**Introduction: (Prepared by CHAN Shing-ho 23414537)**

Predicting mortality risk in the Intensive Care Unit (ICU) is important for several key reasons, including resource allocation, tailored interventions, informed decision-making, and the establishment of realistic care goals. In this project, we focus on three main components: data engineering, model analysis, and prediction. Each component employs various techniques to enhance prediction accuracy, which will be explained in detail below.

**Data Engineering: (Prepared by CHAN Shing-ho 23414537)**

1. Null Values: Our analysis revealed the presence of null values in our data. Since many machine learning algorithms (MLA) cannot process null values, we will address this issue first by using K-Nearest Neighbours (KNN) imputation to modify the missing entries. The KNN imputation leverages the inherent relationships within the data and maintains the natural structure and distribution of the dataset.
2. Feature Engineering: Creating new features ‘range’, ‘min to mean’, and ‘max to mean’ can enhance our dataset by providing additional insights and improving the performance of MLA models. The range provides a simple measure of variability within a feature, indicating how spread out the values are. The ‘min to mean’ and ‘max to mean’ summarize two statistical properties into a single feature. Afterwards, we expanded our dataset to a total of 66 features.
3. Quantile Transformer (QT): During our analysis, we observed that the features have different units and scales. Therefore, we performed QT to ensure that all features follow a uniform or normal distribution. It can enhance the performance of certain algorithms that assume normally distributed data, such as linear regression or logistic regression. Our experiments showed that it yields better result than z-score normalization.
4. A graph with a line

   Description automatically generatedPrincipal Component Analysis (PCA): PCA reduces the number of features while preserving the majority of the variance in the data. Based on the results of the PCA, we typically will select around 27 features for further analysis. However, our experiments found that PCA would decrease our model’s performance, therefore we did not add it to the data pre-processing pipeline at last. Details regarding the decline in performance will be presented in the model evaluation section.
5. Outliers Identification: We employed the Isolation Forest algorithm to identify outliers in our dataset. This machine learning technique utilizes a tree-based approach to detect outliers by randomly selecting features and splitting values; outliers are defined as data points that can be isolated with fewer splits. We performed parameter tuning on the Isolation Forest model to determine the optimal settings, resulting in the identification of 46 outliers. However, the outlier predictions are correlated with our target variable, ‘mortality’, as confirmed by the Spearman rank test. Consequently, we incorporate the ‘is\_outlier’ variable as a new feature. Additionally, we sought to identify outliers using the Interquartile Range (IQR) method. While the IQR method is easy to understand and implement, it can be influenced by skewed data, potentially leading to the misidentification of outliers. This approach resulted in the identification of 3,241 instances as outliers, which are not useful for our project.
6. A graph of a blue and orange bar

   Description automatically generated with medium confidenceOversampling: After removing outliers, we further analysed the class distribution of our target variable. We discovered that the minority class '1' comprises only one-fourth of the dataset, indicating an imbalance. To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to achieve better balance. SMOTE generates synthetic samples for the minority class by interpolating between existing instances, rather than simply replicating them as in traditional oversampling. SMOTE considers the feature space of the minority class, creating new samples that are realistic and consistent with the distribution of the existing data. This process helps maintain the integrity of the dataset while enhancing the representation of the minority class.

**Next Part:**

To Be Continue

Reference:

[QuantileTransformer — scikit-learn 1.7.dev0 documentation](https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.QuantileTransformer.html)

[IsolationForest — scikit-learn 1.7.dev0 documentation](https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.IsolationForest.html)

[SMOTE — Version 0.12.4 (imbalanced-learn.org)](https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html)