



Binomial Logistic Regression (Parkinson's disease)

จัดทำโดย

เลขที่ 61	รหัสสถิติ 66365325	นายสิรภพ วิเศษศรี
เลขที่ 62	รหัสสถิติ 66365417	นางสาวสุชิราภรณ์ เปานาเรียง
เลขที่ 63	รหัสสถิติ 66365486	นางสาวสุพรรณษา ถิระพันธ์
เลขที่ 64	รหัสสถิติ 66365493	นางสาวสุพรรณษา โฉมหาญ
เลขที่ 66	รหัสสถิติ 66365592	นางสาวหยกผกา บวบหอม

เสนอ

ผศ.ดร. ศิริพร เดชะศิลารักษ์

รายงานนี้เป็นส่วนหนึ่งของรายวิชา 305231 Applied Statistics

คณะวิศวกรรมศาสตร์ สาขาวิชาวิศวกรรมคอมพิวเตอร์

ภาคเรียนที่ 2 ปีการศึกษา 2567

มหาวิทยาลัยนเรศวร

คำนำ

รายงานนี้เป็นส่วนหนึ่งของรายวิชา 305231 Applied Statistics จัดทำขึ้นโดยมีวัตถุประสงค์เพื่อศึกษาและเลือกเนื้อหา Dataset และอัลกอริทึมที่เหมาะสมที่จะนำมาใช้ในการสร้าง model และนำเสนอข้อมูลออกมาโดยใช้หลักการของ Data Science ทั้ง 5 steps และสรุปผลข้อมูล

ทั้งนี้ คณะผู้จัดทำหวังเป็นอย่างยิ่งว่าข้อมูลที่นำเสนอออกมาจะเป็นประโยชน์ต่อผู้ที่สนใจไม่มากนักน้อย หากรายงานนี้มีข้อผิดพลาดประการใดทางคณะผู้จัดทำต้องขออภัยมา ณ ที่นี้ด้วย

คณะผู้จัดทำ

สารบัญ

เรื่อง	หน้า
คำนำ	ก
สารบัญ	ข
ที่มาและความสำคัญของปัญหา	1
หลักการและทฤษฎีที่เกี่ยวข้อง	2
ขั้นตอนการแก้โจทย์ปัญหาตามหลักการของ Data Science	3
Step 1: Dataset	3-5
Step 2: Data preparation , Data visualization	5-7
Step 3: Model Building	8
Step 4: Results and Evaluation	8
Step 5: Deployment	9
สรุปผลและวิจารณ์ผล (Discussion)	10
แหล่งที่มาและเอกสารอ้างอิง	11

ที่มาและความสำคัญของปัญหา

การวินิจฉัยโรคพาร์กินสันในระยะเริ่มต้นเป็นเรื่องที่ทำหาย เนื่องจากอาการอาจไม่ชัดเจนและคล้ายคลึงกับโรคอื่นๆ และการวินิจฉัยมักต้องพึ่งแพทย์ผู้เชี่ยวชาญ ซึ่งอาจใช้เวลานาน การใช้ Machine Learning และการวิเคราะห์ข้อมูลเสียงพูดสามารถช่วยในการวินิจฉัยโรคได้อย่างแม่นยำและรวดเร็วขึ้น หากสามารถพัฒนาโมเดลที่มีประสิทธิภาพ จะช่วยลดภาระของบุคลากรทางการแพทย์ และเพิ่มโอกาสให้ผู้ป่วยได้รับการดูแลที่เร็วขึ้น

1. การวินิจฉัยโรคพาร์กินสันผ่านเสียงพูด

ชุดข้อมูลนี้มีวัตถุประสงค์เพื่อช่วยแยกแยะผู้ป่วยพาร์กินสันออกจากบุคคลปกติ โดยใช้คุณลักษณะที่เกี่ยวข้องกับเสียงพูด เช่น jitter , shimmer และ harmonic-to-noise ratio (HNR) ซึ่งมีแนวโน้มเปลี่ยนแปลงในผู้ป่วยพาร์กินสัน

2. การพัฒนาโมเดล Machine Learning สำหรับการแพทย์

ชุดข้อมูลนี้เป็นที่นิยมในการนำไปใช้พัฒนาโมเดล Machine Learning และ AI สำหรับช่วยในการวินิจฉัยและติดตามอาการของโรคพาร์กินสัน ซึ่งเป็นโรคทางระบบประสาทที่พบได้บ่อยในผู้สูงอายุ

3. การประยุกต์ใช้กับระบบ Telemonitoring

งานวิจัยที่ใช้ชุดข้อมูลนี้แสดงให้เห็นว่า การตรวจสอบอาการของผู้ป่วยผ่านเสียงพูดสามารถช่วยให้แพทย์เฝ้าติดตามอาการของโรคพาร์กินสันได้จากระยะไกล โดยไม่ต้องให้ผู้ป่วยเดินทางมาพบแพทย์บ่อยครั้ง

การนำสถิติประยุกต์ (Applied Statistics) มาใช้ในการแก้ไขปัญหา

นำสถิติประยุกต์มาช่วยในการวิเคราะห์ข้อมูลจาก dataset ที่ได้ทั้งหมดตามหลักการ Binomial Logistic Regression และ Data Science ทั้ง 5 steps

หลักการและทฤษฎีที่เกี่ยวข้อง

ใช้ทฤษฎี Binomial Logistic Regression ในการจำแนกผู้ป่วยโรคพาร์กินสัน โดยจะเป็นการจำแนกข้อมูลออกเป็นสองกลุ่ม (Binary Classification) โดยที่ตัวแปรเป้าหมายมีค่าเป็น 0 หรือ 1 โดยมีหลักการดังนี้

1. การสำรวจข้อมูล (Exploratory Data Analysis - EDA)

การวิเคราะห์การกระจายตัวของข้อมูล : ใช้สถิติเช่น ค่าเฉลี่ย (Mean) , ค่ามัธยฐาน (Median) , ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) เพื่อสรุปการกระจายตัวของข้อมูลทุก Feature

2. การเตรียมข้อมูล (Data Preparation)

มีการจัดการกับข้อมูลที่หายไป (Missing Values) การกำจัด Outliers โดยการใช้ Interquartile Range (IQR) การแปลงข้อมูล (Feature Scaling) โดยการใช้ Standardization (Z-score Normalization) เพื่อปรับสเกลของข้อมูลให้เหมาะสม และมีการแยกข้อมูลออกเป็นชุดข้อมูล train และ test

3. การฝึกและสร้างโมเดล (Model Building and Training)

สร้างโมเดล Binomial Logistic Regression และสร้าง Simple linear regression in each 2 class ด้วยชุดข้อมูลผ่านการ Training แล้ว

4. การประเมินประสิทธิภาพโมเดล (Model Evaluation)

ประเมินประสิทธิภาพโมเดลโดยการทำ Confusion Matrix และดูค่า Accuracy , Precision , Recall และ F1-score

5. การนำโมเดลไปใช้งาน (Deployment)

นำโมเดลที่ผ่านการประเมินประสิทธิภาพแล้ว มาใช้ทำนายว่าบุคคลใดมีโอกาสที่จะเป็นโรคพาร์กินสัน

ขั้นตอนการแก้โจทย์ปัญหาตามหลักการของ Data Science

Step 1 : Dataset

รายละเอียดของชุดข้อมูล (Dataset Details) นำมาจาก UC Irvine Machine Learning Repository ประกอบไปด้วย จำนวนตัวอย่าง (Instances) : 195 ตัวอย่าง , จำนวนคุณลักษณะ (Features) : 23 คุณลักษณะ ที่เกี่ยวข้องกับการวิเคราะห์เสียง และมีทั้งหมด 24 columns โดยมีรายละเอียดของแต่ละ attribute ดังนี้ :

1. name

รหัสหรือชื่อของตัวอย่าง (Subject Identifier) ซึ่งใช้สำหรับแยกแยะตัวอย่างแต่ละรายการ แต่ไม่ใช่เป็น feature ในการวิเคราะห์

2. MDVP:Fo(Hz)

ค่าเฉลี่ยของความถี่พื้นฐาน (Fundamental Frequency) ของเสียงที่ถูกวัดจากผู้เข้าร่วม (เสียงพูด)

3. MDVP:Fhi(Hz)

ค่าสูงสุดของความถี่พื้นฐานที่ถูกบันทึกในช่วงการออกเสียง

4. MDVP:Flo(Hz)

ค่าต่ำสุดของความถี่พื้นฐานที่ถูกบันทึกในช่วงการออกเสียง

5. MDVP:Jitter(%)

ค่าวัดความผันผวนของช่วงเวลาการสั่น (Jitter) โดยแสดงในรูปเปอร์เซ็นต์ เพื่อบ่งบอกถึงความไม่เสถียรในการสั่นของเส้นเสียง

6. MDVP:Jitter(Abs)

การวัดความผันผวนของช่วงเวลาการสั่นในรูปแบบค่าสัมบูรณ์

7. MDVP:RAP

ค่าที่ใช้วัดความผันผวนของเสียงในช่วงเวลาสั้น ๆ โดยการวัดอัตราส่วนการเปลี่ยนแปลงของรอบสั้นในลักษณะเฉพาะ

8. MDVP:PPQ

การคำนวณค่า jitter จากค่าเฉลี่ยของช่วงเวลาการสั่นในหลายๆ รอบ ช่วยให้เห็นภาพรวมของความผันผวน

9. Jitter:DDP

อีกรูปแบบหนึ่งของการวัด jitter ที่เน้นการเปลี่ยนแปลงระหว่างจุดในสัญญาณเสียง

10. MDVP:Shimmer

ค่าวัดความผันผวนของความดังของเสียง (Amplitude Variation) ที่บ่งบอกถึงความไม่สม่ำเสมอของระดับเสียง

11. MDVP:Shimmer(dB)

การวัด shimmer โดยใช้หน่วย dB ซึ่งช่วยให้เห็นความแตกต่างในระดับความดังของเสียงได้ชัดเจนขึ้น

12. Shimmer:APQ3

มาตรการความผันผวนของความดังในระดับสั้น (3 รอบ) ที่ใช้เปรียบเทียบค่าความดังในช่วงเวลาสั้น

13. Shimmer:APQ5

มาตรการความผันผวนของความดังในระดับ 5 รอบ ซึ่งให้ภาพรวมของความไม่เสถียรในความดังของเสียง

14. MDVP:APQ

การวัดค่าเฉลี่ยของความผันผวนในความดังของเสียงโดยรวม

15. Shimmer:DDA

มาตรการอีกตัวหนึ่งสำหรับวัดความแปรปรวนของความดังเสียงในรูปแบบที่แตกต่างจากวิธีอื่น ๆ

16. NHR (Noise-to-Harmonics Ratio)

อัตราส่วนระหว่างเสียงรบกวน (noise) กับเสียงที่มีลักษณะเป็น harmonic ซึ่งช่วยบ่งบอกถึงความชัดของเสียง

17. HNR (Harmonics-to-Noise Ratio)

อัตราส่วนระหว่างส่วนของ harmonic ต่อเสียงรบกวน ยิ่งค่าสูงหมายถึงเสียงมีความชัดเจนและมี harmonic มากกว่า noise

18. RPDE (Recurrence Period Density Entropy)

มาตรการความไม่แน่นอน (entropy) ในการปรากฏซ้ำของรูปแบบช่วงเวลาการสั่นในสัญญาณเสียง

19. DFA (Detrended Fluctuation Analysis)

เทคนิควัดลักษณะ self-similarity ของสัญญาณเสียงที่ช่วยให้เห็นความสัมพันธ์ระยะยาวของข้อมูล

20. spread1

มาตรการที่วัดความแปรปรวนของข้อมูลเสียงในแง่มุมที่อาจมาจากการวิเคราะห์แบบ nonlinear

21. spread2

อีกตัวชี้วัดหนึ่งที่เกี่ยวข้องกับความแปรปรวนหรือการกระจายของข้อมูลเสียงในมิติที่แตกต่าง

22. D2 (Correlation Dimension)

ค่าที่วัดความซับซ้อนของสัญญาณเสียง โดยพิจารณาจากมิติของข้อมูลในเชิงฟร็กทัล

23. PPE (Pitch Period Entropy)

มาตรการความไม่แน่นอนในช่วงของ pitch period ซึ่งแสดงถึงความแปรปรวนและความไม่สม่ำเสมอของช่วงเวลาการสั่น

24. status

ตัวแปรเป้าหมาย (Target variable) สำหรับการจำแนกประเภท

ค่า 1 ระบุว่าเป็นผู้ป่วยโรค Parkinsons

ค่า 0 ระบุว่าเป็นบุคคลปกติ (Healthy)

Step 2 : Data preparation , Data visualization

การจัดการข้อมูลสูญหาย (Missing values)

มีการจัดการข้อมูลสูญหายในแต่ละคอลัมน์โดยใช้ `df2.isnull().sum()` พบว่าข้อมูล Dataset นี้ไม่มีข้อมูลใดสูญหาย

	0
MDVP:F0(Hz)	0
MDVP:Fhi(Hz)	0
MDVP:Flo(Hz)	0
MDVP:Jitter(%)	0
MDVP:Jitter(Abs)	0
MDVP:RAP	0
MDVP:PPQ	0
Jitter:DDP	0

MDVP:Shimmer	0
MDVP:Shimmer(dB)	0
Shimmer:APQ3	0
Shimmer:APQ5	0
MDVP:APQ	0
Shimmer:DDA	0
NHR	0
HNR	0

RPDE	0
DFA	0
spread1	0
spread2	0
D2	0
PPE	0
status	0

การเปลี่ยนชนิดข้อมูลให้เหมาะสม

ทำการ info เพื่อดูสรุปเกี่ยวกับ DataFrame พบว่าคอลัมน์ status ซึ่งเอาไว้แสดงสถานะการเป็นโรค Parkinsons มี type เป็น int64 จึงทำการเปลี่ยน type เป็น category จากนั้นขยับคอลัมน์ status ไปไว้ที่ท้ายสุดของ Dataframe และทำการลบคอลัมน์ name ออก ดังนี้

Before				After			
#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype
0	name	195 non-null	object	0	MDVP:F0(Hz)	195 non-null	float64
1	MDVP:F0(Hz)	195 non-null	float64	1	MDVP:F1(Hz)	195 non-null	float64
2	MDVP:F1(Hz)	195 non-null	float64	2	MDVP:F2(Hz)	195 non-null	float64
3	MDVP:F2(Hz)	195 non-null	float64	3	MDVP:F3(Hz)	195 non-null	float64
4	MDVP:F3(Hz)	195 non-null	float64	4	MDVP:F4(Hz)	195 non-null	float64
5	MDVP:F4(Hz)	195 non-null	float64	5	MDVP:F5(Hz)	195 non-null	float64
6	MDVP:F5(Hz)	195 non-null	float64	6	MDVP:F6(Hz)	195 non-null	float64
7	MDVP:F6(Hz)	195 non-null	float64	7	MDVP:F7(Hz)	195 non-null	float64
8	MDVP:F7(Hz)	195 non-null	float64	8	MDVP:F8(Hz)	195 non-null	float64
9	MDVP:F8(Hz)	195 non-null	float64	9	MDVP:F9(Hz)	195 non-null	float64
10	MDVP:F9(Hz)	195 non-null	float64	10	MDVP:F10(Hz)	195 non-null	float64
11	MDVP:F10(Hz)	195 non-null	float64	11	MDVP:F11(Hz)	195 non-null	float64
12	MDVP:F11(Hz)	195 non-null	float64	12	MDVP:F12(Hz)	195 non-null	float64
13	MDVP:F12(Hz)	195 non-null	float64	13	MDVP:F13(Hz)	195 non-null	float64
14	MDVP:F13(Hz)	195 non-null	float64	14	MDVP:F14(Hz)	195 non-null	float64
15	MDVP:F14(Hz)	195 non-null	float64	15	MDVP:F15(Hz)	195 non-null	float64
16	MDVP:F15(Hz)	195 non-null	float64	16	MDVP:F16(Hz)	195 non-null	float64
17	MDVP:F16(Hz)	195 non-null	float64	17	MDVP:F17(Hz)	195 non-null	float64
18	MDVP:F17(Hz)	195 non-null	float64	18	MDVP:F18(Hz)	195 non-null	float64
19	MDVP:F18(Hz)	195 non-null	float64	19	MDVP:F19(Hz)	195 non-null	float64
20	MDVP:F19(Hz)	195 non-null	float64	20	MDVP:F20(Hz)	195 non-null	float64
21	MDVP:F20(Hz)	195 non-null	float64	21	MDVP:F21(Hz)	195 non-null	float64
22	MDVP:F21(Hz)	195 non-null	float64	22	MDVP:F22(Hz)	195 non-null	float64
23	MDVP:F22(Hz)	195 non-null	float64	23	MDVP:F23(Hz)	195 non-null	float64

การแปลงข้อมูลประเภทกลุ่ม (Categorical data) เป็นตัวเลข (Numeric data)

เนื่องจาก Dataset ที่ได้อามีเพียงคอลัมน์ status ที่แต่เดิมเป็น Numeric data จึงได้ทำการเปลี่ยนเป็น Categorical data แล้ว ส่วนการแปลงข้อมูลประเภทกลุ่ม (Categorical data) เป็นตัวเลข (Numeric data) นั้นจึงไม่จำเป็น

การเลือกลักษณะที่สำคัญของข้อมูลหรือใช้ทั้งหมดในการสร้าง model (Feature selection)

Model นี้ใช้ Feature ทั้งหมดในการสร้าง model เนื่องจากทุก Feature มีความสำคัญกับข้อมูลทั้งหมด จึงไม่ได้มีการตัด Feature ใดออก

การแปลงข้อมูลให้อยู่ในรูปมาตรฐาน (Feature scaling)

มีการแปลงข้อมูลให้อยู่ในรูปมาตรฐานเดียวกัน โดยใช้การแปลงเป็น z-score เพื่อให้สามารถนำ Dataset ไปใช้คำนวณ model ให้ได้ค่าที่มีความแม่นยำมากขึ้น

Step 3 : Model Building

การเลือก model ที่ดีที่สุดสำหรับ dataset

การเลือก model ที่ดีที่สุดสำหรับ dataset นี้คือ Model Binomial Logistic Regression โดยจะแยก feature กับ target และใช้ feature ทั้งหมดในการสร้าง model ทำการ Train ข้อมูลและสร้าง Simple linear regression in each 2 class ได้ดังนี้

Simple linear ----> $1.9445 + (-0.1561)x_1 + (-0.1520)x_2 + 0.1998x_3 + (-0.7434)x_4 + (-0.5465)x_5 + 0.3395x_6 + (-0.2935)x_7 + 0.3394x_8 + 0.0342x_9 + 0.3408x_{10} + (-0.1322)x_{11} + 0.1418x_{12} + 0.2364x_{13} + (-0.1329)x_{14} + 1.2220x_{15} + 0.0650x_{16} + (-0.3407)x_{17} + 0.1524x_{18} + 0.5907x_{19} + 0.1862x_{20} + 0.4162x_{21} + 0.6748x_{22}$

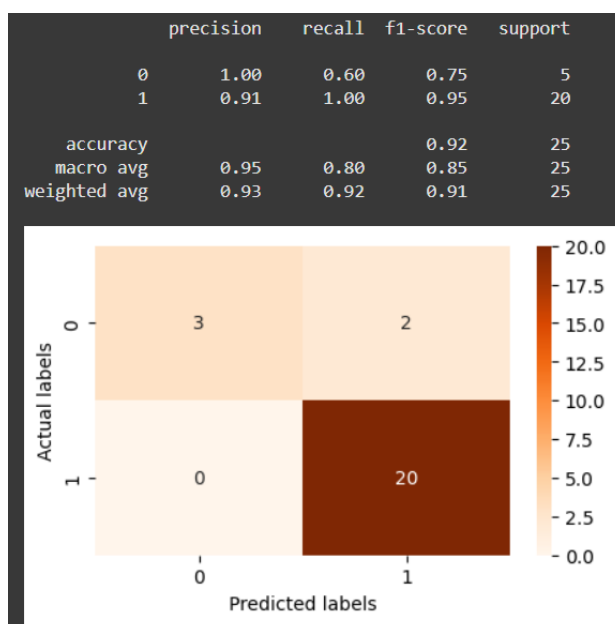
การปรับพารามิเตอร์ของ model ที่เหมาะสม

ไม่มีการปรับพารามิเตอร์ของ model เนื่องจาก model ไม่ได้มีความซับซ้อนมากนัก ทำให้ได้ model ที่เหมาะสมและมีประสิทธิภาพแล้ว

Step 4 : Results and Evaluation

การเลือกตัววัดให้เหมาะสมกับงานด้าน regression , classification หรือ clustering

จาก model นี้ได้เลือกใช้ตัววัด classification โดยมีการทำนายผลลัพธ์แบบคลาสใช้ `model.predict(x_test)` และมีการทำนายผลลัพธ์แบบค่าความน่าจะเป็น (probability) ใช้ `model.predict_proba(x_test)` โดยค่า 0 จะเท่ากับไม่เป็นโรค Parkinsons และค่า 1 จะเท่ากับเป็นโรค Parkinsons จากนั้นนำข้อมูลที่ได้อมา Updated dataframe for `x_test` และทำ Classification report โดยการสร้าง Confusion matrix ได้ข้อมูลออกมาดังนี้



Step 5 : Deployment

ก่อนนำ model ไปใช้งานได้มีการประเมินประสิทธิภาพของ model โดยการดูค่า Accuracy , Precision , Recall และ F1 Score พบว่าได้ค่าดังนี้

- Accuracy : 0.9200
- Precision : 0.9545
- Recall : 0.8000
- F1 Score : 0.8512

จะเห็นได้ว่า model นี้มีประสิทธิภาพมาก เนื่องจากค่า Accuracy มีความถูกต้องถึง 92 % และค่า Precision , Recall , F1 Score ก็มีค่าที่สูง โดยรวมแล้ว model จึงมีการทำนายที่มีความแม่นยำสูง และรักษาสมดุลระหว่าง Precision กับ Recall ได้ดี ทำให้ model มีประสิทธิภาพมากที่จะนำไปใช้งานต่อไป

การนำ model ไปใช้งาน

เมื่อได้ model มีประสิทธิภาพมาแล้ว จะนำ model มาใช้ทำนายว่าบุคคลใดมีโอกาสที่จะเป็นโรคพาร์กินสันบ้าง และ model ที่ได้นี้ยังสามารถใช้เป็นตัวช่วยของแพทย์ในการใช้วินิจฉัยโรคพาร์กินสันได้ง่ายและแม่นยำได้ยิ่งขึ้น

สรุปผลและวิจารณ์ผล (Discussion)

การใช้โมเดล Binomial Logistic Regression เป็นเครื่องมือในการทำนายโรคพาร์กินสันนั้นเป็นแนวทางที่มีประโยชน์และเป็นที่ยอมรับในวงการวินิจฉัยทางการแพทย์ ถูกใช้ในการคาดการณ์ความน่าจะเป็นของโรคพาร์กินสันในระยะเริ่มต้น โดยอาศัยตัวแปรต่างๆที่มีอยู่ในชุดข้อมูล เช่น ค่าเฉลี่ยของความถี่เสียงพูดที่วัดจากผู้เข้าร่วม , การสั่นของเสียง (Jitter,Shimmer) , ระดับเสียงพื้นฐาน (Fundamental Frequency,F0) , ความเข้มของเสียง (Intensity) แม้ว่าอาจมีข้อจำกัดเกี่ยวกับคุณภาพของข้อมูลที่ใช้ในแบบจำลองเช่น ความซับซ้อนของโรคพาร์กินสันคือโรคนี้มีหลายปัจจัยที่เกี่ยวข้องไม่ได้ขึ้นอยู่กับข้อมูลเสียงพูดเพียงอย่างเดียว หรือความแม่นยำของโมเดลหากไม่มีการเลือกตัวแปรที่เหมาะสมหรือการทำความสะอาดข้อมูลที่ดี ที่อาจทำให้โมเดลมีประสิทธิภาพต่ำได้เป็นต้น แต่อย่างไรก็ตาม การสร้างโมเดล Binomial Logistic Regression ครั้งนี้ได้มีการทดสอบประสิทธิภาพในการทำนายผลไปเบื้องต้นแล้ว พบว่ามีประสิทธิภาพค่อนข้างดีมาก จึงสามารถนำโมเดลที่ได้นี้มาใช้จำแนกบุคคลที่มีโอกาสเสี่ยงเป็นโรคพาร์กินสันในระยะเริ่มต้นได้ โดยอ้างอิงจากค่าความน่าจะเป็นที่ได้จากแบบจำลองโมเดล Binomial Logistic Regression เพื่อช่วยให้แพทย์สามารถวินิจฉัยโรคพาร์กินสันได้ง่ายยิ่งขึ้น

แหล่งที่มาและเอกสารอ้างอิง

UCI Machine Learning Repository. (n.d.). Parkinson's dataset. สืบค้นเมื่อ 3 มีนาคม 2568.

จาก <https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data>

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009).

Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015-1022.