การประยุกต์การเรียนรู้ด้วยคอมพิวเตอร์เพื่อจำแนกรายการบัญชีในบัญชีแยกประเภท

อารีวรรณ สันติสิริพงศ์[1,*], สมพร สำอางค์ศรี[2] และศุภฤกษ์ มานิตพรสุทธ์[2]

[1]หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต วิศวกรรมคอมพิวเตอร์และเทคโนโลยีการเงิน
[2]สาขาวิชาวิศวกรรมคอมพิวเตอร์และปัญญาประดิษฐ์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยหอการค้าไทย
**areewan2450@hotmail.com*

บทคัดย่อ

องค์กรขนาดกลางและขนาดใหญ่ มักมีจำนวนรายการในบัญชีแยกประเภทจำนวนมากในแต่ละเดือน ทำให้ขั้นตอนในการแยกรายการออกเป็นกลุ่มต่าง ๆ ใช้เวลานาน เพื่อบรรเทาปัญหาดังกล่าว ในบทความนี้ได้นำเสนอการประยุกต์ใช้การเรียนรู้ด้วยคอมพิวเตอร์เพื่อจำแนกรายการบัญชีแบบอัตโนมัติ ได้มีการใช้การประมวลผลภาษาธรรมชาติเพื่อสกัดคุณลักษณะออกจากรายการบัญชีที่อาจเป็นข้อความที่มีทั้งภาษาไทย ภาษาอังกฤษ ตัวเลข หรืออักขระพิเศษปนกัน จากนั้นได้ใช้การเรียนรู้ด้วยคอมพิวเตอร์แบบมีการสอนเพื่อจำแนกรายการทำบัญชีออกเป็น 3 ประเภท ได้แก่ ค่าใช้จ่าย สินทรัพย์ ซื้อวัตถุดิบและวัสดุ ในบทความนี้ได้นำเสนอผลลัพธ์ของการเรียนรู้ด้วยคอมพิวเตอร์ 3 อัลกอริทึมได้แก่ Decision Tree, Logistic Regression และ Naïve Bayes ผลการศึกษาเชิงทดลองปรากฏว่า อัลกอริทึม Naïve Bayes , Logistic Regression, และDecision Tree มีความถูกต้องเฉลี่ย 89.88% ,88.99% และ 79.40% ตามลำดับ

คำสำคัญ: การจำแนกประเภท, การเรียนรู้ด้วยคอมพิวเตอร์, การประมวลผลภาษาธรรมชาติ

An Application of Machine Learning for Bookkeeping Entry Classification in Accounting Ledger

Areewan Santisiripong[1,*], Somporn Samangsri[2] and Suparerk Manitpornsut[2]

[1]*Master of Engineering in Computer Engineering and Financial Technology*
[2]*Computer Engineering and Artificial Intelligence*
School of Engineering, University of the Thai Chamber of Commerce
**areewan2450@hotmail.com*

Abstract

Each month medium and large organizations usually cope with a large number of bookkeeping entries in accounting ledgers. This is a time-consuming process. To alleviate the problem, we propose an application of machine learning to classify bookkeeping entries. The natural language processing is employed to extract the features out of each entry, which may be in the combination of Thai, English, numbers, and/or special characters. Then, supervised machine learning algorithms are applied, i.e. Decision Tree, Logistic Regression and Naïve Bayes. The experimental results show that the accuracy of Naïve Bayes, Logistic Regression and Decision Tree is 89.88%, 88.99% and 79.40%, respectively.

Keywords: Classification, Machine Learning, Natural Language Processing

1. บทนำ

องค์กรทุกองค์กร ไม่ว่าจะเป็นหน่วยงานของรัฐหรือเอกชนจำเป็นต้องมีระบบบัญชี ในแต่ละประเทศมีกฎหมายควบคุมหรือมีระเบียบปฏิบัติในการทำบัญชี ซึ่งรวมทั้งประเทศไทยด้วย ซึ่งระบบบัญชีมีความซับซ้อนและมีเอกสารที่เกี่ยวข้องมากมาย

การทำบัญชีเริ่มต้นจากการจัดบันทึกรายการค้าในบัญชีแยกประเภท ข้อมูลเหล่านี้สามารถนำมาประมวลผลเพิ่มเติม เช่น การสรุปออกเป็นงบการเงิน เป็นต้น ด้วยลักษณะงานบัญชีที่มีการทำเป็นรายเดือน ทำให้อาจมีการสะสมของงานเอกสารจำนวนมาก ทำให้ต้องใช้ทรัพยากรทั้งบุคลากรและเวลาจำนวนมากในการทำงานดังกล่าว

ในปัจจุบันมีการใช้เครื่องมือทางดิจิทัล เช่น การใช้โปรแกรมบนคอมพิวเตอร์ และแท็บเลต เป็นต้น เพื่อช่วยในการทำบัญชี ข้อมูลทางบัญชีในเครื่องมือเหล่านี้มักอยู่ในรูปแบบดิจิทัล แต่อย่างไรก็ตาม เครื่องมือเหล่านี้มักต้องอาศัยความรู้ความชำนาญของนักบัญชีในการกรอกข้อมูลให้ถูกต้อง และยังต้องใช้เวลาในการปฏิบัติงานด้วยมือ ซึ่งอาจนำไปสู่ข้อผิดพลาดได้

เพื่อให้เกิดการไหลของงานแบบดิจิทัล (Digital Workflow) ที่นำไปสู่กระบวนการแบบอัตโนมัติ (Automation) ควรมีขั้นตอนดังนี้

1. รับเอกสารที่เกี่ยวข้อง เช่น ใบเสร็จรับเงิน ใบเรียกเก็บเงิน เป็นต้น หากเอกสารไม่อยู่ในรูปแบบดิจิทัล ต้องมีการแปลงให้อยู่ในรูปแบบดิจิทัลที่สามารถประมวลได้ เช่น การใช้สแกนเนอร์และ OCR (Optical Character Recognition) เพื่อรู้จำตัวอักษร แล้วบันทึกอยู่ในรูปแบบตัวอักขระ เป็นต้น
2. บันทึกข้อมูลในรูปแบบดิจิทัลลงในบัญชีทั่วไป ข้อมูลเหล่านี้ควรอยู่ในรูปแบบตัวอักษร และ/หรือ ตัวเลขที่สามารถประมวลผลได้อย่างรวดเร็ว
3. จำแนกข้อมูลแต่ละรายการ โดยแยกเป็นประเภทต่าง ๆ เช่น ค่าใช้จ่าย สินทรัพย์ ซื้อวัตถุดิบและวัสดุ เป็นต้น
4. สรุปข้อมูลทางบัญชีในรูปแบบที่เหมาะสม เช่น งบการเงิน เป็นต้น

หากองค์กรใดสามารถวางระบบให้สามารถทำตามขั้นตอนข้างต้นได้ จะเกิดการไหลของงานแบบดิจิทัลได้อย่างอัตโนมัติ ทำให้ระบบบัญชีมีความรวดเร็ว รู้ข้อมูลสถานะทางการเงินได้แบบเวลาจริง และสามารถตรวจสอบได้ในระยะเวลาอันสั้น

อย่างไรก็ตามการไหลของงานแบบดิจิทัลที่ยกตัวอย่างไว้ข้างต้นนี้ มีความซับซ้อนและมีประเด็นในการทำวิจัยหลายประเด็นในแต่ขั้นตอน ดังนั้น ในงานวิจัยนี้จึงมีสมมติฐานเพื่อให้ขอบเขตของงานวิจัยอยู่ในกรอบงานและเวลาที่เหมาะสมดังนี้

1. เอกสารที่เกี่ยวข้องอยู่ในรูปแบบดิจิทัล ที่สามารถนำไปประมวลผลได้ทันที เช่น อยู่ในรูปแบบ CSV หรือ XLS/XLSX เป็นต้น ข้อความในแต่ละรายการ ประกอบไปด้วยอักขระภาษาไทย ภาษาอังกฤษ ตัวเลข และ/หรือ อักขระพิเศษ ที่อาจผสมกัน
2. การจำแนกข้อความแต่ละรายการออกเป็นประเภทต่าง ๆ ได้แก่ ค่าใช้จ่าย สินทรัพย์ ซื้อวัตถุดิบและวัสดุ เท่านั้น ไม่มีการนำข้อมูลที่จำแนกแล้วไปประมวลผลในระดับถัดไป

จากสมมติฐานข้างต้น ประโยชน์ของงานวิจัยนี้คือ การนำผลลัพธ์จากการจำแนกรายการออกเป็นประเภทต่าง ๆ ไปประมวลผลเพิ่มเติม เพื่อให้ได้รายงานทางบัญชีที่ต้องการ เช่น งบกำไรขาดทุน เป็นต้น

ในบทความนี้ มีองค์ประกอบหลักดังนี้ ในหัวข้อที่ 2 เป็นความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง ในขณะที่ผู้วิจัยได้อภิปรายสิ่งที่นำเสนอไว้ในหัวข้อที่ 3 และผลการทดสอบในหัวข้อที่ 4 ส่วนหัวข้อที่ 5 ซึ่งเป็นหัวข้อสุดท้ายคือ การสรุป อภิปรายผล และแนวทางในการทำวิจัยในอนาคต

2. ความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง

2.1 ความรู้พื้นฐาน

2.1.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติคือ การใช้เทคนิคทางปัญญาประดิษฐ์ในการประมวลผลเพื่อให้เข้าใจภาษาพูดหรือภาษาเขียนของคน (อัศนีย์และคณะ, 2541) ตัวอย่างพื้นฐานของการประมวลผลภาษาธรรมชาติได้แก่

1) การตัดคำในประโยค (Tokenization) คือ กลไกในการแบ่งตัวอักขระที่อยู่ในประโยคออกเป็นคำ (Token) เช่น “ฉันพูดภาษาไทย” อาจสามารถแบ่งออกเป็นคำได้ดังนี้ “ฉัน”, “พูด”, “ภาษา”, “ไทย” เป็นต้น

2) การกำกับหน้าที่ของคำ (Part of Speech Tagging) คือ กลไกในการหาหน้าที่ของคำ (Part of Speech) ในประโยค เช่น “I am a boy.” สามารถแบ่งประโยคเป็นคำแต่ละคำและกำกับหน้าที่ได้ดังนี้ I_PNP am_VBB a_AT0 boy_NN1 ._PUN โดยหน้าที่ของคำเป็นดังตารางที่ 1 โดยที่หน้าที่ของคำทั้งหมดสามารถดูเพิ่มเติมได้ที่ UCREL (http://ucrel.lancs.ac.uk/claws5tags.html)

ตารางที่ 1 ตัวอย่างหน้าที่ของคำ

คำย่อ	หน้าที่ของคำ
PNP	personal pronoun (e.g. YOU, THEM)
VBB	the "base forms" of the verb "BE"
AT0	singular article (e.g. a, an, every)
NN1	singular common noun (e.g. book, girl)
PUN	punctuation - general mark (i.e. . ! , : ; - ? ...)

นอกจากการประมวลผลภาษาธรรมชาติแบบพื้นฐานข้างต้นแล้ว การประมวลผลภาษาธรรมชาติสามารถพัฒนาไปสู่ระบบที่มีความซับซ้อนมากขึ้น เช่น การแปลภาษา (Machine Translation) การย่อความ (Text Summarization) การวิเคราะห์อารมณ์ความรู้สึก (Sentiment Analysis) ที่มีอยู่ในข้อความ เป็นต้น

2.1.2 การเรียนรู้ด้วยคอมพิวเตอร์ (Machine Learning)

การเรียนรู้ด้วยคอมพิวเตอร์เป็นศาสตร์ย่อยในด้านปัญญาประดิษฐ์ เป็นกลไกในการใช้อัลกอริทึมของคอมพิวเตอร์ที่มีแบบจำลองทางคณิตศาสตร์ในการสร้างผลลัพธ์ซึ่งอาจเป็นการทำนาย การจัดกลุ่ม เป็นต้น

แบบจำลองคณิตศาสตร์ดังกล่าว อาจจัดทำโดยมีการเรียนรู้แบบมีการสอน (Supervised Learning) การเรียนรู้แบบไม่มีการสอน (Unsupervised Learning) การเรียนรู้แบบสร้างเสริม (Reinforcement Learning) และอาจมีการเรียนรู้แบบอื่นหรือแบบใหม่ที่ถูกพัฒนาขึ้น การเรียนรู้แต่ละแบบเหมาะกับลักษณะการนำไปประยุกต์ใช้ที่แตกต่างกัน เช่น การเรียนรู้แบบมีการสอนสามารถนำไปใช้ในการจำแนกข้อมูล (Data Classification) และการรู้จำวัตถุ (Object Recognition) เป็นต้น ส่วนการเรียนรู้แบบไม่มีการสอนสามารถนำไปประยุกต์ใช้กับการจัดกลุ่มข้อมูล (Data Clustering) เป็นต้น ในขณะที่การเรียนรู้แบบสร้างเสริมสามารถนำไปประยุกต์ใช้กับปัญหาที่ใช้ซอฟต์แวร์เอเจนท์หลายตัวในการคำนวณ เช่น การควบคุมสัญญาณไฟจราจรบนถนนที่มีหลายแยก แต่ละแยกใช้ซอฟต์แวร์เอเจนท์ในการตัดสินใจเพื่อควบคุมเป็นอิสระแต่สามารถสื่อสารกันได้ เป็นต้น

2.2 งานวิจัยที่เกี่ยวข้อง

อมรา ติรศรีวัฒน์ (2561) กล่าวถึงการพัฒนาหลักสูตรบัญชี ให้นักศึกษาต้องมีความพร้อมรองรับการทำงานในยุคดิจิทัล ประกอบด้วยความรู้ ทักษะด้านเทคโนโลยีสารสนเทศ ระบบบัญชี โดยใช้เครื่องมือที่ทันสมัยในการประมวลผลในรูปแบบของระบบปัญญาประดิษฐ์ สามารถเรียนรู้และปรับปรุงรูปแบบการทำงานได้ เช่น การเก็บรายรับ รายจ่าย การลงบัญชีอาจถูกแทนด้วยระบบอัตโนมัติ การค้นหาข้อมูล การจัดทำรายงานด้วยระบบดิจิทัล รวมทั้งทักษะและการประยุกต์ใช้เครื่องมืออุปกรณ์ทางเทคโนโลยีสารสนเทศ การสืบค้น การวิเคราะห์ข้อมูล โปรแกรมคอมพิวเตอร์ในการประมวลผล แบบจำลองข้อมูล เป็นต้น สิ่งเหล่านี้แสดงให้เห็นถึงแนวโน้มของการเกิดการไหลของงานแบบดิจิทัล

มีงานวิจัยจำนวนมากนำการเรียนรู้ด้วยคอมพิวเตอร์มาประยุกต์ใช้ในการประมวลผลข้อมูล เช่น กานดา แผ่วัฒนากุล และ ดร.ปราโมทย์ ลือนาม (2556), G. Ozdagoglu et al. (2560) และ ยุทธ ไกยวรรณ์ (2555) ได้ใช้อัลกอริทึมพื้นฐาน เช่น Decision Tree, Logistic Regression, และ Naïve Bayes เพื่อจำแนกข้อมูล เป็นต้น งานที่มีความซับซ้อนมากขึ้นมีการใช้การวิเคราะห์ข้อความและการเรียนรู้ของคอมพิวเตอร์ เพื่อหาข้อมูลที่ไม่มีโครงสร้างในด้านการเงินและบัญชีโดยใช้ Artificial Neural Network (Li Guo, et al., 2562), นอกจากนั้น ในกรณีที่ข้อมูลมีมากกว่า 2 กลุ่ม การใช้การเรียนรู้ของคอมพิวเตอร์

เพื่อจำแนกข้อมูลยังทำได้อย่างมีประสิทธิภาพ (Shiny Abraham,et al., 2020) และ (ทองสา บุตรงามและบุญญสิทธิ์ วรจันทร์ (2562)

3. งานวิจัยที่เสนอ

การจำแนกรายการบัญชีในบัญชีแยกประเภทที่นำเสนอมีขั้นตอนในการประมวลผลดังนี้

1. คัดแยกรายการบัญชีออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลสำหรับการสอน (Training Data Set) และชุดข้อมูลสำหรับการทดสอบ (Testing Data Set)
2. นำข้อมูลแต่ละชุดมาสกัดคุณลักษณะ (Feature Extraction)
3. นำคุณลักษณะของชุดข้อมูลสำหรับการสอนมาป้อนเป็นอินพุตให้กับอัลกอริทึมการจำแนกข้อมูล ซึ่งได้ผลลัพธ์เป็นแบบจำลอง
4. นำแบบจำลองที่ได้มาทดสอบความถูกต้อง โดยทดสอบด้วยชุดข้อมูลสำหรับการทดสอบ

ขั้นตอนข้างต้นมีรายละเอียดดังนี้

3.1 ชุดข้อมูล

รายการบัญชีต้นฉบับ แสดงได้ดังตารางที่ 2 อย่างไรก็ตาม เมื่อใช้เป็นชุดข้อมูลสำหรับงานวิจัยนี้ จำเป็นต้องมีการทำความสะอาดข้อมูล (Data Cleansing) และมีการกำกับประเภทของข้อมูล (Label) ซึ่งได้ดังผลลัพธ์ในตารางที่ 3

ตารางที่ 2 ตัวอย่างของรายการบัญชีต้นฉบับ

shipdate	docudate	docuno	vatamnt	sumgoodamnt	netamnt	goodname	goodamnt	vendornameeng	goodunitname
20/01/2017 00:00	11/01/2017 00:00	POD6001-053	xx	xx	xx	EF-4018AT Metal Pipe Clamp, Flat	xx	SIAM TAISEI INDUSTRY CO., LTD	ชิ้น
20/01/2017 00:00	11/01/2017 00:00	POD6001-053	xx	xx	xx	EF-4018BT Metal Pipe Clamp, Flat	xx	SIAM TAISEI INDUSTRY CO., LTD	ชิ้น
23/02/2017 00:00	21/02/2017 00:00	POD6002-110	xx	xx	xx	ถุงขยะดำหนา 36"x45"	xx	บริษัท สกายแมพ เทรดดิ้ง จำกัด	ชิ้น
18/05/2017 00:00	11/05/2017 00:00	POD6005-047	xx	xx	xx	เทปพิมพ์อักษร Brother สีเหลือง TZ-241 (18 mm.)	xx	บริษัท สกายแมพ เทรดดิ้ง จำกัด	ชิ้น
10/02/2017 00:00	08/02/2017 00:00	POD6002-051	xx	xx	xx	#4003582 เมาส์ออฟติคัลไร้สาย น้ำเงิน Logitech M187	xx	บริษัท ซีโอแอล จำกัด	ชิ้น
18/10/2017 00:00	12/10/2017 00:00	POD6010-065	xx	xx	xx	4000614 เครื่องสำรองไฟ L-850Aเพาเวอร์ คอมCtranlin	xx	บริษัท ซีโอแอล จำกัด	ชิ้น
13/07/2017 00:00	11/07/2017 00:00	POD6007-052	xx	xx	xx	8030330 ถ่านอัลคาไลน์ E91-BP2 AA (แพ็ค 2 ก้อน) En	xx	บริษัท ซีโอแอล จำกัด	แพค
25/01/2017 00:00	24/01/2017 00:00	POD6001-125	xx	xx	xx	0001794 แปรงขัดโถสุขภัณฑ์ ด้ามยาว แดง NCL	xx	บริษัท ซีโอแอล จำกัด	ชิ้น
10/11/2017 00:00	07/11/2017 00:00	POD6011-042	xx	xx	xx	0000285 น้ำยาล้างห้องน้ำ เป็ด ฟิงค์-900ML	xx	บริษัท ซีโอแอล จำกัด	ขวด
27/11/2017 00:00	27/11/2017 00:00	POD6011-155	xx	xx	xx	EVA FOAM + กาว Size :1200x2300 xT=1.5mm.	xx	CRV PACKAGING CO., LTD.	ชิ้น
30/01/2018 00:00	13/12/2017 00:00	POD6012-087	xx	xx	xx	เสื้อแขนยาว	xx	บริษัท มิลาโน ยูนิฟอร์ม จำกัด	ตัว
15/02/2018 00:00	13/02/2018 00:00	POD6102-079	xx	xx	xx	AL6H-P4-A Push button & Pilot Light	xx	QUALITECH CONTROL CO., LTD	ชิ้น
09/09/2018 00:00	05/09/2018 00:00	POD6109-025	xx	xx	xx	Plywood+Formica(black) 375x565xT15mm. R4มุม (	xx	LUANG THAWORN FURNITURE (	ชิ้น
04/07/2018 00:00	03/07/2018 00:00	POD6107-006	xx	xx	xx	0001521 นาฬิกาแขวนผนังดำ JIMKO 548	xx	บริษัท ซีโอแอล จำกัด	ชิ้น
04/09/2018 00:00	31/08/2018 00:00	POD6108-171	xx	xx	xx	0001165 ผงซักฟอกบรีส เพาเวอร์ 900 กรัม	xx	บริษัท ซีโอแอล จำกัด	ถุง

ตารางที่ 3 ตัวอย่างชุดข้อมูลที่ผ่านการทำความสะอาดข้อมูล

รายการบัญชี	ประเภท
EF-4018BT Metal Pipe Clamp, Flat	material
EF-4018AT Metal Pipe Clamp, Flat	material
ถุงขยะสีดำ 36*45	expense
เทปพิมพ์อักษร Brother สีเหลือง TZ-241 (18 mm.)	expense
4003582 เมาส์ออฟติคัลไร้สาย น้ำเงิน Logitech M187	asset

ชุดข้อมูลทั้งสิ้นมีจำนวน 1,875 รายการ โดยแบ่งเป็นรายการประเภทค่าใช้จ่าย, ประเภทสินทรัพย์และประเภทซื้อวัตถุดิบหรือวัสดุ อย่างละ 625 รายการ การแบ่งชุดข้อมูลมีการแบ่ง 2 วิธีตามกลไกในการทดสอบคือ

1) การทดสอบแบบ Train-Test Split ได้แบ่งข้อมูลออกเป็นชุดข้อมูลสำหรับการสอน 70% และชุดข้อมูลสำหรับการทดสอบ 30%

2) การทดสอบแบบ 10-Fold Cross Validation ทำโดยการแบ่งเป็นชุดข้อมูลสำหรับการสอน 90% และชุดข้อมูลสำหรับการทดสอบ 10% แต่มีการปรับเปลี่ยนชุดข้อมูล 10 ครั้งตามกลไกการทดสอบ

3.2 การสกัดคุณลักษณะของข้อมูล

รายการบัญชีเป็นข้อมูลที่เป็นตัวอักษรภาษาไทย ภาษาอังกฤษ ตัวเลข และ/หรืออักขระพิเศษ ซึ่งอาจมีปนกันได้หลายรูปแบบ เพื่อให้สามารถนำข้อมูลไปประมวลผลด้วยอัลกอริทึมการจำแนกข้อมูลได้ จำเป็นต้องมีการสกัดคุณลักษณะจากรายการบัญชีก่อน

กลไกที่ใช้ในการสกัดคุณลักษณะของข้อมูลคือ การตัดคำ (Word Tokenization) ซึ่งได้ใช้ API ของ PyThaiNLP (Wannapong et al., 2016) และใช้เทคนิคถุงคำ (Bag of Words) เพื่อสร้างเว็กเตอร์ของข้อมูลจากคลังศัพท์ที่ได้จากชุดข้อมูลสำหรับการสอน

3.3 อัลกอริทึมการจำแนกข้อมูล

การเรียนรู้ด้วยคอมพิวเตอร์มีอัลกอริทึมที่สามารถนำมาใช้ในการจำแนกข้อมูลได้หลายอัลกอริทึม ในงานวิจัยนี้ได้นำอัลกอริทึม Decision Tree, Logistic Regression และ Naïve Bayes มาใช้ในการทดสอบ เนื่องจากทั้ง 3 อัลกอริทึมนั้นสามารถใช้ได้กับงาน Multiple-Class Classification ซึ่งตรงกับลักษณะของการจำแนกรายการบัญชีซึ่งต้องจำแนกออกเป็น 3 กลุ่มคือ ค่าใช้จ่าย (Expense), สินทรัพย์ (Asset) และวัสดุ (Material)

การพัฒนาโปรแกรมเพื่อทดสอบได้ใช้ผ่านเครื่องมือ Google Colab โดยใช้โปรแกรมภาษาไพธอนและไลบรารี scikit-learn (Pedregosa et al., 2011) ซึ่งมี API สำหรับการสร้างแบบจำลองโดยใช้อัลกอริทึม Decision Tree, Logistic Regression และ Naïve Bayes

3.4 การทดสอบ

การทดสอบประสิทธิภาพของแต่ละอัลกอริทึมโดยการวัดค่าความแม่นยำ (Precision), ค่าความระลึก (Recall), ค่าความถ่วงดุล (F-Measure) และค่าความถูกต้อง (Accuracy) โดยแต่ละค่าสามารถคำนวณได้ดังนี้

ค่าความแม่นยำ (Precision):

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

ค่าความระลึก (Recall):

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

ค่าความถ่วงดุล (F-Measure):

$$F = \frac{2 \; x \; Precision \; x \; Recall}{Precision + Recall} \quad (3)$$

ความถูกต้อง (Accuracy):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

โดยที่แต่ละค่าสามารถอธิบายได้ดังภาพที่ 1 ดังนี้ ในกรณีเมื่อพิจารณาประเภทค่าใช้จ่าย TP หรือ True Positive คือ จำนวนข้อมูลที่จำแนกได้ถูกต้องตามประเภทค่าใช้จ่าย ในขณะที่ TN หรือ True Negative (TN) คือ จำนวนข้อมูลที่จำแนกได้ถูกต้องว่าไม่เป็นไปตามประเภทค่าใช้จ่าย, False Positive (FP) คือ จำนวนข้อมูลที่จำแนกได้ผิดพลาด โดยทำนายว่าเป็นประเภทค่าใช้จ่ายซึ่งไม่ตรงกับประเภทของข้อมูลจริงที่เป็นข้อมูลประเภทอื่น และ False Negative (FN) คือ จำนวนข้อมูลที่จำแนกได้ผิดพลาด โดยทำนายข้อมูลว่าเป็นประเภทอื่น แต่ประเภทของข้อมูลจริงคือค่าใช้จ่าย ในกรณีประเภททรัพย์สินและวัสดุสามารถคำนวณค่า TP, TN, FP และ FN ได้ในแนวทางเดียวกัน

การทำนาย	ประเภทของข้อมูลจริง		
	ค่าใช้จ่าย	ทรัพย์สิน	วัสดุ
ค่าใช้จ่าย	TP	FP	FP
ทรัพย์สิน	FN	TN	TN
วัสดุ	FN	TN	TN

ภาพที่ 1 Confusion Matrix สำหรับประเภทค่าใช้จ่าย

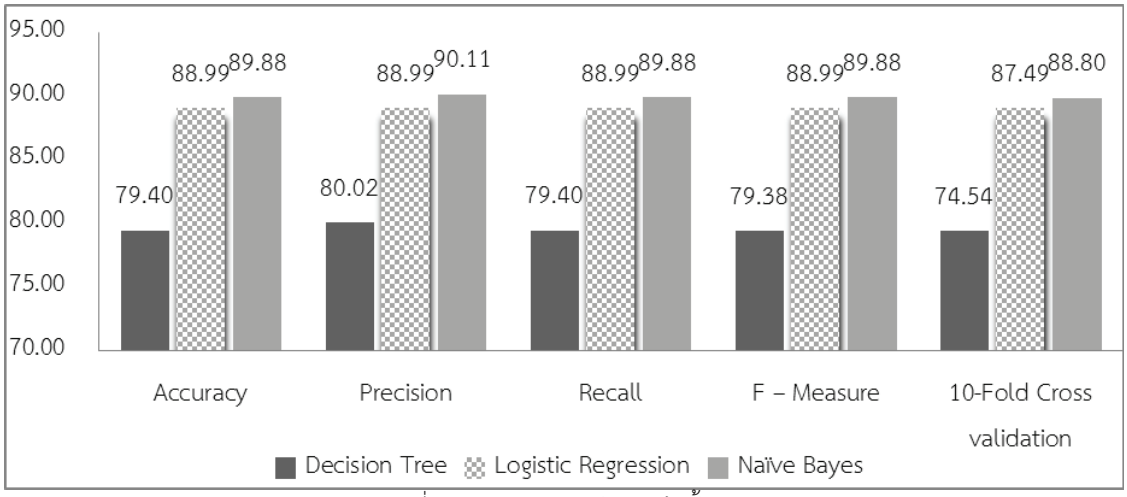
4. ผลการทดสอบ

ผลการทดสอบอัลกอริทึมจำแนกรายการบัญชีในบัญชีแยกประเภทที่ได้กล่าวไว้ในหัวข้อที่แล้วมีผลการทดสอบ ซึ่งเปรียบเทียบค่าความถูกต้อง ความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของการทดสอบทั้ง 3 อัลกอริทึม ดังตารางที่ 4

ตารางที่ 4 ผลการทดสอบเมื่อเปรียบเทียบทั้ง 3 เทคนิค

อัลกอริทึม	Accuracy (%)	Precision (%)	Recall (%)	F – Measure (%)
Decision Tree	79.40	80.02	79.40	79.38
Logistic Regression	88.99	88.99	88.99	88.99
Naïve Bayes	89.88	90.11	89.88	89.84

ผลการทดสอบ 3 อัลกอริทึมจากการทดสอบแบบ Train-Test Split พบว่า Naïve Bayes มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลมากที่สุด รองลงมาคือ Logistic Regression และ Decision Tree โดยให้ค่าความถูกต้องเท่ากับ 89.88 , 88.99 และ 79.40 ตามลำดับ



ภาพที่ 2 ผลการทดสอบอัลกอริทึมทั้ง 3 เทคนิค

ส่วนการทดสอบแบบ 10-Fold Cross Validation อัลกอริทึม Naïve Bayes ยังคงให้ความถูกต้องสูงกว่า Logistic Regression และ Decision Tree ซึ่งมีค่าเท่ากับ 88.80%, 87.49% และ 74.54% ตามลำดับ

5. สรุป

จากผลทดสอบอัลกอริทึมในการจำแนกรายการบัญชีที่ได้นำเสนอ สามารถสรุปได้ว่าอัลกอริทึม Naïve Bayes มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ที่สูงกว่าอัลกอริทึมอื่น ทั้งนี้อาจเป็นเพราะลักษณะของการใช้ความน่าจะเป็นมาประกอบตามทฤษฎีของเบย์ส (Bayes' Theorem) และการสกัดคุณลักษณะโดยใช้ถุงคำ ซึ่งจะให้ความน่าจะเป็นสูงขึ้นเมื่อรายการบัญชีที่ใช้ทดสอบมีคำที่คล้ายกับรายการบัญชีที่ใช้สำหรับสอน

อย่างไรก็ตาม อัลกอริทึมในการจำแนกข้อมูลที่สามารถนำมาประยุกต์ได้งานวิจัยนี้และที่ยังไม่ได้นำเสนอยังมีอีกมาก ซึ่งอยู่นอกเหนือจากขอบเขตของงานวิจัยนี้ เช่น SVM, ANN และโดยเฉพาะ Deep Learning ซึ่งกำลังได้รับความนิยมเป็นอย่างมากในปัจจุบัน ดังนั้น แนวทางในการวิจัยต่อไปในการอนาคตคือ การพิจารณาอัลกอริทึมอื่นดังที่กล่าวมาข้างต้น นอกจากนั้น หากต้องการสร้างการไหลของงานแบบดิจิทัลเพื่อให้รู้สถานการณ์ทางบัญชีแบบเวลาจริง สามารถนำระบบ RPA (Robot Process Automation) และระบบการสร้างรายงานมาประกอบในขั้นตอนต่าง ๆ ที่ได้อธิบายไว้ในบทนำ

เอกสารอ้างอิง (References)

อัศนีย์ ก่อตระกูล, กมลา นาคะศิริ, วิสมัย มโนมัยพิบูลย์, ศิริพร แตงเที่ยง, วิภากร วงศ์ไทย และทัศนาลัย บูรพาชีพ. (2541). การประมวลผลภาษามนุษย์ด้วยคอมพิวเตอร์. **วารสารมนุษยศาสตร์วิชาการ**, มหาวิทยาลัยเกษตรศาสตร์, 2541 (6), 94 – 104.

อมรา ติรศรีวัฒน์ (2561). รายงานวิจัยเรื่อง **การบัญชีดิจิทัลและการเตรียมความพร้อมในการเรียนการสอนนักศึกษาในยุคเศรษฐกิจดิจิทัล** : วารสารประชุมวิชาการและนำเสนอผลงานวิชาการระดับชาติ UTCC Academic Day ครั้งที่ 2 June 8,2018 ของ มหาวิทยาลัยหอการค้าไทย

กานดา แผ่วัฒนากุล และ ดร.ปราโมทย์ ลือนาม (2556). รายงานวิจัยเรื่อง **การวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายสังคมออนไลน์** วารสารการจัดการสมัยใหม่ ปีที่ 11 ฉบับที่ 2 เดือน กรกฎาคม – ธันวาคม 2556 นิสิต,สาขาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, มหาสารคาม, 44150

G.Ozdagoglu, A. Ozdagoglu , Y.Gumus and G. Kurt-Gumus (2560) : เรื่อง **การประยุกต์ใช้เทคนิคการขุดข้อมูลในการจัดการจำแนกงบการเงิน** จาก Journal of AI and Datamining Vol5, No. 1 , 2017 , p. 67-77

ยุทธ ไกยวรรณ์ (2555) งานวิจัยเรื่อง **หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย.** วารสารวิจัย มหาวิทยาลัยราชมลคลศรีวิชัย 4(1) : 1-12 (2555)

Li Guo ,Feng Shi and Jun Tu (2562). **Textual analysis and machine leaning: Crack unstructured data I finance and accounting .**(ออนไลน์) ค้นเมื่อ วันที่ 19 สิงหาคม 2562 http://www.keaipublishing.com/en/journals/jfds/

Shiny Abraham, Chau Huynh and Huy Vu. (2020) . **Classification of Soils into Hydrologic Groups Using Machine Learning** , (ออนไลน์) ค้นเมื่อ วันที่ 25 สิงหาคม 2562 , https://www.mdpi.com/search?q=soil+into+hydrologic&authors=&journal=&article_type=&search=Search&section=&special_issue=&volume=&issue=&number=&page=

ทองสา บุตรงาม และ บุญญสิทธิ์ วรจันทร์ (2562) งานวิจัยเรื่อง **การเปรียบเทียบประสิทธิภาพวิธีการพยากรณ์การเกิดภาวะแทรกซ้อนทางไตในผู้ป่วยโรคเบาหวาน ชนิดที่ 2 กรณีศึกษา : โรงพยาบาลแห่งหนึ่งในจังหวัดบุรีรัมย์.** วารสารวิจัย งานประชุมวิชาการระดับชาติ ครั้งที่ 11 มหาวิทยาลัยราชภัฏนครปฐม วันที่ 11-12 กรกฎาคม 2562

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, & Pattarawat Chormai. (2016). **PyThaiNLP: Thai Natural Language Processing in Python**. Zenodo, (ออนไลน์) ค้นเมื่อ วันที่ 26 สิงหาคม 2562 , http://doi.org/10.5281/zenodo.3519354

Pedregosa et al. (2011). **Scikit-learn: Machine Learning in Python**. JMLR 12 pp.2825-2830, 2011.