Logo, company name

Description automatically generated

**รายงานประมวลความรู้รายวิชา DSI204**

เรื่อง การพยากรณ์ปริมาณการปล่อยก๊าซคาร์บอนไดออกไซด์ของรถยนต์ในประเทศแคนาดา

**นำเสนอ**

ผศ.ดร.บุญฤทธิ์ ชูประดิษฐ์

**สมาชิก**

นายธนารักษ์ ลีนานนท์ 6524650030

นายศิรภพ จุลละภมร 6524650089

นายวัชรนันท์ พันมูล 6524650071

นายวสันต์ อารัมภ์สกุล 6524651400

นายณิชพน รัถยาบัณฑิต 6524651244

รายงานนี้เป็นส่วนหนึ่งของการศึกษารายวิชา DSI204 Probability Thinking

ตามหลักสูตรวิทยาศาสตร์บัณฑิต สาขาวิทยาศาสตร์และนวัตกรรมข้อมูล

วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์

**การพยากรณ์ปริมาณการปล่อยก๊าซคาร์บอนไดออกไซด์ของรถยนต์ในประเทศแคนาดา**

1.ลักษณะของกลุ่มข้อมูล (Meta Data)

Data set ที่ทางกลุ่มคณะผู้จัดทำได้นำมาศึกษาและทำการทดลองคือ Co2 Emission\_Canada หรือปริมาณการปลดปล่อย Co2 โดยรถยนต์สันดาปประเภทต่าง ๆ ในประเทศแคนาดา ซึ่งมีเนื้อหาข้อมูลเกี่ยวกับรายการองค์ประกอบของเครื่องยนต์, ประเภทเชื้อเพลิง, รุ่นของรถยนต์และองค์ประกอบอื่น ๆ เป็นต้น โดยกลุ่มคณะผู้จัดทำได้เลือกใช้โปรแกรม R-studio ในการทำการทดลองและวิเคราะห์หาประเด็นสำคัญต่าง ๆ ในชุดข้อมูล

นำชุดข้อมูลมาจาก

https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

โดย Data set ที่นำมาชื่อว่า Co2 Emission\_Canada ซึ่งมีรายละเอียดดังนี้

* Make = ชื่อบริษัทที่เป็นผู้ผลิตของรถยนต์
* Model = ชื่อ Model ของรถประจำบริษัทที่ผลิตนั้นๆ ซึ่งประกอบไปด้วย
  + 4WD/4X4 = Four-wheel Drive
  + AWD = All-wheel drive
  + FFV = Flexible-fuel vehicle
  + SWB = Short wheelbase
  + LWB = Long wheelbase
  + EWB = Extended wheelbase
* Vehicle Class = ประเภทของยานพาหนะซึ่งอ้างอิงตามประโยชน์ใช้สอย, ความจุและน้ำหนัก
* Engine size = ขนาดของเครื่องยนต์โดยใช้หน่วยเป็นลิตร
* Cylinders = จำนวนลูกสูบ
* Transmission = ประเภทเกียร์และจำนวน
  + A = Automatic
  + AM = Automated manual
  + AS = Automatic with select shift
  + AV = Continuously variable
  + M = Manual
  + 3 - 10 = Number of gears
* Fuel Type = ประเภทเชื้อเพลิง
  + X = Regular gasoline
  + Z = Premium gasoline
  + D = Diesel
  + E = Ethanol (E85)
  + N = Natural gas
* Fuel consumption in city roads (L/100 Km) = อัตราสิ้นเปลืองเชื้อเพลิงบนถนนในเมือง (ลิตร/100 กม.)
* Fuel consumption in highways (L/100 km) = อัตราสิ้นเปลืองเชื้อเพลิงบนทางหลวง (ลิตร/100 กม.)
* Fuel Consumption Comb (mpg) = อัตราสิ้นเปลืองเชื้อเพลิงแบบผสม (ในเมือง 55% ทางหลวง 45%) แสดงเป็น L/100 กม
* Co2\_Emission = ปริมาณการปลดปล่อยก๊าซ Co2

**การใช้สถิติพรรณนา (Descriptive statistics) เพื่อหาประเด็นสำคัญต่าง ๆ**

ทางกลุ่มคณะผู้จัดทำได้มีการใช้สถิติพรรณนาในเรื่องของการวัดตำแหน่งข้อมูลโดยกลุ่มคณะผู้จัดทำรายงานได้เลือกใช้ Quartile และ Inter Quartile Range (IQR) มาเป็นหลักการในการกำจัดค่านอกเกณฑ์ (Outlier) ของ Feature data ที่ชื่อว่า Engine size และตรวจสอบลักษณะของข้อมูลลว่ามีการแจกแจงแบบปกติมาตรฐานหรือไม่โดยใช้กราฟฮิสโทแกรม

*ตัวอย่าง* code *ภาษา* R

Text

Description automatically generated

*A picture containing diagram

Description automatically generated*

*ทำให้เราทราบว่าใน* Feature Engine size *มี* Q1 = 2.0 *และ* Q3 = 3.7 *และ* IQR = 1.7

*ทำการกำจัด* Outlier *จะพบว่าชุดข้อมูลจะเหลือ* 7248 records *จาก* 7385records

*หลังจากทำการ* cleaned outlier *นำมา* plot histogram *เพื่อดูการกระจายตัวที่เกิดขึ้น*

*Text

Description automatically generated*

*Chart, histogram

Description automatically generated*

***3.*Data preparation**

*จากการพล็อตกราฟฮิสโทแกรมเพื่อดูการกระจายของข้อมูลทำให้ทางกลุ่มคณะผู้จัดทำทราบว่าชุดข้อมูลนี้ไม่มีการกระจายตัวเป็นปกติมาตรฐานเพื่อให้สามารถใช้เทคนิคการถดถอย (*regression) *ได้ทางคณะผู้จัดทำจึงต้องมีการสร้าง* data frame *ชุดใหม่ขึ้นมาโดยเรียกว่า* X *และ* Y *ซึ่งมีรายละเอียด ดังนี้*

X *คือ* data frame *ที่ประกอบไปด้วย* feature *ดังนี้* fuel.type, engine.size.L, fuel consumption *ทั้ง 3 รูปแบบ*

Y *คือ* data frame *ที่เป็น* Labeled data *ที่มีชื่อว่า* Co2.Emission.g.km

*ทำการแปลง* X *ที่ไม่ใช่* Numeric feature *ให้เป็น* Dummy variable

Text

Description automatically generated

*ทำการ* Plot *เพื่อดูสหสัมพันธ์ (*Correlation*) ระหว่าง* X, Y

Text

Description automatically generated

Chart

Description automatically generated

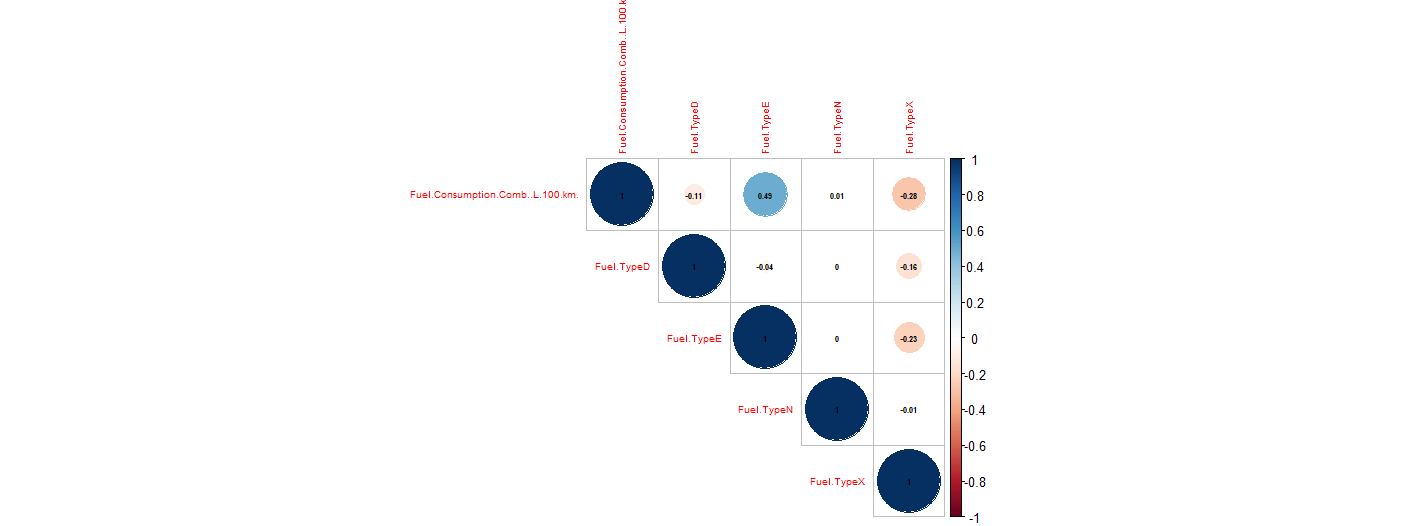
***4.*Feature selection**

*จาก* correlation *ที่เกิดขึ้นทำให้เราสามารถเลือก* X *ที่ส่งผลกับ* Y *ได้โดยอิงจากค่า* correlation *ที่เกิดขึ้น*

*ตรวจสอบค่าความสัมพันธ์ที่เกิดขึ้นหากตัด* feature fuel.type, engine.size.L, fuel consumption *ทั้ง 3 รูปแบบ ออกและสร้างเป็นคอลัมน์ใหม่แทนที่ค่าในคอลัมน์เดิม*

*Text

Description automatically generated*

**

*จาก* correlation *ที่เกิดขึ้นทำให้เราสามารถสร้าง* data frame *ชุดใหม่ที่สามารถนำไปใช้ต่อในเทคนิค* Linear regression *ได้และหากนำไปตรวจสอบความ* linearity *จะได้ดังนี้*

*Text

Description automatically generated*

*Graphical user interface, application

Description automatically generated*

**5.** **Linear Regression**

*ใช้เทคนิค* Linear regression *โดยอิงข้อมูลจาก* new\_df *ในการแบ่งส่วน* train-test *ออกเป็น 70*:30

Text

Description automatically generated

*ทำการสร้าง* model Linear regression *จะได้ว่า*

Text

Description automatically generated

สมการรูปทั่วไปของ Linear regression คือ

*ซึ่งสมการที่เกิดขึ้นของ* data set *ชุดข้อมูล* train *เป็นดังนี้*

Text

Description automatically generated

*กำหนดให้*

Fuel.Consumption.Comb..L.100.Km = X1 Fuel.TypeD = X2 Fuel.TypeE = X3 Fuel.TypeN = X4 Fuel.TypeX = X5

*เราจึงสามารถเขียนสมการรูปทั่วไปของโมเดลได้ดังนี้*

*ซึ่งสามารถตีความผลจาก* Linear regression model *ได้ดังนี้*

*b0 = เมื่อให้ตัวแปรอิสระ* x1-x5 *มีค่าเป็น 0 จะส่งผลให้ y มีค่าเพิ่มขึ้น 6.12994*

*b1 = เมื่อ* Fuel.Consumption.Comb..L.100.Km *มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ y มีค่าเพิ่มขึ้น 22.74454*

*b2 = เมื่อ* Fuel.TypeD *มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ y มีค่าเพิ่มขึ้น* 30.46087

*b3 = เมื่อ* Fuel.TypeE *มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ y มีค่าลดลง 114.54171*

*b4 = เมื่อ* Fuel.TypeN *มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ y มีค่าลดลง 81.98560*

*b5 = เมื่อ* Fuel.TypeX *มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ y มีค่าลดลง 0.37966*

*หากใช้ชุดข้อมูล* test *จะได้ผลของ* model *ดังนี้*

|  |  |
| --- | --- |
| Metrics | value |
| MSE | 30.55181 |
| RMSE | 5.527369 |
| R-squared | 0.9902685 |

1.R-squared หมายถึง ตัวแปรอิสระX1,X2,X3,X4,X5 สามารถอธิบาย y ได้อย่างถูกต้องร้อยละ99.02% ส่วนที่เหลือสามารถอธิบายได้ด้วยปัจจัยอื่นๆ

2.MSE หมายถึง ผลรวมค่าเฉลี่ยสัมบูรณ์ของ Residuals มีค่าอยู่ที่ 30.55181หรือประมาณตามหลักนัยสำคัญคือ 31

3.RMSE หมายถึง ผลรวมค่าเฉลี่ยสัมบูรณ์ที่ไม่ถูกยกกำลังสองของ ของ Residuals มีค่าอยู่ที่ 5.527369หรือประมาณตามหลักนัยสำคัญคือ 6

*เพื่อพิสูจน์* Consumption *ทางคณะผู้จัดทำรายงานจึงได้ทำการใช้* Shapiro-wilk test *และ* plot residual *ซึ่งได้ผลลัพธ์ดังนี้*

*Text

Description automatically generated*

*Chart, scatter chart

Description automatically generated*

*Chart, histogram

Description automatically generated*

*Chart, histogram

Description automatically generated*

***6.สรุปผล***

*จากการทดลอง จะพบว่าโมเดลของเราสามารถใช้ในการพยากรณ์ชุดข้อมูล* Co2 Emission\_Canada *ได้เนื่องจาก* residuals *มีการกระจายตัวอย่างเป็นปกติมาตรฐาน (*Normal Distribution) *ซึ่งแปลว่าโมเดลนี้ผ่าน* Assumption *ในเบื้องต้น และจากสมการรูปทั่วไปทำให้เราทราบว่าปัจจัยที่มีผลต่อการปลดปล่อยก๊าซ* Co2 *มากที่สุดคือ* Fuel type D *หรือ Diesel*