# Lab KNN and Naïve Bayes

Please submit your answers in this word document, including detailed discussion for each question with supportive screenshots of your analysis in software.

On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

The Titanic Passengers.xlsx file contains data for 1,309 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived (nominal or binary), passenger class (ordinal), their sex (binary), their age (interval), siblings and spouses (interval), parents and children (interval), the fare they paid (interval), the port (nominal), and home/destination (nominal). Please note that this data set is already partitioned with a validation column as well. The data table does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

In this assignment, “Survived” is your dependent variable while all the other variables except name are possible explanatory/input variables. Create a file titled “YourLastName KNN and Naïve Bayes Homework”.

## KNN

1. Develop a diagram that evaluates *k* values from 5 to 15, based on misclassification rate. What is the optimal *k* value and what is the misclassification rate? Provide a screen shot of the Model Selection based on Valid: Misclassification Rate to support your answer.

**Answer:**

The below screenshot shows the diagram that evaluates k values from 5 to 15.

From the screenshot we can clearly see that for k as 7 we get the least misclassification value. Hence it is the optimal k value for the given data.

Encoding, null filtering and appropriate data conversion is done before processing the data to get optimal results.

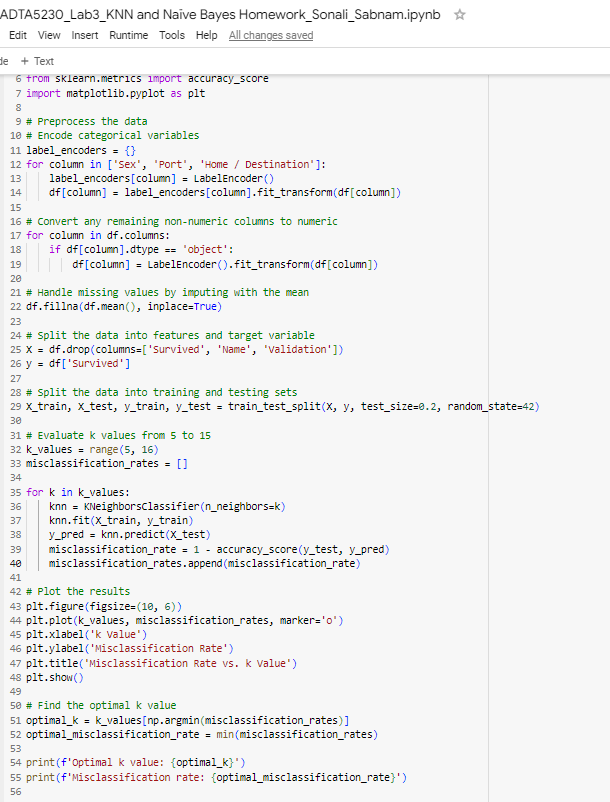
**Optimal k value: 7**

**Misclassification rate: 0.316793893129771**

A graph with blue dots

Description automatically generated

The below screenshot shows the code used to identify the K neighbor value for the data.



1. Using the validation results for the selected model, which category is your model more likely to classify correctly (survived or did not survive)? Provide a screen shot of the information from the output that you used as well as any calculations that you performed to justify your answer.

**Answer:**

The confusion matrix can be described as:

* **123**: True Negatives (correctly predicted “Not Survived”)
* **21**: False Positives (incorrectly predicted “Survived”)
* **62**: False Negatives (incorrectly predicted “Not Survived”)
* **56**: True Positives (correctly predicted “Survived”)

The below formula was used o calculate the recall value

A black text on a white background

Description automatically generated

* Recall for “Not Survived” (0): 0.85
* Recall for “Survived” (1): 0.47

We can see that the “recall” for “Not Survived” is 0.85 which is much higher than recall for “Survived” which is 0.47. This means that the category “**Not Survived**” is more likely to be classified **correctly**.

The below screenshot shows the confusion matrix and classification report.

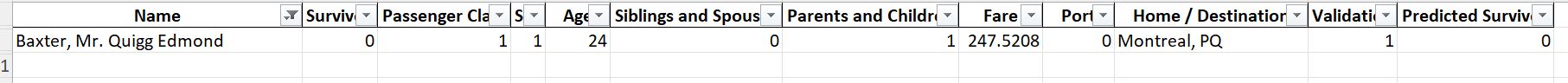
**A screenshot of a computer program

Description automatically generated**

1. Save the predicted values for the training and validation sets. Was Baxter, Mr. Quigg Edmond from the Validate set predicted to survive? Note, although there is no new data to score, we can use Score node to obtain predictions for training and validation data only. Please submit the Excel sheet containing predictions for the validation set. (Because there is no new data to score, you can simply connect the selected model node to a ‘Save Data’ node and select Excel in Output Format in SAS EM. The column ‘Into: Survived’ in this output excel shows the predicted survival.)

**Answer:** Baxter, Mr. Quigg Edmond from the Validate set was not predicted to survive.

The below screenshot shows from the validation output excel shows the same.

****

**Note:** The validation output will be submitted as part of the assignment.

1. Now, only evaluate passengers from Port “S”. Based on this selection of data, what was the optimal *k* value (from 5 to 15) and its misclassification rate on validation data? How does this misclassification rate compare to the optimal model prepared using the observations from all ports (better, worse, or the same)?

**Answer:** The below screenshot shows the misclassification rate and the optimal k value for passengers from port “S”.

**Optimal k value: 12**

**Misclassification rate: 0.3079847908745247**

A graph with blue lines and dots

Description automatically generated

**Comparison:** We can see that the optimal k value is increased from 7 to 12 but the misclassification rate has decreased from 0.3167 to 0.3079. The lower the misclassification rate the better the model. Hence this model with k as 12 and misclassification rate as 0.3079 is a better model.