Step 1 Import Libraries

```
import pandas as pd
import numpy as np
print(pd.__version__)
```

```
1.3.5
```

Step 2 Download And Load Dataset into Dataframe

```
path = "/content/drive/MyDrive/Colab Notebooks/DBBD/hepatitis_csv.csv"
df = pd.read_csv(path)
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm
0	30	male	False	False	False	False	False	False	False
1	50	female	False	False	True	False	False	False	False
2	78	female	True	False	True	False	False	True	False
3	31	female	NaN	True	False	False	False	True	False
4	34	female	True	False	False	False	False	True	False

Step 3 Data Preprocessing

```
# Null Check
df.isna()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	True	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	..
150	False	False	False	False	False	False	False	False	False

```
df.isnull().sum()
```

```
age                0
sex                0
steroid            1
antivirals         0
fatigue            1
malaise            1
anorexia           1
liver_big         10
liver_firm        11
spleen_palpable    5
spiders            5
ascites            5
varices            5
bilirubin          6
alk_phosphate     29
sgot               4
albumin           16
protime           67
histology          0
class              0
dtype: int64
```

```
p = df.isnull().sum()*100/len(df)
print(p)
```

```
age                0.000000
sex                0.000000
steroid            0.645161
antivirals         0.000000
fatigue            0.645161
malaise            0.645161
anorexia           0.645161
liver_big          6.451613
liver_firm         7.096774
spleen_palpable    3.225806
spiders            3.225806
ascites            3.225806
```

```
varices             3.225806
bilirubin           3.870968
alk_phosphate      18.709677
sgot                2.580645
albumin            10.322581
protime            43.225806
histology           0.000000
class               0.000000
dtype: float64
```

```
# Remove the Nan Values
df = df.dropna()
```

```
df.isna()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm
5	False	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False	False	False
13	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	..
139	False	False	False	False	False	False	False	False	False
143	False	False	False	False	False	False	False	False	False
145	False	False	False	False	False	False	False	False	False
153	False	False	False	False	False	False	False	False	False
154	False	False	False	False	False	False	False	False	False

80 rows × 20 columns

```
df.isnull().sum()
```

```
age               0
sex               0
steroid           0
antivirals        0
fatigue           0
malaise           0
anorexia          0
liver_big         0
liver_firm        0
```

```
spleen_palpable    0
spiders            0
ascites            0
varices            0
bilirubin          0
alk_phosphate      0
sgot               0
albumin            0
protime            0
histology          0
class              0
dtype: int64
```

```
df.describe()
```

	age	bilirubin	alk_phosphate	sgot	albumin	protime
count	80.00000	80.000000	80.000000	80.000000	80.000000	80.000000
mean	40.66250	1.221250	102.912500	82.025000	3.843750	62.512500
std	11.28003	0.875213	53.684779	71.599974	0.576292	23.427774
min	20.00000	0.300000	26.000000	14.000000	2.100000	0.000000
25%	32.00000	0.700000	68.250000	30.750000	3.500000	46.000000
50%	38.50000	1.000000	85.000000	56.500000	4.000000	62.000000
75%	49.25000	1.300000	133.500000	102.750000	4.200000	77.250000
max	72.00000	4.800000	280.000000	420.000000	5.000000	100.000000

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 80 entries, 5 to 154
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              80 non-null     int64
 1   sex              80 non-null     object
 2   steroid          80 non-null     object
 3   antivirals       80 non-null     bool
 4   fatigue          80 non-null     object
 5   malaise          80 non-null     object
 6   anorexia         80 non-null     object
 7   liver_big        80 non-null     object
 8   liver_firm       80 non-null     object
 9   spleen_palpable  80 non-null     object
 10  spiders          80 non-null     object
 11  ascites          80 non-null     object
 12  varices          80 non-null     object
 13  bilirubin        80 non-null     float64
```

```
 14  alk_phosphate    80 non-null     float64
 15  sgot             80 non-null     float64
 16  albumin          80 non-null     float64
 17  protime          80 non-null     float64
 18  histology        80 non-null     bool
 19  class            80 non-null     object
dtypes: bool(2), float64(5), int64(1), object(12)
memory usage: 12.0+ KB
```

```
df.dtypes
```

```
age                 int64
sex                object
steroid            object
antivirals           bool
fatigue            object
malaise            object
anorexia           object
liver_big          object
liver_firm         object
spleen_palpable    object
spiders            object
ascites            object
varices            object
bilirubin         float64
alk_phosphate     float64
sgot              float64
albumin           float64
protime           float64
histology            bool
class              object
dtype: object
```

```
df["sex"].value_counts()
```

```
female    69
male      11
Name: sex, dtype: int64
```

```
clean_up = {"sex":{'female':0,"male":1}}
```

```
df = df.replace(clean_up)
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	sp
5	34	0	True	False	False	False	False	True	False	
10	39	0	False	True	False	False	False	False	True	

```
df["steroid"] = df["steroid"].astype(int)
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	sp
5	34	0	1	False	False	False	False	True	False	
10	39	0	0	True	False	False	False	False	True	
11	32	0	1	True	True	False	False	True	True	
12	41	0	1	True	True	False	False	True	True	
13	30	0	1	False	True	False	False	True	True	

```
df["antivirals"] = df["antivirals"].astype(int)

df["malaise"] = df["malaise"].astype(int)
df["fatigue"] = df["fatigue"].astype(int)
df["anorexia"] = df["anorexia"].astype(int)
df["liver_big"] = df["liver_big"].astype(int)

df["liver_firm"] = df["liver_firm"].astype(int)

df["spleen_palpable"] = df["spleen_palpable"].astype(int)


df["spiders"] = df["spiders"].astype(int)

df["ascites"] = df["ascites"].astype(int)

df["varices"] = df["varices"].astype(int)

df["histology"] = df["histology"].astype(int)


df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	sp
5	34	0	1	0	0	0	0	1	0	
10	39	0	0	1	0	0	0	0	1	

```
df["class"].value_counts()
```

```
live    67
die     13
Name: class, dtype: int64
```

```
clean_up = {"class":{'live':0,"die":1}}
df= df.replace(clean_up)
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	sp
5	34	0	1	0	0	0	0	1	0	
10	39	0	0	1	0	0	0	0	1	
11	32	0	1	1	1	0	0	1	1	
12	41	0	1	1	1	0	0	1	1	
13	30	0	1	0	1	0	0	1	1	

Step 4 Normalizing Data

```
df_normalized = df.copy()
for i in df_normalized.columns:
  df_normalized[i] = df_normalized[i]/df_normalized[i].abs().max()
```

```
df_normalized.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_fir
5	0.472222	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.
10	0.541667	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.
11	0.444444	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.
12	0.569444	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.
13	0.416667	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.

```
df_normalized = df.copy()
for i in df_normalized.columns:
  df_normalized[i] = (df_normalized[i] - df_normalized[i].abs().min())/(df_normalized[i].abs(
```

```
df_normalized.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_fir
5	0.269231	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.
10	0.365385	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.
11	0.230769	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.
12	0.403846	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.
13	0.192308	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.

```
df_1 = df.iloc[:,0]
df_1.head()
```

```
5     34
10    39
11    32
12    41
13    30
Name: age, dtype: int64
```

```
df_1 = df_1 / df_1.abs().max()
```

```
df_1.head()
```

```
5     0.472222
10    0.541667
11    0.444444
```

```
12    0.569444
13    0.416667
Name: age, dtype: float64
```

```
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	sp
5	34	0	1	0	0	0	0	1	0	
10	39	0	0	1	0	0	0	0	1	
11	32	0	1	1	1	0	0	1	1	
12	41	0	1	1	1	0	0	1	1	
13	30	0	1	0	1	0	0	1	1	

```
df.head()
```

	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm	
0	30	male	False	False	False	False	False	False	False	
1	50	female	False	False	True	False	False	False	False	
2	78	female	True	False	True	False	False	True	False	
3	31	female	NaN	True	False	False	False	True	False	
4	34	female	True	False	False	False	False	True	False	

```
f = []
m = []
df_new = df[["sex"]].copy()
for i in range(len(df["sex"])):
  if df["sex"][i] == "male":
    m.append(1)
    f.append(0)
  else:
    m.append(0)
    f.append(1)
df_new.head()
```

	sex
0	male
1	female
2	female
3	female
4	female

```
df_new["female"] = f
df_new["male"] = m
```

```
df_new.head()
```

	sex	female	male
0	male	0	1
1	female	1	0
2	female	1	0
3	female	1	0
4	female	1	0