NANYANG TECHNOLOGICAL UNIVERSITY

BS6200 Final presentation

The hospital mortality prediction for ICU- admitted HF (Heart Failure) patients using MIMIC III dataset

presented by

Han Wenhao
School of Biological Science

17 / 10 / 2022

Table of content

1. Dataset description
2. Problem statement
3. Dataset preprocessing
4. Training and testing procedure
5. Experimental Study and analysis
6. Summary of the Project achievement
7. Future direction for further improvement

1. Dataset description

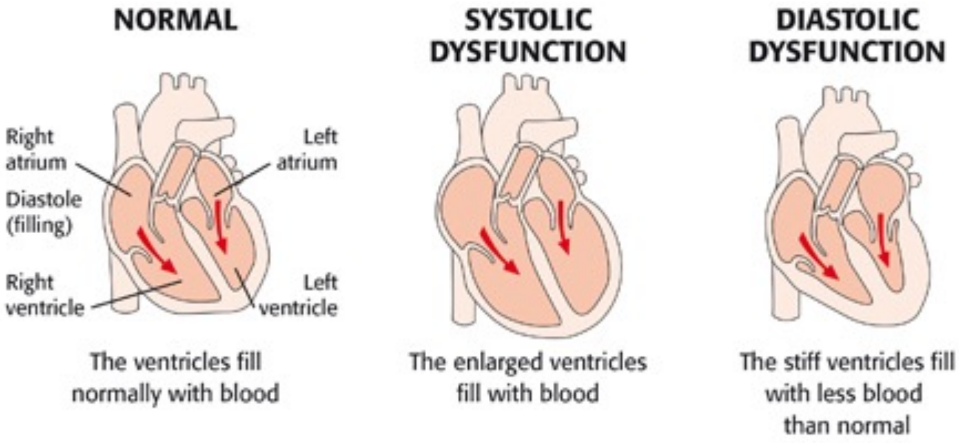
MIMIC-III ('Medical Information Mart for Intensive Care')

- Comprising Information relating to patients admitted to Intensive care units at a large tertiary care hospital
- Covering 38,597 distinct adult patients and 49,785 hospital admissions between 2001 and 2012.
- Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers.

2. Problem statement

Disease introduction:

- Heart failure (HF) is a complex clinical syndrome that causes a patient's heart to not pump enough blood (ventricular insufficiency) to meet the oxygen needs of vital organs and tissues in the body.
- HF disease worsens over time, causing progressive remodeling of the heart (change the size and shape of the heart), finally may lead to the death of the patients.


NORMAL
SYSTOLIC DYSFUNCTION
DIASTOLIC DYSFUNCTION
Right atrium
Left atrium
Diastole (filling)
Right ventricle
Left ventricle
The ventricles fill normally with blood
The enlarged ventricles fill with blood
The stiff ventricles fill with less blood than normal

2. Problem statement

Problem:

- The predictors of in-hospital mortality for intensive care units (ICU)-admitted HF patients remain poorly characterized.

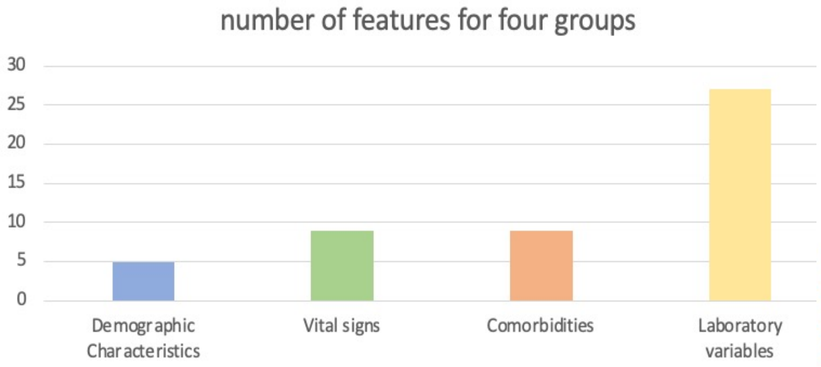
Aim of the Project:

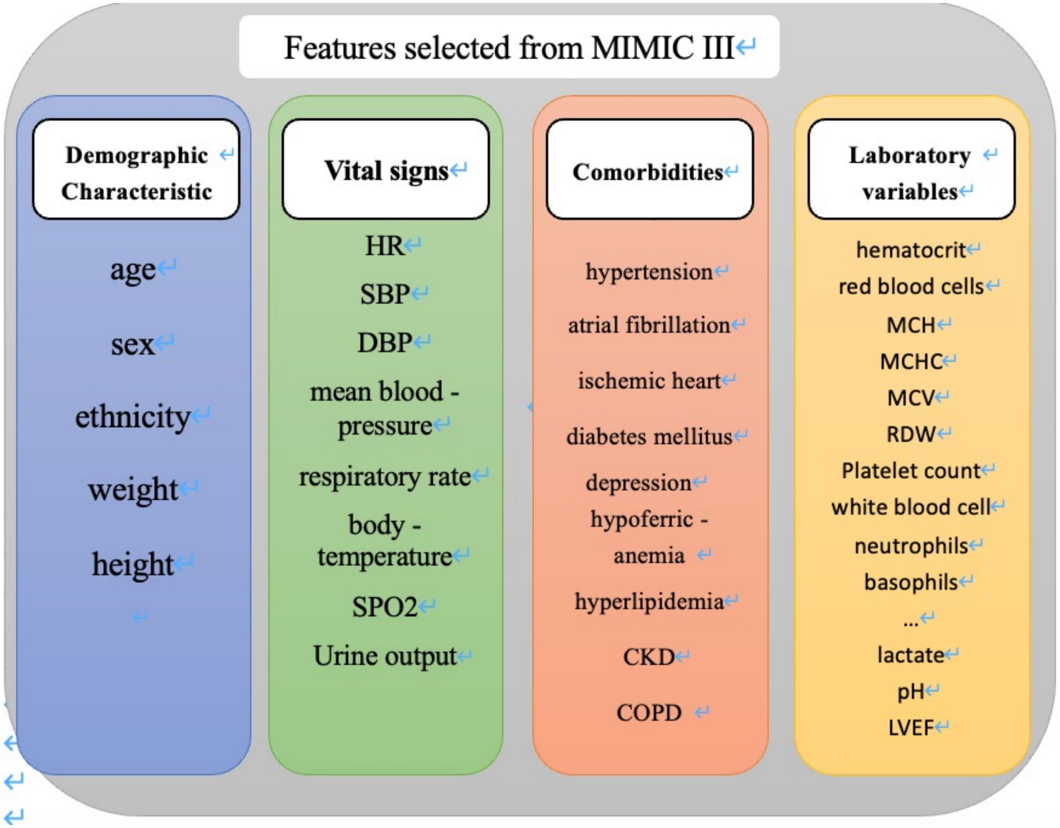
- Develop and validate prediction models for all-cause in hospital mortality among ICU-admitted HF patients.

3. Data preprocessing

1) Feature engineering

- **1177 Patients** with a diagnosis of HF, identified by manual review of ICD-9 codes, and who were >15 years old at the time of ICU admission.
- **50 features** related to the the cause of HF are selected using Structured Query Language queries(PostgreSQL, version 9.6) from MIMIC III.





Features selected from MIMIC III

Demographic Characteristic

age

sex

ethnicity

weight

height

Vital signs

HR

SBP

DBP

mean blood - pressure

respiratory rate

body - temperature

SPO2

Urine output

Comorbidities

hypertension

atrial fibrillation

ischemic heart

diabetes mellitus

depression

hypoferric - anemia

hyperlipidemia

CKD

COPD

Laboratory variables

hematocrit

red blood cells

MCH

MCHC

MCV

RDW

Platelet count

white blood cell

neutrophils

basophils

...

lactate

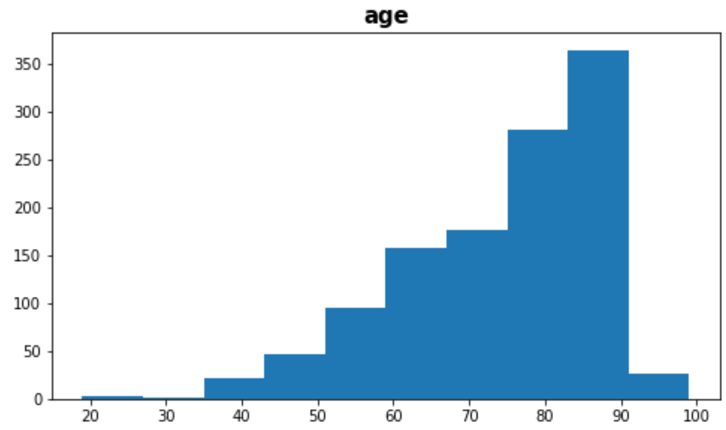
pH

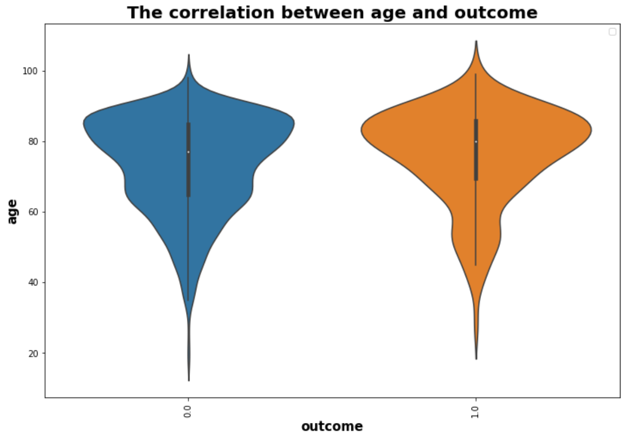
LVEF

3. Data preprocessing

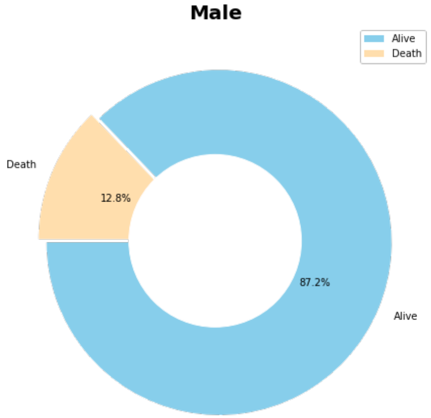
2) Data exploration

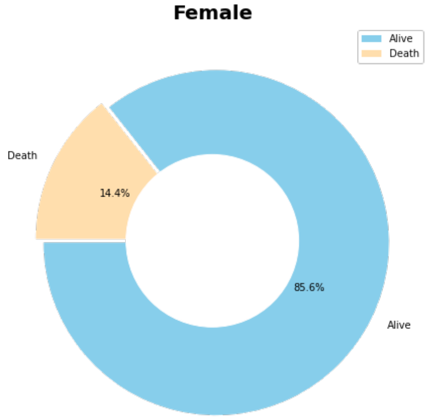
- Age vs. outcome





- Gender vs. outcome

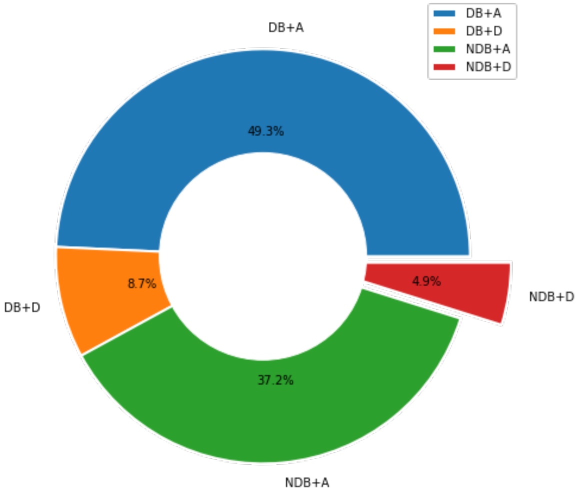




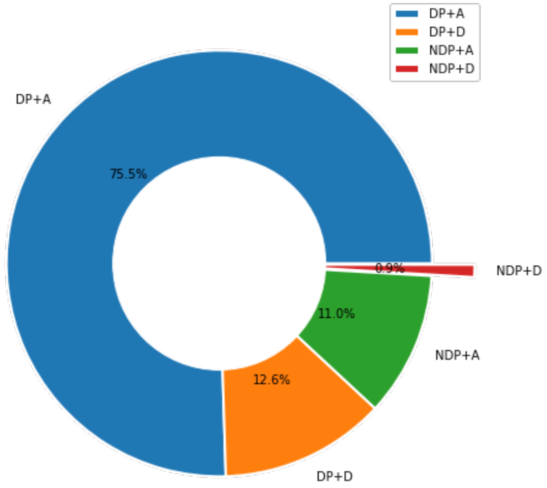
3. Data preprocessing

- 4 complications improve the mortality rate of HF（might be lethal factor）

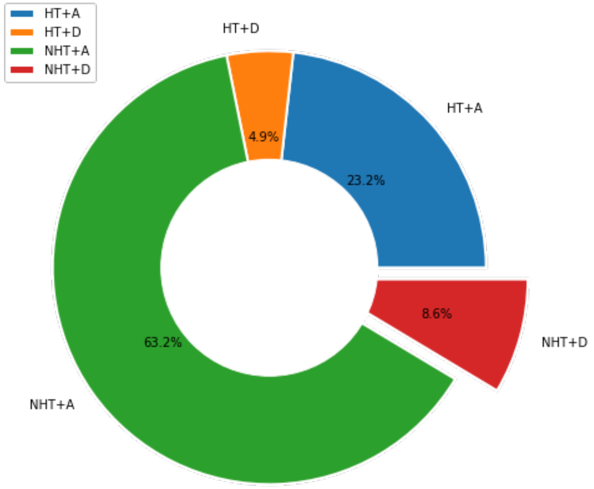
Diabetes Vs. Outcome



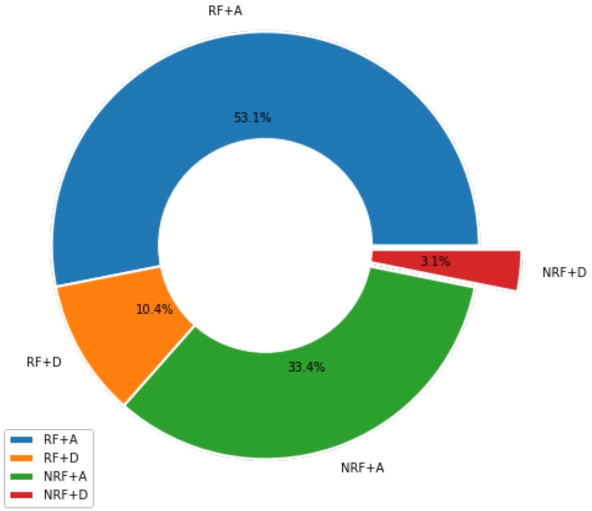
Depression Vs. Outcome



Hypertensive Vs. Outcome

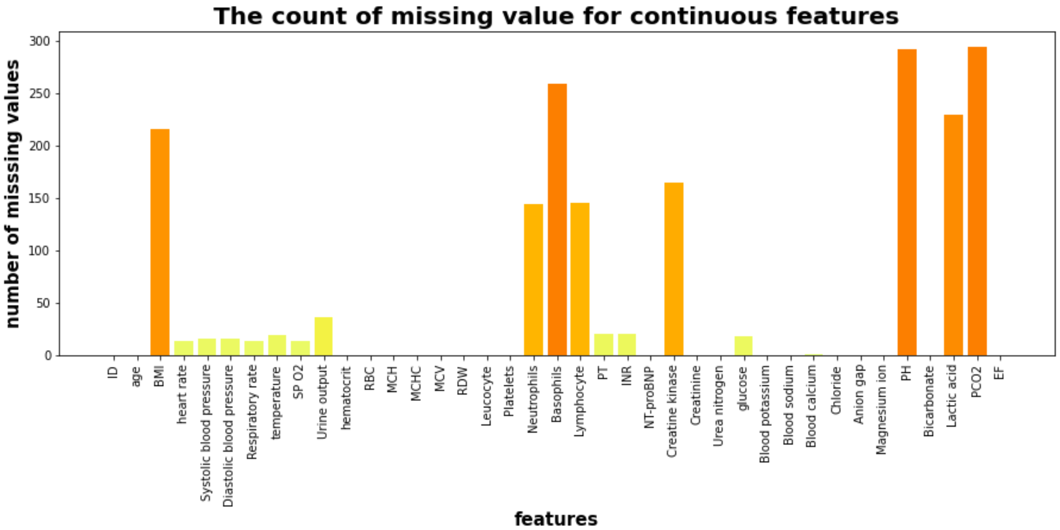


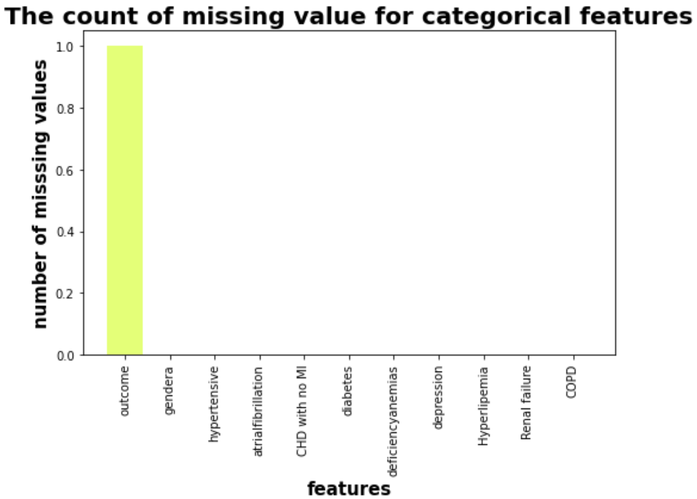
Renal Failure Vs. Outcome



3. Data preprocessing

3) Data cleaning





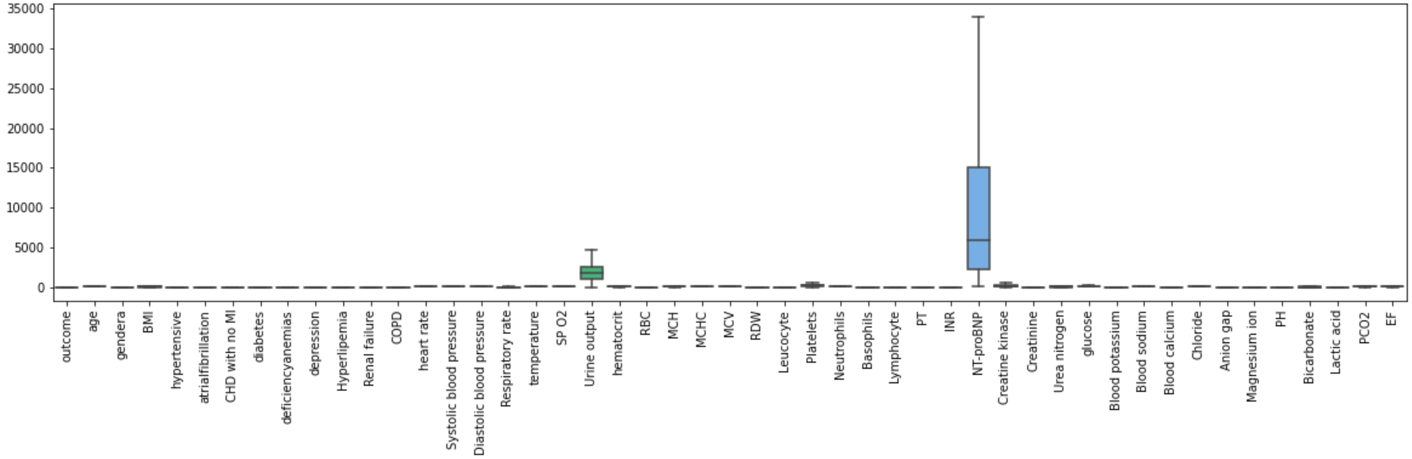
Strategy to handle missing values:

1) Eliminating the features (‘ID’ and ‘group’) .
2) Fill the missing values in numerical data with the mean of the available data in each column.
3) Eliminating the samples with missing values in categorical data.

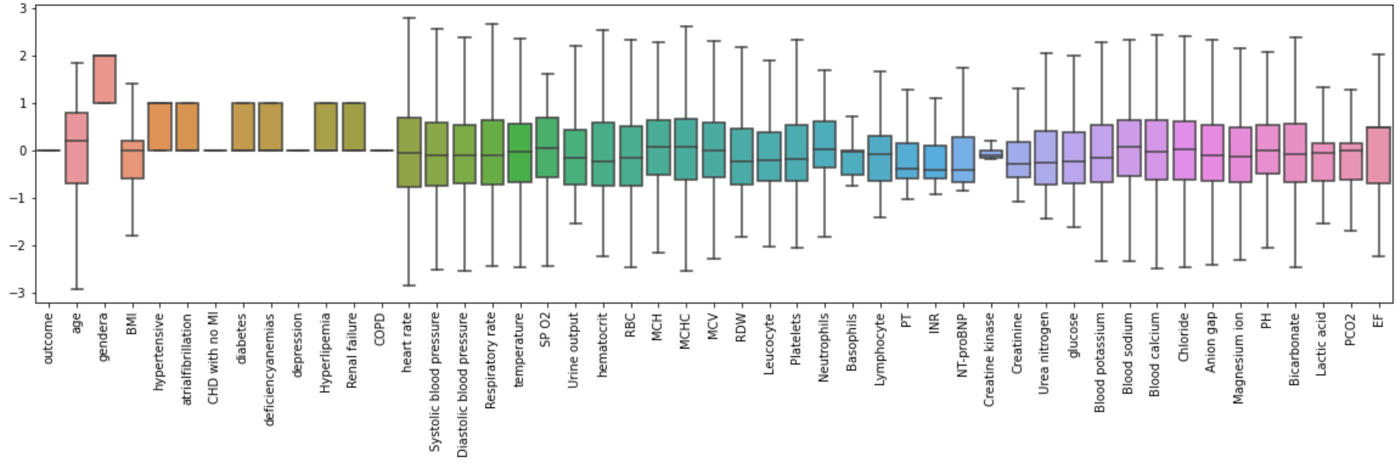
3. Data preprocessing

- Perform z-score normalization to convert data to the same scope

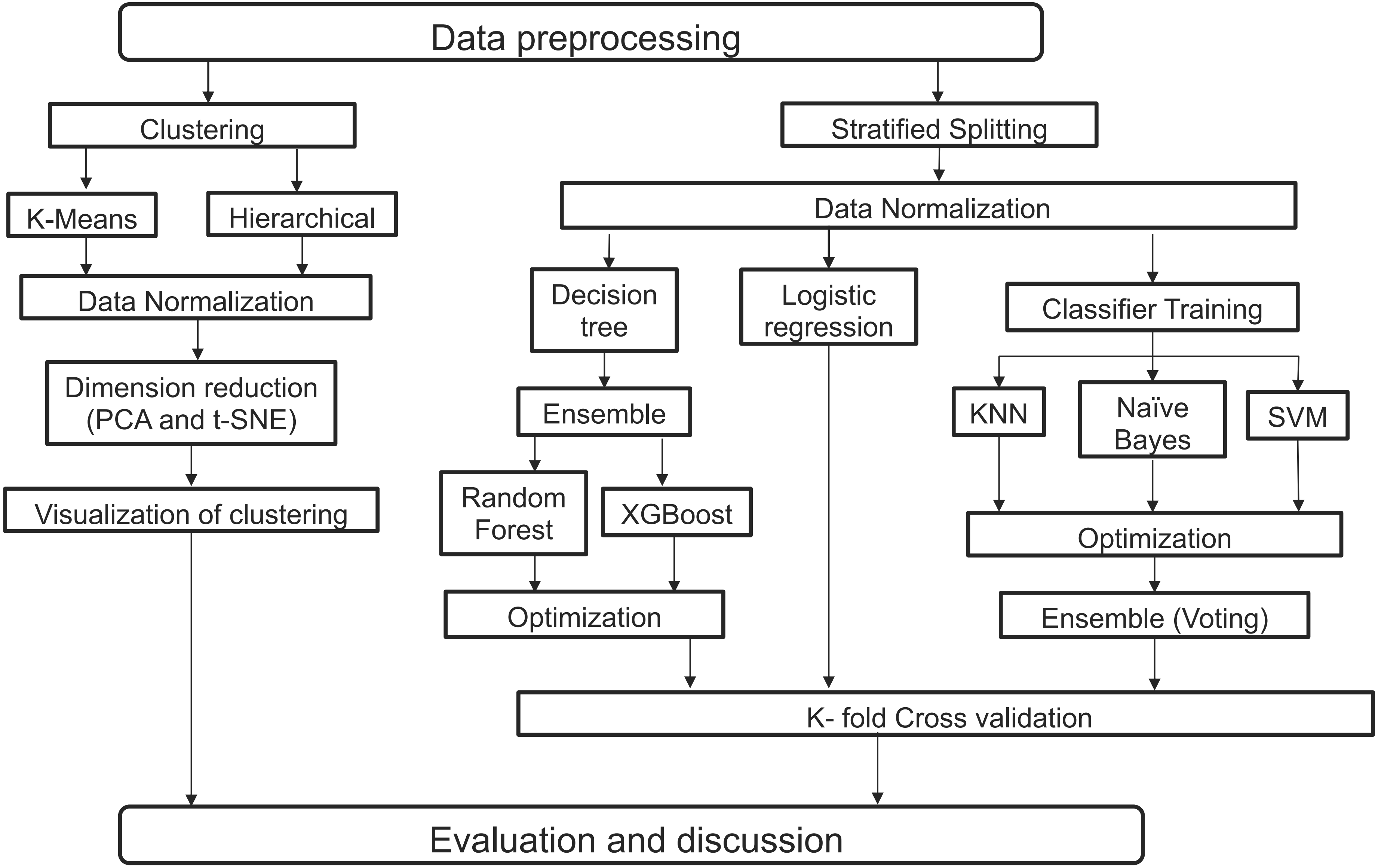
The distribution of features before normalization



The distribution of features after normalization



4. Training and testing procedure



NANYANG TECHNOLOGICAL UNIVERSITY

5. Experimental study and analysis

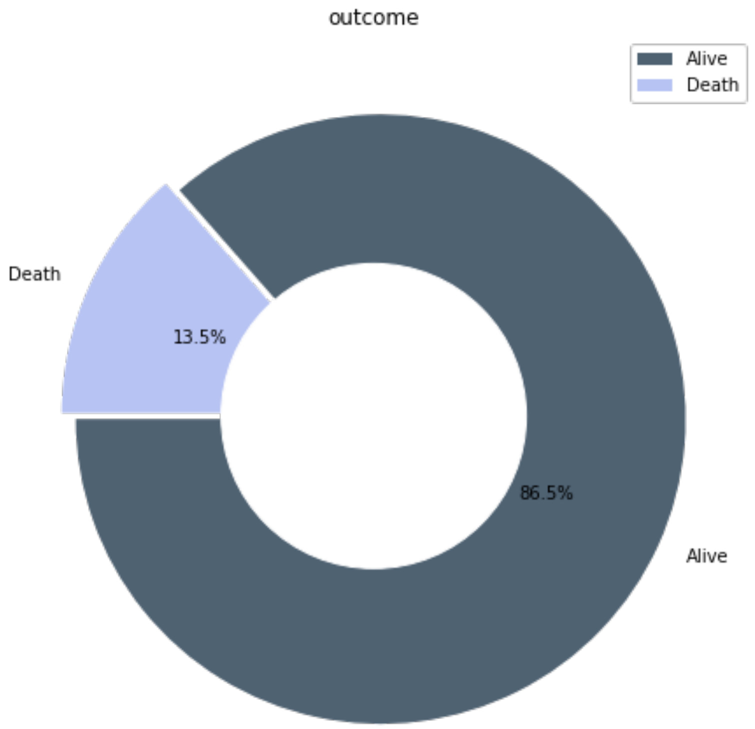
- The class label of the dataset is imbalanced

Strategies to handle imbalanced:

- Stratified splitting

- Evaluate model using AUC

- Add class weight to different label



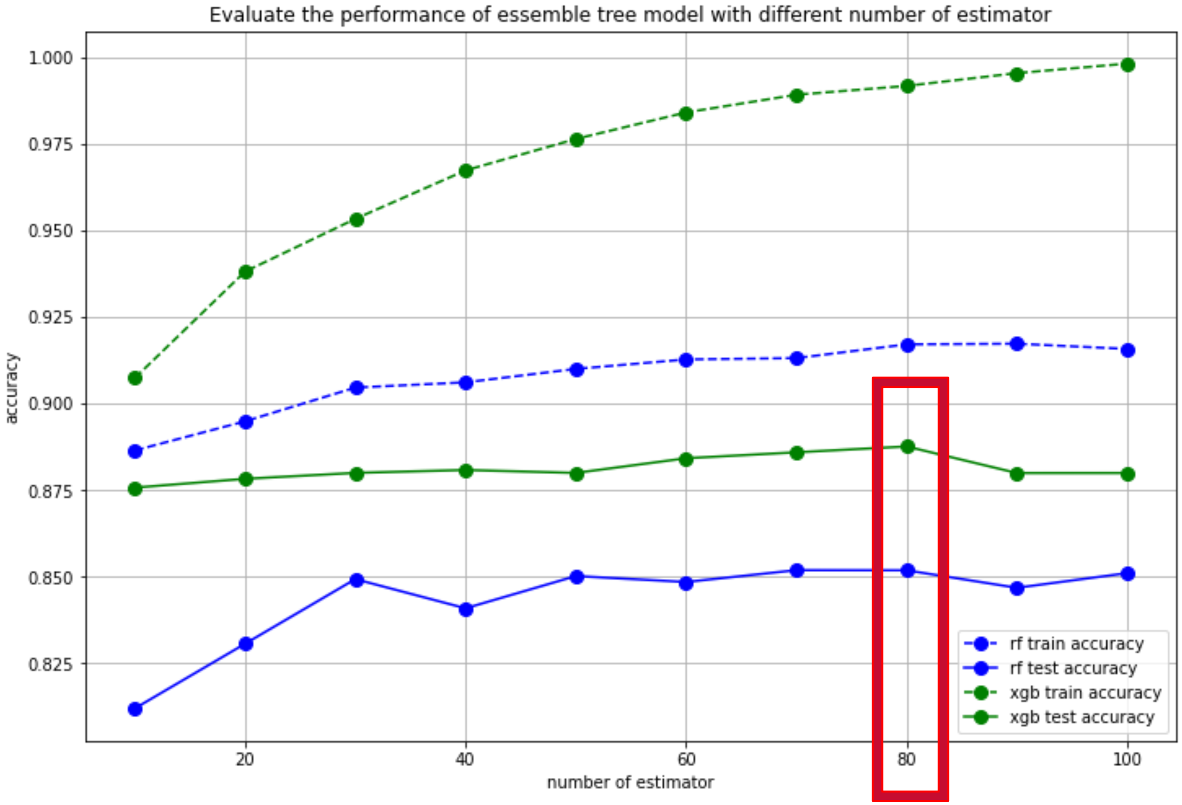
Weight 0= number of samples / (number of classes∗ number of sample with label 0))

5. Experimental study and analysis

1) Tree based model

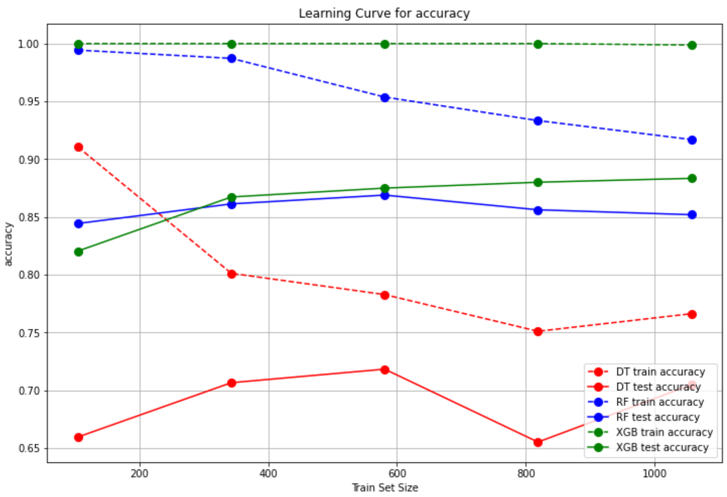
- Ensemble tree based model optimization

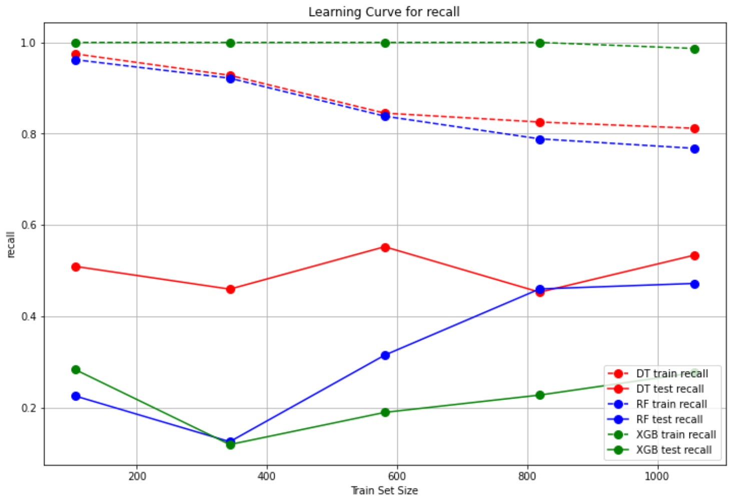
The optimal number of estimator is 80.

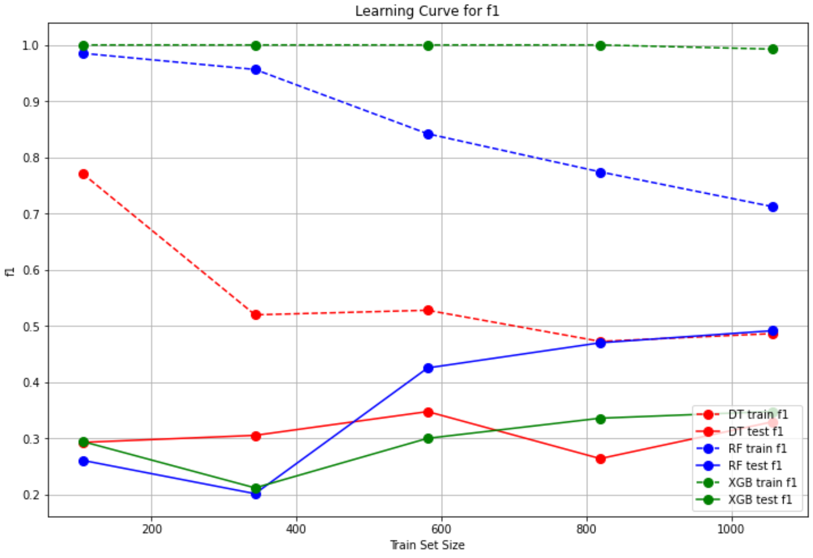


5. Experimental study and analysis

1) Tree based model

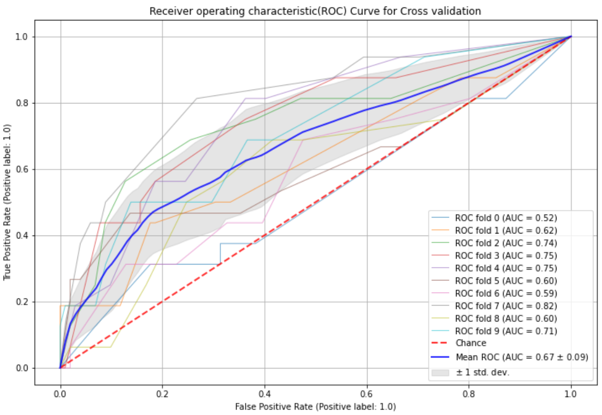


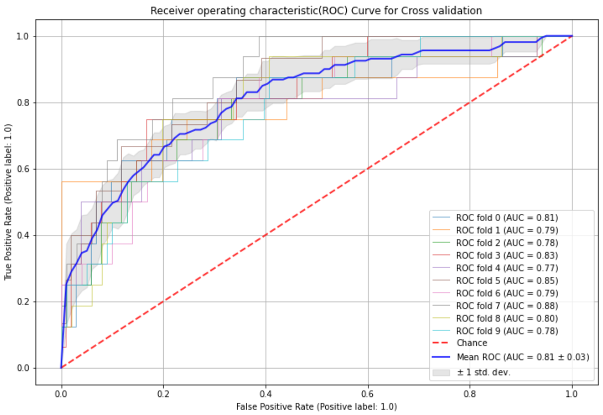


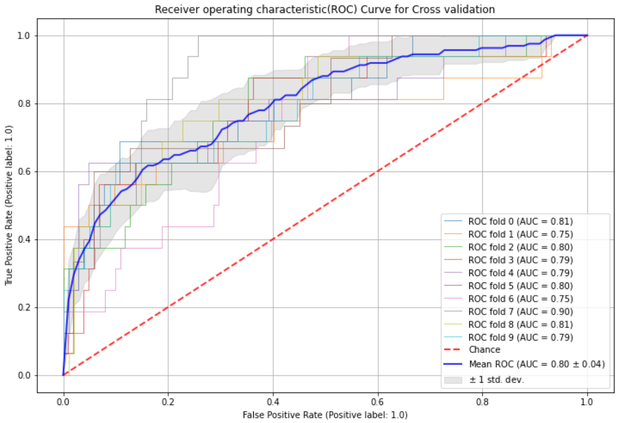


5. Experimental study and analysis

1) Tree based model







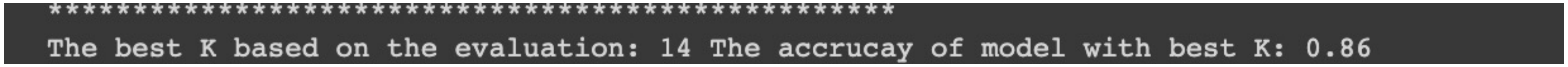
Evaluation scores for tree based model

	accuracy	precision	sensitivity	specificiy	f1-score	AUC
decision tree	0.49	**0.96**	0.43	**0.88**	0.6	0.67
random forest	0.86	0.93	0.91	0.56	0.92	**0.81**
XGBoost	**0.88**	0.89	**0.98**	0.22	**0.93**	0.8

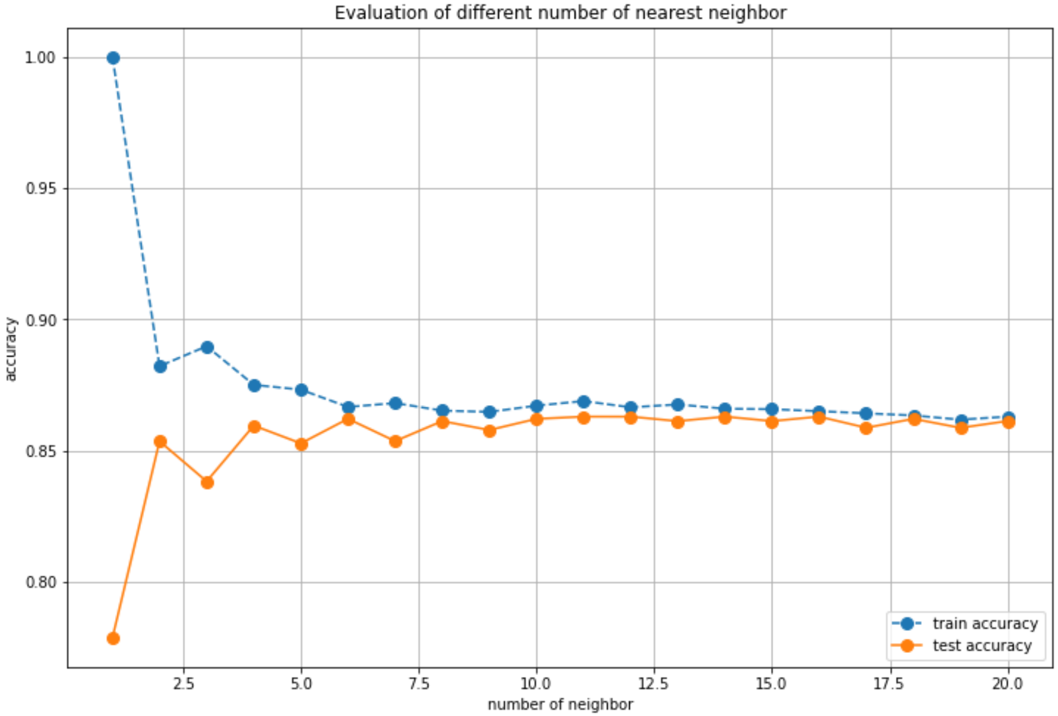
NANYANG
TECHNOLOGICAL
UNIVERSITY

5. Experimental study and analysis

2) KNN optimization

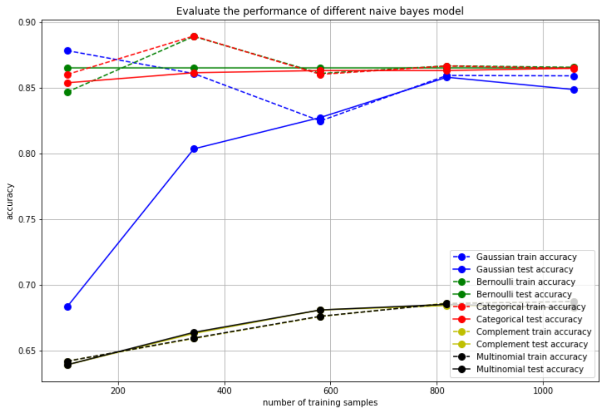


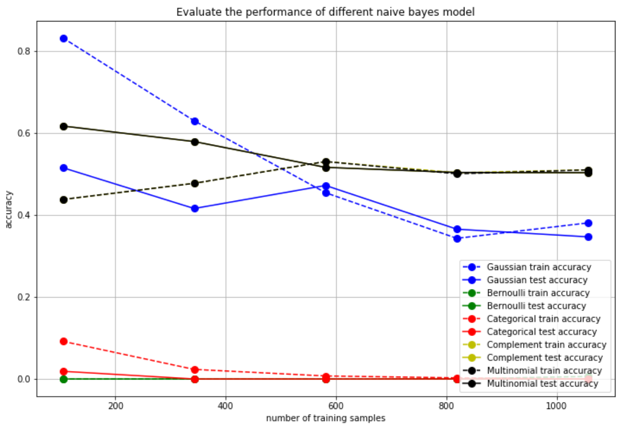
```
**************************************************
The best K based on the evaluation: 14 The accrucay of model with best K: 0.86
```

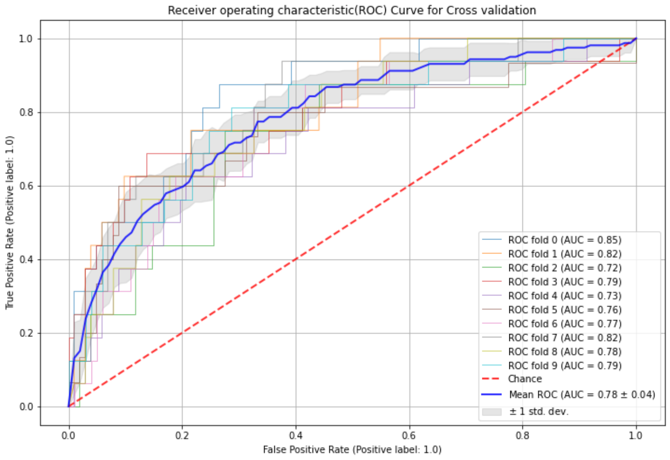


5. Experimental study and analysis

3) Naïve bayes model evaluation


Evaluate the performance of different naive bayes model


Evaluate the performance of different naive bayes model

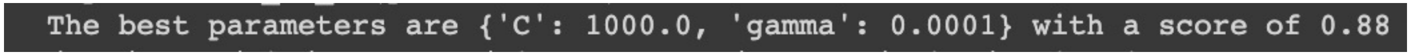

Receiver operating characteristic(ROC) Curve for Cross validation

Evaluation scores for tree based model

	accuracy	precision	sensitivity	f1-score	AUC
Gaussian	0.85	0.45	0.35	0.38	**0.78**
Bernoulli	**0.86**	0	**0**	0	0.64
Categorical	0.86	0	**0**	0	0.64
Complement	0.68	**0.22**	0.50	0.30	0.64
Multinomial	0.68	**0.22**	0.50	0.30	0.64

5. Experimental study and analysis

4) SVM model optimization



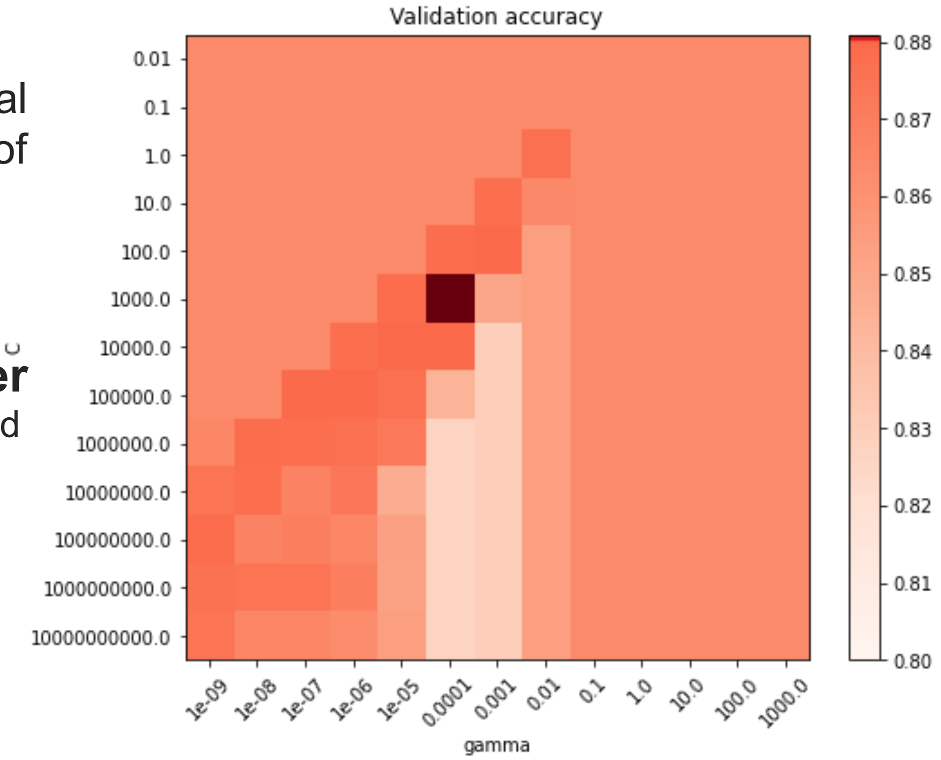
```
The best parameters are {'C': 1000.0, 'gamma': 0.0001} with a score of 0.88
```

- **C：penalty coefficient**

Balance the classification interval margin and the misclassification of samples;

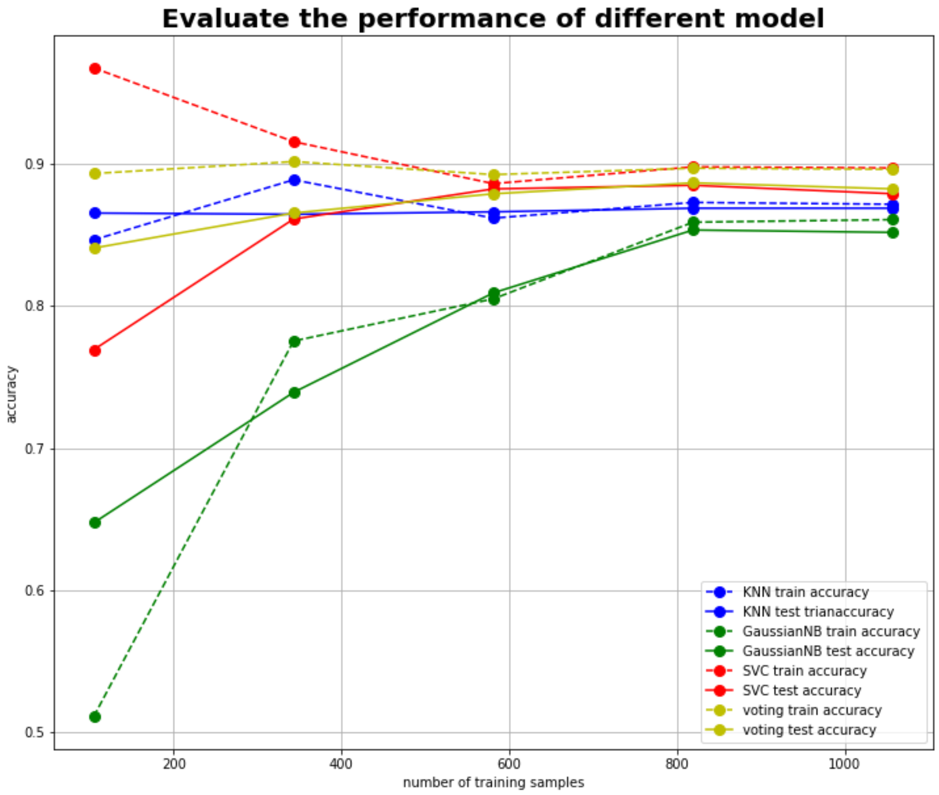
- **gamma: A kernel parameter**

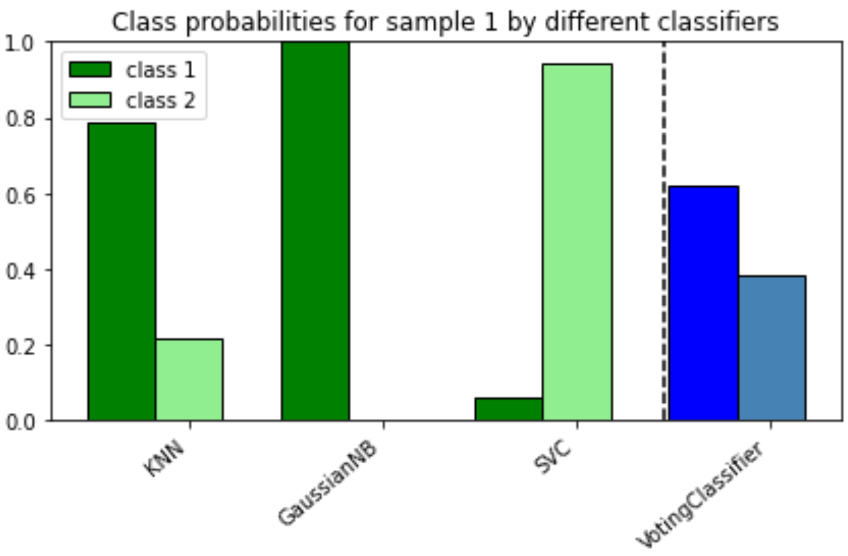
It control the ’spread’ of kernel(how broad the decision region is).



5. Experimental study and analysis

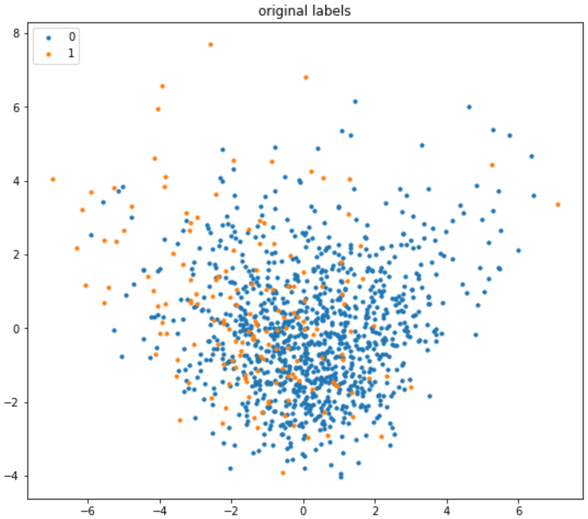
5) Voting for KNN, GaussianNB, and SVC(rbf)



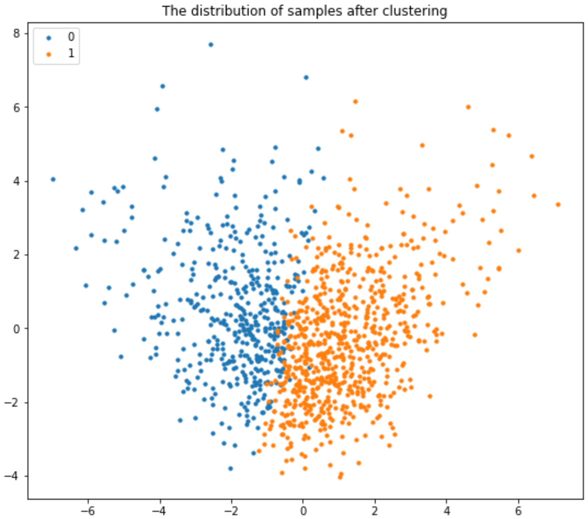


5. Experimental study and analysis

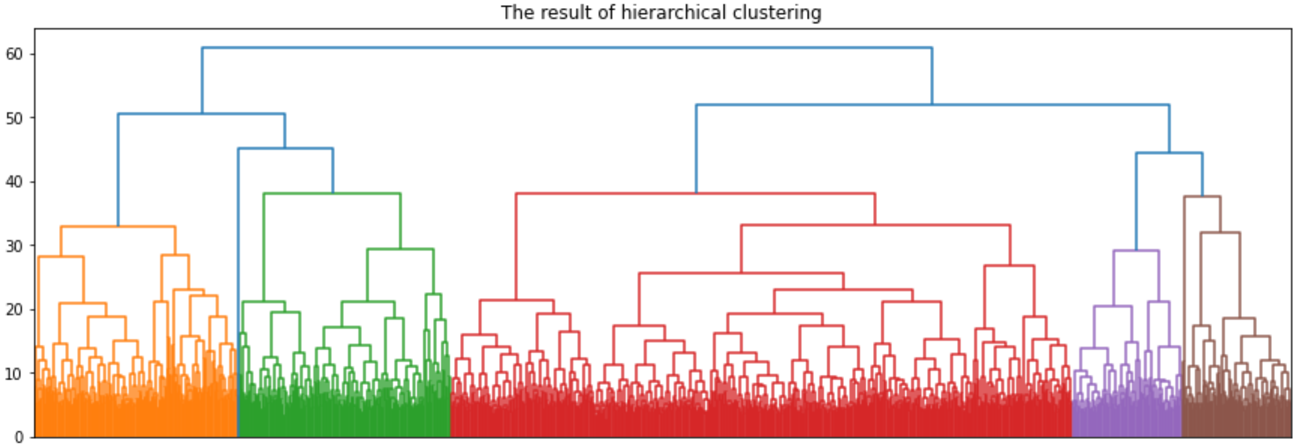
- Result of Clustering



original labels



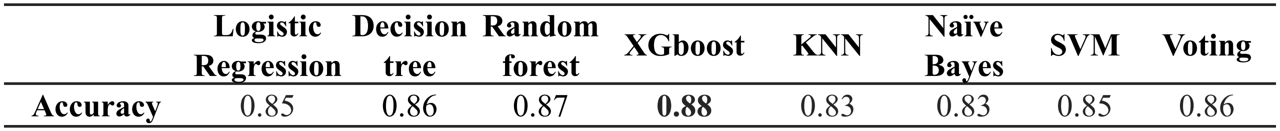
The distribution of samples after clustering



The result of hierarchical clustering

5. Experimental study and analysis

- Final result



	Logistic Regression	Decision tree	Random forest	XGboost	KNN	Naïve Bayes	SVM	Voting
Accuracy	0.85	0.86	0.87	**0.88**	0.83	0.83	0.85	0.86

6) Summary of project achievement

- Implement data exploration and get some insights of the dataset.

- Applied different model on the HF prediction and performed comparison and analysis.

- Achieve the optimal prediction model with 88% prediction accuracy.

7) Future direction for further improvement

- Perform further analysis on the instability of the model performance.

- Perform feature selection.

- Applied other ensemble model to improve the classification accuracy.

Thank you !

NANYANG
TECHNOLOGICAL
UNIVERSITY