

หน้า 1

Hello everyone,

Today, I'll share the progress of my project titled "**Machine Learning Models for Head Injury Survival Prediction in Motorcycle Accidents.**"

I will cover data preparation, model selection, and initial evaluation results.

สวัสดีครับ

วันนี้จะมาอัปเดตความคืบหน้าของโปรเจกต์ "Machine Learning Models for Head Injury Survival Prediction in Motorcycle Accidents"

ในสัปดาห์นี้ จะมาพูดถึงการเตรียมข้อมูล, การเลือกโมเดล, และผลการประเมินเบื้องต้น

หน้า 2

According to the plan for Week 7, we are still in the phase of refining the models and have started writing the manuscript.

ตามแผนสำหรับสัปดาห์ที่ 7, เรายังคงอยู่ในขั้นตอนของการปรับปรุงโมเดล และเริ่มเขียนต้นฉบับ

หน้า 3

In the next step, we will modify the data in the "**Dead**" column. Originally, the value **1** represented deceased and **0** represented survived.

However, since our model focuses on predicting the probability of **survival**, we have reversed these values as follows:

- **1 now indicates survival**

- **0 now indicates death**

This change will help make the model training and interpretation clearer and better aligned with the research objectives.

ในขั้นตอนถัดไป, เราจะแก้ไขข้อมูลในคอลัมน์ "**Dead**". เดิมที, ค่า **1** แสดงถึงผู้เสียชีวิต และ **0** แสดงถึงผู้รอดชีวิต

อย่างไรก็ตาม, เนื่องจากโมเดลของเราเน้นการพยากรณ์ความน่าจะเป็นของการ **รอดชีวิต**, เราจึงได้สลับค่าเหล่านี้ดังต่อไปนี้:

• **1** ตอนนี้บ่งชี้การรอดชีวิต

• **0** ตอนนี้บ่งชี้การเสียชีวิต

- วัน นาว อินดิเคเตอร์ เซอร์ไวฟัล

- ซีโร นาว อินดิเคเตอร์ เดธ

การเปลี่ยนแปลงนี้จะช่วยให้การฝึกโมเดลและการตีความผลลัพธ์ชัดเจนขึ้น และสอดคล้องกับวัตถุประสงค์ของการวิจัย

Next, we will show the data **before and after the modification** of the values in the "Dead" column.

On the left side, the original data is displayed, where **1 indicates deceased** and **0 indicates survived**.

On the right side, the modified data is shown, where the values have been reversed:

- **1 now represents survived**
- **0 now represents deceased**

This change aligns with the model's focus on predicting survival.

เน็กซ์, วี วิล โชว์ เดอะ เดต้า บีฟอร์ แอนด์ แอฟเตอร์ เดอะ โมดิฟิเคชัน ออฟ เดอะ แวลูส อิน เดอะ "เดด" คอลัมน์

ออน เดอะ เลฟท์ไซด์, เดอะ ออริจินัล เดต้า อีส ดิสเพลย์ด, แวร์ วัน อินดิเคตส์ ดีเซสส์ แอนด์ ซีโร อินดิเคตส์ เซอร์ไวฟด์

ออน เดอะ ไรท์ไซด์, เดอะ โมดิฟายด์ เดต้า อีส โชว์น, แวร์ เดอะ แวลูส แฮฟ บีน รีเวิร์สด:

- วัน นาว รีเพรเซนต์ เซอร์ไวฟด์
 - ซีโร นาว รีเพรเซนต์ ดีเซสส์
- ดิส เซนจ์ ออนไลน์ วิธ เดอะ โมเดลส ฟอกัส ออน พรดิคทิง เซอร์ไววัล

This page presents the two most accurate models selected from last week's experiments for comparison during the refinement phase.

The images shown here display results based on the original data, prior to any modifications.

ดิส เพจ เพรสเซนท์ส เดอะ ทู โมส แอคเคอเรท โมเดลส ซีเล็คเตด ฟรอม ลาสต์ วีคส์ เอกซ์เพริเมนต์ส ฟอร์ คอมแพริสัน ดัวริง เดอะ รีไฟน์เมนต์ เฟส

เดอะ อิมเมจเซส โชว์น เฮียร์ ดิสเพลย์ รีซัลต์ส เบสด์ ออน เดอะ ออริจินัล เดต้า, ไพเรเออร์ ทู เอนี่ โมดิฟิเคชันส

The next slide presents data improvement using **Method 1**, focusing on addressing the problem of **imbalanced data**.

This step utilizes **SMOTE (Synthetic Minority Over-sampling Technique)**, which generates synthetic samples for the minority class by sampling nearby data points, thereby increasing the number of samples in the **deceased** group, which is smaller than the survived group.

After balancing the data, this dataset is used with the **two selected models** for comparing their performance under balanced data conditions.

เดอะ เน็กซ์ สไลด์ เปรสเซนส์ เดต้า อิมพรวเมนต์ ยูซิง **เมธอด วัน**, โฟกัสซิง ออน แอดเดรสซิง เดอะ โปรบเล็ม ออฟ **อิมบาลานซ์ เดต้า**

ดิส สเลป ยูทิลไลส์ **SMOTE (ซินเทติก ไมเนอร์ตี โอเวอร์-แซมพลิง เทคนิก)** วิช เจเนอเรตส ซินเทติก แซมเพิลส์ ฟอว์ เดอะ ไมเนอร์ตี คลาส บาย แซมพลิง เนียร์บาย เดต้า พอยต์ส, แดร์บาย อินครีซซิง เดอะ นัมเบอร์ ออฟ แซมเพิลส อิน เดอะ **ดิเซสท์** กรุป, วิช อีส สมอลเลอร์ แธน เดอะ เซอร์ไวฟด์ กรุป

แอฟเตอร์ บาลานซิง เดอะ เดต้า, ดิส เดต้าเซต อีส ยูสด์ วิท เดอะ **ทู ซีเล็คเตด โมเดลส** ฟอว์ คอมแพริง แธร์ เพอร์ฟอร์แมนซ์ อันเดอร์ บาลานซ์ เดต้า คอนดิชันส์

Method 2 uses **Stratified K-Fold Cross-Validation**, a technique that splits the data into **K folds** while maintaining the **class proportions** close to the original dataset.

The model is trained **K times**, each time using a different fold as the test set, and the results are averaged.

This method provides a more **accurate and fair** evaluation, especially when dealing with imbalanced data.

เมธอด ทู ยูซส **สเตรทิไฟด์ เค-โฟลด์ ครอส-แวลิดชัน**, อะ เทคนิก แธท สปลิตส เดอะ เดต้า อินทู เค โฟลด์ส ไวล ไมเทนนิ่ง เดอะ **คลาส โปรพอร์ชันส** คลอส ทู เดอะ ออริจินัล เดต้าเซต

เดอะ โมเดล อีส เทรนด เค **ไทม์ส**, อีช ไทม์ ยูซิง อะ ดิฟเฟอเรนท โฟลด์ แอส เดอะ เทสต์ เซต, แอนด์ เดอะ รีซัลต์ส อาร์ แอฟเวอริจด์

ดิส เมธอด โปรไวดส์ อะ มอร์ **แอคเคอเรท แอนด์ แฟร์** อีวาลูเอชัน, เอสเปเชียลลี เว็น ดิลลิ่ง วิท อิมบาลานซ์ เดต้า

This slide shows the results from using **Stratified K-Fold Cross-Validation** to compare the **Random Forest** and **Logistic Regression** models.

Both models achieved over **94% Accuracy** and **F1-score**, indicating strong overall performance.

- Logistic Regression has the highest **Recall** for the survived class (class 1) at 97.2%.
- Random Forest has a higher **Recall** for the deceased class (class 0) at 44.0% compared to 39.5%.
- The **AUC** of Random Forest (0.9019) is slightly higher than Logistic Regression (0.8985).

In summary, Logistic Regression excels at predicting survival, while Random Forest better distinguishes between classes overall.

ดิส สไลด์ โชว์ เดอะ รีซัลต์ส ฟรอม ยูซิง **สแตรทิไฟด์ เค-โฟลด์ ครอส-แวลิดേഷัน** ทุ คอมแพร์ เดอะ แรนดอม ฟอเรสต์ แอนด์ ลอจิสติก รีเกรสชัน โมเดลส

โบท โมเดลส แอชีฟด์ โอเวอร์ **94% แอคคิวราซี** แอนด์ เอฟวัน-สกอร์, อินดิเคติง สตรอง โอเวอร์อล เพอร์ฟอร์แมนซ์

- ลอจิสติก รีเกรสชัน แอส เดอะ ไฮเอสท์ **รีคอล** ฟอร์ เดอะ เซอร์ไวฟด์ คลาส (คลาส วัน) แอท **97.2%**
- แรนดอม ฟอเรสต์ แอส อะ ไฮเออร์ **รีคอล** ฟอร์ เดอะ ดีเซสท์ คลาส (คลาส ซีโร) แอท **44.0%** คอมแพร์ด ทุ **39.5%**
- เดอะ **เอยูซี** ออฟ แรนดอม ฟอเรสต์ (0.9019) อิส สไลต์ลี ไฮเออร์ แธน ลอจิสติก รีเกรสชัน (0.8985)

อิน ซัมเมอรี, ลอจิสติก รีเกรสชัน เอ็กเซลแอต พรีดิกทิง เซอร์ไววัล, ไวล แรนดอม ฟอเรสต์ เบ็ทเทอร์ ดิสทิงกวิช บีทวีน คลาสเซส โอเวอร์อล

From the two methods tested, **Stratified K-Fold Cross-Validation** demonstrates better performance than **SMOTE (Synthetic Minority Over-sampling Technique)**.

ฟรอม เดอะ ทุ เมธอดส เทสเต็ด, **สแตรทิไฟด์ เค-โฟลด์ ครอส-แวลิดേഷัน** เดมอนสเตรทส เบ็ทเทอร์ เพอร์ฟอร์แมนซ์ แธน **SMOTE (ซินเทติก ไมนอร์ตี โอเวอร์-แซมปลิง เทคนิก)**

I have started researching how to write the manuscript and have begun drafting a small portion of it.

ไอ แฮฟ สตาร์ทเท็ด รีเสิร์ชอิง ฮาว ทุไรท์ เดอะ แมนูสคริปต์ แอนด์ แฮฟ บีกัน ดราฟทิง อะ สมอลล์ พอร์ชัน ออฟ อิท