MSE 491 Assignment 2: Classification

By Tony Nguyen - 301286979

(Group 17)

Prepared for:

Mohammed Narimani, PhD, P.Eng
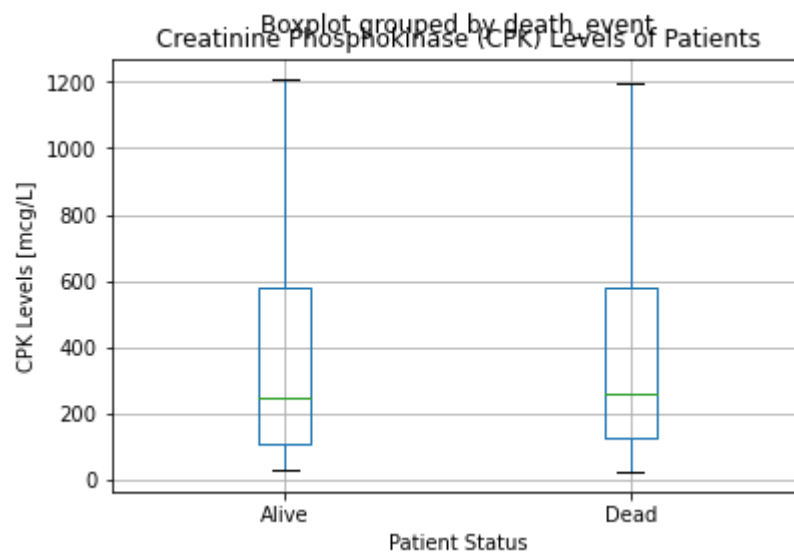
School of Mechatronic Systems Engineering

Simon Fraser University - Surrey Campus

1 Introduction

2 Results and Discussion

2.1 Getting Started

1. Percent of deceased patients that had anaemia and were smokers: 12.5%
2.

    a.

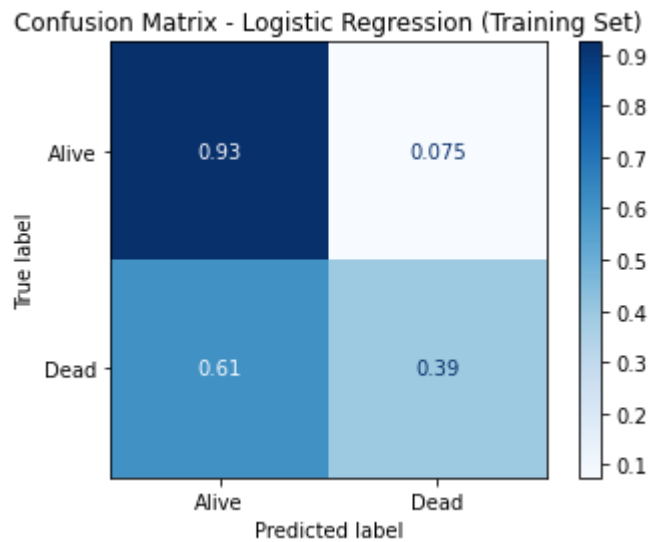   b. p-value = 0.27961124573614, therefore not significant

2.2 Logistic Regression

Training Set

Accuracy: 0.7489539748953975
Precision: 0.7209302325581395
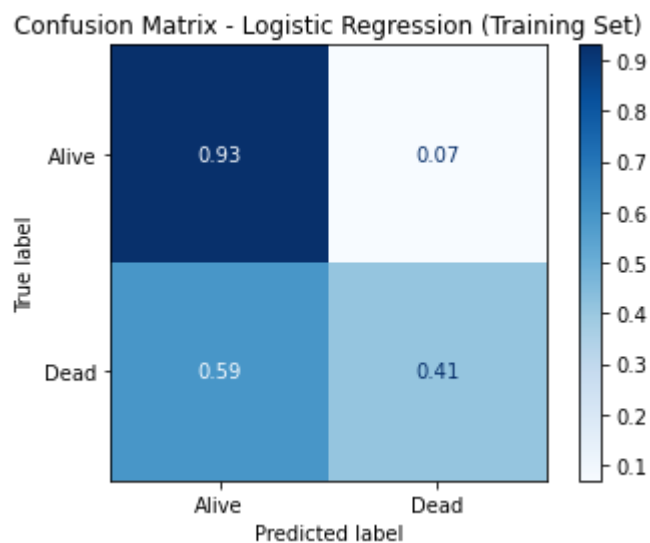Recall: 0.3924050632911392
F1 Score: 0.5081967213114754

Confusion Matrix - Logistic Regression (Training Set)

Testing Set

Accuracy: 0.7833333333333333

Precision: 0.7

Recall: 0.4117647058823529

F1 Score: 0.5185185185185185

Confusion Matrix - Logistic Regression (Training Set)

3) I would, personally, not use this model in a real world application. The large inaccuracies, and relatively large false positives this model predicted makes me daily hesitant, especially if this data is meant to be forwarded to the patient.
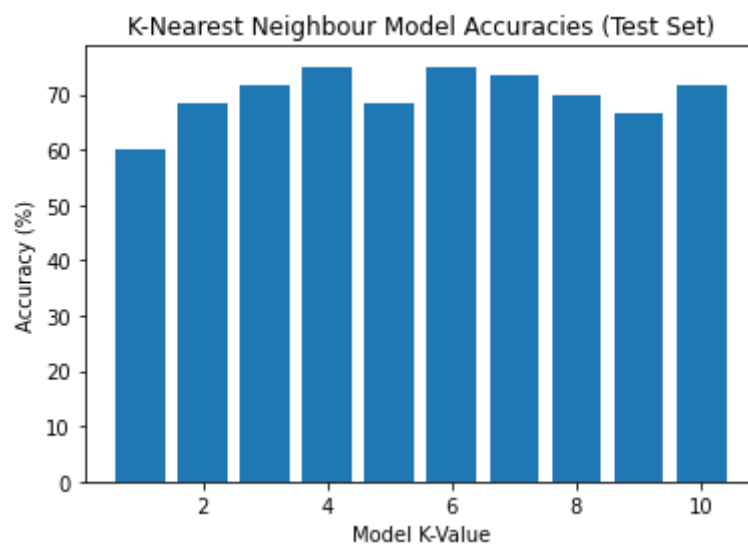
2.3 K-Nearest Neighbor

For performance metrics and confusion matrices for all the KNN models, please refer to the code.

Training Set
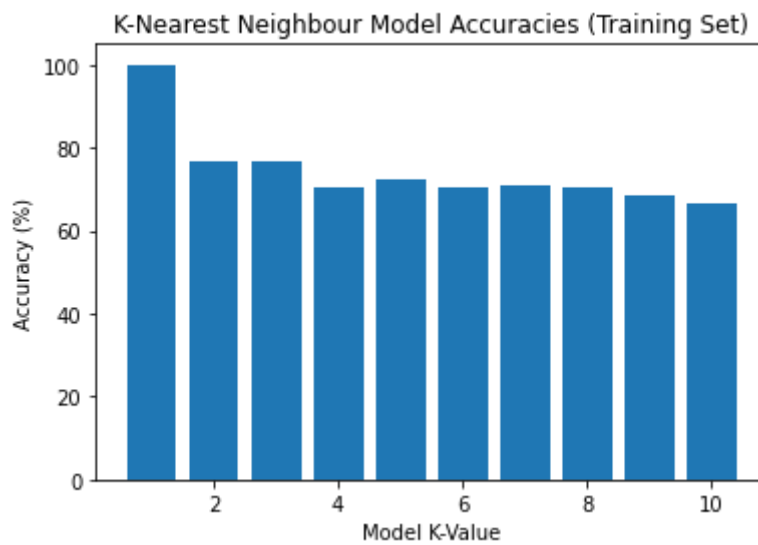
From the graph above, the model where k = 1 was the most accurate; this makes sense as the inputted training data would perfectly

Testing Set

2.4 Feature Extraction

2.5 Multiclass Evaluation

KNN

```
Classification Report (K-Nearest Neighbour)
              precision    recall  f1-score   support

         APC       0.92      0.92      0.92       410
        LBBB       0.94      0.95      0.94       414
      NORMAL       0.90      0.91      0.91       381
         PVC       0.93      0.92      0.93       393
        RBBB       0.96      0.95      0.96       402

    accuracy                          0.93      2000
   macro avg       0.93      0.93      0.93      2000
weighted avg       0.93      0.93      0.93      2000
```


Normalized Confusion Matrix (K-Nearest Neighbour)
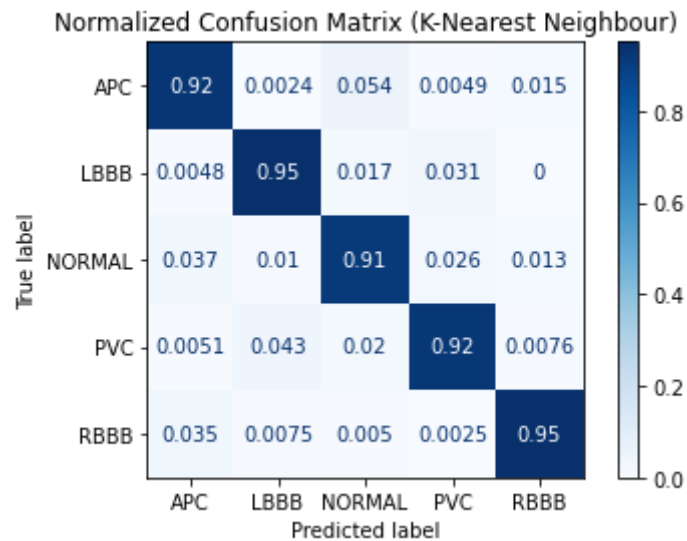
Decision Tree

```
Classification Report (Decision Tree)
              precision    recall  f1-score   support

         APC       0.84      0.87      0.85       410
        LBBB       0.93      0.89      0.91       414
      NORMAL       0.82      0.83      0.83       381
         PVC       0.88      0.91      0.89       393
        RBBB       0.95      0.92      0.93       402

    accuracy                          0.88      2000
   macro avg       0.88      0.88      0.88      2000
weighted avg       0.89      0.88      0.88      2000
```


Normalized Confusion Matrix (Decision Tree)
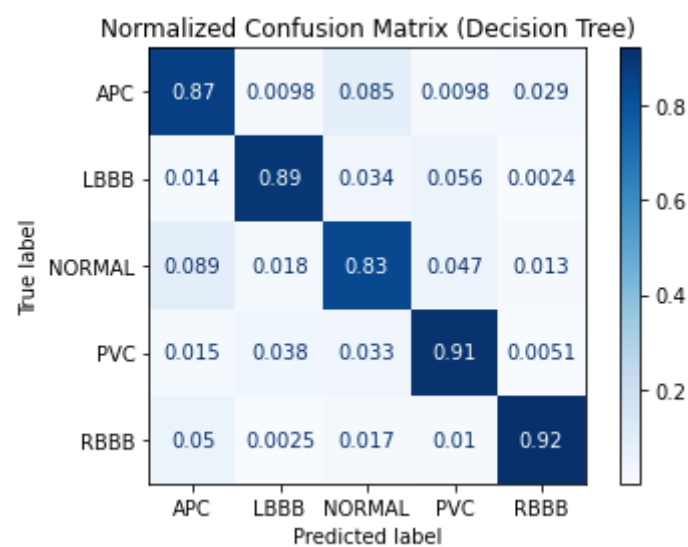
```
Classification Report (Gaussian Naive Bayes)
              precision    recall  f1-score   support

         APC       0.45      0.23      0.30       410
        LBBB       0.75      0.51      0.60       414
      NORMAL       0.61      0.60      0.60       381
         PVC       0.62      0.78      0.69       393
        RBBB       0.57      0.91      0.70       402

    accuracy                          0.60      2000
   macro avg       0.60      0.60      0.58      2000
weighted avg       0.60      0.60      0.58      2000
```
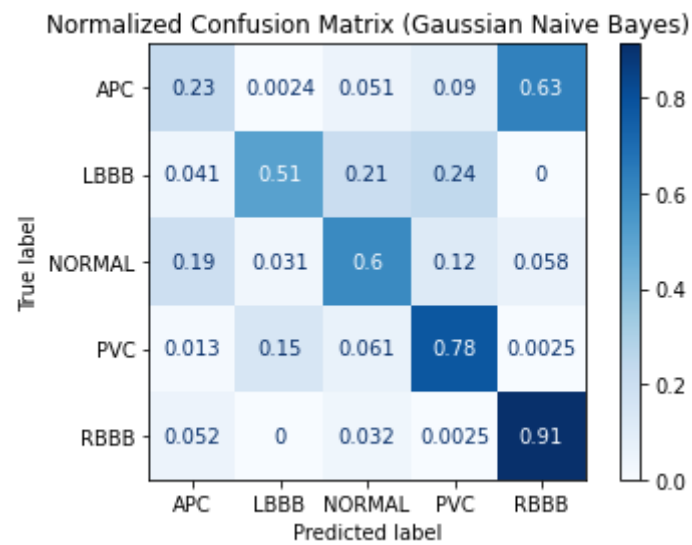
Gaussian Naive Bayes

Normalized Confusion Matrix (Gaussian Naive Bayes)

True label	APC	LBBB	NORMAL	PVC	RBBB
APC	0.23	0.0024	0.051	0.09	0.63
LBBB	0.041	0.51	0.21	0.24	0
NORMAL	0.19	0.031	0.6	0.12	0.058
PVC	0.013	0.15	0.061	0.78	0.0025
RBBB	0.052	0	0.032	0.0025	0.91

Predicted label

Support Vector Machine

```
Classification Report (Support Vector Machine)
              precision    recall  f1-score   support

         APC       0.71      0.90      0.80       410
        LBBB       0.90      0.91      0.90       414
      NORMAL       0.84      0.80      0.82       381
         PVC       0.89      0.90      0.90       393
        RBBB       0.97      0.73      0.83       402

    accuracy                           0.85      2000
   macro avg       0.86      0.85      0.85      2000
weighted avg       0.86      0.85      0.85      2000
```

Normalized Confusion Matrix (Support Vector Machine)
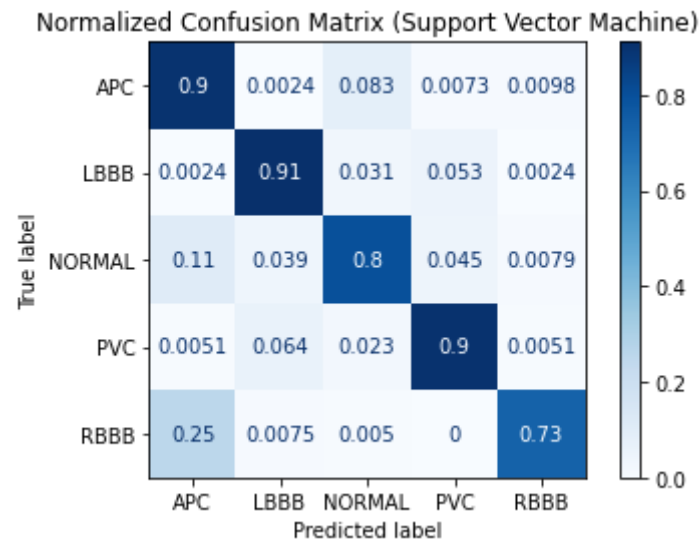
The best model to use would be the KNN mode; it provided the most accurate readings of any of the models. Given the uniform size of the data (240 points per beat), KNN wqould provide the best results.

4 Conclusion

5 References

Narimani, M. (2022). *MSE 491 Lab 2 Classification* [Class Handout]. Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC.