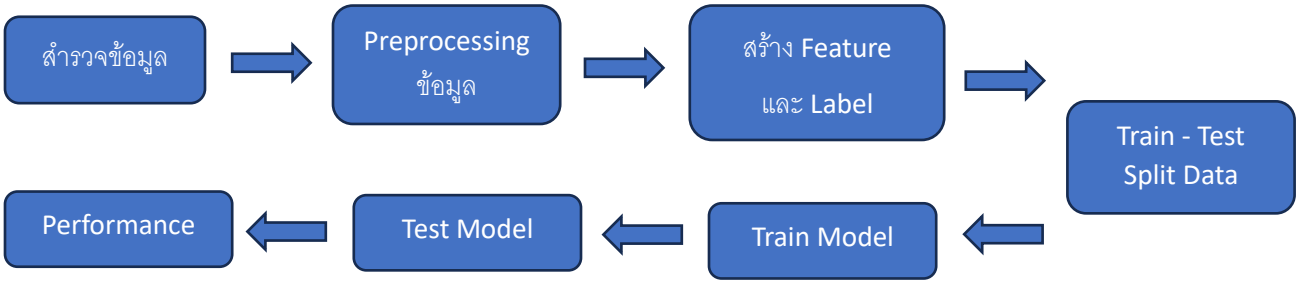
กระบวนการจำแนกผู้ป่วยโรคอัลไซเมอร์

<u>แผนภาพการดำเนินการทดลอง</u>



1. สำรวจชุดข้อมูล Alzheimer.csv ว่ามี Feature - Label อะไรบ้าง และมีชนิดข้อมูลเป็นอย่างไร

	HN	Education	age	hypertension	heart_disease	family history	Occupation	Triglycerides	Blood Sugar	BMI	smoking_status	alzheimer
0	11046	Secondary School	67.0	0	1	Yes	Private	Normal	228.69	36.6	formerly smoked	1
1	51676	Bachelor	61.0	0	0	Yes	Freelance	High	202.21	NaN	never smoked	1
2	35112	Secondary School	80.0	0	1	Yes	Private	High	105.92	32.5	never smoked	1
3	67182	Bachelor	49.0	0	0	Yes	Private	Normal	171.23	34.4	smokes	1
4	1665	Bachelor	79.0	1	0	Yes	Freelance	High	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
4095	68398	Secondary School	82.0	1	0	Yes	Freelance	High	71.97	28.3	never smoked	0
4096	14180	Bachelor	13.0	0	0	No	Farmer	High	103.08	18.6	Unknown	0
4097	44873	Bachelor	81.0	0	0	Yes	Freelance	Normal	125.20	40.0	never smoked	0
4098	19723	Bachelor	35.0	0	0	Yes	Freelance	High	82.99	30.6	never smoked	0
4099	44679	Bachelor	44.0	0	0	Yes	Officer	Normal	85.28	26.2	Unknown	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4100 entries, 0 to 4099
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   HN              4100 non-null   int64
 1   Education       4100 non-null   object
 2   age             4100 non-null   float64
 3   hypertension    4100 non-null   int64
 4   heart_disease   4100 non-null   int64
 5   family history  4100 non-null   object
 6   Occupation      4100 non-null   object
 7   Triglycerides   4100 non-null   object
 8   Blood Sugar     4100 non-null   float64
 9   BMI             3936 non-null   float64
 10  smoking_status  4100 non-null   object
 11  alzheimer       4100 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 384.5+ KB
```

2. ตรวจสอบชุดข้อมูลและทำ Preprocessing Data โดยการแปลง Feature ที่เป็น Nominal ให้อยู่ในรูปของตัวเลขผ่านการใช้ LabelEncoder และจัดการ Missing Data โดยการ Drop row ที่ไม่มีข้อมูล BMI เนื่องจากไม่สามารถคำนวณหาค่า BMI ได้จาก Feature ที่มีอยู่ได้

3. แบ่งข้อมูลออกเป็น Feature และ Label โดย กำหนดให้ Feature มี column ดังนี้ 'age', 'heart_disease', 'family history', 'BMI', 'smoking_status', 'hypertension', 'Triglycerides', 'Blood Sugar' และ Label คือ alzheimer column

4. หลังจากสร้าง Feature – Label แล้ว ทำการแบ่งชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ โดยกำหนดชุดข้อมูลฝึกฝนมี 70% และชุดข้อมูลทดสอบมี 30%

```
y_train.value_counts()

0    2617
1     138
Name: alzheimer, dtype: int64
```

```
y_test.value_counts()

0    1117
1      64
Name: alzheimer, dtype: int64
```

5. ทำการฝึกฝนโมเดลผ่านชุดข้อมูลฝึกฝน โดยใช้ 3 โมเดล คือ Decision Tree, K-nearest neighbors และ Neural Network และใช้ GridSearchCV ในการปรับ Parameter ของแต่ละโมเดลให้เหมาะสม

- Decision Tree ใช้ Parameter ดังนี้ criterion = 'entropy', class_weight = {0: 1, 1: 5}, max_depth = 3

- K-nearest neighbors ใช้ Parameter ดังนี้ n_neighbors = 3, metric= 'manhattan', weights = 'distance'

- Neural Network ใช้ Parameter ดังนี้ hidden_layer_sizes = (24, 12, 6), activation = 'relu', solver = 'adam', learning_rate = 'adaptive', max_iter = 700, batch_size = 16

6. วัดประสิทธิภาพการจำแนกโรคผู้ป่วยอัลไซเมอร์ของแต่ละโมเดลบนชุดข้อมูลทดสอบ ได้ผลดังนี้

	accuracy	precision	recall	f1
Decision Tree	0.85	0.93	0.85	0.88
K-nearest neighbors	0.81	0.91	0.81	0.85
Neural Network	0.92	0.91	0.92	0.92

จากผลการทดลองทั้งหมด พบว่า Neural Network สามารถจำแนกโรคอัลไซเมอร์ได้ดีที่สุดจากทั้ง 3 โมเดลของการทดลองนี้ และปัญหาของชุดข้อมูลนี้คือเรื่อง Imbalance Class ซึ่งจะส่งผลกับประสิทธิภาพของโมเดล K-nearest neighbors ได้ โดยผู้ทำการทดลองได้มีการปรับให้ชุดข้อมูลฝึกฝนมีจำนวนคลาสที่เท่ากันด้วยวิธี Over Sampling ใช้ SMOTE และทำการคัดเลือก Feature เพื่อให้เหมาะสมกับการทำงานของแต่ละโมเดล