a) selected KNN & neural network to build my classifier.

b) Excluded the attributes (passenger ID, Embarked, Name, Ticket, Cable) as most of them have low correlation and confident level to survival. And rest of them having too many missing values. Excluded these attributes to prevent noise.

For the missing data in age, used multivariable linear regression to predict the missing value to fill in the blank, as there are not sufficient data that allow deleting the missing entries.

c) KNN is a very simple model that give me a sense of accuracy that I should reach. Neural Network is a very robust model that can understand the problem in a deep level through the layers. I Split the training dataset to training set and validation set with ratio 7:3. Optimized the hyperparameter by fitting the validation set. I used fully connected neural network with 6-hidden layer and 100 neural per layer.

Normalized the feature before feeding into training, mean=0 & variance=1. If all the feature inputs are always positive, it make the gradient of feature inputs all positive or all negative, result in updating all in the same direct thus inefficient updating. Besides, controlled variance among feature. We can speed up gradient descent by having each of our input values in roughly the same range. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

It controlled the activation of neurons along with weight initialization. As the deep networks seem overfitting the training set nor the validation set, after fine tuning I set the regulation a=1, resulted in a similar accuracy between training and validation set. I used 5-Nearest Neighbor after tuning against validation set, it gave the best accuracy and F1 value.

d) Accuracy: KNN - 86% Neural Network - 85%.

e) Both KNN & Neural Network gave similar accuracy on training set and validation set, it is because the complicity of the problem is not high. Neural network robust on complicated system and relation between feature. As, the problem is relatively simple, thus result in similar accuracy across model. But Neural network give better accuracy in test dataset most of the time, it seems it has a relatively higher understanding of the problem, but it’s only >5%. To give a reasonable and time efficient solution, KNN seems a better model as it is easy and fast to build and give a reasonable accuracy on prediction.

f) Survivor: 127

(i) female: 112

(ii) below 18: 23(missing value not counted)

(iii) without any other family members on board: 49

(iv) 1st: 54/107

2nd: 33/93

3rd: 40/436 -> the least chance of surviving the tragedy

(v ) Cherbourg: 41/102

Queenstown: 9/46

Southampton: 77/541 -> the least chance of surviving the tragedy

(g) accuracy: KNN – 72.22% Neural Network – 77.9%

(h) Team Name: kenneth319

Rank: 5184

Score: 0.77990

(i) both models have better accuracy from training set. Because the model is overfitting the training set and result in lower prediction accuracy in new dataset. Neural network has better generalization capability as it can apply to general prediction problem both simple and complex, with the same architecture, while KNN can only handle simple system.