SNRU JST

SNRU Journal of Science and Technology

Journal home page: snrujst.snru.ac.th

การเปรียบเทียบประสิทธิภาพอัลกอริธึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง

A comparative efficiency of data mining algorithms for analysis of factors affecting the cancer

ณัฐวุฒิ ศรีวิบูลย์*

Nattavut Sriwiboon*

สาขาวิทยาการคอมพิวเตอร์ คณะศิลปศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยกาฬสินธุ์ กาฬสินธุ์ 46230 ประเทศไทย

Computer Science Program, Faculty of Liberal Arts and Science, Kalasin University, Kalasin, 46230 Thailand

* **Corresponding Author**: nattavut.sri@gmail.com

Received: 1 September 2016; **Revised:** 6 October 2016; **Accepted:** 7 October 2016; **Available online:** XXXX

บทคัดย่อ

โรคมะเร็งเป็นสาเหตุการเสียชีวิตของประชากรไทยเป็นอันดับ 1 ก่อนหน้านี้การวินิจฉัยการเกิดโรคมะเร็งทำได้เช่นการสอบถามประวัติ และการทดสอบเลือดหรือของเหลวภายในร่างกายโดยทดสอบในห้องปฏิบัติการเป็นต้น ดังนั้นงานวิจัยนี้จึงนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์การเกิดโรคมะเร็ง โดยเปรียบเทียบอัลกอริธึมการทำเหมืองข้อมูลประกอบด้วย อัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และอัลกอริธึม Naïve Bayes ผลการเปรียบเทียบพบว่าอัลกอริธึม C4.5 มีประสิทธิภาพสูงสุดที่ 98.63% แล้วนำแบบจำลองที่มีค่าประสิทธิภาพสูงที่สุดเป็นแบบจำลองสำหรับวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งพบว่าผู้ที่สูบบุหรี่จะมีความเสี่ยงต่อการเป็นโรคมะเร็ง และสามารถนำกฎการจำแนกข้อมูลที่ได้ไปพัฒนาเป็นระบบตรวจวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งได้

คำสำคัญ: โรคมะเร็ง; การทำเหมืองข้อมูล; อัลกอริธึม

Abstract

Cancer is the most leading cause of death in Thailand. Previous diagnosis of the cancer has been done by several methods, namely, medical history, blood testing and body fluids. In this paper, we have proposed a cancer diagnostic technique by using data mining. By comparing C4.5, k-Nearest Neighbor and Naïve Bayes, performance comparisons have showed that the C4.5 outperforms k-Nearest Neighbor and Naïve Bayes. The C4.5 performs 98.63% of accuracy, and is the most efficient algorithms for analyzing the cause of cancer. In addition, smokers are riskiness for cancer using our analytical model. Furthermore, our analytical model can be used to apply for developing the analytical system for cancer.

Keywords: Cancer; Data mining; Algorithm

1. บทนำ

โรคมะเร็งเป็นกลุ่มโรคที่ทำให้ประชากรไทยเสียชีวิตเป็นอันดับ 1 ของประเทศ [1] ชนิดของมะเร็งที่พบประกอบด้วยมะเร็งปากมดลูกและมะเร็งเต้านมพบในผู้หญิง ส่วนมะเร็งตับและมะเร็งปอดพบในผู้ชาย โดยโรคมะเร็งมีสาเหตุเกิดจากหลายๆ ปัจจัย เช่น การบริโภคอาหารและเครื่องที่มีส่วนประกอบของสารเคมี การได้รับรังสี หรือพยาธิบางชนิดเป็นต้น

การทำเหมืองข้อมูล (data mining) เป็นการนำระบบคอมพิวเตอร์มาประมวลผลกับข้อมูลที่มีอยู่เป็นจำนวนมากเพื่อให้ได้ความรู้ในฐานข้อมูลนั้นโดยอาศัยหลักการค้นหารูปแบบความสัมพันธ์ใหม่ที่มีความหมายและแนวโน้มจากข้อมูลจำนวนมากที่เก็บไว้โดยใช้การจดจำรูปแบบทางสถิติและคณิตศาสตร์ ซึ่งการทำเหมืองข้อมูลมีประโยชน์สำหรับการวิเคราะห์กับข้อมูลที่มีอยู่เป็นจำนวนมากในหลายๆ สาขาเช่น ในองค์กรธุรกิจ สถาบันการศึกษา และองค์กรเกี่ยวกับวิทยาศาสตร์สุขภาพ

จากปัญหาการตรวจค้นหามะเร็งระยะเริ่มแรกทำได้โดยการสอบถามประวัติโดยละเอียด การตรวจร่างกายโดยละเอียด การตรวจทางห้องปฏิบัติการเป็นต้นซึ่งวิธีการดังกล่าวเป็นวิธีที่ไม่สะดวกสำหรับประชาชนที่ต้องการทราบการตรวจค้นหาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง โดยจากที่ปัจจุบันเทคโนโลยีสารสนเทศมีความก้าวหน้าสามารถนำมาใช้ในการตรวจวิเคราะห์เพื่อตรวจค้นหาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งได้และจากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าการทำเหมืองข้อมูลเป็นวิธีที่สามารถนำมาใช้สำหรับการตรวจวิเคราะห์หาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง

ดังนั้นในงานวิจัยนี้จึงได้รวบรวมข้อมูลและออกแบบขั้นตอนการวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง โดยการสำรวจประชากรกลุ่มที่เป็นโรคมะเร็งในจังหวัดกาฬสินธุ์เพื่อนำข้อมูลที่ได้มาเปรียบเทียบประสิทธิภาพด้วยการทำเหมืองข้อมูล โดยสำหรับอัลกอริธึมการทำเหมืองข้อมูลในงานวิจัยนี้ใช้อัลกอริธึม C4.5 [2] อัลกอริธึม k-Nearest Neighbor [3] และอัลกอริธึม Naïve Bayes [4] มาวิเคราะห์ปัจจัยที่มีผลต่อการเกิดโรคมะเร็งโดยวัดประสิทธิภาพเพื่อเปรียบเทียบค่าความแม่นยำ (accuracy) [5] สำหรับนำไปสร้างแบบจำลอง (model) พยากรณ์แล้วนำแบบจำลองที่ได้ค้นหาปัจจัยเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง ผลที่ได้แสดงให้เห็นว่าอัลกอริธึม C4.5 มีประสิทธิภาพสูงสุดสำหรับการสร้างแบบจำลองพยากรณ์และผลจากการค้นหาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งพบว่า การสูบบุหรี่เป็นปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งมากที่สุด สามารถนำผลที่ได้จากงานวิจัยนี้เป็นข้อมูลเพื่อสนับสนุนการวิเคราะห์สาเหตุการเกิดโรคมะเร็งได้ อีกทั้งสามารถนำกฎการจำแนกข้อมูลและแบบจำลองที่ได้จากงานวิจัยนี้ไปพัฒนาเป็นระบบตรวจวิเคราะห์การเกิดโรคมะเร็ง

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โรคมะเร็ง [6] คือกลุ่มของโรคที่เกิดขึ้นจากการแบ่งเซลล์ที่ไม่สามารถควบคุมได้และการที่เซลล์เหล่านี้เข้าไปทำลายเนื้อเยื่ออื่นๆ อาจโดยการที่เซลล์เจริญเติบโตเข้าไปยังเนื้อเยื่ออื่นๆ หรือการกระจายเซลล์ไปยังที่อื่นๆ ที่เป็นการแพร่กระจายของเนื้อร้าย การเจริญเติบโตที่ไม่สามารถควบคุมได้นี้อาจเกิดจากการกลายพันธุ์ของ DNA ภายในเซลล์ทำให้ข้อมูลทางพันธุกรรมที่ควบคุมการทำงานของเซลล์สูญเสียไป ในประเทศไทยพบบ่อยอยู่ 4 ชนิด [1] คือมะเร็งปากมดลูกและมะเร็งเต้านมพบในผู้หญิง ส่วนมะเร็งตับและมะเร็งปอดพบในผู้ชาย โดยสาเหตุและปัจจัยเสี่ยงของการเกิดมะเร็ง อย่างเช่นการบริโภคอาหารและเครื่องดื่มในปัจจุบันส่วนใหญ่มีวัตถุดิบมาจากสารเคมี สารก่อมะเร็งที่ปนเปื้อน ในอาหาร อากาศ เครื่องดื่ม ยารักษาโรคเป็นต้น รวมทั้งการได้รับรังสี เชื้อไวรัส เชื้อแบคทีเรียและพยาธิบางชนิด

2.2 เหมืองข้อมูล ถูกนิยามไว้หลายความหมายเช่น Daniel [3] กล่าวว่าการทำเหมืองข้อมูลคือขั้นตอนการค้นหารูปแบบความสัมพันธ์ใหม่ที่มีความหมายและแนวโน้มจากข้อมูลจำนวนมากที่เก็บไว้โดยใช้การจดจำรูปแบบทางสถิติและคณิตศาสตร์ ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านเศรษฐศาสตร์และวิทยาศาสตร์สุขภาพเป็นต้น การทำเหมืองข้อมูลคือวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บ

ข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปของฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

2.3 อัลกอริธึม C4.5 ถูกเสนอโดย Quinlan [2] เป็นอัลกอริธึมที่สามารถสร้างต้นไม้ตัดสินใจ (decision tree) ใช้งานง่าย กระบวนการทำงานไม่ซับซ้อนและได้ผลลัพธ์ที่ดีในการนำไปใช้งาน หลักการทำงานของอัลกอริธึมจะคัดเลือก Attribute ที่สำคัญที่สุดที่มีค่า Gain Ratio สูงสุดมาเป็นโหนดราก (root note) ซึ่งโครงสร้างการทำงานของอัลกอริธึม C4.5 สามารถคัดแยกข้อมูล (Classification) ได้

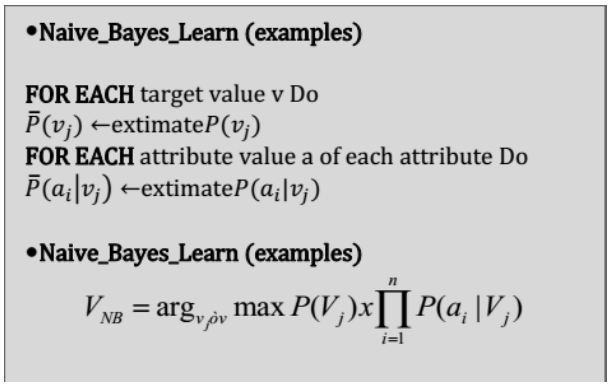
2.4 อัลกอริธึม k-Nearest Neighbor (k-NN) [3] เป็นวิธีที่ใช้สำหรับจัดแบ่งคลาสของข้อมูล โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่สามารถแทนเงื่อนไขหรือกรณีใหม่ๆ ได้ โดยใช้วิธีการตรวจสอบจำนวนบางจำนวน (k) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (count up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส แล้วกำหนดเงื่อนไขใหม่ๆ ให้กับคลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด โดยขั้นตอนการหาระยะที่ใกล้ที่สุดด้วยสมการ Euclidean Distance ดังสมการ (1)

$$d_{\text{Euclidean}}(x_i, y_i) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - y_{i,k})^2} \quad (1)$$

โดย $d_{\text{Euclidean}}(x_i, y_i)$ คือระยะห่างระหว่าง ตัวอย่าง x_i และตัวอย่าง y_i

$x_{i,k}$ คือสมบัติตัวที่ k ของตัวอย่าง x_i

2.5 อัลกอริธึม Naïve Bayes [4] เป็นอัลกอริธึมที่ใช้สำหรับการจำแนกประเภทข้อมูล โดยลักษณะของอัลกอริธึมใช้งานได้ดีกับลักษณะข้อมูลที่มี Attribute ของตัวอย่างไม่ขึ้นต่อกันและเซตตัวอย่างข้อมูลที่มีจำนวนมาก นิยมนำไปใช้กับการจำแนกประเภทข้อความเป็นต้น โดยอัลกอริธึม Naïve Bayes เป็นวิธีการจำแนกที่ไม่ซับซ้อนดัง ภาพที่ 1



ภาพที่ 1 Naïve Bayes Learning Algorithms [4]

2.6 งานวิจัยที่เกี่ยวข้อง

Cheewaprakobkit [7] ได้เสนองานวิจัยที่ใช้เทคนิคต้นไม้ตัดสินใจด้วยอัลกอริธึม C4.5 สำหรับเปรียบเทียบประสิทธิภาพกับโครงข่ายประสาทเทียมเพื่อแยกประเภทของปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับปริญญาตรีในหลักสูตรนานาชาติ จากงานวิจัยแสดงให้เห็นว่า C4.5 ให้ความแม่นยำในการพยากรณ์ที่ 85.13% มากกว่าโครงข่ายประสาทเทียมที่ได้ผลเป็น 83.87%

Kotsiantis และคณะ [8] ได้เสนองานวิจัยที่เปรียบเทียบประสิทธิภาพของอัลกอริธึมเพื่อพยากรณ์ประสิทธิภาพของนักศึกษาในระบบการศึกษาทางไกลด้วยอัลกอริธึม C4.5 อัลกอริธึม Naïve Bayes อัลกอริธึม Ripper และ อัลกอริธึม k-Nearest Neighbor ผลการวิจัยแสดงให้เห็นว่าอัลกอริธึม Naïve Bayes ให้ค่าประสิทธิภาพ 74.70% สูงกว่าอัลกอริธึมอื่นๆ

Pansumret และคณะ [9] ได้เสนองานวิจัยที่เปรียบเทียบอัลกอริธึม C4.5 อัลกอริธึม Naïve Bayes และ อัลกอริธึม k-Nearest Neighbor สำหรับวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาโดยงานวิจัยแสดงให้เห็นว่าอัลกอริธึม C4.5 ให้ค่าประสิทธิภาพ 73.55% สูงกว่าอัลกอริธึมอื่นๆ

Muntham และ Ingsrisawang [10] ได้เสนองานวิจัยที่ใช้ต้นไม้ตัดสินใจด้วยอัลกอริธึม C4.5 เพื่อวินิจฉัยโรคระบบการหายใจโดยใช้ข้อมูลจากเวชระเบียนจำนวน 7,327 ราย โดยแบ่งออกเป็นโรคติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลันพบว่าใช้ตัวแปรที่คัดเลือก 7 ตัวแปรกับชุดข้อมูลเรียนรู้ต่อชุดข้อมูลทดสอบ 70:30 ได้ค่าความถูกต้องของการจำแนกเท่ากับ 92.32% โรคปอดอักเสบพบว่าใช้ตัวแปรที่คัดเลือก 8 ตัวแปรกับชุดข้อมูลเรียนรู้ต่อชุดข้อมูลทดสอบ 70:30 ได้ค่าความถูกต้องของการจำแนกเท่ากับ 94.70% และโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลันพบว่าใช้ตัวแปรที่คัดเลือก 7 ตัวแปรกับชุดข้อมูลเรียนรู้ต่อชุดข้อมูลทดสอบ 50:50 ได้ค่าความถูกต้องของการจำแนกเท่ากับ 94.69%

ตารางที่ 1 Attribute

Attribute	Description	Attribute	Description
Age	อายุ (ปี) 26 - 35 ปี = 1; 36 - 45 ปี = 2; 46 - 55 ปี = 3; 56 - 65 ปี = 4; 66 ปี ขึ้นไป = 5	Alcohol	ดื่มเครื่องดื่มที่มีส่วนผสมของแอลกอฮอล์ 0 = เคย, 1 = ไม่เคย
Sex	0 = หญิง, 1 = ชาย	Smoking	0 = เคย, 1 = ไม่เคย
Status	0 = โสด, 1 = สมรสหรือเคยสมรส	Salt	รับประทานอาหารที่มีไขมันเป็นประจำ 0 = ใช่, 1 = ไม่ใช่
Career	อาชีพ 0 = อื่นๆ, 1= เกี่ยวกับสุขภาพ	High-fat	รับประทานอาหารที่มีไขมันเป็นประจำ 0 = ใช่, 1 = ไม่ใช่
Cripple	0 = พิการ, 1 = ไม่พิการ	HIV	เป็นผู้ติดเชื้อ HIV 0 = ใช่, 1 = ไม่ใช่
Heredity	กรรมพันธุ์ 0 = มี, 1 = ไม่มี		

3. วิธีดำเนินการวิจัย

3.1 การเตรียมข้อมูล

งานวิจัยนี้มุ่งเน้นเปรียบเทียบประสิทธิภาพค่าความแม่นยำของอัลกอริธึมที่ใช้ทำเหมืองข้อมูลเพื่อนำผลที่ได้สร้างแบบจำลองพยากรณ์ โดยเปรียบเทียบระหว่างอัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และ อัลกอริธึม Naïve Bayes ซึ่งจากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าอัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และอัลกอริธึม Naïve Bayes มีประสิทธิภาพค่า

ความแม่นยำสำหรับสร้างแบบจำลองพยากรณ์ และในงานวิจัยนี้ได้วิเคราะห์ปัจจัยที่มีผลต่อการเกิดโรคมะเร็งโดยปัจจัยเสี่ยงที่จะเกิดโรคมะเร็งได้รับข้อมูลจากสถาบันมะเร็งแห่งชาติ [11] ซึ่งประกอบด้วย อายุ (age) เพศ (sex) สถานะภาพ (status) อาชีพ (career) ความผิดปกติในร่างกายตั้งแต่กำเนิด (cripple) กรรมพันธุ์ (heredity) เคยดื่มเครื่องดื่มที่มีส่วนผสมของแอลกอฮอล์ (alcohol) สูบบุหรี่ (smoking) รับประทานอาหารเค็มเป็นประจำ (salt) รับประทานอาหารที่มีไขมันเป็นประจำ (high-fat) และเป็นผู้ติดเชื้อ HIV

3.2 การจัดเก็บข้อมูล

งานวิจัยนี้ดำเนินการจัดเก็บข้อมูลโดยใช้แบบสอบถามพฤติกรรมผู้ป่วยโรคมะเร็งในจังหวัดกาฬสินธุ์ 3 อำเภอประกอบด้วยพื้นที่อำเภอสมเด็จ อำเภอนามนและอำเภอกุฉินารายณ์ที่มีอายุ 26 ปีขึ้นไปจำนวน 517 ราย ระหว่างเดือนตุลาคม พ.ศ. 2557 ถึง เดือนกันยายน พ.ศ. 2558 โดยมีโครงสร้างข้อมูลและรหัสข้อมูลดังตารางที่ 1

3.3 การวัดประสิทธิภาพของอัลกอริธึม

เพื่อเปรียบเทียบค่าความแม่นยำสำหรับนำไปสร้างแบบจำลองพยากรณ์ในงานวิจัยนี้ใช้การวัดค่า k-Fold cross-validation [12] ค่าความแม่นยำ [5] และค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย (mean absolute error : MAE) [13] มีรายละเอียดดังนี้

k-Fold cross-validation คือวิธีการวัดประสิทธิภาพสำหรับพยากรณ์ตัวอย่างของแบบจำลอง โดยพื้นฐานของเทคนิคนี้คือการสุ่มตัวอย่างมีกระบวนการทำงานเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วนๆ เท่าๆ กัน โดยแบ่งข้อมูลออกเป็น k ชุดเท่าๆ กันเพื่อนำข้อมูลบางส่วนมาใช้สำหรับเรียนรู้ (training set) แล้วคำนวณค่าความแม่นยำจากการพยากรณ์ k รอบ โดยแต่ละรอบจะมีการสร้างแบบจำลองจำแนกประเภท 1 ตัวแล้วนำข้อมูลบางส่วนมาใช้ทดสอบ (testing set) กับแบบจำลองที่ได้จากการเรียนรู้

ค่าความแม่นยำ คือค่าที่ได้จากวิธีการทดสอบหาค่าพยากรณ์ความถูกต้องของข้อมูลว่ามีความถูกต้องมากน้อยเพียงใดโดยคิดเป็นค่าร้อยละ (%) โดยใช้สูตรการคำนวณดังสมการ (2)

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

โดย TP คือ ค่าที่พยากรณ์ถูกต้องเชิงบวก

TN คือ ค่าที่พยากรณ์ถูกต้องเชิงลบ

FP คือ ค่าที่พยากรณ์ผิดพลาดเชิงบวก

FN คือ ค่าที่พยากรณ์ผิดพลาดเชิงลบ

ค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย คือค่าที่ได้จากวิธีการวัดความแตกต่างระหว่างค่าความจริงกับค่าพยากรณ์ ซึ่งหากคำนวณแล้วค่า MAE น้อยแสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับความจริงซึ่งหากได้ผลการคำนวณ MAE เท่ากับ 0 แสดงว่าไม่เกิดความคลาดเคลื่อนในแบบจำลองโดยการคำนวณหาค่า MAE ใช้สมการดังสมการ (3)

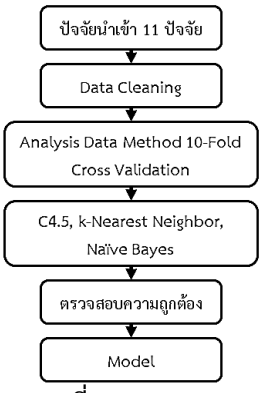
$$MAE = \frac{1}{n}\sum_{i=1}^{n} | e_i | \tag{3}$$

โดย e_i คือ ผลต่างระหว่างค่าข้อมูลจริงกับค่าพยากรณ์

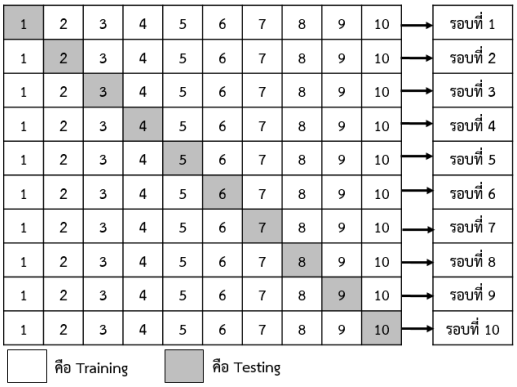
n คือ จำนวนข้อมูลที่ใช้พยากรณ์

3.4 ขั้นตอนการสร้างแบบจำลองพยากรณ์

ขั้นตอนการสร้างแบบจำลองพยากรณ์ใช้อัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และอัลกอริธึม Naïve Bayes ดังภาพที่ 2



ภาพที่ 2 Work Flow



ภาพที่ 3 Data Method 10-Fold cross-validation

จากภาพที่ 2 แสดงขั้นตอนการสร้างแบบจำลองเพื่อใช้พยากรณ์การเกิดโรคมะเร็งโดยวัดประสิทธิภาพค่าความแม่นยำเพื่อหาแบบจำลองที่ดีที่สุดใช้เป็นแบบจำลองสำหรับพยากรณ์โรคมะเร็งโดยมีรายละเอียดดังนี้

1) จากการจัดเก็บข้อมูลที่มีปัจจัยนำเข้า 11 ปัจจัยการดำเนินการในขั้นตอนนี้จะแทนข้อมูลที่ได้จากแบบสอบถามด้วยตัวเลขตามคำอธิบายในตารางที่ 1 ยกตัวอย่างเช่นมีกรรมพันธุ์ที่เป็นโรคมะเร็งถ้ามีให้แทนข้อมูลด้วยตัวเลข 0 ถ้าไม่มีให้แทนข้อมูลด้วยตัวเลข 1 เป็นต้น

2) Data Cleaning เมื่อรวบรวมแบบสอบถามโดยจัดเก็บข้อมูลที่มีปัจจัยนำเข้า 11 ปัจจัยและแทนค่าด้วยตัวเลขแล้วตรวจสอบว่าข้อมูลมีค่าว่าง (missing value) และสิ่งรบกวน (noisy data) หรือไม่โดยแทนค่าที่เป็นตัวเลขด้วยค่าเฉลี่ยของข้อมูล

3) เลือกข้อมูลที่ใช้สำหรับเรียนรู้ และทดสอบ ตามวิธีการ k-Fold cross-validation ที่เป็นวิธีสุ่มเลือกข้อมูลแบบความเที่ยงตรง k กลุ่ม โดยในงานวิจัยนี้กำหนด k = 10 โดยการทดสอบครั้งแรกข้อมูลชุดที่ 1 จะเป็นข้อมูลชุดทดสอบส่วนที่เหลือจะเป็นชุดเรียนรู้ ครั้งที่ 2 ข้อมูลชุดที่ 2 จะเป็นข้อมูลชุดทดสอบส่วนที่เหลือจะเป็นชุดเรียนรู้ ทำแบบนี้จนถึงการทดลองที่ k โดยอธิบายขั้นตอนดังภาพที่ 3

4) สร้างแบบจำลองเพื่อใช้สำหรับการพยากรณ์โดยนำข้อมูลมาวิเคราะห์ตามอัลกอริธึม C4.5 อัลกอริธึม Naïve Bayes และอัลกอริธึม k-Nearest Neighbor (k = 10) ซึ่งในงานวิจัยนี้ใช้โปรแกรม Weka 3.8 สำหรับการทดสอบอัลกอริธึมเพื่อวัดประสิทธิภาพความแม่นยำ

5) ตรวจสอบความถูกต้องโดยการทดสอบแบบจำลองด้วยชุดข้อมูลทดสอบเพื่อวัดประสิทธิภาพค่าความแม่นยำและค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ยแล้วเปรียบเทียบประสิทธิภาพแต่ละอัลกอริธึมเพื่อหาแบบจำลองที่ดีที่สุดเพื่อใช้สำหรับเป็นแบบจำลองพยากรณ์โรคมะเร็ง

6) ค้นหาปัจจัย ซึ่งในขั้นตอนนี้เป็นกระบวนการวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งโดยนำอัลกอริธึมที่มีประสิทธิภาพที่ดีที่สุดในการสร้างแบบจำลองมาค้นหาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งด้วยวิธีการตรวจสอบปัจจัยย้อนกลับแล้วลดการนำเข้าปัจจัยนำเข้าทีละหนึ่งปัจจัย [14] จากนั้นตรวจสอบค่าความแม่นยำ โดยปัจจัยนำเข้าใดมีค่าลดลงมากที่สุดแสดงให้เห็นว่าปัจจัยนั้นมีความสำคัญที่สุด

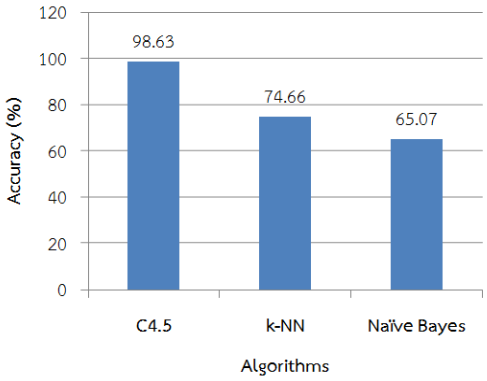
4. ผลและการอภิปรายผลการวิจัย

4.1 ผลการสร้างแบบจำลองพยากรณ์

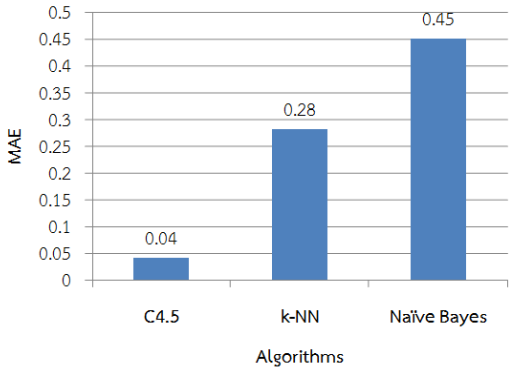
ตารางที่ 2 Performance of Mode

Algorithms	Accuracy	MAE
C4.5	98.63	0.04
k-NN	74.66	0.28
Naïve Bayes	65.07	0.45

จากตารางที่ 2 แสดงค่าความแม่นยำ และค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย ของแต่ละอัลกอริธึม โดยอัลกอริธึม C4.5 เป็นอัลกอริธึมที่มีความแม่นยำสูงสุดที่ 98.63 % ค่า MAE เท่ากับ 0.04 แสดงให้เห็นว่า C4.5 มีความแม่นยำมากกว่า k-Nearest Neighbor (k=10) ที่มีความแม่นยำ 74.66 % ค่า MAE เท่ากับ 0.28 และ Naïve Bayes ที่มีความแม่นยำ 65.07 % ค่า MAE เท่ากับ 0.45 ดังภาพที่ 4



ภาพที่ 4 Comparison Accuracy of Model



ภาพที่ 5 Comparison Mean Absolute Error (MAE) of Model

อีกทั้งค่า MAE ที่แสดงความคลาดเคลื่อนเฉลี่ยของอัลกอริธึม C4.5 มีค่าน้อยกว่าอัลกอริธึม k-Nearest Neighbor (k=10) และอัลกอริธึม Naïve Bayes ดังภาพที่ 5 แสดงให้เห็นว่าอัลกอริธึม C4.5 สามารถประมาณค่าได้ใกล้เคียงกับความจริง ดังนั้นจึงนำแบบจำลองที่ได้จากอัลกอริธึม C4.5 มาใช้สำหรับค้นหาปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง

จากผลการสร้างแบบจำลองพยากรณ์ในงานวิจัยนี้ที่มีการกำหนดปัจจัยนำเข้า 11 ปัจจัยที่มีลักษณะชุดข้อมูลเป็นตัวเลข โดยอัลกอริธึม C4.5 มีประสิทธิภาพสามารถประมาณค่าได้ใกล้เคียงกับความจริงมากที่สุด และเมื่อเปรียบเทียบกับงานวิจัยของ Kotsiantis และคณะ [8] ผลการสร้างแบบจำลองพยากรณ์ในงานวิจัยนี้มีประสิทธิภาพดีกว่าอัลกอริธึม Naïve Bayes และจากการกำหนด k-Fold cross-validation ที่ใช้สุ่มเลือกข้อมูลแบบความเที่ยงตรง k กลุ่ม โดยในงานวิจัยนี้กำหนด k = 10 แสดงให้เห็นว่าลักษณะข้อมูลปัจจัยนำเข้าของงานวิจัยนี้สามารถให้ค่าความแม่นยำมากกว่างานวิจัยของ Cheewaprakobkit [7] งานวิจัยของ Pansumret และคณะ [9] และงานวิจัยของ Muntham และ Ingsrisawang [10] ที่มีการกำหนดค่า k = 3

4.2 ผลการค้นหาปัจจัย

ผลการค้นหาปัจจัยซึ่งแสดงให้เห็นถึงผลการวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งโดยการนำอัลกอริธึม C4.5 ที่มีประสิทธิภาพในการสร้างแบบจำลองสูงสุดมาค้นหาปัจจัย โดยแสดงปัจจัยนำเข้าที่มีความสำคัญ 5 ลำดับ ดังนี้

ตารางที่ 3 Factors Affecting the Cancer

Attribute	Accuracy	No.
Smoking	90.19	1
Age	91.23	2
Alcohol	93.69	3
HIV	94.15	4
Heredity	95.66	5

จากตารางที่ 3 แสดงให้เห็นว่าปัจจัยที่มีผลต่อการเกิดโรคมะเร็งมากที่สุดคือการสูบบุหรี่ โดยพิจารณาจากความแม่นยำที่ลดลงมากที่สุดซึ่งปัจจัยที่มีผลต่อการเกิดโรคมะเร็งลำดับถัดไปประกอบด้วย อายุ เคยดื่มเครื่องดื่มที่มีส่วนผสมของแอลกอฮอล์ เป็นผู้ติดเชื้อ HIV และกรรมพันธุ์ ตามลำดับ

5. สรุปผล

จากการเปรียบเทียบประสิทธิภาพความแม่นยำของอัลกอริธึมในการสร้างแบบจำลองพยากรณ์ประกอบด้วยอัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และอัลกอริธึม Naïve Bayes โดยอัลกอริธึม C4.5 เป็นอัลกอริธึมที่มีความแม่นยำสูงสุดที่ 98.63 % และมีค่า MAE ที่แสดงความคลาดเคลื่อนเฉลี่ยของอัลกอริธึมเท่ากับ 0.04 แสดงให้เห็นว่าอัลกอริธึม C4.5 มีประสิทธิภาพความแม่นยำและสามารถประมาณค่าได้ใกล้เคียงกับความจริงกับข้อมูล 11 ปัจจัยที่งานวิจัยนี้กำหนด โดยข้อมูลที่ใช้สร้างแบบจำลองพยากรณ์งานวิจัยนี้ได้ดำเนินการจัดเก็บข้อมูลโดยใช้แบบสอบถามพฤติกรรมผู้ป่วยโรคมะเร็งในจังหวัดกาฬสินธุ์ 3 อำเภอประกอบด้วยพื้นที่อำเภอสมเด็จ อำเภอนามนและอำเภอกุฉินารายณ์ที่มีอายุ 26 ปีขึ้นไปจำนวน 517 ราย ระหว่างเดือนตุลาคม พ.ศ. 2557 ถึง เดือนกันยายน พ.ศ. 2558 ผลจากการสร้างแบบจำลองพยากรณ์โรคมะเร็งด้วยอัลกอริธึม C4.5 ที่มีประสิทธิภาพความแม่นยำสูงสุด งานวิจัยนี้จึงนำแบบจำลองพยากรณ์มาค้นหาปัจจัยเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งแสดงให้เห็นว่าการสูบบุหรี่เป็นปัจจัยที่มีผลต่อการเกิดโรคมะเร็งมากที่สุดและลำดับถัดไปประกอบด้วยปัจจัยเรื่อง อายุ เคยดื่มเครื่องดื่มที่มีส่วนผสมของแอลกอฮอล์ เป็นผู้ติดเชื้อ HIV และกรรมพันธุ์ตามลำดับ โดยสามารถนำผลที่ได้จากงานวิจัยนี้เป็นข้อมูลเพื่อสนับสนุนการวิเคราะห์หาสาเหตุการเกิดโรคมะเร็งทั่วไปที่มีโอกาสเกิดขึ้นกับมนุษย์ได้ และนำไปแนะนำให้กับผู้ป่วยสำหรับการรักษาโรคมะเร็งรวมถึงนำไปแนะนำให้กับประชาชนเพื่อหลีกเลี่ยงปัจจัยเสี่ยงต่อการเกิดโรคมะเร็ง

6. ข้อเสนอแนะ

ในอนาคตจะนำแบบจำลองที่มีประสิทธิภาพพัฒนาเป็นระบบตรวจวิเคราะห์การเกิดโรคมะเร็งที่ระบบและวิธีการสร้างแบบจำลองพยากรณ์สามารถตรวจวิเคราะห์และแสดงผลการวิเคราะห์กับโรคมะเร็งแต่ละชนิด

7. กิตติกรรมประกาศ

โครงการนี้ได้รับทุนสนับสนุนการวิจัยจากสำนักงานคณะกรรมการวิจัยแห่งชาติ (วช.) ประจำปีงบประมาณ 2559 และขอขอบคุณมหาวิทยาลัยกาฬสินธุ์ที่สนับสนุนสถานที่สำหรับทำวิจัย

8. References

[1] Number of in - patients by 75 cause groups according from health service units, ministry of public health, kanchana buri province: 2003-2012, http://service.nso.go.th/nso/web/statseries/statseries09.html, 1 January 2016.

[2] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Francisco, 1993.

[3] T. Daniel, Discovering Knowledge in Data, A JOHN WILEY & SONS, New Jersey, 2005.

[4] J. Rennie, L. Shih , J. Teevan , D. Karger, Tackling the Poor Assumptions of Naive Bayes Text Classifiers, The Twentieth International Conference on Machine Learning 2003, Washington. 21-24 August 2003, 1 - 8.

[5] B. Tilmann, The Business Impact of Predictive Analytics, IGI Global. 3(2007) 118–119.

[6] Cancer, http://www.who.int/mediacentre/factsheets/fs297/en/, 10 January 2016.

[7] P. Cheewaprakobkit, Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program, IMECS, Hong Kong. 2013, 1 - 8.

[8] S. Kotsiantis, C. Pierrakeas, I. Zaharakis, P. Pintelas, Efficiency of Machine Learning Techniques in Predicting Student's Performance in Distance Learning System, Recent Advances in Mechanics and Related Fields University of Patras Greece, 2003.

[9] Y. Pansumret, J. Phuboon-ob, W. Pongsiri, On Comparison of Data Mining Algorithms for Analysis of Factors Affecting the Academic Performance of Students, J. Sci. Technol MSU. 1(2)(2013) 281-289.

[10] D. Muntham, L. Ingsrisawang, An Application of Decision Tree Algorithms for Diagnosis of the Respiratory System: A Case Study of Pranakorn Sri Ayudthaya Hospital, HSRI. 4(1) (2010) 73-81.

[11] Causes and risk factors of cancer can be divided into 2 categories, http://www.nci.go.th/th/Knowledge/ reasonrisk.html, 29 January 2016.

[12] G. Seymour, Predictive Inference, New York, Chapman and Hall, 1993.

[13] R. Hyndman, A. Koehler, Another look at measures of forecast accuracy, IIF. 22(4)(2006) 679–688.

[14] S. Rizvi, L. Wang, N. Nasrabadi, Nonlinearvector prediction using feed-forward neural networks, IEEE Trans Image Process. (1997) 1431-1436.