๑. **เรื่อง:** การเปรียบเทียบตัวแบบสถิติในการทำนายค่ารักษาพยาบาลผู้ป่วยปอดบวม

๒. **หน่วยงาน:** นายเอกมาศ วงศ์ไพรินทร์ กลุ่มงานสารสนเทศทางการแพทย์ โรงพยาบาลสตูล

๓. **กลุ่มเป้าหมายกับผู้ใช้:**ทีมเศรษฐกิจโรงพยาบาล ทีมนำคุณภาพ ศูนย์ข้อมูลและสารสนเทศโรงพยาบาล หน่วยงานวิจัยและ R๒R

๔. **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 ๔๓ table ตลอดระยะเวลา ๑๐ ปีที่ผ่านมาพบว่าข้อมูลในระบบมีการเพิ่มขึ้นอย่างมาก สิ่งเหล่านี้พัฒนาขึ้นเพื่อตอบสนองตามความต้องการเชิงนโยบายที่จะขับเคลื่อนการพัฒนาประเทศด้านสุขภาพ บนพื้นฐานข้อมูลและข้อเท็จจริงที่ปรากฏในสังคม แต่กลับพบว่าการวัดวิเคราะห์ข้อมูลส่วนใหญ่ที่ใช้กับข้อมูลประเภท big data ยังคงใช้สถิติพื้นฐานอย่าง linear regression ในการดำเนินงานมาอย่างต่อเนื่อง ในขณะที่วิธีการ machine learning มีการใช้อย่างแพร่หลายมากกว่า ๒๕ ปี ในการจัดการข้อมูล big data กลับถูกนำมาใช้ในการประมวลผลข้อมูลทางด้านสุขภาพน้อยมาก

การศึกษาวิจัยเพื่อเปรียบเทียบประสิทธิภาพของการทำนายระหว่างสถิติพื้นฐานอย่างกับวิธีการ machine learning จึงมีความสำคัญอย่างมาก เพื่อให้สามารเข้าใจถึงความแตกต่าง เพื่อนำไปสู่การพัฒนาเครื่องมือในการวัดวิเคราะห์ health big data ประสิทธิภาพของการทำนายที่แตกต่างกันสามารถส่งผลต่อความผิดพลาดในการกำหนดนโยบาย ทิศทางและงบประมาณในการพัฒนาประเทศ การตัดสินใจของผู้บริหาร ร่วมถึงการคาดการณ์สถานการณ์ที่ผิดพลาด การวิจัยครั้งนี้จึงได้ทำการเปรียบเทียบประสิทธิภาพการทำนายระหว่าง linear regression กับ random forest เพื่อเป็นข้อมูลสนับสนุนการจัดทำ research methods สำหรับนักวิจัยทางด้านสาธารณสุขและการจัดทำข้อมูลเชิงนโยบายในอนาคต

๕. **วัตถุประสงค์:** เพื่อเปรียบเทียบ fitted value ระหว่าง linear regression กับ random forest

๖. **โครงสร้างและสาระสำคัญ:** การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการทำนายค่ารักษาพยาบาล โรคปอดบวม ระหว่าง linear regression กับ random forest รูปแบบการศึกษา cross-sectional analytical study รวบรวมข้อมูลจากฐานข้อมูล satun hospital: hospital information system ระหว่าง พุทธศักราช ๒๕๕๓ – ๒๕๕๗ จำนวน ๒๐๘๒ ราย ผลการศึกษาพบว่า ค่ารักษาพยาบาลโรคปอดบวมโดยเฉลี่ย ๔๙,๓๗๐ บาท และค่ารักษามีการแจกแจงไม่ปกติ การวิเคราะห์ข้อมูลจึงต้องทำการปรับค่าด้วย log

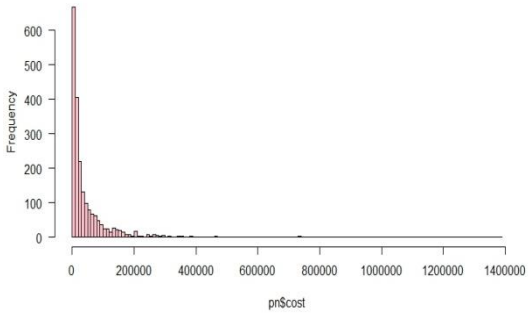

Figure๑: distribution cost of treatment

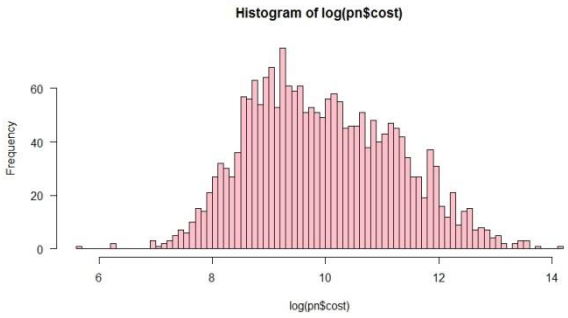

Figure๒: log cost of treatment

การเปรียบเทียบประสิทธิภาพการทำนาย พบว่า เมื่อนำ fitted value มาเปรียบเทียบกับ cost of treatment ในการวิเคราะห์ด้วย linear regression ได้ค่า r^2 เท่ากับ ๔๗.๗% ในขณะที่การวิเคราะห์ random forest ได้ค่า r^2 เท่ากับ ๖๗.๓% และเมื่อจำแนกการวิเคราะห์ออกเป็นช่วงขนาดตัวอย่าง พบว่า ค่า r^2 จาก linear regression จะมีการเปลี่ยนแปลงในระดับกลุ่มตัวอย่างที่น้อย แต่เมื่อมีกลุ่มตัวอย่างเพิ่มขึ้นค่าดังกล่าวกลับมีการเปลี่ยนแปลงน้อยมาก ในขณะที่การวิเคราะห์ด้วย random forest พบการเปลี่ยนแปลงในการระดับที่มีจำนวนกลุ่มตัวอย่างน้อย แต่จะพบการเปลี่ยนแปลงค่า r^2 อย่างมากในกรณีที่มีกลุ่มตัวอย่างเพิ่มขึ้น

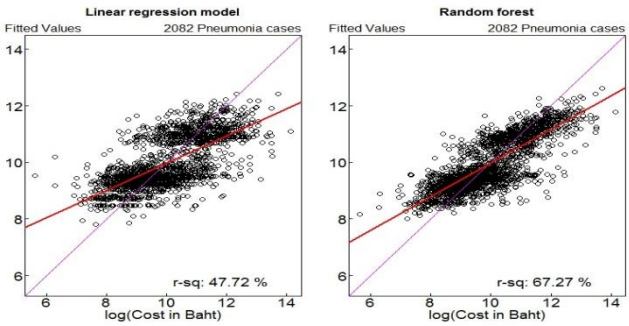

Figure๓: การเปรียบเทียบการทำนาย n=๒๐๘๒

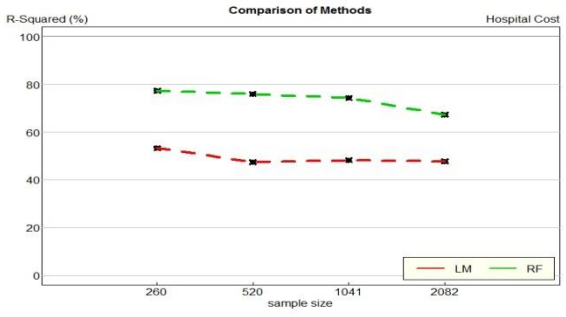

Figure๓: การเปรียบเทียบการทำนายแยกตามช่วง

๗. **การนำไปใช้:** ผลการวิจัยชี้ให้เห็นว่า linear regression มีข้อจำกัดในการประมาณค่ารักษาพยาบาลในกรณีที่มีข้อมูลจำนวนมาก เนื่องจากไม่สามารถเพิ่มประสิทธิภาพการทำนายถึงแม้จะมีการเพิ่มจำนวนมากขึ้นไปเท่าใด ในขณะที่การวิเคราะห์ random forest จะเหมาะสมกับการวิเคราะห์ข้อมูล big data และไม่เหมาะสมกับการวิเคราะห์ข้อมูลที่มีจำนวนตัวอย่างน้อย ดังนั้นในการวิเคราะห์ข้อมูลเพื่อจัดทำรายงานวิจัยหรือนโยบายเชิงพัฒนา การทำนายค่ารักษา ต้นทุนการรักษา จึงควรเลือกสถิติให้เหมาะสมกับระดับจำนวนตัวอย่างที่มี โดยเฉพาะในสถานการณ์ปัจจุบันที่โรงพยาบาลและหน่วยงานราชของการกระทรวงสาธารณสุข มีการรวบรวมข้อมูลข้อมูลระดับ big data จึงควรเพิ่มความระมัดระวังในการเลือกใช้เครื่องมือวัดวิเคราะห์ข้อมูล เนื่องจากผลการจัดทำนายงานอาจมีความคลาดเคลื่อนจากความเป็นจริง

๘. **แหล่งอ้างอิง**

Kim, G. H., An, S. H., & Kang, K. I. (๒๐๐๔). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning.*Building and environment*, ๓๙(๑๐), ๑๒๓๕-๑๒๔๒.

Liaw, A., & Wiener, M. (๒๐๐๒). Classification and regression by randomForest. *R news*, ๒(๓), ๑๘-๒๒.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (๒๐๐๗). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, ๘(๑), ๒๕.

Pal, M. (๒๐๐๕). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, ๒๖(๑), ๒๑๗-๒๒๒.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (๒๐๐๓). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, ๔๓(๖), ๑๙๔๗-๑๙๕๘.