**การพยากรณ์น้ำที่สามารถดื่มได้โดยใช้การเรียนรู้ของเครื่อง**

**The Prediction of Drinking Water Potability using Machine Learning**

**โดย**

**นาย****คมสัน เกตุแก้ว         รหัส** **1630702650**

**นายธีรภพ ตันจิรวัฒนา      รหัส** **1630703039**

**นางสาว****ทิพย์มินตรา ชมชื่น รหัส** **1630706149**

**อาจารย์ที่ปรึกษา**

**ดร.สรพรรค ภักดีศรี**

**หลักสูตรวิทยาศาสตรบัณฑิต**

**ภาคการศึกษาที่ 2 ปีการศึกษา 2565**

**ภาควิชาวิทยาการคอมพิวเตอร์ คณะเทคโนโลยีสารสนเทศและนวัตกรรม**

**มหาวิทยาลัยกรุงเทพ**

**บทที่ 1**

**บทนำ**

## **ความเป็นมาของโครงการ**

น้ำเป็นปัยจัยสำคัญในการใช้ชีวิตของมนุษย์ ในปัจจุบันมนุษย์ได้ให้ความสำคัญกับน้ำที่สะอาด ปลอดภัย โดยองค์กรอนามัยโลก (WHO) เป็นทบวงการชำนาญพิเศษของสหประชาชาติ ซึ่งรับผิดชอบการประสานงานด้านสาธารณสุขระหว่างประเทศ ได้มีการระบุว่าการใช้น้ำที่ปนเปื้อนและการสุขภิบาลไม่ดีเป็นแหล่งกระจายของโรคต่าง ๆ เช่น อหิวาตกโรค ท้องร่วง ตับอักเสบ ไทฟอยด์ และโปลิโอ ซึ่งเกิดจากคุณภาพของน้ำและสุขอนามัยที่มีการจัดการที่ไม่ดีพอทำให้บุคคลมีความเสี่ยงต่อสุขภาพ จึงได้มีการกำหนดมาตรฐานการผลิตน้ำที่สะอาดเพื่อจัดการและควบคุมคุณภาพของน้ำ โดยในโครงงานนี้จะนำเอาแมชชีนเลิร์นนิ่ง (Machine learning) มาใช้ในการประเมินคุณภาพของน้ำ

แมชชีนเลิร์นนิ่งเป็นปัญญาประดิษฐ์ประเภทหนึ่งที่ช่วยให้คอมพิวเตอร์เรียนรู้และปรับปรุงประสิทธิภาพได้โดยไม่ต้องตั้งโปรแกรมไว้อย่างชัดเจน สามารถใช้ในการพยากรณ์ข้อมูลและคาดการณ์หรือตัดสินใจจากข้อมูลในการตรวจสอบคุณภาพของน้ำ อัลกอริทึมการเรียนรู้ของเครื่องสามารถฝึกฝนกับข้อมูลค่า Water Quality ของ ADITYA KADIWAL จากเว็บไซต์ Kaggle มาใช้สำหรับเทรนแมชชีนเลิร์นนิ่งโมเดล เนื่องจากคุณภาพของน้ำเกิดจากการวัดค่าของน้ำและนำมาเปรียบเทียบกับเกณฑ์จึงสามารถทราบได้ว่าน้ำนั้นสามารถดื่มได้หรือไม่

วัตถุประสงค์หลักของโครงงานนี้คือการสร้างและประเมินผลโมเดลแมชชีนเลิร์นนิ่ง ซึ่งโมเดลแมชชีนเลิร์นนิ่งที่ผ่านการเรียนรู้มาแล้วสามารถใช้ทำนายว่าค่าของน้ำที่ตรวจสอบมานั้นสามารถดื่มได้หรือไม่ โดยอ้างอิงจากข้อมูลที่ผ่านการตรวจสอบและผ่านการยืนยันว่าสามารถดื่มได้ ซึ่งจะเป็นประโยชน์ในด้านการตรวจสอบและด้านการพยากรณ์ โดยจะสามารถทราบได้ว่าน้ำนั้นมีคุณภาพมากพอที่จะดื่มได้หรือมีคุณภาพไม่ดีจนไม่สามารถดื่มได้ อย่างไรก็ตามสิ่งสำคัญที่ควรทราบคือการเรียนรู้ของเครื่องเป็นเพียงเครื่องมือหนึ่งที่สามารถใช้ในการตรวจวัดคุณภาพน้ำเบื้องต้นได้ ดังนั้น จึงควรใช้ร่วมกับวิธีการอื่น ๆ จากผู้เชี่ยวชาญของการวัดคุณภาพของน้ำ

## **วัตถุประสงค์ของโครงการ**

1. เพื่อพัฒนากรอบแนวคิดการเรียนรู้ของเครื่องมาประยุกต์ใช้กับข้อมูลคุณภาพของน้ำเป็นตัวช่วยในการประเมิณคุณภาพของน้ำเบื้องต้น
2. เพื่อพัฒนาโมเดลแมชชีนเลิร์นนิ่งที่ผ่านการเรียนรู้ข้อมูลคุณภาพของน้ำที่ผ่านการตรวจสอบแล้ว
3. เพื่อประเมินผลลัพธ์โมเดลแมชชีนเลิร์นนิ่งที่ผ่านการเรียนรู้ข้อมูลคุณภาพของน้ำที่ผ่านการตรวจสอบแล้ว
4. เพื่อพัฒนาเว็ปแอปพลิเคชัน (Responsive Web-based application) ที่สามารถแสดงผลได้บนอุปกรณ์ที่มีพื้นที่การแสดงผลที่หลากหลาย มีความสามารถในการรับข้อมูลที่ใช้ในการพยากรณ์ และแสดงผลการประเมิณคุณภาพของน้ำรวมถึงการป้อนข้อมูลเข้ามาในโมเดลโมเดลแมชชีนเลิร์นนิ่ง

## **ขอบเขตของโครงการ**

พัฒนาเว็ปแอปพลิเคชัน (Responsive Web-based application) ที่สามารถทำนายคุณภาพของน้ำโดยแสดงผลเป็นสามารถดื่มได้และดื่มไม่ได้ โดยแบ่งการทำงานออกเป็น 2 ส่วนดังนี้

1. ผู้พัฒนาระบบ
   1. นำ Dataset ที่ได้มาทำการแทนที่ในส่วนที่ไม่สมบูรณ์
   2. ทำการแบ่งข้อมูลออกเป็น 2 ส่วนโดยใช้ K-Fold Cross Validation คือ การที่เราแบ่งข้อมูลเป็นจำนวน K ส่วน โดยในแต่ละส่วนจะต้องมาจากการสุ่มเพื่อที่จะให้ข้อมูลกระจายเท่า ๆ กัน
   3. ทำการเทรนโมเดลโดยการนำ Training dataset เข้าสู่กระบวนการเรียนรู้ เพื่อสร้าง model โดยใช้อัลกอรึทึม Logistic Regression, Support Vector Machines, Random Forest และ Decision Tree โดยแต่ละอัลกอริทึมจะสร้างโมเดลของตัวเอง
   4. ประเมินผลประสิทธิภาพของ model มาทำการเปรียบเทียบเพื่อเลือกโมเดลที่ดีที่สุดสำหรับ dataset
   5. พัฒนาเว็ปแอปพลิเคชัน โดยใช่โมเดลที่ให้ความแม่นยำในการทำนายมากที่สุด
2. ผู้พัฒนาระบบ
   1. นำ Dataset ที่ได้มาทำการแทนที่ในส่วนที่ไม่สมบูรณ์
   2. ทำการแบ่งข้อมูลออกเป็น 2 ส่วนโดยใช้ K-Fold Cross Validation คือ การที่เราแบ่งข้อมูลเป็นจำนวน K ส่วน โดยในแต่ละส่วนจะต้องมาจากการสุ่มเพื่อที่จะให้ข้อมูลกระจายเท่า ๆ กัน
   3. ทำการเทรนโมเดลโดยการนำ Training dataset เข้าสู่กระบวนการเรียนรู้ เพื่อสร้าง model โดยใช้อัลกอรึทึม Logistic Regression, Support Vector Machines, Random Forest และ Decision Tree โดยแต่ละอัลกอริทึมจะสร้างโมเดลของตัวเอง
   4. ประเมินผลประสิทธิภาพของ model มาทำการเปรียบเทียบเพื่อเลือกโมเดลที่ดีที่สุดสำหรับ dataset
   5. พัฒนาเว็ปแอปพลิเคชัน โดยใช่โมเดลที่ให้ความแม่นยำในการทำนายมากที่สุด
3. ผู้พัฒนาระบบ
   1. นำ Dataset ที่ได้มาทำการแทนที่ในส่วนที่ไม่สมบูรณ์
   2. ทำการแบ่งข้อมูลออกเป็น 2 ส่วนโดยใช้ K-Fold Cross Validation คือ การที่เราแบ่งข้อมูลเป็นจำนวน K ส่วน โดยในแต่ละส่วนจะต้องมาจากการสุ่มเพื่อที่จะให้ข้อมูลกระจายเท่า ๆ กัน
   3. ทำการเทรนโมเดลโดยการนำ Training dataset เข้าสู่กระบวนการเรียนรู้ เพื่อสร้าง model โดยใช้อัลกอรึทึม Logistic Regression, Support Vector Machines, Random Forest และ Decision Tree โดยแต่ละอัลกอริทึมจะสร้างโมเดลของตัวเอง
   4. ประเมินผลประสิทธิภาพของ model มาทำการเปรียบเทียบเพื่อเลือกโมเดลที่ดีที่สุดสำหรับ dataset
   5. พัฒนาเว็ปแอปพลิเคชัน โดยใช่โมเดลที่ให้ความแม่นยำในการทำนายมากที่สุด
4. ผู้ใช้
   1. การกรอกข้อมูลของตัวแปร 9 ตัวแปรดังต่อไปนี้
      * pH
      * Hardness
      * Solids (TDS)
      * Chloramines
      * Sulfate
      * Conductivity
      * Organic Carbon
      * Trihalomethanes
      * Turbidity
   2. การทำนายผลคุณภาพของน้ำโดยนำตัวแปรทั้ง 9 มาทำนายโดยจะแสดงผลเป็นข้อมูลดังต่อไปนี้
      * สามารถดื่มได้ หรือ ดื่มไม่ได้
      * เปอร์เซ็นการดื่มได้ และ ดื่มไม่ได้
      * มีตัวแปรไดที่ส่งผลให้ไม่สามารถดื่มได้

## **ซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ (Implementation Software)**

ซอฟต์แวร์ที่นำในใช้ในการพัฒนาระบบ แบ่งออกเป็น 2 ส่วน ดังนี้

1. ซอฟต์แวร์ที่นำมาใช้ในการพัฒนาระบบ
   1. ภาษา Python ใช้ในการพัฒนาระบบ Machine Leanings และ Web-based application
   2. Visual Studio Code ใช้ในการเขียนโค้ด
2. ซอฟต์แวร์ที่นำมาใช้ในการพัฒนาระบบส่วนอื่น ๆ
   1. Terminal ในการติดตั้ง Script ต่าง ๆ
   2. Microsoft Word ในการจัดทำเล่มรายงานนำเสนอโครงการ
   3. Lucidchart ในการสร้างแผนภาพ Data flow diagram
   4. GitHub ในการจัดเก็บ และแชร์ไฟล์ Source code ในการทำงานร่วมกัน
   5. Microsoft Team ในการสนทนา และประชุมงาน
   6. Google Chrome ในการทดสอบและแสดงผลของแอพพลิเคชั่น

## **ประโยชน์ที่ได้รับ**

1. ช่วยในเรื่องการทดสอบคุณภาพน้ำเบื้องต้น
2. ช่วยในการประเมินเพื่อช่วยลดภาระงานแก่การแยกคุณภาพน้ำโดยผู้เชี่ยวชาญ
3. เพิ่มความรวดเร็วในการแยกแยะข้อมูลคุณภาพน้ำที่สามารถดื่มได้หรือดื่มไม่ได้
4. ช่วยเพิ่มโอกาสในการเข้าถึงน้ำดื่มของพื้นที่ขาดแคลนน้ำเป็นไปได้มากขึ้น

## **ระยะเวลาในการทำโครงการ**

**ตารางที่ 1.1** ตารางระยะเวลาดำเนินงาน

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ขั้นตอนการดำเนินงาน** | **ระยะเวลาการดำเนินการ** | | | | | | | | | | |
| **2566** | | | | | | **2566** | | | | |
| **พ.ย.** | **ธ.ค.** | **ม.ค.** | **ก.พ.** | **มี.ค.** | **เม.ย.** | **พ.ค.** | **มิ.ย.** | **ก.ค.** | **ส.ค.** | **ก.ย.** |
| 1. เสนอหัวข้อโครงงาน |  |  |  |  |  |  |  |  |  |  |  |
| 1. ศึกษาและรวบรวมข้อมูลเพื่อนำมาเป็นแนวทางในการพัฒนา |  |  |  |  |  |  |  |  |  |  |  |
| 1. ออกแบบส่วนการติดต่อกับผู้ใช้ |  |  |  |  |  |  |  |  |  |  |  |
| 1. เขียนโปรแกรมในแต่ละส่วนของการทำงาน |  |  |  |  |  |  |  |  |  |  |  |
| 1. ทดสอบ แก้ไขและปรับปรุงระบบ |  |  |  |  |  |  |  |  |  |  |  |
| 1. ทดสอบการติดตั้ง |  |  |  |  |  |  |  |  |  |  |  |
| 1. จัดทำเอกสารคู่มือการใช้งาน |  |  |  |  |  |  |  |  |  |  |  |

**บทที่ 2**

**ทฤษฎีและความรู้ต่าง ๆ ที่ใช้ประกอบในการทำโครงงาน**

ในการทำวิจัยครั้งนี้ ผู้พัฒนาได้ศึกษาเอกสารและข้อมูลที่เกี่ยวข้องและได้นำเสนอตามหัวข้อเรียงตามลำดับดังนี้

1. ค่าที่ใช้วัดคุณภาพของน้ำดื่มและเครื่องมือที่ใช้ในการวัด
2. ความสำคัญของน้ำ
3. สารปนเปื้อนที่มาจากสารเคมีในน้ำ
4. วิทยาการข้อมูล (Data Science)
5. ปัญญาประดิษฐ์ และ การเรียนรู้ของเครื่อง (AI and Machine Learning)
6. งานวิจัยที่เกี่ยวข้อง
7. ขั้นตอนการตรวจน้ำดื่ม หรือ น้ำบริโภค
8. **ค่าที่ใช้วัดคุณภาพของน้ำดื่ม**
9. ค่า pH หรือค่าความเป็นกรดและด่าง

ค่า pH ของสารละลายคือค่าลอการิทึมทั่วไปที่เป็นค่าลบของไฮโดรเจนไอออน : pH = −log (H+) ซึ่งเป็นตัววัดค่าความเข้มข้นของไฮโดรเจนไอออน ใช้บอกความเป็นกรดหรือด่างที่อยู่ในน้ำ โดยมีค่าจะอยู่ในช่วงตั้งแต่ 0 จนถึง 14 ระดับที่ค่า pH มีความเป็นกลางจะอยู่ในช่วง pH = 6.5 – 8.5 แต่น้ำที่มีก๊าซคาร์บอนไดออกไซด์อยู่จะมีค่า pH ที่ต่ำกว่า 5 ส่วนน้ำที่มีคาร์บอเนตอยู่จะมีค่า pH สูงกว่า 9 โดยค่า pH ที่เหมาะสมมีการอ้างอิงมาจาก World Health Organization

ค่า pH ที่มีความสะอาดและเหมาะสมในการดื่มมีค่า pH = 7 หรือ pH ที่มีค่าเป็นกลาง เนื่องจากไม่มีคุณสมบัติการเป็นกรดหรือด่าง จึงไม่มีผลเสียต่อร่างกาย ในกรณีที่ค่า pH มีค่าต่ำกว่า 7 ลงไปสามารถมีค่าเป็นกรดได้ โดยที่สารที่มีค่าเป็นกรดสูงสามารถมีค่า pH เป็น 0 ได้ ซึ่งแบตเตอรี่จัดอยู่ในประเภทนี้ และในกรณีที่ค่า pH มีค่ามากกว่า 8 ขึ้นไปสามารถมีค่าเป็นด่างได้ โดยสารที่มีค่าเป็นด่างรุนแรง อย่างเช่น น้ำด่างสามารถมีค่า pH เท่ากับ 14 ได้ ซึ่งค่า pH ของน้ำจืดทั่วโลกจะมีความแตกต่างกันตามสภาพแวดล้อม สภาพภูมิอากาศ กิจกรรมต่าง ๆ ของมนุษย์ และกระบวนการทาง ธรรมชาติ โดยน้ำที่มีค่า pH ต่ำหรือสูง

อาจจะเป็นสัญญาณของมลพิษจากสารเคมีหรือโลหะหนักได้ และน้ำที่ค่าความเป็นด่างสูงอาจทำให้น้ำมีกลิ่นหรือรสชาติที่ไม่พึงประสงค์และอาจทำให้ท่อที่ลำเลียงน้ำมาเกิดการเสียหายได้ ดังนั้นน้ำที่มีค่าความเป็นกรดมีแนวโน้มที่จะปนเปื้อนสารพิษได้ทำให้ไม่ปลอดภัยต่อการบริโภคเพราะมันสามารถกัดกร่อนท่อโลหะได้ (Butler. 2019: ออนไลน์) (World Health Organization. 2022: 245)

1. ค่า Hardness หรือ ค่าความกระด้าง

Hardness หรือค่าความกระด้างคือ เมื่อมีแคลเซียมและแมกนีเซียมละลายอยู่ในน้ำจะทำให้น้ำไม่เกิดฟองกับสบู่และจะเกิดการตกตะกอน ความกระด้างในน้ำที่เกิดตามธรรมชาติส่วนมากจะมาจากโลหะหลายชนิดที่ละลายอยู่ในน้ำ ซึ่งส่วนใหญ่เป็นไอออนบวกของแคลเซียมและแมกนีเซียม โดยในบางกรณีผู้บริโภคนั้นสามารถทนต่อความกระด้างของน้ำได้เกินกว่า 500 มิลลิลิตร/ลิตร ร่วมกับปัจจัยอย่างพวกค่า pH และมีเกณฑ์ที่ยอมรับจาก World Health Organization คือ 100-300 มิลลิกรัม/ลิตร โดยความกระด้างแบ่งออกเป็น 2 ส่วน ได้แก่

* Temporary Hardness หรือความกระด้างชั่วคราว เกิดจากสารละลายของแคลเซียมหรือแมกนีเซียมไบคาร์บอเนต เมื่อโดนความร้อนจะตกตะกอน
* Permanent Hardness หรือความกระด้างถาวร เกิดจากเกลือคลอไรด์หรือซัลเฟตของแคลเซียมหรือแมกนีเซียม ที่ไม่สามารถตกตะกอนเมื่อโดนความร้อน

ระดับความกระด้างของน้ำ

* ความกระด้างที่ 0 – 75 มล./ล. น้ำมีความกระด้างอ่อน
* ความกระด้างที่ 75 – 150 มล./ล. น้ำมีความกระด้างปานกลาง
* ความกระด้างที่ 150 – 300 มล./ล. น้ำกระด้าง
* ความกระด้างที่ 300 ขึ้นไป มล./ล. น้ำมีความกระด้างสูง

ความกระด้างของน้ำนั้นไม่มีผลต่อสุขภาพในเชิงลบ โดยการดื่มน้ำที่มีความกระด้างนั้นก็เท่ากับการดื่มน้ำที่มีแคลเซียมและแมกนีเซียมปนอยู่ด้วยซึ่งธาตุทั้ง 2 มีผลกับร่างกายดังนี้

* แคลเซียมช่วยในการหดตัวของกล้ามเนื้อโดยเฉพาะกล้ามเนื้อหัวใจ ซึ่งจะเป็นตัวเร่งการทำงานของเอนไซม์ในร่างกายช่วยป้องกันโรคกระดูกพรุนและรักษาความสมดุลของความด่างในร่างกาย
* แมกนีเซียมช่วยในการคลายตัวของกล้ามเนื้อโดยเฉพาะกล้ามเนื้อหัวใจ ซึ่งจะทำงานร่วมกับแคลเซียมเป็นตัวเร่งเอนไซม์ที่มีหน้าที่เผาผลาญอาหาร เป็นองค์ประกอบของเลซิ-ตินป้องกันคอลเลสเตอรอลในเลือดสูงขึ้น อีกทั้งยังป้องกันการเกิดนิ่วได้อีกด้วย (World Health Organization. 2010: 3) (ศูนย์ห้องปฏิบัติการกรมอนามัย. 2563: ออนไลน์)

1. ค่า Solids (TDS) หรือค่าของแข็งละลายน้ำ

Solids (TDS) คือ การวัดปริมาณของแข็ง ซึ่งเป็นคำอธิบายของเกลืออนินทรีย์ที่มีขนาดเล็ก ปริมาณสารอินทรีย์ที่อยู่ในสารละลายที่เป็นน้ำจะมีส่วนประกอบหลักคือ แคลเซียม แมกนีเซียม และโซเดียม Solids (TDS) นั้นจะไม่ถือว่าเป็นการวัดค่ามลพิษแต่เป็นตัวบ่งชี้ถึงลักษณะคุณภาพของน้ำดื่มและบ่งชี้ถึงสารปนเปื้อนในน้ำ โดยค่าที่แสดงออกมาจะแสดงเป็น มิลลิกรัม/ลิตร หรือในส่วนต่อล้านจะมีค่าเป็น ppm อย่างไรก็ตามก็มีบางรายงานกล่าวถึงความสัมพันธ์แบบผกผันระหว่างความเข้มข้นของ TDS ในน้ำดื่มและอุบัติการณ์ของมะเร็ง โรคหลอดเลือดหัวใจ หัวใจหลอดเลือดแดงแข็งโรค และโรคหัวใจและหลอดเลือด โดยมีอัตราการเสียชีวิตทั้งหมด ในรายงานสรุปผลของการศึกษาในออสเตรเลียพบว่าการตายจากทุกประเภทของโรคหัวใจ มักมีอาการขาดเลือดและมีอาการกล้ามเนื้อหัวใจตายเฉียบพลันเพิ่มขึ้นในชุมชนด้วยปริมาณของแข็งที่ละลายน้ำได้สูง (World Health Organization .2022: 246)

Solids (TDS) จะส่งผลต่อรสชาติของน้ำ โดยจะมีระดับดังนี้

* ดีเยี่ยม จะอยู่ที่น้อยกว่า 300 มิลลิกรัม/ลิตร
* ดี จะอยู่ที่ระหว่าง 300 – 600 มิลลิกรัม/ลิตร
* พอใช้ จะอยู่ที่ระหว่าง 600 – 900 มิลลิกรัม/ลิตร
* ไม่ดี จะอยู่ที่ระหว่าง 900 – 1200 มิลลิกรัม/ลิตร
* ยอมรับไม่ได้ จะอยู่ที่ 1200 มิลลิกรัม/ลิตร ขึ้นไป

1. ค่า chloramines

Monochloramine หรือที่มักจะเรียกกันสั้นๆว่า Chloramines เป็นสารประกอบทางเคมี ร่วมกับ dichloramine และไนโตรเจนไตรคลอไรด์ เกิดจากปฏิกิริยาของคลอรีนกับแอมโมเนีย โดยที่โมโนคลอรามีนเป็นคลอรีนฆ่าเชื้อที่มีประโยชน์เพียงชนิดเดียว และระบบคลอรามีนทำงานเพื่อลดการก่อตัวของไดคลอรามีนและไตรคลอรามีน คลอรามีนมักพบในน้ำดื่มเป็นส่วนใหญ่ ซึ่งเป็นผลจากการบำบัด ทั้งโดยตั้งใจให้เป็นสารฆ่าเชื้อในระบบจำหน่าย และโดยไม่ได้ตั้งใจเป็นผลพลอยได้จากคลอรีนในน้ำดื่มที่มีแอมโมเนียตามธรรมชาติ เนื่องจากโมโนคลอรามีนมีความเสถียรมากกว่าและให้การฆ่าเชื้อที่ยาวนานกว่า จึงเป็นที่นิยมใช้ในระบบจ่ายน้ำ คลอรามีนที่สูงขึ้นโดยเฉพาะไตรคลอรามีนมีแนวโน้มที่จะก่อให้เกิดรสชาติและกลิ่นที่ความเข้มข้นต่ำมาก สำหรับโมโนคลอรามีนไม่พบกลิ่นหรือรสที่ความเข้มข้นระหว่าง 0.5 และ 1.5 มิลลิกรัม/ลิตร และสำหรับไดคลอรามีนจะมีผลต่อประสาทสัมผัสอยู่ที่ 0.1 – 0.5 มิลลิกรัม/ลิตร ซึ่งถือว่าอยู่ในขั้นที่เล็กน้อยและยอมรับได้ (World Health Organization. 2022: 223) (Health of Canada. 2018: online)

1. Sulfate

ซัลเฟตนั้นจะเกิดขึ้นตามธรรมชาติในแร่ธาตุต่าง ๆ การที่มีซัลเฟตในน้ำดื่มสามารถทำให้เกิดรสชาติแปลก ๆ ในน้ำได้ ในกรณีที่มีปริมาณที่มากอาจทำให้เกิดฤทธิ์เป็นยาระบายได้ในหมู่ผู้ที่ไม่คุ้นเคย เกณฑ์การรับรสสำหรับโซเดียมซัลเฟตจะอยู่ที่ปริมาณตั้งแต่ 250 มิลลิกรัม/ลิตร ถึง 1,000 มิลลิกรัม/ลิตร (World Health Organization. 2004: 4)

* ความเข้มข้นของเกณฑ์การรับรสชาติในน้ำดื่มคือ 250–500 มิลลิกรัม/ลิตร มีค่ามัธยฐานอยู่ที่ 350 มิลลิกรัม/ลิตร
* โซเดียมซัลเฟต 250–1000 มิลลิกรัม/ลิตร มีค่ามัธยฐานอยู่ที่ 525 มิลลิกรัม/ลิตร
* แมกนีเซียมซัลเฟต 400-600 มิลลิกรัม/ลิตร มีค่ามัธยฐานอยู่ที่ 525 มิลลิกรัม/ลิตร

1. ค่า Conductivity หรือ ค่าการนำไฟฟ้า

Conductivity คือวิธีที่ใช้วัดค่าปริมาณไอออนิก (เช่น คลอไรด์ ไนเตรต ซัลเฟต โซเดียม แมกนีเซียม แคลเซียม หรือเหล็ก) ของน้ำโดยการวัดค่าการนำไฟฟ้าของน้ำ เนื่องจากน้ำที่มีความบริสุทธิ์จะเป็นตัวนำไฟฟ้าได้ไม่ดี เพราะน้ำปกติจะมีสิ่งเจือปนจำพวกจากไอออนิกอยู่ และไอออนเหล่านี้ก็คือตัวนำไฟฟ้าในน้ำเนื่องจากกระแสไฟฟ้าในน้ำถูกส่งโดยไอออนที่มีอยู่ในน้ำ ค่าการนำไฟฟ้าสามารถใช้พารามิเตอร์วัดคุณภาพน้ำได้ สามารถใช้ตรวจสอบความเค็มของน้ำระหว่างการบำบัดน้ำได้โดยค่าของ Conductivity ในการนำไฟฟ้าคือซีเมนส์ต่อเมตร (S/m) ถึงแม้ว่าหน่วยมาตรฐาน SI คือซีเมนส์ต่อเมตร (S/m) แต่ในทางปฎิบัติค่าของ Conductivity ในน้ำไม่สูงมากนัก จึงใช้หน่วยเป็นไมโครซีเมนส์/เซ็นติเมตร (µS/cm) และ mS/cm หรือมิลลิซีเมนส์ต่อเซนติเมตร (BYJU'S. 2023: online)

ค่าทั่วไปของการนำไฟฟ้า

* น้ำประปา 50 – 800 ไมโครซีเมนส์/เซนติเมตร
* น้ำดื่ม 30 – 1500 ไมโครซีเมนส์/เซนติเมตร
* ลำธารน้ำจืด 100 – 1,000 ไมโครซีเมนส์/เซนติเมตร

1. ค่า Organic Carbon หรือค่าคาร์บอนอินทรีย์

Total Organic Carbon (TOC) คือการวัดปริมาณคาร์บอนทั้งหมดในสารประกอบอินทรีย์ในระบบน้ำ โดยน้ำจะมีสารประกอบอินทรีย์ตามธรรมชาติอยู่แล้ว เว้นแต่ว่าน้ำจะมีความบริสุทธิ์มาก

TOC ได้กลายเป็นตัวแปรสำคัญที่ใช้ตรวจสอบระดับโดยรวมของสารประกอบอินทรีย์ที่มีอยู่ สิ่งนี้จะเกิดขึ้นได้แม้ว่าจะไม่มีความสัมพันธ์เชิงปริมาณโดยตรงระหว่างคาร์บอนอินทรีย์ทั้งหมดและความเข้มข้นรวมของสารประกอบอินทรีย์ที่มีอยู่ สะท้อนให้เห็นถึงความสำคัญของการมีตัวบ่งชี้ทั่วไปที่ง่ายต่อการวัดระดับการปนเปื้อนอินทรีย์

Total Organic Carbon วัดอะไรบ้าง

* TC – คาร์บอนทั้งหมด
* TIC – คาร์บอนอนินทรีย์ทั้งหมด
* POC – คาร์บอนอินทรีย์ที่สามารถล้างทำความสะอาดได้
* NPOC – คาร์บอนอินทรีย์ที่ล้างไม่ได้
* DOC – คาร์บอนอินทรีย์ที่ละลายน้ำ
* NDOC – คาร์บอนอินทรีย์ที่ไม่ละลายน้ำ

คาร์บอนอินทรีย์ที่ละลายน้ำ (DOC) โดยทั่วไปจะถือว่าผ่านตัวกรองขนาด 0.45um TOC จะมีอนุภาคขนาดใหญ่ถูกจัดประเภทเป็นอนุภาคหรือไม่ละลายน้ำ (NDOC) ประมาณ 50 ถึง 75% ของ DOC ในน้ำธรรมชาติจะอยู่ในรูปของกรดอินทรีย์โพลิเมอร์ – กรดฟุลวิคและกรดฮิวมิก ประมาณ 10% ของ TOC อยู่ในคอลลอยด์ ส่วนใหญ่เป็นกรดฮิวมิกและแร่ธาตุต่าง ๆ อีก 10 ถึง 20% เป็นโมเลกุลขนาดเล็กจากการย่อยสลายสารอินทรีย์ (DR PAUL WHITEHEAD. 2021: online)

1. ค่า Trihalomethanes

Trihalomethanes (THMs) เป็นกลุ่มสารเคมีที่สามารถปนเปื้อนในน้ำดื่มได้ THMs เกิดขึ้นเมื่อคลอรีนที่ใช้ในการฆ่าเชื้อในน้ำทำปฏิกิริยากับสารอินทรีย์ตามธรรมชาติ การปนเปื้อน THMs จะเกิดขึ้นโดยเฉพาะอย่างยิ่งในระบบน้ำดื่มที่มาจากน้ำผิวดิน ในคลอโรฟอร์มจะมีกลิ่นเฉพาะเกณฑ์กลิ่นของมันจะมีค่าปริมาณอยู่ที่ 2.4 มิลลิกรัม/ลิตร ในน้ำ และ 420 มิลลิกรัม/ลูกบาศก์เมตร ในอากาศ

มาตรฐานสำหรับความเข้มข้นของ THM ในควิเบกคือ 80 ไมโครกรัม/ลิตร การได้รับสาร THMs ในน้ำที่มีความเข้มข้นสูงเป็นเวลานาน ไม่ว่าจะจากทางการดื่มน้ำที่ปนเปื้อน สูดดมไอระเหยของน้ำที่ปนเปื้อน หรือร่างกายสัมผัสกับน้ำที่ปนเปื้อนล้วนแต่มีความเสี่ยงต่อการเป็นมะเร็งกระเพาะปัสสาวะได้ (Gouvernement du Québec. 2023: online) ( World Health Organization. 2022: 475)

1. ค่า Turbidity หรือ ค่าความขุ่นของน้ำ

ความขุ่นของน้ำเกิดจากอนุภาคแขวนลอยหรือสารคอลลอยด์ที่ขวางกั้นการส่งแสงผ่านน้ำ อาจเกิดจากสารอนินทรีย์หรือสารอินทรีย์หรือทั้งสองอย่างรวมกัน โดยทั่วไปแล้วจุลินทรีย์ (แบคทีเรีย ไวรัส และโปรโตซัว) จะติดมากับอนุภาค นอกจากนี้ ความขุ่นอาจรบกวนประสิทธิภาพการฆ่าเชื้อโรคได้ ความขุ่นนั้นส่งผลต่อการยอมรับของผู้บริโภคต่อน้ำเพราะค่าความขุ่นนั้นสามารถมองเห็นได้โดยเฉพาะเมื่อมีค่าความขุ่นสูงแบบเห็นได้ชัดเจน ความขุ่นสามารถให้อาหารและที่พักพิงแก่เชื้อโรคได้ หากไม่กำจัดออก สาเหตุของความขุ่นสูงสามารถส่งเสริมการเจริญเติบโตของเชื้อโรคในน้ำ ซึ่งนำไปสู่การระบาดของโรคที่มากับน้ำ ซึ่งทำให้เกิดกรณีสำคัญของการเจ็บป่วยในลำไส้ ความขุ่นจะมีค่าคือ NTU โดยมีเกณฑ์เหมาะสมคือ 5 NTU (Water Science School. 2018: online)

น้ำตามธรรมชาติจะมีคสามขุ่นอยู่เสมอ

* ในน้ำใสจะมีค่าความขุ่นไม่เกิน 25 NTU
* ในน้ำขุ่นปานกลางจะมีค่าความขุ่นไม่เกิน 25 – 100 NTU
* ในน้ำขุ่นมากจะมีค่าความขุ่นเกิน 100 NTU

1. วิธีการวัดข้อมูลของค่าต่างๆในน้ำ
   * วัดค่า pH หรือ ค่าความเป็นกรดและด่าง



**ภาพที่ 2.1** pH Meter (ที่มา : [www.tools.in.th](http://www.tools.in.th))

* + วัดค่า Hardness หรือ ค่าความกระด้าง



**ภาพที่ 2.2** HI3812 ชุดตรวจสอบความกระด้างของน้ำ (ที่มา : www.neonics.co.th)

* + วัดค่า Solids (TDS) หรือ ค่าของแข็งละลายน้ำ

Graphical user interface

Description automatically generated

**ภาพที่ 2.3** เครื่องวัดค่า TDS และค่า EC (ที่มา : [www.neonics.co.th](http://www.neonics.co.th))

* + วัดค่า chloramines หรือ chlorine



**ภาพที่ 2.4** HI97701 เครื่องวัดคลอรีนอิสระ (Free Chlorine) (ที่มา : [www.neonics.co.th](http://www.neonics.co.th))

* + วัดค่า Sulfate



**ภาพที่ 2.5** เครื่องวัดปริมาณซัลเฟตในน้ำ Sulfate Meter (ที่มา : [www.neonics.co.th](http://www.neonics.co.th))

* + วัดค่า Conductivity หรือ ค่าการนำไฟฟ้า



**ภาพที่ 2.6** เครื่องวัด EC Meter (ที่มา : [www.neonics.co.th](http://www.neonics.co.th))

* + วัดค่า Organic Carbon หรือ ค่า คาร์บอนอินทรีย์



**ภาพที่ 2.7** เครื่อง Total organic carbon รุ่น 1030C (ที่มา : [www.entech.co.th](http://www.entech.co.th))

* + วัดค่า Turbidity หรือ ค่าความขุ่นของน้ำ



**ภาพที่ 2.8** เครื่อง Turbidity meter (ที่มา : [www.spscience.com](http://www.spscience.com))

1. **ความสำคัญของน้ำ**

น้ำเป็นทรัพยากรที่มีสำคัญอย่างมากในการดำรงชีวิตอย่างไรก็ตามแม้น้ำจะมีปริมาณมากแต่ปริมาณที่สามารถนำมาใช้ได้นั้นกลับมีปริมาณที่จำกัด ลักษณะของน้ำโดยทั่วไปนั้นเป็นแบบไดนามิกสามารถเปลี่ยนแปลงได้ทั้งลักษณะทางกายภาพและทางเคมีตามกาลเวลา

น้ำเป็นองค์ประกอบสำคัญของสิ่งมีชีวิตทั้งหมด เซลล์ของสิ่งมีชีวิตทั้งหมดก็สร้างจากสารละลายในน้ำ สารแขวนลอย และอิมัลชันที่มีน้ำอยู่ในช่วง 25–85% โดยทั่วไปแล้วการดำรงอยู่พืชก็ขึ้นอยู่กับน้ำเช่นกัน พวกเขาจะใช้น้ำและเกลือแร่อื่น ๆ จากดิน พืชประกอบด้วยน้ำ 50–75% ในขณะที่มนุษย์มีน้ำเป็นองค์ประกอบ 60–65% น้ำในธรรมชาตินั้นจะประบอกไปด้วยแร่ธาตุต่าง ๆ สารมลพิษอินทรีย์ และก๊าซที่ละลายปะปนอยู่ นอกจากนี้โรคภัยไข้เจ็บหลายอย่างยังสามารถส่งต่อกันผ่านทางน้ำได้อีกด้วย

1. **สารปนเปื้อนที่มาจากสารเคมีในน้ำ**
   1. อะลูมิเนียม (Aluminum)

อะลูมิเนียมที่เกิดขึ้นตามธรรมชาติและเกลืออะลูมิเนียมที่ใช้เป็นสารตกตะกอนในการบำบัดน้ำดื่ม ในกรณีที่อะลูมิเนียมมีความเข้มข้นเกิน 0.1–0.2 มิลลิกรัม/ลิตร มักนำไปสู่ข้อร้องเรียนของผู้บริโภคอันเป็นผลมาจากการสะสมของอะลูมิเนียมไฮดรอกไซด์ฟล็อกและอาการกำเริบของการเปลี่ยนสีของน้ำด้วยธาตุเหล็ก ดังนั้นจึงเป็นเรื่องสำคัญที่จะต้องเพิ่มประสิทธิภาพในกระบวนการบำบัดน้ำเพื่อลดอะลูมิเนียมที่หลงเหลือเข้าสู่ระบบจำหน่าย ภายใต้สภาวะการทำงานที่ดีส่งผลให้ความเข้มข้นของอะลูมิเนียมลดน้อยซึ่งสามารถทำได้มากกว่า 0.1 มิลลิกรัม/ลิตร

* 1. แอมโมเนีย (Ammonia)

ความเข้มข้นของกลิ่นตามเกณฑ์ของแอมโมเนียที่ pH อัลคาไลน์อยู่ที่ประมาณ 1.5 มิลลิกรัม/ลิตร และเกณฑ์การรับรสอยู่ที่ประมาณ 35 มิลลิกรัม/ลิตร สำหรับแอมโมเนียมไอออนบวก แอมโมเนียไม่มีผลกระทบต่อสุขภาพ อย่างไรก็ตามแอมโมเนียก็สามารถทำปฏิกิริยากับคลอรีนเพื่อลดคลอรีนอิสระและสร้างคลอรามีนได้

* 1. คลอรามีน (Chloramines)

คลอรามีน เช่น โมโนคลอรามีน ไดคลอรามีน และไตรคลอรามีน (ไนโตรเจนไตรคลอไรด์) เกิดจากปฏิกิริยาของคลอรีนกับแอมโมเนีย (สามารถดูเพิ่มเติมได้ที่ 2.1.4)

* 1. คลอไรด์ (Chloride)

คลอไรด์ที่มีความเข้มข้นสูงจะทำให้น้ำมีรสเค็ม เกณฑ์รสชาติสำหรับคลอไรด์ไอออนขึ้นอยู่กับไอออนบวกที่มีและอยู่ในช่วงของ 200–300 มิลลิกรัม/ลิตร สำหรับโซเดียม โพแทสเซียม และแคลเซียมคลอไรด์ ความเข้มข้นที่มีปริมาณมากเกิน 250 มิลลิกรัม/ลิตร มีแนวโน้มที่จะตรวจพบได้มากขึ้นจากรสชาติ แต่ผู้บริโภคบางรายอาจตรวจพบได้จากความคุ้นเคยกับรสชาติที่เกิดจากคลอไรด์ในระดับต่ำ

* 1. คลอรีน (Chlorine)

บุคคลส่วนใหญ่สามารถลิ้มรสหรือได้กลิ่นคลอรีนในน้ำดื่มที่มีความเข้มข้นต่ำกว่า 5 มิลลิกรัม/ลิตร และบางชนิดมีระดับต่ำถึง 0.3 มิลลิกรัม/ลิตร เกณฑ์รสชาติสำหรับคลอรีนต่ำกว่าค่าแนวทางด้านสุขภาพอยู่ที่ 5 มิลลิกรัม/ลิตร

* 1. คลอโรเบนซีน (Chlorobenzenes)

คลอโรเบนซีนมีเกณฑ์การรับรสชาติและกลิ่นอยู่ที่ 10–20 µg/l และเกณฑ์การรับกลิ่นตั้งแต่ 40-120 µg/l สำหรับโมโนคลอโรเบนซีน ค่าแนวทางด้านสุขภาพไม่ได้มาจากโมโนคลอโรเบนซีนแม้ว่าค่าตามสุขภาพที่สามารถได้รับเกินกว่ารสชาติที่รายงานต่ำสุดและเกณฑ์กลิ่นในน้ำมีรายงานเกณฑ์กลิ่นที่ 2–10 และ 0.3–30 µg/l ไดคลอโรเบนซีน ตามลำดับ มีการรายงานเกณฑ์การรับรสที่ 1 และ 6 ไมโครกรัม/ลิตร 1,2- และ 1,4-ไดคลอโรเบนซีน ตามลำดับ ค่าแนวทางด้านสุขภาพของ 1 มก./ลิตร สำหรับ 1,2-ไดคลอโรเบนซีน และ 0.3 มก./ลิตร สำหรับ 1,4-ไดคลอโรเบนซีน

* 1. คลอโรฟีนอล (Chlorophenols)

คลอโรฟีนอลโดยทั่วไปมีเกณฑ์การรับรสและกลิ่นที่ต่ำมาก เกณฑ์รสชาติในน้ำสำหรับ 2-คลอโรฟีนอล, 2,4-ไดคลอโรฟีนอล และ 2,4,6-ไตรคลอโรฟีนอล คือ 0.1, 0.3 และ 2 ไมโครกรัม/ลิตร ตามลำดับ ค่าเกณฑ์ของกลิ่นคือ 10, 40 และ 300 µg/l ตามลำดับถ้าน้ำที่มี 2,4,6-ไตรคลอโรฟีนอลไม่มีรสชาติ จะไม่แสดงความเสี่ยงที่สำคัญต่อสุขภาพ จุลินทรีย์ในการกระจายบางครั้งระบบอาจเมทิลเลตคลอโรฟีนอลเพื่อผลิตคลอรีนอนิโซลซึ่งเกณฑ์การรับกลิ่นจะต่ำกว่ามาก

* 1. สี (Color)

น้ำดื่มไม่ควรมีสีที่มองเห็นได้ สีในน้ำดื่มมักเกิดจากการมีสารอินทรีย์ที่มีสี (กรดฮิวมิกและกรดฟุลวิคเป็นหลัก) เกี่ยวข้องกับเศษซากพืชในดิน สียังได้รับอิทธิพลอย่างมากจากการปรากฏของเหล็กและโลหะอื่น ๆ ไม่ว่าจะเป็นสิ่งเจือปนตามธรรมชาติหรือเป็นผลิตภัณฑ์ที่กัดกร่อน นอกจากนี้ยังอาจเกิดจากการปนเปื้อนของแหล่งน้ำอุตสาหกรรมด้วย คนส่วนใหญ่สามารถตรวจจับสีได้มากกว่า 15 หน่วยสีจริง (TCU) ในแก้วน้ำระดับสีที่ต่ำกว่า 15 TCU มักจะเป็นที่ยอมรับของผู้บริโภคจากคาร์บอนอินทรีย์ธรรมชาติ (เช่น ฮิวมิกส์) ยังสามารถบ่งชี้ถึงแนวโน้มสูงที่จะผลิตผลพลอยได้จากกระบวนการฆ่าเชื้อโรค ในด้านแนวทางด้านสุขภาพไม่มีการเสนอค่าด้านสุขภาพสำหรับสีในน้ำดื่ม

* 1. ทองแดง (Copper)

ทองแดงในน้ำดื่มมักเกิดจากการกัดกร่อนของน้ำที่ชะล้างทองแดงจากท่อทองแดงในอาคารและความเข้มข้นอาจแตกต่างกันไปอย่างมีนัยสำคัญตามระยะเวลาที่น้ำได้สัมผัสกับท่อทองแดงตัวอย่างเช่น น้ำที่ดึงออกมาครั้งแรกคาดว่าจะมีความเข้มข้นของทองแดงสูงกว่าตัวอย่างที่ชะล้างจนหมด ความเข้มข้นสูงสามารถรบกวนการวัตถุประสงค์ของการใช้น้ำในครัวเรือนที่ความเข้มข้นของทองแดงมากกว่า 1 มก./ล. ที่ระดับสูงกว่า 5 มก./ล. ทองแดงยังให้สีและรสขมที่ไม่ต้องการลงในน้ำ แม้ว่าทองแดงจะทำให้เกิดรสชาติได้ ค่าควรยอมรับได้ที่ค่าแนวทางด้านสุขภาพอยู่ที่ 2 มก./ล

* 1. ออกซิเจนที่ละลายน้ำ (Dissolved oxygen)

ปริมาณออกซิเจนที่ละลายในน้ำได้รับอิทธิพลจากแหล่งที่มา อุณหภูมิของน้ำดิบ การบำบัด และกระบวนการทางเคมีหรือชีวภาพที่เกิดขึ้นในระบบจำหน่าย การลดลงของออกซิเจนที่ละลายในน้ำสามารถกระตุ้นให้เกิดการลดจุลินทรีย์จากไนเตรตเป็นไนไตรต์และซัลเฟตเป็นซัลไฟด์ นอกจากนี้ยังสามารถทำให้เกิดเพิ่มความเข้มข้นของเหล็กในสารละลาย โดยจะมีการเปลี่ยนสีที่ก๊อกน้ำภายหลังเมื่อเติมอากาศ ไม่แนะนำให้ใช้ค่าแนวทางด้านสุขภาพ อย่างไรก็ตาม ออกซิเจนที่ละลายในน้ำในระดับที่สูงมากอาจทำให้การกัดกร่อนรุนแรงขึ้นท่อโลหะ

* 1. เอทิลเบนซีน (Ethylbenzene)

เอทิลเบนซีนมีกลิ่นที่หอม เกณฑ์กลิ่นที่รายงานในช่วงน้ำตั้งแต่ 2 ถึง 130 ไมโครกรัม/ลิตร เกณฑ์กลิ่นที่รายงานต่ำสุดต่ำกว่า 100 เท่าค่าแนวทางด้านสุขภาพที่ 0.3 มก./ลิตร เกณฑ์รสชาติอยู่ระหว่าง 72 ถึง 200 µg/l

* 1. ความแข็งกระด้างของน้ำ (Hardness)

ความกระด้างที่เกิดจากแคลเซียมและแมกนีเซียมมักจะแสดงโดยการตกตะกอนของคราบสบู่ผู้บริโภคคือมีแนวโน้มที่จะสังเกตเห็นการเปลี่ยนแปลงของความแข็ง การยอมรับของสาธารณะในระดับความแข็งของน้ำอาจแตกต่างกันมากจากชุมชนหนึ่งไปยังอีกชุมชนหนึ่ง (สามารถดูเพิ่มเติมได้ที่ 2.1.2)

* 1. ไฮโดรเจนซัลไฟด์ (Hydrogen sulfide)

ไฮโดรเจนซัลไฟด์ค่าขีดจำกัดของรสชาติและกลิ่นของไฮโดรเจนซัลไฟด์ในน้ำอยู่ที่ประมาณ 0.05 ถึง 0.1 มก./ล. โดยเฉพาะอย่างยิ่งกลิ่น “ไข่เน่า” ของก๊าซไข่เน่าสังเกตได้ในน้ำบาดาลบางแห่งและในน้ำดื่มอันเป็นผลมาจากการลดลงของออกซิเจนและการลดลงของซัลเฟตตามมากิจกรรมของแบคทีเรียซัลไฟด์จะถูกออกซิไดซ์อย่างรวดเร็วเป็นซัลเฟตในน้ำที่มีบ่อเติมอากาศหรือน้ำที่มีคลอรีน และระดับไฮโดรเจนซัลไฟด์ในน้ำที่มีออกซิเจนมักจะต่ำมากผู้บริโภคสามารถตรวจจับไฮโดรเจนซัลไฟด์ในน้ำดื่มได้ง่ายและต้องดำเนินการแก้ไขทันที มีความเป็นไปได้น้อยที่บุคคลจะบริโภคไฮโดรเจนซัลไฟด์ในปริมาณที่จะเป็นอันตรายดังนั้น จึงไม่มีค่าแนวทางด้านสุขภาพสำหรับสารประกอบนี้

* 1. เหล็ก (Iron)

น้ำบาดาลแบบไม่ใช้ออกซิเจนอาจมีธาตุเหล็กที่ความเข้มข้นสูงถึงหลายมิลลิกรัมต่อลิตรโดยไม่มีการเปลี่ยนสีหรือความขุ่นในน้ำอย่างไรก็ตามเมื่อน้ำสัมผัสกับบรรยากาศ ธาตุเหล็กในน้ำจะออกซิไดซ์เป็นเหล็กเฟอริก ทำให้น้ำมีสีน้ำตาลแดงที่ไม่เหมาะสม เหล็กยังส่งเสริมการเจริญเติบโตของ "แบคทีเรียเหล็ก" ซึ่งได้รับพลังงานจากปฏิกิริยาออกซิเดชันของเหล็กเฟอร์รัสกับเหล็กเฟอริก และในกระบวนการนี้จะทำให้เกิดสารเคลือบที่ลื่นไหลบนท่อ ที่ระดับค่าสูงกว่า 0.3 มก./ล.โดยปกติแล้วจะไม่มีรสชาติที่สังเกตเห็นได้ที่ความเข้มข้นของธาตุเหล็กต่ำกว่า 0.3 มก./ล ความขุ่นและสีอาจพัฒนา ไม่มีการเสนอค่าแนวทางด้านสุขภาพของธาตุเหล็กในน้ำ

* 1. แมงกานีส (Manganese)

ที่ระดับต่ำถึง 0.02 มก./ล. แมงกานีสซึ่งเป็นแมงกานีสออกไซด์ที่ไม่ละลายน้ำในแหล่งน้ำอาจทำให้น้ำเปลี่ยนสีการมีแมงกานีสในน้ำดื่ม เช่น เหล็ก อาจนำไปสู่การสะสมของตะกอนในระบบจำหน่ายแม้ที่ความเข้มข้น 0.02 มก./ล. แมงกานีสอาจสะสมอยู่ในท่อในรูปของแมงกานีสออกไซด์ ซึ่งอาจหลุดลอกออกมาเป็นตะกอนสีดำ ในทางตรงกันข้าม แมงกานีสที่ละลายน้ำได้นั้นไม่มีสีและมองไม่เห็นด้วยตาเปล่าที่ความเข้มข้นสูงถึง 506 มก./ล. ซึ่งเป็นความเข้มข้นสูงสุดที่ทดสอบค่าแนวทางด้านสุขภาพสำหรับแมงกานีส 0.08 มก./ลิตรสำหรับแมงกานีสนั้นสูงกว่าเกณฑ์ที่ยอมรับได้ของแมงกานีสที่ไม่ละลายน้ำ อย่างไรก็ตาม เนื่องจากแมงกานีสที่ละลายน้ำสามารถถูกปล่อยออกมา ณ จุดใช้งานหรือการรวบรวมได้ของน้ำที่เปลี่ยนสีไม่สามารถนำมาใช้ประเมินได้อย่างน่าเชื่อถือว่ามีแมงกานีสอยู่หรือไม่ดังนั้นจึงควรพิจารณาด้านสุขภาพ ในการกำหนดหลักเกณฑ์และมาตรฐานคุณภาพน้ำดื่ม

* 1. น้ำมันปิโตรเลียม (Petroleum oils)

น้ำมันปิโตรเลียมสามารถก่อให้เกิดการมีน้ำหนักโมเลกุลต่ำจำนวนหนึ่งได้ไฮโดรคาร์บอนที่มีกลิ่นในน้ำดื่ม เบนซีน , โทลูอีน , เอทิลเบนซีนและไซลีน ด้รับการพิจารณาแยกจากกันในส่วนนี้ เนื่องจากได้รับค่าแนวทางด้านสุขภาพสำหรับสารเคมีเหล่านี้ อย่างไรก็ตาม จำนวนของไฮโดรคาร์บอนอื่นๆ โดยเฉพาะอัลคิลเบนซีน เช่น ไตรเมทิลเบนซีน อาจให้มีกลิ่น คล้ายน้ำมันดีเซล ที่ไม่พึงประสงค์มากที่ระดับความเข้มข้นไม่กี่ไมโครกรัมต่อลิตรมีการระบุว่าเกณฑ์รสชาติของส่วนผสมต่ำไฮโดรคาร์บอนอะโรมาติกน้ำหนักโมเลกุลต่ำกว่าเกณฑ์

* 1. ค่า pH และการกัดกร่อน (pH and corrosion)

แม้ว่าค่า pH มักจะไม่ส่งผลกระทบโดยตรงต่อผู้บริโภค แต่ก็เป็นหนึ่งในพารามิเตอร์คุณภาพน้ำที่สำคัญที่สุดในการปฏิบัติงาน ความเอาใจใส่อย่างรอบคอบต่อการควบคุมค่า pH เป็นสิ่งจำเป็นในทุกขั้นตอนของการบำบัดน้ำเพื่อให้แน่ใจว่าน้ำมีความใสน่าพอใจและการฆ่าเชื้อโรค (สามารถดูเพิ่มเติมได้ที่ 2.1.1)

* 1. โซเดียม (Sodium)

ความเข้มข้นของเกณฑ์รสชาติของโซเดียมในน้ำขึ้นอยู่กับประจุลบที่เกี่ยวข้องและอุณหภูมิของสารละลาย ที่อุณหภูมิห้อง เกณฑ์รสชาติเฉลี่ยสำหรับโซเดียมประมาณ 200 มก./ล. ไม่ได้รับค่าแนวทางด้านสุขภาพเนื่องจากปริมาณน้ำที่บริโภคต่อวันมีปริมาณน้อย

* 1. สไตรีน (Styrene)

สไตรีนมีกลิ่นฉุนและมีรายงานเกณฑ์กลิ่นสำหรับสไตรีนในน้ำอยู่ในช่วงตั้งแต่ 0.004 ถึง 2.6 มก./ลิตร ขึ้นอยู่กับอุณหภูมิ สไตรีนจึงอาจจะตรวจพบในน้ำที่ความเข้มข้นต่ำกว่าค่าแนวทางสุขภาพ 0.02 มก./ล

* 1. ซัลเฟต (Sulfate)

การมีซัลเฟตในน้ำดื่มสามารถทำให้เกิดรสชาติที่สังเกตได้และสูงมากระดับอาจทำให้เกิดฤทธิ์เป็นยาระบายในผู้บริโภคที่ไม่คุ้นเคยการด้อยค่าของรสชาติแปรผันตามลักษณะของไอออนบวกที่เกี่ยวข้อง เกณฑ์การรับรสได้รับ (สามารถดูเพิ่มเติมได้ที่ 2.1.5)

* 1. ผงซักฟอกสังเคราะห์ (Synthetic detergents)

ในหลายประเทศ ผงซักฟอกที่เป็นประจุลบชนิดถาวรถูกแทนที่ด้วยชนิดอื่นที่ย่อยสลายได้ง่ายกว่า ดังนั้นระดับที่พบในแหล่งน้ำจึงลดลงอย่างมาก ไม่ควรมีความเข้มข้นของผงซักฟอกในน้ำดื่ม ไม่ควรปล่อยให้ความเข้มข้นของผงซักฟอกในน้ำดื่มสูงถึงระดับที่ก่อให้เกิดปัญหาฟองหรือรสชาติ การปรากฏตัวของผงซักฟอกใด ๆ อาจบ่งบอกถึงการปนเปื้อนของแหล่งน้ำกับสิ่งปฏิกูลหรือการไหลเข้าของน้ำยาซักผ้าเข้าสู่ระบบ

* 1. โทลูอีน (Toluene)

โทลูอีนมีกลิ่นฉุนคล้ายเบนซิน ช่วงเกณฑ์รสชาติที่รายงานมีตั้งแต่ 0.04 ถึง 0.12 มก./ล. เกณฑ์กลิ่นที่รายงานสำหรับโทลูอีนในน้ำตั้งแต่ 0.024 ถึง 0.17 มก./ล. โทลูอีนจึงอาจส่งผลต่อการยอมรับของน้ำที่ความเข้มข้นต่ำกว่าค่าแนวทางด้านสุขภาพที่ 0.7 มก./ล

* 1. ของแข็งที่ละลายทั้งหมด (Total dissolved solids)

ความอร่อยของน้ำที่มีระดับของแข็งที่ละลายน้ำได้ทั้งหมดน้อยกว่าประมาณ600 มก./ลิตร โดยทั่วไปถือว่าเป็นน้ำดื่มที่ดีอย่างมีนัยสำคัญ และไม่อร่อยมากขึ้นที่ระดับ TDS มากกว่า 1,000 มก./ล. (สามารถดูเพิ่มเติมได้ที่ 2.1.3)

* 1. ความขุ่น (Turbidity)

ความขุ่น โดยทั่วไปแสดงเป็นหน่วยความขุ่นแบบเนฟีโลเมตริกอธิบายถึงความขุ่นของน้ำเกิดจากอนุภาคแขวนลอย เช่น ดินเหนียวและตะกอนดิน การตกตะกอนของสารเคมี เช่น แมงกานีสและเหล็ก อนุภาคอินทรีย์ เช่น เศษพืช และสิ่งมีชีวิต(สามารถดูเพิ่มเติมได้ที่ 2.1.9)

* 1. ไซลีน (Xylenes)

ความเข้มข้นของไซลีนในช่วง 0.3 มก./ล. ทำให้เกิดรสชาติและกลิ่นที่ตรวจจับได้เกณฑ์กลิ่นสำหรับไอโซเมอร์ไซลีนในน้ำมีรายงานว่าอยู่ในช่วงตั้งแต่ 0.02ถึง 1.8 มก./ล. เกณฑ์กลิ่นต่ำสุดต่ำกว่าค่าแนวทางด้านสุขภาพ0.5 มก./ล. สำหรับไซลีน

* 1. สังกะสี (Zinc)

สังกะสีให้รสฝาดที่ไม่ต้องการแก่น้ำที่ความเข้มข้นของเกณฑ์การรับรสประมาณ 4 มก./ล.(ในรูปของซิงค์ซัลเฟต)น้ำที่มีสังกะสีที่ความเข้มข้นเกิน 3-5 มก./ล. แม้ว่าน้ำดื่มไม่ค่อยมีสังกะสีที่ความเข้มข้นสูงกว่า 0.1 มก./ล. ระดับในก๊อกน้ำอาจสูงขึ้นมากเนื่องจากสังกะสีที่ใช้ในวัสดุประปาสังกะสีรุ่นเก่า นี่อาจเป็นตัวบ่งชี้ถึงแคดเมียมที่สูงขึ้นจากอายุที่มากขึ้น ไม่มีการเสนอค่าแนวทางด้านสุขภาพสำหรับสังกะสีในน้ำดื่ม

1. **วิทยาศาสตร์ข้อมูล (Data Science)**
   * 1. อะลูมิเนียม (Aluminum)

วิทยาการข้อมูลคือการผสมผสานกันระหว่างคณิตศาสตร์และสถิติ การเขียนโปรแกรมเฉพาะทาง การวิเคราะห์ขั้นสูง ปัญญาประดิษฐ์ (AI) และการเรียนรู้ของ machine learning ที่มีความเชี่ยวชาญเฉพาะด้านเพื่อเปิดเผยข้อมูลเชิงลึกที่นำไปใช้ได้จริงซึ่งซ่อนอยู่ในข้อมูลขององค์กร ข้อมูลเชิงลึกเหล่านี้สามารถใช้เป็นแนวทางในการตัดสินใจและการวางแผนเชิงกลยุทธ์ได้

ปริมาณแหล่งข้อมูลที่เพิ่มขึ้นอย่างรวดเร็ว และข้อมูลที่ตามมา ทำให้วิทยาการข้อมูลเป็นหนึ่งในสาขาที่เติบโตเร็วที่สุดในทุกอุตสาหกรรม ด้วยเหตุนี้ จึงไม่น่าแปลกใจที่บทบาทของนักวิทยาศาสตร์ข้อมูลได้รับการขนานนามว่าเป็น " sexiest job of the 21st century " เพราะเป็นอาชีพที่มีความต้องการตัวมากที่สุดและหายากมากที่สุดอาชีพหนึ่งในโลกของเทคโนโลยี โดย Harvard Business Review องค์กรต่างพึ่งพา Data Science มากขึ้นในการตีความข้อมูลและให้คำแนะนำที่นำไปปฏิบัติได้เพื่อปรับปรุงผลลัพธ์ทางธุรกิจ

* + 1. การรวบรวมข้อมูลของวิทยาศาสตร์ข้อมูล
  + Data ingestion เริ่มต้นด้วยการรวบรวมข้อมูล ทั้งข้อมูลดิบที่มีโครงสร้างและไม่มีโครงสร้างจากแหล่งข้อมูลที่เกี่ยวข้องทั้งหมดโดยใช้วิธีการที่หลากหลาย วิธีการเหล่านี้อาจรวมถึงการป้อนข้อมูลด้วยตนเอง ดึงข้อมูลจากเว็บต่างๆและข้อมูลที่ถูกสร้างขึ้นอย่างต่อเนื่องจากระบบและอุปกรณ์ต่างๆ แหล่งข้อมูลอาจรวมถึงข้อมูลที่มีโครงสร้าง เช่น ข้อมูลลูกค้า พร้อมด้วยข้อมูลที่ไม่มีโครงสร้าง เช่น ไฟล์บันทึก วิดีโอ เสียง รูปภาพ Internet of Things (IoT) โซเชียลมีเดีย และอื่นๆ
  + Data storage and data processing เนื่องจากข้อมูลสามารถมีรูปแบบและโครงสร้างที่แตกต่างกันได้ จึงจำเป็นต้องพิจารณาระบบจัดเก็บข้อมูลที่แตกต่างกันตามประเภทของข้อมูลที่ต้องจัดเก็บ ซึ่งช่วยอำนวยความสะดวกในขั้นตอนการทำงานเกี่ยวกับการวิเคราะห์ การเรียนรู้ของ machine learning และโมเดลการเรียนรู้เชิงลึก ขั้นตอนนี้รวมถึงการล้างข้อมูล การทำซ้ำ การแปลง และการรวมข้อมูลโดยใช้ งาน ETL (แยก แปลง โหลด) หรือเทคโนโลยีการรวมข้อมูลอื่นๆ การเตรียมข้อมูลนี้จำเป็นสำหรับการส่งเสริมคุณภาพข้อมูลก่อนที่จะโหลดลงในคลังข้อมูล
  + Data analysis นักวิทยาศาสตร์ข้อมูลทำการวิเคราะห์ข้อมูลเชิงสำรวจเพื่อตรวจสอบข้อมูล รูปแบบ ช่วง และการกระจายของค่าภายในข้อมูล การสำรวจวิเคราะห์ข้อมูลนี้ขับเคลื่อนการสร้างสมมติฐานสำหรับการทดสอบ a หรือ b นอกจากนี้ยังช่วยให้นักวิเคราะห์สามารถระบุความเกี่ยวข้องของข้อมูลเพื่อใช้ในความพยายามสร้างแบบจำลองสำหรับการวิเคราะห์เชิงคาดการณ์ การเรียนรู้ของ machine learning หรือการเรียนรู้เชิงลึก ขึ้นอยู่กับความถูกต้องของแบบจำลอง
  + Communicate ข้อมูลเชิงลึกจะแสดงเป็นรายงานและการแสดงภาพข้อมูลอื่นๆ ในการทำความเข้าใจ นักวิทยาศาสตร์ข้อมูลสามารถใช้เครื่องมือสร้างภาพข้อมูลโดยเฉพาะได้

1. **การถดถอยของ Logistic และ Machine Learning**

ภายในแมชชีนเลิร์นนิงการถดถอยโลจิสติกเป็นส่วนหนึ่งของโมเดลแมชชีนเลิร์นนิงภายใต้การดูแล นอกจากนี้ยังถือเป็นรูปแบบการเลือกปฏิบัติ ซึ่งหมายความว่าจะพยายามแยกแยะระหว่างชั้นเรียนซึ่งแตกต่างจากอัลกอริธึมเชิงกำเนิด เช่นnaïve bayesซึ่งไม่สามารถสร้างข้อมูล เช่น รูปภาพ ของคลาสที่พยายามทำนายได้กล่าวถึงวิธีที่การถดถอยโลจิสติกเพิ่มฟังก์ชันความน่าจะเป็นของล็อกสูงสุดเพื่อกำหนดค่าสัมประสิทธิ์เบต้าของโมเดล สิ่งนี้เปลี่ยนแปลงเล็กน้อยภายใต้บริบทของการเรียนรู้ของเครื่อง ภายในแมชชีนเลิร์นนิง ความน่าจะเป็นของบันทึกเชิงลบที่ใช้เป็นฟังก์ชันการสูญเสีย โดยใช้กระบวนการไล่ระดับสีลงเพื่อหาค่าสูงสุดทั่วโลกนี่เป็นอีกวิธีหนึ่งในการได้รับค่าประมาณเดียวกันกับที่กล่าวไว้ข้างต้นนอกจากนี้ การถดถอยโลจิสติกยังมีแนวโน้มที่จะเกิดoverfittingโดยเฉพาะอย่างยิ่งเมื่อมีตัวแปรทำนายจำนวนมากภายในโมเดล การทำให้เป็นมาตรฐานมักใช้เพื่อลงโทษค่าสัมประสิทธิ์ขนาดใหญ่ของพารามิเตอร์เมื่อโมเดลมีมิติสูง

* 1. กรณีของการใช้การถดถอยโลจิสติก

การถดถอยโลจิสติกมักใช้สำหรับการทำนายและการจำแนกปัญหา กรณีการใช้งานบางส่วน ได้แก่

* การตรวจจับการฉ้อโกง:แบบจำลองการถดถอยโลจิสติกสามารถช่วยทีมระบุความผิดปกติของข้อมูล ซึ่งเป็นการคาดการณ์การฉ้อโกง พฤติกรรมหรือลักษณะบางอย่างอาจมีความเกี่ยวข้องกับกิจกรรมการฉ้อโกงสูงกว่า ซึ่งเป็นประโยชน์อย่างยิ่งต่อการธนาคารและสถาบันการเงินอื่น ๆ ในการปกป้องลูกค้าของตน บริษัทที่ใช้ SaaS ได้เริ่มนำแนวทางปฏิบัติเหล่านี้มาใช้เพื่อกำจัดบัญชีผู้ใช้ปลอมออกจากชุดข้อมูลของตนเมื่อทำการวิเคราะห์ข้อมูลเกี่ยวกับประสิทธิภาพทางธุรกิจ
* การทำนายโรค: ในทางการแพทย์ วิธีการวิเคราะห์นี้สามารถใช้เพื่อทำนายความเป็นไปได้ในการเกิดโรคหรือการเจ็บป่วยสำหรับประชากรที่กำหนด องค์กรด้านการดูแลสุขภาพสามารถตั้งค่าการดูแลเชิงป้องกันสำหรับบุคคลที่มีแนวโน้มสูงขึ้นสำหรับความเจ็บป่วยที่เฉพาะเจาะจง
* การคาดคะเนการเลิกใช้งาน : พฤติกรรมเฉพาะอาจบ่งบอกถึงการเลิกใช้งานในหน้าที่ต่างๆ ขององค์กร ตัวอย่างเช่น ฝ่ายทรัพยากรบุคคลและทีมผู้บริหารอาจต้องการทราบว่ามีพนักงานที่มีประสิทธิภาพสูงภายในบริษัทที่มีความเสี่ยงที่จะลาออกจากองค์กรหรือไม่ ข้อมูลเชิงลึกประเภทนี้สามารถกระตุ้นการสนทนาเพื่อทำความเข้าใจปัญหาภายในบริษัท เช่น วัฒนธรรมหรือค่าตอบแทน หรืออีกทางหนึ่ง องค์กรการขายอาจต้องการเรียนรู้ว่าลูกค้ารายใดที่มีความเสี่ยงที่จะพาธุรกิจของตนไปที่อื่น สิ่งนี้สามารถกระตุ้นให้ทีมตั้งค่ากลยุทธ์การรักษาเพื่อหลีกเลี่ยงการสูญเสียรายได้
  1. ซัพพอร์ตเวกเตอร์ (Support Vector Machine : SVM)

เป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่ม ดังสมการที่ 1

………(1)

เมื่อ

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ดังสมการที่ 2

………(2)

……….(3)

กลุ่มข้อมูล x จากสมการที่ 3 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงแต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (kernel function) ดังสมการที่ 4

………(4)

เมื่อ แทน ฟังก์ชันสาหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่

อยู่ในรูปเชิงเส้นสามารถแบ่งแยกได้

แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature space ไปสู่ output space

**b** แทน ค่าโน้มเอียง (bias)

แทน ซัพพอร์ตเวกเตอร์ โดย

แทน จำนวนซัพพอร์ตเวกเตอร์

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลใน feature space ที่ใกล้ที่สุดดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุดและเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ต เวกเตอร์” (support vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์

ดังนั้นจึงต้องมีการก าหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร (slack variable) ดังสมการที่ 5 และ 6 ดังนี้

เมื่อกำหนดให้ y = 1 ………(5)

เมื่อกำหนดให้ y = -1 ………(6)

จากการก าหนดค่า > 0 ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมทชีนบรรลุวัตถุประสงค์ใน2 ส่วนคือการเพิ่มระยะแบ่งแยกให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังสมการที่ 7

Minimize ………(7)

โดยที่ :

0,i = 1,2,…,N

และมีเคอร์เนลฟังก์ชัน ที่นิยมใช้อยู่ 3 ชนิดด้วยกันคือ

* โพลิโนเมียล (polynomial) :

* เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF) :

* ซิกมอยด์ (sigmoid) :
  1. Random Forest

Random Forest คืออัลกอริธึมการเรียนรู้ของเครื่องที่ใช้กันทั่วไปซึ่งรวมเอาต์พุตของแผนผังการตัดสินใจหลาย ๆ อันเพื่อให้ได้ผลลัพธ์เดียว ความสะดวกในการใช้งานและความยืดหยุ่นได้กระตุ้นให้เกิดการนำไปใช้ เนื่องจากสามารถจัดการกับปัญหาการจัดหมวดหมู่และการถดถอยได้

เนื่องจากแบบจำลองฟอเรสต์แบบสุ่มประกอบด้วยต้นไม้การตัดสินใจหลายต้น จึงควรเริ่มต้นด้วยการอธิบายอัลกอริทึมแผนผังการตัดสินใจโดยสังเขป โครงสร้างการตัดสินใจเริ่มต้นด้วยคำถามพื้นฐาน สามารถถามคำถามหลายชุดเพื่อหาคำตอบคำถามเหล่านี้ประกอบขึ้นเป็นโหนดการตัดสินใจในแผนผัง ซึ่งทำหน้าที่เป็นวิธีการแยกข้อมูลคำถามแต่ละข้อช่วยให้แต่ละคนตัดสินใจขั้นสุดท้ายได้ ซึ่งจะแสดงด้วยโหนดปลายใบ การสังเกตที่เหมาะกับเกณฑ์จะเป็นไปตามสาขา "ใช่" และที่ไม่เป็นไปตามเกณฑ์อื่น แผนผังการตัดสินใจพยายามค้นหาการแยกส่วนที่ดีที่สุดเพื่อจัดชุดข้อมูลย่อย และโดยทั่วไปแล้วต้นไม้เหล่านี้จะได้รับการฝึกอบรมผ่านอัลกอริทึมการจัดหมวดหมู่และแผนผังการถดถอย (CART) เมตริก

แม้ว่าแผนผังการตัดสินใจจะเป็นอัลกอริทึมการเรียนรู้แบบมีผู้สอนทั่วไป แต่ก็มีแนวโน้มที่จะเกิดปัญหาได้ เช่น ความลำเอียงและการเกินพอดี อย่างไรก็ตาม เมื่อต้นไม้การตัดสินใจหลายต้นประกอบกันเป็นชุดในอัลกอริทึมของฟอเรสต์แบบสุ่ม พวกมันทำนายผลลัพธ์ที่แม่นยำยิ่งขึ้น โดยเฉพาะอย่างยิ่งเมื่อต้นไม้แต่ละต้นไม่มีความสัมพันธ์กัน

วิธีการเรียนรู้ทั้งมวลประกอบด้วยชุดตัวแยกประเภทเช่น ต้นไม้การตัดสินใจและการคาดคะเนจะถูกรวมเข้าด้วยกันเพื่อระบุผลลัพธ์ที่ได้รับความนิยมสูงสุด วิธีการรวมที่รู้จักกันดีที่สุดคือการบรรจุถุงหรือที่เรียกว่าการรวมบูตสแตรปและการเพิ่มได้แนะนำวิธีการบรรจุถุง ในวิธีนี้ จะมีการเลือกตัวอย่างข้อมูลแบบสุ่มในชุดการฝึกด้วยการแทนที่ ซึ่งหมายความว่าสามารถเลือกจุดข้อมูลแต่ละจุดได้มากกว่าหนึ่งครั้ง หลังจากสร้างตัวอย่างข้อมูลหลายตัวอย่างแล้ว โมเดลเหล่านี้จะได้รับการฝึกอบรมโดยอิสระ และขึ้นอยู่กับประเภทของงาน เช่น การถดถอยหรือการจำแนกประเภท ค่าเฉลี่ยหรือส่วนใหญ่ของการคาดคะเนเหล่านั้นจะให้ค่าประมาณที่แม่นยำยิ่งขึ้น วิธีการนี้มักใช้เพื่อลดความแปรปรวนภายในชุดข้อมูลที่มีสัญญาณรบกวน

อัลกอริทึมของฟอเรสต์แบบสุ่มเป็นส่วนเสริมของวิธีการบรรจุถุงเนื่องจากใช้ทั้งการบรรจุถุงและการสุ่มลักษณะพิเศษเพื่อสร้างฟอเรสต์ของการตัดสินใจที่ไม่สัมพันธ์กัน การสุ่มคุณลักษณะ หรือที่เรียกว่าคุณลักษณะบรรจุถุงหรือ “ วิธีการสุ่มพื้นที่ย่อย ” สร้างชุดย่อยของคุณลักษณะแบบสุ่ม ซึ่งรับประกันความสัมพันธ์ที่ต่ำระหว่างแผนผังการตัดสินใจ นี่เป็นข้อแตกต่างที่สำคัญระหว่างต้นไม้ตัดสินใจและฟอเรสต์แบบสุ่ม ในขณะที่แผนผังการตัดสินใจพิจารณาการแบ่งคุณลักษณะที่เป็นไปได้ทั้งหมด ฟอเรสต์แบบสุ่มจะเลือกเฉพาะส่วนย่อยของคุณลักษณะเหล่านั้นเท่านั้น

* 1. มันทำงานอย่างไร

อัลกอริทึมฟอเรสต์แบบสุ่มมีไฮเปอร์พารามิเตอร์หลักสามตัว ซึ่งจำเป็นต้องตั้งค่าก่อนการฝึก ซึ่งรวมถึงขนาดโหนด จำนวนทรี และจำนวนคุณลักษณะที่สุ่มตัวอย่าง จากตรงนั้น สามารถใช้ตัวจำแนกฟอเรสต์แบบสุ่มเพื่อแก้ปัญหาการถดถอยหรือการจำแนกประเภท

อัลกอริทึมของฟอเรสต์แบบสุ่มประกอบด้วยคอลเล็กชันของต้นไม้ตัดสินใจ และต้นไม้แต่ละต้นในชุดประกอบด้วยตัวอย่างข้อมูลที่ดึงมาจากชุดการฝึกที่มีการแทนที่ ซึ่งเรียกว่าตัวอย่าง Bootstrap จากตัวอย่างการฝึกนั้น หนึ่งในสามของตัวอย่างนั้นถูกกันไว้เป็นข้อมูลทดสอบ ซึ่งเรียกว่าตัวอย่างที่แกะออกจากถุง (oob) ซึ่งเราจะกลับมาพูดถึงในภายหลัง อีกตัวอย่างหนึ่งของการสุ่มจะถูกแทรกผ่านคุณสมบัติการบรรจุถุง เพิ่มความหลากหลายให้กับชุดข้อมูลและลดความสัมพันธ์ระหว่างแผนผังการตัดสินใจ การกำหนดคำทำนายจะแตกต่างกันไปขึ้นอยู่กับประเภทของปัญหา สำหรับงานการถดถอย ต้นไม้การตัดสินใจแต่ละรายการจะถูกหาค่าเฉลี่ย และสำหรับงานการจำแนกประเภท การลงคะแนนเสียงส่วนใหญ่—เช่น ตัวแปรตามหมวดหมู่ที่พบบ่อยที่สุด จะให้คลาสที่คาดการณ์ไว้ สุดท้าย ตัวอย่าง oob จะใช้สำหรับการตรวจสอบข้าม

รูปภาพประกอบด้วย แผนภูมิ

คำอธิบายที่สร้างโดยอัตโนมัติ

**ภาพที่ 2.9** (ที่มา : https://www.ibm.com/topics/ramdom-forest)

* 1. ประโยชน์และความท้าทายของ Random Forest

มีข้อดีและความท้าทายที่สำคัญหลายประการที่ random forest นำเสนอเมื่อใช้สำหรับการจำแนกประเภทหรือปัญหาการถดถอย

ประโยชน์หลัก

* ลดความเสี่ยงของการ overfitting: แผนผังการตัดสินใจจะเสี่ยงต่อการ overfitting เนื่องจากพวกมันมีแนวโน้มที่จะพอดีกับตัวอย่างทั้งหมดภายในข้อมูลการฝึกอบรม อย่างไรก็ตาม เมื่อมีจำนวนต้นไม้การตัดสินใจที่แข็งแกร่งในฟอเรสต์แบบสุ่ม ตัวแยกประเภทจะไม่ทำให้โมเดลมากเกินไป เนื่องจากค่าเฉลี่ยของต้นไม้ที่ไม่สัมพันธ์กันจะลดความแปรปรวนโดยรวมและข้อผิดพลาดในการทำนาย
* ให้ความยืดหยุ่น: เนื่องจากฟอเรสต์แบบสุ่มสามารถจัดการทั้งงานการถดถอยและการจำแนกประเภทด้วยความแม่นยำสูง จึงเป็นวิธีที่นิยมในหมู่นักวิทยาศาสตร์ข้อมูล การรวมคุณสมบัติยังทำให้ตัวจำแนกประเภทฟอเรสต์แบบสุ่มเป็นเครื่องมือที่มีประสิทธิภาพในการประมาณค่าที่ขาดหายไป เนื่องจากจะรักษาความถูกต้องเมื่อข้อมูลบางส่วนขาดหายไป
* ง่ายต่อการกำหนดความสำคัญของคุณลักษณะ: Random Forest ทำให้ง่ายต่อการประเมินความสำคัญของตัวแปรหรือการมีส่วนร่วมกับแบบจำลอง มีสองสามวิธีในการประเมินความสำคัญของฟีเจอร์ ความสำคัญของ Gini และค่าเฉลี่ยการลดลงของสิ่งเจือปน (MDI) มักจะใช้เพื่อวัดว่าความแม่นยำของแบบจำลองลดลงมากเพียงใดเมื่อไม่รวมตัวแปรที่กำหนด อย่างไรก็ตาม ความสำคัญของการเปลี่ยนรูปหรือที่เรียกว่าค่าเฉลี่ยความแม่นยำในการลดลง (MDA) เป็นอีกหนึ่งมาตรการที่สำคัญ MDA ระบุค่าความแม่นยำที่ลดลงโดยเฉลี่ยโดยการสุ่มเปลี่ยนค่าคุณลักษณะในตัวอย่าง oob
  1. ความท้าทายที่สำคัญ
* กระบวนการที่ใช้เวลานาน: เนื่องจากอัลกอริทึมของฟอเรสต์แบบสุ่มสามารถจัดการกับชุดข้อมูลขนาดใหญ่ได้ จึงสามารถคาดการณ์ได้แม่นยำมากขึ้น แต่อาจประมวลผลข้อมูลได้ช้า เนื่องจากเป็นการประมวลผลข้อมูลสำหรับแผนผังการตัดสินใจแต่ละรายการ
* ต้องการทรัพยากรมากขึ้น: เนื่องจากฟอเรสต์สุ่มประมวลผลชุดข้อมูลที่มีขนาดใหญ่ขึ้น พวกเขาจึงต้องใช้ทรัพยากรมากขึ้นในการจัดเก็บข้อมูลนั้น
* ซับซ้อนกว่า: การคาดคะเนของแผนผังการตัดสินใจแบบเดี่ยวจะตีความได้ง่ายกว่าเมื่อเทียบกับแบบฟอเรสต์
  1. ต้นไม้แห่งการตัดสินใจ (Decision Trees : DT)

แผนผังการตัดสินใจเป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนแบบไม่มีพารามิเตอร์ ซึ่งใช้สำหรับทั้งงานการจัดหมวดหมู่และการถดถอย มีโครงสร้างแบบลำดับชั้น

รูปภาพประกอบด้วย แผนภาพ

คำอธิบายที่สร้างโดยอัตโนมัติ

**ภาพที่ 2.10** (ที่มา : <https://www.ibm.com/topics/decision-trees>)

จากแผนภาพด้านบน ต้นไม้การตัดสินใจเริ่มต้นด้วยโหนดรูท ซึ่งไม่มีสาขาใดๆ ที่เข้ามา สาขาขาออกจากรูทโหนดจากนั้นฟีดไปยังโหนดภายในหรือที่เรียกว่าโหนดการตัดสินใจ ตามคุณสมบัติที่มีอยู่ โหนดทั้งสองประเภทจะทำการประเมินเพื่อสร้างชุดย่อยที่เป็นเนื้อเดียวกัน ซึ่งแสดงโดยโหนดลีฟหรือโหนดเทอร์มินัล โหนดปลายสุดแสดงถึงผลลัพธ์ที่เป็นไปได้ทั้งหมดภายในชุดข้อมูล ตัวอย่างเช่น สมมติว่าคุณกำลังพยายามประเมินว่าคุณควรไปเล่นเซิร์ฟหรือไม่ คุณอาจใช้กฎการตัดสินใจต่อไปนี้เพื่อตัดสินใจเลือก

รูปภาพประกอบด้วย แผนภาพ

คำอธิบายที่สร้างโดยอัตโนมัติ

**ภาพที่ 2.11** (ที่มา : https://www.ibm.com/topics/decision-trees)

โครงสร้างผังงานประเภทนี้ยังสร้างการแสดงการตัดสินใจที่ย่อยง่าย ช่วยให้กลุ่มต่างๆ ทั่วทั้งองค์กรเข้าใจได้ดีขึ้นว่าเหตุใดการตัดสินใจจึงเกิดขึ้น

การเรียนรู้แผนผังการตัดสินใจใช้กลยุทธ์การแบ่งแยกและพิชิตโดยการค้นหาอย่างละโมบเพื่อระบุจุดแยกที่เหมาะสมที่สุดภายในต้นไม้ จากนั้นกระบวนการแยกนี้จะถูกทำซ้ำในลักษณะจากบนลงล่าง วนซ้ำจนกว่าเรคคอร์ดทั้งหมดหรือส่วนใหญ่จะถูกจัดประเภทภายใต้ป้ายกำกับระดับเฉพาะ ไม่ว่าจุดข้อมูลทั้งหมดจะถูกจัดประเภทเป็นชุดที่เป็นเนื้อเดียวกันหรือไม่นั้นขึ้นอยู่กับความซับซ้อนของแผนผังการตัดสินใจเป็นส่วนใหญ่ ต้นไม้ขนาดเล็กสามารถรับโหนดใบบริสุทธิ์ได้ง่ายกว่า นั่นคือจุดข้อมูลในคลาสเดียว อย่างไรก็ตาม เมื่อต้นไม้มีขนาดโตขึ้น การรักษาความบริสุทธิ์นี้ก็ยิ่งยากขึ้นเรื่อยๆ และมักจะส่งผลให้ข้อมูลตกหล่นภายในทรีย่อยที่กำหนดน้อยเกินไป เมื่อสิ่งนี้เกิดขึ้น จะเรียกว่าการกระจายตัวของข้อมูล และมักจะนำไปสู่การใช้มากเกินไปเป็นผลให้ต้นไม้ตัดสินใจชอบต้นไม้ขนาดเล็ก ซึ่งสอดคล้องกับหลักการของ parsimony ใน Occam's Razor คือ "ไม่ควรเพิ่มจำนวนเอนทิตีเกินความจำเป็น" ต้นไม้การตัดสินใจควรเพิ่มความซับซ้อนในกรณีที่จำเป็นเท่านั้น เนื่องจากคำอธิบายที่ง่ายที่สุดมักจะดีที่สุด เพื่อลดความซับซ้อนและป้องกันการใส่มากเกินไป มักใช้การตัดแต่งกิ่ง นี่เป็นกระบวนการที่จะลบสาขาที่แบ่งคุณลักษณะที่มีความสำคัญต่ำออก ความพอดีของแบบจำลองสามารถประเมินผ่านกระบวนการตรวจสอบข้ามได้ อีกวิธีหนึ่งที่แผนผังการตัดสินใจสามารถรักษาความถูกต้องได้คือการสร้างกลุ่มผ่านอั ลกอริธึมฟ อเรสต์แบบสุ่ม ลักษณนามนี้ทำนายผลลัพธ์ได้แม่นยำกว่า โดยเฉพาะอย่างยิ่งเมื่อต้นไม้แต่ละต้นไม่มีความสัมพันธ์กัน

1. ประเภทของต้นไม้การตัดสินใจ

อัลกอริทึมของ Hunt ซึ่งพัฒนาขึ้นในทศวรรษที่ 1960 เพื่อจำลองการเรียนรู้ของมนุษย์ในด้านจิตวิทยา เป็นรากฐานของอัลกอริทึมแผนผังต้นไม้การตัดสินใจที่เป็นที่นิยมมากมาย เช่น

* **ID3**: Ross Quinlan ได้รับเครดิตในการพัฒนา ID3 ซึ่งเป็นชวเลขของ "Iterative Dichotomiser 3" อัลกอริทึมนี้ใช้ประโยชน์จากเอนโทรปีและการรับข้อมูลเป็นเมตริกเพื่อประเมินการแยกผู้สมัคร งานวิจัยบางส่วนของ Quinlan เกี่ยวกับอัลกอริทึมนี้จากปี 1986
* **C4.5**: อัลกอริทึมนี้ถือเป็นการทำซ้ำในภายหลังของ ID3 ซึ่งพัฒนาโดย Quinlan สามารถใช้การรับข้อมูลหรือเพิ่มอัตราส่วนเพื่อประเมินจุดแยกภายในแผนผังการตัดสินใจ
* **CART**: คำว่า CART เป็นคำย่อของ "ต้นไม้การจำแนกประเภทและการถดถอย" และได้รับการแนะนำโดย Leo Breiman อัลกอริทึมนี้มักใช้สิ่งเจือปน Gini เพื่อระบุคุณลักษณะในอุดมคติที่จะแยกออก สิ่งเจือปน Gini วัดว่าแอตทริบิวต์ที่สุ่มเลือกนั้นจัดประเภทผิดบ่อยเพียงใด เมื่อประเมินโดยใช้สิ่งเจือปน Gini ค่าที่ต่ำกว่าจะเหมาะสมกว่า

1. วิธีเลือกแอตทริบิวต์ที่ดีที่สุดในแต่ละโหนด

แม้ว่าจะมีหลายวิธีในการเลือกแอตทริบิวต์ที่ดีที่สุดในแต่ละโหนด แต่สองวิธีคือ การรับข้อมูลและการไม่บริสุทธิ์ของ Gini ทำหน้าที่เป็นเกณฑ์การแยกที่เป็นที่นิยมสำหรับแบบจำลองแผนผังการตัดสินใจ ช่วยในการประเมินคุณภาพของแต่ละเงื่อนไขการทดสอบและความสามารถในการจำแนกกลุ่มตัวอย่างในชั้นเรียนได้ดีเพียงใด

* + - เอนโทรปีและการรับข้อมูล

เป็นการยากที่จะอธิบายการรับข้อมูลโดยไม่พูดถึงเรื่องเอนโทรปีก่อน เอนโทรปีเป็นแนวคิดที่เกิดจากทฤษฎีสารสนเทศซึ่งใช้วัดความไม่บริสุทธิ์ของค่าตัวอย่าง ถูกกำหนดโดยสูตรต่อไปนี้ โดยที่

สูตรเอนโทรปี:

S แทน ชุดข้อมูลที่คำนวณเอนโทรปี

C แทน คลาสในเซต S

p(c) แสดงถึงสัดส่วนของจุดข้อมูลที่อยู่ในคลาส c ต่อจำนวนจุดข้อมูลทั้งหมดในชุด S

ค่าเอนโทรปีสามารถอยู่ระหว่าง 0 ถึง 1 หากตัวอย่างทั้งหมดในชุดข้อมูล S อยู่ในคลาสเดียวกัน ค่าเอนโทรปีจะเท่ากับศูนย์ ถ้าครึ่งหนึ่งของตัวอย่างถูกจำแนกเป็นคลาสหนึ่งและอีกครึ่งหนึ่งอยู่ในอีกคลาสหนึ่ง เอนโทรปีจะสูงสุดที่ 1 เพื่อเลือกคุณลักษณะที่ดีที่สุดเพื่อแยกและค้นหาแผนผังการตัดสินใจที่เหมาะสมที่สุด คุณลักษณะที่มีขนาดเล็กที่สุด ควรใช้จำนวนเอนโทรปี ข้อมูลที่ได้รับแสดงถึงความแตกต่างของเอนโทรปีก่อนและหลังการแยก

แอตทริบิวต์ที่กำหนด แอตทริบิวต์ที่ได้รับข้อมูลสูงสุดจะสร้างการแยกที่ดีที่สุดเนื่องจากทำหน้าที่ได้ดีที่สุดในการจำแนกข้อมูลการฝึกอบรมตามการจัดประเภทเป้าหมาย การรับข้อมูลมักจะแสดงด้วยสูตรต่อไปนี้ โดยที่

สูตรการรับข้อมูล :

a แสดงถึงแอตทริบิวต์หรือป้ายกำกับคลาสเฉพาะ

Entropy(S) คือเอนโทรปีของชุดข้อมูล S

|Sv|/ |S| แสดงถึงสัดส่วนของค่าใน ต่อจำนวนของค่าในชุดข้อมูล S

Entropyคือเอนโทรปีของชุดข้อมูล

* + - สิ่งเจือปน Gin

สิ่งเจือปน Gini คือความน่าจะเป็นของการจำแนกจุดข้อมูลแบบสุ่มในชุดข้อมูลอย่างไม่ถูกต้อง หากได้รับการติดป้ายตามการแจกแจงคลาสของชุดข้อมูล เช่นเดียวกับเอนโทรปี ถ้าตั้งค่า S คือบริสุทธิ์ ดังนั้น สิ่งเจือปนจะเป็นศูนย์ นี่แสดงโดยสูตรต่อไปนี้

สูตรสิ่งเจือปน Gini:

1. ข้อดีและข้อเสียของ Decision Tree

ในขณะที่แผนผังการตัดสินใจสามารถใช้ได้ในกรณีการใช้งานที่หลากหลายขั้นตอนวิธีอื่นๆ มักจะมีประสิทธิภาพดีกว่าขั้นตอนวิธี แผนผังการตัดสินใจ ที่กล่าวว่า ต้นไม้การตัดสินใจมีประโยชน์อย่างยิ่งสำหรับ การทำเหมืองข้อมูล และงานค้นหาความรู้ เรามาสำรวจประโยชน์และความท้าทายที่สำคัญของการใช้แผนผังการตัดสินใจเพิ่มเติมด้านล่าง

**ข้อดี**

* + ง่ายต่อการตีความ: ตรรกะบูลีนและการแสดงภาพต้นไม้การตัดสินใจทำให้เข้าใจและบริโภคได้ง่ายขึ้น ลักษณะลำดับชั้นของแผนผังการตัดสินใจยังทำให้ง่ายต่อการดูว่าแอตทริบิวต์ใดที่สำคัญที่สุด ซึ่งอัลกอริทึมอื่นๆ อาจไม่ชัดเจนเสมอไป
  + ไม่จำเป็นต้องเตรียมข้อมูลเพียงเล็กน้อยหรือไม่มีเลย: แผนผังการตัดสินใจมีลักษณะเฉพาะหลายอย่าง ซึ่งทำให้มีความยืดหยุ่นมากกว่าตัวแยกประเภทอื่นๆ สามารถจัดการประเภทข้อมูลต่างๆ เช่น ค่าที่ไม่ต่อเนื่องหรือค่าต่อเนื่อง และค่าต่อเนื่องสามารถแปลงเป็นค่าหมวดหมู่ได้โดยใช้เกณฑ์ นอกจากนี้ยังสามารถจัดการค่าที่มีค่าขาดหายไป ซึ่งอาจเป็นปัญหาสำหรับตัวแยกประเภทอื่นๆ
  + ยืดหยุ่นมากขึ้น: ต้นไม้การตัดสินใจสามารถใช้ประโยชน์จากทั้งการจัดหมวดหมู่และงานการถดถอย ทำให้มีความยืดหยุ่นมากกว่าอัลกอริทึมอื่นๆ นอกจากนี้ยังไม่ไวต่อความสัมพันธ์พื้นฐานระหว่างแอตทริบิวต์ ซึ่งหมายความว่าหากตัวแปรสองตัวมีความสัมพันธ์กันสูง อัลกอริทึมจะเลือกเพียงหนึ่งในคุณสมบัติที่จะแยกออก

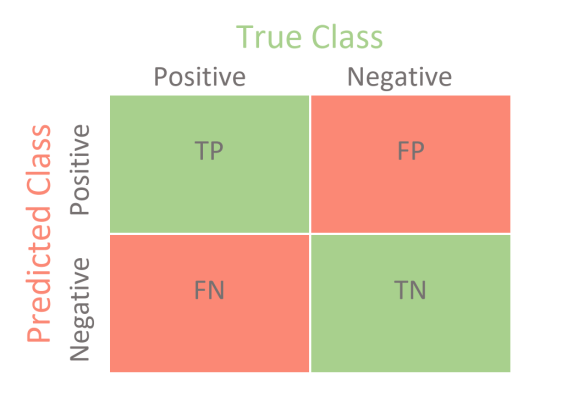
**ข้อเสีย**

* + มีแนวโน้มที่จะ overfitting: โครงสร้างการตัดสินใจที่ซับซ้อนมีแนวโน้มที่จะ overfit และไม่สามารถสรุปได้ดีกับข้อมูลใหม่ สถานการณ์นี้สามารถหลีกเลี่ยงได้ผ่านกระบวนการก่อนการตัดแต่งกิ่งหรือหลังการตัดแต่งกิ่ง การตัดแต่งก่อนการตัดแต่งจะหยุดการเจริญเติบโตของต้นไม้เมื่อมีข้อมูลไม่เพียงพอ ในขณะที่การตัดแต่งภายหลังการตัดแต่งจะลบทรีย่อยที่มีข้อมูลไม่เพียงพอหลังจากการสร้างต้นไม้
  + ตัวประมาณค่าความแปรปรวนสูง: ความแปรผันเล็กน้อยภายในข้อมูลสามารถสร้างแผนผังการตัดสินใจที่แตกต่างกันมาก การใส่ถุงหรือการหาค่าเฉลี่ยของค่าประมาณ สามารถเป็นวิธีการลดความแปรปรวนของแผนผังการตัดสินใจ อย่างไรก็ตาม วิธีการนี้มีข้อจำกัดเนื่องจากอาจนำไปสู่ตัวทำนายที่สัมพันธ์กันสูง
  + มีราคาแพงกว่า: เนื่องจากแผนผังการตัดสินใจใช้วิธีค้นหาแบบละโมบในระหว่างการก่อสร้าง การฝึกอบรมจึงมีราคาแพงกว่าเมื่อเทียบกับอัลกอริทึมอื่นๆ
  + ไม่รองรับอย่างสมบูรณ์ใน scikit-learn: Scikit-learn เป็นไลบรารีการเรียนรู้ของเครื่องยอดนิยมที่ใช้ Python แม้ว่าไลบรารีนี้จะมี โมดูล Decision Tree) การใช้งานในปัจจุบันไม่สนับสนุนตัวแปรที่เป็นหมวดหมู่
  1. การประเมินแบบจำลอง

การประเมินผลนั้นจำเป็นกับการทำ Machine Learning อย่างมาก เพราะมันเป็นสิ่งที่เอาไว้วัดว่าโมเดลนั้น ๆ ทำงานเป็นอย่างไร ประเมินผลว่าทำงานได้น่าพึงพอใจแล้วหรือไม่ รวมถึงยังไว้ใช้เปรียบเทียบระหว่างโมเดลเพื่อตอบคำถามที่ว่าเราควรเลือกใช้โมเดลไหนดี

* + - 1. Confusion Matrix

มันคือตารางวัดประสิทธิภาพของโมเดล classification ที่มีขนาด C\*C โดยที่ค่า C ก็คือประเภทของผลลัพธ์ที่เราได้ (Classifiers) เช่น ถ้าหากเราสร้างโมเดลในการทำนายว่า (ใช่, ไม่ใช่) Confusion Matrix ของเราก็จะมีขนาด 2\*2 โดยลักษณะตารางของ Confusion Matrix จะแบ่งสองส่วนคือส่วนที่เป็นผลลัพธ์จริง (True/Actual Class) และผลลัพธ์ที่ได้จากการทำนายผลด้วยโมเดล (Predicted Class)



**ภาพที่ 2.12** ตัวอย่างตาราง Confusion Matrix ขนาด 2x2

(ที่มา : https://datascihaeng.medium.com)

* True Positive (TP): หมายถึงจำนวนข้อมูลที่เป็น positive และโมเดลสามารถทำนายออกมาได้ถูกต้องว่าเป็น positive
* True Negative (TN): หมายถึงจำนวนข้อมูลที่เป็น negative และโมเดลสามารถทำนายออกมาได้ถูกต้องว่าเป็น negative
* False Positive (FP): หมายถึงจำนวนข้อมูลที่เป็น negative แต่โมเดลทำนายออกมาว่าเป็น positive ปัญหานี้ในทางสถิตินิยมเรียกกันว่า Type I error
* False Negative (FN): หมายถึงจำนวนข้อมูลที่เป็น positive แต่โมเดลทำนายออกมาว่าเป็น negative ปัญหานี้ในทางสถิตินิยมเรียกกันว่า Type I error

นั่นแปลว่าข้อมูลที่ทายได้ถูกต้องนั้นจะต้องอยู่ในพื้นที่สีเขียว ซึ่งก็คือ True Negative, True Positive และข้อมูลที่โมเดลทายผิดจะไปอยู่ในพื้นที่สีแดง ซึ่งก็คือ False Negative, False Positive

Accuracy >> (TP + TN) / (TP + FP + TN + FN)

สำหรับวัดว่าโมเดลนั้นทำนายได้ถูกต้องแค่ไหน โดยนำค่าที่ทำนายได้ถูกต้องทั้งหมดหารด้วยค่าทั้งหมดที่ให้โมเดลทำนาย แต่ปัญหามีอยู่ว่าถ้าเป็นการจำแนกประเภทที่มากกว่า 2 คลาสขึ้นไปวิธีนี้จะไม่สามารถวัดค่าความแม่นยำบนแต่ละคลาสได้ และถ้าชุดข้อมูลไม่บาลานซ์กันนั้นก็จะส่งผลให้ผลลัพธ์ของวิธีนี้ผิดพลาดได้ เช่นว่า ถ้าชุดข้อมูลมีทุเรียนหมอนทองมากกว่าทุเรียนก้านยาว แน่นอนว่ามันก็ต้องทำนายเก่งบนทุเรียนหมอนทองและส่งผลให้ Accuracy สูงตามไปด้วยเพราะเฉลยเองก็มีแต่หมอนทองเป็นส่วนใหญ่ จึงทำให้ไม่ทราบเลยว่ามันสามารถแยกแยะได้จริงหรือไม่

Precision/Recall >> Precision = TP/(TP + FP), Recall = TP/(TP + FN)

สองอย่างนี้จะต้องรายงานด้วยกันเสมอ โดยมีความหมายคือ Precision มองในมุมว่าโมเดลทำนายได้แม่นยำแค่ไหนในค่าที่สนใจ เช่น ในผลการทำนายที่เป็นหมอนทองทั้งหมดโมเดลทำนายว่าเป็นหมอนทองได้อย่างถูกต้องมากน้อยแค่ไหน

Recall

มองในมุมว่าโมเดลทำนายได้ถูกต้องแค่ไหน เช่น ในผลการทำนายที่ทั้งถูกและผิดโมเดลทำนายว่าเป็นหมอนทองได้อย่างถูกต้องมากน้อยแค่ไหน

F1-Score > 2 \* ((precision x recall) / (precision + recall))

โดยการที่ต้องเลือกว่าจะใช้โมเดลไหนดีจำเป็นจะต้องมีค่าที่ใช้สำหรับตัดสินใจได้ ดังนั้นจึงนำ Precision และ Recall มาหาค่าเฉลี่ยแบบฮาร์โมนิก ซึ่งยิ่ง 2 ค่านี้สูงก็จะทำให้ F1 สูงตามไปด้วย

ROC curve > TPR = TP / (TP + FN), FPR = FP / (FP + TN)

เป็นการพล็อตระหว่างค่าอัตราส่วนที่ทำนายได้อย่างถูกต้อง กับ ค่าอัตราส่วนที่ทำนายผิดแต่ตรงกับค่าคาดหวัง โดยไว้ใช้สำหรับ Classification แบบที่มีแค่ 2 คลาส ใช้เพื่อเลือกค่า Threshold ให้กับโมเดล

* + - 1. ROC-AUC Curve

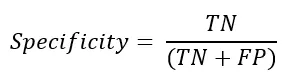
ROC (Receiver Operating Characteristic) Curve เป็นเส้นที่ใช้วัดถึงประสิทธิภาพของโมเดลแบบ Classification ว่าสามารถทำนายประเด็นที่สนใจได้อย่างแม่นยำขนาดไหน (โดยทั่วไปนิยมวัดประสิทธิภาพของโมเดลแบบ Binary) ส่วน AUC ก็คือ Area Under ROC Curve หรือพื้นที่ใต้กราฟ ROC นั่นเอง

* + - 1. Sensitivity & Specificity

ค่าความอ่อนไหว (Sensitivity) มันก็คือ“สัดส่วนความถูกต้องของผลการทำนายที่เราสนใจ (True Positive)” โดยมีค่าเท่ากับ

Text

Description automatically generated

ส่วนค่าความจำเพาะ (Specificity) ก็คือ“สัดส่วนความถูกต้องของผลการทำนายที่เราไม่ได้สนใจ (True Negative)” โดยมีค่าเท่ากับ

1. **งานวิจัยที่เกี่ยวข้อง**
2. คุณภาพของน้ำดื่มโดยใช้การเรียนรู้ของเครื่องเทคนิค (The Quality of Drinkable

Water using Machine Learning Techniques)

ในโครงงานวิจัยนี้พูดถึงปัญหาของมลพิษทางน้ำที่กำลังเป็นอันตรายต่อมนุษย์อย่างรุนแรง โดยมีสาเหตุมาจากปุ๋ยเคมีและยาฆ่าแมลงที่ไหลลงสู่น้ำโดยไม่ผ่านการบำบัดน้ำ น้ำเสียและน้ำทิ้งที่มาจากโรงงานอุตสาหกรรม โดยเป้าหมายของโครงงานนี้คือการพัฒนาวิธีการคำนาณที่มีความสามารถในการประมาณค่าคูณลักษณะคุณภาพของน้ำดื่มเพื่อลดขั้นตอนและค่าใช้จ่ายที่เกี่ยวข้องกับการวัดค่าสิ่งต่างๆในน้ำ

ในโครงงานนี้จะเก็บค่าที่วัดในน้ำทั้งหมด 9 ค่า เพื่อใช้ในการตรวจสอบคุณภาพของน้ำว่าดื่มได้หรือไม่ โดยใช้ Machine Learning Algorithms คือ Random Forest , Deep neural network , Support vector machine , Gaussian naive bayes , Artificial neural network

ผลสรุปของงานวิจัยน้ำที่ดื่มได้จากการตรวจสอบแล้วมีทั้งหมด 89.71% น้ำที่ดื่มไม่ได้มี 10.29% โดยที่โครงข่ายประสาทเทียมได้รับความแม่นยำ 98.12% มีข้อผิดพลาดอยู่ที่ 0.75%

1. แบบจำลองการคาดการณ์ความสามารถในการผลิตน้ำโดยใช้การเรียนรู้ของเครื่อง

โดยการใช้เทคนิคการสุ่มตัวอย่างแบบกลุ่มน้อยสังเคราะห์และAI ที่อธิบายได้ (A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI)

โครงงานนี้พูดถึงปัญหาของน้ำที่ไม่ปลอดภัยนั้นจะมีผลกระทบต่อสุขภาพเป็นผลมาจากอุตสาหกรรมที่ปล่อยสารมลพิษลงสู่น้ำ และขยะที่ผู้คนทิ้งลงน้ำเพราะเหตุนี้จึงทำให้น้ำปนเปื้อนและทำให้เกิดโรคต่างๆโดยเป้าหมายของโครงงานนี้คือประเมินน้ำเบื้องต้นเพื่อกรองคุณภาพน้ำและลดค่าใช้ง่ายที่ต้องเข้าห้องปฏิบัติการที่ต้องใช้ค่าใช้ง่ายที่สูง

ในโครงงานนี้จะจะเก็บค่าที่ใช้วัดในน้ำทั้งหมด 9 ค่า เพื่อใช้ตรวจสอบคุณภาพของน้ำ โดยใช้ Machine Learning Algorithms ที่ใช้ในการตรวจสอบ คือ Random Forest , Gradient Boost , Decision Tree , AdaBoost , Support Vector

ผลสรุปของโครงงานนี้ มีความแตกต่างในระดับ TDS ถึง 40 เท่า สำหรับน้ำดื่ม ตัวอย่างน้ำที่ปลอดภัยสำหรับคลอรามีน คือ 2.72% ตัวอย่างน้ำที่ปลอดภัยสำหรับซัลเฟตมีเพียง 1.77% ตัวอย่างน้ำที่ปลอดภัยสำหรับคาร์บอน คือ 90.57% ตัวอย่างน้ำที่ปลอดภัยสำหรับไตรฮาโลมีเทนคือ 81.62% ตัวอย่างน้ำที่ปลอดภัยต่อการกัดกร่อนคือ 90.42% ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างคุณลักษณะต่างๆอยู่ในระดับที่ ต่ำมาก

1. โมเดลแมชชีนเลิร์นนิงที่ขับเคลื่อนด้วยข้อมูลอย่างมีประสิทธิภาพสำหรับน้ำในการ

ทำนายคุณภาพ (Efficient Data Driven Machine Learning Models for Water Quality Prediction)

โครงงานนี้พูดถึงความสำคัญของน้ำ และพูดถึงการเพิ่มขั้นของมลพิษทางสิ่งแวดล้อมที่มาจากกิจกรรมของมนุษย์คุณภาพน้ำจึงได้รับผลกระทบโดยตรงจากของเสียต่างๆ ในการจัดประเภทของน้ำดื่มจำเป็นต้องมีน้ำตามข้อกำหนด ในการทดสอบคุณภาพน้ำทางเคมีจะทดสอบโดยการทดสอบจุลชีววิทยาในน้ำ

ในโครงงานนี้จะนำชุดข่อมูลที่มีค่าที่ได้จากการวัดค่าในน้ำมาทั้งหมด 20 ค่า เพื่อใช้ตรวจสอบคุณภาพของน้ำ โดยที่ในโครงงานนี้จะเน้นไปที่การประมวลผลข้อมูลล่วงหน้า การวิเคราะห์คุณลักษณะ โดยใช้ Machine Learning Algorithms ที่ใช้ในการตรวจสอบ คือ Ensemble ML algorithms , Bagging , RotF built upon decision trees , RF , AdaBoostM1 , Voting and Stacking were exploited. , ANN

ผลสรุปของโครงงานแสดงให้เห็นว่าแบบจำลองการทดลองการจำแนกประเภทแบบเรียงซ้อนหลังจาก SMOTE ที่มีการตรวจสอบความถูกต้องข้าม 10 เท่ามีประสิทธิภาพดีกว่าแบบจำลองอื่นๆ โดยมีความแม่นยำ และการเรียกคืน 98.1 % 100 % และ 98.1 % ตามลำดับ และ AUC เท่ากับ 99.99 %

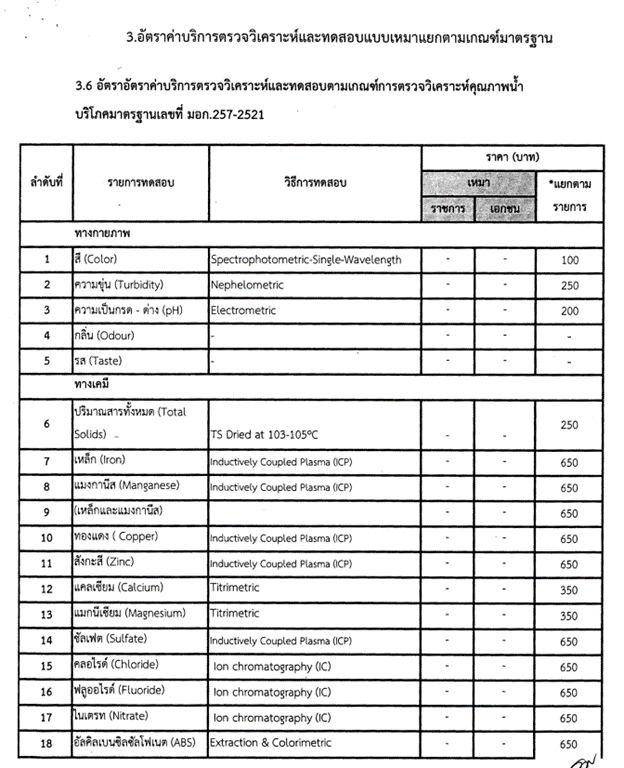
1. **ขั้นตอนการตรวจน้ำดื่ม หรือน้ำบริโภค**

น้ำบริโภค หมายถึง น้ำประปา น้ำบ่อบาดาล น้ำบ่อตื้น น้ำฝน และน้ำดื่มบรรจุขวด

* 1. ทดสอบคุณภาพน้ำบริโภคเบื้องต้นด้วยเครื่องมืออย่างง่าย ตรวจสอบโดยการวัดค่าคลอรีนอิสระคงเหลือในน้ำ
  2. เก็บตัวอย่างน้ำ ตรวจสอบคุณภาพตามเกณฑ์มาตรฐานคุณภาพน้ำประปาดื่มได้ กรมอนามัย ทางห้องปฏิบัติการ (20 พารามิเตอร์) จำนวน 2 ครั้ง โดยครั้งที่ 1และครั้งที่ 2 เป็นจุดเดียวกัน และสุ่มเก็บตัวอย่างน้ำ ตามจุดที่กำหนดอัตราค่าบริการวิเคราะห์และทดสอบคุณภาพน้ำบริโภค

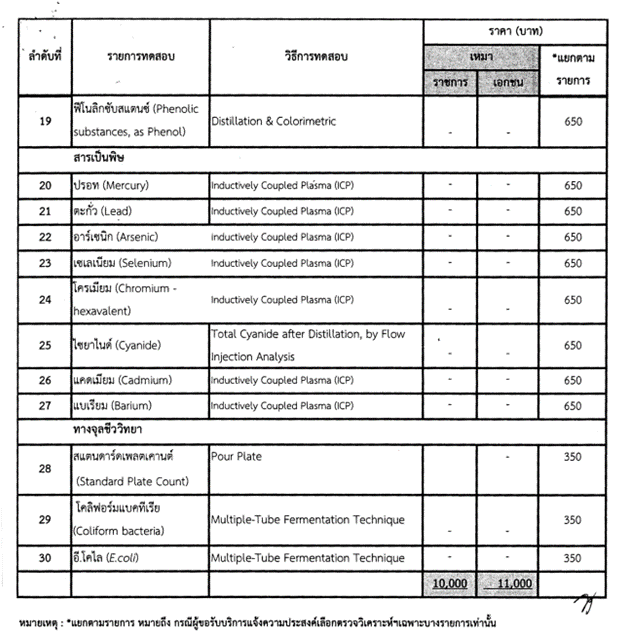
งบประมาณค่าบริการวิเคราะห์คุณภาพน้ำบริโภค มอก.257-2521

1. ค่าตรวจวิเคราะห์ สำหรับหน่วนงานภาครัฐ เหมาจ่าย เป็นเงิน 10,000 บาท
2. ค่าตรวจวิเคราะห์ สำหรับเอกชน เหมาจ่าย เป็นเงิน 11,000 บาท



**ภาพที่ 2.13** อัตราค่าบริการวิเคราะห์และทดสอบคุณภาพน้ำบริโภค มอก.257-2521

(ที่มา : https://rldc.anamai.moph.go.th/th/water-quality-analysis-service-rates/66285#)



ภาพที่ 2.14 อัตราค่าบริการวิเคราะห์และทดสอบคุณภาพน้ำบริโภค มอก.257-2521

(ที่มา : [https://rldc.anamai.moph.go.th/th/water-quality-analysis-service-rates/66285#](https://rldc.anamai.moph.go.th/th/water-quality-analysis-service-rates/66285))

การทดสอบเฉพาะสารเคมีในน้ำ

|  |  |  |
| --- | --- | --- |
| ลำดับ | รายการที่ทดสอบ | อัตราค่าทดสอบ  (บาท) |
| 1 | ความเป็นกรด-ด่าง (pH) | 100 |
| 2 | ความกระด้างทั้งหมด (Total Hardness as ) | 100 |
| 3 | ของแข็งทั้งหมด (Total solids) | 200 |
| 4 | คลอไรด์ (Chloride) | 200 |
| 5 | ซัลเฟต (Sulfate) | 300 |
| 6 | การนำไฟฟ้า (Conductivity) | 100 |
| 7 | สารอินทรีย์คาร์บอนทั้งหมด (Total organic carbon) | 1500 |
| 8 | ไตรฮาโลมีเทน (Trihalomethane)  คลอไรฟอร์ม (Chloroform)  โบรโมไดคลอโรมีเทน (Bromodichloromethane)  ไดโบรโมคลอโรมีเทน (Dibromochloromethane)  โบรโมฟอร์ม (Bromoform) | 3000 |
| 9 | ความขุ่น (Turbidity) | 100 |

**บทที่ 3  
การวิเคราะห์และออกแบบระบบ**

1. **ขอบเขตการดำเนินงานของระบบ**

ขอบเขตของระบบจะแบ่งออกเป็นระบบย่อย 2 ระบบได้แก่ ระบบการจัดการแบบจำลอง (Model Management System) และ ระบบการทำนาย (Prediction System) ที่ระบบการจัดการแบบจำลองจะจำกัดการเข้าใช้งานได้แค่ผู้ใช้ที่เป็นผู้ดูแลระบบเท่านั้น ซึ่งระบบนี้จะเป็นระบบที่ผู้ดูแลระบบสามารถ แก้ไข แบบจำลองบนฐานข้อมูลเพื่อใช้ในการทำนายข้อมูล ส่วนระบบการทำนายสามารถเข้าถึงที่ทุกผู้ใช้งานซึ่งเป็นระบบที่ให้ผู้ใช้กรอกข้อมูลของน้ำเพื่อที่จะทำนายข้อมูลนั้นออกมาว่าเป็นดื่มได้หรือ ดื่มไม่ได้

* + 1. **ขอบเขตการดำเนินงานของระบบการจัดการแบบจำลอง (Model Management System)**
  1. Admin ทำการสร้างโมเดล
  2. ใช้ Water Quality Dataset เพื่อนำไปใช้ใน machine learning เพื่อเพิ่มความแม่นยำและความเร็วในการจำแนกข้อมูลของโมเดล (ADITYA KADIWAL. 2022)
  3. นำ Dataset ที่ได้มาทำการแทนที่ในส่วนที่ไม่สมบูรณ์
  4. ทำการแบ่งข้อมูลออกเป็น 2 ส่วนโดยใช้ K-Fold Cross Validation คือ การที่เราแบ่งข้อมูลเป็นจำนวน K ส่วน โดยในแต่ละส่วนจะต้องมาจากการสุ่มเพื่อที่จะให้ข้อมูลกระจายเท่า ๆ กัน หลังจากที่แบ่งข้อมูลเป็นส่วน ๆ เรียบร้อย จึงจะนำ Training Data ไปหา Regression Line และทดสอบว่าข้อมูลชุดไหนที่ทำให้เส้นทำนายทำได้ดีที่สุดในการหา Training Data ซึ่งโดยวิธีดังกล่าวจึงแบ่ง data เป็น Training dataset และ Test dataset สำหรับการฝึกฝนแบบจำลอง
  5. ทำการฝึกฝนแบบจำลองโดยใช้ชุดข้อมูลสำหรับการฝึกฝน (Training Data) เพื่อเรียนรู้การแยกแยะระหว่างคุณสมบัติ (Features) และคำตอบ (Labels) โดยใช้วิธีการเรียนรู้ของเครื่อง Logistic Regression, Support Vector Machines, Random Forest และ Decision Tree

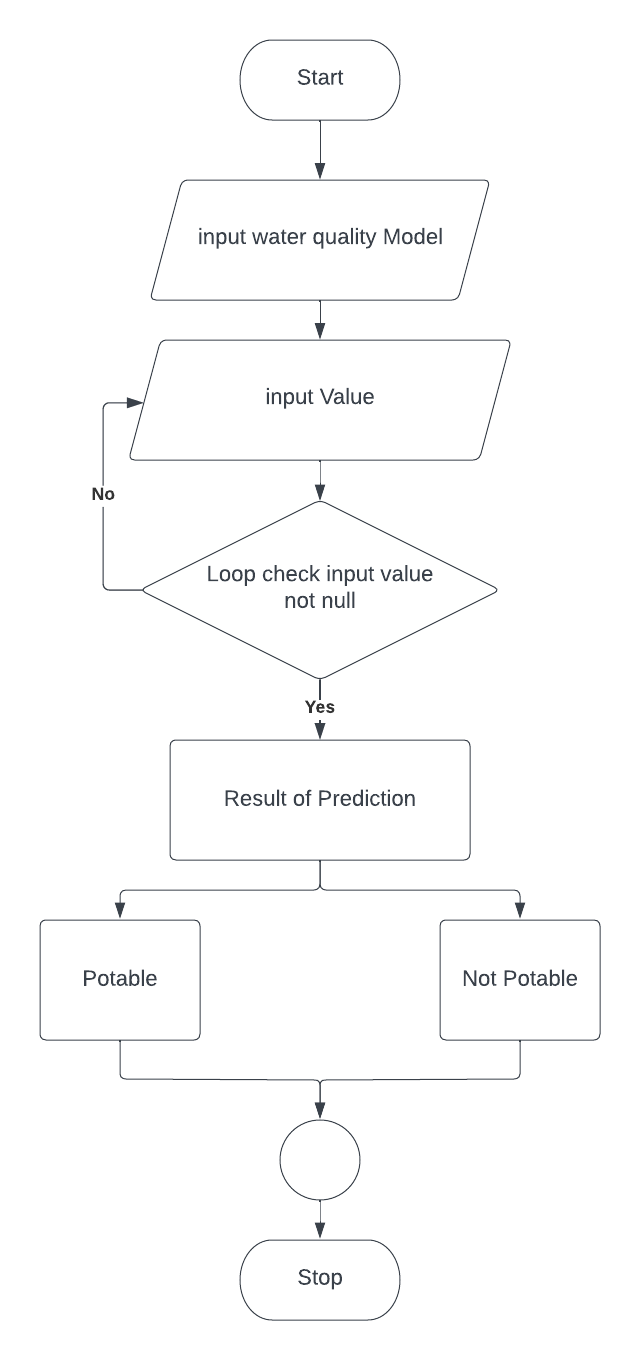
Diagram

Description automatically generated

**ภาพที่ 3.1** ขั้นตอนการการสร้างโมเดล

* 1. การประเมินผลแบบจำลอง

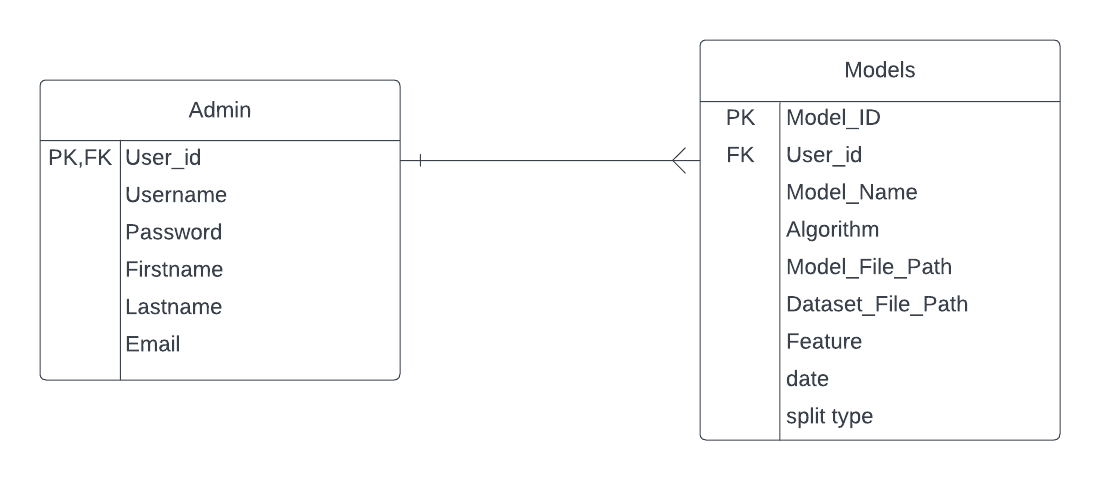
การประเมินผลแบบจำลอง เป็นกระบวนการที่สำคัญในการวิเคราะห์และออกแบบระบบการตรวจจับคุณภาพของน้ำดื่มด้วยการใช้การเรียนรู้ของเครื่อง การประเมินผลเพื่อตรวจสอบว่าแบบจำลองสามารถทำงานได้ดีและมีประสิทธิภาพตามความต้องการ ซึ่งการประเมินผลแบบจำลองจำเป็นต้องทำการแบ่งชุดข้อมูลสำหรับทำการทดสอบ ผู้พัฒนาจึงเลือกวิธีการที่มีชื่อว่า K-Fold Cross Validation ที่เป็นการวัดประสิทธิภาพของแบบจำลองด้วยการสุ่มตัวอย่าง (Resampling) โดยเริ่มต้นจากการแบ่งข้อมูลออกเป็น K ชุดที่มีขนาดเท่าๆ กันเพื่อทำการสร้างและทดสอบแบบจำลองซึ่งต้องคำนวณหาค่าความผิดพลาดของข้อมูล K ชุดก่อนนำแบบจำลองไปทำนายข้อมูลทดสอบ โดยจะต้องแบ่งข้อมูลด้วยการสุ่ม (Random) เพื่อช่วยให้ข้อมูลแต่ละส่วนมีการกระจายตัวที่ใกล้เคียงกัน จำนวนชุดข้อมูล K ที่นิยมใช้กันคือค่า 5 และ 10 เมื่อทำการแบ่งข้อมูล K ชุดออกแล้วจะทำการฝึกฝนแบบจำลอง (Train Folds) และทดสอบแบบจำลอง (Validation Folds) ตามจำนวนรอบของชุดข้อมูล K ที่ทำการแบ่ง และต้องทำการบันทึกค่า Validation Error หลังจบกระบวนการแต่ละรอบ และทำการหาค่าเฉลี่ย Validation Error (Satangmongkol. 2019) โดยในแต่ละ Fold จะทำการคำนวณหาค่า Precision, Recall, F1 Score และ AUC และเมื่อครบจำนวน Fold ตามค่า K ที่กำหนดแล้ว จะนำค่า Precision, Recall, F1 Score และ AUC มาหาค่าเฉลี่ย และจะได้ผลการประเมินเป็น Averaged Precision, Recall, F1 Score และ AUC ซึ่งบ่งบอกถึงประสิทธิภาพของระบบในการจำแนกข้อมูล

* 1. ขอบเขตการดำเนินงานของระบบการทำนาย

**ภาพที่ 3.2** แผนภาพขั้นตอนการนำโมเดลมาใช้ในระบบการทำนาย

* + 1. นำเข้า Water Quality Model ที่ทำการสร้างขึ้นโดย Admin
    2. นำเข้าค่าของตัวแปรทั้ง 9 ค่า ได้แก่ pH, Hardness, Solid, Chloramine, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity
    3. ตรวจสอบว่านำเข้าตัวแปรครบทุกตัวแล้วหรือไม่
    4. ทำการทำนายโดยใช้โมเดลที่ได้สร้างขึ้นโดยผลจะออกมาเป็น Potable และ Not Potable

1. **โครงสร้างฐานข้อมูล (Database Schema)**



**ภาพที่ 3.3** แผนภาพ Database Schema

* + 1. พจนานุกรมข้อมูล (Data Dictionary)

พจนานุกรมข้อมูล (Data Dictionary) ของเว็บแอพพลิเคชั่นทำนายคุณภาพน้ำดื่มเพื่อดึงข้อมูลจากฐานข้อมูลมาใช้งานในตัวเว็ปแอพพลิเคชั่น

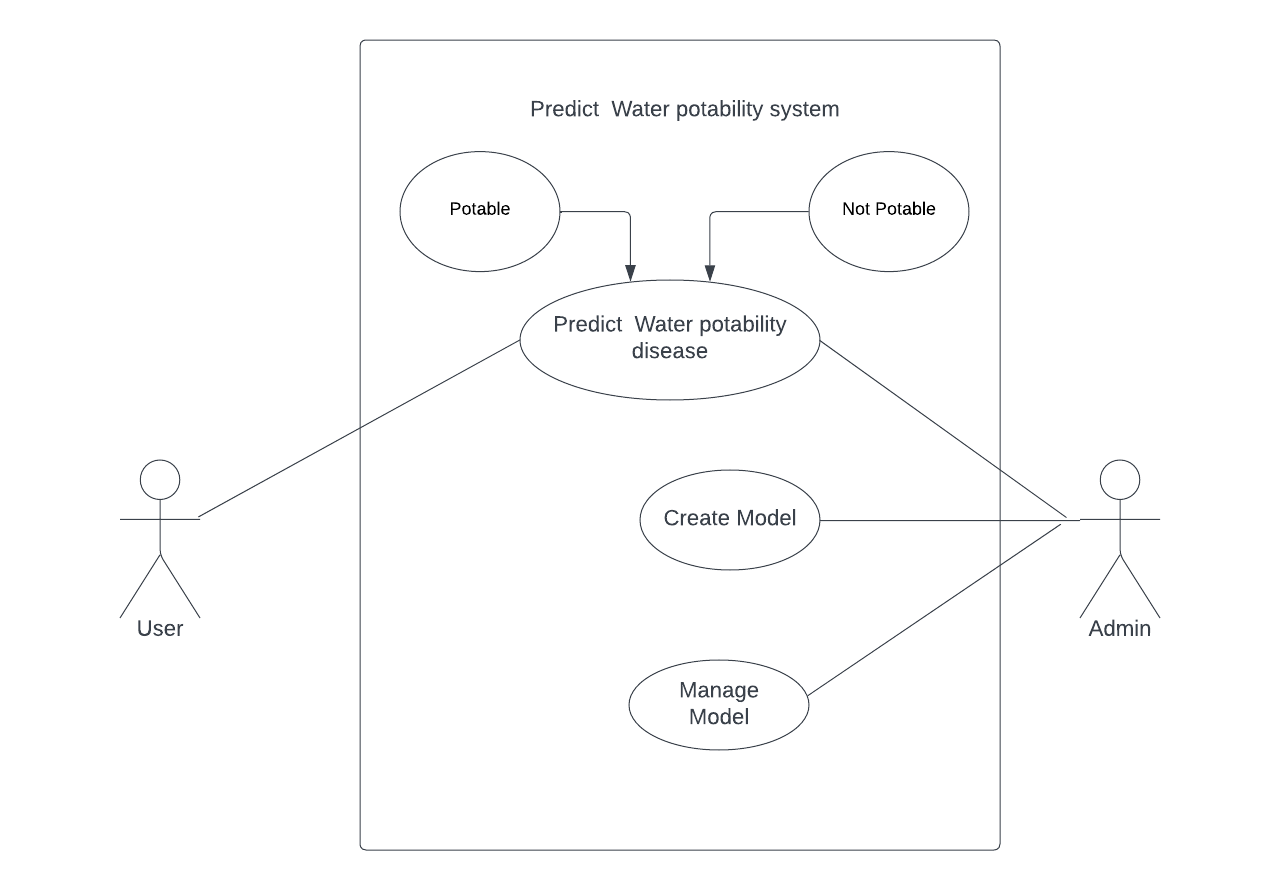
**ตารางที่ 3.1** ตารางเก็บข้อมูล Admin

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** | **TYPE** | **KEY** | **LENGTH** |
| User\_id | ชื่อผู้ใช้ของผู้ดูแลระบบ | VARCHAR | PK, FK | 20 |
| Username | ชื่อผู้ใช้ของผู้ดูแลระบบ | VARCHAR |  | 20 |
| Password | รหัสผ่านของผู้ดูแลระบบ | VARCHAR |  | 20 |
| Firstname | ชื่อของผู้ดูแลระบบ | VARCHAR |  | 30 |
| Lastname | นามสกุลของผู้ดูแลระบบ | VARCHAR |  | 30 |
| Email | ที่อยู่อีเมลของผู้ดูแลระบบ | VARCHAR |  | 30 |
| Phone | เบอร์โทรศัพท์ของผู้ดูแลระบบ | VARCHAR |  | 10 |

**ตารางที่ 3.2** ตารางเก็บข้อมูล Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ATTRIBUTE NAME** | **DESCRIPTION** | **TYPE** | **KEY** | **LENGTH** |
| Model\_ID | หมายเลขของแบบจำลอง | VARCHAR | PK | 10 |
| User\_id | ไอดีผู้ใช้ของผู้สร้างแบบจำลอง | VARCHAR | FK | 20 |
| Model\_Name | ชื่อของแบบจำลอง | VARCHAR |  | 20 |
| Algorithm | ขั้นตอนวิธีที่ใช้ในการสร้างแบบจำลอง | VARCHAR |  | 20 |
| Model\_File\_Path | ที่อยู่ของไฟล์แบบจำลอง | VARCHAR |  | 50 |
| Dataset\_File\_Path | ที่อยู่ของชุดข้อมูลที่ใช้ในการฝึกฝนและทดสอบแบบจำลอง | VARCHAR |  | 50 |
| Feature | เก็บfeatureที่เลือก | VARCHAR |  | 20 |
| date | เก็บวันที่สร้าง | VARCHAR |  | 15 |
| split type | เก็บประเภทการแบ่งข้อมูลและจำนวนการแบ่งข้อมูล | VARCHAR |  | 20 |

1. **Use Case**

****

**ภาพที่ 3.4** Use Case ระบบทำนายคุณภาพน้ำ

* + 1. Use Case Description

**ตารางที่ 3.3** Use Case Description Predict Water

|  |  |
| --- | --- |
| Use case Name | Predict Water potability disease |
| Actor | User |
| Description | Data set จากผู้ใช้ที่ต้องการเพื่อนำมาใช้ทดสอบ |
| Normal Course | 1.ทำการนำเข้า Data set จากผู้ใช้ 2.ทำการพยากรณ์ผลลัพธ์ของการทำนาย |
| Alternate Course | 1.หากมีการใส่ data set ที่ไม่ถูกต้องจะขึ้นข้อความแจ้งเตือน 2.หากมีการใส่ data set ไม่ครบจะขึ้นข้อความแจ้งเตือน 3.หากพยากรณ์ไม่สำเร็จจะขึ้นข้อความแจ้งเตือน |

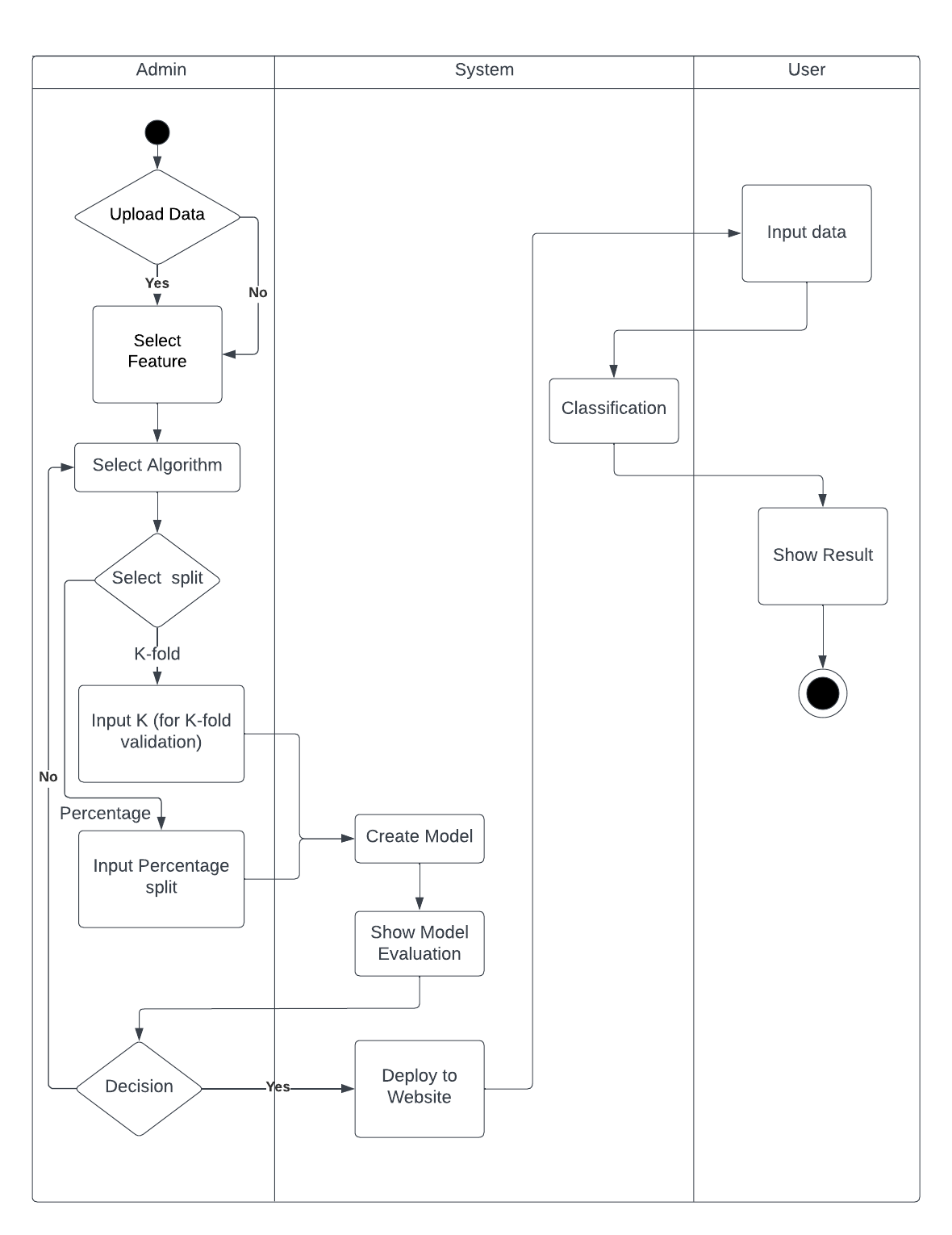
|  |  |
| --- | --- |
| Use case Name | Create Model |
| Actor | Admin |
| Description | สร้างโมเดลใหม่แทนที่โมเดลเดิม |
| Normal Course | 1.เมื่อ Create Model สำเร็จโมเดลใหม่จะแทนที่โมเดลเดิม |
| Alternate Course | 1.ถ้า Algorithm ผิดพลาดจะขึ้นข้อความแจ้งเตือน  2.ถ้า Create Model ไม่สำเร็จจะขึ้นข้อความแจ้งเตือน |

**ตารางที่ 3.4** Use Case Predict Water potability disease

**ตารางที่ 3.5** Use Case Create Model

|  |  |
| --- | --- |
| Use case Name | Login |
| Actor | Admin |
| Description | เมื่อกดปุ่ม Login |
| Normal Course | 1.เมื่อ login สำเร็จจะสามารถสร้างหรือเพิ่ม โมเดลได้ 2.เมื่อ login สำเร็จจะสามารถแก้ไขโมเดลได้ |
| Alternate Course | 1.ถ้าใส่ username หรือ password ผิดจะขึ้นข้อความแจ้งเตือน 2.ถ้าใส่ข้อมูลไม่ครบจะขึ้นข้อความแจ้งเตือน |

1. **Activity Diagram**

****

**ภาพที่ 3.5** แผนภาพ Activity Diagram แสดงการ Create Models ของAdmin และการ Prediction ของ User

1. **การออกแบบส่วนติดต่อกับผู้ใช้ (User Interface Design)**
   * 1. หน้า Login

Graphical user interface

Description automatically generated

**ภาพที่ 3.6** หน้า Login

หน้า Login สำหรับการเข้าสู่ระบบของ Admin โดยจะให้ผู้ใช้กรอก Username และ Password ของตนเองเพื่อเข้าสู่ระบบ ซึ่งจะทำการตรวจสอบข้อมูลว่ามี Username และ Password อยู่ในฐานข้อมูลหรือไม่ หากไม่พบข้อมูลจะทำการแจ้งเตือน และเมื่อทำการ Login สำเร็จจะทำการเข้าสู่หน้า Create Models (ภาพที่ 3.7)

* + 1. หน้า Create Models

A screenshot of a computer

Description automatically generated with medium confidence

**ภาพที่ 3.7** หน้า Create Models

หน้า Create Models เป็นหน้าต่างสำหรับการสร้างโมเดล โดยระบบจะให้ผู้ใช้ตั้งชื่อโมเดล อัพโหลดdataใหม่ได้โดยไฟล์เป็นcsv เลือก Feature ที่ต้องการใช้ในการฝึกฝนแบบจำลอง หลังจากเลือกแล้วให้ผู้ใช้เลือกวิธีการฝึกของแบบจำลอง และ ทำการเลือกการแบ่งข้อมูลพร้อมกับเลือกจำนวนข้อมูลที่ต้องการจะแบ่ง แล้วจึงกดปุ่ม Create Models หากพบข้อผิดพลาดในการสร้างจะมีข้อความแจ้งเตือนและเมื่อสร้างสำเร็จจะแสดงผลลัพธ์ (ภาพที่ 3.8)



**ภาพที่ 3.8** หน้า Create Models เมื่อกด Create

จากภาพที่ 3.8 ในหน้านี้จะแสดงผลลัพธ์ของโมเดลที่ดีที่สุด ผลของโมเดลใหม่และผลของโมเดลที่มีอยู่ในระบบทั้งหมดเพื่อให้ผู้ใช้ทำการเปรียบเที่ยบ และกดปุ่มsaveเพื่อทำการบันทึกโมเดลลงระบบ

* + 1. หน้า Predictionc

A screenshot of a computer screen

Description automatically generated with low confidence

**ภาพที่ 3.9** หน้า Prediction Values

หน้า Prediction เป็นหน้าต่างที่ใช้ได้ทั้ง User และ Admin เป็นการทำนายคุณภาพของน้ำว่าสามารถดื่มได้หรือไม่ โดยจะให้ผู้ใช้ทำการกรอกข้อมูลดังต่อไปนี้ pH, Hardness, Solid, Chloramine, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity เมื่อกรอกข้อมูลแล้วกดปุ่มPrediction จะทำการตรวจสอบว่ามีการกรอกข้อมูลครบทุกช่องหรือไม่หากไม่ครบจะมีข้อความแจ้งเตือนและเมื่อตรวจสอบว่ากรอกข้อมูลครบแล้วจะทำการแสดงผลการประเมินออกมาทางช่อง result โดยจะแสดงผลออกเป็นสองค่าคือ Potable และ Not Potable และแสดงค่าตัวแปรที่มีผลทำให้น้ำไม่สามารถดื่มได้ นอกจากนี้ผู้ใช่สามารถทำการPredictionแบบหลายชุดข้อมุลได้โดยกดปุ่มData Prediction

เพื่อนไปยังหน้า Prediction data (ภาพที่ 3.10)

A screenshot of a computer screen

Description automatically generated with low confidence

**ภาพที่ 3.10** หน้า Prediction data

จากภาพที่ 3.10 ในหน้านี้จะมีปุ่มDownload Forms สำหรับโหลดFormsสำหรับการนำมาใช้ทำการprediction และมีปุ่ม Upload Forms สำหรับอับFormsที่ทำการใส่ข้อมูลเรียบร้อยแล้ว เมื่ออัพแล้วจะมีการแสดงข้อมูลบนหน้าจอแล้วทำการกดPrediction เพื่อทำการทำนายและจะนำไปสู่หน้า Result(ภาพที่ 3.11)

A screenshot of a computer screen

Description automatically generated with medium confidence

**ภาพที่ 3.11** หน้า Result

จากภาพที่ 3.11 จะแสดงผลการทำนายโดยจะมีชื่อสถานที่ที่ตรวจวัดและผลการทำนาย สามารถกดปุ่มDownloadเพื่อโหลดไฟล์csv

**บรรณานุกรม**

การประปาส่วนภูมิภาค. (2562). หลักเกณฑ์การให้บริการทดสอบตัวอย่างน้ำ สารเคมี สารกรองและท่อน้ำ. สืบค้น

4 พฤษภาคม 2566, จาก https://www.pwa.co.th/download/testing-rate.pdf

กรมอนามัยส่งเสริมให้คนไทยสุขภาพดี. (2556). การจัดการน้ำบริโภคในโรงเรียน. สืบค้น 4 พฤษภาคม 2566,

จาก <https://hp.anamai.moph.go.th/th/manual-of-official/download?id=72908&mid=35018&mkey=m_document&lang=th&did=23704>

กองห้องปฏิบัติการสาธารณสุขกรมอนามัย. (2565). ค่าบริการวิเคราะห์คุณภาพน้ำบริโภค มอก.257-2521. สืบค้น

4 พฤษภาคม 2566, จาก [https://rldc.anamai.moph.go.th/th/water-quality-analysis-service-rates/66285#](https://rldc.anamai.moph.go.th/th/water-quality-analysis-service-rates/66285)

สำนักสุขาภิบาลอาหารและน้ำ กรมอนามัย. (2553). คู่มือพัฒนาคุณภาพน้ำบริโภคในโรงพยาบาล. สืบค้น

4 พฤษภาคม 2566, จาก <https://apps.hpc.go.th/dl/web/upFile/2019/12-5027-> 20191227122624/4bb5cf02960604709e10b3f51c0c3860.pdf

Health Canada. (2018). Chloramines in Drinking Water. Retrieved May 22, 2023, from

<https://www.canada.ca/en/health-canada/programs/consultationchloramines-drinking-water/document.html>

Water Science School. (2018). Turbidity and Water. Retrieved May 20, 2023, from

<https://www.usgs.gov/special-topics/water-science-school/science/turbidity-and-water>

World Health Organization. (2004). Sulfate in Drinking-water. Retrieved May 21, 2023, from

<https://cdn.who.int/media/docs/default-source/wash-documents/wash-> chemicals/sulfate.pdf?sfvrsn=b944d584\_4

World Health Organization. (2004). Trihalomethanes in Drinking-water. Retrieved May 21, 2023,

from <https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/trihalomethanes.pdf?sfvrsn=3d3a90e3_4>

World Health Organization. (2010). Hardness in Drinking-water. Retrieved May 21, 2023, from

<https://apps.who.int/iris/bitstream/handle/10665/70168/WHO_HSE_WSH_10.01_10_Rev1_eng.pdf>

World Health Organization. (2022). Drinking-water. Retrieved May 21, 2023, from

<https://www.who.int/news-room/fact-sheets/detail/drinking-water>

World Health Organization. (2022). Guidelines for drinking-water quality. Retrieved May 21, 2023,

from <https://www.who.int/publications/i/item/9789240045064>