

วิธีจำแนกกลุ่มการเป็นโรคมะเร็งเต้านมด้วยวิธีการทางสถิติ

Classification Groups For Breast Cancer With Statistical Method

เกรียง กิจบำรุงรัตน์ (Krieng Kitbumrungrat)*

บทคัดย่อ

การวิจัยครั้งนี้เป็นการหาตัวแบบการจำแนกกลุ่มผู้ป่วยที่ได้รับการตรวจจากแพทย์พบว่าผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม โดยพิจารณาตามลักษณะเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านม ซึ่งมีตัวแปรอิสระ คือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ได้แก่ ความหนาของก้อนเนื้อ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of Cell Shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)), และการขยายตัวของเซลล์ (Mitoses (X_9)), ส่วนตัวแปรตามคือผลการตรวจจากแพทย์พบว่าผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม โดยใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) จะสรุปได้ว่า ตัวแบบ Logistic Regression มีอำนาจจำแนกได้ถูกต้องร้อยละ 96.90 สูงกว่าวิธี Discriminant Analysis มีอำนาจจำแนกได้ถูกต้องร้อยละ 96.10 ซึ่งตัวแบบ Logistic Regression ใช้ตัวแปรพยากรณ์ 4 ชนิดคือความหนาของก้อนเนื้อ (Clump Thickness (X_1)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และโครมาตินเฉพาะ (Bland Chromatin(X_7)) ส่วนตัวแบบ Discriminant Analysis ใช้ตัวแปรพยากรณ์ทั้ง 9 ชนิดในการพยากรณ์จัดกลุ่ม

คำสำคัญ: ตัวแบบการถดถอยลอจิสติก, ตัวแบบการวิเคราะห์การจำแนกกลุ่ม, สถิติทดสอบวอลต์, สถิติทดสอบวิลค์ แลมด้า.

* คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏธนบุรี

Faculty of Science and Technology, Dhonburi Rajabat University, Corresponding Author,

E-mail: Kriengstat@yahoo.com

Abstract

This research is a classification group, the patient is detected at any breast cancer or non-breast cancer. Assessment of characteristics of abnormal growth of breast cancer cells such as Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8), and Mitoses (X_9) are independent variables. The dependent variable is the result which is detected at any breast cancer or non-breast cancer by using Logistic Regression Model and Discriminant Model. Conclude that Logistic Regression Model has 96.90% classification higher than Discriminant Model has 96.10% classification. Logistic Regression Model can used predicted variables 4 variables are Clump Thickness (X_1), Marginal Adhesion (X_4), Bare Nuclei (X_6) and Bland Chromatin (X_7). The study results reveal that the discriminant analysis can used predicted variables 9 variables for classifying groups of breast cancer and non- breast cancer.

Keywords: Logistic Regression Model, Discriminant Model, Wald statistic, Wilks' Lambda

1.บทนำ

โรคมะเร็งจัดเป็น 1 ใน 10 ของสาเหตุการตายของประชากรโลก องค์การอนามัยโลก (WHO) รายงานว่า ในปี ค.ศ. 2008 มีผู้เสียชีวิตทั่วโลก 58 ล้านคน สาเหตุมาจากโรคมะเร็ง 7.6 ล้านคน คิดเป็นร้อยละ 13 ของจำนวนผู้เสียชีวิตทั้งหมด ปัจจุบันโรคมะเร็งเป็นโรคที่ก่อความสูญเสียให้กับชีวิตคนเราเป็นจำนวนมาก เช่น มะเร็งเต้านม มะเร็งตับ เป็นต้น มะเร็งเป็นเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ แผลแทรกซึม เบียดเข้าไปในเซลล์เนื้อเยื่อปกติ แบ่งตัวเร็วกว่าเซลล์ปกติหลายเท่าอยู่นอกเหนือการควบคุมของร่างกาย สามารถแพร่กระจายไปยังบริเวณอื่นของร่างกายที่ห่างไกลออกไป และไม่ติดต่อกับก้อนมะเร็งเดิม โดยแพร่เข้าไปในหลอดเลือดและน้ำเหลือง คุณสมบัติของก้อนเนื้อเยื่อ มะเร็งโตเร็วมาก บริเวณก้อนตรงกลางมักขาดสารอาหารทำให้เกิดการตายของเนื้อเยื่อมะเร็งเกิดขึ้น ผลต่อมาก็คือ เกิดการติดเชื้อได้ง่าย เพราะเนื้อตายและน้ำเหลืองเป็นอาหารที่ดีสำหรับเชื้อโรค การอักเสบติดเชื้อสู่กระแสเลือดและส่งผล ทำให้เกิดการเสียชีวิตในที่สุด มะเร็งทำให้ร่างกายขาดออกซิเจน เพราะมะเร็งแย่งออกซิเจนไปใช้ในการแบ่งตัวของเซลล์ ทำให้เม็ดเลือดขาวทำงานหนักคอยกำจัดเซลล์มะเร็งจนทำให้ภูมิคุ้มกันต่ำลงทำให้ไวต่อโรคที่เป็นมะเร็งสูญเสียการทำงาน ผลการขยายตัวทำให้เบียดอวัยวะข้างเคียงให้เสียหายได้ เซลล์มะเร็งจะทำการสร้างสารแองจิโอเจนเนซิสมีผลทำให้เกิดเส้นท่อเลือดมาเลี้ยงก้อนมะเร็งเพิ่มมากขึ้น เพื่อให้เพียงพอกับการเจริญแบ่งตัวของเนื้อเยื่อมะเร็ง โรคมะเร็งเต้านมเป็นโรคที่พบมากที่สุดอันดับหนึ่งของผู้หญิงสามารถถ่ายทอดทางพันธุกรรม ซึ่งเกิดจากฮอร์โมนที่ผิดปกติ ความอ้วน อาหารที่รับประทานมีไขมันสูง การกลูกลามของมะเร็งจากอวัยวะอื่นมายังเต้านม (ชัยวุฒิ ศรีสุข ,2554)

การรักษาผู้ป่วยมะเร็งเต้านมตามแนวทางของ สถาบันมะเร็งแห่งชาติ กรมการแพทย์ กระทรวงสาธารณสุข มีการรักษา 4 แบบ คือ การผ่าตัด การฉายรังสี การรักษาด้วยฮอร์โมนและระบบบำบัดด้วย systemic therapy และการรักษาด้วยเคมีบำบัด ในการรักษาโรคมะเร็งเต้านมมักจะใช้การรักษาหลายวิธีร่วมกัน เพื่อป้องกันการกลับมาเป็นซ้ำ และผลที่เกิดขึ้นจากการรักษาด้วยวิธีต่างๆ เหล่านี้คือ อาการข้างเคียงที่ไม่พึงประสงค์ เช่น คลื่นไส้อาเจียน ผื่นคัน ท้องเสีย โลหิตจาง เป็นต้น จากผลที่เกิดขึ้นดังกล่าวก่อให้เกิดความวิตกกังวล ทุกข์ทรมานทั้งกายและใจต่อผู้ป่วยเป็นอันมาก ค่าใช้จ่ายต่างๆ ที่มีราคาสูงมากจากทั้งการรักษาเอง การต้องหยุดพักงานของญาติที่ดูแลผู้ป่วยก็เป็นเหตุที่ทำให้ขาดรายได้เช่นกัน แม้ว่าผลการรักษาทางการแพทย์จะทำให้ผู้ป่วย

มะเร็งหายขาดได้ ถ้าตรวจพบในระยะเริ่มแรกแต่สำหรับผู้ป่วยที่เพิ่งตรวจพบว่าเป็นระยะท้ายๆ แล้ว การรักษาที่ไม่อาจทำให้หายขาดได้เนื่องจากมะเร็งเป็นโรคเรื้อรัง (สถาบันมะเร็งแห่งชาติกรมการแพทย์ กระทรวงสาธารณสุข, 2550).

ดังนั้นในการสร้างระบบสภาวะสุขภาพของผู้ป่วยโรคมะเร็งเต้านมนั้น จะพิจารณาจาก เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านม ซึ่งเป็นตัวบ่งบอกการลุกลาม ความรุนแรงของโรคมะเร็งเต้านม การที่แพทย์ทราบว่าเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมมีผลต่อการเป็นโรคมะเร็งเต้านม แพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ที่เกิดจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมแบบเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมเป็นตัวกำหนดการเป็นโรคมะเร็งเต้านม ซึ่งเป็นเซลล์มะเร็งที่แบ่งตัว เติบโตนอกเหนือการควบคุมจากร่างกาย สามารถเติบโตลุกลามเข้าไปในอวัยวะใกล้เคียงโดยเข้ากระแสเลือด ต่อมน้ำเหลืองได้ ซึ่งเรียกว่าการลุกลามแพร่กระจาย (Metastasis)

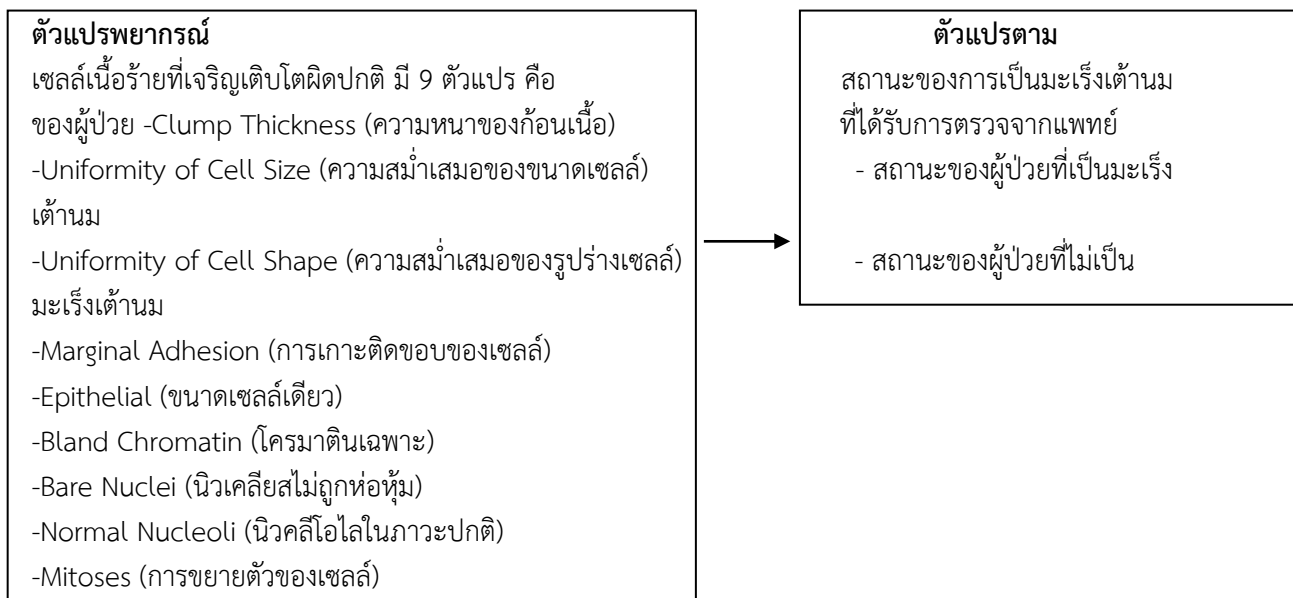
การสร้างตัวแบบจำแนกทางสถิติเป็นการช่วยให้แพทย์ทราบว่าเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมชนิดใดที่มีผลกระทบต่อระยะของโรคมะเร็งเต้านม ซึ่งแพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ในแบบเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น ด้วยเหตุผลดังกล่าวนี้ การวิจัยนี้จึงมีความสนใจที่จะศึกษาวิธีการสำหรับการพยากรณ์ความน่าจะเป็นที่ผู้ป่วยจะเป็นโรคมะเร็งเต้านมโดยอาศัยข้อมูลจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านม โดยใช้วิธีการทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธี ได้แก่ การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) อย่างไรก็ตาม ประสิทธิภาพของวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) ขึ้นอยู่ลักษณะการแจกแจงของข้อมูลที่นำมาวิเคราะห์ว่าเป็นไปตามข้อสมมติของตัวแบบหรือไม่โดยวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) มีข้อสมมติเกี่ยวกับตัวแบบค่อนข้างมาก เช่น ความสัมพันธ์ระหว่างกลุ่มและตัวแปรพยากรณ์ต้องเป็นแบบเชิงเส้น ตัวแปรพยากรณ์ควรมีการแจกแจงร่วมแบบปกติพหุและไม่มีความสัมพันธ์ระหว่างกันมากจนก่อให้เกิดปัญหาภาวะแนวร่วมเส้นตรง เป็นต้น สำหรับการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) มีข้อจำกัดน้อยกว่าวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis)

2.วัตถุประสงค์ของการวิจัย

เพื่อศึกษาว่าวิธีการจำแนกกลุ่ม 2 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) เพื่อนำมาใช้ในการพยากรณ์ความน่าจะเป็นที่ผู้ป่วยตรวจเจออาการแล้ว พบว่าจะเป็นโรคมะเร็งเต้านมหรือไม่เป็นโรคมะเร็งเต้านม โดยใช้เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติเป็นตัวแปรพยากรณ์ จะให้ผลในการพยากรณ์ได้ถูกต้องแตกต่างกันอย่างไรระหว่างวิธีการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) กับวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) และเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติใดบ้างที่สามารถใช้ในการพยากรณ์ได้

กรอบแนวคิดของการวิจัย

การวิเคราะห์การพยากรณ์เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับความน่าจะเป็นที่ผู้ป่วยจะเป็นมะเร็งเต้านมหรือไม่เป็นโรคมะเร็งเต้านม โดยใช้ชุดฐานข้อมูลจาก Wisconsin Breast Cancer Database (WBCD) (<http://www.archive.ics.uci.edu/ml/dataset.html>). ซึ่งลักษณะของตัวแปรของข้อมูลเป็นดังต่อไปนี้



ภาพที่ 1 กรอบแนวคิดของการวิจัย

3.วิธีการดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงทดลอง (Experiment Research) เพื่อศึกษาการพยากรณ์ของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับความน่าจะเป็นที่ผู้ป่วยโรคมะเร็งเต้านมที่ตรวจพบว่าเป็นโรคมะเร็งเต้านมหรือพบว่าเป็นโรคมะเร็งเต้านม โดยใช้ตัวอย่างชุดฐานข้อมูลการตรวจรักษาโรคมะเร็งเต้านม Wisconsin Breast Cancer Database (WBCD) จำนวน 683 ตัวอย่าง ซึ่งมีตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพ คือ ผลการตรวจจากแพทย์ที่พบว่าผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม ส่วนตัวแปรอิสระเป็นข้อมูลเชิงปริมาณของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติทั้ง 9 ชนิด ได้แก่ Clump Thickness (X_1), Uniformity of Cell Size (X_2), Uniformity of Cell Shape (X_3), Marginal Adhesion (X_4), Single Epithelial Cell Size (X_5), Bare Nuclei (X_6), Bland Chromatin (X_7), Normal Nucleoli (X_8), และ Mitoses (X_9) โดยใช้วิธีการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis)

ทฤษฎีและวิธีการศึกษา

1. ตัวแบบการถดถอยลอจิสติก (Logistic Regression Model)

ในการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) (วีรพันธ์ พงศาภักดี. 2541) การจำแนกกลุ่มทำได้โดยการประมาณค่าความน่าจะเป็นของการที่ค่าสังเกตจะอยู่ในกลุ่มที่กำหนด เช่น ในกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมเมื่อค่าของตัวแปรพยากรณ์เป็นค่าที่สังเกตจากค่าประมาณความน่าจะเป็นที่ได้ ค่าสังเกตจะถูกจำแนกให้อยู่ในกลุ่มที่กำหนดหรือกลุ่มอื่นขึ้นอยู่กับค่าความน่าจะเป็นที่เป็นจุดตัด (cut-off probability) โดยในการประมาณสัมประสิทธิ์การถดถอยลอจิสติกเพื่อประมาณความน่าจะเป็นนั้นใช้วิธีประมาณแบบภาวะน่าจะเป็นสูงสุด (maximum likelihood estimation) ตัวแบบของการถดถอยลอจิสติก คือ

$$\ln \left(\frac{P_1}{1-P_1} \right) = \alpha + \beta_1 x_1, \text{ เมื่อ } P_1 \text{ คือความน่าจะเป็นเหตุการณ์ที่ผู้ป่วยเป็นโรคมะเร็งเต้านม}$$

α คือ ค่าคงที่สำหรับกลุ่มของตัวแปรตาม

β_i คือ จำนวนสัมประสิทธิ์การถดถอย

$\beta_i x_i$ คือ $\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$

และ $x_1, x_2, \dots, x_i$ คือเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติจะเป็นตัวแปรพยากรณ์จำนวนกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) อาจนำมาใช้ในการพยากรณ์ได้ดีกว่าการวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) ในแง่ของข้อสมมติเบื้องต้นของตัวแบบมีเพียงความคลาดเคลื่อนต้องเป็นอิสระกัน ซึ่งไม่เท่ากับของการวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis)

2. ตัวแบบการวิเคราะห์การจำแนกกลุ่ม (Discriminant Model)

การวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) (กัลยา วานิชย์บัญชา, 2551). เป็นวิธีการจำแนกกลุ่มโดยการสร้างฟังก์ชันจำแนกกลุ่ม ซึ่งเป็นผลบวกเชิงเส้นของตัวแปรอิสระหรือตัวแปรพยากรณ์ $x_1, x_2, \dots, x_p$ เพื่อให้ความแปรปรวนระหว่างกลุ่มเมื่อเทียบกับความแปรปรวนภายในกลุ่มมีค่าสูงสุด นั่นคือหาค่าสัมประสิทธิ์ $a_1, a_2, \dots, a_p$ ของ $x_1, x_2, \dots, x_p$ ที่ทำให้ได้ฟังก์ชันจำแนกที่มีคุณสมบัติดังกล่าว ทำให้ได้ $Z_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_p x_{ip}$ เป็นคะแนนจำแนก (Discriminant score) สำหรับค่าสังเกตที่ i คะแนนที่ใช้เป็นจุดตัดในการจำแนก (cut-off score) คำนวณจากความน่าจะเป็นก่อน (a-priori probabilities) ของการเป็นสมาชิกกลุ่มและต้นทุนในการจำแนกผิด จากคะแนนจำแนกที่ได้ของแต่ละค่าสังเกตและคะแนนจุดตัด ค่าสังเกตจะถูกจำแนกว่าอยู่กลุ่มใด ตัวแบบ Discriminant Analysis มีข้อสมมติเบื้องต้นเกี่ยวกับตัวแปรพยากรณ์ว่า ในแต่ละกลุ่มตัวแปรพยากรณ์ต้องมีการแจกแจงร่วมแบบปกติพหุที่มีเวกเตอร์ค่าเฉลี่ยแตกต่างกัน แต่เมตริกซ์ความแปรปรวนร่วมเท่ากัน

4. ผลการวิจัย

1. การวิเคราะห์การถดถอยลอจิสติก

ขั้นตอนที่ 1 การวิเคราะห์ข้อมูล จะได้ว่าการตรวจพบว่าเป็นโรคมะเร็งเต้านม หรือตรวจพบว่าเป็นโรคมะเร็งเต้านมของผู้ป่วยมีสัดส่วนเท่ากันหรือไม่ ผลการตรวจสอบพบว่าผู้ป่วยทั้งหมด 683 คน เมื่อผู้ป่วยมาตรวจแล้วพบว่าไม่เป็นโรคมะเร็งเต้านมมีจำนวน 444 คน และพบว่าเป็นโรคมะเร็งเต้านมมีจำนวน 239 คน เมื่อทำการทดสอบความสัมพันธ์ระหว่างชุดของตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับตัวแปรตามของความน่าจะเป็นที่จะเป็นมะเร็งเต้านมหรือไม่ โดยใช้ Logit เป็น Link function ในการวิเคราะห์ข้อมูล ได้ผลการวิเคราะห์ ดังตารางที่ 1

ตารางที่ 1 ผลการทดสอบความสัมพันธ์ของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับการจะเป็นมะเร็งเต้านม

Model		-2Log Likelihood	Chi-Square	df	p-value
Intercept	Only	884.350			
	Final	102.888	781.462	9	0.000

จากตารางที่ 1 ผลการทดสอบความสัมพันธ์ของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติกับการจะเป็นมะเร็งเต้านม โดยใช้ Logit เป็น Link function ในการวิเคราะห์ข้อมูล พบว่าถ้าตัวแบบประกอบด้วยค่าคงที่เพียงอย่างเดียว จะให้ค่า -2 Log Likelihood (Intercept Only) = 884.350 และถ้าตัวแบบประกอบด้วยค่าคงที่และตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติจะให้ค่า -2 Log Likelihood (Final) = 102.888 นั่นคือ ตัว

แบบที่มีตัวแปรอิสระของเซลล์เนื้อ ร้ายที่เจริญเติบโตผิดปกติรวมอยู่ด้วยมีความเหมาะสมมากกว่าตัวแบบที่มีค่าคงที่เพียงอย่างเดียว นั่นคือมีตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอย่างน้อย 1 ตัว มีความสัมพันธ์กับตัวแปรตามของความเป็นมะเร็งเต้านมหรือไม่ ซึ่งจะได้ว่าจำนวนเหตุการณ์ตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติมีความสัมพันธ์กับสถานะการที่เป็นมะเร็งเต้านมหรือไม่ (Chi-Square = $[-2LL(\text{Intercept Only})] - [-2LL(\text{Final})] = (884.350 - 102.888) = 781.462$ ซึ่งได้ค่า P-value = $0.000 < 0.05$) จึงปฏิเสธ H_0 นั่นคือ ตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอย่างน้อย 1 ตัว มีความสัมพันธ์กับตัวแปรตาม คือ ผลการตรวจจากแพทย์ที่พบว่า ผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม

ขั้นตอนที่ 2 การทดสอบความมีนัยสำคัญของสัมประสิทธิ์การถดถอยแต่ละตัวของตัวแบบผู้ป่วยเป็นโรคมะเร็งเต้านมโดยการทดสอบ Likelihood Ratio Test

ตารางที่ 2 การทดสอบสัมประสิทธิ์การถดถอยของตัวแบบโดยการทดสอบ Likelihood Ratio Test

Likelihood Ratio Tests				
Effect	-2log Likelihood	Chi-Square	df	p-value
Intercept	518.600	415.712	1	(0.000)*
X1 (Clump Thickness)	120.719	17.830	1	(0.000)*
X2 (Uniformity of cell size)	102.889	0.001	1	0.976
X3 (Uniformity of cell shape)	104.738	1.850	1	0.174
X4 (Marginal Adhesion(x4))	110.312	7.424	1	(0.006)*
X5 (Single Epithelial cell size)	103.265	0.377	1	0.539
X6 (Bare nuclei)	122.067	19.179	1	(0.000)*
X7 (Bland Chromatin)	110.331	7.442	1	(0.006)*
X8 (Normal nucleoli)	106.663	3.775	1	0.052
X9 (Mitoses)	106.608	3.720	1	0.054

*: p-value < 0.05 มีนัยสำคัญทางสถิติระดับ 0.05

จากตารางที่ 2 พบว่า การวิเคราะห์ความมีนัยสำคัญทางสถิติของสัมประสิทธิ์การถดถอยลอจิสติก (Coefficient of Logistic Regression Model) ของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลต่อการเป็นมะเร็งเต้านมของผู้ป่วย มีค่า P-value < 0.05 จึงปฏิเสธ H_0 นั่นคือ ตัวแปรอิสระของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติอย่างน้อย 1 ตัว มีความสัมพันธ์กับผลการตรวจจากแพทย์ที่พบว่าผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม ซึ่งตัวแบบในการพยากรณ์ความน่าจะเป็นเพื่อการจำแนกกลุ่มผู้ป่วยจากการตรวจพบว่าผู้ป่วยเป็นโรคมะเร็งเต้านมหรือไม่ โดยการพิจารณาจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติมีดังนี้คือ ความหนาของก้อนเนื้อ (Clump Thickness (X_1)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และโครมาตินเฉพาะ (Bland Chromatin (X_7)) ดังนั้นจะได้ตัวแบบในการพยากรณ์ความน่าจะเป็นเพื่อการจำแนกกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม โดยมีกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมเป็นระดับอ้างอิงของตัวแปรตาม แล้วจะได้ตัวแบบในการพยากรณ์ในการจำแนกกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมดังตารางที่ 3

ตารางที่ 3 การทดสอบสัมประสิทธิ์การถดถอยแต่ละตัวของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลการตรวจจากแพทย์ที่พบว่าผู้ป่วยที่ตรวจแล้วเป็นมะเร็งเต้านม

Dependent variable (ผู้ป่วยที่ได้รับการตรวจจากแพทย์)	Independent variable (เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ)	B	Wald-test	P-value
พบว่าผู้ป่วยเป็นโรคมะเร็งเต้านม	Intercept	10.104	73.958	(0.000)*
	X ₁ (Clump Thickness)	-0.535	14.192	(0.000)*
	X ₂ (Uniformity of Cell Size)	0.006	0.001	0.976
	X ₃ (Uniformity of Cell Shape)	-0.323	1.958	0.162
	X ₄ (Marginal Adhesion)	-0.331	7.173	(0.007)*
	X ₅ (Single Epithelial Cell Size)	-0.097	0.381	0.537
	X ₆ (Bare Nuclei)	-0.383	16.659	(0.000)*
	X ₇ (Bland Chrom)	-0.447	6.808	(0.009)*
	X ₈ (Normal Nucleoli)	-0.213	3.562	0.059
	X ₉ (Mitoses)	-0.535	2.646	0.104

a : ผู้ไม่เป็นโรคมะเร็งเป็นตัวแปรอ้างอิงตามสภาวะสุขภาพจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ

*: P-value < 0.05 มีนัยสำคัญทางสถิติที่ระดับ 0.05

ผลการวิเคราะห์ข้อมูลเป็นดังนี้

จากตารางที่ 3 จะได้ตัวแบบในการพยากรณ์ความน่าจะเป็นเพื่อการจำแนกกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม โดยมีกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมเป็นระดับอ้างอิงของตัวแปรตาม แล้วจะได้ตัวแบบในการพยากรณ์ในการจำแนกกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม

$$\ln \left(\frac{P_1}{1-P_1} \right) = \alpha + \beta_i x_i$$

ซึ่งสามารถสร้างสมการตัวแบบถดถอยลอจิสติก (Logistic Regression Model) ได้ดังนี้

$$\ln \left(\frac{P_1}{1-P_1} \right) = 10.104 - 0.535 X_1 - 0.331 X_4 - 0.383 X_6 - 0.447 X_7$$

โดย p_1 คือ ความน่าจะเป็นที่ผู้ป่วยจะอยู่กลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านม ($Y = 1$)

การทดสอบสัมประสิทธิ์การถดถอยแต่ละตัวของเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลการตรวจจากแพทย์ที่พบว่าผู้ป่วยที่ตรวจแล้วเป็นมะเร็งเต้านม มีค่า P-value < 0.05 จึงปฏิเสธ H_0 นั่นคือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีความสัมพันธ์กับผลการตรวจจากแพทย์ที่ผู้ป่วยที่ตรวจแล้วเป็นมะเร็งเต้านม โดยการพิจารณาจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติได้แก่ ความหนาของก้อนเนื้อ (Clump Thickness (X_1)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และโครมาตินเฉพา (Bland Chromatin (X_7))

2. การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis)

การวิเคราะห์ข้อมูลด้วยการวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) ในการศึกษาปัจจัยเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติที่มีผลต่อผู้ป่วยที่ได้รับการตรวจจากแพทย์ พบว่า ผู้ป่วยเป็นโรคมะเร็งเต้านมหรือผู้ป่วยไม่เป็นโรคมะเร็งเต้านม เมื่อทำการทดสอบว่าแต่ละปัจจัยที่นำมาใช้จำแนกกลุ่มผู้ป่วยมีความแตกต่างกันระหว่างกลุ่มหรือไม่ด้วยสถิติทดสอบ Wilks' Lambda พบว่า ปัจจัยที่มีค่าเฉลี่ยแตกต่างกันระหว่างกลุ่มผู้ป่วยที่

เป็นโรคมะเร็งและผู้ป่วยไม่เป็นโรคมะเร็งเท่านั้น ที่ระดับนัยสำคัญ 0.05 ได้แก่ ความหนาของก้อนเนื้อ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin(X_7)), นิวคลีโอลัสในภาวะปกติ (Normal Nucleoli (X_8)) และการขยายตัวของเซลล์ (Mitoses (X_9)) ดังแสดงในตารางที่ 4

ตารางที่ 4 ค่าสถิติทดสอบการเท่ากันของค่าเฉลี่ยของแต่ละปัจจัย

Effect	Wilks' Lambda	F-test	p-value
X_1 (Clump Thickness)	0.489	711.423	(0.000)*
X_2 (Uniformity of cell size)	0.326	1406.123	(0.000)*
X_3 (Uniformity of cell shape)	0.324	1417.644	(0.000)*
X_4 (Marginal Adhesion(x4)	0.501	677.878	(0.000)*
X_5 (Single Epithelial call size)	0.523	622.158	(0.000)*
X_6 (Bare nuclei)	0.323	1426.240	(0.000)*
X_7 (Bland Chromatin)	0.425	921.010	(0.000)*
X_8 (Normal nucleoli)	0.484	727.471	(0.000)*
X_9 (Mitoses)	0.821	148.788	(0.000)*

*: P-value < 0.05 มีนัยสำคัญทางสถิติที่ระดับ 0.05

จากตารางที่ 4 ค่าสถิติทดสอบการเท่ากันของค่าเฉลี่ยของแต่ละปัจจัย เมื่อคัดเลือกปัจจัยที่ใช้ในการจำแนกกลุ่มพบว่า มีค่า P-value < 0.05 จึงปฏิเสธ H_0 นั่นคือปัจจัยที่ใช้ในการจำแนกกลุ่มคือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ทั้ง 9 ชนิด ได้แก่ ความหนาของก้อนเนื้อเซลล์ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลัสในภาวะปกติ (Normal Nucleoli (X_8)) และการขยายตัวของเซลล์ (Mitoses (X_9)) ดังนั้นจะได้ตัวแบบปัจจัยที่ใช้ในการจำแนกกลุ่ม Fishes's Linear discriminant function (FLDF) ดังตารางที่ 5

ตารางที่ 5 การทดสอบตัวแบบปัจจัยที่ใช้ในการจำแนกกลุ่ม Fishes's Linear discriminant function (FLDF)

Eigenvalue			
Function	Eigenvalue	% of Variance	Canonical Correlation
D (ค่าคะแนนการจำแนกกลุ่ม)	5.383	100.0	0.918

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	p-value
D (ค่าคะแนนการจำแนกกลุ่ม)	0.157	1254.832	8	(0.000)*

*: P-value < 0.05 มีนัยสำคัญทางสถิติที่ระดับ 0.05

Canominal Discriminant Function Coefficient

Effect	Function
	D (ค่าคะแนนการจำแนกกลุ่ม)
Intercept	-3.442
X1 (Clump Thickness)	0.183
X2 (Uniformity of cell size)	0.126
X3 (Uniformity of cell shape)	0.900
X4 (Marginal Adhesion(x4))	0.047
X5 (Single Epithelial call size)	0.058
X6 (Bare nuclei)	0.261
X7 (Bland Chromatin)	0.110
X8 (Normal nucleoli)	0.107
X9 (Mitoses)	0.006

ผลการวิเคราะห์ข้อมูลเป็นดังนี้

จากตารางที่ 5 ปัจจัยที่ใช้ในการจำแนกกลุ่มคือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ทั้ง 9 ชนิด ได้แก่ ความหนาของก้อนเนื้อเซลล์ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการขยายตัวของเซลล์ (Mitoses (X_9)) ซึ่งจะได้ตัวแบบปัจจัยที่ใช้ในการจำแนกกลุ่ม Fishes's Linear discriminant function (FLDF) คือ

$D = -3.442 + 0.183X_1 + 0.126X_2 + 0.90X_3 + 0.047X_4 + 0.058X_5 + 0.261X_6 + 0.110X_7 + 0.107X_8 + 0.006X_9$ เมื่อ D เป็นค่าคะแนนการจำแนกกลุ่มและมีค่าจุดตัดเป็น 5.383 ในการจำแนกกลุ่มผู้ป่วยใดที่มีคะแนนการจำแนกต่ำกว่าจุดตัดจะจัดให้อยู่ในกลุ่มที่ 1 (กลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม) และผู้ป่วยใดที่มีคะแนนการจำแนกกลุ่มมากกว่าหรือเท่ากับจุดตัดจะจัดให้อยู่ในกลุ่มที่ 2 (กลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านม) ส่วนสมการจำแนกที่ได้พบว่าที่ระดับนัยสำคัญ 0.05 มี Wilks' Lamda = 0.157, P-value=0.000 และ Chi-square = 1254.832 ดังนั้นจะได้ตัวแบบการจำแนกแบ่งกลุ่มของกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมและกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านม ได้ดังตารางที่ 6

ตารางที่ 6 การทดสอบตัวแบบปัจจัยที่ใช้ในการจำแนกกลุ่มผู้ป่วยเป็นโรคมะเร็งเต้านมและกลุ่มผู้ป่วยไม่เป็นโรคมะเร็งเต้านม (Classification Function Coefficients)

Effect	Y	
	D ₁ (กลุ่มของผู้ป่วยที่เป็นโรคมะเร็งเต้านม)	D ₂ (กลุ่มของผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านม)
Intercept	-23.536	-3.288
X1 (Clump Thickness)	1.630	0.742
X2 (Uniformity of cell size)	0.312	0.301
X3 (Uniformity of cell shape)	0.495	0.057
X4 (Marginal Adhesion(x4))	0.339	0.106
X5 (Single Epithelial cell size)	0.995	0.708
X6 (Bare nuclei)	1.457	0.188
X7 (Bland Chromatin)	1.198	0.664
X8 (Normal nucleoli)	0.054	0.026

ผลการวิเคราะห์ข้อมูลเป็นดังนี้

จากตารางที่ 6 ปัจจัยที่ใช้ในการจำแนกกลุ่มของผู้ป่วยที่เป็นโรคมะเร็งเต้านมและกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านม (Y) ด้วยตัวแบบปัจจัยที่ใช้จำแนกกลุ่ม Fishes's linear discriminant functions คือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ 8 ชนิด ได้แก่ ความหนาของก้อนเนื้อเซลล์ (Clump Thickness (X₁)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X₂)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X₃)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X₄)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X₅)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X₆)), โครมาตินเฉพาะ (Bland Chromatin (X₇)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X₈)) ซึ่งจะยกเว้น การขยายตัวของเซลล์ (Mitoses (X₉))

ดังนั้นสามารถสร้างตัวแบบการจำแนกประเภทแบ่งกลุ่มของผู้ป่วยที่เป็นโรคมะเร็งเต้านมและกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมด้วยตัวแบบปัจจัยที่ใช้จำแนกกลุ่ม Fishes's linear discriminant functions ได้ดังนี้

ตัวแบบกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมคือ

$$D_1 = -23.536 + 1.63X_1 + 0.312X_2 + 0.495X_3 + 0.339X_4 + 0.995X_5 + 1.457X_6 + 1.198X_7 + 0.054X_8$$

ตัวแบบกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมคือ

$$D_2 = -3.288 + 0.742X_1 + 0.301X_2 + 0.057X_3 + 0.106X_4 + 0.708X_5 + 0.188X_6 + 0.664X_7 + 0.026X_8$$

3. ผลการเปรียบเทียบประสิทธิภาพ

ประสิทธิภาพของตัวแบบ Discriminant Analysis และ Logistic Regression Analysis พิจารณาจากอัตราการจำแนกกลุ่มได้ถูกต้องซึ่งแสดงไว้ในตารางที่ 7

ตารางที่ 7 จำนวนและร้อยละของผู้ป่วยที่จำแนกกลุ่มได้ถูกต้องสำหรับตัวแบบ Discriminant Analysis และ Logistic Regression Analysis เมื่อใช้กับกลุ่มตัวอย่างผู้ป่วย 683 คน

ตัวแบบ	ตัวแปรพยากรณ์	เป็นมะเร็ง (%)	ไม่เป็นมะเร็ง (%)	รวม (%)
Discriminant	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$	224 (92.90)	448 (97.80)	672 (96.10)
Logistic	$X_1, X_4, X_6, X_7,$	228 (95.40)	434 (97.70)	662 (96.90)

จากตารางที่ 7 พบว่า ตัวแบบ Discriminant Analysis มีอำนาจจำแนกค่อนข้างสูง สามารถจำแนกกลุ่มผู้ป่วยที่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 92.90 และสามารถจำแนกกลุ่มผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 97.80 โดยภาพรวมสามารถจำแนกได้ถูกต้องร้อยละ 96.10

ตัวแบบ Logistic Regression มีอำนาจจำแนกสูงกว่าวิธี Discriminant ซึ่งสามารถจำแนกกลุ่มผู้ป่วยที่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 95.40 และสามารถจำแนกกลุ่มผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 97.70 โดยภาพรวมสามารถจำแนกได้ถูกต้องร้อยละ 96.90

5.สรุปและอภิปรายผล

การพยากรณ์ความน่าจะเป็นเพื่อการจำแนกกลุ่มผู้ป่วยจากการตรวจพบว่าเป็นโรคมะเร็งเต้านมหรือไม่เป็นโรคมะเร็งเต้านม ซึ่งถือว่าเป็นสิ่งจำเป็น เพราะจะได้เป็นเครื่องมือในการสร้างระบบสภาวะสุขภาพจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ ในการศึกษาที่ใช้เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ 9 ชนิดได้แก่ ความหนาของก้อนเนื้อเซลล์ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพาะ (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) และการขยายตัวของเซลล์ (Mitoses (X_9)) โดยใช้วิธีการทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธี ได้แก่ การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) ผลการวิเคราะห์พบว่า ตัวแบบ Logistic Regression Analysis มีอำนาจจำแนกสูงกว่าวิธี Discriminant Analysis ซึ่งสามารถจำแนกกลุ่มผู้ป่วยที่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 95.40 และสามารถจำแนกกลุ่มผู้ป่วยที่ไม่เป็นมะเร็งเต้านมได้ถูกต้องร้อยละ 97.70 โดยภาพรวมสามารถจำแนกได้ถูกต้องร้อยละ 96.90 ในการเลือกตัวแปรพยากรณ์เข้าสู่ตัวแบบ จะได้ว่าตัวแบบ Logistic Regression Analysis จะใช้จำนวนตัวแปรพยากรณ์น้อยและให้ความแม่นยำในการพยากรณ์สูงกว่าวิธี Discriminant Analysis โดยพบว่าตัวแบบ Logistic Regression Analysis ใช้ตัวแปรพยากรณ์ 4 ชนิด คือ ความหนาของก้อนเนื้อ (Clump Thickness (X_1)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)) และ โครมาตินเฉพาะ (Bland Chromatin (X_7)) ซึ่งจะสามารถสร้างสมการถดถอยลอจิสติก (Logistic Regression) ได้ดังนี้

$$\ln \left(\frac{P_1}{1-P_1} \right) = 10.104 - 0.535 X_1 - 0.331 X_4 - 0.383 X_6 - 0.447 X_7$$

ส่วนตัวแบบ Discriminant Analysis ใช้ตัวแปรพยากรณ์ทั้ง 9 ชนิด ซึ่งจะแบ่งกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านม และกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมด้วยตัวแบบปัจจัยที่ใช้จำแนกกลุ่ม Fishes's linear discriminant functions

ซึ่งจะสามารถสร้างสมการตัวแบบได้ดังนี้

ตัวแบบกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมคือ

$$D_1 = -23.536 + 1.63X_1 + 0.312X_2 + 0.495X_3 + 0.339X_4 + 0.995X_5 + 1.457X_6 + 1.198X_7 + 0.547X_8$$

และตัวแบบกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านมคือ

$$D_2 = -3.288 + 0.742X_1 + 0.301X_2 + 0.057X_3 + 0.106X_4 + 0.708X_5 + 0.188X_6 + 0.664X_7 + 0.026X_8$$

ปัจจัยที่ใช้ในการจำแนกกลุ่มของกลุ่มผู้ป่วยที่เป็นโรคมะเร็งเต้านมและกลุ่มผู้ป่วยที่ไม่เป็นโรคมะเร็งเต้านม (Y) ด้วยตัวแบบปัจจัยที่ใช้จำแนกกลุ่ม Fishes's linear discriminant functions คือ เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติ 8 ชนิด ได้แก่ ความหนาของก้อนเนื้อเซลล์ (Clump Thickness (X_1)), ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size (X_2)), ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of cell shape (X_3)), การเกาะติดขอบของเซลล์ (Marginal Adhesion (X_4)), ขนาดเซลล์เดี่ยว (Single Epithelial Cell Size (X_5)), นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei (X_6)), โครมาตินเฉพา (Bland Chromatin (X_7)), นิวคลีโอลในภาวะปกติ (Normal Nucleoli (X_8)) ซึ่งจะยกเว้น การขยายตัวของเซลล์ (Mitoses (X_9))

ดังนั้นในการสร้างระบบสุขภาพของผู้ป่วยโรคมะเร็งเต้านมนั้น จะพิจารณาจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านม ซึ่งเป็นตัวบ่งบอกการลุกลาม ความรุนแรงของโรคมะเร็งเต้านม การที่แพทย์ทราบว่าเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมมีผลต่อการเป็นโรคมะเร็งเต้านม แพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ที่เกิดจากเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมแบบเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น เซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมเป็นตัวกำหนดการเป็นโรคมะเร็งเต้านม ซึ่งเป็นเซลล์มะเร็งที่แบ่งตัว เติบโตนอกเหนือการควบคุมจากร่างกาย สามารถเติบโตลุกลามเข้าไปในอวัยวะใกล้เคียงโดยเข้ากระแสเลือด ต่อมาน้ำเหลืองได้ ซึ่งเรียกว่าการลุกลามแพร่กระจาย (Metastasis)

การสร้างตัวแบบจำแนกทางสถิติเป็นการช่วยให้แพทย์ทราบว่าเซลล์เนื้อร้ายที่เจริญเติบโตผิดปกติของโรคมะเร็งเต้านมชนิดใดที่มีผลกระทบต่อระยะของโรคมะเร็งเต้านม ซึ่งแพทย์จะสามารถวางแผนการรักษาและพยากรณ์โรคได้และยังสามารถนำไปเทียบเคียงกับผลการวิจัยที่ทำกับคนไข้ในแบบเดียวกัน เพื่อให้ทราบข้อมูลเกี่ยวกับโรคและการรักษาได้ดียิ่งขึ้น ตัวแบบจำแนกทางสถิติจะมีประสิทธิภาพก็ต่อเมื่อข้อมูลที่วิเคราะห์มีลักษณะการแจกแจงสอดคล้องกับข้อสมมติ วิธีการจำแนกกลุ่มแบบพารามตริกทางสถิติที่นำมาใช้ในการพยากรณ์ 2 วิธี คือ การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) และวิธีการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) ในภายหลังควรมีการนำวิธีการทางนอนพารามตริกมาใช้กับระบบสุขภาพของผู้ป่วย เช่น เครือข่ายประสาทเทียม (Neural Networks) ต้นไม้การตัดสินใจ (Decision Trees) หรือการรู้จำลักษณะเฉพาะ (Trait Recognition Approach) เป็นต้น

เอกสารอ้างอิง

ภาษาไทย

- ชัยวุฒิ ศรีสุโข (2554). โรคมะเร็งเต้านม: อมรินทร์พริ้นติ้งแอนด์พับลิชชิ่ง กรุงเทพฯ; 2554; หน้า 5-10.
- สถาบันมะเร็งแห่งชาติ (2550). แนวทางการตรวจวินิจฉัยและรักษาโรคมะเร็งเต้านม. พิมพ์ครั้งที่ 1 พ.ศ. 2550. กรุงเทพฯ
- กรมการแพทย์ กระทรวงสาธารณสุข (2546). การประเมินเทคโนโลยีการตรวจวินิจฉัยมะเร็งเต้านมในระยะเริ่มแรกที่ เหมาะสมสำหรับประเทศไทย: ตุลาคม 2546. หน้า28-30.
- ชุดฐานข้อมูลจาก Wisconsin Breast Cancer Database (WBCD).
<http://www.archive.ics.uci.edu/ml/dataset.html>.
- วีรานันท์ พงศาภักดี (2541). การวิเคราะห์ข้อมูลเชิงกลุ่ม : ทฤษฎีและการประยุกต์ (GLIM และ SPSS/FW). พิมพ์ครั้งที่ 2. นครปฐม: ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์, 2541.
- กัลยา วานิชย์บัญชา. (2551) การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- กัลยา วานิชย์บัญชา. (2551) การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. พิมพ์ครั้งที่ 6. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
- เกรียง กิจบำรุงรัตน์. (2555). "ตัวแบบการถดถอยเชิงอันดับสำหรับวิเคราะห์ระยะการเป็นโรคมะเร็งเต้านม" วารสารวิศวกรรมและเทคโนโลยี มหาวิทยาลัยรังสิต วารสารเล่มปีที่ 14 ฉบับที่ 2 (วารสาร RSU JET Vol.14, No. 2, 2011) หน้า 27 – 34
- เกรียง กิจบำรุงรัตน์. (2556). "ตัวแบบการถดถอยลอจิสติกเชิงอันดับสำหรับคณะวิชาและการวิเคราะห์เกรดนักศึกษา"(Ordinal Logistic Regression Model for Faculty and Student Grade Analysis) วารสารวิชาการ(Veridian E- Journal) มหาวิทยาลัยศิลปากร สาขามนุษยศาสตร์และสังคมศาสตร์ ปีที่ 6 ฉบับเดือนมกราคม – เมษายน 2556.
- เกรียง กิจบำรุงรัตน์. (2560). "ตัวแบบการถดถอยแบบมัลติโนเมียลลอจิสติกเพื่อการจัดกลุ่มการเรียนรู้และตัวแบบการถดถอยลอจิสติกเชิงอันดับสำหรับการวิเคราะห์เกรดนักศึกษา" (Multinomial Logistic

ภาษาต่างประเทศ

- Regression Model for Learning Classification and Ordinal Logistic Regression Model for Student Grade Analysis) Veridian E- Journal Science and Technology Silpakorn University Vol.4 No. 2 March–April 2017
- Krieng Kitbumrungrat. (2012) "Comparison Logistic Regression and Discriminant Analysis in classification groups for Breast Cancer." IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.5, MAY 2012. pages 111-115.
- Doungporn Maiprasert, Krieng Kitbumrungrat. (2012) "Comparison Multinomial Logistic Regression and Discriminant Analysis in predicting the stage of breast cancer." IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.4, April 2012. pages 44-52.

Doungporn Maiprasert, Krieng Kitbumrungrat. (2012) “Prediction the stages of breast cancer using PCA-ANN algorithm compare with multinomial logistic regression.” IJITNA International Journal of Information Technology and Network Application, Vol. 1, No. 1, May 2012. pages 38-44.

Doungporn Maiprasert, Krieng Kitbumrungrat. (2012) “Multinomial Logistic Regression of Breast Cancer” RANGSIT JOURNAL OF ARTS AND SCIENCES (RJAS), Volume 2, Number 1: January-June 2012. pages 25-35.