

รายงาน

โครงงานการวิเคราะห์ข้อมูลประเภทจัดกลุ่ม (Categorical) และตัวเลข (Numerical)

เรื่อง “การทำนายการรอดชีวิตจากภัยเรือไททานิค”

โดย

|  |  |  |
| --- | --- | --- |
| นางสาวณัฐชานันท์ | ล้อดี | รหัสประจำตัวนิสิต 6530200118 |
| นางสาววีรญา | เล็กชะอุ่ม | รหัสประจำตัวนิสิต 6530200479 |
| นางสาวกฤติมา | เชาวน์ดี | รหัสประจำตัวนิสิต 6530200568 |
| นายนนท์ปวิธ | ศิลาบำรุงราษฎร์ | รหัสประจำตัวนิสิต 6530200665 |

รายงานเล่มนี้เป็นส่วนนึงของการศึกษารายวิชา 01418362 การเรียนรู้ของเครื่องเบื้องต้น

หลักสูตรวิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์ ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

ภาคปลาย ปีการศึกษา 2567

**# ตัวอย่างข้อมูลที่ใช้ในโครงงาน**

**A screenshot of a computer

AI-generated content may be incorrect.**

**รูปภาพ 1 ภาพรวมข้อมูลที่ใช้ในการทำโครงงาน**

ในโครงงานนี้ ผู้จัดทำได้นำข้อมูลมาจากเว็บไซต์ Kaggle.com โดยใช้ชุดข้อมูลในหัวข้อ “Titanic – Machine Learning from Disaster” ซึ่งเป็นข้อมูลที่รวบรวมรายชื่อผู้โดยสารที่อยู่บนเรือ Titanic ในเหตุการณ์อุบัติเหตุทางทะเลที่เกิดขึ้นในปี 1912 ข้อมูลดังกล่าวประกอบด้วยคุณสมบัติต่างๆ ของผู้โดยสาร เช่น ชื่อ, อายุ, เพศ, สถานการณ์รอดชีวิต รวมถึงข้อมูลเกี่ยวกับการเดินทางบนเรือ เช่น ตั๋วโดยสาร, ค่าธรรมเนียมการโดยสาร, และตำแหน่งของผู้โดยสารในเรือ ข้อมูลนี้ได้รับการนำมาใช้ในการพัฒนาโมเดลแบบมีผู้ฝึกสอน สำหรับการทำนายการรอดชีวิตของผู้โดยสารในเหตุการณ์ดังกล่าว

**# คำสั่งและผลลัพธ์**

# เรียกดูตัวอย่างข้อมูลที่จะใช้ในการฝึกสอนโมเดล เราพบว่าข้อมูลมีทั้งหมด 12 ฟีเจอร์ ได้แก่ เลขลำดับผู้โดยสารบนเรือ, สถานะการเสียชีวิตจากอุบัติเหตุ, ชั้นโดยสาร, ชื่อผู้โดยสาร, เพศ, อายุ, จำนวนพี่น้องหรือคู่สมรสที่เดินทางมาด้วย, จำนวนพ่อแม่หรือบุตรที่เดินทางมาด้วย, หมายเลขตั๋วโดยสาร, ค่าโดยสาร, หมายเลขห้องโดยสาร และท่าเรือที่ขึ้นเรือ

**A screenshot of a computer

AI-generated content may be incorrect.**

# ลบข้อมูลบางฟีเจอร์ที่ไม่ได้ส่งผลกระทบโดยตรงต่อการเรียนรู้ของโมเดล เราพิจารณาว่า ข้อมูลบางฟีเจอร์ ได้แก่ เลขลำดับผู้โดยสารบนเรือ, ชื่อผู้โดยสาร, หมายเลขตั๋วโดยสาร และหมายเลขห้องโดยสาร ไม่ได้ส่งผลกระทบโดยตรงต่อการเรียนรู้ของโมเดล เราจึงตัดฟีเจอร์เหล่านั้นออกจากข้อมูลสำหรับการฝึกสอน

**A screen shot of a computer

AI-generated content may be incorrect.**

# จัดการข้อมูลที่สูญหาย หลังจากเราตัดฟีเจอร์ที่ไม่ส่งผลกระทบโดยตรงต่อการฝึกสอนโมเดลแล้ว เราตรวจสอบพบว่ามีบางฟีเจอร์ที่มีข้อมูลสูญหาย เช่น อายุของผู้โดยสาร และท่าเรือที่ขึ้นเรือ เราจึงแก้ไขข้อมูลสูญหายเหล่านั้นโดยแบ่งออกเป็น 2 กรณีดังนี้ 1. ข้อมูลที่มีลักษณะเป็น Categorical data หรือข้อมูลประเภทกลุ่ม เราจัดการข้อมูลที่สูญหายโดยการหาฐานนิยมของข้อมูลทั้งหมดที่มีอยู่ เช่น ฟีเจอร์ท่าเรือขึ้นโดยสาร 2. ข้อมูลที่มีลักษณะเป็น Numerical data หรือข้อมูลประเภทตัวเลข เราจัดการข้อมูลที่สูญหายโดยการหาค่าเฉลี่ยของข้อมูลทั้งหมดที่มีอยู่ เช่น ฟีเจอร์อายุของผู้โดยสาร

**A screen shot of a computer

AI-generated content may be incorrect.**

# แปลงข้อมูลที่มีลักษณะเป็น Categorical data หรือข้อมูลประเภทกลุ่ม ให้เป็น Numerical data หรือข้อมูลประเภทตัวเลขเพื่อให้โมเดลสามารถเรียนรู้จากข้อมูลเหล่านั้นได้ รวมถึงการปรับช่วงค่าของข้อมูลให้มีลักษณะที่มีช่วงค่าลดลง (อยู่ระหว่าง 0 ถึง 1) เพื่อให้โมเดลสามารถเรียนรู้ข้อมูลฝึกสอนได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

**A screenshot of a computer program

AI-generated content may be incorrect.**

# ตรวจสอบความสมดุลของข้อมูล เราตรวจสอบความสมดุลของข้อมูลสำหรับการเรียนรู้ของโมเดล เราพบว่า ข้อมูลประเภทผู้โดยสารเสียชีวิตมีมากถึง 549 เรคคอร์ด แต่ข้อมูลที่มีลักษณะเป็นผู้รอดชีวิตจากภัยอุบัติเหตุมีเพียง 342 เรคคอร์ด ซึ่งอาจส่งผลต่อการเรียนรู้ของโมเดลได้

**A screen shot of a computer screen

AI-generated content may be incorrect.**

# แบ่งชุดข้อมูล เราแบ่งชุดข้อมูลออกเป็น 2 ชุดข้อมูลได้แก่ 1. ข้อมูลสำหรับการฝึกสอน 2. ข้อมูลสำหรับการทดสอบและประเมินผลโมเดล โดยการแบ่งข้อมูล เราแบ่งข้อมูลสำหรับการฝึกสอนเป็น 80 เปอร์เซ็นต์ และแบ่งข้อมูลสำหรับการทดสอบและประเมินผลโมเดลเป็น 20 เปอร์เซ็นต์ จากข้อมูลทั้งหมด

**A screen shot of a computer code

AI-generated content may be incorrect.**

# เริ่มทำการฝึกสอนโมเดล หลังจากจัดการข้อมูลและแบ่งชุดข้อมูลสำหรับการฝึกสอนและทดสอบแล้ว เราเริ่มทำการฝึกสอนโมเดล โดยเราใช้อัลกอรึทึมทั้งหมด 3 รูปแบบดังนี้ 1. Logistic Regression 2. K-Nearest Neighbors และ 3. Naïve Bayes โดยข้อมูลสำหรับการฝึกสอนและทดสอบ จะใช้ข้อมูลเดียวกันทั้งหมดเพื่อความเท่าเทียมสำหรับการประเมินประสิทธิภาพของโมเดลแต่ละชนิด

**A computer screen shot of a code

AI-generated content may be incorrect.A black background with purple text

AI-generated content may be incorrect.A black rectangular object with a white stripe

AI-generated content may be incorrect.**

**# การประเมินผลของโมเดล (Model Evaluation)**

# การประเมินผลของโมเดล Logistic Regression พบว่ามีอัตราความถูกต้องในการทำนายอยู่ที่ 0.798 หรือคิดเป็น 79.8% จากข้อมูลทั้งหมด

**A screenshot of a computer

AI-generated content may be incorrect.**

# สร้างตาราง Confusion Matrix โดยกำหนดให้คลาส Negative หมายถึง เสียชีวิต และคลาส Positive หมายถึง รอดชีวิต

**A screenshot of a computer

AI-generated content may be incorrect.**

# การประเมินผลของโมเดล K-Nearest Neighbor พบว่ามีอัตราความถูกต้องในการทำนายอยู่ที่ 0.832 หรือคิดเป็น 83.2% จากข้อมูลทั้งหมดA screenshot of a computer

AI-generated content may be incorrect.

# สร้างตาราง Confusion Matrix โดยกำหนดให้คลาส Negative หมายถึง เสียชีวิต และคลาส Positive หมายถึง รอดชีวิต**A screen shot of a computer

AI-generated content may be incorrect.**

# การประเมินผลของโมเดล Naïve Bayes พบว่ามีอัตราความถูกต้องในการทำนายอยู่ที่ 0.765 หรือคิดเป็น 76.5% จากข้อมูลทั้งหมดA screenshot of a computer

AI-generated content may be incorrect.

# สร้างตาราง Confusion Matrix โดยกำหนดให้คลาส Negative หมายถึง เสียชีวิต และคลาส Positive หมายถึง รอดชีวิต**A screenshot of a computer

AI-generated content may be incorrect.**