โครงงาน
เรื่อง การคาดเดาประเภทของกระจก

จัดทำโดย

นายภีมภัช พจน์สุนทร 6209650081
นางสาวพรกนก ศรีสังสิทธิสันติ 6209650214
นางสาวพรไพลิน สวนสิน 6209650537

เสนอ

อาจารย์ ดร. วนิดา พฤทธิวิทยา

โครงงานนี้เป็นส่วนหนึ่งของรายวิชา
หลักการวิทยาการข้อมูล (CS245)
ภาคเรียนที่ 1 ปีการศึกษา 2563
มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต

บทนำ

ชุดข้อมูลที่กลุ่มเราได้นำมาศึกษานั้นเป็นชุดข้อมูลที่เกี่ยวกับเรื่อง glass โดยสาเหตุที่เลือกชุดข้อมูลนี้เพราะเป็นชุดข้อมูลที่สามารถวิเคราะห์ได้ง่ายและมีความสนใจเกี่ยวกับส่วนประกอบของ glass ว่า glass แต่ละชนิดประกอบด้วยธาตุอะไรบ้าง และข้อมูลชุดนี้ยังสามารถนำไปใช้ในเรื่องของ machine learning ได้ ซึ่งในปัจจุบันเรื่องของ glass (กระจก) มักพบบ่อยมากในชีวิตประจำวัน จึงเห็นได้ว่า glass เป็นสิ่งที่สำคัญและยังนำไปใช้ได้หลากหลายด้าน ซึ่งการนำไปใช้นั้นมักจะแตกต่างกันไปตามแต่ละประเภทของธาตุและคุณภาพของแต่ละธาตุ

ในปัจจุบันเป็นยุคที่เทคโนโลยีมีความก้าวหน้าและ glass ก็เป็นตัวเลือกนึงที่สามารถทำให้เทคโนโลยีพัฒนาได้อย่างสมบูรณ์โดยนำไปใช้ในด้านของ

1.ด้านเทคโนโลยี อย่างเช่น หน้าจอโทรศัพท์ หน้าจอคอมพิวเตอร์หรือแม้กระทั่งสามารถนำมาสร้างแทนโปรเจคเตอร์เพื่อให้เกิดภาพที่ดูทันสมัยและแปลกใหม่

2.ด้านอุตสาหกรรม อย่างเช่น ประตู หน้าต่าง โคมไฟซึ่งใช้ได้ทั้งภายในและภายนอกบ้าน

3.ด้านประติมากรรม เช่น สร้างรูปทรง 3 มิติ การแกะสลักบนแผ่นกระจกเพื่อให้เกิดความสวยงาม

จากที่กล่าวมาข้างต้นจะเห็นว่า glass หรือกระจกนั้นมีประโยชน์มากและสำคัญกับในยุคสมัยปัจจุบันมากเพราะต้องมีการนำ glass มาเป็นส่วนประกอบในหลาย ๆ ด้าน เราจึงสนใจที่จะนำข้อมูลนี้มาศึกษาโดยเจาะลึกรายละเอียดว่าในกระจกแต่ละประเภทประกอบด้วยธาตุอะไรบ้างและศึกษาว่าในกระจกแต่ละประเภทมีส่วนประกอบของธาตุนั้นอยู่เท่าไหร่

ข้อมูลที่กล่าวมาข้างต้นนี้ในเรื่องของ glass เป็นความสนใจของสมาชิกในกลุ่มที่อยากจะนำชุดข้อมูลนี้มาศึกษาเพื่อเป็นแนวทางให้กับผู้ที่สนใจในเรื่องนี้เป็นข้อมูลในการนำไปศึกษาต่อสำหรับผู้ที่ต้องการที่จะศึกษาเพิ่มเติมและเพื่อให้เข้าใจถึงหลักการของการที่จะนำชุดข้อมูลไปทำ machine learning โดยเริ่มศึกษาหรือดูชุดตัวอย่างข้อมูลได้จากหน้าถัดไปโดยในรายงานเล่มนี้จะบอกถึงรายละเอียดของโค้ดด้วยเพื่อให้กับผู้ที่อยากนำ dataset ชุดนี้ไปเรียนรู้และนำไปประยุกต์ใช้ต่อไป

ชุดข้อมูล

เป็นชุดข้อมูลของเรื่อง กระจก (Glass) จากเว็บ https://www.kaggle.com/uciml/glass ซึ่งบุคคลผู้สร้างชุดข้อมูลนี้คือ B. German จาก Central Research Establishment โดยภายในชุดข้อมูลมี Rows ทั้งหมด 214 rows และมี Columns ทั้งหมด 10 columns

Rows แต่ละอันจะมีข้อมูลทั้ง 10 อย่าง โดยมี RI, Na, Mg, Al, Si, K, Ca, Ba, Fe เป็นเลขทศนิยมและมี Type เป็นเลขจำนวนเต็ม 1 ถึง 7

Columns ทั้ง 10 ได้แก่ 1.RI: refractive index ค่าดัชนีหักเหของ glass 2.Na: Sodium โซเดียม 3.Mg: Magnesium แมกนีเซียม 4.Al: Aluminum อลูมิเนียม 5.Si: Silicon ซิลิคอน 6.K: Potassium โพแทสเซียม 7.Ca: Calcium แคลเซียม 8.Ba: Barium แบเรียม 9.Fe: Iron เหล็ก 10.Type ประเภทของ glass ซึ่งมีด้วยกันทั้งหมด 7 ประเภท ได้แก่

1.building windows float processed กระจกโฟลตสำหรับอาคาร

2.building windows non float processed กระจกอื่นๆ สำหรับอาคาร

3.vehicle windows float processed กระจกโฟลตสำหรับยานพาหนะ

4.vehicle windows non float processed (none in this database) กระจกอื่นๆ สำหรับยานพาหนะ (ประเภทนี้ไม่มีในข้อมูลชุดนี้)

5.containers กระจกสำหรับทำเป็นที่ใส่ของ

6.tableware กระจกสำหรับชุดเครื่องใช้บนโต๊ะอาหาร

7.headlamps กระจกสำหรับไฟหน้า

การวิเคราะห์ข้อมูล

ส่วนของการวิเคราะห์ข้อมูลได้ทำการเขียนไว้ใน colab notebook สามารถเข้าไปดูรายละเอียดของการวิเคราะห์และการทำงานเกี่ยวกับชุดข้อมูลนี้ได้ตาม link ที่ได้ใส่ไว้ให้ด้านล่างนี้

ลิงก์ : https://colab.research.google.com/drive/1NumoGvZRIbpkiSkdLVwPex99-UwDfTYj?usp=sharing

สรุปผลการดำเนินงาน

จากการวิเคราะห์ชุดข้อมูลที่ได้เลือกมาและนำมาสรุปหัวข้อที่ได้สามารถ แบ่งออกเป็น 5 ส่วนได้แก่ 1.Data Importing 2.Data Cleaning 3.Data Wrangling 4.Data Visualization และ 5.Machine Learning

ส่วนที่ 1 Data Importing

ได้มีการ import ข้อมูลจาก google drive และใช้คำสั่งต่าง ๆ ในการแสดงข้อมูลพื้นฐาน เช่น คำสั่ง head, info, describe, ndim, shape และ size

ส่วนที่ 2 Data Cleaning

ได้มีการตรวจสอบข้อมูลว่ามีค่า Nan หรือค่า NULL หรือไม่ โดยผลลัพธ์ที่ได้คือข้อมูลไม่มีค่า Nan และค่า NULL หลังจากนั้นได้มีการตรวจสอบข้อมูลว่ามีค่าซ้ำกันหรือไม่ โดยผลลัพธ์ที่ได้คือ 1 ซึ่งหมายถึงข้อมูลมีค่าซ้ำอยู่ 1 แถว จึงทำการ drop แถวที่ซ้ำทิ้งไป จากเดิมข้อมูลมี 214 แถว เหลือเพียง 213 แถวแทน และค่าที่ซ้ำคือข้อมูลตำแหน่งที่ 39 หรือข้อมูลตัวที่ 40 ซึ่งเป็นกระจกประเภทที่ 1

ส่วนที่ 3 Data Wrangling

ได้มีการแบ่งกลุ่มตาม Type ซึ่งจากการแบ่งกลุ่มตาม Type ทำให้ทราบว่ากระจกประเภทที่ 1 มีจำนวน 69 ข้อมูล

กระจกประเภทที่ 2 มีจำนวน 76 ข้อมูล กระจกประเภทที่ 3 มีจำนวน 17 ข้อมูลกระจกประเภทที่ 5 มีจำนวน 13 ข้อมูลกระจกประเภทที่ 6 มีจำนวน 9 ข้อมูล และกระจกประเภทที่ 7 มีจำนวน 29 ข้อมูล ซึ่งรวมทั้งสิ้น 213 ข้อมูล และได้เตรียมข้อมูลไว้ใช้ในส่วนอื่น ๆ ยกตัวอย่างเช่น เก็บข้อมูลของแต่ละ Type ลงในตัวแปรของ Type นั้น ๆ เช่น เก็บข้อมูล Type ที่ 1 ลงในตัวแปร Type1 และเก็บชื่อของ column ลงในตัวแปร col_groupType และ col_dataset โดย col_groupType คือเก็บชื่อ column ที่ไม่มี column Type และ col_dataset คือเก็บชื่อ column ที่มี column Type

ส่วนที่ 4 Data Visualization

ได้นำข้อมูลมาทำเป็นกราฟต่าง ๆ เช่น Bar Chart, Pie Chart, Heatmap, scatter matrix, kde (Kernel Density Estimation), Histogram เป็นต้น โดยการทำ Hearmap นั้น ทำให้สามารถดูความสัมพันธ์ในข้อมูลได้ ซึ่งผลลัพธ์ที่ได้นั้น มีหลากหลาย แต่ที่เห็นได้ชัดเจนคือ 1.ความสัมพันธ์ระหว่าง Ca กับ RI มีความสัมพันธ์เชิงบวกอย่างมาก ซึ่งหมายถึงยิ่ง Ca มีค่ามาก RI ก็จะมีค่ามากตาม จึงทำให้สามารถใช้ Ca หรือ RI แทนกันได้ 2.ความสัมพันธ์ระหว่าง Ca, K กับ Type คือไม่มีความสัมพันธ์กันหรือสัมพันธ์กันน้อยมาก ๆ ซึ่งหมายถึง Ca, K ไม่ได้ส่งผลต่อการแบ่ง Type จึงทำให้สามารถ drop column Ca และ K ได้

ส่วนที่ 5 Machine Learning

ได้มีการแบ่งข้อมูลเป็น 2 ส่วน ส่วนแรกไว้ใช้เป็นข้อมูลในการคาดเดาประเภทของกระจกแก่ระบบคอมพิวเตอร์ ส่วนที่สองเป็นข้อมูลที่ให้ระบบคอมพิวเตอร์ใช้คาดเดาประเภทของกระจก โดยแบ่งข้อมูลออกเป็น ส่วนแรก 80% (170 แถว) ส่วนที่สอง 20% (43 แถว) โดยจะให้ระบบคอมพิวเตอร์คาดเดาข้อมูล ต้องเลือก method ให้กับระบบคอมพิวเตอร์ จึงจะทำการหา method ที่มีความแม่นยำสูงที่สุด ผลลัพธ์ที่ได้คือ method CART (Decision Trees) มีความแม่นยำสูงที่สุด จึงใช้ method CART (Decision Trees) ในการคาดเดา ผลลัพธ์ในการคาดเดาของคอมพิวเตอร์เป็นดังนี้

Accuracy Score: 0.7441860465116279

Confusion Matrix

		Predicted					
		ประเภทที่ 1	ประเภทที่ 2	ประเภทที่ 3	ประเภทที่ 5	ประเภทที่ 6	ประเภทที่ 7
A	ประเภทที่ 1	12	2	2	0	0	0
C	ประเภทที่ 2	1	12	0	1	0	0
T	ประเภทที่ 3	2	1	0	0	0	0
U	ประเภทที่ 5	0	0	0	2	0	0
A	ประเภทที่ 6	1	0	0	0	0	0
L	ประเภทที่ 7	0	0	0	0	1	6

จาก Confusion Matrix ทำให้สรุปได้ดังนี้

มีข้อมูลกระจกทั้งหมด 43 บาน

ประเภทที่ 1 ที่คอมพิวเตอร์คาดเดามีจำนวน 16 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภทที่ 1 ผิดไป 4 บาน โดยคาดเดาผิดเป็นประเภทที่ 2 จำนวน 1 บาน ประเภทที่ 3 จำนวน 2 บาน และประเภทที่ 6 จำนวน 1 บาน ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ กระจกประเภทที่ 2 จำนวน 1 บาน อาจมีปัญหาเพราะกระจกประเภทที่ 1 เป็น float processed แต่ประเภทที่ 2 เป็น non float processed ประเภทที่ 3 อาจไม่เป็นปัญหาเพราะเป็น float processed เหมือนกัน แต่เป็นคนละชนิดกัน ประเภทที่ 1 เป็น building windows ประเภทที่ 3 เป็น vehicle windows อาจทำให้ไม่ได้มาตรฐานกระจก ส่วนประเภทที่ 6 จะมีปัญหามากที่สุด เพราะประเภทที่ 6 เป็น tableware ไม่สามารถนำมาใช้เป็น builiding windows ได้

ประเภทที่ 2 ที่คอมพิวเตอร์คาดเดา มีจำนวน 14 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภทที่ 2 ผิดไป 2 บาน โดยคาดเดาผิดเป็นประเภทที่ 1 จำนวน 2 บาน ประเภทที่ 3 จำนวน 1 บาน ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ กระจกประเภทที่ 1 จำนวน 2 บาน อาจมีปัญหาเพราะกระจกประเภทที่ 1 เป็น float processed แต่ประเภทที่ 2 เป็น non float processed ประเภทที่ 3 อาจเป็นปัญหาด้วยเพราะประเภที่ 3 เป็น float processed แต่ประเภทที่ 2 เป็น non float processed

ประเภทที่ 3 ที่คอมพิวเตอร์คาดเดา มีจำนวน 2 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภทที่ 3 ผิดหมด อาจเป็นเพราะในข้อมูลมีกระจกประเภทที่ 3 น้อยจึง

ทำให้คอมพิวเตอร์คาดเดาผิดหมด โดยคาดเดาเป็นประเภทที่ 1 จำนวน 2 บาน ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ ก็อาจไม่เป็นปัญหา เพราะ ประเภทที่ 1 และประเภทที่ 3 เป็น float precessed เหมือนกันแต่ ประเภทที่ 1 เป็น building windows ประเภทที่ 3 เป็น vehicle windows

ประเภทที่ 5 ที่คอมพิวเตอร์คาดเดา มีจำนวน 3 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภทที่ 5 ผิดไป 1 บาน โดยคาดเดาผิดเป็นประเภทที่ 2 ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ อาจเป็นปัญหาใหญ่เพราะประเภทที่ 2 เป็น window glass แต่ประเภทที่ 5 ไม่ใช่ window glass จึงอาจเป็นปัญหาใหญ่

ประเภทที่ 6 ที่คอมพิวเตอร์คาดเดา มีจำนวน 1 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภทที่ 6 ผิดหมด อาจะเป็นเพราะในข้อมูลมีกระจกประเภทที่ 6 น้อยจึงทำให้คอมพิวเตอร์คาดเดาผิดหมด โดยคาดเดาเป็นประเภทที่ 7 จำนวน 1 บาน ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ อาจเป็นปัญหาเพราะกระจกประเภทที่ 6 คือ tableware แต่กระจกประเภทที่ 7 คือ headlamps ซึ่งเป็นคนละประเภทกัน อาจทำให้เกิดปัญหาได้

ประเภทที่ 7 ที่คอมพิวเตอร์คาดเดา มีจำนวน 6 บาน จะเห็นว่าคอมพิวเตอร์คาดเดากระจกประเภที่ 7 ถูกต้องหมด อาจเป็นเพราะกระจกประเภทที่ 7 มีคุณลักษณะของธาตุที่โดดเด่น จึงทำให้คอมพิวเตอร์คาดเดากระจกประเภทที่ 7 ได้ถูกต้องหมด ถ้าหน่วยงานนำข้อมูลนี้ไปใช้ ก็จะใช้ถูกประเภท ไม่เกิดปัญหาอะไรทั้งสิ้น

Classification Report

	Precision	Recall	F1-score	Support
ประเภทที่ 1	0.75	0.75	0.75	16
ประเภทที่ 2	0.86	0.86	0.86	14
ประเภทที่ 3	0.00	0.00	0.00	3
ประเภทที่ 5	0.67	1.00	0.80	2
ประเภทที่ 6	0.00	0.00	0.00	1
ประเภทที่ 7	1.00	0.86	0.92	7

Accuracy			0.74	43
Macro avg	0.55	0.58	0.56	43
Weighted avf	0.75	0.75	0.75	43

จาก Classification Report ทำให้สรุปได้ดังนี้

การที่หน่วยงานจะสามารถนำข้อมูลไปใช้และเกิดข้อผิดพลาดน้อยที่สุดได้ หน่วยงานต้องดูข้อมูลทั้งหมดในช่อง Recall เพราะช่อง Recall เป็นช่องที่คำนวณจำนวนผล Actual ของประเภทนั้น ๆ ที่ระบบคอมพิวเตอร์คาดเดาถูกต้อง ยกตัวอย่าง เช่น Actual ของประเภทที่ 1 มี 16 บาน แต่ระบบคอมพิวเตอร์คาดเดาประเภทที่ 1 ถูกจำนวน 12 บาน และคาดเดาผิดเป็นประเภทที่ 2 จำนวน 2 บาน และประเภทที่ 3 จำนวน 2 บาน ซึ่งหน่วยงานจำเป็นต้องนำข้อมูลประเภทที่ 1 ไปใช้อยู่แล้ว แต่หน่วยงานจะต้องนำข้อมูลประเภทที่ 2 และ 3 ไปใช้งานร่วมด้วย เพราะข้อมูลประเภทที่ 2 และประเภทที่ 3 แท้จริงแล้วคือข้อมูลประเภทที่ 1 ที่ระบบคอมพิวเตอร์คำนวณผิดพลาด รวมทั้งประเภทอื่น ๆ ด้วย เป็นต้น

จากการวิเคราะห์ข้อมูลทั้งหมดจะเห็นได้ว่าจะมีการทำ Data Cleaning และ Data Wrangling เพื่อจัดการกับชุดข้อมูล เพื่อให้สามารถนำชุดข้อมูลนี้ไปใช้งานได้จริง ซึ่งจากที่กล่าวมานี้เราสามารถนำชุดข้อมูลไปใช้ในการทำ Data Visualization และ Machine Learning ได้ โดย Data Visualization เป็นการใช้ชุดข้อมูลเพื่อแสดงออกมาเป็นกราฟความสัมพันธ์ต่าง ๆ ซึ่งสามารถนำข้อมูลมาเปรียบเทียบกันได้ ซึ่งกราฟต่าง ๆ เป็นการนำชุดข้อมูลที่เกี่ยวกับเรื่องของแร่ธาตุในกระจกออกมาแสดงเป็นกราฟแสดงความสัมพันธ์ต่าง ๆ และ machine learning เป็นการทำให้ระบบคอมพิวเตอร์เรียนรู้ได้ด้วยตนเองโดยเรียนรู้จากข้อมูลที่มีให้ ซึ่งความแม่นยำที่ได้จะขึ้นอยู่กับข้อมูลที่มีให้ว่ามีค่ามากแค่ไหน ข้อมูลที่มีเป็นเรื่องของแร่ธาตุในกระจก ข้อมูลจึงเป็น Supervised Learning ซึ่งการทำ Machine Learning ของข้อมูลชุดนี้จึงทำเพื่อให้ระบบคอมพิวเตอร์สามารถคำนวณแยกประเภทของกระจกได้ โดยหวังว่าการทำโครงานนี้จะเป็นประโยชน์แก่ผู้ที่มาศึกษาหรือนำข้อมูลไปใช้เพื่อนำไปต่อยอดหรือดัดแปลงต่อไป