การเปรียบเทียบตัวแบบทำนายเงินเดือน ของบัณฑิตสาขาวิชาที่เกี่ยวข้องกับสถิติ จากมหาวิทยาลัยในประเทศไทย

A Comparative model for predicting salary of Statistics and related field graduate from Thailand universities

วิษณุวีร์ สุริยอมร[1] และนุชนาถ คงช่วย[2]

[1] ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์, 5810210445@psu.ac.th

[2] ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์, nootchanath.k@psu.ac.th

----

บทคัดย่อ

ข้อมูลการมีงานทำของบัณฑิตในประเทศไทย มีการเก็บรวบรวมอย่างต่อเนื่องโดยกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย และนวัตกรรม จากสถาบันอุดมศึกษาทั่วประเทศไทย ข้อมูลที่ได้มีจำนวนมาก มีมิติข้อมูลสูง รวมถึงมีความไม่สมบูรณ์ ในการวิเคราะห์การทำงาน รายได้ เป็นหนึ่งในปัจจัยสำคัญที่คนส่วนใหญ่ตัดสินใจเลือกทำงาน ในบทความนี้เป็นการศึกษาตัวแบบทำนายเงินเดือนของบัณฑิตที่จบจากสาขาวิชาที่เกี่ยวข้องกับสถิติ ซึ่งเป็นสายงานหนึ่งที่ได้รับความสนใจและมีรายได้สูงในปัจจุบัน โดยเปรียบเทียบความแม่นยำในการทำนายของตัวแบบการถดถอยเชิงเส้น และตัวแบบทาง Machine learning ด้วยเกณฑ์รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย และร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ผลการศึกษาพบว่าตัวแบบ Random forest มีความแม่นยำในการทำนายมากที่สุด แต่ใช้อธิบายการทำนายได้ยากเมื่อเปรียบเทียบกับตัวแบบการถดถอยเชิงเส้นที่สามารถอธิบายได้ง่ายที่สุด โดยพบว่าปัจจัยสำคัญต่อการทำนายเงินเดือนของบัณฑิต 5 อันดับแรกในตัวแบบ Random forest ประกอบด้วย ระดับการศึกษาระดับปริญญาโท ประเภทงานที่ไม่ได้ระบุ สาขาวิชาสถิติ ประเภทงานธุรกิจเอกชน และสาขาวิชาสถิติประยุกต์ และปัจจัยที่สำคัญ 5 อันดับแรกของตัวแบบการถดถอยเชิงเส้น ประกอบด้วยระดับการศึกษาปริญญาโท ระดับการศึกษาปริญญาเอก ปีการศึกษา 2560 ปีการศึกษา 2559 และระยะเวลาในการหางาน 1–2 เดือน ผลการทำนายที่ได้ช่วยประกอบการตัดสินใจของบัณฑิต หรือนักศึกษาในการตัดสนใจเข้าทำงาน

คำสำคัญ: การทำงานของบัณฑิต, เงินเดือน, การวิเคราะห์การถดถอย

ABSTRACT

Thailand's graduate employment data have been collected continuously from the Ministry of Higher Education, Science, Research and Innovation. The data are collected from different universities in Thailand but it a huge amount of data, incomplete and high dimensional. An analysis of graduate employment, salary is one of the important factors that make graduates decide for work. This paper is to study a predictive model to predict the salary of Statistics and related fields graduate by comparing a model's performance from statistical and machine learning models with RMSE and MAPE. The result got the most accuracy for predict graduate salary is a Random

forest model but Random forest is complicated than linear regression to describe a model. Top 5 of importance variables from Random forest are Master level, Undefined job, Statistics major, Private job and Applied statistics major. Top 5 of importance variables from linear regression are Master level, Ph.D. level, Year 2017, Year 2016 and Time find work 1 - 2 months. The conclusions of the study are useful to help graduates decide a direction for their work.

Keywords: Graduate employment, Salary, Regression analysis

1. บทนำ

จากระบบการศึกษาของประเทศไทย เมื่อจบการศึกษาในระดับขั้นพื้นฐานแล้ว ผู้เรียนส่วนใหญ่ยังคงมุ่งหวังที่จะเข้าศึกษาต่อในระดับอุดมศึกษา แต่ในกลุ่มผู้เข้าศึกษาเหล่านั้นยังมีบางส่วนที่เลือกเข้าศึกษาในหลักสูตรที่ไม่สอดคล้องกับความสามารถของตัวเอง ตามความต้องการของคนในครอบครัว หรือแนวโน้มในปัจจุบัน รวมถึงสาเหตุอื่น ๆ เช่น ปัญหาสุขภาพ หรือสถานะทางการเงิน ทำให้เมื่อเข้าศึกษาในระดับอุดมศึกษาแล้วเกิดปัญหาในการเรียนรวมไปถึงการออกกลางคัน (Tentsho et al., 2019) (Khongchai & Songmuang, 2016b) และจากความต้องการที่จะมีอาชีพการงานที่ดี มีรายได้ที่มั่นคง ซึ่งเป็นหนึ่งในความต้องการของนักศึกษาหลาย ๆ คน ทำให้การศึกษารายได้ของอาชีพที่จบการศึกษาไป สามารถช่วยให้นักศึกษาพิจารณาเลือกหลักสูตรที่จะเข้าศึกษา รวมถึงกระตุ้นตนเองให้ขยันหมั่นเพียรในการศึกษาต่อไป โดยจากการศึกษาแนวโน้มความต้องการแรงงานในประเทศไทย (สำนักงานสถิติแห่งชาติ, 2560) พบว่า อาชีพที่มีความต้องการแรงงานเพิ่มขึ้นในวุฒิการศึกษาระดับอุดมศึกษามากที่สุดคือ อาชีพในกลุ่มนักคณิตศาสตร์ และสถิติ โดยในอาชีพกลุ่มนี้สามารถประกอบอาชีพได้หลากหลาย เช่น นักบัญชี นักคณิตศาสตร์ประกันภัย เจ้าหน้าที่วิเคราะห์ข้อมูล เป็นต้น และมีรายได้ที่ค่อนข้างสูงในสังคมปัจจุบันที่ให้ความสำคัญกับงานด้านการวิเคราะห์ข้อมูล โดยจากรายงานคู่มือฐานเงินเดือนของบริษัท Adeco (Adeco Thailand, 2020) พบว่าเงินเดือนเฉลี่ยของบัณฑิตจบใหม่อยู่ที่ 15,000 บาท และในสายอาชีพนักวิเคราะห์ข้อมูลมีเงินเดือนอยู่ระหว่าง 35,000 – 40,000 บาท ถือเป็นสายอาชีพที่มีรายได้สูง นอกจากนั้นสายงานเหล่านี้ในต่างประเทศ (Glassdoor.com, 2020) มีเงินเดือนเฉลี่ย 70,000 – 100,000 USD ต่อปี ซึ่งเป็นอาชีพที่มีรายได้ที่สูงมากอาชีพหนึ่ง ทั้งนี้กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม ได้มีการเก็บรวบรวมข้อมูลการทำงานของบัณฑิตที่จบการศึกษาในแต่ละปีจากสถาบันอุดมศึกษาทั่วประเทศไทย สามารถนำข้อมูลนี้มาใช้วิเคราะห์การทำงาน และทราบถึงรายได้ของบัณฑิตหลังจบการศึกษาได้ แต่เนื่องจากมีการเก็บข้อมูลจากแหล่งที่แตกต่างกัน ทำให้ข้อมูลที่เก็บมีขนาดใหญ่ มีความซับซ้อน และอาจมีข้อมูลที่ไม่สมบูรณ์ นอกจากวิธีการทางสถิติที่ใช้ศึกษา สามารถใช้วิธีการทางวิทยาการข้อมูลมาประยุกต์ใช้เพื่อจัดการกับข้อมูลให้เหมาะสมได้ ซึ่งวิธีการนี้เป็นที่ได้รับการยอมรับว่าเหมาะสมต่อการประยุกต์ใช้ข้อมูลในชีวิตจริง เช่น การประกันภัย การวางแผนธุรกิจ อุตุนิยมวิทยา หรือวิทยาศาสตร์สุขภาพ เป็นต้น

จากงานวิจัยที่เกี่ยวข้องกับการศึกษารายได้ของบัณฑิต เช่น (หลายพสุ et al., 2018) ได้ศึกษาความแตกต่างระหว่างเงินเดือนของบัณฑิตคณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ด้วยสถิติทดสอบ Mann-Whitney Test และ (Khongchai & Songmuang, 2016a) (Khongchai & Songmuang, 2016c) ได้พัฒนาระบบการทำนายเงินเดือนด้วยตัวแบบทำนายเงินเดือนของนักศึกษาจากเทคนิคการทำเหมืองข้อมูล โดยใช้ข้อมูลส่วนตัวของนักศึกษาแต่ละคน ตัวแบบทำนายเหล่านี้สามารถใช้ข้อมูลจากกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย

และนวัตกรรม มาใช้สร้างตัวแบบทำนายโดยใช้ตัวแปรที่สัมพันธ์กับเงินเดือนของบัณฑิตผ่านวิธีการทางสถิติ และวิธีวิทยาการข้อมูล

2. วัตถุประสงค์การวิจัย

ประยุกต์ใช้การวิเคราะห์ข้อมูลทางสถิติ และวิทยาการข้อมูล เพื่อทำนายเงินเดือนของบัณฑิตที่จบจากสาขาวิชาที่เกี่ยวข้องกับสถิติ ในประเทศไทย และเปรียบเทียบประสิทธิภาพของตัวแบบทำนายวิธีทางสถิติ และวิทยาการข้อมูล

3. การดำเนินการวิจัย

การวิเคราะห์การถดถอย (Regression analysis) เป็นเครื่องมือที่ใช้วิเคราะห์ข้อมูลเชิงสถิติ สำหรับการศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่สองตัวขึ้นไป โดยพิจารณารูปแบบที่เป็นไปได้ของความสัมพันธ์ และเพื่อทำนายข้อมูลเชิงปริมาณ วิธีการนี้เป็นวิธีการทางสถิติที่มีประโยชน์ และนิยมใช้กันอย่างแพร่หลาย ทั้งยังเป็นจุดเริ่มต้นสำหรับการวิเคราะห์ข้อมูลในสมัยใหม่ที่เป็นส่วนขยายของการวิเคราะห์การถดถอย (จั่นสกุล & วิกยานนท์, 2012) (Gareth James et al., 2013) สำหรับการศึกษาครั้งนี้ใช้โปรแกรมวิเคราะห์ข้อมูลทางสถิติ R Software version 3.6.2 เพื่อวางแผน และวิเคราะห์ข้อมูลด้วยตัวแบบทางสถิติ และตัวแบบ Machine learning

ตัวแบบการถดถอยเชิงเส้น (Linear regression model) เป็นวิธีการที่ใช้ทำนายตัวแปรตาม Y จากตัวแปรทำนาย X จากการประมาณค่าด้วยความสัมพันธ์เชิงเส้น ดังสมการที่ (1) หรือเรียกว่า ตัวแบบการถดถอยเชิงเส้นเชิงเดียว พารามิเตอร์ β_0 และ β_1 ในตัวแบบการถดถอย คือสัมประสิทธิ์การถดถอยแสดงถึงค่าคงที่ และความชันของสมการ โดยตัวแบบการถดถอยเชิงเส้นเป็นตัวแบบพื้นฐานที่ใช้ทำนายข้อมูลเชิงปริมาณ ทำให้สามารถแปรผลของตัวแบบได้ง่ายเมื่อเทียบกับตัวแบบที่มีความซับซ้อน เช่นงานวิจัยของ (Hamlen & Hamlen, 2016) ที่ได้ศึกษาตัวแบบทำนายเงินเดือนของบัณฑิตที่จบจากหลักสูตร MBA ด้วยตัวแบบการถดถอยเชิงเส้น โดยสามารถแสดงโครงสร้างของตัวแบบ และระบุตัวแปรที่สำคัญต่อเงินเดือนของบัณฑิตได้ชัดเจน

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X \tag{1}$$

เมื่อ Y คือ ตัวแปรตาม และพารามิเตอร์ β_0 และ β_1 คือ ค่าสัมประสิทธิ์การถดถอย

ตัวแบบต้นไม้ตัดสินใจ (Decision tree) เป็นตัวแบบที่มีลักษณะคล้ายกับแผนภาพต้นไม้ โดยมีแต่ละจุดของแผนภาพแสดงถึงตัวแปรที่ใช้ในการทำนายค่า แต่ละกิ่งแสดงค่าที่ทำการจำแนก และใบแสดงผลการทำนาย (Gareth James et al., 2013) และในงานวิจัยของ (Khongchai & Songmuang, 2016b) ได้ประยุกต์ใช้ตัวแบบต้นไม้ตัดสินใจจากข้อมูลผลการศึกษา และการทำงานเพื่อทำนายเงินเดือนของบัณฑิตที่จะได้รับ โดยได้สร้างเป็นแอพพลิเคชั่นเพื่อให้ผู้ใช้งานสามารถเพิ่มข้อมูล และแปลผลการทำนายให้เข้าใจง่ายขึ้น นอกจากนั้นตัวแบบนี้สามารถประยุกต์เข้ากับการศึกษาในหัวข้ออื่น ๆ ได้มากมาย และเป็นที่นิยมอย่างแพร่หลาย เช่น (Hamoud et al., 2018) ได้ประยุกต์ใช้ตัวแบบต้นไม้ตัดสินใจทำนายผลการเรียนของนักศึกษา หรือในงานวิจัยของ (Quadri & N.V. Kalyankar, 2010) ได้ใช้ตัวแบบต้นไม้ตัดสินใจทำนายการออกกลางคันของนักศึกษา และแสดงแผนภาพต้นไม้ทำให้ง่ายต่อการตีความตัวแบบ

ตัวแบบ Random forest เป็นตัวแบบที่พัฒนาต่อมาจากตัวแบบต้นไม้ตัดสินใจ โดยมีแนวคิดจากการสร้างตัวแบบต้นไม้ตัดสินใจหลายตัวแบบ ซึ่งแต่ละตัวแบบได้รับข้อมูล และตัวแปรที่ใช้แตกต่างกัน จากนั้นทำการรวมผลที่ได้จากทุกตัวแบบเพื่อหาค่าที่ทำนายที่เหมาะสมที่สุด เนื่องจากตัวแบบนี้เกิดจากการสุ่มของชุดข้อมูล จึงต้องหาค่าที่เหมาะสมในการปรับค่าพารามิเตอร์ของตัวแบบ เพื่อให้ได้ตัวแบบที่เหมาะสมที่สุด จากงานของ (Parmar et al., 2014) ที่ศึกษาการวิเคราะห์ความพึงพอใจของการชมภาพยนตร์โดยใช้ตัวแบบ Random Forest ผลที่ได้พบว่า ตัวแบบนี้สามารถทำนายได้ดี และจำเป็นต้องให้ความสนใจเป็นพิเศษต่อการปรับค่าพารามิเตอร์ การประยุกต์ใช้ตัวแบบ Random forest กับการศึกษาการทำงานของบัณฑิตในงานวิจัย (Khongchai & Songmuang, 2016c) ที่ได้เปรียบเทียบผลการทำนายเงินเดือนของบัณฑิตของตัวแบบต้นไม้ตัดสินใจ และตัวแบบ Random forest พบว่า ตัวแบบ Random forest มีความแม่นยำในการทำนายมากที่สุด และมีค่าสูงถึง 90% แต่หากเปรียบเทียบกับตัวแบบประเภท Deep learning ในงานของ (Viroonluecha & Kaewkiriya, 2018) ที่เปรียบเทียบตัวแบบทำนายเงินเดือนด้วยตัวแบบ DNN ตัวแบบ Random forest และตัวแบบ Gradient boosted trees พบว่า ตัวแบบโครงข่ายประสาทเทียมเชิงลึก มีประสิทธิภาพในการทำนายมากที่สุด ตัวแบบเหล่านี้ใช้อธิบายการทำงานของตัวแบบได้ยากกว่าตัวแบบการถดถอยเชิงเส้น หรือตัวแบบต้นไม้ตัดสินใจ แต่มีประสิทธิภาพในการทำนายของตัวแบบที่สูงมาก

ตัวแบบ k - Nearest neighbors (KNN) เป็นตัวแบบอย่างง่ายรูปแบบหนึ่งที่ใช้เรียนรู้จากชุดข้อมูล การทำนายค่าของตัวแปรที่ศึกษาจะใช้การเปรียบเทียบจากข้อมูลที่ใช้ศึกษา และใช้ชุดข้อมูลที่มีความคล้ายกันมากที่สุด เพื่อทำนายตัวแปรทำนายให้มีค่าเดียวกับข้อมูลชุดนั้น หรือใช้การเฉลี่ยค่าที่ได้จากข้อมูลที่ทำการเปรียบเทียบด้วย (Gareth James et al., 2013) และการกำหนดความคล้ายของแต่ละชุดข้อมูล โดยตัวแบบนี้จะใช้การวัดระยะห่างของจุดด้วย Euclidean distance โดยมีรูปแบบของฟังก์ชันดังสมการที่ (2) นอกจากนั้นตัวแบบ KNN จะพิจารณาจำนวณของชุดข้อมูลที่ใช้ทำนาย โดยปกติจำนวณของชุดข้อมูลที่ใช้พิจารณาจะเป็นจำนวนคี่ และทำการหาค่าที่ทำนายเหมือนกันมากที่สุด การหาค่าที่เหมาะสมของของจำนวนชุดข้อมูลที่ใช้ ส่งผลให้ได้ตัวแบบที่มีประสิทธิภาพเช่นกัน ในงานวิจัย (Khongchai & Songmuang, 2016a) ได้ศึกษาการทำนายเงินเดือนของบัณฑิตโดยเปรียบเทียบตัวแบบต้นไม้ตัดสินใจ ตัวแบบ Naïve Bayes ตัวแบบ KNN ตัวแบบ SVM และตัวแบบ ANN พบว่าตัวแบบ KNN มีประสิทธิภาพในการทำนายมากที่สุดเมื่อเปรียบเทียบกับตัวแบบอื่น ๆ ที่ทำการศึกษา และจากงานวิจัย (Rahman et al., 2017) ที่ศึกษาตัวแบบทำนายภาวะการทำงานของบัณฑิตด้วยตัวแบบต้นไม้ตัดสินใจ ตัวแบบ Naïve Bayes และตัวแบบ KNN พบว่าตัวแบบ KNN มีความแม่นยำมากที่สุดเช่นเดียวกัน

$$d_{euclidean}(x, y) = \sqrt{\Sigma_i (x_i - y_i)^2} \quad (2)$$

3.1 การจัดการข้อมูลเบื้องต้น

จากวัตถุประสงค์เพื่อสร้างตัวแบบทำนายเงินเดือนของบัณฑิตที่จบจากสาขาวิชาที่เกี่ยวข้องกับสถิติในประเทศไทย โดยใช้ข้อมูลทุติยภูมิจากฐานข้อมูลภาวะการทำงานของบัณฑิตจากกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย และนวัตกรรม ข้อมูลที่ได้มาจากสถาบันอุดมศึกษาทั่วประเทศไทย ระหว่างปีการศึกษา 2556 – 2560 จำนวนทั้งสิ้น 703,777 ข้อมูล โดยขั้นตอนการจัดการข้อมูลเบื้องต้น จะทำการรวมข้อมูลทั้งหมดเข้าด้วยกัน และคัดกรองข้อมูลบัณฑิตที่จบจากสาขาวิชาที่เกี่ยวข้องกับสถิติ เช่น สถิติศาสตร์ สถิติประยุกต์ คณิตศาสตร์ประยุกต์ เป็น

ต้น และจัดเรียงตามปีที่เก็บข้อมูล จากนั้นทำให้ข้อมูลสมบูรณ์ โดยการลบข้อมูลที่ไม่สามารถใช้ได้ หรือการแทนค่าข้อมูลสูญหายด้วยวิธี KNN Imputation ซึ่งจากงานวิจัย (Minakshi et al., 2014) การแทนค่าข้อมูลสูญหายด้วยวิธี KNN เป็นวิธีที่ดีกว่าการแทนค่าด้วยวิธีอื่น ๆ และทำให้ตัวแบบที่ได้มีความแม่นยำในการทำนายเพิ่มขึ้นเช่นกัน

ข้อมูลบัณฑิตที่ได้จากการคัดกรองประกอบไปด้วยข้อมูลบัณฑิตที่ทำงานแล้ว และบัณฑิตที่ยังไม่มีงานทำ ซึ่งในส่วนของบัณฑิตที่ยังไม่มีงานทำ ข้อมูลรายได้ และการทำงาน จะไม่ถูกบันทึกไว้ หรือมีค่าเป็นศูนย์ ในส่วนนี้จะเลือกข้อมูลบัณฑิตที่ทำงานแล้วมาใช้ในการวิเคราะห์เท่านั้น และข้อมูลที่บันทึกจากแต่ละสถาบันอุดมศึกษานั้นแตกต่างกันตามการเก็บข้อมูลของสถาบันศึกษา ทำให้ข้อมูลที่ได้ยังมีส่วนของการตอบแบบสอบถามที่แตกต่างจากแบบฟอร์มที่กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย และนวัตกรรม ได้จัดทำไว้ จึงต้องคัดกรองเฉพาะข้อมูลที่ตรงกับแบบฟอร์มหลักเท่านั้น และเพิ่มตัวแปรที่ใช้ในการศึกษาเพิ่มเติม คือ ประเภทของมหาวิทยาลัย อันดับของมหาวิทยาลัย สถานะมหาวิทยาลัยวิจัยแห่งชาติ และภูมิภาคของมหาวิทยาลัย โดยตัวแปรที่ใช้ในการศึกษาทั้งหมดแสดงได้ดังตารางที่ 1

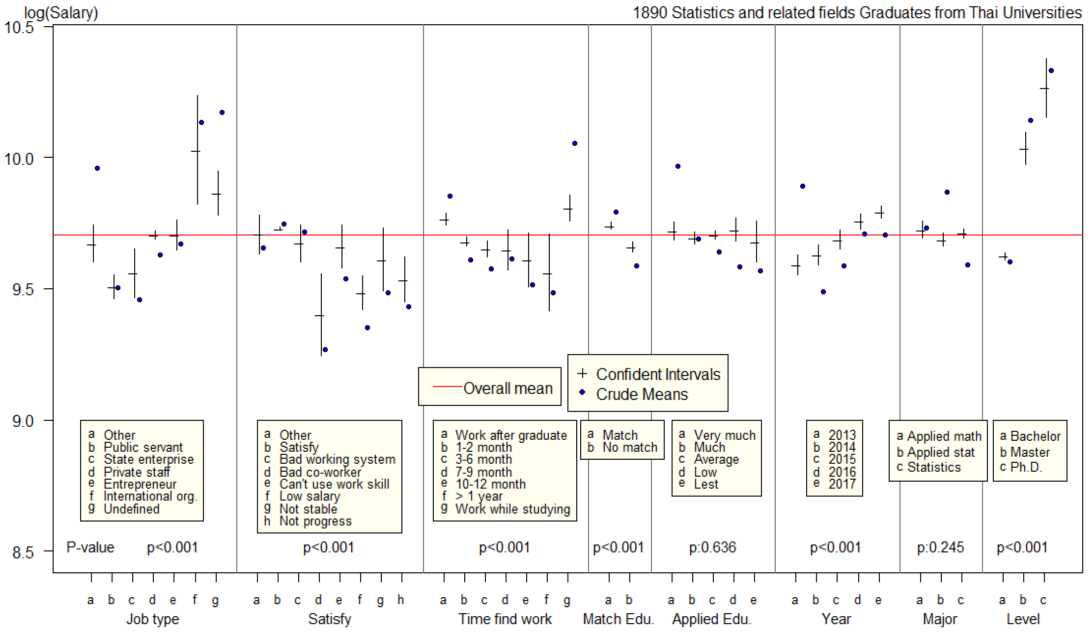
ตารางที่ 1 ตัวแปรที่ใช้ในการศึกษา

No.	ตัวแปร	ค่าของตัวแปร	No.	ตัวแปร	ค่าของตัวแปร
1.	เงินเดือน	3,000, 5,000, …, 60,000	9.	ระดับการศึกษา	ปริญญา ตรี, โท, เอก
2.	ประเภทงานที่ทำ	ราชการ, เอกชน, ธุรกิจ, …	10.	สาขาวิชาที่จบ	สถิติ, สถิติประยุกต์, คณิตประยุกต์
3.	จังหวัดที่ทำงาน	กระบี่, กทม., กาญจนบุรี, …	11.	มหาวิทยาลัยที่จบ	BUU, CU, CMU, …, TU, TSU
4.	ความพึงพอใจในการทำงาน	พอใจ, ระบบงานไม่ดี, เพื่อนร่วมงานไม่ดี,…	12.	ปีการศึกษาที่จบ	2556, 2557, 2558, 2559, 2560
5.	ระยะเวลาหางาน	1-2 เดือน, 3-6 เดือน, …	13.	ประเภทของมหาวิทยาลัย	ในกำกับรัฐ, รัฐบาล
6.	ตรงสาขาวิชาที่จบ	ตรงสาขา, ไม่ตรงสาขา	14.	อันดับของมหาวิทยาลัย	อยู่ใน THE Ranking, ไม่อยู่ใน THE Ranking
7.	ประยุกต์ใช้ความรู้กับงาน	มากที่สุด, มาก, ปานกลาง, น้อย, น้อยที่สุด	15.	มหาวิทยาลัยวิจัยแห่งชาติ	มหาวิทยาลัยวิจัยแห่งชาติ, ไม่ใช่
8.	ความต้องการศึกษาต่อ	ต้องการศึกษาต่อ, ไม่ต้องการ	16.	ภูมิภาคของมหาวิทยาลัย	เหนือ, กลาง, ใต้, ตะวันออก, ตะวันออกเฉียงเหนือ

การคัดเลือกตัวแปรที่เหมาะสมสำหรับการศึกษาครั้งนี้ ใช้การพิจารณาด้วยกราฟของช่วงความเชื่อมั่นที่ปรับด้วยค่าเฉลี่ย (Confident interval for adjusted mean) และเลือกตัวแปรที่มีนัยสำคัญเพื่อนำไปสร้างตัวแบบทำนายต่อไป ข้อดีของวิธีการนี้จะสามารถเห็นภาพของช่วงความเชื่อมั่นของตัวแปรแต่ละระดับได้อย่างชัดเจน และสำหรับตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพจะไม่ต้องสร้างตัวแปรหุ่นเพื่อใช้เทียบระดับของแต่ละตัวแปร (Tongkumchum & McNeil, 2009) โดยสามารถประยุกต์ใช้ร่วมกับตัวแบบการถดถอยเชิงเส้น หรือการวิเคราะห์จำแนกที่แสดงช่วงความเชื่อมั่นจากการปรับค่าสัดส่วน (Kongchouy & Sampantarak, 2010) ตัวแปรที่ผ่านการคัดเลือกจากการพิจารณาช่วงความเชื่อมั่นประกอบไปด้วย ประเภทงานที่ทำ ความพึงพอใจในการทำงาน ระยะเวลาหางาน ทำงานตรงกับสาขาวิชา การประยุกต์ใช้ความรู้กับงาน ปีการศึกษาที่จบ ระดับการศึกษา และสาขาวิชา โดยชุดข้อมูลที่ใช้ใน

การศึกษาครั้งนี้ ประกอบไปด้วย 8 ตัวแปร ตัวอย่างบัณฑิตที่จบการศึกษาในช่วง 5 ปี จำนวน 1,890 คน ตัวแปรที่ผ่านการคัดเลือกจากการพิจารณาช่วงความเชื่อมั่น แสดงดังรูปที่ 1 จะสังเกตว่าถ้าช่วงความเชื่อมั่นของตัวแปรในระดับใดไม่ซ้อนทับกัน และไม่ตัดกับเส้นค่าเฉลี่ยรวมของเงินเดือน แสดงให้เห็นว่าปัจจัยนั้นส่งผลให้เงินเดือนของบัณฑิตคนนั้นแตกต่างกับปัจจัยอื่นอย่างชัดเจน

การแปลงค่าข้อมูลสำหรับการสร้างตัวแบบการถดถอยเชิงเส้น ข้อมูลเงินเดือนบัณฑิตที่ใช้ในการศึกษาครั้งนี้ไม่มีการแจงแจงแบบปรกติ และจากการแปลงค่าข้อมูลด้วยวิธี Box-Cox พบว่า การแปลงค่าตัวแปรเงินเดือนของบัณฑิตให้อยู่ในรูปค่าลอการิทึม เป็นค่าที่เหมาะสมที่สุดในการสร้างตัวแบบการถอถอยต่อไป ซึ่งในรูปที่ 1 แสดงเงินเดือนของบัณฑิตในรูปลอการิทึม และปรับปรุงประสิทธิภาพของตัวแบบในการศึกษานี้เพื่อลดการเกิดปัญหา Over fitting โดยใช้วิธีการแบ่งชุดข้อมูลเป็น 2 ส่วน คือ ชุดที่ใช้สร้างตัวแบบ และชุดที่ใช้ทดสอบตัวแบบในอัตราส่วน 70 : 30 จากนั้นข้อมูลชุดที่ใช้สร้างตัวแบบใช้การปรับปรุงประสิทธิภาพด้วยวิธี k – fold cross validation และการปรับค่าพารามิเตอร์ให้เหมาะสมเพื่อให้ได้ตัวแบบที่เหมาะสมที่สุดในการทำนายเงินเดือนของบัณฑิต



รูปที่ 1 กราฟช่วงความเชื่อมั่นของตัวแบบการถดถอยเชิงเส้น

3.2 การวัดประสิทธิภาพของตัวแบบ

ในการศึกษาครั้งนี้ เลือกใช้เกณฑ์การวัดค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error: MAPE) เป็นเกณฑ์การวัดประสิทธิภาพของตัวแบบการถดถอยเชิงเส้น เพื่อวัดความแม่นยำในการทำนายค่าของตัวแปรตามว่าใกล้เคียงกับค่าจริงของตัวแปรตามที่กำหนดไว้หรือไม่ ในการวิเคราะห์การถดถอยโดยทั่วไปนิยมใช้เกณฑ์การวัดค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย โดยเป็นการวัดค่าความแตกต่างระหว่างค่าของตัวแปรที่ได้จากชุดข้อมูลจริง และค่าที่ได้

จากการทำนาย ดังสมการที่ (3) และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยเป็นการหาค่าร้อยละของความคลาดเคลื่อนในทำนาย ที่ไม่คำนึงถึงเครื่องหมายดังสมการที่ (4) หากเกณฑ์การวัดทั้งสองค่ามีค่าน้อย แสดงให้เห็นว่าตัวแบบทำนายมีความแม่นยำในการทำนายสูงเข้าใกล้ค่าจริง

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{predict} - y_{actual}\right)^2} \quad (3)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_{predict} - y_{actual}|}{y_{actual}} \quad (4)$$

4. ผลการวิจัย

จากรูปที่ 1 แสดงระดับนัยสำคัญของแต่ละตัวแปรด้านล่างของรูป และช่วงความเชื่อมั่นที่ปรับด้วยค่าเฉลี่ยของค่าลอการิทึมของเงินเดือน จะเห็นว่าช่วงความเชื่อมั่นในระดับการศึกษาปริญญาเอกมีค่าเฉลี่ยของเงินเดือนมากที่สุด เมื่อเทียบกับตัวแปรอื่น ๆ และมีค่าสูงกว่าค่าเฉลี่ยรวม รองลงมาคือระดับปริญญาโท แต่ในระดับปริญญาตรี มีช่วงความเชื่อมั่นของเงินเดือนต่ำกว่าค่าเฉลี่ยรวม นั่นคือ ตัวแปรระดับการศึกษาแสดงให้เห็นการเปลี่ยนแปลงของเงินเดือนจากตัวแบบการถดถอยเชิงเส้นได้ชัดเจนมากที่สุด สำหรับตัวแปรการทำงานตรงสาขา มีช่วงความเชื่อมั่นแตกต่างกัน และมีช่วงที่แคบทั้ง 2 กลุ่ม แต่เป็นตัวแปรที่มีค่าเฉลี่ยของเงินเดือนใกล้เคียงกับค่าเฉลี่ยรวม และในตัวแปรประเภทการทำงานองค์กรต่างประเทศ มีค่าเฉลี่ยของเงินเดือนสูงใกล้เคียงระดับปริญญาโท แต่มีช่วงความเชื่อมั่นกว้างมาก เนื่องจากมีจำนวนข้อมูลน้อย และในตัวแปรการทำงานประเภทอื่น ๆ มีช่วงความเชื่อมั่นที่ทับซ้อนกัน และคาบเกี่ยวกับค่าเฉลี่ยรวม จึงพิจารณาการเปลี่ยนแปลงของเงินเดือนได้ยาก

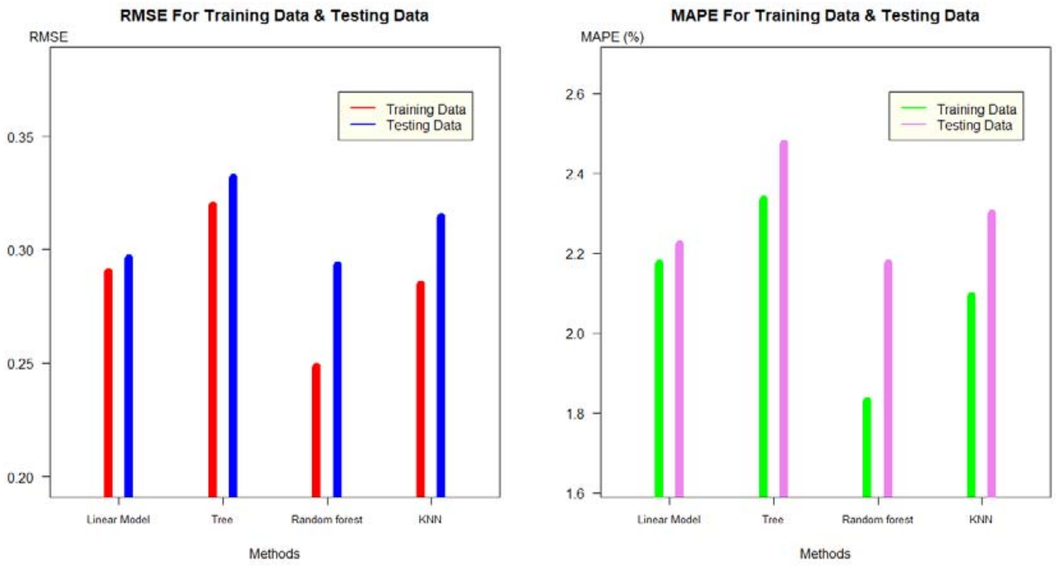
ผลการศึกษาตัวแบบทำนายเงินเดือนของบัณฑิต จะเปรียบเทียบผลที่ได้จากการทำนายของตัวแบบทั้ง 4 วิธี โดยระหว่างกระบวนการสร้างตัวแบบใช้การปรับปรุงประสิทธิภาพของตัวแบบวิธี 10 – Fold cross validation ด้วยโปรแกรมทางสถิติ R Program ซึ่งในขั้นตอน Cross validation จะแบ่งข้อมูลออกเป็น 10 ส่วนเท่ากัน จากนั้นใช้ข้อมูล 9 ส่วนแรกเป็นชุดข้อมูลสำหรับสร้างตัวแบบ และส่วนสุดท้ายเป็นข้อมูลสำหรับทดสอบ จากนั้นทำซ้ำโดยเปลี่ยนข้อมูลชุดทดสอบไปจนครบทุกชุดข้อมูลที่แบ่งไว้ และเปรียบเทียบการสร้างตัวแบบ 4 วิธีคือ ตัวแบบการถดถอยเชิงเส้น ตัวแบบต้นไม้ตัดสินใจ ตัวแบบ Random forest และตัวแบบ KNN โดยเปรียบเทียบผลการทำนายของตัวแบบด้วยเกณฑ์การวัดค่า RMSE และค่า MAPE

ตารางที่ 2 แสดงค่า RMSE และค่า MAPE ของตัวแบบการถดถอยทั้ง 4 ตัวแบบโดยจะเปรียบเทียบผลที่ได้ของข้อมูลชุดที่ใช้สร้างตัวแบบ และข้อมูลชุดที่ใช้ทดสอบตัวแบบ ซึ่งตัวแบบที่ดีควรจะสามารถทำนายได้ดีทั้งข้อมูลที่ใช้สร้าง และข้อมูลที่ไม่รู้จักมาก่อน โดยผลที่ได้จากการศึกษา พบว่า ตัวแบบที่มีค่า RMSE ที่น้อยที่สุดคือตัวแบบ Random forest ซึ่งในข้อมูลชุดสร้างตัวแบบมีค่า 0.2484 และในข้อมูลชุดทดสอบมีค่า 0.2932 นั่นคือตัวแบบนี้สามารถทำนายเงินเดือนของบัณฑิตคลาดเคลื่อนเฉลี่ย 0.2484 และ 0.2932 หน่วย ของค่าลอการิทึมของเงินเดือน รองลงมา คือตัวแบบ KNN ในข้อมูลชุดสร้างตัวแบบมีค่า 0.2846 และในข้อมูลชุดทดสอบมีค่า 0.3146 ทั้งสองตัวแบบข้างต้นมีค่า RMSE ที่ได้ในข้อมูลชุดสร้างตัวแบบ และชุดทดสอบมีความแตกต่างกันอย่างเห็นได้ชัด เมื่อพิจารณาที่ตัวแบบการถดถอยเชิงเส้น จะเห็นว่าค่า RMSE ที่ได้ไม่แตกต่างกันมากนัก ซึ่งสามารถพิจารณาได้จากรูปที่ 2

ในส่วนของค่า MAPE ดังตารางที่ 2 แสดงร้อยละของค่าคลาดเคลื่อนเฉลี่ยของการทำนายจากตัวแบบทั้ง 4 วิธี ผลการศึกษาที่ได้พบว่า ตัวแบบที่มีค่า MAPE ที่น้อยที่สุดคือตัวแบบ Random forest ซึ่งในข้อมูลชุดสร้างตัวแบบมีค่า 1.83% และในข้อมูลชุดทดสอบมีค่า 2.17% นั่นคือ ตัวแบบนี้สามารถทำนายเงินเดือนของบัณฑิตคลาดเคลื่อนเฉลี่ยเพียงร้อยละ 1.83% และ 2.17% ตามลำดับ รองลงมาในข้อมูลชุดสร้างตัวแบบคือตัวแบบ KNN มีค่า MAPE 2.09% แต่ในข้อมูลชุดทดสอบ ตัวแบบการถดถอยเชิงเส้นมีค่า MAPE 2.22% ซึ่งน้อยกว่าตัวแบบ KNN เมื่อพิจารณารูปที่ 2 จะเห็นว่าตัวแบบทั้ง 4 วิธี มีค่า MAPE ของข้อมูลในชุดสร้างตัวแบบ และชุดทดสอบตัวแบบแตกต่างกัน ซึ่งมีเพียงตัวแบบการถดถอยเชิงเส้นที่มีประสิทธิภาพในการทำนายใกล้เคียงกันมากที่สุด

ตารางที่ 2 ค่า RMSE และ MAPE จากตัวแบบทั้ง 4 วิธี

Method	RMSE		MAPE (%)	
	Train data	Test data	Train data	Test data
Linear regression	0.2923	0.2959	2.1752	2.2232
Decision tree	0.3197	0.3318	2.3353	2.4737
Random forest	0.2484	0.2932	1.8315	2.1745
k – Nearest neighbors	0.2846	0.3146	2.0920	2.2986



รูปที่ 2 เปรียบเทียบค่า RMSE และ MAPE ของตัวแบบทั้ง 4 วิธี

จากตัวแบบที่มีความแม่นยำในการทำนายมากที่สุด คือ ตัวแบบ Random forest ซึ่งตัวแบบนี้เป็นตัวแบบทาง Machine learning โดยใช้หลักการ Bootstrapping เพื่อสร้างตัวแบบต้นไม้ตัดสินใจ แล้วจึงทำการรวบรวมผลที่ได้เพื่อตัดสินผลสุดท้าย ทำให้ไม่สามารถรู้ถึงโครงสร้างของตัวแบบที่ใช้ได้อย่างชัดเจน แต่สามารถแสดงตัวแปรหลักที่ตัวแบบใช้พิจารณาจากค่า Gini impurity index และร้อยละความสำคัญของตัวแปร (Overall) 5 อันดับแรก ดังตารางที่ 3

จากตัวแบบ Random forest จะเห็นว่าตัวแปรระดับการศึกษาในระดับปริญญาโท เป็นตัวแปรที่ส่งผลต่อการทำนายเงินเดือนของบัณฑิตมากที่สุด รองลงมาคือประเภทของงานที่ไม่ได้ระบุ สาขาวิชาสถิติ ประเภทงานธุรกิจเอกชน และสาขาวิชาสถิติประยุกต์ ตามลำดับ นอกจากนั้นหากพิจารณาตัวแบบการถดถอยเชิงเส้น ที่มีประสิทธิภาพด้านความแม่นยำของการทำนายในข้อมูลทั้ง 2 ชุดใกล้เคียงกันมากที่สุด สามารถแสดงตัวแปรที่สำคัญของตัวแบบ ซึ่งพิจารณาจากค่า P-value และร้อยละความสำคัญของตัวแปร (Overall) 5 อันดับแรก จะได้ว่าตัวแปรระดับการศึกษาปริญญาโท เป็นตัวแปรที่ส่งผลต่อการทำนายเงินเดือนของบัณฑิตมากที่สุดเช่นกัน รองลงมาคือ ระดับการศึกษาปริญญาเอก ปีการศึกษา 2560 ปีการศึกษา 2559 และระยะเวลาในการหางาน 1 – 2 เดือน ตามลำดับ

ตารางที่ 3 ตัวแปรที่สำคัญ 5 อันดับแรกของตัวแบบการถดถอยเชิงเส้น และตัวแบบ Random forest

Random forest model			Linear regression model			
Variable Importance	Gini index	Overall	Variable Importance	P-value	Overall	$\hat{\beta}$
Education level is Master	20.422	100.00	Education level is Master	6.37×10^{-17}	100.00	0.384
Job type is undefined	10.470	57.192	Education level is Ph.D.	4.39×10^{-15}	93.48	0.633
Major is Statistics	6.064	31.949	Year 2560	1.50×10^{-09}	70.96	0.116
Job type is private staff	5.615	28.554	Year 2559	1.57×10^{-06}	55.57	0.087
Major is Applied Statistics	5.126	25.261	Get a job in 1 – 2 months	1.07×10^{-04}	44.14	-0.083

5. การอภิปรายผล

จากตารางที่ 3 ตัวแปรที่มีความสำคัญมากที่สุดของตัวแบบ Random forest คือ ตัวแปรระดับการศึกษาปริญญาโท รองลงมาคือ ประเภทของงานที่ไม่ได้ระบุ และสาขาวิชาสถิติ ตามลำดับ แต่เนื่องจากตัวแบบนี้ไม่อาจทราบโครงสร้างได้ชัดเจน จึงทำให้การสรุปผลของตัวแบบได้ยาก เช่น หากพิจารณาที่ตัวแปรระดับการศึกษาระดับปริญญาโท ที่เป็นตัวแปรที่สำคัญที่สุดในการทำนายเงินเดือน แต่ไม่สามารถทราบได้ว่าจะส่งผลให้เงินเดือนเพิ่มขึ้นหรือลดลงมากน้อยเพียงใด ทำให้ควรพิจารณาจากรูปที่ 1 ประกอบกัน จะเห็นว่าที่ระดับการศึกษาปริญญาโทมีช่วงความเชื่อมั่นของเงินเดือนอยู่ในระดับสูง อาจสรุปได้ว่าตัวแปรนี้ส่งผลต่อการได้รับเงินเดือนของบัณฑิตมากขึ้น เช่นเดียวกับตัวแปรประเภทการทำงานที่ไม่ได้ระบุ ที่มีช่วงความชื่อมั่นของเงินเดือนอยู่ในระดับสูง ขณะที่ตัวแปรสาขาวิชาสถิติ ประเภทงานธุรกิจเอกชน มีช่วงความเชื่อมั่นอยู่ในระดับเท่ากับค่าเฉลี่ยรวมของเงินเดือนทั้งหมด จึงไม่อาจสรุปได้ชัดเจนว่าจะส่งผลต่อการเพิ่มขึ้น หรือลดลงของเงินเดือนมากน้อยเพียงใด ตัวแปรในตัวแบบ Random forest เป็นตัวแปรที่ให้ค่าสารสนเทศมากที่สุด ทำให้ตัวแบบเลือกใช้ตัวแปรเหล่านี้ทำนายเงินเดือนของบัณฑิต

เมื่อพิจารณาที่ผลของตัวแบบการถดถอยเชิงเส้นที่แสดงในตารางที่ 3 พบว่าตัวแปรที่มีความสำคัญมากที่สุด คือ ตัวแปรระดับการศึกษาปริญญาโท รองลงมาคือ ระดับการศึกษาปริญญาเอก และปีการศึกษา 2560 ตามลำดับ ตัวแบบนี้สามารถอธิบายโครงสร้างของตัวแบบได้ง่ายกว่าตัวแบบ Random forest โดยพิจารณาจากค่าสัมประสิทธิ์การถดถอยของตัวแบบ $(\hat{\beta})$ ในตารางที่ 3 ซึ่งจากตัวแปรที่สำคัญของตัวแบบการถดถอยเชิงเส้นนี้ จะได้ว่าตัวแปรระดับการศึกษาปริญญาโทมีค่าสัมประสิทธิ์การถดถอยเท่ากับ 0.384 ตัวแปรระดับการศึกษาปริญญาเอกมีค่าสัมประสิทธิ์การถดถอยเท่ากับ 0.633 และตัวแปรปีการศึกษา 2560 มีค่าสัมประสิทธิ์การถดถอยเท่ากับ 0.116 ซึ่งมีค่าสัมประสิทธิ์

เป็นบวก บัณฑิตที่มีคุณสมบัติตรงกับปัจจัยเหล่านี้จะทำให้การทำนายเงินเดือนที่ได้รับของบัณฑิตมีโอกาสเพิ่มสูงขึ้น จะเห็นว่าที่ตัวแปรระดับการศึกษาปริญญาเอกมีสัมประสิทธิ์การถดถอยมากกว่า แต่มีค่า p-value น้อยกว่าระดับปริญญาโท อาจมีสาเหตุมากจากจำนวนข้อมูลของบัณฑิตปริญญาเอกมีจำนวนน้อย และมีช่วงของเงินเดือนที่สูงกว่ากลุ่มอื่น ๆ ทำให้การประมาณค่าพารามิเตอร์ที่ได้จึงมีค่าสูง แต่การทดสอบนัยสำคัญของตัวแปรมีค่าต่ำกว่า

ผลการเปรียบเทียบความแม่นยำในการทำนายเงินเดือน พบว่า ตัวแบบ Random forest เป็นตัวแบบที่มีความแม่นยำในการทำนายมากที่สุด โดยมีค่า RMSE และค่า MAPE ต่ำที่สุดในทั้ง 2 ชุดข้อมูล ถึงแม้ว่าตัวแบบนี้จะมีความแม่นยำมากที่สุด แต่มีความแม่นยำในข้อมูลทั้ง 2 ชุดแตกต่างกันมากที่สุด ขณะที่ตัวแบบการถดถอยเชิงเส้นมีความแตกต่างกันน้อยที่สุด รวมถึงสามารถอธิบายโครงสร้างของตัวแบบในการทำนายเงินเดือนของบัณฑิตได้ง่ายที่สุด

6. บทสรุปและข้อเสนอแนะ

การศึกษาครั้งนี้เป็นการประยุกต์ใช้การวิเคราะห์ข้อมูลทางสถิติ และวิทยาการข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อเงินเดือนของบัณฑิตสาขาวิชาที่เกี่ยวข้องกับสถิติ ในประเทศไทย โดยแสดงผลในรูปแบบที่เข้าใจง่ายขึ้น และเปรียบเทียบผลการทำนายเงินเดือนของตัวแบบ 4 วิธี คือ ตัวแบบการถดถอยเชิงเส้น ตัวแบบต้นไม้ตัดสินใจ ตัวแบบ Random forest และตัวแบบ KNN จากข้อมูลที่ได้รับจากกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย และนวัตกรรม ซึ่งพบว่าตัวแบบ Random forest เป็นตัวแบบที่มีความแม่นยำในการทำนายเงินเดือนมากที่สุด และตัวแบบที่สามารถใช้อธิบายการทำนายเงินเดือนของบัณฑิตได้ดีกว่าคือตัวแบบการถดถอยเชิงเส้น

จะเห็นได้ชัดว่าตัวแบบทั้งสองมีตัวแปรที่สำคัญอันดับแรกคือ ระดับการศึกษาปริญญาโท นั่นคือหากบัณฑิตที่กำลังจะจบการศึกษาต้องการหางานที่รายได้สูง ควรจบการศึกษาในระดับปริญญาโทเพื่อใช้เป็นวุฒิการศึกษาในการสมัครงานที่เหมาะสมกับความต้องการ นอกจากนั้นจากตัวแปรประเภทของงาน สาขาวิชาที่จบ และปีการศึกษา จะเห็นว่าประเภทงานที่ไม่ได้ระบุ อาจเป็นธุรกิจส่วนตัว หรืองานที่ไม่มีประเภทในแบบสอบถาม และงานธุรกิจเอกชนเป็นงานที่มีรายได้สูงเช่นกัน ปีการศึกษาที่จบแสดงให้เห็นว่าผู้ที่จบการศึกษาในสาขาวิชาที่เกี่ยวข้องกับสถิติ มีแนวโน้มของเงินเดือนเพิ่มสูงขึ้น แต่ในส่วนของสาขาวิชาหากพิจารณาจากรูปที่ 1 ร่วมด้วยอาจตัดสินได้ยากเนื่องจากมีค่าเฉลี่ยเงินเดือนที่ได้รับใกล้เคียงกับค่าเฉลี่ยรวม

ตัวแบบที่ได้จากการศึกษาครั้งนี้ช่วยให้บัณฑิตสามารถตัดสินใจเลือกเข้าทำงานตามความต้องการ รวมถึงยังอาจเป็นประโยชน์ต่อสถานศึกษา และสถานประกอบการที่สามารถวางแผนจัดการหลักสูตร หรือโครงการรับสมัครงานของสถานประกอบการที่รับสมัครนักศึกษาโดยตรงเพื่อให้มีคุณสมบัติตรงกับความสามารถที่สำเร็จการศึกษา

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณ บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ ที่ให้การสนับสนุนทุนผลการเรียนดีเด่น (SOGPA 18/2561) สำหรับการทำวิจัยครั้งนี้ และ Prof. Don McNeal สำหรับการให้คำปรึกษาในการทำวิจัย

เอกสารอ้างอิง

จั่นสกุล, น., & วิกยานนท์, ท. (2012). การวิเคราะห์การถดถอย โดยใช้โปรแกรม R (1st ed.). โรงพิมพ์ดิจิตอล หน่วยโสตทัศนศึกษา.

หลายพสุ, พ., และคณะ. (2018). การสำรวจภาวะการทำงานของบัณฑิต คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ปีการศึกษา 2556-2557. *วารสารวิทยาศาสตร์ ลาดกระบัง*, *27*(1), 88–98.

สำนักงานสถิติแห่งชาติ. (2560). *การศึกษาแนวโน้มความต้องการแรงงานของตลาดแรงงานในประเทศไทย ช่วงปี 2560—2564*. สำนักสถิติพยากรณ์ สำนักงานสถิติแห่งชาติ. http://www.nso.go.th/sites/2014/DocLib13

Adeco Thailand. (2020). *Adeco Thailand Salary Guide 2019*. https://adecco.co.th/salary-guide

Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, NY.

Glassdoor.com. (2020). *Data Analyst Salaries*. https://www.glassdoor.com/Salaries

Hamlen, K. R., & Hamlen, W. A. (2016). Faculty salary as a predictor of student outgoing salaries from MBA programs. *Journal of Education for Business*, *91*(1), 38–44.

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26–31.

Khongchai, P., & Songmuang, P. (2016a). Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique. *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, 1–6.

Khongchai, P., & Songmuang, P. (2016b). Improving Students' Motivation to Study using Salary Prediction System. *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–6.

Khongchai, P., & Songmuang, P. (2016c). Random Forest for Salary Prediction System to Improve Students' Motivation. *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems*, 638–642.

Kongchouy, N., & Sampantarak, U. (2010). Confidence Intervals for Adjusted Proportions Using Logistic Regression. *Modern Applied Science*, *4*(6), 2–7.

Minakshi, Rajan Vohra, & Gimpy. (2014). Missing Value Imputation in Multi Attribute Data Set. *International Journal of Computer Science and Information Technologies*, *5*(4), 5315–5321.

Parmar, H., Bhanderi, S., & Shah, G. (2014). Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters. *International Conference on Information Science, Kerala*.

Quadri, M., & N.V. Kalyankar. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology*, *10*(2), 2–5.

Rahman, N. A. B. A., Tan, K. L., & Lim, C. K. (2017). Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students. *Journal of Telecommunication, Electronic and Computer Engineering*, *9*(2–12), 155–161.

Tentsho, K., McNeil, R., & Tongkumchum, P. (2019). Determinants of University Dropout: A Case of Thailand. *Canadian Center of Science and Education*, *15*(7), 49–56.

Tongkumchum, P., & Don McNeil. (2009). Confidence intervals using contrasts for regression model. *Songklanakarin Journal of Science and Technology*, *31*(2), 151–156.

Viroonluecha, P., & Kaewkiriya, T. (2018). Salary Predictor System for Thailand Labour Workforce using Deep Learning. *The 18th International Symposium on Communications and Information Technologies (ISCIT 2018)*, 473–478.