บทนำ

โรคไฮโปไทรอยด์ คือ ภาวะที่ต่อมไทรอยด์ (Thyroid Gland) ผลิตฮอร์โมนไทรอยด์ ออกมาไม่เพียงพอ โดยฮอร์โมนไทรอยด์นั้นจะควบคุมกระบวนการใช้พลังงานของเซลล์ต่าง ๆ ในร่างกายหรือที่เรียกว่ากระบวนการเมตาบอลิซึม โดยกระบวนการเมตาบอลิซึมส่งผลต่ออุณหภูมิ อัตราการเต้นหัวใจ และการเผาผลาญพลังงาน หากร่างกายผลิตฮอร์โมนไทรอยด์ไม่เพียงพอ จะส่งผลให้กระบวนการทำงานของร่างกายช้าลง ซึ่งจะส่งผลให้ อ่อยเพลีย น้ำหนักขึ้น หนาวง่าย ท้องผูก เป็นต้น โดยฮอร์โมนที่เกี่ยวกับโรคนี้ ที่งานของเรามีการนำข้อมูลมาวิเคราะห์ประกอบด้วย

1.triiodothyronine (T3) เป็นฮอร์โมนจากต่อมไทรอยด์ ช่วยไปกระตุ้นการเจริญเติบโตและพัฒนาการของร่างกาย

2.Thyroid stimulating hormone (TSH) เป็นฮอร์โมนที่สร้างจากต่อมใต้สมอง pituitary gland ทำหน้าที่กระตุ้นให้ต่อมไทรอยด์ Thyroid gland สร้างฮอร์โมน T3

3.thyroxine (T4) มีหน้าที่ควบคุมอัตราเร็วของเมแทบอลิซึมเพื่อผลิตพลังงานให้แก่ร่างกาย

4.Free thyroxine index(FTI) บ่งบอกถึงปริมาณฮอร์โมนไทรอยด์ในร่างกาย TSH ซึ่งเป็นฮอร์โมนที่หลั่งมาจากต่อมใต้สมอง เพื่อดูว่าในร่างกายมีปริมาณฮอร์โมนไทรอยด์เพียงพอหรือไม่ ถ้าฮอร์โมนไทรอยด์มีไม่พอจะทำระดับ TSH สูงขึ้น

5.Thyroxine – binding globulin (TBG) test เป็นการวัดระดับของ Thyroxine – binding globulin ในซีรั่ม ซึ่งเป็นโปรตีนที่สร้างจากตับ มีหน้าที่จับไทรอยด์ฮอร์โมนเพื่อป้องกันมิให้ไตขับไทรอยด์ฮอร์โมนออกแต่ถ้าไทรอยด์ฮอร์โมนในเลือดมีน้อยลง ฮอร์โมนที่จับก็จะหลุดออกมาในกระแสเลือด

6.Total T4 (TT4) = ไทรอกซินทั้งหมดในเลือด ทั้งส่วนที่จับกับโปรตีนและส่วนที่เป็นอิสระ

7.Thyroxine utilization rates (T4U) = อัตราการใช้ไทรอกซิน

ข้อมูลที่เราจะนำมาวิเคราะห์เกี่ยวกับการเรียนรู้ของเครื่องจักรคือ ข้อมูลของผู้ป่วยโรค Hypothyroid เป็นหนึ่งในชุดข้อมูลจาก Thyroid Disease Data Set ของ Ross Quinlan ที่อยู่ใน University of California at Irvine ในปีคริสต์ศักราช 1987 ซึ่งเราจะแดสงให้เห็นเกี่ยวกับความสัมพันธ์ระหว่างฮอร์โมนต่างๆที่ส่งผลต่อกัน และตัวโรคโดยใช้การนำเสนอข้อมูลทางสถิติ และความรู้ในเรื่อง machine learning

คำอธิบายเกี่ยวกับข้อมูล

ข้อมูลจะมีทั้งหมด 26 คอลัมภ์และ 3163 ประกอบด้วย

- result = ผลตรวจว่าเป็น hypoั้yroid หรือไม่
- age = อายุ
- sex = เพศ
- on_thyroxine = เกี่ยวกับ thyroxine
- query_on_thyroxine = แบบทดสอบเกี่ยวกับ thyroxine
- on_antithyroid_medication = การรับยาต้านไทรอยด์
- thyroid_surgery = การผ่าตัดไทรอยด์
- query_hypothyroid = แบบทดสอบ hypothyroid
- query_hyperthyroid = แบบทดสอบ hyperthyroid
- pregnant = การตั้งครรย์
- sick = มีการแสดงอาการ
- tumor = มีเนื้องอก
- lithium = ได้รับ lithium (ยา)
- goitre = อาการคอพอก
- TSH_measured = วัด THS แล้ว
- TSH = ปริมาณ TSH
- T3_measure = วัด T3 แล้ว
- T3 = ปริมาณ T3
- TT4_measured = วัด TT4 แล้ว
- TT4 = ปริมาณ TT4
- T4U_measured = วัด T4U แล้ว

- T4U = ปริมาณ T4U
- FTI_measured = วัด FTI แล้ว
- FTI = ปริมาณ FTI
- TBG_measured = วัด TBG แล้ว
- TBG = ปริมาณ TBG

หลักการและขั้นตอน การสกัด การเลือก และการเตรียมลักษณะ

เราจะเลือกบางคอลัมภ์มาใช้ในการพิจารณาข้อมูล ซึ่งจะไปกอบไปด้วย result , age , sex on_thyroxine , query_on_thyroxine , on_antithyroid_medication , thyroid_surgery , query_hypothyroid , query_hyperthyroid , pregnant , sick , tumor , lithium , goitre , TSH , T3 , TT4 , T4U , FTI เนื่องจากคอลัมภ์ TBG เป็นข้อมูลที่เสียหายจึงจะตัดออก และเราจะมีการจัดข้อมูลก่อนนำเข้าไปประมวลผล ดังนี้

1.นำค่า NULL Value ออก

2.ทำ one hot encoding โดย

2.1 ถ้าเป็นเพศหญิงค่าเป็น 1 และชายเป็น 0

2.2 ถ้าเป็นโรคให้ค่าเป็น 0 ถ้าไม่เป็นให้เป็น 1

2.3 ส่วนข้อมูลอื่นๆที่มีค่า True จะให้เป็น 1 ถ้า False เป็น 0

ก่อน

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3163 entries, 0 to 3162
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   result                    3163 non-null   object
 1   age                       3163 non-null   object
 2   sex                       3163 non-null   object
 3   on_thyroxine              3163 non-null   object
 4   query_on_thyroxine        3163 non-null   object
 5   on_antithyroid_medication 3163 non-null   object
 6   thyroid_surgery           3163 non-null   object
 7   query_hypothyroid         3163 non-null   object
 8   query_hyperthyroid        3163 non-null   object
 9   pregnant                  3163 non-null   object
 10  sick                      3163 non-null   object
 11  tumor                     3163 non-null   object
 12  lithium                   3163 non-null   object
 13  goitre                    3163 non-null   object
 14  TSH                       3163 non-null   object
 15  T3                        3163 non-null   object
 16  TT4                       3163 non-null   object
 17  T4U                       3163 non-null   object
 18  FTI                       3163 non-null   object
dtypes: object(19)
memory usage: 469.6+ KB
```

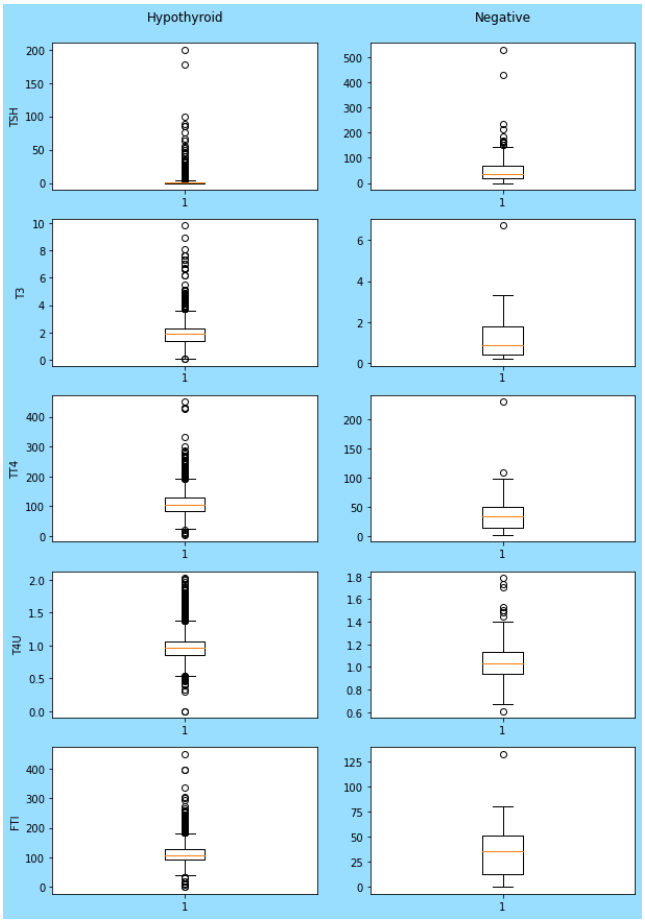
หลัง

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2000 entries, 0 to 3162
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   result                    2000 non-null   category
 1   age                       2000 non-null   int64
 2   sex                       2000 non-null   category
 3   on_thyroxine              2000 non-null   category
 4   query_on_thyroxine        2000 non-null   category
 5   on_antithyroid_medication 2000 non-null   category
 6   thyroid_surgery           2000 non-null   category
 7   query_hypothyroid         2000 non-null   category
 8   query_hyperthyroid        2000 non-null   category
 9   pregnant                  2000 non-null   category
 10  sick                      2000 non-null   category
 11  tumor                     2000 non-null   category
 12  lithium                   2000 non-null   category
 13  goitre                    2000 non-null   category
 14  TSH                       2000 non-null   float64
 15  T3                        2000 non-null   float64
 16  TT4                       2000 non-null   float64
 17  T4U                       2000 non-null   float64
 18  FTI                       2000 non-null   float64
dtypes: category(13), float64(5), int64(1)
memory usage: 136.3 KB
```

เทคนิคที่ใช้

Data information

ตรวจสอบการรกระจายของข้อมูลเบื้องต้นด้วย boxplot



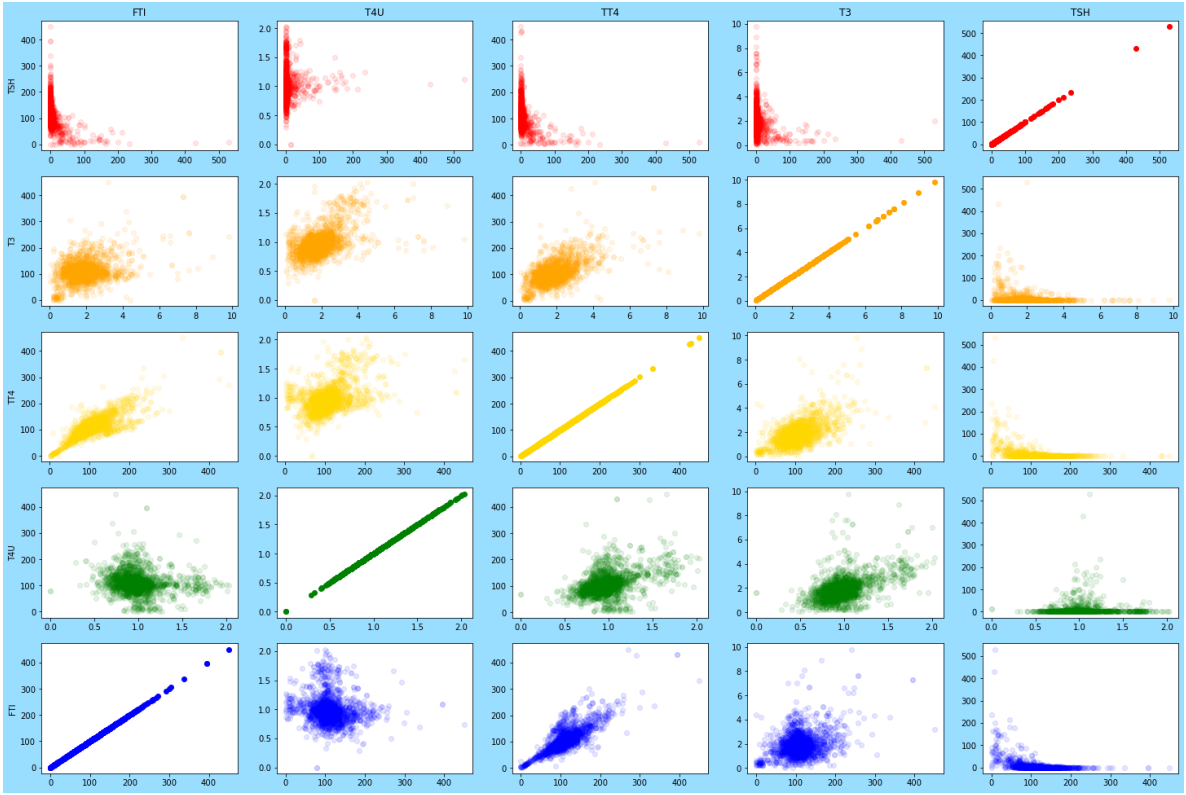
ตรวจสอบข้อมูลเบื้องต้นโดยจะเปรียบเทียบข้อมูลของฮอร์โมนต่างๆในร่างกายของผู่ที่เป็นโรคไฮโปไทรอยและไม่เป็นจากกราฟโดยได้ผลลัพธ์ว่า

- ในคนที่เป็นให้โรคจะมีค่าของฮอร์โมน TSH มากกว่า คนที่ไม่เป็น โดยคนที่เป็นจะเกาะกลุ่มอยู่ในช่วง 20-70 ในขณะคนที่ไม่เป้นจะอยู่ในช่วง 0-1
- ในคนที่เป็นให้โรคจะมีค่าของฮอร์โมน T3 ใกล้เคียงกับคนที่ไม่เป็นโรคโดยจะเฉลี่ยอยู่ที่ 1.1598 และ 1.9746 ตามลำดับ
- ในคนที่เป็นให้โรคจะมีค่าของฮอร์โมน TT4 น้อยกว่า คนที่ไม่เป็น โดยคนที่เป็นจะเกาะกลุ่มอยู่ในช่วง 15-50 ในขณะคนที่ไม่เป้นจะอยู่ในช่วง 86-128

- ในคนที่เป็นให้โรคจะมีค่าของฮอร์โมน T4U ใกล้เคียงกับคนที่ไม่เป็นโรคโดยจะเฉลี่ยอยู่ที่ 1.0543 และ 0.9837 ตามลำดับ8

- ในคนที่เป็นให้โรคจะมีค่าของฮอร์โมน FTI น้อยกว่า คนที่ไม่เป็น โดยคนที่เป็นจะเกาะกลุ่มอยู่ในช่วง 13-51 ในขณะคนที่ไม่เป้นจะอยู่ในช่วง 93-129

ตรวจสอบความสัมพันธ์ระหว่างตัวแปรด้วย scatter plot



จากข้อมูลสามารถกล่าวได้ว่า

- TSH มีความสัมพันธ์เชิงเส้นตรงกับฮอโมนอื่นๆ ค่อนข้างน้อย

- T3 มีความสัมพันธิเชิงเส้นตรงกับ T4U และ TT4

- TT4 มีความสัมพันธิเชิงเส้นตรงกับ FTI

- T4U มีความสัมพันธ์เชิงเส้นตรงกับ FTI น้อย

Machine Learning

ใช้ฟังก์ชั่น train_test_split แบ่งข้อมูลเป็นข้อมูลทดสอบ 20% และข้อมูล train 80% เมื่อนำไปรันกับ model จะได้ผลลัพธ์ออกมาอยู่ในรูปของ confusion matrix โดยแต่ละโมเดลจะได้ผลลัพธ์ ดังนี้

Logistic regression

	Predicted		
Actual		Positive	Negative
	Positive	17	7
	Negative	0	379

Accurrancy = 98.25%	Precition	Recall
Positive	100%	71%
Negative	98%	100%

Support Vector machine

	Predicted		
Actual		Positive	Negative
	Positive	5	19
	Negative	0	379

Accurrancy = 98.25%	Precition	Recall
Positive	100%	21%
Negative	95%	100%

Multilayer perceptron

	Predicted		
Actual		Positive	Negative
	Positive	20	4
	Negative	0	379

Accurrancy = 98.25%	Precition	Recall
Positive	100%	83%
Negative	99%	100%

Naïve Bay

	Predicted		
Actual		Positive	Negative
	Positive	24	0
	Negative	157	219

Accurrancy = 98.25%	Precition	Recall
Positive	13%	100%
Negative	100%	58%

การประเมินประสิทธิภาพของแบบจำลอง

โดยเราได้เลือกตัวแบบมาทั้งหมด 4 ตัวแบบคือ

1. Logistic regression(LR)

สามารถทายคนที่เป็นโรคได้ 98% และไม่เป็นโรคได้ถึง 100% แต่เมื่อพิจารณาค่า recall จะพบว่าจะมีค่าไม่สูงมากเนื่องจากมีทายผิดอยู่ 7 คน จากทั้งหมด 24 คนของคนเป็นโรคทั้งหมด

2. Support vector machine(SVM)

ในตัวแบบนี้มีโอกาสทายผิดค่อนข้างสูง เมื่อพิจารณาจาก confusion metric จะพบว่าค่าที่ทำนายเมื่อเทียบกับค่าที่เกิดขึ้นจริง จะมีโอกาสทำนายได้ถูกค่อนข้างน้อย

3. Multilayer perceptron(MLP)

ในตัวแบบนี้มีโอกาสทายผิดที่ต่ำมาก โดยเมื่อพิจารณาค่า recall และ precision แล้วมีค่าที่ค่อนข้างสูง เมื่อพิจารณาจาก confusion metric จะพบว่าทายถูกจาก 20 คนจาก 4 คน

4. Naive Bay(NB)

ตัวแบบนี้มีโอกาสทายคนที่ไม่ป่วยว่าป่วย(False Positive) สูงแต่ทายคนที่เป็นโรคถูกทั้งหมด เมื่อพิจารณาในมุมด้านการแพทย์แล้วตัวแบบนี้จะดีกว่าตัวแบบอื่นๆ เพราะการเป็นโรคจะส่งผลต่อชีวิตของคนไข้ ดังนั้นการที่ตัวแบบทำนายออกมาเป็น False Positive จะดีกว่ามีค่าออกมาเป็น False Negative