Data Mining Assignment 2
Classification

Bo-Han Chen (陳柏翰)
Student ID:312551074
bhchen312551074.cs12@nycu.edu.tw

Experiment Environment & Usage

Environment

- OS: Windows 10 22H2
- Hardware: Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40GHz
- Python 3.12.0

Data Preprocessing

Data Overview

The given dataset contains 44939 patients' information, including 81 features and 1 label representing whether the patient has died. Among the 81 features, 23 features are categorical and the rest are numerical.

Categorical Features

Categorical features contain two types of data, including *bool* and *object*. After visualizing the data, I found several features that are highly related to the result of death. The details are shown as follows:

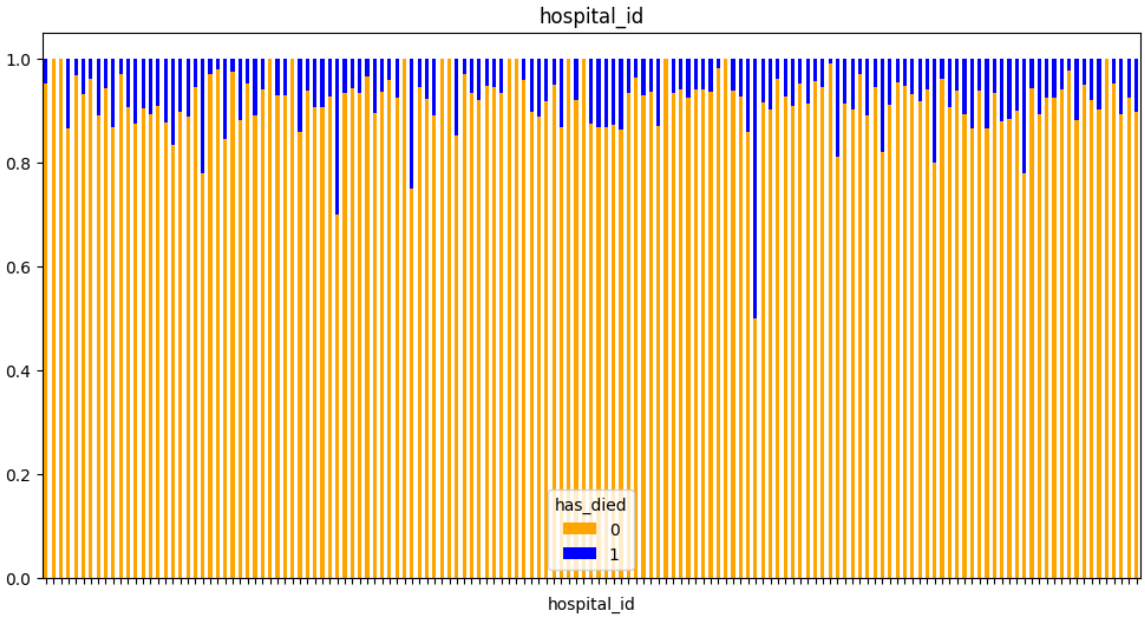

Figure 1: Percentage of Death in Different Hospital

From the figure 1 and figure 2, we can see that some of the hostipal and icu are has higher death rate than others. So it's reasonable to assume that the hostipal and icu information are related to the result of death in this dataset, and it is necessary to keep these features.
Since there are some missing values in the categorical features, so we can first analyze the relationship between missing values and result of death, and then decide the way to transform the

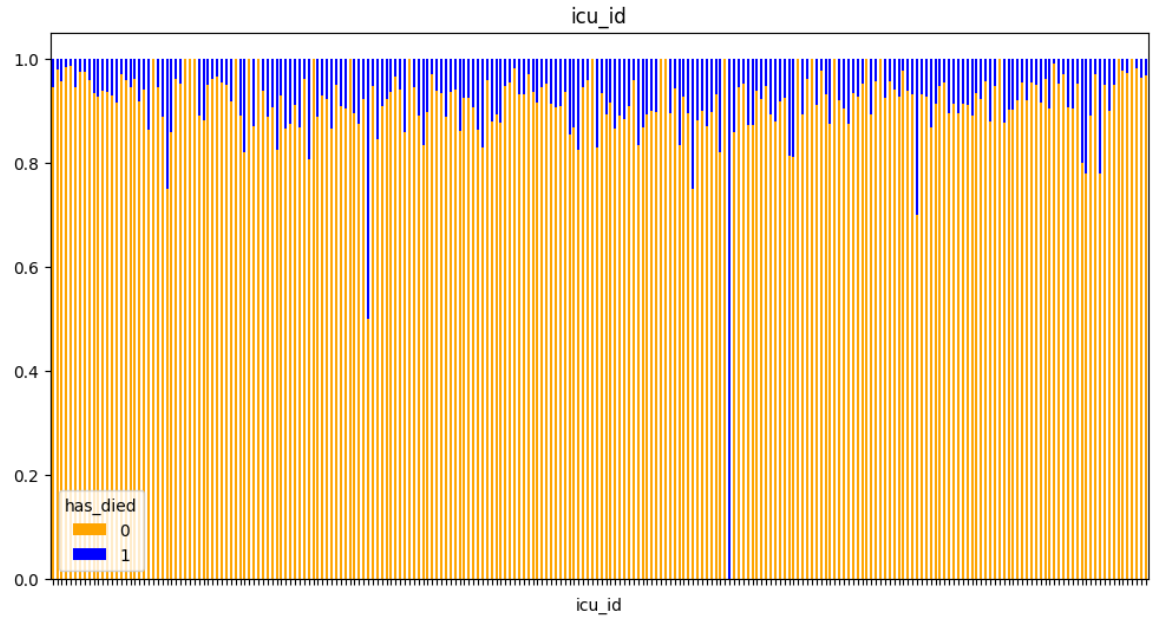

Figure 2: Percentage of Death in Different ICU

missing values. First, I analyze the percentage of missing values in each feature in figure 3. From the figure, we can see that the missing values accounts for proportion of 0.75% to 1.75% in most of the categorical features, now we can look into the features with high percentage of missing values.

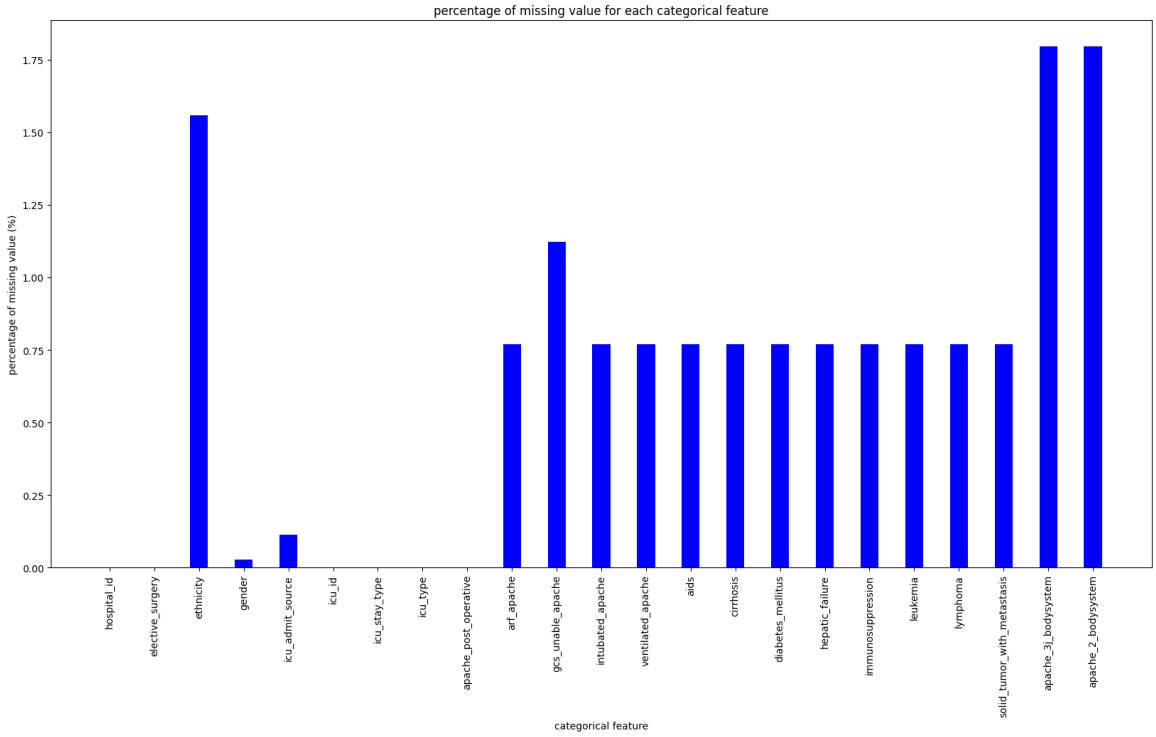

Figure 3: Percentage of Missing Values in Categorical Features

For feature *apache_2_bodysystem* and *apache_3j_bodysystem* and *ethnicity*, the relationship between missing values and death rate is shown in figure 4, figure 5 and figure 6 From the figure, we can see that the missing values in these three features still contains some information about the result of death, so it's not reasonable to simply drop these missing values. For the rest of categorical features that also contains missing values, the missing values are also accounts for a small proportion of death rate. Although these related death result represents a very small proportion of the whole dataset, it's still necessary to keep these missing values to avoid losing important information.

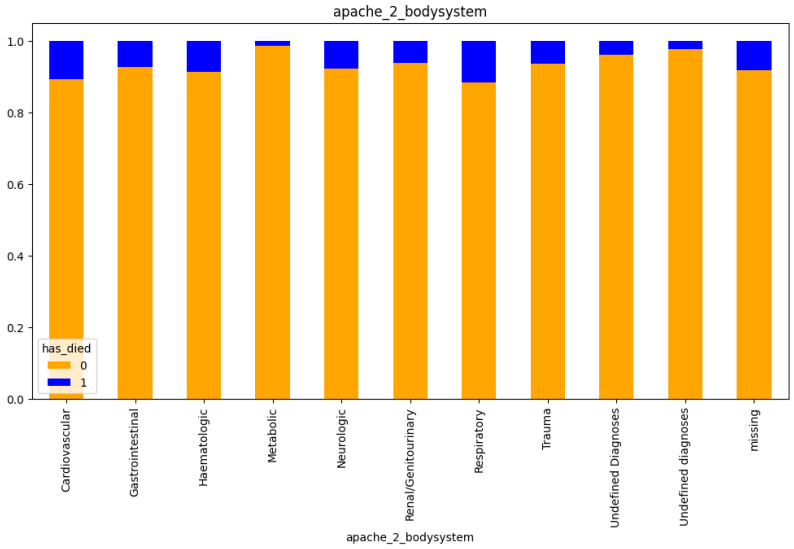

Figure 4: Percentage of Death in Different apache_2_bodysystem

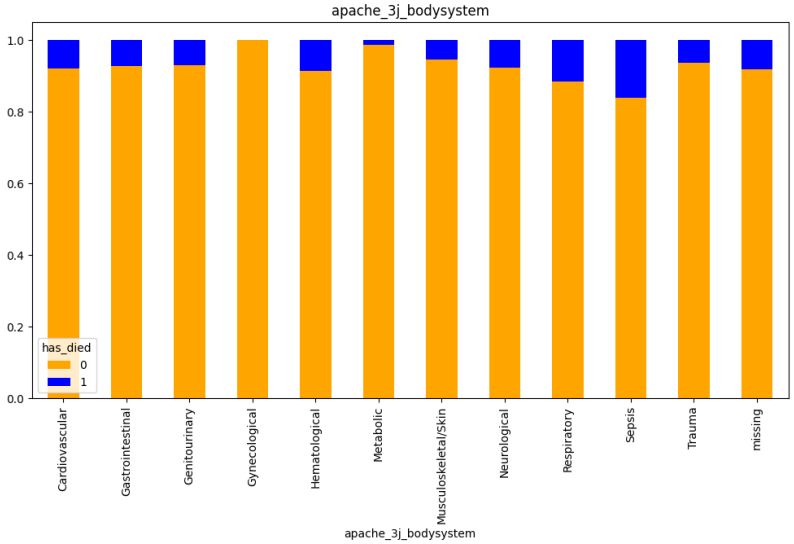

Figure 5: Percentage of Death in Different apache_3j_bodysystem

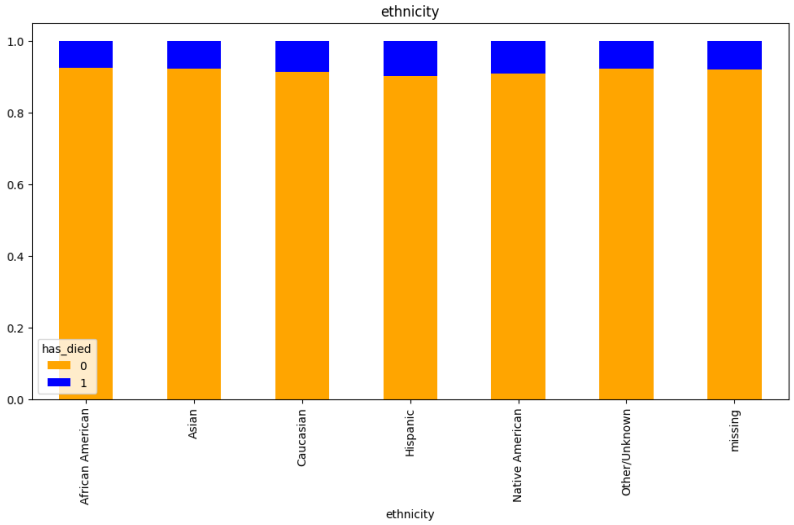

Figure 6: Percentage of Death in Different ethnicity

Numerical Features

For numerical features, I first analyze the percentage of missing values in figure 7. From the figure we can see the proportion of missing values in most of numerical features is much higher than the categorical features. For further analysis, I fill the missing values with the mean value of each feature, which can minimize the impact of missing values.

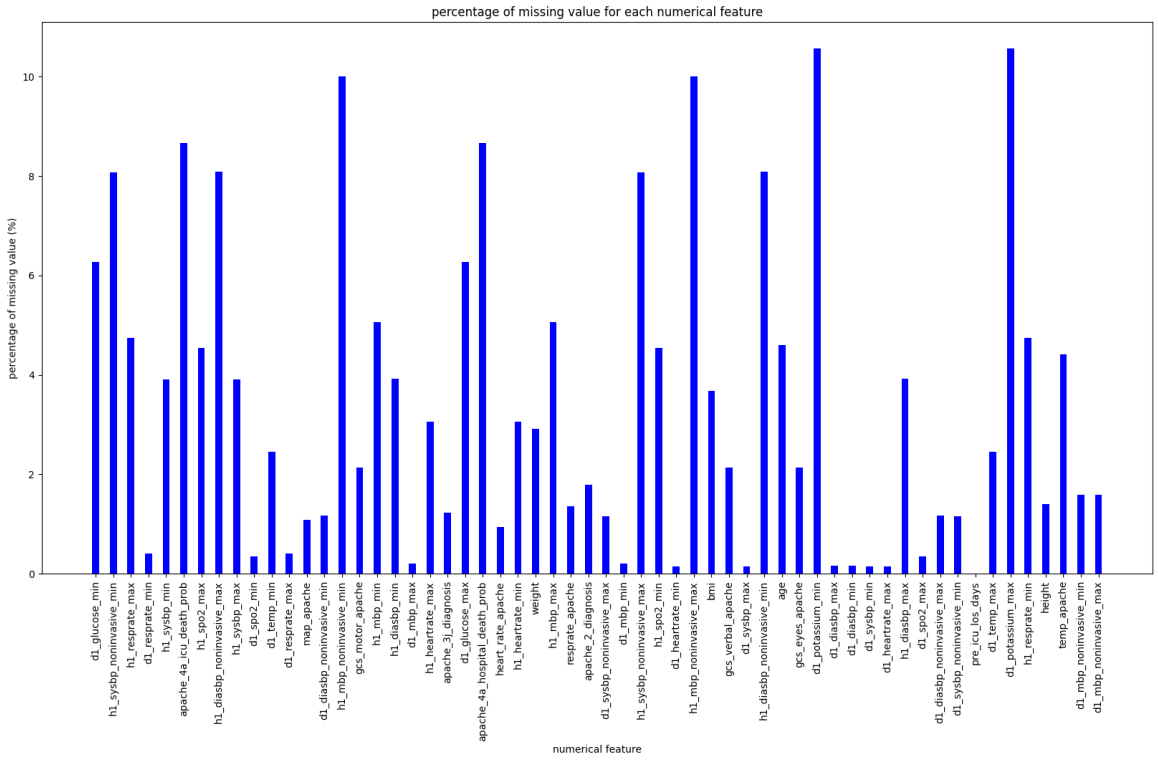

Figure 7: Percentage of Missing Values in Numerical Features

Figure 8 show the distribution of bmi in different result of death. From the figure, we can see the trend of bmi value among our dataset, and the percentage of death is quite similar in different bmi value. An other interesting discovery is that there are some abnormal distribution near the bmi value 70, which I think is the outlier of this dataset at the first glance. However, after further analysis, I found that the same distribution trend can be found in feature *weight* from figure 9, which means the abnormal distribution may be casued by realistic reasons, such as some diagnosis or treatment may cause the patient to gain weight.

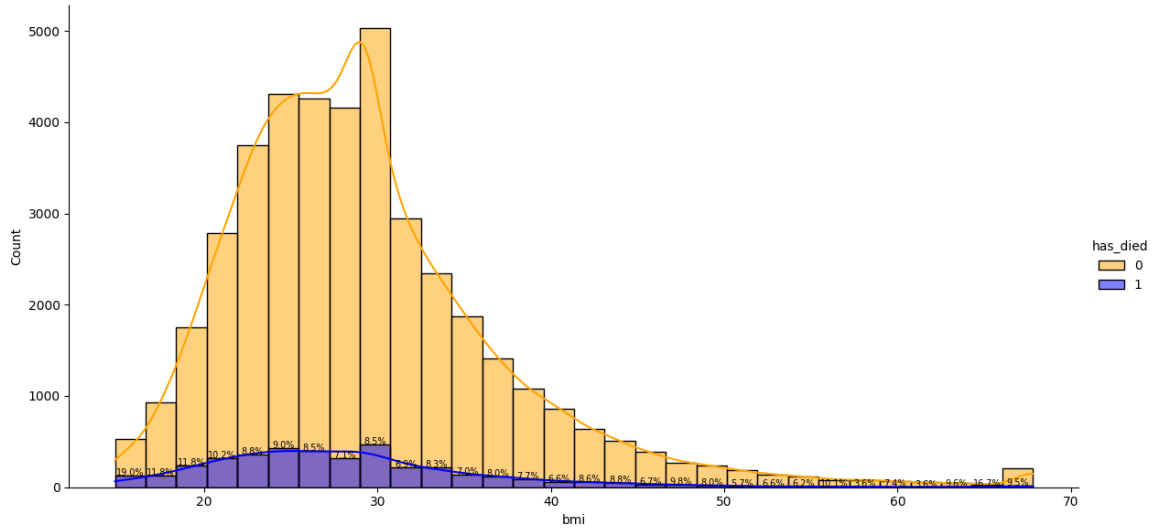

Figure 8: Distribution of bmi

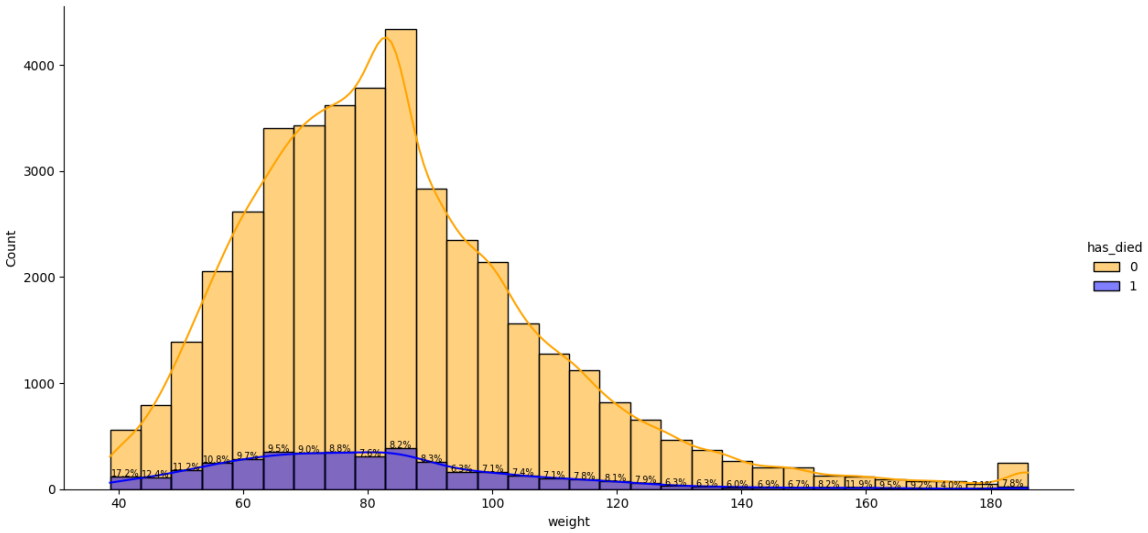

Figure 9: Distribution of weight

The discovery about outliers can be found in feature *apache_4a_hospital_death_prob* and *apache_4a_icu_death_prob*. From figure 10 and figure 11, we can see there are some negative values in these two features, which is obviously the outliers when representing the probability of death. To dealing with these outliers, I replace them with NaN values and use imputation for restoring, expect this operation can let the new value be more realistic. The details of imputation will be mentioned in the following section.

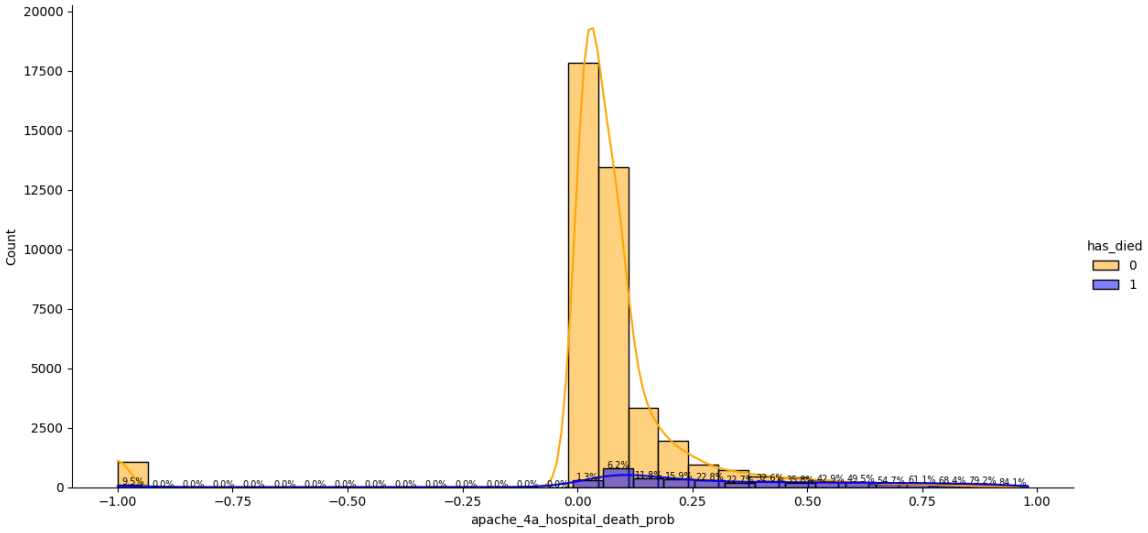

Figure 10: Distribution of apache_4a_hospital_death_prob

Feature Encoding

For this dataset, I tried several encoding method for categorical feature transformation, since data transformation will also lead to different feature selected for training, so we can

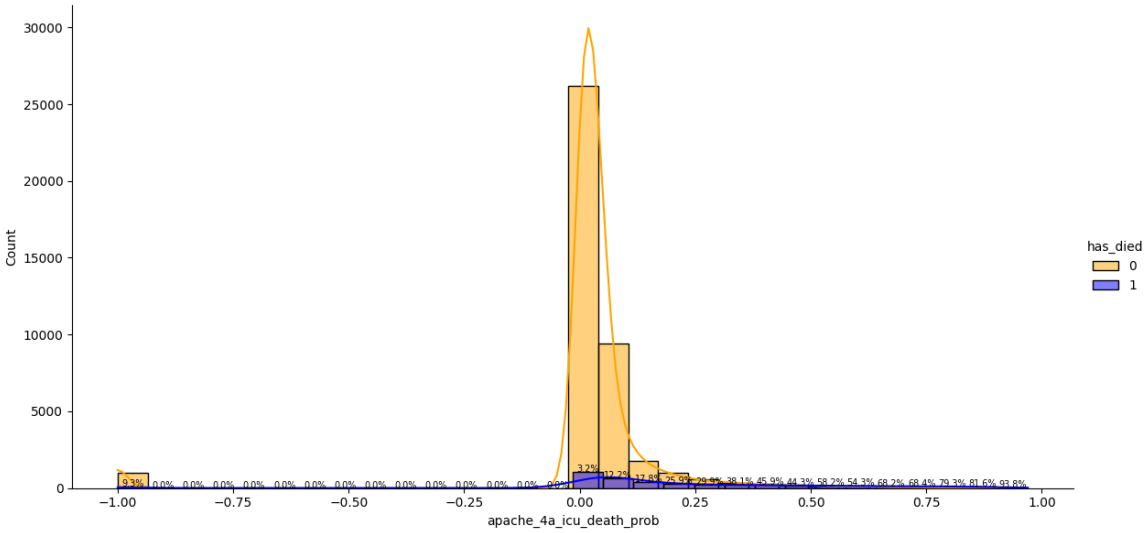

Figure 11: Distribution of apache_4a_icu_death_prob