การวิเคราะห์ปัจจัยที่มีผลต่อโรคเบาหวาน และสร้างโมเดลการทำนายโรคเบาหวาน โดย Decision Tree

URL ของ app: https://sodavytong.shinyapps.io/App_DM/

By Sodavy Tong

ID: 6314400325

ที่มาและความสำคัญ

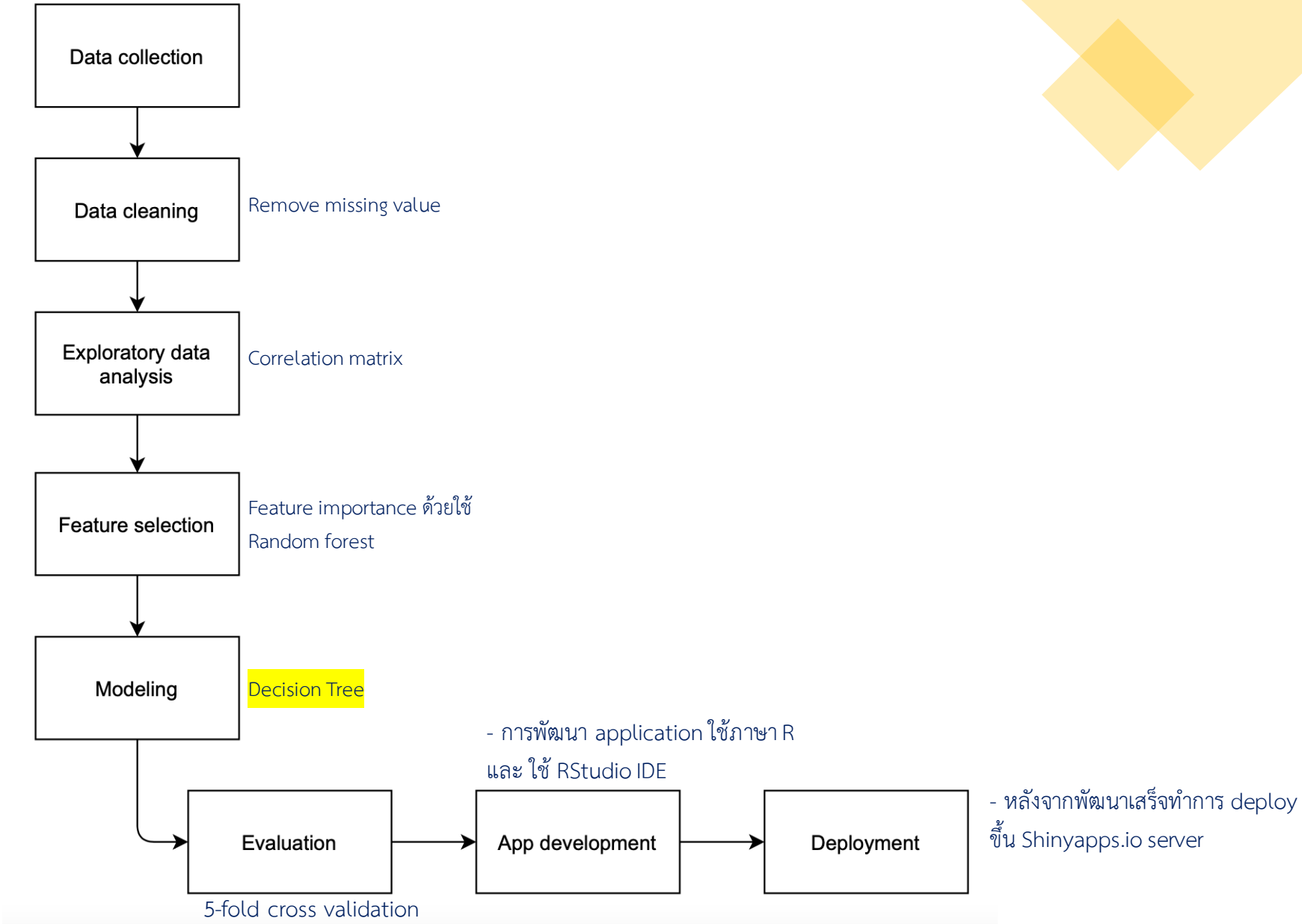
- โรคเบาหวานเป็นโรคเรื้อรังที่เป็นปัญหาคุกคามของคนทั่วโลก และเป็นปัญหาทางสาธารณสุขที่ส่งผลต่อผลลัพธ์ทางเศรษฐกิจโดยรวมต่อการพัฒนาประเทศ ซึ่งโรคเบาหวานเกิดจากการทำงานของฮอร์โมนอินซูลินของร่างกายผิดปกติ ส่งผลให้กระบวนการ นำน้ำตาลในกระแสเลือดเข้าสู่เซลล์ได้ไม่เต็มประสิทธิภาพทำให้น้ำตาลในกระแสเลือดสูงกว่าปกติ [1]
- จากสถิติของสหพันธ์ เบาหวานนานาชาติ (International Diabetes Federation; IDF) ผู้ป่วยโรคเบาหวานทั่วโลกมีประมาณ 463 ล้านราย ในปี 2019 โดยมีคนเสียชีวิตประมาณ 4.2 ล้านคน และในปี 2030 คาดว่าผู้ป่วยโรคเบาหวานมีถึง 578 ล้านคน [2]
- เนื่องจากจำนวนผู้ป่วยโรคเบาหวานเพิ่มขึ้นอย่างรวดเร็ว และส่งผลกระทบต่อกำดำรงชีวิตของประชาชน ดังนั้นการติดตามและตรวจสอบโรคเบาหวานเป็นเรื่องที่จำเป็นในการรักษาสภาวะสุขภาพให้ดี ในยุคปัจจุบัน เทคนิคการเรียนรู้ของเครื่องมีบทบาทสำคัญในการช่วยแก้ปัญหาทางการแพทย์ร่วมถึงการวิเคราะห์หาโรคเบาหวานโดยสามารถทำได้ง่ายและลดค่าใช้จ่ายในการตรวจสอบทางการแพทย์
- เป้าหมายหลัก คือ สร้าง application เพื่อช่วยวิเคราะห์หาปัจจัยต่างๆ ที่มีผลต่อความเสี่ยงต่อการเกิดโรคเบาหวานและสร้างโมเดลทำนายโรคเบาหวาน
- ใช้ชุดข้อมูลจากฐานข้อมูลของ Kaggle และใช้วิธีการเรียนรู้ของ Decision Tree เพื่อสร้างโมเดลทำนายโรคเบาหวาน
- **วัตถุประสงค์** เพื่อวิเคราะห์หาปัจจัยที่มีผลต่อการเกิดโรคเบาหวานและสร้างโมเดลทำนายโรคเบาหวาน

Why Decision Tree?

- สามารถจัดการได้ทั้งข้อมูลตัวเลข(numerical) และหมวดหมู่(categorical)
- มันง่ายในการแปลงผล และมีประโยชน์สำหรับการอธิบายโมเดล

	Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	Smoking	Alcohol	Sleep	SoundSleep	RegularMedicine	JunkFood	Stress	BPLevel	Pregancies	UriationFreq	Diabetic
1	50-59	Male	no	yes	one hr or more	39	no	no	8	6	no	occasionally	sometimes	high	0	not much	no
2	50-59	Male	no	yes	less than half an hr	28	no	no	8	6	yes	very often	sometimes	normal	0	not much	no
3	40-49	Male	no	no	one hr or more	24	no	no	6	6	no	occasionally	sometimes	normal	0	not much	no
4	50-59	Male	no	no	one hr or more	23	no	no	8	6	no	occasionally	sometimes	normal	0	not much	no

ขั้นตอนการทำงาน



1. Data collection

- ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลผู้ป่วยโรคเบาหวานจากฐานข้อมูลของ Kaggle ที่เก็บรวบรวมโดย Dr. Shruti Grag สาขาวิทยาการคอมพิวเตอร์และวิสวะกรรมในปี 2019 [3]
- Dataset name: “diabetes_dataset__2019”
- ชุดข้อมูลมีทั้งหมด 952 ตัวอย่าง ซึ่งแต่ละตัวอย่างมี 18 feature
  - Positive sample มี 266 ตัวอย่าง
  - Negative sample มี 685 ตัวอย่าง

ลำดับ	Attribute	Description	Type	Missing value
1.	Age	ช่วงอายุ	ordinal	0
2.	Gender	เพศ	binary	0
3.	Family_Diabetes	มีพันธุกรรมโรคเบาหวานหรือไม่	binary	0
4.	highBP	มีโรคความดันสูงหรือไม่	binary	0
5.	PhysicallyActive	ความถี่ในการออกกำลังกาย	ordinal	0
6.	BMI	ค่า BMI	numeric	0.42%
7.	Smoking	สูบบุหรี่หรือไม่	binary	0
8.	Alcohol	ดื่ม alcohol หรือไม่	binary	0
9.	Sleep	จำนวนชั่วโมงนอนต่อวัน	interval	0
10.	SoundSleep		interval	0
11.	RegularMedicine	รับประทานยาอื่น ๆ เป็นประจำหรือไม่	binary	0
12.	JunkFood	ความถี่ในการรับประทานอาหารขยะ	binary	0
13.	Stress	ความถี่ในการเครียด	ordinal	0
14.	BPLevel	ระดับความดัน	ordinal	0
15.	Pregancies	ระยะเวลาในการตั้งครรภ์	numeric	4.41%
16.	Pdiabetes		binary	0.11%
17.	UriationFreq	ความถี่ในการปัสสาวะ	ordinal	0
18.	Diabetic	เป็นโรคเบาหวานหรือไม่(ผลเฉลย)	binary	0.11%

2. Data cleaning

- ชุดข้อมูลเดิมมีจำนวน 952 ตัวอย่าง เนื่องจากแถวที่มีค่าว่างมีจำนวนน้อย ดังนั้นการทำความสะอาดข้อมูลด้วยการตัดแถวที่มีค่าว่าง (missing value) ออกไป
- ลบบางแอททริบิวท์ที่ไม่มีความสำคัญในการวิเคราะห์ข้อมูล (เช่น Pdiabetes)
- ปรับ column ที่เป็น text ให้เป็น factor ทั้งแบบมีลำดับ และไม่มีลำดับ (เช่น Diabetic, Age, Gender, Family_Diabetes, highBP, PhysicallyActive, Smoking, Alcohol, RegularMedicine, JunkFood, Stress, BPLevel, UriationFreq)
- หลังจากตัดแถวว่างออกไปแล้ว ชุดข้อมูลมีจำนวน 906 ตัวอย่าง (Positive sample มี 263 ตัวอย่าง & Negative sample มี 642 ตัวอย่าง)
- ตัวอย่าง code ทำความสะอาดข้อมูล

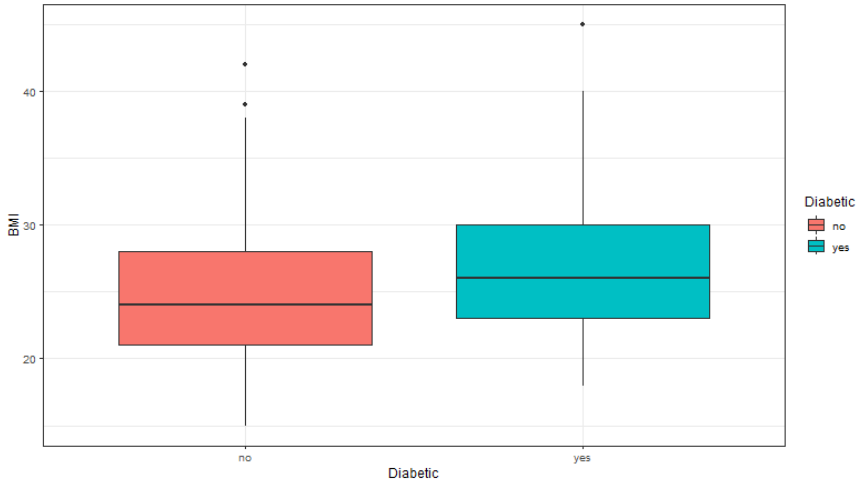
```
df.dm2019 <- dplyr::select(df.dm2019, -Pdiabetes)
df.dm2019 <- df.dm2019 %>% filter(!is.na(Diabetic))
df.dm2019 <- df.dm2019 %>% drop_na()

df.dm2019$Diabetic <- factor(df.dm2019$Diabetic)
df.dm2019$Age <- factor(df.dm2019$Age, ordered = T, levels = c("less than 40", "40-49", "50-59", "60 or older"))
df.dm2019$Gender <- factor(df.dm2019$Gender)
df.dm2019$Family_Diabetes <- factor(df.dm2019$Family_Diabetes)
df.dm2019$highBP <- factor(df.dm2019$highBP)
df.dm2019$PhysicallyActive <- factor(df.dm2019$PhysicallyActive, ordered = T, levels = c("none", "less than half an hr",
                                                                                         "more than half an hr", "one hr or more"))

df.dm2019$Smoking <- factor(df.dm2019$Smoking)
df.dm2019$Alcohol <- factor(df.dm2019$Alcohol)
df.dm2019$RegularMedicine <- factor(df.dm2019$RegularMedicine)
df.dm2019$JunkFood <- factor(df.dm2019$JunkFood, ordered = T, levels = c("occasionally", "often", "very often", "always"))
df.dm2019$Stress <- factor(df.dm2019$Stress, ordered = T, levels = c("not at all", "sometimes", "very often", "always"))
df.dm2019$BPLevel <- factor(df.dm2019$BPLevel, ordered = T, levels = c("low", "normal", "high"))
df.dm2019$UriationFreq <- factor(df.dm2019$UriationFreq, ordered = T, levels = c("not much", "quite often"))
```

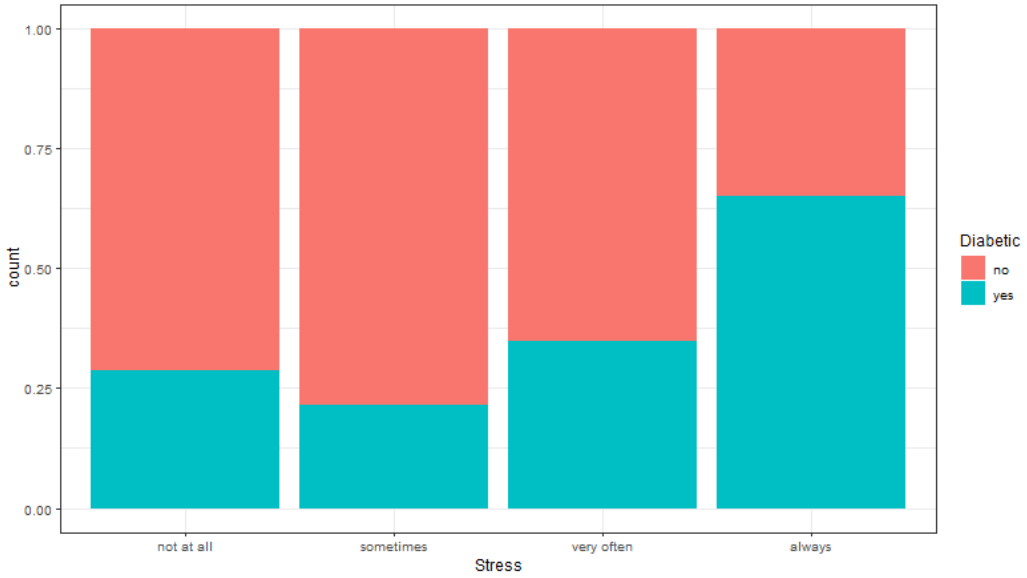
3. Exploratory data analysis

1. correlation matrix

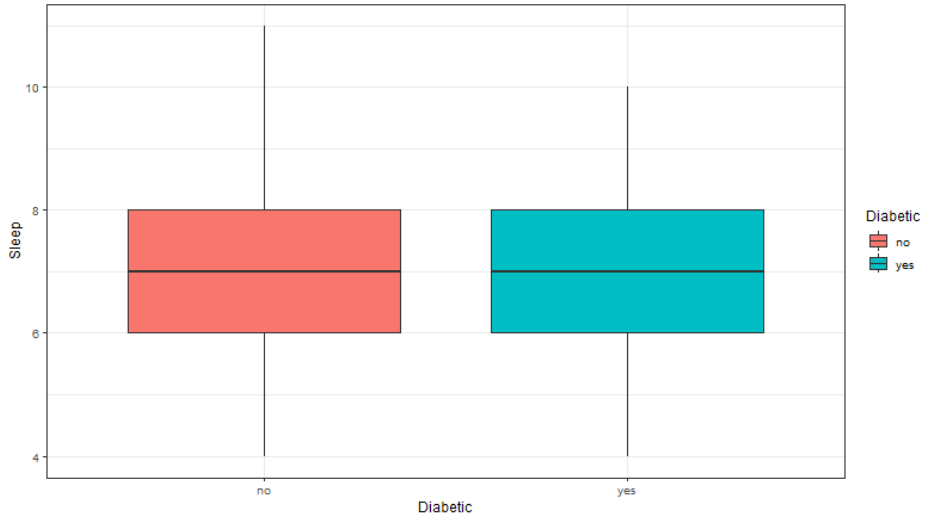
- จากตาราง correlation เห็นว่า ตัวแปรที่มีความสัมพันธ์กับความเป็นโรคเบาหวานสูงมี
  - RegularMedicine
  - Age
  - highBP
  - BPlevel
- ขณะที่ตัวแปรที่มีความสัมพันธ์ต่ำมี
  - Sleep
  - Smoking
  - Gender
  - PhysicallyActive

3. **Exploratory data analysis** (cont.)





- จาก plot box เห็นได้ว่า BMI ของผู้ป่วยโรคเบาหวานมีค่าโดยรวมสูงกว่า BMI ของบุคคลปกติ
- ขณะที่จำนวนชั่วโมงนอน (Sleep) ของผู้ป่วยโรคเบาหวานและบุคคลปกติไม่ได้แตกต่างกัน
- ดังนั้น BMI น่าจะมีผลต่อการเป็นโรคเบาหวานมากกว่า Sleep



- จาก bar plot เห็นว่า
- คนที่มีความถี่ในการเครียด (Stress) สม่ำเสมอ จะมีโอกาสเป็นเบาหวานสูงกว่าคนที่มีความเครียดบ่อยมาก คนที่มีความเครียดบางครั้ง และคนที่ไม่มีความเครียดเลย

4. Feature selection

- การหาความสำคัญของ feature (Feature importance) ด้วยใช้ Random forest
- จากตารางความสำคัญของ feature เห็นว่า feature ที่มีความสำคัญต่อการทำนายสูงมี RegularMedicine, Age, และ BMI ที่สามารถลดค่า Gini ได้สูงกว่า feature อื่น

สรุป จากการหาค่า correlation และ feature importance ได้ทำการคัดเลือก feature 9 ตัวคือ

- ✓ RegularMedicine
- ✓ Age
- ✓ BMI
- ✓ highBP
- ✓ PhysicallalyActive
- ✓ Stress
- ✓ Sleep
- ✓ Family_diabetes และ
- ✓ JunkFood ในการสร้างโมเดล

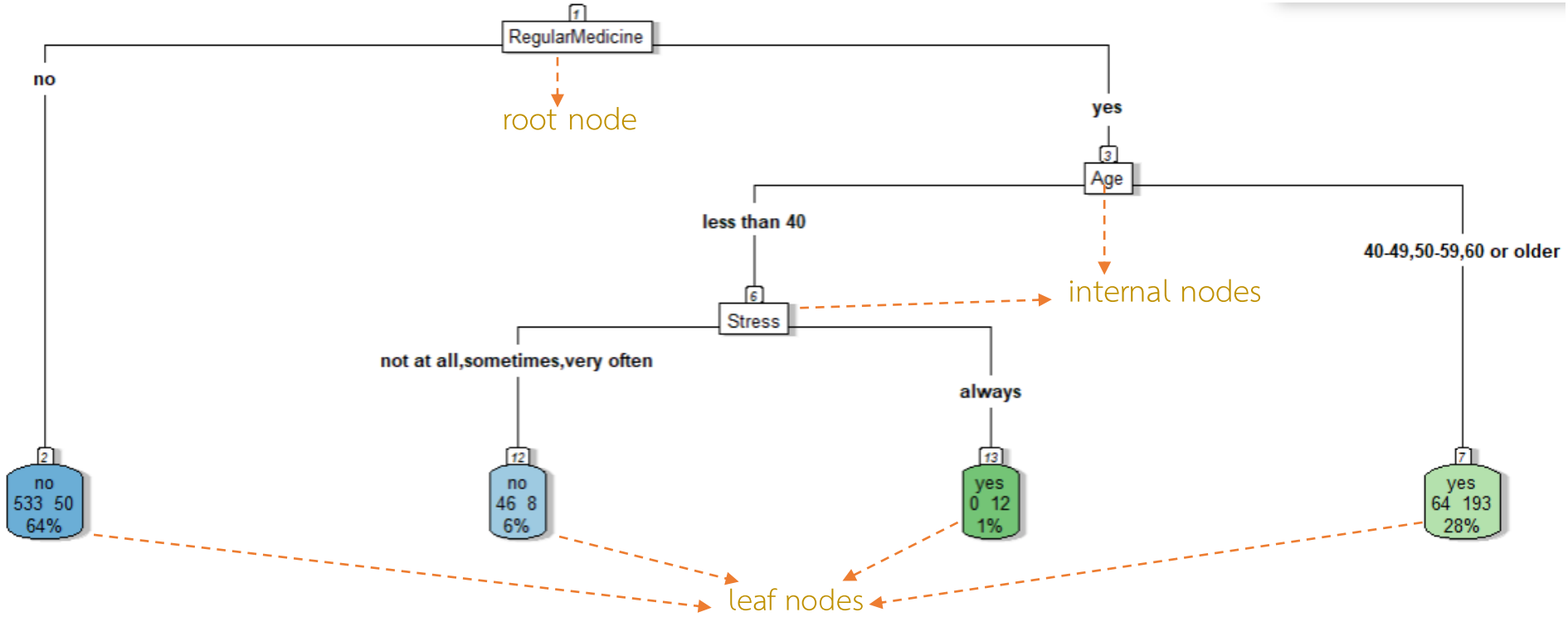
	no	yes	MeanDecreaseAccuracy	MeanDecreaseGini
Age	27.364563	28.25804	37.202476	62.426662
Gender	10.636842	11.82035	14.740361	5.163570
Family_Diabetes	20.482125	18.64170	22.220992	14.669130
highBP	12.127155	10.55006	13.340566	14.770999
PhysicallyActive	19.501999	18.90040	24.815977	18.524190
BMI	24.410779	21.04184	28.465522	30.314527
Smoking	7.395625	6.90139	9.377938	2.330934
Alcohol	10.816592	10.27918	12.463086	5.172307
Sleep	16.999713	19.11545	22.006247	16.756270
SoundSleep	19.689958	22.89769	27.429117	21.777571
RegularMedicine	25.939043	33.17577	39.209533	80.358189
JunkFood	12.341972	13.50885	16.285088	8.614740
Stress	16.485624	19.74123	21.568409	19.471028
BPLevel	13.136379	14.56227	16.272043	21.577091
Pregancies	13.961301	16.42738	18.300823	11.002138
UriationFreq	11.662663	11.85672	14.045719	6.266490

➢ Code:

```
tree_RF2 <- randomForest(Diabetic~., data=df, na.action = na.omit, importance=TRUE, ntree=170)

#Show important Feature
 output$importantFeature <- renderPrint(importance(tree_RF2) )
```

5. Modeling

- Algorithm: Decision Tree C4.5
- Package: RWeka, rpart.plot
- Parameter: Pruning = True
  (กระบวนการตัดเล็มต้นไม้)
- Code:




```
df <- df.dm2019
tree_c45_prune <-
    J48(Diabetic ~ RegularMedicine+Age+BMI+Family_Diabetes+Sleep+highBP+PhysicallyActive+JunkFood+Stress,
        data = df, control = Weka_control(R = T))
tree_rpart <- rpart(Diabetic~RegularMedicine+Age+BMI+Family_Diabetes+Sleep+highBP+PhysicallyActive+JunkFood+
                    Stress, data= df, cp = .02, xval = 10, parms = list(split = "information"))
```

- จากภาพต้นไม้ตัดสินใจเห็นได้ว่า
- ถ้าผู้ป่วยมีการใช้ยาเป็นประจำและมีอายุมาก(>40) หรืออายุน้อย(less than 40) แต่มีความเครียดอยู่ตลอดเวลา(always) มีโอกาสเป็นโรคเบาหวานสูง ขณะที่บุคคลที่ไม่มีการใช้ยาใดๆ เป็นประจำและอายุยังน้อย(<40) และไม่มีความเครียด มีโอกาสเป็นโรคเบาหวานต่ำ

6. Evaluation

- การประเมินผลแบบ: 5-fold cross validation
- จากภาพเห็นได้ว่า โมเดลมีค่า Accuracy ที่ 91.72 % โมเดลทำนายข้อมูลที่เป็นคลาสลบ (ไม่เป็นโรคเบาหวาน) ถูกต้องมากกว่าข้อมูลที่เป็นคลาสบวก (เป็นโรคเบาหวาน) ที่สามารถเห็นได้จากค่า Recall หรือ Confusion Matrix
- และโดยรวมแล้วมีค่า
  - TP 91.7%
  - Recall 91.7%
  - F-Measure 91.6%
  - ROC Area 92.9%
- Code:

```
output$tree_summary <- renderPrint(
    evaluate_Weka_classifier(tree_c45_prune, numFolds = 5, complexity = T,
                             class = T, seed = 1234)
)
```

```
=== 5 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         831               91.7219 %
Incorrectly Classified Instances        75                8.2781 %
Kappa statistic                          0.7947
K&B Relative Info Score                 71.8131 %
K&B Information Score                  565.463  bits      0.6241 bits/instance
Class complexity | order 0             787.4093 bits      0.8691 bits/instance
Class complexity | scheme            21710.5834 bits     23.9631 bits/instance
Complexity improvement     (Sf)     -20923.1741 bits    -23.094  bits/instance
Mean absolute error                      0.1173
Root mean squared error                  0.2701
Relative absolute error                 28.4507 %
Root relative squared error             59.5134 %
Total Number of Instances              906

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.956    0.179    0.929      0.956    0.943      0.796    0.929     0.953     no
               0.821    0.044    0.885      0.821    0.852      0.796    0.929     0.872     yes
Weighted Avg.  0.917    0.139    0.916      0.917    0.916      0.796    0.929     0.930

=== Confusion Matrix ===

   a   b   <-- classified as
 615  28 |   a = no
  47 216 |   b = yes
```

Reference

- [1] ผศ.พญ. พิมพ์ใจ อันทานนท์ 2017, โรคเบาหวาน, สมาคมโรคเบาหวานแห่งประเทศไทย, viewed 22 October 2021, <https://www.dmthai.org/index.php/knowledge/for-normalperson/health-information-and-articles/health-information-and-articles-old-3/846-2019-04-20-01-49-18?fbclid=IwAR0gcP_7v9Q1pgbpJqo5ZvEs4SQjRdXDbvBOHr5MxHHn5r7NbynwYpD2QbA>.
- [2] Worldwide toll of diabetes (2019), The International Diabetes Federation (IDF), viewed 22 October 2021, <https://www.diabetesatlas.org/en/sections/worldwide-toll-of-diabetes.html>.
- [3] Neha Prerna Tigga & Dr. Shruti Garg (2019), Diabetes Dataset 2019, Kaggle, viewed 22 October 2021, <https://www.kaggle.com/tigganeha4/diabetes-dataset-2019>.