สรุปบทเรียนจากการ
พัฒนาความรู้
รอบการประเมินที่ 1 : ปีงบประมาณ พ.ศ. 2569

หลักสูตร
Data Analytics

จัดทำโดย นายวีระ ปะทะขีนัง
ตำแหน่ง นักวิชาการแผนที่ภาพถ่ายชำนาญการพิเศษ

Data Analytics

เป้าหมายการเรียนรู้

1. เพื่อให้สามารถอธิบายความหมายของการวิเคราะห์ข้อมูล (Data Analytics) ได้
2. เพื่อให้สามารถอธิบาย กระบวนการและความเข้าในข้อมูลได้
3. เพื่อให้สามารถอธิบาย การสำรวจข้อมูลและการทำ Visualization ได้
4. เพื่อให้สามารถอธิบาย การเรียนรู้ของคอมพิวเตอร์ หรือ Machine Learning ได้
5. เพื่อให้สามารถอธิบายการวิเคราะห์แบบพื้นฐาน (Descriptive Analytics) ได้
6. เพื่อให้สามารถอธิบาย การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) ได้
7. เพื่อให้สามารถอธิบาย ระบบข้อเสนอแนะ (Suggestion System) ได้

1. Introduction to Data Analytics

Data Analytics คือ การวิเคราะห์ข้อมูลโดยนำหลักการทางวิทยาศาสตร์พื้นฐานมาใช้ในการวิเคราะห์ชุดข้อมูลที่มีขนาดใหญ่และซับซ้อน ซึ่ง Data Analytics มักเกี่ยวข้องกับการศึกษาข้อมูลที่ผ่านมาในอดีต เพื่อหาแนวโน้มที่อาจเป็นไปได้ในการวิจัย เพื่อการวิเคราะห์ผลกระทบของการตัดสินใจ และ Data Analytics เป็นศาสตร์แห่งการวิเคราะห์ข้อมูลเพื่อแปลงข้อมูลให้เป็นประโยชน์ เช่น วิเคราะห์ หาแนวโน้ม สร้างแบบจำลอง ใช้สนับสนุนการตัดสินใจ เป็นต้น

1.1 ระดับของ Data Analytics

1) Query Processing การประมวลผลคำสั่งสืบค้นข้อมูล เป็นขั้นตอนการดึงข้อมูลจากฐานข้อมูลเพื่อเตรียมข้อมูลสำหรับการวิเคราะห์

2) Summary Statistics (Descriptive) การสรุปข้อมูลเชิงสถิติ เป็นการสรุปข้อมูลให้เข้าใจง่าย โดยใช้ค่าสถิติพื้นฐาน

3) Exploration (Descriptive) การสำรวจข้อมูล เป็นการสำรวจข้อมูลเพื่อค้นหารูปแบบ แนวโน้ม หรือความสัมพันธ์ เพื่อคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต โดยใช้ข้อมูลในอดีตและแบบจำลอง

4) Modeling (Descriptive, Predictive and Prescriptive) การสร้างแบบจำลอง เป็นการสร้างโมเดลเพื่อคาดการณ์ หรืออธิบายข้อมูล โดยใช้สถิติหรือ Machine Learning เพื่อ “แนะนำว่าควรทำอะไร” เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด

1.2 Analytics Team

คือ ทีมที่รับผิดชอบในการนำข้อมูลมาวิเคราะห์ เพื่อสร้าง Insight และสนับสนุนการตัดสินใจ โดยทั่วไปประกอบด้วยบทบาทหลักดังนี้

- Business analyst/Expert นักวิเคราะห์ธุรกิจ มีหน้าที่ในการเชื่อมโยงข้อมูลกับการตัดสินใจทางธุรกิจ
- Machine-learning expert/Data scientist/Data Analyst คือ ผู้เชี่ยวชาญด้าน Machine learning/นักวิทยาศาสตร์ข้อมูล/นักวิเคราะห์ข้อมูล มีหน้าที่วิเคราะห์ข้อมูลและสร้างรายงาน สร้างโมเดลคาดการณ์ และใช้ Machine Learning
- Data Engineer หรือ วิศวกรข้อมูล สร้างและดูแลโครงสร้างพื้นฐานด้านข้อมูล

1.3 Data Lake

Data Lake คือ ระบบจัดเก็บข้อมูลขนาดใหญ่ที่สามารถเก็บข้อมูลได้ทุกประเภท ทั้งแบบมีโครงสร้างและไม่มีโครงสร้าง โดยเก็บในรูปแบบข้อมูลดิบ เพื่อรองรับการวิเคราะห์ข้อมูลขั้นสูง เช่น Big Data Analytics และ Machine Learning

ลักษณะสำคัญของ Data Lake

- เก็บข้อมูลได้ทุกประเภท ทั้ง Structured เช่น ตารางฐานข้อมูล Semi-structured เช่น JSON, XML Unstructured เช่น รูปภาพ วิดีโอ ไฟล์ข้อมูลต่างๆ
- เก็บข้อมูลในรูปแบบดิบ (Raw Data)
- ยังไม่ต้องผ่านการแปลง (Transform)
- สามารถนำไปใช้วิเคราะห์ได้หลายรูปแบบในอนาคต
- รองรับข้อมูลขนาดใหญ่มาก (Big Data)
- มีความยืดหยุ่นสูง

1.4 Big Data Analytics

Big Data Analytics คือ การวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data) เพื่อค้นหารูปแบบ แนวโน้ม หรือความรู้ใหม่ ที่ช่วยในการตัดสินใจ

Big Data Analytics ทำอะไรได้บ้าง

- หารูปแบบ (Pattern) เช่น พื้นที่ใดมีแนวโน้มเกิดภัยแล้ง
- คาดการณ์ (Prediction) เช่น คาดการณ์ผลผลิตทางการเกษตร
- วิเคราะห์แนวโน้ม (Trend) เช่น การเปลี่ยนแปลงการใช้ที่ดิน
- ช่วยตัดสินใจ (Decision Making) เช่น ควรวางแผนใช้น้ำอย่างไร

2. Data Understanding and Processing

Data Understanding & Processing คือขั้นตอนสำคัญก่อนการวิเคราะห์ข้อมูล เพื่อให้เข้าใจโครงสร้าง คุณภาพ และเตรียมข้อมูลให้พร้อมใช้งาน

- Data Source หรือ แหล่งที่มาของข้อมูล หมายถึงแหล่งที่ข้อมูลถูกสร้างหรือจัดเก็บ เช่น ฐานข้อมูล (Database)
- Types of Data Elements หรือ ประเภทของข้อมูล คือประเภทของตัวแปร (Variable) ในชุดข้อมูล แบ่งหลัก ๆ เป็น 4 ประเภท ได้แก่ Numerical (ตัวเลข) Categorical (หมวดหมู่) Ordinal (ข้อมูลมีลำดับ) และ Boolean (ข้อมูลจริง/เท็จ)
- Missing Values คือ ข้อมูลที่หายไป หรือ ข้อมูลบางช่องไม่มีค่า
- Outlier ข้อมูลสุโต่ง หรือข้อมูลผิดปกติ ข้อมูลที่แตกต่างจากข้อมูลส่วนใหญ่ เช่น การวิเคราะห์เงินเดือนของพนักงาน
- Standardization คือ การปรับมาตรฐานข้อมูล เป็นการปรับข้อมูลให้อยู่ในรูปแบบมาตรฐานเดียวกัน เพราะข้อมูลแต่ละตัวมีหน่วยต่างกัน

3. Data Exploration and Visualization

Data Exploration and Visualization คือ มุมมองการมองภาพในหลายมิติ การทำความเข้าใจข้อมูลก่อนสร้างโมเดล ดูว่า ข้อมูลกระจายแบบไหน มี Outlier ไหม ตัวแปรสัมพันธ์กันไหม โดยการนำเสนอในรูปแบบต่างๆ เช่น การใช้ Scatter plot เพื่อดูความสัมพันธ์ของข้อมูลใน 2 มิติ การใช้ Pie chart ซึ่งเป็นการมองภาพในลักษณะ 1 มิติ การใช้ Bar plot ซึ่งเป็นการนำเสนอข้อมูลในรูป 1 มิติ ในรูปแบบกราฟแท่ง การวิเคราะห์ความหนาแน่น เพื่อดูความหนาแน่นของข้อมูล เช่น Histogram plot, Density plot เป็นต้น การใช้ Percentiles and Box plots, Radar/Spider plot, Map

Dimension Reduction เป็นการลดมิติ โดยใช้วิธี PCA เป็นการลดแกนของข้อมูล หรือเปลี่ยนแกนในการแสดงข้อมูล

4. Machine Learning

Machine Learning คือ การทำให้คอมพิวเตอร์ เรียนรู้จากข้อมูล ประกอบด้วย

- Supervised Machine Learning เป็นการสอนให้เครื่องคอมพิวเตอร์เรียนรู้และสามารถทำนายข้อมูลใหม่ได้จากข้อมูลที่มี คำตอบอยู่แล้ว (Label) เช่น การแยกภาพหมาและแมว Supervised Machine Learning ใช้ในการจำแนกประเภท (Classification) และการการพยากรณ์ค่า (Regression)
- Unsupervised Machine Learning เป็นการให้เครื่องคอมพิวเตอร์เรียนรู้จากข้อมูลที่ไม่มีคำตอบ (Label) โดยคอมพิวเตอร์จะค้นหารูปแบบหรือจัดกลุ่มข้อมูลที่เหมือนกันอยู่ในกลุ่มเดียวกัน Unsupervised Machine Learning ใช้ในการแบ่งกลุ่ม (Clustering)

5. Descriptive Analytics

Descriptive Analytics การวิเคราะห์เชิงพรรณนา คือ การวิเคราะห์ข้อมูลเพื่ออธิบายว่า เกิดอะไรขึ้นแล้วในอดีต โดยการสรุป จัดกลุ่ม และนำเสนอข้อมูลให้อยู่ในรูปแบบที่เข้าใจง่าย เช่น รายงาน ตาราง หรือ Dashboard

- Similarity Measurement การวัดความเหมือน และการวัดความต่าง เป็นวิธีวัดว่า ข้อมูล 2 ตัว มีความคล้ายหรือแตกต่างกันมากแค่ไหน การวัด Similarity Measurement เช่น Manhattan Distance (ระยะทางแบบตาราง) Euclidean Distance (ระยะห่างแบบเส้นตรง) เป็นวิธีที่นิยมมากที่สุด Jaccard Similarity และ Cosine Similarity (วัดมุมระหว่างข้อมูล) และ Edit Distance
- Segmentation Clustering (การจัดกลุ่มข้อมูล) เป็นการจัดกลุ่มข้อมูลที่ คล้ายกันให้อยู่ในกลุ่มเดียวกัน โดยจะต้องตอบคำถาม 4 คำถามเกี่ยวกับ จะแบ่งข้อมุลออกเป็นกลุ่มได้อย่างไร จะเลือกจำนวนกลุ่มข้อมูลเท่าไหร่ จะอธิบายกลุ่มข้อมูลได้อย่างไร และจะบอกได้อย่างไรว่าเมื่อมีข้อมูลใหม่เข้ามาข้อมูลนั้นดีอย่างไร วิธีการจัดกลุ่ม เช่น K-means Clustering (นิยมมากที่สุด) Density-based model เป็นต้น
- Anomaly Detection Strategy การตรวจจับข้อมูลผิดปกติ เป็นการหาข้อมูลที่ แตกต่างจากข้อมูลส่วนใหญ่ วิธี Detect Anomaly เช่น Univariate anomaly detection, Multivariate anomaly detection

6. Predictive Analytics-Regression

6.1 Predictive Analytics-Regression เป็นการใช้ทำนาย ค่าตัวเลขต่อเนื่อง หรือ Continuous value ด้วย Regression เช่น ขนาดของห้อง ทำนาย ราคาบ้าน การทำนายดัชนีมวลกาย (BIM) เพื่อทำนายปริมาณไขมัน การทำนายการขึ้นลงของหุ้นโดยใช้หุ้นตัวอื่นๆ

- Linear Regression เป็นสมการเชิงเส้น ซึ่งในกรณีที่มี 1 มิติ (Univariate linear regression) จะใช้สมาการ $h_w(x) = w_0 + w_1x$ และในกรณีที่มีมากกว่า 1 มิติ จะใช้สมาการ $x_0 = 1$ โดย $h_w(x) = w_0x_0 + w_1x_{1+} w_2x_{2+\ldots+}w_mx_m$
- Polynomial Regression ใช้เมื่อข้อมูลเป็นโค้ง เช่น ยอดขายโตเร็วช่วงแรก แล้วชะลอ ระวัง Overfitting ถ้า degree สูงเกินไป

ตัววัดผล Regression

- R^2 (ยิ่งใกล้ 1 ยิ่งดี)
- MSE
- RMSE

6.2 Outlier คือ ค่าข้อมูลที่แตกต่างจากข้อมูลส่วนใหญ่อย่างชัดเจน หรือเป็นค่าที่ “ผิดปกติ” เมื่อเทียบกับข้อมูลอื่นในชุดเดียวกัน

สาเหตุของ Outlier

- ความผิดพลาดในการบันทึกข้อมูล เช่น กรอก 6.5 เป็น 65
- เครื่องมือวัดผิดพลาด เช่น Sensor เสีย
- ความผิดปกติจริง เช่น พื้นที่มีค่าผิดปกติจริง

- ความผิดพลาดในการประมวลผลข้อมูล

6.3 ผลกระทบของ Outlier

- ทำให้ค่าเฉลี่ยผิดเพี้ยน
- ทำให้ผลวิเคราะห์ผิด
- ทำให้ Machine Learning ผิดพลาด

7. Predictive analytics Classification problem

Predictive analytics Classification problem คือ การประมินควาถูกต้องของโมเดลในการทำนายข้อมูล เช่น การทำนายว่าเป็นโรค / ไม่เป็นโรค เป็นต้น

โมเดลยอดนิยม

- Decision Tree คือ โมเดลที่ใช้โครงสร้างแบบ “ต้นไม้” เพื่อจำแนกข้อมูล โดยใช้เงื่อนไขในการตัดสินใจทีละขั้น
- Logistic Regression เป็นโมเดลที่ใช้ “สมการทางคณิตศาสตร์” เพื่อทำนายความน่าจะเป็นของแต่ละประเภท แม้ชื่อจะมี Regression แต่ใช้สำหรับ Classification
- K-Nearest Neighbor (KNN) เป็นโมเดลที่จำแนกข้อมูล โดยดูจาก “เพื่อนบ้านที่ใกล้ที่สุด” โดย K = จำนวนเพื่อนบ้าน
- Neural Network โครงข่ายประสาทเทียม คือ โมเดล Machine Learning ที่เลียนแบบการทำงานของสมองมนุษย์ โดยใช้ “โหนด (Neuron)” หลาย ๆ ตัวเชื่อมต่อกัน เพื่อเรียนรู้รูปแบบของข้อมูล และใช้ในการจำแนกหรือทำนายผล

8. Recommendation System

Recommendation System หรือ ระบบแนะนำ คือ ระบบที่ใช้ข้อมูลและอัลกอริทึมในการวิเคราะห์พฤติกรรมหรือความสนใจของผู้ใช้ เพื่อแนะนำสิ่งที่เหมาะสมหรือที่ผู้ใช้น่าจะสนใจ

1) Association Rule หากฎความสัมพันธ์ของสิ่งที่มักเกิดร่วมกัน” ในชุดข้อมูล โดยนิยมใช้มากในงานวิเคราะห์พฤติกรรม เช่น การซื้อสินค้า

2) Collaborative Filtering คือ เทคนิคที่ใช้ “พฤติกรรมของผู้ใช้อื่น” มาช่วยแนะนำสิ่งที่เหมาะกับผู้ใช้

- User-Based Collaborative Filtering คือการแนะนำโดยดูจาก “ผู้ใช้คนอื่นที่มีพฤติกรรมคล้ายกับคุณ
- Model-Based Collaborative Filtering คือการใช้ Machine Learning Model สร้างโมเดลเพื่อทำนายว่า user จะชอบอะไร

การประยุกต์ใช้ในการทำงาน

Data Analytics สามารถนำมาประยุกต์ใช้ในการทำงานเพื่อรวบรวม จัดการและวิเคราะห์ข้อมูลให้เกิดประโยชน์ต่อการตัดสินใจ เช่น การวิเคราะห์ข้อมูลดิน ข้อมูลการใช้ที่ดิน และข้อมูลภูมิสารสนเทศ เพื่อให้เข้าใจสภาพพื้นที่และรูปแบบการเปลี่ยนแปลง นอกจากนี้ยังสามารถใช้เทคนิค Machine Learning เพื่อจำแนกประเภท ทำนายแนวโน้ม และแนะนำการใช้ประโยชน์พื้นที่ให้เหมาะสม ผลลัพธ์ที่ได้สามารถนำไปใช้ในการวางแผน การบริหารจัดการทรัพยากร และสนับสนุนการตัดสินใจเชิงนโยบายได้อย่างมีประสิทธิภาพ และช่วยเพิ่มความถูกต้อง รวดเร็ว และความน่าเชื่อถือของการวิเคราะห์ข้อมูลในองค์กร

แหล่งที่มา

หลักสูตร : Data Analytics
ด้านการพัฒนา : ☑ **ทักษะด้านดิจิทัล**
บรรยายโดย : ชื่อ-สกุล ดร.ไพรสันต์ ผดุงเวียง ตำแหน่ง อาจารย์
หน่วยงานผู้รับผิดชอบ : OCSC Learning Space สำนักงานคณะกรรมการข้าราชการพลเรือน
วิธีการพัฒนาตนเอง : e-learning
วันที่ได้รับการฝึกอบรม : 20 ก.พ. 2569 **สถานที่ :** 52/1 บ้านสมบูรณ์อพาร์ทเม้น ถนนพหลโยธิน 53 แยก 4 แขวงอนุสาวรีย์ เขตบางเขน 10220