Assignment 3

สมาชิกกลุ่ม

1. จิตติณณ์ จินดานรเศรษฐ์ 6110405949 (หัวหน้ากลุ่ม)
2. รินลณี วัชรนิมมานกุล 6110401633
3. กัญญาณัฐ อินทรโชติ 6110402737
4. ณัฐณิชา คงสุนทร 6110402753
5. นิรัติศัย คงศักดิ์ 6110406066
6. พิมพ์ลภัส ตันธนกุล 6110406171
7. สุรยุทธ์ บุญคล้าย 6110406252

ชุดข้อมูล คือ Contraceptive Method Choice ซึ่งเป็นชุดข้อมูลที่ระบุวิธีการคุมกำเนิดที่เลือกใช้ โดยมีผลสำรวจมาจากผู้หญิงประเทศอินโดนีเซียที่แต่งงานในปี ค.ศ. 1987

Download Dataset: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

Feature คือ คุณลักษณะของข้อมูล ซึ่งแต่ละคอลัมน์หมายถึงดังนี้

- wife_age คือ อายุของภรรยา
- wife_edu คือ ระดับการศึกษาของภรรยา
- husband_edu คือ ระดับการศึกษาของสามี
- children คือ จำนวนเด็กในครอบครัวนั้น ๆ
- wife_islam คือ ภรรยานับถือศาสนาอิสลามหรือไม่
- wife_working คือ ภรรยาทำงานหรือไม่
- husband_job คือ ระยะเวลาของการทำงาน
- living_index คือ ค่าครองชีพ
- media_exposure คือ การเปิดรับสื่อ

Target class คือ คอลัมน์ method โดยที่ method คือ การวางแผนการคุมกำเนิด [No use (ไม่มีการวางแผนการคุมกำเนิด), short-term (การวางแผนการคุมกำเนิดระยะสั้น), long-term (การวางแผนการคุมกำเนิดระยะยาว)]

การออกแบบการทดลอง

ขั้นตอนการออกแบบการทดลองมีดังนี้

1. ทำการเปลี่ยนแปลงชนิดของข้อมูลในตารางให้เหมาะสมกับลักษณะข้อมูล
2. แก้ไขข้อมูลชนิดข้อมูลจากตัวเลขให้เป็นแบบ Category หรือ Boolean
3. เขียนโปรแกรมให้เรียนรู้และทำนายผล

วิธีการ Preprocess

1. ติดตั้ง Anaconda Navigator เพื่อใช้เครื่องมือ Jupyter Notebook ในการออกแบบการทดลอง โดยเรียกใช้ Library ชื่อ Pandas ในการจัดการข้อมูล
2. ทำการ import library ที่ต้องการใช้งานเข้ามา

```
In [1]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plot
        from sklearn import preprocessing
        from collections import Counter
        from sklearn.preprocessing import MinMaxScaler
```

3. ทำการอ่านไฟล์ cmc.data เข้ามา เนื่องจากชุดข้อมูลไม่มีหัวตารางมาดังนั้นต้องการทำการกำหนด header = none เพื่อป้องแถวของชุดข้อมูลในแถวแรกกลายเป็นชื่อคอมลัมน์

```
In [2]: df = pd.read_csv('cmc.data', header=None)
```

4. ทำการสร้างชื่อคอลัมน์ให้กับชุดข้อมูลดังนี้

```
In [3]: df.columns = ['wife_age', 'wife_edu', 'husband_edu', 'children', 'wife_islam', 'wife_working'
                      ,'husband_job', 'living_index', 'media_exposure', 'method']
```

5. เนื่องจากชนิดของข้อมูลในทุก ๆ คอลัมน์เมื่อทำการ import เข้ามาจะกลายเป็น int64 ทั้งหมด จึงทำการปรับเปลี่ยนชนิดข้อมูลแต่ละคอลัมน์ให้เหมาะสม

```
In [5]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
wife_age          1473 non-null int64
wife_edu          1473 non-null int64
husband_edu       1473 non-null int64
children          1473 non-null int64
wife_islam        1473 non-null int64
wife_working      1473 non-null int64
husband_job       1473 non-null int64
living_index      1473 non-null int64
media_exposure    1473 non-null int64
method            1473 non-null int64
dtypes: int64(10)
memory usage: 115.2 KB
```

5.1 ภาพด้านล่างนี้คือคอลัมน์ที่จะปรับให้เป็น Category โดยที่จะทำการเปลี่ยนข้อมูลตัวเลขให้เป็นนามบัญญัติ

```
In [6]: df.wife_edu = pd.Categorical(df.wife_edu).rename_categories({1:'low', 2:'mid-low', 3:'mid-high', 4:'high'})
        df.husband_edu = pd.Categorical(df.husband_edu).rename_categories({1:'low', 2:'mid-low', 3:'mid-high', 4:'high'})
        df.husband_job = pd.Categorical(df.husband_job)
        df.living_index =  pd.Categorical(df.living_index).rename_categories({1:'low', 2:'mid-low', 3:'mid-high', 4:'high'})
        df.method =  pd.Categorical(df.method).rename_categories({1:'No-use', 2:'Long-term', 3:'Short-term'})
```

5.2 ภาพข้างล่างนี้คือคอลัมน์ที่จะปรับให้เป็น Boolean เนื่องจากมีค่าที่เป็นไปได้เพียงสองค่า

```
In [7]: df.wife_islam = df.wife_islam.astype(bool)
        df.wife_working = df.wife_working.astype(bool)
        df.media_exposure = df.media_exposure.astype(bool)
```

5.3 คอลัมน์อายุมีค่าที่หลากหลายจึงทำการปรับให้เป็นช่วง ๆ ดังภาพข้างล่าง

```
In [8]: df.wife_age = pd.cut(df.wife_age, bins=[0,17,31,46,100], right=False,
                             labels=["Children", "Young", "Middle", "Old"])
```

5.4 ในคอลัมน์ children จะทำการปรับให้ทุกค่าในคอลัมน์นี้อยู่ในช่วง [0,1] ด้วยการใช้ MinMaxScaler ดังภาพข้างล่าง

```
In [9]: x = df['children'].values.reshape(-1, 1)
        min_max_scaler = MinMaxScaler()
        x_scaled = min_max_scaler.fit_transform(x)
        df['children'] = x_scaled
```

6. ลักษณะของตารางและชนิดของข้อมูลในแต่ละคอลัมน์หลังจากการทำ preprocess ตามข้อที่ผ่านมา

	wife_age	wife_edu	husband_edu	children	wife_islam	wife_working	husband_job	living_index	media_exposure	method
0	Young	mid-low	mid-high	0.1875	True	True	2	mid-high	False	No-use
1	Middle	low	mid-high	0.6250	True	True	3	high	False	No-use
2	Middle	mid-low	mid-high	0.4375	True	True	3	high	False	No-use
3	Middle	mid-high	mid-low	0.5625	True	True	3	mid-high	False	No-use
4	Middle	mid-high	mid-high	0.5000	True	True	3	mid-low	False	No-use
...	...	...	...	...	...	...	...	...	...	...
1468	Middle	high	high	0.1250	True	False	2	high	False	Short-term
1469	Middle	high	high	0.1875	True	True	1	high	False	Short-term
1470	Middle	mid-high	mid-high	0.5000	True	False	1	high	False	Short-term
1471	Middle	mid-high	mid-high	0.2500	True	False	2	mid-low	False	Short-term
1472	Young	mid-high	mid-high	0.0625	True	True	2	high	False	Short-term

```
In [11]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
wife_age          1473 non-null category
wife_edu          1473 non-null category
husband_edu       1473 non-null category
children          1473 non-null float64
wife_islam        1473 non-null bool
wife_working      1473 non-null bool
husband_job       1473 non-null category
living_index      1473 non-null category
media_exposure    1473 non-null bool
method            1473 non-null category
dtypes: bool(3), category(6), float64(1)
memory usage: 25.6 KB
```

การเรียนรู้และการทำนายผล

โดยเราจะแบ่ง dataset เป็นสองส่วนคือ train dataset 80% และ test dataset 20% โดยที่ X จะคือ feature และ y คือผลเฉลยดังภาพด้านล่าง

```
In [12]: from sklearn.model_selection import train_test_split
```

```
In [13]: X = pd.get_dummies(df.loc[:, 'wife_age':'media_exposure'])
         y = df.method
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

1. SVM (Support Vector Machine)

หาค่า hyperparameter ที่เหมาะสมสำหรับ SVM โดยการใช้ GridSearchCV และกำหนด kernel และ gamma ตามที่ต้องการดังภาพด้านล่าง In[16] ต่อมาทำการแสดงค่ากลุ่ม hyperparameter ที่ให้ผลลัพธ์ที่ดีที่สุดตามในส่วนของ Out[17]

```
In [14]: from sklearn.model_selection import GridSearchCV
```

SVM

```
In [15]: from sklearn import svm
```

```
In [16]: param_grid = {
             'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
             'gamma': ['scale', 'auto', 2**-5]
         }
```

```
In [17]: svc = svm.SVC()
         SVM = GridSearchCV(svc, param_grid)
         SVM.fit(X_train, y_train)
         SVM.best_params_
```

```
Out[17]: {'gamma': 'scale', 'kernel': 'rbf'}
```

โดยเมื่อปรับค่า hyperparameter ตาม GridSearchCV จะให้ค่าประสิทธิภาพคือ 0.4915254237288136

```
In [18]: svm_model = svm.SVC(kernel='rbf', gamma='scale')

In [19]: svm_model.fit(X_train, y_train)
         svm_model.score(X_test,y_test)

Out[19]: 0.4915254237288136
```

2. Bagging

ขั้นตอนวิธีที่ใช้ในการเรียนรู้คือ Decision tree เป็น base classifier โดยใช้วิธีวัดสัดส่วนการปะปนคือจีนีอินเด็กซ์ และกำหนดให้มี base classifier 100 ตัว และมีการสุ่มแบบ bootstrap

Bagging

```
In [20]: from sklearn.tree import DecisionTreeClassifier
         from sklearn.ensemble import BaggingClassifier

In [21]: dtc = DecisionTreeClassifier(criterion="gini")
         bag_model = BaggingClassifier(base_estimator=dtc, n_estimators=100, bootstrap=True, random_state=0)
         bag_model = bag_model.fit(X_train, y_train)
         bag_model.score(X_test,y_test)

Out[21]: 0.5152542372881356
```

ค่าประสิทธิภาพที่ได้คือ 0.5152542372881356

3. ADABOOST

Boosting

```
In [22]: from sklearn.ensemble import AdaBoostClassifier

In [23]: ada_model = AdaBoostClassifier(DecisionTreeClassifier(max_depth=1), n_estimators=100)
         ada_model.fit(X_train, y_train)
         ada_model.score(X_test,y_test)

Out[23]: 0.5322033898305085
```

ค่าประสิทธิภาพที่ได้คือ 0.5322033898305085

สรุปผลการทดลอง

จากผลการทดลองการทำ preprocessing ข้อมูลและวัดค่าประสิทธิภาพการเรียนรู้ของแต่ละอัลกอริทึมในการแบ่งข้อมูลฝึกและการสุ่มตัวอย่าง

- SVM วัดค่าประสิทธิภาพการเรียนรู้ได้ 0.4915254237288136
- Bagging วัดค่าประสิทธิภาพการเรียนรู้ได้ 0.5152542372881356
- ADABOOST วัดค่าประสิทธิภาพการเรียนรู้ได้ 0.5322033898305085

จากการวัดค่าประสิทธิภาพการเรียนรู้อัลกอริทึมที่ให้ค่าประสิทธิภาพที่ดีที่สุดคือ ADABOOST เนื่องจากเป็นอัลกอริทึมที่มีการให้ค่าน้ำหนักกับข้อมูลฝึกที่จำแนกผิดบ่อยจึงทำให้มีการเรียนรู้ที่ดีขึ้น และจะเห็นได้ว่าทั้งสามอัลกอริทึม ให้ค่าประสิทธิภาพการเรียนรู้ที่ค่อนข้างต่ำโดยอ้างอิงจากเอกสารที่เกี่ยวข้อง มีเพียงไม่กี่อัลกอริทึมที่สามารถใช้งานได้ แต่ก็ยังคงมีค่า Error rates ค่อนข้างสูงคือ ต่ำสุดที่ 0.43 และสูงสุดที่ 0.60

Table 4. Minimum, maximum, and 'naive' plurality rule error rates for each dataset. A '√'-mark indicates that the algorithm has an error rate within one standard error of the minimum for the dataset. A '**X**'-mark indicates that the algorithm has the worst error rate for the dataset. The mean error rate for each algorithm is given in the second row.

	Decision trees and rules																						Statistical algorithms									Nets		Error rates		
	QUO	QU1	QLO	QL1	FTU	FTL	C4T	C4R	IB	IBO	IM	IMO	ICO	IC1	OCU	OCL	OCM	STO	ST1	LMT	CAL	T1	LDA	QDA	NN	LOG	FM1	FM2	PDA	MDA	POL	LVQ	RBF	Min	Max	Naive
Mean	.221	.226	.208	.211	.238	.234	.220	.220	.229	.219	.220	.219	.215	.227	.227	.260	.230	.232	.233	.220	.270	.354	.208	.301	.281	.204	.242	.217	.213	.207	.195	.269	.257			
#√	8	8	10	9	4	12	7	8	3	8	5	6	4	5	9	4	4	5	5	1	4	4	10	4	4	13	12	12	10	9	15	4	8			
#X	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	1	0	1	0	0	11	0	3	4	0	2	1	0	1	0	4	0			
cmc	√	√																							X						√			.43	.60	.57
cmc+	√	√																							X						√			.43	.58	

Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning