**บทที่ 1**

**บทนำ**

**1.1 ความเป็นมาและความสำคัญของปัญหา**

มหาวิทยาลัยราชภัฏจันทรเกษมเป็นหน่วยงานราชการประเภทสถานศึกษา ในทุกๆปีจะมีการรับสมัครผู้ขอทุนการศึกษา แบ่งออกเป็น 7 ประเภท ได้แก่ 1.ทุนรางวัลผลการเรียนดี 2.ทุนช่วยเหลือผู้ขาดแคลน โดยเรามุ่งเน้นไปที่ทุนช่วยเหลือผู้ขาดแคลนประเภทกองทุนเงินให้กู้ยืมเพื่อการศึกษา (กยศ.) ทุนที่มีวัตถุประสงค์ช่วยการสนับสนุนและส่งเสริมการศึกษา โดยการให้กู้ยืมแก่นักเรียนนักศึกษา ที่ขาดแคลนทุนทรัพย์ และนักเรียนนักศึกษาต้องชำระหนี้คืนพร้อมดอกเบี้ยตามอัตราที่กำหนดเมื่อจบการศึกษาแล้ว (อ้างอิง 2563 : กองทุนเงินให้กู้ยืมเพื่อการศึกษา กยศ.) การรับสมัครมีหลายรูปแบบ อาทิเช่น การยื่นผลคะแนนสอบวัดระดับการศึกษา และการสอบสัมภาษณ์ 3.ทุนส่งเสริมการศึกษาเฉพาะทาง 4.ทุนสนับสนุนกิจกรรมเสริมหลักสูตร 5.ทุนสำหรับผู้ด้อยโอกาสทางสังคมและวัฒนธรรม 6.ทุนเงินยืมเพื่อการลงทุนพัฒนาบุคคล 7.ทุนการศึกษาเพื่อส่งเสริมธุรกิจ

การสัมภาษณ์ขอทุนเงินให้กู้ยืมเพื่อการศึกษาในปัจจุบัน จากสถิติข้อมูล กยศ. เดือนมิถุนายน 2563 มีผู้เข้าสอบสัมภาษณ์ปริมาณ 5,771,655 ราย เข้าสัมภาษณ์ 10 รายต่อวัน (อ้างอิง 2563 : กองทุนเงินให้กู้ยืมเพื่อการศึกษา กยศ www.studentloan.or.th ) ผู้ถูกสัมภาษณ์ประสบปัญหาการรอสอบสัมภาษณ์ทุนเป็นเวลานาน เสียค่าเดินทาง และเสี่ยงต่อการแพร่เชื้อระบาดโควิดเนื่องจากเป็นที่รวมตัวกันของคนหมู่มาก ในส่วนของปัญหาผู้รับสัมภาษณ์ คือ ผู้ถูกสัมภาษณ์มีจำนวนมากก่อให้เกิดการนั่งเป็นเวลานาน ส่งผลกระทบต่อกล้ามเนื้อของผู้รับสัมภาษณ์ อาการปวด อาทิเช่น ศีรษะ คอ ไหล่ หลัง ขา อาจส่งผลไปถึงกระดูกสันหลังทับเส้นประสาทได้ (อ้างอิง 2561 : Coach Bank Chira: เคล็ดลับง่ายๆรักษาโรคด้วยอาหาร และการเป็นเศรษฐีความสุข ) ปัญหาของการรับฟังผู้ถูกสัมภาษณ์นำไปสู่ผลกระทบต่อสภาวะทางจิตใจของผู้รับสัมภาษณ์ จากทฤษฎีโดยธรรมชาติ ภาวะทางจิตใจมีการปรับเปลี่ยนไปตามสิ่งที่ได้รับรู้จากการฟังปัญหาเชิงลบของบุคคลหนึ่ง ( อ้างอิง 2563 : จากเว็บไซต์ RamaMental.mahidol.ac.th ภาควิชาจิตเวชศาสตร์คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล ) จากสถานการณ์ปัญหาข้างต้น อารมณ์ความรู้สึกของผู้รับสัมภาษณ์ ส่งผลถึงการประเมินคะแนนผู้ถูกสัมภาษณ์ทำให้เกิดความคลาดเคลื่อนได้ และยังเสี่ยงต่อการแพร่ระบาดโควิด

ผู้วิจัยเล็งเห็นถึงความสำคัญดังกล่าวจึงได้นำเอาวิทยาการทางคอมพิวเตอร์สมัยใหม่ที่เรียกว่าแมชชีนเลิร์นนิ่ง (Machine Learning) ซึ่งสามารถตรวจจับคำตอบจากผู้สัมภาษณ์ทางการออกเสียง และตรวจจับตัวเลข ประเภทของคำตอบ พร้อมประเมินคะแนนโดยเกณฑ์สัมภาษณ์กองทุนเงินให้กู้ยืมเพื่อการศึกษา (กยศ.) เข้ามาช่วยแก้ปัญหาโดยการนำมาสร้างเว็บแอปพลิเคชันหุ่นยนต์ประเมินคะแนนสำหรับการสอบสัมภาษณ์ เพื่อจัดการคะแนนของผู้ถูกสัมภาษณ์อย่างถูกต้องตามมาตรฐาน อีกทั้งรองรับปัญหาข้างต้น และท้ายที่สุดนำผลที่ได้รับการประเมินเข้ายื่นคะแนนต่อกับทางมหาวิทยาลัยราชภัฏจันทรเกษม

**4.วัตถุประสงค์**

เพื่อพัฒนาหุ่นยนต์ประเมินคะแนนให้มีความสามารถทางด้านด้านการคัดแยกคำตอบ ประเภท ตัวเลข จากการเรียนรู้ด้วยเสียง ใช้ประโยชน์ในการเข้าเกณฑ์มาตรฐานของ กยศ. ทั้งในด้านความแม่นยำของการประเมินคะแนนและเวลาที่ใช้สัมภาษณ์

**5. หลักการ ทฤษฎี เหตุผล**

5.1 แปลงเสียงพูดให้เป็นข้อความ (Speech to Text) เป็นการแปลงเสียงพูดให้เป็นข้อความ โดยในบทความนี้เลือกใช้โมดูล speech\_recognition ที่เรียกใช้ API มาจากคลาวด์ของ Google โดยการใช้งานต้องติดตั้งไลบรารี react-speech-recognition สามารถเรียนรู้จากเสียงคำพูดภาษามนุษย์ รองรับทั้งภาษาไทยและภาษาอังกฤษ แล้วแปลงเป็นข้อความ

5.2 Text to Speech เป็นกระบวนการแปลงจากข้อความเป็นเสียงพูด โดยในตัวอย่างบทความนี้เลือกใช้บริการของ "วาจา" ผู้ผลิตโดยองกรณ์ Nectac เพื่ออ่านข้อความออกมาเป็นเสียง ซึ่งผลลัพธ์ที่ได้จาก วาจา เป็นไฟล์เสียง MP3 โดยหลักการเลียนแบบเสียงมนุษย์ ใช้เสียงสังเคราะห์

5.3 การรูจําชื่อเฉพาะภาษาไทย (Conditional Random Fields) วิธีสำหรับการทำ Labeling และการทำ Segmentation สำหรับข้อมูลแบบลำดับ (Sequence Data) คือการใช้ Hidden Markov Models (HMMs) หรือ เครื่องที่ทำงานตามสถานะ จำกัดตามความจะเป็น (Probabilistic Finite-state Automata) เพื่อที่จะระบุลำดับที่เป็นไปได้จากประเภทข้อมูลนำเข้า ซึ่งอาจแทนด้วยรูปแบบ, สัญลักษณ์หรือ คุณสมบัติ(Features) ที่กําหนดโดย ที่รูปแบบที่ถูกสร้างจาก Hidden Markov Models (HMMs) นั้นสามารถถูกกำหนดด้วยการกระจาย ความน่าจะเป็นร่วมกันของตัวแปรเชิงสุ่มX และY โดยที่ X คือข้อมูลนำเข้า (ObservationSequence) ส่วน Y คือลักษณะเฉพาะหรือผลลัพธ์ (label Sequence) ซึ่งตรงกับคุณสมบัติของข้อมูลนำเข้า X แบบจำลองฮิดเดนมาร์คอฟนี้เป็น Generative Model และใช้คำความน่าจะเป็นร่วม(Joint Probability) ระหว่างข้อมูลนําเข้า(X) และ ลักษณะเฉพาะหรือผลลัพธ์(Y) ซึ่งอาจพบปัญหาว่าไม่ สามารถจับคู่ความสัมพันธ์ระหว่างคุณสมบัติของลักษณะเฉพาะต่างๆกับข้อมูลนำเข้า

5.4 การจำแนกประเภทหลายฉลาก (TAG SUGGESTION) การจำแนกรูปแบบนี้ คลาสทั้งหมดที่เป็นคำตอบได้มีมากกว่าสองคลาส และข้อมูลอาจถูกจำแนกให้อยู่ได้มากกว่าหนึ่งคลาส อาทิเช่น การจำแนกคำพูดที่ได้จากการสัมภาษณ์ คำตอบผู้ถูกสัมภาษณ์จัดได้ในหลายประเภท หรือการจำแนกประเภทรูปภาพรูปหนึ่งอาจจะเป็นไปได้ทั้งรูปขาวดำ รูปทิวทัศน์ รูปวาด เป็นต้น

**6. ระยะเวลาดำเนินการ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ลำดับ | กิจกรรม | ระยะเวลา | | | |
| มกราคม 2565 | กุมภาพัน 2565 | มีนาคม 2565 | เมษายน 2565 |
| 1 | รวบรวมข้อมูล |  |  |  |  |
| 2 | เตรียมข้อมูล |  |  |  |  |
| 3 | สร้างโมเดล |  |  |  |  |
| 4 | สร้างเว็บแอปพลิเคชัน |  |  |  |  |
| 6 | ประเมินผล |  |  |  |  |
| 7 | นำไปใช้งาน |  |  |  |  |

**7. แผนการดำเนินงาน ขอบเขตการศึกษา**

**7.1 แผนการดำเนินงาน**

**7.1.1** **รวบรวมข้อมูล**

- ศึกษารูปแบบของการสมัครขอทุนการศึกษาของผู้สมัครที่ทำให้เกิดคะแนนของผู้ได้รับทุนการศึกษา

- ศึกษาข้อมูลและใช้เทคนิคการจัดหมวดหมู่ (Classification) ในการสร้างเว็บแอปพลิเคชันเพื่อวิเคราะห์และหาสิ่งสำคัญจากข้อมูล และนำมาสู่การให้คะแนนอย่างเที่ยงตรงแม่นยำ

- ตั้งคำถามที่อยู่ในเกณฑ์การประเมินที่ใช้ในการกู้ กยศ.

**7.1.2** **เตรียมข้อมูล**

- คำตอบให้เป็นตัวเลขทั้งหมด เพื่อให้เข้าถึงได้ง่ายและเรียกใช้ได้อย่างรวดเร็ว เป็นประโยชน์ต่อการทำงานขั้นถัดไป ในการบวกคะแนน

**7.1.3** **สร้างโมเดล**

- จัดทำโมเดลเกณฑ์ให้คะแนนการรับสมัครตามเกณฑ์การประเมิน กยศ.

**7.1.4 สร้างเว็บแอปพลิเคชัน**

- พัฒนาเทคโนโลยีเว็บแอปพลิเคชันคัดกรองผู้สอบเข้าสัมภาษณ์ขอทุนการศึกษาด้วยระบบการให้คะแนนจำแนกตัวเลขจากคำตอบ (Conditional Random Fields) มหาวิทยาลัยราชภัฏจันทรเกษม

**7.1.5 ประเมินผลเว็บแอปพลิเคชัน**

- การประเมินผลโดยผู้เชี่ยวชาญที่นำหุ่นยนต์ไปใช้จริง

**7.2 ขอบเขตการศึกษา**

ข้อมูลที่นำมาใช้ทดสอบการจำแนกในการวิจัยนี้ คือ ข้อมูลเชิงทั่วไป ประกอบด้วย ชุดข้อมูลเรียนรู้ และชุดข้อมูลที่ไม่ได้นำมาเรียนรู้ อาจจัดอยู่ได้ในหลายคลาส คลาสเหล่านี้มีความสัมพันธ์ตามโครงสร้างที่กำหนดมาให้ โดยอัลกอริทึมการจัดการคำตอบคือ (Conditional Random Fields)

**7.3 ฮาร์ดแวร์ (Hardware) ที่ใช้**

**7.3.1** **คอมพิวเตอร์ Spec ขั้นต่ำ**

OS : Windows, MacOS

Memory : 4 GB RAM

Storage : 20 GB available space

**7.4 ซอฟท์แวร์ (Software) ที่ใช้**

7.4.1 Editor : Visual Studio Code

7.4.2 Language programming Frontend : React, Typescript

7.4.3 Language programming Backend : NodeJs, Javascript

7.4.4 Hosting : Firebase

**8. ประโยชน์ที่คาดว่าจะได้รับ**

8.1 ได้อัลกอริทึมวิเคราะห์ข้อมูลผู้สอบสัมภาษณ์ทุนการศึกษาที่มีประสิทธิภาพมากที่สุด

8.2 ได้ระบบการวิเคราะห์ข้อมูลการสมัครเข้าร่วมรับทุนเพิ่มนำไปช่วยในการตัดสินใจของเจ้าหน้าที่ทางมหาวิทยาลัยราชภัฏจันทรเกษม

8.3 ได้ความสะดวกสบายในการคาการณ์ให้คะแนน ลดระยะดวลาในการสอบสัมภาษณ์ของนักศึกษาในแต่ละปีการศึกษา

8.4 สามารถนำการวิเคราะห์มาช่วยในการสอบสัมภาษณ์ทุนการศึกษาตามมารตรฐาน ในแต่ละปีการศึกษา

8.5 สามารถประหยัดจำนวนผู้รับสัมภาษณ์

8.6 ลดการแพร่ระบาดของโรคโควิด

**9. คำนิยามศัพท์เฉพาะ**

9.1 การเรียนรู้ของเครื่อง (Machine Learning) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ที่พัฒนามาจากการศึกษาการรู้จำรูปแบบ เกี่ยวข้องกับการศึกษาและการสร้างอัลกอริทึมที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ อัลกอริทึมนั้นจะทำงานโดยอาศัยโมเดลที่สร้างมาจากชุดข้อมูลตัวอย่างขาเข้าเพื่อการทำนายหรือตัดสินใจในภายหลัง แทนที่จะทำงานตามลำดับของคำสั่งโปรแกรมคอมพิวเตอร์

9.2 การจัดหมวดหมู่ (Classification) เป็น Model ประเภท Supervised Model หมายถึง Model ที่ต้องมี Target หรือ ตัวแปรที่ใช้วัดเป้าหมาย เป็นตัวตั้งต้นให้เรียนรู้ โดย Target ของ Classification จะเป็นแบ่งออกเป็นกลุ่ม หรือมีลักษณะเป็น Discrete เช่น yes/no, A/B/C เป็นต้น ดังนั้น ในการประเมินผลลัพธ์ที่ได้จาก Classification Model จะสามารถวัดค่าความแม่นยำ หรือ Accuracy ได้

9.3 การจำแนกประเภทหลายฉลาก (TAG SUGGESTION) การจำแนกรูปแบบนี้ คลาสทั้งหมดที่เป็นคำตอบได้มีมากกว่าสองคลาส และข้อมูลอาจถูกจำแนกให้อยู่ได้มากกว่าหนึ่งคลาส อาทิเช่น การจำแนกคำพูดที่ได้จากการสัมภาษณ์ คำตอบผู้ถูกสัมภาษณ์จัดได้ในหลายประเภท หรือการจำแนกประเภทรูปภาพรูปหนึ่งอาจจะเป็นไปได้ทั้งรูปขาวดำ รูปทิวทัศน์ รูปวาด เป็นต้น

9.4 การรูจําชื่อเฉพาะภาษาไทย (Conditional Random Fields) วิธีสำหรับการทำ Labeling และการทำ Segmentation สำหรับข้อมูลแบบลำดับ (Sequence Data) คือการใช้ Hidden Markov Models (HMMs) หรือ เครื่องที่ทำงานตามสถานะ จำกัดตามความจะเป็น (Probabilistic Finite-state Automata) เพื่อที่จะระบุลำดับที่เป็นไปได้จากประเภทข้อมูลนำเข้า ซึ่งอาจแทนด้วยรูปแบบ, สัญลักษณ์หรือ คุณสมบัติ(Features) ที่กําหนดโดย ที่รูปแบบที่ถูกสร้างจาก Hidden Markov Models (HMMs) นั้นสามารถถูกกำหนดด้วยการกระจาย ความน่าจะเป็นร่วมกันของตัวแปรเชิงสุ่มX และY โดยที่ X คือข้อมูลนำเข้า (ObservationSequence) ส่วน Y คือลักษณะเฉพาะหรือผลลัพธ์ (label Sequence) ซึ่งตรงกับคุณสมบัติของข้อมูลนำเข้า X แบบจำลองฮิดเดนมาร์คอฟนี้เป็น Generative Model และใช้คำความน่าจะเป็นร่วม(Joint Probability) ระหว่างข้อมูลนําเข้า(X) และ ลักษณะเฉพาะหรือผลลัพธ์(Y) ซึ่งอาจพบปัญหาว่าไม่ สามารถจับคู่ความสัมพันธ์ระหว่างคุณสมบัติของลักษณะเฉพาะต่างๆกับข้อมูลนำเข้า

**บทที่ 2**

**เอกสารและงานวิจัยที่เกี่ยวข้อง**

การประเมินคะแนนผู้สอบสัมภาษณ์ โดยการสร้างโมเดลในการจำแนกกลุ่มคะแนนของการตอบคำถามและวัดผลที่อาศัยหลักการ การจัดหมวดหมู่ (Classification) ในการวิเคราะห์ข้อมูลที่มีขนาดใหญ่ ด้วยอัลกอริทึมการจำแนกประเภทหลายฉลาก (Multi-Label Classification) ผู้วิจัยได้รวบรวมข้อมูลเอกสาร และงานวิจัยที่เกี่ยวข้อง เพื่อเป็นข้อมูลในการวิจัย ซึ่งมีรายละเอียดดังนี้

2.1 การเรียนรู้ของเครื่อง (Machine Learning)

2.2 การจัดหมวดหมู่ (Classification)

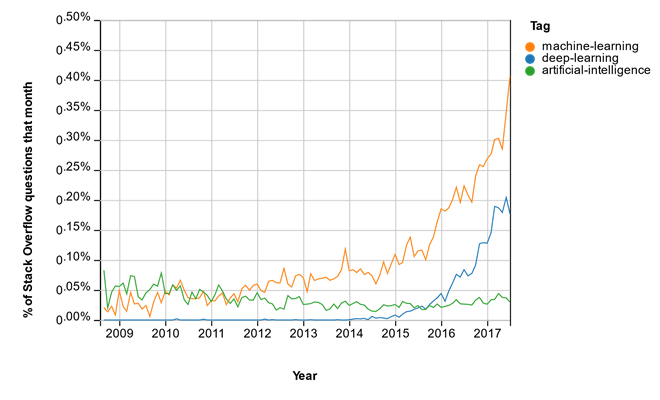
2.3 อัลกอริทึมการจำแนกประเภทหลายฉลาก (Multi-Label Classification)

2.4 งานวิจัยที่เกี่ยวข้อง

**2.1 การเรียนรู้ของเครื่อง (Machine Learning)**

Machine Learning เป็นเครื่องมือที่ดีที่สุด ณ ปัจจุบันเพื่อวิเคราะห์, เข้าใจ และหา pattern ของข้อมูล หนึ่งในแนวคิดหลักภายใต้ Machine Learning คือการที่คอมพิวเตอร์สามารถถูก train อย่างอัตโนมัติซึ่งสามารถทำได้อย่างหมดจดหรือเป็นไปไม่ได้สำหรับที่มนุษย์จะทำ และยังมีช่องโหว่ที่ชัดเจนจากการวิเคราะห์ยุคก่อนคือการที่ Machine Learning สามารถตัดสินใจได้ด้วยการแทรกแซงจากมนุษย์เพียงเล็กน้อย

Machine Learning ใช้ข้อมูลเพื่อส่งต่อเข้าไปในอัลกอริทึมซึ่งสามารถเข้าใจความสัมพันธ์ระหว่างข้อมูลขาเข้าและขาออกได้ เมื่อ machine สิ้นสุดการเรียนรู้แล้ว มันสามารถทำนายมูลค่าหรือประเภทของข้อมูลใหม่ได้



ภาพที่ 2-1 กระกระบวนการเรียนรู้ของเครื่อง

ที่มา : <https://www.guru99.com/images/tensorflow/083018_0454_MachineLear2.png>

กระบวนการของ Machine Learning

สมมติว่าคุณต้องการจะสร้างโปรแกรมที่มีการจดจำวัตถุขึ้นมา เพื่อที่จะฝึก(train) model นั้น,คุณจะต้องใช้ตัวแบ่งประเภท(classifier) classifier ใช้คุณลักษณะ(feature)ของวัตถุเพื่อพยายามหาประเภทของวัตถุนั้น

ในตัวอย่าง, classifier จะถูก train เพื่อตรวจจับ(detect)ถ้ารูปเป็น : จักรยาน, เรือ, รถยนต์, เครื่องบิน

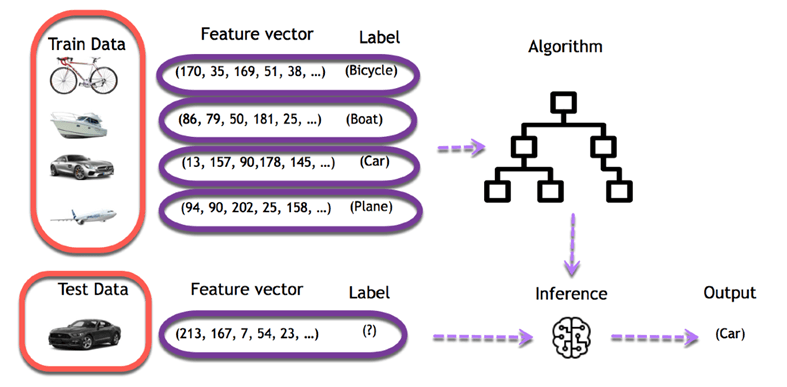
4 วัตถุนี้คือประเภทของวัตถุที่แตกต่างกัน classifierจำเป็นต้องจดจำ เพื่อที่จะสร้าง classifier ขึ้น,คุณจำเป็นต้องมีข้อมูลจำนวนหนึ่งเป็นข้อมูลขาเข้าและกำหนดให้เลยว่าข้อมูลแต่ละอันเป็นประเภทอะไร อัลกอริทึมนี้จะนำข้อมูลไปหารูปแบบ(pattern)แล้วแบ่งประเภทของข้อมูลตามประเภทต่าง ๆ

การกระทำรูปแบบนี้เรียกว่า การเรียนรู้แบบมีผู้สอน(Supervised Learning) ใน Supervised Learning,ข้อมูลสำหรับการฝึก (training data) คุณจะส่งต่อไปในอัลกอริทึมพร้อมกับทำสัญลักษณ์เพื่อบอกผลลัพธ์ไว้แล้ว

การ train อัลกอริทึมจำเป็นต้องทำตามวิธีการดังนี้ : เก็บข้อมูล, Train classifier, ทำการทำนาย

ขั้นตอนแรกเป็นขั้นตอนที่จำเป็นมาก การเลือกข้อมูลที่ถูกต้องจะนำมาซึ่งอัลกอริทึมที่ประสบความสำเร็จหรือล้มเหลว ข้อมูลซึ่งคุณเลือกมาเพื่อ train จะถูกเรียกว่า คุณลักษณะ (feature) ในตัวอย่างของวัตถุ,feature คือพิกเซล(pixel)ของรูป

แต่ละรูปในแถวของข้อมูลขณะที่แต่ละ pixel เป็นหลักแทน ถ้ารูปของคุณมีขนาด 28 x 28 ชุดข้อมูลจะมีขนาด 784 หลัก ในรูปภาพด้านล่าง,แต่ละรูปจะถูกแปลงเป็นเวกเตอร์ของคุณลักษณะ(feature vector) การทำสัญลักษณ์ไว้(label)เป็นการบอกคอมพิวเตอร์ว่าอะไรอยู่ในรูปภาพนั้น



ภาพที่ 2-2 ประเภทของตัวแบบอัลกอริทึมในการเรียนรูปของเครื่อง

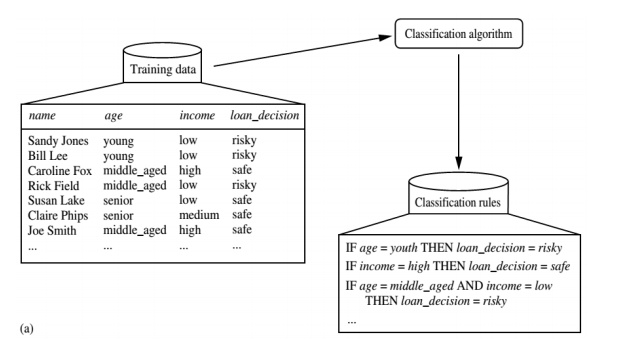
ที่มา : <https://www.guru99.com/images/tensorflow/083018_0454_MachineLear3.png>

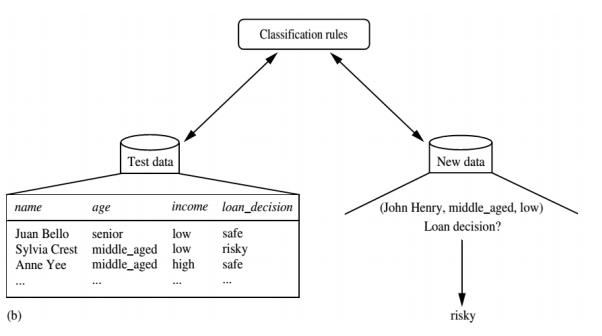
วัตถุประสงค์หลักจะเปนการใช้ training data เพื่อแบ่งประเภทของชนิดของวัตถุ ในขั้นตอนแรกประกอบไปด้วยการสร้าง feature เป็นหลัก แล้วต่อมา,ขั้นตอนที่ 2 เกี่ยวข้องกับการเลือกอัลกอริทึมเพื่อ train model นั้น เมื่อ train เสร็จแล้ว,model ดังกล่าว จะทำนายว่ามีสิ่งใดอยู่ในรูปภาพบ้าง

หลังจากนั้น , มันง่ายมากที่จะใช้ model นั้นไปทำนายรูปภาพอื่น ๆ ต่อไป สำหรับการนำรูปภาพใหม่ ๆ เข้าไปสู่ model นั้น ,machine นั้นจะทำนายประเภทของวัตถุนั้นว่าอยู่ประเภทไหน ยกตัวอย่างเช่น คุณมีรูปภาพใหม่ทั้งหมดอยู่โดยปราศจากการ label ไว้ จึงนำไปใช้กับ model ดังกล่าว สำหรับมนุษย์มันเป็นสิ่งง่ายดายมากตอบว่ารูปภาพนั้นมีรถอยู่แต่ machine ใช้ความรู้ที่เพิ่งกล่าวมาทำได้มากสุดแค่ทำนายว่ามีรถอยู่ในรูปภาพนั้นเท่านั้นเอง

**2.2 การจัดหมวดหมู่ (Classification)**

การจำแนกข้อมูลจะประกอบไปด้วยสองกระบวนการหลัก ดังแสดงตัวอย่างในรูปที่ 6.1 ที่เป็นการกู้-ยืมเงิน โดยจากรูปที่ 6.1(a) จะเป็นกระบวนการสร้างตัวจำแนกข้อมูลจากชุดของข้อมูลที่เป็นอินพุต ที่ซึ่งแต่ละเรคคอร์ดของข้อมูลที่ทำการพิจารณาจะประกอบไปด้วยเซตของแอทริบิวที่บ่งบอกถึงคุณลักษณะของบุคคลที่ทำการกู้-ยืมเงิน และหมวดหมู่ของบุคคลนั้นๆว่ามีความปลอดภัยหรือมีความเสี่ยงในการให้กู้-ยืมเงินหรือไม่โดยกระบวนการสร้างตัวจำแนกข้อมูลมักถูกเรียกว่า ‘learning’ หรือ ‘training’ ที่เกิดจากการนำเอาขั้นตอนวิธีสำหรับการจำแนกข้อมูลมาดำเนินการกับข้อมูล ข้อมูลเรคคอร์ด X หนึ่งๆ ในชุดข้อมูลที่ทำการพิจารณา





ภาพที่ 2-3 การจัดหมวดหมู่

ที่มา หน้าที่ 2: https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-Classification.pdf

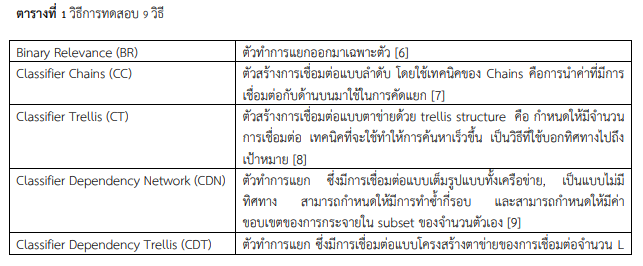
รูปที่ 6.1 ตัวอย่างกระบวนการในการจำแนกข้อมูล (a) การเรียนรู้จากข้อมูลเพื่อสร้างตัวจำแนกข้อมูล (b) การทดสอบตัวจำแนกข้อมูลเพื่อวัดความถูกต้อง

ประกอบไปด้วยเซตของแอทริบิว 𝑋 = (x1, x2, … , x𝑛), 𝑛 แอทริบิวที่บ่งบอกถึงคุณลักษณะต่างๆของข้อมูลเรค-คอร์ด 𝑋 นอกจากนั้นเรคคอร์ด 𝑋 ยังมีข้อมูลอีกหนึ่งแอทริบิวที่บ่งบอกถึงหมวดหมู่ของข้อมูล (class label attribute) โดยแอทริบิวหมวดหมู่ข้อมูลจะเป็นข้อมูลแบบไม่ต่อเนื่อง (discrete-valued) โดยชุดข้อมูลที่เป็นอินพุตสำหรับทำการสร้างตัวจำแนกข้อมูลจะถูกเรียกว่า “ชุดข้อมูลสำหรับสอน (training data)ตัวอย่าง (samples/instances) ชุดข้อมูล (data points) or สิ่งของ (objects)” เป็นต้น หมายเหตุ—เนื่องจากแต่ละเรคคอร์ดในชุดข้อมูลที่เป็นอินพุตจะมีแอทริบิวหมวดหมู่ข้อมูลแนบอยู่ด้วย ดังนั้น การจำแนกข้อมูลด้วยข้อมูลลักษณะนี้จะเรียกว่า การเรียนรู้แบบมีผู้สอน (Supervised learning—คือ การสร้างตัวจำแนกข้อมูลจะถูกสอนโดยแอทริบิวหมวดหมู่ข้อมูลต่างๆที่ถูกแนบอยู่ในแต่ละเรคคอร์ดของชุดข้อมูล โดยการเรียนรู้แบบมีผู้สอนจะแตกต่างกับการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning หรือ clustering) ที่จะไม่ทราบถึงหมวดหมู่ของข้อมูล ตัวอย่างเช่น ในการวิเคราะห์ข้อมูลการกู้-ยืมเงินที่ไม่มีหมวดหมู่ข้อมูลที่บ่งบอกว่าการกู้ยืมครั้งหนึ่งๆมีความเสี่ยงหรือไม่ เราจะสามารถวิเคราะห์ข้อมูลได้จากการเชื่อมโยงเรคคอร์ดของการกู้-ยืมเงินที่มีลักษณะใกล้เคียงกันหรือเหมือนกันให้อยู่ในกลุ่มเดียวกัน เป็นต้นขั้นตอนที่สองของการจำแนกข้อมูล (ดังแสดงในรูปที่ 6.1(b)) จะเป็นการเรียกใช้ตัวจำแนกข้อมูลที่สร้างขึ้นจากขั้นตอนที่หนึ่งเพื่อทำการจำแนกข้อมูล โดยในตอนเริ่มต้น ตัวจำแนกข้อมูลจะถูกทดสอบและประเมินค่าความถูกต้อง (ถ้าเราใช้ชุดข้อมูลสำหรับสอนในการทดสอบตัวจำแนกข้อมูลจะทำให้ความถูกต้องจำมีค่าค่อนข้างสูง เนื่องจากตัวจำแนกข้อมูลที่สร้างขึ้นจะเหมาะกับข้อมูลชุดนั้นเป็นอย่างมาก (overfit) แต่ถ้าเราใช้ชุดข้อมูลที่แตกต่างออกไปในการทดสอบ (test set) โดยชุดข้อมูลที่ใช้จะต้องมีแอทริบิวหมวดหมู่ข้อมูลแนบอยู่ด้วย จะทำให้เราทราบค่าความถูกต้องของตัวจำแนกข้อมูลได้) โดยค่าความถูกต้องของตัวจำแนกข้อมูลที่ถูกสร้างขึ้นจะเป็นเปอร์เซ็นต์ของตัวจำแนกข้อมูลที่สามารถจำแนกข้อมูลได้อย่างถูกต้อง (ตัวจำแนกข้อมูลบ่งบอก

ถึงหมวดหมู่ข้อมูลได้เหมือนกับหมวดหมู่ข้อมูลที่ถูกแนบมากับข้อมูลเรคคอร์ดหนึ่งๆ)เมื่อค่าความถูกต้องของตัวจำแนกข้อมูลมีค่าที่น่าพึงพอใจหรือยอมรับได้ เราจะใช้ตัวจำแนกข้อมูลในการจำแนกหรือบ่งบอกถึงหมวดหมู่ข้อมูลที่เข้ามาใหม่ที่ซึ่งเราไม่ทราบหมวดหมู่ข้อมูลมาก่อน (ข้อมูลที่เข้ามาใหม่จะถูกเรียกว่า ‘unknown’ หรือ ‘previously unseen’ data) ตัวอย่างเช่น ตัวจำแนกข้อมูลที่ถูกสร้างขึ้นในรูปที่ 6.1(a) จะถูกใช้เพื่อตัดสินใจการให้กู้-ยืมเงินของเอกสารที่ยื่นเข้ามาใหม่ว่าจะให้กู้-ยืมหรือไม่

**2.3 อัลกอริทึมการจำแนกประเภทหลายฉลาก (Multi-Label Classification)**

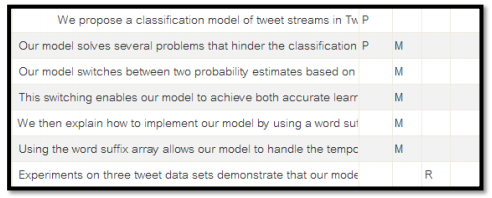
การเรียนรู้จากชุดของตัวอย่าง ปัญหาการจำแนกแบบมัลติเลเบลจัดหมวดหมู่แบบไบนารี(Binary Relevance Methods) ปัญหาการจำแนกประเภทหลายชั้น (multi-label classification) โดยผู้วิจัยเลือกเทคนิคที่ใช้ในการจำแนกข้อมูลแบบไบนารี โดยใช้วิธีการจำแนกแบบมัลติเลเบล จำนวน 9 รูปแบบ ดังนี้





ภาพที่ 2-4 อัลกอริทึมการจำแนกประเภทหลายฉลาก

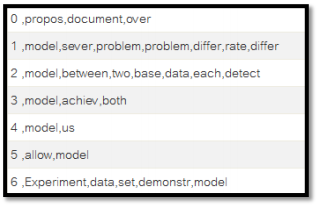
ที่มา หน้าที่ 2: https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-Classification.pdf



ภาพที่ 2-5 ตัวอย่างการจำแนกบริบทหน้าที่แต่ละประโยค

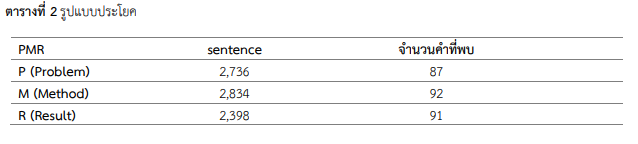
ที่มา หน้าที่ 3: https://www.stou.ac.th/website/PMR/file/A%20Study%20of%20Sentence%20Classification%20Using%20Multi-label%20Classifiers.pdf

การกำหนดคุณลักษณะ ทำการแปลงข้อมูลโดยใช้กระบวนการ NLP (Natural language processing)ให้อยู่ในลักษณะที่เหมาะสมในการวิเคราะห์ข้อมูลสำหรับด าเนินการและการจำแนกบริบทหน้าที่ของประโยคโดยการหารากศัพท์ (Stemming) และการกำจัดคำหยุด (Stopping) เมื่อได้ประโยคจากขั้นตอนการสกัดคำและการหาบริบทหน้าที่แล้ว ต่อไปจึงน าคำเหล่านี้มาหารากศัพท์โดยใช้อัลกอริทึม Porter[16] ซึ่งเป็นอัลกอริทึมที่ได้รับความนิยม โดยปรับเปลี่ยนคำท้าย (Suffix) ในภาษาอังกฤษ แล้วจึงกำหนดให้ตัวอักษรตัวเล็กทั้งหมด เช่น car, cars, car's, cars = car การกำจัดคำหยุด เป็นกระบวนการการกำจัดคำที่ไม่จำเป็นสำหรับการวิเคราะห์ออกให้เหลือเฉพาะคำที่ความสำคัญเท่านั้น คำเหล่านี้เรียกว่า“Stoplist” ตัวอย่างของคำที่เป็นคำหยุด เช่น a,about, above, across, after, again, against เป็นต้น จากนั้นผู้วิจัยทำการลดขนาดคุณลักษณะด้วยค่าความถี่ (frequency)ของคำโดยการเลือกคำที่เกิดขึ้นมากที่มากที่สุด 100 คำในบทคัดย่อที่ได้รวบรวมเพื่อเป็นตัวคัดเลือกคุณลักษณะ จากนั้นกำหนดขนาดคุณลักษณะโดย D={w1,w2,…wn} และน าข้อมูลที่ทำการแปลงค่าเรียบร้อยแล้วน าไปสู่ขั้นตอนที่ 3.2 ต่อไป



ภาพที่ 2-6 ตัวอย่างการกำหนดคุณลักษณะ

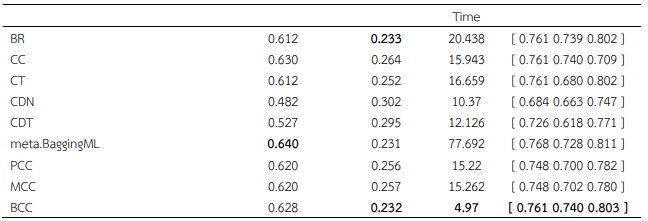
ที่มา หน้าที่ 4: https://www.stou.ac.th/website/PMR/file/A%20Study%20of%20Sentence%20Classification%20Using%20Multi-label%20Classifiers.pdf

โดยแบ่งข้อมูลออกเป็น 3 กลุ่มคือ P (Problem)= 2,736 M (Method)= 2,834 R(Result)= 2,398 ทดสอบด้วย 9 วิธีดังกล่าวข้างต้น โดยผู้ศึกษาได้ทดลองกับค่าพารามิเตอร์ให้เหมาะสมกับการจำแนกข้อมูลในแต่ละวิธี การทดสอบได้ค่า 4 ค่าคือ ค่าความถูกต้อง (Accuracy), ค่าระยะที่หายไป (Hamming loss), เวลาในการสร้าง (BuildTime), ค่าความถูกต้องในแต่ละ label (Accuracy (per label)) มาประเมินประสิทธิ์ภาพอัตราการจำแนกโดยคิดเป็นร้อยละความถูกต้องเมื่อเทียบกับจำนวนที่ถูกต้องโดยใช้เครื่องมือในการทดสอบคือโปรแกรม Meka [15] และใช้วิธี10 Cross-validationในการหาค่าต่างๆเพื่อประเมินประสิทธิภาพในครั้งนี้

ภาพที่ 2-6 ตัวอย่างการกำหนดคุณลักษณะ

ที่มา หน้าที่ 4: https://www.stou.ac.th/website/PMR/file/A%20Study%20of%20Sentence%20Classification%20Using%20Multi-label%20Classifiers.pdf





ตารางที่3 แสดงผลการทดลอง Multi-label 9 วิธี**สรุปผลการทดลอง**

การใช้วิธีการ Multi-label ในแต่ล่ะวิธีจำเป็นที่จะต้องเลือกวิธี classification ที่เหมาะสมในแต่ล่ะวิธีด้วยเช่นกัน โดยผลการทดสอบนั้นได้แสดงให้เห็นถึงการเลือกใช้การ classification ในรูปแบบที่แตกต่างกันออกไปร่วมกับวิธีMulti-label โดยเลือกใช้ทั้ง 9 วิธีเป็นตัวทดสอบกระบวนการ การทดลองในครั้งนี้ในแต่ละวิธีมีข้อดีข้อเสียที่แตกต่างกันออกไปขึ้นอยู่กับผู้ใช้งานต้องการที่จะเน้นค่าอะไรเป็นพิเศษ อาจจะนำวิธีอื่นๆเพื่อมาทดสอบโดยการเปลี่ยนแปลงค่าของวิธี Multi-label เป็นวิธีอื่นๆหรือเปรียบเทียบกับข้อมูลชนิดอื่นๆ เพื่อเปรียบเทียบประสิทธิภาพกับวิธี classification เพื่อให้ได้วิธีที่เหมาะสมและได้ค่าความถูกต้องที่ดีที่สุดในการทดลองกับข้อมูลในแต่ละประเภทที่แตกต่างกันออกไปในงานวิจัยในครั้งต่อไป ซึ่งงานวิจัยชิ้นนี้ต้องการเน้นที่เรื่องของค่าAccuracy ซึ่งจากการทดสอบ ค่า Accuracy อยู่ในช่วง 0.482-0.640 ซึ่งยังให้ผลที่ไม่น่าพอใจสำหรับการทดลองกับข้อมูลชุดนี้ดังนั้นผู้วิจัยจึงต้องท าการทดสอบเพื่อหาวิธีอื่นๆที่สามารถให้ค่า Accuracy ที่สูงกว่าเดิมเพื่อให้ได้การท านายข้อมูลที่ถูกต้องแม่นยำมากยิ่งขึ้น

**2.4 งานวิจัยที่เกี่ยวข้อง**

วัชเรศ ขันธิโชต (2560) การจำแนกประโยคด้วยตัวจำแนกแบบมัลติเลเบล A Study of Sentence Classification Using Multi-label Classifiers วันที่ 25 พฤษภาคม (2560) วิทยานิพนธ์เล่มนี้มุ่งเน้นในการสร้างข้อมูลสนับสนุนการค้นหาเอกสารงานวิจัยเพื่อให้ตรงตามความต้องการของผู้อ่านมากที่สุด โดยมีมุมมองในประเด็นของคำสำคัญ (keyword) ที่ถูกน ามาใช้ในการสืบค้นว่า โดยส่วนใหญ่แล้วมนุษย์มักใช้คำที่คุ้นเคย หรือคำที่เป็นที่รู้จักโดยทั่วไปในการค้นหา ซึ่งคำเหล่านี้เมื่อปรากฏในเอกสารต่างๆ อาจมีบริบทหน้าที่ที่แตกต่างกัน เช่น คำ neural networkอาจเป็นปัญหาที่พบในเอกสารหรืองานวิจัยกลุ่มหนึ่ง แต่อาจเป็นวิธีการที่น ามาแก้ปัญหาในอีกกลุ่มของเอกสารหรืองานวิจัยอื่นได้เป็นต้น ด้วยเหตุนี้ผู้วิจัยมีความเห็นว่าหากระบบสืบค้นสามารถระบุบริบทหน้าที่ของคำสำคัญ อันได้แก่ ปัญหา วิธีการ ผลลัพธ์หรือข้อสรุป ในเอกสารแต่ละชิ้นได้ ย่อมเกิดประโยชน์ต่อผู้อ่านโดยตรง

อัครา ประโยชน (2553) การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวเคราะห์โรคอัตโนมัติ วัตถุประสงค์ของงานวิจัยครั้งนี้ มุ่งเน้นการค้นหาเทคนิคด้านเหมืองข้อมูล เพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติเพื่อค้นหาอลกอริทึมที่เหมาะสมที่สุดสสำหรับฐานข้อมลทางการแพทย์ โดยใช้การ Classification function Network, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Ripperรวมถึงการศกษาเปรียบเทียบการลดคุณลักษณะที่เหมาะสมด้วย วิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) รวมถึงทดสอบกับอัลกอริทึมประเภท Single learning กับ Multiple learning โดยเพิ่มประสทธิภาพด้วยวิธี Bagging และ Boosting โดยทดสอบกับข้อมูลทางการแพทย์ทั้ง 13 ชุด

ดร.นิตยา เกิดประสพ (2560) เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอาย งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาตัวแบบการพยากรณ์ภาวะข้อเข่าเสื่อมในผู้สูงอายุ จากข้อมูลแบบบันทึกการประเมินข้อเข่าเสื่อม โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช จำนวน370 เรคอร์ด และมีข้อมูล 4 คลาส ได้แก่คลาส 0 ยังไม่พบอาการผิดปกติ 200 เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 115 เรคอร์ดคลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 39 เรคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 16 เรคอร์ด สำหรับการวินิจฉัยทางการแพทย์ ข้อมูลกลุ่มน้อย คือ ข้อมูลที่สนใจและการจำแนกผิดพลาดเกิดขึ้นได้สูงกว่าข้อมูลกลุ่มมาก ซึ่งข้อมูลชุดนี้มีจำนวนรวมของคลาส 0 และคลาส 1 สูงกว่าคลาส 2 และคลาส 3 เป็นจำนวนมาก จึงเกิดความไม่สมดุลของข้อมูล ส่งผลให้การจำแนกข้อมูลผิดพลาดได้การปรับความไม่สมดุลของข้อมูลคลาส 2 และคลาส 3 ทำได้ด้วยเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม โดยใช้วิธี ADASYN และSMOTE และใช้งานวิธีการตรวจสอบไขว้แบบ 10 กลุ่ม ในการแบ่งเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ จากนั้นจำแนกข้อมูลด้วย multi-class imbalanced data classification

พุทธิพร ธนธรรมเมธ (2560) ตัวแบบการพยากรณ์ที่นำเสนอในงานวิจัยนี้เป็นหนึ่งในเทคนิคของการทำ Classification ข้อมูลที่มีจุดเด่นคือ ค้นหาตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับการพยากรณ์ภาวะโรคข้อเข่าเสื่อมในผู้สูงอายุเพื่อทำให้สามารถวินิจฉัยภาวะโรคข้อเข่าเสื่อมในระดับปานกลางและในระดับรุนแรงได้ถูกต้องเพิ่มขึ้นและรวดเร็ว โดยใช้ข้อคำถามเพียง 7 ข้อ จาก 14 ข้อ (ข้อคำถามในแบบประเมินข้อเข่าเสื่อม 12 ข้อ รวมเพศและ BMI) ซึ่งการลดจำนวนข้อคำถามยังคงทำให้พยากรณ์ภาวะโรคข้อเข่าเสื่อมได้ถูกต้อง และมีข้อดีคือ ทำให้ลดจำนวนข้อมูลที่ต้องเก็บรวบรวมจากผู้ป่วยลดเวลาการทำงานของหน่วยพยาบาลปฐมภูมิ และลดเวลาการกรอกข้อมูลของผู้ป่วย แต่มีข้อจำกัดของจำนวนชุดข้อมูล ดังนั้นการพัฒนางานวิจัยให้มีประสิทธิภาพเพิ่มขึ้นจะต้องเพิ่มปริมาณข้อมูลที่นำมาใช้สร้างตัวแบบเพื่อให้มีข้อมูลที่หลากหลาย ซึ่งจะส่งผลให้ตัวแบบมีประสิทธิภาพเพิ่มขึ้น

เยาวเรศ ศิริสถิตกุล (2560) การประเมินเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มโดยนำวิธี Classification และ Multi-Label มาใช้ปรับสมดุลข้อมูล ซึ่งเป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อย พบว่าวิธี ADASYN สามารถเพิ่มจำนวนข้อมูลกลุ่มน้อยของคลาส 2 และคลาส 3 ได้สูงกว่าวิธี SMOTE เนื่องจากในการเพิ่มข้อมูลเทียมนั้น วิธี ADASYN จะพิจารณาจากข้อมูลจริงที่ยากต่อการแบ่งกลุ่มข้อมูล ซึ่งไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย ถ้าข้อมูลตัวใดยากต่อการแบ่งกลุ่มก็ให้ค่าน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาบริเวณนั้น ๆ ซึ่งทำให้มีการปรับขอบเขตของเส้นการตัดสินใจในการแบ่งกลุ่มให้ดี[27] และพบว่าการปรับเพิ่มข้อมูลด้วยวิธีสุ่มด้วยวิธี ADASYN ทำให้ประสิทธิภาพของตัวแบบสูงกว่าวิธี SMOTE และสอดคล้องกับ Wang และคณะซึ่งพบว่าวิธี SMOTE ส่งผลต่อการทำนายต่ำกว่าวิธี CSC และในงานของ Shoorangiz และคณะ [34]ได้นำวิธี SMOTE และ ADASYN มาใช้ในการทำนายEEG ของภาวะหลับในหรือการหลับระยะสั้น ๆ ผลการทำนายแสดงค่า ROC ของทั้ง 2 วิธีอยู่ในระดับ 90 %เท่ากัน วิธี ADASYN ให้ค่า sensitivity ที่ 76 % แต่วิธีSMOTE อยู่ที่ 70 % ดังนั้น วิธี SMOTE จึงไม่ได้เป็นวิธีที่มีประสิทธิภาพดีที่สุดในการเพิ่มจำนวนข้อมูลสำหรับการทำนาย

**บทที่ 3**

**วิธีดำเนินการวิจัย**

ส่วนการดำเนินงานพัฒนาหุ่นยนต์ประเมินคะแนนจากการสัมภาษณ์โดยจำแนกแบบมัลติเลเบล มหาวิทยาลัยราชภัฏจันทรเกษม ผู้วิจัยได้กำหนดวิธีดำเนินงานออกเป็น 2 ส่วน ดังนี้

3.1 สร้างโมเดล

3.2 การพัฒนาซอฟต์แวร์ด้วยการจำแนกแบบมัลติเลเบล

**3.1 สร้างโมเดล**

**3.1.1 การจัดเตรียมข้อมูล**

การเตรียมข้อมูลการตั้งคำถามเพื่อนำมาทำการทดสอบโดยหลักการให้คะแนน กยศ. เรื่องคำถามทั่วไป จากเว็บไซต์ กองทุนเงินให้กู้ยืมเพื่อการศึกษา (กยศ.)

(1.) สาขาวิชาที่เรียน

(2.) สถานะครอบครัวผู้สัมภาษณ์

(3.) ผู้สัมภาษณ์มีสัญชาติประเทศอะไร

(4.) ผลการเรียนเกรดเฉลี่ย

(5.) เคยสำเร็จการศึกษาจากที่มหาลัยใดมาก่อน

(6.) ครอบครัวมีรายได้ 1 ปี ทั้งหมด

(7.) ผู้สัมภาษณ์ได้ทำงานพิเศษ หรือไม่

(8.) ทำไมถึงเลือกมาเรียนสาขานั้น

(9.) เคยทำประโยชน์ต่อสังคม / สาธารณะ อะไรบ้าง รวมทั้งสิ้น กี่ ชั่วโมง

(10.) ผู้สัมภาษณ์มีจำนวนกี่คนในครอบครัว

เพื่อทำการศึกษาหาวิธีจำแนกบริบทหน้าที่ของประโยคโดยแบ่งออกเป็น 2 ขั้นตอนหลัก คือ ขั้นตอนที่1 คือการเตรียมข้อมูลจำแนก

บริบทหน้าที่ของประโยคโดยผู้เชี่ยวชาญ และ ขั้นตอนที่ 2 คือกระบวนการปรับข้อมูลเพื่อเข้าสู่การทดลอง