HW2 黃偉祥 X1136010

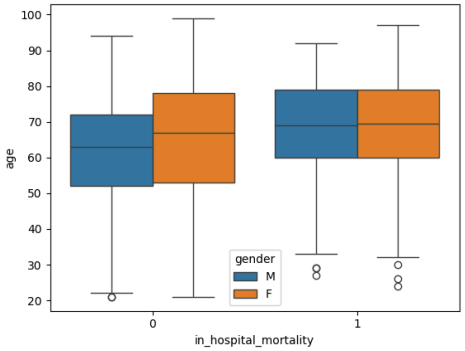
Features Table

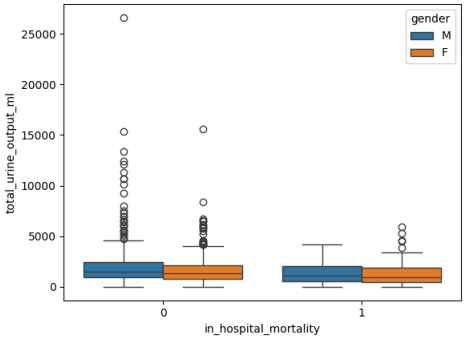
		Missing Data	Outcome		P-value
			Alive	Die	
Number of Patients		0	1445	291	
Urine output (mean)		599	1878.1338	1352.1867	<0.001
Vasopressin usage during ICU stay (number of patients)	0	0	451	115	0.007
Vasopressin usage during ICU stay (number of patients)	1	0	994	176	0.007
Gender (number of patients)	Male		735	137	0.265
Gender (number of patients)	Female		710	154	0.265
Age (mean, std)		0	(62.6,16.9)	(68.5,14.7)	<0.001
hemoglobin_min (mean, std)		500	(2.8,13.1)	(4,15.5)	0.001
hemoglobin_max (mean, std)		500	(97.9,6.1)	(97.1,7.4)	0.009
resp_rate_mean (mean, std)		491	(64.7,11)	(66.5,13.3)	0.366
glucose_max (mean, std)		74	(146.4,65)	(185.8,118.1)	<0.001
heart_rate_max (mean, std)		491	(226,49)	(237,54)	<0.001
glucose_avg (mean, std)		74	(134.4,47.9)	(160.7,92.6)	<0.001
glucose_min (mean, std)		74	(123.5,42)	(137.5,79.8)	0.353
Race	Asian		28	11	0.004
Race	Black		185	38	0.004
Race	Hispanic		45	8	0.004
Race	Portuguese		6	0	0.004
Race	White		987	173	0.004
Race	Other		194	61	0.004

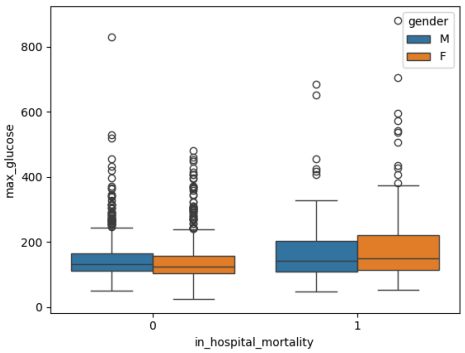
Data Preprocessing Description

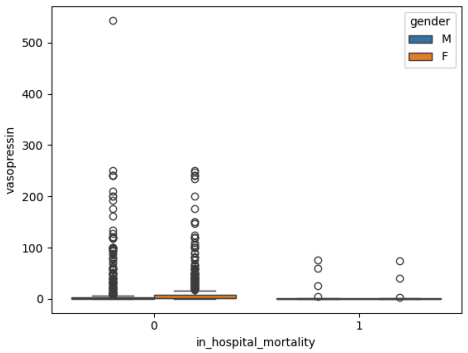
Outliers

- age > 90, this is special case, because too old may die anyway regardless the diseases
- total_urine_output_ml > 5000, too much urine output within 30 hours is not really possible
- max_glucose > 600, use box plot to determine
- vasopressin > 100, use box plot to determine

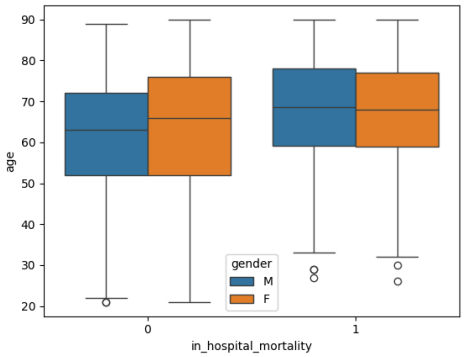


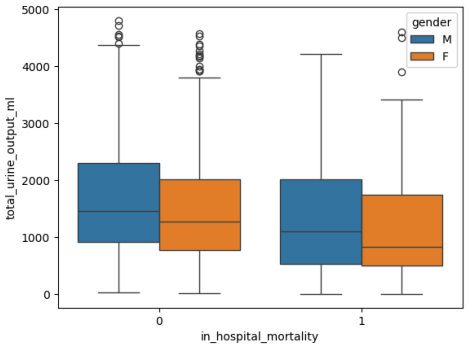


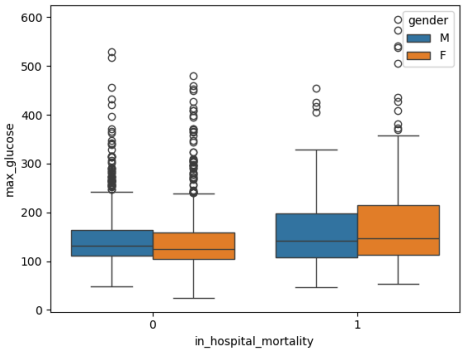


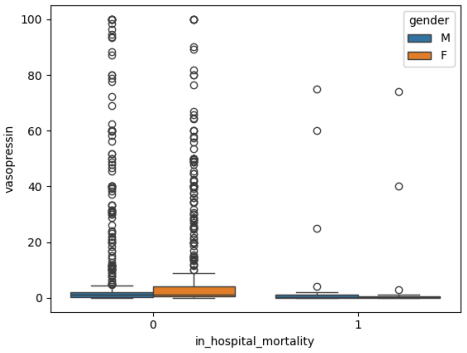


After delete outliers









Apply Z-score

```
outlier_mask = (z.abs() >= 3).any(axis=1)
df_outliers  = df[outlier_mask]
df_clean     = df[~outlier_mask]
```

- Z-score ≥ 3 does not detect any outliers

Fill in missing

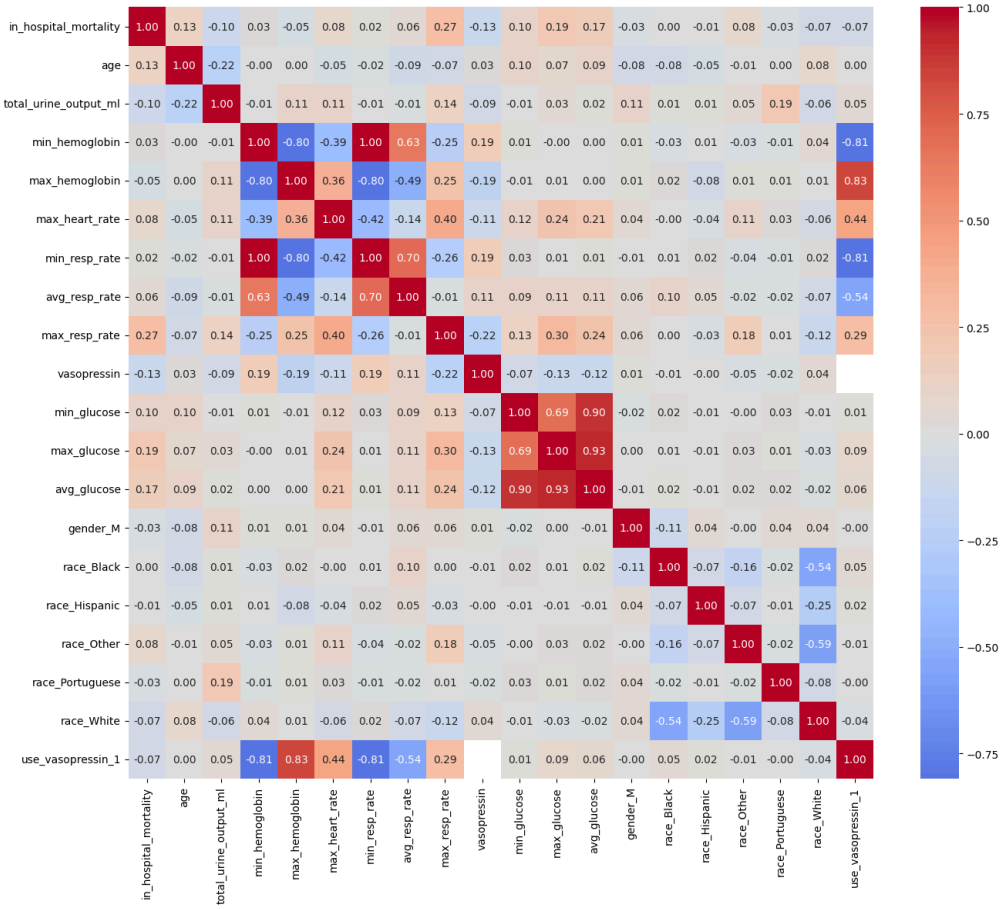
Use **K-Nearest Neighbors** to fill in missing value based on different classes

- Because the classes are very imbalanced, so we need to fill in missing value based on different classes or it will make class 1 missing value filled in with class 0 features
- First separate the dataset into 2 dataset (class_0, class_1) then fill in missing value using K-Nearest Neighbors
- Then concat both and done

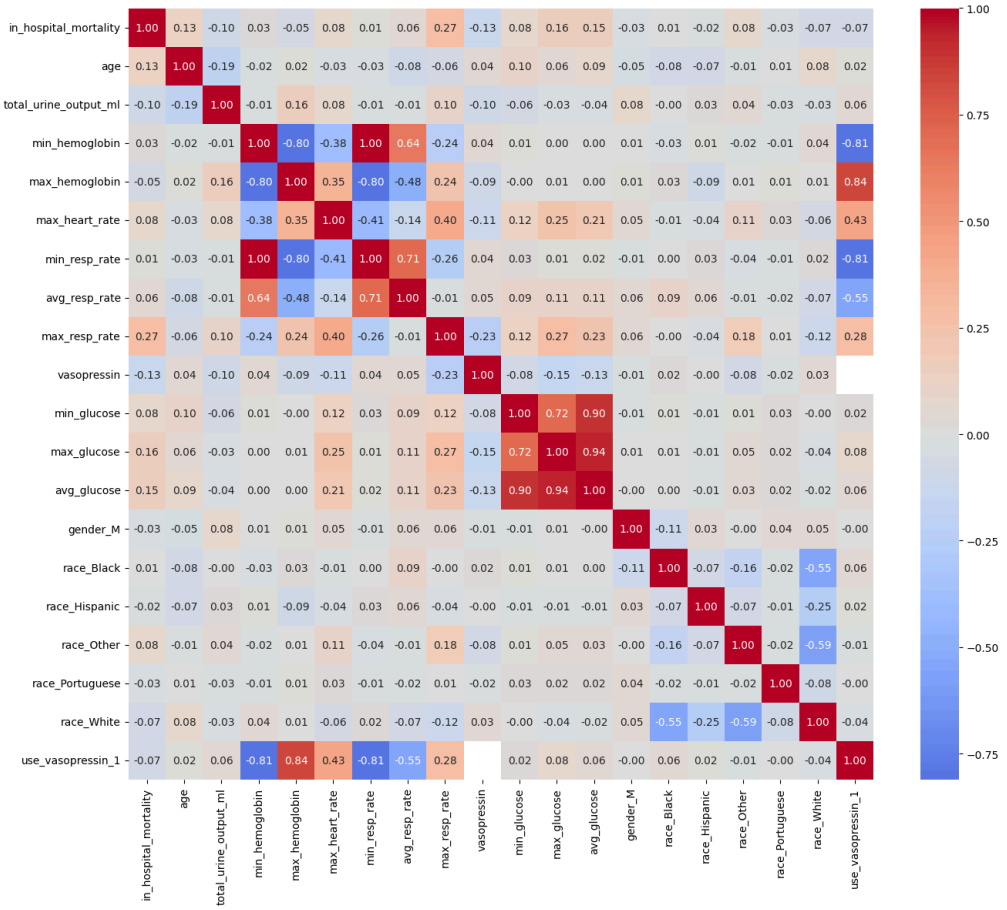
Effective Feature Representation

Correlation Map

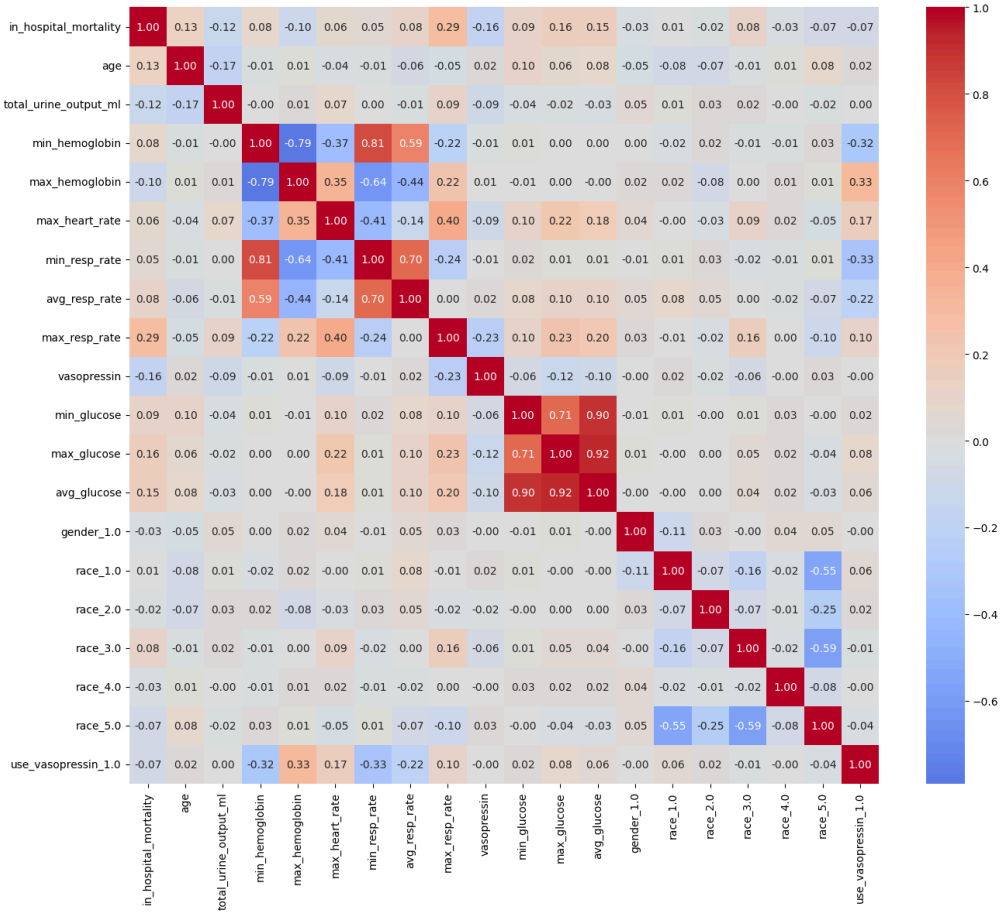
- Original



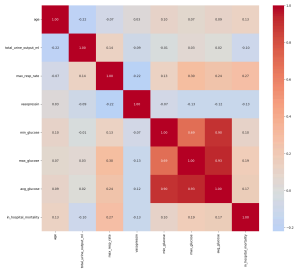
- Clean outliers



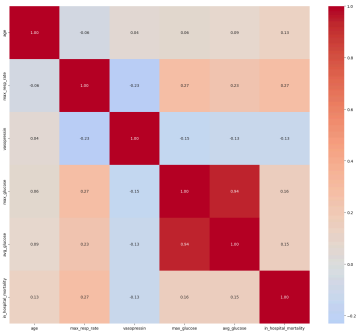
- Fill in missing



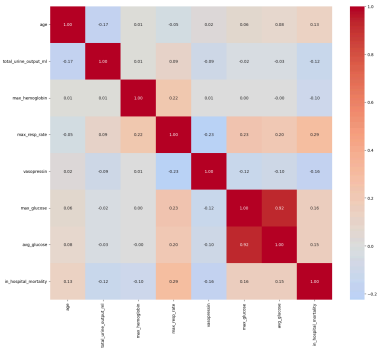
- Original strong map



- No outliers strong map

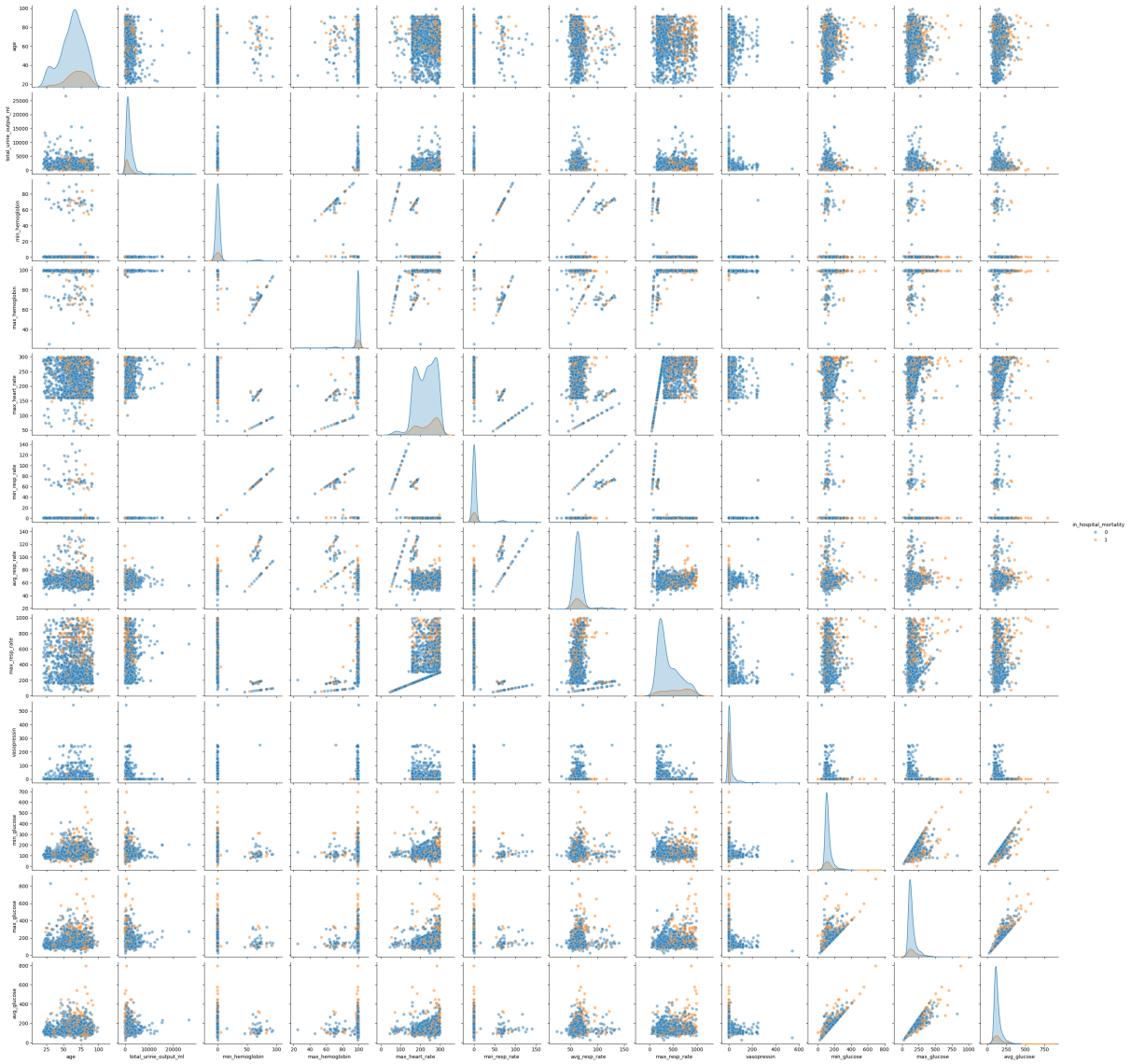


- Filled in strong map

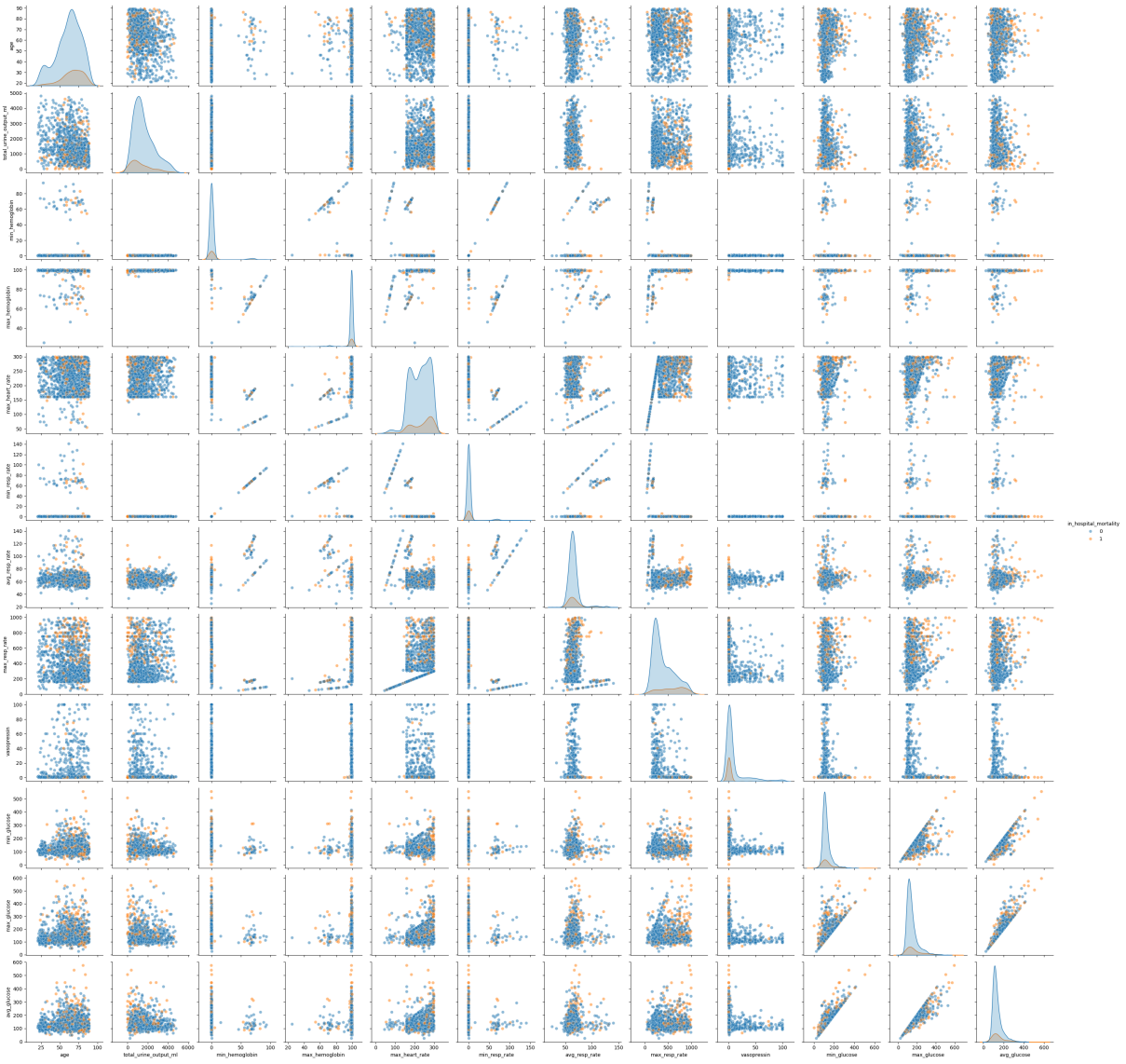


Pairplot

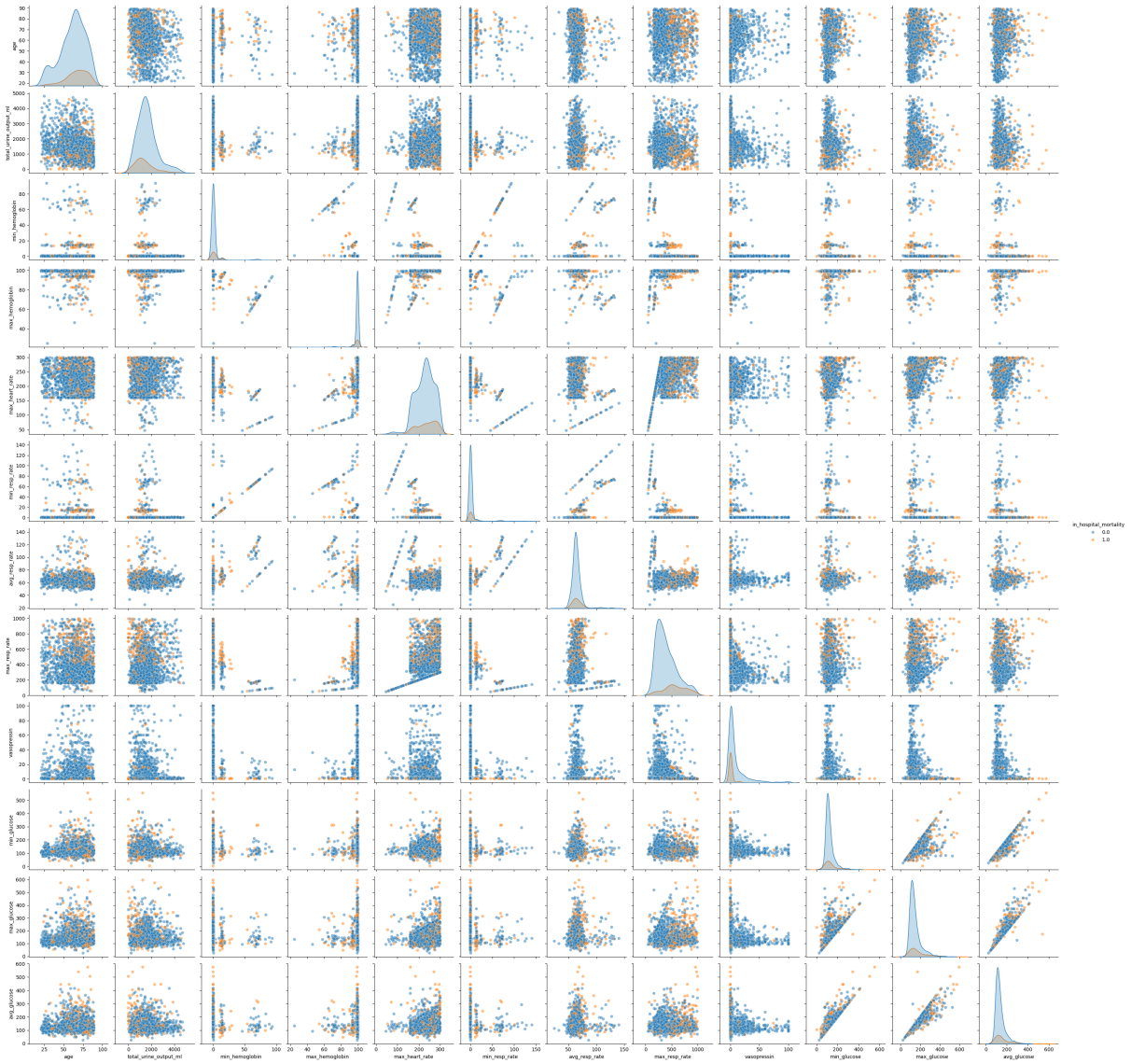
- Original



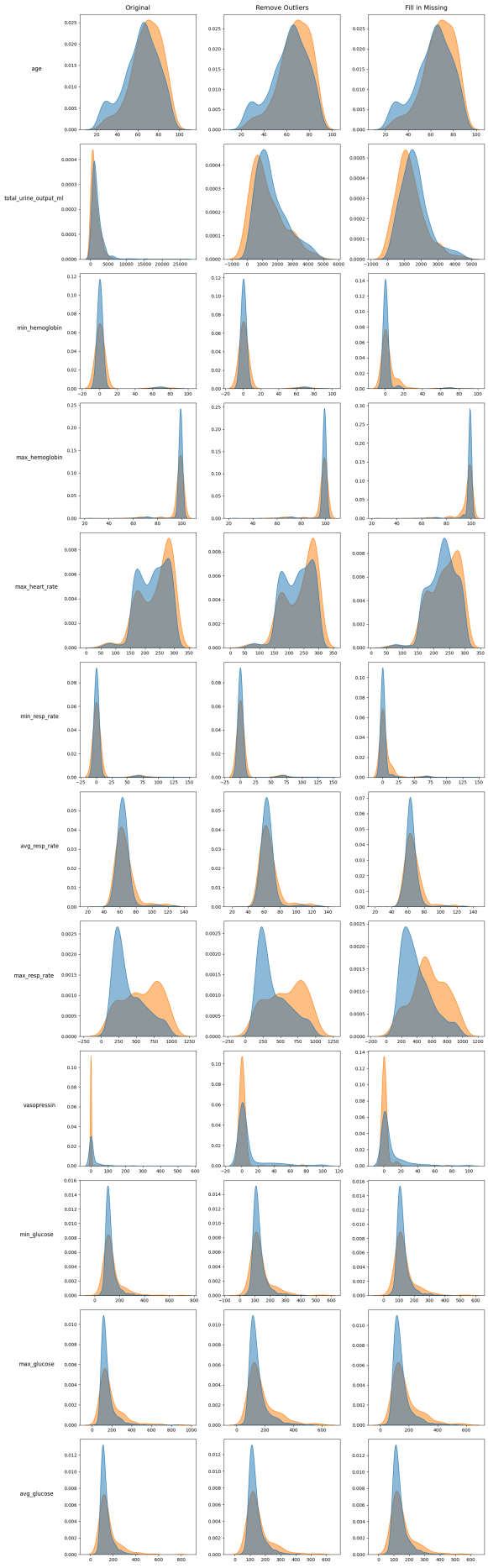
- Clean outliers



- Fill in missing



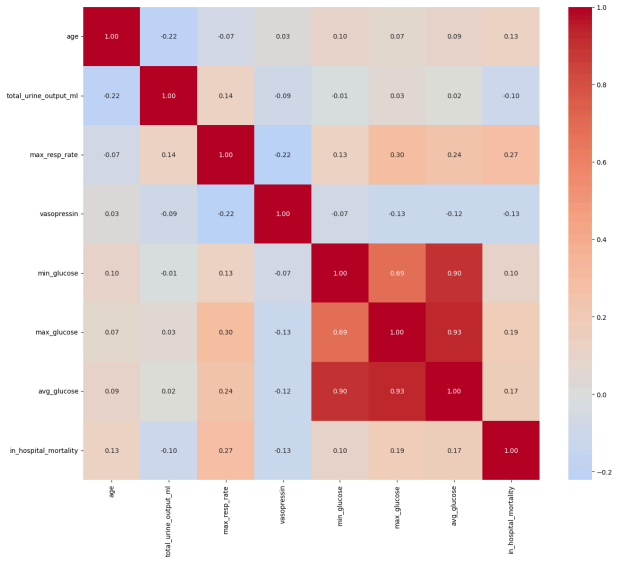
Normalised Distribution

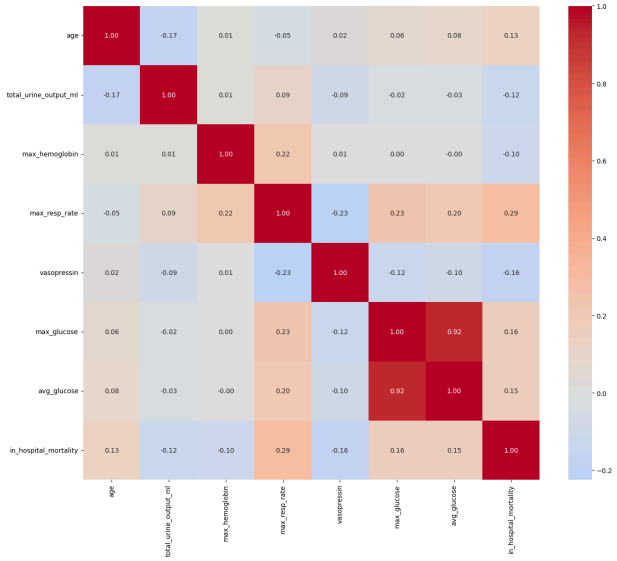

Original
Remove Outliers
Fill in Missing
age
total_urine_output_ml
min_hemoglobin
max_hemoglobin
max_heart_rate
min_resp_rate
avg_resp_rate
max_resp_rate
vasopressin
min_glucose
max_glucose
avg_glucose

Preprocessing Impact

Correlation map

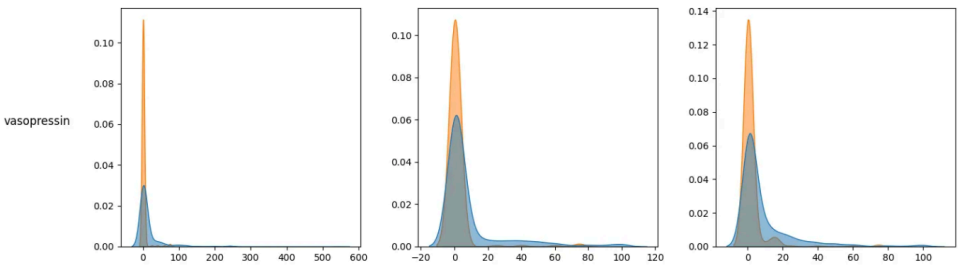
- The strong correlation change, and some correlation become stronger after preprocessing
  - This means we may get the more accurate correlation after preprocessing



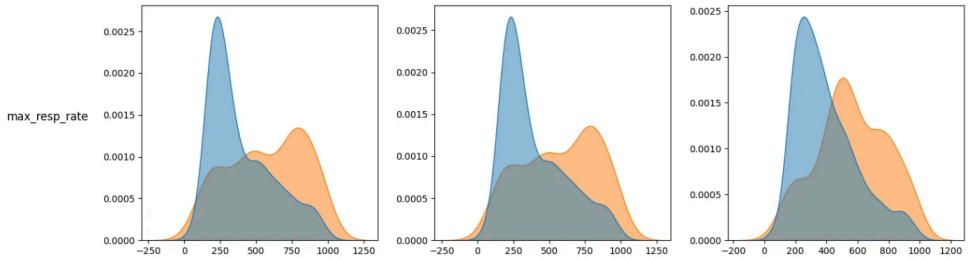


About normalised distribution

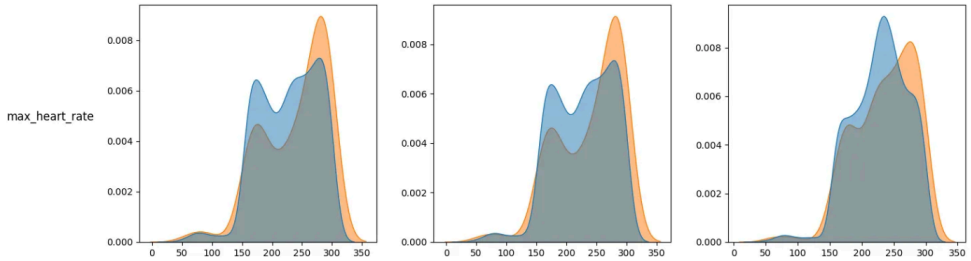
Original → Remove outliers → Fill in missing value



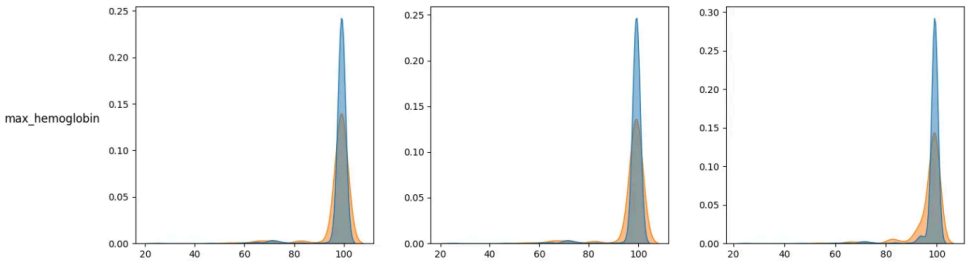
- After preprocessing the distribution change quite much, it looks more sense than the original distribution



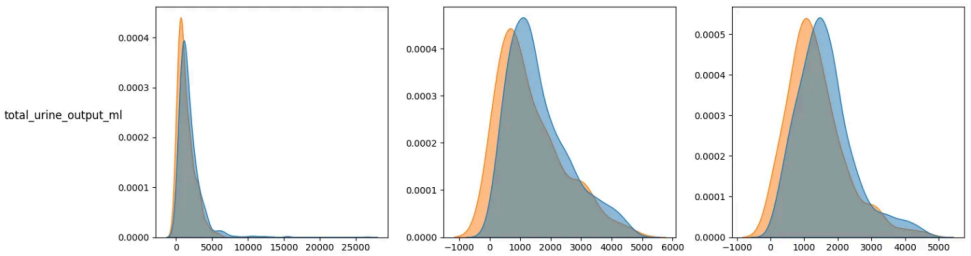
- For resp_rate, it has a lot of missing value, so the original may not represent the **accurate** distribution.



- Heart rate also has a lot of missing value and distribution for both class 0 & class 1 change a lot



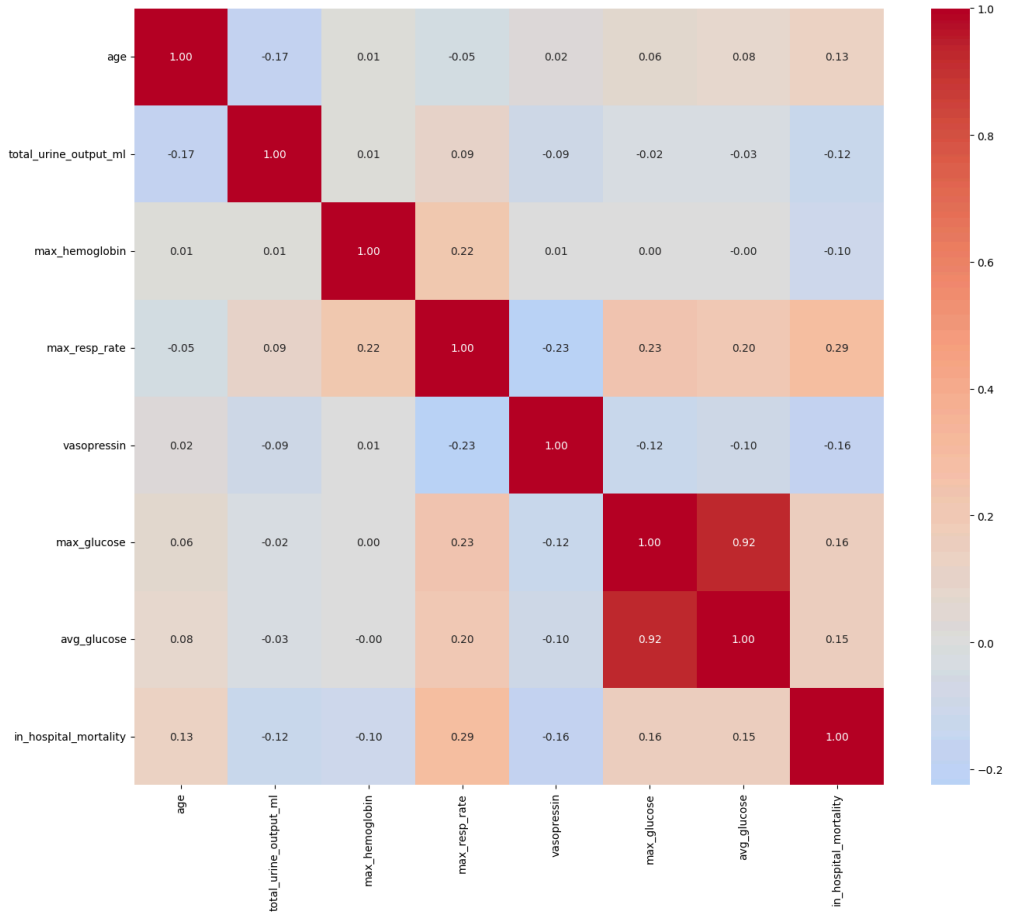
- This change a little at the left corner part but this little change make the distribution of the 2 classes more different



- Urine output have a lot of in-common sense data, such as 5000 ml above within 30 hours which is very impossible, after preprocessed the distribution make more sense also.

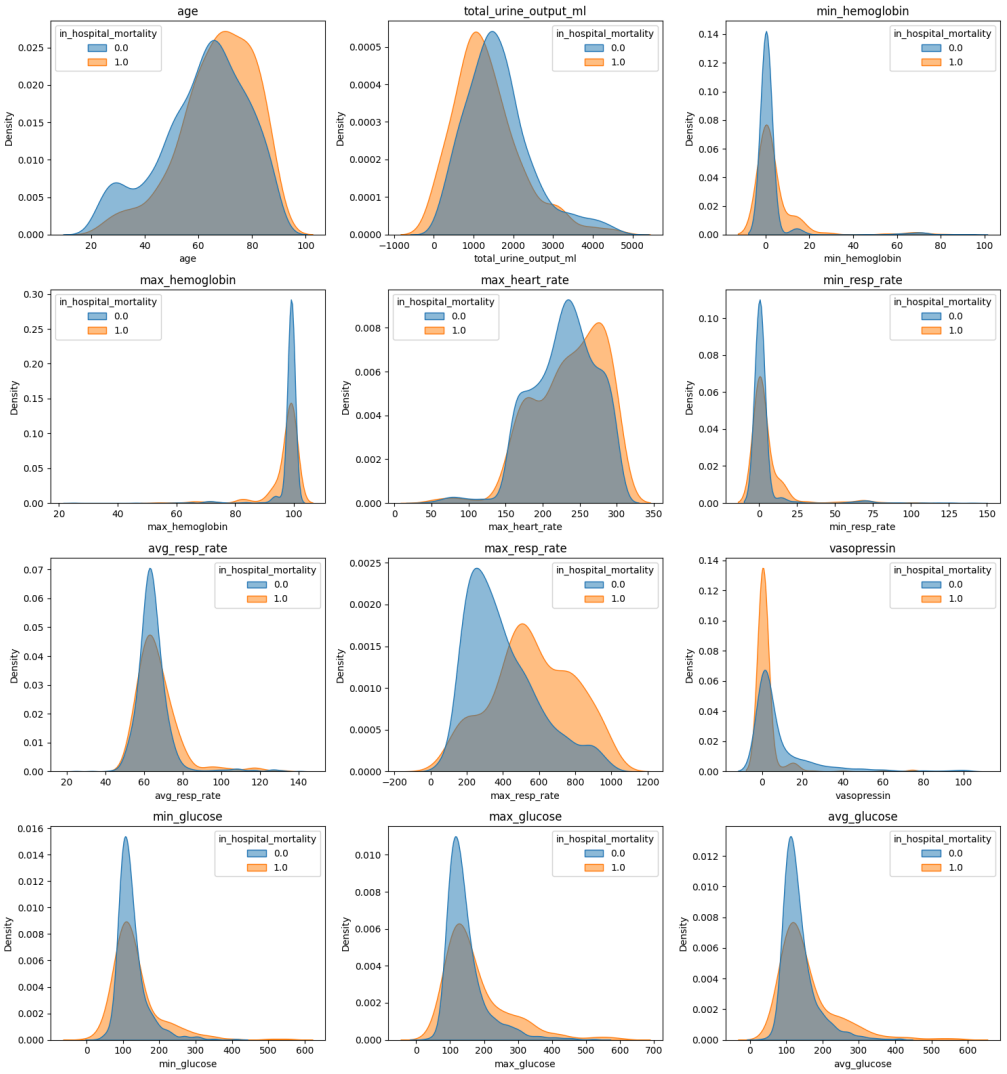
Feature Relevance Analysis

Correlation map



- From the correlation map we can conclude that the strong correlation order are:
  - (highest)max_resp_rate → vasopressin → max_glucose → avg_glucose → age → urine_output → max_hemoglobin (lowest)
  - This only show the features that has correlation more than 0.1 with in-hospital mortality

Normalised distribution



- while from the normalised distribution we can conclude that max_resp_rate and max_heart_rate has big difference distribution
- Also age and urine output has “shifted” left/right distribution for class 0 & 1
- Also max hemoglobin, max glucose, and vasopressin also has 2 different distribution but in difference of the **high** and **width**

Code

In ipynb file and sql file