การเพิ่มประสิทธิภาพโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส

LMPROVING PREDICTION MODELS OF STUDENT BUSINESS CAREER USING SAMPLING TECHNIQUES FOR LEARNING IN MULTI-CLASS LMBALANCE DATA SET

สำราญ วานนท์
Samran Wanon
รจนา เมืองแสน
Rojjana Muangsan
คณะบริหารธุรกิจ มหาวิทยาลัยราชภัฏชัยภูมิ
Faculty of Business Administration, Chaiyaphum Rajabhat University
E-mail: samran@cpru.ac.th
วันที่รับบทความ (Received) : 2 มีนาคม 2564
วันที่แก้ไขบทความ (Revised) : 28 เมษายน 2564
วันที่ตอบรับบทความ (Accepted) : 29 เมษายน 2564

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ ดังนี้ 1) เพื่อศึกษาเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส และ 2) เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส โดยใช้โปรแกรม WEKA ปรับความสมดุลด้วยเทคนิค Over sampling เทคนิค Under sampling และเทคนิค SMOTE สร้างโมเดลและการประเมินโมเดลพยากรณ์ด้วยเทคนิคต้นไม้ตัดสินใจและเทคนิคแรนดอมฟอเรสเปรียบเทียบโมเดลที่ให้ผลการพยากรณ์ที่ดีที่สุด

ผลการวิจัยพบว่า ชุดข้อมูลที่ปรับความสมดุลด้วยเทคนิค Over sampling สร้างโมเดลจำแนกข้อมูล และเปรียบเทียบโมเดลการจำแนก 2 เทคนิค คือต้นไม้ตัดสินใจและแรนดอมฟอเรส ผลปรากฏว่าเทคนิคแรนดอมฟอเรสเป็นเทคนิคที่ให้ประสิทธิภาพมากที่สุด ค่าความถูกต้องร้อยละ 67.17 ค่าความแม่นยำ 0.66 ค่าความระลึก 0.67 และค่าเอฟเมเชอร์ 0.66
คำสำคัญ: ข้อมูลไม่สมดุล, การจำแนกข้อมูล, โมเดลพยากรณ์

Abstract

The purposes of the research are 1) to study the sampling techniques for learning in multiclass imbalanced datasets and 2) to compare the efficiency of the sampling techniques for learning in multiclass imbalanced datasets. The WEKA program is used to adjust the balance by the techniques of Over-sampling, Under-sampling and SMOTE. The prediction models are assessed by the techniques of Decision Tree and Random Forest to compare the highest forecasting result.

The results of the research indicate that the datasets balanced by Over-sampling and compared by Decision Tree and Random Forest techniques show that Random Forest is the most efficient with accuracy 67.17 %, precision 0.66 %, recall 0.67 % and F-measure 0.66 %.

Keywords: Imbalance Data, Data Classification, Prediction Model

บทนำ

ในโลกความเป็นจริงการค้นหาความรู้จากข้อมูลขนาดใหญ่ (Big Data) ด้วยเทคนิคเหมืองข้อมูลนั้น ในธรรมชาติแล้วชุดข้อมูลจะมีความไม่สมดุล (Elhassan T, Aljurf M, Al-Mohanna F & Shoukri M., 2016: 1) เป็นส่วนใหญ่ ปัญหาการแบ่งคลาสข้อมูลที่ไม่สมดุลจากการที่ข้อมูล 2 คลาสหรือมากกว่า โดยคลาสของข้อมูลหลัก (Majority) มีจำนวนข้อมูลมากกว่าคลาสของข้อมูลรอง (Minority) ซึ่งส่งผลต่อการเรียนรู้และรูปแบบการยอมรับโมเดลการพยากรณ์ ดังนั้นจึงมีความจำเป็นที่จะต้องปรับชุดข้อมูลที่ไม่สมดุลก่อนนำไปทำการเรียนรู้และทดสอบ ถ้านำข้อมูลทั้งสองชุดหรือมากกว่าเข้าสู่ขั้นตอนการ เรียนรู้พร้อมกันทั้งหมด จะทำให้ผลการจำแนกข้อมูลเกิดความผิดพลาด ในงานวิจัยนี้ผู้วิจัยจะนำเทคนิคการปรับชุดข้อมูลที่ไม่สมดุลแบบหลายคลาสมาใช้ 3 วิธีด้วยกันคือ 1) เทคนิค SMOTE (Synthetic Minority Over-sampling Technique) 2) เทคนิค ROS (Random over Sampling) และ 3) เทคนิค RUS (Random under Sampling) (Raisul Islam Rashu, Naheena Haq and Rashedur M Rahman, 2014: 14) เลือกจากหลายๆ เทคนิคที่ใช้เรียนรู้และทดสอบกับการจำแนกข้อมูล (Guillaume Lemaˆıtre, Fernando Nogueira & Christos K. Aridas, 2017: 3)

ปัจจุบันสภาพเศรษฐกิจและสังคมของโลกมีการเปลี่ยนแปลงอย่างรวดเร็ว ซึ่งส่งผลทำให้เกิดภาวะการแข่งขันกันอย่างรุนแรงของตลาดแรงงาน ในยุคของการปฏิวัติอุตสาหกรรมครั้งที่ 4 โลกของการทำงานจะมีทิศทางไปในแนวทางใด ในเมื่อผู้ประกอบการพยายามลดต้นทุนด้านแรงงานคนลงพร้อมกับนำเครื่องจักรและเทคโนโลยีเข้ามาแทนคน (กองวิจัยตลาดแรงงาน, 2559) ซึ่งภาวะการมีงานทำของคนในประเทศก็ถือเป็นตัวชี้วัดทางด้านเศรษฐกิจและสังคมอย่างหนึ่งมาก เนื่องจากตลาดแรงงานเป็นส่วนหนึ่งของเศรษฐกิจ ถ้าประชากรของประเทศมีงานทำมากย่อมส่งผลทำให้เศรษฐกิจขยายตัว ทำให้วงจรของเศรษฐกิจมีการขับเคลื่อนไปด้วยดี

สถาบันอุดมศึกษาเป็นแหล่งผลิตบัณฑิตระดับปริญญาตรีออกสู่ตลาดแรงงานในแต่ละปีการศึกษา จำนวนบัณฑิตที่จบการศึกษามีจำนวนมาก ดังนั้นสถาบันอุดมศึกษาจะต้องทราบถึงสภาวะของตลาดอาชีพว่าจะมีแนวโน้มไปอย่างไร เพื่อจะได้มีข้อมูลสำหรับการเตรียมวางแผนเร่งผลิตบัณฑิตที่มีคุณภาพในสาขาวิชาใด จึงได้ทำแบบสำรวจภาวะการมีงานทำของบัณฑิต จากข้อมูลอาชีพของผู้สำเร็จการศึกษาทางด้านธุรกิจนั้นมีอาชีพที่หลากหลาย ในบางอาชีพมีจำนวนผู้ประกอบอาชีพนั้นมากและในบางอาชีพก็มีผู้ประกอบอาชีพน้อย จากความไม่สมดุลของชุดข้อมูลดังกล่าวมาแล้ว ถ้านำไปประมวลผลตามขั้นตอนทางเทคนิคเหมืองข้อมูลแล้ว จะส่งผลต่อการเรียนรู้และความเหมาะสมต่อการยอมรับโมเดลพยากรณ์ ดังนั้นจะต้องมีการปรับชุดข้อมูลที่ไม่สมดุลดังกล่าวมาข้างต้นเพื่อเพิ่มประสิทธิภาพของโมเดลพยากรณ์อาชีพของผู้สำเร็จการศึกษาจากมหาวิทยาลัยราชภัฏชัยภูมิ

วัตถุประสงค์ของการวิจัย

1) เพื่อศึกษาเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส

2) เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส

แนวคิดเทคนิคการเรียนรู้

1) เทคนิคการทำเหมืองข้อมูล เนื่องจากการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่างๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะ ใช้ ตัวอย่างศาสตร์ที่ใช้เช่น เทคโนโลยีฐานข้อมูล (Database technology) วิทยาศาสตร์ สารสนเทศ (Information science) สถิติ (Statistics) และระบบการเรียนรู้ (Machine learning) เป็นต้น ซึ่งศาสตร์ต่างๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่างๆ (อุกฤษ ปัจฉิม, 2546)

2) เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นรูปแบบการรวบรวมนำเอาชุดข้อมูลในอดีต นำมาวิเคราะห์โดยพิจารณาความสัมพันธ์ หรือรูปแบบเฉพาะของชุดข้อมูลนั้นๆ เพื่อนำเอารูปแบบผลลัพธ์จากการวิเคราะห์ข้อมูลที่ได้ไปใช้ในการคาดการณ์ หรือ ทำนายสิ่งว่าจะเกิดขึ้นในอนาคต อาจกล่าวได้ว่า เป็นการวิเคราะห์ข้อมูลจากอดีต เป็นสิ่งที่นำเหมือนการสั่งสอน ให้ปรากฏเป็นรูปแบบ และนำรูปแบบ โมเดล ที่ได้ ไปใช้กับข้อมูลชุดใหม่ ดังนั้นจึงเรียกว่าเป็นการเรียนรู้แบบมีผู้สอนนั้นเอง โดยเทคนิควิธีการวิเคราะห์ข้อมูลลักษณ์นี้ คือ การวิเคราะห์จำแนกประเภทข้อมูล

3) การวิเคราะห์จำแนกประเภทข้อมูล (Classification Analysis) เป็นการวิเคราะห์เพื่อค้นหารูปแบบในการจัดการข้อมูลให้จำแนกตามกลุ่มที่ถูกกำหนดขึ้นโดยการสร้างตัวแบบเพื่อช่วยสนับสนุนการตัดสินใจจำแนกกลุ่มข้อมูลในอดีตที่มีอยู่เพื่อนำไปใช้ในการคาดการณ์ หรือ ทำนายแนวโน้มการเกิดขึ้นของข้อมูลชุดใหม่ โดยตัวแบบที่ได้จากการวิเคราะห์ข้อมูล อาจอยู่ในรูปแบบของกฎการจำแนกประเภท (Classification Rules) หรือต้นไม้ตัดสินใจ (Decision Tree) และการประมาณค่าข้อมูล (Regression)

3.1) ต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้ตัดสินใจถูกพัฒนาโดย J.R. Quinlan (1986: 1) เป็นวิธีการที่ได้รับความนิยมอย่างแพร่หลายเนื่องจากโมเดลที่ได้สามารถแปลความหมายและเข้าใจง่าย ลักษณะของรูปแบบข้อมูล (Pattern) ซึ่งแต่ละโหนด (Node) จะแสดงคุณลักษณะหรือแอตทริบิวต์ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งจะแสดงผลในการทดสอบ และลีฟโหนด (Leaf Node) จะแสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ (ชัชชฎา วันดี, 2557)

3.2) Random Forest เป็นวิธีการที่คล้ายกับ Bagging มากแต่เพิ่มการสร้างความหลากหลายของโมเดลด้วยการสุ่มแอตทริบิวต์ด้วยแทนที่จะเป็นการสุ่มเฉพาะข้อมูลตัวอย่างเพียงอย่างเดียวเหมือน Bagging และเทคนิคที่ใช้ในการสร้างโมเดลก็เป็นเพียงแค่ Decision Tree อย่างเดียว แม้ว่าจะเป็นเทคนิค Decision Tree เหมือนกันแต่ข้อมูลและแอตทริบิวต์ที่ใช้ในการสร้างโมเดลต่างกันก็ทำให้โมเดลที่สร้างขึ้นมาได้มีลักษณะที่ต่างกัน

4) การสุ่มตัวอย่างบนข้อมูลที่ไม่สมดุล ข้อมูลไม่สมดุล คือข้อมูลที่มีจำนวนข้อมูลของบางกลุ่มมีมากกว่า จำนวนข้อมูลของอีกกลุ่มอยู่เป็น จำนวนมาก (Chawla, 2005) โดยแบ่งเป็นข้อมูลส่วนมาก (Majority Class) และข้อมูลส่วนน้อย (Minority Class) (Chawla et al., 2002) ข้อมูลที่ไม่สมดุลจะส่ง ผลต่อการจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนน้อยได้ไม่ถูกต้อง หรือถูกต้องน้อย แต่

ในขณะเดียวกันจะ สามารถจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนมาก ได้ถูกต้องกว่า แต่ในความเป็นจริงกลุ่มข้อมูลที่มีขนาดน้อยนั้นอาจจะสำคัญมาก ส่งผลให้ตัวแบบในการจำแนกประเภทของข้อมูลมีประสิทธิภาพที่ต่ำลง งานวิจัยมุ่งเน้นการเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลที่มีขนาดน้อยให้มีความแม่นยำสูงขึ้น โดยวิธีการ sampling ข้อมูลด้วยการทำการสุ่มเพิ่มชุดข้อมูลตัวอย่าง

4.1) วิธีสุ่มเพิ่ม (Over sampling) วิธีสุ่มเพิ่ม (กิตติพงษ์, 2016) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิม

4.2) วิธีสุ่มลด (Under sampling) วิธีสุ่มลด (กิตติพงษ์, 2016) เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับ จำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย

4.3) วิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) วิธีสังเคราะห์ข้อมูลใหม่ (Chawla et al., 2002; Chawla, 2003, Chawla et al., 2004) เป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูล เดิมที่มีอยู่หลักการเพื่อนบ้านที่อยู่ใกล้ที่สุดในการขยายขอบเขตการตัดสินใจของตัวแบบ ซึ่งการสังเคราะห์ข้อมูลใหม่มีขั้นตอนดังนี้คือระบุเพื่อนบ้านที่ใกล้เคียงที่สุด k ค่าของข้อมูลเดิม สำหรับข้อมูลเดิม M หา k=l ที่มีระยะทางใกล้เคียงกับข้อมูลเดิมโดย l คือจำนวนเพื่อนบ้านที่ใกล้เคียงกับจุด M แล้วสุ่มเลือกจุด ระหว่างสองจุดและสร้างกรณีใหม่

5) ประเภทของอาชีพ กองวิจัยตลาดแรงงานและกองส่งเสริมการมีงานทำ กรมการจัดหางาน กระทรวงแรงงาน (2557: 27) ได้แบ่งประเภทอาชีพออกเป็นลักษณะใหญ่ ๆ ได้ 2 ลักษณะดังนี้

5.1) งานอาชีพที่ทำให้ผู้อื่น (Work for Others/Organization) งานอาชีพที่ทำให้ผู้อื่นหรือเป็นลูกจ้าง/พนักงานทำให้องค์กรที่เป็นรัฐบาลหรือองค์กรธุรกิจเอกชนหรือองค์กรการกุศลสาธารณะต่าง ๆ

5.2) งานอาชีพอิสระ (Self–Employment) งานอาชีพอิสระนี้เป็นการทำงานที่ไม่ต้องมาสมัครงานแต่เป็นการทำงานที่เกิดจากการตัดสินใจของผู้ที่จะกระทำเพื่อให้ตนเองหรือกระทำด้วยตนเองเป็นการจ้างตนเองโดยอาศัยปัจจัยความรู้ ความสามารถ โอกาส

วิธีดำเนินการวิจัย

1) การจัดเตรียมข้อมูล

1.1) ทำความสะอาดข้อมูล งานวิจัยนี้ได้มีการจัดเตรียมข้อมูลโดยใช้ข้อมูลภาวะการมีงานทำของบัณฑิตที่จบจากมหาวิทยาลัยราชภัฏชัยภูมิ ตั้งแต่ปีการศึกษา 2559 – 2561 เป็นจำนวน 2,005 รายการ

ตำแหน่งครูพี่เลี้ยงเด็ก	513120	โรงเรียนบ้านกุดน้ำใส(๓ พระครูอนุสร	P	1036100133	3	โรงเรียนบ้านกุดน้ำใส(๓ พระครูอนุสรณ์)
	419010	ไปรษณีย์ไทย	H	333	1	ไปรษณีย์จัตุรัส
	412230	กรุงเทพประกันชีวิต	K	270/9-11	6	
การพัฒนาชุมชน	111040	บ้านเนินสง่า	N	-	14	-
	123130	บริษัท คลังแบตเตอรี่ศรีพิพิธ เจริญอิ	G	51/2	5	-
	412230	ธนาคารออมสิน	K	344	1	ธนาคารออมสินจัตุรัส
ด้านการเกษตร	611120	ธุรกิจส่วนตัว	A	8	8	หมู่บ้านใหม่นาดี
	419010	องค์การบริหารส่วนตำบลรังงาม	S			
	522090	-	G	149/11	9	
	513120	โรงเรียนบ้านดลุคหิน	P	176	3	-
	131430	อู.กล้า การช่าง	G	171	1	-
ครูพี่เลี้ยง	513120	โรงเรียน	P	555	10	โรงเรียนศรีแก้งคร้อ
	522090	เจริญผลพลังไทย	G	59	59	-
	112020	องค์การบริหารส่วนตำบลหนองไผ่	O	302	3	-
	911120	ร้านอาหาร	S	-	-	-

ภาพที่ 1 ตัวอย่างข้อมูลที่ไม่สมบูรณ์

1.2) รวบรวมข้อมูลที่แยกการเก็บตามปี ซึ่งจะอยู่ในรูปแบบของ Excel File ดังตัวอย่างภาพที่

1.3) คัดเลือกคุณลักษณะที่ต้องการเพื่อนำไปสู่ขบวนการแปลงข้อมูลและจัดให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ ผู้วิจัยได้ทำการคัดเลือกคุณลักษณะดังตารางที่ 1

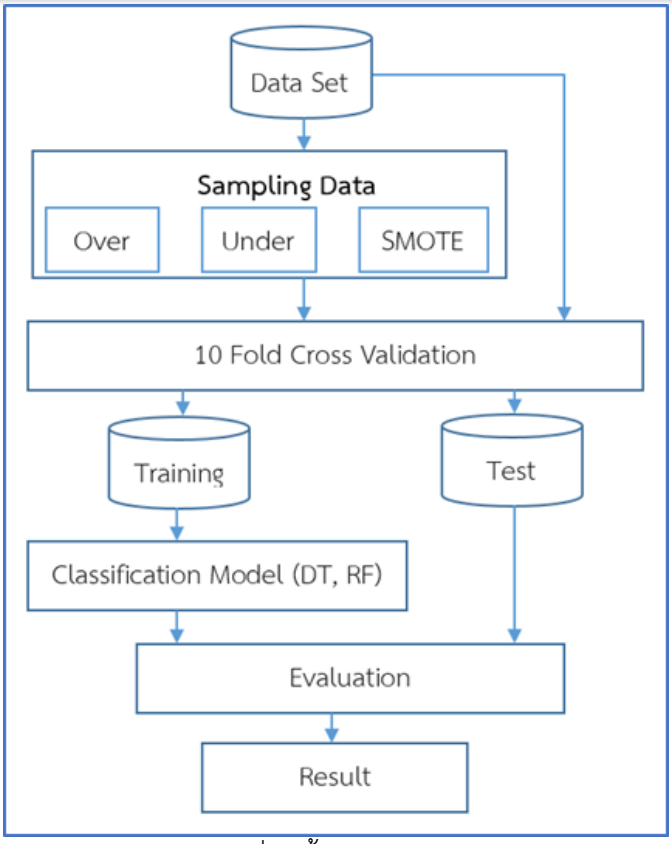
ตารางที่ 1 คุณลักษณะที่คัดเลือกเพื่อการวิเคราะห์สำหรับการจำแนกข้อมูล

ลำดับที่	คุณลักษณะ	ตัวอย่างข้อมูล
1	เพศ	ชาย
2	สาขาวิชา	วิทยาการคอมพิวเตอร์
3	ความสามารถพิเศษ	เล่นดนตรี
4	ผลการเรียนเฉลี่ย	3.54
5	ความสอดคล้องสาขาวิชา	ตรง
6	ประเภทอาชีพ/ตำแหน่ง	Class

1.4) จัดเตรียมข้อมูลโดยใช้ข้อมูลที่มีผลต่อการพยากรณ์อาชีพ ระหว่างปี พ.ศ. 2559 – 2561 จำนวน 2,005 รายการ ประกอบด้วย เพศ สาขาวิชา ความสามารถพิเศษ ผลการเรียนเฉลี่ย ความสอดคล้องสาขาวิชา และประเภทอาชีพ/ตำแหน่ง ข้อมูลดังตารางที่ 1 จากนั้น ผู้วิจัยได้นำข้อมูลไปปรับปรุงในส่วนของค่าที่ขาดหายไป โดยใช้วิธีการกำหนดค่า การเลือกคุณลักษณะเฉพาะที่เกี่ยวข้อง มีการปรับรูปแบบของข้อมูลบางคุณลักษณะเพื่อลดขนาดของข้อมูลที่ใช้ในการประมวลผล และทำการแบ่งช่วงข้อมูล (Discretization) เพื่อเปลี่ยนแปลงข้อมูลให้อยู่ในช่วงที่กำหนดจัดเป็นการลดระยะห่างของข้อมูล จากการปรับปรุงพบว่าข้อมูลที่พร้อมสำหรับการนำไปวิเคราะห์ครั้งนี้ประกอบด้วย 6 คุณลักษณะ

2) การปรับความสมดุลของชุดข้อมูลที่ไม่สมดุล ด้วยเทคนิค Over sampling เทคนิค Under sampling และเทคนิค SMOTE โดยใช้โปรแกรม WEKA

3) การสร้างโมเดลพยากรณ์และการประเมินโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส การสร้างโมเดลและการประเมินโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส โดยนำข้อมูลที่เตรียมไว้ตามคุณลักษณะตามตารางที่ 1 มาทำการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคแรนดอมฟอเรส (Random Forest) โดยใช้โปรแกรม WEKA จากนั้นทำการเปรียบเทียบผลจากทั้งสองวิธีเพื่อเลือกโมเดลที่ให้ผลการพยากรณ์ที่ดีที่สุด ด้วยมาตรวัดที่ใช้วัดประสิทธิภาพโมเดลพยากรณ์ด้วยค่า F-measurement, Precision, Recall และ Accuracy



ภาพที่ 2 ขั้นตอนการวิจัย

ผลการวิจัย

1) ผลการปรับความสมดุลของชุดข้อมูลที่ไม่สมดุล จากข้อมูลทั้งหมด โดยทำการสำเนาข้อมูลออกเป็น 4 ชุด กำหนดเป็น D1 คือข้อมูลที่ไม่ปรับความไม่สมดุล D2 คือชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Over sampling D3 คือชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Under sampling และ D4 คือชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค SMOTE แสดงดังตารางต่อไปนี้

ลำดับที่	คลาส	จำนวนข้อมูล
1	ดำเนินธุรกิจอิสระ/เจ้าของกิจการ (C04)	735
2	ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ (C01)	540
3	พนักงานบริษัท/องค์กรธุรกิจเอกชน (C03)	387
4	รัฐวิสาหกิจ (C02)	343
5	พนักงานองค์การต่างประเทศ/ระหว่างประเทศ (C05)	0

จากตารางที่ 2 จะพบว่าแต่ละคลาสจะมีจำนวนรายการของข้อมูลที่แตกต่างกันส่วนคลาสพนักงานองค์การต่างประเทศ/ระหว่างประเทศนั้นไม่มีจำนวนรายการดังนั้น งานวิจัยนี้จะเหลือจำนวนคลาสทั้งหมดอยู่ 4 คลาสดังต่อไปนี้ 1) ดำเนินธุรกิจอิสระ/เจ้าของกิจการ จำนวน 735 รายการ 2) ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ จำนวน 540 รายการ 3) พนักงานบริษัท/องค์กรธุรกิจเอกชน จำนวน 387 รายการ และ 4) รัฐวิสาหกิจ จำนวน 343 รายการ

ตารางที่ 3 ชุดข้อมูลที่ปรับความไม่สมดุล (D2)

ลำดับที่	คลาส	จำนวนข้อมูล
1	ดำเนินธุรกิจอิสระ/เจ้าของกิจการ (C04)	501
2	ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ (C01)	501
3	พนักงานบริษัท/องค์กรธุรกิจเอกชน (C03)	501
4	รัฐวิสาหกิจ (C02)	501

จากตารางที่ 3 จะพบว่าแต่ละคลาสจะมีจำนวนรายการของข้อมูลที่แตกต่างกันส่วนคลาสพนักงานองค์การต่างประเทศ/ระหว่างประเทศนั้นไม่มีจำนวนรายการดังนั้น งานวิจัยนี้จะเหลือจำนวนคลาสทั้งหมดอยู่ 4 คลาสดังต่อไปนี้ 1) ดำเนินธุรกิจอิสระ/เจ้าของกิจการ จำนวน 501 รายการ 2) ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ จำนวน 501 รายการ 3) พนักงานบริษัท/องค์กรธุรกิจเอกชน จำนวน 501 รายการ และ 4) รัฐวิสาหกิจ จำนวน 501 รายการ ปรับความไม่สมดุลด้วยเทคนิค Over sampling ปรับค่าพารามิเตอร์ดังนี้ weka.filters.supervised.instance.Resample -B 1.0 -S 1 -Z 100.0

ตารางที่ 4 ชุดข้อมูลที่ปรับความไม่สมดุล (D3)

ลำดับที่	คลาส	จำนวนข้อมูล
1	ดำเนินธุรกิจอิสระ/เจ้าของกิจการ (C04)	343
2	ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ (C01)	343
3	พนักงานบริษัท/องค์กรธุรกิจเอกชน (C03)	343
4	รัฐวิสาหกิจ (C02)	343

จากตารางที่ 4 จะพบว่าแต่ละคลาสจะมีจำนวนรายการของข้อมูลที่แตกต่างกันส่วนคลาสพนักงานองค์การต่างประเทศ/ระหว่างประเทศนั้นไม่มีจำนวนรายการดังนั้น งานวิจัยนี้จะเหลือจำนวนคลาสทั้งหมดอยู่ 4 คลาสดังต่อไปนี้ 1) ดำเนินธุรกิจอิสระ/เจ้าของกิจการ จำนวน 343 รายการ 2) ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ จำนวน 343 รายการ 3) พนักงานบริษัท/องค์กรธุรกิจเอกชน จำนวน 343 รายการ และ 4) รัฐวิสาหกิจ จำนวน 343 รายการ ปรับความไม่สมดุลด้วยเทคนิค Under sampling ปรับค่าพารามิเตอร์ดังนี้ weka.filters.supervised.instance.SpreadSubsample -M 1.0 -X 0.0 -S 1

ตารางที่ 5 ชุดข้อมูลที่ปรับความไม่สมดุล (D4)

ลำดับที่	คลาส	จำนวนข้อมูล
1	ดำเนินธุรกิจอิสระ/เจ้าของกิจการ (C04)	735
2	ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ (C01)	702
3	พนักงานบริษัท/องค์กรธุรกิจเอกชน (C03)	657
4	รัฐวิสาหกิจ (C02)	686

จากตารางที่ 5 จะพบว่าแต่ละคลาสจะมีจำนวนรายการของข้อมูลที่แตกต่างกันส่วนคลาสพนักงานองค์การต่างประเทศ/ระหว่างประเทศนั้นไม่มีจำนวนรายการดังนั้น งานวิจัยนี้จะเหลือจำนวนคลาสทั้งหมดอยู่ 4 คลาสดังต่อไปนี้ 1) ดำเนินธุรกิจอิสระ/เจ้าของกิจการ จำนวน 735 รายการ 2) ข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ จำนวน 702 รายการ 3) พนักงานบริษัท/องค์กรธุรกิจเอกชน จำนวน 657 รายการ และ 4) รัฐวิสาหกิจ จำนวน 686 รายการปรับความไม่สมดุลด้วย

เทคนิค SMOTE ปรับค่าพารามิเตอร์ดังนี้ weka.filters.supervised.instance.SMOTE -C 2 -K 5 -P 100.0 -S 1

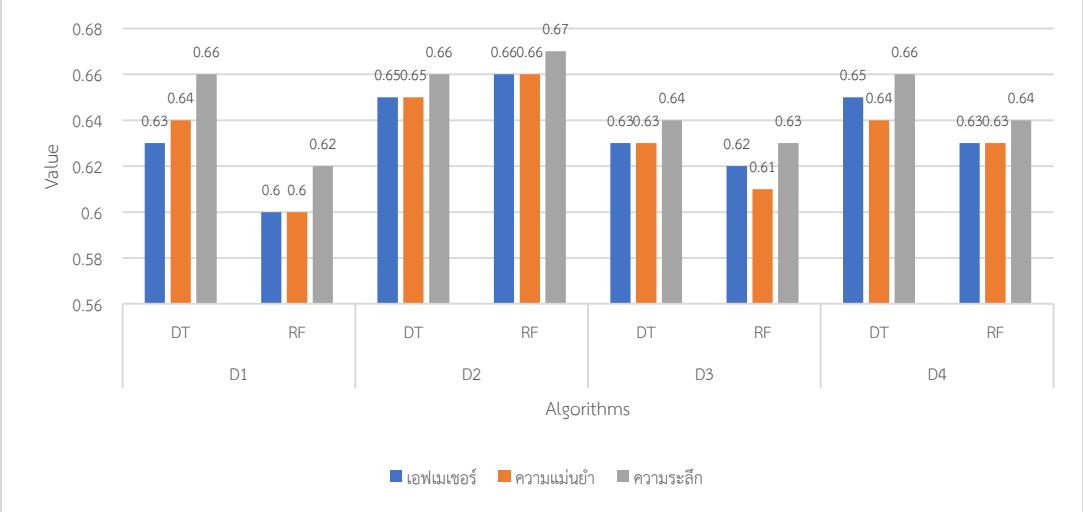
แต่ละคลาสจะมีจำนวนรายการของข้อมูลที่แตกต่างกันส่วนคลาสพนักงานองค์การต่างประเทศ/ระหว่างประเทศนั้นไม่มีจำนวนรายการเลยดังนั้น งานวิจัยนี้จะเหลือจำนวนคลาสทั้งหมดอยู่ 4 คลาสดังแสดงตารางที่ 3 - 5

2) ผลการสร้างโมเดลพยากรณ์และการประเมินโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส โดยนำข้อมูลที่เตรียมไว้ D1, D2, D3 และ D4 มาทำการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจและ เทคนิคแรนดอมฟอเรส โดยใช้โปรแกรม WEKA แบ่งข้อมูลออกเป็น 2 ส่วน 1) ส่วนเรียนรู้ (Training Set) และ 2) ส่วนทดสอบ (Test Set) แบบ Cross Validation 10-Folds จากนั้นทำการเปรียบเทียบผลการจำแนกข้อมูลจากทั้งสองวิธีพบว่า การเปรียบเทียบการจำแนกข้อมูลแยกตามชุดข้อมูลที่ไม่สมดุลและข้อมูลที่ผ่านการปรับสมดุลของข้อมูลแล้ว ด้วยเทคนิคต้นไม้ตัดสินใจและเทคนิคแรนดอมฟอเรส เปรียบเทียบประสิทธิภาพกันโดยใช้ค่าความแม่นยำ ค่าความระลึก ค่าเอฟเมเชอร์ และค่าความถูกต้อง ผลแสดงดังตารางที่ 6 ภาพที่ 3 และภาพที่ 4

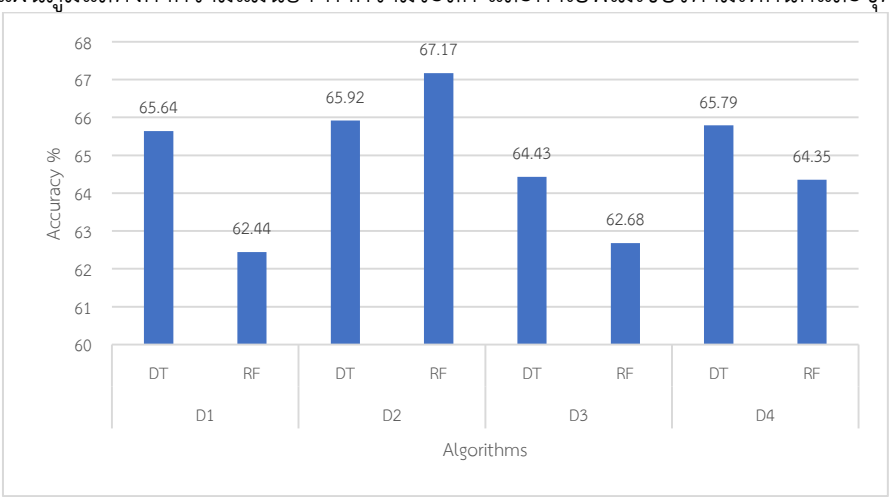
ตารางที่ 6 ตารางการเปรียบเทียบผลการวิเคราะห์การจำแนกข้อมูลของโมเดลพยากรณ์

ที่	ชุดข้อมูล	เทคนิค	ความถูกต้อง (%)	เอฟเมเชอร์	ความแม่นยำ	ความระลึก
1	D1	DT	65.64	0.63	0.64	0.66
		RF	62.44	0.60	0.60	0.62
2	D2	DT	65.92	0.65	0.65	0.66
		RF	67.17	0.66	0.66	0.67
3	D3	DT	64.43	0.63	0.63	0.64
		RF	62.68	0.62	0.61	0.63
4	D4	DT	65.79	0.65	0.64	0.66
		RF	64.35	0.63	0.63	0.64

จากตารางที่ 6 พบว่าโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Over sampling จะให้ประสิทธิภาพในการพยากรณ์ที่สูงกว่า ชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Under sampling ชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค SMOTE และชุดข้อมูลที่ไม่ปรับความไม่สมดุล ผลการพยากรณ์ของโมเดลจากชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Over sampling จะพบว่าเทคนิคแรนดอมฟอเรสให้ค่าถูกต้องสูงที่สุด คือ 67.17% รองลงมาคือเทคนิคต้นไม้ตัดสินใจค่าความถูกต้อง 65.92% ผลการพยากรณ์ของโมเดลจากชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Under sampling จะพบว่าเทคนิคต้นไม้ตัดสินใจให้ค่าถูกต้องสูงที่สุด คือ 64.43% รองลงมาคือเทคนิคแรนดอมฟอเรสค่าความถูกต้อง 62.68% ผลการพยากรณ์ของโมเดลจากชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค SMOTE จะพบว่าเทคนิคต้นไม้ตัดสินใจให้ค่าถูกต้องสูงที่สุด คือ 65.79% รองลงมาคือเทคนิคแรนดอมฟอเรสค่าความถูกต้อง 64.35% และผลการพยากรณ์ของโมเดลจากชุดข้อมูลที่ไม่ปรับความไม่สมดุลของข้อมูลจะพบว่าเทคนิคต้นไม้ตัดสินใจให้ค่าถูกต้องสูงที่สุด คือ 65.64% รองลงมาคือเทคนิคแรนดอมฟอเรสค่าความถูกต้อง 62.44%



ภาพที่ 3 แผนภูมิแสดงค่าความแม่นยำ ค่าความระลึก และค่าเอฟเมเซอร์ตามเทคนิคและชุดข้อมูล



ภาพที่ 4 แผนภูมิแสดงค่าความถูกต้องตามเทคนิคและชุดข้อมูล

อภิปรายผล

ผลการศึกษาโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส ใช้ข้อมูลภาวะการมีงานทำของบัณฑิตที่จบจากมหาวิทยาลัยราชภัฏชัยภูมิ ตั้งแต่ปีการศึกษา 2559 – 2561 เป็นจำนวน 2,005 รายการ ทำการปรับสมดุลของชุดข้อมูลที่ไม่สมดุลของคลาสทั้ง 4 คลาสด้วย 1) เทคนิค Over sampling ได้ชุดข้อมูลทั้งหมด 2,004 รายการ แยกตามคลาสดำเนินธุรกิจอิสระ/เจ้าของกิจการ 501 รายการ คลาสข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ 501 รายการ คลาสพนักงานบริษัท/องค์กรธุรกิจเอกชน 501 รายการ และคลาสรัฐวิสาหกิจ 501 รายการ 2) เทคนิค Under sampling ได้ชุดข้อมูลทั้งหมด 1,372 รายการ แยกตามคลาสดำเนินธุรกิจอิสระ/เจ้าของกิจการ 343 รายการ คลาสข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ 343 รายการ คลาสพนักงานบริษัท/องค์กรธุรกิจเอกชน 343 รายการ และคลาสรัฐวิสาหกิจ 343 รายการ และ 3) เทคนิค SMOTE ได้ชุดข้อมูลทั้งหมด 2,780 รายการ แยกตามคลาสดำเนินธุรกิจอิสระ/เจ้าของกิจการ 735 รายการ คลาสข้าราชการ /เจ้าหน้าที่หน่วยงานของรัฐ 702 รายการ คลาสพนักงานบริษัท/องค์กรธุรกิจเอกชน 657 รายการ และคลาสรัฐวิสาหกิจ 686 รายการ และสร้างโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วย 2 เทคนิคคือ เทคนิคต้นไม้ตัดสินใจและ เทคนิคแรนด

อมฟอเรส แล้วประเมินประสิทธิภาพโมเดลจากทั้ง 2 เทคนิคพบว่าเทคนิคแรนดอมฟอเรส ให้ประสิทธิภาพในการพยากรณ์สูงที่สุด บนชุดข้อมูลที่ปรับความไม่สมดุลด้วยเทคนิค Over sampling คือให้ ค่าความถูกต้องร้อยละ 67.17 ค่าความแม่นยำ 0.66 ค่าความระลึก 0.67 และค่าเอฟเมเชอร์ 0.66

การศึกษาโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส ผู้วิจัยได้นำชุดข้อมูลแบบไม่ปรับสมดุลและชุดข้อมูลที่ปรับสมดุลมาทำการเปรียบเทียบการจำแนกข้อมูลระหว่างชุดข้อมูลทั้งสองประเภทชุดข้อมูล โดยวิธีการทำเหมืองข้อมูลโดยใช้ 2 เทคนิค ได้แก่เทคนิควิธีต้นไม้ตัดสินใจ (Decision Tree : DT) และเทคนิคแรนดอมฟอเรส (Random Forest : RF) พบว่าชุดข้อมูลที่ปรับความไม่สมดุลของข้อมูลด้วยเทคนิค Over Sampling และใช้เทคนิคการจำแนกข้อมูลด้วยเทคนิคแรนดอมฟอเรสเป็นเทคนิคที่ให้ประสิทธิภาพมากที่สุด สามารถนำโมเดลนี้ไปเป็นแนวทางในการประยุกต์ใช้เพื่อจำแนกหมวดหมู่ของข้อมูลประเภทอื่น ๆ ได้

องค์ความรู้ที่ได้จากการศึกษา

การจำแนกข้อมูลด้วยเทคนิคเหมืองข้อมูลที่ใช้ชุดข้อมูลแบบหลายคลาส สามารถเพิ่มประสิทธิภาพการพยากรณ์ของโมเดลให้สูงขึ้นได้ด้วยการปรับสมดุลของชุดข้อมูลก่อน ด้วยเทคนิคการสุ่มข้อมูล และเทคนิคการจำแนกข้อมูล โดยใช้เทคนิคการสุ่มข้อมูลแบบ Over sampling ร่วมกับเทคนิคการจำแนกข้อมูลแบบแรนดอมฟอเรส

ข้อเสนอแนะ

การศึกษาโมเดลพยากรณ์อาชีพนักศึกษาทางธุรกิจด้วยเทคนิคการสุ่มตัวอย่างเรียนรู้บนชุดข้อมูลที่ไม่สมดุลแบบหลายคลาส จากโมเดลการพยากรณ์อาชีพ สามารถนำโมเดลนี้ไปเป็นแนวทางในการประยุกต์ใช้เพื่อจำแนกหมวดหมู่ของข้อมูลประเภทอื่น ๆ หรือนำไปพัฒนาระบบสารสนเทศเพื่อการแนะแนวอาชีพ การทำวิจัยในครั้งต่อไป ควรพิจารณาเทคนิคการสร้างตัวแบบการพยากรณ์ที่หลากหลาย การปรับแต่งค่าพารามิเตอร์ และจำนวนข้อมูลที่มากขึ้น เพื่อให้การพยากรณ์มีความถูกต้องและแม่นยำยิ่งขึ้น

บรรณานุกรม

กองวิจัยตลาดแรงงาน. (2559). [ออนไลน์]. การเตรียมความพร้อมการปฏิวัติอุตสาหกรรม ครั้งที่ 4.[สืบค้นเมื่อ 20 พฤศจิกายน 2559]. จาก http://lmi.doe.go.th/index.php/research/525-research48.

กองวิจัยตลาดแรงงานและกองส่งเสริมการมีงานทำ. (2557). แนวโน้มอาชีพอิสระในอนาคต 3 ปีข้างหน้า พ.ศ. 2558-2560. รายงานวิจัย. กรมการจัดหางาน กระทรวงแรงงาน.

กิตติพงษ์ ชมบุญ. (2016). เทคนิคการค้นหาคลาสที่ค้นพบได้ยากสำหรับข้อมูลที่ขนาดแตกต่างกันมาก. วิทยานิพนธ์วิศวกรรมศาสตร์ดุษฎีบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

ชัชชฎา วันดี. (2557). การศึกษาปจจัยที่มีผลตอการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จ

การศึกษา โดยใช้เทคนิคเหมืองข้อมูล. ปริญญานิพนธ์ หลักสูตรวิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยามหาสารคาม.

อุกฤษ ปัจฉิม. (2546). การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการทำนายระดับน้ำสูงสุด. วิทยานิพนธ์หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมทรัพยากรน้ำ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets : An Overview. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook ,853-867. Boston, MA: Springer US.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res., 16(1), 321-357.

Elhassan T, Aljurf M, Al-Mohanna F & Shoukri M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. Journal of Informatics and Data Mining, 1, 1 – 12.

Guillaume Lemaˆıtre, Fernando Nogueira & Christos K. Aridas. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18, 1 – 5.

Raisul Islam Rashu, Naheena Haq & Rashedur M Rahman. (2014). Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem. International Conference on Computer and Information Technology (ICCIT), 17, 14 – 19.