P-ISSN 0858-4435
E-ISSN 2651-1231
วารสารวิทยาศาสตร์และเทคโนโลยี Homepage : *http://li01.tci-thaijo.org/index.php/tstj*
ปีที่ 30 ฉบับที่ 5 (กันยายน-ตุลาคม 2565) หน้า [26-36]

การเปรียบเทียบวิธีการจำแนกคุณภาพน้ำในประเทศไทย

Comparison of Water Quality Classification Methods in Thailand

กัลยา บุญหล้า*, วรวุฒิ มหาโพธิ์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร พิษณุโลก 65000
Kanlaya Boonlha*, Worawut Mahapho
Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok 65000
Received 26 June 2022; Received in revised 11 September 2022; Accepted 4 October 2022

บทคัดย่อ

การวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบวิธีการจำแนกคุณภาพน้ำด้วยการวิเคราะห์การถดถอยลอจิสติกทวิภาค ต้นไม้การตัดสินใจ และเพื่อนบ้านใกล้ที่สุด โดยใช้ค่าความแม่นในการจำแนกกลุ่มเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของแต่ละวิธี ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลจากแหล่งน้ำทั่วประเทศไทยเก็บรวบรวมโดยกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ตั้งแต่วันที่ 1 มกราคม 2561 ถึง วันที่ 1 มกราคม 2564 ทำการจำแนกคุณภาพน้ำเป็น 2 กลุ่ม ได้แก่ กลุ่มน้ำที่มีคุณภาพได้ตามมาตรฐานและไม่ได้ตามมาตรฐาน ตัวแปรอิสระที่ใช้ในการศึกษามีทั้งสิ้น 14 ตัวแปร แบ่งข้อมูลด้วยวิธีการตรวจสอบไขว้แบ่งเป็น10 กลุ่ม พบว่า ต้นไม้การตัดสินใจที่เลือกตัวแปรอิสระตามการวิเคราะห์การถดถอยลอจิสติกทวิภาคที่ประกอบด้วย ความขุ่นของน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แอมโมเนีย-ไนโตรเจน แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ออกซิเจนที่ละลายในน้ำ สารอินทรีย์ในน้ำ ให้ค่าความแม่นสูงที่สุดร้อยละ 89.64 ต้นไม้การตัดสินใจที่เลือกตัวแปรอิสระทุกตัวแปรให้ค่าความแม่นร้อยละ 88.71 การถดถอยลอจิสติกทวิภาคให้ค่าความแม่นร้อยละ 87.25 และเพื่อนบ้านใกล้ที่สุดที่เลือกตัวแปรอิสระทุกตัวแปรให้ค่าความแม่นน้อยที่สุดร้อยละ 79.05

คำสำคัญ: คุณภาพน้ำ; เหมืองข้อมูล; การถดถอยลอจิสติกทวิภาค; ต้นไม้การตัดสินใจ; เพื่อนบ้านใกล้ที่สุด

Abstract

The objective of this research was to compare the classification of water quality with binary logistic regression analysis, decision tree, and K-Nearest Neighbors using the accuracy group classification as a criterion to compare the efficiency of each method. The data used in the study were from water sources throughout Thailand collected by the Water Quality Management Division, Pollution Control Department from January 1, 2018, to January 1, 2021. The water quality is classified

***ผู้รับผิดชอบบทความ:** kanlayab@nu.ac.th **doi:**

into two groups; the water quality is standards and is not standards. A total of 14 independent variables were used in the study. The datasets were divided with a 10-fold cross-variation. The results showed that according to the binary logistic regression analysis, the decision tree selected variables, including water turbidity, total coliform bacteria, ammonia-nitrogen, fecal coliform bacteria, dissolved oxygen, and organic water, have the highest accuracy of 89.64%. The decision tree selected for all independent variables has an accuracy value of 88.71 %. While K-Nearest Neighbors selected all independent variables have the lowest accuracy value of 79.05 %

Keywords: Water quality; Data mining; Binary logistic regression; Decision tree; K-Nearest Neighbor

1. บทนำ

น้ำเป็นสารประกอบที่พบมากถึง 3 ใน 4 ส่วน ของพื้นโลก โดยส่วนใหญ่อยู่ในสภาพน้ำเค็มในทะเล และมหาสมุทรประมาณร้อยละ 97 เป็นน้ำแข็งตาม ขั้วโลกประมาณร้อยละ 2 และเป็นน้ำจืดตามแม่น้ำ ลำคลองต่างๆ ประมาณร้อยละ 1 ถ้าโลกปราศจากน้ำ สิ่งมีชีวิตต่างๆ จะไม่สามารถดำรงชีวิตอยู่ได้ ปัจจุบัน แม่น้ำลำคลองหลายแห่งกลายเป็นน้ำเสีย ซึ่งเกิดจาก ชุมชน อุตสาหกรรม และเกษตรกรรม ในประเทศไทยมี หน่วยงานที่ตรวจสอบคุณภาพแหล่งน้ำที่สำคัญ คือ สำนัก จัดการคุณภาพน้ำ กรมควบคุมมลพิษ [1] ได้ประเมิน คุณภาพน้ำของแหล่งน้ำผิวดินโดยทั่วไปหรือแหล่งน้ำจืด ในแม่น้ำลำคลอง โดยใช้ดัชนีคุณภาพน้ำทั่วไป (Water Quality Index, WQI) ที่มีช่วงคะแนนจาก 0 ถึง 100 คะแนน แบ่งคะแนนเป็น 5 ระดับ ได้แก่ ช่วง 91 – 100 คะแนน จัดว่าคุณภาพน้ำอยู่ในเกณฑ์ดีมาก ช่วง 71 – 90 คะแนน จัดว่าคุณภาพน้ำอยู่ในเกณฑ์ดี ช่วง 61 – 70 คะแนน จัดว่าคุณภาพน้ำอยู่ในเกณฑ์พอใช้ ช่วง 31 – 60 คะแนน จัดว่าคุณภาพน้ำอยู่ในเกณฑ์เสื่อมโทรม และช่วง 0 – 30 คะแนน จัดว่าคุณภาพน้ำอยู่ในเกณฑ์เสื่อมโทรม มาก ปี พ.ศ. 2564 สำนักจัดการคุณภาพน้ำ [2] ได้รายงาน ว่า แหล่งน้ำสำคัญทั่วประเทศ 65 แหล่งน้ำ อยู่ในเกณฑ์ ดีมากเพิ่มขึ้นร้อยละ 2 (เท่ากับปี พ.ศ. 2563) เกณฑ์ ดีเพิ่มขึ้นร้อยละ 40 (เพิ่มขึ้นจาก พ.ศ. 2563 ร้อยละ 3) เกณฑ์พอใช้เพิ่มขึ้น ร้อยละ 44 (เพิ่มขึ้นจากปี พ.ศ. 2563 ร้อยละ 1) และเกณฑ์เสื่อมโทรมลดลงร้อยละ 14 (ลดลงจากปี พ.ศ. 2563 ร้อยละ 4) ไม่มีแหล่งน้ำที่ มีคุณภาพน้ำอยู่ในเกณฑ์เสื่อมโทรมมาก สาเหตุหลักมา จากการระบายน้ำเสียที่เกิดจากการใช้น้ำในชุมชนและ การท่องเที่ยว กระบวนการผลิตในโรงงานอุตสาหรรม ระบบบำบัดน้ำเสียบางแห่งยังไม่มีประสิทธิภาพ และมี ไม่เพียงพอครอบคลุมพื้นที่ให้บริการ

การศึกษาการจำแนกประเภทข้อมูลคุณภาพ ของน้ำโดยใช้เทคนิคเหมืองข้อมูลมีการศึกษาอย่าง กว้างขวาง เช่น Gakii และ Jepkoech [3] ศึกษาการ วิเคราะห์คุณภาพน้ำโดยใช้ต้นไม้ตัดสินใจ (Decision Tree) เพื่อวิเคราะห์ข้อมูลคุณภาพน้ำในประเทศเคนยา โดยตัวแปรอิสระ ระดับกรด-เบส ความเป็นด่าง การนำ ไฟฟ้า และสี ในการทดลองใช้ชุดข้อมูลร้อยละ 80 เป็น ชุดข้อมูลเรียนรู้และร้อยละ 20 เป็นชุดข้อมูลทดสอบ ผล ที่ได้จากการศึกษา พบว่า วิธีต้นไม้ตัดสินใจอัลกอริทึม J48 มีความแม่นสูงสุดร้อยละ 94 และวิธีตอไม้ตัดสิน ใจ (Decision Stump) มีความแม่นต่ำสุดร้อยละ 83 ปี ค.ศ. 2021 Ramadhani และ Rahmawita [4] ศึกษา การจำแนกคุณภาพน้ำ แม่น้ำ 3 แห่งในจังหวัด Riau ประเทศอินโดนีเซีย เก็บรวบรวมข้อมูล 624 รายการ โดยใช้ระดับกรด-เบส แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด ของแข็งที่ละลายน้ำทั้งหมด ค่าออกซิเจนที่ละลายในน้ำ ค่าความต้องการออกซิเจนทางชีวภาพ ค่าความต้องการ

ออกซิเจนทางเคมี ปริมาณไนเตรทไดออกไซด์ ปริมาณไนเตรท-ไนโตรเจน ปริมาณแอมโมเนีย ปริมาณคลอรีนอิสระ ปริมาณฟอสเฟตทั้งหมด ปริมาณฟีนอล ปริมาณน้ำมันและไขมัน ปริมาณผงซักฟอก ปริมาณแบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ปริมาณแบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด ปริมาณไฮโดรเจนซัลไฟด์ ปริมาณเหล็ก ปริมาณแคดเมียม ปริมาณสังกะสี ปริมาณทองแดง และปริมาณสารตะกั่ว โดยใช้เทคนิคการจำแนกเพื่อนบ้านใกล้ที่สุดแบบปรับปรุง (Modified K-Nearest Neighbor, MKNN) ที่ K = 5 ค่าความแม่นสูงสุดร้อยละ 85.10 Wechmongkhonkon และคณะ [5] ทำการศึกษาข้อมูลน้ำ 11 แหล่งของคลองในเขตดุสิต กรุงเทพฯ ประเทศไทย โดยใช้ข้อมูลปี พ.ศ. 2550 - 2554 ด้วยการประยุกต์ใช้โครงข่ายประสาทเทียมกับการจำแนกคุณภาพน้ำผิวดิน โดยตัวแปรอิสระ 6 ตัวแปร ได้แก่ ระดับกรด-เบส ออกซิเจนละลายน้ำ ความต้องการออกซิเจนทางชีวเคมี ไนเตรตไนโตรเจน แอมโมเนียไนโตรเจน และแบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด พบว่าโครงข่ายประสาทเทียมแบบหลายชั้นเครือข่ายแสดงอัตราการรับรู้หลายชั้นมีความแม่นสูงร้อยละ 96.52 นอกจากนี้ Najah และคณะ [6] การประยุกต์ใช้โครงข่ายประสาทเทียมในการวิเคราะห์ทำนายคุณภาพน้ำในประเทศมาเลเซีย Diamantopoulou และคณะ [7] ทำนายค่ารายเดือนของพารามิเตอร์คุณภาพน้ำของแม่น้ำในประเทศกรีซ โดยใช้โครงข่ายประสาทเทียม และ Areerachakul และ Sanguansintukul [8] การประยุกต์ใช้โครงข่ายประสาทเทียมในการวิเคราะห์ทำนายคุณภาพน้ำของคลองในกรุงเทพฯ จำนวน 229 คลอง โดยใช้ข้อมูลจากจากกรมระบายน้ำและระบายน้ำทิ้งกรุงเทพมหานคร ในช่วงปี พ.ศ. 2546-2550

ในปี พ.ศ. 2560 Pudchaya [9] ทำการศึกษาประเมินค่าคุณภาพน้ำของแม่น้ำบางปะกง โดยพิจารณาจากปริมาณความเข้มข้นของพารามิเตอร์ที่เป็นตัวบ่งบอกคุณภาพน้ำของแม่น้ำบางปะกง ตั้งแต่ปี พ.ศ. 2540 - 2558 โดยตัวแปรที่ใช้ ประกอบด้วย อุณหภูมิ การนำไฟฟ้า สารแขวนลอย ระดับกรด-เบส ออกซิเจนละลายน้ำ บีโอดี ไนเตรท ฟอสเฟต ของแข็งทั้งหมด ฟีคอลโคลิฟอร์มแบคทีเรีย แบคทีเรียโคลิฟอร์มทั้งหมด ความขุ่น ความเค็ม แบคทีเรียกลุ่มโคลิฟรอมทั้งหมด ฟอสฟอรัสทั้งหมด ไนไตรท์-ไนโตรเจน สร้างสมการจำแนกกลุ่มเพื่อทำนายกลุ่มคุณภาพน้ำของแม่น้ำบางปะกง โดยแบ่งกลุ่มคุณภาพน้ำของแม่น้ำบางปะกง ออกเป็น 2 กลุ่มคือ กลุ่มน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $0 \leq WQI \leq 60$ และกลุ่มน้ำที่มีคุณภาพตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $61 \leq WQI \leq 100$ ด้วยการวิเคราะห์จำแนกกลุ่ม โดยพิจารณาคะแนนการจำแนกกลุ่มเชิงเส้น (Linear Discriminant Score) เพื่อใช้ในการจำแนกประเมินค่าสมการจำแนกกลุ่มที่ได้ในการทำนายกลุ่มคุณภาพน้ำของแม่น้ำบางปะกง พบว่าสัดส่วนความถูกต้องในการจำแนกกลุ่มสำหรับชุดข้อมูลเรียนรู้ที่เท่ากับร้อยละ 85.76 และความถูกต้องในการจำแนกกลุ่มสำหรับชุดข้อมูลทดสอบร้อยละ 68.75

ในการศึกษานี้ผู้วิจัยสนใจที่ศึกษาการจำแนกประเภทข้อมูลของคุณภาพน้ำโดยใช้ข้อมูลจากกรมควบคุมมลพิษ จำแนกน้ำออกเป็น 2 กลุ่มคือกลุ่มน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $0 \leq WQI \leq 60$ และกลุ่มน้ำที่มีคุณภาพตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $61 \leq WQI \leq 100$ ตามการแบ่งกลุ่มของ Pudchaya [8] ด้วยเทคนิคเหมืองข้อมูล วิธีต้นไม้การตัดสินใจ และเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) นำมาเปรียบเทียบกับการวิเคราะห์ทางสถิติการวิเคราะห์ถดถอยลอจิสติกทวิภาค (Binary Logistic Regression Model) โดยมีเป้าหมายหลักเพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ นำสมการถดถอยที่ได้ไปพยากรณ์และทำการเปรียบเทียบประสิทธิภาพของต้นไม้ตัดสินใจเพื่อนบ้านใกล้ที่สุด ที่ใช้ทุกตัวแปรในการจำแนกกลุ่มและใช้ตัวแปรที่เลือกตัวแปรอิสระตามการวิเคราะห์การถดถอยลอจิสติกทวิภาคในการจำแนกกลุ่ม โดยพิจารณาจากค่าความแม่น (Accuracy) เป็นเกณฑ์ ในการศึกษานี้

แบ่งข้อมูลแต่ละเทคนิคด้วยวิธีการตรวจสอบไขว้โดยแบ่งเป็น k กลุ่ม (k-fold Cross-variation) โดยกำหนด k=10

2. วิธีการวิจัย

การวิจัยในครั้งนี้ เป็นการศึกษาการจำแนกคุณภาพของน้ำจากแหล่งน้ำทั่วประเทศไทย จากจุดตรวจวัดคุณภาพน้ำ 366 ซึ่งเป็นข้อมูลจาก กรมควบคุมมลพิษ โดยมีขั้นตอนในการดำเนินงานวิจัยดังนี้

ขั้นที่ 1 ศึกษาและรวบรวมข้อมูลที่เกี่ยวข้อง จากการเก็บรวบรวมข้อมูลคุณภาพของน้ำจากแหล่งน้ำทั่วประเทศไทย 59 แหล่งน้ำสายหลัก และ 6 แหล่งน้ำนิ่ง จากจุดตรวจวัดคุณภาพน้ำ 366 รวบรวมข้อมูลจากสำนักจัดการคุณภาพน้ำ กรมควบคุมมลพิษ [2] ตั้งแต่วันที่ 1 มกราคม พ.ศ. 2561 ถึง วันที่ 1 มกราคม พ.ศ. 2564 จำนวนข้อมูลทั้งสิ้น 4,453 ชุด โดยตัวแปรที่ใช้ในการศึกษามีทั้งสิ้น 15 ตัวแปร จำแนกเป็น ตัวแปรอิสระ จำนวน 14 ตัวแปร ได้แก่ อุณหภูมิอากาศ (temp A) อุณหภูมิน้ำ (temp W) กรด-เบส (pH) ความขุ่นของน้ำ (tur) การนำไฟฟ้า (Cond) ออกซิเจนที่ละลายในน้ำ (DO) สารอินทรีย์น้ำ(BOD) แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด (Total Coli) แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม (Fecal Coli) ฟอสฟอรัสทั้งหมด (TP) ไนเตรท-ไนโตรเจน (NO3-N) แอมโมเนีย-ไนโตรเจน (NH3-N) สารแขวนลอยที่เป็นของแข็ง (SS) ของแข็งที่ละลายน้ำทั้งหมด (TS) ตัวแปรตามจำนวน 1 ตัวแปร ได้แก่ ดัชนีคุณภาพน้ำซึ่งจำแนกเป็น 2 กลุ่มตามเกณฑ์ โดยผู้วิจัยจำแนกออกเป็น 2 กลุ่มคือกลุ่มน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $0 \leq WQI \leq 60$ และกลุ่มน้ำที่มีคุณภาพตามมาตรฐาน ซึ่งมีค่าดัชนีคุณภาพน้ำเท่ากับ $61 \leq WQI \leq 100$ ตามการแบ่งกลุ่มของ Pudchaya [8] กำหนด แทน กลุ่มน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน และ $y = 1$ แทน กลุ่มน้ำที่มีคุณภาพได้ตามมาตรฐาน

ขั้นที่ 2 เตรียมข้อมูลสำหรับการนำไปวิเคราะห์ โดยผู้วิจัยทำการตรวจสอบข้อมูล ตัดข้อมูลที่มีค่าผิดปกติและข้อมูลที่มีค่าสูญหาย พบว่ามีข้อมูลค่าผิดปกติ จำนวน 211 ชุด และมีข้อมูลที่มีค่าสูญหาย จำนวน 2,533 ชุด ซึ่งเกิดจากอุปกรณ์ และภัยพิบัติทางธรรมชาติ เช่น การใช้การใช้อุปกรณ์ที่มีความละเอียดต่างกัน อุปกรณ์เกิดการชำรุดระหว่างเก็บข้อมูล เกิดภัยแล้งทำให้ไม่สามารถเก็บข้อมูลได้ จึงเหลือข้อมูลที่ใช้วิเคราะห์ทั้งสิ้น 1,709 ชุด ทำการวิเคราะห์ข้อมูลเบื้องต้นโดยใช้ค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน

ขั้นที่ 3 ทำการวิเคราะห์ข้อมูลโดยโปรแกรม RapidMiner โดยแบ่งข้อมูลที่ใช้ในการศึกษาแต่ละเทคนิคด้วยวิธีการตรวจสอบไขว้กำหนด k=10 วิเคราะห์ข้อมูลโดยวิธีการถดถอยลอจิสติกทวิภาค กำหนดระดับนัยสำคัญทางสถิติ 0.05 ทำการวิเคราะห์ต้นไม้การตัดสินใจ และการวิเคราะห์วิธีเพื่อนบ้านใกล้ที่สุด โดยใช้ตัวแปรทั้งหมดที่ทำการศึกษา และตัวแปรที่ผ่านการเลือกจากการวิเคราะห์การถดถอยลอจิสติกทวิภาค ในการวิเคราะห์วิธีเพื่อนบ้านใกล้ที่สุด ผู้วิจัยกำหนดขนาดของ K โดยใช้ Operators Optimize Parameter เมื่อ K คือ จำนวนข้อมูลที่อยู่ใกล้เคียงกันจำนวน K ตัว ซึ่งผู้วิจัยใช้ K ตั้งแต่ 5 ถึง 25 ทำการวิเคราะห์ค่า K ที่มีค่าความแม่นมากที่สุด รายละเอียดการวิเคราะห์แต่ละวิธีดังนี้

2.1 การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression Analysis)

การวิเคราะห์การถดถอยลอจิสติกทวิภาค [10] เป็นการวิเคราะห์เพื่อทำนายโอกาสที่เหตุการณ์ที่สนใจจะเกิดขึ้น และสมการถดถอยลอจิสติกที่ดีจะต้องประกอบด้วยตัวแปรอิสระที่เหมาะสมที่จะทำให้ค่าทำนายโอกาสที่จะเกิดนั้นใกล้เคียงกับความเป็นจริง ที่ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ มีค่าได้เพียง 2 ค่า ส่วนตัวแปรอิสระอาจจะเป็นข้อมูลเชิงปริมาณหรือข้อมูลเชิงคุณภาพ จะใช้ประมาณโอกาสที่จะเกิดเหตุการณ์ ผลของการวิเคราะห์ขึ้นอยู่กับปัจจัยที่สำคัญคือการใช้แบบจำลองที่เหมาะสมในการเลือกปัจจัยที่สำคัญ การแบ่งกลุ่มย่อยและ

จำนวนตัวอย่างในแต่ละกลุ่มย่อยของปัจจัยนั้น โดยมีเป้าหมายหลักเพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรตาม 1 ตัว กับตัวแปรอิสระ p ตัว และนำสมการถดถอยที่ได้ไปพยากรณ์ตัวแปรตาม โดยกำหนดให้

$$P_y = \frac{e^{b_0+b_1x_1+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+\cdots+b_px_p}} \quad (1)$$

$$Q_y = 1 - P_y = \frac{1}{1+e^{b_0+b_1x_1+\cdots+b_px_p}} \quad (2)$$

เมื่อ

P_y แทน ความน่าจะเป็นของการเกิดเหตุการณ์

Q_y แทน ความน่าจะเป็นของการไม่เกิดเหตุการณ์

e แทน ค่าคงตัว มีค่าประมาณ 2.71828

$b_1,\ldots,b_p$ แทน ค่าสัมประสิทธิ์ถดถอยลอจิสติกของตัวแปรอิสระตัวที่ 1 ถึง p

$x_1,\ldots,x_p$ แทน ค่าตัวแปรอิสระตัวที่ 1 ถึง p

การเขียนตัวแบบลอจิสติกจะอยู่ในรูป log ของ Odds เรียกว่า ลอจิต (Logit) หรือฟังก์ชันตอบสนองลอจิต ดังนี้

$$Odds = \frac{P_y}{Q_y} \quad (3)$$

$$log(Odds) = log\left(\frac{P_y}{Q_y}\right) \quad (4)$$

$$log(Odds) = e^{b_0+b_1x_1+\cdots+b_px_p} \quad (5)$$

ถ้าค่า Odds มีค่ามากกว่า 1 แสดงว่าเหตุการณ์นั้นมีโอกาสเกิดขึ้นมากกว่าที่จะไม่เกิดขึ้น

เทคนิคการเลือกตัวแปรอิสระเข้าสมการการถดถอยลอจิสติกมีหลากหลายวิธี เช่น การเลือกตัวแปรด้วยวิธีพิจารณาทุกรูปแบบ (Enter Method) การเลือกตัวแปรด้วยวิธีการเพิ่มตัวแปร (Forward Method) การเลือกตัวแปรด้วยวิธีการลดตัวแปร (Backward Method) สำหรับการวิจัยครั้งนี้ผู้วิจัยทำการวิเคราะห์โดยเลือกตัวแปรด้วยวิธีการเพิ่มตัวแปร ซึ่งเป็นเทคนิคหนึ่งโดยที่ทดสอบเพื่อเลือกตัวแปรของสมการถดถอยลอจิสติก โดยจะเริ่มด้วยรูปแบบที่ไม่มีตัวแปรอิสระอยู่ ทำการเพิ่มตัวแปรอิสระทีละตัวเข้าไปในตัวแบบการถดถอย ในแต่ละขั้นตอนที่มีการเพิ่มตัวแปรอิสระจะพิจารณาจากค่าความน่าจะเป็นของตัวทดสอบอัตราส่วนน่าจะเป็น (likelihood ratio test statistic) และถ้าตัวแปรอิสระใดไม่มีผลต่อความน่าจะเป็นของตัวแปรตาม จะทำการตัดตัวแปรอิสระที่ไม่มีนัยสำคัญ จนได้สมการสุดท้ายที่มีเฉพาะตัวแปรที่มีนัยสำคัญต่อตัวแปรตามเท่านั้น และทำการตรวจสอบความเหมาะสมของสมการถดถอยลอจิสติกที่ได้โดยวิธีทดสอบความเหมาะสมของตัวแบบโดยใช้วิธีสารูปสนิทดี Hosmer-Lemeshow ที่ระดับนัยสำคัญ 0.05

2.2 ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจเป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (Hierarchy) ใช้ สนับสนุนการตัดสินใจ โดยจะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดราก (Root Node) อยู่ด้านบนสุดและใบ (Leaf) อยู่ล่างสุดของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ (Attribute) เป็นตัวทดสอบ กิ่งของต้นไม้ (Branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือก ทดสอบ และใบ (Leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (Class) หรือ นั่นก็คือ ผลลัพธ์ที่ได้จากการทำนาย

โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดรากหลังจากนั้นจะทำการหาตัวแปรอิสระที่มีความสัมพันธ์ในลำดับถัดมาเรื่อยๆ เพื่อหาปมตัดสินใจ (Decision Node) สำหรับโครงสร้างต่อๆ ไป จนถึงใบ ซึ่งเป็นค่าของตัวแปรตาม ซึ่งขั้นตอนวิธีที่ใช้ในการสร้างต้นไม้ตัดสินใจจะนำหลักการทฤษฎีสารสนเทศ (Information Theory) มาใช้โดยมีการวัดปริมาณสารสนเทศของข้อมูลด้วยค่าเอนโทรปี (Entropy) ซึ่งเป็นปริมาณที่บ่งบอกความไม่แน่นอน ซึ่งค่าสารสนเทศของข้อมูลจะขึ้นอยู่กับความน่าจะเป็นของข้อมูล [10, 11] สามารถเขียนในรูปสมการ ได้ดังนี้

$$info(D) = -\sum_{i=1}^{m} plog_2(p_i) \quad (6)$$

เมื่อ

p_i แทน ความน่าจะเป็นที่ข้อมูลในฐานข้อมูล D อยู่ในกลุ่ม C_i ซึ่งมีค่า $|C_iD|/|D|$

p แทน จำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

C_i แทน กลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง m

$|C_iD|$ แทน จำนวนข้อมูลในฐานข้อมูล D ที่อยู่ในกลุ่ม C_i

2.3 เพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors, K-NN)

เพื่อนบ้านที่ใกล้ที่สุดหลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัว จากชุดข้อมูลเรียนรู้ทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่หรือข้อมูลที่ป้อนถามกับชุดข้อมูลทดสอบ จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกค่าที่สมาชิกส่วนใหญ่ K ดังกล่าวอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูล K ตัวประกอบด้วยแอตทริบิวหลายตัวแปรจะนำมาใช้ในการแบ่งกลุ่ม โดยระบุค่าตัวเลขจำนวนเต็ม K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณีที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ K-NN การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝนมาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทางใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบยูคลิเดียน (Euclidean Distance) มีหลักการคือการวัดระยะทางระหว่างวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้ามีระยะทางระหว่างวัตถุน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก [11]

ขั้นที่ 4 พิจารณาประสิทธิภาพของข้อมูล หลังจากสร้างแบบจำลองวิเคราะห์ข้อมูลแล้ว นำตัวแบบของแบบจำลองที่ได้มาทดสอบประสิทธิภาพของข้อมูลดังสมการที่ (7)

$$\text{ค่าความแม่น (Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

เมื่อ

TP แทน ค่าคลาสเป้าหมาย คือ น้ำที่มีคุณภาพตามมาตรฐานและตัวแบบทำนายว่าน้ำที่มีคุณภาพตามมาตรฐาน

FP แทน ค่าคลาสเป้าหมาย คือ น้ำที่มีคุณภาพตามมาตรฐานและตัวแบบทำนายว่าน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน

TN แทน ค่าคลาสเป้าหมาย คือ น้ำที่มีคุณภาพไม่ได้ตามมาตรฐานและตัวแบบทำนายว่าน้ำที่มีคุณภาพไม่ได้ตามมาตรฐาน

FN แทน ค่าคลาสเป้าหมาย คือ น้ำที่มีคุณภาพไม่ได้ตามมาตรฐานและตัวแบบทำนายว่าน้ำที่มีคุณภาพตามมาตรฐาน

โดยในการศึกษาครั้งนี้จะพิจารณาจากค่าความแม่นที่ใช้ในการจำแนกประเภทข้อมูล ที่ให้ค่าความแม่นสูงสุด แสดงว่ามีประสิทธิภาพมากที่สุด

3. ผลการวิจัยและวิจารณ์

การวิเคราะห์การถดถอยลอจิสติกทวิภาค ต้นไม้การตัดสินใจ เพื่อนบ้านใกล้ที่สุด ด้วยโปรแกรม Rapid-Miner ผลการวิจัยสรุปได้ดังนี้

3.1 การศึกษาและการจำแนกคุณภาพของน้ำจากแหล่งน้ำทั่วประเทศไทยจำนวน 1,709 ชุด จำแนกเป็นคุณภาพน้ำไม่ได้ตามมาตรฐาน 570 ชุด และคุณภาพน้ำได้ตามมาตรฐานจำนวน 1,139 ชุดผลการวิเคราะห์ข้อมูลเบื้องต้นจากตารางที่ 1 พบว่า ตัวแปรความขุ่นของน้ำ การนำไฟฟ้า สารอินทรีย์ในน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ฟอสฟอรัสทั้งหมด ไนเตรท-ไนโตรเจน แอมโมเนีย-ไนโตรเจน สารแขวนลอยที่เป็นของแข็ง ของแข็งที่ละลายน้ำทั้งหมด มีค่าส่วนเบี่ยงเบนมาตรฐานสูงกว่าค่าเฉลี่ยเนื่องจากข้อมูลของตัวแปรดังกล่าวมีความแตกต่างกันมาก

3.2 การวิเคราะห์การถดถอยลอจิสติกทวิภาค จากตารางที่ 2 พบว่าตัวแปรอิสระที่มีผลต่อการจำแนกคุณภาพน้ำมีจำนวน 6 ตัวแปร ได้แก่ ความขุ่นของน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แอมโมเนีย-ไนโตรเจน แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ออกซิเจนที่ละลายในน้ำ สารอินทรีย์ในน้ำ มีผลต่อการจำแนกคุณภาพน้ำได้ ที่ระดับนัยสำคัญ 0.05 ค่า Cox & Snell R^2 เท่ากับ 0.600 หมายความว่า ตัวแบบการถดถอยลอจิสติกทวิภาคสามารถอธิบายความผันแปรในการวิเคราะห์จำแนกคุณภาพน้ำได้ร้อยละ 60.0 การตรวจสอบความเหมาะสมของสมการถดถอยลอจิสติกที่ได้โดยใช้วิธีสารูปสนิทดี Hosmer-Lemeshow พบว่าค่าไคกำลังสองเท่ากับ 2.387 และค่าพีเท่ากับ 0.967 มากกว่าระดับนัยสำคัญ 0.05 แสดงว่าตัวแบบที่ได้มีความเหมาะสมที่ระดับนัยสำคัญ 0.05

3.3 การเปรียบเทียบประสิทธิภาพของการจำแนกคุณภาพน้ำ จากตารางที่ 3 พบว่า การวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจโดยการนำตัวแปรที่ได้จากการวิเคราะห์การถดถอยลอจิสติกทวิภาคที่ประกอบด้วยตัวแปร ความขุ่นของน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แอมโมเนีย-ไนโตรเจน แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ออกซิเจนที่ละลายในน้ำ สารอินทรีย์ในน้ำ มีค่าความแม่นสูงที่สุดร้อยละ 89.64 รองลงมาเป็นการวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจโดยการทุกตัวแปรมาพิจารณาการวิเคราะห์การถดถอยลอจิสติกทวิภาค มีค่าความแม่นร้อยละ 88.71และ 87.25 ตามลำดับ

Table 1 The descriptive statistics of water quality classification (n= 1,709).

Variables (Unit)	Minimum	Maximum	Mean	Standard Deviation
temp A (^{O}C)	16.00	45.00	31.31	3.54
temp W (^{O}C)	17.90	39.0	28.97	2.36
pH	5.60	9.24	7.60	0.52
Tur (NTU)	0.13	1,615.0	59.51	122.54
Cond (µS)	11.80	50,835.00	1,408.78	5,311.45
DO (mg/l)	0.03	13.90	5.62	1.75
BOD (mg/l)	0.10	75.00	1.85	2.26
Total Coli (MPN/100ml)	20.00	16,000.00	8,833.53	15,215.26
Fecal Coli (MPN/100ml)	2.00	92,000.00	2,553.51	6,252.77
TP (mg/l)	0.01	2.40	0.21	0.26
NO3-N (mg/l)	0.01	4.86	0.30	0.33
NH3-N (mg/l)	0.02	11.50	0.31	0.61
SS (mg/l)	0.40	1,415.00	42.20	95.52
TS (mg/l)	18.00	36,635.00	726.14	2,979.52

Table 2 The binary logistic regression result of water quality classification.

Variables	B	S.E.	Z	p-value
Tur (NTU)	-0.003	0.001	-3.023	0.003
Total Coli (MPN/100ml)	0.000	0.000	9.095	0.003
NH3-N (mg/l)	-0.409	0.139	-2.944	<0.001
Fecal Coli (MPN/100ml)	0.000	0.000	8.313	<0.001
DO (mg/l)	-0.476	0.050	-9.477	<0.001
BOD (mg/l)	1.886	0.115	16.417	<0.001
Intercept	-2.738	0.338	-8.108	<0.001
$\chi^2 = 2.387$ p – value 0.967	-2log likelihood = 487.353	Cox & Snell R^2 = 0.600		

Table 3 The classification performance.

Methods	Accuracy
Binary Logistic Regression	87.25%
Decision tree	88.71%
Decision tree with Binary Logistic Regression	89.64%
K-Nearest Neighbors (K=22)	79.05%
K-Nearest Neighbors with Binary Logistic Regression (K=22)	79.70%

4. สรุปผลการวิจัย

จากการศึกษาและรวบรวมข้อมูลคุณภาพของน้ำจากแหล่งน้ำทั่วประเทศไทยตั้งแต่ วันที่ 1 มกราคม พ.ศ. 2561 ถึง วันที่ 1 มกราคม พ.ศ. 2564 จำนวน 1,709 ชุด พบว่า คุณภาพน้ำไม่ได้ตามมาตรฐาน 570 ชุด และคุณภาพน้ำได้ตามมาตรฐานจำนวน 1,139 ชุด แบ่งข้อมูลแต่ละเทคนิคด้วยวิธีการตรวจสอบไขว้ กำหนด k=10 พบว่าที่ระดับนัยสำคัญ 0.05

ตัวแปรอิสระที่มีผลต่อการจำแนกคุณภาพน้ำโดยการวิเคราะห์การถดถอยลอจิสติกทวิภาค มีจำนวน 6 ตัวแปร ได้แก่ ความขุ่นของน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แอมโมเนีย-ไนโตรเจน แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ออกซิเจนที่ละลายในน้ำ สารอินทรีย์ในน้ำ มีค่าความแม่นร้อยละ 87.25 ซึ่งสอดคล้องกับงานวิจัยการประเมินค่าคุณภาพน้ำของแม่น้ำบางปะกง [6] ตัวแปรอิสระที่มีผลต่อการจำแนกคุณภาพน้ำประกอบไปด้วยตัวแปร 5 ตัวแปร ได้แก่ ออกซิเจนที่ละลายในน้ำ น้ำที่มีสารอินทรีย์ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม และแอมโมเนีย – ไนโตรเจน มีค่าความแม่นร้อยละ 85.76 เมื่อทำการวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจโดยการทุกตัวแปรมาพิจารณาค่าความแม่นร้อยละ 88.71 สอดคล้องกับการศึกษาของ การวิเคราะห์คุณภาพน้ำโดยใช้ต้นไม้การตัดสินใจเพื่อวิเคราะห์ข้อมูลคุณภาพน้ำในประเทศเคนยา [4] ซึ่งต้นไม้ตัดสินใจมีค่าความแม่นร้อยละ 94 การวิเคราะห์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดของตัวแปรตามตัวแบบลอจิสติกทวิภาค จำนวนค่า K ที่ทำให้ค่าความแม่นมากที่สุด คือ ค่า K = 22 ค่าความแม่นเท่ากับร้อยละ 79.70 การวิเคราะห์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดของข้อมูลทุกตัวแปร ค่า K = 22 ค่าความแม่นร้อยละ 79.05 สอดคล้องกับการศึกษาการจำแนกคุณภาพน้ำในประเทศอินโดนีเซีย [4] โดยใช้เทคนิคการจำแนกประเภทการทำเหมืองข้อมูลเพื่อนบ้านใกล้ที่สุดแบบปรับปรุง (Modified K-Nearest) ที่มีความแม่นร้อยละ 85.10

การเปรียบเทียบประสิทธิภาพของการจำแนกคุณภาพน้ำโดยใช้ค่าความแม่นเป็นเกณฑ์ พบว่า การวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจโดยการนำตัวแปรที่ได้จากการวิเคราะห์การถดถอยลอจิสติกทวิภาคมาพิจารณา ซึ่งประกอบไปด้วยตัวแปร 5 ตัวแปร ได้แก่ ความขุ่นของน้ำ แบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด แอมโมเนีย-ไนโตรเจน แบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม ออกซิเจนที่ละลายในน้ำ สารอินทรีย์ในน้ำ มีค่าความแม่นร้อยละ 89.64 มากที่สุด รองลงมาเป็น การวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจโดยการทุกตัวแปรมาพิจารณาค่าความแม่นร้อยละ 88.71 และการวิเคราะห์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดของตัวแปรตามตัวแบบลอจิสติกทวิภาค จำนวนค่า K ที่ทำให้ค่าความแม่นมากที่สุด คือ ค่า K = 22 ค่าความแม่นเท่ากับร้อยละ 79.70 และการวิเคราะห์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดของข้อมูลทุกตัวแปร ค่า K = 22 ค่าความแม่นน้อยที่สุดร้อยละ 79.05

อย่างไรก็ตามจากการศึกษาในครั้งนี้ พบว่ามีบางตัวแปรที่ใช้ในการศึกษามีค่าส่วนเบี่ยงเบนมาตรฐานสูง

กว่าค่าเฉลี่ยเนื่องจากข้อมูลของตัวแปรดังกล่าวมีความแตกต่างกันมาก ในการศึกษาครั้งต่อไปอาจมีการตรวจสอบความผิดปกติของข้อมูล ศึกษาหาสาเหตุของความผิดปกติของข้อมูล และในการศึกษาครั้งนี้ข้อมูลที่ใช้วิเคราะห์เป็นคุณภาพน้ำไม่ได้ตามมาตรฐาน 570 ชุด และคุณภาพน้ำได้ตามมาตรฐานจำนวน 1,139 ซึ่งเป็นข้อมูลที่ไม่สมดุล (Imbalance data) ในการศึกษาครั้งต่อไปอาจศึกษาประสิทธิภาพของการจัดการข้อมูลให้มีความสมดุล เช่น วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ เป็นต้น เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกประเภทของน้ำ นอกจากนี้อาจทำกลุ่มของคุณภาพใหม่ให้มากกว่า 2 กลุ่มและใช้การวิเคราะห์การถดถอยลอจิสติกเชิงพหุ (Multinomial Logistic Regression Analysis) รวมถึงการศึกษาวิธีต้นไม้ตัดสินใจที่มีหลากหลาย เช่น ต้นไม้ตัดสินใจตอไม้ (Decision Stump) ป่าสุ่ม (Random Forest) โครงข่ายประสาทเทียม เป็นต้น

5. กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ สำนักจัดการคุณภาพน้ำ กรมควบคุมมลพิษ สำหรับการอนุเคราะห์ข้อมูลในการวิจัย

6. References

[1] Pollution Control Department. Thailand State of Pollution Report 2021, Available Source: https://www.pcd.go.th/publication/26626, April 2, 2022. (in Thai)

[2] Pollution Control Department. Total score of water quality 5 parameters, Available Source: http://iwis.pcd.go.th/module/wqi_calculate/wqi.pdf, February 2, 2022. (in Thai)

[3] Gakii, C., & Jepkoech, J. (2019). A classification model for water quality analysis using decision tree. *European Journal of Computer Science and Information Technology*, vol. 7(3), 1-8.

[4] Ramadhani, D., Afdal, M., & Rahmawita, M. (2021, February). The Classification Status of River Water Quality in Riau Province Using Modified K-Nearest Neighbor Algorithm with STORET Modeling and Water Pollution Index. In *Journal of Physics: Conference Series* (Vol. 1783, No. 1, p. 012020). IOP Publishing.

[5] Wechmongkhonkon, S., Poomtong, N., & Areerachakul, S. (2012). Application of artificial neural network to classification surface water quality. *World Academy of Science, Engineering and Technology*, 6(9), 574-578.

[6] Najah, A., Elshafie, A., Karim, O. A., & Jaffar, O. (2009). Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of scientific research, 28*(3), 422-435.

[7] Diamantopoulou, M. J., Papamichail, D. M., & Antonopoulos, V. Z. (2005). The use of a neural network technique for the prediction of water quality parameters. *Operational Research*, 5(1), 115-125.

[8] Areerachakul, S., & Sanguansintukul, S. (2009, November). Water quality classification using neural networks: Case study of canals in Bangkok, Thailand. In 2009 *International Conference for Internet Technology and Secured Transactions, (ICITST)* (pp. 1-5). IEEE.

[9] Pudchaya, S. (2017). Assessment of water quality of Bang Pakong River using multivariate analysis. *Burapha Science Journal*. Vol. 22 No. 2. 183-193. (in Thai)

[10] Saichon, S.(2016), *Multivariate analysis.* Bangkok. Chamchuree Products Co., Ltd. (in Thai)

[11] Anupong, S. (2021). *Data Mining Guide with RapidMiner Studio.* Mahasarakham Business School, Mahasarakham University, Mahasarakham. (in Thai)