# ETC3550 Project – ICU Mortality Prediction

## Name: Minh Chung Ngo

### Student ID: 34667830

### Data Visualization Summary

Initial visualizations highlighted a class imbalance in the HOSPITAL_EXPIRE_FLAG variable, with more patients recorded as deceased (1) than survived (0). The distribution of vital signs and demographic features revealed clear differences between the two groups. Notably, patients who died tended to have higher heart and respiratory rates, elevated glucose levels, and were generally older at admission. Conversely, they exhibited lower systolic and diastolic blood pressure and lower minimum oxygen saturation ($SpO_2$). A correlation heatmap revealed strong inter-correlations among groups of related variables, including Glucose (Mean, Max, Min), $SpO_2$ (Mean, Max, Min), Temperature (Mean, Max, Min), Blood Pressure measures (Systolic, Diastolic, Mean), and Heart Rate (Mean, Max, Min). Moreover, some words in text (Diagnosis) also have relationship with death and survive – which was calculated by CNDE in the preprocessing step. These findings suggest that physiological instability, age and diagnosis that they have are important indicators of mortality risk.

### Data Preprocessing

#### Select statistically significant predictors

To reduce multicollinearity and prepare structured data for modeling, generalized linear models (GLMs) were used to identify statistically significant predictors among highly correlated numeric features. For each group of correlated variables, the one with the lowest p-value was retained. The selected predictors—such as HeartRate_Mean, SysBP_Min, DiasBP_Mean, MeanBP_Min, RespRate_Mean, TempC_Mean, SpO2_Min, and Glucose_Mean were chosen for use in models sensitive to high dimensionality.

#### Text Analysis

I applied diagnosis-based text analysis using a modified version of the Crucial Nursing Description Extractor (CNDE). This involved tokenizing and removing stopwords from the DIAGNOSIS, SHORT_DIAGNOSE, and LONG_DIAGNOSE columns, computing word-level log-likelihood ratios (LLR) to measure association with mortality (HOSPITAL_EXPIRE_FLAG), and generating a text score for each patient by summing the LLR values of words in their text. These scores were added as structured numeric features to both the training and testing datasets.

#### Diagnosis Analysis

To deal with ICD9_diagnosis table, I decided to select these codes appearing more than 20 times before transforming into a wide binary count matrix via pivoting and then merged with the patient-level datasets. All missing values were filled with zeros, and categorical variables such as admission type, insurance type, and first care unit were one-hot encoded

#### Scale Data

Finally, continuous numeric features were standardized using z-score scaling (mean = 0, standard deviation = 1) to ensure uniformity in feature magnitude, which is particularly important for distance-based and regularized models (Logistic regression, neural network).

### Cross-validation

To evaluate out-of-sample performance, the training data was further split into 5 folds using stratified cross-validation to ensure a balanced proportion of survivors and non-survivors in each fold. Two sets of folds were created: one that includes the high-dimensional ICD9 diagnosis features and one that excludes them. This approach was necessary because some models may struggle with a large number of sparse categorical variables. By separating the folds accordingly, models could be trained and validated fairly while accounting for their limitations.

## Model Evaluation

### Logistic Regression

The logistic regression model demonstrated poor performance both in-sample and out-of-sample. While the in-sample accuracy was around 0.89, the ROC AUC was only approximately 0.23, indicating that the model failed to effectively distinguish between patients who survived and those who did not. This large discrepancy suggests that the model is likely overfitting to the majority class and lacks generalizability. As a result, the predictions produced by logistic regression are not considered reliable for this dataset.

### Decision Tree

The decision tree model achieved strong performance with an in-sample accuracy of 0.91 and ROC AUC of 0.86, indicating excellent fit on the training data. When evaluated on the out-of-sample test set, the model maintained good generalization with an accuracy of 0.89 and ROC AUC of 0.76. The results demonstrate the model's solid predictive ability on unseen data, making it a reliable tool for predicting hospital mortality risk.

### Neural Network

I trained a fully connected (dense) neural network for binary classification using Keras. The architecture included: An input layer (128 units, using kernel regularization (0.001), RELU activation, and dropout 0.2 rate for regularization. Following that, some hidden layers were added (64, 64, 32, 32 layers), each followed by batch normalization and RELU activations as well, to improve the  stability and expressiveness in model. The final layer is 1 with sigmoid activation to achieve the probabilities for the binary outcome. The model was compiled with Adam optimizer – learning rate 0.00005, binary cross-entropy loss, and monitored both accuracy and AUC metrics.

Before applying dataset to the model, I also converted it into matrix format. The performance results are better than the previous models, which were 0.96 in accuracy and 0.98 in ROC AUC when doing in-sample. The out-sample results are little lower, 0.92 in accuracy and 0.93 in ROC AUC, which indicate the model fits the training data very well and generalizes strongly to unseen data with only a slight decrease in performance. The high AUC values suggest excellent discrimination capability for predicting hospital mortality.

## Conclusion

The Neural Network model provided the best result in ROC AUC when performing out-sample; therefore, it is the most appropriate model using to predicting the probability of death in an ICU. Along with these models, I also apply Random Forest, Boosted Tree method, however, the results of them when conducting out of sample are similar, around 0.8, lower than Neural Network result.