Hepatitis C Virus (HCV) for Egyptian patients

Dataset overview

ข้อมูลจาก UCI Machine Learning Repository ถูกเปิดเผยเมื่อปี 2019 ภายใต้ใบอนุญาต CC BY 4.0 ที่สามารถนำข้อมูลไปใช้ได้แม้กระทั่งในเชิงพาณิชย์ เพียงแค่อ้างอิงถึงเจ้าของข้อมูล ซึ่งข้อมูลนี้ถูกใช้เพื่อทำนายระดับความรุนแรงของผังผืดในตับของผู้ป่วย HCV

Feature dictionary *ข้อมูลไม่มี missingness และช่วงของข้อมูลที่เป็นไปได้จะอ้างอิงจากข้อมูลจริงใน CSV

ชื่อ	ประเภท	Medical meaning	หน่วย	ช่วงของข้อมูล
Age	numerical	อายุ	ปี	32 - 61
Gender	binary	เพศ	-	1=ชาย, 2=หญิง
BMI	numerical	ดัชนีมวลกาย	กิโลกรัมต่อตารางเมตร	22 - 35
Fever	binary	บ่งบอกว่าผู้ป่วยมีไข้หรือไม่	-	1=ไม่มีไข้, 2=มีไข้
Nausea/Vomting	binary	มีอาการคลื่นไส้ หรืออาเจียน หรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
Headache	binary	มีอาการปวดศีรษะ หรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
Diarrhea	binary	มีอาการท้องร่วง หรือท้องเสีย หรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
Fatigue & generalized bone ache	binary	มีอาการอ่อนเพลีย ร่วมกับอาการปวดเมื่อยตามกระดูก หรือตามตัวโดยทั่วไป หรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
Jaundice	binary	มีอาการตัวเหลืองตาเหลือง หรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
Epigastric pain	binary	มีอาการปวดบริเวณลิ้นปี่ หรือยอดอกหรือไม่	-	1=ไม่มีอาการ, 2=มีอาการ
WBC	numerical	จำนวนเซลล์เม็ดเลือดขาวในเลือดของผู้ป่วย	เซลล์ต่อไมโครลิตร	2991 - 12101
RBC	numerical	จำนวนเซลล์เม็ดเลือดแดงในเลือดของผู้ป่วย	เซลล์ต่อไมโครลิตร	3816422 - 5018451
HGB	numerical	ความเข้มข้นของฮีโมโกลบิน ในเลือดของผู้ป่วย	กรัมต่อเดซิลิตร	2 - 20
Plat	numerical	จำนวนเกล็ดเลือด	เซลล์ต่อไมโครลิตร	93013 - 226464
AST 1	numerical	ค่าเอนไซม์ AST (Aspartate Transaminase) ในเลือดของสัปดาห์ที่ 1	ยูนิตต่อลิตร	39 - 128
ALT 1, 4, 12, 24, 36, 48, after 24 w	numerical	ค่าเอนไซม์ ALT (Alanine Transaminase) ในเลือดของผู้ป่วย ในเวลาที่แตกต่างกัน ตลอดกระบวนการรักษา หมายเลย 1, 4, 12 ฯลฯ หมายถึงสัปดาห์ของการรักษา	ยูนิตต่อลิตร	5 - 128
RNA Base	numerical	ปริมาณไวรัส HCV (Baseline Viral Load) ในกระแสเลือดก่อนที่จะได้รับยาต้านไวรัส	ยูนิตต่อมิลลิลิตร	0 - 1201086
RNA 4, 12	numerical	ปริมาณไวรัส HCV (Viral Load) ในกระแสเลือดของสัปดาห์ที่ 4 และ 12	ยูนิตต่อมิลลิลิตร	0 - 3731527
RNA EOT	numerical	ปริมาณไวรัส HCV (Viral Load) ในกระแสเลือดเมื่อสิ้นสุดการรักษา	ยูนิตต่อมิลลิลิตร	0 - 808450
RNA EF	numerical	ปริมาณไวรัส HCV (Viral Load) ในกระแสเลือดเมื่อสิ้นสุดการติดตามผล	ยูนิตต่อมิลลิลิตร	0 - 810333
Baseline histological Grading	categorical	ระดับการอักเสบ และการตายของเซลล์ตับที่ประเมินจากชิ้นเนื้อตับ ที่ประเมินเมื่อเริ่มรักษา	-	1 - 16
Baselinehistological staging	categorical	target ระดับความรุนแรงของพังผืดในตับ	-	1 - 4

Problem statement with causal reasoning

การติดเชื้อไวรัสตับอักเสบซีเรื้อรัง (HCV) เป็นสาเหตุที่ทำให้เกิดการอักเสบของตับอย่างต่อเนื่อง และส่งผลต่อค่าเลือดต่างๆ ความเสียหายที่สะสมนี้ส่งผลกระทบให้เกิดพังผืดในตับ (Fibrosis) ซึ่งเป็นแผลเป็นที่นำไปสู่ภาวะตับแข็ง เป้าหมายของการศึกษานี้คือการพยายามทำนายระยะของพังผืดในตับ 1-4 (Classification Task) โดยใช้ข้อมูล เช่น ผลเลือดและอาการ เพื่อหลีกเลี่ยงการเจาะชิ้นเนื้อตับ

Methods

ทดลองฝึกโมเดลจากข้อมูลต้นฉบับ และทดลองเตรียมข้อมูลโดยการเปลี่ยนข้อมูลประเภท Numerical ทั้งหมดให้เป็น Categorical โดยอ้างอิงจากงานวิจัย ในตารางที่ 1 ที่พัฒนาโมเดลในข้อมูลชุดเดียวกัน ในงานนี้ทดลองฝึกกับโมเดล Random Forest, Support Vector Machine, Multilayer Perceptron, Bagging Classifierและ Hist Gradient Boosting สาเหตุที่เลือกโมเดลเหล่านี้มทดลองเพราะมีความเหมาะกับ Classification Task ที่มีความซับซ้อนของข้อมูลสูงจะเห็นได้ว่ามีการใช้โมเดลประเภท Ensemble (โมเดลเดียวกันแต่หลายตัว) เพื่อช่วยให้การตอบมีความแม่นยำมากยิ่งขึ้น

Evaluation

ใช้ตัววัดประสิทธิภาพเป็น Accuracy, F1-Score เพราะว่าแต่ละคลาสข้อมูล balance กันอย่างดี ซึ่งผมลัพธ์ที่ได้จากการฝึกโมเดลทั้งหมดได้ Accuracy และ F1-Score อยู่ในช่วง 0.2-0.31 ทั้งๆ ที่การเตรียมข้อมูล และโมเดลที่ทำอ้างอิงมากจากงานวิจัย แต่ในงานวิจัยมีความแม่นยำสูงถึง 0.9-1.0 นั่นจึงสรุปได้ว่าการทดลองในงานวิจัยนั้นไม่สามรถทำซ้ำได้ และอาจจะมีการฝึกโมเดลให้จดจำรูปแบบของข้อมูลโดยไม่สนใจ Overfitting ก็เป็นได้

จากการฝึกโมเดลทำให้ทราบว่าโมเดลให้ความสำคัญกับ Feature Age, BMI, HBG, WBC โดย Feature ที่กล่าวมาทั้งหมดอยู่ในรูปแบบเป็น Categorical และมีความสำคัญจากมากไปน้อยตามลำดับ

Conclusion

จากผลการทดลองสรุปได้ว่าข้อมูลผลเลือดและอาการที่นำมาฝึกโมเดลไม่เพียงพอต่อการทำนายระยะของพังผืดในตับ 1-4 เพื่อให้ความแม่นยำเพิ่มมากยิ่งขึ้นก็ยังคงต้องเจาะชิ้นเนื้อตับเพื่อนำไปตรวจ แต่ในการนำชิ้นเนื้อมาตรวจกยังสามารถนำปัญญาประดิษฐ์เข้าไปช่วยในการลดเวลาการวินิจฉัยได้ แต่จะเป็นการส่วนของ Image Classification Task นั่นเอง

References

Dataset: https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients

Paper: https://ieeexplore.ieee.org/document/8289800

Author

Tanarat Saehia 663380035-4 CP-AI @KKU