

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหา

นักวิทยาศาสตร์ข้อมูล (Data Scientist) เป็นหนึ่งในอาชีพที่เติบโตอย่างรวดเร็วในช่วงทศวรรษที่ผ่านมา เนื่องจากการเพิ่มขึ้นของข้อมูลขนาดใหญ่ (Big Data) และความต้องการในการวิเคราะห์ข้อมูลเพื่อสร้างมูลค่าและตัดสินใจทางธุรกิจอาชีพ Data Scientist จุดเริ่มต้นและวิวัฒนาการ ยุค ค.ศ.1960-1970 การใช้คอมพิวเตอร์ในการวิเคราะห์ข้อมูลเริ่มขึ้นในช่วงนี้ ด้วยการใช้เทคนิคเชิงสถิติและการวิเคราะห์ข้อมูลเพื่อช่วยในการตัดสินใจทางธุรกิจยุค ค.ศ.1980-1990 การพัฒนาเทคโนโลยีฐานข้อมูลและการประมวลผลข้อมูลเริ่มมีบทบาทสำคัญการเกิดขึ้นของฐานข้อมูลเชิงสัมพันธ์ (Relational Databases) และการพัฒนาภาษาสำหรับการสืบค้นข้อมูล เช่น SQL ทำให้การจัดการและการวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้นยุค ค.ศ.2000 การเกิดขึ้นของอินเทอร์เน็ตและเทคโนโลยีเว็บทำให้ข้อมูลขนาดใหญ่เพิ่มขึ้นอย่างรวดเร็ว การพัฒนาเทคโนโลยีคลาวด์และการประมวลผลแบบกระจาย (Distributed Computing) ทำให้การจัดการข้อมูลขนาดใหญ่เป็นไปได้การเกิดขึ้นของ Big Data และการขยายตัวของอาชีพ Data Scientist ปี ค.ศ.2008-2010 การขยายตัวของข้อมูลที่เกิดจากการใช้งานอินเทอร์เน็ตและอุปกรณ์ดิจิทัล

อาชีพ Data Scientist ยังถือได้ว่าอยู่ในระดับที่ขาดแคลนในขณะที่ความต้องการคนที่มีความรู้ความสามารถในด้านนี้มีสูงมาก และยังมีแนวโน้มที่จะมีสูงมากขึ้นไปเรื่อยๆ เพราะเทรนด์การนำข้อมูลที่ใช้ประโยชน์ในธุรกิจ นับวันก็จะยิ่งทวีความสำคัญมากขึ้น เรียกได้ว่า Data Scientist เป็นอีกหนึ่งอาชีพที่มีอนาคต เพราะทั้งเป็นที่ต้องการตัวและมีค่าตอบแทนในระดับสูงเมื่อเปรียบเทียบกับข้อมูลจากต่างประเทศ ซึ่งทำให้การวิเคราะห์ข้อมูลนี้สามารถช่วยให้เข้าใจความแตกต่างและแนวโน้มที่เกิดขึ้นในตลาดงาน การจัดการบุคลากรบริษัทต่างๆ ต้องการข้อมูลที่ชัดเจนเกี่ยวกับเงินเดือนเพื่อวางแผนการจ่ายเงินเดือนและจัดการทรัพยากรบุคคลได้อย่างมีประสิทธิภาพ การเข้าใจความแตกต่างของเงินเดือนตามประสบการณ์และทักษะจะช่วยในการกำหนดระดับเงินเดือนที่เหมาะสม

ดังนั้นผู้จัดทำได้ใช้เทคนิคการวิเคราะห์ความสัมพันธ์ (Correlation Analysis) เพื่อคัดเลือกแอททริบิวต์หรือพารามิเตอร์ที่สำคัญในการสร้างโมเดล จากนั้นจึงนำข้อมูลผ่านการคัดเลือกมาใช้ในการจำแนกประเภท (Classification) โดยใช้เทคนิค Decision Tree ซึ่งเป็นวิธีการที่ได้รับความนิยมและมีความสามารถในการอธิบายปัจจัยที่มีผลต่อการแบ่งกลุ่มข้อมูลได้อย่างชัดเจน เพื่อทำนายและทำความเข้าใจรูปแบบของเงินเดือน Data Scientist ในแต่ละช่วงชั้น (เช่น ระดับต่ำ ปานกลาง และสูง) อย่างเป็นระบบ ผลลัพธ์ที่ได้คาดว่าจะสะท้อนให้เห็นถึงความแตกต่างเชิงโครงสร้างของตลาดแรงงานสาย Data Science และช่วยให้สามารถนำข้อมูลที่ได้ไปประยุกต์ใช้ทั้งในการวางแผนกลยุทธ์การบริหารทรัพยากรบุคคลและการพัฒนาทักษะอาชีพของบุคคลได้อย่างเหมาะสม

1.2 วัตถุประสงค์

1.2.1 เพื่อวิเคราะห์และจำแนกเงินเดือนของ Data Scientist โดยพิจารณาจากปัจจัยด้านประสบการณ์การทำงาน ตำแหน่งงาน ทักษะ ความเชี่ยวชาญเฉพาะด้าน และคุณลักษณะอื่น ๆ ที่เกี่ยวข้อง

1.2.2 เพื่ออธิบายลักษณะเฉพาะของแต่ละกลุ่มที่ได้จากการจำแนก เช่น ระดับเงินเดือน ความสัมพันธ์กับประสบการณ์ และปัจจัยสำคัญที่ส่งผลต่อความแตกต่าง เพื่อให้เข้าใจความหลากหลายของตลาดแรงงานสาย Data Scientist ได้ชัดเจนยิ่งขึ้น

1.3 ประโยชน์ที่ได้รับ

1.3.1 ได้ข้อมูลเชิงลึกที่ผ่านการวิเคราะห์และจัดกลุ่มเงินเดือนของ Data Scientist โดยอ้างอิงจากปัจจัยด้านเงินเดือน ประสบการณ์ ทักษะ ความเชี่ยวชาญเฉพาะด้าน และคุณลักษณะอื่น ๆ

1.3.2 ได้คำอธิบายลักษณะเฉพาะของแต่ละกลุ่มที่เกิดจากการจัดกลุ่ม เช่น ระดับเงินเดือน แนวโน้มตามประสบการณ์ และทักษะที่เกี่ยวข้อง ซึ่งจะช่วยให้เห็นภาพรวมและเข้าใจความหลากหลายของตลาดแรงงานสาย Data Scientist ได้ชัดเจนยิ่งขึ้น

1.4 ขอบเขต

การวิเคราะห์ข้อมูลเพื่อศึกษาความหลากหลายของสายงาน Data Scientist โดยใช้ เทคนิคการวิเคราะห์การจำแนกประเภท (Classification Analysis) เพื่อจำแนกกลุ่มตามลักษณะต่าง ๆ เช่น เงินเดือน ประสบการณ์ ทักษะ และความเชี่ยวชาญเฉพาะด้าน โดยการสร้างโมเดลทำนาย

(Predictive Model) ที่สามารถคาดการณ์หรือจัดประเภทบุคคลตามปัจจัยที่เกี่ยวข้อง จากนั้นนำผลลัพธ์ที่ได้ไปพัฒนาเป็น ระบบวิเคราะห์และแสดงผลบนเว็บไซต์ เพื่อให้ผู้ใช้งานสามารถเข้าถึงข้อมูลได้สะดวก

ขั้นตอนการดำเนินงานทั้งหมดอยู่บนกรอบแนวคิด CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งประกอบด้วย การทำความเข้าใจธุรกิจ การทำความเข้าใจข้อมูล การเตรียมข้อมูล การสร้างโมเดลการจำแนกประเภท การประเมินผล และการนำไปใช้ เพื่อให้กระบวนการวิเคราะห์มีความเป็นระบบและสามารถอธิบายผลลัพธ์ได้ชัดเจน

1.4.1 ขอบเขตด้านเนื้อหา

1. 4. 1. 1 ศึกษา ข้อมูล เงินเดือน ของ Data Scientists (Business Understanding) ขั้นตอนแรกคือการศึกษาและทำความเข้าใจปัจจัยที่มีผลต่อเงินเดือนของ Data Scientist เพื่อกำหนดวัตถุประสงค์และขอบเขตของการวิเคราะห์ ข้อมูลจากการศึกษาพบว่าปัจจัยสำคัญ ได้แก่ 1) ประสบการณ์ (Experience) – Data Scientist ที่มีประสบการณ์สูงมักได้รับเงินเดือนมากกว่าผู้ที่เพิ่งเริ่มงานหรือ มีประสบการณ์น้อย 2) การศึกษาและคุณวุฒิ (Education & Qualification) – การศึกษาที่สูงขึ้นหรือปริญญาในสาขาที่เกี่ยวข้อง เช่น ปริญญาตรีหรือปริญญาโท มีผลต่อระดับเงินเดือน 3) สถานที่ทำงาน (Work Location) – เงินเดือนแตกต่างกันตามภูมิภาค เช่น เมืองใหญ่หรือพื้นที่ที่มีค่าครองชีพสูงมักมีเงินเดือนสูงกว่า 4) ประเภทของบริษัท (Company Type) – บริษัทขนาดใหญ่หรือบริษัทที่มีเทคโนโลยีทันสมัยมักจ่ายเงินเดือนสูงกว่าบริษัทขนาดเล็กหรือเพิ่งเริ่มต้น 5) ทักษะเฉพาะ (Specialized Skills) ความเชี่ยวชาญในเทคโนโลยีหรือเครื่องมือเฉพาะ เช่น การวิเคราะห์ข้อมูลทางการเงินหรือการใช้เครื่องมือวิเคราะห์ขั้นสูง สามารถเพิ่มมูลค่าเงินเดือน

1.4.1.2 รวบรวมและทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนการจัดเก็บ รวบรวมข้อมูล ตรวจสอบความถูกต้องของข้อมูลที่มาจาก เว็บไซต์ Kaggle.com โดยเลือกที่จะใช้ข้อมูลบางส่วนในการวิเคราะห์ให้สอดคล้องกับวัตถุประสงค์ที่กำหนดไว้ ชุดข้อมูล “data_science_salaries” ขั้นตอนนี้ประกอบด้วย การรวบรวมข้อมูลและตรวจสอบความถูกต้อง การเลือกเฉพาะส่วนข้อมูลที่สอดคล้องกับวัตถุประสงค์ เช่น เงินเดือน, ประสบการณ์, ทักษะ และตำแหน่งงาน การตรวจสอบความสมบูรณ์ของข้อมูล เช่น การลบค่าที่ขาดหายหรือผิดปกติ การทำความเข้าใจโครงสร้างข้อมูลเพื่อกำหนดแนวทางการวิเคราะห์

1.4.1.3 คัดเลือกข้อมูล กลั่นกรองข้อมูลและแปรงรูปแบบข้อมูล (Data Preparation) การคัดเลือกข้อมูล ขั้นตอนนี้มุ่งเน้นการทำความสะอาดและเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์คัดเลือกแอตทริบิวต์ที่สำคัญ เลือกเฉพาะคอลัมน์ที่เกี่ยวข้อง เช่น เงินเดือน, ประสบการณ์, ทักษะเฉพาะด้านทำความสะอาดข้อมูล (Data Cleaning) ตรวจสอบและแก้ไขข้อมูลผิดพลาด ลบข้อมูลซ้ำหรือค่าผิดปกติ แปรงรูปแบบข้อมูล (Data Transformation) แปลงตัวแปรเชิงหมวดหมู่เป็นตัวเลข (Categorical to Numerical) เตรียมข้อมูลสำหรับโมเดล (Data Formatting) สร้างชุดข้อมูลที่พร้อมสำหรับการทำ Decision Tree

1.4.1.4 สร้างแบบจำลอง (Modeling) ขั้นตอนนี้ได้เลือกใช้เทคนิค Decision Tree (Classification) เพื่อจำแนกระดับเงินเดือนของ Data Scientist ตามปัจจัยต่าง ๆ เช่น ประสบการณ์ ระดับการศึกษา สถานที่ทำงาน และทักษะเฉพาะด้าน ก่อนสร้างโมเดล ได้ทำการวิเคราะห์ความสัมพันธ์ของตัวแปร (Correlation Analysis) เพื่อคัดเลือกตัวแปรสำคัญสำหรับการสร้างโมเดล และเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ การสร้างโมเดล ดำเนินการผ่าน โปรแกรม RapidMiner โดยใช้ Operator หลัก เช่น Decision Tree และตั้งค่า Parameter ที่สำคัญ เช่น Maximum Depth, Criterion สำหรับการแบ่งกิ่ง (Split) ของต้นไม้ การกำหนด Target Attribute คือระดับเงินเดือนที่ถูกจัดกลุ่มเป็นช่วง เช่น Low, Medium, High ส่วน Attribute อื่น ๆ ที่เกี่ยวข้องถูกตั้งเป็น Feature Input เพื่อให้โมเดลสามารถสร้างกฎการตัดสินใจได้อย่างมีประสิทธิภาพ

1.4.1.5 การประเมินผล (Evaluation) หลังจากสร้างโมเดล Decision Tree ขั้นตอนการประเมินผลเป็นการตรวจสอบความถูกต้องและความน่าเชื่อถือของโมเดล โดยใช้ตัวชี้วัดมาตรฐานในการจำแนกประเภท ได้แก่ Accuracy ความแม่นยำของโมเดลในการทำนายระดับเงินเดือน Precision, Recall และ F1-Score เพื่อวัดความถูกต้องและความสมดุลของโมเดลในแต่ละคลาส Confusion Matrix เพื่อตรวจสอบว่ามีการจำแนกคลาสผิดพลาดมากน้อยเพียงใด ใน RapidMiner สามารถใช้ Cross Validation เพื่อตรวจสอบประสิทธิภาพของโมเดลอย่างเป็นระบบ โดยแบ่งข้อมูลออกเป็นชุดฝึก (Training) และชุดทดสอบ (Testing) และรันโมเดลซ้ำหลายรอบเพื่อให้มั่นใจว่าผลลัพธ์ไม่ขึ้นอยู่กับการสุ่มของข้อมูล หากพบว่าโมเดลมีความแม่นยำต่ำหรือคลาสบางกลุ่มจำแนกไม่ชัดเจน จะทำการปรับปรุงโมเดล เช่น การเลือกตัวแปรใหม่ การปรับพารามิเตอร์ของ Decision Tree หรือการใช้เทคนิคการปรับสมดุลของข้อมูล (Data Balancing) จนได้โมเดลที่เหมาะสมและพร้อมสำหรับขั้นตอนการเผยแพร่

1.4.1.6 การเผยแพร่ผลวิเคราะห์ (Deployment) เผยแพร่ผลวิเคราะห์หรือการนำผลลัพธ์ที่ได้ไปใช้งานเพื่อให้สามารถนำไปใช้ประโยชน์ได้อย่างมีประสิทธิภาพ หลังจากประเมินและปรับปรุงโมเดลจนได้ผลลัพธ์ที่เหมาะสม ผลลัพธ์ทั้งหมดถูกนำไปเผยแพร่ผ่าน เว็บไซต์เพื่อให้ผู้ใช้งานทั่วไปสามารถเข้าถึงข้อมูลและเข้าใจแนวโน้มตลาดแรงงานสาย Data Scientist ได้สะดวกเว็บไซต์จะแสดงข้อมูลการจัดกลุ่ม Data Scientist พร้อมกราฟและแผนภูมิที่สรุปลักษณะเฉพาะของแต่ละคลัสเตอร์ เช่น ระดับเงินเดือนเฉลี่ย ประสบการณ์เฉลี่ย และทักษะโดดเด่นของสมาชิกในกลุ่ม การนำเสนอในรูปแบบกราฟิกและตารางช่วยให้ผู้ใช้งานสามารถเปรียบเทียบและทำความเข้าใจความหลากหลายของตลาดแรงงานได้อย่างชัดเจน

1.4.2 ขอบเขตผู้ใช้งานทั่วไปบนเว็บไซต์

8.2.1 สามารถดูข้อมูลการวิเคราะห์เงินเดือนของ Data Scientist ผ่านเว็บไซต์

8.2.2 สามารถเข้าถึงข้อมูลการจัดจำแนกกลุ่มของ Data Scientist ตามลักษณะต่าง ๆ เช่น ประสบการณ์ ทักษะ และความเชี่ยวชาญเฉพาะด้าน

8.2.3 สามารถดูภาพรวมของความหลากหลายในตลาดแรงงานสาย Data Scientist ผ่านกราฟหรือแผนภูมิที่แสดงผลบนเว็บไซต์

8.2.4 สามารถเปรียบเทียบลักษณะของแต่ละกลุ่ม เพื่อทำความเข้าใจแนวโน้มและคุณลักษณะเฉพาะของ Data Scientist แต่ละกลุ่ม

1.5 เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล

1.5.1 Hardware

– Notebook: Acer A315-22 AMD A9-9420e RADEON R5 RAM 4 GB

1.5.2 Software

1.5.2.1 ภาษา Python นำมาใช้ในการพัฒนาเว็บไซต์เพื่อแสดงผลการวิเคราะห์ข้อมูลจากโมเดล RapidMiner

1.5.2.2 โปรแกรม FileZillaใช้สำหรับเชื่อมต่อกับ FTP Server เพื่ออัปโหลดไฟล์เว็บไซต์และฐานข้อมูล ทำให้เว็บไซต์สามารถออนไลน์และเข้าถึงได้จากผู้ใช้งานทุกที่ โปรแกรมนี้ช่วยให้การจัดการไฟล์และการปรับปรุงเว็บไซต์เป็นไปอย่างรวดเร็วและปลอดภัย

1.5.2.3 Microsoft Office Excel 2019ใช้ในการทำความสะอาดข้อมูล จัดเตรียมข้อมูลเบื้องต้น และจัดการตารางข้อมูลสำหรับการวิเคราะห์ เช่น การลบข้อมูลซ้ำ การจัดรูปแบบ

ตัวเลข และตรวจสอบค่าผิดปกติ การใช้ Excel ทำให้กระบวนการตรวจสอบและเตรียมข้อมูลเป็นไปอย่างรวดเร็วและสะดวก

1.5.2.4 RqpidMiner ใช้เพื่อสร้างและรันโมเดลวิเคราะห์ข้อมูลแบบกราฟิกโดยไม่จำเป็นต้องเขียนโค้ดเอง ทำให้ผู้ใช้สามารถออกแบบกระบวนการวิเคราะห์ (Process) ตั้งแต่การเตรียมข้อมูล การสร้างโมเดล การประเมินผล จนถึงการนำผลลัพธ์ไปใช้งานได้ภายในโปรแกรมเดียว

1.5.2..5 Visual Studio Code (VS Code) VS Code ใช้ในการเขียนและจัดการโค้ด Python สำหรับโปรเจกต์ขนาดใหญ่ ช่วยให้การจัดการไฟล์หลายไฟล์ ตรวจสอบโค้ด และใช้ส่วนขยาย (Extension) ต่าง ๆ เป็นไปอย่างมีประสิทธิภาพ

1.5.2.6 Microsoft Word ใช้ในการจัดทำเอกสารรายงานผลการวิเคราะห์ ทำให้เอกสารมีรูปแบบที่เป็นระเบียบ อ่านง่าย และพร้อมนำเสนอ

1.5.2.7 XAMPP เป็นโปรแกรมที่ใช้สำหรับสร้างและจัดการเซิร์ฟเวอร์เว็บแบบท้องถิ่น (Local Server) เพื่อใช้ในการทดสอบการทำงานของเว็บไซต์และฐานข้อมูลบนเครื่องคอมพิวเตอร์ของผู้พัฒนา ก่อนนำไปเผยแพร่และใช้งานบนเซิร์ฟเวอร์จริง

1.5.2.8 ภาษา PHP ใช้ในการพัฒนาเว็บไซต์ฝั่งเซิร์ฟเวอร์และเชื่อมต่อฐานข้อมูล ช่วยให้เว็บไซต์สามารถประมวลผลข้อมูลและแสดงผลได้อย่างยืดหยุ่น

1.5.2.9 ภาษา HTML ถูกใช้สร้างโครงสร้างหน้าเว็บและแสดงข้อมูลที่วิเคราะห์ ทำให้ผู้ใช้งานสามารถเข้าถึงข้อมูลผ่านเว็บเบราว์เซอร์ได้อย่างง่ายดาย

1.6 สถานที่ใช้ในการดำเนินการศึกษาและรวบรวมข้อมูล

สถานที่ใช้ในการดำเนินการศึกษา มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ที่อยู่ 128 ถ. ห้วยแก้ว ตำบล ช้างเผือกอำเภอ เมืองเชียงใหม่ จังหวัดเชียงใหม่

1.7 ระยะเวลาในการดำเนินการ

ตารางที่ 1.1 ระยะเวลาในการดำเนินการ

แผนการดำเนินการ	2567					2568				
	ก.ค.	ส.ค.	ก.ย.	ต.ค.	พ.ย.	มิ.ย.	ก.ค.	ส.ค.	ก.ย.	ต.ค.
1. ศึกษาและกำหนดความต้องการ	←→									
2. การเตรียมข้อมูล			←→							
3. ตรวจสอบความถูกต้อง				←→						
4. วิเคราะห์ ออกแบบฐานข้อมูล							←→			
5. การสร้างโมเดล								←→		
6. การประเมินประสิทธิภาพ								←→		
7. นำข้อมูลไปแสดงผลทางเว็บไซต์									←→	
8. จัดทำเอกสารประกอบโครงการ	←→									

ตารางที่ 1.1 ระยะเวลาในการดำเนินการ

1.8 นิยามศัพท์เฉพาะ

1.9 บทสรุป

ผู้จัดทำเห็นความสำคัญของการประยุกต์ใช้เทคโนโลยีในการวิเคราะห์ข้อมูลเพื่อทำความเข้าใจความหลากหลายของ Data Scientist โครงการนี้ใช้กรอบแนวคิด CRISP-DM ในการวิเคราะห์ ตั้งแต่การทำความเข้าใจธุรกิจ ข้อมูล การเตรียมข้อมูล การสร้างโมเดล และการประเมินผล เลือกใช้ Decision Tree (Classification) ผ่าน RapidMiner เพื่อจำแนกระดับเงินเดือนตามประสบการณ์ การศึกษา สถานที่ทำงาน และทักษะเฉพาะด้านโมเดลช่วยระบุปัจจัยสำคัญที่ส่งผลต่อระดับเงินเดือน และอธิบายลักษณะเฉพาะของแต่ละกลุ่มอย่างชัดเจนข้อมูลที่ได้นำมาพัฒนาเป็น เว็บแอปพลิเคชัน ด้วย Python และ HTML ผู้ใช้งานสามารถเข้าถึงข้อมูล การจำแนกประเภท และกราฟสรุปผลได้สะดวกและทันสมัย