In [2]:
```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
```

In [3]:
```
os.chdir('C:\\Excel Class')
```

In [4]:
```
df=pd.read_csv('horse.csv')
df
```

Out[4]:

	surgery	age	hospital_number	rectal_temp	pulse	respiratory_rate	temp_of_extremities	pe
0	no	adult	530101	38.5	66.0	28.0	cool	
1	yes	adult	534817	39.2	88.0	20.0	NaN	
2	no	adult	530334	38.3	40.0	24.0	normal	
3	yes	young	5290409	39.1	164.0	84.0	cold	
4	no	adult	530255	37.3	104.0	35.0	NaN	
...	...	...	...	...	...	...	...	
294	yes	adult	533886	NaN	120.0	70.0	cold	
295	no	adult	527702	37.2	72.0	24.0	cool	
296	yes	adult	529386	37.5	72.0	30.0	cold	
297	yes	adult	530612	36.5	100.0	24.0	cool	
298	yes	adult	534618	37.2	40.0	20.0	NaN	

299 rows × 28 columns

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   surgery                299 non-null    object
 1   age                    299 non-null    object
 2   hospital_number        299 non-null    int64
 3   rectal_temp            239 non-null    float64
 4   pulse                  275 non-null    float64
 5   respiratory_rate       241 non-null    float64
 6   temp_of_extremities    243 non-null    object
 7   peripheral_pulse       230 non-null    object
 8   mucous_membrane        252 non-null    object
 9   capillary_refill_time  267 non-null    object
 10  pain                   244 non-null    object
 11  peristalsis            255 non-null    object
 12  abdominal_distention   243 non-null    object
 13  nasogastric_tube       195 non-null    object
 14  nasogastric_reflux     193 non-null    object
 15  nasogastric_reflux_ph  53 non-null     float64
 16  rectal_exam_feces      197 non-null    object
 17  abdomen                181 non-null    object
```

```
 18  packed_cell_volume     270 non-null    float64
 19  total_protein          266 non-null    float64
 20  abdomo_appearance      134 non-null    object
 21  abdomo_protein         101 non-null    float64
 22  outcome                299 non-null    object
 23  surgical_lesion        299 non-null    object
 24  lesion_1               299 non-null    int64
 25  lesion_2               299 non-null    int64
 26  lesion_3               299 non-null    int64
 27  cp_data                299 non-null    object
dtypes: float64(7), int64(4), object(17)
memory usage: 65.5+ KB
```

In [6]:
```
print(df.describe())
```

```
       hospital_number  rectal_temp       pulse  respiratory_rate  \
count     2.990000e+02   239.000000  275.000000        241.000000
mean      1.087733e+06    38.168619   72.000000         30.460581
std       1.532032e+06     0.733744   28.646219         17.666102
min       5.184760e+05    35.400000   30.000000          8.000000
25%       5.289040e+05    37.800000   48.000000         18.000000
50%       5.303010e+05    38.200000   64.000000         25.000000
75%       5.347360e+05    38.500000   88.000000         36.000000
max       5.305629e+06    40.800000  184.000000         96.000000

       nasogastric_reflux_ph  packed_cell_volume  total_protein  \
count              53.000000          270.000000     266.000000
mean                4.707547           46.307407      24.274436
std                 1.982311           10.436743      27.364194
min                 1.000000           23.000000       3.300000
25%                 3.000000           38.000000       6.500000
50%                 5.000000           45.000000       7.500000
75%                 6.500000           52.000000      56.750000
max                 7.500000           75.000000      89.000000

       abdomo_protein      lesion_1     lesion_2     lesion_3
count      101.000000    299.000000   299.000000   299.000000
mean         3.039604   3659.709030    90.528428     7.387960
std          1.967947   5408.472421   650.637139   127.749768
min          0.100000      0.000000     0.000000     0.000000
25%          2.000000   2111.500000     0.000000     0.000000
50%          2.300000   2322.000000     0.000000     0.000000
75%          3.900000   3209.000000     0.000000     0.000000
max         10.100000  41110.000000  7111.000000  2209.000000
```

In [7]:
```
df['outcome'].unique()
```

Out[7]:
```
array(['died', 'euthanized', 'lived'], dtype=object)
```

In [8]:
```
target=df['outcome']
target
```

Out[8]:
```
0           died
1      euthanized
2           lived
3           died
4           died
         ...
294    euthanized
295    euthanized
296         died
297         lived
298    euthanized
Name: outcome, Length: 299, dtype: object
```

In [9]:
```
df=df.drop(['outcome'],axis=1)
df
```

Out[9]:

	surgery	age	hospital_number	rectal_temp	pulse	respiratory_rate	temp_of_extremities	pe
0	no	adult	530101	38.5	66.0	28.0	cool	
1	yes	adult	534817	39.2	88.0	20.0	NaN	
2	no	adult	530334	38.3	40.0	24.0	normal	
3	yes	young	5290409	39.1	164.0	84.0	cold	
4	no	adult	530255	37.3	104.0	35.0	NaN	
...	...	...	...	...	...	...	...	
294	yes	adult	533886	NaN	120.0	70.0	cold	
295	no	adult	527702	37.2	72.0	24.0	cool	
296	yes	adult	529386	37.5	72.0	30.0	cold	
297	yes	adult	530612	36.5	100.0	24.0	cool	
298	yes	adult	534618	37.2	40.0	20.0	NaN	

299 rows × 27 columns

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
surgery                  0
age                      0
hospital_number          0
rectal_temp             60
pulse                   24
respiratory_rate        58
temp_of_extremities     56
peripheral_pulse        69
mucous_membrane         47
capillary_refill_time   32
pain                    55
peristalsis             44
abdominal_distention    56
nasogastric_tube       104
nasogastric_reflux     106
nasogastric_reflux_ph  246
rectal_exam_feces      102
abdomen                118
packed_cell_volume      29
total_protein           33
abdomo_appearance      165
abdomo_protein         198
surgical_lesion          0
lesion_1                 0
lesion_2                 0
lesion_3                 0
cp_data                  0
dtype: int64
```

In [22]:
```
print(df.dtypes)
```

```
surgery                object
age                    object
hospital_number         int64
```

```
rectal_temp              float64
pulse                    float64
respiratory_rate         float64
temp_of_extremities       object
peripheral_pulse          object
mucous_membrane           object
capillary_refill_time     object
pain                      object
peristalsis               object
abdominal_distention      object
nasogastric_tube          object
nasogastric_reflux        object
nasogastric_reflux_ph    float64
rectal_exam_feces         object
abdomen                   object
packed_cell_volume       float64
total_protein            float64
abdomo_appearance         object
abdomo_protein           float64
surgical_lesion           object
lesion_1                   int64
lesion_2                   int64
lesion_3                   int64
cp_data                   object
dtype: object
```

In [25]:

```
cateogry=pd.concat([df.isnull().sum(),df.dtypes],axis=1)
cateogry.columns=['Null_type','DataType']
cateogry
```

Out[25]:

	Null_type	DataType
surgery	0	object
age	0	object
hospital_number	0	int64
rectal_temp	60	float64
pulse	24	float64
respiratory_rate	58	float64
temp_of_extremities	56	object
peripheral_pulse	69	object
mucous_membrane	47	object
capillary_refill_time	32	object
pain	55	object
peristalsis	44	object
abdominal_distention	56	object
nasogastric_tube	104	object
nasogastric_reflux	106	object
nasogastric_reflux_ph	246	float64
rectal_exam_feces	102	object
abdomen	118	object
packed_cell_volume	29	float64

	Null_type	DataType
total_protein	33	float64
abdomo_appearance	165	object
abdomo_protein	198	float64
surgical_lesion	0	object
lesion_1	0	int64
lesion_2	0	int64
lesion_3	0	int64
cp_data	0	object

In [31]:
```
cat=cateogry[(cateogry.Null_type>0)&(cateogry.DataType=='object')]
cat
```

Out[31]:

	Null_type	DataType
temp_of_extremities	56	object
peripheral_pulse	69	object
mucous_membrane	47	object
capillary_refill_time	32	object
pain	55	object
peristalsis	44	object
abdominal_distention	56	object
nasogastric_tube	104	object
nasogastric_reflux	106	object
rectal_exam_feces	102	object
abdomen	118	object
abdomo_appearance	165	object

In [32]:
```
cat_variables=list(cat.index)
cat_variables
```

Out[32]:
```
['temp_of_extremities',
 'peripheral_pulse',
 'mucous_membrane',
 'capillary_refill_time',
 'pain',
 'peristalsis',
 'abdominal_distention',
 'nasogastric_tube',
 'nasogastric_reflux',
 'rectal_exam_feces',
 'abdomen',
 'abdomo_appearance']
```

In [33]:
```
from sklearn.impute import SimpleImputer
```

In [38]:
```
sm=SimpleImputer(strategy='most_frequent')
df[cat_variables]=sm.fit_transform(df[cat_variables])
df
```

Out[38]:

	surgery	age	hospital_number	rectal_temp	pulse	respiratory_rate	temp_of_extremities	pe
0	no	adult	530101	38.5	66.0	28.0	cool	
1	yes	adult	534817	39.2	88.0	20.0	cool	
2	no	adult	530334	38.3	40.0	24.0	normal	
3	yes	young	5290409	39.1	164.0	84.0	cold	
4	no	adult	530255	37.3	104.0	35.0	cool	
...	...	...	...	...	...	...	...	
294	yes	adult	533886	NaN	120.0	70.0	cold	
295	no	adult	527702	37.2	72.0	24.0	cool	
296	yes	adult	529386	37.5	72.0	30.0	cold	
297	yes	adult	530612	36.5	100.0	24.0	cool	
298	yes	adult	534618	37.2	40.0	20.0	cool	

299 rows × 27 columns

In [40]:
```
df.isnull().sum()
```

Out[40]:
```
surgery                  0
age                      0
hospital_number          0
rectal_temp             60
pulse                   24
respiratory_rate        58
temp_of_extremities      0
peripheral_pulse         0
mucous_membrane          0
capillary_refill_time    0
pain                     0
peristalsis              0
abdominal_distention     0
nasogastric_tube         0
nasogastric_reflux       0
nasogastric_reflux_ph  246
rectal_exam_feces        0
abdomen                  0
packed_cell_volume      29
total_protein           33
abdomo_appearance        0
abdomo_protein         198
surgical_lesion          0
lesion_1                 0
lesion_2                 0
lesion_3                 0
cp_data                  0
dtype: int64
```

In [44]:
```
float_col=cateogry[(cateogry.Null_type>0)]
float_col
```

Out[44]:

	Null_type	DataType
rectal_temp	60	float64
pulse	24	float64
respiratory_rate	58	float64
nasogastric_reflux_ph	246	float64
packed_cell_volume	29	float64
total_protein	33	float64
abdomo_protein	198	float64

In [45]:

```
float_columns=list(float_col.index)
float_columns
```

Out[45]:

```
['rectal_temp',
 'pulse',
 'respiratory_rate',
 'nasogastric_reflux_ph',
 'packed_cell_volume',
 'total_protein',
 'abdomo_protein']
```

In [47]:

```
sms=SimpleImputer(strategy='mean')
df[float_columns]=sms.fit_transform(df[float_columns])
df
```

Out[47]:

	surgery	age	hospital_number	rectal_temp	pulse	respiratory_rate	temp_of_extremities	pe
0	no	adult	530101	38.500000	66.0	28.0	cool	
1	yes	adult	534817	39.200000	88.0	20.0	cool	
2	no	adult	530334	38.300000	40.0	24.0	normal	
3	yes	young	5290409	39.100000	164.0	84.0	cold	
4	no	adult	530255	37.300000	104.0	35.0	cool	
...	...	...	...	...	...	...	...	
294	yes	adult	533886	38.168619	120.0	70.0	cold	
295	no	adult	527702	37.200000	72.0	24.0	cool	
296	yes	adult	529386	37.500000	72.0	30.0	cold	
297	yes	adult	530612	36.500000	100.0	24.0	cool	
298	yes	adult	534618	37.200000	40.0	20.0	cool	

299 rows × 27 columns

In [48]:

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

Out[48]:

```
surgery                0
age                    0
hospital_number        0
rectal_temp            0
```

```
pulse                   0
respiratory_rate        0
temp_of_extremities     0
peripheral_pulse        0
mucous_membrane         0
capillary_refill_time   0
pain                    0
peristalsis             0
abdominal_distention    0
nasogastric_tube        0
nasogastric_reflux      0
nasogastric_reflux_ph   0
rectal_exam_feces       0
abdomen                 0
packed_cell_volume      0
total_protein           0
abdomo_appearance       0
abdomo_protein          0
surgical_lesion         0
lesion_1                0
lesion_2                0
lesion_3                0
cp_data                 0
dtype: int64
```

In [49]:
```
df[float_columns]
```

Out[49]:

	rectal_temp	pulse	respiratory_rate	nasogastric_reflux_ph	packed_cell_volume	total_protein	a
0	38.500000	66.0	28.0	4.707547	45.0	8.400000	
1	39.200000	88.0	20.0	4.707547	50.0	85.000000	
2	38.300000	40.0	24.0	4.707547	33.0	6.700000	
3	39.100000	164.0	84.0	5.000000	48.0	7.200000	
4	37.300000	104.0	35.0	4.707547	74.0	7.400000	
...	...	...	...	...	...	...	
294	38.168619	120.0	70.0	4.707547	55.0	65.000000	
295	37.200000	72.0	24.0	4.707547	44.0	24.274436	
296	37.500000	72.0	30.0	4.707547	60.0	6.800000	
297	36.500000	100.0	24.0	4.707547	50.0	6.000000	
298	37.200000	40.0	20.0	4.707547	36.0	62.000000	

299 rows × 7 columns

In [53]:
```
all_category=cateogry[cateogry.DataType=='object']
all_category=list(all_category.index)
all_category
```

Out[53]:
```
['surgery',
 'age',
 'temp_of_extremities',
 'peripheral_pulse',
 'mucous_membrane',
 'capillary_refill_time',
 'pain',
 'peristalsis',
```

```
'abdominal_distention',
'nasogastric_tube',
'nasogastric_reflux',
'rectal_exam_feces',
'abdomen',
'abdomo_appearance',
'surgical_lesion',
'cp_data']
```

In [55]:
```
df1=df
for i in all_category:
    df1=pd.get_dummies(df1,columns=[i],drop_first=True)
df1
```

Out[55]:

	hospital_number	rectal_temp	pulse	respiratory_rate	nasogastric_reflux_ph	packed_cell_volume
0	530101	38.500000	66.0	28.0	4.707547	45.0
1	534817	39.200000	88.0	20.0	4.707547	50.0
2	530334	38.300000	40.0	24.0	4.707547	33.0
3	5290409	39.100000	164.0	84.0	5.000000	48.0
4	530255	37.300000	104.0	35.0	4.707547	74.0
...	...	...	...	...	...	..
294	533886	38.168619	120.0	70.0	4.707547	55.0
295	527702	37.200000	72.0	24.0	4.707547	44.0
296	529386	37.500000	72.0	30.0	4.707547	60.0
297	530612	36.500000	100.0	24.0	4.707547	50.0
298	534618	37.200000	40.0	20.0	4.707547	36.0

299 rows × 51 columns

In [57]:
```
target.unique()
```

Out[57]:
```
array(['died', 'euthanized', 'lived'], dtype=object)
```

In [60]:
```
from sklearn.preprocessing import LabelEncoder
lb=LabelEncoder()
target=lb.fit_transform(target)
target
```

Out[60]:
```
array([0, 1, 2, 0, 0, 2, 2, 0, 1, 2, 2, 2, 2, 0, 2, 0, 0, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 0, 2, 0, 1, 2, 2, 2, 1, 1, 2, 2, 0, 0, 2, 2, 1,
       1, 0, 2, 2, 0, 0, 2, 0, 2, 2, 1, 0, 2, 0, 0, 0, 2, 2, 0, 1, 2, 2,
       2, 2, 2, 2, 1, 2, 0, 0, 0, 1, 2, 2, 0, 2, 0, 2, 2, 0, 2, 2, 0, 1,
       2, 2, 2, 0, 2, 0, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 0, 2, 2, 2, 1, 0,
       0, 2, 2, 0, 2, 2, 2, 1, 2, 2, 2, 1, 2, 0, 1, 2, 2, 2, 1, 2, 0, 0,
       2, 0, 0, 0, 2, 2, 2, 2, 0, 0, 2, 1, 2, 2, 0, 2, 2, 2, 0, 2, 2, 2,
       2, 2, 1, 2, 2, 2, 0, 2, 1, 2, 1, 2, 2, 1, 2, 0, 2, 0, 0, 2, 2,
       0, 2, 1, 0, 2, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 2, 2, 1, 0, 2,
       2, 2, 0, 1, 2, 2, 1, 2, 0, 1, 2, 0, 0, 2, 2, 2, 2, 2, 2, 1, 2, 1,
       2, 2, 2, 2, 2, 0, 2, 1, 2, 2, 0, 2, 2, 2, 0, 2, 2, 1, 1, 0, 2, 0,
       2, 2, 0, 2, 0, 2, 2, 1, 0, 0, 2, 2, 0, 1, 0, 2, 0, 2, 2, 2, 0, 2,
       1, 2, 2, 2, 2, 0, 2, 2, 0, 1, 0, 1, 2, 2, 2, 2, 0, 0, 2, 2, 2, 1,
       2, 2, 2, 2, 0, 0, 1, 2, 1, 1, 0, 2, 1], dtype=int64)
```

In [62]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(df1,target,test_size=0.2,random_state
```

In [65]:
```
from sklearn.tree import DecisionTreeClassifier
```

In [66]:
```
dt=DecisionTreeClassifier()
dt.fit(x_train,y_train)
```

Out[66]:
```
DecisionTreeClassifier()
```

In [67]:
```
pred=dt.predict(x_test)
pred
```

Out[67]:
```
array([1, 2, 0, 1, 0, 2, 0, 0, 2, 2, 1, 2, 2, 1, 0, 1, 0, 0, 2, 2, 0, 2,
       1, 0, 1, 2, 2, 1, 2, 1, 2, 0, 2, 1, 1, 0, 2, 2, 2, 2, 0, 2, 2, 1,
       0, 2, 0, 0, 0, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2], dtype=int64)
```

In [69]:
```
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test,pred))
```

```
0.6333333333333333
```

In [70]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [71]:
```
rf=RandomForestClassifier()
rf.fit(x_train,y_train)
```

Out[71]:
```
RandomForestClassifier()
```

In [72]:
```
predy=rf.predict(x_test)
predy
```

Out[72]:
```
array([1, 0, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 0, 1, 0, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 0, 2, 2, 2, 0, 2, 0, 2, 0, 2, 2, 2, 2, 2, 2, 2, 2,
       0, 0, 0, 2, 0, 2, 2, 2, 0, 0, 2, 2, 2, 2, 0, 2], dtype=int64)
```

In [73]:
```
print(accuracy_score(predy,y_test))
```

```
0.7333333333333333
```