

2. B)

DEAD:

1. Average Event Count =982.014

2. Max Event Count=8635

3. Min Event Count=1

1. Average Encounter Count=23.038

2. Max Encounter Count=203

3. Min Encounter Count=1

1. Average Record Length=127.5

2. Max Record Length=1972

3. Min Record Length=0

ALIVE:

1. Average Event Count=498.118

2. Max Event Count=12627

3. Min Event Count=1

1. Average Encounter Count=15.452

2. Max Encounter Count=391

3. Min Encounter Count=1

1. Average Record Length

2. Max Record Length=291

3. Min Record Length=0

4.1 B)

Accuracy AUC Precision Recall F-Score

Logistic Regression

SVM

Decision Tree

	Acc	AUC	Precision	Recall	F-Score		
Logistic Regression	0.955	0.954	0.987	0.899	0.941		
SVM	0.994	0.995	0.988	0.997	0.993		
Decision Tree	0.776	0.748	0.792	0.601	0.684		

```

0.6865768463073852 0.6816963204756004
(base) george@george-VirtualBox: ~/BD4H/homework1/src$ cd ..
(base) george@george-VirtualBox: ~/BD4H/homework1$ cd src
(base) george@george-VirtualBox: ~/BD4H/homework1/src$ /home/george/anaconda3/bin/python /home/george/BD4H/homework1/src/models/partb.py
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

Classifier: Logistic Regression
Accuracy: 0.9545454545454546
AUC: 0.9454047619047619
Precision: 0.9869281045751634
Recall: 0.8988095238095238
F1-score: 0.9408099680473521

Classifier: SVM
Accuracy: 0.9940191387559809
AUC: 0.9945119047619048
Precision: 0.9882065899705014
Recall: 0.9970238095238095
F1-score: 0.9925925925925925

Classifier: Decision Tree
Accuracy: 0.7763157894736842
AUC: 0.7475952380952382
Precision: 0.792156862745098
Recall: 0.6011904761904762
F1-score: 0.6835871404399323

(base) george@george-VirtualBox: ~/BD4H/homework1/src$

```

4.1 C)

	Acc	AUC	Precision	Recall	F-Score		
Logistic Regression	0.738	0.738	0.680	0.733	0.706		
SVM	0.738	0.739	0.677	0.744	0.709		
Decision Tree	0.671	0.657	0.633	0.556	0.592		

```
(base) george@george-VirtualBox:~/BD4H/homework1/src$ /home/george/anaconda3/bin/python /home/george/BD4H/homework1/src/models/partc.py
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify
is warning.
  FutureWarning)

Classifier: Logistic Regression
Accuracy: 0.7380952380952381
AUC: 0.7375
Precision: 0.6804123711340206
Recall: 0.7333333333333333
F1-score: 0.7058823529411765

Classifier: SVM
Accuracy: 0.7380952380952381
AUC: 0.7388888888888889
Precision: 0.6767676767676768
Recall: 0.7444444444444445
F1-score: 0.708994708994709

Classifier: Decision Tree
Accuracy: 0.6714285714285714
AUC: 0.6569444444444444
Precision: 0.6329113924050633
Recall: 0.5555555555555556
F1-score: 0.591715976331361

(base) george@george-VirtualBox:~/BD4H/homework1/src$
```

4.1 D) It seems more data could definitely help with these model results. I think it is also worth tuning the observation window – it may be that longer histories of the patients have a more significant impact on the mortality of a patient as histories of medical issues typically span many, many years. Grid search or Bayesian hyperparameter tuning should also be implemented to get better results.

4.2 B)

Avg AUC for KCV: 0.7058

Avg ACC for KCV: 0.7213

Avg AUC for Random CV: 0.7188

Avg ACC for Random CV: 0.7357

```
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
Average Accuracy in KFold CV: 0.7213216424294269
Average AUC in KFold CV: 0.7075773303028468
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
/home/george/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning:
is warning.
FutureWarning)
Average Accuracy in Randomised CV: 0.7357142857142858
Average AUC in Randomised CV: 0.7188220160244053
(base) george@george-VirtualBox:~/BD4H/homework1/src$ ^C
(base) george@george-VirtualBox:~/BD4H/homework1/src$
```

4.3 My best model – Gradient boosting classifier didn't perform as well as I had hoped. I did 5 fold Kfold and used basic hyperparameters (default) and a max depth of 5 so I could compare to that given to us from you guys and the results were not promising. I repeated this same procedure for RandomForest Classifier. Gradient Boosting Classifier performed even better in this case so that was my best model. I thought that the more sophisticated models would actually perform better as they have more room to represent data with higher variance. This was not the case. The main reason for this I can think of is that these tree based models are much more complex than logistic regression and are much better suited for larger more complex datasets (there's an argument made by Andrew NG that larger/complex models are better suited for larger datasets). It may be that this dataset is rather straightforward