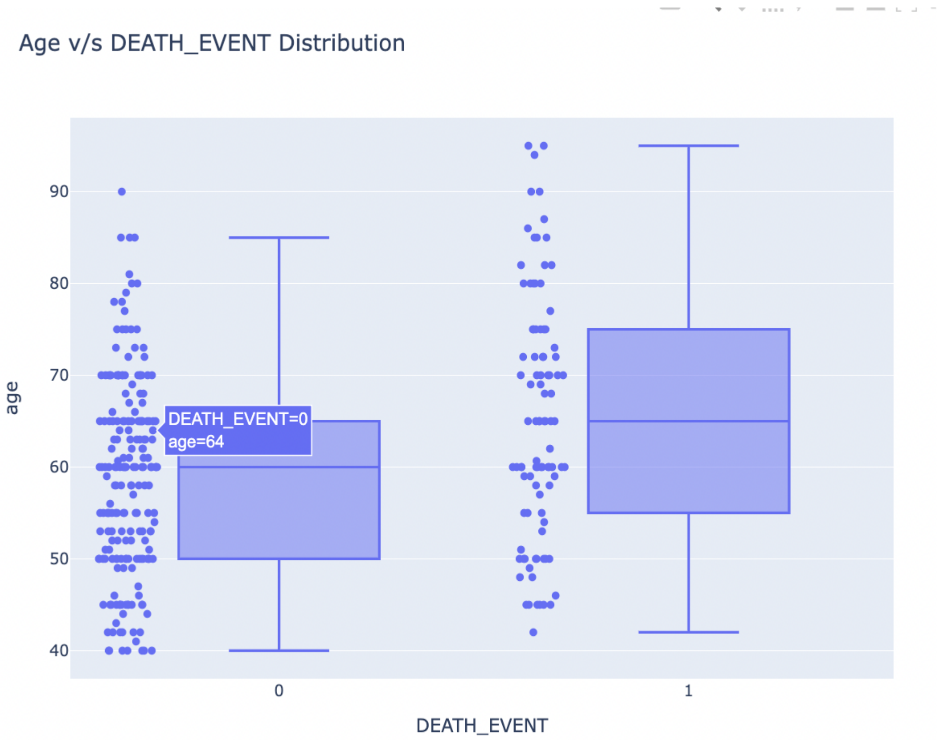
2. **Bivariate analysis:**

In this bivariate analysis it means that bi means two which is the analysis between 2 samples in this bivariate we observe the relationship among different features.

1. Smoking & Death Event distribution plot



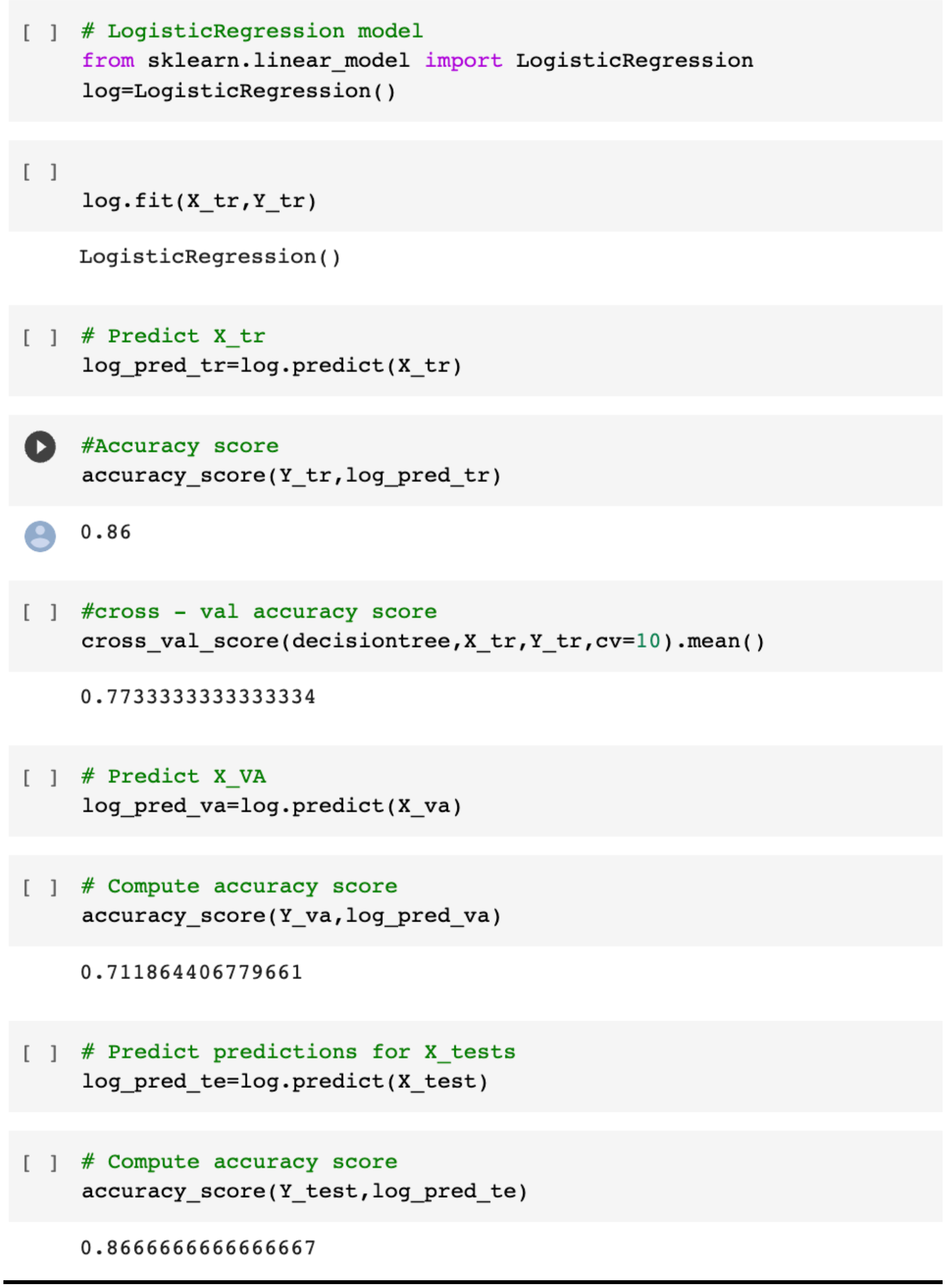
2. Age & Death Event distribution plot



Implementation:

We are implementing this model using both logistic regression and decision tree below are the pseudocodes for the algorithms

Pseudocode for the logistic regressions



```
[ ] # LogisticRegression model
    from sklearn.linear_model import LogisticRegression
    log=LogisticRegression()
```

```
[ ]
    log.fit(X_tr,Y_tr)

    LogisticRegression()
```

```
[ ] # Predict X_tr
    log_pred_tr=log.predict(X_tr)
```

```
    #Accuracy score
    accuracy_score(Y_tr,log_pred_tr)

    0.86
```

```
[ ] #cross - val accuracy score
    cross_val_score(decisiontree,X_tr,Y_tr,cv=10).mean()

    0.7733333333333334
```

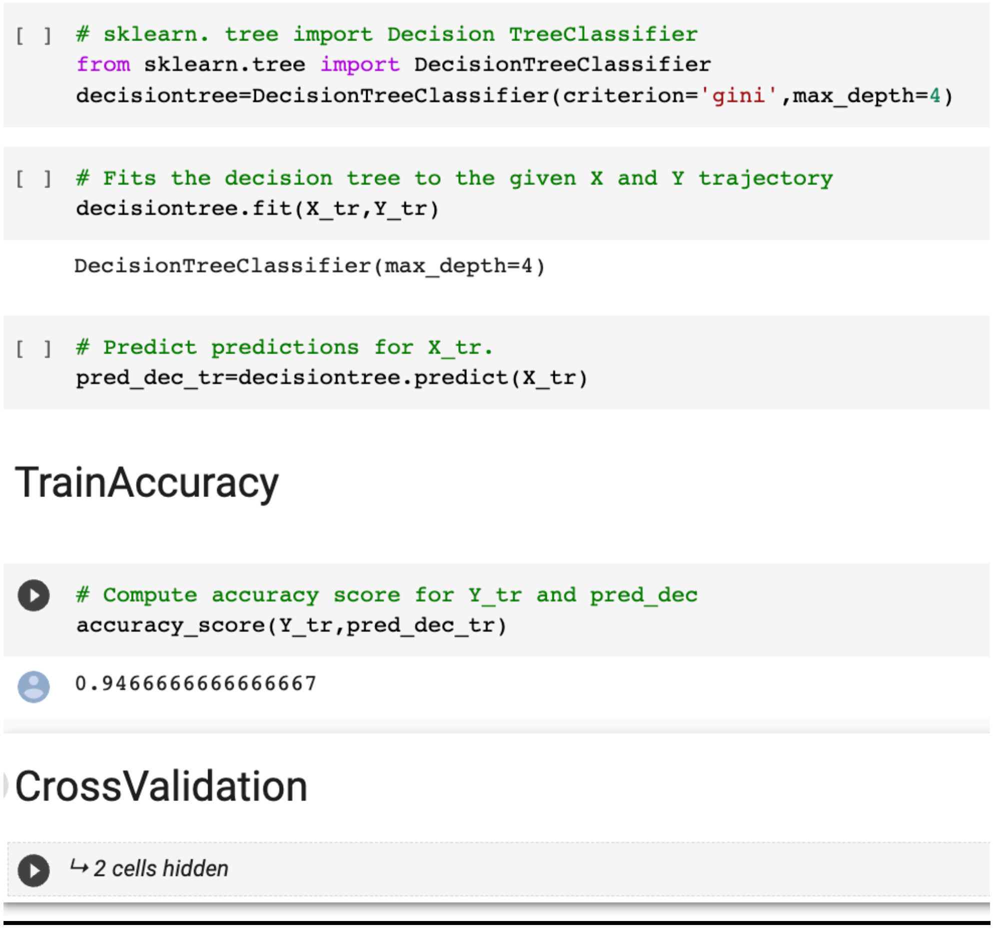
```
[ ] # Predict X_VA
    log_pred_va=log.predict(X_va)
```

```
[ ] # Compute accuracy score
    accuracy_score(Y_va,log_pred_va)

    0.711864406779661
```

```
[ ] # Predict predictions for X_tests
    log_pred_te=log.predict(X_test)
```

```
[ ] # Compute accuracy score
    accuracy_score(Y_test,log_pred_te)

    0.8666666666666667
```

Pseudocode for decision algorithm:



```
[ ] # sklearn. tree import Decision TreeClassifier
    from sklearn.tree import DecisionTreeClassifier
    decisiontree=DecisionTreeClassifier(criterion='gini',max_depth=4)
```

```
[ ] # Fits the decision tree to the given X and Y trajectory
    decisiontree.fit(X_tr,Y_tr)

    DecisionTreeClassifier(max_depth=4)
```

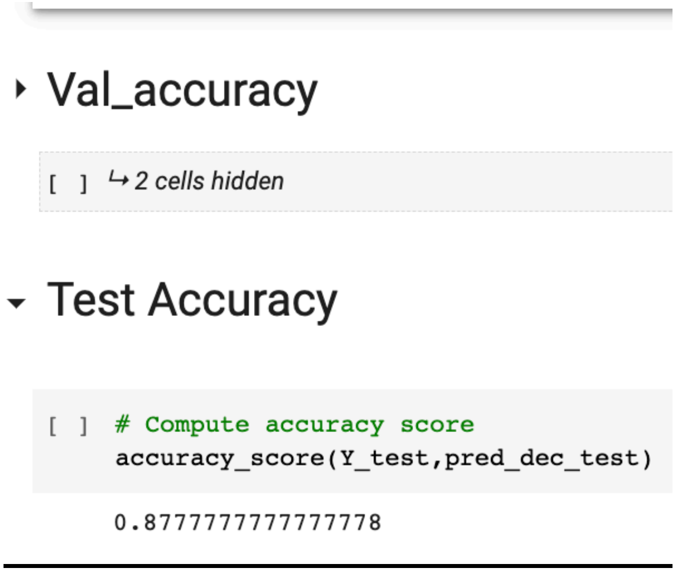
```
[ ] # Predict predictions for X_tr.
    pred_dec_tr=decisiontree.predict(X_tr)
```

TrainAccuracy

```
# Compute accuracy score for Y_tr and pred_dec
accuracy_score(Y_tr,pred_dec_tr)

0.9466666666666667
```

CrossValidation

↳ *2 cells hidden*



Val_accuracy

[ ] ↳ *2 cells hidden*

Test Accuracy

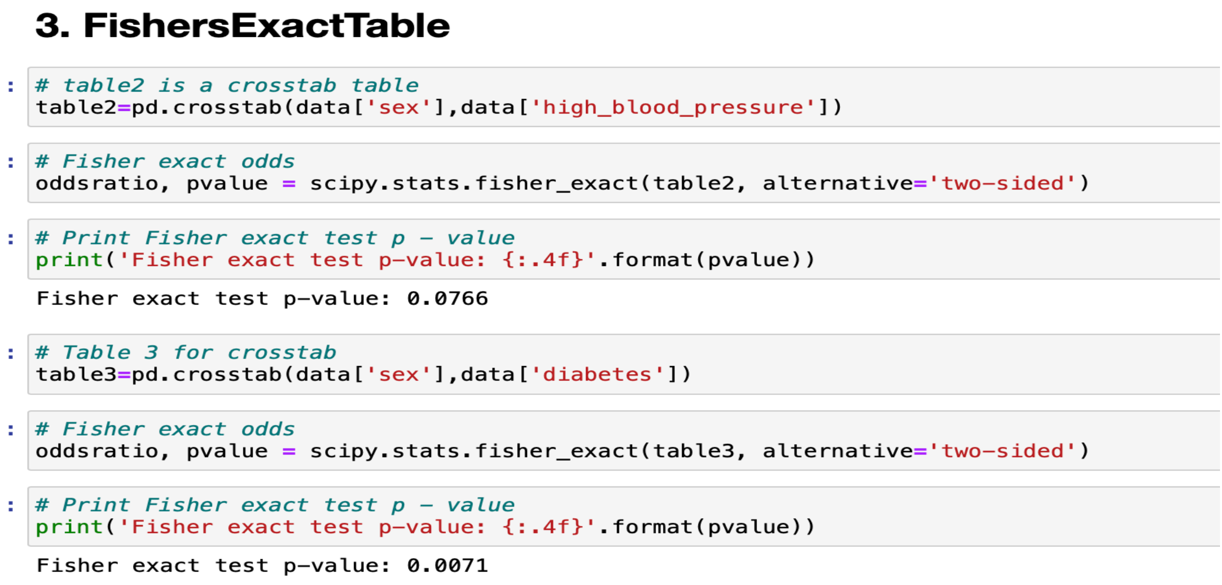
```
[ ] # Compute accuracy score
    accuracy_score(Y_test,pred_dec_test)

    0.8777777777777778
```

Statistical Tests:

We are perform different statistical tests to estimate the sample mean. Among the population they are

1. One sample t-test
2. Two sample paired t-test
3. Two sample unpaired t-test
4. Fisher exact table
5.

In addition to the previous work we implement one more test which is Fisher Exact Table.



3. FishersExactTable

```
: # table2 is a crosstab table
  table2=pd.crosstab(data['sex'],data['high_blood_pressure'])
```

```
: # Fisher exact odds
  oddsratio, pvalue = scipy.stats.fisher_exact(table2, alternative='two-sided')
```

```
: # Print Fisher exact test p - value
  print('Fisher exact test p-value: {:.4f}'.format(pvalue))

  Fisher exact test p-value: 0.0766
```

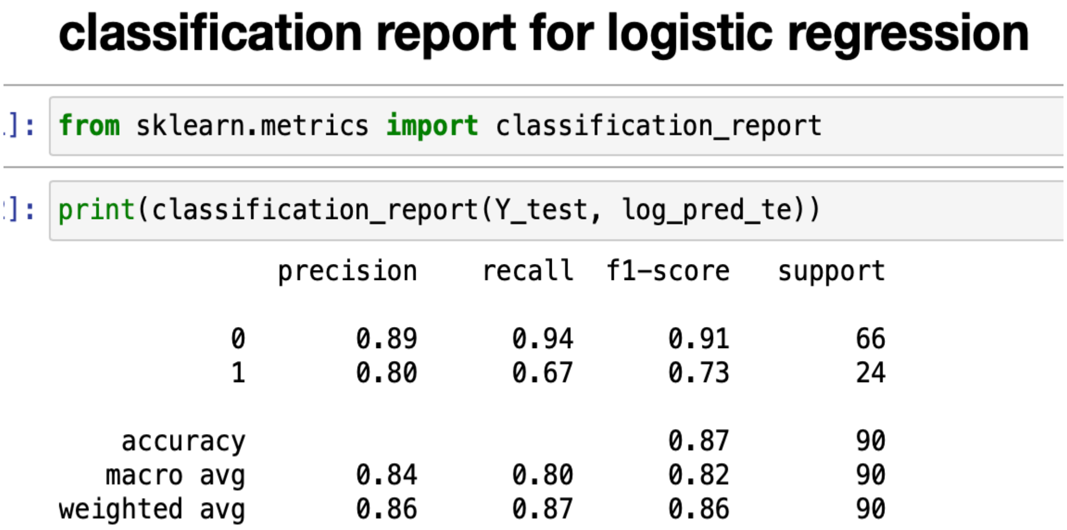
```
: # Table 3 for crosstab
  table3=pd.crosstab(data['sex'],data['diabetes'])
```

```
: # Fisher exact odds
  oddsratio, pvalue = scipy.stats.fisher_exact(table3, alternative='two-sided')
```

```
: # Print Fisher exact test p - value
  print('Fisher exact test p-value: {:.4f}'.format(pvalue))

  Fisher exact test p-value: 0.0071
```

Results:

As we perform different models, we are use cross validation techniques also to get better accuracy and we got the highest test accuracy for the decision tree than logistic regression we obtained 87 accuracy for decision tree and 86 for logistic regression.

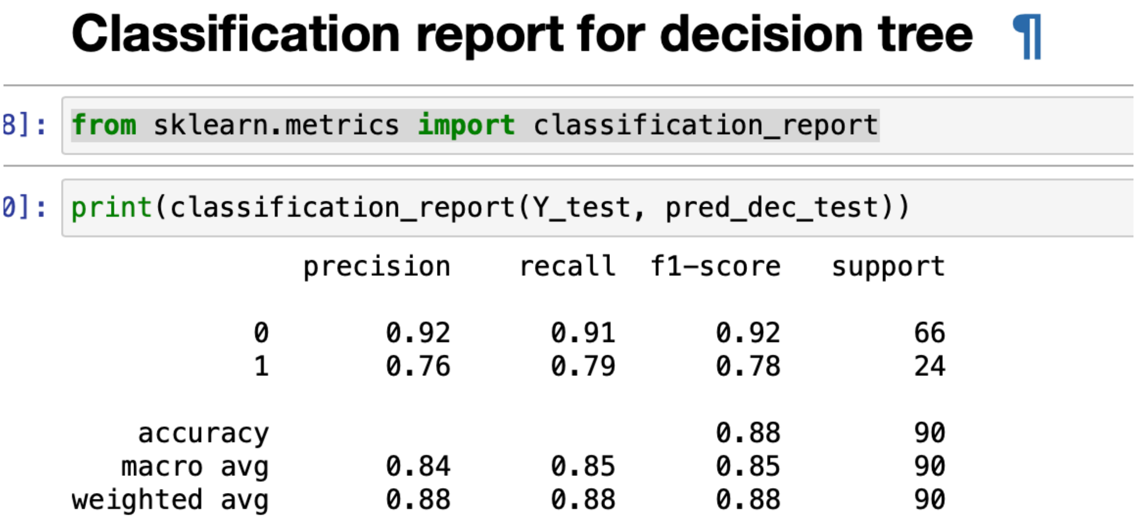


classification report for logistic regression

```
from sklearn.metrics import classification_report
```

```
print(classification_report(Y_test, log_pred_te))
```

```
              precision    recall  f1-score   support

           0       0.89      0.94      0.91        66
           1       0.80      0.67      0.73        24

    accuracy                           0.87        90
   macro avg       0.84      0.80      0.82        90
weighted avg       0.86      0.87      0.86        90
```



Classification report for decision tree

```
from sklearn.metrics import classification_report
```

```
print(classification_report(Y_test, pred_dec_test))
```

```
              precision    recall  f1-score   support

           0       0.92      0.91      0.92        66
           1       0.76      0.79      0.78        24

    accuracy                           0.88        90
   macro avg       0.84      0.85      0.85        90
weighted avg       0.88      0.88      0.88        90
```

Project management:

Work completed:

we performed data preprocessing and data cleaning and also, we use data visualization techniques to show the results and also we use different models for the prediction of the accuracies and also we perform different statistical tests to test sample means on whole population.

Contribution:

Ananya Samala – 25%
Naveen Vutukuri – 25%
Bindu Gadusu – 25%
Phani Krishna Yadlapati – 25%

Work Completed:

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

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

++++