DIC Phase 2

Abdul Wasi Lone

50609995

Introduction: This study aims to identify factors that impact the **length of hospitalization** for patients and explore the relationship between **patient characteristics and ASA ratings**.

We employed two machine learning algorithms, **Decision Tree and Random Forest** Classifiers, to analyze a comprehensive medical dataset and derive insights.

Dataset Overview: Our analysis utilized a dataset containing various medical features, including:

• BIRTH_DATE (patient age)

• GENDER

• ASA_RATING_C (ASA physical status classification)

• AN_TYPE (type of anesthesia)

• ICU_ADMIN_FLAG (ICU admission status)

• AN_LOS_HOURS (length of stay in hours)

Methodology:

Data Preprocessing: We implemented the following preprocessing steps:

Categorical variable encoding using OneHotEncoder

Numerical feature scaling using StandardScaler

Data splitting: 80% training, 20% testing

Model Selection and Justification :

Decision Tree Classifier: We chose the Decision Tree Classifier for its:

• Interpretability, aligning well with medical decision-making processes

• Ability to handle both numerical and categorical data effectively

• Capacity to capture non-linear relationships between features

Random Forest Classifier: We selected the Random Forest Classifier to:

• Reduce overfitting and improve generalization as an ensemble method

• Provide robust feature importance rankings

• Effectively handle high-dimensional data Model Tuning

We used **GridSearchCV for hyperparameter** optimization:

Decision Tree Parameters tuned:

• max_depth: [3, 5, 7, 9]

• min_samples_split: [2, 5, 10]

• min_samples_leaf: [1, 2, 5, 10]

Best parameters: max_depth=5, min_samples_split=2, min_samples_leaf=5

Random Forest Parameters tuned:

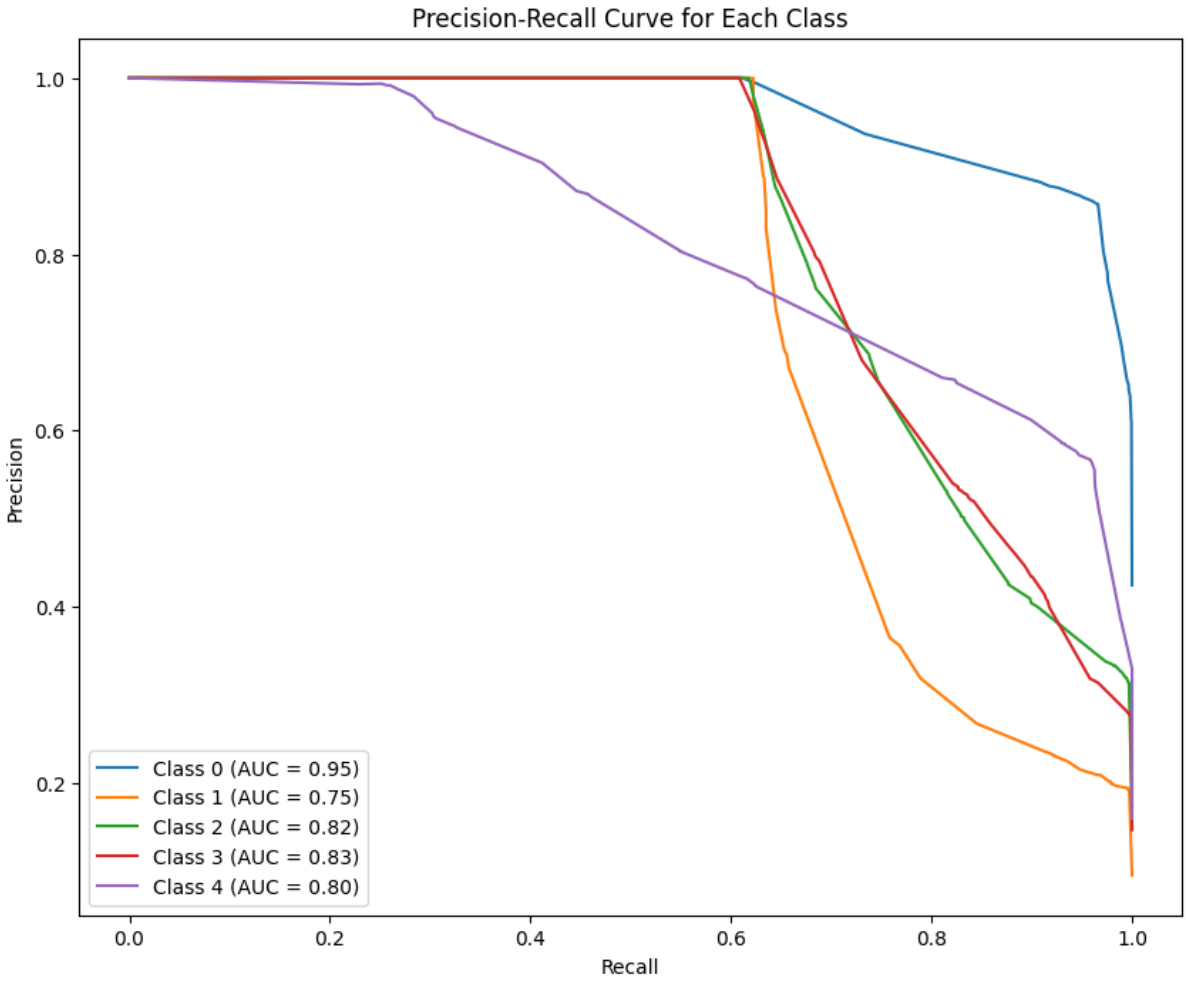
• n_estimators: [100, 200]

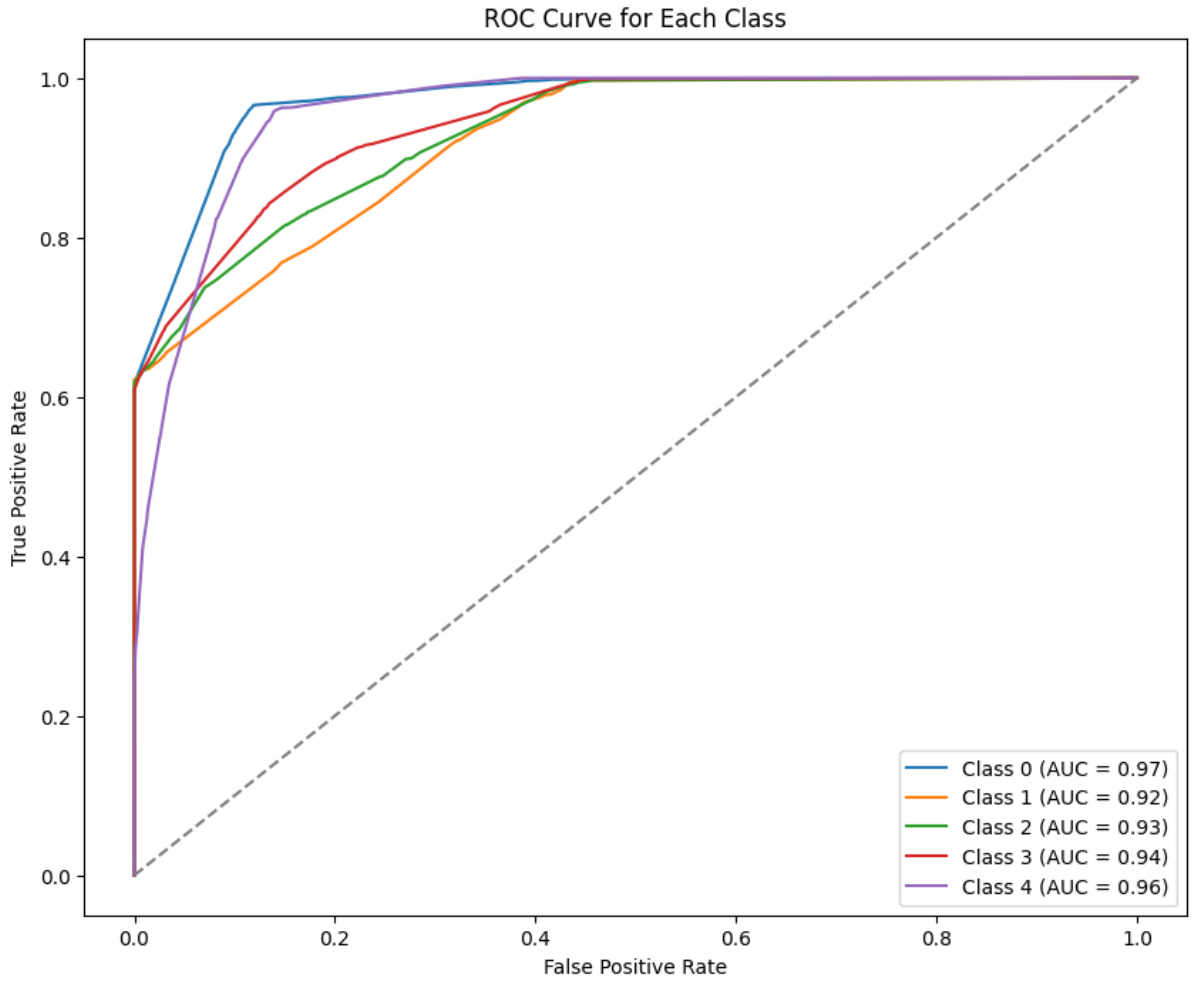
• max_features: ['auto', 'sqrt']

• max_depth: [10, 20, 30, None]

Best parameters: n_estimators=200, max_features='sqrt', max_depth=10

Results and Analysis Model Performance Decision Tree:


Precision-Recall Curve for Each Class
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.95)
Class 1 (AUC = 0.75)
Class 2 (AUC = 0.82)
Class 3 (AUC = 0.83)
Class 4 (AUC = 0.80)

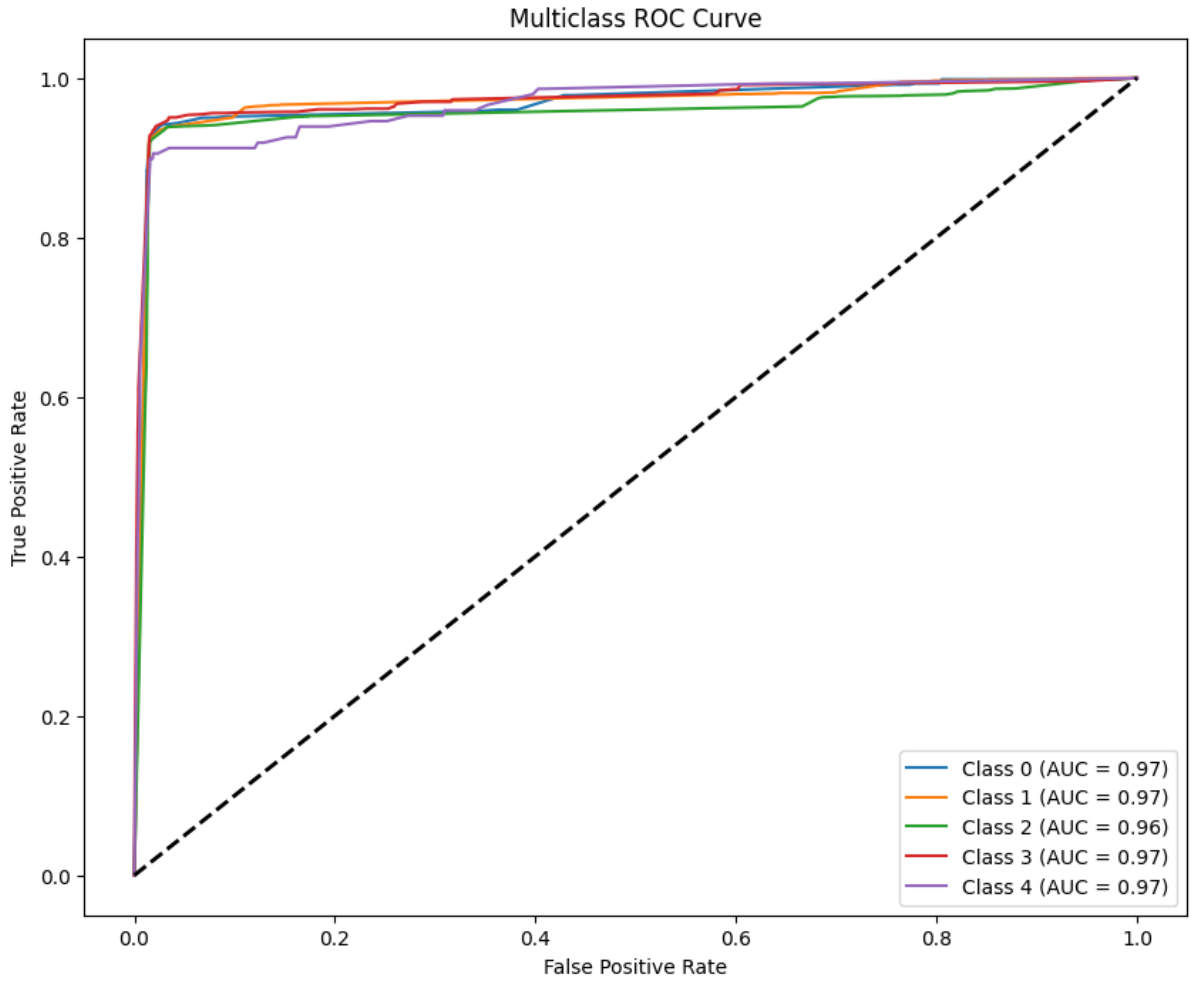


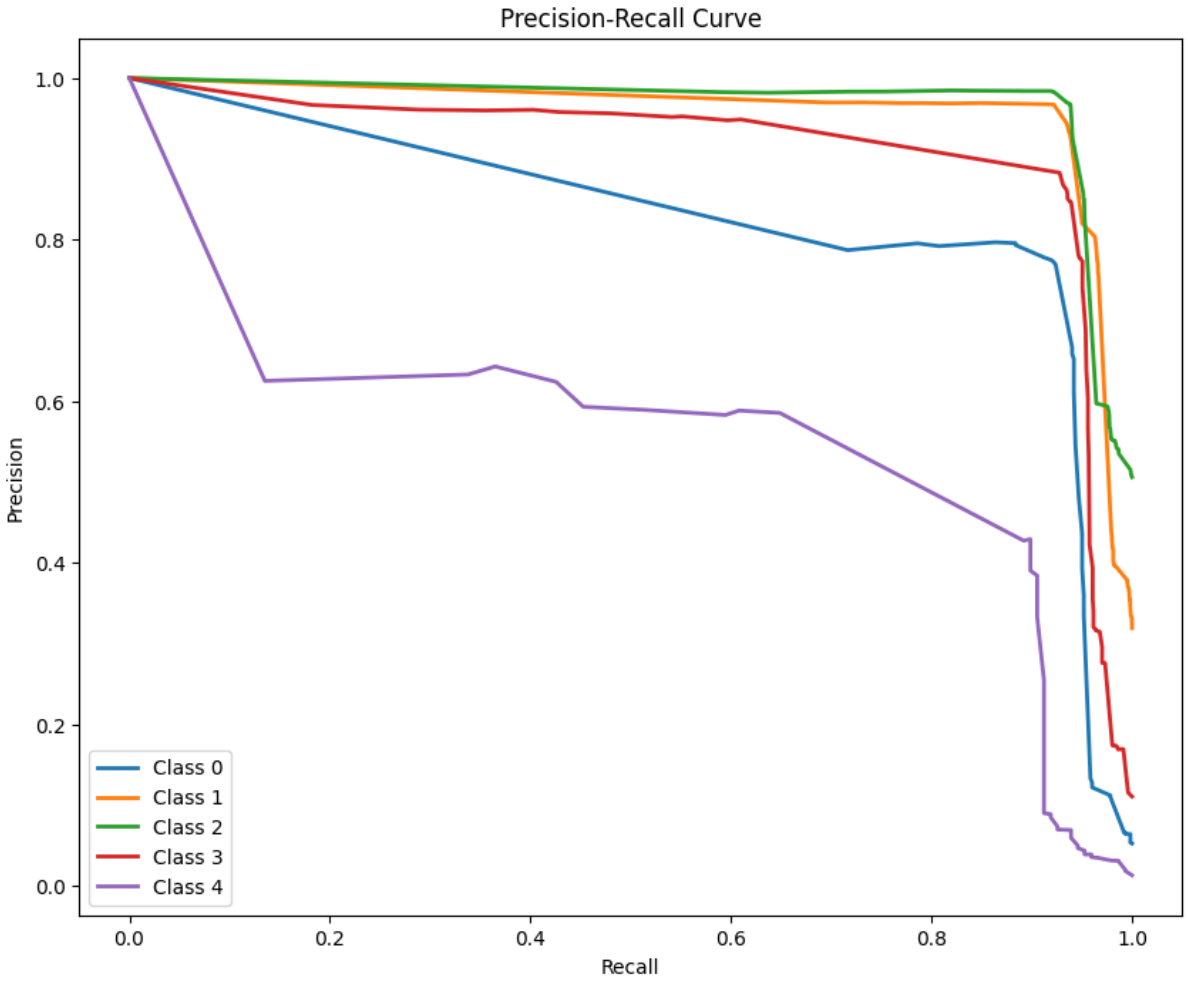

```
Best hyperparameters: {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 2}
Accuracy: 0.818378471921779
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.97      0.91      4907
           1       1.00      0.62      0.77      1104
           2       1.00      0.62      0.76      1997
           3       0.92      0.63      0.75      1700
           4       0.59      0.93      0.72      1849

    accuracy                           0.82     11557
   macro avg       0.87      0.75      0.78     11557
weighted avg       0.86      0.82      0.82     11557
```

Results and Analysis Model Performance Random Forest Tree:


Multiclass ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.97)
Class 1 (AUC = 0.97)
Class 2 (AUC = 0.96)
Class 3 (AUC = 0.97)
Class 4 (AUC = 0.97)




```
Best hyperparameters: {'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__n_estimators': 100}
Accuracy: 0.926060606060606
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       603
           1       0.95      0.93      0.94      3683
           2       0.98      0.92      0.95      5842
           3       0.85      0.94      0.89      1274
           4       0.42      0.90      0.57       148

    accuracy                           0.93     11550
   macro avg       0.79      0.92      0.84     11550
weighted avg       0.94      0.93      0.93     11550
```