Neha Dantuluri
HST 953/MIT 6.8850

PROBLEM SET 1 - DIAGRAMS AND WRITTEN REPORT

PART 1

A) in codebook.
B) in codebook.

C) logistic regression results:

logistic regression performance metrics

	Metric	Train	Test
0	Accuracy	0.927791	0.926457
1	Precision	0.693609	0.673977
2	Recall	0.291561	0.290485
3	F1-score	0.410547	0.405989
4	AUC score	0.883158	0.878863

D) random forest results:

random forest classifier performance metrics:

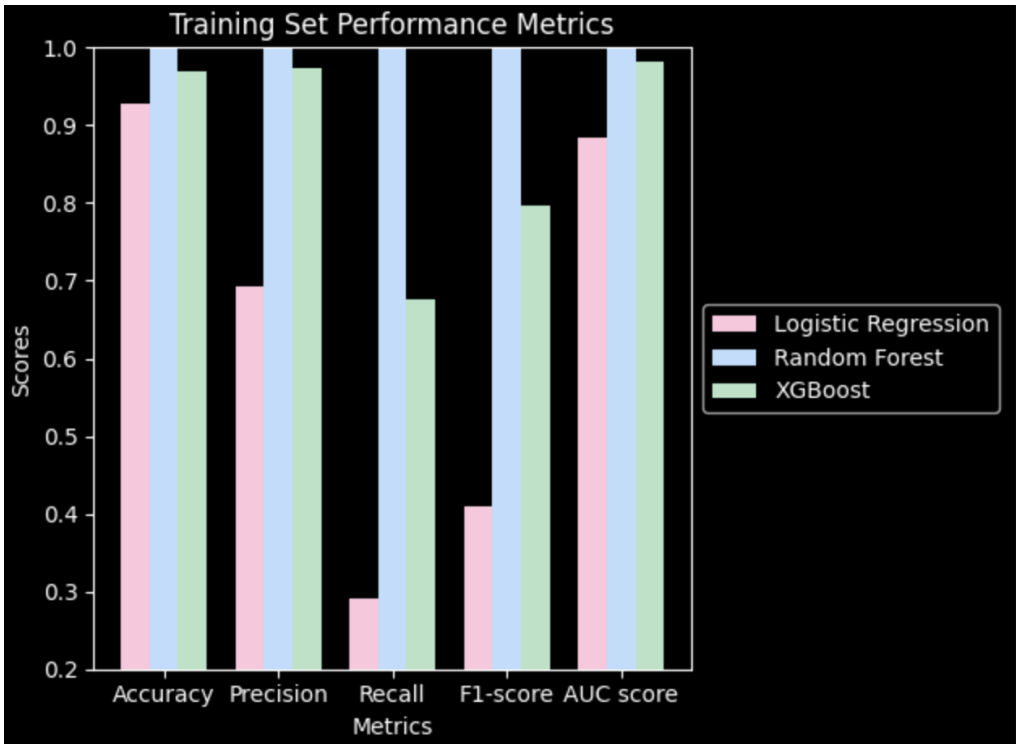
	Metric	Train	Test
0	Accuracy	0.999973	0.929128
1	Precision	1.000000	0.773333
2	Recall	0.999684	0.255829
3	F1-score	0.999842	0.384470
4	AUC score	1.000000	0.873930

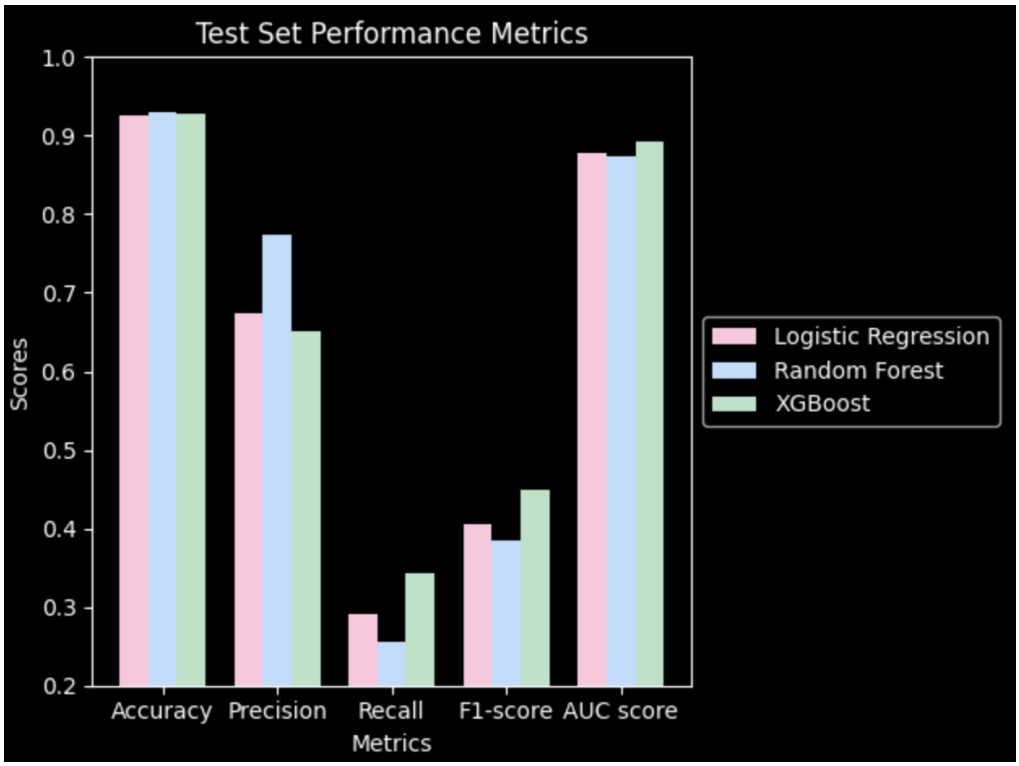
xgboost model metrics

xgboost performance metrics:

	Metric	Train	Test
0	Accuracy	0.970410	0.927275
1	Precision	0.973786	0.651861
2	Recall	0.675095	0.342155
3	F1-score	0.797387	0.448760
4	AUC score	0.982688	0.892596

E) Training + Test Set Histograms

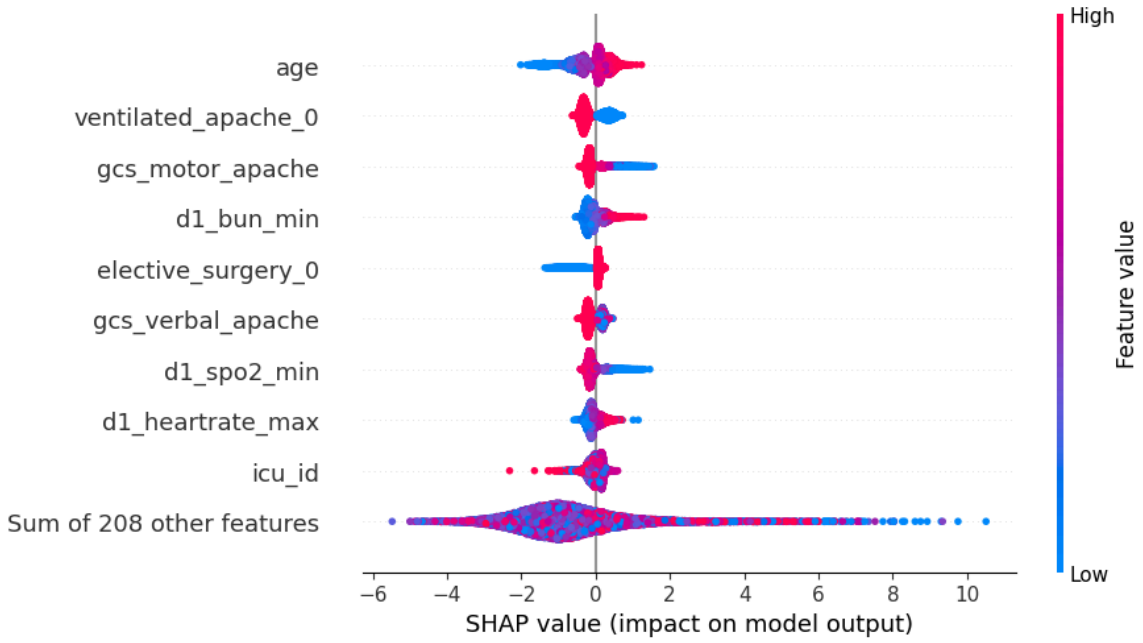




F) Which model has the best performance? Briefly explain your answer

XGBoost is the best-performing model. This is evident from the test set performance metrics histogram, as XGBoost shows better performance in recall, F1-score, and AUC. However, Random Forest and Logistic Regression performed better than XGBoost in terms of precision and Random Forest slightly outperformed XGBoost in accuracy.. Despite this, XGBoost demonstrates the overall best performance.

G) Beeswarm Plot:



***Note: the output plot in my Jupyter notebook has a dark background. I like the aesthetic more but some might discern the text to be a little difficult to read. So, I created an additional white-brackground beeswarm plot for this written report!

Which features contribute the most to the model's predictions?

Age, elective surgery, gcs_motor_apache, d1_bun_min, and elective_surgery_0 contribute the most to the model's predictions because they are features towards the top of the plot with the widest spreads.

Do they seem like reasonable features that the model can rely on, or is the model basing its predictions on spurious correlations?

These features seem reasonable for predicting ICU mortality as patient's risk of death increases if they are older (high age), ventilated (yes to ventilated_apache_0), have impaired brain function (low gcs_motor_apcache) , have impaired kidney function (low d1_bun_min), and are receiving emergency surgery (no to elective_surgery_0).

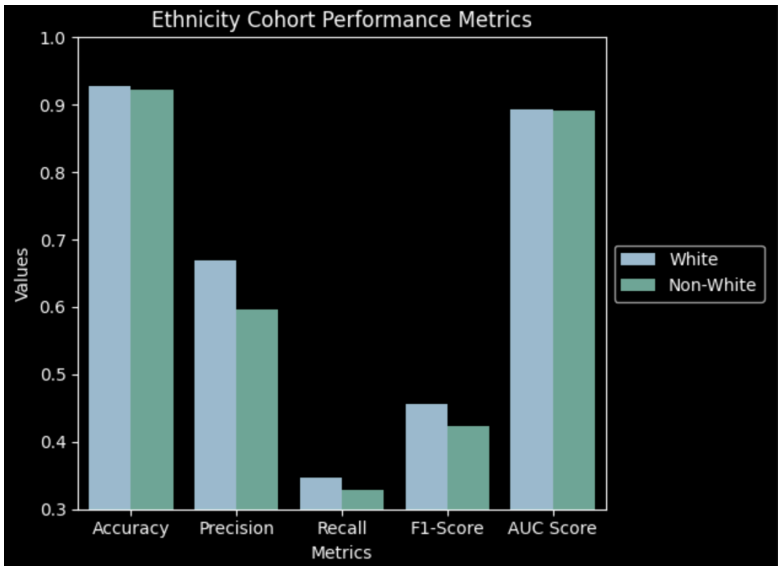
H) in codebook.

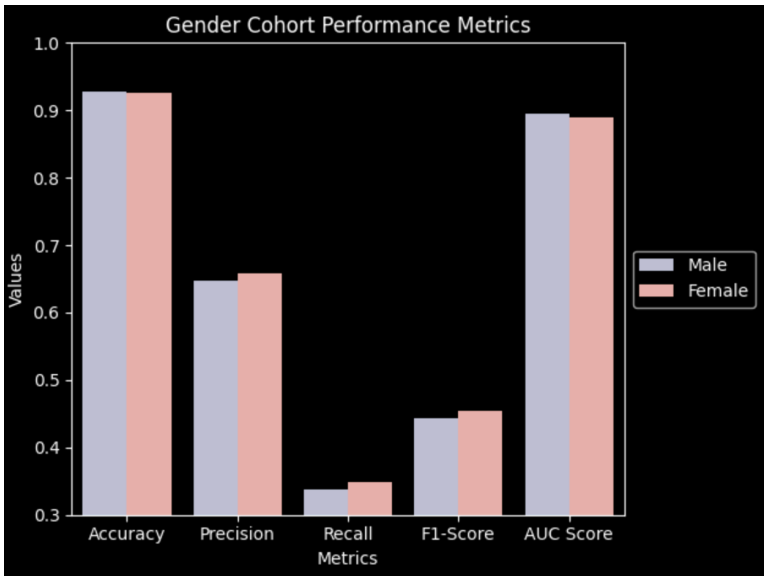
I) **Xgboost model prediction metrics on different cohorts:**

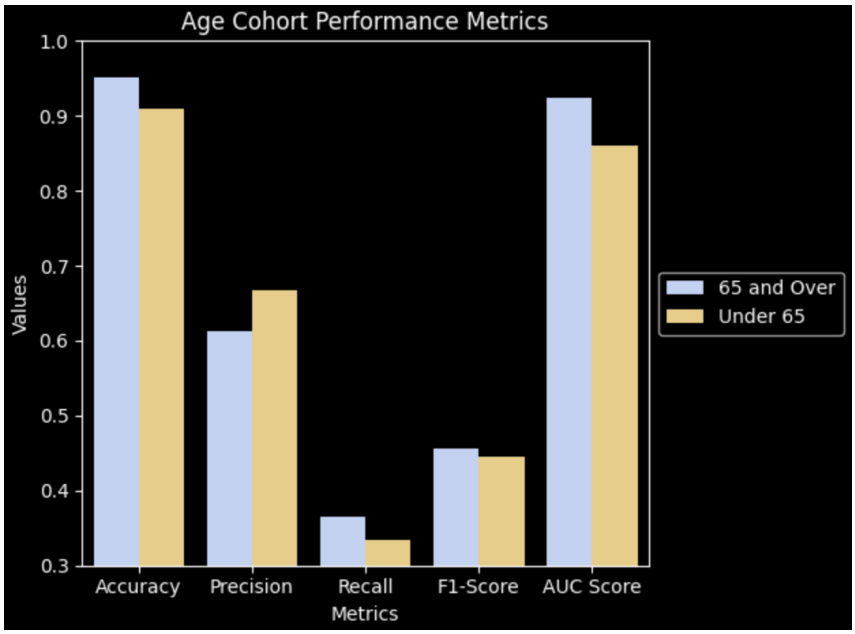
```
XGBoost performance metrics across cohorts:
      Metric     White  Nonwhite      Male    Female  Age Above 65  \
0   Accuracy  0.928668  0.922680  0.927973  0.926505      0.951796
1  Precision  0.669841  0.596059  0.646925  0.657360      0.611888
2     Recall  0.346470  0.327913  0.337292  0.348118      0.364583
3   F1-score  0.456710  0.423077  0.443404  0.455185      0.456919
4  AUC score  0.892925  0.891871  0.894784  0.889987      0.923881

   Age Below 65
0      0.909716
1      0.667360
2      0.333680
3      0.444906
4      0.860946
```

Xgboost model performance metrics histograms:







J) How well does the model perform on the two cohorts in each split?

Across all cohorts, the model performs pretty well (closest to ~0.900) in terms of accuracy and AUC score. It performs worse in terms of precision, f1 score, and recall.

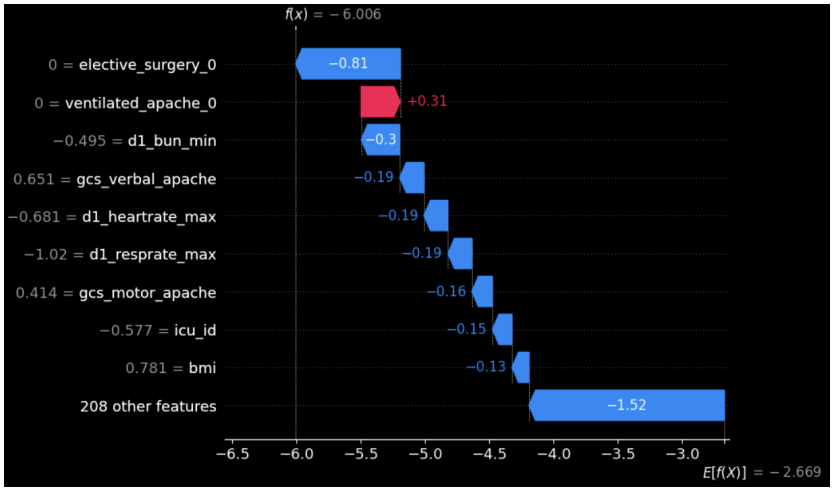
In the ethnicity cohort, the predictions for white patients score higher than those for non-white patients. White patients scored significantly better on precision, recall, and f1 scores. However, white patients scored better (albeit minimally) than black patients in terms of accuracy and AUC score.

In the gender cohort, the predictions for females perform slightly better than those for males. For accuracy, precision, recall, and f1, females performed slightly better than males. However, males scored slightly better than females in terms of AUC score.

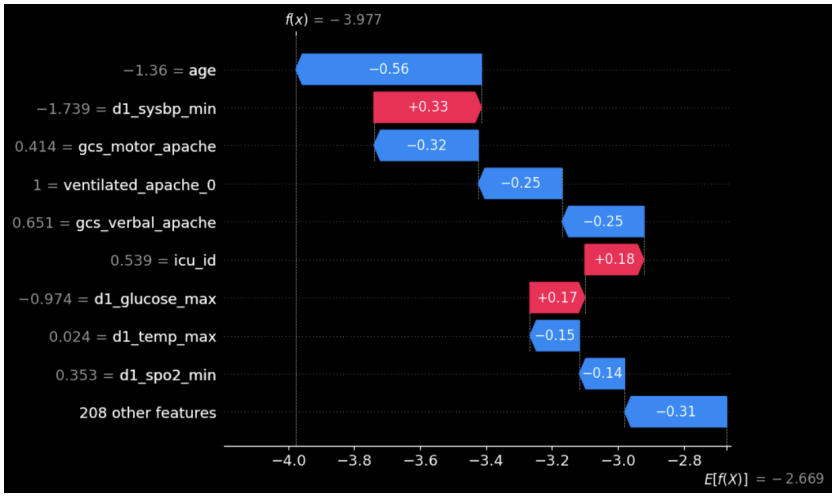
In the age cohort, the model performs better on patients 65 and older rather than patients that are younger than 65. The older patient cohort scored significantly better than younger patients in terms of accuracy, recall, f1 score, and AUC score. However, younger patients scored slightly better than older patients in terms of precision.

K) Shapley Plots:

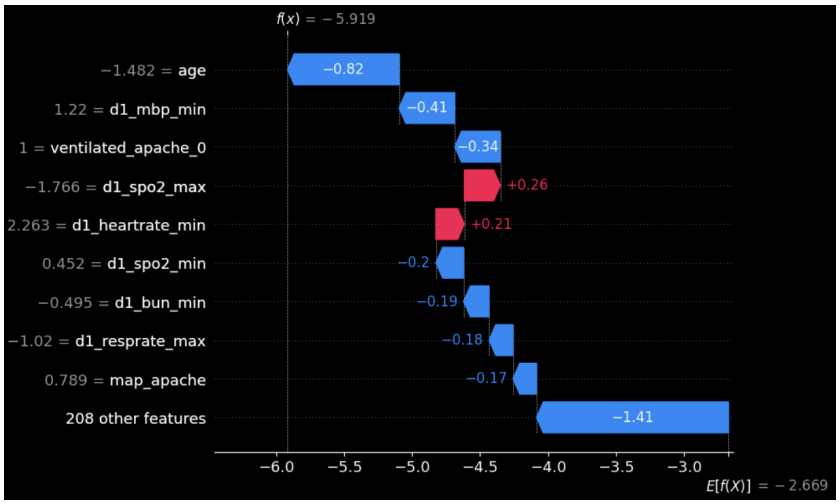
white_sample



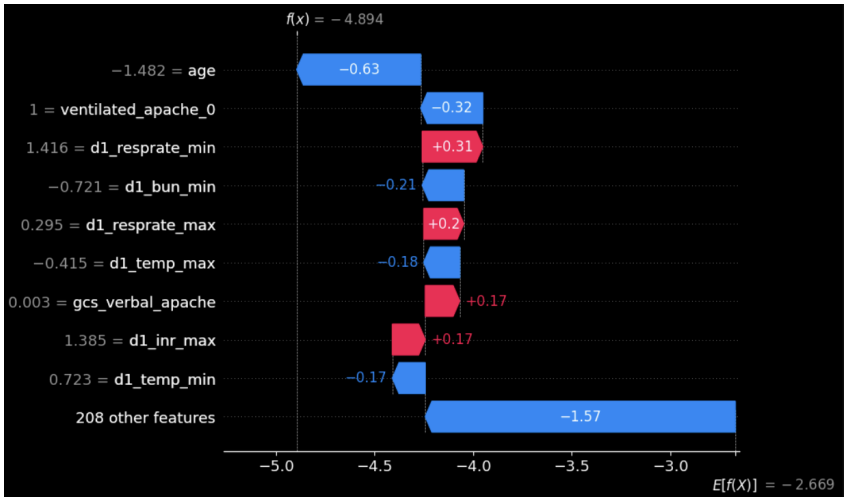
nonwhite_sample



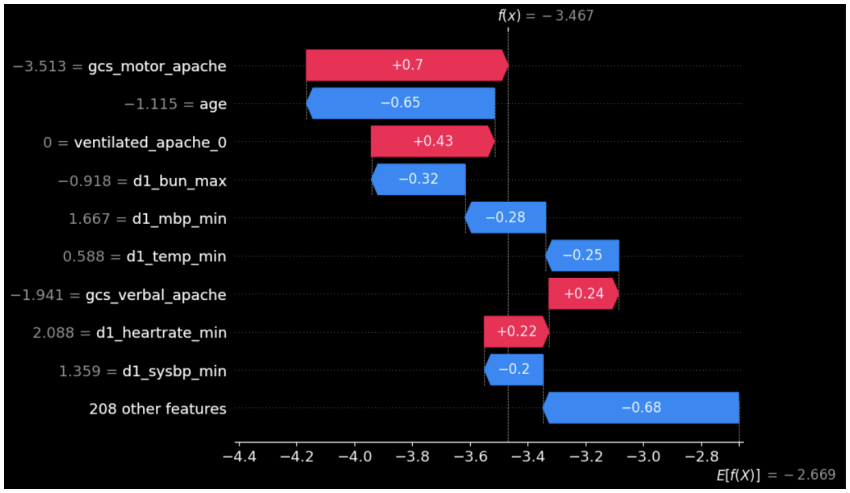
male_sample



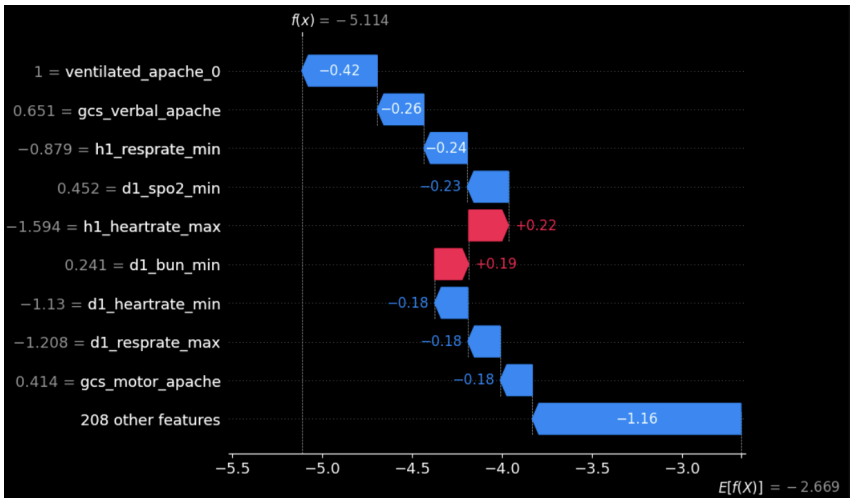
female_sample



old_sample



young_sample



L) Do you notice any discrepancies in the features used by the model to make predictions for the two cohorts in each split? If yes, briefly describe those discrepancies. If not, briefly explain why you think such discrepancies were not observed.

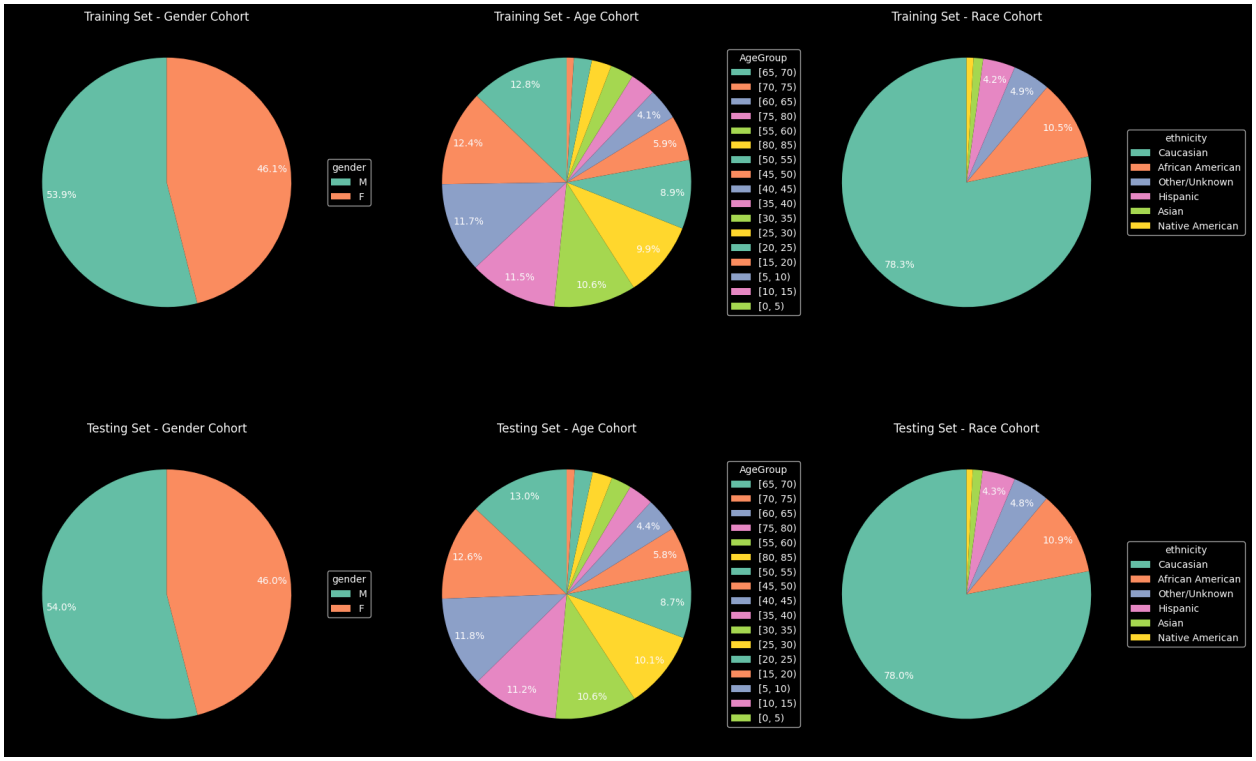
I noticed discrepancies in the ethnicity cohort. In white_sample, the top feature was elective_surgery_0, while in nonwhite_sample, the top feature was age. Interestingly, d1_resprate_max and bmi are contributing features in white_sample while being missing in nonwhite_sample. However, d1_sysbp_min and d1_glucose_max are contributing features in nonwhite_sample but not in white_sample.

I also noticed some discrepancies and non-discrepancies in the gender cohort. Age (as a top feature) was the same for male_sample and female_sample. To add, d1_mbp_min and map_apache are contributing features in male_sample but are missing in female_sample. However, d1_inr_max and d1_temp_max are contributing features in female_sample but are missing in male_sample.

Lastly, I also noticed discrepancies in the age cohort. In old_sample, the top feature was gcs_motor_apache. The top feature in young_sample was ventilated_apache_0. Additionally, d1_mbp_min and d1_sysbp_min are contributing features in old_sample but are missing in young_sample. However, h1_resprate_min and d1_spo2_min are contributing features in young_sample but are missing in old_sample.

PART 2

A) Pie Charts:



Do you notice any imbalances in the data?

I notice slight and/or significant imbalances in the training and testing sets across all cohorts.

In the gender cohort, the distribution is close to being equivalent in both the testing and training sets. In general, men are slightly more represented than women. Specifically, there are more men (+0.1%) in the testing set compared to the training set.

In the age cohort, both the testing and training sets have a much higher proportion of older patients being represented, while younger patients are much less represented.

In the race cohort, Caucasians are represented significantly more than all other ethnicities in both the testing and training sets. Representation of African Americans, Hispanics, and “other” races is low, and representation of Native Americans and Asians is extremely minimal to none. Specifically, there is a slight increase for Caucasians (+0.3%), Hispanics (+0.1%), and African Americans (+0.4%) in the testing set compared to the training set.

B) Xgboost performance after reducing training data points

	Reduction Percentage	Cohort	Accuracy	Precision	Recall	F1 Score	AUC
0	0.2	General	0.928256	0.663845	0.345936	0.454847	0.890143
1	0.2	Male Cohort	0.928680	0.656682	0.338480	0.446708	0.891408
2	0.2	Female Cohort	0.927759	0.671756	0.354362	0.463972	0.888636
3	0.4	General	0.928256	0.663450	0.346566	0.455298	0.889848
4	0.4	Male Cohort	0.928074	0.645089	0.343230	0.448062	0.889333
5	0.4	Female Cohort	0.928470	0.685039	0.350336	0.463588	0.890394
6	0.6	General	0.927166	0.651023	0.340895	0.447477	0.884680
7	0.6	Male Cohort	0.927973	0.641758	0.346793	0.450270	0.887461
8	0.6	Female Cohort	0.926220	0.662234	0.334228	0.444246	0.881417
9	0.8	General	0.927983	0.666667	0.335224	0.446122	0.880644
10	0.8	Male Cohort	0.928377	0.655012	0.333729	0.442172	0.884493
11	0.8	Female Cohort	0.927523	0.680217	0.336913	0.450628	0.876067

C) How does reducing the number of female patient datapoints affect the performance of the model? Why do you think that is?

In the general cohort, I observed slightly decreased performance across all measures and consistent performance metrics in the male cohort. However, the female cohort showed improved precision but had decreased F1 scores and recall.

In summary, I think that decreasing the number of female data points decreases the performance of the general cohort, doesn't significantly change the performance of the male cohort, and increases precision but decreases recall in the female cohort.

I think this occurs because reducing female data points skews the data to be less representative or generalizable for females. The initial testing and training sets already had slightly more males. Removing female data points increased the representativeness of males. While the overall metrics for the female cohort showed better precision, the decreased recall indicates that the model is predicting positive cases more accurately but is missing more true positives among females.

D) Do you expect to see the same or different results when varying the degree of missingness of some populations in other datasets or cohort splits? Briefly explain why you think so.

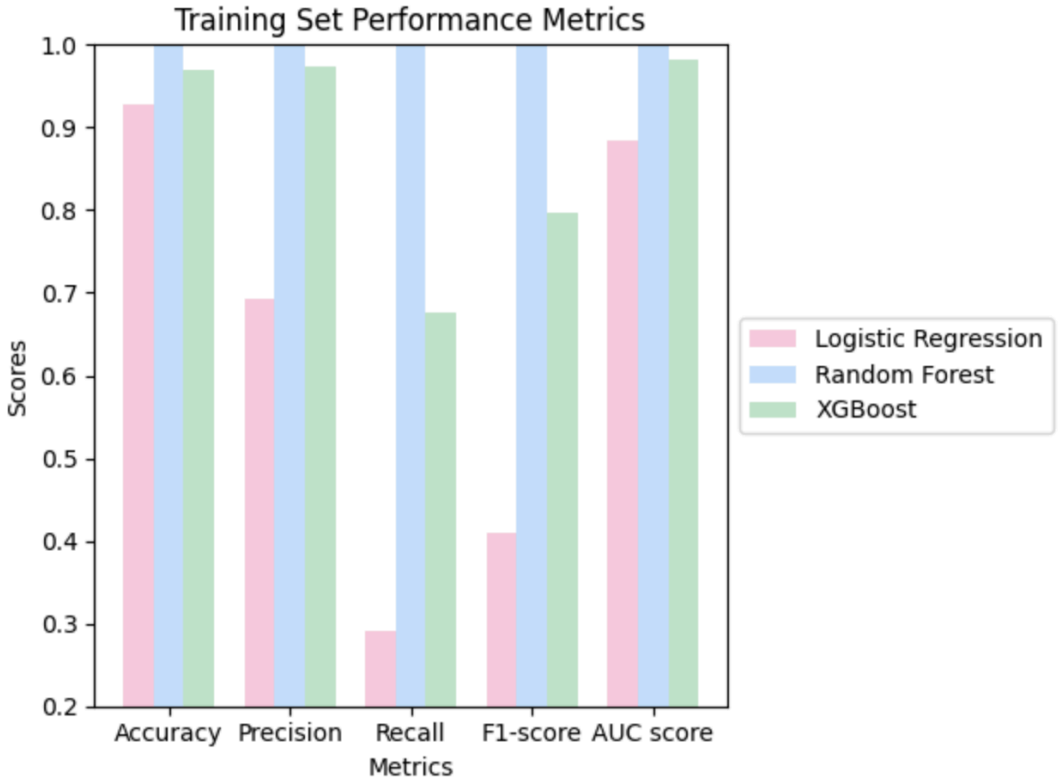
I would expect to see much more drastic results for the ethnicity and age cohorts. For ethnicity, Caucasians are significantly more represented than other ethnicities. I would expect worse model generalizability when reducing the data points of other

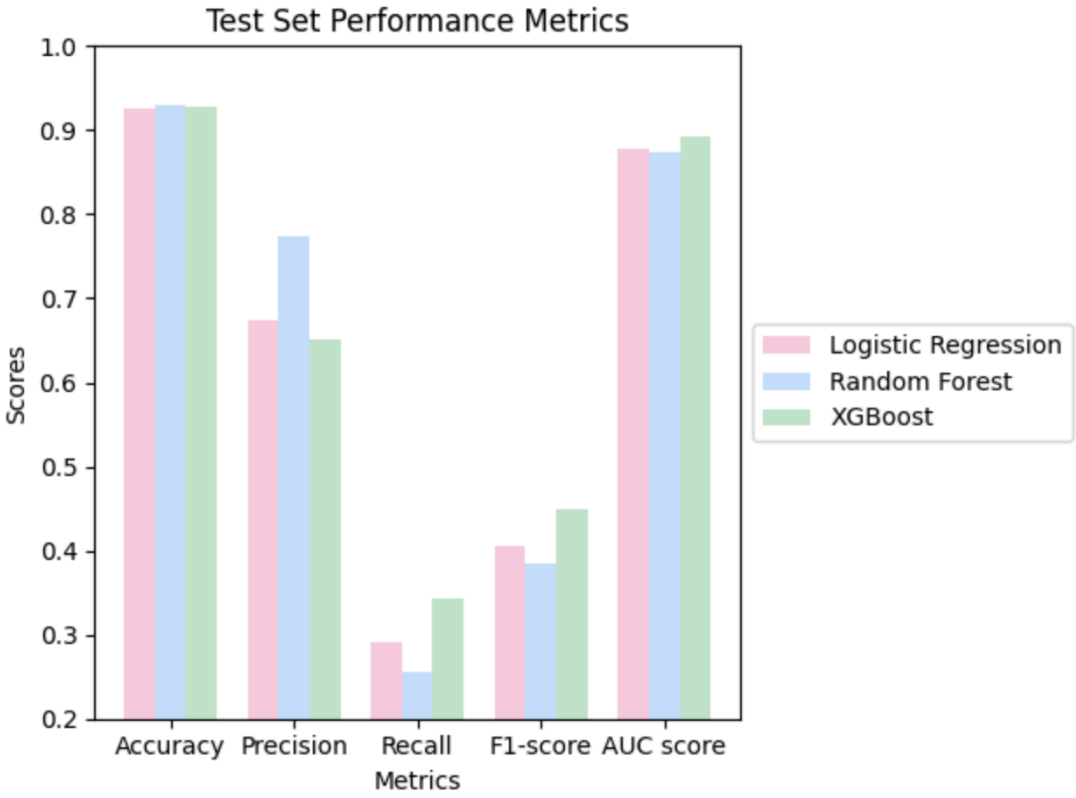
ethnicities. The same is true for the age cohort because it already doesn't represent younger people well. Decreasing data points in the age and ethnicity cohorts makes performance much worse because it makes the model more generalizable to the already most overrepresented groups.

E) **Another form of data imbalance is more implicit, and happens due to some patients undergoing fewer tests even if the number of patients is the same in each cohort. How do you expect increasing the degree of missingness of test results in some training datapoints would affect the performance of your model? Briefly justify your answe**r

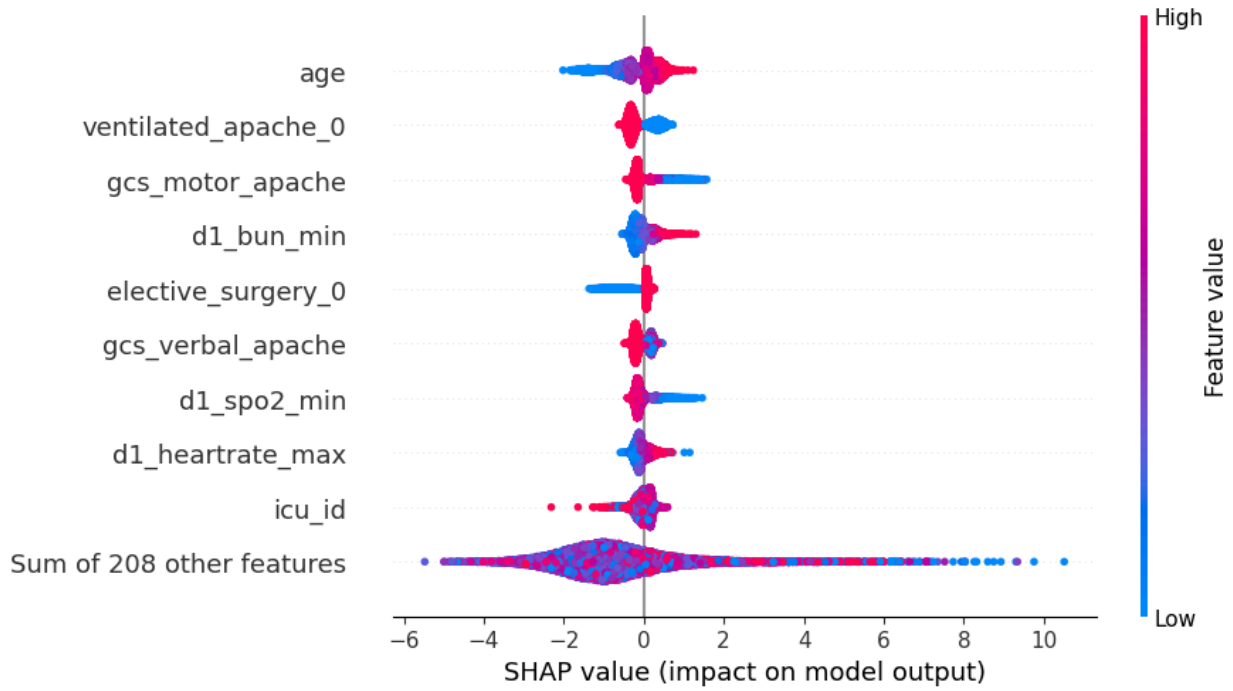
If more test results are missing for some patients, I would expect the model's performance to decrease because the model has **less information** to learn from. Missing test results mean the model can't fully understand the relationships between important health factors and the outcome. Alas, the model could have less accurate predicts that can contribute to worse health care outcomes due to inadequate performance.

F) Suggest one potential way of handling missingness of test results.
Multiple imputation!


Training Set Performance Metrics
Scores
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Accuracy
Precision
Recall
F1-score
AUC score
Metrics
Logistic Regression
Random Forest
XGBoost


Test Set Performance Metrics
Scores
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Accuracy
Precision
Recall
F1-score
AUC score
Metrics
Logistic Regression
Random Forest
XGBoost

Beeswarm plot



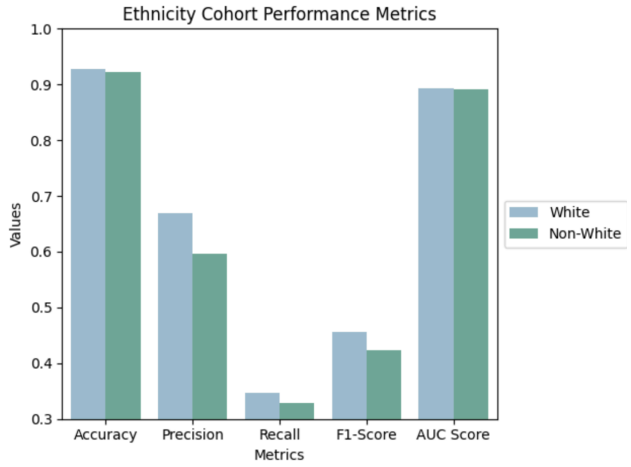
XGBoost cohort metrics

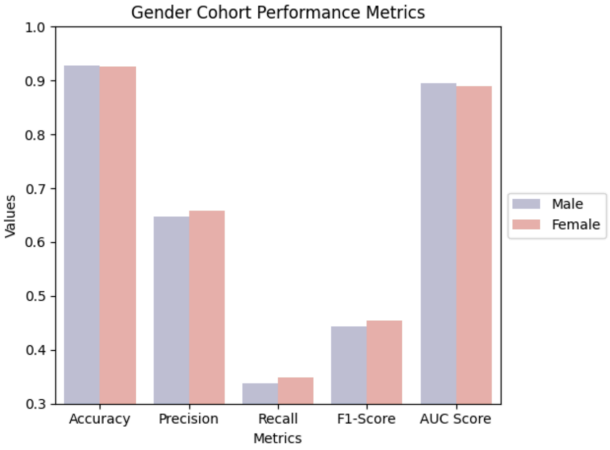
XGBoost performance metrics across cohorts:

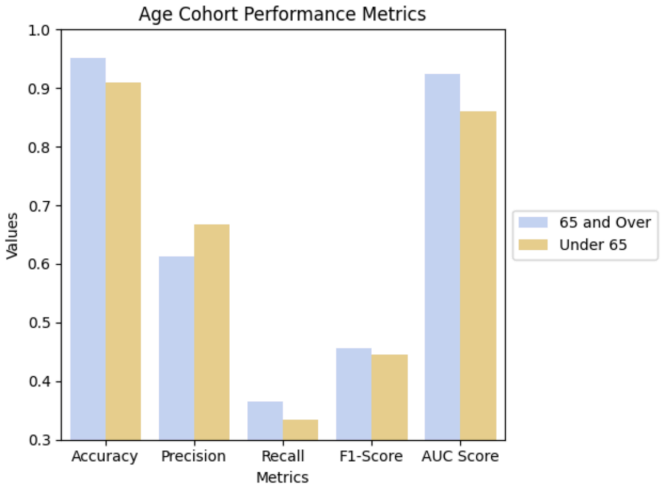
	Metric	White	Nonwhite	Male	Female	Age Above 65
0	Accuracy	0.928668	0.922680	0.927973	0.926505	0.951796
1	Precision	0.669841	0.596059	0.646925	0.657360	0.611888
2	Recall	0.346470	0.327913	0.337292	0.348118	0.364583
3	F1-score	0.456710	0.423077	0.443404	0.455185	0.456919
4	AUC score	0.892925	0.891871	0.894784	0.889987	0.923881

	Age Below 65
0	0.909716
1	0.667360
2	0.333680
3	0.444906
4	0.860946

Cohort Metrics Histogram:

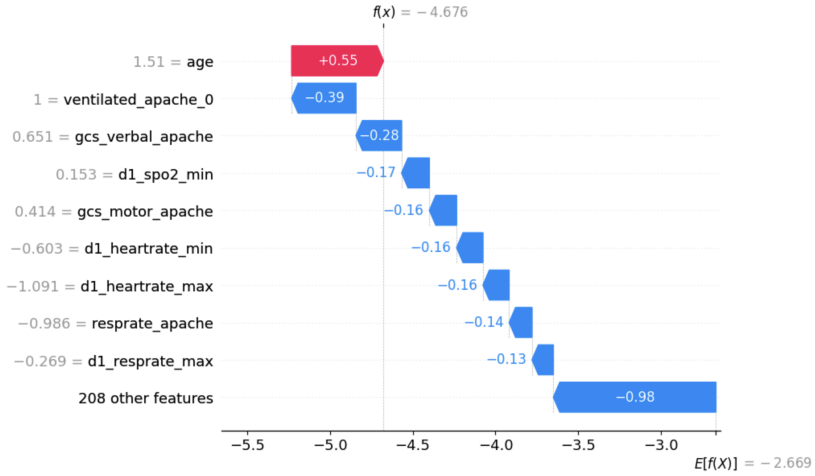




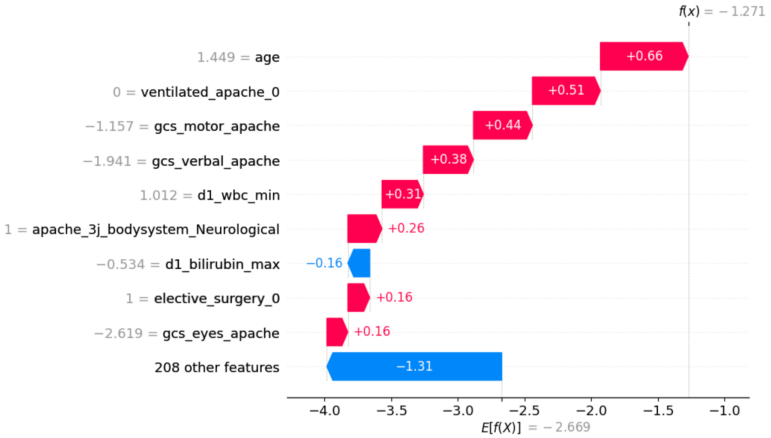


Waterfall plots

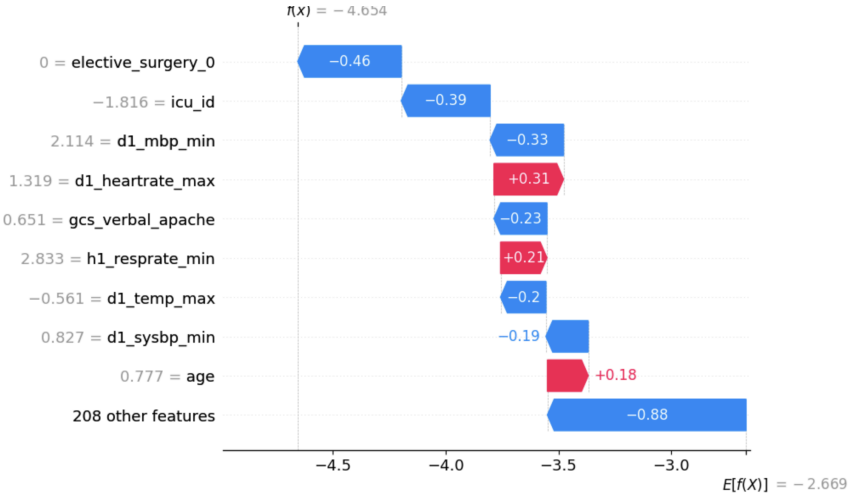
white_sample



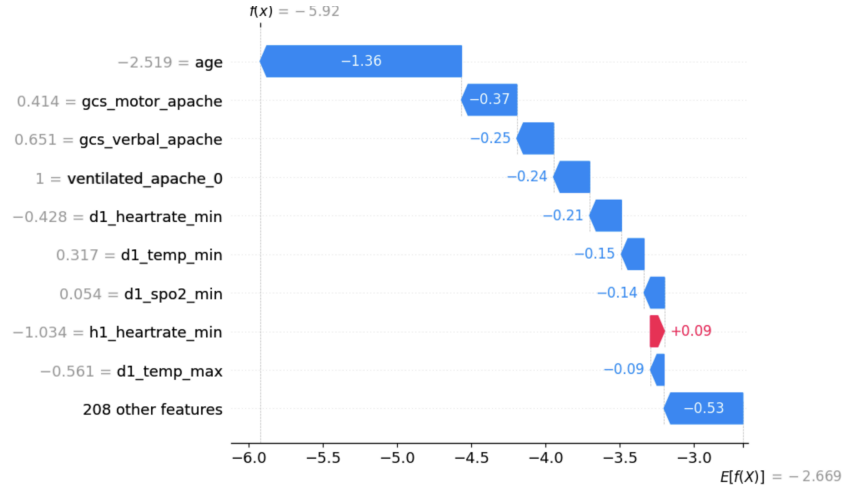
nonwhite_sample



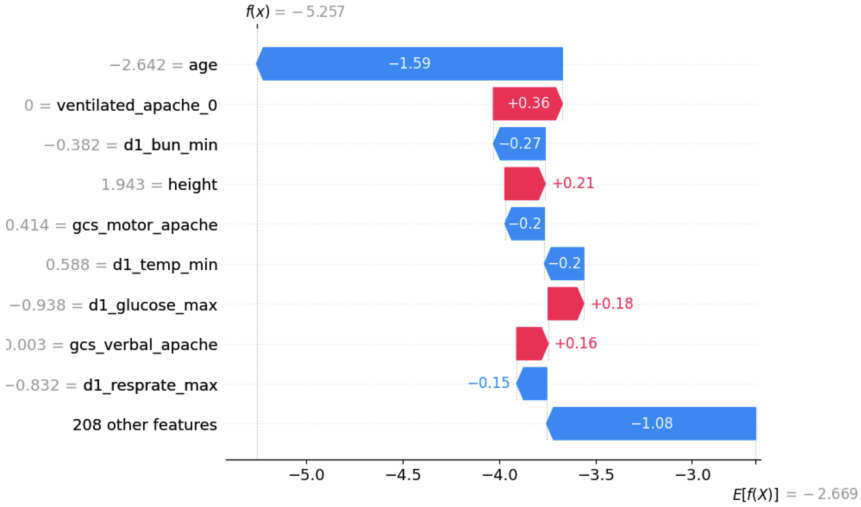
male_sample



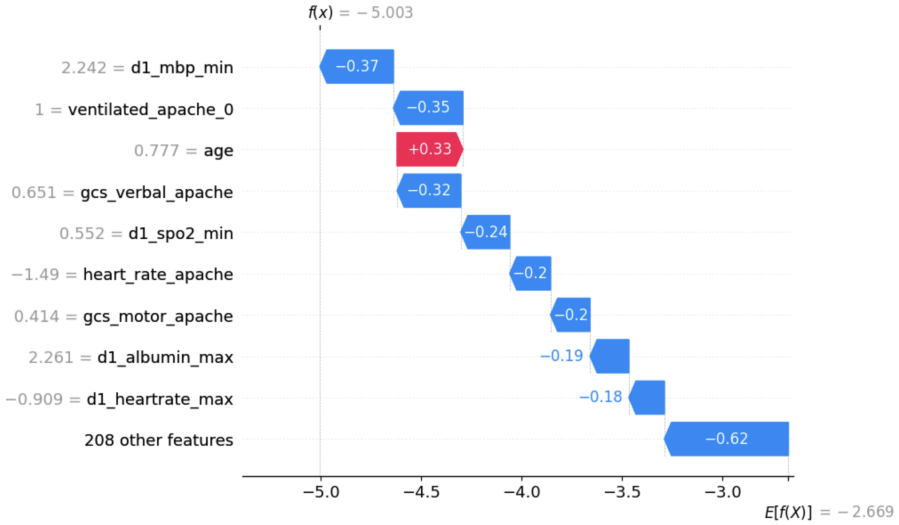
female_sample



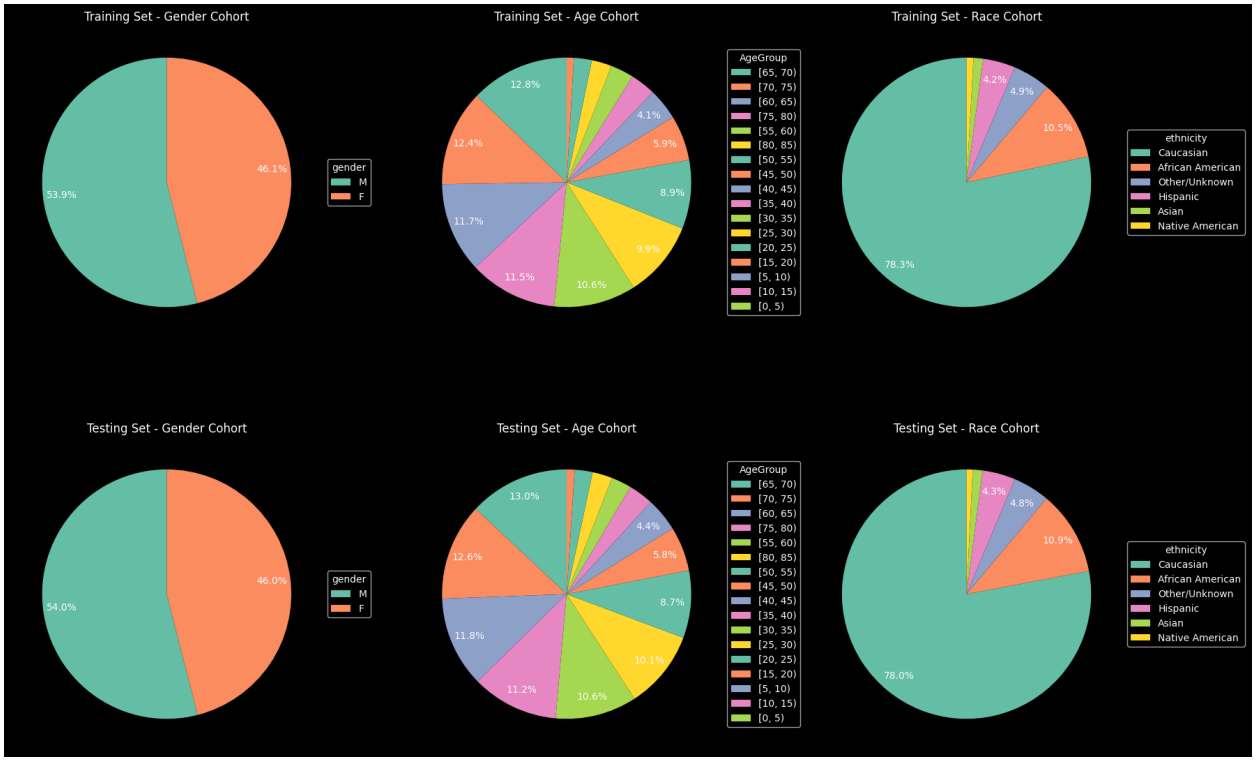
old_sample



young_age



Piecharts



Final table!!!!

	Reduction Percentage	Cohort	Accuracy	Precision	Recall	F1 Score	AUC
0	0.2	General	0.928256	0.663845	0.345936	0.454847	0.890143
1	0.2	Male Cohort	0.928680	0.656682	0.338480	0.446708	0.891408
2	0.2	Female Cohort	0.927759	0.671756	0.354362	0.463972	0.888636
3	0.4	General	0.928256	0.663450	0.346566	0.455298	0.889848
4	0.4	Male Cohort	0.928074	0.645089	0.343230	0.448062	0.889333
5	0.4	Female Cohort	0.928470	0.685039	0.350336	0.463588	0.890394
6	0.6	General	0.927166	0.651023	0.340895	0.447477	0.884680
7	0.6	Male Cohort	0.927973	0.641758	0.346793	0.450270	0.887461
8	0.6	Female Cohort	0.926220	0.662234	0.334228	0.444246	0.881417
9	0.8	General	0.927983	0.666667	0.335224	0.446122	0.880644
10	0.8	Male Cohort	0.928377	0.655012	0.333729	0.442172	0.884493
11	0.8	Female Cohort	0.927523	0.680217	0.336913	0.450628	0.876067