2.2b

Metric	Deceased patients	Alive patients	Function to complete
Event Count			event_count_metrics
1. Average Event Count	982.014	498.118	
2. Max Event Count	8635	12627	
3. Min Event Count	1	1	
Encounter Count			encounter_count_metrics
1. Average Encounter Count	23.038	15.452	
2. Max Encounter Count	203	391	
3. Min Encounter Count	1	1	
Record Length			record_length_metrics
1. Average Record Length	127.532	159.2	
2. Max Record Length	1972	2914	
3. Min	0	0	

4.1b Training Data

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.954545	0.961389	0.898810	0.986928	0.940810
SVM	0.994019	0.993094	0.997024	0.988201	0.992593
Decision Tree	0.776316	0.780760	0.601190	0.792157	0.683587

4.1c Test Data

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.738095	0.734011	0.733333	0.680412	0.705882
SVM	0.738095	0.734780	0.744444	0.676768	0.708995
Decision Tree	0.671429	0.663784	0.555556	0.632911	0.591716

4.1d

Some strategies to improve the models' performance are as follows:

1. Outlier detection: Identify if there are outliers present and understand the impact of removing it on our model. Could Cooks distance to identify major outliers.
2. Adding more data: Its always better to have more data rather relying on assumptions.

3. Exploratory data analysis using graphs is very important before doing any modeling. This helps us understand what models might be suitable for our data set.
4. Use parameter tuning to find optimal value for each parameter which will improve the model accuracy.
5. Correlated features in general don't improve models and can yield unstable values. Use principle component analysis for eliminating correlated features.

4.2b

CV strategy	Accuracy	AUC
K-Fold	0.664671	0.708907
Randomized	0.714286	0.730846

4.3d

In my model, I experimented with random forest classification and Decision Tree Regressor and kept the same features. I chose random forest classifier because it is an ensemble algorithm which combines multiple algorithms (like support vector machine, decision tree etc.) for classification. I chose Decision Tree Regressor since its related to decision classifier. Below are my results of cross validation and AUC. Based on the results, my model decision tree regressor has done better overall when compared to the previous sections.

Results of Random Forest Classifier

CV strategy	Accuracy	AUC
K-Fold	0.592814	0.664063
Randomized	0.669856	0.666155

Results of Decision Tree Regressor

CV strategy	Accuracy	AUC
K-Fold	0.70658683	0.67807327
Randomized	0.68181818	0.65792202