รายงาน

การสืบค้นงานวิจัยการวิเคราะห์ประสิทธิภาพของแนวทางการจำแนกด้วยการเรียนรู้ของเครื่องสำหรับการทำนายโรคฝีดาษลิง

นายกฤษติชัย รัตนวัน

นายชิตณุพงศ์ นาควงศ์

นายธนดล สวนงาม

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาเหมืองข้อมูล

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ

ปีการศึกษา 2568

คำนำ

ในยุคปัจจุบันที่ข้อมูลดิจิทัลเพิ่มขึ้นอย่างก้าวกระโดด (Big Data) การทำเหมืองข้อมูล (Data Mining) ได้กลายเป็นเครื่องมือสำคัญที่ขาดไม่ได้ในการสร้างความได้เปรียบทางการแข่งขันและการตัดสินใจที่มีประสิทธิภาพ Data Mining คือกระบวนการที่ซับซ้อนในการค้นหารูปแบบ ความสัมพันธ์ และข้อมูลเชิงลึกที่มีความหมายซึ่งซ่อนอยู่ในชุดข้อมูลขนาดใหญ่ ด้วยการประยุกต์ใช้เทคนิคทางสถิติ, ปัญญาประดิษฐ์ และการเรียนรู้ของเครื่อง (Machine Learning) ทำให้องค์กรสามารถเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ที่นำไปใช้ได้จริง ตั้งแต่การทำนายแนวโน้มตลาด การปรับปรุงการบริการลูกค้า การตรวจจับการฉ้อโกง ไปจนถึงการสนับสนุนการวินิจฉัยทางการแพทย์ ดังนั้น ความเชี่ยวชาญในด้าน Data Mining จึงถือเป็นทักษะหลักในการขับเคลื่อนนวัตกรรมและสร้างมูลค่าทางธุรกิจ

เพื่อให้บรรลุความเข้าใจอย่างลึกซึ้งในหลักการและแนวปฏิบัติของ Data Mining การศึกษาจากตำราเรียนและบทความเบื้องต้นเพียงอย่างเดียวอาจไม่เพียงพอ การทำความเข้าใจอย่างถ่องแท้จำเป็นต้องอาศัยการเข้าถึงองค์ความรู้ที่ทันสมัยและเป็นปัจจุบันที่สุด ซึ่งมักถูกนำเสนออยู่ใน งานวิจัยและบทความทางวิชาการ การอ่านงานวิจัยไม่เพียงแต่ช่วยให้ผู้เรียนทราบถึงเทคนิคและอัลกอริทึมล่าสุดที่ถูกพัฒนาขึ้นเท่านั้น แต่ยังช่วยให้เห็นภาพรวมของความท้าทายในปัจจุบัน (Current Challenges) และช่องว่างขององค์ความรู้ (Knowledge Gaps) ในสาขานี้ การวิเคราะห์แนวคิด วิธีการทดลอง และผลลัพธ์จากงานวิจัยที่ตีพิมพ์ช่วยเสริมสร้างทักษะการคิดเชิงวิพากษ์และเตรียมความพร้อมในการประยุกต์ใช้เครื่องมือ Data Mining อย่างมีประสิทธิภาพในสถานการณ์จริง

การศึกษาวิเคราะห์งานวิจัยเชิงลึกเกี่ยวกับ Data Mining นี้ คาดว่าจะก่อให้เกิดประโยชน์ที่สำคัญหลายประการ ประการแรก ผู้ศึกษาจะได้รับความรู้ที่ทันสมัยและเป็นระบบเกี่ยวกับประเภทของงานวิจัย วิธีการวิเคราะห์ข้อมูล และการตีความผลลัพธ์ในบริบทของ Data Mining ประการที่สอง จะช่วยพัฒนาความสามารถในการระบุและประเมินประสิทธิภาพของแบบจำลองการทำเหมืองข้อมูลที่เหมาะสมกับปัญหาที่แตกต่างกัน ซึ่งเป็นการยกระดับทักษะการวิเคราะห์ข้อมูลเชิงลึก และ ประการสุดท้าย ผลจากการเรียนรู้ผ่านการอ่านงานวิจัยนี้จะปูทางไปสู่การประยุกต์ใช้เทคนิค Data Mining เพื่อแก้ไขปัญหาจริงในสาขาที่สนใจ หรืออาจนำไปสู่การพัฒนางานวิจัยต่อยอดในอนาคตได้อย่างมีคุณภาพ

สารบัญ

เนื้อหา	หน้า
คำนำ	ข
สารบัญ	ค
บทที่ 1 ที่มาและความสำคัญของปัญหา	1
บทที่ 2 วิธีดำเนินการ	2
บทที่ 3 เทคนิค Data mining ที่ใช้ในงานวิจัย	3
บทที่ 4 ผลลัพธ์	5
บรรณานุกรม	ง

บทที่ 1

ที่มาและความสำคัญของปัญหา

นับตั้งแต่เดือนพฤษภาคม ปี 2022 ไวรัสฝีดาษลิง (Monkeypox) จำนวนมากได้ส่งผลกระทบต่อผู้คนในหลายประเทศ รวมถึงสหราชอาณาจักร สเปน โปรตุเกส และสหรัฐอเมริกา และล่าสุดในอินเดีย ไวรัสฝีดาษลิงก็เริ่มส่งผลกระทบต่อผู้คนเช่นกัน ด้วยเหตุนี้ ศูนย์ควบคุมและป้องกันโรคแห่งสหรัฐอเมริกา (US Centers for Disease Control and Prevention) และองค์การอนามัยโลก (WHO) จึงได้เพิ่มมาตรการเฝ้าระวังการติดเชื้อออร์โธพอกซ์ไวรัส (orthopoxvirus) และเรียกร้องให้ผู้ให้บริการด้านสุขภาพเฝ้าระวังผู้ป่วยที่มีอาการคล้ายผื่นไข้ทรพิษ

จากการวินิจฉัยของผู้เชี่ยวชาญด้านฝีดาษลิง แพทย์ได้ดำเนินการโครงการรักษาหลากหลายรูปแบบสำหรับการบำบัดรักษา เมื่อเกิดการวินิจฉัยผิดพลาด (misdiagnosis) จะนำไปสู่การรักษาที่ไม่เหมาะสมและทำให้ผู้ป่วยพลาดช่วงเวลาที่ดีที่สุดสำหรับการรักษา ซึ่งจะนำไปสู่ผลลัพธ์ที่ร้ายแรง ดังนั้น การเลือกแบบจำลองเพื่อทำนายลักษณะของโรคฝีดาษลิงจึงมีความสำคัญอย่างยิ่ง

ในบทความนี้ ได้มีการเลือกใช้แบบจำลองการจำแนกประเภท (classification models) ที่แตกต่างกันแปดแบบเพื่อจำแนกลักษณะของโรคฝีดาษลิงสำหรับผู้ป่วย ได้แก่ Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, KNN, SVM, และขั้นตอนวิธีแบบ Ensemble คือ Ada Boosting และ Gradient Boosting ชุดข้อมูลไวรัสฝีดาษลิงได้มาจาก Kaggle] และถูกนำมาใช้สำหรับการตรวจสอบความถูกต้องและการสาธิต ผลลัพธ์สุดท้ายแสดงให้เห็นว่า Gradient Boosting มีความสามารถในการทำนายไวรัสฝีดาษลิงสำหรับผู้ป่วยได้ดีกว่า และให้ประสิทธิภาพที่ดีกว่าแบบจำลองทั้งเจ็ดแบบในด้านความแม่นยำในการจำแนกประเภท

บทที่ 2

วิธีดำเนินการ

การศึกษาวิจัยเรื่อง "การวิเคราะห์ประสิทธิภาพของแนวทางการจำแนกด้วยการเรียนรู้ของเครื่องสำหรับการทำนายโรคฝีดาษลิง" มีขั้นตอนการดำเนินงานที่เป็นระบบเพื่อให้ได้ผลลัพธ์ที่น่าเชื่อถือ โดยแบ่งออกเป็น 4 ขั้นตอนหลัก

1. การรวบรวมข้อมูล (Data Collection)
2. การเตรียมและประมวลผลข้อมูล (Data Preprocessing)
3. การสร้างและจำแนกด้วยโมเดล (Classification Modeling)
4. การวัดและประเมินผลประสิทธิภาพ (Performance Evaluation)

บทที่ 3

เทคนิค Data mining ที่ใช้ในงานวิจัย

3.1 การแยกข้อมูล

ในการทดลองนี้ ผู้วิจัยได้ทำการแบ่งชุดข้อมูล (Dataset) ออกเป็น 2 ส่วนหลัก เพื่อให้มั่นใจว่าโมเดลสามารถเรียนรู้และทำนายผลได้อย่างถูกต้อง โดยไม่เกิดโอเวอร์ฟิตติ้ง (Overfitting)

1. ชุดข้อมูลสำหรับฝึกสอน (Training Set): เป็นข้อมูลส่วนใหญ่ที่ใช้ป้อนให้คอมพิวเตอร์เพื่อ "เรียนรู้" รูปแบบความสัมพันธ์ระหว่างอาการต่างๆ กับผลการติดเชื้อ (Positive/Negative)
2. ชุดข้อมูลสำหรับทดสอบ (Testing Set): เป็นข้อมูลส่วนที่แยกไว้ต่างหาก ไม่เคยถูกนำไปสอนโมเดลมาก่อน ใช้สำหรับ "วัดผล" ว่าเมื่อโมเดลเจอข้อมูลผู้ป่วยรายใหม่ จะสามารถทำนายได้ถูกต้องหรือไม่

3.2 การเตรียมข้อมูล

ก่อนนำข้อมูลเข้าสู่อัลกอริทึม ข้อมูลดิบจะต้องผ่านกระบวนการเตรียมความพร้อมเพื่อให้มีคุณภาพและอยู่ในรูปแบบที่คอมพิวเตอร์ประมวลผลได้ (Machine Readable)

1. การทำความสะอาดข้อมูล (Data Cleaning): ตรวจสอบค่าที่สูญหาย (Missing Values) หรือข้อมูลรบกวน (Noise) เพื่อไม่ให้เกิดข้อผิดพลาดในการคำนวณ
2. การคัดเลือกคุณลักษณะ (Feature Selection): เลือกเฉพาะข้อมูลอาการที่เกี่ยวข้องกับโรคฝีดาษลิง เช่น ไข้ ผื่น ต่อมน้ำเหลืองโต เพื่อลดความซับซ้อนที่ไม่จำเป็น

3. การแปลงข้อมูล (Data Transformation/Encoding): เนื่องจากคอมพิวเตอร์คำนวณด้วยตัวเลข ข้อมูลที่เป็นข้อความ เช่น "มีอาการ" "ไม่มีอาการ" จะถูกแปลงเป็นตัวเลข 1 และ 0 เพื่อให้สามารถนำไปคำนวณทางคณิตศาสตร์ในขั้นตอนต่อไปได้

3.3 อัลกอริทึมจำแยก (Classification Algorithm)

งานวิจัยนี้ได้เลือกใช้เทคนิคการจำแนกข้อมูล (Classification) ซึ่งเป็นเทคนิคหลักของ Data Mining ในการทำนายผลลัพธ์แบบกลุ่ม (เป็นโรค หรือ ไม่เป็นโรค) โดยมีการนำ 8 อัลกอริทึม มาทดสอบประสิทธิภาพ

1. Decision Tree (DT): สร้างโมเดลการตัดสินใจในรูปแบบโครงสร้างต้นไม้ โดยแตกกิ่งก้านสาขาตามเงื่อนไขของอาการต่างๆ
2. Random Forest (RF): สร้าง Decision Tree หลายๆ ต้น แล้วนำผลลัพธ์มาโหวตร่วมกันเพื่อความแม่นยำที่สูงขึ้น
3. Naïve Bayes (NB): ใช้หลักความน่าจะเป็นทางสถิติ (Bayes' Theorem) ในการคำนวณโอกาสที่จะเป็นโรคจากอาการที่ปรากฏ
4. K-Nearest Neighbor (KNN): จำแนกผู้ป่วยโดยดูจาก "เพื่อนบ้าน" หรือข้อมูลผู้ป่วยคนอื่นๆ ในฐานข้อมูลที่มีอาการใกล้เคียงกันที่สุด
5. Support Vector Machine (SVM): พยายามหาเส้นแบ่ง (Hyperplane) ที่แบ่งแยกกลุ่มผู้ป่วยที่เป็นโรคและไม่เป็นโรคออกจากกันให้ชัดเจนที่สุด
6. Logistic Regression (LR): ใช้วิธีการทางสถิติเพื่อหาความสัมพันธ์ระหว่างตัวแปรอาการกับโอกาสการเกิดโรค
7. Ada Boosting (AB): เทคนิคการเรียนรู้แบบกลุ่ม (Ensemble) ที่สร้างโมเดลหลายตัว โดยโมเดลตัวหลังจะเน้นแก้ไขข้อผิดพลาดของโมเดลตัวก่อนหน้า
8. Gradient Boosting (GB): (อัลกอริทึมที่ได้ผลลัพธ์ดีที่สุดในงานวิจัยนี้) เป็นเทคนิคการเรียนรู้แบบกลุ่มที่ค่อยๆ สร้างโมเดลเพื่อลดค่าความผิดพลาด (Loss Function) ลงไปเรื่อยๆ ตามลำดับความชัน

บทที่ 4

ผลลัพธ์

ในการศึกษานี้ แบบจำลอง DT, SVM, LR, NB, KNN, RF, AB และ GB ถูกนำมาใช้ในการจำแนกประเภทเพื่อทำนายการมีอยู่ของไวรัสฝีดาษลิง ผลลัพธ์ของการทำนายช่วยลดอัตราการวินิจฉัยที่ผิดพลาดและช่วยให้สามารถให้การรักษาที่เหมาะสมสำหรับโครงการบำบัด ข้อมูลชุดไวรัสฝีดาษลิงประกอบด้วย 11 คุณลักษณะ (attributes) และ 25,000 ตัวอย่าง (instances) เพื่อค้นหาไวรัสฝีดาษลิง

หลังจากเปรียบเทียบความแม่นยำ (Accuracy) และ F-Score ของแบบจำลองการจำแนกประเภททั้งแปดแบบ ผลลัพธ์แสดงให้เห็นว่า Gradient Boosting ถูกเลือกเป็นแบบจำลองการจำแนกประเภทหลักในการศึกษานี้ ค่าความแม่นยำและ F-Score ของอัลกอริทึม Gradient Boosting คือ 71% และ 79.44% ตามลำดับ ดังนั้นผลลัพธ์ของการศึกษานี้จึงเป็นข้อมูลอ้างอิงสำหรับผู้เชี่ยวชาญในการจำแนกลักษณะของไวรัสฝีดาษลิง ในอนาคตเทคนิคการเรียนรู้เชิงลึก (Deep Learning) สามารถนำมาใช้เพื่อปรับปรุงความแม่นยำของการจำแนกประเภทได้

บรรณานุกรม

Manjunath, P. T. V. P., Pradeep, B., Naidu, K. R. R. A. V. S. R., & Dheeraj, A. J. S. V. (2022). Performance analysis of machine learning classification approaches for monkey pox disease prediction. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1656–1661). IEEE. https://doi.org/10.1109/ICECA55336.2022.10009407