

Data Source  
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

GOAL- Use machine learning to determine which physicochemical properties make a wine 'good'!  
Assumption: 1) I used for random\_state = 42

```
In [1]: import matplotlib
import pylab as plt
import numpy as pd

In [2]: from numpy import random

df= pd.read_csv("winequality-red.csv",na_values=['NA','?','']) # Read data file # na values can be represent
np.random.seed(42) # undo the random error function to not change ur data after shuffling
#if ur comment on the above function it will keep on manipulating
#pick a random number everytime
wine = df.reindex(np.random.permutation(df.index)) #shuffle data
#permutation of index
#index of pandas randomly select indices, reindex our index
wine.reset_index(inplace=True, drop=True) # Reset index

Wine[0:5] # Display top five rows
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.7	0.56	0.08	2.50	0.114	14.0	46.0	0.9971	3.24	0.66	9.6	6
1	7.8	0.50	0.17	1.60	0.082	21.0	102.0	0.9960	3.39	0.48	9.5	0
2	10.7	0.67	0.22	2.70	0.107	17.0	34.0	1.0004	3.28	0.98	9.9	0
3	8.5	0.46	0.31	2.25	0.078	32.0	58.0	0.9980	3.33	0.54	9.8	5
4	6.7	0.46	0.24	1.70	0.077	18.0	34.0	0.9948	3.39	0.60	10.6	6

```
In [3]: wine.shape

Out[3]: (1599, 12)
```

There are 1599 rows and 12 columns

COMMENT - All the data is regression so no one-hot coding is required.

```
In [4]: wine.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
# Column Non-Null Count Dtype
---
0 fixed acidity 1599 non-null float64
1 volatile acidity 1599 non-null float64
2 citric acid 1599 non-null float64
3 residual sugar 1599 non-null float64
4 chlorides 1599 non-null float64
5 free sulfur dioxide 1599 non-null float64
6 total sulfur dioxide 1599 non-null float64
7 density 1599 non-null float64
8 pH 1599 non-null float64
9 sulphates 1599 non-null float64
10 alcohol 1599 non-null float64
11 quality 1599 non-null int64
dtypes: float64(11), int64(1)
memory usage: 130.0 KB
```

CONCLUSION - The data has no missing values and no imputation is needed.

```
In [5]: wine.describe()

Out[5]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538906	0.087467	15.874922	46.467792	0.996747	3.31113
std	1.741096	0.179960	0.194001	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000
25%	7.100000	0.390000	0.090000	1.900000	0.030000	7.000000	22.000000	0.995600	3.210000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000

Look at distribution make a histogram

```
In [6]: columns=['fixed acidity','volatile acidity','citric acid','residual sugar','chlorides','free sulfur dioxide','total sulfur dioxide','density','pH','sulphates','alcohol','quality']
wine.columns.hist(bins=15, layout=(4, 3), figsize=(15,10)) # 3 rows and 3 columns # figure size is the axes
plt.show()
```

Target Variable - Quality since the goal is to determine which physicochemical properties make a wine 'good'.

Assign good quality wine where wine\_quality is greater than 6.

The remaining wine is classified as poor quality wine

METHOD 1

```
In [7]: wine["wine_quality"]!=wine["quality"]>6*1

wine[10:20]
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_quality
10	10.5	0.510	0.64	2.4	0.107	6.0	15.0	0.99730	3.09	0.66	11.80	7	1
11	6.7	0.760	0.02	1.8	0.078	6.0	12.0	0.99600	3.55	0.63	9.95	3	0
12	6.7	0.480	0.08	2.1	0.064	18.0	34.0	0.99552	3.33	0.64	9.70	5	0
13	7.0	0.560	0.13	1.6	0.077	25.0	42.0	0.99629	3.34	0.59	9.20	5	0
14	7.8	0.340	0.37	2.0	0.082	24.0	58.0	0.99649	3.34	0.59	9.40	6	0
15	8.9	0.280	0.45	1.7	0.067	7.0	12.0	0.99354	3.25	0.55	12.30	7	1
16	6.7	0.675	0.07	2.4	0.089	17.0	82.0	0.99580	3.35	0.54	10.10	5	0
17	10.1	0.430	0.40	2.6	0.092	13.0	62.0	0.99834	3.22	0.64	10.00	7	1
18	9.4	0.300	0.56	2.8	0.080	6.0	17.0	0.99640	3.15	0.92	11.70	8	1
19	8.4	0.745	0.11	1.9	0.090	16.0	63.0	0.99650	3.19	0.82	9.60	5	0

METHOD 2

```
In [9]: wine["wine_quality_cat"]=np.ceil(wine["quality"])*7 # 10 leads to most data in categories in 4 arbitrary #
#12 will work as well, just looking for an even distribution
#wine["wine_quality_cat"]>where(wine["quality"]< 7, 1, inplace=True) #anything more than 4 assigns to 4
#any number will work, 4 is generally chosen, bigger the # more the computation time
wine[10:20]
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_quality	wine_qual
10	10.5	0.510	0.64	2.4	0.107	6.0	15.0	0.99730	3.09	0.66	11.80	7	1	1
11	6.7	0.760	0.02	1.8	0.078	6.0	12.0	0.99600	3.55	0.63	9.95	3	0	0
12	6.7	0.480	0.08	2.1	0.064	18.0	34.0	0.99552	3.33	0.64	9.70	5	0	0
13	7.0	0.560	0.13	1.6	0.077	25.0	42.0	0.99629	3.34	0.59	9.20	5	0	0
14	7.8	0.340	0.37	2.0	0.082	24.0	58.0	0.99649	3.34	0.59	9.40	6	0	0
15	8.9	0.280	0.45	1.7	0.067	7.0	12.0	0.99354	3.25	0.55	12.30	7	1	1
16	6.7	0.675	0.07	2.4	0.089	17.0	82.0	0.99580	3.35	0.54	10.10	5	0	0
17	10.1	0.430	0.40	2.6	0.092	13.0	62.0	0.99834	3.22	0.64	10.00	7	1	1
18	9.4	0.300	0.56	2.8	0.080	6.0	17.0	0.99640	3.15	0.92	11.70	8	1	1
19	8.4	0.745	0.11	1.9	0.090	16.0	63.0	0.99650	3.19	0.82	9.60	5	0	0

We see that by both of the methods any wine with quality greater than 7 has been assigned as good i.e. binary as 1 in wine\_quality and wine\_quality\_cat and vice versa

```
In [10]: df=wine.copy()

In [11]: df = Pandas dataframe
df : n in the equation m=sd
name: Column name
***
mean=df[name].mean() # Calculate mean of column
std=df[name].std() # Calculate standard deviation of column
drop_x = df.index[(mean - n * sd < df[name]) | (mean + n * sd < df[name])] #vertical line is or
df.drop(drop_x, axis=0, inplace=True) # dropping rows that dont satisfy the code
df.reset_index(inplace=True, drop=True) # Reset index

# Drop outliers in last column "Oil Prod. (m3/month)"
outlier_remove(df, n=3, name="quality") #based on oil production
df.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_quality	wine_qual
count	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000	1589.000000
mean	8.319383	0.525756	0.271605	2.538200	0.087246	15.905601	46.603524	0.996742	3.310566					
std	1.741471	0.175602	0.194337	1.410398	0.046866	10.460067	32.929860	0.001886	0.154337					
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000					
25%	7.100000	0.390000	0.100000	1.900000	0.030000	7.000000	22.000000	0.995600	3.210000					
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000					
75%	9.200000	0.635000	0.420000	2.600000	0.090000	21.000000	63.000000	0.997820	3.400000					
max	15.900000	1.630000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000					

The number of rows have been removed so the number of rows have been reduced from 1599 to 1571.

```
In [12]:
```

BINARY CLASSIFICATION

Now removing quality and wine\_quality

```
In [13]: df_binary=df.copy()
df_binary.drop(["wine_quality_cat"], axis=1, inplace=True)
df_binary.drop(["quality"], axis=1, inplace=True)
df_binary.reset_index(inplace=True, drop=True) # Reset index
df_binary[0:5]
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	wine_quality
0	7.7	0.56	0.08	2.50	0.114	14.0	46.0	0.9971	3.24	0.66	9.6	0
1	7.8	0.50	0.17	1.60	0.082	21.0	102.0	0.9960	3.39	0.48	9.5	0
2	10.7	0.67	0.22	2.70	0.107	17.0	34.0	1.0004	3.28	0.98	9.9	0
3	8.5	0.46	0.31	2.25	0.078	32.0	58.0	0.9980	3.33	0.54	9.8	0
4	6.7	0.46	0.24	1.70	0.077	18.0	34.0	0.9948	3.39	0.60	10.6	0

```
In [12]: font = {'size' : 10}
matplotlib.rc('font', **font)

fig = plt.subplots(figsize=(10, 3), dpi= 100, facecolor='w', edgecolor='k')

df_binary["wine_quality"].hist(bins=15)
plt.title("Original dataset")
plt.xlabel("Classification Labels")
plt.ylabel("Frequency")

Out[12]: Text(0, 0.5, 'Frequency')
```

```
In [13]: df_binary["wine_quality"].value_counts()

Out[13]:
0    1372
1     217
Name: wine_quality, dtype: int64
```

The original dataset has approximately 1599 rows. The original dataset has approximately 1372 rows classified as poor quality wine i.e. 87.3% of the total cases.

```
In [14]: df_binary
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	wine_quality
0	7.7	0.660	0.08	2.50	0.114	14.0	46.0	0.99710	3.24	0.66	9.6	0
1	7.8	0.500	0.17	1.60	0.082	21.0	102.0	0.99600	3.39	0.48	9.5	0
2	10.7	0.670	0.22	2.70	0.107	17.0	34.0	1.00040	3.28	0.98	9.9	0
3	8.5	0.460	0.31	2.25	0.078	32.0	58.0	0.99800	3.33	0.54	9.8	0
4	6.7	0.460	0.24	1.70	0.077	18.0	34.0	0.99480	3.39	0.60	10.6	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1584	9.1	0.600	0.00	1.90	0.058	5.0	10.0	0.99770	3.18	0.63	10.4	0
1585	8.2	0.635	0.10	2.10	0.073	25.0	60.0	0.99638	3.29	0.75	10.9	0
1586	7.2	0.620	0.06	2.70	0.077	15.0	85.0	0.99746	3.51	0.54	9.5	0
1587	7.9	0.200	0.26	1.70	0.054	7.0	15.0	0.99458	3.32	0.80	11.9	1
1588	5.8	0.280	0.26	1.70	0.063	3.0	11.0	0.99150	3.29	0.54	13.5	0

1589 rows x 12 columns

SPLIT DATA INTO TEST AND TRAINING DATA BY STRATIFIED SAMPLING

```
In [15]: from sklearn.model_selection import StratifiedShuffleSplit

# Training and Test
spt = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)#always use shuffle
train_idx, test_idx = spt.split(df_binary, df_binary["wine_quality"])#split training and testing data
train_set_strat = df_binary.loc[train_idx]
test_set_strat = df_binary.loc[test_idx]

# For dataset
X_train, y_train = train_set_strat, test_set_strat
dataset.reset_index(inplace=True, drop=True) # Reset index
dataset[0:4]
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	wine_quality
0	7.6	0.685	0.23	2.3	0.111	20.0	84.0	0.99640	3.21	0.61	9.3	0
1	5.9	0.440	0.00	1.6	0.042	3.0	11.0	0.99440	3.48	0.85	11.7	0
2	7.9	0.690	0.21	2.1	0.080	33.0	141.0	0.99620	3.25	0.51	9.9	0
3	9.1	0.660	0.15	3.2	0.097	9.0	59.0	0.99976	3.28	0.54	9.6	0

```
In [16]: font = {'size' : 10}
matplotlib.rc('font', **font)

fig = plt.subplots(figsize=(10, 3), dpi= 100, facecolor='w', edgecolor='k')

ax1=plt.subplot(1,2,1)
train_set_strat["wine_quality"].hist(bins=15)
plt.title("Training Set")
plt.xlabel("Classification Labels")
plt.ylabel("Frequency")

ax2=plt.subplot(1,2,2)
test_set_strat["wine_quality"].hist(bins=15)
plt.title("Test Set")
plt.xlabel("Classification Labels")
plt.ylabel("Frequency")
plt.show()
```

STRAIFIED SAMPLING IN TRAINING DATASET

```
In [17]: train_set_strat["wine_quality"].value_counts()

Out[17]:
0    1097
1     174
Name: wine_quality, dtype: int64
```

The training dataset has approximately 1599 good quality wine i.e. 12.6% of the total cases. The training dataset has approximately 1097 as poor quality wine i.e. 87.4% of the total cases.

STRATIFIED SAMPLING IN TEST DATASET

```
In [18]: test_set_strat["wine_quality"].value_counts()

Out[18]:
0     275
1      43
Name: wine_quality, dtype: int64
```

The test dataset has approximately 40 as good quality wine i.e. 12.6% of the total cases. The test dataset has approximately 275 as poor quality wine i.e. 87.4% of the total cases.

conclusion

Both the training and testing dataset as well as the original dataset have almost the same ratio for good and poor quality wine. There is similar split in all the datasets.

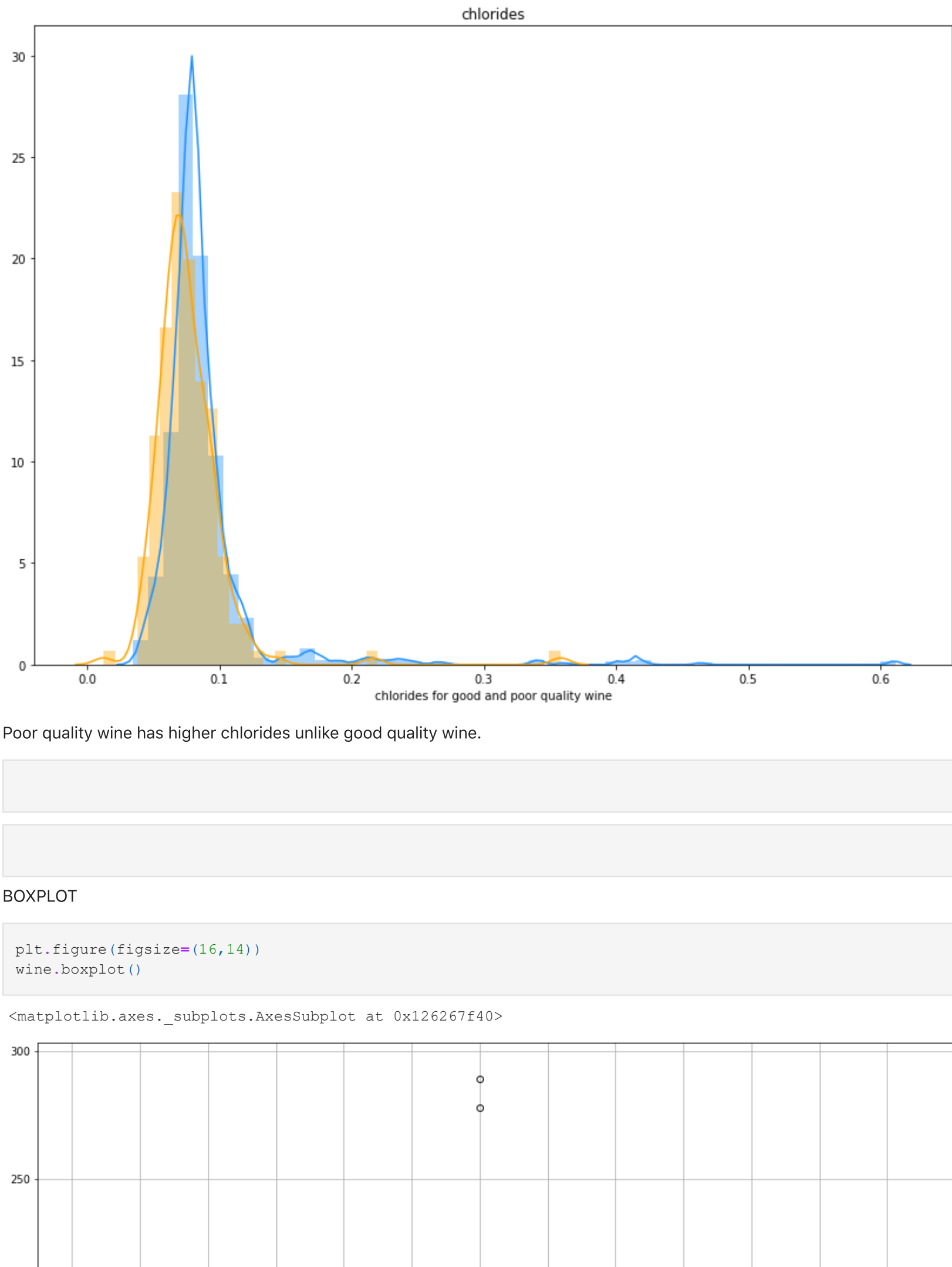
Now setting Classification column as target

```
In [19]: # Note that drop() creates a copy and does not affect train_set_strat
X_train = train_set_strat.drop("wine_quality", axis=1) # remove target
y_train = train_set_strat["wine_quality"].values # only target values

In [20]: X_train
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.6	0.685	0.23	2.3	0.111	20.0	84.0	0.99640	3.21	0.61	9.3
1	5.9	0.440	0.00	1.6	0.042	3.0	11.0	0.99440	3.48	0.85	11.7
2	7.9	0.690	0.21	2.1	0.080	33.0	141.0	0.99620	3.25	0.51	9.9
3	9.1	0.660	0.15	3.2	0.097	9.0	59.0	0.99976	3.28	0.54	9.6
4	7.9	0.720	0.17	2.6	0.096						





Poor quality wine has higher chlorides unlike good quality wine.

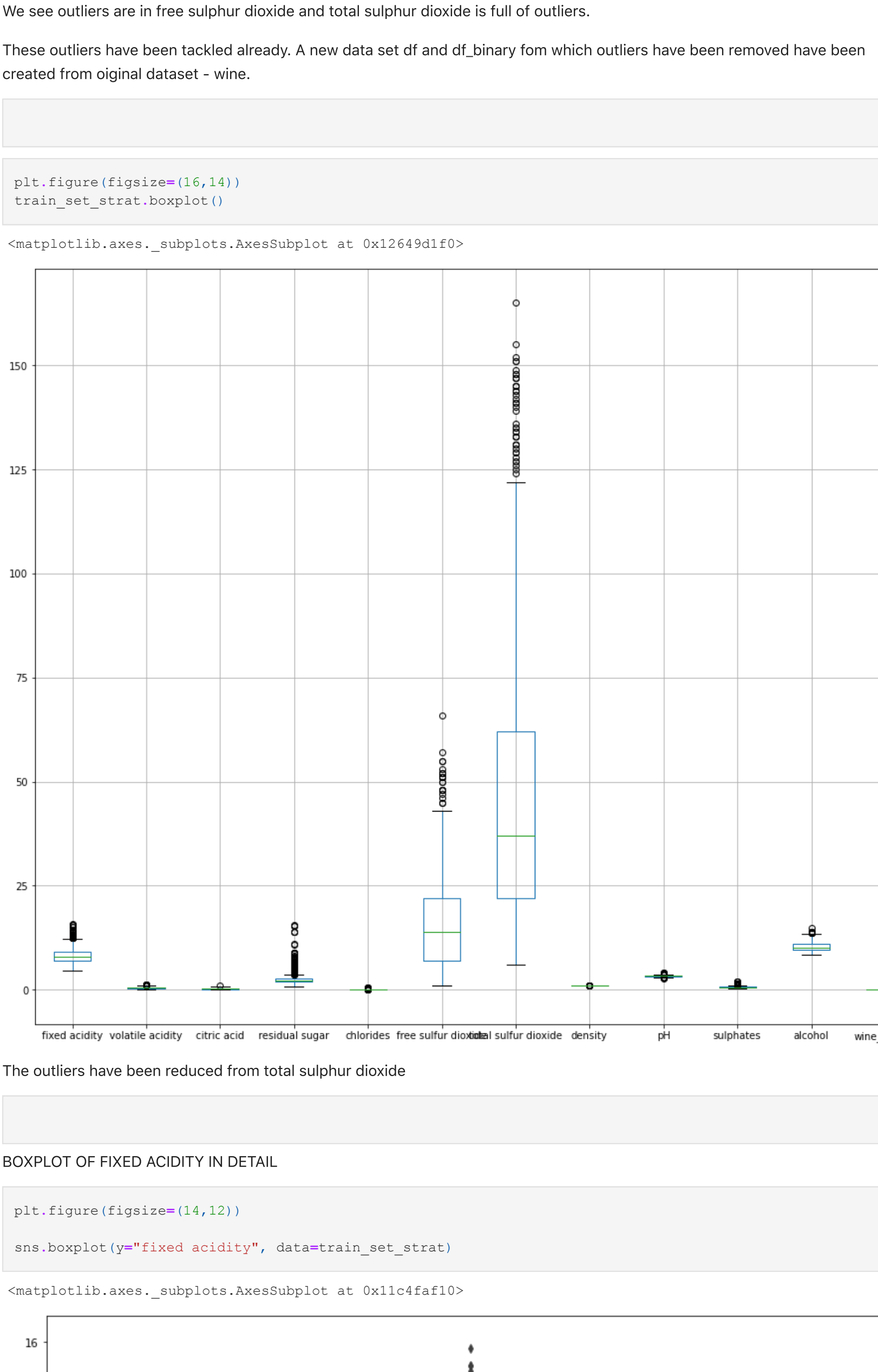
```
In [ ]:
```

```
In [ ]:
```

### BOXPLOT

```
In [33]: plt.figure(figsize=(16,14))
wine.boxplot()
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x12626740>
```



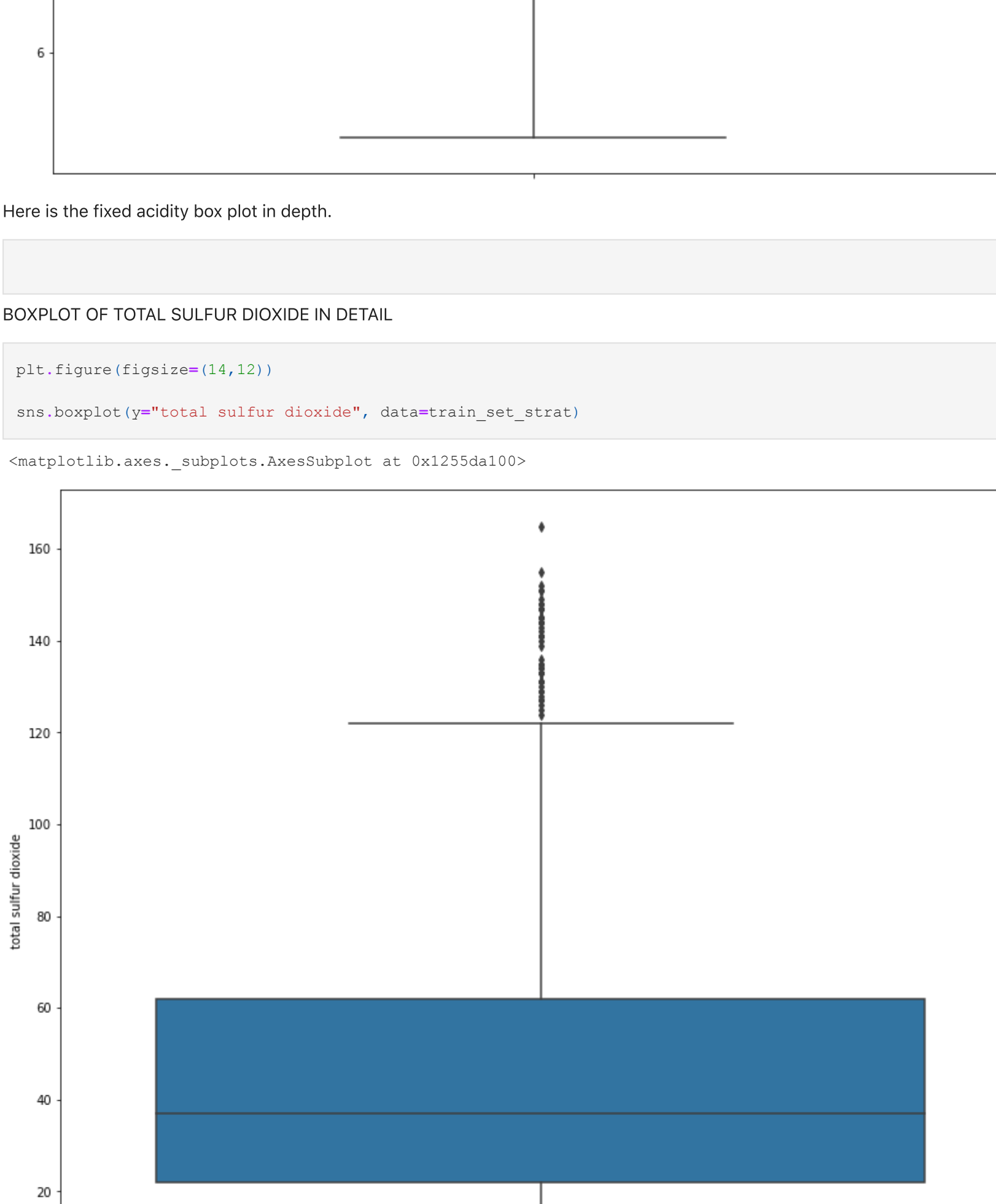
We see outliers are in free sulfur dioxide and total sulphur dioxide is full of outliers.

These outliers have been tackled already. A new data set df and df\_binary from which outliers have been removed have been created from original dataset - wine.

```
In [ ]:
```

```
In [34]: plt.figure(figsize=(16,14))
train_set_strat.boxplot()
```

```
Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x12649d1f0>
```



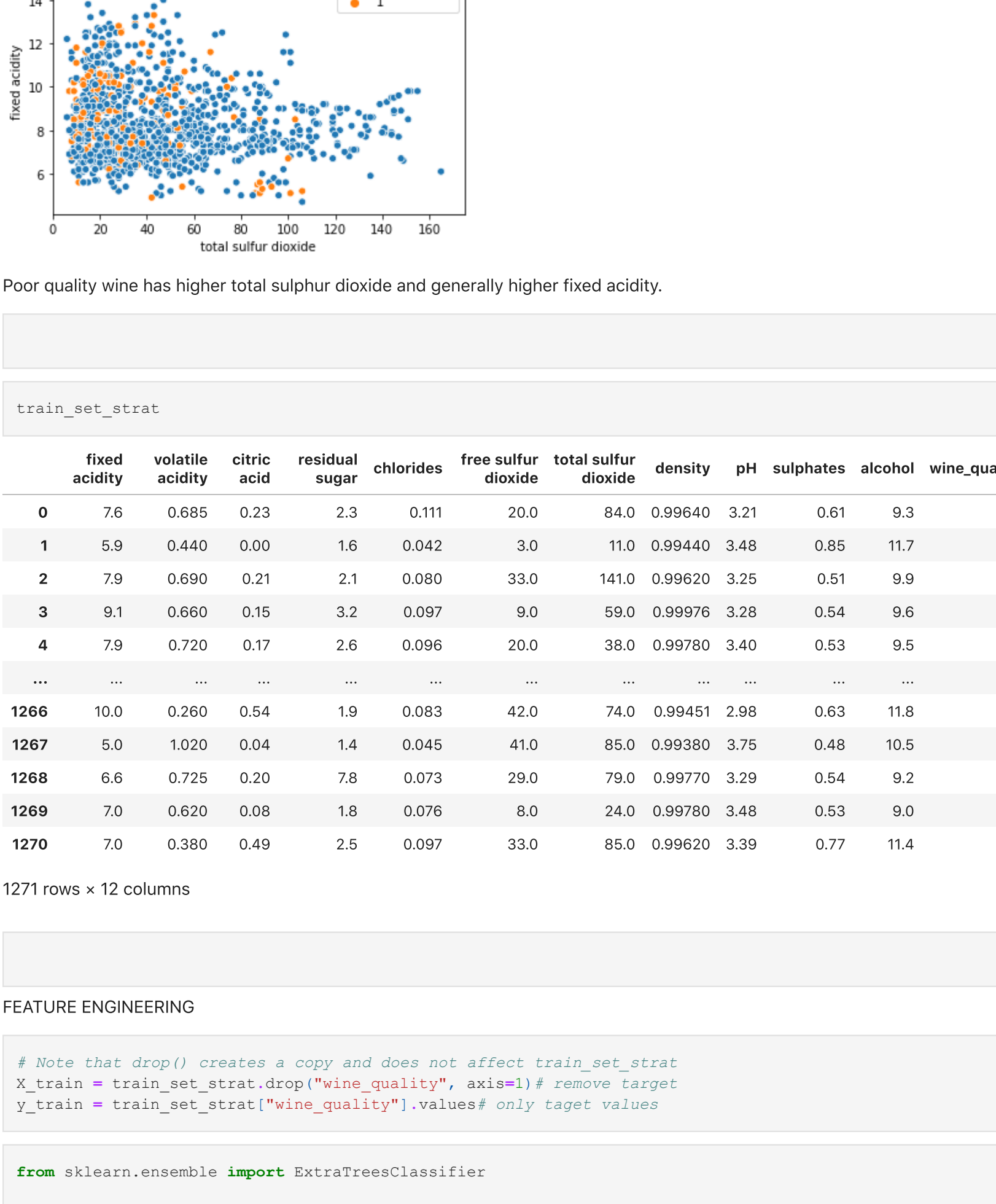
The outliers have been reduced from total sulphur dioxide

```
In [ ]:
```

### BOXPLOT OF FIXED ACIDITY IN DETAIL

```
In [35]: plt.figure(figsize=(14,12))
sns.boxplot(y='fixed acidity', data=train_set_strat)
```

```
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x1c4c6a1f0>
```



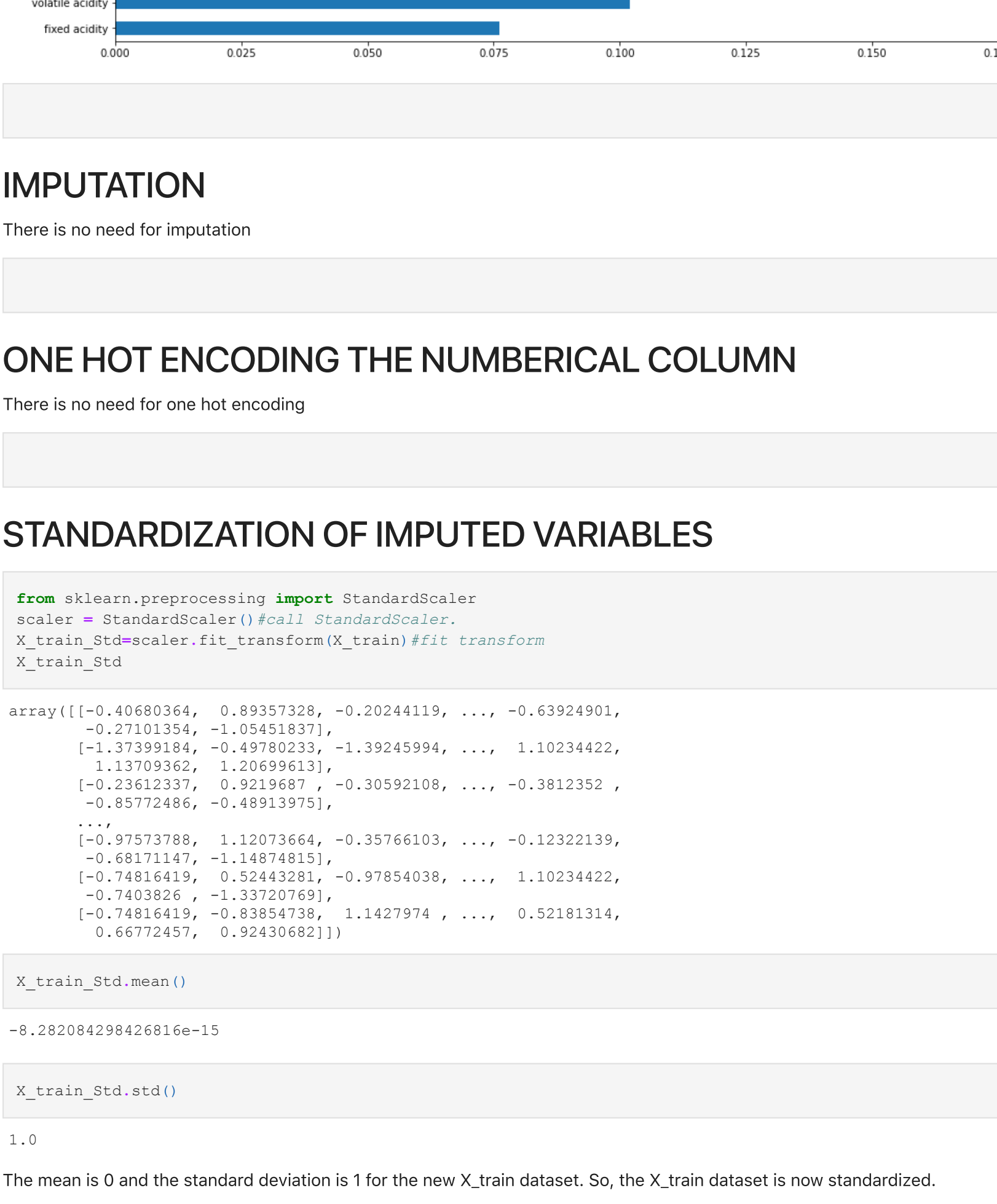
Here is the fixed acidity box plot in depth.

```
In [ ]:
```

### BOXPLOT OF TOTAL SULFUR DIOXIDE IN DETAIL

```
In [36]: plt.figure(figsize=(14,12))
sns.boxplot(y='total sulfur dioxide', data=train_set_strat)
```

```
Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x125da1d00>
```



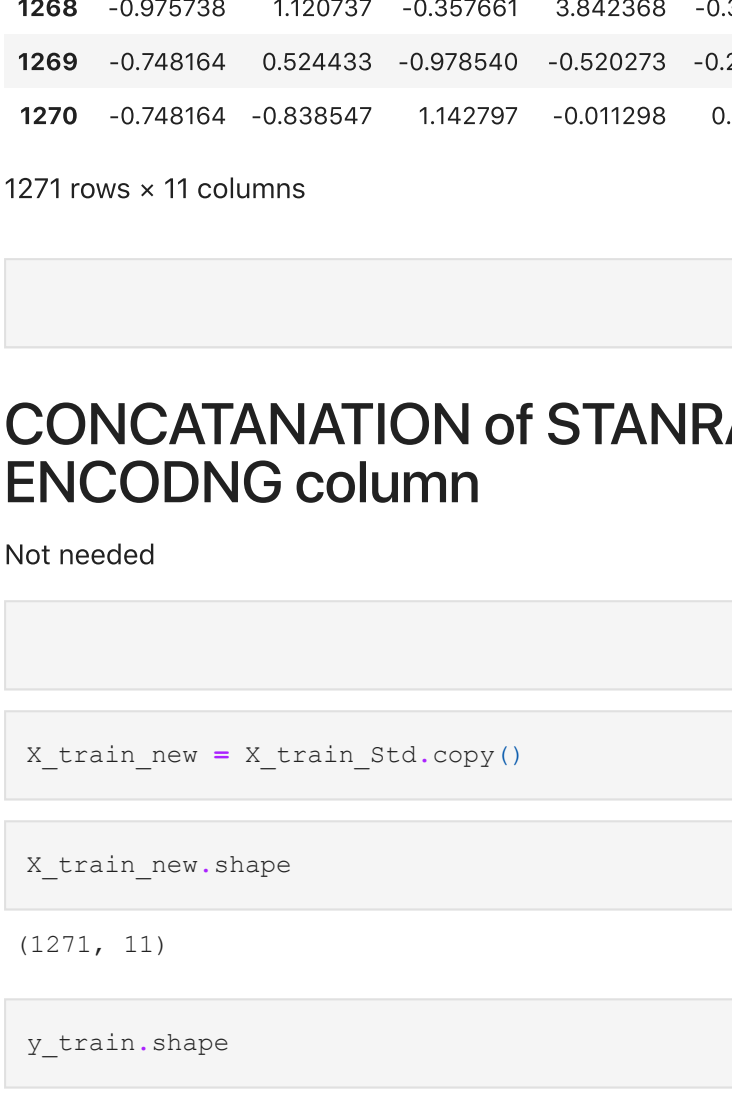
Here is the total sulfur dioxide box plot in depth.

```
In [ ]:
```

### SCATTERPLOT

```
In [37]: sns.scatterplot(x='total sulfur dioxide', y='fixed acidity', hue='wine_quality', data=train_set_strat)
plt.xlim(0, 175)
plt.title("Yards vs Week into the season")
```

```
Out[37]: Text(0.5, 1.0, 'Yards vs Week into the season')
```



Poor quality wine has higher total sulphur dioxide and generally higher fixed acidity.

```
In [ ]:
```

```
In [38]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	wine_quality
0	7.6	0.685	0.23	2.3	0.111	20.0	84.0	0.99640	3.21	0.61	9.3	0
1	5.9	0.440	0.00	1.6	0.042	3.0	11.0	0.99440	3.48	0.85	11.7	0
2	7.9	0.680	0.21	2.1	0.080	33.0	141.0	0.99620	3.28	0.51	9.9	0
3	9.1	0.660	0.15	3.2	0.097	9.0	59.0	0.99976	3.25	0.54	9.6	0
4	7.9	0.720	0.17	2.6	0.096	20.0	38.0	0.99780	3.40	0.53	9.5	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1266	10.0	0.260	0.54	1.9	0.083	42.0	74.0	0.99451	2.98	0.63	11.8	1
1267	5.0	1.020	0.04	1.4	0.045	41.0	85.0	0.99380	3.75	0.48	10.5	0
1268	6.6	0.725	0.20	7.8	0.073	29.0	79.0	0.99770	3.29	0.54	9.2	0
1269	7.0	0.620	0.08	1.8	0.076	8.0	24.0	0.99780	3.48	0.53	9.0	0
1270	7.0	0.380	0.49	2.5	0.097	33.0	85.0	0.99620	3.39	0.77	11.4	0

1271 rows x 12 columns

```
In [ ]:
```

### FEATURE ENGINEERING

```
In [39]: # Note that drop() creates a copy and does not affect train_set_strat
X_train = train_set_strat.drop("wine_quality", axis=1) # remove target
y_train = train_set_strat["wine_quality"].values # only target values
```

```
In [40]: from sklearn.ensemble import ExtraTreesClassifier

X = X_train
y = y_train
bestFeatures2 = ExtraTreesClassifier()
fit2 = bestFeatures2.fit(X,y) # model's fit(x,y)
model = pd.Series(fit2.feature_importances_, index=X.columns)
plt.figure(figsize = (16,5))
model.plot(kind='bar')
plt.title("Importance of features")
plt.show()
```



```
In [ ]:
```

### IMPUTATION

There is no need for imputation

```
In [ ]:
```

### ONE HOT ENCODING THE NUMERICAL COLUMN

There is no need for one hot encoding

```
In [ ]:
```

### STANDARDIZATION OF IMPUTED VARIABLES

```
In [26]: from sklearn.preprocessing import StandardScaler
scaler = StandardScaler() #call StandardScaler
X_train_std = scaler.fit_transform(X_train) #fit transform
X_train_std
```

```
Out[26]: array([[ -0.46880164,  0.89307339, -0.20244119, ..., -0.63924901,
-0.27103354,  1.03994531],
[ -1.37399184, -0.49780233, -1.39245994, ...,  1.10234422,
 1.13709362,  1.20599637],
[ -0.23612337,  0.9219687 , -0.30592108, ..., -0.38123352,
-0.85772486, -0.48913975],
...,
[ -0.97573788,  1.12073664, -0.35766103, ..., -0.12322139,
-0.68711149, -1.14874815],
[ -0.74816419,  0.52443231, -0.97854038, ...,  1.10239422,
-0.7403826 , -1.33207093],
[ -1.00702453, -0.82310683,  1.54981025, ..., -0.24827627,
 0.29875465,  0.27353737],
...,
[ -0.84511439, -1.45353442, -0.09159319, ...,  2.04149855,
 0.80159483, -0.76711368],
[ -1.17196243,  1.24306864, -0.303080 , ...,  2.04149855,
-1.27262092,  1.57394618],
[ -2.20546737, -0.36646469,  1.10215476, ..., -0.57539696,
 0.42446469,  0.36063931]])
```

Out[27]: -8.282084298426916e-15

Out[28]: X\_train\_std.std()

Out[28]: 1.0

The mean is 0 and the standard deviation is 1 for the new X\_train dataset. So, the X\_train dataset is now standardized.

```
In [29]: X_train_std = pd.DataFrame(X_train_std, columns=X_train.columns) #numpy(merical data) to pd
# split again, and we should see the same split
X_train_std, X_test_std, y_train_std, y_test_std = train_test_split(X_train_std, y_train_std, test_size=0.2, random_state=42)
```

```
Out[29]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	-0.46880164	0.89307339	-0.20244119	-0.156720	0.521691	0.386334	1.175382	-0.170105	-0.63924901	-0.27103354	-1.054518
1	-1.37399184	-0.49780233	-1.392460	-0.666695	-0.983945	-1.240546	-1.101113	-1.227544	1.102344	1.137094	1.205996
2	-0.236123	0.921969	-0.305921	-0.302141	-0.154754	1.630419	2.952919	1.606391	-0.187235	-0.857725	-0.489140
3	0.446598	0.751596	-0.161361	0.497676	-0.164764	-0.666353	0.395761	1.606391	-0.187235	-0.857725	-0.489140
4	-0.236123	1.092341	-0.512181	0.061412	0.194379	0.386334	-0.259121	0.570101	0.586317	-0.740383	-0.866059
...	...	...	...	...	...	...	...	...	...	...	...
1266	0.956839	-1.520037	1.401497	-0.447563	-0.089292	2.491708	0.863534	-1.169384	-2.122828	-0.153671	1.301226
1267	-1.886033	2.796066	-1.185500	-0.811116	-0.916483	2.396009	1.205667	-1.544775	2.843937	-1.033738	0.076239
1268	-0.567738	1.207367	-0.357661	3.842368	-0.307500	1.247623	1.019458	0.517230	-0.123221	-0.687111	-1.148748
1269	-0.748164	0.524433	-0.978540	-0.520273	-0.242038	-0.762052	-0.695709	0.570101	1.102344	-0.740383	-1.337208
1270	-0.748164	-0.838547	1.142797	-0.012198	0.216199	1.630419	1.205667	-0.275849	0.521813	0.667725	0.923407

1271 rows x 11 columns

```
In [ ]:
```

### CONCATANATION of STANRRARDIZED DATA AND ONE-HOT ENCODNG column

Not needed

```
In [ ]:
```

```
In [30]: X_train_new = X_train_std.copy()
```

```
In [31]: X_train_new.shape
```

Out[31]: (1271, 11)

```
In [32]: y_train.shape
```

Out[32]: (1271,)

### UPSAMPLING

```
In [33]: df_binary['wine_quality'].value_counts()
```

Out[33]:

```
0    1372
1     177
Name: wine_quality, dtype: int64
```

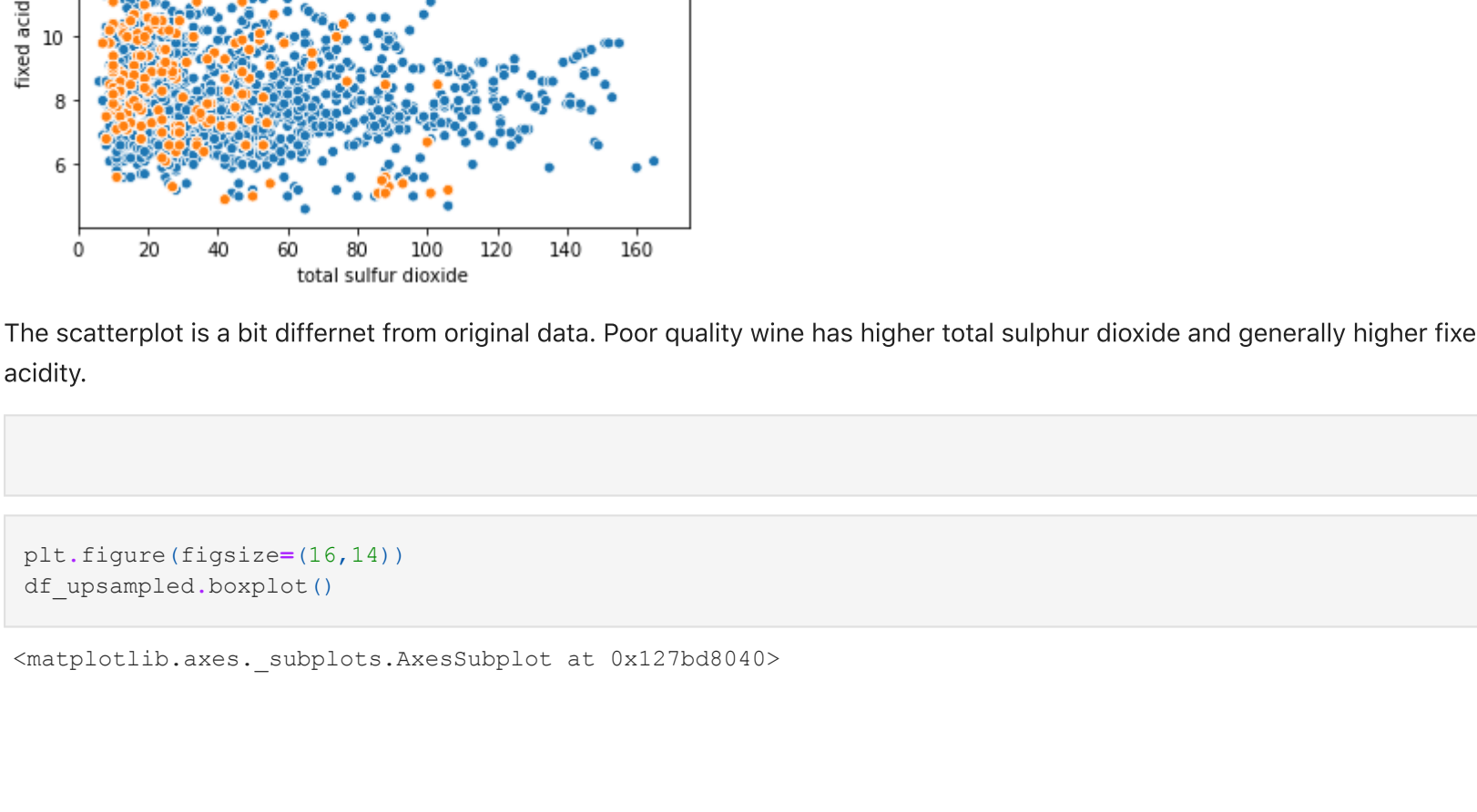
```
In [34]: df_minority = df_binary[df_binary['wine_quality']==1]
#df_major = df_majority['wine_quality']
#df_major.value_counts()
df_majority = df_binary[df_binary['wine_quality']==0]
```

```
In [35]: from sklearn.utils import resample
df_minority_upsampled = resample(df_minority,
                                replace=True, # sample with replacement
                                n_samples=1372, # to match majority class
                                random_state=42) # reproducible results
```

```
In [36]: font = {'size': 10}
matplotlib.rc('font', **font)
fig = plt.subplots(figsize=(10, 3), dpi= 100, facecolor='w', edgecolor='k')
```

```
df_minority_upsampled['wine_quality'].hist(bins=15)
plt.title("Original dataset")
plt.xlabel("Classification Labels")
plt.ylabel("Frequency")
```

Out[36]: Text(0, 0.5, 'Frequency')



```
In [37]: df_upsampled = pd.concat([df_majority, df_minority_upsampled])
df_upsampled['wine_quality'].value_counts()
```

Out[37]:

```
0    1372
1     177
Name: wine_quality, dtype: int64
```

The upsampled dataset has approximately 1372 is classified as good quality wine i.e. 12.6% of the total cases. The original dataset has approximately 1380 is classified as poor quality wine i.e. 86.4% of the total cases.

```
In [38]: df_upsampled
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.7	0.58	0.08	2.50	0.114	14.0	46.0	0.99710	3.24	0.66	9.6
1	7.8	0.50	0.17	1.60	0.082	21.0	102.0	0.99600	3.39	0.48	9.5
2	10.7	0.67	0.22	2.70	0.107	17.0	34.0	1.00040	3.28	0.98	9.9
3	8.5	0.46	0.31	2.25	0.078	32.0	58.0	0.99800	3.33	0.54	9.8
4	6.7	0.46	0.24	1.70	0.077	18.0	34.0	0.99480	3.39	0.60	10.6
...	...	...	...	...	...	...	...	...	...	...	...
404	6.4	0.57	0.12	2.30	0.120	25.0	36.0	0.99519	3.47	0.71	11.3
812	11.1	0.31	0.53	2.20	0.060	3.0	10.0	0.99572	3.02	0.83	10.9
1174	7.9	0.30	0.68	8.30	0.050	37.5	288.0	0.99316	3.01	0.51	12.3
1016	10.5	0.24	0.42	1.80	0.077	6.0	22.0	0.99760	3.21	1.05	10.8
1045	9.2	0.41	0.50	2.50	0.055	12.0	25.0	0.99520	3.34	0.79	13.3

2744 rows x 12 columns

```
In [39]: X_up = df_upsampled.drop("wine_quality", axis=1) # remove target
y_up = df_upsampled["wine_quality"].values # only target values
```

```
In [40]: X_up
```

```
Out[40]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.7	0.58	0.08	2.50	0.114	14.0	46.0	0.99710	3.24	0.66	9.6
1	7.8	0.50	0.17	1.60	0.082	21.0	102.0	0.99600	3.39	0.48	9.5
2	10.7	0.67	0.22	2.70	0.107	17.0	34.0	1.00040	3.28	0.98	9.9
3	8.5	0.46	0.31	2.25	0.078	32.0	58.0	0.99800	3.33	0.54	9.8
4	6.7	0.46	0.24	1.70	0.077	18.0	34.0	0.99480	3.39	0.60	10.6
...	...	...	...	...	...	...	...	...	...	...	...
404	6.4	0.57	0.12	2.30	0.120	25.0	36.0	0.99519	3.47	0.71	11.3
812	11.1	0.31	0.53	2.20	0.060	3.0	10.0	0.99572	3.02	0.83	10.9
1174	7.9	0.30	0.68	8.30	0.050	37.5	288.0	0.99316	3.01	0.51	12.3
1016	10.5	0.24	0.42	1.80	0.077	6.0	22.0	0.99760	3.21	1.05	10.8
1045	9.2	0.41	0.50	2.50	0.055	12.0	25.0	0.99520	3.34	0.79	13.3

2744 rows x 11 columns

```
In [41]: y_up
```

Out[41]: array([0, 0, 0, ..., 1, 1, 1, 1])

```
In [42]: y_up.shape
```

Out[42]: (2744,)

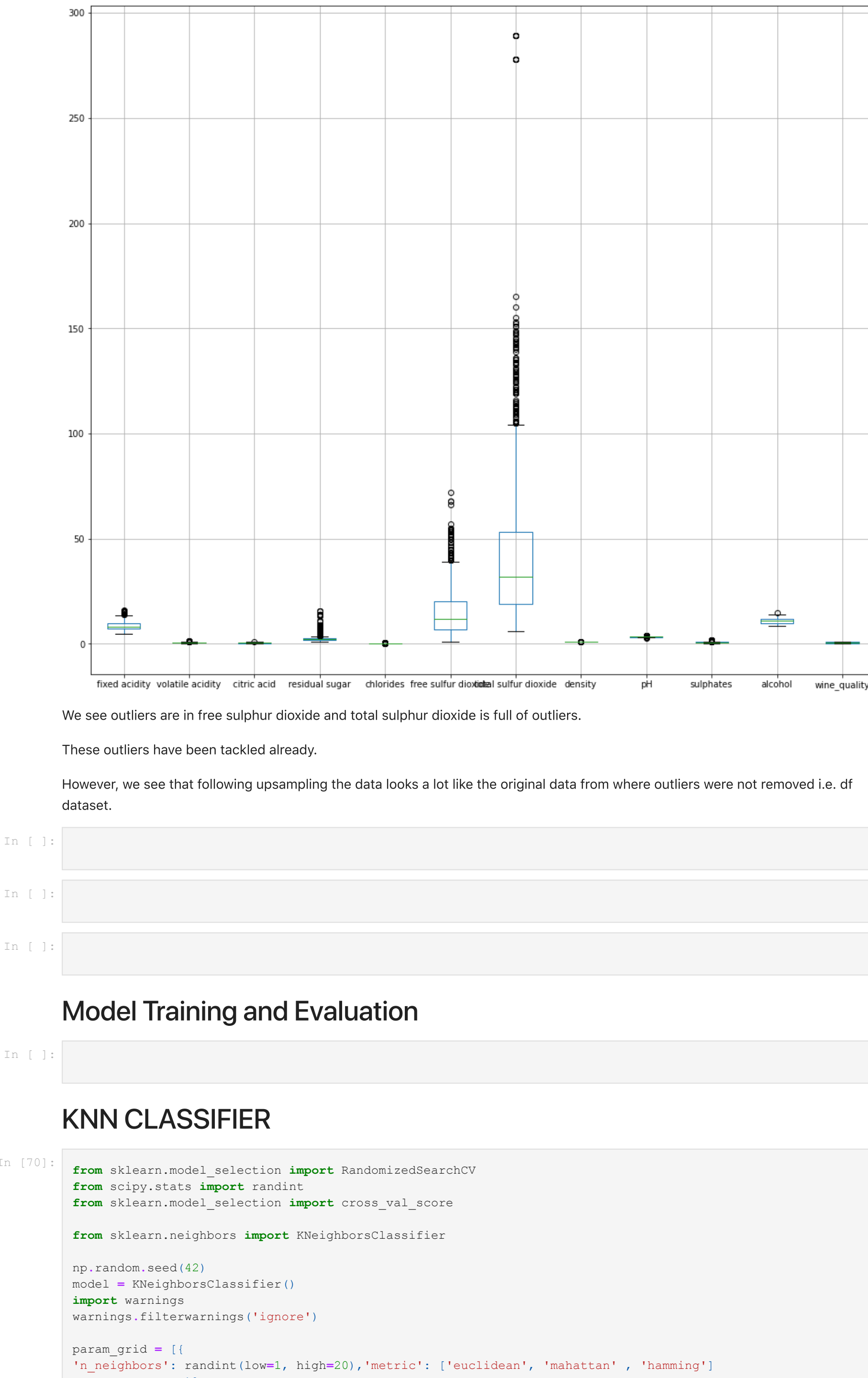
### SPLIT DATA INTO TEST AND TRAINING DATA

```
In [43]: from sklearn.model_selection import train_test_split
# split again, and we should see the same split
X_train_up, X_test_up, y_train_up, y_test_up = train_test_split(X_up, y_up, test_size=0.2, random_state=42)
```

```
In [44]: X_train_up
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
214	7.3	0.550	0.01	1.8	0.093	9.0	15.0	0.99514	3.35	0.58	11.0
933	9.6	0.380	0.31	2.5	0.096	16.0	49.0	0.99820	3.19	0.70	10.0
611	10.4	0.830	0.63	2.8	0.084	5.0	22.0	0.99980	3.26	0.74	11.2
1308	13.3	0.290	0.75	2.8	0.084	23.0	43.0	0.99860			

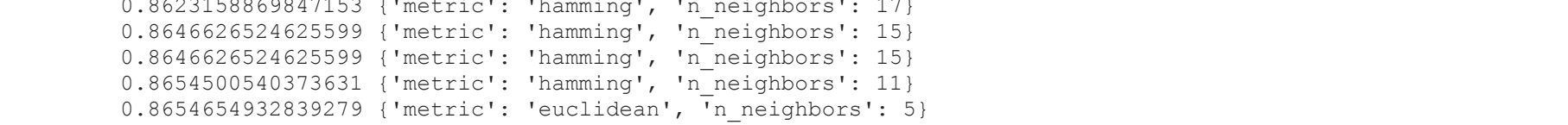
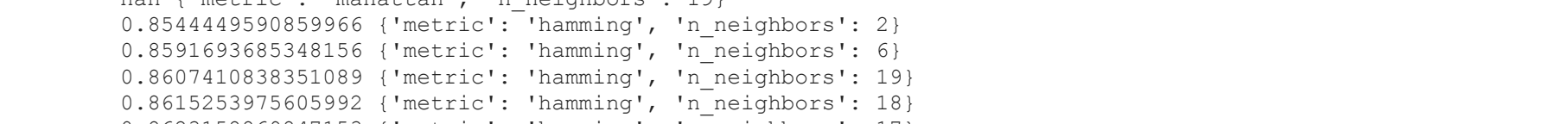
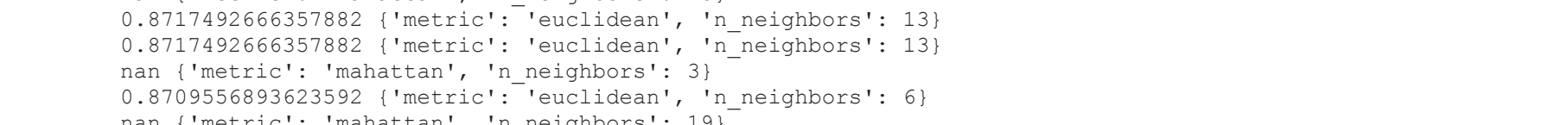
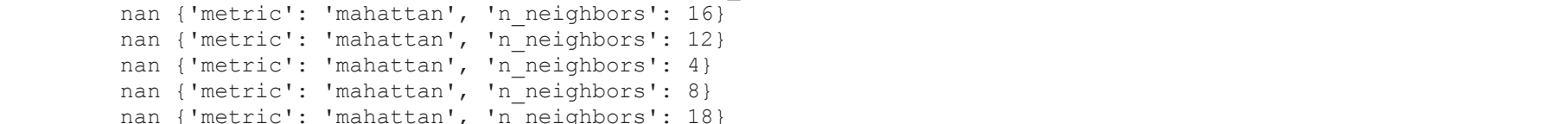
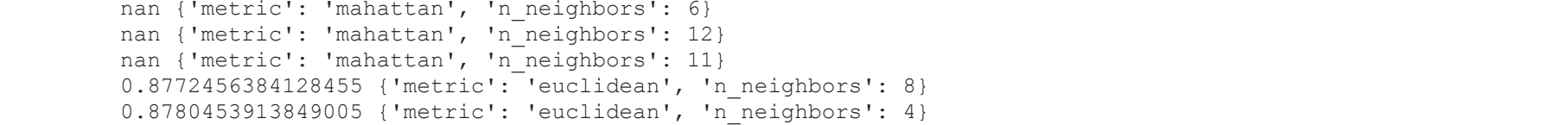
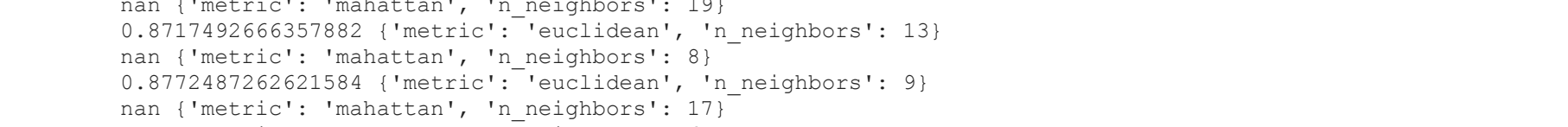
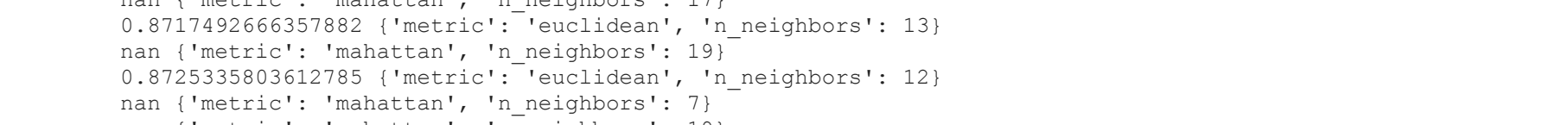
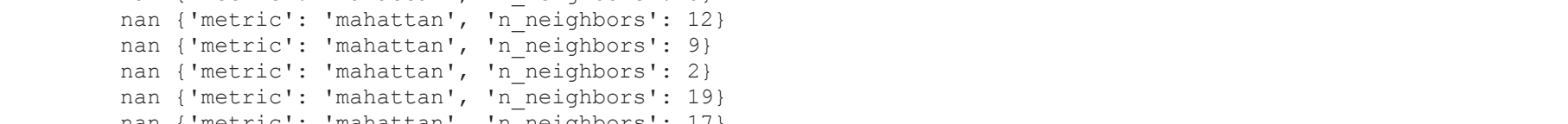
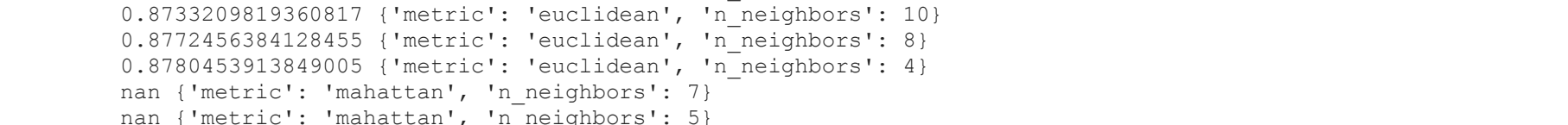
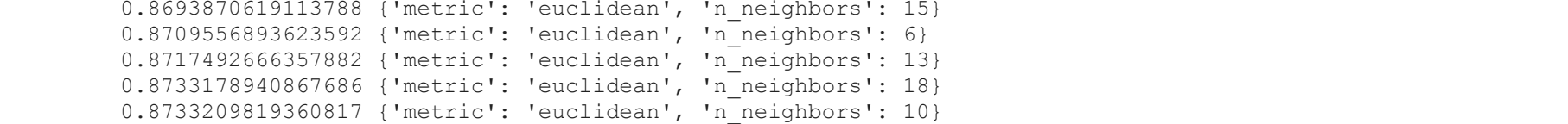
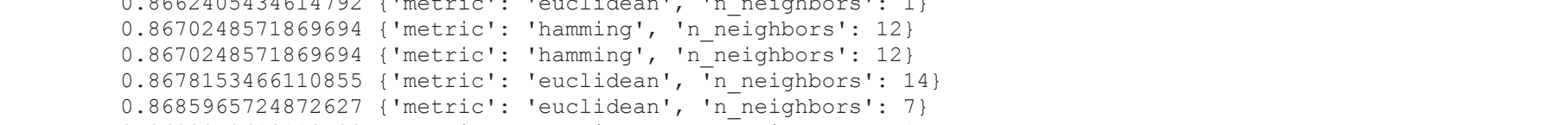
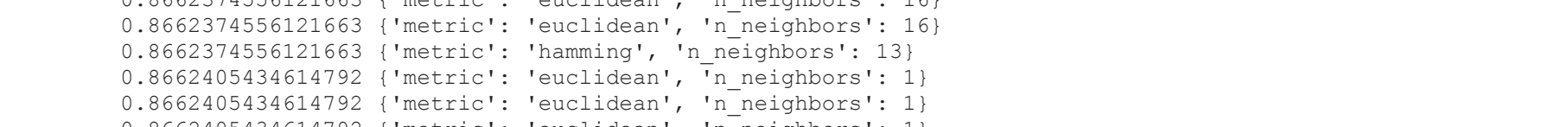
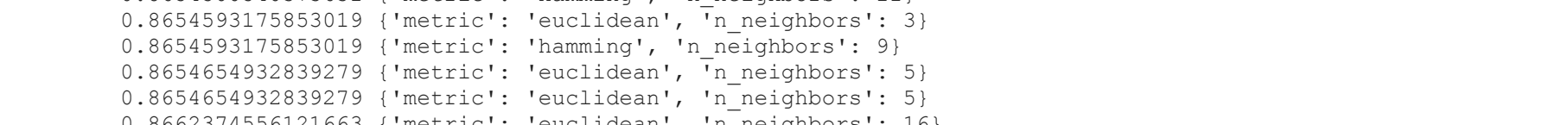
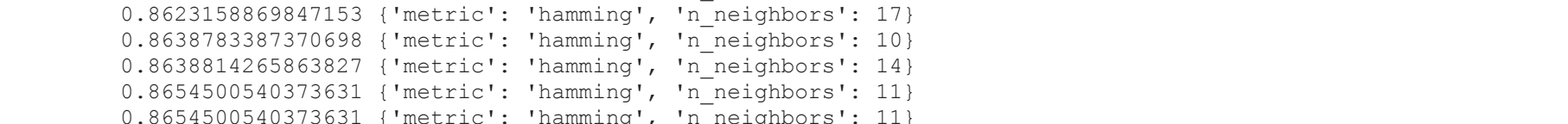
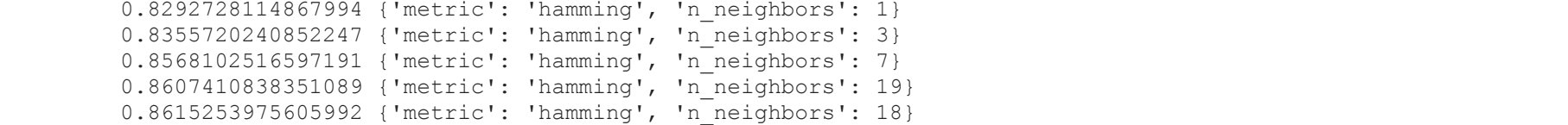
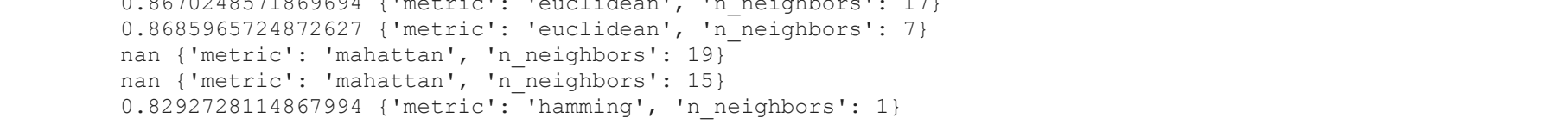
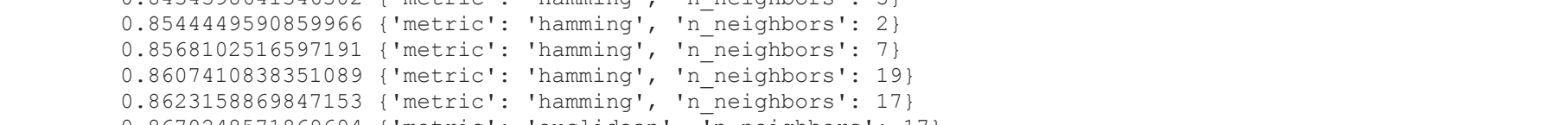
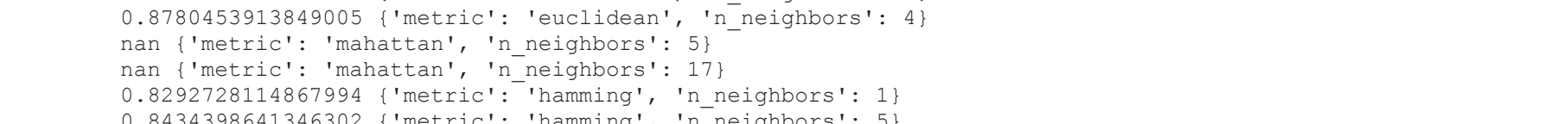
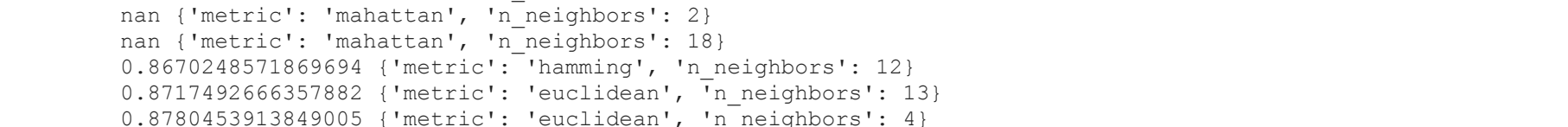
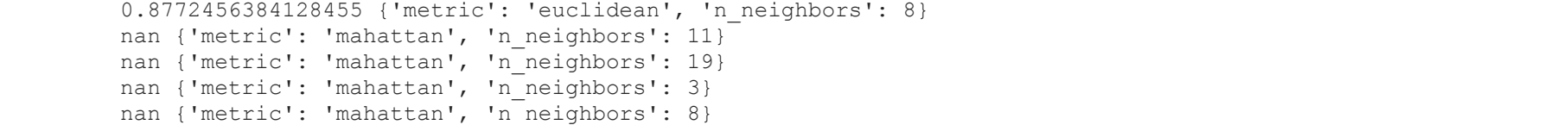
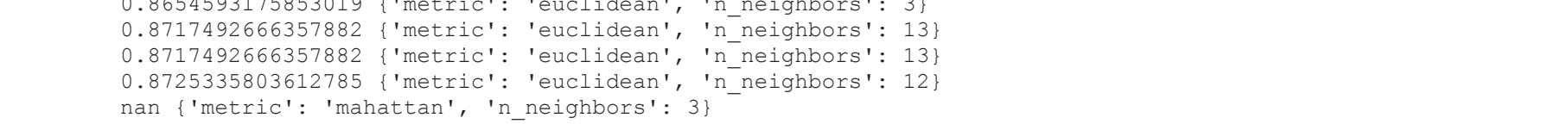
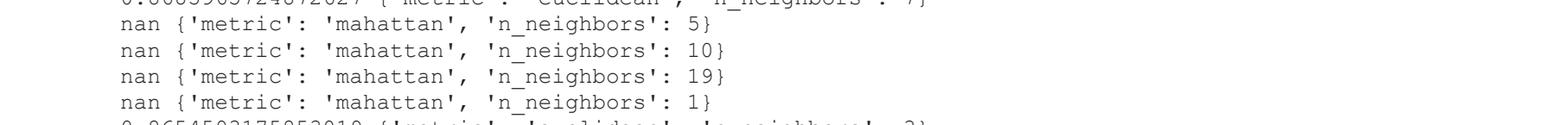
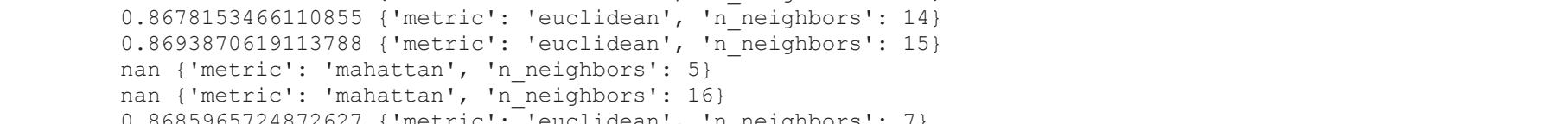
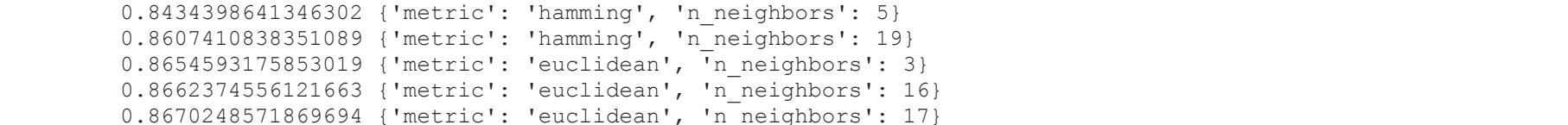
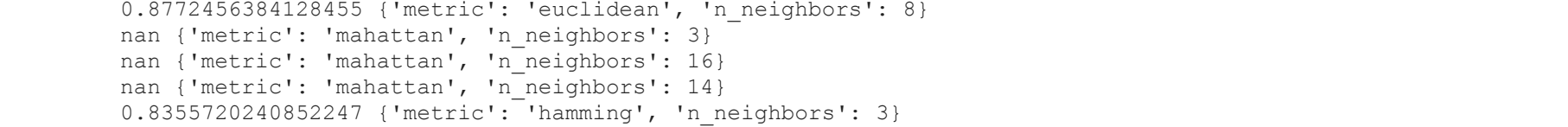
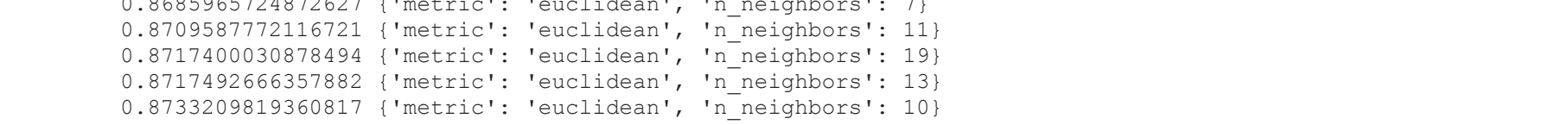
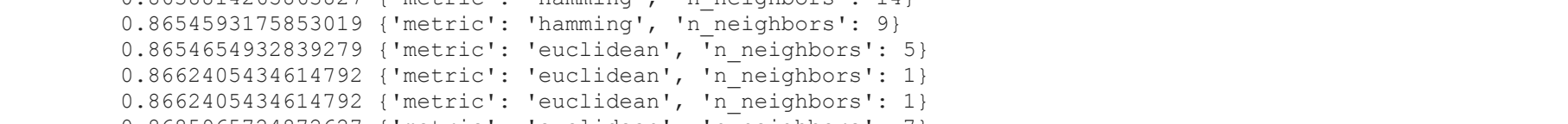
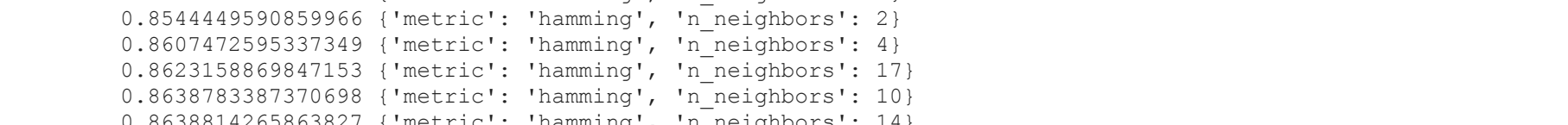
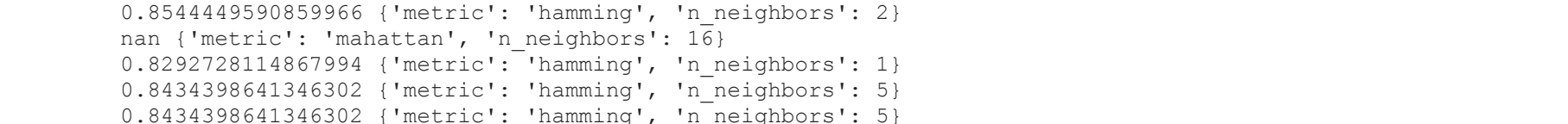
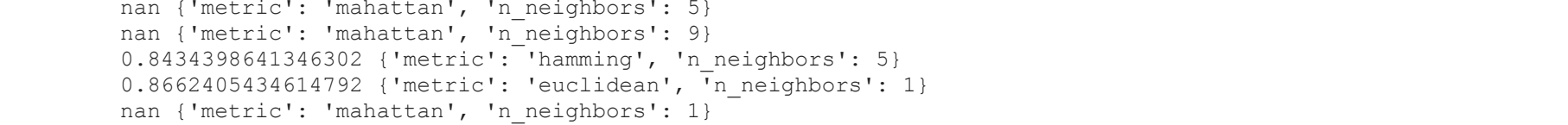
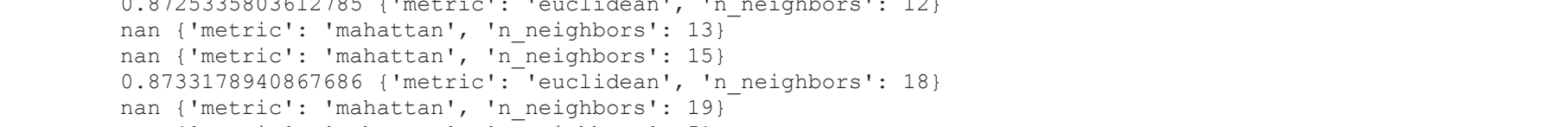
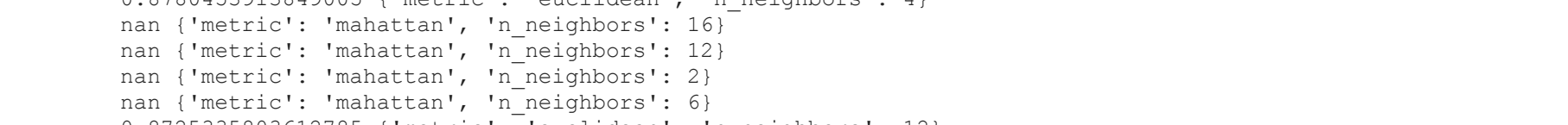
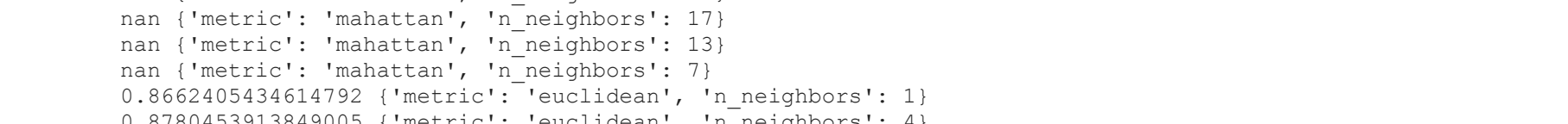
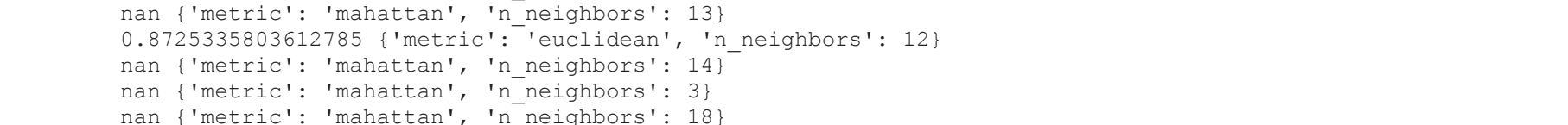
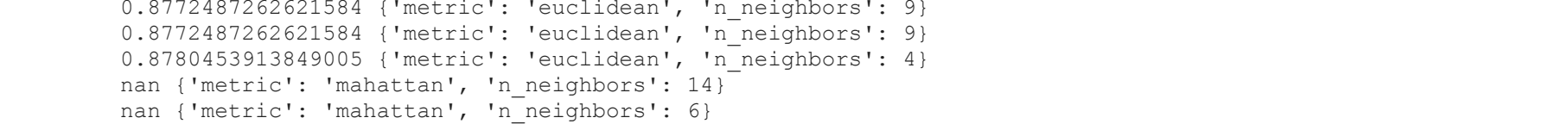
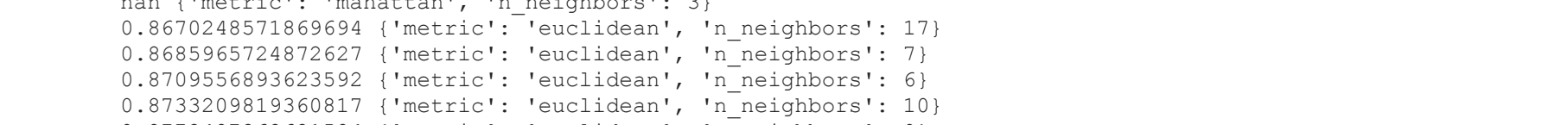
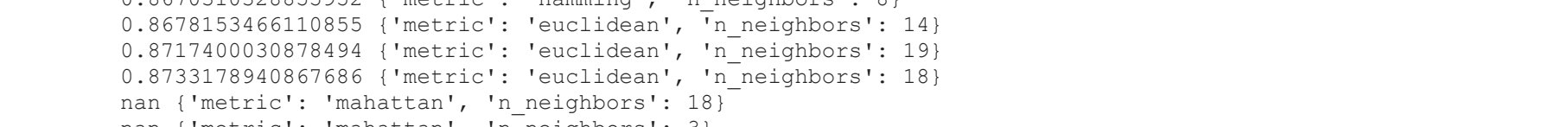
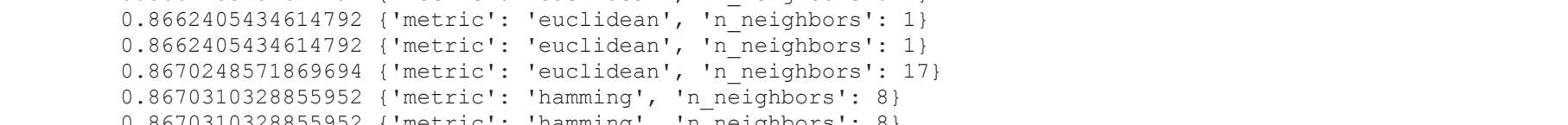
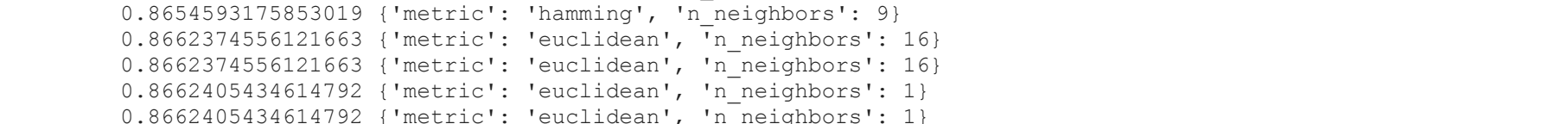
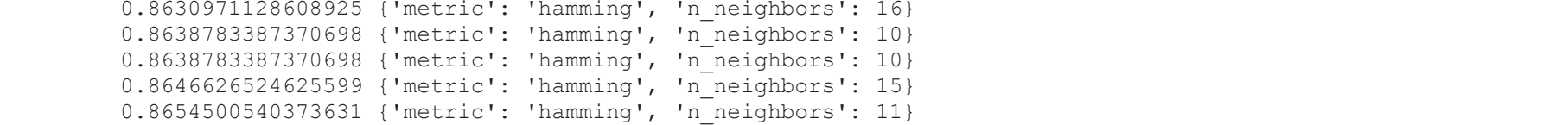
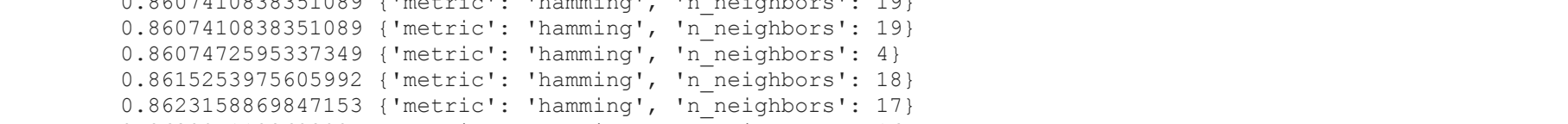
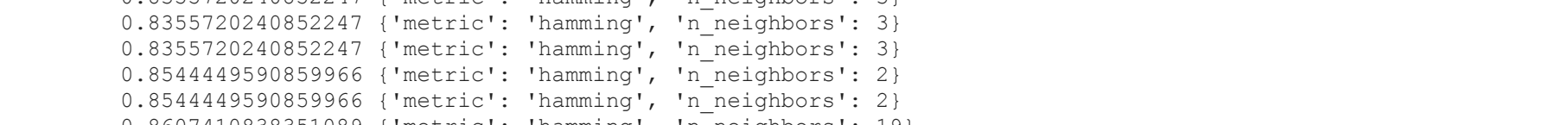
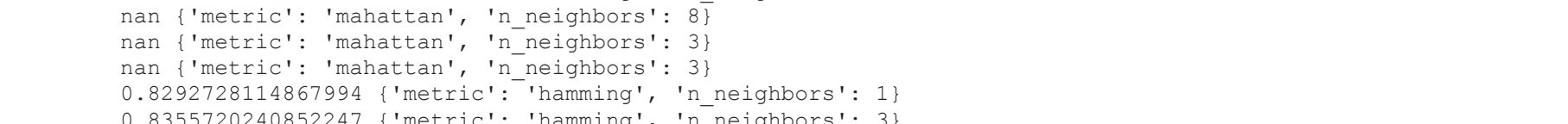
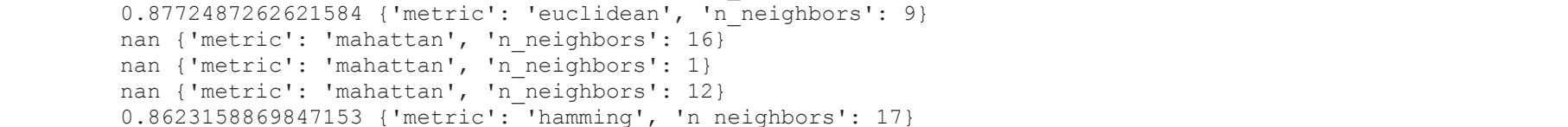
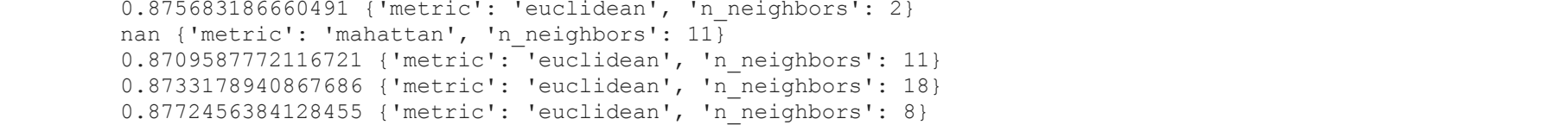
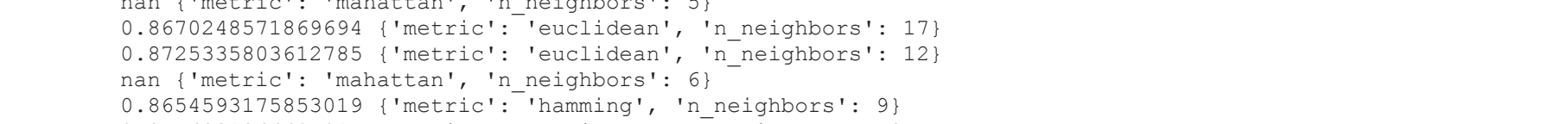
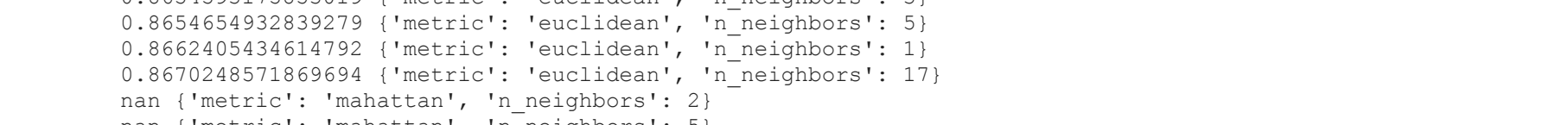
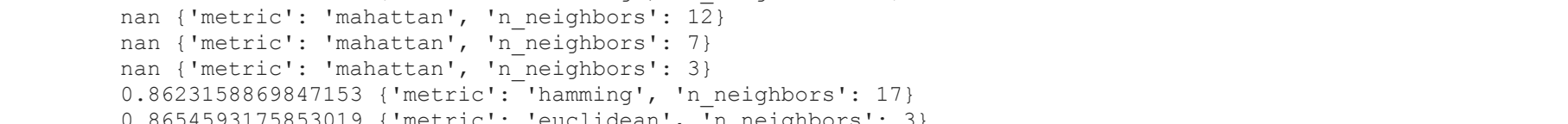
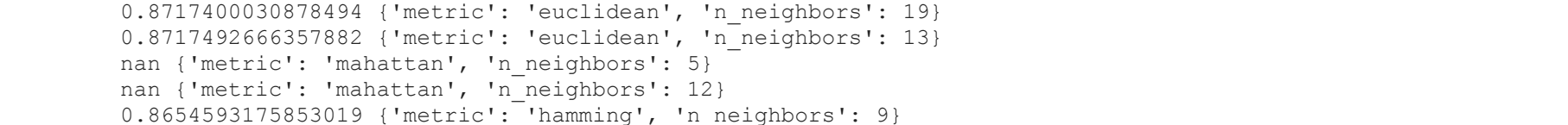
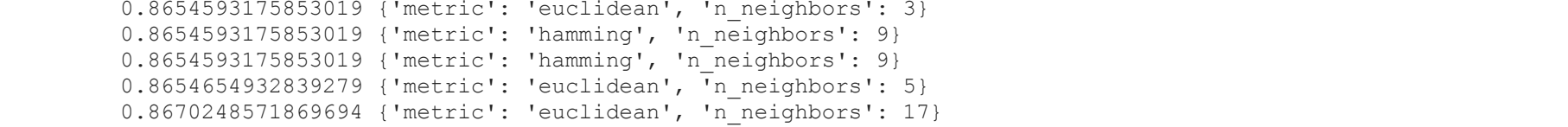
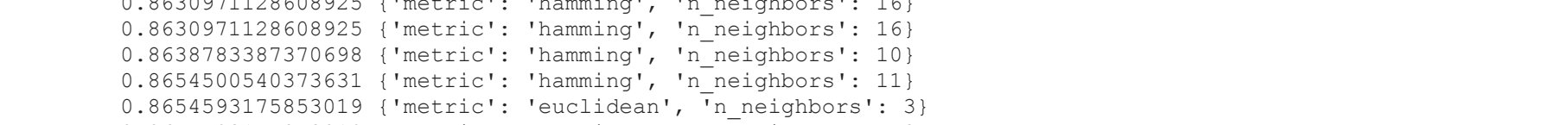
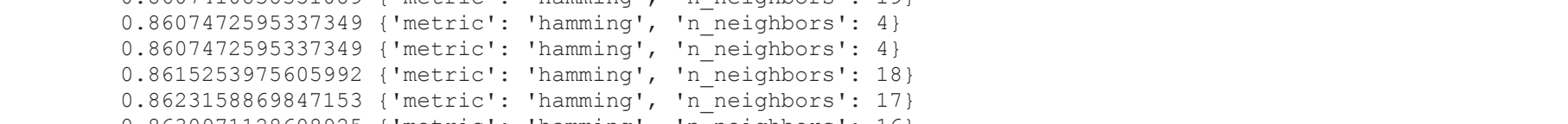
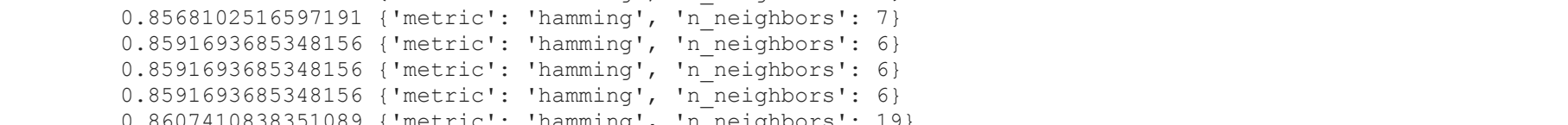
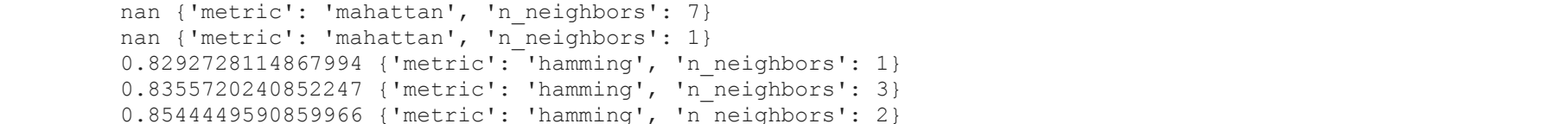
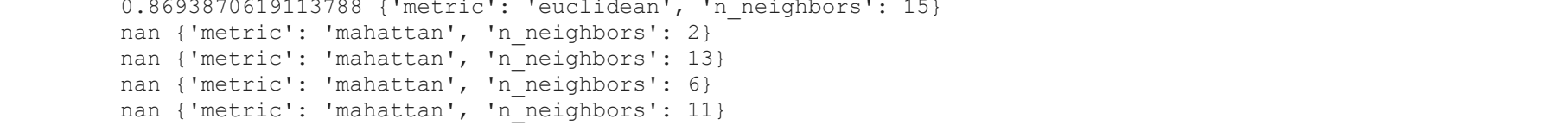
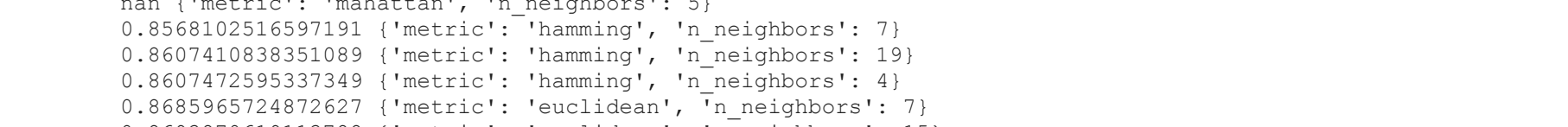
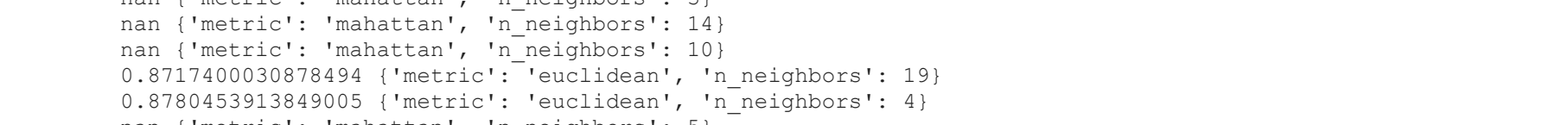
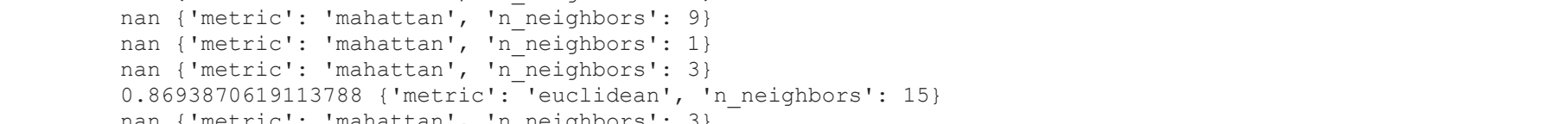
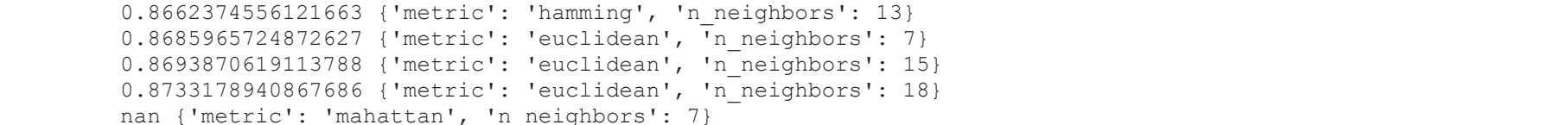
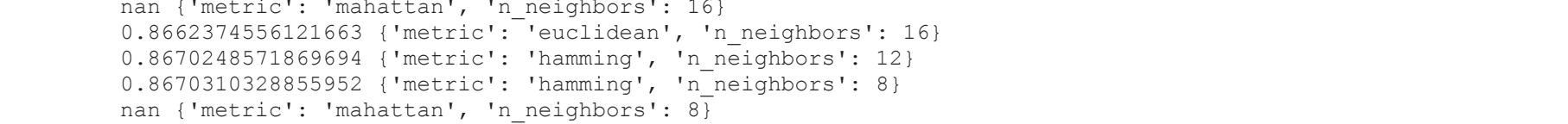
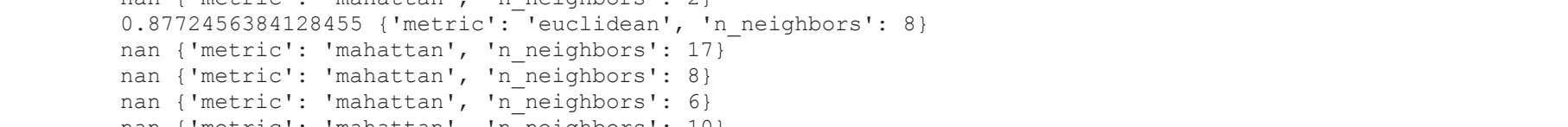
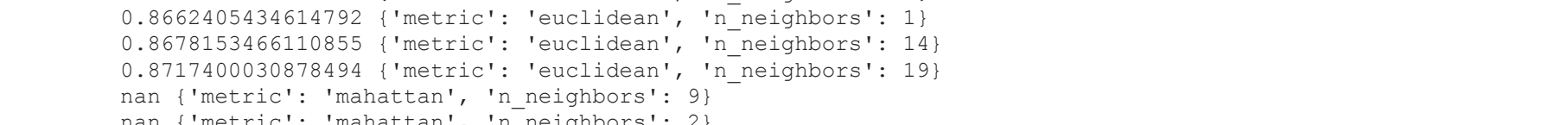
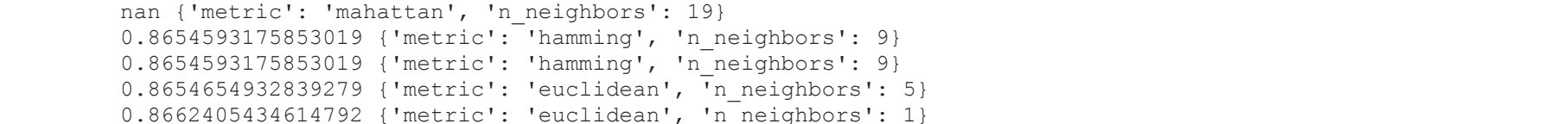
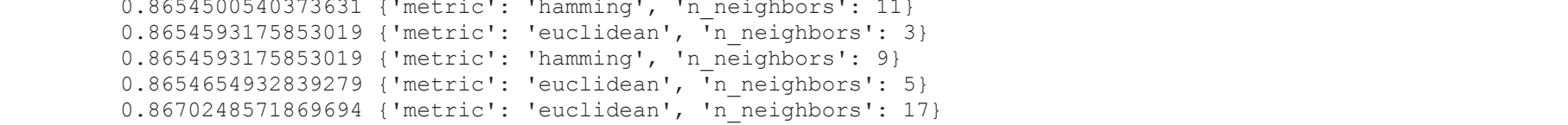
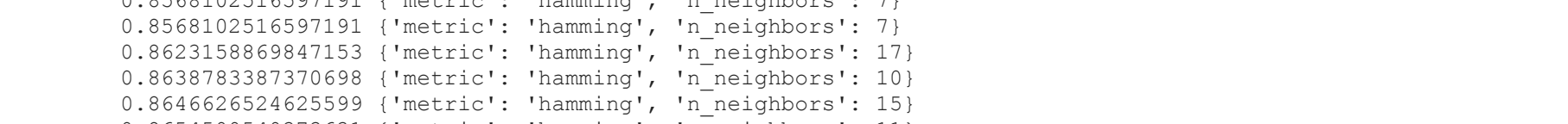
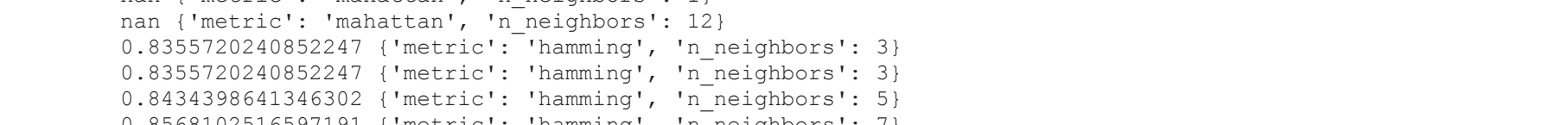
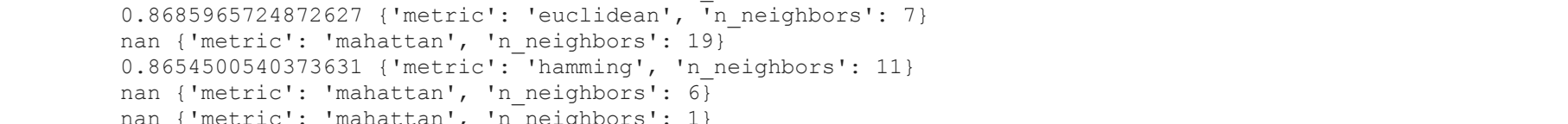
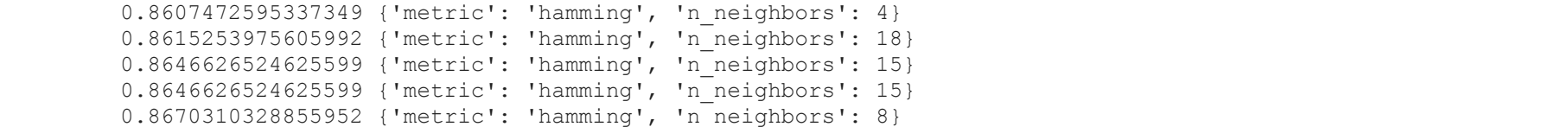
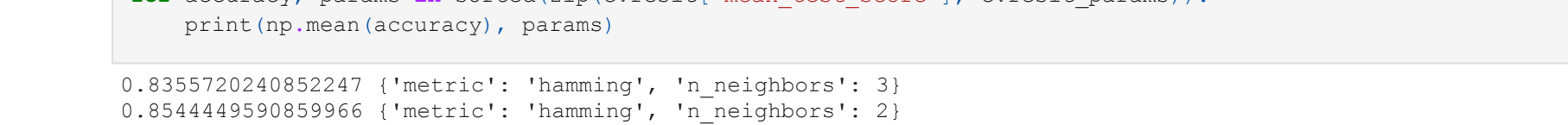




We see outliers are in sulphur dioxide and total sulphur dioxide is full of outliers.

These outliers have been tackled already.

However, we see that following upsampling the data looks a lot like the original data from where outliers were not removed i.e. df













[illegible]

## DECISION TREE - UPSAMPLED

```

from sklearn.tree import DecisionTreeClassifier
from scipy.stats import randint

np.random.seed(42)

model = DecisionTreeClassifier(random_state=42)

# Import warnings
#warnings.filterwarnings('ignore')

param_grid = {
    'max_depth': randint(low=1, high=100), 'min_samples_leaf': randint(low=0, high=200),
    'min_samples_split': randint(low=1, high=200)}

gs = RandomizedSearchCV(model, param_grid, cv=5, n_iter = 100, scoring='accuracy', random_state=42)
gs.fit(X_train, y_train)
gs.best_params_

Out[86]:
{'max_depth': 97, 'min_samples_leaf': 3, 'min_samples_split': 3}

In [87]:
cvresult=gs.cv_results_

cvresult_params=[str(i) for i in cvresult['params']]

for accuracy, params in sorted(zip(cvresult['mean_test_score'], cvresult_params)):
    print(np.mean(accuracy), params)

0.7462414578578999 ('max_depth': 2, 'min_samples_leaf': 137, 'min_samples_split': 159)
0.755303751708429 ('max_depth': 1, 'min_samples_leaf': 129, 'min_samples_split': 62)
0.764446462829156 ('max_depth': 2, 'min_samples_leaf': 137, 'min_samples_split': 159)
0.764446462829156 ('max_depth': 42, 'min_samples_leaf': 190, 'min_samples_split': 14)
0.764446462829156 ('max_depth': 54, 'min_samples_leaf': 190, 'min_samples_split': 145)
0.764446462829156 ('max_depth': 62, 'min_samples_leaf': 189, 'min_samples_split': 87)
0.764446462829156 ('max_depth': 62, 'min_samples_leaf': 189, 'min_samples_split': 174)
0.764446462829156 ('max_depth': 62, 'min_samples_leaf': 189, 'min_samples_split': 20)
0.764446462829156 ('max_depth': 62, 'min_samples_leaf': 189, 'min_samples_split': 72)
0.764446462829156 ('max_depth': 12, 'min_samples_leaf': 189, 'min_samples_split': 39)
0.764446462829156 ('max_depth': 14, 'min_samples_leaf': 189, 'min_samples_split': 74)
0.764446462829156 ('max_depth': 8, 'min_samples_leaf': 189, 'min_samples_split': 123)
0.764446462829156 ('max_depth': 96, 'min_samples_leaf': 189, 'min_samples_split': 128)
0.764446462829156 ('max_depth': 71, 'min_samples_leaf': 171, 'min_samples_split': 7)
0.77153986332574 ('max_depth': 33, 'min_samples_leaf': 166, 'min_samples_split': 173)
0.77153986332574 ('max_depth': 79, 'min_samples_leaf': 166, 'min_samples_split': 173)
0.77153986332574 ('max_depth': 33, 'min_samples_leaf': 166, 'min_samples_split': 62)
0.77153986332574 ('max_depth': 99, 'min_samples_leaf': 171, 'min_samples_split': 103)
0.77255125487381 ('max_depth': 33, 'min_samples_leaf': 166, 'min_samples_split': 173)
0.783143507972652 ('max_depth': 13, 'min_samples_leaf': 159, 'min_samples_split': 70)
0.783143507972652 ('max_depth': 41, 'min_samples_leaf': 156, 'min_samples_split': 14)
0.783143507972652 ('max_depth': 39, 'min_samples_leaf': 189, 'min_samples_split': 42)
0.785102055694761 ('max_depth': 3, 'min_samples_leaf': 141, 'min_samples_split': 102)
0.78601478935908 ('max_depth': 24, 'min_samples_leaf': 130, 'min_samples_split': 149)
0.78601478935908 ('max_depth': 69, 'min_samples_leaf': 129, 'min_samples_split': 479)
0.78926400911161 ('max_depth': 51, 'min_samples_leaf': 134, 'min_samples_split': 20)
0.78926400911161 ('max_depth': 4, 'min_samples_leaf': 110, 'min_samples_split': 87)
0.80318906052524 ('max_depth': 92, 'min_samples_leaf': 110, 'min_samples_split': 180)
0.805589058065036 ('max_depth': 86, 'min_samples_leaf': 34, 'min_samples_split': 172)
0.805589058065036 ('max_depth': 54, 'min_samples_leaf': 100, 'min_samples_split': 194)
0.805589058065036 ('max_depth': 89, 'min_samples_leaf': 88, 'min_samples_split': 59)
0.80683712984054 ('max_depth': 49, 'min_samples_leaf': 85, 'min_samples_split': 27)
0.80683712984054 ('max_depth': 59, 'min_samples_leaf': 85, 'min_samples_split': 27)
0.807744874751619 ('max_depth': 3, 'min_samples_leaf': 71, 'min_samples_split': 11)
0.808200455580865 ('max_depth': 11, 'min_samples_leaf': 80, 'min_samples_split': 135)
0.808200455580865 ('max_depth': 62, 'min_samples_leaf': 62, 'min_samples_split': 87)
0.808200455580865 ('max_depth': 83, 'min_samples_leaf': 74, 'min_samples_split': 87)
0.808200455580865 ('max_depth': 43, 'min_samples_leaf': 64, 'min_samples_split': 161)
0.810933940744784 ('max_depth': 52, 'min_samples_leaf': 92, 'min_samples_split': 9)
0.811389521440091 ('max_depth': 4, 'min_samples_leaf': 14, 'min_samples_split': 133)
0.811389521440091 ('max_depth': 39, 'min_samples_leaf': 10, 'min_samples_split': 153)
0.81296623463902 ('max_depth': 53, 'min_samples_leaf': 23, 'min_samples_split': 159)
0.81412306833713 ('max_depth': 45, 'min_samples_leaf': 64, 'min_samples_split': 88)
0.81543236236217 ('max_depth': 33, 'min_samples_leaf': 25, 'min_samples_split': 173)
0.816562920273349 ('max_depth': 90, 'min_samples_leaf': 41, 'min_samples_split': 123)
0.816562920273349 ('max_depth': 33, 'min_samples_leaf': 62, 'min_samples_split': 90)
0.821867881489749 ('max_depth': 39, 'min_samples_leaf': 50, 'min_samples_split': 21)
0.821867881489749 ('max_depth': 62, 'min_samples_leaf': 50, 'min_samples_split': 107)
0.825512055113994 ('max_depth': 28, 'min_samples_leaf': 27, 'min_samples_split': 107)
0.825512055113994 ('max_depth': 89, 'min_samples_leaf': 49, 'min_samples_split': 58)
0.825512055113994 ('max_depth': 96, 'min_samples_leaf': 49, 'min_samples_split': 103)
0.826436920902013 ('max_depth': 36, 'min_samples_leaf': 49, 'min_samples_split': 103)
0.826436920902013 ('max_depth': 96, 'min_samples_leaf': 49, 'min_samples_split': 95)
0.826436920902013 ('max_depth': 36, 'min_samples_leaf': 32, 'min_samples_split': 87)
0.834660462829156 ('max_depth': 71, 'min_samples_leaf': 8, 'min_samples_split': 87)
0.834660462829156 ('max_depth': 34, 'min_samples_leaf': 3, 'min_samples_split': 87)
0.834660462829156 ('max_depth': 79, 'min_samples_leaf': 2, 'min_samples_split': 102)
0.8401138952144009 ('max_depth': 53, 'min_samples_leaf': 0, 'min_samples_split': 3)
nmin ('max_depth': 76, 'min_samples_leaf': 130, 'min_samples_split': 0)
0.7462414578578999 ('max_depth': 2, 'min_samples_leaf': 137, 'min_samples_split': 159)
0.7462414578578999 ('max_depth': 2, 'min_samples_leaf': 129, 'min_samples_split': 53)
0.7462414578578999 ('max_depth': 2, 'min_samples_leaf': 16, 'min_samples_split': 103)
0.7462414578578999 ('max_depth': 2, 'min_samples_leaf': 137, 'min_samples_split': 159)
0.755303751708429 ('max_depth': 1, 'min_samples_leaf': 15, 'min_samples_split': 188)
0.755303751708429 ('max_depth': 1, 'min_samples_leaf': 33, 'min_samples_split': 98)
0.755303751708429 ('max_depth': 1, 'min_samples_leaf': 137, 'min_samples_split': 159)
0.764446462829156 ('max_depth': 12, 'min_samples_leaf': 184, 'min_samples_split': 37)
0.764446462829156 ('max_depth': 11, 'min_samples_leaf': 184, 'min_samples_split': 36)
0.764446462829156 ('max_depth': 12, 'min_samples_leaf': 156, 'min_samples_split': 176)
0.764446462829156 ('max_depth': 16, 'min_samples_leaf': 184, 'min_samples_split': 171)
0.764446462829156 ('max_depth': 19, 'min_samples_leaf': 188, 'min_samples_split': 107)
0.764446462829156 ('max_depth': 19, 'min_samples_leaf': 188, 'min_samples_split': 163)
0.764446462829156 ('max_depth': 20, 'min_samples_leaf': 190, 'min_samples_split': 85)
0.764446462829156 ('max_depth': 35, 'min_samples_leaf': 191, 'min_samples_split': 48)
0.764446462829156 ('max_depth': 54, 'min_samples_leaf': 191, 'min_samples_split': 100)
0.764446462829156 ('max_depth': 63, 'min_samples_leaf': 190, 'min_samples_split': 124)
0.764446462829156 ('max_depth': 64, 'min_samples_leaf': 190, 'min_samples_split': 68)
0.764446462829156 ('max_depth': 64, 'min_samples_leaf': 194, 'min_samples_split': 144)
0.764446462829156 ('max_depth': 70, 'min_samples_leaf': 199, 'min_samples_split': 154)
0.764446462829156 ('max_depth': 74, 'min_samples_leaf': 185, 'min_samples_split': 144)
0.764446462829156 ('max_depth': 74, 'min_samples_leaf': 185, 'min_samples_split': 92)
0.764446462829156 ('max_depth': 82, 'min_samples_leaf': 193, 'min_samples_split': 53)
0.764446462829156 ('max_depth': 94, 'min_samples_leaf': 189, 'min_samples_split': 196)
0.764446462829156 ('max_depth': 98, 'min_samples_leaf': 189, 'min_samples_split': 173)
0.77153986332574 ('max_depth': 12, 'min_samples_leaf': 177, 'min_samples_split': 162)
0.77153986332574 ('max_depth': 33, 'min_samples_leaf': 179, 'min_samples_split': 122)
0.77153986332574 ('max
```



















```

for accuracy, params in sorted(cv_results["mean_test_score"], cvresult_params):
    print(np.mean(accuracy), params)

0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 150]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 50]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 100]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 125]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75]
0.894568473055148 ['algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 50]

clf = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm='SAMME.R', learning_rate=1, n_estimators=100)

Accuracies=cross_val_score(clf, x_train_new_y_train, cv=5, scoring='accuracy')

np.mean(Accuracies)

0.901658175810561

ADAPTIVE BOOSTING WITH RADOM FOREST UPSAMPLED

from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from scipy.stats import randint
from sklearn.model_selection import GridSearchCV
import numpy

_ = forest = AdaBoostClassifier(bootstrap = False, max_depth=49, min_samples_split=2, n_estimators = 44,
                                learning_rate=0.1, n_estimators=100,
                                random_state=42)

params = {'learning_rate': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1],
          'algorithm': ['SAMME', 'SAMME.R'],
          'n_estimators': [50, 100, 125, 150]}

ada_tree_search_cv = GridSearchCV(forest, params, cv=5, scoring='accuracy')

ada_tree_search_cv.fit(x_train_std_up, y_train_up)

params_ada_tree = ada_tree_search_cv.best_params_
params_ada_tree

{'algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 50}

cvresult=ada_tree_search_cv.cv_results_

cvresult_params=[str(i) for i in cvresult["params"]]

for accuracy, params in sorted(cvresult["mean_test_score"], cvresult_params):
    print(np.mean(accuracy), params)

0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.1, 'n_estimators': 75]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.2, 'n_estimators': 75]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.3, 'n_estimators': 75]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 75]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 75]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 100]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 125]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 50]
0.975854214123068 ['algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 75]
0.9758542141
```

```

0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.4, 'n_estimators': 751
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 2001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.6, 'n_estimators': 2251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.7, 'n_estimators': 2001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.8, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 0.9, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.2, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.4, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.6, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.7, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.8, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 0.9, 'n_estimators': 1751
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 1001
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 1251
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 1501
0.9785842141213068 | 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 1751

clf = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm='SAMME.R', learning_rate=1, n_estimators=2000)
clf.fit(X_train_std_up, y_train_up)

Accuracies=cross_val_score(clf, X_train_std_up, y_train_up, cv=5, scoring='accuracy')

np.mean(Accuracies)

0.961275624236901

RANDOM FOREST IN BAGGING

import warnings
warnings.filterwarnings('ignore')
from sklearn.ensemble import BaggingClassifier

np.random.seed(42)

model = BaggingClassifier(RandomForestClassifier(random_state=42), random_state=42)

param_grid = [
    {'max_samples': randint(low=0.1, high=30), 'n_estimators': randint(low=0.1, high=30), 'bootstrap': [True, False]},
    {'gs': RandomisedSearchCV(model, param_grid, cv=5, n_iter=60, scoring='accuracy', random_state=42)},
    {'gs_fit': X_train_new, y_train}
]

gs.best_params_

{'bootstrap': False, 'max_samples': 26, 'n_estimators': 2}

cvresult=gs.cv_results_

cvresult_params=[str(i) for i in cvresult['params']]

for accuracy, params in sorted(zip(cvresult['mean_test_score'], cvresult_params)):
    print(np.mean(Accuracies), params)

0.8591639328516397 ('bootstrap': False, 'max_samples': 20, 'n_estimators': 1)
0.860377995895279 ('bootstrap': True, 'max_samples': 13, 'n_estimators': 2)
0.861537157352925 ('bootstrap': True, 'max_samples': 26, 'n_estimators': 6)
0.8621157155157155 ('bootstrap': True, 'max_samples': 21, 'n_estimators': 10)
0.8623127991354023 ('bootstrap': False, 'max_samples': 23, 'n_estimators': 4)
0.8623127991354023 ('bootstrap': False, 'max_samples': 27, 'n_estimators': 20)
0.862315886984713 ('bootstrap': True, 'max_samples': 23, 'n_estimators': 125)
0.862315886984713 ('bootstrap': True, 'max_samples': 23, 'n_estimators': 25)
0.862315886984713 ('bootstrap': False, 'max_samples': 23, 'n_estimators': 29)
0.863102070102054 ('bootstrap': False, 'max_samples': 1, 'n_estimators': 91)

```

```

0.86310020207102054    'bootstrap': False, 'max_samples': 13, 'n_estimators': 22
0.86310020207102054    'bootstrap': False, 'max_samples': 15, 'n_estimators': 15
0.86310020207102054    'bootstrap': False, 'max_samples': 14, 'n_estimators': 14
0.86310020207102054    'bootstrap': False, 'max_samples': 18, 'n_estimators': 11
0.86310020207102054    'bootstrap': False, 'max_samples': 18, 'n_estimators': 12
0.86310020207102054    'bootstrap': False, 'max_samples': 2, 'n_estimators': 21
0.86310020207102054    'bootstrap': False, 'max_samples': 20, 'n_estimators': 13
0.86310020207102054    'bootstrap': False, 'max_samples': 21, 'n_estimators': 28
0.86310020207102054    'bootstrap': False, 'max_samples': 21, 'n_estimators': 9
0.86310020207102054    'bootstrap': False, 'max_samples': 22, 'n_estimators': 23
0.86310020207102054    'bootstrap': False, 'max_samples': 24, 'n_estimators': 29
0.86310020207102054    'bootstrap': False, 'max_samples': 27, 'n_estimators': 14
0.86310020207102054    'bootstrap': False, 'max_samples': 28, 'n_estimators': 11
0.86310020207102054    'bootstrap': False, 'max_samples': 29, 'n_estimators': 14
0.86310020207102054    'bootstrap': False, 'max_samples': 3, 'n_estimators': 13
0.86310020207102054    'bootstrap': False, 'max_samples': 6, 'n_estimators': 11
0.86310020207102054    'bootstrap': False, 'max_samples': 7, 'n_estimators': 23
0.86310020207102054    'bootstrap': False, 'max_samples': 8, 'n_estimators': 27
0.86310020207102054    'bootstrap': True, 'max_samples': 1, 'n_estimators': 231
0.86310020207102054    'bootstrap': True, 'max_samples': 10, 'n_estimators': 181
0.86310020207102054    'bootstrap': True, 'max_samples': 10, 'n_estimators': 17
0.86310020207102054    'bootstrap': True, 'max_samples': 10, 'n_estimators': 7
0.86310020207102054    'bootstrap': True, 'max_samples': 11, 'n_estimators': 7
0.86310020207102054    'bootstrap': True, 'max_samples': 11, 'n_estimators': 161
0.86310020207102054    'bootstrap': True, 'max_samples': 14, 'n_estimators': 121
0.86310020207102054    'bootstrap': True, 'max_samples': 15, 'n_estimators': 141
0.86310020207102054    'bootstrap': True, 'max_samples': 16, 'n_estimators': 261
0.86310020207102054    'bootstrap': True, 'max_samples': 17, 'n_estimators': 171
0.86310020207102054    'bootstrap': True, 'max_samples': 17, 'n_estimators': 251
0.86310020207102054    'bootstrap': True, 'max_samples': 19, 'n_estimators': 24
0.86310020207102054    'bootstrap': True, 'max_samples': 19, 'n_estimators': 241
0.86310020207102054    'bootstrap': True, 'max_samples': 19, 'n_estimators': 281
0.86310020207102054    'bootstrap': True, 'max_samples': 20, 'n_estimators': 81
0.86310020207102054    'bootstrap': True, 'max_samples': 20, 'n_estimators': 171
0.86310020207102054    'bootstrap': True, 'max_samples': 27, 'n_estimators': 271
0.86310020207102054    'bootstrap': True, 'max_samples': 9, 'n_estimators': 7
0.86310020207102054    'bootstrap': True, 'max_samples': 4, 'n_estimators': 181
0.86310020207102054    'bootstrap': True, 'max_samples': 7, 'n_estimators': 141
0.86310020207102054    'bootstrap': True, 'max_samples': 7, 'n_estimators': 21
0.86310020207102054    'bootstrap': True, 'max_samples': 8, 'n_estimators': 7
0.86310020207102054    'bootstrap': True, 'max_samples': 9, 'n_estimators': 91
0.86310020207102054    'bootstrap': True, 'max_samples': 9, 'n_estimators': 61
0.86388425863637      'bootstrap': False, 'max_samples': 23, 'n_estimators': 5
0.86468829161128      'bootstrap': True, 'max_samples': 29, 'n_estimators': 131
nan ('bootstrap': True, 'max_samples': 23, 'n_estimators': 0)
nan ('bootstrap': True, 'max_samples': 0, 'n_estimators': 0)
nan ('bootstrap': False, 'max_samples': 0, 'n_estimators': 0)
0.870175513354949     'bootstrap': False, 'max_samples': 26, 'n_estimators': 21

clf = BaggingClassifier(RandomForestClassifier(random_state=42, n_estimators=2), random_state=42, bootstrap
clf.fit(X_train,y_train)

Accuracies across val_score(clf,X_train,y_train, cv=5, scoring='accuracy')

np.mean(Accuracies)

0.870175513354949

Random Forest in Bagging - UPSAMPLED

from sklearn.ensemble import BaggingClassifier

np.random.seed(42)

model = BaggingClassifier(RandomForestClassifier(random_state=42), random_state=42)

param_grid = [
    'max_samples': randint(low=0.1, high=100), 'n_estimators': randint(low=1, high=200), 'bootstrap': {True, False}

]

gs = RandomizedSearchCV(model,param_grid,cv=5, n_iter= 60, scoring='accuracy', random_state=42)

gs.fit(X_train,y_train)

gs.best_estimator_

('bootstrap': True, 'max_samples': 26, 'n_estimators': 136)

cvresults = cv_results

cvresults.params

for accuracy, params in sorted(zip(cvresults['max_test_score'], cvresults.params)):
    print(np.mean(accuracy), params)

0.49794988610478363    'bootstrap': False, 'max_samples': 1, 'n_estimators': 321
0.506833712968004     'bootstrap': False, 'max_samples': 1, 'n_estimators': 191
0.782493757609979     'bootstrap': False, 'max_samples': 6, 'n_estimators': 171
0.613667425949094     'bootstrap': True, 'max_samples': 2, 'n_estimators': 101
0.619389972205968     'bootstrap': False, 'max_samples': 2, 'n_estimators': 162
0.687473209398989     'bootstrap': False, 'max_samples': 7, 'n_estimators': 81
0.756198177676537     'bootstrap': False, 'max_samples': 9, 'n_estimators': 31
0.785876991662869     'bootstrap': False, 'max_samples': 13, 'n_estimators': 8
0.800355886656036     'bootstrap': True, 'max_samples': 14, 'n_estimators': 141
0.787693162870161     'bootstrap': False, 'max_samples': 6, 'n_estimators': 71
0.800355886656036     'bootstrap': True, 'max_samples': 14, 'n_estimators': 141
0.791795944419344     'bootstrap': False, 'max_samples': 6, 'n_estimators': 40
0.792170501503447     'bootstrap': True, 'max_samples': 2, 'n_estimators': 174
0.792170501503447     'bootstrap': True, 'max_samples': 2, 'n_estimators': 174
0.795632574031891     'bootstrap': False, 'max_samples': 7, 'n_estimators': 621
0.7956443193393      'bootstrap': False, 'max_samples': 12, 'n_estimators': 40
0.800355886656036     'bootstrap': True, 'max_samples': 14, 'n_estimators': 141
0.800911617312002     'bootstrap': False, 'max_samples': 9, 'n_estimators': 189
0.801360743624444     'bootstrap': True, 'max_samples': 7, 'n_estimators': 151

```

```
0.8022779043280182 ('bootstrap': True,
0.8022779043280182 ('bootstrap': True,
0.8027334851936218 ('bootstrap': False,
0.8027334851936218 ('bootstrap': True,
0.8031890660592254 ('bootstrap': False,
```

```

0.801389066592256 'bootstrap': False, 'max_samples': 13, 'n_estimators': 94)
0.801464693505925 'bootstrap': True, 'max_samples': 11, 'n_estimators': 103)
0.8036446469248292 'bootstrap': False, 'max_samples': 18, 'n_estimators': 74)
0.8036446469248292 'bootstrap': False, 'max_samples': 20, 'n_estimators': 81)
0.804102277990328 'bootstrap': False, 'max_samples': 20, 'n_estimators': 129)
0.804102277990328 'bootstrap': True, 'max_samples': 18, 'n_estimators': 23)
0.804102277990328 'bootstrap': True, 'max_samples': 26, 'n_estimators': 41)
0.804558086580364 'bootstrap': False, 'max_samples': 25, 'n_estimators': 43)
0.804558086580364 'bootstrap': True, 'max_samples': 13, 'n_estimators': 130)
0.80501338952164 'bootstrap': True, 'max_samples': 18, 'n_estimators': 134)
0.80501338952164 'bootstrap': True, 'max_samples': 13, 'n_estimators': 80)
0.805466970387249 'bootstrap': True, 'max_samples': 21, 'n_estimators': 52)
0.8059255251258474 'bootstrap': False, 'max_samples': 23, 'n_estimators': 157)
0.8063791321845411 'bootstrap': False, 'max_samples': 22, 'n_estimators': 150)
0.8063791321845411 'bootstrap': False, 'max_samples': 26, 'n_estimators': 178)
0.8068377129840547 'bootstrap': False, 'max_samples': 20, 'n_estimators': 160)
0.8068377129840547 'bootstrap': True, 'max_samples': 27, 'n_estimators': 163)
0.807282938496593 'bootstrap': False, 'max_samples': 24, 'n_estimators': 130)
0.807282938496593 'bootstrap': False, 'max_samples': 24, 'n_estimators': 187)
0.807282938496593 'bootstrap': False, 'max_samples': 22, 'n_estimators': 189)
0.808200455808655 'bootstrap': False, 'max_samples': 28, 'n_estimators': 190)
0.8086560346446492 'bootstrap': False, 'max_samples': 17, 'n_estimators': 103)
0.8086560346446492 'bootstrap': False, 'max_samples': 27, 'n_estimators': 110)
0.8086560346446492 'bootstrap': True, 'max_samples': 19, 'n_estimators': 92)
0.809200455808655 'bootstrap': True, 'max_samples': 14, 'n_estimators': 89)
0.8100227790428202 'bootstrap': True, 'max_samples': 20, 'n_estimators': 102)
0.81093324607744874 'bootstrap': True, 'max_samples': 24, 'n_estimators': 70)
0.81093324607744874 'bootstrap': True, 'max_samples': 23, 'n_estimators': 128)
0.812030683712985 'bootstrap': False, 'max_samples': 23, 'n_estimators': 36)
0.8129526424360202 'bootstrap': True, 'max_samples': 26, 'n_estimators': 58)
0.814378387693916 'bootstrap': True, 'max_samples': 23, 'n_estimators': 116)
0.8164009116171311 'bootstrap': True, 'max_samples': 26, 'n_estimators': 136)

clf = BaggingClassifier(RandomForestClassifier(random_state=42), n_estimators=22, random_state = 42, bootstrap
clf.fit(X_train_std_up, y_train_up)

Accuracies=cross_val_score(clf,X_train_std_up, y_train_up, cv=5, scoring="accuracy")

np.mean(Accuracies)

0.8127564234369022

FINDING THE TOP THREE CLASSIFIERS FOR HARD AND SOFT VOTING

from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import roc_auc_score
from sklearn.pipeline import Pipeline

knn = KNeighborsClassifier(n_neighbors=1, metric="euclidean")
svm = SVC(kernel="rbf", gamma=0.5, C=1)
dt = DecisionTreeClassifier(max_depth=11, min_samples_leaf=39, min_samples_split=10, random_state = 42)
rf = RandomForestClassifier(bootstrap = True, max_depth=13, min_samples_split=9, max_samples = 64, random
XGBClassifier(learning_rate=0.1, n_estimators=219, max_depth = 129, random_state = 42)
adaptive_boosting = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = "SAMME.R", learn
adaptive_boosting_boosting = BaggingClassifier(RandomForestClassifier(random_state=42), n_estimators=2, random

clf=[('K_neighbour', 'Poly_SVC',
      'Decision tree', 'Random Forest', 'Xgb', 'Adaptive_boosting', '
      'Bagging_random_forest', '

i=0
pre_acc=[]
for officer in (knn,svm, dt,rf,xgb,
                adaptive_boosting, bagging_adaptive_boosting):
    clfier =fit(X_train_new, y_train)
    y_pred = clfier.predict(X_train_new)
    fpr, roc_auc_score =roc_auc_score(y_train, y_pred)
    acc = cross_val_score(clfier,X_train_new,y_train, cv=5, scoring="accuracy")
    mean_acc, std_dev, mean_acc
    pre_acc.append(mean_acc)
    print(clfier, mean_acc)
    i+=1

F_neighbour = 0.866204344614792
Poly_SVC = 0.874989872350009
Decision_tree = 0.885116563135641
Random_Forest = 0.8914283103072611
Xgb = 0.8945715690798277
Adaptive_boosting = 0.9016887705011061
Bagging_random_forest = 0.870377551334949

Here we see that Adaptive Boosting, Random Forest, and XGB

HARD VOTING

from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

rf = RandomForestClassifier(bootstrap = True, max_depth=17, min_samples_split=10, n_estimators = 64, random
xgb = xgbboost.XGBClassifier(learning_rate=0.1, n_estimators=219, max_depth = 129, random_state = 42)
adaptive_boosting = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = "SAMME.R", learn
hard_voting = VotingClassifier(estimators=[('random_forest', rf), ('xgb', xgb),
                                          ('adaptive_boosting', adaptive_boosting)], voting="hard")
hard_voting.fit(X_train_new, y_train)

clf=[('Random Forest', 'Xgb', 'Adaptive Boosting', 'Hard Voting')]

```

```
for clfier in (rf, xgb, adaptive_boost):
    clfier.fit(X_train_new, y_train)
    Accuracies=cross_val_score(clfier,
    print(clf[i], np.mean(Accuracies))
```

```

i=1
RandomForest: 0.8914281303072411
Xgb: 0.8945715609078277
Adaptive_Boosting: 0.9016581750810561
Hard Voting: 0.9024424898065463

SOFT VOTING

from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

rf = RandomForestClassifier(bootstrap = True, max_depth=7, min_samples_split=9, n_estimators = 64, random_xgb = xgbost.XGBClassifier(learning_rate=1, n_estimators=219, max_depth = 129, random_state = 42)
adaptive_boosting = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = 'SAMME.R', learn

soft_voting = VotingClassifier(estimators=[('random_forest', rf), ('xgb', xgb)],
                              ('adaptive_boosting', adaptive_boosting)], voting='soft')

soft_voting.fit(X_train_new, y_train)

clf=[('Random_forest', 'Xgb', 'Adaptive_boosting', 'Soft Voting')]

for clfier in (rf,xgb,adaptive_boosting, soft_voting):
    clfier.fit(X_train_new, y_train)
    Accuracies=cross_val_score(clfier,X_train_new,y_train, cv=5, scoring="accuracy")
    print(clf(i), np.mean(Accuracies))

i+=1

RandomForest: 0.8914281303072411
Xgb: 0.8945715609078277
Adaptive_Boosting: 0.9016581750810561
Soft Voting: 0.8985023910832176

Hard Voting performs better than soft voting.

MAKING PLOT TOP CLASSIFIERS

from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from scipy.stats import randint
from sklearn.pipeline import Pipeline

knn = KNeighborsClassifier(n_neighbors=1, metric='euclidean')
svm = Pipeline([("svm_clrf",SVC(kernel='poly', degree=2, coef0=2, C=8, random_state=42))])
dt = DecisionTreeClassifier(max_depth=1, min_samples_leaf=5, min_samples_split=10, random_state = 42)
rf = RandomForestClassifier(bootstrap = True, max_depth=7, min_samples_split=9, n_estimators = 64, random_xgb = xgbost.XGBClassifier(learning_rate=1, n_estimators=219, max_depth = 129, random_state = 42)
adaptive_boosting = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = 'SAMME.R', learn
bagging_adaptive_boosting = BaggingClassifier(RandomForestClassifier(random_state=4),n_estimators=2, random
soft_voting = VotingClassifier(estimators=[('random_forest', rf), ('xgb', xgb)],
                              ('adaptive_boosting', adaptive_boosting)], voting='soft')
hard_voting = VotingClassifier(estimators=[('random_forest', rf), ('xgb', xgb)],
                              ('adaptive_boosting', adaptive_boosting)], voting='hard')

clf=[('K_neighbour', 'Poly_SVC',
'Decision tree', 'Random_Forest', 'Xgb', 'Adaptive_boosting',
'Bagging_random_forest', 'Soft_voting', 'Hard_voting')]

i=0
pre_acc=[]
for clfier in (knn,svm, dt,rf,xgb,
               adaptive_boosting,bagging_adaptive_boosting, soft_voting,hard_voting ):
    clfier.fit(X_train_new, y_train)
    #pred = clfier.predict(X_train_new)
    facc=accuracy_score(y_train, y_pred)
    acc = cross_val_score(clfier,X_train_new,y_train, cv=5, scoring="accuracy")

    mean_acc = np.mean(acc)
    pre_acc.append(mean_acc)
    #print(clf(i), mean_acc)
    #pre_acc.append(acc)
    print(clf(i), mean_acc)

i+=1

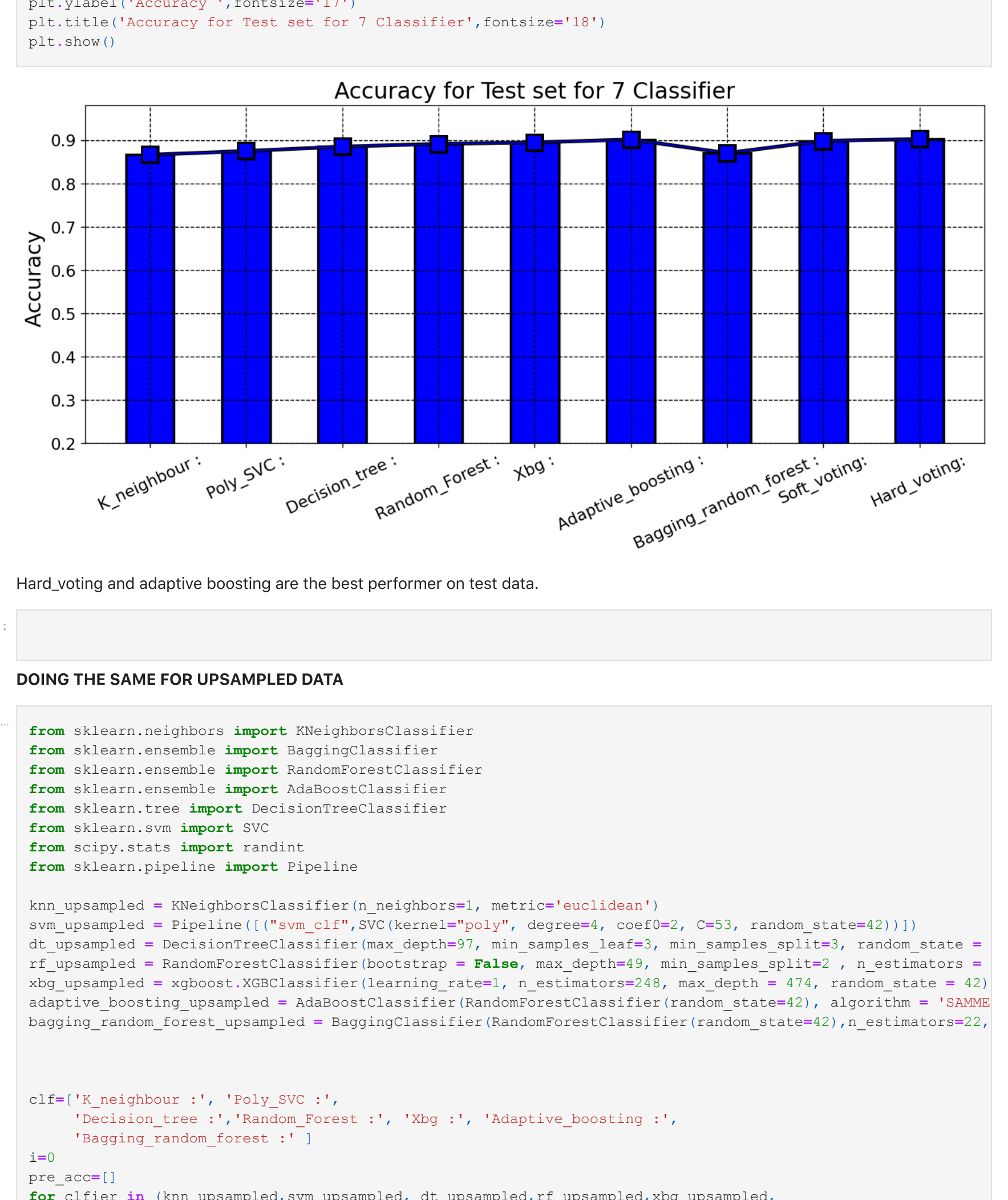
F_neighbour : 0.8662045436414792
Poly_SVC : 0.8748988729150009
Decision tree : 0.885114566315641
Random Forest : 0.8914281303072411
Xgb : 0.8945715609078277
Adaptive_Boosting : 0.9016581750810561
Bagging_Random_Forest : 0.870777551335494
Soft_voting : 0.8985023910832176
Hard_voting : 0.9024424898065463

import matplotlib.pyplot as plt

font = {'size' : 13}
matplotlib.rc('font', **font)
fig, ax1 = plt.subplots(figsize=(13, 5), dpi= 120, facecolor='w', edgecolor='k')

plt.plot(clf,pre_acc,linewidth=3,path_effects=[pe.Stroke(linewidth=3, foreground='k'), pe.Normal()],
         markersize=12,label='Accuracy',markeredgecolor=
ax1.bar(clf,pre_acc,iv=i+2, align='center',width=0.5, alpha=1,color='black', edgecolor='k',capsize=5,color=
plt.ylim(0.7, 0.9)
ax1.set_xticklabels(clf, rotation=25)
ax1.xaxis.grid(color='k', linestyle='--', linewidth=0.8)

```



```
clfier.fit(X_train_Std_up, y_train_up)
#y_pred = clfier.predict(X_train_new)
#acc=accuracy score(y train, y pred)
```

```
mean_acc = np.mean(acc)
pre_acc.append(mean_acc)
#print(clf[i], mean_acc)
```

```

print(cif(i), mean_ac,
      i=i)

K_neighbour : 0.948063781321846
Poly_SVC : 0.9334851936218678
Decision_tree : 0.93565480378131
Random_Forest : 0.9735763097949887
Xgb : 0.961731207289284
Adaptive_boosting : 0.9612756264236901
Bagging_random_forest : 0.8127562642369022

Here we see that Adaptive_boosting, Random_Forest, and Xgb

HARD VOTING

from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

rf_upsampled = RandomForestClassifier(bootstrap = False, max_depth=49, min_samples_split=2, n_estimators =
xbg_upsampled = xgbost.XGBClassifier(learning_rate=1, n_estimators=248, max_depth = 474, random_state = 42)
adaptive_boosting_upsampled = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = 'SAMME')

hard_voting = VotingClassifier(estimators=[('random_forest', rf_upsampled), ('xgb', xbg_upsampled),
('adaptive_boosting', adaptive_boosting_upsampled)], voting='hard')

hard_voting.fit(X_train_std_up, y_train_up)

cif=[('Random_forest:', 'Xgb:', 'Adaptive_boosting:', 'Hard Voting:')]
i=0
for cifier in (rf_upsampled,xbg_upsampled,adaptive_boosting_upsampled, hard_voting):
    cifier.fit(X_train_std_up, y_train_up)
    Accuracies=cross_val_score(cifier,X_train_std_up, y_train_up, cv=5, scoring="accuracy")
    print(cif(i), np.mean(Accuracies))

    i+=1

Random_forest: 0.9735763097949887
Xgb: 0.961731207289284
Adaptive_boosting: 0.9612756264236901
Hard Voting: 0.9699316628701595

SOFT VOTING

from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

rf_upsampled = RandomForestClassifier(bootstrap = False, max_depth=49, min_samples_split=2, n_estimators =
xbg_upsampled = xgbost.XGBClassifier(learning_rate=1, n_estimators=248, max_depth = 474, random_state = 42)
adaptive_boosting_upsampled = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = 'SAMME')

soft_voting = VotingClassifier(estimators=[('random_forest', rf_upsampled), ('xgb', xbg_upsampled),
('adaptive_boosting', adaptive_boosting_upsampled)], voting='soft')

soft_voting.fit(X_train_std_up, y_train_up)

cif=[('Random_forest:', 'Xgb:', 'Adaptive_boosting:', 'Soft Voting:')]
i=0
for cifier in (rf_upsampled,xbg_upsampled,adaptive_boosting_upsampled, soft_voting):
    cifier.fit(X_train_std_up, y_train_up)
    Accuracies=cross_val_score(cifier,X_train_std_up, y_train_up, cv=5, scoring="accuracy")
    print(cif(i), np.mean(Accuracies))

    i+=1

Random_forest: 0.9735763097949887
Xgb: 0.961731207289284
Adaptive_boosting: 0.9612756264236901
Soft Voting: 0.966742596810934

MAKING PLOT FOR UPSAMPLED DATA

from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from scipy.stats import randint
from sklearn.pipeline import Pipeline

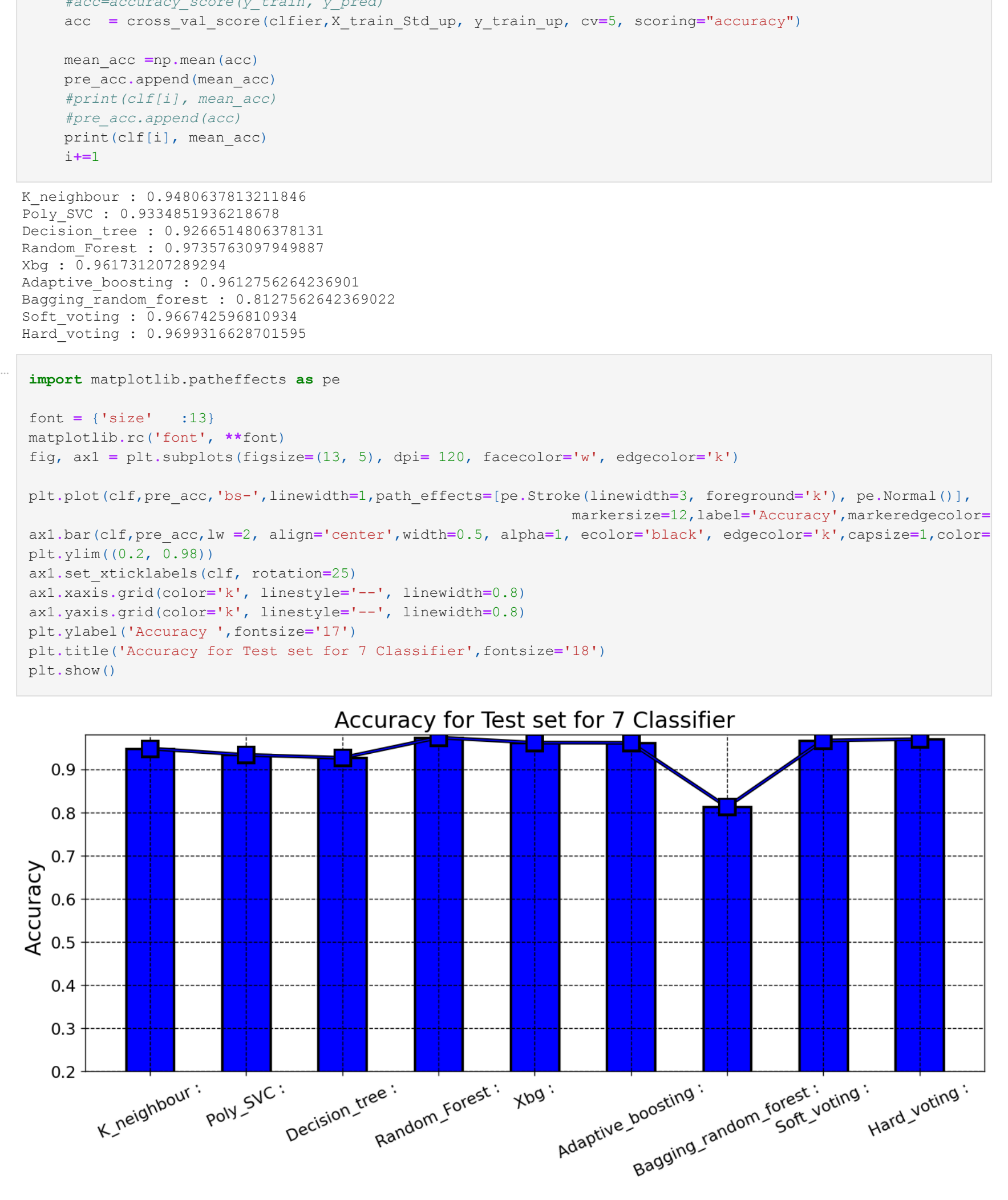
knn_upsampled = KNeighborsClassifier(n_neighbors=1, metric='euclidean')
svm_upsampled = Pipeline([('svm_clf',SVC(kernel='poly', degree=4, coef0=2, C=33, random_state=42))])
dt_upsampled = DecisionTreeClassifier(max_depth=97, min_samples_leaf=3, min_samples_split=3, random_state =
rf_upsampled = RandomForestClassifier(bootstrap = False, max_depth=49, min_samples_split=2, n_estimators =
xbg_upsampled = xgbost.XGBClassifier(learning_rate=1, n_estimators=248, max_depth = 474, random_state = 42)
adaptive_boosting_upsampled = AdaBoostClassifier(RandomForestClassifier(random_state=42), algorithm = 'SAMME')
bagging_random_forest_upsampled = BaggingClassifier(RandomForestClassifier(random_state=42), n_estimators=22,
soft_voting = VotingClassifier(estimators=[('random_forest', rf_upsampled), ('xgb', xbg_upsampled),
('adaptive_boosting', adaptive_boosting_upsampled)], voting='soft')

hard_voting = VotingClassifier(estimators=[('random_forest', rf_upsampled), ('xgb', xbg_upsampled),
('adaptive_boosting', adaptive_boosting_upsampled)], voting='hard')

cif=[('K_neighbour :', 'Poly_SVC :',
'Decision_tree :', 'Random_Forest :', 'Xgb :', 'Adaptive_boosting :',
'Bagging_Random_forest :', 'Soft_voting :', 'Hard_voting :')]

i=0
for_acoef in (knn_upsampled,svm_upsampled, dt_upsampled,rf_upsampled,xbg_upsampled,
adaptive_boosting_upsampled,bagging_random_forest_upsampled , soft_voting , hard_voting):
    cifier.fit(X_train_std_up, y_train_up)

```



Handwriting speed and handwriting are the best performers.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	wine_quality
0	6.6	0.735	0.02	7.9	0.122	68.0	124.0	0.99940	3.47	0.53	9.9	0
1	6.9	0.550	0.15	2.2	0.076	19.0	40.0	0.99610	3.41	0.59	10.1	0



