บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหา

โรคมะเร็งถือเป็นหนึ่งในปัญหาทางสาธารณสุขระดับโลกที่ส่งผลกระทบอย่างลึกซึ้งต่อคุณภาพชีวิตของประชากร ระบบเศรษฐกิจ และภาระของระบบบริการสาธารณสุขทั่วโลก ข้อมูลจากองค์การอนามัยโลก (WHO) รายงานว่าในปี พ.ศ. 2565 มีผู้เสียชีวิตจากโรคมะเร็งมากกว่า 10 ล้านคนทั่วโลก โดยโรคมะเร็งกลายเป็นสาเหตุการเสียชีวิตอันดับต้น ๆ รองจากโรคหัวใจ โดยเฉพาะอย่างยิ่งในประเทศกำลังพัฒนาและประเทศที่มีรายได้น้อยและปานกลาง ซึ่งมักประสบปัญหาการเข้าถึงการวินิจฉัยและการรักษาที่มีประสิทธิภาพและทันท่วงที ปัจจัยที่มีผลต่อการอยู่รอดของผู้ป่วยโรคมะเร็งนั้นมีความหลากหลายและซับซ้อน ไม่ได้ขึ้นอยู่กับชนิดของมะเร็งเพียงอย่างเดียว แต่ยังรวมถึงปัจจัยทางประชากรศาสตร์ เช่น อายุ เพศ เชื้อชาติ พฤติกรรมสุขภาพ เช่น การสูบบุหรี่ การดื่มแอลกอฮอล์ โรคประจำตัว รวมถึงปัจจัยด้านสิ่งแวดล้อม ความเสี่ยงทางพันธุกรรม และความพร้อมของระบบสาธารณสุขในแต่ละพื้นที่ ความหลากหลายของตัวแปรเหล่านี้ทำให้การพยากรณ์โอกาสในการรอดชีวิตของผู้ป่วยมะเร็งกลายเป็นเรื่องที่มีความซับซ้อนและท้าทายมากยิ่งขึ้น

ในยุคดิจิทัลที่ข้อมูลด้านสุขภาพมีปริมาณเพิ่มขึ้นอย่างต่อเนื่อง การวิเคราะห์ข้อมูลด้วยเทคนิค Machine Learning จึงกลายเป็นเครื่องมือที่ทรงพลังในการจัดการข้อมูลขนาดใหญ่ และสามารถช่วยค้นหารูปแบบความสัมพันธ์เชิงลึกที่อาจไม่สามารถมองเห็นได้ด้วยการวิเคราะห์ทางสถิติแบบดั้งเดิม Machine Learning ยังสามารถจัดการกับข้อมูลที่ไม่เป็นเชิงเส้น มีมิติมาก (high-dimensional) หรือมีข้อมูลที่ขาดบางส่วนได้ดี จึงเหมาะอย่างยิ่งสำหรับการนำมาใช้พยากรณ์ระยะเวลาการอยู่รอดของผู้ป่วยโรคมะเร็ง โดยเฉพาะชุดข้อมูล “Global Cancer Patients 2015–2024” ที่เผยแพร่บนเว็บไซต์ Kaggle ซึ่งเป็นแหล่งรวบรวมข้อมูลขนาดใหญ่จากผู้ป่วยมะเร็งทั่วโลก โดยมีรายละเอียดข้อมูลที่หลากหลาย อาทิ อายุ เพศ เชื้อชาติ ระดับความเสี่ยงทางพันธุกรรม สภาพแวดล้อม ชนิดของมะเร็ง ระยะของโรค วิธีการรักษา ไปจนถึงจำนวนปีที่ผู้ป่วยสามารถมีชีวิตอยู่หลังได้รับการวินิจฉัย ข้อมูลเหล่านี้จึงเหมาะสมอย่างยิ่งสำหรับการ

นำมาพัฒนาโมเดลเพื่อพยากรณ์ระยะเวลาการอยู่รอดของผู้ป่วยด้วยเทคนิคทาง Machine Learning

ด้วยเหตุนี้ ผู้ศึกษาตระหนักถึงความสำคัญของการพยากรณ์ระยะเวลาการอยู่รอดของผู้ป่วยโรคมะเร็ง ซึ่งถือเป็นข้อมูลที่มีคุณค่าอย่างยิ่งในการวางแผนการรักษา การติดตามผล และการจัดสรรทรัพยากรทางการแพทย์ได้อย่างมีประสิทธิภาพ โดยเฉพาะในยุคที่ข้อมูลสุขภาพขนาดใหญ่ (Big Data) สามารถนำมาใช้เพื่อเสริมสร้างการตัดสินใจเชิงคลินิกได้อย่างแม่นยำยิ่งขึ้น

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาตัวแบบที่เหมาะสมสำหรับการระยะเวลาอยู่รอดผู้ป่วยโรคมะเร็ง

1.2.2 เพื่อพยากรณ์ระยะเวลาอยู่รอดของผู้ป่วย

1.2.3 เพื่อสร้างระบบเว็บไซต์ที่สามารถให้ผู้ใช้งานทั่วไปทดลองกรอกข้อมูลและดูผลการพยากรณ์ได้ พร้อมทั้งนำเสนอข้อมูลภาพรวมผ่าน Dashboard เพื่อการเรียนรู้

1.3 ประโยชน์ที่จะได้รับ

1.3.1 ได้รับแบบจำลองที่มีเหมาะสมสำหรับการพยากรณ์ระยะเวลาการอยู่รอด

1.3.2 ได้รับผลพยากรณ์ระยะเวลาอยู่รอดของผู้ป่วย

1.3.3 ได้เผยแพร่สารสนเทศผ่านเว็บไซต์

1.4 ขอบเขต และเครื่องมือที่ใช้พัฒนาระบบ

1.4.1 ขอบเขตวิเคราะห์ข้อมูล

1.4.1.1 วิเคราะห์ข้อมูลรวบรวมข้อมูลผู้ป่วยโรคมะเร็งที่ได้จากเว็บไซต์ Kaggle เพื่อนำมาวิเคราะห์ ข้อมูลที่ได้นำมาวิเคราะห์ ได้แก่ Cancer Type Cancer Stage Diagnosis Date

1.4.1.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) โดยข้อมูลบางส่วนอาจมีความผิดพลาดหรือขาดหาย จึงจำเป็นต้องดำเนินการทำความสะอาด เช่น การลบข้อมูลที่ขาด, การเติมค่าที่หายไป, การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม และการตัดข้อมูลที่ไม่จำเป็นออก เพื่อให้ข้อมูลมีความถูกต้องและสมบูรณ์สำหรับการวิเคราะห์

1.4.1.3 ดำเนินการวิเคราะห์ข้อมูลเพื่อพยากรณ์ ระยะเวลาการอยู่รอดของผู้ป่วยโรคมะเร็ง โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) และอัลกอริธึม Machine Learning ผ่านโปรแกรม RapidMiner ซึ่งมีแนวทางการวิเคราะห์ ดังนี้

1) ใช้การวิเคราะห์หาปัจจัยที่เหมาะสมสำหรับการพยากรณ์ เช่น ปริมาณการสูบ ความเสี่ยงมลพิษทางอากาศ ค่าความอ้วนปริมาณการดื่ม

2) ใช้เทคนิค Machine Learning ในการพยากรณ์ ได้แก่ Random Forest , Gradient Boosted Trees, Neural Networks เพื่อเปรียบเทียบประสิทธิภาพของแต่ละโมเดลโดยใช้ค่า MAE และ RMSE เป็นเกณฑ์การประเมินประสิทธิภาพของโมเดล

1.4.1.4 ดำเนินการนำโมเดล Random Forest ซึ่งเป็นโมเดลที่ดีที่สุดนำมาใส่ในเว็บไซต์โดยใช้ Python สร้างตัวโมเดลมาพยากรณ์ภายในเว็บไซต์

1.4.2 ขอบเขตผู้ใช้ทั่วไป

1.4.2.1 สามารถกรอกข้อมูลเบื้องต้น เช่น อายุ เพศ ประเทศ ประเภทของมะเร็ง และพฤติกรรมสุขภาพ เพื่อให้ระบบทำการประเมินและพยากรณ์ระยะเวลาการอยู่รอดโดยอัตโนมัติ

1.4.2.2 สามารถดูผลลัพธ์ผ่าน Dashboard ที่แสดงข้อมูลในรูปแบบกราฟ เช่น กราฟอัตราการอยู่รอดตามกลุ่มอายุ เพศ หรือประเทศ

1.4.2.3 สามารถ ดาวน์โหลดข้อมูลเพื่อใช้ในการศึกษาหรือทำวิจัยต่อยอด

1.4.3 ขอบเขตผู้ดูแลระบบ

1.4.3.1 สามารถเข้าสู่ระบบโดยใช้ชื่อผู้ใช้และรหัสผ่าน

1.4.3.2 สามารถ เพิ่ม แก้ไข หรือลบข้อมูล ที่แสดงบนเว็บไซต์ได้

1.4.3.3 สามารถอัปโหลดข้อมูลขึ้นไปบนเว็บไซต์ได้

1.4.4 ชุดข้อมูล

	Patient_ID	Age	Gender	Country_R	Year	Genetic_Risk	Air_Pollution	Alcohol_Use	Smoking	Obesity_Level	Cancer_Type	Cancer_Stage	Treatment_Cost_USD	Survival_Years	Target_Severity_Score
2	PT0000000	71	Male	UK	2021	6.4	2.8	9.5	0.9	8.7	Lung	Stage III	62913.44	5.9	4.92
3	PT0000001	34	Male	China	2021	1.3	4.5	3.7	3.9	6.3	Leukemia	Stage 0	12573.41	4.7	4.65
4	PT0000002	80	Male	Pakistan	2023	7.4	7.9	2.4	4.7	0.1	Breast	Stage II	6984.33	7.1	5.84
5	PT0000003	40	Male	UK	2015	1.7	2.9	4.8	3.5	2.7	Colon	Stage I	67446.25	1.6	3.12
6	PT0000004	43	Female	Brazil	2017	5.1	2.8	2.3	6.7	0.5	Skin	Stage III	77977.12	2.9	3.62
7	PT0000005	22	Male	Germany	2018	9.5	6.4	3.3	3.9	5.1	Cervical	Stage IV	33468.99	9.5	5.98
8	PT0000006	41	Male	Canada	2021	5.1	8.2	0.3	3.7	2.1	Cervical	Stage 0	9790.83	1	5.05
9	PT0000007	72	Female	Canada	2018	6	8.2	6.4	0.6	8.5	Prostate	Stage I	17161.4	6.2	6.02
10	PT0000008	21	Male	USA	2022	4.3	3.8	1	0.3	8.5	Lung	Stage II	56458.48	6.5	3.36
11	PT0000009	49	Female	Canada	2016	8.1	0.8	7.8	5.2	9.3	Prostate	Stage II	56133.45	5.7	5.76
12	PT0000010	57	Other	Brazil	2022	1.9	1.9	4.6	4	0.2	Skin	Stage I	15093.39	1	3.87
13	PT0000011	21	Female	Brazil	2021	5.2	1.7	7.2	3.1	8.3	Prostate	Stage I	72315.19	6	4.38
14	PT0000012	83	Male	Canada	2016	3.5	1.5	8.1	5	1.5	Leukemia	Stage II	99120.52	8	3.31
15	PT0000013	79	Female	USA	2021	8.5	9.6	3.6	9.8	8.7	Cervical	Stage II	94210.93	7.1	6.63
16	PT0000014	40	Male	UK	2023	4.6	3.6	3.5	6.2	3.4	Breast	Stage IV	58397.96	8.3	4.4
17	PT0000015	52	Male	Germany	2024	2.3	5.8	6.3	5.6	1.9	Lung	Stage II	19910.36	7	5.19
18	PT0000016	77	Other	UK	2017	8.9	4.3	1.9	8.2	3.7	Colon	Stage III	59285.13	0.5	5.53
19	PT0000017	41	Male	Germany	2016	5.4	9.1	9.2	4	5.1	Liver	Stage 0	56875.63	1.9	6
20	PT0000018	68	Male	UK	2021	8.4	7.4	7.8	7	7.2	Leukemia	Stage II	10360.2	4.6	7.87

ภาพที่ 1.1 แสดงข้อมูลของผู้ป่วยที่เป็นโรคมะเร็ง

จาก Dataset ข้อมูลผู้ป่วยโรคมะเร็งทั่วโลกก่อนทำการกลั่นกรองข้อมูลจำนวนข้อมูล 50,000 แถว จำนวนคอลัมน์ 53 คอลัมน์ สามารถอธิบายชุดข้อมูลได้ ดังนี้

ตารางที่ 1.1 แสดงข้อมูลความหมายแต่ละคอลลัมน์ของข้อมูลผู้ป่วยโรคมะเร็ง

คอมลัมน์	ความหมาย
Gender	เพศ
Country_Region	ประเทศที่อยู่
Year	ปีที่ตรวจพบโรคมะเร็ง
Genetic_Risk	ความเสี่ยงทางพันธุกรรม
Air_Pollution	มลพิษทางอากาศ
Alcohol_use	การดื่มแอลกอฮอล์
Smoking	การสูบบุหรี่
Obesity_Level	ระดับความอ้วน
Cancer_Type	ประเภทของมะเร็ง
Treatment_Cost_USD	ค่าใช้จ่ายในการรักษา
Survival_Years	ปีที่อยู่รอด
Target_Serverity_Score	คะแนนความรุนแรงของโรค

หลังจากทำการกลั่นกรองข้อมูลแล้วเหลือจำนวนข้อมูล 9,062 แถว จำนวนคอลัมน์ คอลัมน์ โดยคอลัมน์ที่ทำการกลั่นกรองข้อมูลที่มีว่าง และข้อมูลที่ไม่สมบูรณ์ออก

1.5 ขอบเขต และเครื่องมือที่ใช้พัฒนาระบบ

1.5.1 Hardware

1.5.1.1 โน๊ตบุ๊ค Asus CPU AMD Ryzen 5 5600h 4.2 GHZ Ram 8 GB SSD 512 GB

1.5.2 Software

1.5.2.1 โปรแกรม Rapid Miner ใช้ในการสร้างโมเดล

1.5.2.2 โปรแกรม Visual Studio Code ใช้ในการสร้างเว็บไซต์

1.5.2.3 โปรแกรม Microsoft Word ใช้ในการจัดทำเอกสาร

1.5.2.4 โปรแกรม Microsoft Excel ใช้ในการจัดการชุดข้อมูล

1.5.2.5 ระบบปฏิบัติการ Windows 10 ใช้ในการเปิดซอฟต์แวร์

1.6 สถานที่ใช้ในการดำเนินการศึกษา และรวบรวมข้อมูล

1.6.1 มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนาเชียงใหม่ 128 ถนนห้วยแก้ว ตำบลช้างเผือก อำเภอเมือง จังหวัดเชียงใหม่ 50300

1.7 ระยะเวลาในการดำเนินการ

ตารางที่ 1.2 ตารางแสดงระยะเวลาดำเนินงาน

แผนการดำเนินการ	2568							2569		
	มิ.ย.	ก.ค.	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.
1. ศึกษาและกำหนดความต้องการ	←	→								
2. วิเคราะห์ออกแบบระบบและสร้างฐานข้อมูล		←	→							
3. เขียนและทดสอบโปรแกรม			←	→	→					
4. ติดตั้ง ทดสอบ และปรับปรุงระบบ					←	→				
5. ตรวจสอบระบบโดยรวม							↔			
6. ประเมินการใช้งานระบบ								↔		
7. จัดทำคู่มือการใช้งาน			←	→	→	→	→	→	→	→
8. จัดทำเอกสารประกอบโครงงาน	←	→	→	→	→	→	→	→	→	→

1.8 นิยามศัพท์เฉพาะ

1.8.1 Cancer (มะเร็ง) หมายถึง โรคที่เกิดจากความผิดปกติของการแบ่งตัวของเซลล์ในร่างกาย ซึ่งทำให้เซลล์เติบโตอย่างไม่สามารถควบคุมได้ และสามารถแพร่กระจาย (metastasize) ไปยังส่วนอื่น ๆ ของร่างกายได้ มะเร็งมีหลายชนิด เช่น มะเร็งปอด มะเร็งตับ มะเร็งเต้านม ซึ่งแต่ละชนิดมีปัจจัยเสี่ยง ระยะของโรค และอัตราการรอดชีวิตที่แตกต่างกัน

1.8.2 Neural Networks(NN) หมายถึง แบบจำลองทางคณิตศาสตร์ที่เลียนแบบการทำงานของสมองมนุษย์ ใช้ในการเรียนรู้รูปแบบซับซ้อนของข้อมูล โดยประกอบด้วยหลายชั้น (Layers) และโหนด (Nodes)

1.8.3 Random Forest (RF) คือแบบจำลองการเรียนรู้ของเครื่องจักรที่ใช้ต้นไม้ตัดสินใจจำนวนมากในการดำเนินการวิเคราะห์การถดถอยและการจำแนกประเภท โดยให้การเลือกคุณลักษณะโดยปริยายและตัวบ่งชี้ความสำคัญของคุณลักษณะ

1.8.4 Gradient-boosted trees (GBT)คือ เทคนิคการเรียนรู้ของเครื่องจักรที่ใช้ในการเพิ่มประสิทธิภาพค่าการพยากรณ์ ของแบบจำลองผ่านขั้นตอนต่างๆ ในกระบวนการเรียนรู้ ในแต่ละรอบของการสร้างต้นไม้ตัดสินใจ จะมีการปรับค่าสัมประสิทธิ์ น้ำหนัก หรือไบแอสที่ใช้กับตัวแปรอินพุตแต่ละตัวที่ใช้ในการพยากรณ์

1.8.5 Mean Absolute Error (MAE) คือ ผลรวมสัมบูรณ์ของค่าความคลาดเคลื่อนทั้งหมดที่ได้จากความแตกต่างระหว่างค่าที่ประมาณการและค่าที่วัดได้ หารด้วยจำนวนการสังเกต MAE เป็นตัวบ่งชี้ในการประเมินว่าการประมาณการใกล้เคียงกับค่าที่วัดได้มากน้อยเพียงใด

1.8.6 RMSE (Root Mean Square Error) คือ มาตรวัดประสิทธิภาพของโมเดลพยากรณ์หรือ Machine Learning (Regression) ที่บอกว่าค่าที่พยากรณ์ คลาดเคลื่อนจากค่าจริงเฉลี่ยเท่าใด โดยนำ MSE มาถอดรากที่สองเพื่อให้หน่วยเหมือนข้อมูลต้นฉบับ ยิ่งค่าต่ำแสดงว่าแม่นยำสูง