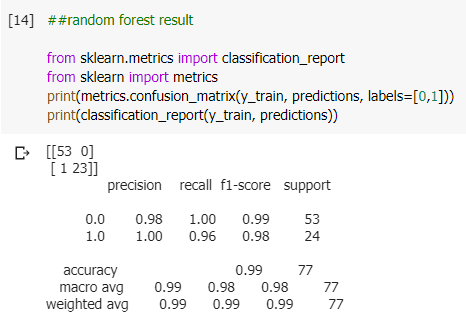
สำหรับ Model นี้เราเลือกใช้ Random Forest ที่ Max depth = 4 เนื่องจากข้อมูลที่เป็น ground truth ที่ต้องการ predict มีน้อยเหลือเกินเราเลยไม่อยากให้มันแตกแขนงกิ่งก้านเยอะเกินไป

ส่วนการ Clean Data เราทำการตัด column Patient\_ID ที่ไม่เกี่ยวและ Exposure\_Time ออก เพราะ Exposure\_Time น่าจะเป็น Bias Column ที่สูงมาก เพราะ Exposure Time ไม่น่าจะเกี่ยวกับการแพ้ยาลองจินตนาการดูว่า คนที่ระยะเวลาการ Test หลังจากใช้ยาผ่านมา 40,000 หน่วยแล้ว จริงๆอาจจะแพ้ยาก็ได้ แต่ระยะเวลาผ่านมานานเกินไปจนผลตรวจไม่พบเป็น Positive ซึ่งการนำ Feature นี้มาทำนายผล แพ้ยาคิดว่าถ้ามี Domain Knowledge น่าจะเป็นสิ่งที่อันตรายมาก เช่นเราต้องการ Test ผู้ป่วยที่ ผ่านการใช้ยามามากกว่า 40,000 หน่วยแล้ว model ย่อมทำนายผลว่าไม่แพ้ยา ทั้งที่จริงๆแล้วคนๆนั้นแค่ไม่แสดงอาการเพราะระยะเวลาผ่านไปนานแล้ว

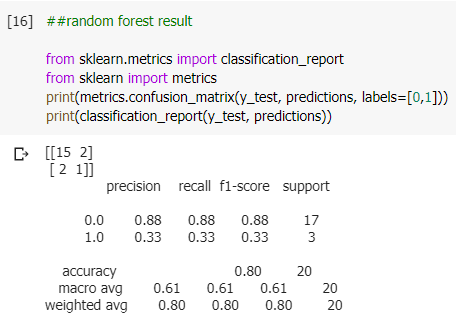
สำหรับผลของ model ออกมากับ Train Set ค่อนข้างดีมาก มีค่า precision recall และ f1 score ที่ดี

(Train set ตัดข้อมูลที่เราไม่รู้ ground truth ออกไป เพราะเราไม่อยากให้โมเดลมันทำนายผลผิดเลย โดยเฉพาะเรื่องอันตรายอย่างแพ้ยา อีกทั้งเราจึงต้องมุ่งเน้นกับ recall มาก เพราะในกรณีการแพ้ยานี้ recall ย่อมสำคัญกว่า precision การทำนายผลว่าแพ้ยาตัวนี้ แต่ในความเป็นจริงไม่แพ้ ในความเป็นจริง เราอาจหลีกเลี่ยงไปใช้ยาตัวอื่นที่ให้สรรพคุณเดียวกันได้ แต่การที่คนแพ้ยาแล้ว model ทำนายว่าไม่แพ้ ความเสียหายย่อมมากกว่า เราสามารถทำการ ปรับ threshold ของ model ลงมาให้มุ่งเน้นที่ recall มากๆได้ แต่ในโค้ดเราใช้ที่ 0.5 ก็ได้ recall ที่ค่อนข้าง ok แล้ว)



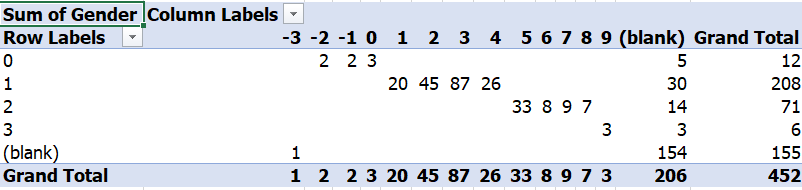
สำหรับผล predict Validation Set ค่อนข้างมีน้อย (Validation set มีคนแพ้ยาแค่ 2 คนเพราะต้องเอา Data ไปให้ trainset ใช้หมด)

เราอาจสามารถทำ Cross Validation เพื่อเพิ่ม Test Set ได้แต่คิดว่าผลลัพธ์ไม่น่าต่างกันเสียเท่าไหร่ กับจำนวน Data ที่น้อยขนาดนี้)



* Address the issue of missing feature values in some way.

สำหรับ Missing Value Issue หลักๆแล้วจะอยู่ที่ Naranjo\_Score และ Naranjo\_Category ซึ่งจากการสำรวจ Data พบว่า 2 feature นี้เกี่ยวเนื่องกัน คนที่อยู่ใน Category 1 จะมี Score อยู่ที่ 1-4 ไม่มีการ intersect ร่วมกับ Category อื่น



ทำให้ในที่นี้เราสามารถ clean data ได้ด้วยการใช้ mode imputation โดยแยกกรุ๊ป สำหรับคนที่ อยู่ Category กรุ๊ป 2 ก็ impute ให้ Score = 5 (เพราะคนส่วนใหญ่ของ Category 2 จะมี Score = 5) ซึ่งทำให้เราได้ Score ของคนที่ Blank ไปได้

ในส่วนของคนที่เป็น null ทั้ง score และ category เราใช้เทคนิคที่เรียกว่า knn nearest neighbor โดยใช้ feature อื่น(ยกเว้น Id คนไข้)มาสร้าง Euclidian Distance และดูว่าคนที่ Score ขาดหายไป อยู่ใกล้เคียงกับคน Score ไหนจำนวนมากที่สุดก็ impute score ให้เท่ากับคนที่ใกล้เคียง (ซึ่งสามารถดูได้ใน code ของ exercise one)

สำหรับการใช้ missing ground truth labels ให้เกิดประโยชน์ ก็คือการนำมาใช้เป็น feature ของ unsupervised learning ในการ clustering เพื่อใช้ หา distance ระยะห่างกับกลุ่มตัวอย่างอื่น เพื่อจัดกลุ่มให้มีค่า Naranjo Score ซึ่งเปิดให้โอกาสให้เราสามารถใช้ค่า Naranjo Score นี้มาสร้าง model predict ground truth label ได้อีกทีหนึ่ง