# 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.

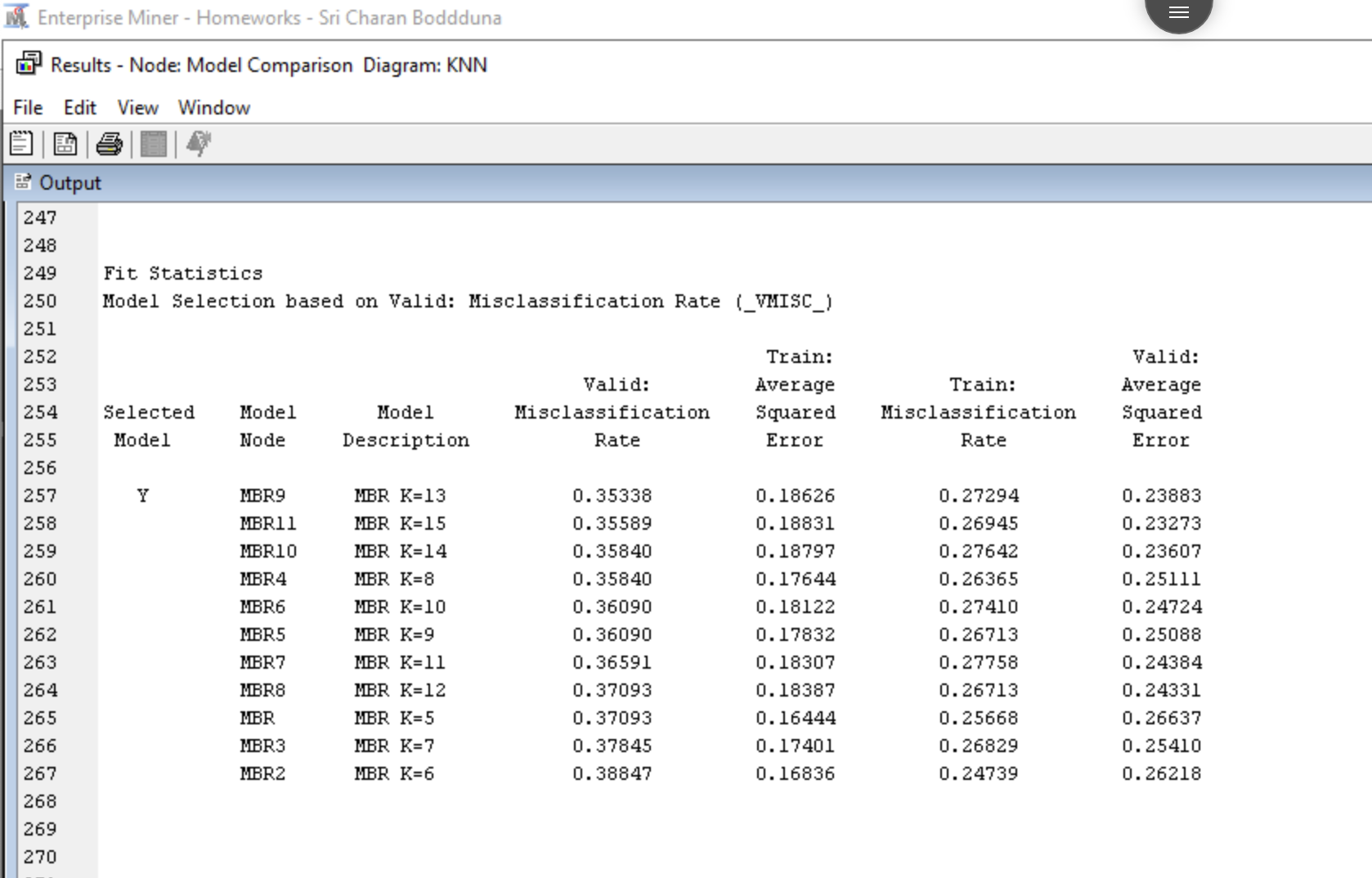
The Titanic Passengers.xlsx file contains data for 1,309 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived (nominal or binary), passenger class (ordinal), their sex (binary), their age (interval), siblings and spouses (interval), parents and children (interval), the fare they paid (interval), the port (nominal), and home/destination (nominal). The Validation column is the indicator of the validation observations and so it should be “Rejected”. The Name is the “ID” and the Survived is the “Target”. The data table does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

## KNN

1. Evaluates *k* values from 5 to 15, based on misclassification rate of validation data. What is the optimal *k* value and what is the misclassification rate? Provide a screen shot of the Model Selection based on validation Misclassification Rate to support your answer.

After building the KNN model with *k* equal to 5 to 15. I got the optimal *k* value of 13 which resulted validation data misclassification rate as 0.353.

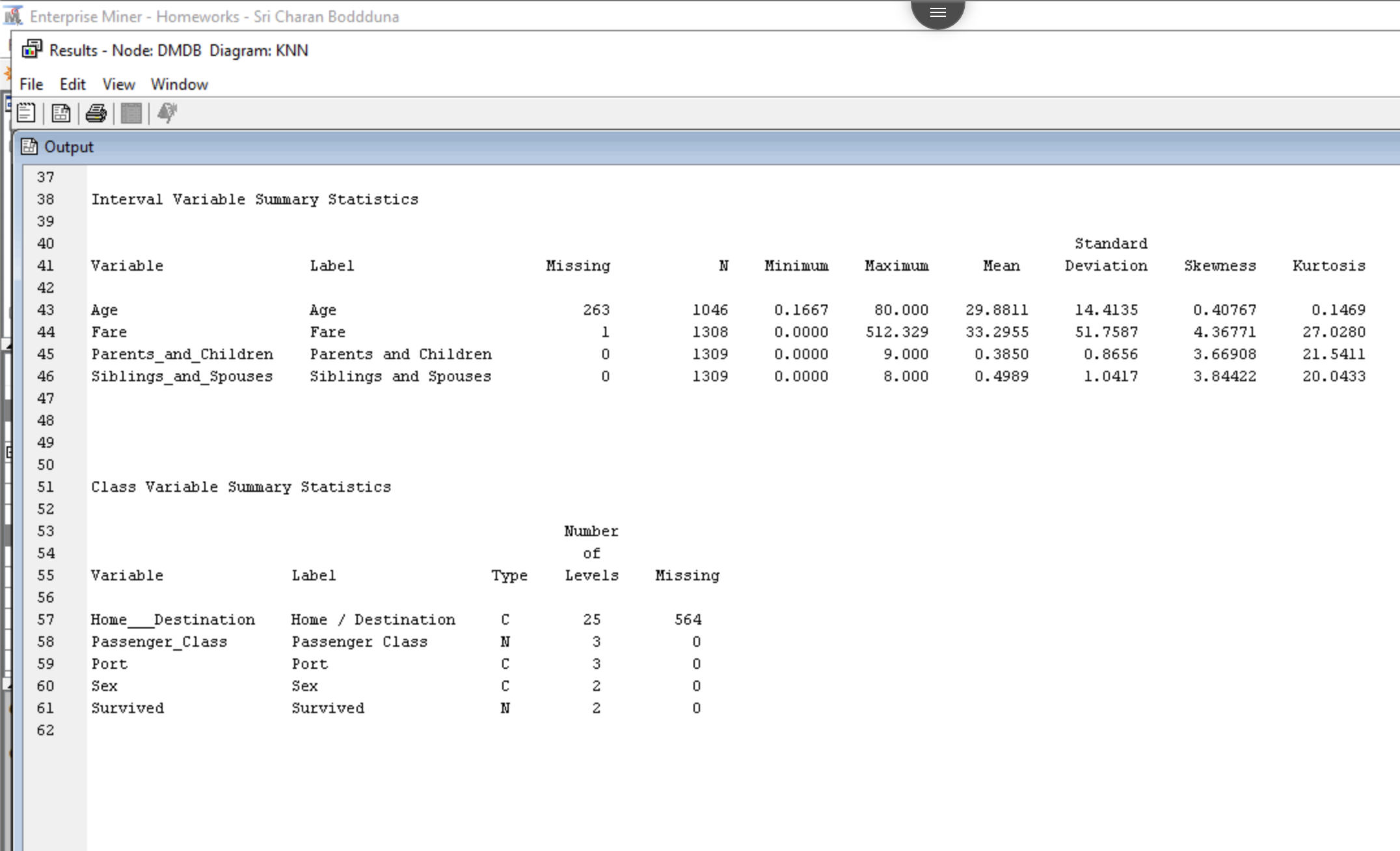
My model is built after removing null values from the columns.





*Removing null values in specific columns:*

We can see Age, Fare have null values.





Using impute node, I have replaced the null values with mean value of the respective columns. Then I developed model.

A screenshot of a computer

Description automatically generated



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.

In the below screenshot showing validation dataset classification results, we can see there classifcations.

Out of total predicted: Survived = 102

Correctly predicted: Survived = 63

TP / (TP+FP)

* 63/(63+39) = 61%

Out of total predicted: Not Survived = 297

Correctly predicted: Not Survived = 195

TN / (TN+FN)

* 195/(195+102) = 65.65%

Which means my model have predicted “Not Survived” more correctly.

A screenshot of a computer

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.)

In the below screenshot taking from the results of Score node. We can see that Baxter, Mr. Quigg Edmond have been classified as did not survive. However, the data shows survived is 0 (did not survived in real).

A screenshot of a computer

Description automatically generated



1. Create a filter to 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)?

A screenshot of a computer

Description automatically generated

After applying the filter to allow only passengers from Port “S”, I got the optimal *k* value as 15 with the misclassification rate of 0.363 (increased by 0.1). Both the evaluation parameters have been increased when compared to unfiltered results.

In my case, as the results are nearer similar hence, I can conclude that the new model have performed like the old model.

A screenshot of a computer

Description automatically generated



## Naïve Bayes

1. Continuing with the Titanic Passengers data (without filtering), you are going to build a Naïve Bayes classifier to predict if a passenger survived. What is the misclassification rate of the Naive Bayes Classifier on the validation data for predicting if a passenger survived? Provide a screen shot of your output to support your answer. How does that compare to the misclassification rate of the *k*NN model?

After building the model using Naïve bayes algorithm, I got validation misclassification rate as 0.233. which is lesser than the KNN model (0.35).

A screenshot of a computer

Description automatically generated



1. Using the validation results of the naïve bayes classifier, 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.

Out of total predicted: Survived = 124

Correctly predicted: Survived = 98

TP / (TP+FP)

* 98/(98+26) = 79%

Out of total predicted: Not Survived = 275

Correctly predicted: Not Survived = 208

TN / (TN+FN)

* 208/(67+208) = 75.63%

Which means my model have predicted “Survived” more correctly.

A screenshot of a computer

Description automatically generated



1. Prepare a second Naïve Bayes model without the predictor Sex. Compare the two Naïve Bayes models (with and without Sex) using ROC as the selection statistic on the validation data. Which model performs better? Provide a screen shot of your output to support your answer.

We can see below screenshot that shows better model when compared 2 naïve bayes models based on ROC index of validation data with ROC of 0.82 (82%), model with Sex have performed well in my scenario.

A screenshot of a computer

Description automatically generated

