การวิเคราะห์ทักษะทางเทคนิคและจรณทักษะของสายงานอาชีพด้วยวิธีการทำเหมืองข้อความในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ANALYSIS OF TECHNICAL SKILLS AND SOFT SKILLS OF CAREER PATH USING TEXT MINING IN MACHINE LEARNING AND DEEP LEARNING METHODS

ปรัชญ์ธิดา พรประเสริฐ

PRACHTIDA PORNPRASERT

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์

ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2566

KMITL-2023-SC-M-017-019

ANALYSIS OF TECHNICAL SKILLS AND SOFT SKILLS OF CAREER PATH USING TEXT MINING IN MACHINE LEARNING AND DEEP LEARNING METHODS

PRACHTIDA PORNPRASERT

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE

IN DATA SCIENCE AND ANALYTICS

KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2023

KMITL-2023-SC-M-017-019

หัวข้อการค้นความอิสระ การวิเคราะห์ทักษะทางเทคนิคและจรณทักษะของสายงานอาชีพด้วยวิธีการทำเหมืองข้อความในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ชื่อนักศึกษา นางสาวปรัชญ์ธิดา พรประเสริฐ

รหัสประจำตัว 64605067

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์) ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง

พ.ศ. 2566

อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์สายชล สินสมบูรณ์ทอง

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อสำรวจความนิยมและศึกษาทักษะที่จำเป็นของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูลและวิศวกรข้อมูลโดยแบ่งประเภทของทักษะออกเป็น 2 ประเภทคือ ทักษะทางเทคนิคและจรณะทักษะ โดยการสร้างแบบจำลองเพื่อทำนายความถูกต้องจากข้อความประกาศรับสมัครงาน โดยใช้วิธีการทำเหมืองข้อความร่วมกับการสร้างแบบจำลองการเรียนรู้ 2 วิธี ซึ่งแบ่งเป็นการเรียนรู้ของเครื่อง 2 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีการถดถอยลอจิสติก เปรียบเทียบกับการเรียนรู้เชิงลึก 2 วิธีคือ วิธีเพอร์เซ็ปตรอนหลายชั้น และวิธีโครงข่ายประสาทแบบคอนโวลูชั่น เปรียบเทียบประสิทธิภาพโดยใช้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และคะแนนเอฟ1 ผลการสำรวจพบว่าอาชีพนักวิเคราะห์ข้อมูลได้รับความนิยมสูงที่สุด และผลการเปรียบเทียบประสิทธิภาพของแบบจำลองสรุปได้ว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพสูงที่สุดโดยมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1สูงที่สุดคือ 85.62%, 85.37%, 85.62% และ 84.78% ตามลำดับ

คำสำคัญ : เหมืองข้อความ ทักษะทางเทคนิค จรณะทักษะ

Independent Study Title Analysis of Technical Skills and Soft Skills of Career Path using Text Mining in Machine Learning and Deep Learning Methods

Student Name Ms. Prachtida Pornprasert

Student ID 64605067

Degree Master of Science Program in Data Science and Analytics
KMITL Digital Analytics and Intelligence Center

Year 2023

Independent Study Advisor Assoc.Prof.Saichon Sinsomboonthong

Abstract

The purpose of this research is to explore the popularity and study the skills required of the Data Analysts, Data Scientists and Data Engineers categorize their skills into two categories: technical skills and soft skills. A model is created to predict the correctness of job posting messages. A text mining method consists of two-step modeling learning algorithm, which is divided into two machine learning methods, namely the support vector machine and a logistic regression methods those methods are compared with two deep learning models, the multilayer perceptron and the convolutional neural network methods. The performance is compared using accuracy, precision, recall, and F1-score. The survey found that the most popular career is the Data Analysts. The results of comparing the model performance concluded that the support vector machine method is the best performance with the highest accuracy, precision, recall and F1-score of 85.62%, 85.37%, 85.62% and 84.78%, respectively.

Keywords: Text mining, Technical skill, Soft skill

กิตติกรรมประกาศ

งานวิจัยเรื่อง การวิเคราะห์ทักษะทางเทคนิคและจรณทักษะของสายงานอาชีพด้วยวิธีการทำเหมืองข้อความในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปด้วยดีเนื่องจากได้รับความอนุเคราะห์และสนับสนุนเป็นอย่างดียิ่งจาก รศ.สายชล สินสมบูรณ์ทอง อาจารย์ที่ปรึกษาการค้นคว้าอิสระที่ได้กรุณาให้คำปรึกษา ให้ความรู้ ข้อคิด คำแนะนำ และปรับปรุงแก้ไขข้อบกพร่องต่างๆ จนกระทั่งงานวิจัยครั้งนี้สำเร็จเรียบร้อยด้วยดี ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณรศ.ดร.ฉัฐไชย์ ลีนาวงศ์ และผศ.ดร.บุษยมาส พิมพ์พรรณชาติ คณะกรรมการการค้นคว้าอิสระที่กรุณาให้คำปรึกษา และคำแนะนำเพื่อความสมบูรณ์ยิ่งขึ้น

ขอขอบคุณครอบครัวและญาติพี่น้องทุกคนที่ช่วยเหลือสนับสนุนทั้งด้านกำลังใจและกำลังทรัพย์ด้วยดีตลอดมา

ขอขอบคุณอาจารย์และเพื่อนคณะวิทยาศาสตร์สาขาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ให้ความรู้และช่วยเหลือในการวิจัยครั้งนี้ นอกจากนี้ยังมีผู้ที่ให้ความร่วมมือช่วยเหลืออีกหลายท่าน ซึ่งผู้วิจัยไม่สามารถกล่าวนาม ณ ที่นี้ได้หมด จึงขอขอบคุณทุกท่านเหล่านั้นไว้ ณ โอกาสนี้ด้วย

สุดท้ายนี้ผู้วิจัยหวังว่างานวิจัยฉบับนี้จะเป็นประโยชน์สำหรับหน่วยงานที่เกี่ยวข้อง และผู้ที่สนใจศึกษาต่อไป

นางสาวปรัชญ์ธิดา พรประเสริฐ

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันสายงานที่เกี่ยวข้องกับการจัดการข้อมูลและวิเคราะห์ข้อมูลได้รับความนิยมมากขึ้นอันเนื่องมาจากกระแสที่มาแรงของข้อมูลขนาดใหญ่ (Big Data) ข้อมูลขนาดใหญ่ถูกสร้างขึ้นอย่างมากมายมหาศาลจากหลากหลายแหล่งที่มา นิยามของข้อมูลขนาดใหญ่ประกอบไปด้วย 3V คือ ปริมาณ (Volume) ความเร็ว (Velocity) และความหลากหลาย (Variety) (Yaqoob et al., 2016) อีกหนึ่งงานวิจัยที่กล่าวถึงข้อมูลขนาดใหญ่คืองานวิจัยของ Nocker and Sena (2019) ได้กล่าวว่าหลายๆ องค์กรเริ่มตระหนักว่าการใช้ประโยชน์จากข้อมูลขนาดใหญ่จะช่วยเพิ่มมูลค่าให้กับองค์กร สามารถนำมาใช้วิเคราะห์เพื่อหาโอกาสทางธุรกิจใหม่ๆ และตัดสินใจในเรื่องสำคัญ ในช่วงทศวรรษที่ผ่านมาการใช้ประโยชน์จากข้อมูลขนาดใหญ่นั้นได้กลายเป็นที่นิยมอย่างแพร่หลาย บริษัทหรือองค์กรต่างๆ เริ่มรับสมัครบุคคลที่จะทำหน้าที่ในการจัดการและวิเคราะห์ข้อมูลเพื่อให้ได้ประโยชน์สูงสุดต่อธุรกิจ จึงเป็นสาเหตุที่อาชีพของสายงานข้อมูลกำลังเป็นที่ต้องการในตลาดแรงงานอย่างมาก ในปัจจุบันมีการแบ่งสายงานการวิเคราะห์ข้อมูล (Data Analytics) ออกได้เป็น 3 สายหลัก (Johari, 2022) ดังต่อไปนี้คือ นักวิเคราะห์ข้อมูล (Data Analyst) นักวิทยาศาสตร์ข้อมูล (Data Scientist) และวิศวกรข้อมูล (Data Engineer) สำหรับ 3 สายงานดังกล่าวถึงแม้ว่าจะเกี่ยวข้องกับการจัดการข้อมูลหรือการวิเคราะห์ข้อมูล แต่จะมีรายละเอียดของงานและทักษะที่จำเป็นต่อการประกอบอาชีพแตกต่างกัน ซึ่งโดยปกติแล้วรายละเอียดของงานและทักษะของตำแหน่งงานที่ระบุในประกาศรับสมัครงานนั้นโดยส่วนใหญ่จะเขียนในรูปแบบของข้อความที่เป็นประโยค การจัดการกับข้อมูลที่เป็นลักษณะข้อความหรือประโยคปัจจุบันมีการใช้เทคนิคการทำเหมืองข้อความ (Text Mining) ร่วมกับแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) และแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) เข้ามาช่วยในการวิเคราะห์ข้อมูลที่อยู่ในรูปของข้อความหรือประโยคต่างๆมาวิเคราะห์เพื่อหาข้อมูลเชิงลึก (Insight) มากขึ้น

ในงานวิจัยของ Kobayashi et al. (2018) ได้ศึกษากระบวนการทำเหมืองข้อความโดยใช้เทคนิคการจัดสรรไดริชเลทแฝง (Latent Dirichlet Allocation : LDA) ร่วมกับการทำเหมืองข้อมูลโดยใช้แบบจำลองการเรียนรู้ของเครื่องเข้ามาช่วยวิเคราะห์ข้อความเพื่อวิเคราะห์หาหัวข้อที่อยู่ภายใต้กลุ่มของบทความ (Topic Modelling) ซึ่งการเรียนรู้ของเครื่องจะช่วยในการตรวจสอบรูปแบบของความถี่ของคำ งานวิจัยของ Hiranrat and Harncharnchai (2018) ได้ศึกษาทักษะทางเทคนิค (Technical Skill) และจรณทักษะที่จำเป็นสำหรับอุตสาหกรรมซอฟต์แวร์ในประเทศไทยเพื่อนำไปใช้ในการออกแบบหลักสูตรเพื่อพัฒนาทักษะของนักเรียนนักศึกษาโดยใช้วิธีการทำเหมืองข้อความด้วยชุดโปรแกรม NLTK (Natural

Language Toolkit) จากผลการศึกษาพบว่าทักษะทางเทคนิคที่พบมากที่สุดในประกาศรับสมัครงานสูงสุด 3 อันดับแรกคือ ภาษา JavaScript ภาษา SQL และภาษาจาวา โดยคิดเป็น 32.34, 26.44 และ 25.72 เปอร์เซ็นต์ ตามลำดับ และจรณทักษะ 3 อันดับแรกคือ ภาษาอังกฤษ การสื่อสาร และการวิเคราะห์โดยคิดเป็น 35.80, 31.14 และ 15.97 เปอร์เซ็นต์ ตามลำดับ

ในปีต่อมา Shirani (2019) ได้ศึกษาทักษะที่จำเป็นของสายงานการวิเคราะห์ข้อมูลโดยวิเคราะห์ทักษะสองประเภทดังต่อไปนี้คือ ทักษะทางเทคนิคและจรณทักษะ โดยมีขั้นตอนการดำเนินงานวิจัย 2 ขั้นตอนหลักคือ ขั้นตอนแรกวิธีการทำเหมืองข้อความโดยดำเนินการตัดคำ (Tokenization) และกำจัดคำที่ไม่สำคัญ (Stop Words) ออกจากข้อมูล เช่น คำสันธานและคำทั่วไปอื่นๆ ที่ไม่เกี่ยวข้อง ขั้นตอนที่สองใช้ขั้นตอนวิธีแบ่งกลุ่มข้อมูลแบบค่าเฉลี่ย (K-means Clustering) เพื่อหาค่า k ที่ดีที่สุดในการแบ่งแยกทักษะทั้งสองประเภท โดยทดลองกำหนดค่า k ทั้งหมดสามค่าดังต่อไปนี้คือ 2, 3 และ 4 วัดประสิทธิภาพโดยใช้ค่าความเหมือน (Similarity) จากผลการศึกษาพบว่าค่า k ที่ดีที่สุดในการแบ่งทักษะทั้งสองประเภทของแต่ละอาชีพมีดังต่อไปนี้ อาชีพนักวิเคราะห์ข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: การแสดงข้อมูล โปรแกรม Tableau โปรแกรม Qlik การสร้างแบบจำลองข้อมูล ฐานข้อมูลเชิงสัมพันธ์ ภาษา SQL โปรแกรม Microsoft Office โปรแกรม Excel โปรแกรม Excel Solver โปรแกรม Power BI การบริหารจัดการโครงการ ระบบการจัดการลูกค้าสัมพันธ์ ระบบบริหารจัดการทรัพยากรภายในองค์กร และแพลตฟอร์ม Salesforce จรณทักษะ: การโน้มน้าว การฟัง การจูงใจ การต่อรอง การทำงานเป็นทีม การแก้ปัญหา การมุ่งเน้นลูกค้า อาชีพนักวิทยาศาสตร์ข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และ 4 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: การวิเคราะห์ทางสถิติ การสร้างแบบจำลองการทำนาย การทดสอบ การวิจัยเชิงสาเหตุ ภาษาไพทอน การใช้แพ็กเกจ Pandas แพ็กเกจ NumPy และแพ็กเกจอื่นๆ ซอฟต์แวร์ Apache Spark ซอฟต์แวร์ Hadoop การเรียนรู้ของเครื่อง ภาษาสกาลา การทำให้เห็นได้ โปรแกรม MATLAB ฐานข้อมูลเชิงสัมพันธ์ การเรียนรู้เชิงลึก การหาค่าเหมาะที่สุด และการสร้างแบบจำลองมิติ จรณทักษะ: การสื่อสาร มุมมองระดับโลก การวิจัย การทำงานเป็นทีม การแก้ปัญหา และการทำงานร่วมกับผู้อื่น อาชีพวิศวกรข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และ 4 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: ซอฟต์แวร์ Apache Hadoop ซอฟต์แวร์ Spark Hive ซอฟต์แวร์ Sqoob โปรแกรม NiFi ภาษาจาวา ภาษาสกาลา เครื่องมือ HBase เครื่องมือ PySpark เครื่องมือ Flume เครื่องมือ Impala เครื่องมือ Parquet เครื่องมือ Oozie เครื่องมือ Storm การเก็บข้อมูลแบบ Avro ฐานข้อมูลเชิงสัมพันธ์ เครื่องมือ Bitbucket ฐานข้อมูลแบบ Columnar และแบบ NoSQL แพลตฟอร์ม Pig แพลตฟอร์ม Yarn และโปรแกรม Docker จรณทักษะ: การเขียนและการสื่อสาร การทำงานเป็นทีม และการแก้ปัญหา ในปีเดียวกันนั้น Maer-Matei et al. (2019) มีการใช้เทคนิคการทำเหมืองข้อความโดยใช้เทคนิคการจัดสรรได

ริชเลทแฝง เพื่อหาทักษะที่จำเป็นต่ออาชีพนักวิจัยและพัฒนาโดยการดึงข้อมูลจากประกาศรับสมัครงาน ซึ่งไม่เพียงแต่ได้ทักษะที่จำเป็นต่ออาชีพนักวิจัยและพัฒนาเท่านั้น แต่ยังมีประโยชน์ต่อการพัฒนาหลักสูตรการเรียนการสอนของสถาบันการศึกษาต่างๆ ในอนาคตอีกด้วย นอกจากนี้ Gurcan and Cagiltay (2019) ใช้เทคนิคการทำเหมืองข้อความเพื่อศึกษาเกี่ยวกับความรู้และทักษะที่จำเป็นในอาชีพที่เกี่ยวข้องกับการจัดการข้อมูลขนาดใหญ่จากข้อความประกาศรับสมัครงาน โดยใช้เทคนิคการจัดสรรไดริชเลทแฝง จากผลการศึกษาพบว่าทักษะทางภาษาคอมพิวเตอร์ (Programing Language) ที่สำคัญ 3 อันดับแรกคือ ภาษาจาวา ภาษาไพทอน และภาษาสกาลา ตามลำดับ ทักษะการใช้เครื่องมือในการจัดการข้อมูลขนาดใหญ่ (Big Data Tool) 3 อันดับแรกคือ เครื่องมือ Hadoop, Spark และ Kafka ตามลำดับ และจรณทักษะ (Soft Skill) 3 อันดับแรกคือ การสื่อสาร การแก้ไขปัญหา และการบริหารจัดการโครงการ ซึ่งสามารถกล่าวได้ว่านอกจากทักษะทางเทคนิคแล้ว จรณทักษะก็เป็นทักษะที่สำคัญที่องค์กรต่างๆ ให้ความสำคัญเป็นอย่างมาก หนึ่งในปัญหาที่ผู้หางานหรือผู้ที่กำลังคิดจะเปลี่ยนสายงานและนักเรียนนักศึกษาที่กำลังค้นหาว่าต้องการจะประกอบอาชีพใดหลังจากจบการศึกษาคือ ไม่ทราบว่าทักษะที่มีนั้นควรจะไปทำในตำแหน่งใด ทักษะที่มีอยู่นั้นเหมาะกับตำแหน่งใด หรือควรจะเสริมทักษะด้านใดในการจะก้าวเข้าสู่อาชีพที่เกี่ยวข้องกับการจัดการข้อมูลและการวิเคราะห์ข้อมูล จากความแตกต่างในรายละเอียดของงานและทักษะต่างๆ การวิเคราะห์ว่าทักษะใดบ้างที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูลและวิศวกรข้อมูล จึงมีความสำคัญเป็นอย่างมากในการช่วยประกอบการตัดสินใจ

Fareri et al. (2021) ศึกษาจรณทักษะโดยใช้วิธีการทำเหมืองข้อความและใช้เทคนิคการตัดคำด้วยชุดโปรแกรม Spacy ร่วมกับเทคนิคการรู้จำชื่อเฉพาะ (Named Entity Recognition : NER) จากนั้นศึกษาว่าวิธีใดสามารถนำมาทำนายทักษะได้มีประสิทธิภาพดีที่สุด โดยเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แ มชชีน (Support Vector Machine : SVM) และวิธีพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron : MLP) วัดประสิทธิภาพโดยใช้ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าคะแนนเอฟ1 (F1-Score) จากผลการศึกษาพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1สูงที่สุดคือ 68.1, 77.8 และ 72.6 ตามลำดับ ในขณะที่วิธีพอร์เซ็ปตรอนหลายชั้นได้ผลลัพธ์คือ 59.1, 65.7 และ 62.2 ตามลำดับ โดยจรณทักษะที่ได้จากชุดข้อมูลทดสอบ (Testing Data) คือ การแก้ปัญหา การให้เหตุผลเชิงรุก การสื่อสาร ความเป็นมืออาชีพ ความเป็นผู้นำ การทำงานเป็นทีม และความยืดหยุ่น ส่วน Wings et al. (2021) ได้กล่าวว่าขั้นตอนการคัดกรองผู้สมัครนั้นใช้เวลาและค่าดำเนินการจำนวนมาก จึงได้ทำการศึกษาการคัดกรองผู้สมัครงานแบบอัตโนมัติ (Automating Screening Process) โดยดึงสมรรถนะทักษะและจรณทักษะจากข้อมูลรายละเอียดของงาน โดยการทำเหมืองข้อความด้วยแบบจำลอง BERT จากนั้นเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก 5 วิธีคือ วิธีการถดถอยลอจิสติก

(Logistic Regression : LR) วิธี Gradient Boosting Classifier (GBC) วิธีป่าสุ่ม (Random Forest : RF) วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีพอร์เซ็ปตรอนหลายชั้น วัดประสิทธิภาพโดยใช้ค่าเรียกคืน ค่าความเที่ยง และค่าคะแนนเอฟ1 จากผลการศึกษาพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าเรียกคืน ค่าความเที่ยง และค่าคะแนนเอฟ1สูงที่สุดคือ 0.90, 0.91 และ 0.93 ตามลำดับ รองลงมาคือวิธีการถดถอยลอจิสติก 0.89, 0.90, 0.90 ตามลำดับ วิธีพอร์เซ็ปตรอนหลายชั้น 0.89, 0.89 และ 0.89 ตามลำดับ วิธีป่าสุ่ม 0.72, 0.95 และ 0.78 ตามลำดับ และ วิธี Gradient boosting Classifier 0.38, 0.40 และ 0.39 ตามลำดับ หลังจากนั้น Florentin et al. (2021) เป็นอีกหนึ่งงานวิจัยศึกษาทักษะโดยใช้โปรแกรมอัตโนมัติสกัดคำออกมาจากรายละเอียดของใบประวัติการทำงาน (Resume) โดยใช้วิธีการเข้ารหัส (Encoding) 3 วิธีคือ Word2vec, Doc2vec และ One-Hot จากนั้นเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีการถดถอยลอจิสติกและวิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional neural networks : CNN) วัดประสิทธิภาพโดยใช้ค่าเรียกคืน ค่าความเที่ยง และค่าความแม่น (Accuracy) จากผลการศึกษาพบว่าวิธี Word2vec ร่วมกับวิธีโครงข่ายประสาทแบบคอนโวลูชันให้ค่าเรียกคืน ค่าความเที่ยง และค่าความแม่นสูงที่สุดคือ 98.79, 91.34 และ 90.22 เปอร์เซ็นต์ ตามลำดับ รองลงมาคือวิธี Doc2vec ร่วมกับวิธีการถดถอยลอจิสติก 94.68, 81.45 และ 78.63 เปอร์เซ็นต์ ตามลำดับ และวิธี One-Hot ร่วมกับวิธีการถดถอยลอจิสติก 92.28, 80.11 และ 78.07 เปอร์เซ็นต์ ตามลำดับ จากความสำคัญและปัญหาที่กล่าวมาข้างต้น การวิเคราะห์ทักษะที่จำเป็นของแต่ละสายอาชีพ นอกจากจะช่วยผู้ที่กำลังหางานในสายอาชีพที่เกี่ยวกับข้อมูลแล้ว ยังช่วยให้บริษัทต่างๆ ที่กำลังรับสมัครบุคคลเข้าไปทำงานในสายอาชีพข้อมูล เขียนรายละเอียดของงานในช่องทางการประกาศรับสมัครงานได้ถูกต้องมากยิ่งขึ้น เพื่อให้ได้คนที่มีทักษะที่เหมาะสมในแต่ละอาชีพอย่างแท้จริง และช่วยลดระยะเวลาในการคัดเลือกบุคคลเข้าทำงานได้อีกด้วย

ดังนั้นผู้วิจัยจึงตั้งเป้าที่จะศึกษากระบวนการที่เหมาะสมในการวิเคราะห์ทักษะทางเทคนิคและจรณทักษะที่จำเป็นของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลด้วยการทำเหมืองข้อความร่วมกับการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก โดยเปรียบเทียบการทำเหมืองข้อมูลด้วยแบบจำลองการเรียนรู้ของเครื่อง 2 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีการถดถอยลอ จิสติก และแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีเพอร์เซ็ปตรอนหลายชั้น และวิธีโครงข่ายประสาทแบบคอนโวลูชัน เปรียบเทียบประสิทธิภาพแบบจำลองด้วยค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อสำรวจความนิยมของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

2) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก

3) เพื่อนำเสนอเชิงลึกเกี่ยวกับทักษะที่จำเป็นของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

1.3 ขอบเขตของงานวิจัย

ข้อมูลจากเว็บไซต์ www.jobsdb.com ซึ่งเป็นแหล่งรวบรวมประกาศรับสมัครงานในประเทศไทย ทำการเก็บข้อมูลจำนวน 2 ครั้ง คือเดือนมกราคม พ.ศ.2566 จำนวน 542 ชุด และเดือนเมษายน พ.ศ.2566 จำนวน 287 ชุด รวมทั้งหมด 829 ชุด หลังจากนั้นดำเนินการลบข้อมูลที่ซ้ำกัน จำนวน 68 ชุด จึงเหลือข้อมูลสำหรับการวิเคราะห์ทั้งหมดจำนวน 761 ชุด ประกอบด้วยตัวแปรอิสระ (X) คือ รายละเอียดของงาน (Description) และตัวแปรตาม (Y) คือชื่อตำแหน่งงาน (Job Title) เพื่อศึกษาทักษะทางเทคนิค (Technical Skill) และจรณทักษะ (Soft Skill) ที่จำเป็นของ 3 อาชีพคือ นักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล โดยในงานวิจัยนี้ผู้วิจัยได้ทำสำรวจข้อมูลเพิ่มเติม เช่น การกำหนดคำที่ไม่สื่อความหมาย และกำหนดทักษะทางเทคนิคเพิ่มเติมในขั้นตอนการวิเคราะห์ทักษะที่จำเป็นต่ออาชีพ โดยทำการแบ่งข้อมูลออกเป็น 2 ชุดคือ ชุดข้อมูลฝึกฝน 80% และชุดข้อมูลทดสอบ 20% (Wings et al., 2021) ใช้วิธีการทำเหมืองข้อความด้วยชุดโปรแกรม NLTK เปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง 2 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีการถดถอยลอจิสติก และแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีเพอร์เซ็ปตรอนหลายชั้น และวิธีโครงข่ายประสาทแบบคอนโวลูชัน โดยวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ทำให้ทราบความนิยมของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

2) ทำให้ได้แบบจำลองที่มีประสิทธิภาพสูงในการวิเคราะห์ทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูลนักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

3) ทำให้ทราบทักษะที่จำเป็นของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้กล่าวถึงทฤษฎีของการทำเหมืองข้อความ การเรียนรู้ของเครื่อง การเรียนรู้เชิงลึก มาตรวัดประสิทธิภาพ การทดสอบสัดส่วน และงานวิจัยที่เกี่ยวข้อง มีรายละเอียดดังต่อไปนี้

2.1 การทำเหมืองข้อความ (Text Mining)

การทำเหมืองข้อความเป็นกระบวนการสกัดข้อมูลในลักษณะข้อความที่มีขนาดใหญ่โดยทำการสกัดคำค้นหารูปแบบและความสัมพันธ์ที่อยู่ภายในข้อมูลเพื่อให้เกิดความหมายและสามารถใช้ประโยชน์จากข้อมูลได้อย่างครอบคลุม วิธีการทำเหมืองข้อความเป็นการสกัดสารสนเทศที่มีคุณค่าและค้นหาความสัมพันธ์ของข้อความจำนวนมาก โดยการวิเคราะห์ข้อมูลจากชุดข้อมูลที่มีโครงสร้างและไม่มีโครงสร้างอย่างมีเป้าหมาย จึงทำให้สามารถนำมาใช้ในการทำนายและจัดกลุ่มข้อมูล การเริ่มต้นวิเคราะห์ข้อความจะได้ข้อมูลที่มีลักษณะไม่มีโครงสร้างจึงต้องปรับให้อยู่ในข้อมูลที่มีโครงสร้าง เพื่อสะดวกต่อการวิเคราะห์โดยผ่านวิธีการวิเคราะห์ การตีความ และการให้ความหมาย เพื่อให้ได้ข้อมูลเชิงลึกจากข้อมูลขนาดใหญ่ที่มีความซับซ้อน นำไปสู่การสร้างแบบจำลองการทำนายการเรียนรู้ของเครื่อง แบบจำลองการทำนายการเรียนรู้เชิงลึก เพื่อให้ได้ประโยชน์แก่องค์กร สังคมและหน่วยงานต่างๆ (Kwartler, 2017)

ในการทำเหมืองข้อความมี 3 ขั้นตอน ก่อนการประมวลผลซึ่งประกอบด้วยการแบ่งคำ (Word Tokenization) การลบคำที่ไม่สื่อความหมาย (Stop Word Removal) และการตัดส่วนท้ายของคำ (Lemmatization) (Nayak et al., 2016)

2.1.1 การแบ่งคำ (Word Tokenization)

การแบ่งคำ คือ ขั้นตอนการจัดเตรียมข้อมูลข้อความโดยทำการแยกคำออกจากกันในประโยค ให้อยู่ในรูปของคำเดี่ยวหรือกลุ่มคำที่มีจำนวนตามที่เราสนใจ โดยเรียกว่า Token จากนั้นทำการรวมกลุ่มคำที่แตกต่างกัน และทำการนับจำนวนคำเหล่านั้นที่ปรากฏอยู่ในชุดข้อมูลซึ่งสามารถนำ Token ที่ได้ไปใช้ในการสกัดความรู้ออกจากชุดข้อมูลในขั้นตอนต่อ ๆ ไป การทำการแบ่งคำนั้น พยางค์หรือคำไม่เพียงแต่มีความหมายในตัวมันเองเท่านั้น แต่เมื่อนำคำศัพท์มาต่อกันจะทำให้เกิดความหมายของคำใหม่ที่แตกต่างจากเดิม ความหมายที่เฉพาะเจาะจงมากขึ้น หรือสามารถเพิ่มความสำคัญให้กับข้อความหรือประโยคนั้นๆ

2.1.2 การลบคำที่ไม่สื่อความหมาย (Stop Word Removal)

การลบคำที่ไม่สื่อความหมายเป็นขั้นตอนการลบคำทั่วๆไปในประโยคที่ไม่ค่อยสื่อความหมาย เช่น คำว่า a an และ the เป็นต้น

2.1.3 การตัดส่วนท้ายของคำ (Lemmatization)

การตัดส่วนท้ายของคำคือการตัดส่วนท้ายของคำภาษาอังกฤษที่ไม่จำเป็นต่อการคำนวณเช่น s หรือ es

2.1.4 ชุดโปรแกรม NLTK

NLTK เป็นแพลตฟอร์มทำงานกับข้อมูลภาษามนุษย์ด้วยชุดของไลบรารีประมวลผลข้อความสำหรับการจัดหมวดหมู่ การแปลงโทเค็น การแยกส่วน การติดแท็ก การแยกวิเคราะห์ และการให้เหตุผลเชิงความหมาย (Bird et al., 2009)

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือ การออกแบบโปรแกรมให้สามารถเรียนรู้และพัฒนาตนเองได้จากประสบการณ์ หลักการของการเรียนรู้ของเครื่อง คือการนำข้อมูลชุดฝึกฝนและข้อมูลออก มาป้อนเข้าไปให้กับคอมพิวเตอร์เพื่อสอนให้คอมพิวเตอร์เรียนรู้และทำให้เกิดการพัฒนาประสบการณ์ของตัวโปรแกรม เป็นการสร้างแบบจำลองการเรียนรู้ให้คอมพิวเตอร์สามารถทำนายหรือตัดสินใจได้ด้วยตนเองอย่างอัตโนมัติคล้ายมนุษย์ การเรียนรู้ของเครื่องแบ่งออกได้เป็น 3 ประเภทหลักๆ ดังต่อไปนี้ (อรพิน, 2564)

2.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ การนำข้อมูลชุดฝึกฝนมาสอนคอมพิวเตอร์ ซึ่งข้อมูลชุดฝึกฝนแต่ละตัวจะมีเลเบล (Lebel) กำกับอยู่ว่าข้อมูลแต่ละตัวมีเลเบลเป็นอะไร จากนั้นคอมพิวเตอร์ก็จะสร้างแบบจำลองการเรียนรู้ขึ้นมา ดังนั้นเมื่อมีข้อมูลใหม่ๆถูกป้อนเข้ามาการเรียนรู้ของเครื่องก็จะสามารถทำนายได้ว่าข้อมูลนั้นจะมีผลลัพธ์เป็นอย่างไร

2.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) จะไม่มีเลเบลกำกับข้อมูลหรือเป้าหมาย (Target) ใดๆมาสอนให้กับคอมพิวเตอร์ แต่คอมพิวเตอร์จะต้องนำข้อมูลชุดฝึกฝนมาสำรวจด้วยตนเองว่าข้อมูลใดบ้างที่มีคุณลักษณะ รูปแบบ หรือโครงสร้างคล้ายคลึงกันจากนั้นนำข้อมูลมาจัดเป็นกลุ่มของข้อมูลเดียวกัน ดังนั้นเมื่อมีข้อมูลใหม่ๆถูกป้อนเข้ามาก็จะสามารถทำนายได้ว่าข้อมูลใหม่นั้นจัดอยู่ในกลุ่มใด

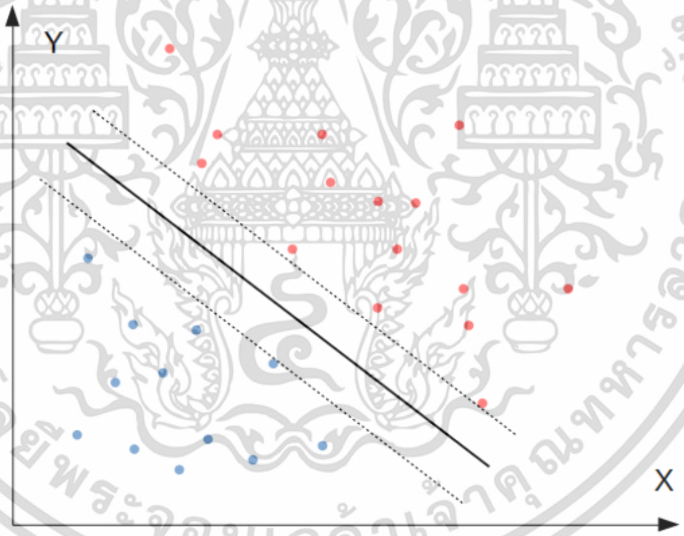
2.2.3 การเรียนรู้แบบลองผิดลองถูก (Reinforcement Learning) จะพิจารณาว่าพฤติกรรมนั้นเป็นสิ่งที่ต้องการหรือไม่ ถ้าเป็นสิ่งที่ต้องการจะให้ค่าเป็นบวก ถ้าเป็นสิ่งที่ไม่ต้องการจะให้ค่าเป็นลบ ซึ่งคอมพิวเตอร์จะต้องเรียนรู้ด้วยการทดลองไปเรื่อยๆจนกระทั่งได้แบบจำลองการเรียนรู้ของเครื่องที่ทำนายหรือตัดสินใจผลลัพธ์ที่ดีที่สุดออกมา

2.2.4 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมชชีนเป็นขั้นตอนวิธีหนึ่งของการเรียนรู้ของเครื่องที่จัดอยู่ในประเภทการเรียนรู้แบบมีผู้สอน หลักการทำงานของขั้นตอนวิธีนี้คือนำข้อมูลตัวอย่างมาพล็อตจุดข้อมูล จากนั้นลากเส้นตรงแบ่งจุดข้อมูล (Hyperplane) เพื่อแบ่งข้อมูลออกเป็นกลุ่มหรือคลาส เส้นตรงแบ่งจุดข้อมูลที่ดีที่สุดคือเส้นที่แบ่งกลุ่มข้อมูลได้ดีที่สุด หลักการคือลากเส้นคู่ขนานขึ้นมา 2 เส้นมาประกบเส้น เส้นตรงแบ่งจุดข้อมูล ซึ่งสามารถทำได้ 2 วิธีคือ

1) **Hard Margin** จะไม่ยอมให้มีจุดข้อมูลใดๆ อยู่บนเส้นคู่ขนาน หากมีข้อมูลบางส่วนกระจายตัวออกมาจากกลุ่มข้อมูล แต่ไม่สามารถลากเส้นผ่านจุดข้อมูลใดๆ ได้ ก็จะทำให้เส้นคู่ขนานแคบลง

2) **Soft Margin** ยอมให้มีจุดอยู่บนเส้นคู่ขนานได้บ้าง ทั้งนี้เพื่อไม่ให้เส้นคู่นานแคบจนเกินไป การพิจารณาว่าเส้นตรงแบ่งจุดข้อมูลที่ดีที่สุดหลักการคือถ้าเส้นคู่ขนานใดกว้างที่สุด มีระยะห่างระหว่างข้อมูลของแต่ละกลุ่มหรือคลาสสูงที่สุดแล้ว จะถือว่าเส้นตรงแบ่งจุดข้อมูลที่อยู่ตรงกลางระหว่างเส้นคู่ขนานนั้นคือเส้นตรงแบ่งจุดข้อมูลที่ดีที่สุด



รูปที่ 2.1 การหาเส้นตรงแบ่งจุดข้อมูล

จากรูปที่ 2.1 เป็นปัญหา Binary Classification เราต้องการจำแนกข้อมูลออกเป็นสองกลุ่มคือ สีน้ำเงินและสีแดง สิ่งที่ SVM ทำคือการหาเส้นแบ่งการตัดสินใจที่เป็นเส้นทึบซึ่งเส้นนี้จะเกิดขึ้นระหว่างกลางของเส้นประด้านซ้ายและขวา โดยมีเงื่อนไขว่าจะต้องหาคู่ของเส้นประที่กว้างที่สุดเท่าที่จะเป็นไปได้ (ชิตพงษ์, 2563)

2.2.5 การถดถอยลอจิสติก (Logistic Regression)

การถดถอยลอจิสติกเป็นขั้นตอนวิธีหนึ่งของการเรียนรู้ของเครื่องที่จัดอยู่ในประเภทการเรียนรู้แบบมีผู้สอน การวิเคราะห์การถดถอยลอจิสติกมีจุดประสงค์เพื่อทำนายโอกาสความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ โดยใช้ตัวแปรอิสระ 1 ตัวหรือมากกว่า 1 ตัวเพื่อนำมาวิเคราะห์ สามารถแบ่งการวิเคราะห์การถดถอยลอจิสติกออกได้เป็น 2 ประเภทคือ (ณรงค์ศักดิ์ และ อัครนันท์, 2564)

1) การวิเคราะห์การถดถอยลอจิสติกทวิภาค คือ การวิเคราะห์ลอจิสติกซึ่งตัวแปรตามจะมี 2 ค่า

2) การวิเคราะห์การถดถอยลอจิสติกพหุนาม คือ การวิเคราะห์ลอจิสติกซึ่งตัวแปรตามจะมีมากกว่า 2 ค่า

2.3 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกมีเป้าหมายเช่นเดียวกับการเรียนรู้ของเครื่องคือทำให้คอมพิวเตอร์เกิดการเรียนรู้แล้วนำความรู้นั้นมาใช้งาน แต่การเรียนรู้เชิงลึกใช้วิธีการหรือเทคนิคลักษณะโครงข่ายประสาทเทียม (Artificial Neural Network) มีความลึกหลายชั้น (Deep Neural Network) ที่เลียนแบบการทำงานของเซลล์โครงข่ายสมอง (กอบเกียรติ, 2565)

2.3.1 เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)

โครงสร้างของเพอร์เซ็ปตรอนหลายชั้นประกอบด้วยชั้น (Layer) ต่างๆ โดยแต่ละชั้นจะประกอบด้วยโหนด (Node) โดยแบ่งเป็น 3 ชั้นหลักมีหลักการทำงานดังต่อไปนี้

1) ชั้นข้อมูลเข้า (Input Layer) เป็นส่วนที่รับข้อมูลเข้า (Input Data) หรือค่าตัวแปร โดยแต่ละโหนดในชั้นนี้จะไม่มีฟังก์ชันกระตุ้น (Activation Function) จะทำหน้าที่รับส่งข้อมูลเข้าไปประมวลผลในชั้นถัดไป

2) ชั้นซ่อน (Hidden Layer) เป็นเพอร์เซ็ปตรอน หรือรับค่ามาจากชั้นข้อมูลเข้าทำการรวมผลและตัดสิน แต่ผลที่ได้จากการตัดสินใจชั้นนี้ยังไม่ใช่ข้อมูลออก (Output Data) สุดท้าย แต่จะส่งให้เพอร์เซ็ปตรอนอีกชั้นประมวลผลต่อไป การที่มีชั้นซ่อนนี้ช่วยให้สามารถตัดแบ่งหรือจำแนกกลุ่มข้อมูลที่ซับซ้อนได้

3) ชั้นข้อมูลออก (Output Layer) เป็นส่วนที่รับข้อมูลต่อจากชั้นซ่อนแล้วประมวลผล (รวมและตัดสิน) จากนั้นให้ค่าผลลัพธ์ข้อมูลออกสุดท้าย

2.3.2 โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

โครงข่ายประสาทแบบคอนโวลูชันมีแนวคิดมาจากการรับรู้ของระบบประสาทการมองเห็นของมนุษย์โดยมองวัตถุหรือภาพเป็นพื้นที่ย่อยๆ ซึ่งการมองเห็นพื้นที่ย่อยของมนุษย์จำทำการแยกคุณลักษณะเด่นเอาไว้เพื่อประกอบการพิจารณา แล้วนำกลุ่มคุณลักษณะเด่นของพื้นที่ย่อยมาพิจารณารวมกัน เพื่อให้

ได้รายละเอียดว่าวัตถุหรือภาพนั้นคืออะไร โครงข่ายประสาทแบบคอนโวลูชันเป็นการรวมสองส่วนคือ การสกัดคุณลักษณะ (Feature Extraction) และโครงข่ายประสาท (Neural Network) โดยกระบวนการสกัดคุณลักษณะอาศัยหลักการประมวลภาพเพื่อทำการจำแนกคุณลักษณะเด่นของวัตถุที่อยู่ในภาพออกมาก่อน เช่น เส้นขอบ เส้นโค้ง เส้นเอียง จากนั้นนำข้อมูลคุณลักษณะเด่นเป็นข้อมูลเข้ามาประมวลผลในโครงข่ายประสาทต่อไป

2.4 มาตรวัดประสิทธิภาพ (Evaluation Metrics)

2.4.1 เมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสนเป็นเครื่องมือในการประเมินผลลัพธ์ของการทำนาย (Prediction) ที่ทำนายจากแบบจำลองที่สร้างขึ้นในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก โดยมีหลักการคือสิ่งที่แบบจำลองทำนายกับสิ่งที่เกิดขึ้นจริงมีสัดส่วนเป็นอย่างไร (Pagon, 2019) ดังแสดงในรูปที่ 2.2

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

รูปที่ 2.2 ตารางเมทริกซ์ความสับสน

บวกจริง (True Positive, TP) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริงในกรณีทำนายว่าจริงและสิ่งที่เกิดขึ้นก็คือจริง

ลบจริง (True Negative, TN) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นในกรณีทำนายว่าไม่จริงและสิ่งที่เกิดขึ้นก็คือไม่จริง

บวกเท็จ (False Positive, FP) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นคือทำนายว่าจริงแต่สิ่งที่เกิดขึ้นคือไม่จริง

ลบเท็จ (False Negative, FN) คือ สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้นคือจริง

2.4.2 ค่าความแม่น (Accuracy)

ค่าความแม่นเป็นค่าความถูกต้องที่ทำนายได้เปรียบเทียบกับค่าที่เกิดขึ้นจริง ดังสมการที่ 2.1

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN) \quad (2.1)$$

2.4.3 ค่าความเที่ยง (Precision)

ค่าความเที่ยงเป็นการเปรียบเทียบการทำนายที่ถูกต้องว่าจริงและก็เกิดขึ้นจริง (TP) กับการทำนายว่าจริง แต่สิ่งที่เกิดขึ้นคือไม่จริง (FP) ดังสมการที่ 2.2

$$\text{Precision} = TP / (TP + FP) \quad (2.2)$$

2.4.4 ค่าเรียกคืน (Recall)

ค่าเรียกคืนเป็นค่าความถูกต้องของการทำนายว่าจะเป็นจริงเทียบกับจำนวนครั้งของเหตุการณ์ทั้งทำนายและเกิดขึ้นว่าเป็นจริง ดังสมการที่ 2.3

$$\text{Recall} = TP / (TP + FN) \quad (2.3)$$

2.4.5 ค่าคะแนนเอฟ1 (F1-Score)

ค่าคะแนนเอฟ1เป็นค่าเฉลี่ยแบบ Harmonic Mean ระหว่างค่าความเที่ยงและค่าเรียกคืน จุดประสงค์ของการสร้างค่าคะแนนเอฟ1ขึ้นมาเพื่อเป็น Single Metric ที่วัดความสามารถของแบบจำลอง ดังสมการที่ 2.4

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (2.4)$$

ในงานวิจัยนี้ใช้มาตรวัดประสิทธิภาพคือ ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1 เนื่องจากชุดข้อมูลรายละเอียดงานส่วนใหญ่ไม่ใช่ทักษะอาจทำให้แบบจำลองทำนายว่าประโยคหรือคำส่วนใหญ่ไม่ใช่ทักษะ ค่าความแม่นอาจจะสูง แต่ยังไม่เพียงพอในการบ่งบอกว่าแบบจำลองนั้นๆ มีประสิทธิภาพดีจริงหรือไม่ (Wings et al., 2021)

2.5 การทดสอบไคกำลังสอง (The Chi–square Test)

2.5.1 การทดสอบข้อมูลจำแนกทางเดียว

การทดสอบข้อมูลจำแนกทางเดียวเป็นการทดสอบข้อมูลที่จำแนกตามลักษณะอย่างใด อย่างหนึ่งเพียงอย่างเดียว คือ จะมีเพียงแถว (row) หรือสดมภ์ (column) เพียงอย่างเดียว เช่น จำนวนนักศึกษาจำแนกตามชั้นปีที่กำลังศึกษา จำนวนรถยนต์จำแนกตามยี่ห้อรถ เป็นต้น (สายชล, 2563)

ให้ O_i แทน จำนวนความถี่สังเกตได้ในกลุ่มหรือชั้นที่ i ; i = 1,..., k

E_i แทน จำนวนความถี่คาดหมายในกลุ่มหรือชั้นที่ i ; i = 1,..., k

n แทน จำนวนความถี่ทั้งหมด

k แทน จำนวนชั้นทั้งหมด

ดังนั้น $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$

และ $E_i = np_i$

โดยที่ p_i แทน ความน่าจะเป็นที่จะเกิดในกลุ่มหรือชั้นที่ i

โดยอาศัยทฤษฎีทางสถิติ จะได้ว่า $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ มีการแจกแจงโดยประมาณใกล้เคียงกับการแจกแจงไคกำลังสอง นั่นคือ

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

โดยมีองศาเสรีเท่ากับจำนวนกลุ่มหรือชั้นลบด้วย 1 และลบด้วยจำนวนพารามิเตอร์ที่ต้องประมาณค่าด้วยค่าจากตัวอย่าง นั่นคือ มีองศาเสรีเท่ากับ $k-1-n_p$ เมื่อ n_p คือ จำนวนพารามิเตอร์ที่ต้องการประมาณค่า แต่ถ้าไม่มีการประมาณค่าพารามิเตอร์ องศาเสรีจะเท่ากับ $k-1$ และจะปฏิเสธ H_0 ถ้า χ^2 ที่คำนวณได้มีค่ามากกว่า $\chi^2_{\alpha;k-1-n_p}$

ข้อจำกัดของการทดสอบไคกำลังสองจำแนกทางเดียว

1. ค่าสังเกตทุกค่าจะต้องเป็นอิสระกัน
2. ข้อมูลในตารางที่จะนำมาวิเคราะห์ควรเป็นข้อมูลที่เป็นความถี่ ไม่ควรอยู่ในรูปสัดส่วนหรือร้อยละหรือเปอร์เซ็นต์ ถ้าข้อมูลอยู่ในรูปสัดส่วนหรือร้อยละหรือเปอร์เซ็นต์ ต้องแปลงข้อมูลเป็นความถี่
3. จำนวนความถี่ทั้งหมดหรือขนาดตัวอย่างทั้งหมดควรมีขนาดใหญ่ คือไม่ต่ำกว่า 50
4. ความถี่คาดหมายในแต่ละกลุ่มจะต้องมีค่าอย่างน้อย 1 และมีจำนวนความถี่คาดหมายที่น้อยกว่า 5 ได้ไม่เกิน 20 % ของจำนวนกลุ่มทั้งหมด แต่ถ้ากลุ่มใดมีค่าความถี่คาดหมายน้อยกว่า 5 อาจแก้ไขได้ดังนี้

4.1 เพิ่มขนาดของตัวอย่าง (n) ให้มากขึ้น หรือเพิ่มค่าสังเกตให้มากขึ้น

4.2 รวมกลุ่มที่อยู่ติดกันเข้าด้วยกัน จนกว่าค่าความถี่คาดหมายไม่น้อยกว่า 5 วิธีนี้จะทำได้เมื่อการรวมกลุ่มเข้าด้วยกันแล้วไม่ทำให้ความหมายของกลุ่มที่นำมารวมกันเปลี่ยนแปลงไป

5. องศาเสรีเท่ากับ 1 และขนาดตัวอย่าง $n < 50$ ตัวสถิติทดสอบไคกำลังสองจะไม่มีความต่อเนื่อง จำเป็นต้องใช้**การปรับให้ต่อเนื่องของเยตส์ (Yate Continuity Correction)** ดังนี้

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i}$$

แต่ถ้า $n \geq 50$ ยังคงใช้สูตร

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

การทดสอบข้อมูลจำแนกทางเดียวแบ่งออกเป็น 2 หัวข้อ คือ

1. การทดสอบสัดส่วนประชากร k กลุ่ม (Test for k population proportions)

2. การทดสอบการแจกแจงหรือการทดสอบภาวะสารูปดี (Test for distribution or Test for goodness of fit)

สำหรับการทดสอบการแจกแจง เช่น การทดสอบการแจกแจงปรกติ ศึกษารายละเอียดได้ในบทที่ 12

2.5.1.1 การทดสอบสัดส่วนประชากร k กลุ่ม (Test for k Population Proportions)

การทดสอบสัดส่วนประชากร k กลุ่ม แบ่งออกเป็น 2 กรณี คือ

1. การทดสอบความแตกต่างระหว่างสัดส่วนประชากร k กลุ่ม $(k \geq 3)$
2. การทดสอบสัดส่วนประชากร k กลุ่ม ว่าเท่ากับค่าที่คาดไว้หรือไม่

1) การทดสอบความแตกต่างระหว่างสัดส่วนประชากร k กลุ่ม $(k \geq 3)$

เราทราบแล้วว่าการทดสอบความแตกต่างระหว่างสัดส่วนประชากร 2 กลุ่ม จะใช้ตัวสถิติทดสอบ Z ส่วนการทดสอบความแตกต่างระหว่างสัดส่วนประชากรตั้งแต่ 3 กลุ่ม ขึ้นไป จะใช้ตัวสถิติทดสอบ χ^2 ซึ่งเป็นการทดสอบสมมติฐานสำหรับข้อมูลจำแนกทางเดียว

สมมติฐาน

$H_0 : p_1 = p_2 = ... = p_k = \frac{1}{k}$

H_1 : มี $p_i \neq p_{i'}$ อย่างน้อย 1 ค่า ; $i \neq i' \; ; \; i, i' = 1, ..., k$

ตัวสถิติทดสอบ

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

โดยที่ $E_i = np_i = \frac{n}{k}$

เขตวิกฤต

จะปฏิเสธ H_0 ถ้า $\chi^2_{cal} > \chi^2_{\alpha,k-1}$

ตัวอย่างที่ 3.13 นักวิจัยตลาดมีความประสงค์ที่จะประเมินความชอบของแม่บ้านเกี่ยวกับเครื่องซักผ้าที่ใช้ภายในบ้านซึ่งมีสีให้เลือกแตกต่างกัน 4 สี ความถี่ต่อไปนี้ได้จากการสุ่มตัวอย่างแม่บ้านจำนวน 200 คน แล้วให้ระบุว่าชอบสีใด

สี	ครีม	น้ำตาล	ขาว	น้ำเงิน
ความถี่	61	55	41	43

ต้องการทราบว่าสีทั้ง 4 สี ได้รับความชอบพอ ๆ กันหรือไม่ ที่ระดับนัยสำคัญ 5%

วิธีทำ สมมติฐาน

H_0 : สัดส่วนของความชอบของแม่บ้านเกี่ยวกับเครื่องซักผ้ามีพอ ๆ กัน

H_1 : สัดส่วนของความชอบของแม่บ้านเกี่ยวกับเครื่องซักผ้าแตกต่างกัน

หรือ $H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ ✓

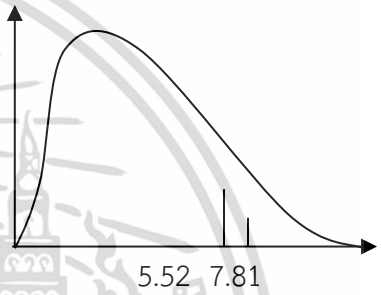
H_1 : มี p_i อย่างน้อย 1 ค่า ที่ไม่เท่ากับ $\frac{1}{4}$

ตัวสถิติทดสอบ

$$\chi^2_{cal} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

สี	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
ครีม	61	$200 \times 1/4 = 50$	2.42
น้ำตาล	55	$200 \times 1/4 = 50$	0.50
ขาว	41	$200 \times 1/4 = 50$	1.62
น้ำเงิน	43	$200 \times 1/4 = 50$	0.98
รวม	200	200	5.52

ดังนั้น χ^2_{cal} = 5.52



<u>วิธีที่ 1</u> พิจารณาจากค่าวิกฤต (critical value)

จากภาคผนวก ตารางที่ 5 ที่ $\alpha = 0.05$ และ k – 1 = 3

ค่าวิกฤตคือ $\chi^2_{0.05,3} = 7.81$

เนื่องจาก $\chi^2_{cal} = 5.52 < 7.81$ ซึ่งไม่ตกอยู่ในเขตวิกฤต จึงไม่สามารถปฏิเสธ H_0

ดังนั้น สีทั้ง 4 สี ได้รับความนิยมจากแม่บ้านพอ ๆ กัน ที่ระดับนัยสำคัญ 5%

<u>วิธีที่ 2</u> พิจารณาจากค่าพี (p-value)

p-value = $P(\chi^2 > 5.52)$

> $\alpha(= 0.05)$

ซึ่งไม่ตกอยู่ในเขตวิกฤต จึงไม่สามารถปฏิเสธ H_0 ที่ $\alpha = 0.05$

2) การทดสอบสัดส่วนประชากร k กลุ่ม ว่าเท่ากับค่าที่คาดไว้หรือไม่

ถ้าต้องการทดสอบว่าสัดส่วนประชากร k กลุ่ม เป็นไปตามที่คาดไว้หรือไม่ เช่น ต้องการทดสอบว่าสัดส่วนของผู้ที่ไม่เห็นด้วยกับการขึ้นค่าโดยสารรถประจำทางเป็น 3 เท่าของผู้ที่เห็นด้วยหรือไม่ หรือต้องการทดสอบว่าอัตราส่วนของกรุ๊ปเลือก A : B : AB : O เป็น 1 : 2 : 1 : 3 หรือไม่ จะใช้ตัวสถิติทดสอบ χ^2 ซึ่งยังคงเป็นการทดสอบสมมติฐานสำหรับข้อมูลจำแนกทางเดียว

สมมติฐาน

$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$

$H_1 :$ มี $p_i \neq p_{i0}$ อย่างน้อย 1 ค่า ; $i = 1, \ldots, k$

หรือ $H_0 : p_1 : p_2 : \ldots : p_k = p_{10} : p_{20} : \ldots : p_{k0}$

$H_1 :$ มี $p_i \neq p_{i0}$ อย่างน้อย 1 ค่า ; $i = 1, \ldots, k$

โดยที่ p_{i0} แทน ค่าสัดส่วนที่คาดว่าควรจะเป็น และ $0 \le p_{i0} \le 1$

ตัวสถิติทดสอบ

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

โดยที่ $E_i = np_{i0}$

เขตวิกฤต

จะปฏิเสธ H_0 ถ้า $\chi^2_{cal} > \chi^2_{\alpha,k-1}$

ตัวอย่างที่ 3.14 ในการผสมพันธ์ไม้ดอกลิสี่สีแดงและสีขาว ตามทฤษฎีทางพันธุกรรมของเมนเดล จะได้ดอกไม้สีขาว : สีชมพู : สีแดง ในอัตราส่วน 1 : 2 : 1 ถ้าผลจากการทดลองผสมพันธ์ไม้ดอก ลิลี่ครั้งหนึ่งได้ไม้ดอกสีขาว 141 ต้น สีชมพู 291 ต้น และสีแดง 132 ต้น จงทดสอบว่าผลการทดลองนี้สนับสนุนทฤษฎีของเมนเดลหรือไม่ ที่ระดับนัยสำคัญ 5%

วิธีทำ สมมติฐาน

H_0 : อัตราส่วนของไม้ดอกลิลี่สีขาว : สีชมพู : สีแดง = 1 : 2 : 1

H_1 : อัตราส่วนของไม้ดอกลิลี่สีขาว : สีชมพู : สีแดง ≠ 1 : 2 : 1

หรือ $H_0 : p_1 : p_2 : p_3 = 0.25 : 0.5 : 0.25$

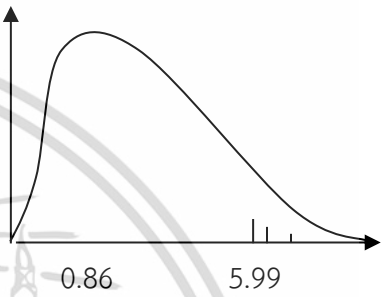
$H_1 : p_1 : p_2 : p_3 \neq 0.25 : 0.5 : 0.25$

ตัวสถิติทดสอบ

$$\chi^2_{cal} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

สีของดอกไม้	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
ขาว	141	$564 \times 1/4 = 141$	0
ชมพู	291	$564 \times 2/4 = 282$	0.29
แดง	132	$564 \times 1/4 = 141$	0.57
รวม	564	564	0.86

ดังนั้น χ^2_{cal} = 0.86



<u>วิธีที่ 1</u> พิจารณาจากค่าวิกฤต (critical value)

จากภาคผนวก ตารางที่ 5 ที่ $\alpha = 0.05$ และ k – 1 = 2

ค่าวิกฤตคือ $\chi^2_{0.05,2} = 5.99$

เนื่องจาก $\chi^2_{cal} = 0.86 < 5.99$ ซึ่งไม่ตกอยู่ในเขตวิกฤต จึงไม่สามารถปฏิเสธ H_0 ดังนั้นอัตราส่วนของไม้ดอกลิลี่เป็น 1 : 2 : 1 หรือผลการทดลองนี้สนับสนุนทฤษฎีของเมนเดล ที่ระดับนัยสำคัญ 5%

<u>วิธีที่ 2</u> พิจารณาจากค่าพี (p-value)

p-value = $P(\chi^2 > 0.86)$

> $\alpha(= 0.05)$

ซึ่งไม่ตกอยู่ในเขตวิกฤต จึงไม่สามารถปฏิเสธ H_0 ที่ $\alpha = 0.05$

2.6 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่ามีงานวิจัยที่ศึกษาเกี่ยวกับการทำเหมืองข้อความ และการสร้างแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก งานวิจัยที่เกี่ยวข้องที่ผู้วิจัยได้ทำการศึกษามีรายละเอียดดังต่อไปนี้

Kobayashi et al. (2018) ได้ศึกษากระบวนการทำเหมืองข้อความโดยใช้ชุดข้อมูลที่เกี่ยวกับการทำงานจำนวน 270,000 ประโยคที่ถูกเลเบลว่าเป็นประโยคที่มีทักษะอยู่หรือไม่มี โดยใช้เทคนิคการจัดสรรไดริชเลทแฝง (Latent Dirichlet Allocation : LDA) ร่วมกับการทำแบบจำลองการเรียนรู้ของเครื่องเพื่อวิเคราะห์หาหัวข้อทักษะที่อยู่ภายใต้กลุ่มของบทความ (Topic Modelling) เพื่อจัดกลุ่มของทักษะ ผลการศึกษาพบว่าหัวข้อทักษะที่ค้นพบมากที่สุดคือทักษะการสื่อสาร โดยคณะผู้วิจัยเปรียบเทียบการทำแบบจำลองการเรียนรู้ของเครื่อง 3 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีป่าสุ่ม และวิธีนาอีฟเบส์ วัดประสิทธิภาพแบบจำลองโดยใช้ค่าความแม่นและค่าคะแนนเอฟ1 จากผลการศึกษาพบว่าวิธีป่าสุ่ม วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีนาอีฟเบส์ให้ค่าความแม่นสูงที่สุดคือ 97.31, 97.30 และ 96.60 เปอร์เซ็นต์ตามลำดับ ในขณะที่วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีป่าสุ่ม และวิธีนาอีฟเบส์ให้ค่าคะแนนเอฟ1สูงที่สุดคือ 0.9751, 0.9750 และ 0.9554 ตามลำดับ

Hiranrat and Harncharnchai (2018) ได้ทำการวิเคราะห์ข้อมูลที่รวบรวมจากเว็บไซต์สมัครงานออนไลน์เพื่อให้ได้ทักษะทางเทคนิค (Technical Skill) และจรณทักษะที่จำเป็นสำหรับสายงานอุตสาหกรรมซอฟต์แวร์ในประเทศไทย โดยมีวัตถุประสงค์เพื่อนำไปใช้ในการออกแบบหลักสูตรเพื่อพัฒนาทักษะของนักเรียนนักศึกษาโดยใช้วิธีการทำเหมืองข้อความด้วยชุดโปรแกรม NLTK (Natural Language Toolkit) จากผลการศึกษาพบว่าทักษะทางเทคนิคที่พบมากที่สุดในประกาศรับสมัครงานสูงสุด 3 อันดับแรกคือ ภาษา JavaScript ภาษา SQL และภาษาจาวา โดยคิดเป็น 32.34, 26.44 และ 25.72 เปอร์เซ็นต์ตามลำดับ และจรณทักษะสามอันดับแรกคือ ภาษาอังกฤษ การสื่อสาร และการวิเคราะห์ โดยคิดเป็น 35.80, 31.14 และ 15.97 เปอร์เซ็นต์ ตามลำดับ

Shirani (2019) ได้ศึกษาทักษะที่จำเป็นของสายงานการวิเคราะห์ข้อมูลโดยวิเคราะห์ทักษะสองประเภทดังต่อไปนี้คือ ทักษะทางเทคนิคและจรณทักษะ โดยมีขั้นตอนการดำเนินงานวิจัย 2 ขั้นตอนหลัก คือขั้นตอนแรกวิธีการทำเหมืองข้อความโดยดำเนินการตัดคำ และกำจัดคำที่ไม่สำคัญออกจากข้อมูล เช่น คำสันธานและคำทั่วไปอื่นๆ ที่ไม่เกี่ยวข้อง ขั้นตอนที่สองใช้ขั้นตอนวิธีแบ่งกลุ่มข้อมูลแบบค่าเฉลี่ยเพื่อหาค่า k ที่ดีที่สุดในการแบ่งแยกทักษะทั้งสองประเภท โดยทดลองกำหนดค่า k ทั้งหมดสามค่าดังต่อไปนี้ 2, 3 และ 4 วัดประสิทธิภาพโดยใช้ค่าความเหมือน (Similarity) จากผลการศึกษาพบว่าค่า k ที่ดีที่สุดในการแบ่งทักษะทั้งสองประเภทของแต่ละอาชีพมีดังต่อไปนี้ อาชีพนักวิเคราะห์ข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: การแสดงข้อมูล โปรแกรม

Tableau โปรแกรม Qlik การสร้างแบบจำลองข้อมูล ฐานข้อมูลเชิงสัมพันธ์ ภาษา SQL โปรแกรม Microsoft Office โปรแกรม Excelโปรแกรม Excel Solver โปรแกรม Power BI การบริหารจัดการโครงการ ระบบการจัดการลูกค้าสัมพันธ์ ระบบบริหารจัดการทรัพยากรภายในองค์กร และแพลตฟอร์ม Salesforce จรณทักษะ: การโน้มน้าว การฟัง การจูงใจ การต่อรอง การทำงานเป็นทีม การแก้ปัญหา การมุ่งเน้นลูกค้า อาชีพนักวิทยาศาสตร์ข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และ 4 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: การวิเคราะห์ทางสถิติ การสร้างแบบจำลองการทำนาย การทดสอบ การวิจัยเชิงสาเหตุ ภาษาไพทอน การใช้แพ็กเกจ Pandas แพ็กเกจ NumPy และแพ็กเกจอื่นๆ ซอฟต์แวร์ Apache Spark ซอฟต์แวร์ Hadoop การเรียนรู้ของเครื่อง ภาษาสกาลา การทำให้เห็นได้ โปรแกรม MATLAB ฐานข้อมูลเชิงสัมพันธ์ การเรียนรู้เชิงลึกการหาค่าเหมาะที่สุด และการสร้างแบบจำลองมิติ จรณทักษะ: การสื่อสาร มุมมองระดับโลก การวิจัย การทำงานเป็นทีม การแก้ปัญหา และการทำงานร่วมกับผู้อื่น อาชีพวิศวกรข้อมูล แบ่งแยกทักษะได้ดีที่สุดด้วยค่า k เท่ากับ 3 และ 4 และทักษะที่จำเป็นต่ออาชีพคือ ทักษะทางเทคนิค: ซอฟต์แวร์ Apache Hadoop ซอฟต์แวร์ Spark Hive ซอฟต์แวร์ Sqoob โปรแกรม NiFi ภาษาจาวา ภาษาสกาลา เครื่องมือ HBase เครื่องมือ PySpark เครื่องมือ Flume เครื่องมือ Impala เครื่องมือ Parquet เครื่องมือ Oozie เครื่องมือ Storm การเก็บข้อมูลแบบ Avro ฐานข้อมูลเชิงสัมพันธ์ เครื่องมือ Bitbucket ฐานข้อมูลแบบ Columnar และแบบ NoSQL แพลตฟอร์ม Pig แพลตฟอร์ม Yarn และโปรแกรม Docker จรณทักษะ: การเขียนและการสื่อสาร การทำงานเป็นทีม และการแก้ปัญหา

Maer-Matei et al. (2019) มีการใช้เทคนิคการทำเหมืองข้อความโดยใช้เทคนิคการจัดสรรไดริชเลทแฝงโดยการดึงข้อมูลจากประกาศรับสมัครงาน เพื่อหาทักษะหรือความรู้ที่จำเป็นต่ออาชีพนักวิจัยและพัฒนา 5 กลุ่ม คือ วิศวกรรม เศรษฐศาสตร์ วิทยาการคอมพิวเตอร์ วิทยาศาสตร์สิ่งแวดล้อม และคณิตศาสตร์ จากผลการศึกษาพบว่าทักษะหรือความรู้ที่จำเป็นต่ออาชีพนักวิจัยและพัฒนากลุ่มวิศวกรรม คือ ฟิสิกส์ พลังงาน วัสดุ เครื่องกล ไฟฟ้า คณิตศาสตร์ และอิเล็กทรอนิกส์ ทักษะหรือความรู้ที่จำเป็นต่ออาชีพนักวิจัยและพัฒนากลุ่มศรษฐศาสตร์คือ การเงิน การจัดการ และธุรกิจ ทักษะหรือความรู้ที่จำเป็นต่ออาชีพอาชีพนักวิจัยและพัฒนากลุ่มวิทยาการคอมพิวเตอร์คือ การเรียนรู้เชิงลึก การเรียนรู้ของเครื่อง ขั้นตอนวิธี ดิจิทัล คอมพิวเตอร์ และปัญญาประดิษฐ์ ทักษะหรือความรู้ที่จำเป็นต่ออาชีพอาชีพนักวิจัยและพัฒนากลุ่มวิทยาศาสตร์สิ่งแวดล้อมคือ น้ำภูมิอากาศ มหาสมุทร โลก ความยั่งยืน ชีววิทยา การเปลี่ยนแปลง ระบบนิเวศ และความยั่งยืน และทักษะหรือความรู้ที่จำเป็นต่ออาชีพอาชีพนักวิจัยและพัฒนากลุ่มคณิตศาสตร์คือ ทฤษฎีความน่าจะเป็น สมการเชิงอนุพันธ์ และฟิสิกส์

Gurcan and Cagiltay (2019) ใช้เทคนิคการทำเหมืองข้อความและเทคนิคการจัดสรรดีไดริชเลทแฝง เพื่อศึกษาเกี่ยวกับความรู้และทักษะที่จำเป็นในอาชีพที่เกี่ยวข้องกับการจัดการข้อมูลขนาดใหญ่

โดยใช้ชุดข้อมูลจากข้อความประกาศรับสมัครงาน จากผลการศึกษาพบว่าทักษะทางภาษาคอมพิวเตอร์ (Programing Language) ที่สำคัญ 3 อันดับแรกคือ ภาษาจาวา ภาษาไพทอน และภาษาสกาลา ตามลำดับ ทักษะการใช้เครื่องมือในการจัดการข้อมูลขนาดใหญ่ (Big Data Tool) 3 อันดับแรกคือ เครื่องมือ Hadoop, Spark, Kafka ตามลำดับ และจรณทักษะ (Soft Skill) 3 อันดับแรกคือ การสื่อสาร การแก้ไขปัญหา และการบริหารจัดการโครงการ

Fareri et al. (2021) ได้ศึกษาจรณทักษะโดยใช้วิธีการทำเหมืองข้อความและใช้เทคนิคการตัดคำด้วยชุดโปรแกรม SPACY ร่วมกับเทคนิคการรู้จำชื่อเฉพาะ จากนั้นศึกษาว่าขั้นตอนวิธีใดสามารถนำมาทำนายทักษะได้มีประสิทธิภาพดีที่สุดโดยเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก 2 ขั้นตอนวิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีพอร์เซ็ปตรอนหลายชั้น วัดประสิทธิภาพโดยใช้ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1 จากผลการศึกษาพบว่าจรณทักษะที่ได้จากชุดข้อมูลทดสอบ คือ การแก้ปัญหา การให้เหตุผลเชิงรุก การสื่อสาร ความเป็นมืออาชีพ ความเป็นผู้นำ การทำงานเป็นทีม และความยืดหยุ่น การเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องพบว่า วิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1สูงที่สุดคือ 68.1, 77.8 และ 72.6 ตามลำดับ ในขณะที่วิธีพอร์เซ็ปตรอนหลายชั้นได้ผลลัพธ์ 59.1, 65.7 และ 62.2 ตามลำดับ

Wings et al. (2021) ได้กล่าวว่าขั้นตอนการคัดกรองผู้สมัครนั้นใช้เวลาและค่าดำเนินการจำนวนมาก จึงได้ทำการศึกษาการคัดกรองผู้สมัครงานแบบอัตโนมัติ (Automating Screening Process) โดยดึงสมรรถนทักษะและจรณทักษะจากชุดข้อมูลรายละเอียดของงาน โดยการทำเหมืองข้อความด้วยแบบจำลอง BERT จากนั้นเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก 5 วิธีคือ วิธีการถดถอยลอจิสติก วิธี Gradient boosting Classifier วิธีป่าสุ่ม วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีพอร์เซ็ปตรอนหลายชั้น วัดประสิทธิภาพโดยใช้ค่าเรียกคืน ค่าความเที่ยง และค่าคะแนนเอฟ1 จากผลการศึกษาพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าเรียกคืน ค่าความเที่ยง และค่าคะแนนเอฟ1สูงที่สุดคือ 0.90, 0.91 และ 0.93 ตามลำดับ รองลงมาคือวิธีการถดถอยลอจิสติก 0.89, 0.90 และ 0.90 ตามลำดับ วิธีพอร์เซ็ปตรอนหลายชั้น 0.89, 0.89 และ 0.89 ตามลำดับ วิธีป่าสุ่ม 0.72, 0.95 และ 0.78 ตามลำดับ และ วิธี Gradient boosting Classifier 0.38, 0.40 และ 0.39 ตามลำดับ

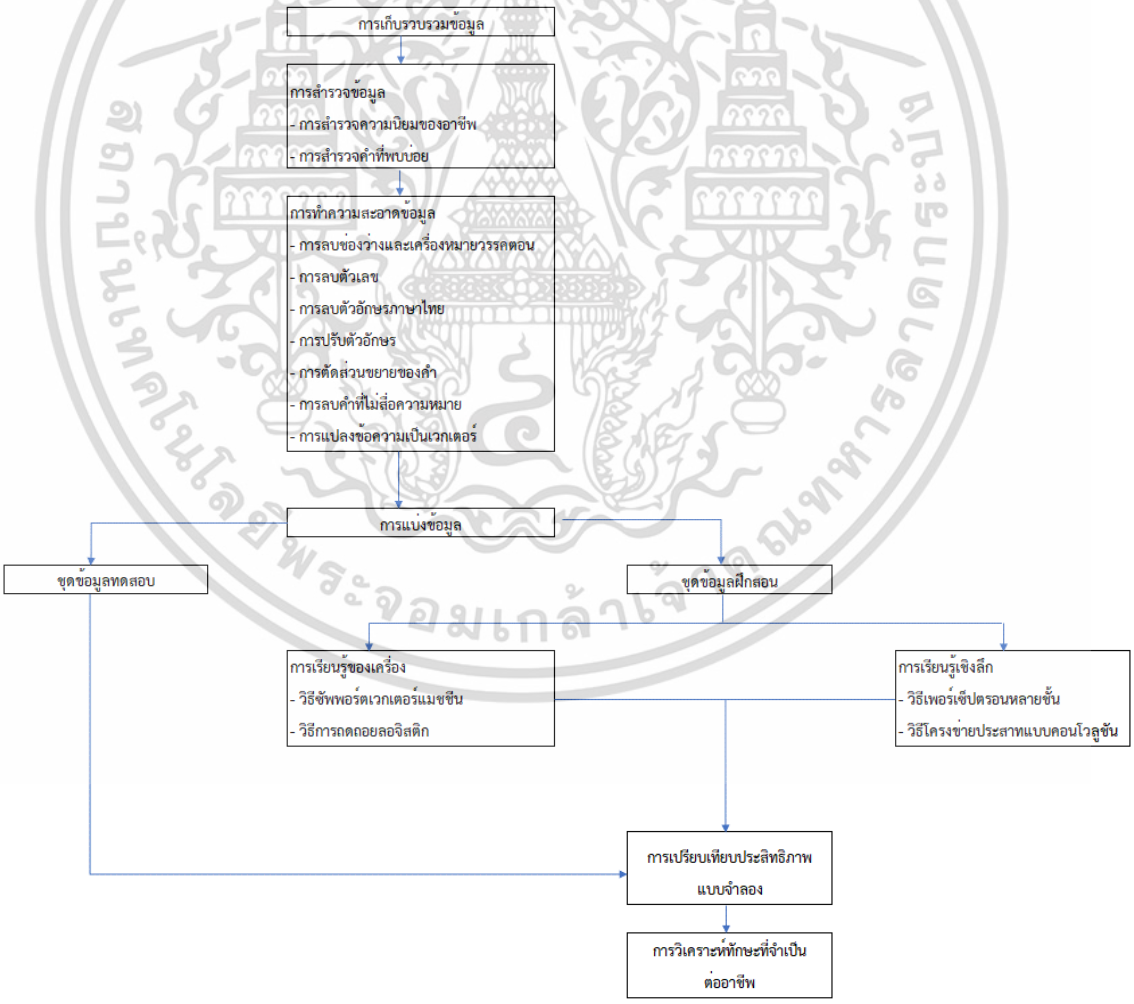
Florentin. et al. (2021) ได้ศึกษาทักษะโดยใช้โปรแกรมอัตโนมัติสกัดคำออกมาจากรายละเอียดในใบประวัติการทำงานโดยใช้วิธีการเข้ารหัส (Encoding) 3 วิธีคือ Word2vec, Doc2vec และ One-Hot จากนั้นเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือวิธีการถดถอยลอจิสติกและวิธีโครงข่ายประสาทแบบคอนโวลูชัน วัดประสิทธิภาพโดยใช้ค่าเรียกคืน ค่าความเที่ยง และค่าความแม่น จากผลการศึกษาพบว่า Word2vec ร่วมกับวิธีโครงข่ายประสาทแบบคอนโวลูชันให้ค่าเรียกคืน ค่าความเที่ยง และค่าความแม่นสูงที่สุดคือ 98.79, 91.34 และ 90.22

เปอร์เซ็นต์ ตามลำดับ รองลงมาคือ Doc2vec ร่วมกับวิธีการถดถอยลอจิสติก 94.68, 81.45 และ 78.63 เปอร์เซ็นต์ ตามลำดับ และ One-Hot ร่วมกับวิธีการถดถอยลอจิสติก 92.28, 80.11 และ 78.07 เปอร์เซ็นต์ ตามลำดับ

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยนี้คือการนำข้อมูลมาวิเคราะห์ทักษะทางเทคนิค (Technical Skill) และจรณทักษะ (Soft Skill) ที่จำเป็นของอาชีพ 3 อาชีพคือ นักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล ขั้นตอนดำเนินงานวิจัยแบ่งออกเป็น 7 ขั้นตอนหลัก ได้แก่ การเก็บรวบรวมข้อมูล การสำรวจข้อมูล การทำความสะอาดข้อมูล การแบ่งข้อมูล การสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก การเปรียบเทียบประสิทธิภาพของแบบจำลอง และการวิเคราะห์ทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล



รูปที่ 3.1 ขั้นตอนในการดำเนินงานวิจัย

จากรูปที่ 3.1 แสดงให้เห็นถึงขั้นตอนการดำเนินงานโดยเริ่มจากการเก็บรวบรวมชุดข้อมูลมาจากเว็บไซต์ www.jobsdb.com จากนั้นทำการสำรวจข้อมูลด้วยโปรแกรม SPSS และชุดคำสั่ง WordClound จากนั้นทำความสะอาดข้อมูลโดยเปลี่ยนตัวอักษรภาษาอังกฤษจากตัวพิมพ์ใหญ่เป็นตัวพิมพ์เล็ก ลบตัวอักขระพิเศษ เช่น ! - # / และลบคำที่เป็นคำที่ไม่สื่อความหมายหรือคำฟุ่มเฟือย (Stop Word) โดยใช้คลังคำศัพท์ภาษาอังกฤษของชุดคำสั่ง NLTK ขั้นตอนต่อไปนำข้อมูลแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอน 80% และชุดข้อมูลทดสอบ 20% นำชุดข้อมูลฝึกสอนมาสร้างแบบจำลองการเรียนรู้ของเครื่อง 2 วิธี คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และวิธีการถดถอยลอจิสติก (Logistic Regression) และสร้างแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron) และวิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network) แล้วนำชุดข้อมูลทดสอบมาวัดประสิทธิภาพของแบบจำลองด้วยค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าคะแนนเอฟ1 (F1-Score) เพื่อมาเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก ในขั้นตอนสุดท้ายเมื่อได้แบบจำลองที่มีประสิทธภาพดีที่สุดจะทำการวิเคราะห์ทักษะทางเทคนิคและจรณทักษะของแต่ละอาชีพ แล้วทำการสรุปและวิเคราะห์ผล

3.1 การเก็บรวบรวมข้อมูล

ผู้วิจัยได้ทำการเก็บรวบรวมชุดข้อมูลประกาศรับสมัครงานจากเว็บไซต์ www.jobsdb.com ซึ่งเป็นแหล่งรวบรวมประกาศรับสมัครงานในประเทศไทยจำนวน 761 ชุด โดยชุดข้อมูลดังกล่าวเปิดให้เข้าใช้งานในรูปแบบของสาธารณะ โดยข้อมูลชุดนี้ประกอบไปด้วยอาชีพของนักวิเคราะห์ข้อมูล 571 ข้อความ คิดเป็น 75% อาชีพวิศวกรข้อมูล 120 ข้อความ คิดเป็น 16% และอาชีพนักวิทยาศาสตร์ข้อมูล 70 ข้อความ คิดเป็น 9% โดยชุดข้อมูลมีลักษณะต่อไปดังนี้

ตารางที่ 3.1 ลักษณะของชุดข้อมูล

ชื่อตัวแปร	คำอธิบาย	ชนิดของตัวแปร
description (x)	รายละเอียดของงาน	ตัวแปรเชิงคุณภาพ
job_title (y)	ชื่อตำแหน่งงาน	ตัวแปรเชิงคุณภาพ

จากตารางที่ 3.1 ภายในชุดข้อมูลประกอบด้วยตัวแปรอิสระ (X) คือ รายละเอียดของงาน (description) และตัวแปรตาม (Y) คือชื่อตำแหน่งงาน (job_title) ของอาชีพของนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูลและวิศวกรข้อมูล โดยชุดข้อมูลทั้งหมดเป็นภาษาอังกฤษและเป็นตัวแปรเชิงคุณภาพ

job_title	description
Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
Data Analyst	Data Analysis Analysis Data Analysis manager
Data Analyst	Opportunity to work in International environment. Have learned digital tools. Office located near BTS and MRT
Data Scientist	Data Analysis role Tableau , Power BI Python , SQL
Data Analyst	Experiences in software development Strong analytical skills, Attention to detail Working hybrid, Flexible working hour
Data Analyst	Microsoft PowerBI, Dashboard, ETL SQL, SAS, R, Python Capacity Planning and Performance Monitoring
Data Engineer	Big Query, SQL, DOMO, Google studio Data pipelines &transformation/ Data visualization Permanent role/ Hybrid Office
Data Engineer	Data Engineer Monitor,plan,create & maintain data architectures Database Management, SQL, Python, MS Excel, Access

รูปที่ 3.2 ตัวอย่างชุดข้อมูลประกาศรับสมัครงาน

จากรูปที่ 3.2 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานที่เก็บรวบรวมมาจากเว็บไซต์ www.jobsdb.com โดยที่ job_title คือ ชื่อตำแหน่งงาน และ description คือ คำอธิบายรายละเอียดของงาน

3.2 การสำรวจข้อมูล (Data Exploration)

หลังจากที่เก็บรวบรวมข้อมูลแล้วมาทำการสำรวจข้อมูลเพื่อหาความนิยมของประกาศรับสมัครงานในแต่ละอาชีพ และเพื่อค้นหาคำที่ปรากฏบ่อยในประกาศรับสมัครงาน

3.2.1 การสำรวจความนิยมของอาชีพ

ในขั้นตอนนี้ผู้วิจัยทำการสำรวจความนิยมของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล ด้วยการตั้งสมมุติฐานการโดยใช้การทดสอบไคกำลังสองจากโปรแกรม SPSS เพื่อทำการสำรวจสัดส่วนของการประกาศรับสมัครงานว่ามีความแตกต่างกันอย่างมีนัยสำคัญหรือไม่ โดยตั้งสมมุติฐานดังต่อไปนี้

H_0 : สัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลไม่แตกต่างกัน

H_1 : สัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลแตกต่างกัน

3.2.2 การสำรวจคำที่พบบ่อย

หลังจากที่ทำการสำรวจความนิยมของอาชีพ แล้วจึงทำการสำรวจคำที่พบบ่อยในแต่ละอาชีพซึ่งประกอบไปด้วยอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล ด้วยการใช้ชุดคำสั่ง

WorldCloud ตามลำดับ ซึ่งการสำรวจในครั้งนี้จะนำไปประกอบการกำจัดคำที่ไม่สื่อความหมายร่วมกันกับขั้นตอนทำความสะอาดข้อมูลในลำดับถัดไป

3.3 การทำความสะอาดข้อมูล (Data Cleaning)

การทำความสะอาดข้อมูล เป็นการแก้ไขรายละเอียดของคำหรือตัดคำที่ไม่สื่อความหมายออกหรือปรับตัวอักษรของคำให้อยู่ในรูปแบบเดียวกัน โดยผู้วิจัยได้ดำเนินการดังต่อไปนี้

3.3.1 การลบช่องว่างและเครื่องหมายวรรคตอน

index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment. Have learned digital tools. Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau , Power BI Python , SQL
4	Data Analyst	Experiences in software development Strong analytical skills, Attention to detail Working hybrid, Flexible working hour
5	Data Analyst	Microsoft PowerBI, Dashboard, ETL SQL, SAS, R, Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query, Python, SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze, monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst

รูปที่ 3.3 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนลบช่องว่างและเครื่องหมายวรรคตอน

index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment Have learned digital tools Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau Power BI Python SQL
4	Data Analyst	Experiences in software development Strong analytical skills Attention to detail Working hybrid Flexible working hour
5	Data Analyst	Microsoft PowerBI Dashboard ETL SQL SAS R Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query Python SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst

รูปที่ 3.4 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังลบช่องว่างและเครื่องหมายวรรคตอน

จากรูปที่ 3.3 และ 3.4 เป็นการลบช่องว่างและเครื่องหมายวรรคตอนโดยประกอบไปด้วย . ? ! , : ; ’ – “ ” () []

3.3.2 การลบตัวเลข

index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment. Have learned digital tools. Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau , Power BI Python , SQL
4	Data Analyst	Experiences in software development Strong analytical skills, Attention to detail Working hybrid, Flexible working hour
5	Data Analyst	Microsoft PowerBI, Dashboard, ETL SQL, SAS, R, Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query, Python, SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze, monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst
11	Data Analyst	Bachelor's degree in Actuarial Sciences or Related Welcome new graduates Excellent in english communication
12	Data Analyst	At least 2 yrs working experience in data analysis Strong analytical skills Can-do attitude

รูปที่ 3.5 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนลบตัวเลข

index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment Have learned digital tools Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau Power BI Python SQL
4	Data Analyst	Experiences in software development Strong analytical skills Attention to detail Working hybrid Flexible working hour
5	Data Analyst	Microsoft PowerBI Dashboard ETL SQL SAS R Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query Python SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst
11	Data Analyst	Bachelor s degree in Actuarial Sciences or Related Welcome new graduates Excellent in english communication
12	Data Analyst	At least yrs working experience in data analysis Strong analytical skills Can do attitude

รูปที่ 3.6 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังลบตัวเลข

จากรูปที่ 3.5 และ 3.6 เป็นการลบตัวเลข 0-9 ออกจากข้อมูล

3.3.3 การลบตัวอักษรภาษาไทย

30	Data Analyst	years in BI and Data Analytics in Finance Advanced in Power BI and Data Visualization Good communication coordination and team work
31	Data Analyst	Strong analytical skills Good communication and interpersonal skills Details oriented and drive for the result
32	Data Analyst	Design and develop data visualizations Design analysis report Diagnose and Troubleshoot Database errors
33	Data Analyst	Hybrid workplace We are a Tech Company of BDMS Group data architecture and pipelines that adhere to ETL
34	Data Analyst	Over yearsไ experience in Data Analyst Understanding of manufacturing process Advance in Excel Power BI
35	Data Analyst	Experienced in Marketing Research Data Analytics Strong presentation skills on verbal presentation days work Work from home Medical Insurance
36	Data Analyst	Collect process and clean data Transform data into structural meaningful format Hybrid working
37	Data Analyst	Business Analyst Assistant Business Analyst Investment
38	Data Analyst	Work From Anywhere Good in English Business Analyst Application Service
39	Data Analyst	Multinational company Attractive Benefits Data Analyst
40	Data Analyst	years experience in Business Analyst Core Bank Good command in English

รูปที่ 3.7 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนลบตัวอักษรภาษาไทย

30	Data Analyst	years in BI and Data Analytics in Finance Advanced in Power BI and Data Visualization Good communication coordination and team work
31	Data Analyst	Strong analytical skills Good communication and interpersonal skills Details oriented and drive for the result
32	Data Analyst	Design and develop data visualizations Design analysis report Diagnose and Troubleshoot Database errors
33	Data Analyst	Hybrid workplace We are a Tech Company of BDMS Group data architecture and pipelines that adhere to ETL
34	Data Analyst	Over years experience in Data Analyst Understanding of manufacturing process Advance in Excel Power BI
35	Data Analyst	Experienced in Marketing Research Data Analytics Strong presentation skills on verbal presentation days work Work from home Medical Insurance
36	Data Analyst	Collect process and clean data Transform data into structural meaningful format Hybrid working
37	Data Analyst	Business Analyst Assistant Business Analyst Investment
38	Data Analyst	Work From Anywhere Good in English Business Analyst Application Service
39	Data Analyst	Multinational company Attractive Benefits Data Analyst
40	Data Analyst	years experience in Business Analyst Core Bank Good command in English

รูปที่ 3.8 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังลบตัวอักษรภาษาไทย

จากรูปที่ 3.7 และ 3.8 เป็นการลบตัวอักษรภาษาไทยออกจากข้อมูล

3.3.4 การปรับตัวอักษรเป็นตัวพิมพ์เล็กทั้งหมด

index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment Have learned digital tools Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau Power BI Python SQL
4	Data Analyst	Experiences in software development Strong analytical skills Attention to detail Working hybrid Flexible working hour
5	Data Analyst	Microsoft PowerBI Dashboard ETL SQL SAS R Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query Python SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst

รูปที่ 3.9 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนปรับตัวอักษรเป็นตัวพิมพ์เล็ก

index	job_title	description
0	Data Analyst	business intelligence analyst ecommerce market insights data
1	Data Analyst	data analysis analysis data analysis manager
2	Data Analyst	opportunity to work in international environment have learned digital tools office located near bts and mrt
3	Data Scientist	data analysis role tableau power bi python sql
4	Data Analyst	experiences in software development strong analytical skills attention to detail working hybrid flexible working hour
5	Data Analyst	microsoft powerbi dashboard etl sql sas r python capacity planning and performance monitoring
6	Data Analyst	out of stock data analyst excellent of analytic skill excellent for using power query python sql
7	Data Analyst	data analysis data visualization good in english communication
8	Data Analyst	analyze monitor and measure department activities provide financial analysis information manage project as project manager
9	Data Analyst	experience as data analyst understand in sql or database and power bi experience as data studio
10	Data Analyst	data management data analyst business analyst

รูปที่ 3.10 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังปรับตัวอักษรเป็นตัวพิมพ์เล็ก

จากรูปที่ 3.9 และ 3.10 เป็นการปรับตัวอักษรเป็นตัวพิมพ์เล็กทั้งหมดเนื่องจากคอมพิวเตอร์จะเข้าใจว่าตัวอักษรพิมพ์ใหญ่และพิมพ์เล็กแตกต่างกัน เช่น ‘A’ กับ ‘a’ เป็นต้น จากชุดคำสั่ง nltk

3.3.5 การตัดส่วนขยายของคำ

index	job_title	description
0	Data Analyst	business intelligence analyst ecommerce market insights data
1	Data Analyst	data analysis analysis data analysis manager
2	Data Analyst	opportunity to work in international environment have learned digital tools office located near bts and mrt
3	Data Scientist	data analysis role tableau power bi python sql
4	Data Analyst	experiences in software development strong analytical skills attention to detail working hybrid flexible working hour
5	Data Analyst	microsoft powerbi dashboard etl sql sas r python capacity planning and performance monitoring
6	Data Analyst	out of stock data analyst excellent of analytic skill excellent for using power query python sql
7	Data Analyst	data analysis data visualization good in english communication
8	Data Analyst	analyze monitor and measure department activities provide financial analysis information manage project as project manager
9	Data Analyst	experience as data analyst understand in sql or database and power bi experience as data studio
10	Data Analyst	data management data analyst business analyst

รูปที่ 3.11 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนตัดส่วนขยายของคำ

index	job_title	description
0	Data Analyst	business intelligence analyst ecommerce market insight data
1	Data Analyst	data analysis analysis data analysis manager
2	Data Analyst	opportunity to work in international environment have learned digital tool office located near bts and mrt
3	Data Scientist	data analysis role tableau power bi python sql
4	Data Analyst	experience in software development strong analytical skill attention to detail working hybrid flexible working hour
5	Data Analyst	microsoft powerbi dashboard etl sql sa r python capacity planning and performance monitoring
6	Data Analyst	out of stock data analyst excellent of analytic skill excellent for using power query python sql
7	Data Analyst	data analysis data visualization good in english communication
8	Data Analyst	analyze monitor and measure department activity provide financial analysis information manage project a project manager
9	Data Analyst	experience a data analyst understand in sql or database and power bi experience a data studio
10	Data Analyst	data management data analyst business analyst

รูปที่ 3.12 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังตัดส่วนส่วนขยายของคำ

จากรูปที่ 3.11 และ 3.12 เป็นการตัดส่วนขยายของคำให้เหลือแต่รูปฟอร์มพื้นฐานเช่น s หรือ es จากชุดคำสั่ง nltk

3.3.6 การลบคำที่ไม่สื่อความหมาย

index	job_title	description
0	Data Analyst	business intelligence analyst ecommerce market insight data
1	Data Analyst	data analysis analysis data analysis manager
2	Data Analyst	opportunity to work in international environment have learned digital tool office located near bts and mrt
3	Data Scientist	data analysis role tableau power bi python sql
4	Data Analyst	experience in software development strong analytical skill attention to detail working hybrid flexible working hour
5	Data Analyst	microsoft powerbi dashboard etl sql sa r python capacity planning and performance monitoring
6	Data Analyst	out of stock data analyst excellent of analytic skill excellent for using power query python sql
7	Data Analyst	data analysis data visualization good in english communication
8	Data Analyst	analyze monitor and measure department activity provide financial analysis information manage project a project manager
9	Data Analyst	experience a data analyst understand in sql or database and power bi experience a data studio
10	Data Analyst	data management data analyst business analyst

รูปที่ 3.13 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานก่อนลบคำที่ไม่สื่อความหมาย

index	job_title	description
0	Data Analyst	business intelligence analyst ecommerce market insight data
1	Data Analyst	data analysis analysis data analysis manager
2	Data Analyst	opportunity work international environment learned digital tool office located near bts mrt
3	Data Scientist	data analysis role tableau power bi python sql
4	Data Analyst	experience software development strong analytical skill attention detail working hybrid flexible working hour
5	Data Analyst	microsoft powerbi dashboard etl sql sa r python capacity planning performance monitoring
6	Data Analyst	stock data analyst excellent analytic skill excellent using power query python sql
7	Data Analyst	data analysis data visualization good english communication
8	Data Analyst	analyze monitor measure department activity provide financial analysis information manage project project manager
9	Data Analyst	experience data analyst understand sql database power bi experience data studio
10	Data Analyst	data management data analyst business analyst

รูปที่ 3.14 ตัวอย่างชุดข้อมูลประกาศรับสมัครงานหลังลบคำที่ไม่สื่อความหมาย

จากรูปที่ 3.13 และ 3.14 เป็นการลบคำที่ไม่สื่อความหมายเช่น a, an และ the ด้วยคลังคำศัพท์จากชุดคำสั่ง nltk

3.3.7 การแปลงข้อความเป็นเวกเตอร์

```
analytical: 0.21693901497281193
business: 0.14674089211724642
applicant: 0.5528703265259215
proficient: 0.4101885940885718
improvement: 0.44247666905696853
command: 0.29979493661961887
excel: 0.3067238432188894
english: 0.22749508002366325
```

รูปที่ 3.15 การแปลงข้อความเป็นเวกเตอร์

จากรูปที่ 3.15 เป็นการแปลงข้อความเป็นเวกเตอร์เพื่อให้คอมพิวเตอร์สามารถนำไปคำนวณและสร้างแบบจำลองได้ต่อไป

3.4 การแบ่งข้อมูล

การแบ่งข้อมุลเพื่อทำการฝึกสอนแบบจำลองโดยการแบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกฝน 80% และชุดข้อมูลทดสอบ 20%

3.5 การสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

สร้างแบบจำลองโดยวิธีการทำเหมืองข้อความด้วยชุดโปรแกรม NLTK จากนั้นเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 2 วิธี คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน (SVM) และวิธีการถดถอยลอจิสติก (LR) และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีเพอร์เซ็ปตรอนหลายชั้น (MLP) และวิธีโครงข่ายประสาทแบบคอนโวลูชัน (CNN) โดยมีรายละเอียดดังต่อไปนี้ คือ ดำเนินการทำเหมืองข้อความด้วยชุดโปรแกรม NLTK และนำผลลัพธ์ที่ได้จากการทำเหมืองข้อความมาดำเนินการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (SVM) และวิธีการถดถอยลอจิสติก (LR) และแบบจำลองการเรียนรู้เชิงลึกด้วยขั้นตอนวิธีเพอร์เซ็ปตรอนหลายชั้น (MLP) และวิธีโครงข่ายประสาทแบบคอนโวลูชัน (CNN) ซึ่งผู้วิจัยได้ทำการแบ่งการทดลองออกเป็นการเรียนรู้ของเครื่อง 2 วิธี และการเรียนรู้เชิงลึก 2 วิธีรวมเป็น 4 วิธีดังต่อไปนี้

1. วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
2. วิธีการถดถอยลอจิสติก (Logistic Regression)
3. วิธีเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)
4. วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

3.6 การเปรียบเทียบประสิทธิภาพแบบจำลอง

หลังจากสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก จากนั้นทำการเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าคะแนนเอฟ1 (F1-Score) โดยพิจารณาจากค่าทั้ง 4 เหล่านี้ที่มีค่าสูงที่สุด

3.7 การวิเคราะห์ทักษะที่จำเป็นต่ออาชีพ

เมื่อได้แบบจำลองที่ดีที่สุดจากหัวข้อ 3.6 ซึ่งแบบจำลองได้แสดงค่าความน่าจะเป็นของทักษะในแต่ละอาชีพโดยเรียงลำดับค่าความน่าจะเป็นจากมากไปน้อย จากนั้นทำการสรุปและวิเคราะห์ผลทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลโดยแบ่งเป็นทักษะทางเทคนิค (Technical Skill) และจรณทักษะ (Soft Skill) โดยใช้ค่าความน่าจะเป็นเรียงลำดับจากมากไปน้อยโดยใช้เครื่องมือ TextBlob ซึ่งเป็นเครื่องมือที่ใช้ในการระบุหน้าที่ของคำในประโยค เช่น นาม กริยา คำคุณศัพท์ เป็นต้น ในงานวิจัยนี้ได้ระบุจรณทักษะเป็นคำคุณศัพท์เนื่องจากจรณทักษะส่วนใหญ่เป็นคำคุณศัพท์ ในส่วนของทักษะทางเทคนิคทางผู้วิจัยได้ระบุทักษะทางเทคนิค (Hiranrat and Harncharnchai, 2018) ดังต่อไปนี้

Technical_Skills =
['python', 'c', 'r', 'c++', 'java','hadoop','scala','flask','pandas','spark','scikit -learn', 'numpy', 'php', 'sql', 'mysql', 'css', 'mongdb', 'nltk', 'fastai' , 'keras','pytorch', 'tensorflow', 'linux', 'Ruby', 'JavaScript', 'django', 'react', 'ai', 'ui', 'reactjs', 'tableau', 'nosql', 'big data', 'cloud', 'powerbi', 'power bi', 'aws', 'qlik', 'excel', 'azure', 'hive', 'sap', 'spss', 'mathematical', 'digitalocean', 'etl', 'nlp','pandas','pyspark']

3.8 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยประกอบด้วยชุดคำสั่ง ฮาร์ดแวร์ ระบบปฏิบัติการ ซอฟต์แวร์ และโปรแกรมประยุกต์ดังต่อไปนี้

3.8.1 ชุดคำสั่งที่ใช้ในการวิจัย (Library)

ชุดคำสั่ง	คำอธิบาย
numpy	NumPy เป็นชุดคำสั่งพื้นฐานที่ใช้คำนวณทางคณิตศาสตร์ด้วยภาษา Python สามารถคำนวณหรือดำเนินการทางตรรกะใน Array หลายมิติ หรือ Matrix ได้อย่างรวดเร็ว
pandas	pandas คือหนึ่งในชุดคำสั่งสำคัญของภาษา Python มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ สามารถใช้การเขียนโค้ด เพื่อปรับแต่ง หรือเชื่อมต่อกับโปรแกรมอื่นๆ เพื่อดู Data set
nltk	Natural Language Toolkit หรือเรียกว่า nltk เป็นการประมวลภาษาธรรมชาติเป็นส่วนหนึ่งของปัญญาประดิษฐ์และภาษาศาสตร์เพื่อให้คอมพิวเตอร์สามารถตีความและเข้าใจภาษามนุษย์ได้
matplotlib	เป็นชุดคำสั่งของภาษา Python เพื่อใช้ในการสร้างหรือแสดงผล Data Visualization ช่วยในการสร้างแผนภูมิและกราฟต่างๆ เพื่อช่วยในการวิเคราะห์ที่ทำให้ดูง่ายขึ้น
sklearn	Scikit-learn หรือ sklearn เป็นชุดคำสั่งของภาษา Python ใช้สำหรับการเรียนรู้ของเครื่องและสร้างตัวแบบทางสถิติการจำแนกประเภท เช่น Regression, Classification และ Clustering

3.8.2 ฮาร์ดแวร์

- หน่วยประมวลผลกลาง Intel i7-10750H
- RAM 16.0 GB
- หน่วยประมวลผลกราฟิก NVIDIA GeForce RTX 2060

3.8.3 ระบบปฏิบัติการ ซอฟต์แวร์ และโปรแกรมประยุกต์

- ระบบปฏิบัติการ Windows 10 Home Single Language 64-bit
- Google Colaboratory

บทที่ 4
ผลการวิจัยและการอภิปรายผล

งานวิจัยนี้คือการนำข้อมูลมาวิเคราะห์ทักษะทางเทคนิคและจรณทักษะที่จำเป็นของอาชีพ 3 อาชีพคือ นักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล ขั้นตอนดำเนินงานวิจัยแบ่งออกเป็น 5 ขั้นตอนหลัก ได้แก่ การเก็บรวบรวมข้อมูล การสำรวจข้อมูล การทำความสะอาดข้อมูล การเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องกับแบบจำลองการเรียนรู้เชิงลึก และการวิเคราะห์ทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล ทำการฝึกสอนแบบจำลองโดยการแบ่งข้อมูลออกเป็น 2 ชุดคือ ชุดข้อมูลฝึกฝน 80% และชุดข้อมูลทดสอบ 20% ซึ่งผู้วิจัยได้ทำการแบ่งการทดลองออกเป็นการเรียนรู้ของเครื่อง 2 วิธี และการเรียนรู้เชิงลึก 2 วิธี รวมเป็น 4 วิธีดังต่อไปนี้

- วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
- วิธีการถดถอยลอจิสติก (Logistic Regression)
- วิธีเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)
- วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

การทดสอบประสิทธิภาพแบบจำลองจะทดสอบด้วยค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าคะแนนเอฟ1 (F1-Score)

เมื่อได้แบบจำลองที่ดีที่สุดจากแบบจำลองทั้ง 4 วิธี จากนั้นจะทำการสรุปและวิเคราะห์ผลทักษะที่จำเป็นต่ออาชีพของนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล โดยแบ่งเป็นทักษะทางเทคนิค และจรณทักษะ

4.1 ผลการเก็บรวบรวมข้อมูล

ผู้วิจัยได้ทำการเก็บรวบรวมชุดข้อมูลมาจากเว็บไซต์ ซึ่งเป็นแหล่งรวบรวมประกาศรับสมัครงานในประเทศไทย จำนวน 761 ชุด โดยชุดข้อมูลดังกล่าวเปิดให้เข้าใช้งานในรูปแบบของสาธารณะภายในชุดข้อมูลประกอบด้วยตัวแปรอิสระคือ รายละเอียดของงาน (Description) และตัวแปรตามคือชื่อตำแหน่งงาน (Job Title) คืออาชีพของนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล โดยชุดข้อมูลทั้งหมดเป็นภาษาอังกฤษ ตัวอย่างชุดข้อมูลดังตารางที่ 4.1 ดังกล่าวไว้ในบทที่ 3 จากนั้นนำชุดข้อมูลรายละเอียดงานจากเว็บไซต์ www.jobsdb.com เข้าโปรแกรม Colab โดยใช้ชุดคำสั่ง pandas ในการอ่านไฟล์นามสกุล .csv

ตารางที่ 4.1 ตัวอย่างชุดข้อมูล

ชื่อตำแหน่งงาน	รายละเอียดของงาน
Data Scientist	Python Machine learning NLP
Data Engineer	Data Lake, Data Warehouse, Big data tools, NoSQL
Data Analyst	Data Analyst 3+ years of experience in relevant fields Advance in MS Excel

4.2 ผลการสำรวจข้อมูล

ผู้วิจัยได้ทำการสำรวจข้อมูลประกาศรับสมัครงานจำนวนทั้งหมด 761 ชุดที่เก็บมาได้จากเว็บไซต์ www.jobsdb.com ดังเขียนในรูปที่ 4.1

รูปที่ 4.1 จำนวนข้อมูลประกาศรับสมัครงานที่เก็บรวบรวม จากเว็บไซต์ www.jobsdb.com

จากรูปที่ 4.1 จะพบว่าตำแหน่งงานทั้งหมด 761 ตำแหน่งประกอบไปด้วยตำแหน่งงาน Data Analyst ทั้งหมด 571 ตำแหน่ง ตำแหน่งงาน Data Scientist ทั้งหมด 70 ตำแหน่งและตำแหน่งงาน Data Engineer ทั้งหมด 120 ตำแหน่ง

ตารางที่ 4.2 จำนวนการประกาศรับสมัครงานของผู้ที่มีอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

อาชีพ	นักวิเคราะห์ข้อมูล	นักวิทยาศาสตร์ข้อมูล	วิศวกรข้อมูล
ความถี่	571	70	120

จากตารางที่ 4.2 พบว่าการประกาศรับสมัครงานของผู้ที่มีอาชีพนักวิเคราะห์ข้อมูลมีจำนวนสูงที่สุดคือ 571 ตำแหน่ง รองลงมาคืออาชีพวิศวกรข้อมูลจำนวน 120 ตำแหน่ง และอาชีพนักวิทยาศาสตร์ข้อมูล 70 ตำแหน่ง

4.2.1 ผลการสำรวจความนิยม

จากการสำรวจความนิยมของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลนำมาทดสอบสัดส่วนโดยตั้งสมมุติฐานดังต่อไปนี้

H_0 : สัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลไม่แตกต่างกัน

H_1 : สัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลแตกต่างกัน

Chi-Square Test

Frequencies

job_title

	Observed N	Expected N	Residual
Data Analyst	571	253.7	317.3
Data Scientist	120	253.7	-133.7
Data Engineer	70	253.7	-183.7
Total	761		

Test Statistics

	job_title
Chi-Square	600.397[a]
df	2
Asymp. Sig.	<.001

รูปที่ 4.2 ผลจากการทดสอบสัดส่วนของการประกาศรับสมัครงาน

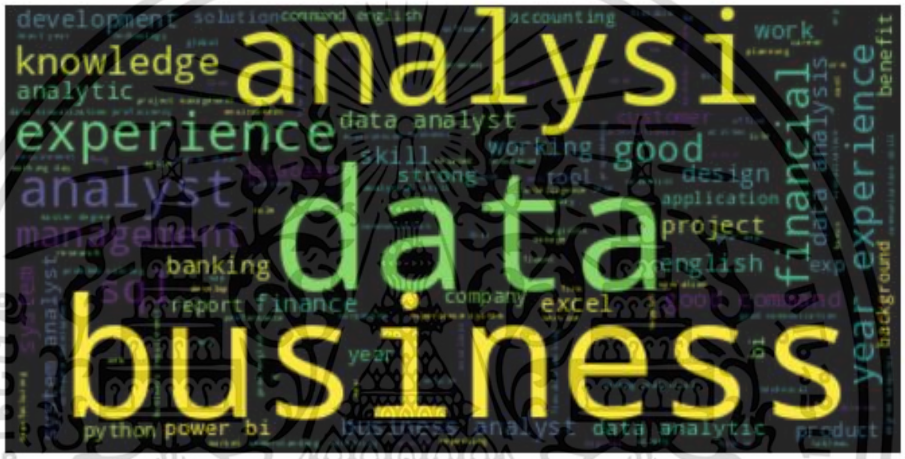
จากรูปที่ 4.2 ผลการทดสอบพบว่าค่าไคกำลังสอง (Chi-Square) = 600.397 และค่า P-Value < 0.001 ซึ่งมีค่าน้อยกว่าค่าแอลฟ่า 0.05 โดยตกอยู่ในบริเวณปฏิเสธ H_0 ดังนั้นสัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลมีความแตกต่างกัน

อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 โดยที่อาชีพนักวิเคราะห์ข้อมูลมีข้อมูลประกาศรับสมัครงานจำนวนมากที่สุด

4.2.2 ผลการสำรวจคำที่พบบ่อย

1) คำที่ปรากฏในรายละเอียดงานของอาชีพนักวิเคราะห์ข้อมูล

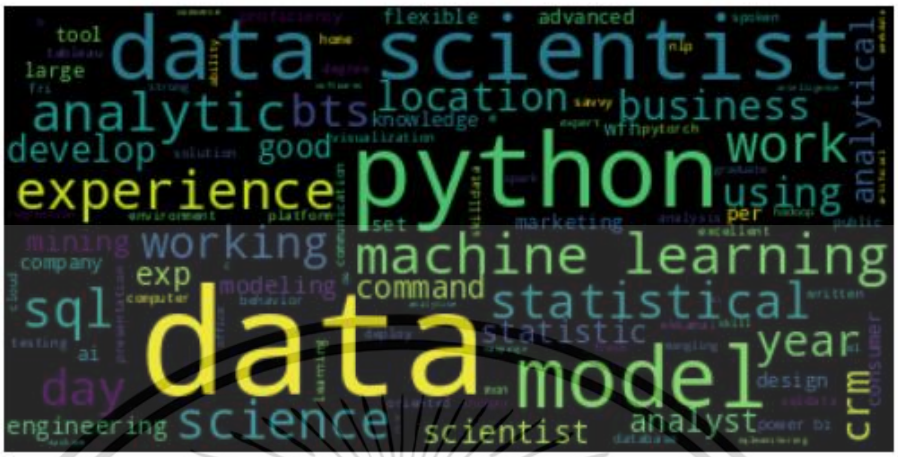
ผู้วิจัยได้ใช้ชุดคำสั่ง “WorldCloud” ซึ่งเป็นชุดคำสั่งที่ใช้ในการแสดงผลคำที่ปรากฏอยู่ในข้อมูลประกาศรับสมัครงานโดยคำที่มีขนาดใหญ่กว่าหมายความว่าปรากฏความถี่ของคำเหล่านั้นจากมากไปน้อยตามลำดับ ซึ่งการสำรวจในครั้งนี้จะนำไปประกอบการกำจัดคำที่ไม่สื่อความหมายของขั้นตอนทำความสะอาดข้อมูลในลำดับถัดไป



รูปที่ 4.3 คำที่ปรากฏในรายละเอียดงานของอาชีพนักวิเคราะห์ข้อมูล

จากรูปที่ 4.3 แสดงผลของคำในประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล ในรูปแบบของการแสดงความถี่ของคำ โดยคำที่มีขนาดใหญ่จะเป็นคำที่ปรากฏบ่อยในประกาศรับสมัครงานเช่น คำว่า analysis, business และ data เป็นต้น

2) คำที่ปรากฏในรายละเอียดงานของอาชีพนักวิทยาศาสตร์ข้อมูล



รูปที่ 4.4 คำที่ปรากฏในรายละเอียดงานของอาชีพนักวิทยาศาสตร์ข้อมูล

จากรูปที่ 4.4 แสดงผลของคำในประกาศรับสมัครงานของอาชีพนักวิทยาศาสตร์ข้อมูลในรูปแบบของการแสดงความถี่ของคำ โดยคำที่มีขนาดใหญ่จะเป็นคำที่ปรากฏบ่อยในประกาศรับสมัครเช่นคำว่า data, data scientist และ python เป็นต้น

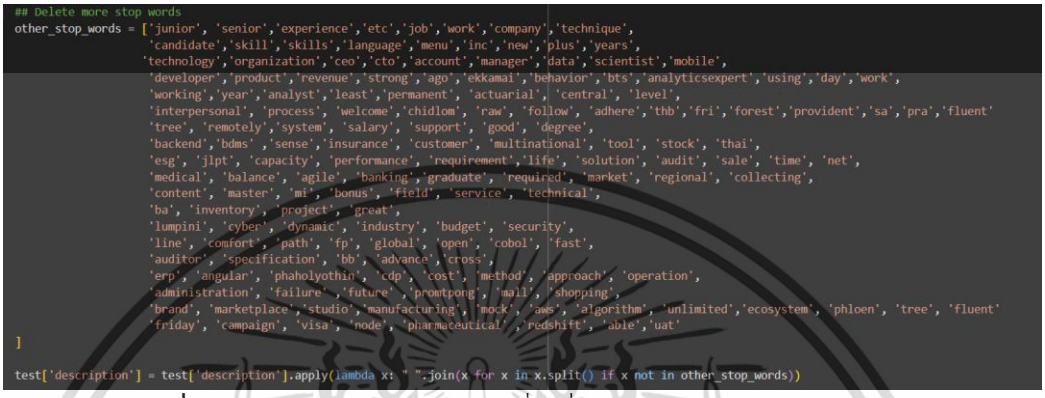
3) คำที่ปรากฏในรายละเอียดงานของอาชีพวิศวกรข้อมูล

รูปที่ 4.5 คำที่ปรากฏในรายละเอียดงานของอาชีพวิศวกรข้อมูล

จากรูปที่ 4.5 แสดงผลลัพธ์ของคำในประกาศรับสมัครงานของอาชีพวิศวกรข้อมูลในรูปแบบของการแสดงความถี่ของคำ โดยคำที่มีขนาดใหญ่จะเป็นคำที่ปรากฏบ่อยในประกาศรับสมัครงานเช่นคำว่า data, python และ experience เป็นต้น

4) กำหนดคำที่ไม่สื่อความหมาย

หลังจากที่ได้ผลการแสดงผลความถี่ของคำที่ปรากฏอยู่ในข้อมูลประกาศรับสมัครงานด้วยชุดคำสั่ง “WorldCloud” ทางผู้วิจัยได้กำหนดคำที่ไม่สื่อความหมายเพิ่มเติมเนื่องจากพบคำที่ไม่เกี่ยวข้องกับทักษะ



```
## Delete more stop words
other_stop_words = ['junior', 'senior','experience','etc','job','work','company','technique',
                    'candidate','skill','skills','language','menu','inc','new','plus','years',
                   'technology','organization','ceo','cto','account','manager','data','scientist','mobile',
                    'developer','product','revenue','strong','ago','ekkamai','behavior','bts','analyticsexpert','using','day','work',
                    'working','year','analyst','least','permanent', 'actuarial', 'central', 'level',
                    'interpersonal', 'process', 'welcome','chidlom', 'raw', 'follow', 'adhere','thb','fri','forest','provident','sa','pra','fluent'
                    'tree', 'remotely','system', 'salary', 'support', 'good', 'degree',
                    'backend','bdms' ,'sense','insurance', 'customer', 'multinational', 'tool', 'stock', 'thai',
                    'esg', 'jlpt', 'capacity', 'performance', 'requirement','life', 'solution', 'audit', 'sale', 'time', 'net',
                    'medical', 'balance', 'agile', 'banking', 'graduate', 'required', 'market', 'regional', 'collecting',
                    'content', 'master', 'mi', 'bonus', 'field', 'service', 'technical',
                    'ba', 'inventory', 'project', 'great',
                    'lumpini', 'cyber', 'dynamic', 'industry', 'budget', 'security',
                    'line', 'comfort', 'path', 'fp', 'global', 'open', 'cobol', 'fast',
                    'auditor', 'specification', 'bb', 'advance','cross',
                    'erp', 'angular', 'phaholyothin', 'cdp', 'cost', 'method', 'approach', 'operation',
                     'administration', 'failure', 'future', 'promtpong', 'mall', 'shopping',
                    'brand', 'marketplace','studio','manufacturing', 'mock', 'aws', 'algorithm', 'unlimited','ecosystem', 'phloen', 'tree', 'fluent'
                    'friday', 'campaign', 'visa', 'node', 'pharmaceutical', 'redshift', 'able','uat'
]

test['description'] = test['description'].apply(lambda x: " ".join(x for x in x.split() if x not in other_stop_words))
```

รูปที่ 4.6 การกำหนดและกำจัดคำที่ไม่สื่อความหมายด้วยภาษาไพทอน

จากรูปที่ 4.6 แสดงถึงขั้นตอนการกำหนดและกำจัดคำที่ไม่สื่อความหมายด้วยภาษาไพทอนที่ผู้วิจัยได้กำหนดเพิ่มเติม

ตารางที่ 4.3 คำที่ไม่สื่อความหมายที่กำหนดโดยผู้วิจัย

คำที่ไม่สื่อความหมาย
'master', 'mi', 'bonus', 'field', 'service', 'technical', 'ba', 'inventory', 'project', 'great', 'lumpini', 'cyber', 'dynamic', 'industry', 'budget', 'security','line', 'comfort', 'path', 'fp', 'global', 'open', 'cobol', 'fast', 'auditor', 'specification', 'bb', 'advance','cross', 'erp', 'angular', 'phaholyothin', 'cdp', 'cost', 'method', 'approach', 'operation', 'administration', 'failure' ,'future' ,'promtpong', 'mall', 'shopping','brand', 'marketplace','studio','manufacturing', 'mock', 'aws', 'algorithm', 'unlimited', 'ecosystem', 'phloen', 'tree', 'fluent', 'friday', 'campaign', 'visa', 'node', 'pharmaceutical' ,'redshift', 'able','uat', 'junior','senior','experience','etc','job','work','company','technique','candidate','skill', 'skills','language','menu','inc','new','plus','years','technology','organization','ceo','cto', 'account','manager','data','scientist','mobile','developer','product','revenue','strong', 'ago','ekkamai','behavior','bts','analyticsexpert','using','day','work','working','year', 'analyst', 'least', 'permanent', 'actuarial', 'central', 'level', 'interpersonal', 'process', 'welcome', 'chidlom', 'raw', 'follow', 'adhere','thb','fri','forest','provident','sa', 'pra','fluent', 'tree', 'remotely','system', 'salary', 'support', 'good', 'degree','backend', 'bdms', 'sense','insurance', 'customer', 'multinational', 'tool', 'stock', 'thai', 'esg', 'jlpt', 'capacity', 'performance', 'requirement', 'life', 'solution', 'audit', 'sale', 'time', 'net', 'medical', 'balance', 'agile', 'banking', 'graduate', 'required', 'market', 'regional', 'collecting', 'content',

จากตารางที่ 4.3 แสดงถึงคำที่ไม่สื่อความหมายที่กำหนดโดยผู้วิจัยหลังจากที่ได้ผลการแสดงผลความถี่ของคำที่ปรากฏอยู่ในข้อมูลประกาศรับสมัครงานด้วยชุดคำสั่ง “WorldCloud” โดยเลือกกำจัดคำที่ไม่สื่อความหมายเพิ่มเติม เช่น คำว่า bts, work และ year เป็นต้น เนื่องจากไม่ได้เป็นคำที่แสดงถึงทักษะ

4.3 ผลการทำความสะอาดข้อมูล

นำข้อมูลที่ได้จากการสำรวจข้อมูลมาแสดงผลดังได้ดังรูปที่ 4.7

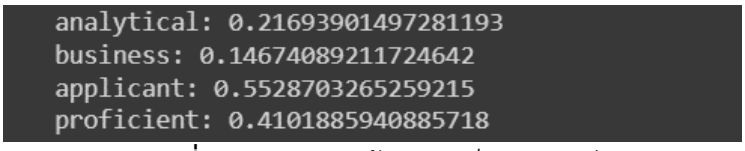
index	job_title	description
0	Data Analyst	Business Intelligence Analyst eCommerce Market Insights Data
1	Data Analyst	Data Analysis Analysis Data Analysis manager
2	Data Analyst	Opportunity to work in International environment. Have learned digital tools. Office located near BTS and MRT
3	Data Scientist	Data Analysis role Tableau , Power BI Python , SQL
4	Data Analyst	Experiences in software development Strong analytical skills, Attention to detail Working hybrid, Flexible working hour
5	Data Analyst	Microsoft PowerBI, Dashboard, ETL SQL, SAS, R, Python Capacity Planning and Performance Monitoring
6	Data Analyst	Out of Stock Data Analyst Excellent of Analytic Skill Excellent for using Power Query, Python, SQL
7	Data Analyst	Data Analysis Data Visualization Good in English Communication
8	Data Analyst	Analyze, monitor and measure Department activities Provide financial analysis information Manage project as Project manager
9	Data Analyst	Experience as Data Analyst Understand in SQL or database and Power BI Experience as Data Studio
10	Data Analyst	Data Management Data Analyst Business Analyst
11	Data Analyst	Bachelor's degree in Actuarial Sciences or Related Welcome new graduates Excellent in english communication
12	Data Analyst	At least 2 yrs working experience in data analysis Strong analytical skills Can-do attitude
13	Data Analyst	Experience of business data analysis (Marketing) Analyze marketing data and present the action plan Design, execute data for business & Implement
14	Data Analyst	Advanced SQL, VBA, Excel skills-Pivot table Experiences with ERP system Excellent in English (written and spoken)
15	Data Analyst	Knowledge of SQL Data analysis Credit modelling business
16	Data Analyst	Data Analysis Design and develop data structure Job security and development opportunity
17	Data Analyst	Data analysis Passionate about football Can do attitude
18	Data Analyst	Supply Chain Analyst/ Logistics Analyst SRM Performance Specialist Power BI query skill is a must
19	Data Analyst	Good Command in English Knowledge of data analytic & visualization Analytical & Problem-solving skill
20	Data Analyst	Data analysis Analytics Tools: Mainly Excel / SAS Credit product knowledge is preferred

รูปที่ 4.7 ชุดข้อมูลก่อนทำความสะอาด

หลังจากนั้นทำความสะอาดข้อมูลด้วย 7 ขั้นตอนดังต่อไปนี้คือ การลบช่องว่างและเครื่องหมายวรรคตอน การลบตัวเลข การลบตัวอักษรภาษาไทย การปรับตัวอักษรเป็นตัวพิมพ์เล็ก การตัดส่วนขยายของคำและการลบคำที่ไม่สื่อความหมายทั้งจากคลังคำศัพท์ของชุดเครื่องมือ nltk คำที่ไม่สื่อความหมายที่กำหนดโดยผู้วิจัย และการแปลงข้อความให้อยู่ในรูปเวกเตอร์ได้ดังรูปที่ 4.8 และ 4.9

index	job_title	description
0	Data Analyst	business intelligence ecommerce insight
1	Data Analyst	analysis analysis analysis
2	Data Analyst	opportunity international environment learned digital office located near mrt
3	Data Scientist	analysis role tableau power bi ython sql
4	Data Analyst	software development analytical attention detail hybrid flexible hour
5	Data Analyst	microsoft powerbi dashboard etl sql r python planning monitoring
6	Data Analyst	excellent analytic excellent power query python sql
7	Data Analyst	analysis visualization english communication
8	Data Analyst	analyze monitor measure department activity provide financial analysis information manage
9	Data Analyst	understand sql database power bi
10	Data Analyst	management business
11	Data Analyst	bachelor science related excellent english communication
12	Data Analyst	yr analys s analytical attitude
13	Data Analyst	business analysis marketing analyze marketing present action plan design execute business mplement
14	Data Analyst	advanced sql vba excel pivot table excellent english written spoken
15	Data Analyst	knowledge sql analysis credit modelling business
16	Data Analyst	analysis design develop structure development opportunity
17	Data Analyst	analysis passionate football attitude
18	Data Analyst	supply chain logistics srm specialist power bi query must
19	Data Analyst	command english knowledge analytic visualization analytical problem solving
20	Data Analyst	analysis analytics mainly excel credit knowledge preferred

รูปที่ 4.8 ชุดข้อมูลหลังทำความสะอาด



รูปที่ 4.9 การแปลงข้อความเป็นเวกเตอร์

จากรูปที่ 4.9 เป็นการแปลงข้อความเป็นเวกเตอร์เพื่อให้คอมพิวเตอร์สามารถนำไปคำนวณได้ต่อไป

4.4 ผลการวิเคราะห์การเรียนรู้ของเครื่อง

4.4.1 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนในการ ทำนายว่าทักษะทางเทคนิคและจรณทักษะใดเหมาะกับอาชีพใด ได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1 ดังตารางที่ 4.4

4.4.2 วิธีการถดถอยลอจิสติก (Logistic Regression)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีการถดถอยลอจิสติกในการทำนายว่าทักษะทางเทคนิคและจรณทักษะใดเหมาะกับอาชีพใด ได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1 ดังตารางที่ 4.4

ตารางที่ 4.4 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าคะแนนเอฟ1
วิธีซัพพอร์ตเวกเตอร์แมชชีน	86.93%	86.18%	86.93%	85.99%
วิธีการถดถอยลอจิสติก	84.31%	83.62%	84.31%	80.93%

จากตารางที่ 4.4 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่องโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่นเท่ากับ 86.93% ค่าความเที่ยงเท่ากับ 86.18% ค่าเรียกคืนเท่ากับ 86.93% และค่าคะแนนเอฟ1เท่ากับ 85.99% และวิธีการถดถอยลอจิสติกพบว่าวิธีการถดถอยลอจิสติกมีค่าความแม่นเท่ากับ 84.31% ค่าความเที่ยงเท่ากับ 83.62% ค่าเรียกคืนเท่ากับ 84.31% และค่าคะแนนเอฟ1เท่ากับ 80.93%

4.5 ผลการวิเคราะห์การเรียนรู้เชิงลึก

4.5.1 วิธีเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)

จากการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีเพอร์เซ็ปตรอนหลายชั้นในการทำนายว่าทักษะทางเทคนิคและจรณทักษะใดเหมาะกับอาชีพใดได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1 ดังตารางที่ 4.5

4.5.2 วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

จากการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีเพอร์เซ็ปตรอนหลายชั้นในการทำนายว่าทักษะทางเทคนิคและจรณทักษะใดเหมาะกับอาชีพใด ได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1 ดังตารางที่ 4.5

ตารางที่ 4.5 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้เชิงลึก

การเรียนรู้เชิงลึก	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าคะแนนเอฟ1
วิธีเพอร์เซ็ปตรอนหลายชั้น	85.62%	85.37%	85.62%	84.78%
วิธีโครงข่ายประสาทแบบคอนโวลูชัน	83.66%	81.32%	83.66%	81.86%

จากตารางที่ 4.5 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้เชิงลึกโดยวิธีเพอร์เซ็ปตรอนหลายชั้นมีค่าความแม่นเท่ากับ 85.62% ค่าความเที่ยงเท่ากับ 85.37% ค่าเรียกคืนเท่ากับ 85.62% และค่าคะแนนเอฟ1เท่ากับ 84.78% และวิธีโครงข่ายประสาทแบบคอนโวลูชันมีค่าความแม่นเท่ากับ 83.66% ค่าความเที่ยงเท่ากับ 81.32% ค่าเรียกคืนเท่ากับ 83.66% และค่าคะแนนเอฟ1เท่ากับ 81.86%

4.6 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกดังแสดงในตารางที่ 4.6 ผลการทดสอบประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

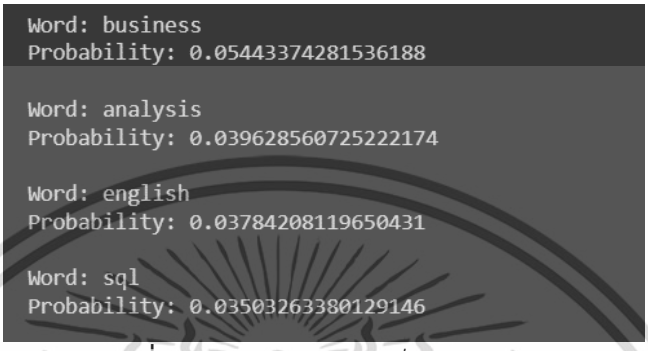
ตารางที่ 4.6 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

การเรียนรู้ของเครื่อง	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าคะแนนเอฟ1
ซัพพอร์ตเวกเตอร์แมชชีน	86.93%	86.18%	86.93%	85.99%
การถดถอยลอจิสติก	84.31%	83.62%	84.31%	80.93%
การเรียนรู้เชิงลึก	**ค่าความแม่น**	**ค่าความเที่ยง**	**ค่าเรียกคืน**	**ค่าคะแนนเอฟ1**
เพอร์เซ็ปตรอนหลายชั้น	85.62%	85.37%	85.62%	84.78%
โครงข่ายประสาทแบบคอนโวลูชัน	83.66%	81.32%	83.66%	81.86%

จากตารางที่ 4.6 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกพบว่าแบบจำลองที่ดีที่สุดคือแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนโดยมีค่าความแม่นเท่ากับ 86.93% ค่าความเที่ยงเท่ากับ 86.18% ค่าเรียกคืนเท่ากับ 86.93% และค่าคะแนนเอฟ1เท่ากับ 85.99% รองลงมาคือแบบจำลองเพอร์เซ็ปตรอน หลายชั้นโดยมีค่าความแม่นเท่ากับ 85.62% ค่าความเที่ยงเท่ากับ 85.37% ค่าเรียกคืนเท่ากับ 85.62% และค่าคะแนนเอฟ1เท่ากับ 84.78%

4.7 การวิเคราะห์ทักษะที่จำเป็นต่ออาชีพ

ผลการการศึกษาทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกร โดยเรียงลำดับค่าความน่าจะเป็นจากมากไปน้อย ได้ผลลัพธ์ดังต่อไปนี้



รูปที่ 4.10 ค่าความน่าจะเป็นของทักษะ

จากรูปที่ 4.10 แสดงค่าความน่าจะเป็นของทักษะทางเทคนิคและจรณะทักษะ เช่น business มีค่าความน่าจะเป็น 0.0544 analysis มีค่าความน่าจะเป็น 0.0396 เป็นต้น

ตารางที่ 4.7 ทักษะทางเทคนิคและจรณทักษะจากแบบจำลองการเรียนรู้ของเครื่องวิธีซัพพอร์ตเวกเตอร์แมชชีน

ชื่อตำแหน่งงาน	ทักษะ	รายละเอียดทักษะ
นักวิเคราะห์ข้อมูล	ทักษะทางเทคนิค	ภาษา SQL ภาษา PHP โปรแกรม Excel การวิเคราะห์ข้อมูล
	จรณทักษะ	การวางแผน ภาษาอังกฤษ ความรู้ทางธุรกิจ
นักวิทยาศาสตร์ข้อมูล	ทักษะทางเทคนิค	สถิติ ภาษา Python การสร้างแบบจำลอง การประมวลผลภาษาธรรมชาติ เครื่องมือ Digitalocean คณิตศาสตร์ เครื่องมือ Pytorch การเรียนรู้ของเครื่อง เครื่องมือ Hive
	จรณทักษะ	ความยืดหยุ่น การนำเสนอ
วิศวกรข้อมูล	ทักษะทางเทคนิค	โปรแกรม Tableau โปรแกรม Spark ภาษา Nosql โปรแกรม Hadoop โปรแกรม MongoDB การทำ ETL โปรแกรม Flask เครื่องมือ Apache ภาษา Python
	จรณทักษะ	การทำงานแบบ Hybrid

จากตารางที่ 4.7 แสดงทักษะทางเทคนิคและจรณทักษะจากแบบจำลองการเรียนรู้ของเครื่องวิธีซัพพอร์ตเวกเตอร์แมชชีน เช่น อาชีพนักวิเคราะห์ข้อมูล ทักษะทางเทคนิคที่จำเป็น เช่น ภาษา SQL ภาษา PHP โปรแกรม Excel และการวิเคราะห์ข้อมูล เป็นต้น ส่วนจรณทักษะที่จำเป็น เช่น การวางแผน ภาษาอังกฤษ และความรู้ทางธุรกิจ เป็นต้น

4.8 อภิปรายผลการวิจัย

ในการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 2 วิธีคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีการถดถอยลอจิสติก และแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีเพอร์เซ็ปตรอนหลายชั้น และวิธีโครงข่ายประสาทแบบคอนโวลูชัน ผลการศึกษาพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1สูงที่สุดเท่ากับ 86.18%, 86.93% และ 85.99% ตามลำดับ รองลงมาคือวิธีเพอร์เซ็ปตรอนหลายชั้นมีค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1เท่ากับ 85.37%, 85.62% และ 84.78% ตามลำดับ ซึ่งสอดคล้องกับงานวิจัยของ Fareri et al (2021) ที่ได้พบว่าวิธีซัพพอร์ตเวคเตอร์แมชชีนให้ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1สูงที่สุดคือ 68.1%, 77.8% และ 72.6% ตามลำดับ

ส่วนการวิเคราะห์ทักษะทางเทคนิคและจรณทักษะ ทักษะทางเทคนิค ผลการศึกษาพบว่าทักษะที่จำเป็นต่อสายงานอาชีพที่เกี่ยวกับข้อมูล คือ ภาษา SQL ภาษา PHP โปรแกรม Excel การวิเคราะห์ข้อมูล สถิติ ภาษาไพทอน การสร้างแบบจำลอง การประมวลผลภาษาธรรมชาติ เครื่องมือ Digitalocean คณิตศาสตร์ เครื่องมือ Pytorch การเรียนรู้ของเครื่อง เครื่องมือ Hive โปรแกรม Tableau โปรแกรม Spark ภาษา Nosql โปรแกรม Hadoop โปรแกรม MongoDB การทำ ETL โปรแกรม Flask และเครื่องมือ Apache ซึ่งสอดคล้องกับงานวิจัยของ Shirani (2019) ที่ได้พบว่าทักษะทางเทคนิคที่จำเป็นต่อสายงานการวิเคราะห์ข้อมูลคือ การวิเคราะห์ทางสถิติ การสร้างแบบจำลองการทำนาย การทดสอบ การวิจัยเชิงสาเหตุ ภาษาไพทอน การใช้แพ็กเกจ Pandas แพ็กเกจ NumPy และแพ็กเกจอื่นๆ ซอฟต์แวร์ Apache Spark ซอฟต์แวร์ Hadoop การเรียนรู้ของเครื่อง ภาษาสกาลา การทำให้เห็นได้ โปรแกรม MATLAB ฐานข้อมูลเชิงสัมพันธ์ การเรียนรู้เชิงลึกการหาค่าเหมาะที่สุด และการสร้างแบบจำลองมิติ ส่วนจรณทักษะ ผลการศึกษาพบว่าจรณะทักษะที่จำเป็นต่อสายงานอาชีพที่เกี่ยวกับข้อมูลคือ การวางแผน ภาษาอังกฤษ ความรู้ทางธุรกิจ ความยืดหยุ่น การนำเสนอ การทำงานแบบ Hybrid ซึ่งผลการทดลองสอดคล้องกับงานวิจัยของ Shirani (2019) ที่ได้พบว่าจรณะทักษะที่จำเป็นต่ออาชีพคือ การสื่อสาร มุมมองระดับโลก การวิจัย การทำงานเป็นทีม การแก้ปัญหา และการทำงานร่วมกับผู้อื่น

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาการวิเคราะห์ทักษะทางเทคนิคและจรณทักษะของสายงานอาชีพข้อมูลด้วยวิธีการทำเหมืองข้อความในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ซึ่งมีวัตถุประสงค์เพื่อเปรียบเทียบสัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล เพื่อศึกษาการทำเหมืองข้อความและเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและแบบจำลองการเรียนรู้เชิงลึก และเพื่อศึกษาเชิงลึกเกี่ยวกับทักษะที่จำเป็นของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูลและวิศวกรข้อมูล โดยแบ่งทักษะออกเป็น 2 ประเภทคือ ทักษะทางเทคนิค และจรณทักษะ การศึกษาครั้งนี้ได้นำชุดข้อมูลมาจากเว็บไซต์ www.jobsdb.com โดยจะมีตัวแปรที่ใช้ทั้งหมด 2 ตัวแปรโดยแบ่งเป็นตัวแปรอิสระคือรายละเอียดของงานและตัวแปรตามคือชื่อตำแหน่งงาน จากข้อมูลประกาศรับสมัครงานทั้งหมด 761 ชุด ทำการจัดการจัดเตรียมข้อมูลให้อยู่ในรูปที่คอมพิวเตอร์สามารถนำไปประมวลผลต่อได้ด้วยการแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ แล้วจึงนำมาสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกเพื่อเปรียบเทียบประสิทธิภาพการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการทำนายทักษะทางเทคนิคและจรณทักษะของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูลและวิศวกรข้อมูล สรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 การสำรวจความนิยมของอาชีพ

จากการสำรวจความนิยมของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล พบว่าสัดส่วนของการประกาศรับสมัครงานของอาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลมีความแตกต่างกันอย่างมีนัยสำคัญ โดยที่อาชีพนักวิเคราะห์ข้อมูลมีข้อมูลประกาศรับสมัครงานจำนวนมากที่สุด

5.1.2 แบบจำลองการเรียนรู้ของเครื่อง

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนนเอฟ1 ดังนี้ 86.93% 86.18% และ 86.93% และ 85.99% ตามลำดับ

5.1.3 แบบจำลองการเรียนรู้เชิงลึก

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูล นักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูลพบว่าวิธีเพอร์เซ็ปตรอนหลายชั้นมีประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าคะแนน เอฟ1 สูงที่สุดคือ 85.62%, 85.37%, 85.62% และ 84.78% ตามลำดับ

จากการเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก จะสรุปได้ว่าการเรียนรู้ของเครื่องมีประสิทธิภาพการทำนายได้ดีกว่าแบบจำลองการเรียนรู้เชิงลึก เนื่องจากวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าคะแนนเอฟ1 สูงที่สุดคือ 86.93%, 86.18%, 86.93% และ 85.99% ตามลำดับ

5.1.4 การวิเคราะห์ทักษะทางเทคนิคและจรณทักษะ

1) ทักษะทางเทคนิค ผลการศึกษาพบว่าอาชีพนักวิเคราะห์ข้อมูล ทักษะที่จำเป็นคือ ภาษา SQL ภาษา PHP โปรแกรม Excel การวิเคราะห์ข้อมูล อาชีพนักวิทยาศาสตร์ข้อมูล ทักษะที่จำเป็นคือ สถิติ ภาษา Python การสร้างแบบจำลอง การประมวลผลภาษาธรรมชาติ เครื่องมือ Digitalocean คณิตศาสตร์ เครื่องมือ Pytorch การเรียนรู้ของเครื่อง เครื่องมือ Hive และอาชีพวิศวกรข้อมูล ทักษะที่จำเป็นคือ โปรแกรม Tableau โปรแกรม Spark ภาษา Nosql โปรแกรม Hadoop โปรแกรม MongoDB การทำ ETL โปรแกรม Flask และเครื่องมือ Apache

2) จรณทักษะ ผลการศึกษาพบว่าจรณะทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูลคือการวางแผน ภาษาอังกฤษ จรณะทักษะที่จำเป็นต่ออาชีพนักวิทยาศาสตร์ข้อมูลคือความรู้ทางธุรกิจ ความยืดหยุ่น การนำเสนอ และจรณะทักษะที่จำเป็นต่ออาชีพวิศวกรข้อมูลคือการทำงานแบบ Hybrid

5.2 ข้อเสนอแนะ

5.2.1 ข้อเสนอแนะ

1) เนื่องจากใช้ชุดข้อมูลจากเว็บไซต์ www.jobsdb.com จากแหล่งเดียว ซึ่งข้อมูลมีจำนวนจำกัด ควรหาข้อมูลจากเว็บไซต์อื่นๆ เพื่อทำให้แบบจำลองได้เรียนรู้กับชุดข้อมูลที่หลากหลายมากขึ้นและจำนวนที่เยอะขึ้น ทำให้ประสิทธิภาพในการทำนายอาจจะดีขึ้นและได้ชุดข้อมูลฝึกฝนที่หลากหลายมากยิ่งขึ้น

2) ควรติดตามทักษะหรือเครื่องมือที่ใช้ในสายงานอาชีพข้อมูลใหม่ๆ อยู่เสมอ เนื่องจากปัจจุบันมีเครื่องมือที่หลากหลายและถูกพัฒนาเพื่อให้ใช้กับงานข้อมูลได้ดียิ่งขึ้น เพื่อนำมาวิเคราะห์ทักษะที่จำเป็นได้ทันต่อเหตุการณ์ปัจจุบัน

3) ควรใช้เทคนิคในการทำให้ข้อมูลมีความสมดุลกันระหว่างตัวแปรตาม เพื่อลดความเอนเอียงของแบบจำลอง

5.2.2 ข้อจำกัด

1) เนื่องจากเวลาในการทำวิจัยที่จำกัด จึงไม่สามารถนำการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกวิธีอื่นมาลองทดสอบได้ทั้งหมดซึ่งอาจจะมีวิธีที่ให้ค่าประสิทธิภาพการทำนายที่ดีกว่า

2) การสร้างแบบจำลองบางวิธีจะต้องใช้เวลาในการประมวลผลและการทำนายที่นาน อุปกรณ์ที่ใช้ในการวิเคราะห์ควรมีประสิทธิภาพที่เหมาะกับชุดข้อมูล

เอกสารอ้างอิง

สายชล สินสมบูรณ์ทอง. 2560. **การทำเหมืองข้อมูล เล่ม 1: การค้นหาความรู้จากข้อมูล.** กรุงเทพฯ : จามจุรี โปรดักส์.

อรพิน ประวัติบริสุทธิ์. 2564. **Python สำหรับงาน Data Science Data Visualization และ Machine Learning.** กรุงเทพฯ : โปรวิชั่น.

กอบเกียรติ สระอุบล. 2565. **เรียนรู้ AI : Deep Learning ด้วย Python.** 1. กรุงเทพฯ : อินเตอร์มีเดีย.

ไกรศักดิ์ เกษร. 2564. **วิทยาศาสตร์ข้อมูล.** พิษณุโลก : ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ.

สายชล สินสมบูรณ์ทอง. 2563. สถิติไม่อิงพารามิเตอร์. พิมพ์ครั้งที่ 2 ฉบับปรับปรุง. กรุงเทพฯ : จามจุรีโปรดักส์.

Yaqoob, I. Hashem, I.A.T. Gani, A. Mokhtar, S. Ahmedm, E. Anuar, N.B. and Vasilako, V.V. 2016. "Big Data: from Beginning to Future." *International Journal of Information Management.* 36 : 1231-1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009.

Nocker, M. and Sena, V 2019. "Big Data and Human Resources Management: The Rise of Talent Analytics." *Social Sciences.* 8(10) : 273. https://doi.org/10.3390/socsci8100273.

Johari, A. 2022. **Data Analyst vs Data Engineer vs Data Scientist: Skills, Responsibilities, Salary.** [Online]. Available : https://www.edureka.co/blog/data-analyst-vs-data-engineer-vs-data-scientist/.

Kobayashi B, V. Mol T, S. Chiarello, F. and Fantoni, G. 2018. "Text Mining in Organizational Research." *Organizational Research Methods.* 21(3) : 733-765. https://doi.org/10.1177/1094428117722619.

Hiranrat, C. and Harncharnchai, A 2018. "Using Text Mining to Discover Skills Demanded in Software Development Jobs in Thailand." *ICEMT 2018: Proceedings of the 2nd International Conference on Education and Multimedia Technology.* : 112-116. https://doi.org/10.1145/3206129.3239426.

เอกสารอ้างอิง (ต่อ)

Shirani, A. 2019. "Upskilling and Retraining in Data Analytics: A Skill-adjacency Analysis for Career Paths." *Issues in Information Systems.* 20(4) : 65-74. https://doi.org/10.48009/4_iis_2019_65-74.

Maer-Matei, M.M. Mocanu, C. Zamfir, A.M. and Georgescu, T.M. 2019. "Skill Needs for Early Career Researchers—A Text Mining Approach." *Sustainability.* 11(10) : 2789. https://doi.org/10.3390/su11102789.

Gurcan, F. and Cagiltay, N.E. 2019. "Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets using LDA-Based Topic Modeling." *IEEE Access.* 7 : 8254-82552. https://doi.org/10.1109/ACCESS.2019.2924075.

Fareri, S. Melluso, N. Chiarello, F. and Fantoni, G. 2021. "SkillNER: Mining and Mapping Soft Skills from any Text." *Expert Systems with Applications.* 184 : 115544. https://doi.org/10.1016/j.eswa.2021.115544.

Wings, I. Nansa, R. and Adebayo, K.J. 2021. "A Context-Aware Approach for Extracting Hard and Soft Skills." *Procedia Computer Science.* 193 : 163-172. https://doi.org/10.1016/j.procs.2021.10.016.

Florentin, K. Melluso, F. Jiechieu, F. and Tsopze, N. 2021. "Skills Prediction Based on Multi-Label Resume Classification using CNN with Model Predictions Explanation." *Neural Computing and Applications.* 33 : 5069–5087. https://doi.org/10.1007/s00521-020-05302-x.

Kwartler, T. 2017. **Text Mining in practice with R.** Chichester : Wiley.

Pagon, G. 2019. **Confusion Matrix เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนายใน Machine learning.** [Online]. Available : https://medium.com/@pagongatchalee/confusion-matrix-เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย-ในmachine-learning-fba6e3f9508c

Nayak, A. S. Kanive, A. P. Chandavekar, N. and Balasubramani, R. 2016. "Survey on Pre-Processing Techniques for Text Mining." *International Journal Of Engineering and Computer Science Neural Computing and Applications.* 5 : 16875-16879.

เอกสารอ้างอิง (ต่อ)

Bird, S. Loper, E. and Klein, E. 2009. **Natural Language Processing with Python**. [Online]. Available : http://www.nltk.org/.

ชิตพงษ์. 2563. **Support Vector Machines**. [Online]. Available : https://guopai.github.io/ml-blog08.html

ณรงค์ศักดิ์ โค้ววิลัยแสง และ อัครนันท์ พงศธรวิวัฒน. 2564. “ตัวแบบการเรียนรู้จำแนกประเภทซัพพลายเออร์แบบมีผู้สอนสำหรับปัญหาการประเมินประสิทธิภาพของซัพพลายเออร์ในระบบ SAP ERP.” Thai Journal of Operations Research. 9(1) : 106-119

Sterling, N. W. Patzer, R. E. Di, M. and Schrager, J. D. 2019. “Prediction of Emergency Department Patient Disposition based on Natural Language Processing of Triage Notes.” *International Journal of Medical Informatics*. 129 : 184-188. https://doi.org/10.1016/j.ijmedinf.2019.06.008.

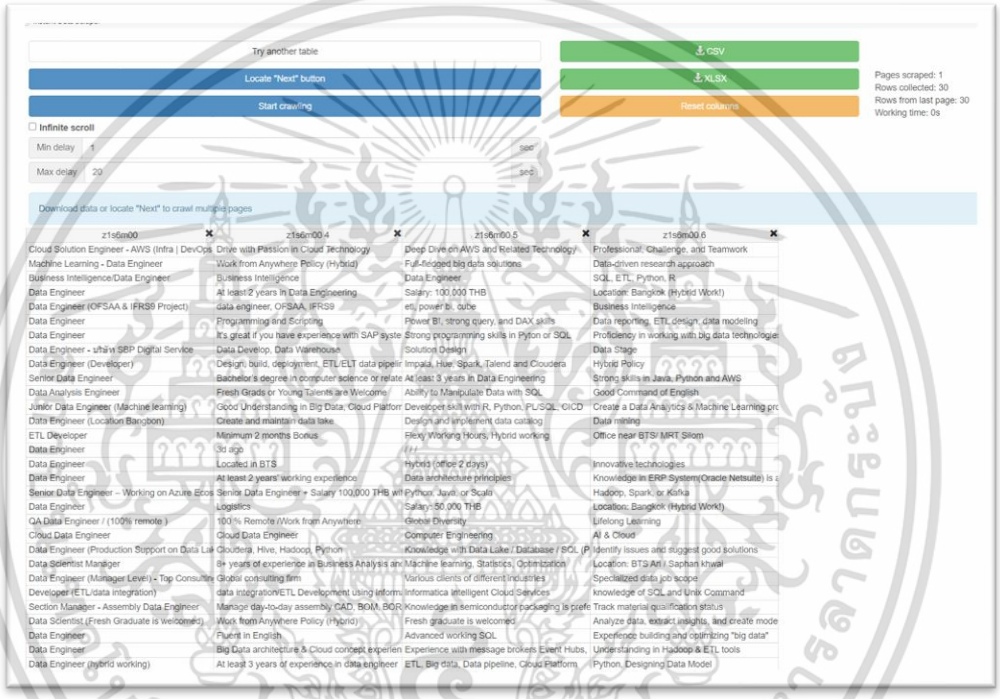
ภาคผนวก

ภาคผนวก ก

ตัวอย่างการเก็บรวบรวมข้อมูลประกาศรับสมัครงานจาก www.jobsdb.com

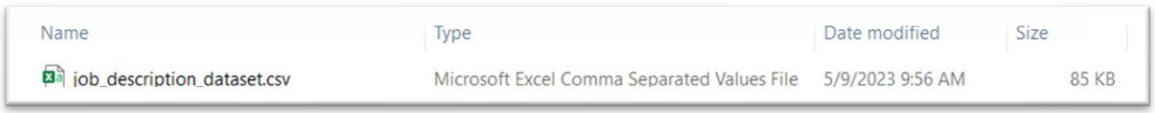
ภาคผนวก ก.1 แสดงตัวอย่างโปรแกรมการรวบรวมภาพจากเว็บไซต์ที่ได้รับอนุญาตแล้ว

ขั้นตอนการรวบรวมข้อมูลภาพจากเว็บไซต์มาจัดสรรลงเครื่องคอมพิวเตอร์ แสดงดับที่รูปที่ ก.1



รูปที่ ก.1 ขั้นตอนการดึงข้อมูลจากเว็บไซต์ผ่านโปรแกรม Instant Data Scraper

จากรูปที่ ก.1 แสดงขั้นตอนการดึงข้อมูลจากเว็บไซต์ผ่านโปรแกรม Instant Data Scraper ซึ่งเป็นโปรแกรม Open Source สามารถใช้งานได้โดยไม่มีค่าใช้จ่าย



รูปที่ ก.2 การจัดเก็บข้อมูลนามสกุล .csv ลงบนเครื่องคอมพิวเตอร์

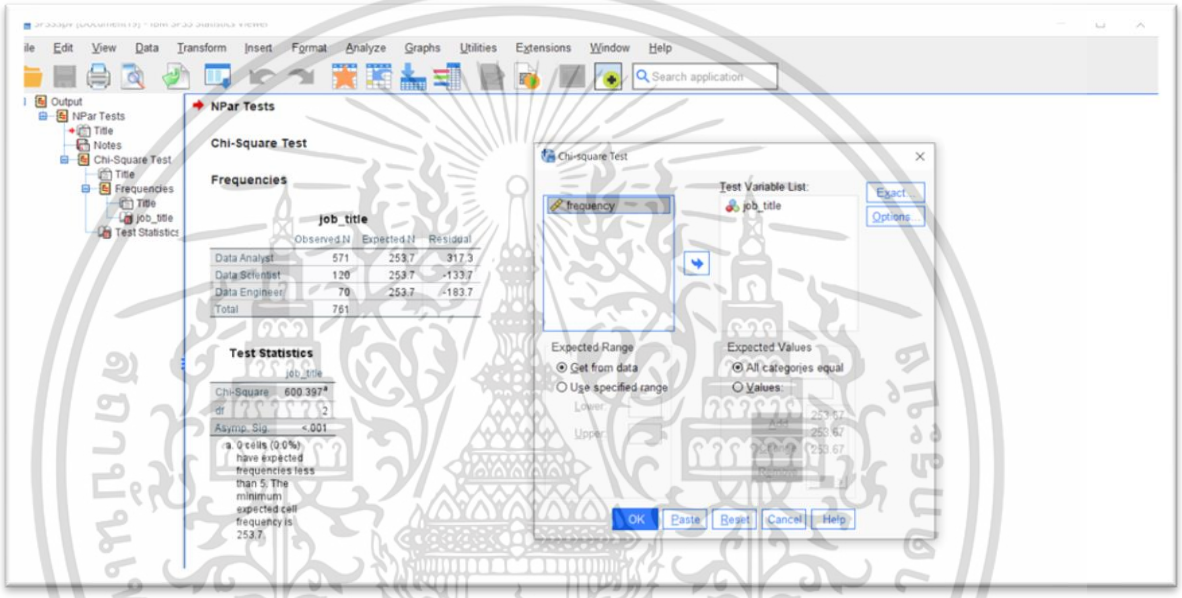
จากรูปที่ ก.2 แสดงการจัดเก็บข้อมูลนามสกุล .csv ลงบนเครื่องคอมพิวเตอร์

ภาคผนวก ข

ตัวอย่างการทดสอบสัดส่วนความนิยมของอาชีพ

ภาคผนวก ข.1 การทดสอบสัดส่วนความนิยมของอาชีพ

ขั้นตอนนี้เป็นขั้นตอนในการทดสอบสัดส่วนความนิยมของอาชีพ



รูปที่ ข.1 การทดสอบสัดส่วนความนิยมของอาชีพ

จากรูปที่ ข.1 ตัวอย่างการตั้งค่าเพื่อการทดสอบสัดส่วนความนิยมของอาชีพ

ภาคผนวก ค

ตัวอย่างการสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ภาคผนวก ค.1 การสร้างแบบจำลองการเรียนรู้ของเครื่อง

ขั้นตอนนี้เป็นขั้นตอนในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยภาษาไพทอน

```
from sklearn import svm
from sklearn.svm import SVC


#clf = svm.SVC(kernel='linear',probability=True)
#clf = svm.SVC(kernel='rbf')
clf = svm.LinearSVC()
clf.fit(X_train, y_train)
SVC()

# Fit model
clf.fit(X_train, y_train)
## Predict
y_predicted = clf.predict(X_test)
```

รูปที่ ค.1 ขั้นตอนในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยภาษาไพทอน

จากรูปที่ ค.1 ตัวอย่างการเขียนโปรแกรมด้วยภาษาไพทอนในการสร้างแบบจำลองการเรียนรู้ของเครื่อง

ภาคผนวก ค.2 การสร้างแบบจำลองการเรียนรู้เชิงลึก

ขั้นตอนนี้เป็นขั้นตอนในการสร้างแบบจำลองการเรียนรู้เชิงลึกของเครื่องด้วยภาษาไพทอน

```
from sklearn.neural_network import MLPClassifier
from sklearn.datasets import make_classification
clf = MLPClassifier(hidden_layer_sizes=(100,10) ,max_iter=300)

# Fit model
clf.fit(X_train,y_train)
# Predict
y_predicted = clf.predict(X_test)
```

รูปที่ ค.2 ขั้นตอนในการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยภาษาไพทอน

จากรูปที่ ค.2 ตัวอย่างการเขียนโปรแกรมด้วยภาษาไพทอนในการสร้างแบบจำลองการเรียนรู้ของเชิงลึก

ภาคผนวก ง

ตัวอย่างการเขียนโปรแกรมเพื่อให้ทักษะทางเทคนิคและจรณทักษะที่จำเป็น

ภาคผนวก ง.1 ตัวอย่างโปรแกรมของขั้นตอนวิธีการเรียนรู้ของเครื่องด้วยแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีน

ขั้นตอนนี้เป็นขั้นตอนการนำคำที่ได้จากแบบจำลองที่ผู้วิจัยสรุปไว้ว่าเป็นวิธีที่ดีที่สุดมาวิเคราะห์ทักษะทางเทคนิคและจรณทักษะที่จำเป็นต่ออาชีพ

```
technical_skills = ['python', 'c','r', 'c++','java','hadoop','scala','flask','pandas','spark','scikit-learn',
                    'numpy','php','sql','mysql','css','mongdb','nltk','fastai' ,'keras', 'pytorch','tensorflow',
                   'linux','Ruby','JavaScript','django','react','reactjs','ai','ui','tableau', 'nosql', 'big data',
                    'clound','powerbi','power bi','aws','qlik','excel','etl','nlp','pandas','pyspark','azure','hive',
                    'mathematical','digitalocean','sap','spss', 'analysis', 'statistic', 'modeling', 'machine']


feature_array = vectorizer.get_feature_names_out()
# number of overall model features
features_numbers = len(feature_array)
## max sorted features number
n_max = int(features_numbers * 0.1)


output = pd.DataFrame()
for i in range(0,len(clf.classes_)):
    print("\n****" ,clf.classes_[i],"****\n")
    class_prob_indices_sorted = clf.coef_[i, :].argsort()[::-1]

    raw_skills = np.take(feature_array, class_prob_indices_sorted[:n_max])
    print("list of unprocessed skills :")
    print(raw_skills)
```

รูปที่ ง.1 ตัวอย่างขั้นตอนการเขียนโปรแกรมเพื่อวิเคราะห์ทักษะที่จำเป็นต่ออาชีพ

จากรูปที่ ง.1 เป็นตัวอย่างขั้นตอนการเขียนโปรแกรมด้วยภาษาไพทอนเพื่อวิเคราะห์ทักษะทางเทคนิคและจรณทักษะที่จำเป็นต่ออาชีพ

```
print(output.T)
```

```
                                          0  \
job_title                      Data Analyst
technical_skills             [sql, php, excel]
soft_skills       [diagram, atmosphere, debug]

                                          1                    2
job_title                      Data Engineer       Data Scientist
technical_skills         [tableau, flask, nosql]  [pyspark, ai, nlp]
soft_skills       [friday, energetic, postgresql]  [exp, hive, digitalocean]
```

รูปที่ ง.2 ผลลัพธ์ของทักษะทางเทคนิคและจรณะทักษะ

จากรูปที่ ง.2 เป็นผลลัพธ์ของทักษะทางเทคนิคและจรณะทักษะที่จำเป็นต่ออาชีพนักวิเคราะห์ข้อมูลนักวิทยาศาสตร์ข้อมูล และวิศวกรข้อมูล

ประวัติผู้เขียน

ชื่อ ปรัชญ์ธิดา พรประเสริฐ

วัน เดือน ปีเกิด 11 กรกฎาคม 2535

ที่อยู่ปัจจุบัน 68/42 หมู่ 1 ถนน สุขุมวิท ต.บ้านฉาง อ.บ้านฉาง จ.ระยอง

ประวัติการศึกษา (2556) วิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมเคมี (มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ)

ทุนการศึกษาที่ได้รับ ไม่มี

ผลงานทางวิชาการ ไม่มี