In [24]:
```
# Data Visualization with Haberman Dataset

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df  = pd.read_csv("G:\Haberman\haberman.csv")
df.head(5)
```

Out[24]:

	Age	Op_Year	axil_nodes_det	Surv_status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

In [31]:
```
# Number of points and number of features

df.shape
```

Out[31]: (306, 4)

There are total 306 datapoints in haberman dataset. There are total 3 features and 1 Output Variable

In [32]:
```
df['Surv_status'].value_counts()
```

Out[32]:
```
1    225
2     81
```

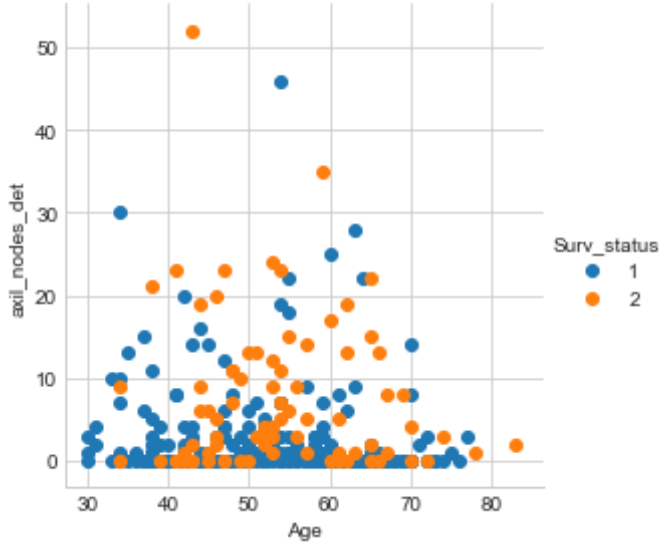
```
Name: Surv_status, dtype: int64
```

There are 2 of classes :- "1" class and "2" class. There are 225 datapoints of class "1" and 81 datapoints of class "2"

Here Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year
Objective: Classify a patient as belonging to one of the 2 classes given the 3 features.

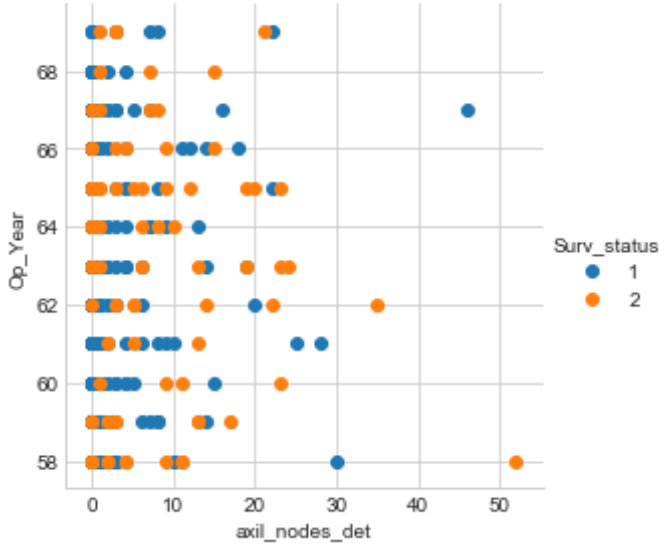
We have to check Bi-variate analysis (scatter plots, pair-plots) to see if combinations of features are useful in classfication

In [35]:
```
# 2-D scatterplot for Age and axil_nodes_det
sns.set_style("whitegrid");
g = sns.FacetGrid(df,hue = 'Surv_status', size = 4)
g = g.map(plt.scatter,'Age','axil_nodes_det').add_legend()
plt.show()
```



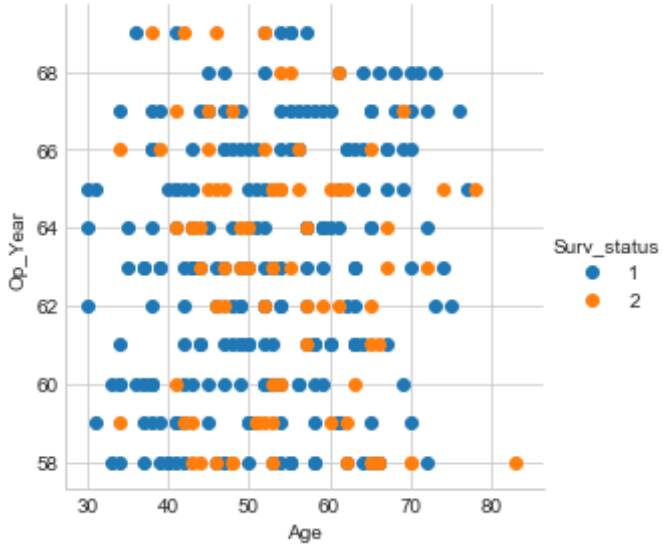
Above scatter plot is combination of two features "Age" and "axil_nodes_det", we are unable to classify cancer patient survived longer or died within 5 years by above plot.

In [37]:
```
# 2-D scatterplot for Op_Year and axil_nodes_det
sns.set_style("whitegrid");
g = sns.FacetGrid(df,hue = 'Surv_status', size = 4)
g = g.map(plt.scatter,'axil_nodes_det','Op_Year').add_legend()
plt.show()
```



Above Scatter plot is combination of two features "axil_nodes_det" and "Op_Year".we are unable to classify cancer patient survived longer or died within 5 years by above plot because of overlapping

In [82]:
```
# 2-D scatterplot for Age and Op_Year
sns.set_style("whitegrid");
g = sns.FacetGrid(df,hue = 'Surv_status', size = 4)
g = g.map(plt.scatter,'Age','Op_Year').add_legend()
plt.show()
```



Above Scatter plot is combination of two features "Age" and "Op_Year".we are unable to classify cancer patient survived longer or died within 5 years by above plot because of overlapping

In [86]:
```
# pair plot
plt.close();
sns.set_style("whitegrid");
sns.pairplot(df, hue="Surv_status", size=3);
plt.show()
```

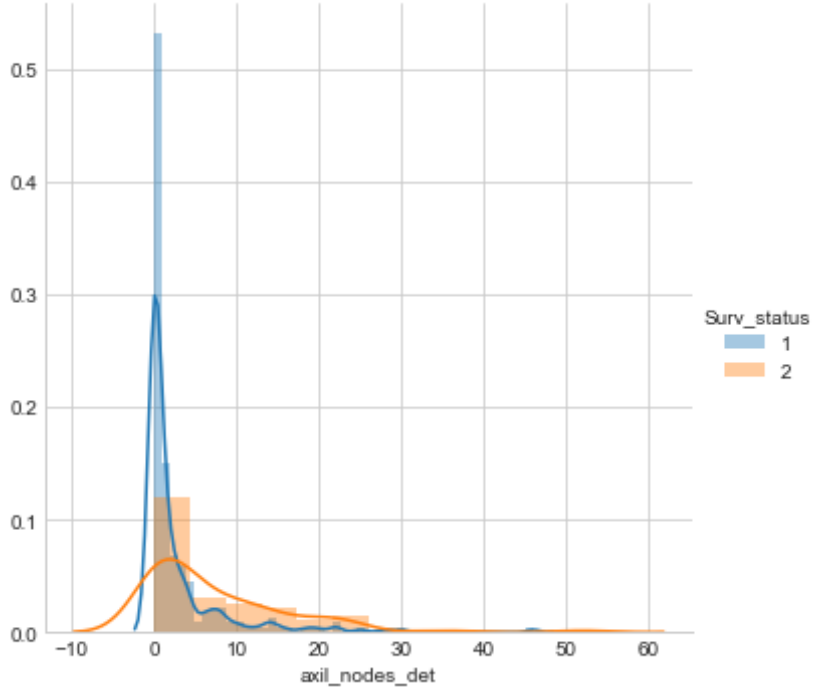
By above pair plot, we are unable to classify cancer patient survived longer or died within 5 years because of overlapping

we have to perform Univaraite analysis(PDF, CDF, Boxplot, Voilin plots) to understand which

features are useful towards classification.

In [88]:
```
# PDF for axil_nodes_det
a = sns.FacetGrid(df,hue = 'Surv_status', size = 5)
a = a.map(sns.distplot,"axil_nodes_det").add_legend()
plt.show()
```

```
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:646
2: UserWarning: The 'normed' kwarg is deprecated, and has been replaced
by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:646
2: UserWarning: The 'normed' kwarg is deprecated, and has been replaced
by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

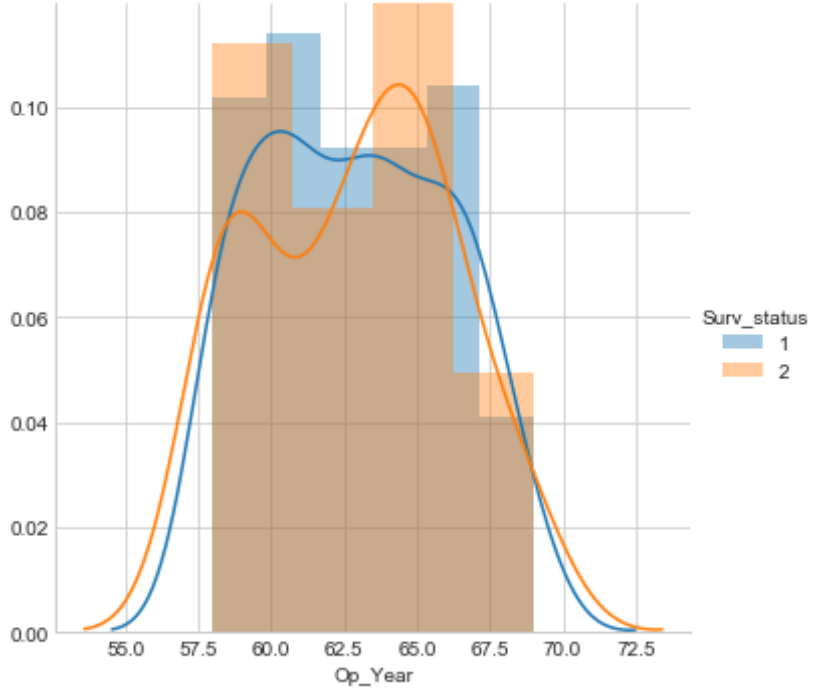


For "axis_nodes_det" features, there is overlap between class 1 and class 2. we are unable to

classify

In [40]:
```
# PDF for Op_Year
a = sns.FacetGrid(df,hue = 'Surv_status',size = 5)
a.map(sns.distplot,'Op_Year').add_legend()
plt.show()
```

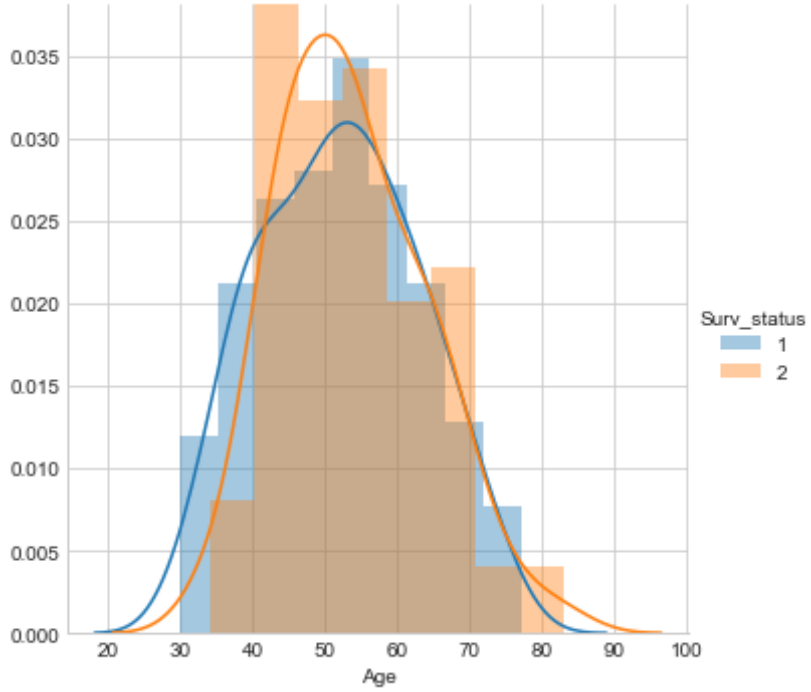
```
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



For "Op_Year" features, there is overlap between class 1 and class 2. we are unable to classify

In [89]:
```
# PDF for Age
a = sns.FacetGrid(df,hue = 'Surv_status',size = 5)
a.map(sns.distplot,'Age').add_legend()
plt.show()
```

```
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\hemant\Anaconda\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

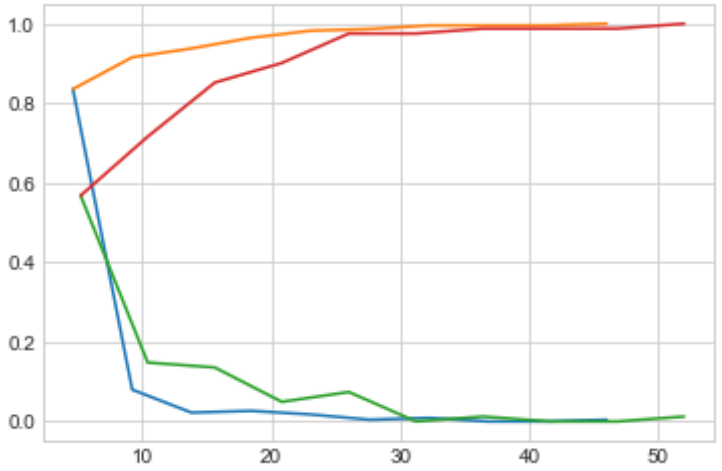


For "Age" features, there is overlap between class 1 and class 2. we are unable to classify

In [71]:
```
#CDF
status_one = df.loc[df["Surv_status"] == 1]
status_two = df.loc[df["Surv_status"] == 2]
status_one.shape
status_zero.shape
```

Out[71]:
```
(0, 4)
```

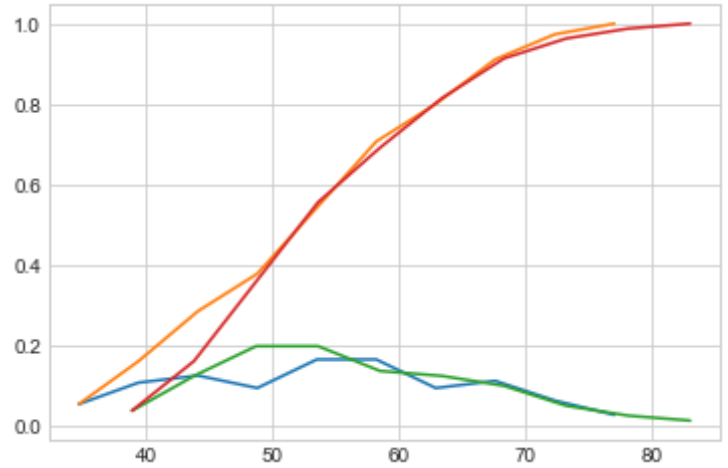
In [74]:
```
#CDF for axil_nodes_det
counts,bin_edges = np.histogram(status_one['axil_nodes_det'],bins = 10,
density =True)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)



counts,bin_edges = np.histogram(status_two['axil_nodes_det'],bins = 10,
density =True)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)



plt.show();
```



Above plot is CDF of 'axil_nodes_det' feature for both class 1 and 2. CDF of both class is overlapping

In [91]:
```
#CDF for Age

counts,bin_edges = np.histogram(status_one['Age'],bins = 10,density =Tr
ue)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)



counts,bin_edges = np.histogram(status_two['Age'],bins = 10,density =Tr
ue)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
```

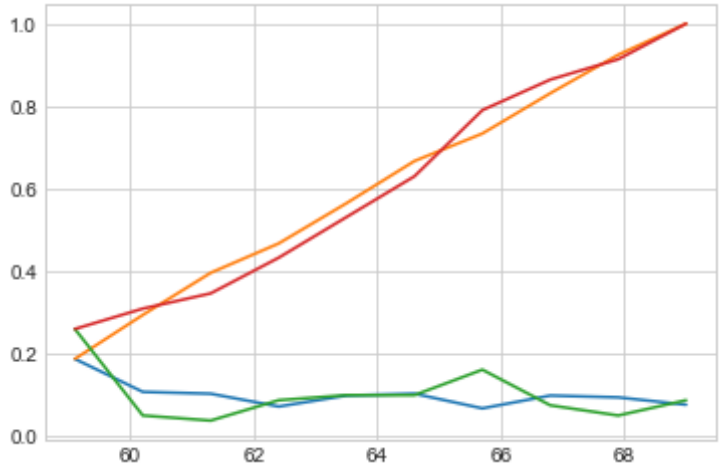
```
plt.plot(bin_edges[1:],cdf)



plt.show();
```



In [ ]:
```
Above plot is CDF of 'Age' feature for both class 1 and 2.  CDF of both
 class is overlapping
```

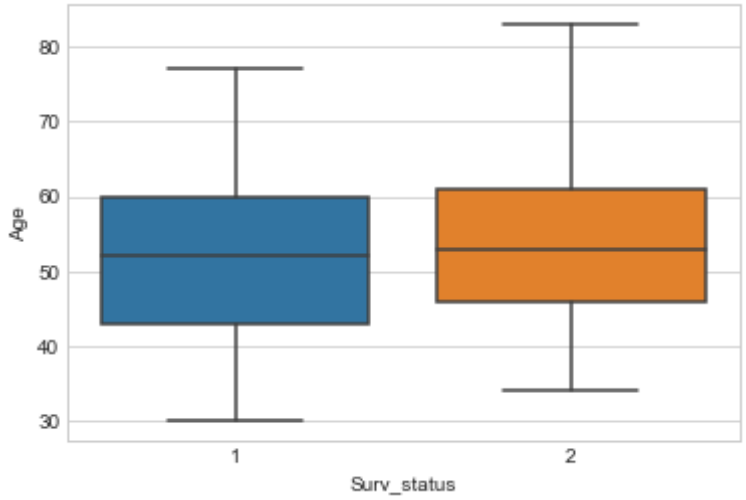
In [75]:
```
#CDF for Op_Year

counts,bin_edges = np.histogram(status_one['Op_Year'],bins = 10,density
 =True)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)



counts,bin_edges = np.histogram(status_two['Op_Year'],bins = 10,density
```

```
 =True)
#print(counts)
#print(bin_edges)
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)



plt.show();
```



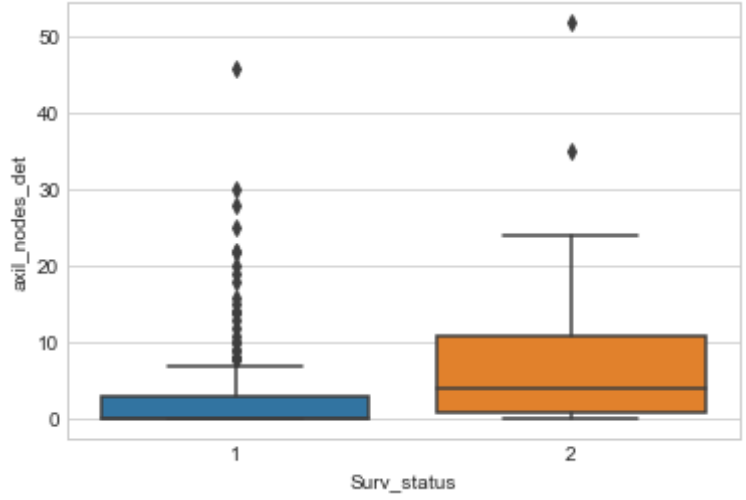
Above plot is CDF of 'Op_Year' feature for both class 1 and 2. CDF of both class is overlapping

In [76]:
```
# boxplot for Age
sns.boxplot(x = 'Surv_status',y = 'Age',data = df)
plt.show()
```



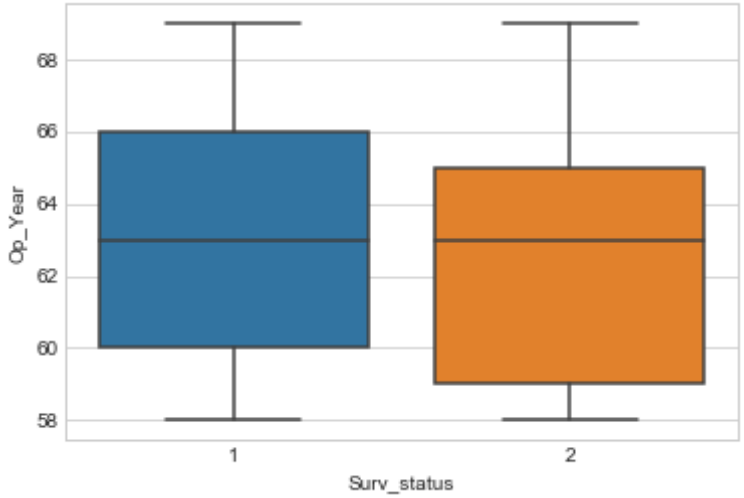
Above is boxplot for Age Feature. for both class 1 and 2,Age data belonging from range 43 to 60.most of data is ovelapping and least amount of Age data from 43 to 45 belonging to class 1

In [93]:
```
# boxplot for axil_nodes_det
sns.boxplot(x = 'Surv_status',y = 'axil_nodes_det',data = df)
plt.show()
```



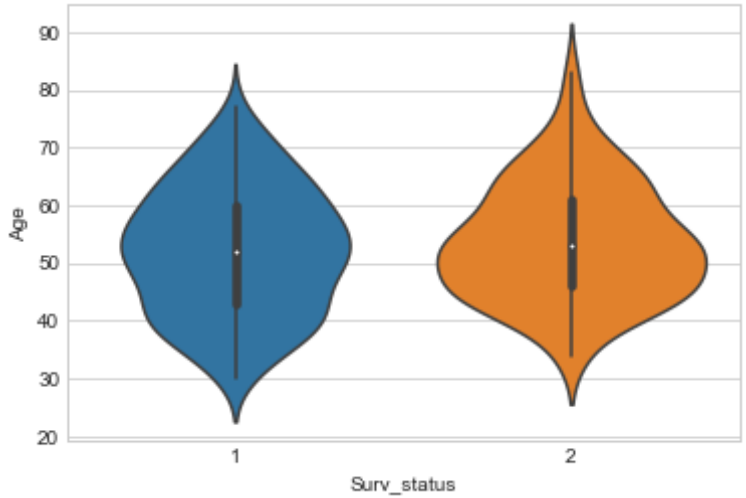
above boxplot for 'axil_nodes_det' feature. Most of data is overlapping for both class 1 and 2. for class 2,

In [78]:
```
#boxplot for Op_Year
sns.boxplot(x = 'Surv_status',y = 'Op_Year',data = df)
plt.show()
```

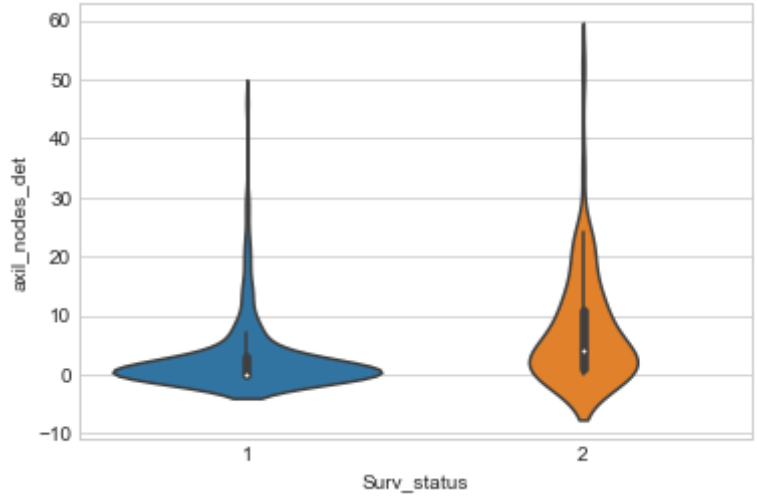


In above boxplot for "Op_Year" feature, for operation year 65 to 66 data is belonging to class 1. and for operation year 59 to 60 data is belonging to class 2, and most of data is overlapping for both class 1 and 2

In [79]:
```
# violinplot for Age
sns.violinplot(x = 'Surv_status',y = 'Age',data = df,size = 8 )
plt.show()
```



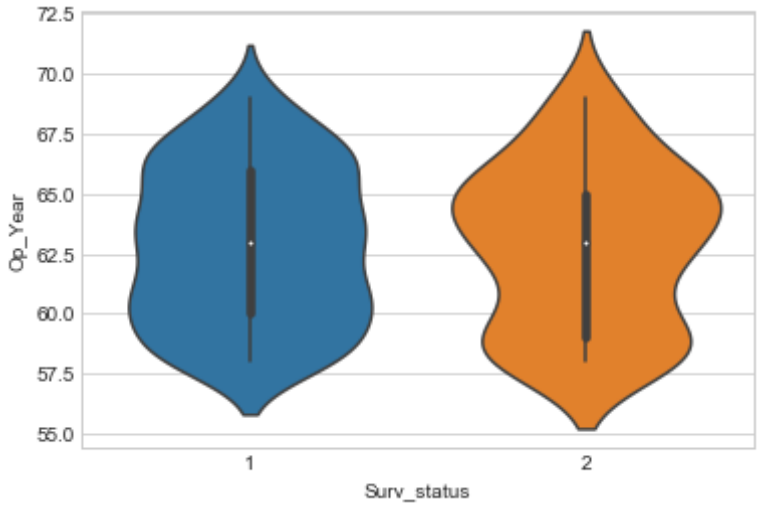
In [80]:
```
# violinplot for axil_nodes_det
sns.violinplot(x = 'Surv_status',y = 'axil_nodes_det',data = df,size = 8 )
plt.show()
```



above violin plot for 'axil_nodes_det' feature. Most of data is overlapping for both class 1 and 2.

data is belonging to 6 to 11 range for class 2 only.

In [81]:
```
# violinplot for Op_Year
sns.violinplot(x = 'Surv_status',y = 'Op_Year',data = df,size = 8 )
plt.show()
```



In above violin for "Op_Year" feature, for operation year 65 to 66 data is belonging to class 1. and for operation year 59 to 60 data is belonging to class 2, and most of data is overlapping for both class 1 and 2

By above both Univaraite analysis and Bi-variate analysis method, we are unable to classify the class 1 and class 2