

Predicting estimated glomerular filtration rate (eGFR) for chronic kidney disease risk assessment using machine learning techniques: A case study of Khu Mueang Hospital

Wannaporn Jaimeetham

Khu Mueang Hospital, Buriram

Abstract

Chronic Kidney Disease (CKD) is a significant health problem affecting populations worldwide. Analyzing and predicting the estimated glomerular filtration rate (eGFR) is essential for planning patient care at various stages, including primary prevention, secondary prevention, and tertiary prevention. In this study, we experimented with various machine learning models, including Linear Regression, Random Forest, Decision Tree, Gradient Boosting Machine (GBM), XGBoost, and LightGBM, to identify the factors influencing eGFR. We selected the most effective model based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 . The results demonstrated that the most effective model was Random Forest, with an MAE of 0.146688, an MSE of 0.260079, and an R^2 of 0.999688, indicating the highest accuracy across all metrics. The R^2 value, being close to 1, reflects the

model's ability to explain the variability in the data effectively. Furthermore, the Random Forest model was refined using Grid Search, resulting in optimal parameters. This model proved to be highly effective in predicting eGFR for patients with chronic kidney disease (CKD) with a low MSE and a high R^2 score, showcasing its capability for accurate outcome predictions. It can be utilized to assess the risk of CKD effectively, aiding in the planning and prevention of chronic kidney disease more efficiently.

Keywords: *Chronic Kidney Disease (CKD), Primary Prevention, Secondary Prevention, Tertiary Prevention, Machine Learning, Risk Factor Prediction, Model Performance*

Received: 10 June 2025, Revised: 25 July 2025, Accepted: 1 September 2025

Correspondence: Wannaporn Jaimeetham Khu Mueang Hospital, 101, Khu Mueang, Khu Mueang District, Buriram 31190, Thailand, Tel.: 044 699 238, email: wjting.954@gmail.com

การเปรียบเทียบต้นทุนประสิทธิภาพของยาปฏิชีวนะแบบฉีดเข้าหลอดเลือดดำ สำหรับการรักษาภาวะติดเชื้อในกระแสเลือด: กรณีศึกษาที่โรงพยาบาลคูเมือง

วรรณพร ใจมีธรรม

โรงพยาบาลคูเมือง บุรีรัมย์

บทคัดย่อ

โรคไตเรื้อรัง (CKD) เป็นปัญหาสุขภาพที่สำคัญซึ่งมีผลกระทบต่อประชากรทั่วโลก การวิเคราะห์และทำนายค่าการกรองของไต (eGFR) เป็นสิ่งจำเป็นในการวางแผนการดูแลผู้ป่วยในระยะต่าง ๆ ทั้งการป้องกันระดับปฐมภูมิ (Primary Prevention) ระดับทุติยภูมิ (Secondary Prevention) และระดับตติยภูมิ (Tertiary Prevention) ในการศึกษาครั้งนี้ เราได้ทดลองใช้โมเดลการเรียนรู้ของเครื่อง (Machine Learning) หลากหลายประเภท ได้แก่ Linear Regression, Random Forest, Decision Tree, Gradient Boosting Machine (GBM), XGBoost และ LightGBM เพื่อค้นหาปัจจัยที่มีผลต่อค่า eGFR โดยเลือกโมเดลที่มีประสิทธิภาพสูงสุดจากการวัดประสิทธิภาพของโมเดลด้วยค่า Mean Absolute Error (MAE), Mean Squared Error (MSE) และ R² ผลการทดลองแสดงให้เห็นว่าโมเดลที่มีประสิทธิภาพสูงสุดคือ Random Forest โดยมี ค่า MAE เท่ากับ 0.146688, MSE เท่ากับ 0.260079 และ R² เท่ากับ 0.999688 ซึ่งแสดงให้เห็นถึงความแม่นยำสูงที่สุดในทุกเมตริกซ์ โดยค่า R² ใกล้เคียง 1 แสดงถึงความสามารถในการอธิบายความแปรปรวน

ในข้อมูลได้ดี นอกจากนี้ โมเดล Random Forest ยังได้รับการปรับปรุงด้วยการใช้ Grid Search ซึ่งส่งผลให้ได้พารามิเตอร์ที่เหมาะสมที่สุดการสร้างโมเดลนี้มีประสิทธิภาพสูงในการทำนายค่า eGFR สำหรับผู้ป่วยโรคไตเรื้อรัง (CKD) ด้วยค่า MSE ที่ต่ำ และ R² Score ที่สูง แสดงถึงความสามารถในการคาดการณ์ผลลัพธ์ได้อย่างแม่นยำ ซึ่งสามารถนำไปใช้ในการประเมินความเสี่ยงของโรคไตเรื้อรังได้อย่างมีประสิทธิภาพ ช่วยในการวางแผนการดูแลและป้องกันโรคไตเรื้อรังอย่างมีประสิทธิภาพยิ่งขึ้น.

คำสำคัญ: โรคไตเรื้อรัง, การป้องกันระดับปฐมภูมิ, การป้องกันระดับทุติยภูมิ, การป้องกันระดับตติยภูมิ, การเรียนรู้ของเครื่อง, การทำนายปัจจัยเสี่ยง, ประสิทธิภาพโมเดล

วันที่รับต้นฉบับ: 10 มิถุนายน 2568, วันที่แก้ไข: 25 กรกฎาคม 2568, วันที่ตอบรับ: 1 กันยายน 2568

บทนำ

โรคไตเรื้อรัง (Chronic Kidney Disease, CKD) เป็นปัญหาสุขภาพที่สำคัญทั่วโลก องค์การอนามัยโลก (WHO) ระบุว่า CKD เป็นหนึ่งในสาเหตุหลักของการเสียชีวิตและความพิการทั่วโลก โดยเฉพาะในกลุ่มประชากรที่มีปัจจัยเสี่ยง เช่น ผู้ป่วยเบาหวานและความดันโลหิตสูง ในรายงานของ WHO เกี่ยวกับการจัดการและการดูแลผู้ป่วยโรคไตในสถานการณ์ฉุกเฉิน ได้เน้นถึงความจำเป็นของการเฝ้าระวังและประเมินการทำงานของไตโดยใช้ค่าการกรองของไต (eGFR) เพื่อวางแผนการดูแลและป้องกันการดำเนินของโรคไตเรื้อรัง [2]

สำหรับประเทศไทย สถานการณ์โรคไตเรื้อรังยังคงเป็นปัญหาที่ทวีความรุนแรงมากขึ้น ข้อมูลจากกรมควบคุมโรคกระทรวงสาธารณสุข ระบุว่า จำนวนผู้ป่วย CKD ในประเทศไทย

เพิ่มขึ้นอย่างต่อเนื่อง โดยมีความเชื่อมโยงกับพฤติกรรมสุขภาพและปัจจัยเสี่ยง เช่น เบาหวาน ความดันโลหิตสูง และโรคอ้วน ซึ่งเป็นปัจจัยเสี่ยงสำคัญในการเกิดโรคไตเรื้อรัง [13]

จากข้อมูลของโรงพยาบาลคูเมือง พบว่าจำนวนผู้ป่วยโรคไตใน Stage 1 และ Stage 2 มีแนวโน้มที่คงที่หรือเพิ่มขึ้นเล็กน้อย ในขณะที่จำนวนผู้ป่วยใน Stage 3 มีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง จาก 942 คนในปี 2562 เป็น 1,417 คนในปี 2566 นอกจากนี้ ผู้ป่วยใน Stage 4 และ Stage 5 ก็พบว่ามีแนวโน้มเพิ่มขึ้นเช่นกัน ซึ่งสะท้อนถึงความท้าทายที่ระบบสาธารณสุขต้องเผชิญในการดูแลผู้ป่วยโรคไตเรื้อรัง

การวิเคราะห์ปัจจัยที่ส่งผลต่อค่า eGFR เป็นสิ่งสำคัญในการวางแผนดูแลและป้องกันโรคไต เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ถูกนำมาใช้เพื่อคาดการณ์และประเมินความเสี่ยงของโรคไตเรื้อรัง (CKD) ช่วยเพิ่มประสิทธิภาพในการวางแผนการดูแลผู้ป่วย โดยการประเมิน eGFR เป็นตัวชี้วัดสำคัญในการจัดการโรคไต การใช้เทคนิคนี้สามารถระบุปัจจัยเสี่ยงที่มีผลต่อ eGFR และพัฒนาแผนการ

ผู้สนับสนุนประสานงาน: วรรณพร ใจมีธรรม โรงพยาบาลคูเมือง 101 หมู่ 6 ตำบลคูเมือง อำเภอคูเมือง จังหวัดบุรีรัมย์ 31190, โทร.: 044 699 238, email: wjting.954@gmail.com

ป้องกันที่เหมาะสมสำหรับผู้ป่วยในแต่ละระยะได้อย่างมีประสิทธิภาพ

วัตถุประสงค์ในการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง (Machine Learning) ที่สามารถคาดการณ์ความเสี่ยงของโรคไตเรื้อรัง (CKD) โดยการประเมินค่าการกรองของไต (eGFR) ในผู้ป่วยซึ่งจะช่วยให้สามารถดำเนินการป้องกันและจัดการได้อย่างมีประสิทธิภาพมากขึ้น โดยมีวัตถุประสงค์หลักดังนี้:

- 1) การประเมินความเสี่ยง: เพื่อประเมินปัจจัยที่มีผลต่อการเกิด CKD ในระยะยาว โดยพิจารณาจากปัจจัยที่เกี่ยวข้องกับสุขภาพ เช่น อายุ เพศ ประวัติการเป็นโรคเบาหวาน และความดันโลหิต
- 2) การพัฒนาโมเดล: เพื่อพัฒนาโมเดลการเรียนรู้ของเครื่องที่มีประสิทธิภาพสูงในการคาดการณ์ค่า eGFR โดยใช้เทคนิคที่หลากหลาย เช่น Linear Regression, Random Forest, Gradient Boosting และ XGBoost
- 3) การเปรียบเทียบประสิทธิภาพ: เพื่อเปรียบเทียบประสิทธิภาพของโมเดลต่าง ๆ โดยใช้เมตริกซ์เช่น Mean Absolute Error (MAE), Mean Squared Error (MSE) และ R^2 เพื่อระบุโมเดลที่มีความแม่นยำสูงสุด
- 4) การสนับสนุนการป้องกัน: เพื่อให้ข้อมูลและเครื่องมือที่จำเป็นแก่ผู้ให้บริการด้านสุขภาพในการระบุบุคคลที่มีความเสี่ยงต่อ CKD และดำเนินการป้องกันได้อย่างทันที่

นทววรรณกรรม

การทบทวนวรรณกรรมเกี่ยวกับปัจจัยที่ส่งผลต่อโรคไตเรื้อรังและการใช้ Machine Learning

1) ปัจจัยที่ส่งผลกระทบต่อโรคไตเรื้อรัง

โรคไตเรื้อรัง (Chronic Kidney Disease, CKD) เป็นปัญหาสุขภาพที่สำคัญที่ได้รับความสนใจจากนักวิจัยและผู้เชี่ยวชาญด้านสุขภาพ เนื่องจากส่งผลกระทบต่อคุณภาพชีวิตของผู้ป่วย และมีอัตราการเกิดที่สูงขึ้นในประชากรหลายกลุ่ม ในการศึกษาวิจัยนี้ ผู้วิจัยมุ่งเน้นไปที่ปัจจัยที่มีอิทธิพลต่อการเกิด CKD และการใช้การเรียนรู้ของเครื่อง (Machine Learning: ML) เพื่อทำนายความเสี่ยงที่เกี่ยวข้องกับโรค

CKD ได้รับอิทธิพลจากหลายปัจจัยที่สำคัญ ได้แก่ โรคร่วม เช่น ความดันโลหิตสูงและโรคเบาหวาน ซึ่งเป็นปัจจัยหลักที่เกี่ยวข้องกับการเกิด CKD โดย Pontes et al. [12] พบว่าผู้ป่วยที่มีความดันโลหิตสูงที่ไม่ได้รับการควบคุมมีความเสี่ยงสูงที่จะพัฒนาเป็น CKD และความเสี่ยงนี้มีแนวโน้มเพิ่มขึ้นในกลุ่มบุคคลที่ขาดความรู้และการศึกษาเกี่ยวกับโรค นอกจากนี้

ยังมีการพบว่าความชุกของโรคเบาหวานในผู้ป่วย CKD มีอัตราสูงถึง 40% ในบางประชากร ซึ่งแสดงถึงความสัมพันธ์ที่แข็งแกร่งระหว่างโรคเบาหวานและ CKD

การศึกษาโดย Napitupulu et al. [11] และ Pontes et al. [12] ชี้ให้เห็นว่าปัจจัยทางพฤติกรรม เช่น การสูบบุหรี่ การบริโภคแอลกอฮอล์ และการบริโภคอาหารที่ไม่สมดุล มีความสัมพันธ์กับการเกิด CKD พฤติกรรมประจำวันและการเลือกวิถีชีวิตที่ไม่เหมาะสมจึงมีผลกระทบต่อความเสี่ยงที่เพิ่มขึ้นของ CKD ทำให้เกิดความจำเป็นในการปรับปรุงพฤติกรรมเพื่อป้องกันโรค อายุและเพศเป็นปัจจัยสำคัญที่มีอิทธิพลต่อการเกิด CKD โดย Xie & Sheng [4] และ Mbah et al. [2] รายงานว่าผู้สูงอายุและเพศชายมีความเสี่ยงสูงต่อการเกิด CKD มากกว่ากลุ่มประชากรอื่น ๆ

กรมควบคุมโรค ได้ศึกษาและพัฒนาคะแนนความเสี่ยงโรคไตใน 10 ปีข้างหน้า (Thai CKD Risk Score) ซึ่งมีการประเมินผ่านโมเดลที่แตกต่างกัน โดยโมเดลแรกเน้นไปที่ปัจจัยเสี่ยงพื้นฐาน เช่น อายุ เพศ รอบเอว ประวัติการป่วยโรคเบาหวาน และระดับความดันโลหิต ในขณะที่โมเดลที่สองรวมถึงค่าการกรองของไต (eGFR) เพื่อให้ได้การประเมินความเสี่ยงที่แม่นยำยิ่งขึ้น [10]

2) การใช้ Machine Learning กำหนดความเสี่ยงการเกิดโรคไตเรื้อรัง

การเรียนรู้ของเครื่อง (Machine Learning: ML) ได้รับการยอมรับว่าเป็นเครื่องมือที่มีศักยภาพสูงในการทำนายผลลัพธ์ของ CKD โดยเฉพาะในการประเมินความเสี่ยงของการพัฒนาไปสู่โรคไตวายระยะสุดท้าย (End-Stage Kidney Disease: ESKD) หลายการศึกษารายงานถึงประสิทธิภาพของอัลกอริทึม ML ในการวินิจฉัย CKD และความแม่นยำในการทำนายอาการ

Bai et al. [10] พบว่าอัลกอริทึม ML เช่น Logistic Regression และ Random Forest มีความไวในการทำนาย ESKD ใกล้เคียงกับ Kidney Failure Risk Equation (KFRE) ที่นิยมใช้ นอกจากนี้ Mahbub et al. [5] ยังรายงานว่าการใช้ Tree-Based Classifiers มีความแม่นยำสูงถึง 100% ในการวินิจฉัย CKD แสดงถึงศักยภาพของเทคโนโลยีนี้ในการให้ผลลัพธ์ที่แม่นยำและรวดเร็ว

การระบุปัจจัยที่สำคัญในการทำนาย CKD ก็มีบทบาทสำคัญ โดย Mahbub et al. [5] พบว่าฮีโมโกลบินและอัลบูมินเป็นตัวบ่งชี้ทางชีวภาพที่สำคัญ ในขณะที่ Forn et al. [8] แสดงให้เห็นว่าตัวบ่งชี้ทางชีวภาพบางชนิดสามารถช่วยเพิ่มความแม่นยำในการประเมินความเสี่ยงได้

ถึงแม้ว่า ML จะมีศักยภาพสูงในการประเมินความเสี่ยงและการตัดสินใจทางคลินิก แต่ก็ยังคงมีความท้าทายในการ

ตรวจสอบความถูกต้องของโมเดลและการบูรณาการข้อมูลทางคลินิกที่หลากหลายเพื่อเพิ่มความแม่นยำในการทำนาย ความร่วมมือระหว่างนักวิทยาศาสตร์ข้อมูลและแพทย์ผู้เชี่ยวชาญจึงเป็นสิ่งจำเป็นสำหรับการพัฒนา ML ในทางปฏิบัติอย่างมีประสิทธิภาพ.

วิธีการศึกษา

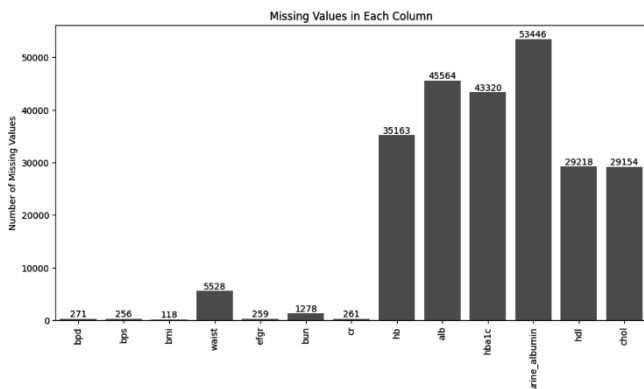
Study design, settings, and data source

1) การรวบรวมข้อมูล (Data collection)

ผู้วิจัยได้รวบรวมจากฐานข้อมูลการระบบ HIS โดยเขียน Query SQL ดึงข้อมูลรายงาน ระหว่างวันที่ 1 ตุลาคม 2563 – 30 สิงหาคม 2567 ทั้งหมด 58,724 รายการ ข้อมูลส่วนบุคคลที่สามารถระบุตัวตน ไม่ได้เป็นส่วนหนึ่งของชุดข้อมูล ข้อมูลสำคัญและเกี่ยวข้องกับการทำนายปัจจัยการเกิดโรคไตวายเรื้อรัง ได้แก่ เพศ อายุ (อายุ 18 ปีขึ้นไป) ค่าความดันโลหิต BMI รอบเอว ประวัติการมีโรคร่วม ได้แก่ เบาหวาน ความดันโลหิตสูง โรคหัวใจ ปอดอุดกั้นเรื้อรัง หัวใจและหลอดเลือด ผลตรวจทางห้องปฏิบัติการ ได้แก่ ค่า eGFR, HbA1C, Urine_albumin, HDL ,Cholesterol ประวัติการดื่มสุรา ประวัติการสูบบุหรี่ คำนวนค่า CKD risk scores, CKD stage

2) เตรียมข้อมูล (Data Cleansing/Data Preprocessing)

a) การตรวจสอบข้อมูลที่ขาดหาย (Missing Data Handling)



รูปที่ 1 ตรวจสอบข้อมูลที่ขาดหาย

b) จัดการข้อมูลที่ขาดหายโดยเติมค่า Mean ในคอลัมน์ที่เป็นตัวเลข

```
# Fill missing values with the mean for each numeric column and display the dataframe
df_filled = df.fillna(df.mean(numeric_only=True))

# Display the first few rows of the dataframe to verify the changes
df_filled.head()
```

bpd	bps	bmi	waist	diabetes	hypertension	heart	copd	ald	hba1c	urine_albumin	hdl	chol	smoking	alcohol	ckd
95.0	161.0	34.627	80.0	No	No	No	No	4.100000	7.775254	0.482908	52.393822	181.067852	No	No	
65.0	101.0	22.052	76.0	No	No	No	No	3.680773	7.775254	0.500000	52.393822	181.067852	No	No	
80.0	144.0	19.141	62.0	No	No	No	No	3.680773	7.775254	0.482908	52.393822	181.067852	No	Yes	
76.0	121.0	21.565	75.0	Yes	No	No	No	3.700000	7.775254	0.482908	52.393822	181.067852	No	No	5
90.0	124.0	24.977	69.0	No	No	No	No	3.300000	7.775254	0.000000	52.393822	181.067852	No	No	

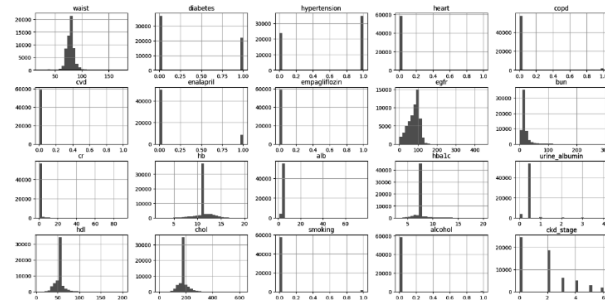
รูปที่ 2 เติมค่าข้อมูลที่ขาดหาย

c) การแปลงข้อมูล (Data Transformation) แปลงค่าคอลัมน์ที่เป็น Object ให้เป็น Numeric

sex	age_y	bpd	bps	bmi	waist	diabetes	hypertension	heart
0	28	95.0	161.0	34.627	80.0	0	0	0
1	43	65.0	101.0	22.052	76.0	0	0	0
0	56	80.0	144.0	19.141	62.0	0	0	0
0	75	76.0	121.0	21.565	75.0	1	1	0

รูปที่ 3 แปลงค่าข้อมูล

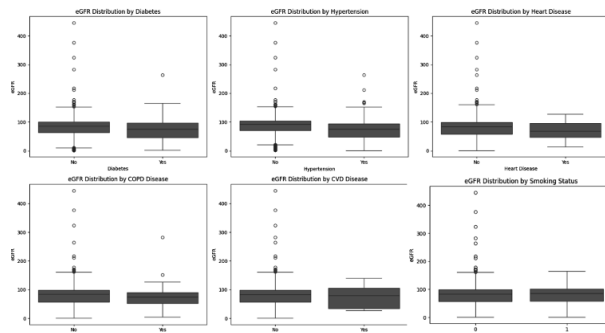
d) ตรวจสอบข้อมูลหลังมีการแปลงข้อมูลเพื่อให้มั่นใจว่าข้อมูลใช้ได้



รูปที่ 4 ตรวจสอบข้อมูลหลังแปลงค่า

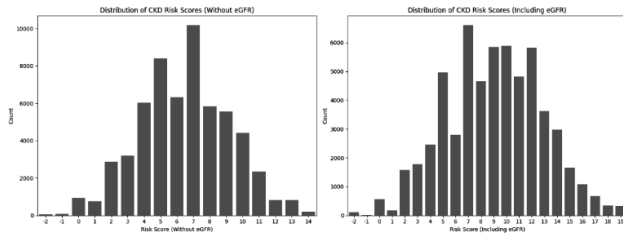
e) ตรวจสอบข้อมูลความสัมพันธ์และวิเคราะห์การกระจายตัวของข้อมูล (Correlation Matrix) หลังแปลงข้อมูล ตรวจสอบความสัมพันธ์ของค่า eGFR และตัวแปรอื่นๆ เพื่อจัดการก่อนเข้า Model

f) relationship between eGFR and chronic conditions

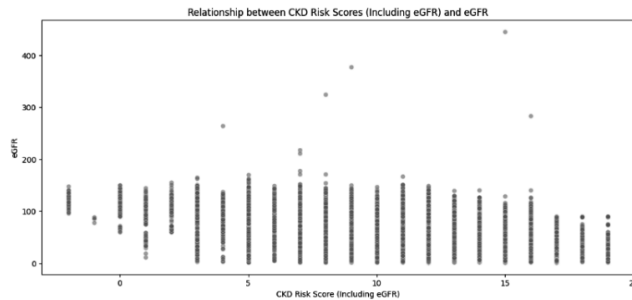


รูปที่ 5 ความสัมพันธ์ของค่าไตและโรคเรื้อรังอื่น

g) relationship between eGFR and CKD Risk score

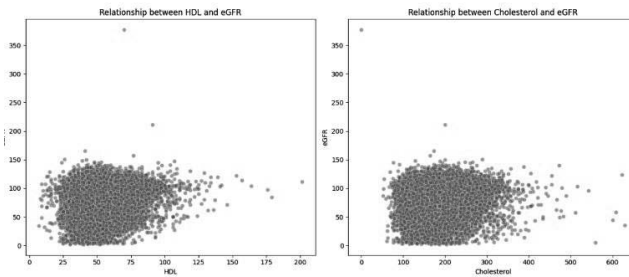


รูปที่ 6 แสดงความสัมพันธ์ค่า eGFR และ Risk score

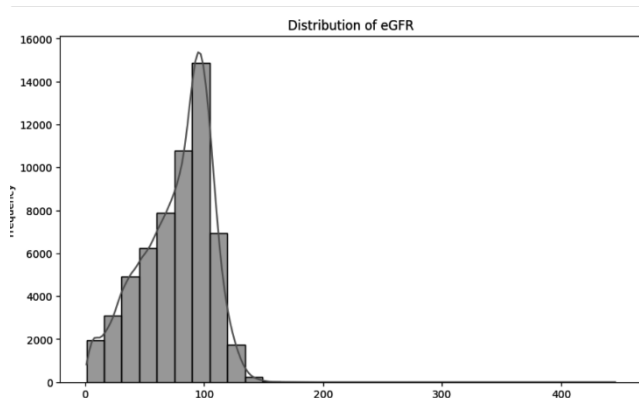


รูปที่ 7 แสดงความสัมพันธ์ค่า eGFR และ Risk score ที่ไม่ใช้ eGFR

c) relationship between eGFR and HDL, Cholesterol



รูปที่ 8 แสดงความสัมพันธ์ค่า eGFR และค่า HDL Cholesterol



รูปที่ 9 แสดงการกระจายตัวค่า eGFR

4) การทดสอบโมเดล

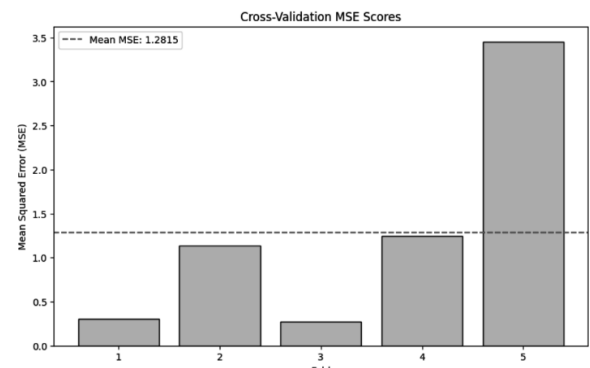
a) แบ่งข้อมูลออกเป็นชุดฝึก (Training Set) และชุดทดสอบ (Test Set) โดยใช้ `train_test_split` เพื่อประเมินประสิทธิภาพของโมเดลได้อย่างถูกต้อง

b) เลือกโมเดล Machine Learning เพื่อเปรียบเทียบประสิทธิภาพของโมเดล ได้แก่ Linear Regression, Random Forest, DecisionTreeRegressor, GradientBoostingRegressor, XGBoost, LightGBM

ตารางที่ 1 ผลการทดสอบโมเดล

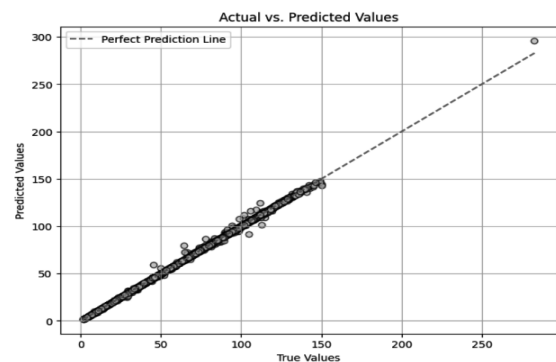
โมเดล	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
Linear Regression	5.978752	60.285084
Random Forest	0.146688	0.260079
Decision Tree	0.230314	0.927988
GBM	1.688123	5.235314
XGBoost	0.459021	1.107133
LightGBM	0.593546	2.408107

5) ปรับปรุงประสิทธิภาพของโมเดลและเมินผลซ้ำ จากการใช้ K-FOLD CROSS-VALIDATION



รูปที่ 10 แสดงผลการปรับปรุงประสิทธิภาพของโมเดล

6) คำนวณค่าคาดการณ์จากโมเดลที่ดีที่สุดและแสดงความสัมพันธ์ระหว่างค่าคาดการณ์และค่าจริง



รูปที่ 11 แสดงความสัมพันธ์ระหว่างค่าคาดการณ์และค่าจริง

7) การนำข้อมูลเข้าเพื่อทดสอบโมเดล

- นำเข้าโมดูล `os` เพื่อใช้งานกับระบบไฟล์
- นำเข้าโมดูล `pickle` สำหรับการแปลง `Object` ใน Python เป็นบิตและการอ่านบิตกลับมาเป็น `Object`
- กำหนดเส้นทางไปยังไฟล์โมเดลที่บันทึกไว้ (`model.pkl`) โดยใช้ `os.getcwd` เพื่อรับตำแหน่งปัจจุบันของโฟลเดอร์ที่ทำงาน
- โหลดโมเดลที่บันทึกไว้ในไฟล์ `model.pkl` โดยเปิดไฟล์ในโหมดอ่านแบบ `Binary`
- ทำนายค่าจากข้อมูลทดสอบ และแสดงผลลัพธ์การทำนายค่า `eGFR`

ผลการศึกษาและการอภิปรายผล

1) ผลการศึกษา

- เปรียบเทียบประสิทธิภาพของ Model

Linear Regression: MAE: 5.978752 MSE: 60.285084 R^2 : 0.927750 โมเดล Linear Regression มีค่าความผิดพลาดสูงและแสดงถึงความแม่นยำน้อยเมื่อเปรียบเทียบกับโมเดลอื่น ๆ

Random Forest: MAE: 0.146688 MSE: 0.260079 R^2 : 0.999688 การวิเคราะห์: Random Forest แสดงให้เห็นถึงความแม่นยำสูงที่สุดในทุกเมทริกซ์ โดยค่า R^2 ใกล้เคียง 1 แสดงถึงความสามารถในการอธิบายความแปรปรวนในข้อมูลได้ดี

Decision Tree: MAE: 0.230314 MSE: 0.927988 R^2 : 0.998888 การวิเคราะห์: แม้ว่า Decision Tree จะแสดงผลลัพธ์ที่ดี แต่ก็มีความแม่นยำน้อยกว่า Random Forest

GBM (Gradient Boosting): MAE: 1.688123 MSE: 5.235314 R^2 : 0.993726 การวิเคราะห์: GBM มีค่าความแม่นยำต่ำกว่า Random Forest และ Decision Tree

XGBoost: MAE: 0.459021 MSE: 1.107133 R^2 : 0.998673 การวิเคราะห์: XGBoost แสดงประสิทธิภาพที่ดี แต่ยังไม่สามารถเทียบเท่ากับ Random Forest หรือ Decision Tree

LightGBM: MAE: 0.593546 MSE: 2.408107 R^2 : 0.997114 การวิเคราะห์: LightGBM มีค่าความแม่นยำสูง แต่ยังคงต่ำกว่า Random Forest และ Decision Tree

ตารางที่ 2 เปรียบเทียบประสิทธิภาพของ Model

โมเดล	MAE	MSE	R^2
Linear Regression	5.978752	60.285084	0.927750
Random Forest	0.146688	0.260079	0.999688
Decision Tree	0.230314	0.927988	0.998888
GBM(Gradient Boosting)	1.688123	5.235314	0.993726
XGBoost	0.459021	1.107133	0.998673
LightGBM	0.593546	2.408107	0.997114

สรุปผลการเปรียบเทียบ Random Forest เป็นโมเดลที่มีประสิทธิภาพสูงที่สุดในการทำนายค่า `eGFR` โดยมีค่า MAE และ MSE ต่ำที่สุด พร้อมกับค่า R^2 ที่ใกล้เคียง 1 Decision Tree และ XGBoost ก็แสดงผลลัพธ์ที่ดี แต่ไม่สามารถแข่งขันกับ Random Forest ได้ Linear Regression เป็นโมเดลที่มีความแม่นยำน้อยที่สุดในกลุ่มนี้ โมเดลเหล่านี้จะใช้ในการจำแนกประเภทได้ดี แต่ก็สามารถนำมาใช้ในงานทำนายค่าต่อเนื่องได้เช่นกัน โดยเฉพาะในกรณีที่ข้อมูลมีความซับซ้อนหรือไม่เป็นเส้นตรง (non-linear relationships) โมเดลเช่น Random Forest และ Decision Tree มีความสามารถในการจัดการข้อมูลที่มีความไม่เป็นเชิงเส้นและความซับซ้อนสูง ซึ่งทำให้มีความเหมาะสมกับงานทำนายค่าต่อเนื่องในงานวิจัยนี้

- ประสิทธิภาพของโมเดลที่ได้รับการปรับแต่ง (Tuned Model Performance)

โมเดล Random Forest ที่ได้รับการปรับแต่งประสิทธิภาพของ Model สรุปผลการปรับแต่งโมเดล Random Forest ทดสอบด้วยการใช้ 3 folds สำหรับ 108 candidates รวมทั้งหมด 324 fits ในการปรับแต่งโมเดล พบพารามิเตอร์ที่ดีที่สุดคือ

`max_depth`: None (ไม่มีการจำกัดความลึกของต้นไม้)
`min_samples_leaf`: 1 (จำนวนตัวอย่างขั้นต่ำในใบไม้)
`min_samples_split`: 2 (จำนวนตัวอย่างขั้นต่ำสำหรับการแบ่ง)
`n_estimators`: 100 (จำนวนต้นไม้ในโมเดล Random Forest)

ผลการประเมินโมเดล Mean Squared Error (MSE): 0.260079 ค่านี้ต่ำมาก แสดงให้เห็นว่าโมเดลมีความแม่นยำสูงในการทำนาย

R^2 Score: 0.999688 Mean Squared Error (MSE): ค่าใกล้เคียง 1 แสดงว่าโมเดลสามารถอธิบายความแปรปรวนในข้อมูลได้อย่างมีประสิทธิภาพ การปรับแต่งโมเดลแสดงให้เห็นว่าโมเดล Random Forest ที่มีความสามารถในการทำนายที่ดีมาก มีพารามิเตอร์ที่เหมาะสมในการทำงาน นอกจากนี้ยังมีความแม่นยำสูงเมื่อประเมินผลด้วย MSE และ R^2 Score

- การวิเคราะห์ผลการตรวจสอบข้าม (Cross-Validation)

ในขั้นตอนการประเมินโมเดล เราได้ใช้การตรวจสอบข้าม (Cross-Validation) เพื่อวัดความสามารถของโมเดลในชุดข้อมูลที่แตกต่างกัน โดยเฉพาะในการทำนายค่า `eGFR` สำหรับผู้ป่วยโรคไตเรื้อรัง (CKD) ผลการตรวจสอบข้ามแสดงรายละเอียดดังนี้:

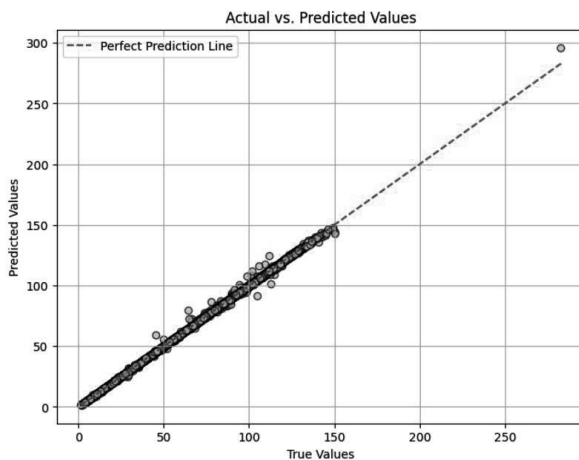
MSE Scores:[0.30052344, 1.13342901, 0.27499768, 1.24587703, 3.45270518]

Mean MSE: ค่า Mean MSE ที่ประมาณ 1.2815 แสดงถึงค่าความผิดพลาดเฉลี่ยในการทำนายของโมเดล แสดงให้เห็น

ว่ามีความแม่นยำในระดับหนึ่ง แต่ยังคงมีความผิดพลาดที่สามารถพัฒนาได้

Standard Deviation of MSE: ค่าเบี่ยงเบนมาตรฐานที่สูง (1.1587) แสดงถึงความแปรผันของผลลัพธ์จากการตรวจสอบข้ามที่ค่อนข้างสูง ซึ่งหมายความว่ามีความแตกต่างระหว่างผลลัพธ์จากการแบ่งข้อมูลในชุดต่าง ๆ ที่อาจส่งผลต่อความมั่นคงของโมเดล

d) ผลของการใช้ Model ในการทำนายข้อมูลโรคไตเรื้อรัง



รูปที่ 12 อธิบายกราฟ Actual vs. Predicted Values

กราฟแสดงให้เห็นว่าจุดส่วนใหญ่กระจายอยู่ใกล้เส้น Perfect Prediction Line แสดงว่าโมเดลมีประสิทธิภาพดีในการทำนาย จุดที่กระจายห่างไกลจากเส้นนั้นมาก อาจแสดงถึงข้อผิดพลาดในการทำนายหรือแสดงให้เห็นว่าข้อมูลบางส่วนอาจมีลักษณะเฉพาะที่โมเดลไม่สามารถจับได้ กราฟนี้ช่วยให้เห็นภาพรวมของการทำงานของโมเดลในการทำนายค่า eGFR สำหรับผู้ป่วย CKD โดยสามารถใช้วิเคราะห์ประสิทธิภาพของโมเดลได้เป็นอย่างดี

Feature Importance จากโมเดลช่วยในการระบุว่าฟีเจอร์ไหนมีบทบาทสำคัญต่อการทำนายค่า eGFR ในผู้ป่วยโรคไตเรื้อรัง สามารถช่วยในการพัฒนาแผนการดูแลผู้ป่วยและการป้องกันโรคไตได้อย่างมีประสิทธิภาพ ค่าความดันโลหิต ค่า Cr. ผู้ป่วยโรคเบาหวาน ค่า BMI ค่ารอบเอว ตามลำดับ

ความดันโลหิตสูงจะสัมพันธ์กับ eGFR ที่ต่ำลง อาจหมายความว่าผู้ป่วยที่มีความดันโลหิตสูงมีแนวโน้มที่จะมีปัญหาเกี่ยวกับการทำงานของไต

ค่า eGFR ที่ลดลงเมื่อค่า HbA1c สูงขึ้น โดยเฉพาะในกลุ่มผู้ที่มีโรคเบาหวาน ซึ่งบ่งบอกว่าระดับน้ำตาลที่ไม่สมดุลมีความสัมพันธ์กับการลดลงของการทำงานของไตซึ่งเน้นถึงความสำคัญของการควบคุมระดับน้ำตาลในเลือดเพื่อป้องกันการเสื่อมสภาพของไตในระยะยาว กลุ่มที่ไม่มีโรคเบาหวาน ค่าของ

eGFR จะค่อนข้างคงที่เมื่อเทียบกับกลุ่มที่มีโรคเบาหวาน

ความดันโลหิตและโรคเบาหวานเป็นปัจจัยที่สำคัญในการวิเคราะห์ความเสี่ยงของ CKD

ระดับคอเลสเตอรอลอาจไม่ใช่ตัวบ่งชี้หลักที่สัมพันธ์กับการทำงานของไต (eGFR) ค่า eGFR ของผู้ป่วยมีความผันแปรไปตามระดับคอเลสเตอรอล โดยไม่มีแนวโน้มที่ชัดเจนว่าระดับคอเลสเตอรอลที่สูงขึ้นจะสัมพันธ์กับการลดลงของ eGFR เสมอไป แต่ก็ยังเป็นปัจจัยที่ต้องได้รับการพิจารณาในการวิเคราะห์สุขภาพโดยรวม โดยเฉพาะอย่างยิ่งในผู้ป่วยที่มีโรคเบาหวาน การจัดการระดับคอเลสเตอรอลอาจเป็นส่วนหนึ่งในการป้องกันและรักษาการทำงานของไตในระยะยาว

ผู้ป่วยที่มีภาวะโรคเบาหวานและความดันโลหิตสูงมีแนวโน้มที่จะมีค่า eGFR ที่ต่ำลง ซึ่งบ่งชี้ว่าการทำงานของไตอาจถูกผลกระทบในทางลบจากภาวะโรคทั้งสองนี้ การเฝ้าระวังและดูแลการทำงานของไตในผู้ป่วยที่มีภาวะโรคเบาหวานและความดันโลหิตสูงจึงเป็นสิ่งสำคัญในการป้องกันและจัดการโรคไตเรื้อรัง

การสูบบุหรี่ไม่ได้มีผลกระทบที่ชัดเจนต่อการทำงานของไต (eGFR) เนื่องจากค่ากลางและช่วงของ eGFR ในกลุ่มที่สูบบุหรี่และไม่สูบบุหรี่มีความใกล้เคียงกัน

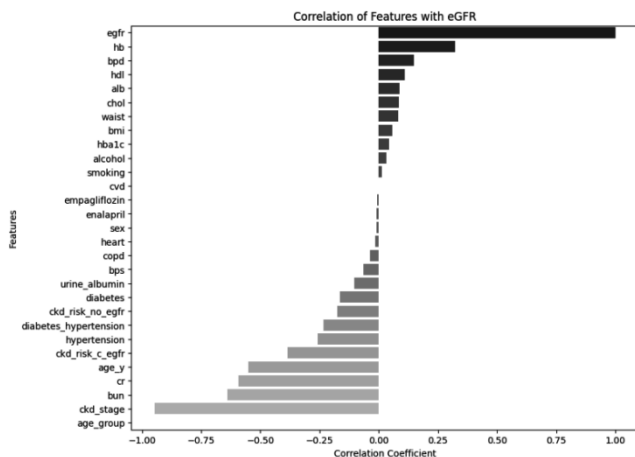
eGFR ในกลุ่มที่ดื่มแอลกอฮอล์มีการกระจายที่คล้ายคลึงกับกลุ่มที่ไม่ดื่ม แต่มีแนวโน้มว่าค่ากลางจะต่ำกว่าเล็กน้อย ซึ่งอาจบ่งบอกถึงความเสี่ยงในการทำงานของไตที่ลดลงในกลุ่มผู้ที่ดื่มแอลกอฮอล์

eGFR มีแนวโน้มลดลงตามอายุที่เพิ่มขึ้น แสดงถึงการเสื่อมสภาพของการทำงานของไตในกลุ่มผู้สูงอายุ การติดตามและประเมินการทำงานของไตในผู้สูงอายุจึงเป็นสิ่งสำคัญในการป้องกันและจัดการโรคไตเรื้อรังอย่างมีประสิทธิภาพ

เพศหญิงและเพศชายมีการกระจายตัวของค่า eGFR ที่ใกล้เคียงกัน ซึ่งแสดงให้เห็นว่าเพศอาจไม่ใช่ปัจจัยหลักที่ส่งผลต่อการทำงานของไต (eGFR) อย่างไรก็ตาม ควรพิจารณาพร้อมกับปัจจัยอื่น ๆ เช่น อายุ โรคเบาหวาน และความดันโลหิตสูง เพื่อประเมินความเสี่ยงและการทำงานของไตอย่างครอบคลุม Hemoglobin (hb) และ eGFR ไม่มีความสัมพันธ์ที่ชัดเจนในทิศทางที่สอดคล้องกัน ซึ่งแสดงว่า Hemoglobin อาจไม่เป็นตัวทำนายที่สำคัญสำหรับการเปลี่ยนแปลงของ eGFR อย่างไรก็ตาม การติดตามระดับ Hemoglobin ยังคงมีความสำคัญในการประเมินสุขภาพโดยรวมของผู้ป่วยโรคไตเรื้อรัง (CKD) โดยเฉพาะในกรณีที่มีภาวะโลหิตจางที่เกิดร่วมกับโรคไต

ค่า BUN และ Creatinine มีความสัมพันธ์เชิงลบกับค่า eGFR ซึ่งแสดงให้เห็นว่าระดับที่สูงขึ้นของทั้ง BUN และ Creatinine เป็นสัญญาณที่ชัดเจนของการเสื่อมสภาพของการทำงานของไต

การใช้ยา Enalapril และการไม่ใช้ยาไม่มีความแตกต่างที่ชัดเจนในด้านค่าการกรองของไต (eGFR) ดังนั้น การใช้หรือไม่ใช้ยา Enalapril อาจไม่เป็นตัวแปรที่มีผลกระทบต่อ eGFR



รูปที่ 13 เปรียบเทียบความสัมพันธ์ของ Features with eGFR

คุณลักษณะ age_group และ ckd_stage มีความสัมพันธ์เชิงลบที่แข็งแกร่งกับค่า eGFR ซึ่งบ่งบอกว่าการเสื่อมสภาพของไตมีแนวโน้มเพิ่มขึ้นเมื่ออายุและ stage ของ CKD เพิ่มขึ้น

ค่า hb (Hemoglobin) มีความสัมพันธ์เชิงบวกกับ eGFR ซึ่งแสดงว่าค่า Hemoglobin ที่สูงสัมพันธ์กับการทำงานของไตที่ดีขึ้น

คุณลักษณะอื่น ๆ เช่น creatinine (cr) และ BUN ยังมีความสัมพันธ์เชิงลบที่สอดคล้องกับการทำงานของไตที่ลดลง กราฟนี้ช่วยให้เราเห็นภาพรวมของปัจจัยที่อาจส่งผลกระทบต่อ eGFR และความเสี่ยงของโรคไตเรื้อรัง (CKD) ซึ่งสามารถนำไปใช้ในการวางแผนการดูแลและป้องกันโรคไตได้อย่างมีประสิทธิภาพมากขึ้น

สรุปผลการศึกษา

การศึกษาในครั้งนี้ได้วิเคราะห์และทำนายค่าการกรองของไต (eGFR) เพื่อประเมินความเสี่ยงของโรคไตเรื้อรัง (Chronic Kidney Disease, CKD) โดยใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) หลากหลายประเภท ได้แก่ Linear Regression, Random Forest, Decision Tree, Gradient Boosting Machine (GBM), XGBoost และ Light GBM พร้อมทั้งประเมินประสิทธิภาพของแต่ละโมเดลด้วยการวัดค่า Mean Absolute Error (MAE), Mean Squared Error (MSE) และค่า R^2

จากการวิเคราะห์ผลลัพธ์ พบว่าโมเดลที่มีประสิทธิภาพสูงสุดในการทำนาย eGFR คือ Random Forest ซึ่งให้ค่า MAE เท่ากับ 0.146688, MSE เท่ากับ 0.260079 และ

ค่า R^2 เท่ากับ 0.999688 ซึ่งแสดงถึงความแม่นยำสูงสุดในการทำนาย โดยค่า R^2 ที่ใกล้เคียง 1 บ่งบอกว่าโมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้อย่างดีเยี่ยม นอกจากนี้การปรับปรุงโมเดลด้วยการใช้ Grid Search ยังช่วยให้ Random Forest มีพารามิเตอร์ที่เหมาะสมยิ่งขึ้น เป็นการใช้ Machine Learning ในการทำนายค่า eGFR เพื่อประเมินความเสี่ยงโรคไตเรื้อรัง โดยโมเดล Random Forest มีประสิทธิภาพสูงสุด สามารถนำไปประยุกต์ใช้ในโรงพยาบาลเพื่อป้องกันโรคไตเรื้อรังได้

การวิเคราะห์ความสัมพันธ์ของปัจจัยต่าง ๆ ต่อค่า eGFR แสดงให้เห็นว่า "ความดันโลหิต" และ "โรคเบาหวาน" เป็นปัจจัยสำคัญที่สัมพันธ์กับการทำงานของไตในผู้ป่วยโรคไตเรื้อรัง นอกจากนี้ ปัจจัยอื่น ๆ เช่น ค่า BUN, Creatinine (Cr), BMI, รอบเอว และอายุยังมียบทบาทสำคัญต่อการเปลี่ยนแปลงของ eGFR ด้วย

การค้นพบนี้สามารถนำไปประยุกต์ใช้ในการประเมินความเสี่ยงของโรคไตเรื้อรังและสนับสนุนการวางแผนการดูแลผู้ป่วยอย่างมีประสิทธิภาพยิ่งขึ้น ช่วยให้ผู้เชี่ยวชาญทางการแพทย์สามารถระบุและป้องกันการเกิดโรคไตเรื้อรังได้อย่างทันท่วงที แนะนำการใช้โมเดลแบบผสมผสาน เช่น ใช้ Random Forest ผสมกับโมเดลอื่นๆ เช่น Neural Networks หรือ XGBoost อาจเพิ่มความแม่นยำในการพยากรณ์ โดยเฉพาะในกรณีที่มีข้อมูลมีความซับซ้อนหรือผันผวนสูง

REFERENCES

- [1] B. P. Ghosh, T. Imam, N. Anjum, M. T. Mia, C. U. Siddiqua, K. Shaharia, Sharif, M. A. I. Mamun, M. Z. Hossain, Md Atikul, and I. Mamun, "Advancing Chronic Kidney Disease Prediction: Comparative Analysis of Machine Learning Algorithms and a Hybrid Model," *Journal of Computer Science and Technology*, pp. 989-993, Jul. 2024
- [2] C. Mbah, A. Johnson, L. Williams, and D. Smith, "Association between Chronic Kidney Disease and Cardiovascular Risk Factors in Diverse Populations," *Journal of Nephrology Research*, vol. 15, no. 2, pp. 123-134, 2023
- [3] D. Vanathi, S. M. Ramesh, K. Tamizharasu, S. N. and K. P., "A Machine Learning Perspective for Predicting Chronic Kidney Disease," *Proceedings*, pp. 989-993, Jul. 2024.
- [4] D. Xie et al., "Global burden and influencing factors of chronic kidney disease due to type 2 diabetes in adults aged 20-59 years, 1990-2019," *Dental Science Reports*, vol. 13, no. 1, pp. 20234-20, Nov. 2023
- [5] E. Mahbub, R. K. Sarker, M. Hasan, and F. Ahmed, "Application of Machine Learning in Predicting Chronic Kidney Disease: A Comprehensive Study," *Healthcare Informatics Research*, vol. 28, no. 4, pp. 315-325, Dec. 2022
- [6] M. Tonelli, V. Nkunu, C. Varghese, A. K. Abu-Alfa, M. N. Alrukhaimi, L. Fox, J. Gill, D. C. H. Harris, F. F. Hou, P. J. O'Connell, H. U. Rashid, A. Niang, S. Ossareh, V. Tesar, E. Zakharova, and C.-W. Yang, "Framework for establishing integrated kidney care programs in low- and middle-income countries," *Kidney International Supplements*, vol. 10, no. 1, pp. e19-e23, 2020.
- [7] M. Rane, M. Derkar, D. Kabra, and T. J. Desai, "Chronic Kidney Disease Prediction Using Machine Learning," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 12, no. 6, pp. 65-70, Jun. 2024
- [8] M. Forň, S. L. Přez-Garća, J. M. Betriu, and L. Fernandez, "Cardiovascular Risk Assessment in Patients with Chronic Kidney Disease: A Comprehensive Review," *Nephrology Dialysis Transplantation*, vol. 34, no. 12, pp. 1915-1923, Dec. 2019.
- [9] N. A. Kikvadze and G. Gorgadze, "The role of environmental factors in the development of chronic kidney disease (CKD) – Literature Review," *Junior Researchers*, vol. 7, no. 1, pp. 1-8, Jun. 2023.
- [10] T. Bai, Y. Zhang, L. Wang, and J. Liu, "Machine Learning Models for Predicting the Progression of Chronic Kidney Disease," *Journal of Medical Systems*, vol. 46, no. 11, pp. 1-10, Nov. 2022.
- [11] T. R. Noviandy, G. M. Idroes, M. Syukri, and R. Idroes, "Interpretable Machine Learning for Chronic Kidney Disease Diagnosis: A Gaussian Processes Approach," *Indonesian Journal of Case Reports*, vol. 2, no. 1, pp. 24-32, Jun. 2024.
- [12] U. O. Pontes, T. L. M. Amaral, C. A. Amaral, and G. T. R. Monteiro, "Prevalence and factors associated with chronic kidney disease in diabetic and hypertensive patients," *Environmental Health Perspectives*, 2023
- [13] กมลทิพย์ วิจิตรสุนทรกุล, "แนวทางการดูแลและจัดการผู้ป่วยโรคไตเรื้อรัง," *กรมควบคุมโรค, กระทรวงสาธารณสุข, สืบค้นเมื่อ 30 กันยายน 2566*. [ออนไลน์]. เข้าถึงได้จาก: <https://ddc.moph.go.th/uploads/pulish/1308820220905025852.pdf>