Homework 1 Report

Author: Yanran Li

Collaborators: Kexuan Liang, Yichen Wang, Mingyang Ma

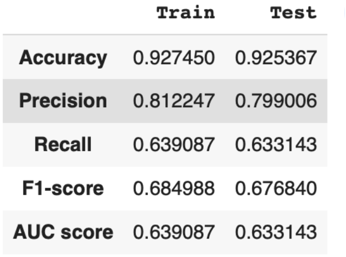
Part 1. Mortality Prediction in the ICU

(a).
Features useful for mortality prediction: all the features other than the ones listed below.

Features that are/should be irrelevant: 'encounter_id', 'patient_id', 'hospital_id', "icu_id", "apache_2_diagnosis", "apache_2_bodysystem", "apache_3j_bodysystem", "apache_3j_diagnosis"
We excluded all "id" related features since they are not important for model prediction while they are just the identifications for different patients. We excluded "apache_2_diagnosis", "apache_2_bodysystem", "apache_3j_bodysystem" and "apache_3j_diagnosis" since they may be the systemic symptoms by patients, which are not affect the mortality prediction by models.

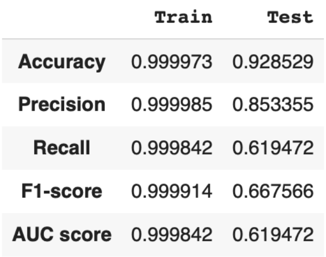
Features that might be predictive but that inadvertently leak data to the model: "apache_4a_hospital_death_prob", "apache_4a_icu_death_prob", h1_blood_culture", "h1_urine_culture", "h1_sputum_culture", "d1_medication_name", "d1_medication_dosage", "d1_medication_name_complete", "h1_serum_immunoglobulins_iga", "h1_serum_immunoglobulins_igg", "h1_serum_immunoglobulins_igm", "h1_anca", h1_serum_complement_total_C3", "h1_serum_complement_total_C4"
We excluded "apache_4a_hospital_death_prob" and "apache_4a_icu_death_prob" since they themselves are probabilities. We excluded " h1_blood_culture", "h1_urine_culture", "h1_sputum_culture", "h1_serum_immunoglobulins_iga", "h1_serum_immunoglobulins_igg", "h1_serum_immunoglobulins_igm", "h1_anca", h1_serum_complement_total_C3", "h1_serum_complement_total_C4"
since they are sample culture results. We also excluded "d1_medication_name", "d1_medication_dosage" since they are obtained as a result of the clinician's prediction of the outcome of the patients.

(c).
logistic regression

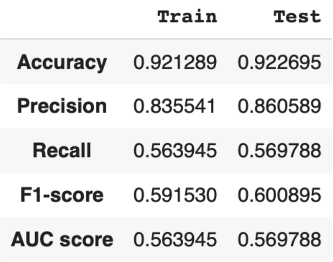


	Train	Test
Accuracy	0.927450	0.925367
Precision	0.812247	0.799006
Recall	0.639087	0.633143
F1-score	0.684988	0.676840
AUC score	0.639087	0.633143

(d).
Random forest Xgboost

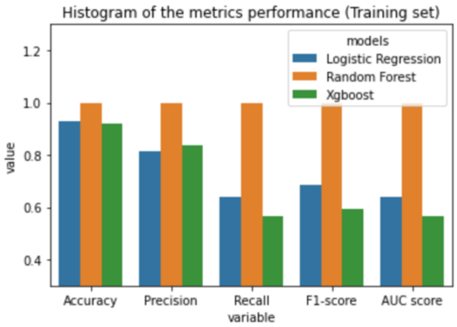


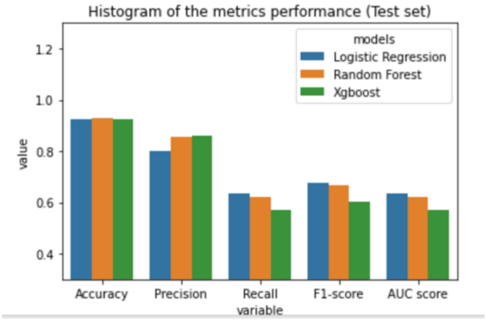
	Train	Test
Accuracy	0.999973	0.928529
Precision	0.999985	0.853355
Recall	0.999842	0.619472
F1-score	0.999914	0.667566
AUC score	0.999842	0.619472



	Train	Test
Accuracy	0.921289	0.922695
Precision	0.835541	0.860589
Recall	0.563945	0.569788
F1-score	0.591530	0.600895
AUC score	0.563945	0.569788

(e).





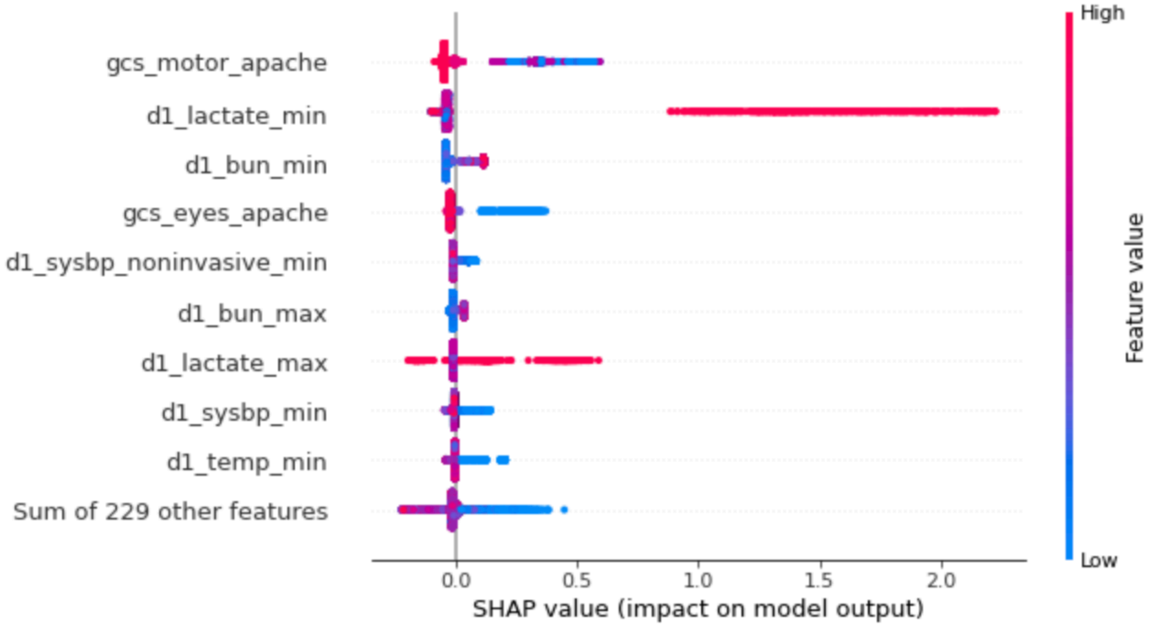
(f).

By looking at the reports from (c)-(d) and histograms from (e), we can conclude that logistic regression model has the best performance among the test set.

For models predicting the mortality rate in the test set, although the logistic regression model has the lowest precision among all 3 models, it have the highest recall, F1-score and AUC score. Precision score (TP/(TP+FP)) reflects the percentage of correct positive predictions in terms of predicting the positive outcomes (ICU death) and suggests the impact of false positives. However, the highest recall (TP/(TP+FN)) and F1 score are more important in ICU death predictions. Moreover, the highest AUC score can better classify death cases and non-death cases. Thus, we can conclude that logistic regression model has the best performance.

(g).

To get a high-level overview of which features contribute the most to our models' predictions, we made a beeswarm plot:

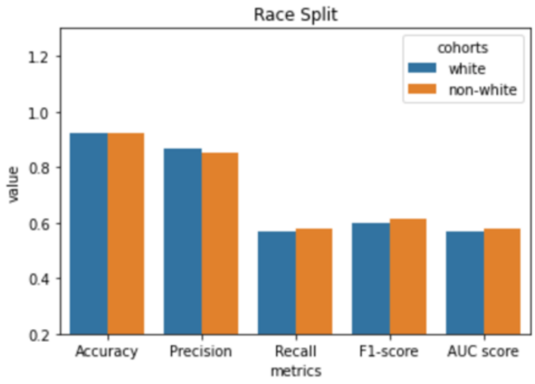


From the beeswarm plot, feature "gcs_motor_apache" contributes the most to the model's predictions. This feature stands for the motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score. It is a reasonable feature that the model can rely on: generally, lower GCS scores are correlated with higher risk of death. In this beewarm SHAP plot, a high "gcs_motor_apache" (motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score) lowers the predicted death probability.

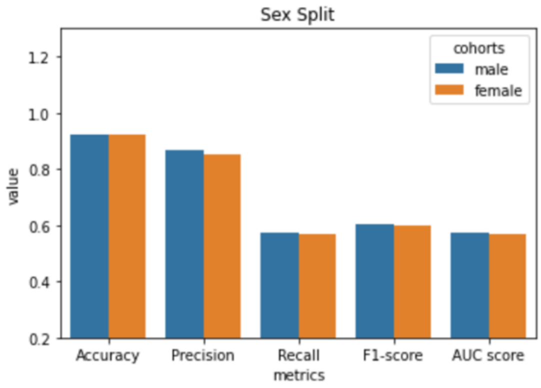
Besides, d1_lactate_min (the lowest lactate concentration), d1_bun_min (the lowest blood urea nitrogen concentration), gcs_eyes_apache (the eye opening component of the Glasgow Coma Scale), d1_sysbp_noninvasive_min (lowest systolic blood pressure non-invasively measured), d1_bun_max (the highest blood urea nitrogen concentration), d1_lactate_max(the highest lactate concentration), d1_sysbp_min (lowest systolic blood pressure) contribute much to the model's predictions.

They all make senses: for lactate value, high lactate value relates to high SHAP positive values. In clinical settings, as an end product of pyruvate metabolism, the levels of lactate can be an important indicator of disturbed metabolism and contribute to higher morbidity and mortality. For bun values, it is a waste product of protein metabolism and should be removed by the kidney, which can be viewed as a biomarker of renal function with its declining indicating renal injury, and is possibly related to higher mortality. Also, eye opening component is an important GCS metrics, which is likely to closely correlate with adverse health outcomes: dangerous health conditions can impair nervous systems and affect the vision function, leading to higher mortality risk.

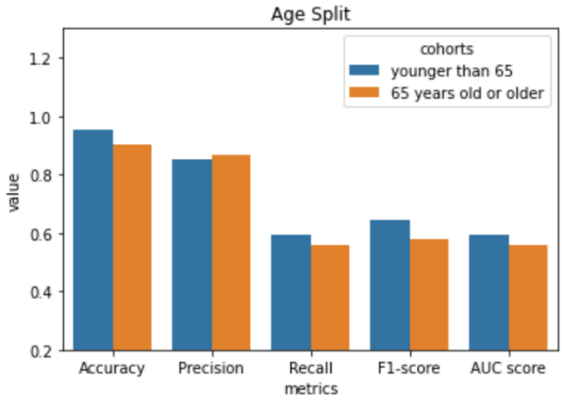
(i)(j).



For the white/non-white split, the xgboost model perform better in the non-white split, for which it has higher Recall, F1-score and AUC score.



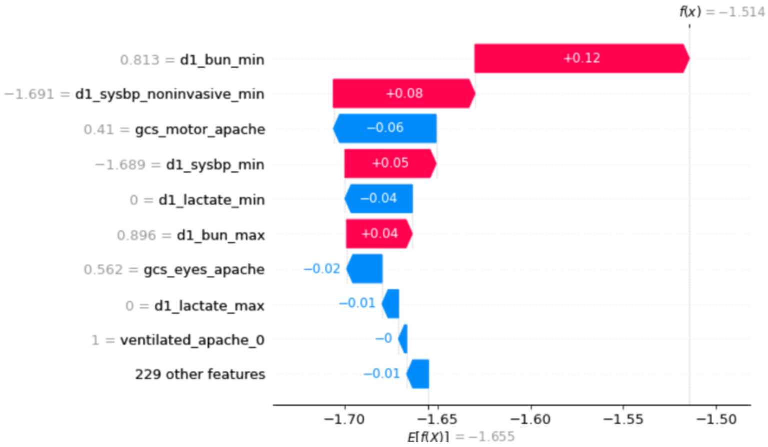
For the age split, the xgboost model perform better in the younger than 65 split, for which it has higher Accuracy, Recall, F1-score and AUC score.



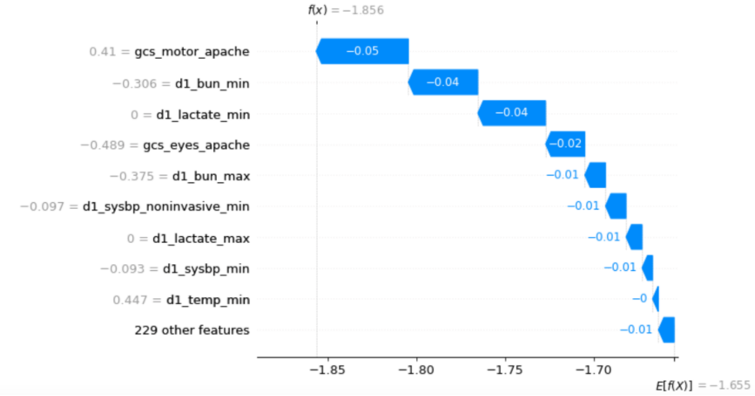
(k).

- Race

white patients:

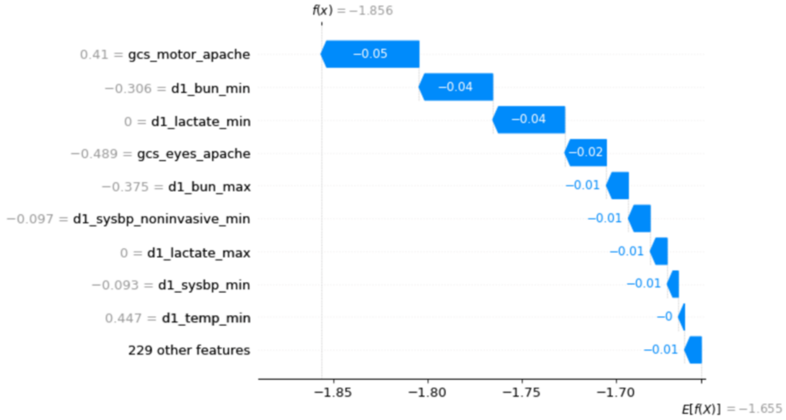


non-white patients:

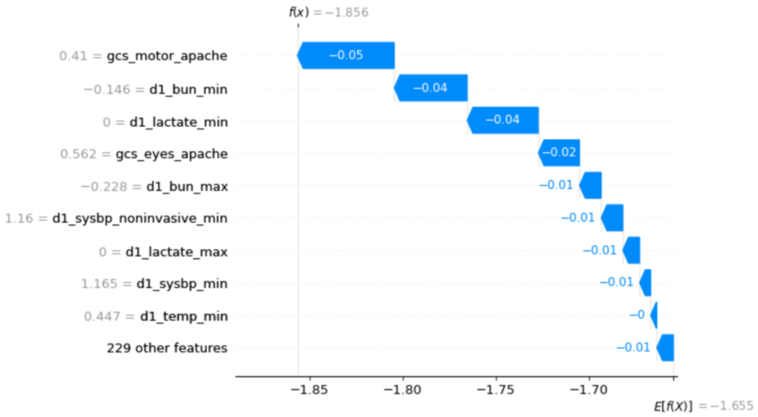


- Sex

male patients:

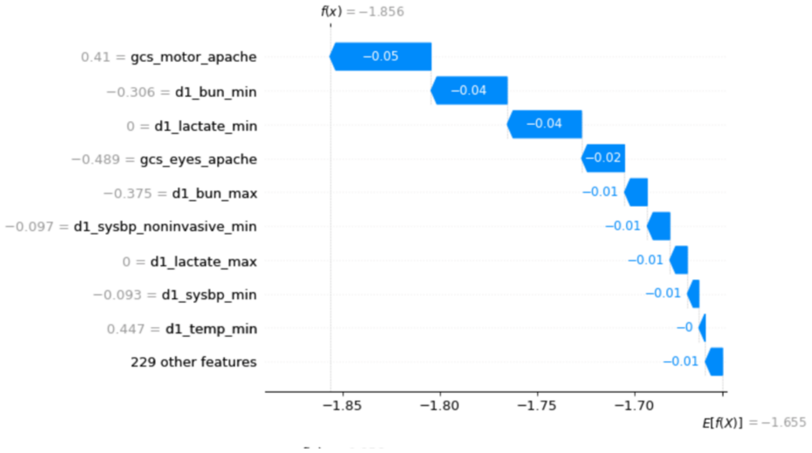


female patients:

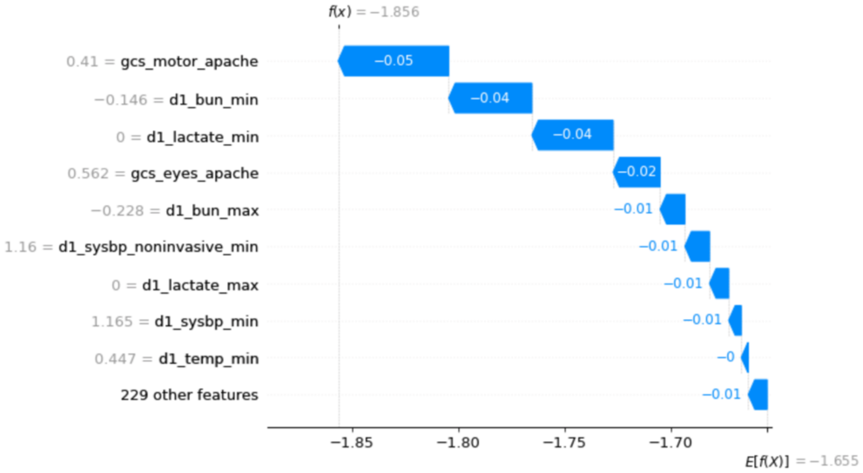


- Age

patients younger than 65:



patients 65 years old or older:



(l).

I’ve noticed discrepancies in the features used by the xgboost model to make predictions for the white/non-white split. Among the white patients, the importance order of contributing features (from high to low) is: d1_bun_min > d1_sysbp_noninvasive_min > gcs_motor_apache > d1_ sysbp_min > d1_lactate_min > d1_bun_max > gcs_eyes_apache >

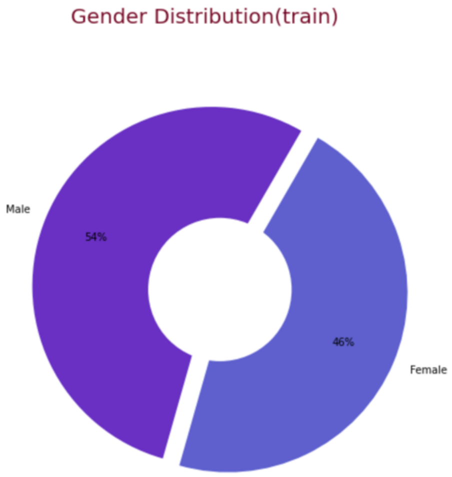
d1_lactate_max > ventilated_apache_0 > (other 229 features). Among the non-white patients, the importance order of contributing features (from high to low) is: gcs_motor_apache > d1_bun_min > d1_lactate_min > gcs_eyes_apache > d1_bun_max > d1_sysbp_noninvasive_min > d1_lactate_max > d1_ sysbp_min > di_temp_min > (other 229 features). Feature "d1_bun_min" (the lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay) contributes to push the model output from the base value (the average model output over the training dataset we passed) to the model prediction output higher in white patients, while it pushes the prediction lower in non-white patients. Similarly, features "d1_sysbp_noninvasive_min" (patient's lowest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured), "d1_ sysbp_min" (patient's lowest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured) and "d1_bun_max" (the highest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay) all contribute to push the model output from the base value to the model prediction output higher in white patients, while they push the prediction lower in non-white patients.

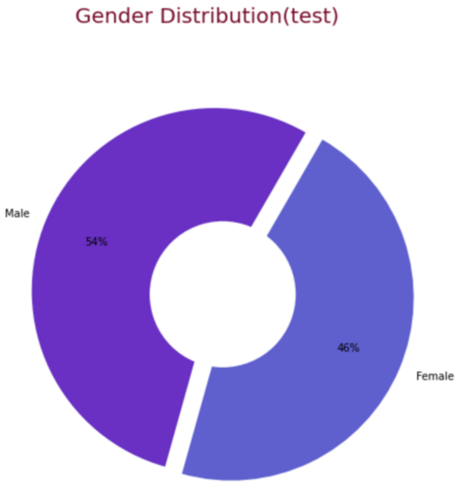
Such discrepancies were not observed among male/female split and patients younger/older than 65 split, which may due to sufficient model generalizability for sexs/ages and there's not too much imbalances among different splits. We expect model to give better predictions on test set since it makes use of almost the same amount of information to predict the two gender/age classes.

Part 2. Delving into Disparities

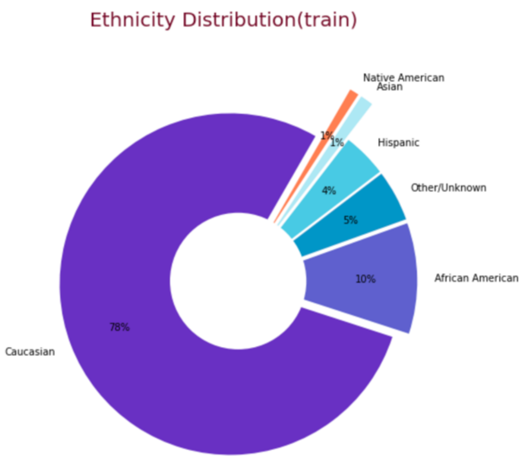
(a).

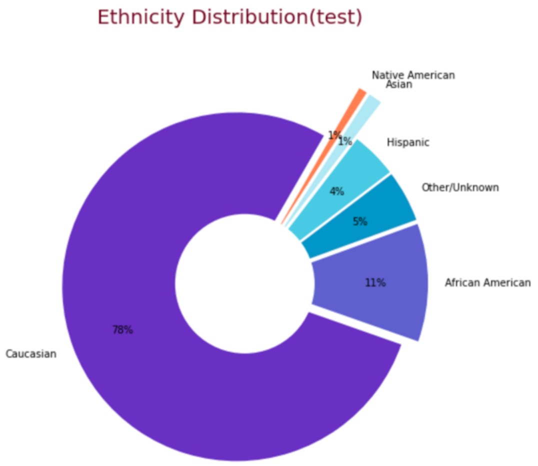
distribution of the patients' gender:



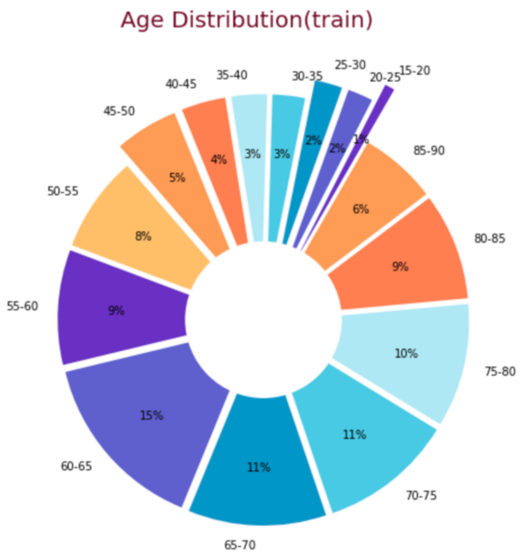


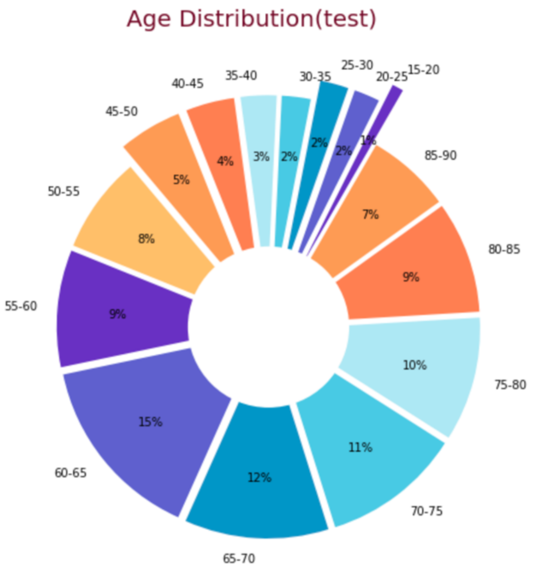
distribution of the patients' race:


Ethnicity Distribution(train)


Ethnicity Distribution(test)

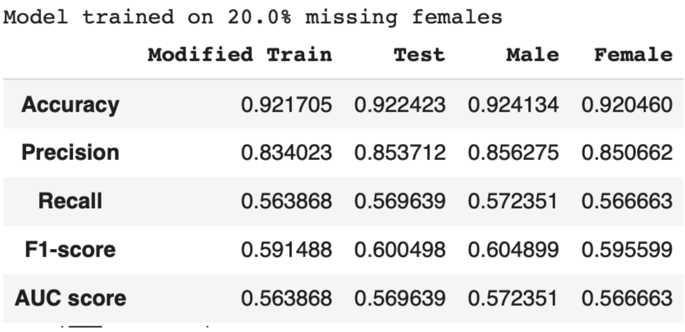
distribution of the patients' age:


Age Distribution(train)


Age Distribution(test)

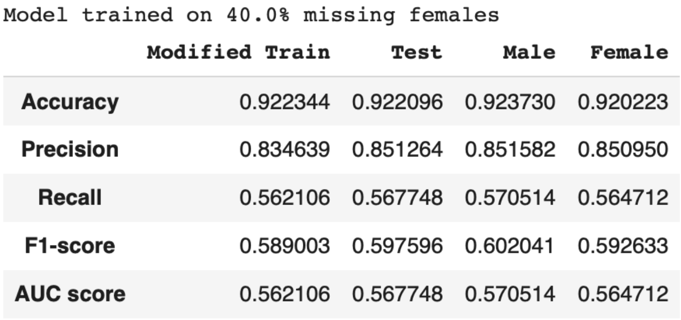
Seeing from the 6 pie plots, we do not see too much differences between the train and test cohorts. For the ethnicity, 78% of the patients are Caucasian, which is a huge imbalance. For the gender, 54% among the patients are male while 46% among the patients are female, and there's not too much imbalance. For the age, patients 65 years old or older roughly account for 48% of the total, and there's not too much imbalance.

(b).



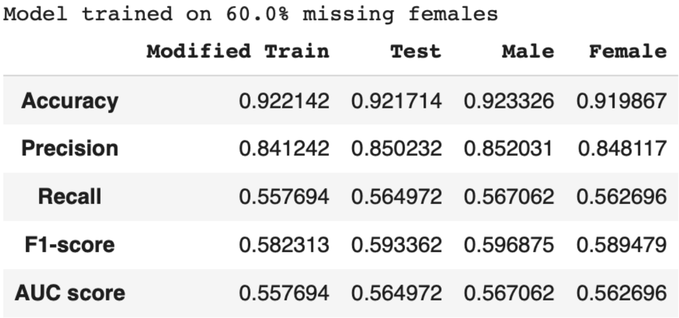
```
Model trained on 20.0% missing females
```

	Modified Train	Test	Male	Female
Accuracy	0.921705	0.922423	0.924134	0.920460
Precision	0.834023	0.853712	0.856275	0.850662
Recall	0.563868	0.569639	0.572351	0.566663
F1-score	0.591488	0.600498	0.604899	0.595599
AUC score	0.563868	0.569639	0.572351	0.566663



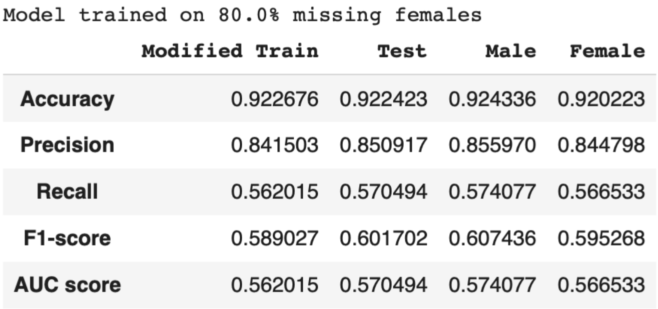
```
Model trained on 40.0% missing females
```

	Modified Train	Test	Male	Female
Accuracy	0.922344	0.922096	0.923730	0.920223
Precision	0.834639	0.851264	0.851582	0.850950
Recall	0.562106	0.567748	0.570514	0.564712
F1-score	0.589003	0.597596	0.602041	0.592633
AUC score	0.562106	0.567748	0.570514	0.564712



```
Model trained on 60.0% missing females
```

	Modified Train	Test	Male	Female
Accuracy	0.922142	0.921714	0.923326	0.919867
Precision	0.841242	0.850232	0.852031	0.848117
Recall	0.557694	0.564972	0.567062	0.562696
F1-score	0.582313	0.593362	0.596875	0.589479
AUC score	0.557694	0.564972	0.567062	0.562696



```
Model trained on 80.0% missing females
```

	Modified Train	Test	Male	Female
Accuracy	0.922676	0.922423	0.924336	0.920223
Precision	0.841503	0.850917	0.855970	0.844798
Recall	0.562015	0.570494	0.574077	0.566533
F1-score	0.589027	0.601702	0.607436	0.595268
AUC score	0.562015	0.570494	0.574077	0.566533

(c).

Reducing the number of female patient datapoints does not affect the performance of the model. This may because our female and male patients are similar in most of the features' distributions. We know from 1(l) that there's no obvious discrepancy in features used to make predictions for the male/female cohorts, which means the associations between these features and the prediction of ICU mortality did not significantly differ between females and males.

(d).

I would expect to see the similar results when varying the degree of missingness of some populations in patients younger/older than 65 split. Because from 1(k) and 1(l), we've notice that discrepancies of features' importance for prediction were not observed among this split.

However, I would expect to see the different results when varying the degree of missingness of some populations in patients white/non-white split. Because from 1(k) and 1(l), we've observed obvious discrepancies of features' importance for prediction among this split.

(e).

Increasing the degree of missingness of test results in some training datapoints would lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain less error. Also, increasing the degree of missingness of test results in some training datapoints would reduce the performance indexes such as F1 score/AUC score. It would lose more valid data that can help us understand how the features can lead to changes in the health outcome and even important predictors. It would impair the true associations and bias the estimation of the features on the outcome of interest and the predicted ICU mortality.

(f).

A potential way of handling missingness of test results is to prevent the missingness by well-planning the study and collecting the test data carefully. To be more specific, before the beginning of the clinical research, form a detailed documentation of the study should be developed in the form of the manual of operations, which includes the tests to screen the participants, protocol to train the nurses, methods to communicate between the research investigators and doctors/nurses, implementation of the treatment, and procedure to collect, enter, and edit data.