

Python Project - Predicting Breast Cancer using two Supervised Learning Models and Comparing their Efficiency - KNN

Naga Harshith Bezwada

Methodology

Classification, which is a data mining function that assigns items in a collection to target categories or classes, will be used to make a predictive model in this project. The goal of classification is to accurately predict the target class for each case in the data. Here, our goal is to predict if a patient has cancer or otherwise.

K-Nearest Neighbours:

This algorithm considers all samples including the new sample/data as points on a graph and finds the 'k' nearest points to the new point using euclidean distance or manhattan distance. Euclidean distance is the length of the line segment connecting two points, it is used when all the features are of the same type. In Cartesian coordinates, if $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ are two points in Euclidean space, then the distance (d) from p to q , or from q to p is given by the Pythagorean formula,

https://en.wikipedia.org/wiki/Euclidean_distance

DATASET

Cancer data set UCI ML Repository.

Now we will import dependencies as needed for our second part of the predictive analysis, using KNN algorithm.

```
In [1]: import numpy as np
import matplotlib.pyplot as plt
from math import sqrt
from collections import Counter
import pandas as pd
import random
```

The dataset is a modified Text version of the same UCI Laboratory data and was obtained from [UCI ML repository](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data).

We will now create a dataframe from the file and drop Patient ID and missing values are replaced before we proceed further.

```
In [2]: df = pd.read_csv('breast-cancer-wisconsin.data.txt')

df.replace('?', -99999, inplace=True)
df.drop(['id'], 1, inplace=True)
full_data = df.astype(float).values.tolist()
print(full_data)
```


10.0, 3.0, 1.0, 1.0, 4.0], [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 2.0],
[1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 2.0], [6.0, 1.0, 1.0, 1.0, 2.0],
1.0, 3.0, 1.0, 1.0, 2.0], [5.0, 8.0, 8.0, 8.0, 5.0, 10.0, 7.0, 8.0, 1.0, 4.0],
[8.0, 7.0, 6.0, 4.0, 4.0, 10.0, 5.0, 1.0, 1.0, 4.0], [2.0, 1.0, 1.0, 1.0, 1.0,
1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 5.0, 8.0, 6.0, 5.0, 8.0, 7.0, 10.0, 1.0, 4.0],
[10.0, 5.0, 6.0, 10.0, 6.0, 10.0, 7.0, 10.0, 4.0], [5.0, 8.0, 4.0, 10.0,
5.0, 8.0, 9.0, 10.0, 1.0, 4.0], [1.0, 2.0, 3.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0,
2.0], [10.0, 10.0, 10.0, 8.0, 6.0, 8.0, 7.0, 10.0, 1.0, 4.0], [7.0, 5.0, 10.0,
10.0, 10.0, 10.0, 4.0, 10.0, 3.0, 4.0], [5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.
0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [3.0, 1.0,
1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0,
1.0, 1.0, 2.0], [8.0, 4.0, 4.0, 5.0, 4.0, 7.0, 7.0, 8.0, 2.0, 2.0], [5.0, 1.0,
1.0, 4.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0,
1.0, 1.0, 2.0], [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 2.0], [9.0, 7.0,
7.0, 5.0, 5.0, 10.0, 7.0, 8.0, 3.0, 4.0], [10.0, 8.0, 8.0, 4.0, 10.0, 10.0, 8.
0, 1.0, 1.0, 4.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [5.0,
1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0,
3.0, 1.0, 1.0, 2.0], [5.0, 10.0, 10.0, 9.0, 6.0, 10.0, 7.0, 10.0, 5.0, 4.0],
[10.0, 10.0, 9.0, 3.0, 7.0, 5.0, 3.0, 5.0, 1.0, 4.0], [1.0, 1.0, 1.0, 1.0, 1.0,
0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 2.
0], [5.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 2.0], [8.0, 10.0, 10.0, 10.
0, 5.0, 10.0, 8.0, 10.0, 6.0, 4.0], [8.0, 10.0, 8.0, 8.0, 4.0, 8.0, 7.0, 7.0,
1.0, 4.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [10.0, 10.0, 1.
0.0, 10.0, 7.0, 10.0, 7.0, 10.0, 4.0, 4.0], [10.0, 10.0, 10.0, 10.0, 10.0, 10.0,
3.0, 10.0, 6.0, 4.0], [8.0, 7.0, 8.0, 7.0, 5.0, 5.0, 5.0, 10.0, 2.0, 4.0],
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
1.0, 3.0, 1.0, 1.0, 2.0], [6.0, 10.0, 7.0, 7.0, 6.0, 4.0, 8.0, 10.0, 2.0, 4.
0], [6.0, 1.0, 3.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0,
2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [10.0, 6.0, 4.0, 3.0, 10.0, 10.0, 9.0, 10.0, 1.
0, 4.0], [4.0, 1.0, 1.0, 3.0, 1.0, 5.0, 2.0, 1.0, 1.0, 4.0], [7.0, 5.0, 6.0,
3.0, 3.0, 8.0, 7.0, 4.0, 1.0, 4.0], [10.0, 5.0, 5.0, 6.0, 3.0, 10.0, 7.0, 9.0,
2.0, 4.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 2.0], [10.0, 5.0, 7.
0, 4.0, 4.0, 10.0, 8.0, 9.0, 1.0, 4.0], [8.0, 9.0, 9.0, 5.0, 3.0, 5.0, 7.0, 7.
0, 1.0, 4.0], [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 2.0], [10.0, 10.0,
10.0, 3.0, 10.0, 10.0, 9.0, 10.0, 1.0, 4.0], [7.0, 4.0, 7.0, 4.0, 3.0, 7.0, 7.
0, 6.0, 1.0, 4.0], [6.0, 8.0, 7.0, 5.0, 6.0, 8.0, 8.0, 9.0, 2.0, 4.0], [8.0,
4.0, 6.0, 3.0, 3.0, 1.0, 4.0, 3.0, 1.0, 2.0], [10.0, 4.0, 5.0, 5.0, 5.0, 10.0,
4.0, 1.0, 1.0, 4.0], [3.0, 3.0, 2.0, 1.0, 3.0, 1.0, 3.0, 6.0, 1.0, 2.0], [3.0,
1.0, 4.0, 1.0, 2.0, -99999.0, 3.0, 1.0, 1.0, 2.0], [10.0, 8.0, 8.0, 2.0, 8.0,
10.0, 4.0, 8.0, 10.0, 4.0], [9.0, 8.0, 8.0, 5.0, 6.0, 2.0, 4.0, 10.0, 4.0, 4.
0], [8.0, 10.0, 10.0, 8.0, 6.0, 9.0, 3.0, 10.0, 10.0, 4.0], [10.0, 4.0, 3.0,
2.0, 3.0, 10.0, 5.0, 3.0, 2.0, 4.0], [5.0, 1.0, 3.0, 3.0, 2.0, 2.0, 2.0, 3.0,
1.0, 2.0], [3.0, 1.0, 1.0, 3.0, 1.0, 1.0, 3.0, 1.0, 1.0, 2.0], [2.0, 1.0, 1.0,
1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 5.0, 5.0,
1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [5.0, 1.0, 1.0, 1.0,
2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [3.0, 1.0, 1.0, 1.0, 2.0,
1.0, 2.0, 1.0, 1.0, 2.0], [5.0, 7.0, 7.0, 1.0, 5.0, 8.0, 3.0, 4.0,
1.0, 2.0], [10.0, 5.0, 8.0, 10.0, 3.0, 10.0, 5.0, 1.0, 3.0, 4.0], [5.0, 10.0,
10.0, 6.0, 10.0, 10.0, 6.0, 5.0, 4.0], [8.0, 8.0, 9.0, 4.0, 5.0, 10.0,
7.0, 8.0, 1.0, 4.0], [10.0, 4.0, 4.0, 10.0, 6.0, 10.0, 5.0, 5.0, 1.0, 4.0],
[7.0, 9.0, 4.0, 10.0, 10.0, 3.0, 5.0, 3.0, 3.0, 4.0], [5.0, 1.0, 4.0, 1.0, 2.
0, 1.0, 3.0, 2.0, 1.0, 2.0], [10.0, 10.0, 6.0, 3.0, 3.0, 10.0, 4.0, 3.0, 2.0,
4.0], [3.0, 3.0, 5.0, 2.0, 3.0, 10.0, 7.0, 1.0, 1.0, 4.0], [10.0, 8.0, 8.0, 2.
0, 3.0, 4.0, 8.0, 7.0, 8.0, 4.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 3.0,
1.0, 2.0], [8.0, 4.0, 7.0, 1.0, 3.0, 10.0, 3.0, 9.0, 2.0, 4.0], [5.0, 1.0, 1.0,
1.0, 2.0], [3.0, 3.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [3.0, 5.0, 2.0,
3.0, 10.0, 7.0, 1.0, 1.0, 4.0], [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0,
1.0, 4.0], [7.0, 2.0, 4.0, 1.0, 3.0, 4.0, 3.0, 1.0, 4.0], [3.0, 1.0, 1.0, 1.0,

2.0], [7.0, 8.0, 8.0, 7.0, 3.0, 10.0, 7.0, 2.0, 3.0, 4.0], [1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [1.0, 1.0, 3.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0], [1.0, 1.0, 3.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [3.0, 1.0, 1.0, 3.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [1.0, 2.0, 1.0, 1.0, 1.0, 2.0], [5.0, 2.0, 2.0, 2.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [5.0, 7.0, 4.0, 1.0, 6.0, 1.0, 7.0, 10.0, 3.0, 4.0], [5.0, 10.0, 10.0, 8.0, 5.0, 5.0, 7.0, 10.0, 1.0, 4.0], [3.0, 10.0, 7.0, 8.0, 5.0, 8.0, 7.0, 4.0, 1.0, 4.0], [3.0, 2.0, 1.0, 2.0, 2.0, 1.0, 3.0, 1.0, 1.0, 2.0], [2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 3.0], [1.0, 1.0, 2.0], [5.0, 3.0, 2.0, 1.0, 3.0, 1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [4.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0], [4.0, 1.0, 4.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [5.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 2.0], [1.0, 1.0, 2.0], [2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [10.0, 10.0, 5.0, 10.0, 10.0, 10.0, 7.0, 4.0], [5.0, 10.0, 10.0, 10.0, 10.0, 4.0, 10.0, 5.0, 6.0, 3.0, 4.0], [5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 2.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0], [4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [3.0, 1.0, 1.0, 1.0, 2.0], [1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 2.0, 1.0, 4.0, 1.0, 4.0], [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0], [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0]

Now, we will Random Shuffle the dataframe, split them in a fashion similar to the last time. We go for a test size of 0.2. The response variable holds a numerical value as well, with 2 representing benign tumors and 4 representing malignant.

We then verify the test set.

```
In [3]: random.shuffle(full_data)
test_size = 0.2
train_set = {2:[], 4:[]} #2 & 4 is output data
test_set = {2:[], 4:[]} #2 is for the benign tumors 4 is for malignant tumors

train_data = full_data[0:-int(test_size*len(full_data))]
test_data = full_data[-int(test_size*len(full_data)):]
```



```
for data in train_data:
    train_set[data[-1]].append(data[0:-1])
for data in test_data:
    test_set[data[-1]].append(data[0:-1])
test_set
```

```
Out[3]: {2: [[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 3.0, 6.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 3.0, 1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
 [4.0, 1.0, 1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
 [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
 [3.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 2.0, 1.0],  
 [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0],  
 [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
 [2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 2.0],  
 [3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
 [5.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
 [1.0, 2.0, 3.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
 [4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],
```

```
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[2.0, 3.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 1.0],  
[4.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[6.0, 3.0, 3.0, 5.0, 3.0, 10.0, 3.0, 5.0, 3.0],  
[3.0, 3.0, 2.0, 1.0, 3.0, 1.0, 3.0, 6.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 2.0, 3.0, 1.0, 1.0],  
[1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[4.0, 2.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 3.0, 2.0, 3.0, 1.0, 1.0, 1.0],  
[5.0, 1.0, 1.0, 6.0, 3.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 4.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 10.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 4.0, 5.0, 3.0, 7.0, 3.0, 4.0, 6.0, 1.0],  
[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[8.0, 2.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 1.0],  
[5.0, 1.0, 1.0, 1.0, 3.0, 2.0, 2.0, 2.0, 1.0],  
[1.0, 1.0, 1.0, 3.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 8.0],  
[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 3.0, 1.0, 2.0, -99999.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  
[2.0, 1.0, 1.0, 2.0, 3.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 5.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 3.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 2.0, 1.0],  
[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[3.0, 1.0, 3.0, 1.0, 3.0, 4.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 1.0, -99999.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[2.0, 2.0, 2.0, 1.0, 1.0, 1.0, 7.0, 1.0, 1.0],  
[6.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[6.0, 1.0, 3.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 2.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  
[3.0, 1.0, 1.0, 1.0, 2.0, 3.0, 3.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[2.0, 1.0, 1.0, 1.0, 3.0, 1.0, 2.0, 1.0, 1.0],  
[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 1.0, 1.0],  
[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[3.0, 1.0, 3.0, 1.0, 2.0, -99999.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 4.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0],  
[1.0, 1.0, 1.0, 1.0, 5.0, 1.0, 3.0, 1.0, 1.0],  
[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 6.0, 1.0],  
[5.0, 1.0, 1.0, 3.0, 2.0, 1.0, 1.0, 1.0, 1.0],  
[5.0, 2.0, 4.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0],
```

```
[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  

[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  

[5.0, 2.0, 2.0, 2.0, 3.0, 1.0, 1.0, 3.0, 1.0],  

[4.0, 1.0, 1.0, 1.0, 2.0, 3.0, 1.0, 1.0, 1.0],  

[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  

[2.0, 3.0, 1.0, 1.0, 3.0, 1.0, 1.0, 1.0, 1.0],  

[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  

[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0],  

[3.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0],  

[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0],  

[2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0],  

[1.0, 2.0, 3.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]],  

4: [[10.0, 6.0, 4.0, 3.0, 10.0, 10.0, 9.0, 10.0, 1.0],  

[5.0, 6.0, 5.0, 6.0, 10.0, 1.0, 3.0, 1.0, 1.0],  

[4.0, 6.0, 6.0, 5.0, 7.0, 6.0, 7.0, 7.0, 3.0],  

[10.0, 9.0, 7.0, 3.0, 4.0, 2.0, 7.0, 7.0, 1.0],  

[5.0, 8.0, 7.0, 7.0, 10.0, 10.0, 5.0, 7.0, 1.0],  

[7.0, 4.0, 5.0, 10.0, 2.0, 10.0, 3.0, 8.0, 2.0],  

[10.0, 6.0, 5.0, 8.0, 5.0, 10.0, 8.0, 6.0, 1.0],  

[10.0, 10.0, 10.0, 7.0, 10.0, 10.0, 8.0, 2.0, 1.0],  

[5.0, 10.0, 10.0, 10.0, 4.0, 10.0, 5.0, 6.0, 3.0],  

[3.0, 6.0, 4.0, 10.0, 3.0, 3.0, 3.0, 4.0, 1.0],  

[9.0, 5.0, 5.0, 2.0, 2.0, 2.0, 5.0, 1.0, 1.0],  

[3.0, 4.0, 5.0, 2.0, 6.0, 8.0, 4.0, 1.0, 1.0],  

[6.0, 10.0, 5.0, 5.0, 4.0, 10.0, 6.0, 10.0, 1.0],  

[6.0, 5.0, 4.0, 4.0, 3.0, 9.0, 7.0, 8.0, 3.0],  

[5.0, 10.0, 10.0, 8.0, 5.0, 5.0, 7.0, 10.0, 1.0],  

[6.0, 6.0, 7.0, 10.0, 3.0, 10.0, 8.0, 10.0, 2.0],  

[10.0, 6.0, 6.0, 3.0, 4.0, 5.0, 3.0, 6.0, 1.0],  

[6.0, 10.0, 10.0, 2.0, 8.0, 10.0, 7.0, 3.0, 3.0],  

[9.0, 10.0, 10.0, 1.0, 10.0, 8.0, 3.0, 3.0, 1.0],  

[10.0, 3.0, 4.0, 5.0, 3.0, 10.0, 4.0, 1.0, 1.0],  

[5.0, 10.0, 10.0, 3.0, 8.0, 1.0, 5.0, 10.0, 3.0],  

[7.0, 6.0, 4.0, 8.0, 10.0, 10.0, 9.0, 5.0, 3.0],  

[5.0, 4.0, 6.0, 7.0, 9.0, 7.0, 8.0, 10.0, 1.0],  

[10.0, 10.0, 10.0, 8.0, 6.0, 1.0, 8.0, 9.0, 1.0],  

[1.0, 5.0, 8.0, 6.0, 5.0, 8.0, 7.0, 10.0, 1.0],  

[10.0, 10.0, 10.0, 1.0, 6.0, 1.0, 2.0, 8.0, 1.0],  

[8.0, 8.0, 9.0, 4.0, 5.0, 10.0, 7.0, 8.0, 1.0],  

[10.0, 7.0, 7.0, 4.0, 5.0, 10.0, 5.0, 7.0, 2.0],  

[5.0, 3.0, 3.0, 3.0, 6.0, 10.0, 3.0, 1.0, 1.0],  

[8.0, 10.0, 10.0, 7.0, 10.0, 10.0, 7.0, 3.0, 8.0],  

[10.0, 8.0, 8.0, 4.0, 10.0, 10.0, 8.0, 1.0, 1.0],  

[9.0, 8.0, 8.0, 5.0, 6.0, 2.0, 4.0, 10.0, 4.0],  

[8.0, 7.0, 8.0, 5.0, 5.0, 10.0, 9.0, 10.0, 1.0],  

[2.0, 5.0, 7.0, 6.0, 4.0, 10.0, 7.0, 6.0, 1.0],  

[5.0, 10.0, 8.0, 10.0, 8.0, 10.0, 3.0, 6.0, 3.0],  

[8.0, 10.0, 10.0, 8.0, 6.0, 9.0, 3.0, 10.0, 10.0],  

[8.0, 10.0, 8.0, 8.0, 4.0, 8.0, 7.0, 7.0, 1.0],  

[8.0, 3.0, 8.0, 3.0, 4.0, 9.0, 8.0, 9.0, 8.0],  

[4.0, 1.0, 1.0, 3.0, 1.0, 5.0, 2.0, 1.0, 1.0],  

[7.0, 4.0, 7.0, 4.0, 3.0, 7.0, 7.0, 6.0, 1.0],  

[4.0, 5.0, 5.0, 8.0, 6.0, 10.0, 10.0, 7.0, 1.0],  

[10.0, 5.0, 10.0, 3.0, 5.0, 8.0, 7.0, 8.0, 3.0],  

[7.0, 5.0, 10.0, 10.0, 10.0, 10.0, 4.0, 10.0, 3.0],  

[9.0, 4.0, 5.0, 10.0, 6.0, 10.0, 4.0, 8.0, 1.0],  

[10.0, 8.0, 8.0, 2.0, 3.0, 4.0, 8.0, 7.0, 8.0],
```

Now we create a function that applies the K-Nearest Neighbors algorithm to our dataframe and run it on our training set, it outputs the number of clusters we have, in this case, Malignant and Benign.

```
In [4]: def knn(data, predict, k=3):
    distances = []
    for group in data:
        for features in data[group]:
            euclidean_distance = np.linalg.norm(np.array(features)-np.array(predict))
            distances.append([euclidean_distance,group])
    votes = [i[1] for i in sorted(distances)[:k]]

    vote_result = Counter(votes).most_common(1)[0][0]
    return vote_result # 2,4
new_pt = [4,1,4,1,2,1,1,1,1]

knn(train_set,new_pt,500)
```

Out[4]: 2

Now, we apply the algorithm on the test set and we check for a confusion matrix styled efficiency calculation.

```
In [5]: #test_set = {2:[ [7,8,8,7,3,10,7,2,3]] , 4:[[7,8,8,7,3,10,7,2,3]]}
correct=0
total =0

for i in test_set:#4
    for pt in test_set[i]:
        pr = knn(train_set,pt,200)
        if (pr == i):
            correct +=1
        total +=1
```

Now we print the predicted result and the actual diagnosis side by side with the features of the data set, the "vote" variable is the prediction and the "group" is the actual result. Here, we calculate and print accuracy, which is the no. of correct results divided by total.

```
In [6]: correct = 0
total=0
for group in test_set:
    for data in test_set[group]:
        print(group,end='->')
        print(data,end='->')
        vote = knn(train_set, data, k=6)
        print(vote)

        if group == vote:
            correct += 1
        total += 1
print('Accuracy:', (correct/total)*100)
```

```
2->[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[5.0, 3.0, 6.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[5.0, 3.0, 1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 1.0]->2
2->[4.0, 1.0, 1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 1.0]->2
2->[3.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0]->2
2->[3.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]->2
2->[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 2.0, 1.0]->2
2->[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0]->2
2->[3.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 1.0]->2
```


2->[1.0, 1.0, 4.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0]->2
2->[1.0, 2.0, 2.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0]->2
2->[1.0, 1.0, 1.0, 1.0, 5.0, 1.0, 3.0, 1.0, 1.0]->2
2->[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 6.0, 1.0]->2
2->[5.0, 1.0, 1.0, 3.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[5.0, 2.0, 4.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]->2
2->[2.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0]->2
2->[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0]->2
2->[5.0, 2.0, 2.0, 2.0, 3.0, 1.0, 1.0, 3.0, 1.0]->2
2->[4.0, 1.0, 1.0, 1.0, 2.0, 3.0, 1.0, 1.0, 1.0]->2
2->[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[2.0, 3.0, 1.0, 1.0, 3.0, 1.0, 1.0, 1.0, 1.0]->2
2->[4.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
2->[3.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0]->2
2->[5.0, 1.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 1.0]->2
2->[2.0, 1.0, 1.0, 1.0, 1.0, 3.0, 1.0, 1.0, 1.0]->2
2->[1.0, 2.0, 3.0, 1.0, 2.0, 1.0, 1.0, 1.0, 1.0]->2
4->[10.0, 6.0, 4.0, 3.0, 10.0, 10.0, 9.0, 10.0, 1.0]->4
4->[5.0, 6.0, 5.0, 6.0, 10.0, 1.0, 3.0, 1.0, 1.0]->4
4->[4.0, 6.0, 6.0, 5.0, 7.0, 6.0, 7.0, 7.0, 3.0]->4
4->[10.0, 9.0, 7.0, 3.0, 4.0, 2.0, 7.0, 7.0, 1.0]->4
4->[5.0, 8.0, 7.0, 7.0, 10.0, 10.0, 5.0, 7.0, 1.0]->4
4->[7.0, 4.0, 5.0, 10.0, 2.0, 10.0, 3.0, 8.0, 2.0]->4
4->[10.0, 6.0, 5.0, 8.0, 5.0, 10.0, 8.0, 6.0, 1.0]->4
4->[10.0, 10.0, 10.0, 7.0, 10.0, 10.0, 8.0, 2.0, 1.0]->4
4->[5.0, 10.0, 10.0, 10.0, 4.0, 10.0, 5.0, 6.0, 3.0]->4
4->[3.0, 6.0, 4.0, 10.0, 3.0, 3.0, 4.0, 1.0]->4
4->[9.0, 5.0, 5.0, 2.0, 2.0, 2.0, 5.0, 1.0, 1.0]->2
4->[3.0, 4.0, 5.0, 2.0, 6.0, 8.0, 4.0, 1.0, 1.0]->4
4->[6.0, 10.0, 5.0, 5.0, 4.0, 10.0, 6.0, 10.0, 1.0]->4
4->[6.0, 5.0, 4.0, 4.0, 3.0, 9.0, 7.0, 8.0, 3.0]->4
4->[5.0, 10.0, 10.0, 8.0, 5.0, 5.0, 7.0, 10.0, 1.0]->4
4->[6.0, 6.0, 7.0, 10.0, 3.0, 10.0, 8.0, 10.0, 2.0]->4
4->[10.0, 6.0, 6.0, 3.0, 4.0, 5.0, 3.0, 6.0, 1.0]->4
4->[6.0, 10.0, 10.0, 2.0, 8.0, 10.0, 7.0, 3.0, 3.0]->4
4->[9.0, 10.0, 10.0, 1.0, 10.0, 8.0, 3.0, 3.0, 1.0]->4
4->[10.0, 3.0, 4.0, 5.0, 3.0, 10.0, 4.0, 1.0, 1.0]->4
4->[5.0, 10.0, 10.0, 3.0, 8.0, 1.0, 5.0, 10.0, 3.0]->4
4->[7.0, 6.0, 4.0, 8.0, 10.0, 10.0, 9.0, 5.0, 3.0]->4
4->[5.0, 4.0, 6.0, 7.0, 9.0, 7.0, 8.0, 10.0, 1.0]->4
4->[10.0, 10.0, 10.0, 8.0, 6.0, 1.0, 8.0, 9.0, 1.0]->4
4->[1.0, 5.0, 8.0, 6.0, 5.0, 8.0, 7.0, 10.0, 1.0]->4
4->[10.0, 10.0, 10.0, 1.0, 6.0, 1.0, 2.0, 8.0, 1.0]->4
4->[8.0, 8.0, 9.0, 4.0, 5.0, 10.0, 7.0, 8.0, 1.0]->4
4->[10.0, 7.0, 7.0, 4.0, 5.0, 10.0, 5.0, 7.0, 2.0]->4
4->[5.0, 3.0, 3.0, 3.0, 6.0, 10.0, 3.0, 1.0, 1.0]->4
4->[8.0, 10.0, 10.0, 7.0, 10.0, 10.0, 7.0, 3.0, 8.0]->4
4->[10.0, 8.0, 8.0, 4.0, 10.0, 10.0, 8.0, 1.0, 1.0]->4
4->[9.0, 8.0, 8.0, 5.0, 6.0, 2.0, 4.0, 10.0, 4.0]->4
4->[8.0, 7.0, 8.0, 5.0, 5.0, 10.0, 9.0, 10.0, 1.0]->4
4->[2.0, 5.0, 7.0, 6.0, 4.0, 10.0, 7.0, 6.0, 1.0]->4
4->[5.0, 10.0, 8.0, 10.0, 8.0, 10.0, 3.0, 6.0, 3.0]->4
4->[8.0, 10.0, 10.0, 8.0, 6.0, 9.0, 3.0, 10.0, 10.0]->4
4->[8.0, 10.0, 8.0, 8.0, 4.0, 8.0, 7.0, 7.0, 1.0]->4
4->[8.0, 3.0, 8.0, 3.0, 4.0, 9.0, 8.0, 9.0, 8.0]->4
4->[4.0, 1.0, 1.0, 3.0, 1.0, 5.0, 2.0, 1.0, 1.0]->2
4->[7.0, 4.0, 7.0, 4.0, 3.0, 7.0, 7.0, 6.0, 1.0]->4
4->[4.0, 5.0, 5.0, 8.0, 6.0, 10.0, 10.0, 7.0, 1.0]->4
4->[10.0, 5.0, 10.0, 3.0, 5.0, 8.0, 7.0, 8.0, 3.0]->4
4->[7.0, 5.0, 10.0, 10.0, 10.0, 10.0, 4.0, 10.0, 3.0]->4
4->[9.0, 4.0, 5.0, 10.0, 6.0, 10.0, 4.0, 8.0, 1.0]->4
4->[10.0, 8.0, 8.0, 2.0, 3.0, 4.0, 8.0, 7.0, 8.0]->4
4->[3.0, 10.0, 8.0, 7.0, 6.0, 9.0, 9.0, 3.0, 8.0]->4
-

Thus we can conclude that KNN produces an accuracy of 97.8%