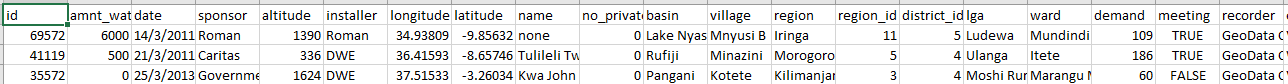
สุรพล กุลเสวตร์ เลขทะเบียน 5709680085

กระบวนการ CRISP-DM ในการวิเคราะห์ข้อมูลเพื่อแก้ไขปัญหาในการพยากรณ์สถานะของบ่อน้ำ

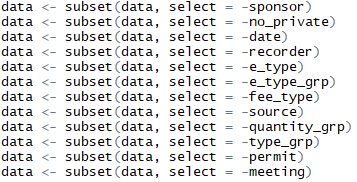
1. ให้น.ศ.เข้าใจปัญหา (Business Understanding)
   1. Scenario
      1. ผู้คนในทวีปแอฟริกาขาดแคลดน้ำดื่ม รัฐบาลแก้ไขปัญหาด้วยการขุดเจาะบ่อน้ำเพื่อจ่ายน้ำ ดังนี้ผู้ใช้ต้องการระบบที่สามารถช่วยดูแลรักษาบ่อน้ำได้ในอยู่ในสภาพใช้งานได้
   2. Business objectives
      1. บ่อน้ำมีเวลาในการใช้งานเพื่อขึ้น
      2. พยากรณ์ถึงสถานะของบ่อจากการใช้จ่ายน้ำในบ่อนั้น ๆ
   3. Business success criteria
      1. สามารถพยากรณ์สถานะของบ่อได้
2. ให้น.ศ.เข้าใจข้อมูล (Data understanding)
   1. ข้อมูลทั้งหมดใน train\_data1.csv มีเป็นข้อมูลที่เกี่ยวข้องกับบ่อน้ำ มีทั้งหมด 5938 แถว มีหลักทั้งหมด 41 หลัก ประกอบไปด้วย
      1. ข้อมูล หมายเลขบ่อน้ำ ปริมาณน้ำที่มีในบ่อน้ำ วันที่บันทึกข้อมูล พิกัด ที่ตั้ง ฯลฯ ในไฟล์ data\_description.pdf ดังภาพที่ 1



ภาพที่ 1 ตัวอย่างข้อมูลของบ่อน้ำ

โดย target ของเราคือ status แถวสุดท้ายของข้อมูลนี้ที่สามารถบอกความสถานะของบ่อน้ำ

* 1. Data Preparation
     1. ในการดูข้อมูล train\_data1 มี attribute มีค่าซ้ำกันและไม่เกี่ยวข้องกับการพยากรณ์ของบ่อน้ำจึงทำการลบออกโดยใช้ภาษา R



ภาพที่ 2 ค่าซ้ำกันและไม่เกี่ยวข้องกับการพยากรณ์

Sponsor ไม่จำเป็นต่อการพยากรณ์

No\_private ค่าที่ไม่ทราบ

Date มีค่าที่ผิดผลาด และ ไม่จำเป็นต่อการพยากรณ์

Recorder เป็นชื่อซ้ำกัน

E\_type และ e\_type\_grp เป็นคำย่อ e\_type\_class

Fee\_type เป็นคำย่องของ fee

Source เป็นค่าซ้ำของ source\_type

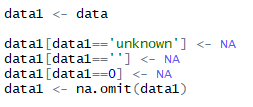
quantity\_grp เป็นค่าซ้ำของ w\_quantity

type\_grp เป็นค่าซ้ำของ type

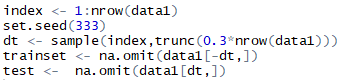
permit ค่าซ้ำกัน

meeting ค่าซ้ำกัน

* + 1. ในไฟล์ train\_data1 มีข้อมูลที่มีค่า 0 ค่าunknown ค่าที่เป็นพื้นที่ว่างจึงต้องทำการลบออกโดยใช้ภาษา R ดังโค็ดตัวอย่าง



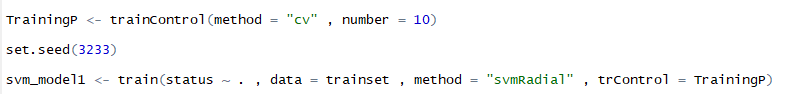
ภาพที่ 3 ข้อมูลที่เป็น missing value จึงต้องทำการลบ

* + - 1. ทำการย้ายข้อมูลไปที่ data1 ข้อมูลในdata1 ทั้งหมดที่มี 0 , unknow ,ค่าว่างแทนด้วย NA จากนั้นนำการลบข้อมูลที่มี NA ทั้งหมด
      2. ข้อมูลที่เหลือทั้งหมดจะเหลือ 1088 แถว มีหลักทั้งหมด 28 หลัก
    1. ทำการแยกข้อมูล data1 เป็น trainset กับ test โดยแบ่งอัตราส่วน 70:30
       1. 

ภาพที่ 4 แบ่งอัตราส่วน 70:30

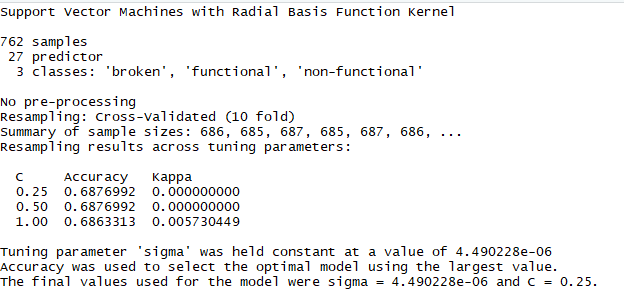
* + - * 1. ตัวแปร trainset มีค่าทั้งหมด 762
        2. ตัวแปร test มีค่าทั้งหมด 326

1. สร้างแบบจำลอง (Modeling)
   * 1. เนื่องจากเรามี target ชัดเจนซึ่งได้แก่ functional , broken , non-functional จึงควรใช้การจำลองในการพยากรณ์แบบ Classification โดยใช้ Support Vector Machine โดยใช้ caret
        1. ทำการจำลองโดยใช้ cross-validation 10 ชั้น เป็นตัวช่วย test performance
        2. ใช้ Support Vector Machine แบบ **Radial Kernel**



ภาพที่ 5 จำลองข้อมูล

**จากการทดสอบโมเดลได้ผลลัพท์ดังนี้**



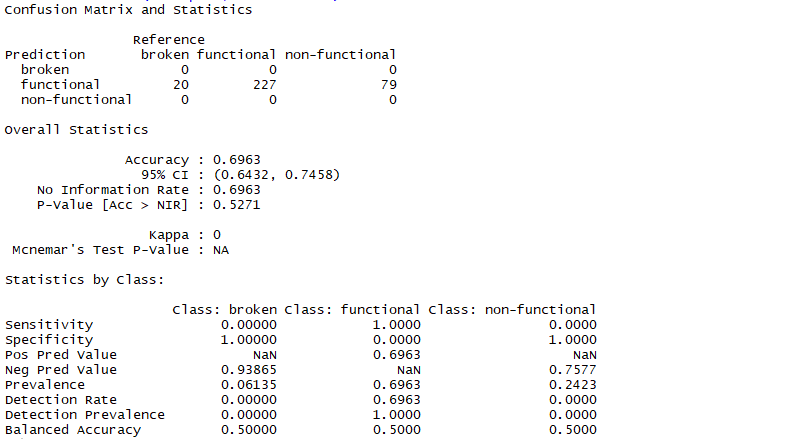
ภาพที่ 6 ผลจาการสร้าง model

* + - 1. **การทดสอบ model กับ test data ที่เหลือ 30**



ภาพที่ 7 **คำสั่ง predict**

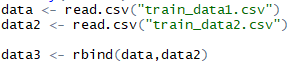
**จากการทำงานคำสั่ง predict และ สร้าง confusion matrix**



ภาพที่ 8 **ผลลัพธ์ confusion matrix**

**จากการทดสอบโมเดลกับtest set ได้ค่า accuracy อยู่ที่ 69% ซึ่งถือว่าค่าน้อย**

4.เพิ่มข้อมูล train\_data2 โดยใช้คำสั่ง

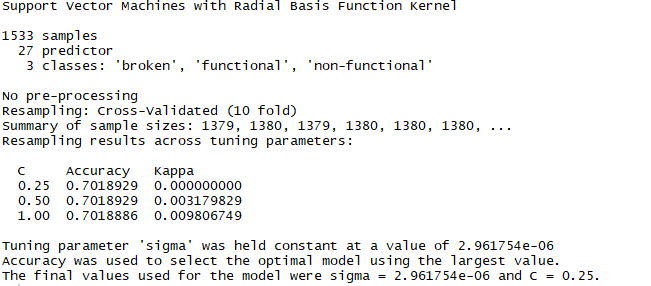


ภาพที่ 9 การเพิ่มข้อมูล

หลังจากการเพิ่มข้อมูลและทำการ Data Preparation เหลือข้อมูลทั้งหมด 2190 row และหลังจากทำการแบ่งข้อมูล 70:30 เป็น train และ test

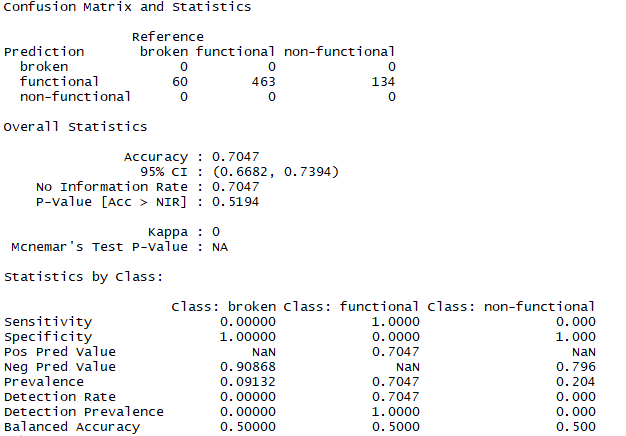
Train set มีจำนวนทั้งหมด 1533 row และ test set มีทั้งหมด 657 row

ทำการโมเดลได้ผลลัพท์ดังนี้



ภาพที่ 10 โมเดล

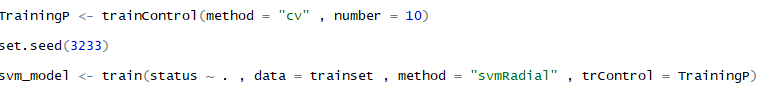
ทำการทดสอบกับ test ได้ผลลัพท์ดังนี้



ภาพที่ 11 **ผลลัพธ์ confusion matrix**

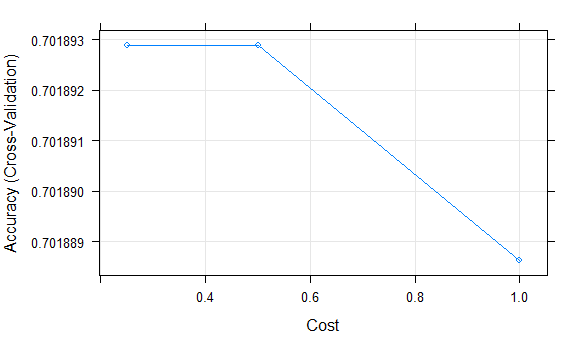
เวลาในการพยากรณ์จะนานขึ้นเนื่องจากข้อมูลที่เพิ่มขึ้น จากการเติม data set เข้าไป ทำให้ค่า accuracy เพิ่มขึ้นเป็น 70 % **ถือว่าค่าน้อยอยู่**

1. จากการจำลอง ในข้อ 2- 4
   1. ข้อมูลทั้งหมด 5940 แถวมี มี 41 หลัก ข้อมูลส่วนใหญ่เป็น 0 ช่องว่าง เว้นวรรก โดยใช้จากคำสั่ง summary
   2. ทำการตัดข้อมูลที่ไม่จำเป็นทั้งหมดทั้งหมดเหลือ 2190 แถว 28 หลัก โดยใช้คำสั่ง subset
   3. ทำการจำลองแบบที่สร้างขึ้นมีลักษณะเป็น Support Vector Machine แบบ **Radial Kernel**
   4. **มีโครงสร้าง**



**TrainingP เป็นตัวแปรที่เก็บค่า cross-validation มีพารามิเตอร์ method คือ ฟังก์ชันนี้จะคำนวณข้อผิดพลาดในการคาด พารามิเตอร์ number คือ รอบของ cross-validation**

Svm\_model **เป็นตัวแปรที่เก็บค่าโมเดล มีพารามิเตอร์ status ¬ . คือค่า target และค่าจาก attribute ที่ต้องการในที่นี้ใช่ . เพื่อเอาทุก attribute พารามิเตอร์ method คือ โมเดลที่ต้องการ พารามิเตอร์ trControl คือตัวแปรควบคุม cross-validation**

* 1. 

ภาพที่ 12 cross-validation

* + 1. จากกราฟด้านบนแสดงให้เห็นว่า C ในช่วง 0.25 – 0.5 มีค่า accuracy สูงสุดคือ 70.1%
  1. ในการนำไปใช้จริงโมเดลนี้ยังไม่สามารถทำได้ดี อยู่ในระดับพอใช้ได้ เนื่องจากค่า accuracy น้อยอยู่ที่ 70.1%

6.สรุป

จากข้อมูลผู้คนในทวีปแอฟริกาขาดแคลดน้ำดื่ม รัฐบาลแก้ไขปัญหาด้วยการขุดเจาะบ่อน้ำเพื่อจ่ายน้ำ ระบบที่สามารถช่วยดูแลรักษาบ่อน้ำได้ในอยู่ในสภาพใช้งานได้ ในการพยากรณ์ข้อมูลโดยใช้โมเดล Support Vector Machine แบบ **Radial Kernel จากข้อมูลทั้งหมด** 5940 รวม train\_data1 และ train\_data2 **และได้ทำการ** Data Preparation จนเหลือ 2190 **ได้ผลความแม่นยำ 70 % จากผลความแม่นยำผมคิดว่าข้อมูลน้อยไปทำให้ความแม่นยำน้อยลง**