Title:

Student ID: s3881257

Student Name and email (contact info): Jagulan Srikaran (s3881257@student.rmit.edu.au)

Affiliations: RMIT University.

Date of Report: 25/05/2022

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Contents for task 4

Task 1

Introduction

The data set I have chosen for this assignment is the clinical record of heart failure. It has 13 attributes and 299 instances. The target feature is DEATH_EVENT.

Goal of the project

The **goal** of this project is to use data science process to predict the most important attributes to identify the death event of a person.

This data set was almost cleaned, it had no missing values, it had no spelling errors. Although there were only few errors (Eg: decimal point in age column) , I used appropriate steps to confirm so I can retrieve and prepare the data

Steps I took to retrieve and prepare the data:

Pd.read_csv('dataset')	Retrieve the data
df.isnull().sum()	To find the number of null values, there were no null values.
Df.shape()	To find the number of instance and features.
age.round(decimals=0)	To get rid of any decimal points in age column
Df.info()	To confirm if the data frame is fully clean and approved for data exploration.

Task 2

Task 2 focuses on data exploration.

Df.astypes()	To convert the default data types to appropriate data types.
Df.’feature’.unique()	To confirm there are only 2 values for all Boolean and binary data type column.

2.1

The purpose of task 2.1, is to explore each column.

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000
mean	60.829431	581.839465	38.083612	263358.029264	1.39388	136.625418	130.260870
std	11.894997	970.287881	11.834841	97804.236869	1.03451	4.412477	77.614208
min	40.000000	23.000000	14.000000	25100.000000	0.50000	113.000000	4.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.90000	134.000000	73.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.10000	137.000000	115.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.40000	140.000000	203.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.40000	148.000000	285.000000

In the above table, which is created by db.describe(), 7 numerical variables are explored and all statistical information is displayed. Whereas for the categorical variable, the frequency of the data is shown below.

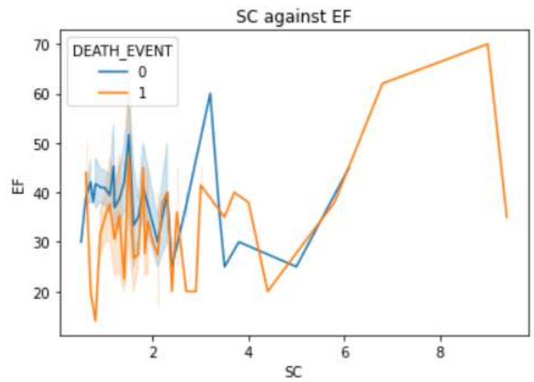
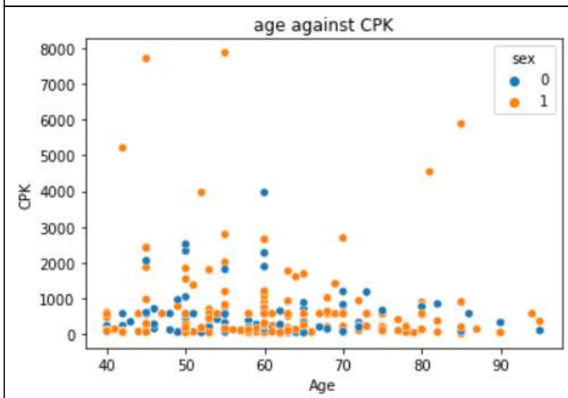
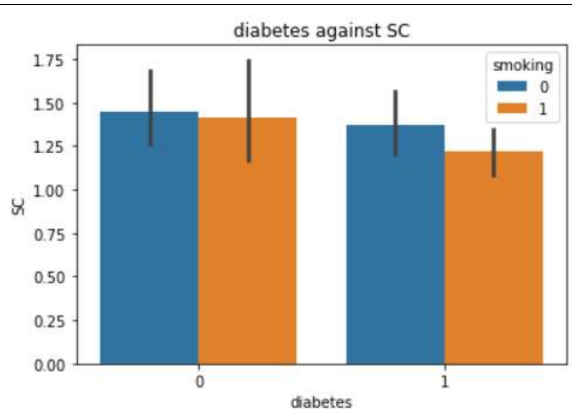
Sex	Frequency
Male	194
Female	105

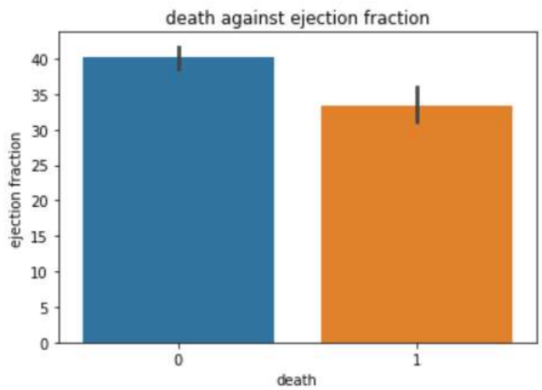
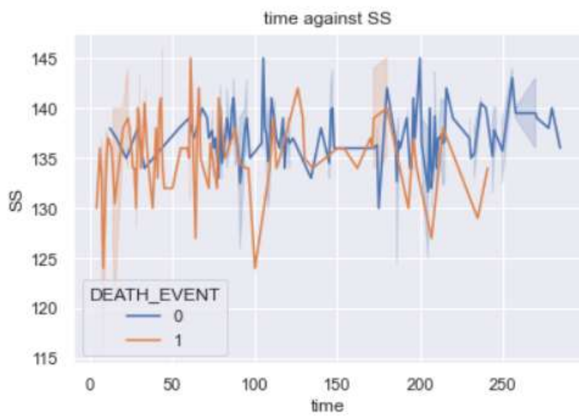
smoking	Frequency
True	203
False	96

High blood pressure	Frequency
True	194
False	105

2.2

Task 2.2, is to display the relationship between the features graphically.

 Hypothesis: if a person has low ejection fraction and low serum creatinine, they would die. They would also die if they have abnormally high EF and SC.	0 displays **not dead** 1 displays **dead** This graph shows the relationship between serum creatinine and ejection fraction. This graph shows that people who are not dead had more ejection fraction and low serum creatine. People who had low ejection fraction and high sodium creatinine died.
 Hypothesis: Men have high CPK levels compared to women.	0 displays **female** 1 displays **Male** This graph displays the relationship between creatinine phosphokinase (CPK) and Age with 2 genders. This graph shows that CPK levels are mostly same with any age for both genders. This shows a lack of relationship since the plot is not too diverse.
 Hypothesis: People who smokes and also has diabetes has low serum creatinine.	0 displays **non-smoking** 1 displays **smoking** This graph shows that people without diabetes who does not smokes have more serum creatinine compared to all others. The people with least serum creatinine has diabetes and they smoke.

 Hypothesis: people with high ejection fraction stays alive.	This graph shows that people with less ejection fraction died where as people who stayed alive have more ejection fraction.
 Hypothesis: People with low serum sodium will die after few follow up period(time).	0 displays **not dead** 1 displays **dead** This graph shows that people with low serum sodium died in the follow up period where as people with high serum sodium stayed alive. No people died after 245 days. There is a lack of relationship between Death event and time.

Above tables have 5 plots with 10 attributes.

Task 3

The purpose of task 3 is to use sklearn and build two models of our chosen method. I chose classification, and the two models I designed are decision tree and K nearest neighbour algorithm.

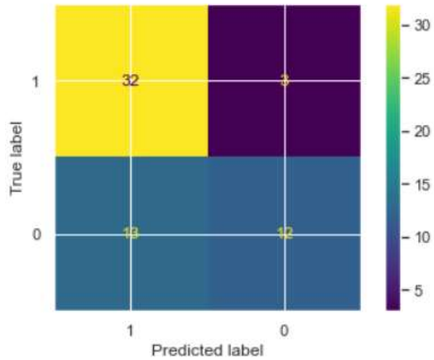
Decision Tree:

I drew 3 decision trees with different depths. I split the dataset into training and sub test datasets. 80% fo training and 20% of test datasets.

Model accuracy score with default parameters: 66.67%
Model accuracy score with entropy and max depth 6 : 70.00%
Model accuracy score with entropy and max depth 3 : 73.33%

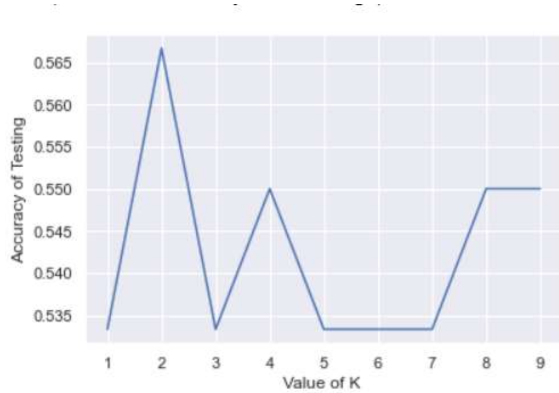
I chose these parameters because deeper tree, means more splits which will capture more Information and therefore will be more accurate.

I also drew a confusion matrix in the python notebook for parameter 3:



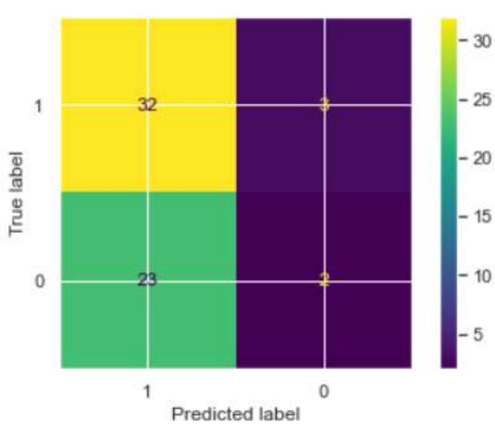
K nearest neighbor:

I first check the best number of neighbors for this algorithm and plotted a graph of value of K against true value.



accuracy of model on the test dataset is : 56.67%

The confusion matrix for this model is :



Results

This shows that decision tree is a better classification model than K nearest neighbor model, since the accuracy of the KNN is 56.67% whereas the decision tree model gives an accuracy of 73.33%.

Conclusion

In conclusion , the main features that predict the target variable is serum creatinine and ejection fraction. Although creatinine phosphokinase, age, sex, serum sodium, smoking factor, high blood pressure are also important to predict the heart failure.

Reference

www.Stackoverflow.com

https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation

www.geeksforgeeks.org

RMIT canvas resources.

https://seaborn.pydata.org/