1. **เรื่อง:** การเปรียบเทียบตัวแบบสถิติในการทำนายค่ารักษาพยาบาลผู้ป่วยปอดบวม
2. **หน่วยงาน:** นางสาวปิยวรรณ เสรีพงค์กลุ่มงานสารสนเทศทางการแพทย์ โรงพยาบาลสตูล
3. **กลุ่มเป้าหมายกับผู้ใช้:**ทีมเศรษฐกิจโรงพยาบาล ทีมนำคุณภาพ ศูนย์ข้อมูลและสารสนเทศโรงพยาบาล หน่วยงานวิจัยและ R2R
4. **ที่มาและความสำคัญของปัญหา:** โรคปอดบวมติดเชื้อทางเดินหายใจและเป็นสาเหตุสำคัญของการเข้ารับบริการในโรงพยาบาลของผู้สูงอายุ ในขณะที่การรักษาพยาบาลจำเป็นต้องได้รับการดูแลเป็นอย่างดี เพื่อลดโอกาสเสียชีวิตจากโรคแทรกซ้อน โดยเฉพาะการใช้ยา antibiotic และ ventilator ในระยะที่ยาวนาน ส่งผลให้การรักษาโรคปอดบวมมีต้นทุนการรักษาที่ค่อนข้างสูง ปัญหาของต้นทุนการรักษาโรคปอดบวมส่งผลกระทบต่อสถานการณ์การเงินของโรงพยาบาล จนนำไปสู่กระบวนการติดตามและประเมินด้วยการวิเคราะห์ข้อมูลทางสถิติ เพื่อนำไปสู่กระบวนการตัดสินใจเชิงนโยบายของการบริหารโรงพยาบาล การวิเคราะห์ข้อมูลขั้นสูงทางการเงินส่วนใหญ่จะใช้สถิติ linear regression ในการวิเคราะห์ข้อมูล เพื่อการประมาณค่ารักษาพยาบาล และหาความสัมพันธ์ระหว่างปัจจัยกับตัวแปรตาม ในขณะที่ปัจจุบันวิวัฒนาการทางด้าน data science กำลังก้าวเข้าสู่ยุค big data ส่งผลให้ข้อมูลที่ใช้ในการจัดการมีปริมาณมหาศาล ข้อมูลเหล่านี้มีคุณค่าอย่างมากมายในการนำไปสู่กระบวนการวิเคราะห์ข้อมูล เพื่อให้ได้มาซึ่งข้อเท็จจริงในการนำเข้าสู่กระบวนการวิเคราะห์หรือจัดทำนโยบาย

   กระทรวงสาธารณสุขได้จัดทำฐานข้อมูลสุขภาพในรูปแบบ health data 43 table ตลอดระยะเวลา 10 ปีที่ผ่านมาพบว่าข้อมูลในระบบมีการเพิ่มขึ้นอย่างมาก สิ่งเหล่านี้พัฒนาขึ้นเพื่อตอบสนองตามความต้องการเชิงนโยบายที่จะขับเคลื่อนการพัฒนาประเทศด้านสุขภาพ บนพื้นฐานข้อมูลและข้อเท็จจริงที่ปรากฏในสังคม แต่กลับพบว่าการวัดวิเคราะห์ข้อมูลส่วนใหญ่ที่ใช้กับข้อมูลประเภท big data ยังคงใช้สถิติพื้นฐานอย่าง linear regression ในการดำเนินงานมาอย่างต่อเนื่อง ในขณะที่วิธีการ machine learning มีการใช้อย่างแพร่หลายมากกว่า 25 ปี ในการจัดการข้อมูล big dataกลับถูกนำมาใช้ในการประมวลผลข้อมูลทางด้านสุขภาพน้อยมาก

   การศึกษาวิจัยครั้งนี้เพื่อเปรียบเทียบประสิทธิภาพของการทำนายระหว่างสถิติพื้นฐานกับวิธีการ machine learning ส่งผลให้สามารเข้าใจถึงความแตกต่าง และข้อจำกัดของแต่ละวิธี นำไปสู่การพัฒนาเครื่องมือในการวัดวิเคราะห์ health big data อย่างมีประสิทธิภาพ ลดความผิดพลาดในในการวิเคราะห์ ซึ่งอาจนำไปสู่การกำหนดนโยบาย งบประมาณ และการประเมินที่ไม่ตรงความเป็นจริง ร่วมถึงการคาดการณ์สถานการณ์ที่ผิดพลาด การวิจัยครั้งนี้จึงได้ทำการเปรียบเทียบประสิทธิภาพการทำนายระหว่าง linear regression กับ random forest ซึ่งเป็นข้อมูล research methods สำหรับนักวิจัยทางด้านสาธารณสุข
5. **วัตถุประสงค์:** เพื่อเปรียบเทียบ fitted value ระหว่าง linear regression กับ random forest

6. **วิธีการศึกษา: การวิจัยครั้งนี้ใช้รูปแบบการศึกษา cross-sectional analytical study** เพื่อเปรียบเทียบประสิทธิภาพการทำนายค่ารักษาพยาบาล โรคปอดบวม รวบรวมข้อมูลจากฐานข้อมูล satun hospital: hospital information system ระหว่าง พุทธศักราช 2553 – 2557 จำนวน 2,082 รายโดยมี power of test เท่ากับ 0.99 สถิติเชิงพรรณนา ได้แก่ ร้อยละ ค่าเฉลี่ย และสถิติเชิงอนุมาน ได้แก่ linear regression โดยทำการเปรียบเทียบการทำนายด้วยวิธีการ data classification แบบ random forest
7. **ผลการศึกษา :** ผลการศึกษาพบว่า ค่ารักษาพยาบาลโรคปอดบวมโดยเฉลี่ย 49,370 บาท และค่ารักษามีการแจกแจงไม่ปกติ การวิเคราะห์ข้อมูลจึงต้องทำการปรับค่าด้วย log10

Figure1: distribution cost of treatment

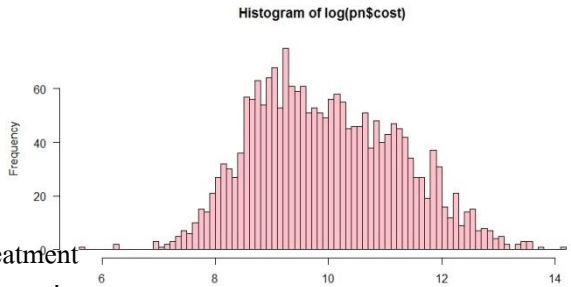
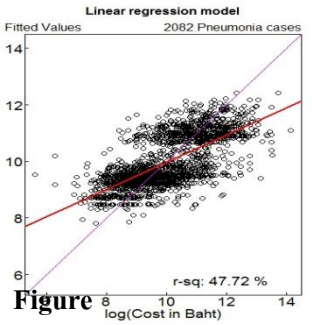


Figure2: log cost of treatment

การเปรียบเทียบประสิทธิภาพการทำนายพบว่า เมื่อนำ fitted value มาเปรียบเทียบกับ cost of treatment ในการวิเคราะห์ด้วย linear regression ได้ค่า r^2 เท่ากับ 47.7% ในขณะที่การวิเคราะห์ random forest ได้ค่า r^2 เท่ากับ 67.3% และเมื่อจำแนกการวิเคราะห์ออกเป็นช่วงขนาดตัวอย่าง พบว่าค่า r^2 จาก linear regression จะมีการเปลี่ยนแปลงในระดับกลุ่มตัวอย่างที่น้อย แต่เมื่อมีกลุ่มตัวอย่างเพิ่มขึ้นค่าดังกล่าวกลับมีการเปลี่ยนแปลงน้อยมาก ในขณะที่การวิเคราะห์ด้วย random forest พบการเปลี่ยนแปลงในการระดับที่มีจำนวนกลุ่มตัวอย่างน้อย แต่จะพบการเปลี่ยนแปลงค่า r^2 อย่างมากในกรณีที่มีกลุ่มตัวอย่างเพิ่มขึ้น



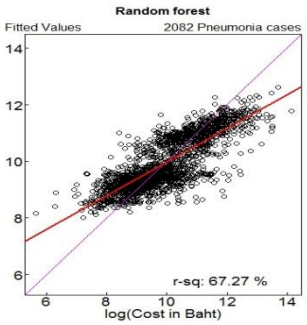


Figure 3: การเปรียบเทียบการทำนาย n=2082

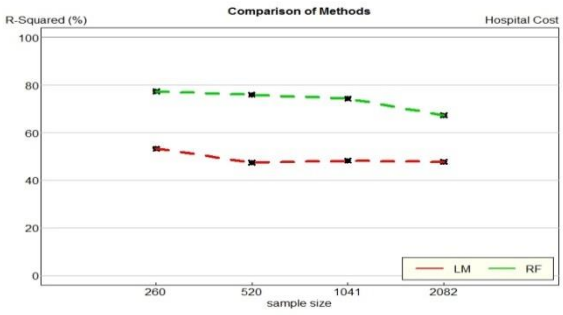


Figure3: การเปรียบเทียบการทำนายแยกตามช่วง

8. **วิจารณ์ผล :** ผลการวิจัยชี้ให้เห็นว่า linear regression มีข้อจำกัดในการประมาณค่ารักษาพยาบาล ในกรณีที่มีข้อมูลจำนวนมาก เนื่องจากไม่สามารถเพิ่มประสิทธิภาพการทำนายถึงแม้จะมีการเพิ่มจำนวนมากขึ้นไปเท่าใด ในขณะที่การวิเคราะห์ random forest จะเหมาะสมกับการวิเคราะห์ข้อมูล

big data และไม่เหมาะสมกับการวิเคราะห์ข้อมูลที่มีจำนวนตัวอย่างน้อย ดังนั้นในการวิเคราะห์ข้อมูลเพื่อจัดทำรายงานวิจัยหรือนโยบายเชิงพัฒนา การทำนายค่ารักษา ต้นทุนการรักษา จึงควรเลือกสถิติให้เหมาะสมกับระดับจำนวนตัวอย่างที่มี โดยเฉพาะในสถานการณ์ปัจจุบันที่โรงพยาบาลและหน่วยงานราชของการกระทรวงสาธารณสุขมีการรวบรวมข้อมูลข้อมูลระดับ big data จึงควรเพิ่มความระมัดระวังในการเลือกใช้เครื่องมือวัดวิเคราะห์ข้อมูล เนื่องจากผลการจัดทำนายงานอาจมีความคลาดเคลื่อนจากความเป็นจริง

9. แหล่งอ้างอิง

Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning.*Building and environment*, *39*(10), 1235-1242.

Liaw, A., & Wiener, M. (2002).Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Strobl, C., Boulesteix, A. L., Zeileis, A., &Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 25.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., &Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.