Homework Assignment Hw 6

보고서 및 논문 윤리 서약

1. 나는 보고서 및 논문의 내용을 조작하지 않겠습니다.
2. 나는 다른 사람의 보고서 및 논문의 내용을 내 것처럼 무단으로 복사하지 않겠습니다.
3. 나는 다른 사람의 보고서 및 논문의 내용을 참고하거나 인용할 시 참고 및 인용 형식을 갖추고 출처를 반드시 밝히겠습니다.
4. 나는 보고서 및 논문을 대신하여 작성하도록 청탁하지도 청탁받지도 않겠습니다.

나는 보고서 및 논문 작성 시 위법 행위를 하지 않고, 명지인으로서 또한 공학인으로서 나의 양심과 명예를 지킬 것을 약속합니다.

학 과 : 융합소프트웨어학부 데이터테크놀로지전공

과 목 : 인공지능

담당교수 : 전종훈

강좌 번호: 6019

학 번 : 60182196

이 름 : 이동혁 (서명)

1.

(a).

```
print("(a)", "₩n")

import pandas as pd

df = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)

print(df.shape)
print(df.head(20), "₩n")
```

(a)

```
(768, 9)
    0   1  2  3   4    5     6  7  8
0   6 148 72 35   0 33.6 0.627 50  1
1   1  85 66 29   0 26.6 0.351 31  0
2   8 183 64  0   0 23.3 0.672 32  1
3   1  89 66 23  94 28.1 0.167 21  0
4   0 137 40 35 168 43.1 2.288 33  1
5   5 116 74  0   0 25.6 0.201 30  0
6   3  78 50 32  88 31.0 0.248 26  1
7  10 115  0  0   0 35.3 0.134 29  0
8   2 197 70 45 543 30.5 0.158 53  1
9   8 125 96  0   0  0.0 0.232 54  1
10  4 110 92  0   0 37.6 0.191 30  0
11 10 168 74  0   0 38.0 0.537 34  1
12 10 139 80  0   0 27.1 1.441 57  0
13  1 189 60 23 846 30.1 0.398 59  1
14  5 166 72 19 175 25.8 0.587 51  1
15  7 100  0  0   0 30.0 0.484 32  1
16  0 118 84 47 230 45.8 0.551 31  1
17  7 107 74  0   0 29.6 0.254 31  1
18  1 103 30 38  83 43.3 0.183 33  0
19  1 115 70 30  96 34.6 0.529 32  1
```

(b).

```
print("(b)", "₩n")

import pandas as pd
import numpy as np

df = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)

df[[1,2,3,4,5]] = df[[1,2,3,4,5]].replace(0, np.NaN)

print(df.head(20), "₩n")
print(df.isnull().sum().sum())
```

(b)

```
    0    1    2    3    4    5    6  7  8
0   6  148.0  72.0  35.0    NaN  33.6  0.627  50  1
1   1   85.0  66.0  29.0    NaN  26.6  0.351  31  0
2   8  183.0  64.0   NaN    NaN  23.3  0.672  32  1
3   1   89.0  66.0  23.0   94.0  28.1  0.167  21  0
4   0  137.0  40.0  35.0  168.0  43.1  2.288  33  1
5   5  116.0  74.0   NaN    NaN  25.6  0.201  30  0
6   3   78.0  50.0  32.0   88.0  31.0  0.248  26  1
7  10  115.0   NaN   NaN    NaN  35.3  0.134  29  0
8   2  197.0  70.0  45.0  543.0  30.5  0.158  53  1
9   8  125.0  96.0   NaN    NaN   NaN  0.232  54  1
10  4  110.0  92.0   NaN    NaN  37.6  0.191  30  0
11 10  168.0  74.0   NaN    NaN  38.0  0.537  34  1
12 10  139.0  80.0   NaN    NaN  27.1  1.441  57  0
13  1  189.0  60.0  23.0  846.0  30.1  0.398  59  1
14  5  166.0  72.0  19.0  175.0  25.8  0.587  51  1
15  7  100.0   NaN   NaN    NaN  30.0  0.484  32  1
16  0  118.0  84.0  47.0  230.0  45.8  0.551  31  1
17  7  107.0  74.0   NaN    NaN  29.6  0.254  31  1
18  1  103.0  30.0  38.0   83.0  43.3  0.183  33  0
19  1  115.0  70.0  30.0   96.0  34.6  0.529  32  1

652
```

(c).

```
print("(c)", "₩n")

import pandas as pd
import numpy as np
from scipy import sparse


df = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)

for c in range(len(df[0])):
    if(df[0][c]==0):
        df[0][c] = 'a'
    elif(1<=df[0][c] and df[0][c]<=3):
        df[0][c] = 'b'
    else:
        df[0][c] = 'c'

df = pd.get_dummies(df[0], sparse = True)
x2 = sparse.csr_matrix(df)

print(x2[80:100]) # sparse matrix
print(type(x2), "₩n")

x3 = x2.toarray() # numpy array
print(x3[80:100])
print(type(x3), "₩n")

x4 = pd.DataFrame(x3) # DataFrame
print(x4[80:100])
print(type(x4), "₩n")

# https://stackoverflow.com/questions/20459536/convert-pandas-dataframe-to-sparse-numpy-matrix-directly
# sparse matrix로 변환하는 법
```

(c)

```
  (0, 1)      1
  (1, 1)      1
  (2, 2)      1
  (3, 0)      1
  (4, 2)      1
  (5, 1)      1
  (6, 2)      1
  (7, 1)      1
  (8, 2)      1
  (9, 1)      1
  (10, 1)     1
  (11, 2)     1
  (12, 2)     1
  (13, 2)     1
  (14, 1)     1
  (15, 2)     1
  (16, 1)     1
  (17, 1)     1
  (18, 2)     1
  (19, 1)     1
<class 'scipy.sparse.csr.csr_matrix'>
```

```
[[0 1 0]
 [0 1 0]
 [0 0 1]
 [1 0 0]
 [0 0 1]
 [0 1 0]
 [0 0 1]
 [0 1 0]
 [0 0 1]
 [0 1 0]
 [0 1 0]
 [0 0 1]
 [0 0 1]
 [0 0 1]
 [0 1 0]
 [0 0 1]
 [0 1 0]
 [0 1 0]
 [0 0 1]
 [0 1 0]]
<class 'numpy.ndarray'>
```

```
    0  1  2
80  0  1  0
81  0  1  0
82  0  0  1
83  1  0  0
84  0  0  1
85  0  1  0
86  0  0  1
87  0  1  0
88  0  0  1
89  0  1  0
90  0  1  0
91  0  0  1
92  0  0  1
93  0  0  1
94  0  1  0
95  0  0  1
96  0  1  0
97  0  1  0
98  0  0  1
99  0  1  0
<class 'pandas.core.frame.DataFrame'>
```

(d).

```
print("(d)", "₩n")

import pandas as pd

dataset = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)

old = dataset.shape

dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, np.NaN)
dataset.dropna(inplace = True)

new = dataset.shape

print(old)
print(new)
```

```
(d)

(768, 9)
(392, 9)
```

(e).

```
print("(e)", "₩n")
# https://jimmy-ai.tistory.com/162 결측치 추출
# https://computer-science-student.tistory.com/375 특정 조건 추출

import pandas as pd
import numpy as np

df = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)
df[[1,2,3,4,5]] = df[[1,2,3,4,5]].replace(0, np.NaN)

df_mean = df.mean()
df.fillna(df_mean, inplace = True) # NaN이 mean으로 대체

# 특정 조건 추출
df = df[(df[1] == df_mean[1]) |(df[2] == df_mean[2]) |(df[3] == df_mean[3]) |(df[4] == df_mean[4]) |(df[5] == df_mean[5])]
print(df.head(10), "₩n")
print(len(df))
```

(e)

```
    0    1       2      3       4       5    6  7  8
0   6  148.0  72.000000  35.00000  155.548223  33.600000  0.627  50  1
1   1   85.0  66.000000  29.00000  155.548223  26.600000  0.351  31  0
2   8  183.0  64.000000  29.15342  155.548223  23.300000  0.672  32  1
5   5  116.0  74.000000  29.15342  155.548223  25.600000  0.201  30  0
7  10  115.0  72.405184  29.15342  155.548223  35.300000  0.134  29  0
9   8  125.0  96.000000  29.15342  155.548223  32.457464  0.232  54  1
10  4  110.0  92.000000  29.15342  155.548223  37.600000  0.191  30  0
11 10  168.0  74.000000  29.15342  155.548223  38.000000  0.537  34  1
12 10  139.0  80.000000  29.15342  155.548223  27.100000  1.441  57  0
15  7  100.0  72.405184  29.15342  155.548223  30.000000  0.484  32  1

376
```

(f).

```
print("(f)", "₩n")

import pandas as pd

df = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/pima-indians-diabetes.csv', header = None)

pos = df[df[8]==1].sample(n = 100, replace = False)
neg = df[df[8]==0].sample(n = 100, replace = False)

# 합치기
train = pd.concat([pos[:50],neg[:50]])
test = pd.concat([pos[50:],neg[50:]])

# 섞기
train = train.sample(frac = 1)
test = test.sample(frac = 1)

train.to_csv('train.csv', index = False)
test.to_csv('test.csv', index = False)

print("mission complete!", "₩n")

# 확인
train = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/train.csv', header = None).drop([0],axis=0)
test = pd.read_csv('/Volumes/GoogleDrive-107262488266475120044/내 드라이브/3-1/인공지능/py/test.csv', header = None).drop([0],axis=0)

# csv로 저장할 때 column 번호가 자동으로 생성됨
# 따라서 .drop([0],axis=0)을 통해 제거한 후 로딩 필요

print("train length :", len(train))
print(train.head(), "₩n")

print("test length :", len(test))
print(test.head(), "₩n")

print(train[8].sum())
print(test[8].sum())
```

(f)

mission complete!

train length : 100

	0	1	2	3	4	5	6	7	8
1	4	132	0	0	0	32.9	0.302	23	1
2	4	117	62	12	0	29.7	0.380	30	1
3	2	105	58	40	94	34.9	0.225	25	0
4	1	124	74	36	0	27.8	0.100	30	0
5	7	107	74	0	0	29.6	0.254	31	1

test length : 100

	0	1	2	3	4	5	6	7	8
1	7	129	68	49	125	38.5	0.439	43	1
2	3	158	76	36	245	31.6	0.851	28	1
3	2	120	76	37	105	39.7	0.215	29	0
4	8	99	84	0	0	35.4	0.388	50	0
5	5	168	64	0	0	32.9	0.135	41	1

50
50