**กระบวนการวิเคราะห์ข้อมูล (CRISP-DM)**

1. **Business Understanding**

การพยากรณ์อากาศในประเทศไทย ส่วนมากจะเป็นการพยากรณ์อากาศโดยภาพรวม โดยในประเทศไทยนั้น สภาพอากาศที่ส่งผลกระทบต่อการดำเนินชีวิตมากที่สุด คือ การเกิดฝนตก ซึ่งหากเราสามารถรู้ล่วงหน้า ว่าฝนจะตกเมื่อไหร่ จะช่วยให้เราสามารถวางแผนการใช้ชีวิตประจำวันได้มีประสิทธิภาพมากยิ่งขึ้น ไม่ว่าจะเป็นการวางแผนกิจกรรมที่จะทำ โดยเฉพาะกิจกรรมกลางแจ้ง, อุปกรณ์ที่ต้องเตรียมไป เช่น ร่ม หรือลักษณะเสื้อผ้าที่เหมาะสมกับสภาพอากาศ เป็นต้น

การสร้าง ดูฝน: ระบบพยากรณ์อากาศเฉพาะที่ ทำการพยากรณ์อากาศปัจจุบัน (Now cast) ซึ่งเป็นการพยากรณ์อากาศเชิงตัวเลข (numerical weather prediction-NWP) เป็นการคาดหมายสภาวะอากาศเวลาไม่เกิน 2 ชั่วโมง จะช่วยรายงานข้อมูลสภาวะอากาศเฉพาะที่ของเวลาปัจจุบันได้แล้ว ยังสามารถพยากรณ์การเกิดฝนตกล่วงหน้าได้ด้วย ซึ่งจะช่วยอำนวยความสะดวกและป้องกันความเสียหายอันเนื่องมาจากการเกิดฝนตก

1. **Data Understanding**
   * ผู้พัฒนาได้เก็บรวมรวบข้อมูลสภาวะอากาศ ผ่านอุปกรณ์ IoT ตั้งแต่วันที่ 26-09-2016 ถึงวันที่ 26-11-2016 มาจำนวน 4516 รายการ
   * ข้อมูลสภาวะอากาศนี้แบ่งออกเป็น 2 กลุ่ม
     + กลุ่มที่ฝนตก จำนวน 277 รายการ
     + กลุ่มที่ฝนไม่ตก จำนวน 4239 รายการ
   * มีแอตทริบิวต์ทั้งหมด 6 แอตทริบิวต์
   * แอททริบิวต์เป้าหมายในการพยากรณ์ คือ rain

|  |  |  |
| --- | --- | --- |
| **แอตทริบิวต์** | **คำอธิบาย** | **ประเภท** |
| temp | อุณหภูมิ(องศาเซลเซียส) | NUMERIC |
| humidity | ความชื้น(%) | NUMERIC |
| dewpoint | อุณหภูมิจุดน้ำค้าง (องศาเซลเซียส) | NUMERIC |
| pressure | ความกดอากาศ (เฮกโตปาสคาล) | NUMERIC |
| light | ความสว่าง | NUMERIC |
| rain | ฝน (ฝนตก,ฝนไม่ตก) | {0,1} |

1. **Data Preparation**

ขึ้นตอนนี้จะทำการแปลงข้อมูลที่ได้เก็บรวบรวมมา ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้ โดยทำการเลื่อนค่าฝนไป 2 ชั่วโมง และทำการแปลงข้อมูลให้อยู่ในรูปแบบนามสกุล .arff เพื่อให้สามารถนำไปใช้ในโปรแกรม Weka ได้

* + ข้อมูลสภาวะอากาศที่สามารถนำไปวิเคราะห์ได้จำนวน 4487 รายการ แบ่งออกเป็น 2 กลุ่ม
    - กลุ่มที่ฝนตก จำนวน 258 รายการ
    - กลุ่มที่ฝนไม่ตก จำนวน 4229 รายการ

1. **Modeling**

ขั้นตอนนี้จะทำการสร้างโมเดลด้วยวิธีการ J48(C4.5), NaiveBayes, Random Forestเพื่อช่วยในการพยากรณ์หาว่าข้อมูลสภาวะอากาศแบบใดมีโอกาสที่ฝนจะตก โดยแบ่งข้อมูลออกเป็น 2 ชุด เป็นชุดข้อมูลสำหรับสร้างโมเดลจำนวน 60% และชุดข้อมูลสำหรับทดสอบโมเดลที่สร้างขึ้นจำนวน 40% เป็นการแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิ์ภาพของโมเดลการจำแนกประเภทข้อมูล

โดยคราสของข้อมูลจะแบ่งเป็น 2 คลาส คือ คราส ฝนตก และ คราส ฝนไม่ตก จะแสดงผลลัพธ์ความน่าจะเป็นที่ฝนตกและฝนไม่ตกตั้งแต่ 0 ถึง 100 เปอร์เซ็นต์

1. **Evaluation**

ขั้นตอนนี้จะทำการประเมินผลประสิทธิภาพของโมเดล J48(C4.5), NaiveBayes, Random Forest หลังจากที่แย่งข้อมูลออกเป็น 2 ส่วนและใช้โมเดลที่สร้างจากข้อมูลสภาวะอากาศจำนวน 60% มาทำการทดสอบพยากรณ์โอกาสที่ฝนจะตกให้กับ ข้อมูลสภาวะอากาศจำนวน 40% ที่เหลือ และใช้ค่า F-measure ของคราสฝนตก ในการพิจารณาหาโมเดลที่ดีที่สุด เนื่องจากเป็นการพยากรณ์โอกาสที่ฝนจะตกจึงให้ความสำคัญของคราสฝนตกมากกว่า ดังนั้น ค่า F-measure ของคราสฝนตก ที่มากที่สุดใน 3 โมเดล เท่ากับ 0.804 คือโมเดล Random Forest

โมเดล Random Forest แบ่งประเภทโอกาสฝนตกกับฝนไม่ตก Min Threshold ดังต่อไปนี้

* + - ความน่าจะเป็นที่ฝนตก อยู่ที่ มากกว่าหรือเท่ากับ 0.44
    - ความน่าจะเป็นที่ฝนไม่ตก อยู่ที่ มากกว่าหรือเท่ากับ 0.57

**Model Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | J48(C4.5) | NaiveBayes -D | Random  Forest (Tree 100) |
| **Accuracy** | 97.0474 | 95.376 | 97.9387 |
| **Kappa statistic** | 0.7207 | 0.55 | 0.7934 |
| **TPR** (Class1) | 0.718 | 0.544 | 0.738 |
| **FPR** (Class1) | 0.014 | 0.021 | 0.006 |
| **Precision** (Class1) | 0.755 | 0.609 | 0.884 |
| **Recall** (Class1) | 0.718 | 0.544 | 0.738 |
| **F-Measure** (Class1) | 0.736 | 0.574 | 0.804 |
| **ROC Area** (Class1) | 0.864 | 0.85 | 0.968 |
| **Min Threshold** (Class1) | 0.871 | 0.7951 | 0.44 |
| **TPR** (Class0) | 0.986 | 0.979 | 0.994 |
| **FPR** (Class0) | 0.282 | 0.456 | 0.262 |
| **Precision** (Class0) | 0.983 | 0.972 | 0.984 |
| **Recall** (Class0) | 0.986 | 0.979 | 0.994 |
| **F-Measure** (Class0) | 0.984 | 0.976 | 0.989 |
| **ROC Area** (Class0) | 0.864 | 0.85 | 0.968 |
| **Min Threshold** (Class0) | 0.1667 | 0.3022 | 0.57 |

1. **Deployment**

จากโมเดล Random Forest ที่สร้างขึ้นมานั้นสามารถนำไปใช้ในการพยากรณ์โอกาสที่ฝนจะตกได้ โดยนำโมเดล Random Forest ไปใช้งานกับข้อมูลสภาวะอากาศ ณ ปัจจุบัน ในการใช้งานจะให้ผลผลลัพธ์คำตอบเป็น ฝนตก หรือ ฝนไม่ตก โดยแสดงความน่าจะเป็นที่ฝนตกและฝนไม่ตก ตั้งแต่ 0 ถึง 100 เปอร์เซ็นต์

ค่าแนะนำเปอร์เซ็นต์ที่ฝนตก Min Threshold คือ >= 0.44 หรือ 44%