

**รายงานสรุปการทำ Model**

**เสนอ**

**ดร. กัญณัฏฐ์ หอมทรัพย์**

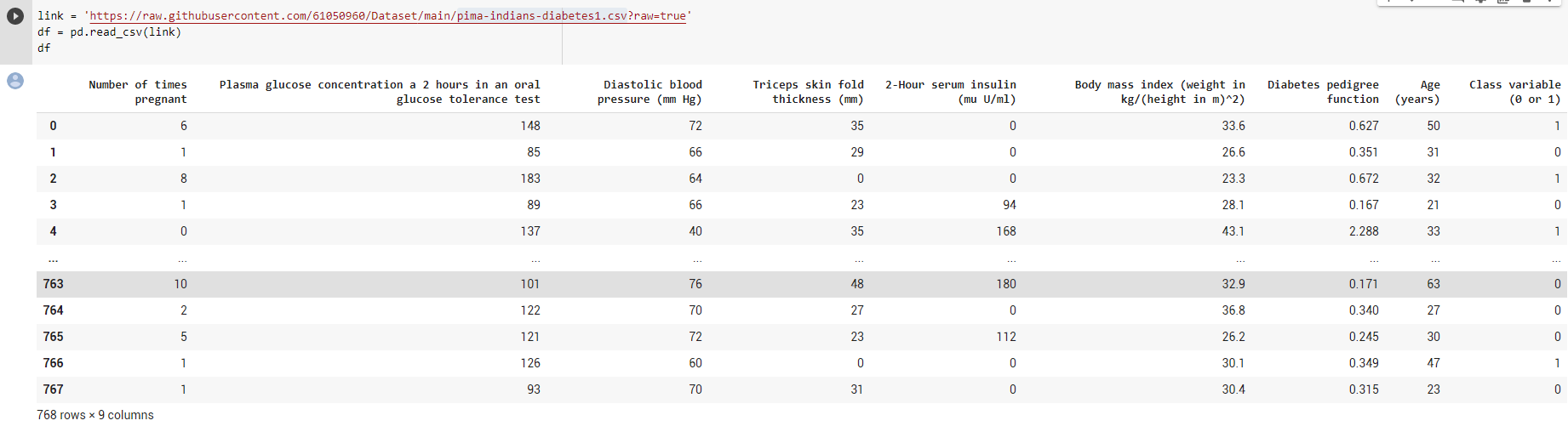
**จัดทำโดย**

**นายฟาริซ หะยีมะสาและ รหัสนักศึกษา 61050258**

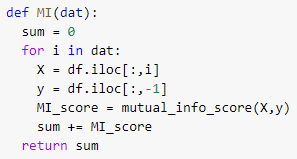
**ชั้นปีที่ 3 ภาคเรียนที่ 1 ปีการศึกษา 2563**

**การทำ model Feature Selection**

เริ่มจากการ insert data set pima-indians-diabetes1.csv

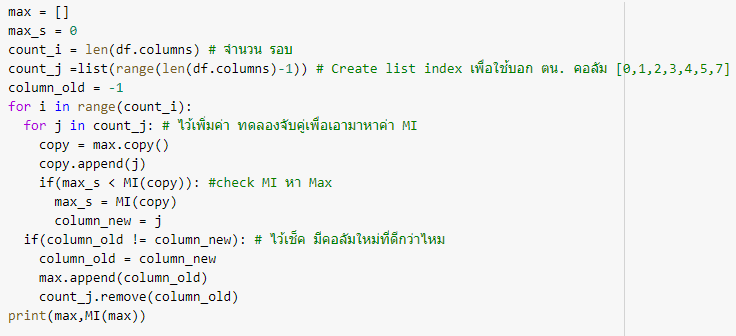


หลังจากนั้นทำการสร้าง Method MI สำหรับทำการวัดคะแนนของคอลัมที่เลือกใช้งาน



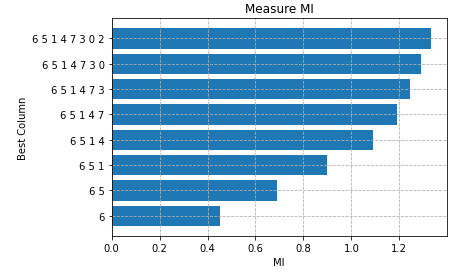
**Sequential Forward Selection (SFS)**

โค้ดส่วนของการทำ Sequential Forward Selection (SFS)



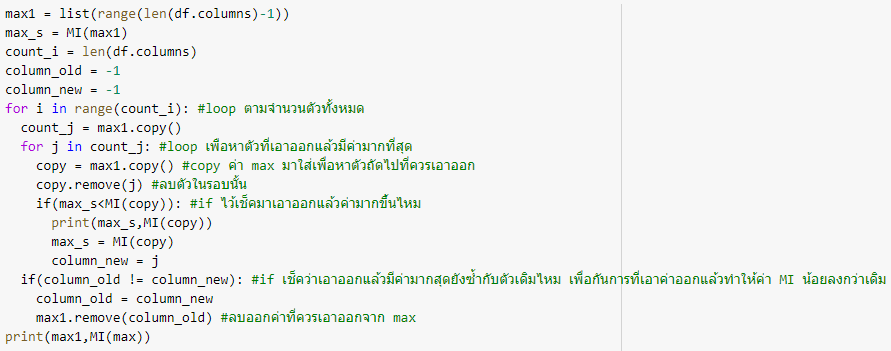
ได้คอลัมและคะแนนของคอลัมทั้งหมดที่ทำการเลือกมาได้คือ

  
กราฟแสดงคะแนนในการเลือกคอลัม



**Sequential Backward Selection (SBS)**

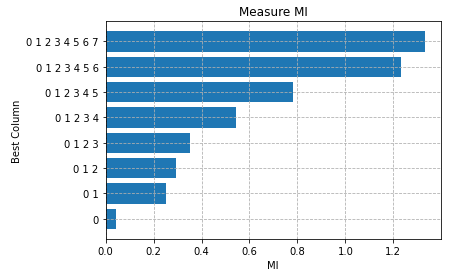
โค้ดส่วนของการทำ Sequential Backward Selection (SFS)



ได้คอลัมและคะแนนของคอลัมทั้งหมดที่ทำการเลือกมาได้คือ

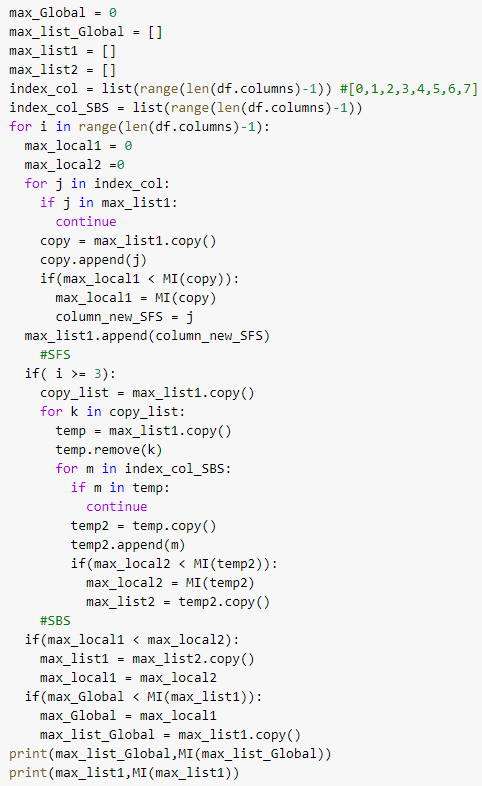


กราฟแสดงคะแนนในการเลือกคอลัม



**Sequential Forward Floating Selection (SFFS)**

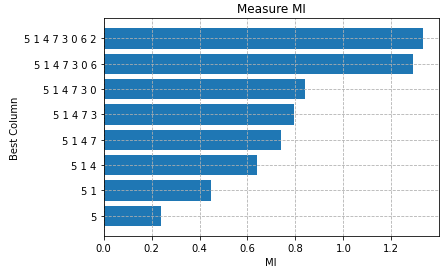
โค้ดส่วนของการทำ Sequential Forward Floating Selection (SFFS)



ได้คอลัมและคะแนนของคอลัมทั้งหมดที่ทำการเลือกมาได้คือ

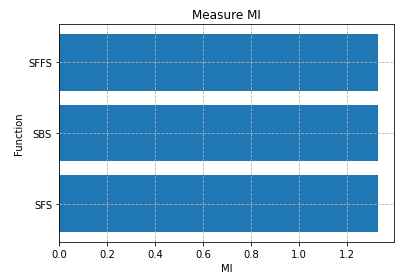


กราฟแสดงคะแนนในการเลือกคอลัม



**การวัดประสิทธิภาพของทั้งสามตัว**

กราฟแสดงคะแนนการเลือกของแต่ละตัว



สรุปการวัดประสิทธิภาพ

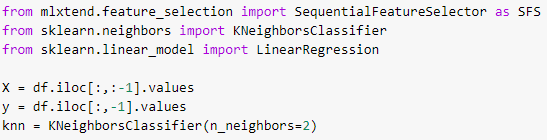
SFS มีคะแนน MI = 1.330679076053227  
SBS มีคะแนน MI = 1.330679076053227  
SFFS มีคะแนน MI = 1.330679076053227  
ทั้งสามวิธีได้คอลัมออกมาเหมือนกันจึงทำให้ คะแนนเท่ากัน  
แต่วิธีที่ดีที่สุดในสามวิธีนี้คือ SFFS เพราะมีการรีเช็คกันระหว่างการทำงานโดยใช้วิธี SFS SBS

ปัญหา SFS SBS เพราะวิธีที่ผมทำจะเป็นแบบ greedy เมื่อเจอการใส่คอลัมแล้วทำให้ค่า MI ลดลงจะหยุด จึงติดปัญหา local maximum อาจจะมีคำตอบที่ดีกว่าแต่ไปไม่ถึงแต่ SFFS จะทำงานไปจนสุดทางและจะมีการจำคำตอบที่ดีที่สุดที่เกิดขึ้นในระหว่างทาง เพื่อแก้ไขปัญหา local maximum

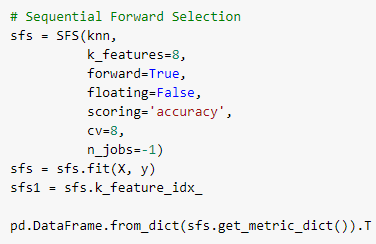
**SFS , SBS , SFFS โดยใช้ Library**

โดยก่อนจะใช้ Library จะต้องทำการ import ตัว SFS และตัว Model ที่จะใช้วัดคะแนนมาก่อน

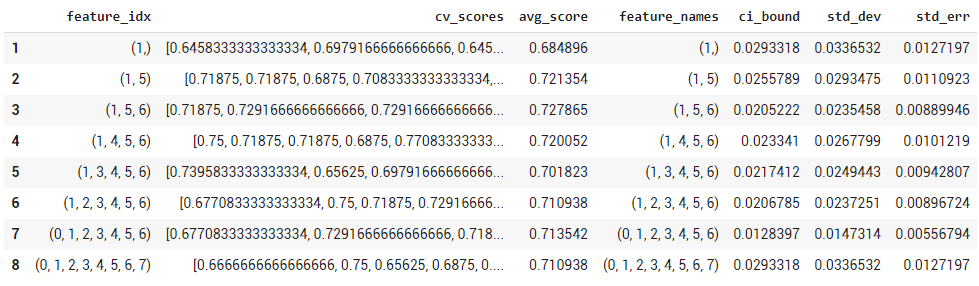
โค้ดส่อนของการ Import



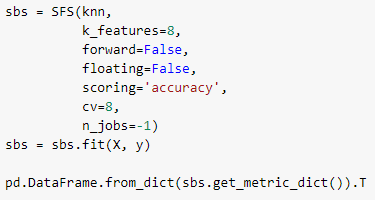
โค้ดส่วนของ SFS



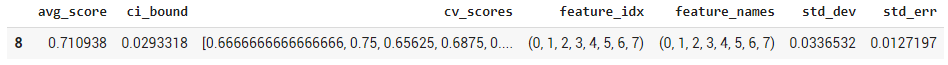
SFS ได้คอลัมและคะแนนดังนี้



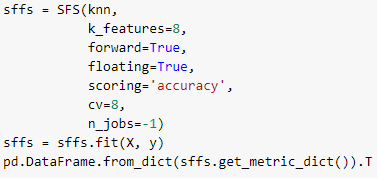
โค้ดส่วนของ SBS



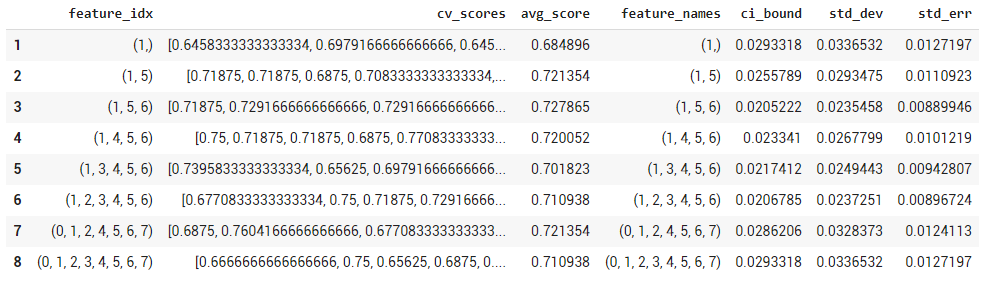
SBS ได้คอลัมและคะแนนดังนี้



โค้ดส่วนของ SFFS



SFFS ได้คอลัมและคะแนนดังนี้

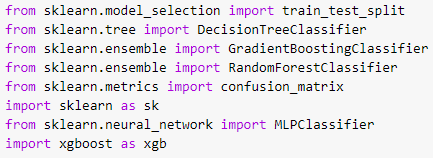


**การนำคอลัมที่เลือกมาสร้างโครงสร้างต้นไม้(Decision Tree)**

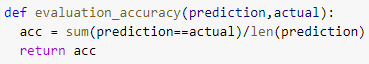
เลือกใช้คอลัม จาก SFFS ที่ได้ Average Score มากที่สุด คือ 0.722656 ได้คอลัม คือ 0, 1, 2, 5, 6, 7

ก่อนเข้า Model Tree ต้องทำการ Import ตัว Model เข้ามาก่อน

โค้ดส่วน Import



โค้ดส่วนของ Method Evaluation Accuracy ไว้หาคะแนน Accuracy

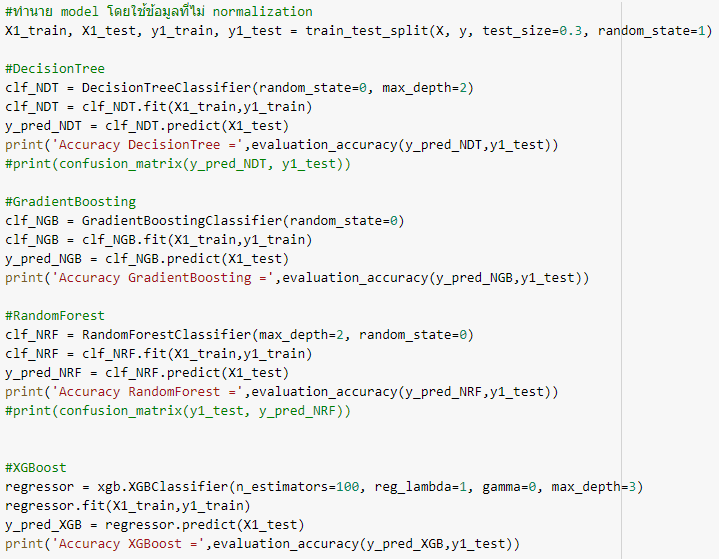


**Non-Normalization (ยังไม่ทำการ Normalization)**

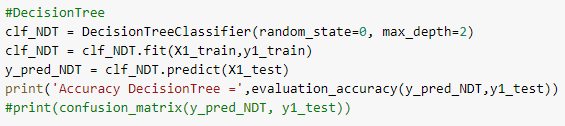
ต่อมาเป็นการแบ่งส่วนของข้อมูลที่เป็นคำตอบและคอลัมที่เลือกใช้



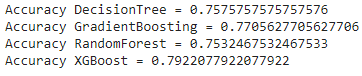
โค้ดของการแบ่งข้อมูล Train และ Test



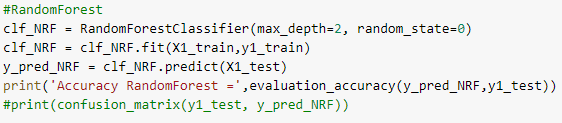
**Decision Tree**



ค่า Accuracy ที่ได้ของ Decision Tree



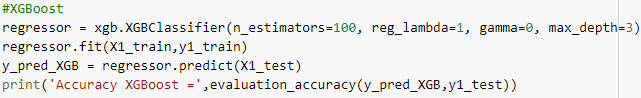
**Random Forest**

****

ค่า Accuracy ที่ได้ของ Random Forest



**XGBoost**

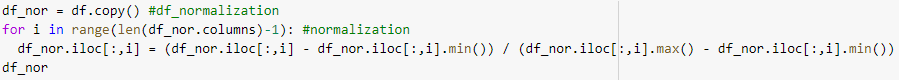
****

ค่า Accuracy ที่ได้ของ XGBoost



**Normalization (ทำการ Normalization)**

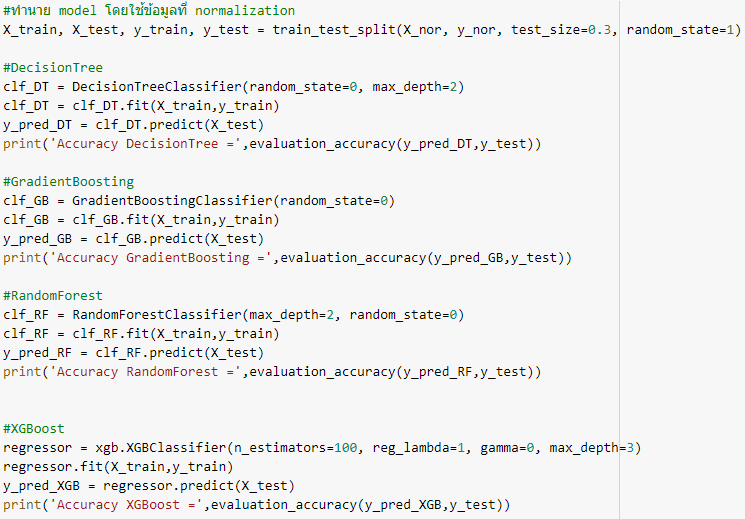
โค้ดของการ Normalization



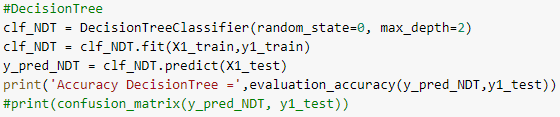
แบ่งข้อมูลคำตอบและคอมลัมที่ใช้



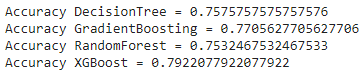
โค้ดส่วนของการแบ่งส่วน Train และ Test



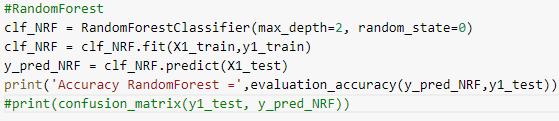
**Decision Tree**



ค่าของ Accuracy ที่ได้ของ Decision Tree



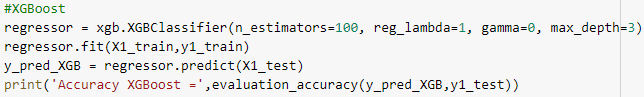
**Random Forest**

****

ค่าของ Accuracy ที่ได้ของ Random Forest



**XGBoost**

****

ค่า Accuracy ที่ได้ของ XGBoost

****

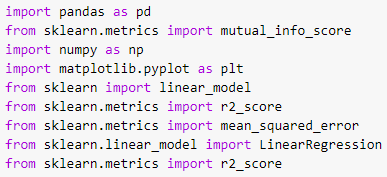
**สรุปการนำข้อมูลเข้า Model**

ค่า Accuracy ของก่อนทำการ Normalize และ หลังทำการ Normalize มีค่าเท่ากัน อาจเกิดจากการที่ข้อมูลมีการจัดการมาแล้วจึงทำให้ค่าของก่อนและหลังทำการ Normalize จึงมีค่าเท่ากัน

**Multiple Regression**

เริ่มจากการ Import Library และ Import Data Set

โค้ดส่วนการ Import Library



โค้ดส่วนของการ Import Data Set

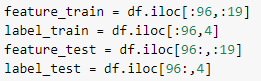


ตัวอย่าง Dataset

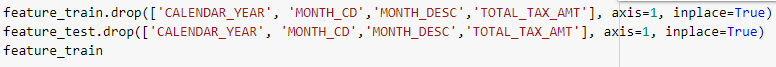


**Non-Normalization**

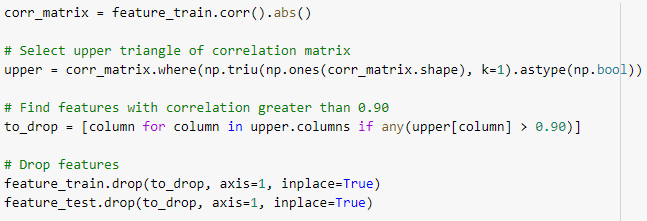
โค้ดส่วนของการแบ่งข้อมูล Train 2554-2561 Test 2562

****

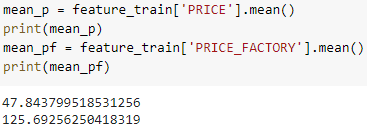
การ Drop คอลัมที่ไม่ได้เกี่ยวข้องกับการคำนวน เช่น ปี เดือน คำตอบ



โค้ดส่วนของการ Drop คอลัมที่มี Correlation มากกว่า 0.90

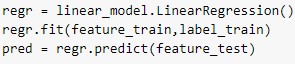


โค้ดส่วนของการจัดการข้อมูลที่มีค่าเป็น 0 โดยการใส่ค่า Mean เข้าไปแทน

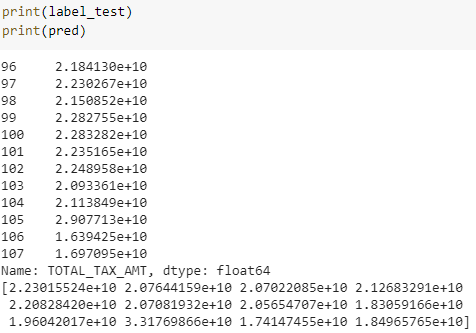




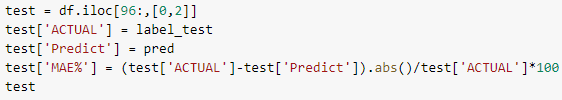
โค้ดการนำข้อมูลเข้า Model เพื่อทำนายค่า predict ของปี 2562

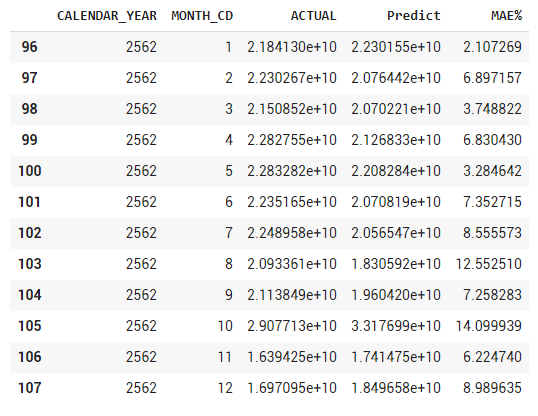


ค่าที่ทำนายออกมา



หลังจากนั้นนำค่าที่ทำนายออกมาไปเทียบกับคำตอบ แล้วคิดค่า MEA% ออกมา





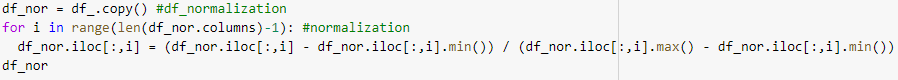
**Normalization**

แบ่งข้อมูลที่จะทำการ Normalization และทำการ Normalization

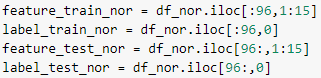
โค้ดการแบ่งข้อมูลที่จะทำ Normalization



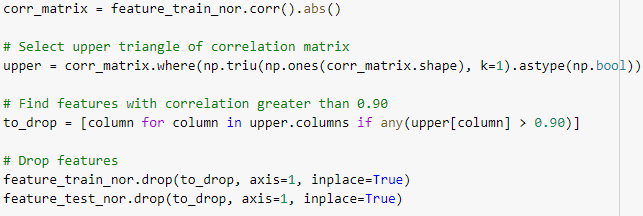
โค้ดการ Normalization



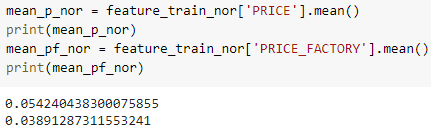
โค้ดการแบ่งข้อมูล Train 2554-2561 Test 2562

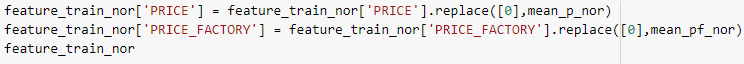


Drop คอลัมที่มี Correlation มากกว่า 0.90



จัดการข้อมูลที่เป็นค่า 0 โดยใส่ค่า Mean

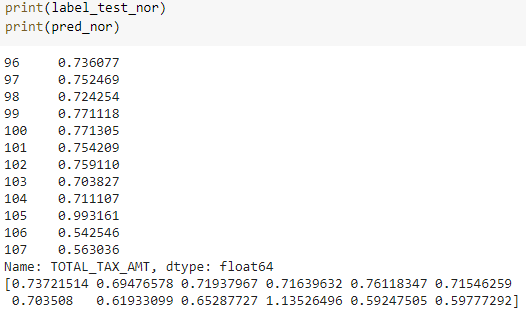




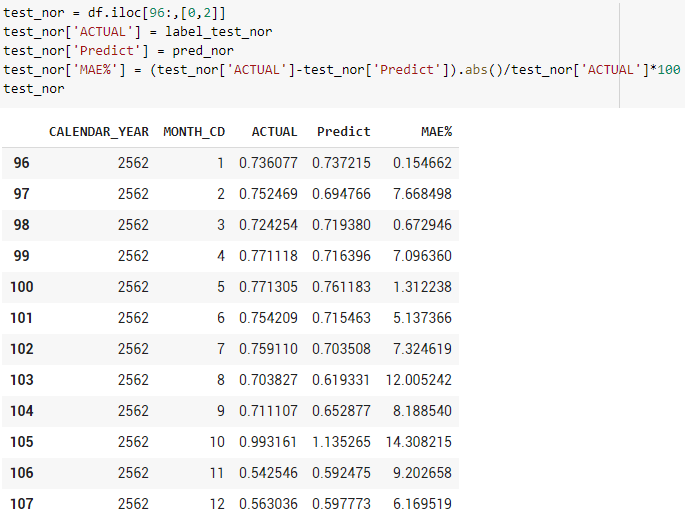
โค้ดการนำข้อมูลเข้า Model เพื่อทำนายผลปี 2562



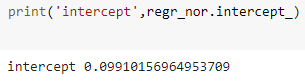
ค่าที่ได้



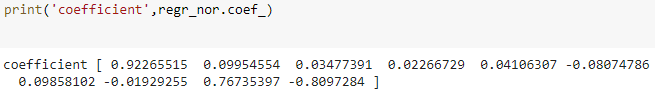
นำค่าที่ได้มาเทียบกับคำตอบเพื่อหาค่า MAE%



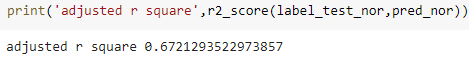
ค่า Intercept



ค่า Coefficient

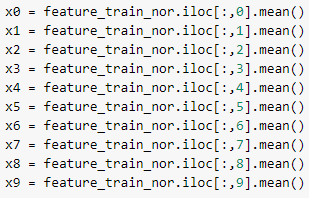


ค่า Adjusted r square

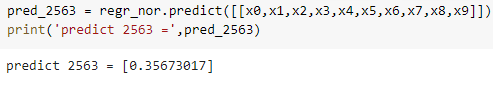


**การทำนายผลปี 2563**

หาค่าเฉลี่ยของแต่ละคอลัม

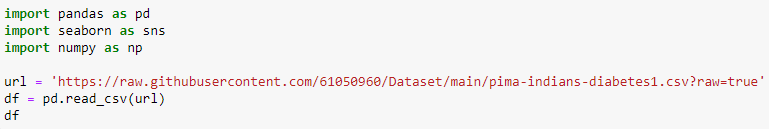


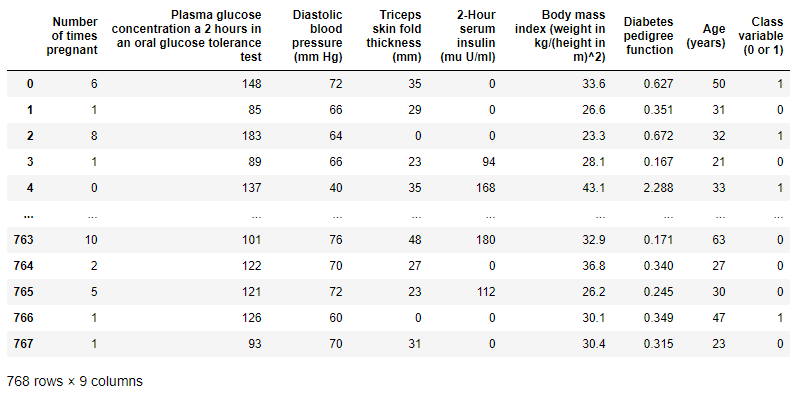
ทำนายผล



**Outlier**

ใช้ Data set pima-indians-diabetes1.csv





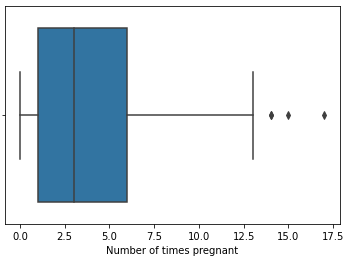
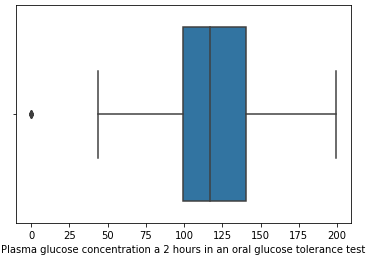
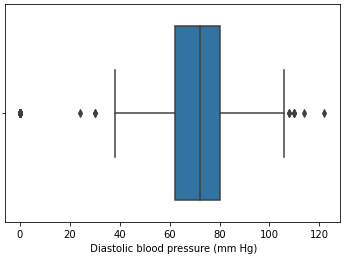
ส่วนของ Box plot แสดงค่าของข้อมูลที่ไม่ได้อยู่ในช่วง หรือข้อมูลที่เป็ยข้อมูลที่ผิดพลาด

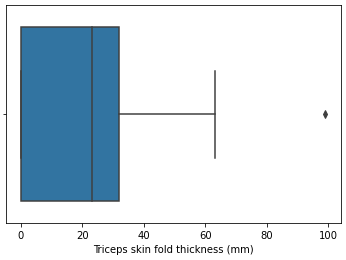
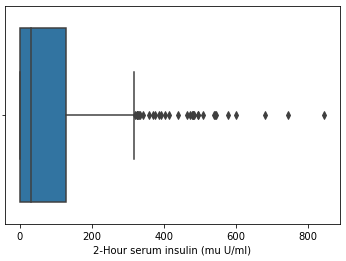
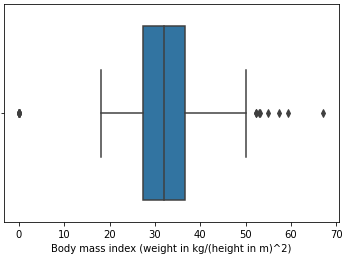
โค้ดของ Box plot ของแต่ละคอลัม

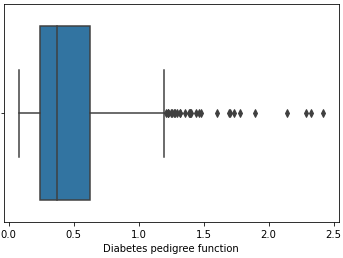
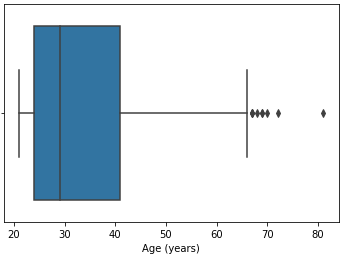
   

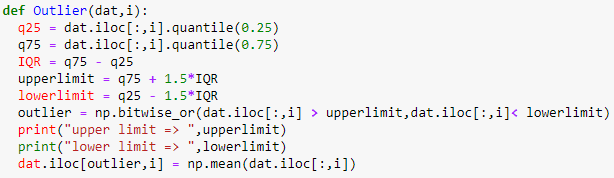
กราฟ Box plot

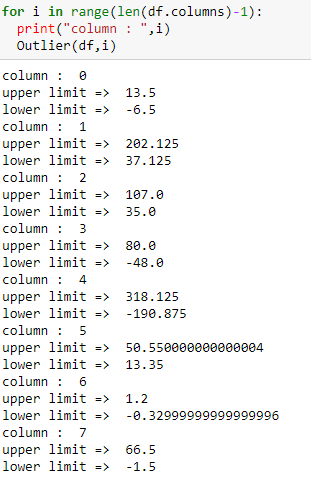
  

โค้ดของ Method การบอก Upper limit และ Lower limit และเลือกเฉพาะค่าที่อยู่ในช่วง

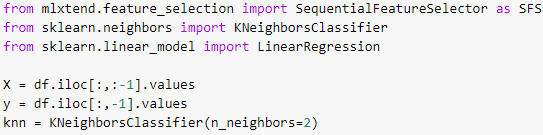


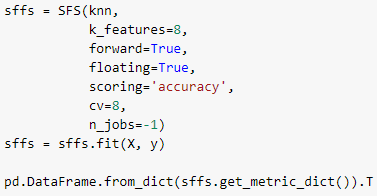
ค่า Upper limit และ Lower limit ของแต่ละคอลัม



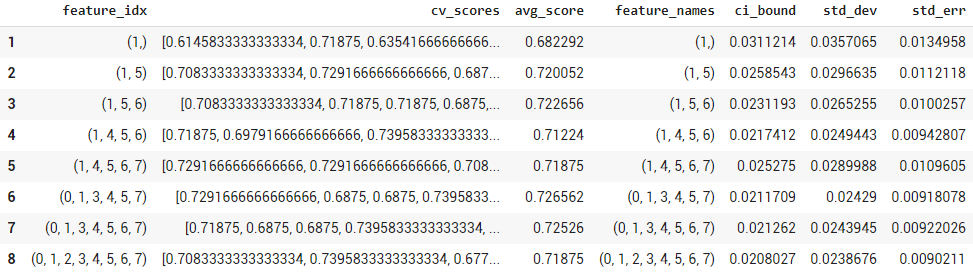
หลังจากเลือกเฉพาะค่าที่อยู่ในช่วงก็จะนำมาทำ Sequential Forward Floating Selection ต่อเพื่อหาคอลัมที่จะนำไปใช้งาน

โค้ดของ Sequential Forward Floating Selection



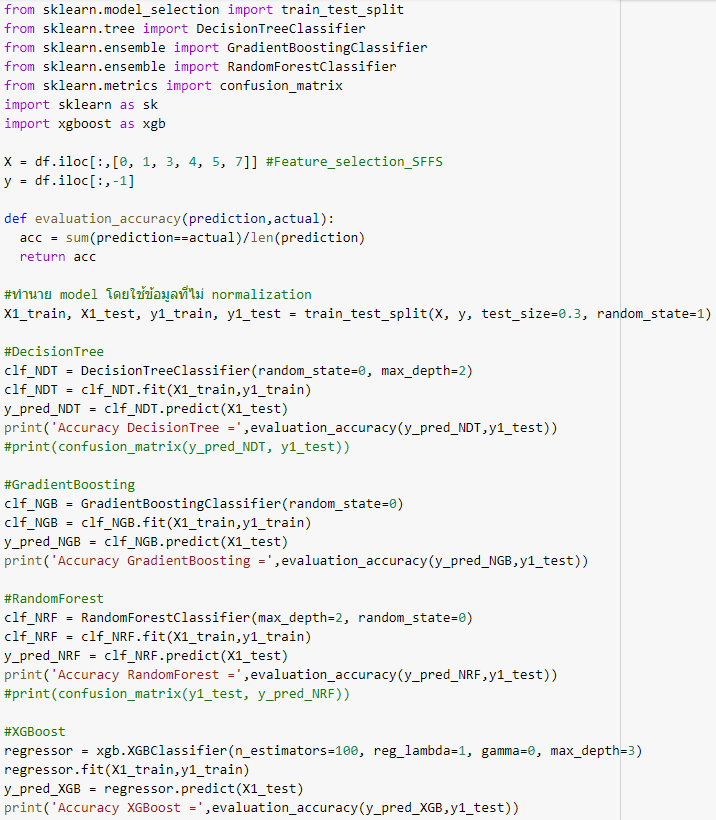


ได้ผลลัพธ์

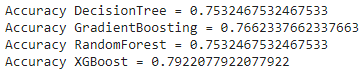


เลือกคอลัมที่มีค่า CV Score มากที่สุดคือ [0,1,3,4,5,7]

นำมาเข้า Model Decision Tree

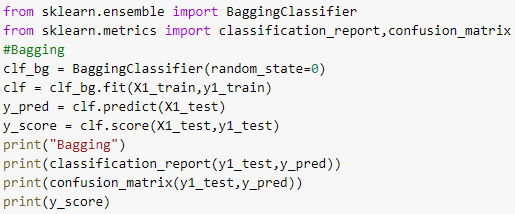


ได้ผลลัพธ์ค่า Accuracy

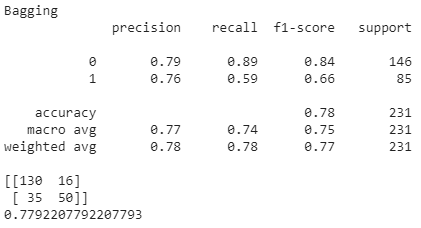


**การทำ Bagging**

โค้ดส่วน Bagging

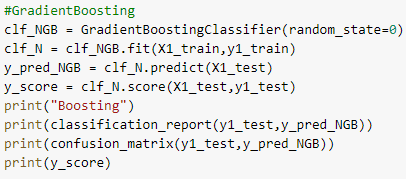


ค่าการทำนายและ confusion matrix



**การทำ Boosting**

โค้ดส่วน Boosting



ค่าการทำนายและ confusion matrix

