# Task 1

Q1-2

Win is encoded as 1 and Lose is encoded as 0.

**Naïve Bayes**

|  |  |
| --- | --- |
| Accuracy | 58.3 % |
| Precision | 75% |
| Recall | 66.7% |
| F1 | 70.6% |
| Outcome | |
| Ground Truth | Predicted |
| 1 | 0 |
| 0 | 0 |
| 1 | 1 |
| 1 | 1 |
| 1 | 1 |
| 1 | 0 |
| 1 | 1 |
| 1 | 1 |
| 1 | 0 |
| 0 | 1 |
| 1 | 1 |
| 0 | 1 |

**KNN**

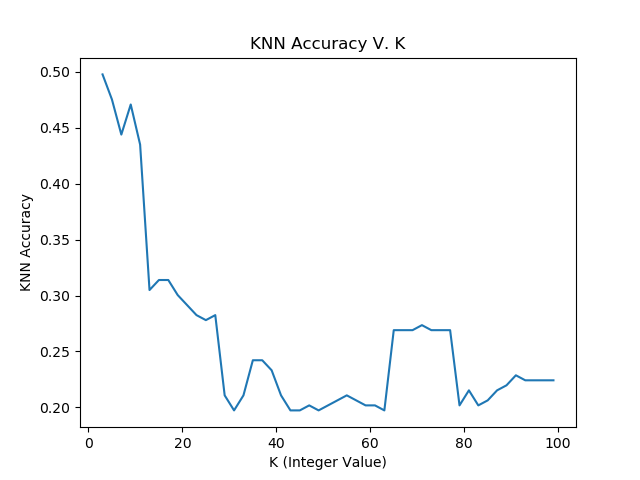
|  |  |
| --- | --- |
| Accuracy | 25.0% |
| Precision | 50.0% |
| Recall | 11.1% |
| F1 | 18.2% |
| Outcome | |
| Ground Truth | Predicted |
| 1 | 1 |
| 0 | 1 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 0 | 0 |
| 1 | 0 |
| 0 | 0 |

# Task 2

**Naïve Bayes**

|  |  |
| --- | --- |
| Accuracy | 80.3% |
| Precision | 75% |
| Recall | 69.5% |
| F1 | 72.2% |

Larger datasets increases all the metric values compared to smaller datasets. This is probably due to larger volume of test data allows for more fine-tuning of probabilistic distribution of P(feature | class)



# Task 3

For smaller datasets, the naïve bayes model works better than KNN. This is due to that with smaller datasets there is less “neighbors” to select from when doing the KNN algorithm thus can quickly shift the classification due to more of the “incorrect” classifier being present.

For larger datasets, naïve bayes also performed better. However, I believe that this is just because of the distance implementation of for this dataset. The values of certain features can take the value of the upper bound of integers. This is a scaling issue with how I am representing this data.

Additional Questions:

* Approximately how many hours did you spend on this assignment?

8 hours

* Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?

Figuring out how to get the titanic data to work. The testing data does not have labels so the training data had to be split up to be used for testing.

* Which aspects of this assignment did you like? Is there anything you would have changed?

I enjoy having to implement these algorithms, especially when making them general so that any dataset could be imported.

If I had to change anything, I would have given some information about how the titanic data should be pre-processed just to show how different types of data pre-processing can alter the outcome of these algorithms.