Problem 1. Logistic regression model

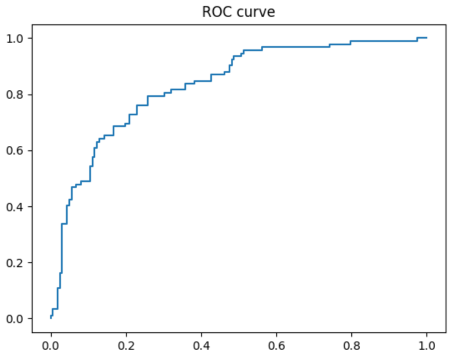
confusion matrix for test data:

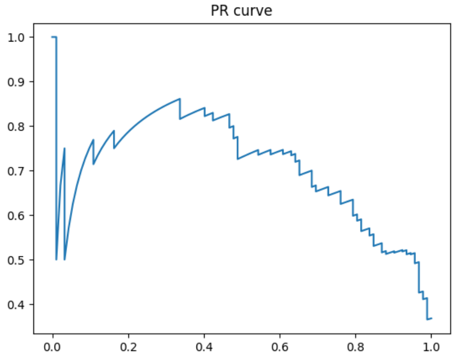
```
 [[142  20]
 [ 34  58]]
```

confusion matrix for train data:

```
 [[299  39]
 [ 77  99]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
Test data	0.2125984251968504	0.6304347826086957	0.8765432098765432	0.7435897435897436	0.8068181818181818
Train data	0.22568093385214008	0.5625	0.8846153846153846	0.717391304347826	0.7952127659574468

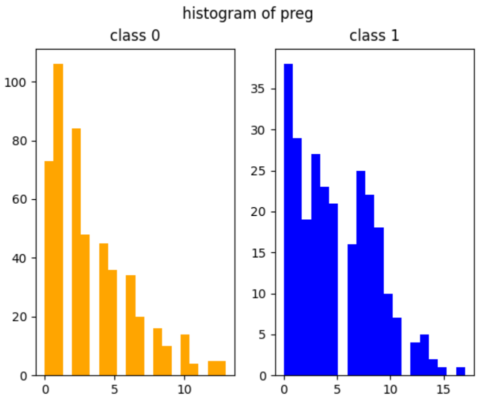


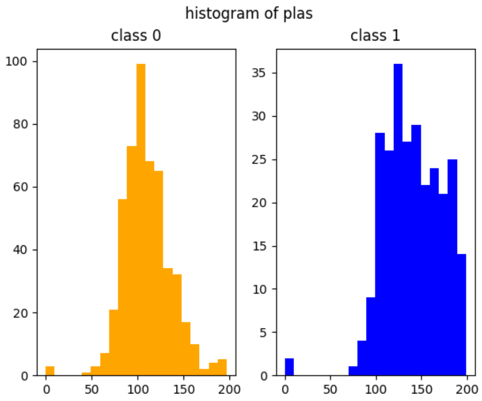


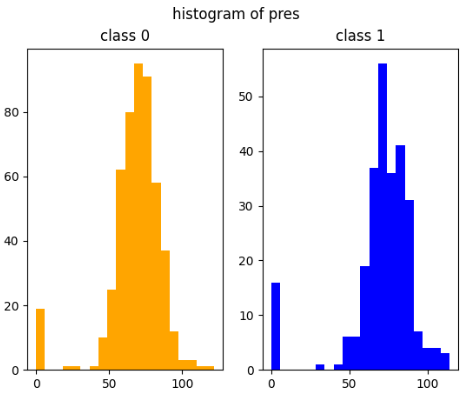
I think the model doesn't overfit the training data. There is not a large discrepancy between the train data and the test data evaluation statistics, and they don't contain extreme/unnormal values. Thus, there is no significant sign that there is an overfitting problem here.

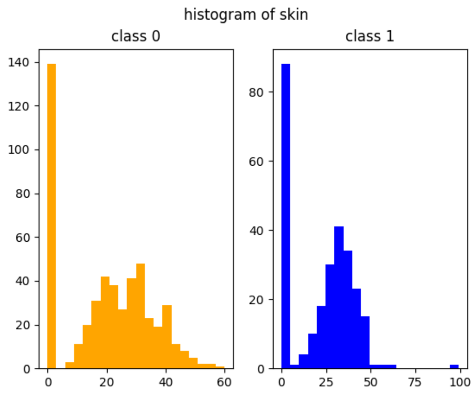
Problem 2. Naive Bayes model

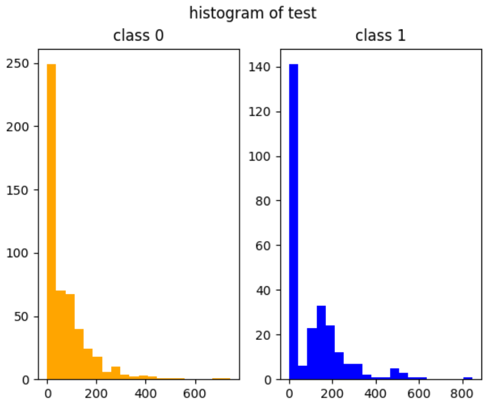
Problem 2.1. Exploratory data analysis

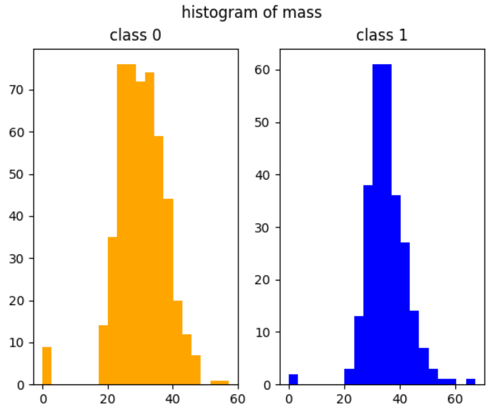


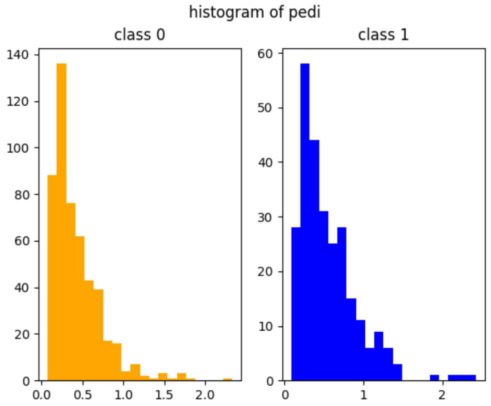


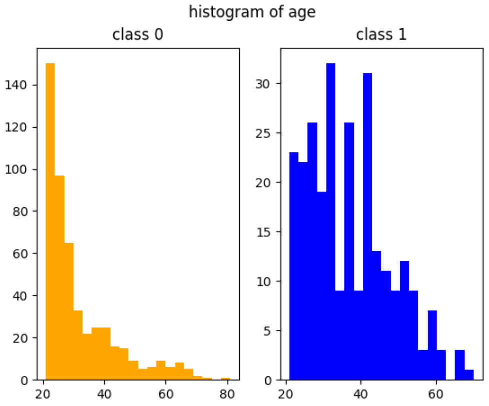












attribute	distribution
preg	Exponential
plas	Normal
pres	Normal
skin	Normal
test	Exponential
mass	Normal
pedi	Exponential
age	Exponential

Part 2.2. The Naive Bayes classifier

(a) the process of finding ML estimate of μ in exponential distribution:

$$\prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{x_i}{\mu}}$$

$$= \frac{1}{\mu^n} \prod_{i=1}^{n} e^{-\frac{x_i}{\mu}}$$

$$take\ log, \ln\left(\frac{1}{\mu^n}\right) - \sum_{i=1}^{n} \frac{x_i}{\mu}$$

$$= \ln\left(\frac{1}{\mu^n}\right) - \frac{1}{\mu}\sum_{i=1}^{n} x_i$$

$$= (-\mathrm{n})\ln \mu - \frac{1}{\mu}\sum_{i=1}^{n} x_i$$

$$take\ derivative, -\frac{n}{\mu} + \frac{\sum_{i=1}^{n} x_i}{\mu^2} = 0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

(b)

Parameters of estimation:

Attribute for class 0	μ	mean	variance
1	3.2544378698224854	-	-
2	-	108.5414201183432	720.6229083630368
3	-	67.5828402366864	307.6800607518482
4	-	19.207100591715978	214.70475655364902
5	69.02366863905326	-	-
6	-	30.399704142011863	62.32620169262375
7	0.4133964497041423	-	-
8	30.72189349112426	-	-

Attribute for class 1	μ	mean	variance
1	4.693181818181818	-	-

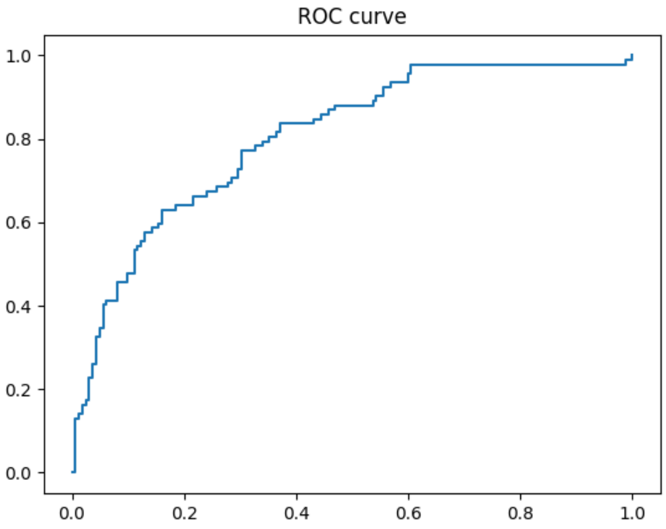
2	-	139.6875	1083.2217857142857
3	-	69.91477272727273	507.00412337662317
4	-	22.363636363636363	330.7127272727273
5	95.32954545454545	-	-
6	-	35.44034090909091	41.58802045454544
7	0.5534545454545456	-	-
8	36.33522727272727	-	-

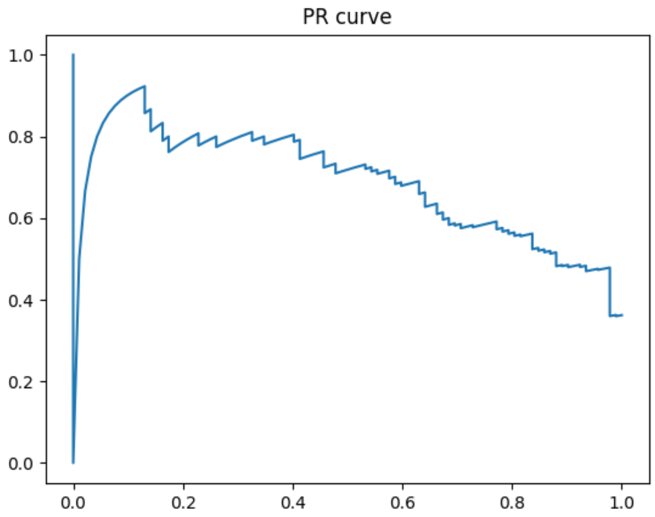
The confusion matrix and other evaluation statistics:

confusion matrix for test data:

```
[[136  26]
[ 35  57]]
```

misclassification error	SENS	SPEC	PPV	NPV
0.24015748031496062	0.6195652173913043	0.8395061728395061	0.6867469879518072	0.7953216374269005





(c)

	AUROC	AUPRC
Problem 1 model	0.83	0.71
Problem 2 model	0.80	0.69

The model in problem 2 (naïve bayes) is better than the logistic regression model in problem 1. As the table shows, both the AUROC and AUPRC in naïve bayes are larger. Larger area means a higher quality and better discriminability between the two classes. Thus, the naïve bayes model is better.