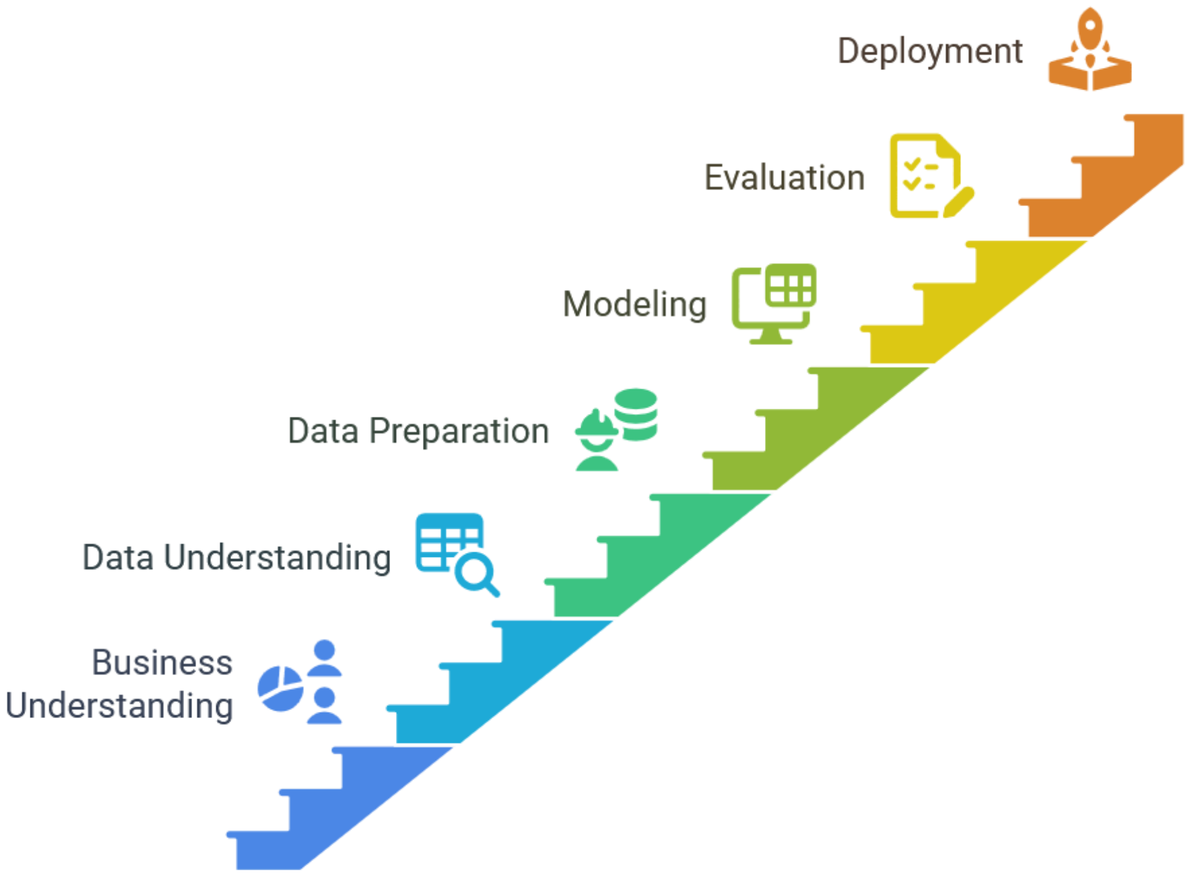
FINAL PRESENTATION

PREDICTION OF TRANSPLANT SURVIVAL RATES EQUITABLY FOR ALLOGENEIC HCT PATIENTS

Group O-เลี้ยง

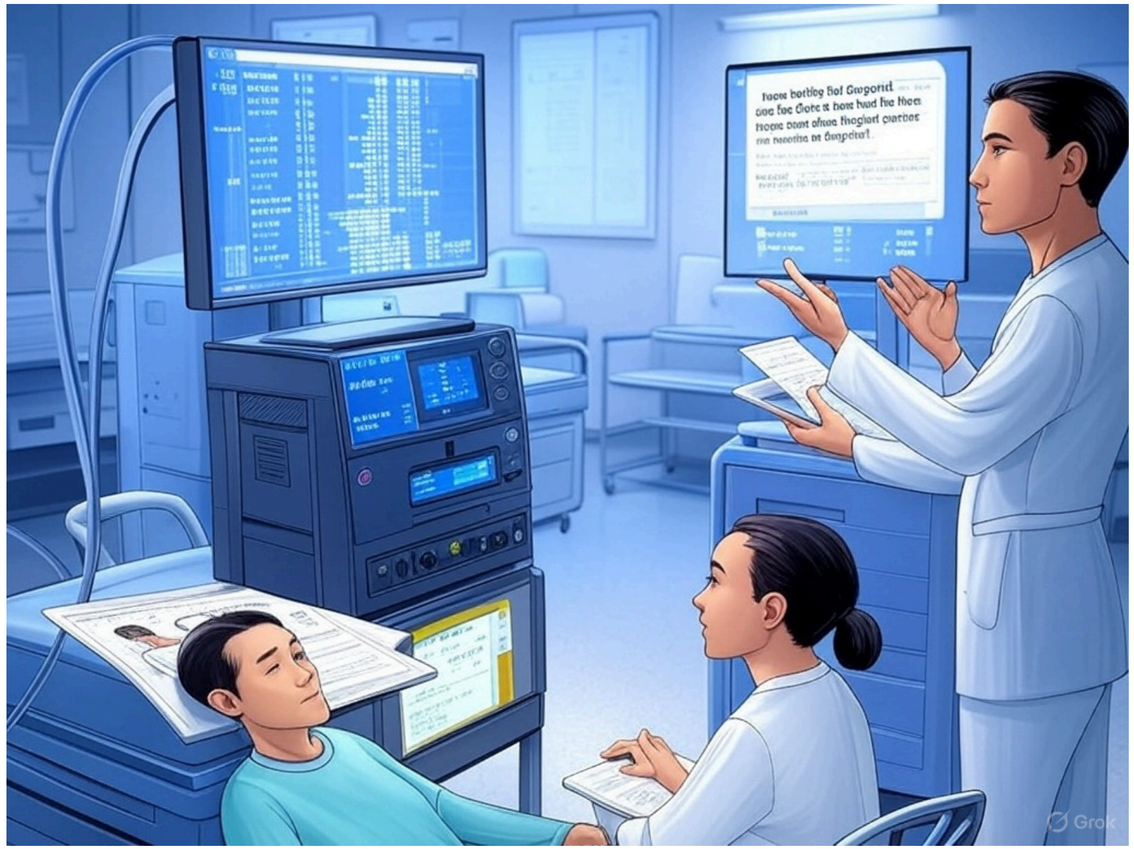
Asst. Prof. Suronapee Phoomvuthisarn, Ph.D.

FRAMEWORK


Deployment
Evaluation
Modeling
Data Preparation
Data Understanding
Business Understanding

BUSINESS UNDERSTANDING

Hematopoietic Stem Cell Transplantation (HSCT) is a critical treatment for patients with blood disorders, but event-free survival (EFS) varies significantly across patient and donor characteristics. Medical institutions and research centers need to optimize treatment strategies and donor selection to improve patient survival outcomes.



BUSINESS UNDERSTANDING

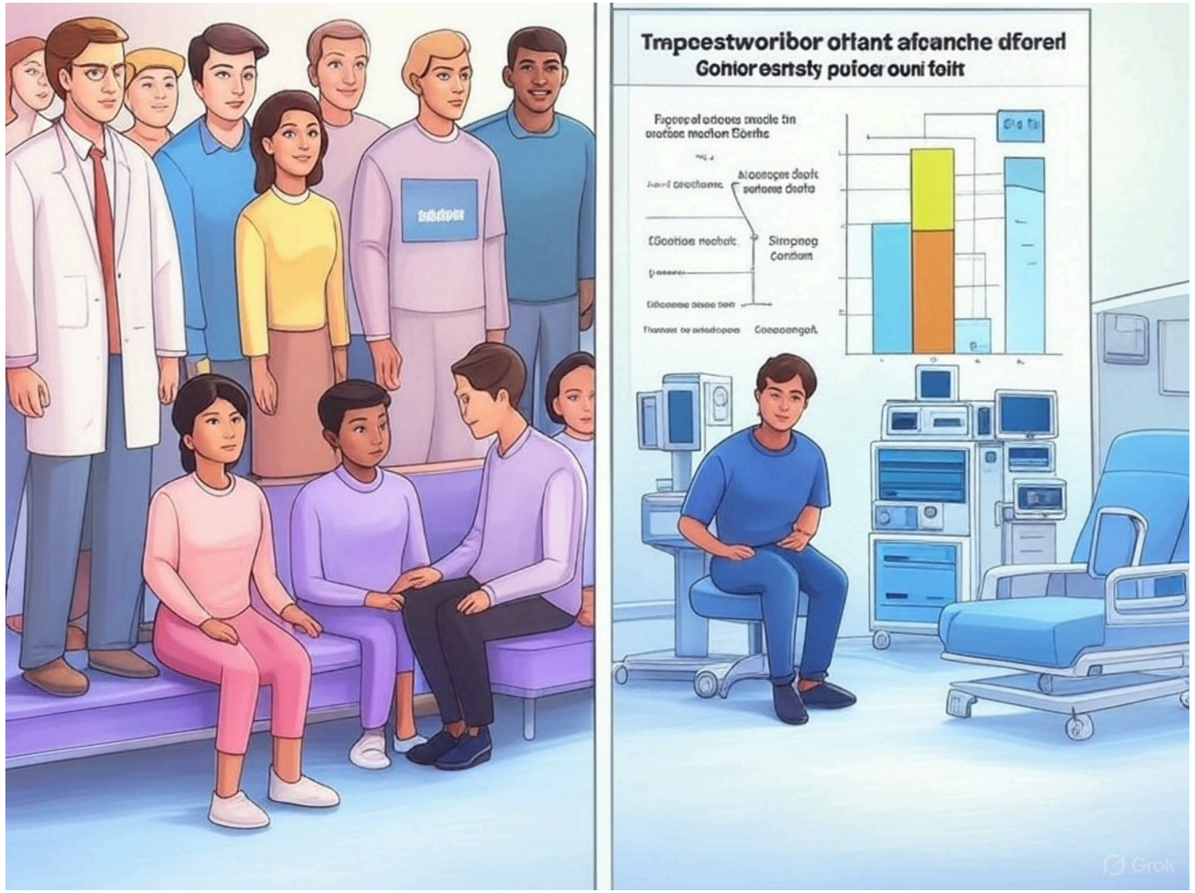
How can we develop a predictive model that accurately ranks survival outcomes for allogeneic Hematopoietic Cell Transplantation (HCT) patients while minimizing racial disparities in prediction accuracy to support equitable healthcare delivery?

Why?

Limited Donor Availability for Minority Groups:
HLA Matching Challenges: Successful HCT requires a close match of human leukocyte antigens (HLA) between donor and recipient. HLA types are inherited and vary widely among racial and ethnic groups.

Underrepresentation in Registries: Minority groups are underrepresented in bone marrow donor registries, reducing the likelihood of finding compatible donors. For example, White patients have a 77% chance of finding a match, while Black patients have only a 23% chance.

BUSINESS UNDERSTANDING



How would Data Science solution help?

1. Improved Donor Matching:

- A predictive model can analyze HLA compatibility more accurately by considering a wider range of genetic markers and demographic factors. This can increase match rates for minority patients.

2. Reducing Bias in Predictions:

- By using fairness-aware metrics like the Stratified Concordance Index, the model can ensure that predictions are equally accurate across different racial groups, supporting fairer clinical decision-making.

3. Efficient Resource Allocation:

- Hospitals can allocate resources, such as transplant slots and post-transplant care, more effectively by predicting which patients are most likely to benefit and which patient are considered to be risky, helping reduce disparities in access.

DATA UNDERSTANDING

Input Variable

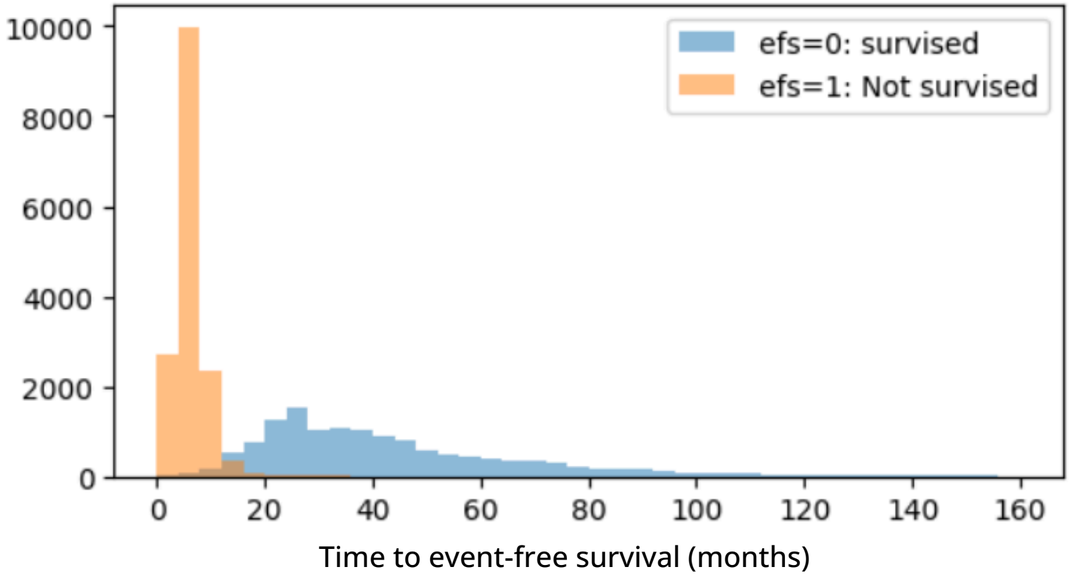
Patient Demographics and Health
ex.ethnicity, diabetes
Disease-Related Factors
ex. tce_match: T-cell Epitope Match
Transplant-Related Factors.
ex. dri_score: Disease Risk Index score

Response Variable

Event-free Survival
(1= not survived,
0 = survived)

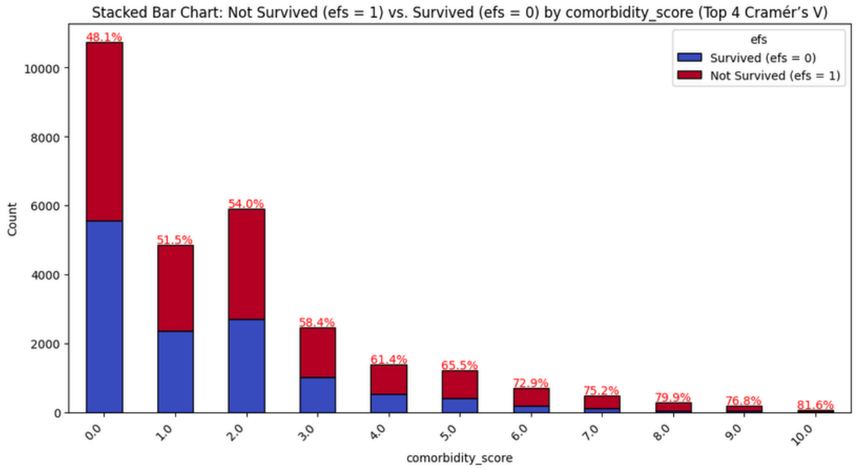
Time to event-free survival (months)

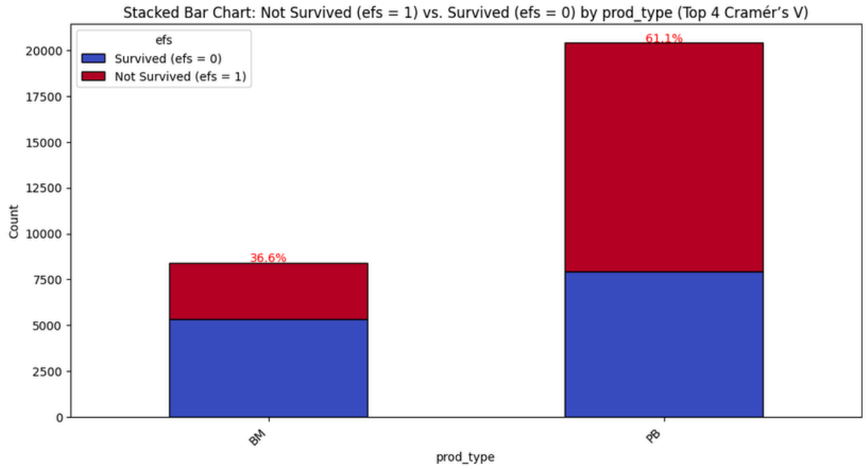
Response Variable Histogram



DATA UNDERSTANDING

Which patient characteristics strongly influence survival outcomes?



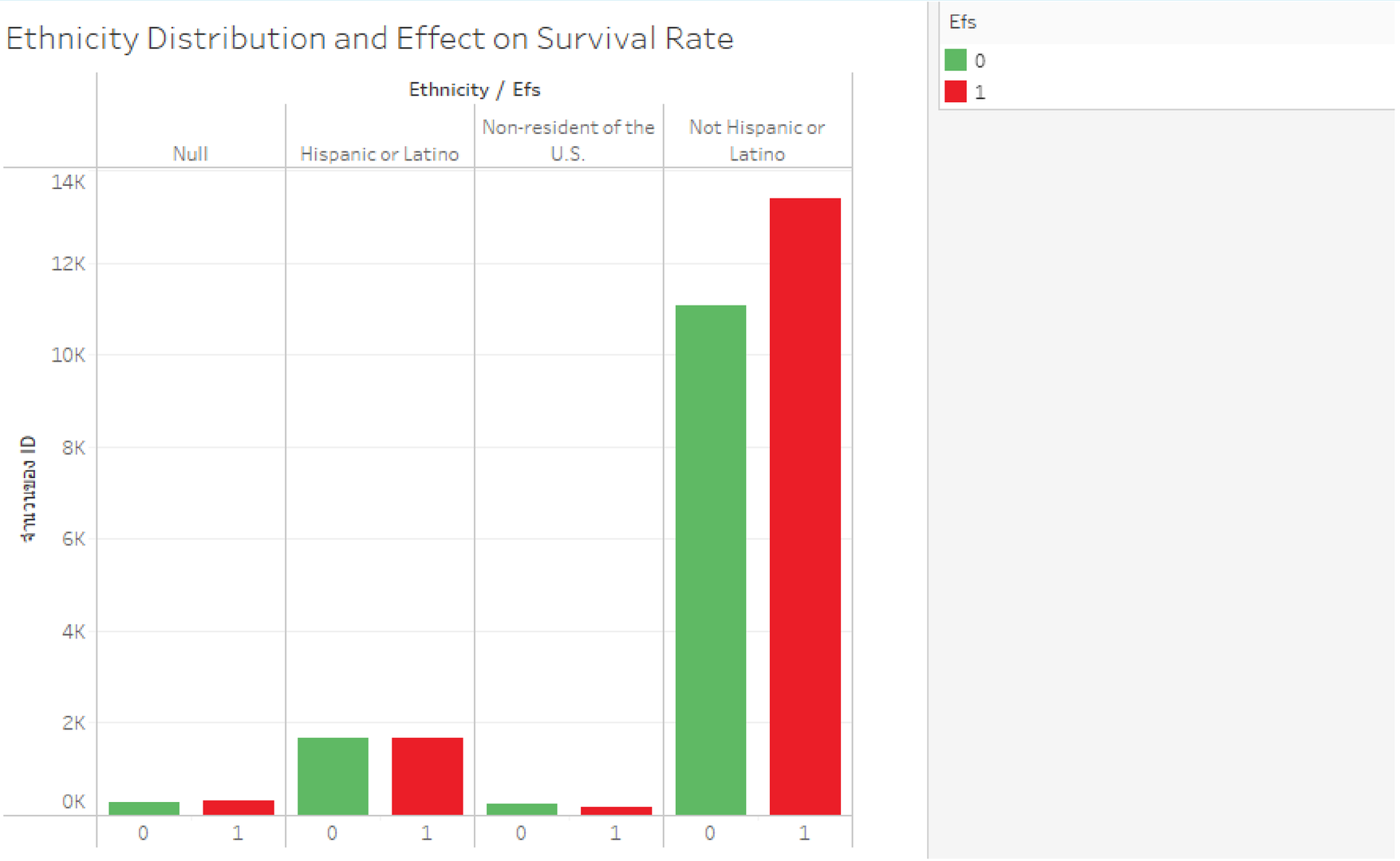


- Bone Marrow (BM) transplants show better survival rates than Peripheral Blood (PB).
- Patients with high comorbidity scores have significantly worse survival outcomes.
- Hospitals should adjust transplant protocols and patient selection criteria to improve survival for high-risk groups.

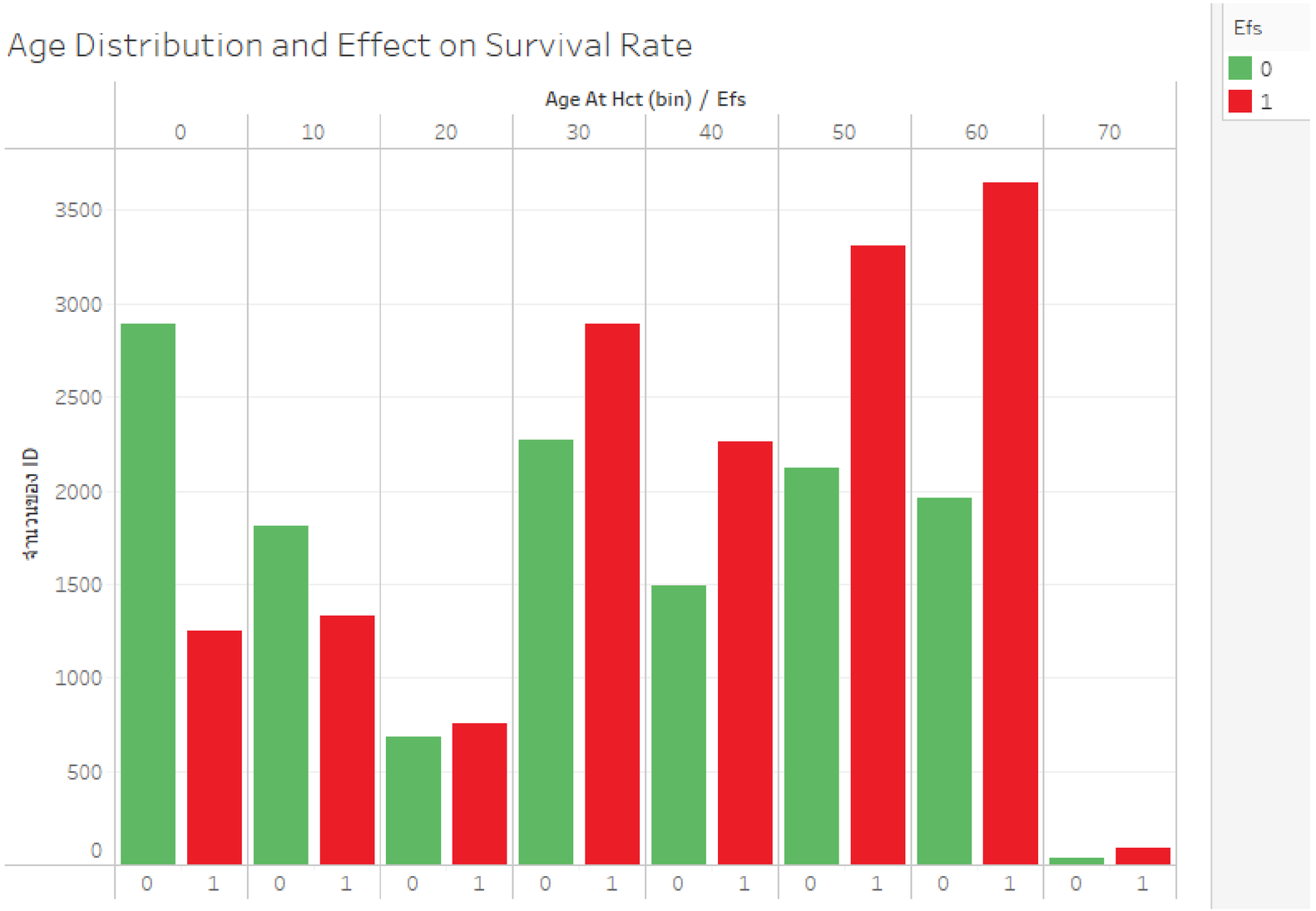
DATA UNDERSTANDING

DATA UNDERSTANDING

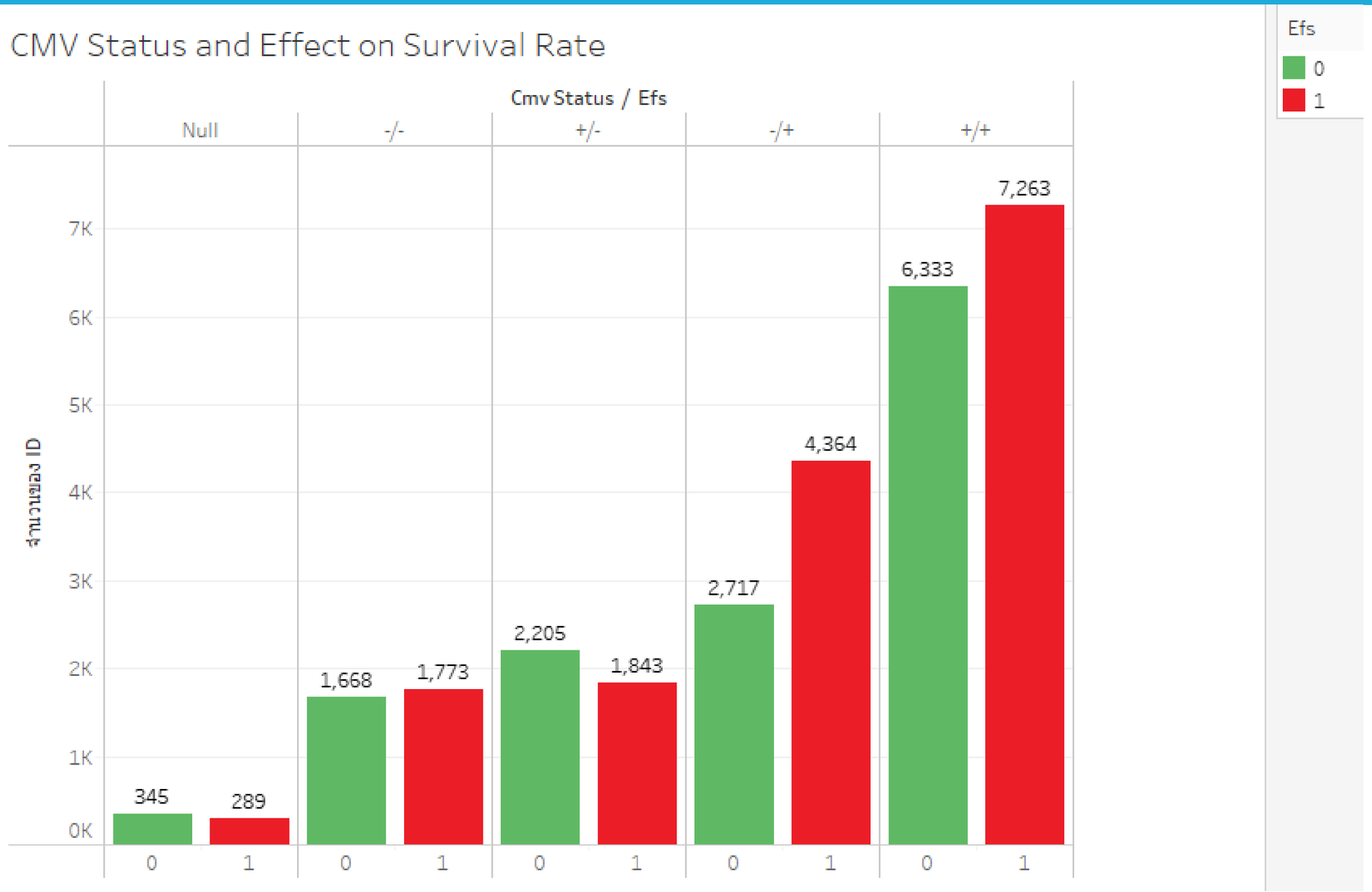
DATA UNDERSTANDING



DATA UNDERSTANDING



DATA UNDERSTANDING



DATA PREPARATION

DATA PREPROCESSING

1. Mapping Categorical Values

- Creates a dictionary to map old group values to new ones.
- Merges data with GVHD prophylaxis mapping.

2. Handling Missing Data

- Drops columns with more than 10% missing values.
- Removes irrelevant columns like 'ID' and 'prod_type'.

3.

SPLIT DATASET	Training	Validation	Test
Percent	60%	20%	20%
Row	17280	5760	5760
Column	45	45	45

9319

DATA PREPARATION

4. Handling Missing Values

- imputing numerical missing values (excluding efs & efs_time).
- Creates _missing indicator columns.
- Replaces missing values with the median of training data.

5. Normalization

- normalize numerical values by scales numerical features to [-1, 1] range.
- Prevents division by zero for constant columns.

6. Encoding Categorical Variables

- One-hot Encoding
- Converts categorical columns into dummy variables (get_dummies).
- Drops the first category to prevent multicollinearity.

7. Race Group Weighting

- Computes inverse frequency weights for race_group.
- Ensures balanced representation in modeling.

MODELING

Baseline Model

- Model Training:Fitting a Cox Proportional Hazards Model
- Target Variables: efs_time (duration), efs (event occurrence).
- Dropped Columns: Drop some missing indicator variables to reduce redundancy.
- Weights: weights_col to adjust for class imbalance.
- Robust Standard Errors: Used to improve model stability.

Model	Stratified C-Index	C-index Asian	C-index Native Hawaii	C-index More than 1	C-index Black	C-index Indian+ Alaska	C-index White
Cox Proportional Hazards Model	0.6437	0.6537	0.6725	0.6703	0.6497	0.6863	0.6335

Stratified Concordance Index $= \overline{\text{C-index}} - \sqrt{\frac{1}{g}\sum_{k=1}^{g}(\text{C-index}_k - \overline{\text{C-index}})^2}$

$$C\text{-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

with:

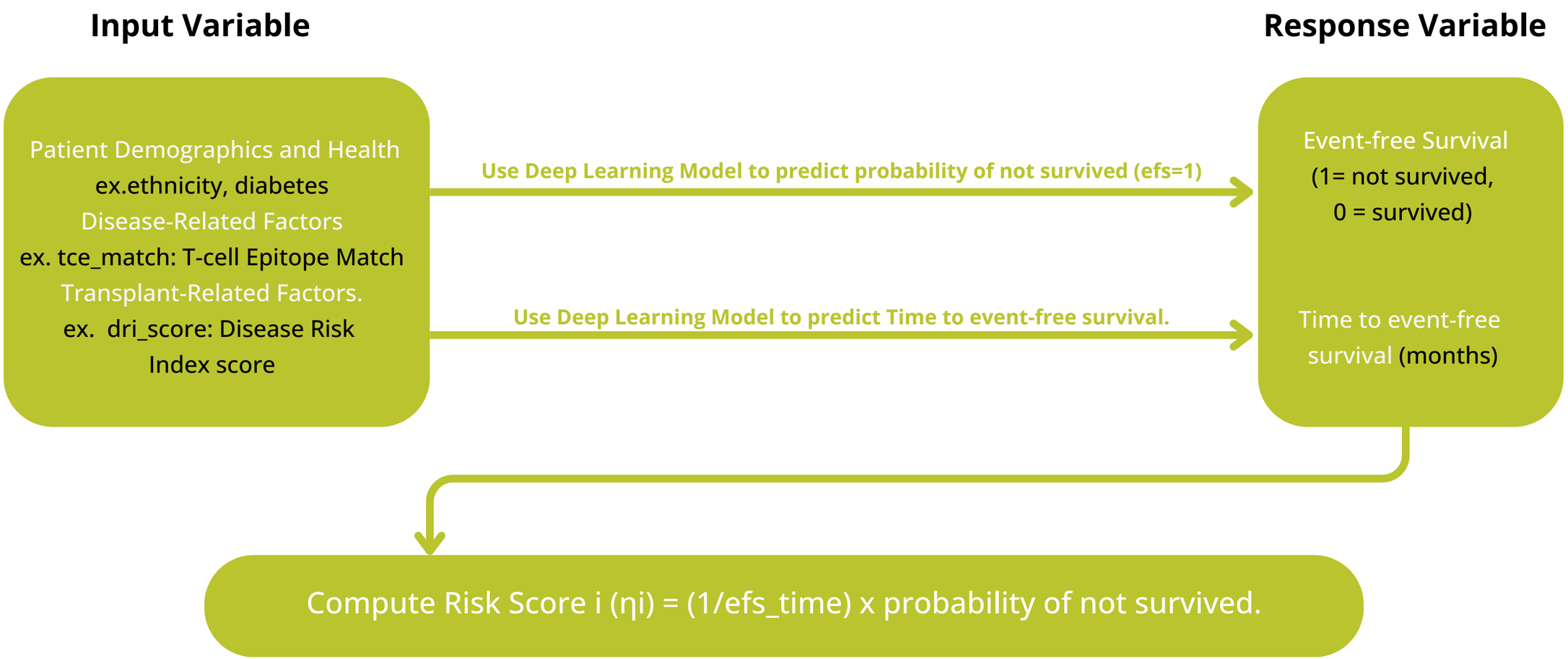
- η_i, the risk score of a unit i
- $1_{T_j<T_i} = 1, if\ T_j < T_i$ else 0
- $1_{\eta_j>\eta_i} = 1, if\ \eta_j > \eta_i$ else 0
- $\delta_j = 1, if\ efs_j = 1$ else 0

$$\overline{\text{C-index}} = \frac{1}{g}\sum_{k=1}^{g} \text{C-index}_k$$

with:

- g = Number of Total Race Group (More than one race, Asian, White, Black or African-American, American Indian or Alaska Native,Native Hawaiian or other Pacific Islander)

MODELING


Input Variable
Response Variable
Patient Demographics and Health
ex.ethnicity, diabetes
Disease-Related Factors
ex. tce_match: T-cell Epitope Match
Transplant-Related Factors.
ex. dri_score: Disease Risk Index score
Use Deep Learning Model to predict probability of not survived (efs=1)
Use Deep Learning Model to predict Time to event-free survival.
Event-free Survival (1= not survived, 0 = survived)
Time to event-free survival (months)
Compute Risk Score i (ηi) = (1/efs_time) x probability of not survived.

MODELING

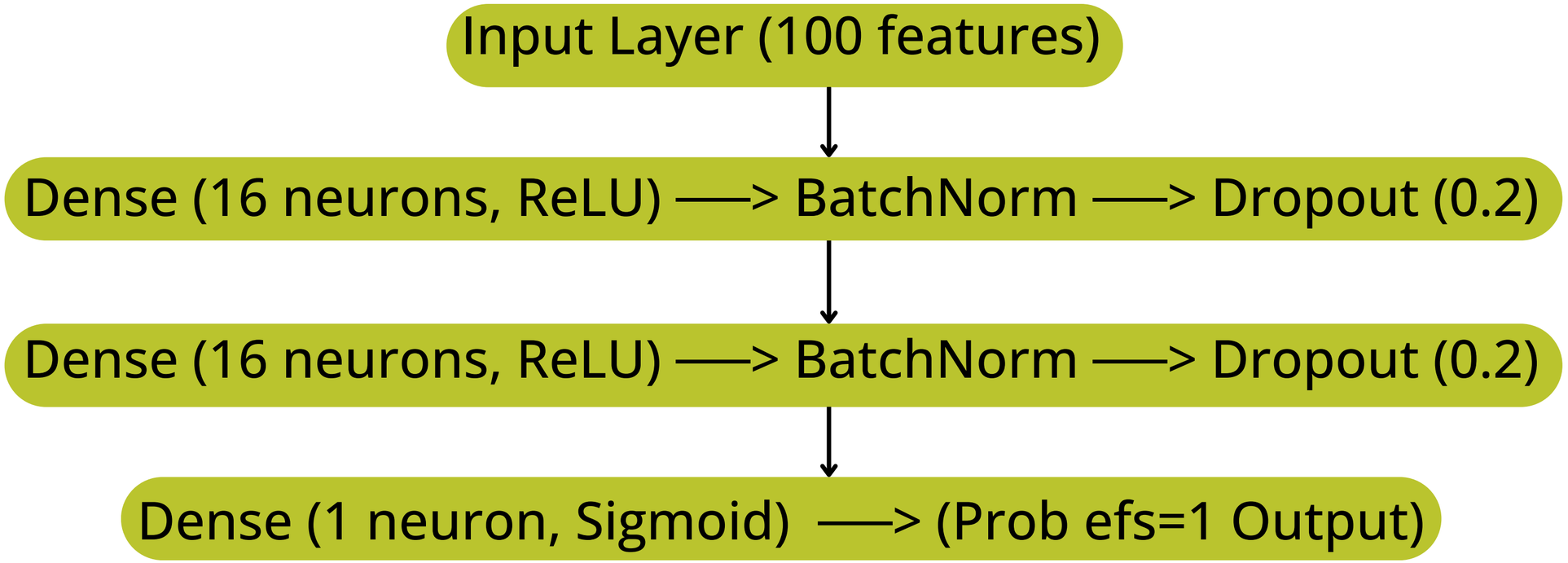
Classification Model (efs=1)

- use the same dataset from the baseline model.

	Training	Validation	Test
Percent	60%	20%	20%
Row	17280	5760	5760
Column	100	100	100

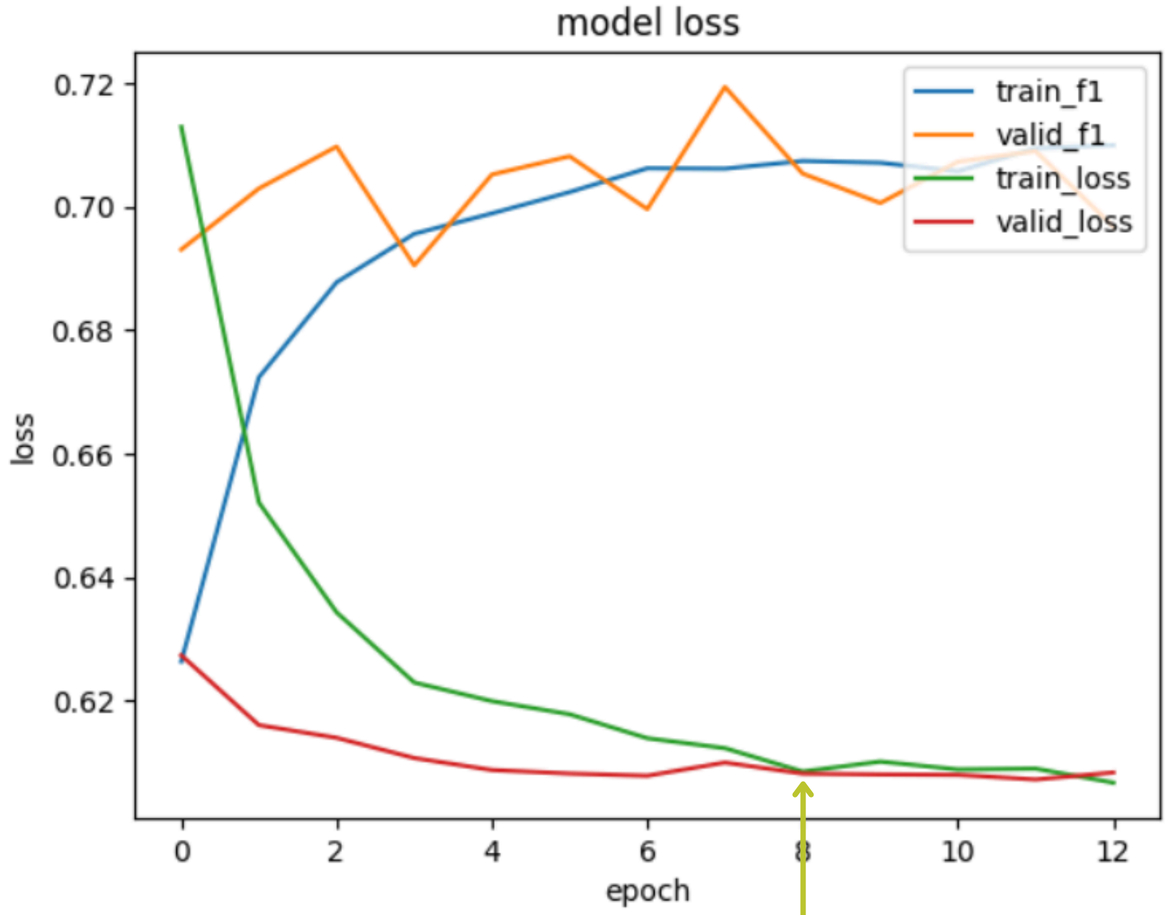
MODELING

Classification Model (efs=1)



Classification Model (efs=1)

- Architecture:
  - Input Layer: 100 features
  - Hidden Layers: 2x Dense (16 units, ReLU)
  - Batch Normalization & Dropout (0.2)
  - Output: Sigmoid Activation
- Training Details:
  - Loss: Binary Crossentropy
  - Optimizer: Adam
  - Metric: F1-Score
  - Early Stopping: patience=5, monitor=val_f1_score
  - Best Model Saved: best_model.h5

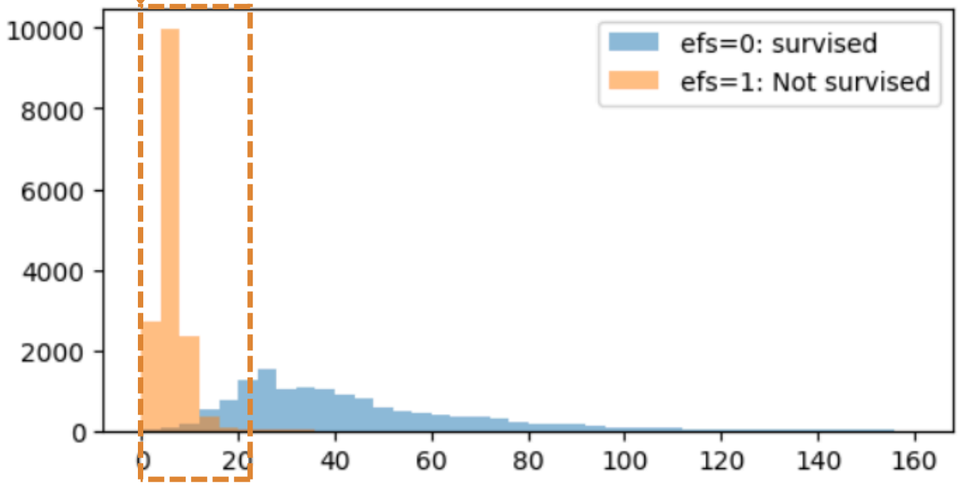


Epoch 8/50: loss: F1 Score: 0.7111 / loss: 0.6101

Test set F1 Score: 0.7211842483472262

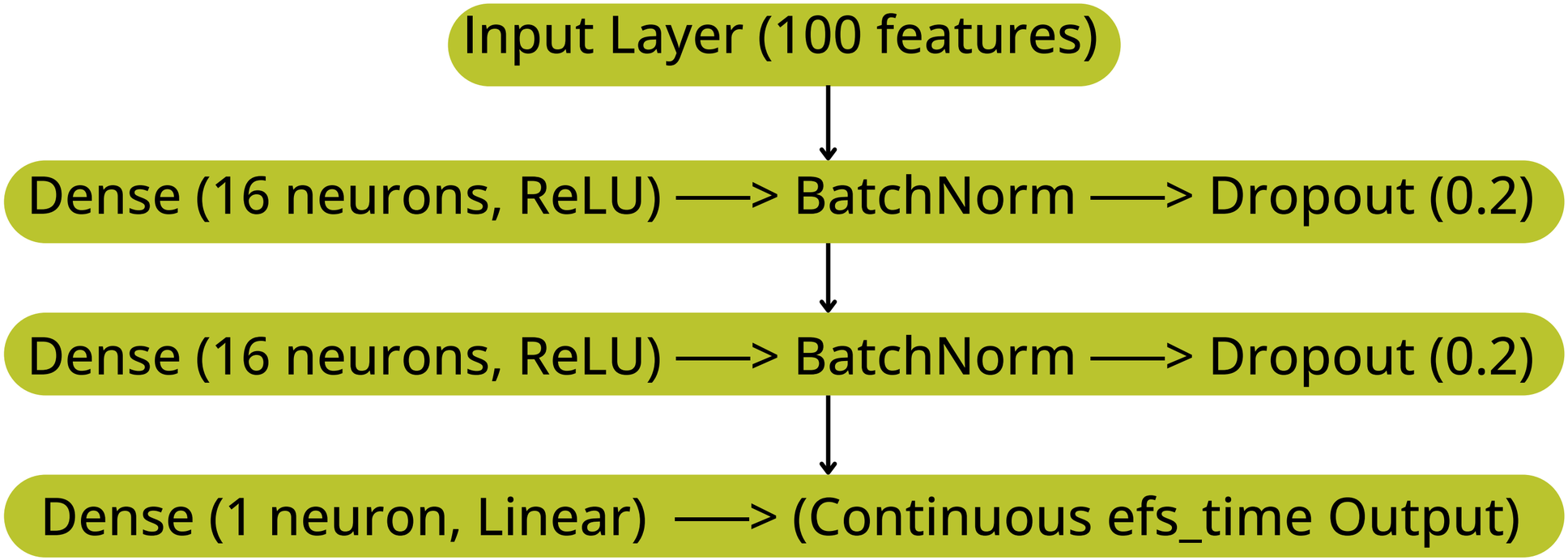
Regression Model (efs_time)

1. Filters Rows Where **efs = 1**
   - The regression model only predicts survival time (efs_time) for cases where the event occurred .
2. Drops Columns
   - efs_time: Target variable (separate for regression).
   - efs: Not needed for regression since we're only keeping efs = 1 cases.
   - weights_col: Used separately for sample weighting.
   - Several _missing columns: Indicators for missing values, likely not useful for regression.
3. The filtered dataset is stored in:
   - X_train_nn_efs_time: Training features
   - X_valid_nn_efs_time: Validation features
   - X_test_nn_efs_time: Test features
4. Extracts Sample Weights (weights_col) for Regression
   - weights_col_efs_time: Weights corresponding to cases where efs = 1.
5. Scales the Target Variable (efs_time) Using MinMaxScaler



MODELING

Regression Model (efs_time)



MODELING

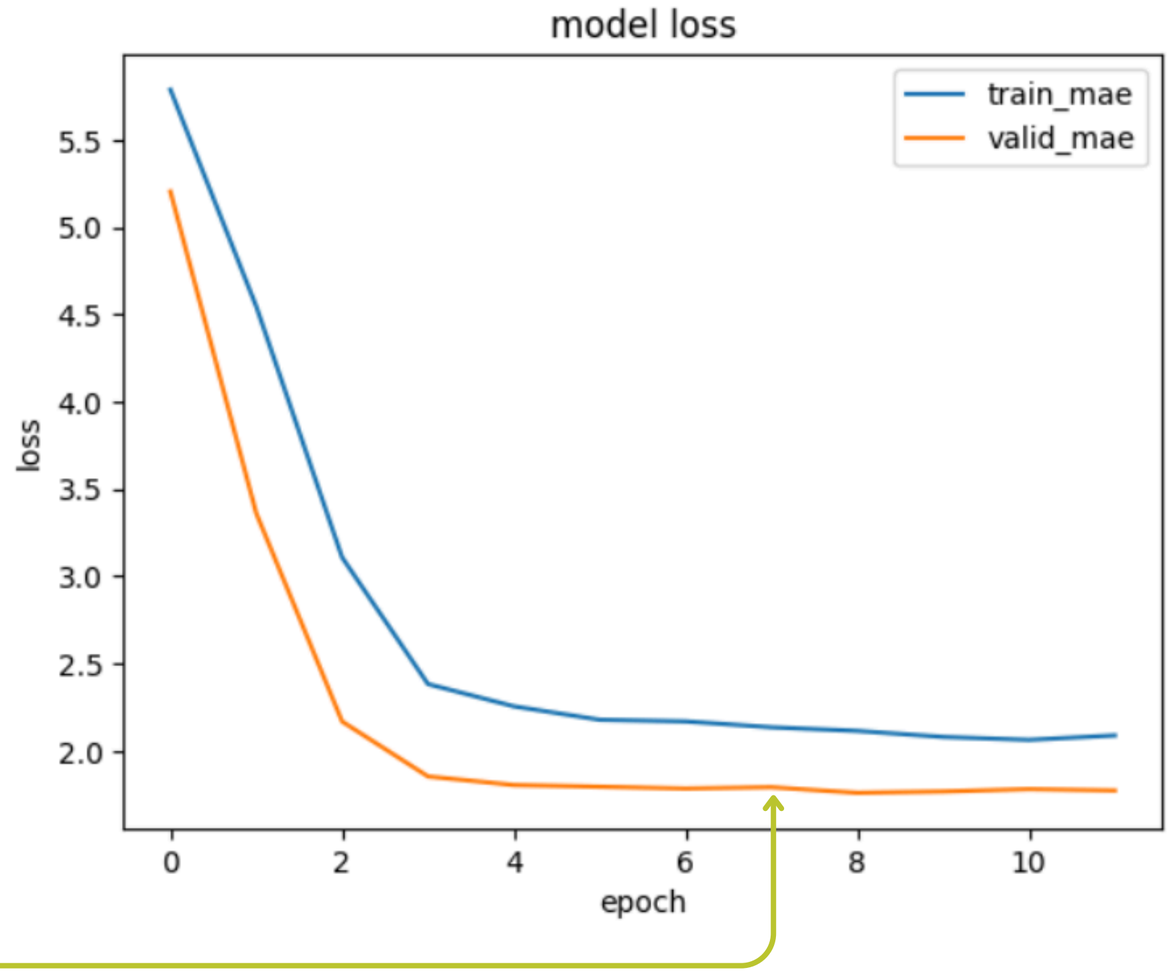
Regression Model (efs_time)

- change the dataset by Filters Rows Where efs = 1

	Training	Validation	Test
Percent	60%	20%	20%
Row	9319	3,106	3,106
Column	100	100	100

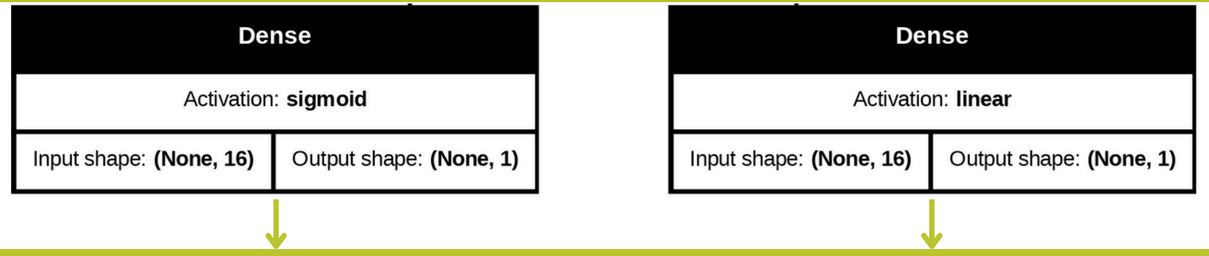
Regression Model (efs_time)

- Architecture:
  - Input Layer: 100 features
  - Hidden Layers: 2x Dense (16 units, ReLU)
  - Batch Normalization & Dropout (0.2)
  - Output: Linear Activation
- Training Details:
  - Loss: Mean Squared Error (MSE)
  - Optimizer: Adam
  - Metric: Mean Absolute Error (MAE)
  - Early Stopping: patience=5, monitor=val_loss
  - Best Model Saved: best_model.h5



Epoch 7/50: loss: 9.9280 / mae: 2.1274

MODELING



Compute Risk Score i (ηi) = (1/efs_time) x probability of not survived.

Model	Stratified C-Index	C-index Asian	C-index Native Hawaii	C-index More than 1	C-index Black	C-index Indian+ Alaska	C-index White
Train	0.6465	0.6767	0.6515	0.6584	0.6458	0.6610	0.6479
Validation	0.6295	0.6500	0.6245	0.6526	0.6253	0.6611	0.6458
Test	0.6435	0.6501	0.6661	0.6607	0.6445	0.6771	0.6399

Model Tuning using keras-tuner

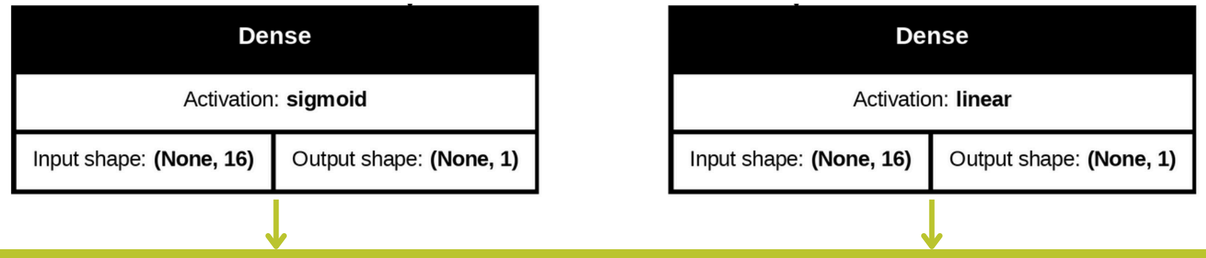
- Hyperparameter Tuning: Optimize model performance using Bayesian Optimization.

Bayesian Optimization (kt.BayesianOptimization)

- Searches optimal units, activation function, dropout rate, learning rate, number of layers.
- Validation Strategy: 60% Train | 20% Validation | 20% Test
- Early Stopping & Model Checkpointing used to prevent overfitting.

Best Hyperparameters	units	activation	dropout	lr
Classification Model (efs=1)	56	Leaky ReLU	0.2	0.0075
Regression Model (efs_time)	32	ReLU	0.0	0.0043

MODELING



Compute Risk Score i (ηi) = (1/efs_time) x probability of not survived.

Model	Stratified C-Index	C-index Asian	C-index Native Hawaii	C-index More than 1	C-index Black	C-index Indian+ Alaska	C-index White
Deep Learning After Hyperparameter Tuning	0.6450	0.6490	0.6772	0.6643	0.6496	0.6769	0.6395

The final model improved by 0.233% compared to the model before tuning.

EVALUATION

Stratified Concordance Index $= \overline{\text{C-index}} - \sqrt{\frac{1}{g}\sum_{k=1}^{g}(\text{C-index}_k - \overline{\text{C-index}})^2}$

$$C\text{-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

with:

- η_i, the risk score of a unit i
- $1_{T_j<T_i} = 1, if\ T_j < T_i$ else 0
- $1_{\eta_j>\eta_i} = 1, if\ \eta_j > \eta_i$ else 0
- $\delta_j = 1, if\ efs_j = 1$ else 0

$$\overline{\text{C-index}} = \frac{1}{g}\sum_{k=1}^{g} \text{C-index}_k$$

with:

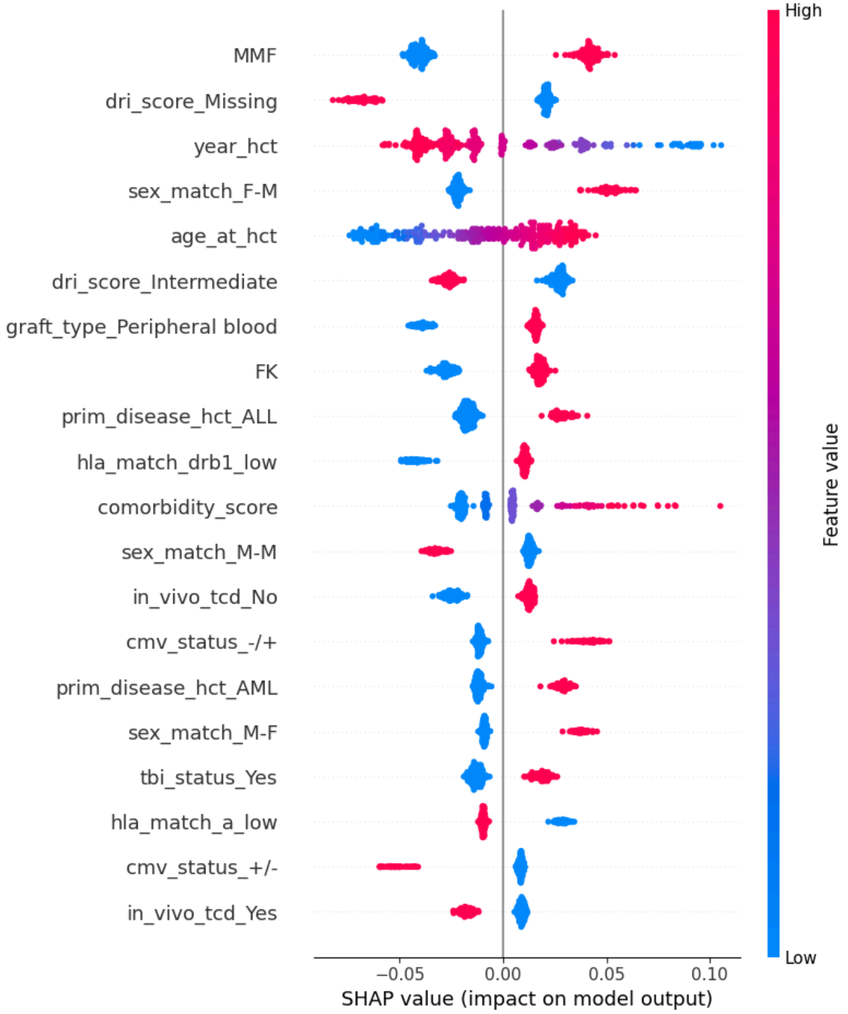
- g = Number of Total Race Group (More than one race, Asian, White, Black or African-American, American Indian or Alaska Native,Native Hawaiian or other Pacific Islander)

EVALUATION

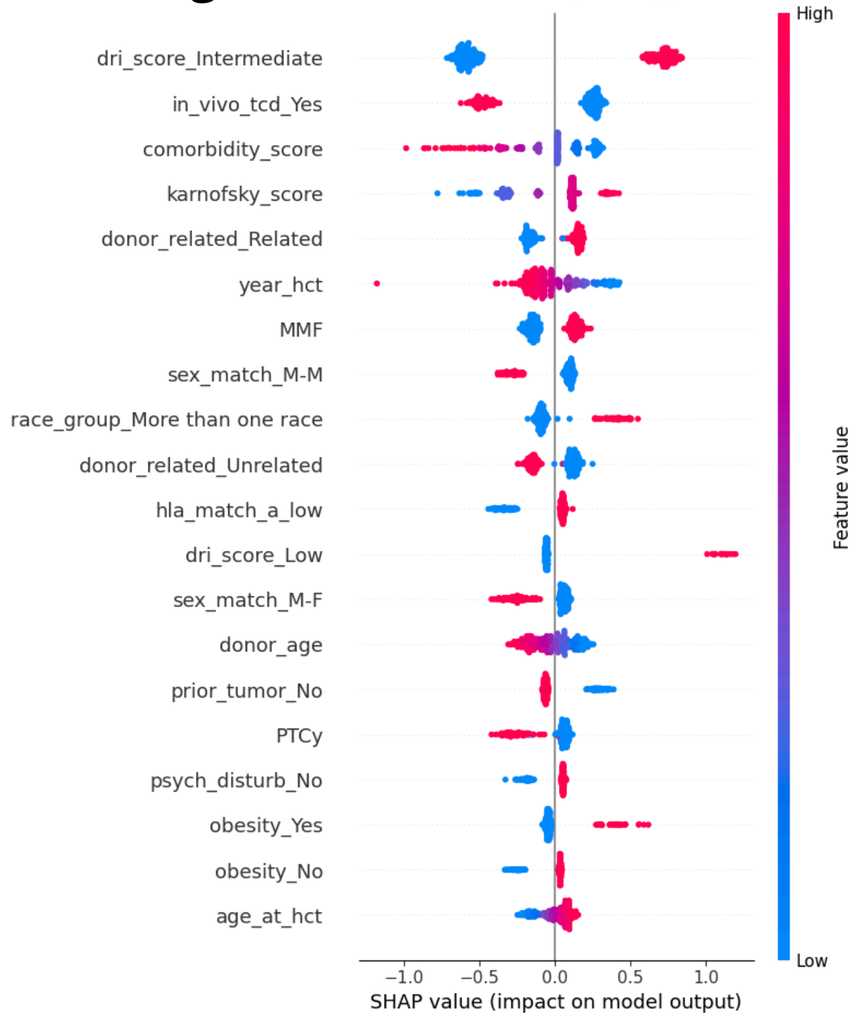
Model	Stratified C-Index	C-index Asian	C-index Native Hawaii	C-index More than 1	C-index Black	C-index Indian+ Alaska	C-index White
Baseline Model	0.6437	0.6537	0.6725	0.6703	0.6497	0.6863	0.6335
Final Model	0.6450	0.6490	0.6772	0.6643	0.6496	0.6769	0.6395

EVALUATION

Shap for: Classification Model (efs=1)



Regression Model (efs_time)



MEMBER
6542091026 พีรพัฒน์ ติ้นเติมทรัพย์
6542111026 วายุ ลิ้มสุวรรณ
6542113226 วิศรุต พีรชัยเดโช
6542116126 สถาปนันท์ เสนารักษ์
6542023426 ชนนน วงษ์คนดี

Q & A

APPENDIX