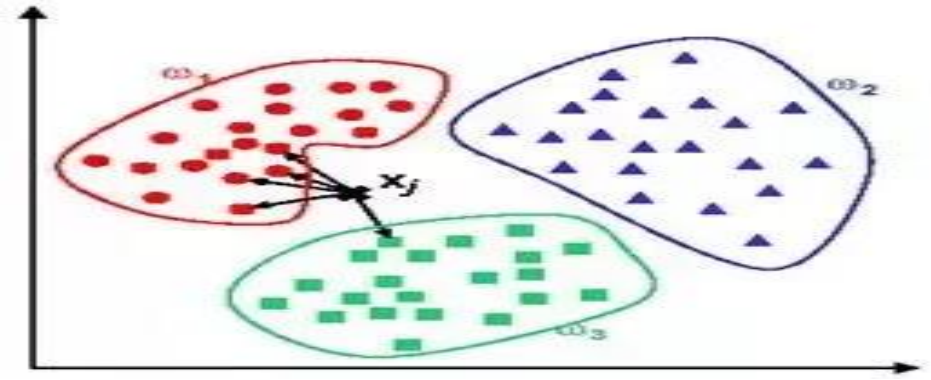
**DATA MINING**

**KNN REPORT**

## 1. Describe Nearest Neighbors method

In a KNN algorithm, a test sample is given as the class of majority of its nearest neighbours. In plain words, if you are similar to your neighbours, then you are one of them. Or if apple looks more similar to banana, orange, and melon (fruits) than monkey, cat and rat (animals), then most likely apple is a fruit. Below is an example, we have three classes and the goal is to find a class label for the unknown example $xjxj$. In this cas,e we use the Euclidean distance and a value of $k$=5 neighbors. Of the 5 closest neighbors, 4 belong to $\omega 1\omega 1$ and 1 belongs to $\omega 3\omega 3$, so $xjxj$ is assigned to $\omega 1\omega 1$, the predominant class. The choice of $k$ is a hyper-parameter that can be tuned or set heuristically.




## 2. Explain what was your criteria for selecting the three attributes

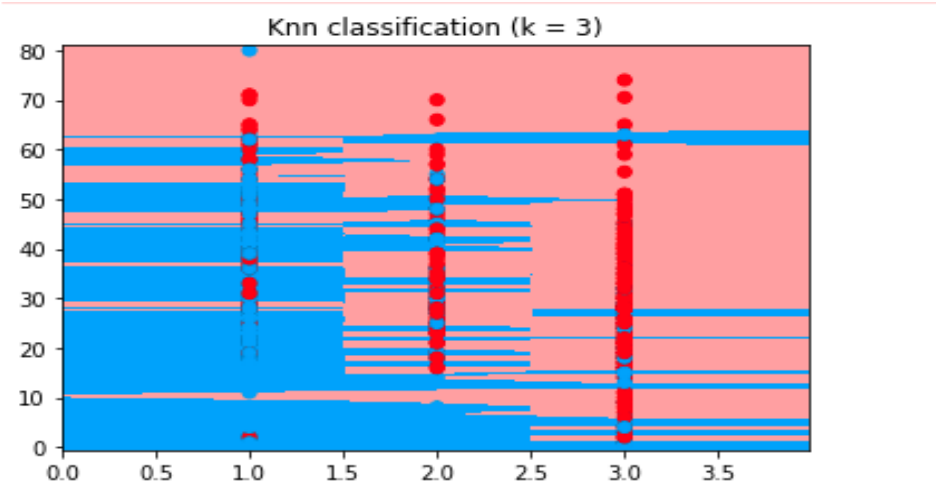
I am not interested in the following features, as I believe they do not hold any logical reason for why the passenger will survive or not

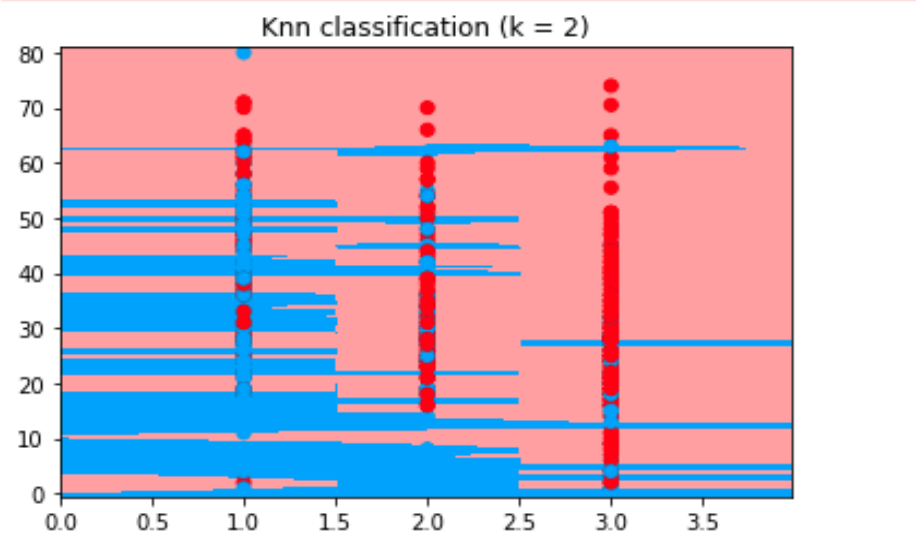
1. Name
2. Fare
3. Ticket
4. Cabin

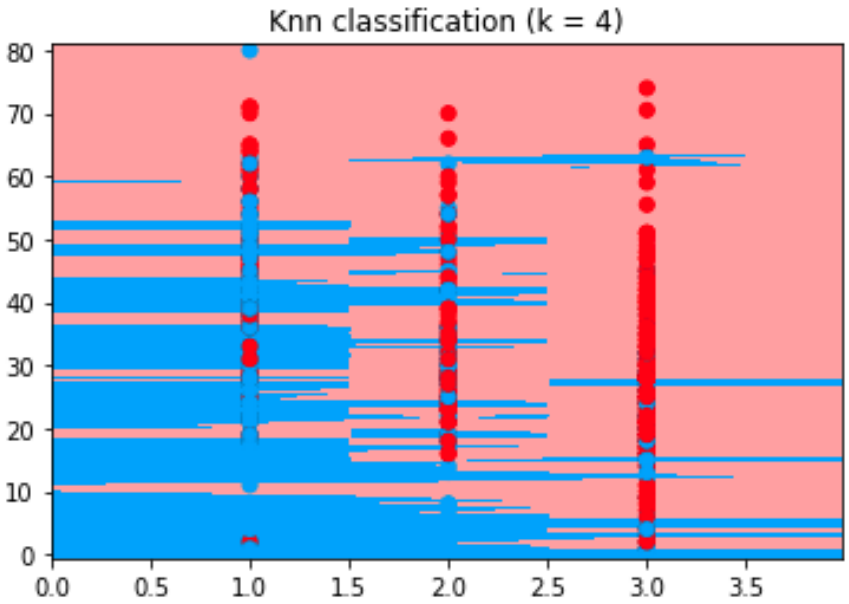
So drop them.
Added parents and siblings column to form 'family' column.
Replace the null values in all the columns to the median value.
Using selectKbest feature of scikit learn, to check the score of the all the attributes in the data set.
Select the best 3 from all of them.
The highest scores were of Sex, pclass and age.

## 3. Visualizations of the classifier in a 2D projection, for all three different number of neighbors.

Here there are 2 different background colors each represents 2 different class labels, represented as 0, 1.
Pink background color is for "0" or 'dead'
Blue background color is for "1" or 'survived'






## 4. Interpret and compare the results

- For k=4:

```
K value: 4
Accuracy: 0.7611940298507462
Confusion Matrix:
 [[160  12]
 [ 52  44]]
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.93      0.83       172
           1       0.79      0.46      0.58        96

    accuracy                           0.76       268
   macro avg       0.77      0.69      0.71       268
weighted avg       0.77      0.76      0.74       268
```

Here, column represent predicted values and rows represent actual values, which means in the 1st row class 0 are of 172 values and among which 160 are correctly classified and 12 are wrongly classified as class 0. In the 2nd row class 1 are of total 96 values among which 52 are correctly
classified and 44 are wrongly classified as class 1.

Accuracy of knn when k=4 is 0.7611940298507462

- For k=2:

```
K value: 2
Accuracy: 0.75
Confusion Matrix:
 [[153  19]
 [ 48  48]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.89      0.82       172
           1       0.72      0.50      0.59        96

    accuracy                           0.75       268
   macro avg       0.74      0.69      0.70       268
weighted avg       0.75      0.75      0.74       268
```

Here, column represent predicted values and rows represent actual values, which means in the 1st row class 0 are of 172 values and among which 153 are correctly classified and 19 are wrongly classified as class 0. In the 2nd row class 1 are of total 96 values among which 48 are correctly
classified and 48 are wrongly classified as class 1.

Accuracy of knn when k=2 is 0.75

- For k=3:

```
K value: 3
Accuracy: 0.7425373134328358
Confusion Matrix:
 [[141  31]
 [ 38  58]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.82      0.80       172
           1       0.65      0.60      0.63        96

    accuracy                           0.74       268
   macro avg       0.72      0.71      0.72       268
weighted avg       0.74      0.74      0.74       268
```

Here, column represent predicted values and rows represent actual values, which means in the 1st row class 0 are of 172 values and among which 141 are correctly classified and 31 are wrongly classified as class 0. In the 2nd row class 1 are of total 96 values among which 38 are correctly
classified and 58 are wrongly classified as class 1.

Accuracy of knn when k=3 is 0.7425373134328358

When we observe the accuracies of three models k=4 has more accuracy of 0.7611940298507462 which
means in this case more number of predicted values are same as actual values than of other models