ปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วย

Influencing Factors with the Effectiveness of Bone Tumors Treatment in Patients

ณัฏฐณิชา เมืองมูล[1*] จุฑาภรณ์ สินสมบูรณ์ทอง[1] และ ผกาเกษ วัตุยา[2]

Nuttanicha Muangmool[1*] Juthaphorn Sinsomboonthong[2] and Pakaket Wattuya[3]

*E-mail: Nuttanicha.m@ku.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วย โดยใช้ 3 วิธีคือ การทดสอบสมมุติฐานด้วยสถิติทดสอบไคสแควร์, การสุ่มป่าไม้ (Random Forest) และ การถดถอยโลจิสติค (Logistic Regression) โดยใช้ข้อมูลจาก Memorial Sloan Kettering Cancer Center (MSKCC) เป็นชุดข้อมูลที่มีการเปิดเผยแหล่งข้อมูลจำนวน 500 เรคคอร์ด ข้อมูลที่ศึกษาประกอบไปด้วยตัวแปรตาม (Y) คือ ผลการรักษาของผู้ป่วย และ ตัวแปรอิสระ (X_i) 7 ตัว ได้แก่ เพศ (X_1), อายุ (X_2), ระดับความรุนแรงของเนื้องอก (X_3), ประเภทเนื้องอกกระดูกทางพยาธิสภาพ (X_4), การแบ่งเนื้องอกกระดูกโดย MSKCC (X_5), ตำแหน่งของเนื้องอกกระดูก (X_6) และวิธีการรักษา (X_7)

ผลการศึกษาพบว่าทั้ง 7 ปัจจัยมีผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกโดยปัจจัยที่มีผลต่อประสิทธิภาพการรักษาของผู้ป่วยมากที่สุดคือวิธีการรักษา ซึ่งผลการวิเคราะห์สอดคล้องกัน 3 วิธี

คำสำคัญ: เนื้องอกกระดูก, การถดถอยโลจิสติค, การสุ่มป่าไม้, ไคสแควร์

Abstract

The purpose of this research is to study the factors that affect the effectiveness of the patient's bone tumor treatment with three methods includes Chi-square test, random forest and logistic regression analysis. Using data set from open source of the Memorial Sloan Kettering Cancer Center (MSKCC) that consists of 500 records. The study data include dependent variable

[1] รองศาสตราจารย์.ดร., ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

[2] ผู้ช่วยศาสตราจารย์.ดร., ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

*E-mail: Nuttanicha.m@ku.th (Corresponding author)

(Y) which is the patient's status and seven independent variables (X_i), such as Sex (X_1), Age (X_2), Grade (X_3), Histological type (X_4), MSKCC type (X_5), Site of primary STS (X_6) and Treatment (X_7).

The results show that all seven factors affect the patient's status and treatment is the most effective factor of this for three methods.

Keywords: Bone tumor, Logistic regression, Random forest, Chi-square

บทนำ

เนื้องอกกระดูก (Bone tumor) คือการเจริญเติบโตที่ผิดปกติของเนื้อเยื่อในกระดูก โดยสามารถแบ่งได้เป็นเนื้องอกที่ไม่เป็นมะเร็ง (Benign) และเนื้องอกมะเร็ง (Malignant) ซึ่งเกิดขึ้นจากเซลล์ของกระดูกโดยตรง และจากการค้นคว้าหาข้อมูลได้มีการบอกว่าระดับความรุนแรงของเนื้องอก (Grade) และชนิดพยาธิวิทยาของเนื้องอก (Histological type) โดยจำแนกชนิดของเนื้องอกกระดูกตาม WHO classification of soft tissue tumor 2020 ได้ใช้ลักษณะทางพยาธิวิทยาในการจำแนกเนื้องอกกระดูกตามแหล่งกำเนิดของเซลล์ ซึ่งประกอบด้วย Osteogenic tumors ,Chondrogenic tumors, Fibrogenic tumors, Ewing sarcoma family, Giant cell tumor of bone และ Undifferentiated or pleomorphic sarcoma of bone (Sbaraglia et al, 2021) มักจะส่งผลต่อผลลัพธ์หลังการรักษามากที่สุด (Lucas et al., 2008) จากงานวิจัยของ Vangala et al. (2021) ได้ทำการวิเคราะห์ความสำคัญของการทำนายโรคของปัจจัยทางคลินิกและลักษณะทางเนื้อเยื่อในผู้ป่วยมะเร็งกระดูกหลังการให้เคมีบำบัดก่อนการผ่าตัด โดยเน้นที่การระบุตัวทำนายการอยู่รอดและผลลัพธ์ของการรักษา ผลการศึกษาที่สำคัญชี้ให้เห็นว่าปัจจัยเฉพาะ เช่น อัตราการตายของเนื้องอก การตอบสนองทางเนื้อเยื่อวิทยา ขนาดของเนื้องอก และระยะทางคลินิก มีความเกี่ยวข้องอย่างมากกับผลลัพธ์ของผู้ป่วย ผู้ป่วยที่มีอัตราการตายที่สูงขึ้นหลังการให้เคมีบำบัดและผู้ป่วยที่มีการตอบสนองทางเนื้อเยื่อวิทยาที่ดีโดยทั่วไปจะมีอัตราการรอดชีวิตที่เพิ่มขึ้น ในงานวิจัย Sbaraglia et al. (2021) ได้อธิบายถึงการเปลี่ยนแปลงที่สำคัญในระบบจำแนกโรคเนื้องอกเนื้อเยื่ออ่อน (Soft Tissue Tumors) เพื่อเพิ่มความแม่นยำในการวินิจฉัยและการพยากรณ์โรค ในงานวิจัยของ Johannes et al. (2023) ได้ทำการพัฒนาระบบตรวจหาปัจจัยเสี่ยงของโรคเบาหวานชนิดที่ 2 เนื่องจากมีอัตราการป่วยและเสียชีวิตจากโรคเบาหวานชนิดที่ 2 (DM)โดยใช้การวิเคราะห์ทางสถิติรวมถึงเทคโนโลยีเรียนรู้เชิงเครื่องจักรการสุ่มป่าไม้ด้วยอัตราความแม่นยำ >80% ผลลัพธ์ที่ได้คือ ระดับน้ำตาลในเลือดเป็นปัจจัยที่สำคัญที่สุดที่มีผลต่อการเกิดโรคเบาหวานชนิดที่ 2 นอกจากนี้งานวิจัยของ Li et al. (2022) ได้ทำการศึกษาย้อนหลังข้อมูลทางคลินิกของผู้ป่วยมะเร็งเต้านมที่ได้รับการรักษาที่โรงพยาบาลในเครือแห่งที่สามของ Sun Yat-sen University และ Liuzhou Women and Children’s Medical Center ตั้งแต่เดือนมกราคม 2013 ถึงเดือนธันวาคม 2020 โดยใช้การขจัดคุณลักษณะแบบวนซ้ำด้วยการสุ่มป่าไม้ และใช้วิธีการสุ่มสุ่มป่าไม้เพื่อพัฒนาแบบจำลองเชิงทำนายสำหรับการเกิด

ซ้ำของมะเร็งเต้านม แบ่งชุดการฝึกอบรมและชุดตรวจสอบเป็นผู้ป่วย 623 และ 151 รายตามลำดับ โดยมีตัวแปร 14 ตัว ได้แก่ ระยะมะเร็งทางพยาธิวิทยา(TNM), แกมมา-กลูตาเมลทรานสเปปติเดส, คอเลสเตอรอลทั้งหมด, Ki-67 จำนวนลิมโฟไซต์, ไลโปโปรตีนความหนาแน่นต่ำ, อายุ, อะพอลิโพโปรตีนB, ไลโปโปรตีนความหนาแน่นสูง, โกลบูลิน, อัตราส่วนจำนวนนิวโทรฟิลต่อจำนวนลิมโฟไซต์, อะลานีนอะมิโนทรานสเฟอเรส, ไตรกลีเซอไรด์, และอัลบูมิน ต่อโกลบูลิน ผลลัพธ์ที่ได้คือ ระยะมะเร็งทางพยาธิวิทยา(TNM) มีผลมากที่สุดในการเกิดซ้ำของมะเร็งเต้านม จากนี้งานวิจัยของ Jin et al. (2023) การศึกษาครั้งนี้มีจุดมุ่งหมายเพื่อพัฒนาและทดสอบตัวแบบการเรียนรู้ของเครื่องจักรที่ใช้การถดถอยโลจิสติค และ การสุ่มป่าไม้เพื่อทำนายการตอบสนองที่ไม่ดีต่อเคมีบำบัดก่อนการผ่าตัดในผู้ป่วยมะเร็งเต้านมโดยใช้ผู้ป่วยทั้งหมด 315 รายผลลัพธ์ที่ได้คือ ตัวแบบป่าไม้มีประสิทธิภาพที่ดีในการทำนายการกลับเป็นซ้ำของเนื้องอก จากงานวิจัยของ Alqahtani et al. (2024) ใช้ข้อมูลจากระบบเฝ้าระวังปัจจัยเสี่ยงทางพฤติกรรม (BRFSS) ปี 2021 และใช้เทคนิคการประมวลผลข้อมูลเบื้องต้นที่ครอบคลุม การวิเคราะห์การถดถอยแบบโลจิสติก, การสุ่มป่าไม้, โครงข่ายประสาทเทียม และอีกหลายวิธีเพื่อความสัมพันธ์ระหว่างคุณลักษณะและความเสี่ยงของภาวะก่อนเบาหวาน ผลลัพธ์ที่ได้พบว่าความสัมพันธ์ที่สำคัญระหว่างภาวะก่อนเป็นเบาหวานคือระดับคอเลสเตอรอส, หมวดหมู่ดัชนีมวลกาย, สถานะความดันโลหิตสูง, กลุ่มอายุ และหมวดหมู่รายได้

การวิจัยในครั้งนี้ผู้วิจัยสนใจศึกษาปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูก ได้แก่ เพศ, อายุ, ระดับความรุนแรงของเนื้องอก, ประเภทเนื้องอกกระดูกทางพยาธิสภาพ, การแบ่งเนื้องอกกระดูกโดย MSKCC, ตำแหน่งของเนื้องอกกระดูก และวิธีการรักษา เพื่อให้สามารถวางแผนการรักษาที่เหมาะสมในอนาคต การเข้าใจความสัมพันธ์ระหว่างปัจจัยเหล่านี้กับผลการรักษาจะช่วยเพิ่มโอกาสในลดอัตราการหายขาดจากโรค (No Evidence of Disease: NED) ลดอัตราการกลับเป็นซ้ำ (Alive With Disease: AWD) และลดอัตราการตาย (Dead:D) โดยใช้วิธีการศึกษาในครั้งนี้ 3 วิธีได้แก่ การทดสอบสมมุติฐานด้วยสถิติทดสอบไคสแควร์, Random Forest และ Logistic Regression Analysis (Géron, 2019) โดยเกณฑ์ในการพิจารณาลำดับความสำคัญของปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วย คือ แมทริกซ์สหสัมพันธ์ และ Feature Importance

ทฤษฎีที่เกี่ยวข้อง

1. การทดสอบไคสแควร์ (Chi-square test)

การทดสอบไคสเป็นการทดสอบแบบไม่ใช้พารามิเตอร์สำหรับตัวแปรเชิงกลุ่ม การทดสอบนี้ใช้เพื่อพิจารณาว่าข้อมูลที่สังเกตได้แตกต่างจากข้อมูลที่คาดหวังอย่างมีนัยสำคัญและเพื่อพิจารณาว่าตัวแปรกลุ่มสองตัวมีความเกี่ยวข้องกันหรือไม่ (Turney, 2023)

วิธีคำนวณสถิติการทดสอบ

สถิติไคสแควร์เป็นสถิติการทดสอบความเป็นอิสระของตัวแปรเชิงกลุ่ม 2 ตัวดังนี้

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

เมื่อ O_{ij} คือ จำนวนจริงที่สังเกตได้ (Observed frequency) ในตำแหน่งแถว i, คอลัมน์ j

E_{ij} คือ จำนวนที่คาดหวัง (Expected frequency) ในตำแหน่งแถว i, คอลัมน์ j

ค่า E_{ij} คำนวณได้จาก

$$E_{ij} = \frac{(\text{row total}_i \text{ X column total}_j)}{\text{total}}$$

เมื่อ row total_i คือ ผลรวมของแถวที่ i

column total_j คือ ผลรวมของคอลัมน์ที่ j

total = ผลรวมของข้อมูลทั้งหมดในตาราง

2. เมทริกซ์สหสัมพันธ์ (Correlation Matrix)

เมทริกซ์สหสัมพันธ์เป็นเทคนิคทางสถิติที่ใช้ในการประเมินความสัมพันธ์ระหว่างตัวแปรสองตัวในชุดข้อมูล โดยค่าสหสัมพันธ์มีค่าตั้งแต่ -1 ถึง +1 ถ้าค่าสหสัมพันธ์เข้าใกล้ -1 หรือ +1 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นตรงกันสูงแต่ถ้าค่าสหสัมพันธ์เข้าใกล้ 0 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นตรงกันน้อย (Wagavkar, 2024)

3. วิธีการสุ่มป่าไม้ (Random Forest)

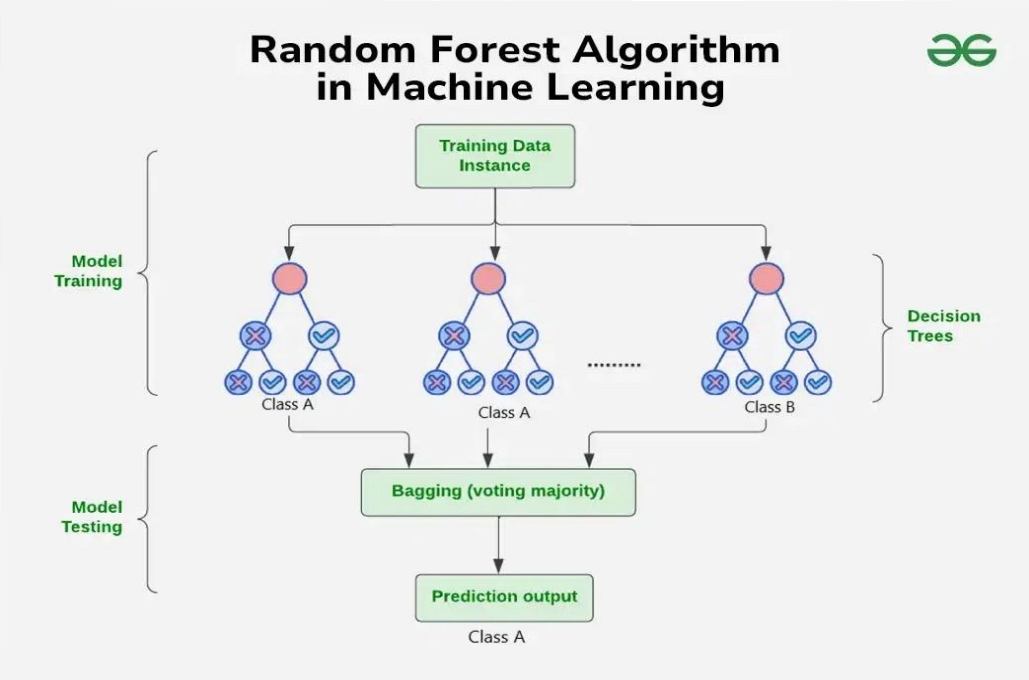
ใช้สำหรับการจำแนกประเภท (Classification) และ การทำนาย (Prediction) โดยใช้หลักการของการสร้าง หลายต้นไม้การตัดสินใจ (Decision Trees) และใช้ผลลัพธ์จากการรวมกันของต้นไม้หลายๆ ต้นเพื่อทำการทำนายผลที่แม่นยำขึ้น

ตัวแบบ Random Forest ใช้การตัดสินใจหลายต้นไม้ตัดสินใจ (Decision Trees) โดยการสุ่มข้อมูลและคุณลักษณะ (Features) ที่จะใช้ในการสร้างแต่ละต้นไม้ สำหรับการสร้างแต่ละต้นไม้จะทำการ สุ่มตัวอย่างข้อมูล (Sampling) โดยการใช้เทคนิคที่เรียกว่า การบูตสแตรป (Bootstrapping) ซึ่งเป็นการสุ่มเลือกข้อมูลบางส่วนจาก

ชุดข้อมูลที่มี โดยอาจมีการเลือกข้อมูลซ้ำกันในการสร้างแต่ละ node การตัดสินใจของต้นไม้ จะเลือกคุณลักษณะบางตัวจากทั้งหมดแต่ไม่ใช้ทั้งหมด เพื่อเพิ่มความหลากหลายในการสร้างต้นไม้และลดการพึ่งพาคุณลักษณะบางตัว

การสร้างต้นไม้การตัดสินใจแต่ละต้นไม้ในการสุ่มป่าไม้จะสร้างการตัดสินใจที่แตกต่างกันไปโดยจะใช้ข้อมูลที่สุ่มเลือกและคุณลักษณะที่สุ่มเลือกในการสร้างกฎการตัดสินใจในแต่ละ node จะสร้างขึ้นจนถึง ความลึก (Depth) ที่สูงสุดหรือเมื่อการแบ่งข้อมูลไม่สามารถทำได้ดีขึ้น

การทำนายจากหลายต้นไม้เมื่อมีการทำนายผลจากการสุ่มป่าไม้การทำนายจะทำโดยการรวมผลลัพธ์จากต้นไม้หลายๆ ต้นในการจำแนกประเภทผลลัพธ์จะถูกกำหนดโดยการโหวต (Voting) ของต้นไม้ทั้งหมด แต่ละต้นไม้จะให้ผลลัพธ์เป็นคลาสที่ทำนาย แล้วเลือกคลาสที่ได้รับการโหวตมากที่สุดในการทำนายค่าตัวเลข ผลลัพธ์จะเป็นค่าเฉลี่ย (Average) ของค่าที่ทำนายจากทุกต้นไม้ (Logunova, 2022)



รูปที่ 1 การสุ่มป่าไม้

4. วิธีถดถอยโลจิสติค (Logistic Regression)

การถดถอยโลจิสติคตามจำนวนกลุ่มของตัวแปรตามจะแบ่งได้ 2 ประเภท นั่นคือ เมื่อตัวแปรตามมีเพียง 2 ค่าหรือกลุ่มข้อมูลแบ่งออกได้เป็น 2 กลุ่มย่อย กรณีนี้จะเรียกว่า การวิเคราะห์การถดถอยโลจิสติคแบบทวิ (Binary Logistic Regression) เช่น การศึกษาการวินิจฉัยการเป็นหรือไม่เป็นโรคตับ มีตัวแปรตาม คือ มีแนวโน้มเป็นโรคตับ และ ไม่มีแนวโน้มเป็นโรคตับ และตัวแปรอิสระ คือ ปริมาณโปรตีนในกระแสเลือด (Total Protein) อัลบูมิน (Albumin) โกลบูลิน (Globulin) ค่าบิลิรูบินทั้งหมด (Total Bilirubin), ค่าบิลิรูบินชนิดละลายน้ำ

(Direct Bilirubin) เอสจีพีที (SGPT) เอสจีโอที (SGOT) และแอลคาไลน์ฟอสฟาเทส (Alkaline Phosphate) เป็นต้น แต่ในทางการเรียนรู้ของเรื่องจักรการถดถอยโลจิสติคแบบทวินั้นจะแบ่งเป็น 0 และ 1 หรือก็คือใช่หรือไม่ใช่ (GeeksforGeeks, 2024) และเมื่อตัวแปรตามมีค่ามากกว่า 2 ค่า หรือ หน่วยข้อมูลสามารถแบ่งออกเป็นกลุ่มย่อยได้มากกว่า 2 กลุ่ม กรณีนี้จะเรียกว่า การวิเคราะห์การถดถอยโลจิสติคอเนกนาม (Multinomial Logistic Regression Analysis) เช่น การศึกษาซุปเปอร์มาร์เก็ตที่ลูกค้าเลือกซื้อ มีตัวแปรตามคือ ท็อปส์ เทสโก้โลตัส บิ๊กซี ฟู้ดแลนด์ และตัวแปรอิสระ คือ ระดับรายได้ การขับรถหรือไม่ ระดับการศึกษา เพศ อายุ เป็นต้น ซึ่งในทางการเรียนรู้ของเครื่องจักรก็ใช้แบบนี้เหมือนกัน

ในกรณีตัวแปรตาม (y) มี 2 ค่า จะเรียกว่า ตัวแบบการถดถอยลอจิสติกทวิภาค (Binary Logistic Model) ซึ่งมีตัวแบบดังนี้

$$\log\left(\frac{p_y}{Q_y}\right) = \log\left(\frac{p_y}{1-P_y}\right) = b_0+b_1x_1+\ldots+b_px_p$$

เมื่อ P_y คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (y=1)

Q_y คือ ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ (y=0)

b_i คือ ตัวประมาณของ β_i ด้วยวิธีภาวะน่าจะเป็นสูงสุด โดยที่ i=0, 1,....,p

ดังนั้น $Q_y = 1 - P_y$

จากความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระของการวิเคราะห์การถดถอยลอจิสติกไม่อยู่ในรูปแบบเชิงเส้น ทำให้ต้องมีการปรับให้อยู่ในรูปแบบของออดส์ (Odds) หรืออัตราส่วนออดส์ (Odd Ratio) นั่นคือเป็นการปรับให้ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระอยู่ในรูปแบบเชิงเส้น ซึ่งออดส์หรืออัตราส่วนออดส์นั้นหมายถึงอัตราส่วนระหว่างโอกาสที่เหตุการณ์จะเกิดเหตุการณ์ที่สนใจ P(Y=1) กับโอกาสที่เหตุการณ์จะไม่เกิดเหตุการณ์ที่สนใจ P(Y=0) เขียนได้ดังนี้

$$\text{Odds} = \frac{P_y}{Q_y}$$

ดังนั้นอัตราส่วนออดส์จึงเป็นค่าที่แสดงถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจว่าเป็นกี่เท่าของโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ ซึ่งถ้าค่าอัตราส่วนออดส์มีค่ามากกว่า 1 แสดงว่าโอกาสที่จะเกิดเหตุการณ์ที่สนใจมีมากกว่า

โอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ ถ้าค่าอัตราส่วนออดส์มีค่าเท่ากับ 1 แสดงว่าโอกาสที่จะเกิดเหตุการณ์ที่สนใจและไม่เกิดเหตุการณ์ที่สนใจมีเท่ากัน และถ้าอัตราส่วนออดส์มีค่าน้อยกว่า 1 แสดงว่าโอกาสที่จะเกิดเหตุการณ์ที่สนใจต่ำกว่าโอกาสที่ไม่เกิดเหตุการณ์ที่สนใจ

แต่ในกรณีตัวแปรตามมีค่ามากกว่า 2 ค่า เช่น $k > 2$ ยุทธ (2555) ได้กล่าวไว้ว่าจะได้โลจิสจำนวนเท่ากับ (k-1) และจะนำโลจิส แต่ละค่ามาเปรียบเทียบกับกลุ่มฐาน (Baseline Category) ซึ่งค่าสัมประสิทธิ์ทั้งหลายของตัวแปรที่เป็นกลุ่มฐานจะเท่ากับ 0 เพื่อเป็นฐานในการเปรียบเทียบให้กับค่าของกลุ่มอื่น หากกรณีที่มีตัวแปรตามจำนวน 3 หรือ 4 ตัว โดยที่กลุ่มที่เป็นฐานคือ k และเมื่อเปรียบเทียบกับกลุ่มที่ i จะได้ตัวแบบโลจิสดังนี้

$$\log\left(\frac{P(\text{กลุ่ม } i)}{P(\text{กลุ่ม } k)}\right) = b_{i0} + b_{i1}x_1 + \ldots + b_{ip}x_p$$

สัมประสิทธิ์ $b_{i0}, b_{i1}, b_{i2}, \ldots, b_{ip}$ ของกลุ่มที่ i และกลุ่มที่เป็นฐานจะมีค่าเป็น $b_{k0} = b_{k1} = \ldots = b_{kp} = 0$ และ การวิเคราะห์จะได้ผลตามนี้ คือ ถ้าตัวแปรตาม (y) มี 3 ค่า หรือ $k = 3$ จะได้สัมประสิทธิ์ 2 ชุด มาจาก k - 1 แต่กลุ่มที่เป็นฐานจะมี 3 ค่าตามตัวแปรตามใน 2 ชุดนั้น ซึ่งในชุดที่ 1 แสดงค่าสัมประสิทธิ์ของ $y = 1$ เปรียบเทียบกับ $y = 3$ และชุดที่ 2 จะแสดงค่าสัมประสิทธิ์ของ $y = 2$ เปรียบเทียบกับ $y = 3$

ตัวอย่างการเปรียบเทียบกลุ่มของผู้วิจัยที่สนใจศึกษานักเรียนมัธยมศึกษาปีที่ 3 จะเลือกเข้ามัธยมศึกษาปีที่ 4 โรงเรียนใดใน 3 โรงเรียน ได้แก่ Aschool Bschool และ Cschool ในการเลือกนี้ผู้วิจัยสงสัยเพศของนักเรียนจะมีผลต่อการเลือกโรงเรียนหรือไม่ โดยให้เพศหญิงเป็นกลุ่มที่เป็นฐาน ค่าสัมประสิทธิ์ของเพศหญิงจะเท่ากับ 0 ผู้วิจัยจะได้ตัวแบบลอจิต 2 ค่าที่ไม่ซ้ำซ้อนกัน ดังนี้

$$g_1 = \left(\frac{P(\text{Aschool})}{P(\text{Cschool})}\right) = b_{10} + b_{11}(\text{ชาย})$$

เลขที่กำกับของตัวแปร

$$g_2 = \left(\frac{P(\text{Bschool})}{P(\text{Cschool})}\right) = b_{20} + b_{21}(\text{ชาย})$$

เลขที่กำกับของ logit

วิธีการดำเนินการ

อุปกรณ์ที่ใช้

1. คอมพิวเตอร์พกพา

2. โปรแกรมที่ใช้ในการวิเคราะห์ผลได้แก่ Python

วิธีการ

1. จัดเตรียมข้อมูล Bone tumor จากเว็บไซส์ Memorial Sloan Kettering Cancer Center (MSKCC) เป็นชุดข้อมูลที่มีการเปิดเผยจำนวน 500 เรคคอร์ด เพื่อการวิเคราะห์

2. ทำการ Exploratory Data Analysis (EDA) ประกอบไปด้วย

1.1 ตัวแปรตาม (Y) 1 แปร

1.2 ตัวแปรอิสระ (X_i) ทั้งหมด 7 ตัวแปร ดังนี้

ตารางที่ 1 แสดงลักษณะและรายละเอียดของตัวแปร

Variable	Label	Measurement scale	Exploratory
Y	Status (NED, AWD, D)	Nominal	0 – NED (no evidence of disease) 1 – AWD (alive with disease) 2 – D (dead)
X_1	Sex	Nominal	0 – Male 1 – Female
X_2	Age_Group	Ordinal	0 - Age <= 41 1 - Age 42 to 50 2 - Age 51 to 63 3 - Age 64 to 70 4 - Age >= 71

X_3	Grade	Ordinal	0 – Intermediate 1 – High
X_4	Histological type	Nominal	0 - Pleiomorphic leiomyosarcoma 1 - Synovial sarcoma 2 - Leiomyosarcoma 3 - Myxofibrosarcoma 4 - Undifferentiated - pleiomorphic 5 - Sclerosing epithelioid fibrosarcoma 6 - Pleiomorphic spindle cell undifferentiated 7 - Malignant solitary fibrous tumor 8 - Undifferentiated pleomorphic liposarcoma 9 - Myxoid fibrosarcoma 10 - Poorly differentiated synovial sarcoma 11 - Pleomorphic sarcoma 12 - Epithelioid sarcoma
X_5	MSKCC type	Nominal	0 – MFH 1 - Leiomyosarcoma 2 - Synovial sarcoma

X_6	**Site of primary STS**	**Nominal**	**0 - Left thigh** **1 - Right thigh** **2 - Right buttock** **3 - Parascapusular** **4 - Left biceps** **5 - Left buttock** **6 - Right parascapusular**
X_7	Treatment	Nominal	0 - Radiotherapy + Surgery 1 - Radiotherapy + Surgery + Chemotherapy 2 - Surgery + Chemotherapy

3. ตรวจสอบความสัมพันธ์ของตัวแปรตาม (Y) และ ตัวแปรอิสระ (X_i) โดยทำการทดสอบไคสแควร์ (Chi-square) และ พิจารณาเมทริกซ์สหสัมพันธ์

4. วิเคราะห์ปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกด้วยวิธี Random Forest และ Logistic Regression โดยพิจารณาจากเกณฑ์ Feature Importance

ผลการวิจัย

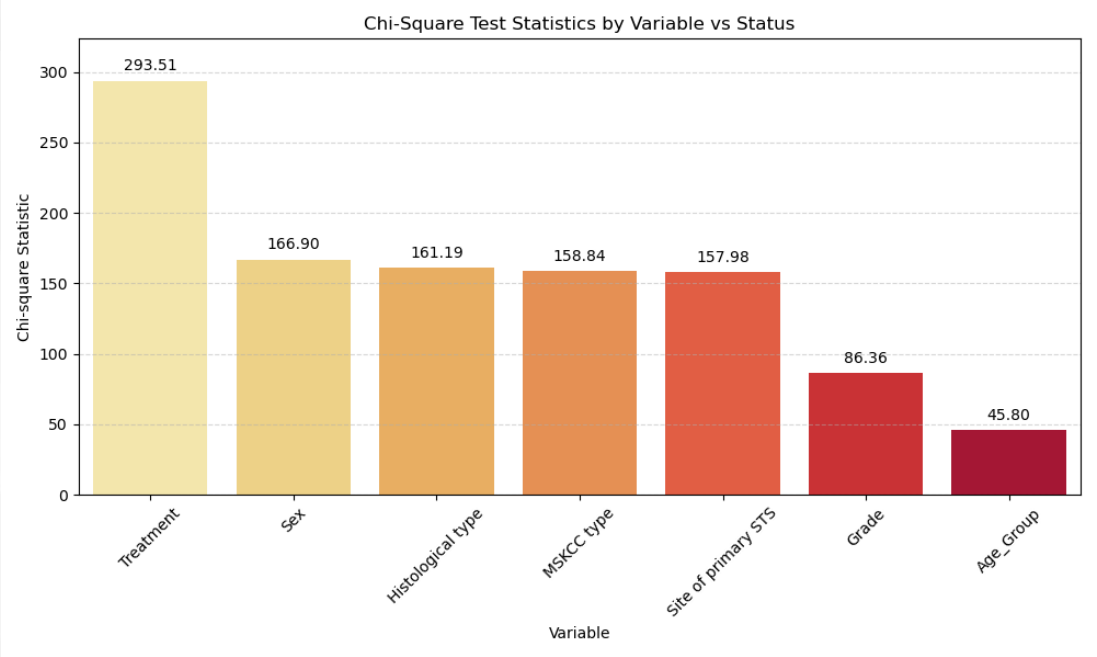
จากตารางที่ 2 ผลการทดสอบ Chi-square พบว่าทุกปัจจัยมีผลต่อผลการรักษาของผู้ป่วยได้แก่ เพศ (Sex), อายุ (Age_Group), ระดับความรุนแรงของเนื้องอก (Grade), ประเภทเนื้องอกกระดูกทางพยาธิสภาพ (Histological type), การแบ่งเนื้องอกกระดูกโดย MSKCC (MSKCC type), ตำแหน่งของเนื้องอกกระดูก (Site of primary STS) และวิธีการรักษา (Treatment) นอกจากนี้จากตารางที่ 3 ผลการรักษาของผู้ป่วยที่หายขาดส่วนใหญ่จะใช้วิธีการรักษา Radiotherapy และ Surgery คิดเป็นร้อยละ 87.04

ตารางที่ 2 ผลการทดสอบความสัมพันธ์ระหว่างตัวแปรอิสระกับผลการรักษาของผู้ป่วยด้วยสถิติทดสอบ Chi-square

Variable	**Chi-square**	**P-value < 0.05**	**ผลการทดสอบ**
Sex	166.90	5.72×10^{-37}	Sex และ Status มีความสัมพันธ์กัน
Age_Group	45.80	2.59×10^{-7}	Age_Group และ Status มีความสัมพันธ์กัน
Grade	86.36	1.77×10^{-19}	Grade และ Status มีความสัมพันธ์กัน
Histological type	161.19	2.68×10^{-22}	Histological type และ Status มีความสัมพันธ์กัน
MSKCC type	158.84	2.59×10^{-33}	MSKCC type และ Status มีความสัมพันธ์กัน
Site of primary STS	157.98	1.36×10^{-27}	Site of primary STS และ Status มีความสัมพันธ์กัน
Treatment	293.51	2.72×10^{-62}	Treatment และ Status มีความสัมพันธ์กัน

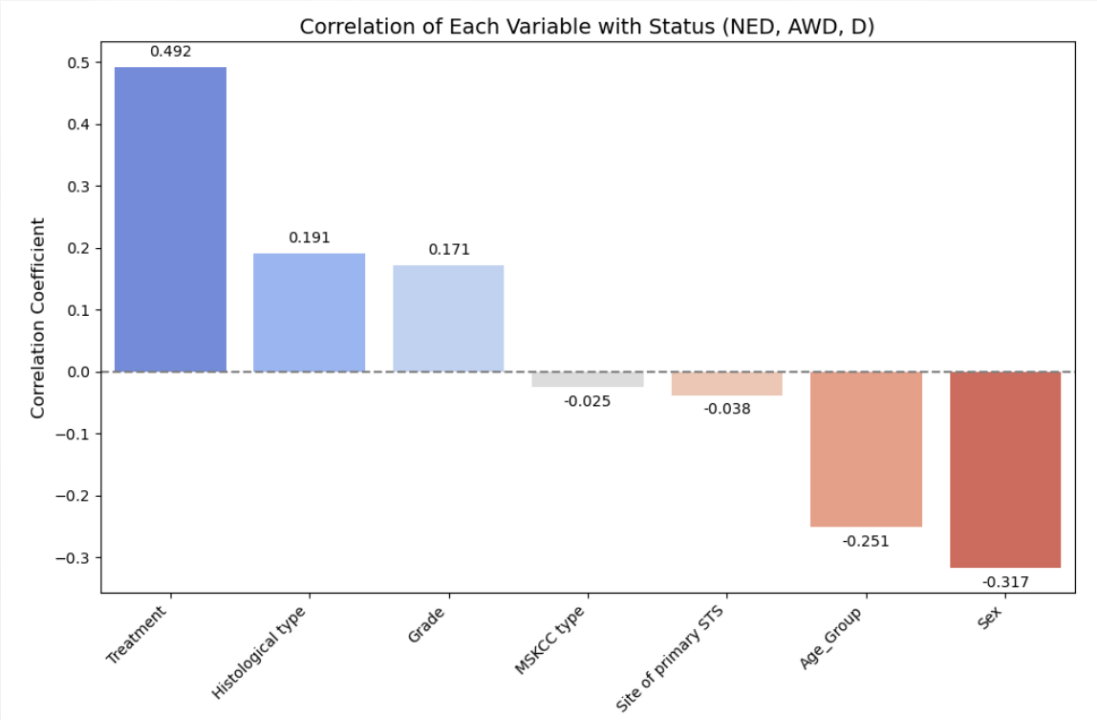
ตารางที่ 3 ร้อยละของวิธีการรักษากับผลการรักษาของผู้ป่วย

Treatment	Status : AWD	Status : D	Status : NED
Radiotherapy + Surgery	16.07	**42.55**	**87.04**
Radiotherapy + Surgery + Chemotherapy	**79.46**	21.28	9.72
Surgery + Chemotherapy	4.47	36.17	3.24



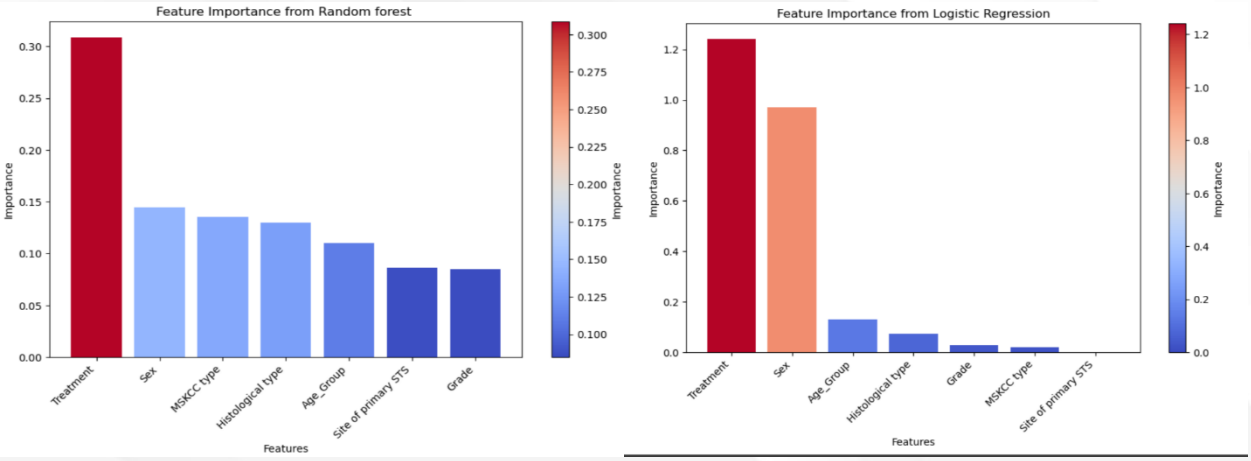
รูปที่ 2 ค่าสถิติทดสอบ Chi-square ของแต่ละตัวแปรอิสระเทียบกับผลการรักษา

ผลการพิจารณาสหสัมพันธ์จากรูปที่ 3 พบว่าปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วยมากที่สุดคือ Treatment รองมาเป็น Sex, Age_Group, Histological type, Grade, Site of primary STS และ MSKCC type ตามลำดับ



รูปที่ 3 ค่าสหสัมพันธ์ของตัวแปรอิสระเทียบกับผลการรักษา

ผลจากการวิเคราะห์ค่า Feature Importance ของวิธี Random Forest และ Logistic Regression พบว่าปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วยมากที่สุดคือ วิธีการรักษา(Treatment)



รูปที่ 4 Feature Importance ของ Random Forest และ Logistic Regression ตามลำดับ

วิจารณ์

จากการศึกษาครั้งนี้ใช้วิธีการทางสถิติและวิธีการเรียนรู้ของเครื่องจักรวิเคราะห์ข้อมูลพบว่าปัจจัยที่มีความสำคัญต่อประสิทธิภาพการรักษาให้ผลลัพธ์สอดคล้องกันทั้ง 3 วิธี กล่าวคือปัจจัยที่ส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วยมากที่สุดคือ วิธีการรักษา โดยผลการรักษาของผู้ป่วยที่หายขาดส่วนใหญ่จะใช้วิธีการรักษา Radiotherapy และ Surgery คิดเป็นร้อยละ 87.04 ซึ่งสอดคล้องกับงานวิจัยของ Vangala et al. (2021)

สรุปผลการวิจัย

ผลการทดสอบสมมติฐานโดยใช้สถิติทดสอบไคสแควร์พบว่าทั้ง 7 ปัจจัยได้แก่ เพศ, อายุ, ระดับความรุนแรงของเนื้องอก, ประเภทเนื้องอกกระดูกทางพยาธิสภาพ,การแบ่งเนื้องอกกระดูกโดย MSKCC, ตำแหน่งของเนื้องอกกระดูก และ วิธีการรักษามีผลต่อผลการรักษาของผู้ป่วย และจากเมทริกซ์สหสัมพันธ์พบว่า วิธีการรักษามีผลต่อผลการรักษาของผู้ป่วยมากที่สุดคิดเป็นร้อยละ 49.2 รองลงมาเป็น เพศร้อยละ 31.7 และอายุคิดเป็นร้อยละ 25.1 ซึ่งสอดคล้องกับ Feature Importance ของการวิเคราะห์ด้วยวิธี Logistic Regression ส่วน Feature Importance ของการวิเคราะห์ด้วยวิธี Random Forest มีเพียง 2 ปัจจัยได้แก่ วิธีการรักษาและเพศที่สอดคล้องกันทำให้สรุปได้ว่า วิธีการรักษาส่งผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วยมากที่สุด

ข้อเสนอแนะ

การรักษาด้วยวิธี Radiotherapy และ Surgery ส่วนใหญ่ผลการรักษาจะหายขาด นอกจากนี้ในการหาปัจจัยมีผลต่อประสิทธิภาพการรักษาโรคเนื้องอกกระดูกของผู้ป่วย พบว่ายังมีวิธีอื่นที่น่าสนใจในการหาความสัมพันธ์ของการรักษา การวิเคราะห์การจำแนกประเภท (Discriminant Analysis) เหมาะกับข้อมูลที่มีจำนวนไม่มากและเป็นวิธีการสถิติอีกหนึ่งวิธีที่นิยมใช้กันและมีประสิทธิภาพดี

เอกสารอ้างอิง

Alqahtani, S. A. M., Alobaid, H. M., Alshammari, J., Alqarzae, S. A., Aloyouni, S. Y., Al-Eidan, A. A., Alhamad, S., Almiman, A., Alkhulaifi, F. M., & Alomar, S. (2024). Feature importance and model performance for prediabetes prediction: A comparative study. *Journal of King Saud University - Science*, 36(11). https://doi.org/10.1016/j.jksus.2024.103583

GeeksforGeeks. (2024). *Logistic Regression in Machine Learning*. https://www.geeksforgeeks.org/understanding-logistic-regression/

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*

(r. Edition, Ed.). O'Reilly Media.

Jin, Y., Lan, A., Dai, Y., Jiang, L., & Liu, S. (2023). Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *Eur J Med Res*, 28(1), 394. https://doi.org/10.1186/s40001-023-01361-7

Li, H., Liu, R. B., Long, C. M., Teng, Y., Cheng, L., & Liu, Y. (2022). Development and Validation of a New Multiparametric Random Survival Forest Predictive Model for Breast Cancer Recurrence with a Potential Benefit to Individual Outcomes. *Cancer Manag Res*, 14, 909-923. https://doi.org/10.2147/CMAR.S346871

Logunova, I. (2022). *Random Forest Classifier: Basic Principles and Applications*. Retrieved june 23 from https://serokell.io/blog/random-forest-classification

Lucas, D. R., Kshirsagar, M. P., Biermann, J. S., Hamre, M. R., Thomas, D. G., Schuetze, S. M., & Baker, L. H. (2008). Histologic alterations from neoadjuvant chemotherapy in high-grade extremity soft tissue sarcoma: clinicopathological correlation. *Oncologist*, 13(4), 451-458. https://doi.org/10.1634/theoncologist.2007-0220

Sbaraglia, M., Bellan, E., & Dei Tos, A. P. (2021). The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives. *Pathologica*, 113(2), 70-84. https://doi.org/10.32074/1591-951X-213

Turney, S. (2023). *Chi-Square Test of Independence | Formula, Guide & Examples*. Retrieved 2022 from https://www.scribbr.com/statistics/chi-square-test-of-independence/

Vangala, N., Uppin, S. G., Rao, K. N., Chandrasekhar, P., & Gundeti, S. (2021). Prognostic Significance of Clinical and Post-Neoadjuvant Chemotherapy Associated Histomorphological Parameters in Osteosarcoma: A Retrospective Study from a Tertiary

Care Center. *Indian Journal of Medical and Paediatric Oncology*, 42(06), 547-553. https://doi.org/10.1055/s-0041-1740113

Wagavkar, S. (2024). *Introduction to the Correlation Matrix*. Retrieved Nov 20 from https://builtin.com/data-science/correlation-matrix

ยุทธ ไกยวรรณ์. (2555). หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย. วารสารวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย, 4(1), 1-12.