เว็บแอปพลิเคชันสำหรับการวิเคราะห์บันทึกข้อมูลจราจรคอมพิวเตอร์ขนาดใหญ่

สลิลทิพย์ ธรรมศิริ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2564

A WEB-BASED APPLICATION FOR BIG DATA LOGS ANALYSIS

SALINTHIP THAMMASIRI

An Independent Study Submitted in Partial Fulfillment of the

Requirements for the Degree of Master of Engineering,

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

Academic Year 2021

ใบรับรองงานสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อสารนิพนธ์ เว็บแอปพลิเคชันสำหรับการวิเคราะห์บันทึกข้อมูลจราจรคอมพิวเตอร์ขนาดใหญ่

เสนอโดย สลิลทิพย์ ธรรมศิริ

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด)

.....กรรมการและอาจารย์ที่ปรึกษา

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ดร.ชัยพร เขมะภาตะพันธ์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2565

หัวข้อสารนิพนธ์	เว็บแอปพลิเคชันสำหรับการวิเคราะห์บันทึกข้อมูลจราจรคอมพิวเตอร์ขนาดใหญ่
ชื่อผู้เขียน	สลิลทิพย์ ธรรมศิริ
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2564

บทคัดย่อ

ในปัจจุบันเว็บแพลตฟอร์มอีคอมเมิร์ซได้รับความสนใจมากขึ้น เพราะความสะดวกสบายในการซื้อขายสินค้า ดังนั้นจึงมีข้อมูลธุรกรรมทางออนไลน์จำนวนมากที่ถูกจัดเก็บไว้ในฐานข้อมูล นอกจากนี้ยังมีไฟล์บันทึกข้อมูลการเข้าใช้งานเว็บไซต์ที่บันทึกข้อมูลการเข้าถึง เช่น IP address ของผู้ใช้งาน, วันที่เข้าใช้งานและสถานะ ที่ถูกจัดเก็บในรูปแบบของ log file ในการใช้งานข้อมูลบันทึกนี้ ทางผู้วิจัยได้ติดตั้งแอปพลิเคชันบนเว็บที่สามารถวิเคราะห์ข้อมูลบันทึกการเข้าใช้งานเว็บไซต์จำนวนมาก และนำเสนอในรูปแบบของ dashboard ด้วยผลลัพธ์จากการวิเคราะห์นี้ ผู้ดูแลระบบสามารถตรวจดูความถี่ของ IP address แต่ละรายการ และดูข้อมูลช่วงเวลารวมไปถึงช่วงโมงเร่งด่วนที่ผู้ใช้งานเข้าใช้เว็บไซต์ นอกจากนี้ ระบบสามารถคาดการณ์จำนวนผู้ใช้งานที่จะเข้ามาใช้งานเว็บไซต์ล่วงหน้าใน 1 ถึง 6 ชั่วโมงข้างหน้าอีกด้วย

Independent Study Title	A WEB-BASED APPLICATION FOR BIG DATA LOGS ANALYSIS
Author	SALINTHIP THAMMASIRI
Independent Study Advisor	Dr. Eakasit Pacharawongsakda
Department	Big Data Engineering
Academic Year	2021

ABSTRACT

Nowadays, an e-commerce platform gets more attention because it's convenient to buy products. Therefore, it has a large number of online transactions stored in a database. Moreover, there is a web access log file that records access data, e.g., client IP address, date, and status. To utilize this log data, we implemented a web-based application that can analyze a large volume of log data, and present it as a dashboard. With this result, an administrator can view the frequency of each IP address and see the peak hours that the many user accessed. Additionally, the system can forecast the number of access in the next 1 to 6 hours.

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้โดยการให้ความช่วยเหลือของ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้คำแนะนำ ตรวจสอบ และแก้ไขข้อบกพร่องต่าง ๆ มาโดยตลอด เพื่อให้สารนิพนธ์ฉบับนี้สมบูรณ์ ผู้เขียนจึงขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ผู้เขียนขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด ที่กรุณาให้เกียรติเป็นประธานโดยมี ผศ.ดร.ดวงใจ จิตคงชื่น เป็นกรรมการในการสอบสารนิพนธ์ ซึ่งได้กรุณาตรวจแก้ไขสารนิพนธ์ฉบับ นี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น และนางสาวกุลธิดา รอดบุญ ที่ให้ความสะดวกด้านอำนวยการ และประสานงาน ในการทำสารนิพนธ์ให้กับผู้เขียนมาโดยตลอด

สุดท้ายนี้ขอขอบคุณครอบครัวและเพื่อนๆ ที่คอยช่วยส่งเสริม สนับสนุนและให้กำลังใจ ทำให้การศึกษาวิจัยในครั้งนี้สำเร็จลุล่วงไปด้วยดี

สลิลทิพย์ ธรรมศิริ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันเว็บไซต์เข้ามามีบทบาทและความสำคัญต่อชีวิตประจำวันเราเป็นอย่างมาก ทุกองค์กรทั้งภาครัฐและเอกชน ได้ให้ความสำคัญของการมีเว็บไซต์เพิ่มมากขึ้น จุดประสงค์เพื่อประชาสัมพันธ์ หรือเพื่อการค้าขาย หลายองค์กรจึงมีการทำเว็บไซต์ เพื่อเป็นการยืนยันตัวตนให้เป็นที่รู้จักและเป็นการโฆษณาไปในตัว ในแง่ของการทำธุรกิจการสามารถใช้เว็บไซต์เพื่อช่วยกระตุ้นยอดขาย รวมถึงสร้างฐานลูกค้าใหม่ ๆ ได้อีกด้วย

โดยทั่วไปไม่ว่าเราจะใช้งานเว็บไซต์ไหนก็จะมีการจัดเก็บข้อมูลจราจรทางคอมพิวเตอร์ (Log file) เป็นข้อมูลที่แสดงเกี่ยวกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ แสดงให้เห็นถึง ต้นทาง, ปลายทาง, วันที่, เวลา หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ การบันทึกการใช้งานข้อมูลด้วย Log file คือวิธีที่จะช่วยแก้ปัญหาที่อาจจะเกิดขึ้น ไม่ว่าจะเป็นการส่งข้อมูลที่ผิดพลาดจากผู้ใช้งาน การขโมยข้อมูล หรือการกระทำที่เกิดจากการถูกโจมตีจากโปรแกรมที่ไม่พึงประสงค์ ที่ทำให้เกิดความเสียหายในระบบ การมีข้อมูลจาก Log file จะสามารถตรวจสอบที่มาที่ไปของข้อมูลได้ ทำให้หาสาเหตุของปัญหาและแก้ไขได้รวดเร็วยิ่งขึ้น

ในด้านของการทำธุรกิจออนไลน์การนำ Log file ของเว็บไซต์มาวิเคราะห์ข้อมูลผู้ใช้งาน จะสามารถทำให้เข้าใจพฤติกรรมของลูกค้าที่ใช้งานเว็บไซต์และเพื่อนำไปพัฒนาเว็บไซต์ให้มีประสิทธิภาพมากยิ่งขึ้น

งานวิจัยนี้จึงนำเสนอข้อมูลการวิเคราะห์ Web access log ผ่านเว็บแอปพลิเคชันในรูปแบบของ Dashboard เพื่อให้มองเห็นภาพรวมและพฤติกรรมของผู้ใช้งานเว็บไซต์ รวมไปถึงการวิเคราะห์ปริมาณการเข้าใช้งานเว็บไซต์ เพื่อใช้สำหรับวางแผนการจัดสรรทรัพยากรของระบบให้รองรับปริมาณการใช้งาน

1.2 วัตถุประสงค์ของการศึกษาหรือวิจัย

1. เพื่อสร้างโมเดลสำหรับพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์
2. เพื่อสร้างเครื่องมือที่แสดงผลการวิเคราะห์ข้อมูลในหลายมุมมอง เข้าใจง่ายในรูปแบบของ Dashboard

1.3 ขอบเขตของงานวิจัย

1. เป็นข้อมูล Web access log จาก Kaggle ที่นำมาใช้งาน
2. เป็นข้อมูลรูปแบบ Semi-structured
3. ข้อมูลเข้าใช้งานเว็บไซต์ตั้งแต่วันที่ 22 มกราคม 2562 จนถึงวันที่ 26 มกราคม 2562 มีจำนวนข้อมูลทั้งหมด 10,365,152 แถว

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้โมเดลที่สามารถพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์
2. ได้เครื่องมือในการนำเสนอข้อมูลที่เข้าใจง่ายในรูปแบบของ Dashboard

1.5 นิยามศัพท์

1. “Web Access log” หมายถึง บันทึกการเข้าใช้งานของเว็บเซิร์ฟเวอร์ ซึ่งมีข้อมูลเชิงลึกมากมายเกี่ยวกับผู้เข้าใช้งานเว็บไซต์ ที่แสดงถึงต้นทาง, ปลายทาง, วันที่, เส้นทาง หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์

บทที่ 2
ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างเครื่องมือที่แสดงผลการวิเคราะห์ข้อมูลในหลายมุมมอง เข้าใจง่ายในรูปแบบของ Dashboard และเพื่อพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.1 การจัดเก็บข้อมูลการจราจรทางคอมพิวเตอร์ (Log file)

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

2.3 Gradient Boosted Trees

2.4 Autoregressive Integrated Moving Average (ARIMA)

2.5 งานวิจัยที่เกี่ยวข้อง

2.1 การจัดเก็บข้อมูลการจราจรทางคอมพิวเตอร์ (Log file)

การจัดเก็บข้อมูลการจราจรทางคอมพิวเตอร์ (Log file) หรือที่เรียกอีกแบบว่าการเก็บข้อมูลการใช้งานคอมพิวเตอร์และอินเทอร์เน็ต คือ ข้อมูลเส้นทางการใช้งานอินเทอร์เน็ต หรือการเชื่อมต่อสื่อสารกันระหว่างเครื่องคอมพิวเตอร์ รวมถึงมือถือและแท็บเล็ต ที่มีไว้บันทึกเหตุการณ์ต่าง ๆ ข้อมูลที่เก็บไว้จะแสดงถึงต้นทาง, ปลายทาง, วันที่, เส้นทาง หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ โดยส่วนใหญ่การให้บริการอินเทอร์เน็ตจะจัดเก็บแบบ Access logs หรือ Logs เหตุการณ์การเข้าถึงเครือข่าย เพราะทุกองค์กรหรือผู้ให้บริการอย่างน้อยจะต้องมี Logs ประเภทนี้เสมอ เพื่อให้ทราบเหตุการณ์เมื่อเข้าเว็บไซต์หรือใช้งานอินเทอร์เน็ต โดยข้อมูลจะประกอบด้วย Timestamp, IP Address, Destination IP, Destination port หรือ Protocol name

```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/~lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"
304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /~lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt
HTTP/1.0" 404 276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET
/teachers/pitas1.html HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET
/~oswinds/publications.html HTTP/1.0" 200 48966
```

ภาพที่ 2.1 ตัวอย่างของการเก็บ Log file ของเว็บเซิร์ฟเวอร์

ที่มา: https://www.researchgate.net/figure/A-sample-of-Web-Server-Log-File_fig1_220773991

ตารางที่ 2.1 ตารางแสดงรูปแบบประเภทของ Web server log

Types	Actions	Format
Access log file	- บันทึกคำขอ (Request) ของผู้ใช้ทั้งหมด ประมวลผลโดยเซิร์ฟเวอร์ และ ข้อมูลเกี่ยวกับผู้ใช้	127.0.0.1 - Scott [10/Dec/2019:13:55:36 -0700] "GET /server-status HTTP/1.1" 200 2326
Agent log file	- บราวเซอร์ของผู้ใช้และเวอร์ชันของบราวเซอร์	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.108 Safari/537.36
Error log file	- รายการข้อผิดพลาดสำหรับผู้ใช้ที่ร้องขอโดยเซิร์ฟเวอร์	[Thu Mar 13 19:04:13 2014] [error] [client 50.0.134.125] File does not exist: /var/www/favicon.ico
Referrer log file	- ข้อมูลเกี่ยวกับลิงค์ และ การเปลี่ยนเส้นทางผู้เยี่ยมชมไปยังไซต์	"http://www.google.com/search?q=keyword", "/page.html"

ตารางที่ 2.2 ตารางแสดงพารามิเตอร์ของ log file

No.	Parameter name	Description
1	IP Address	ระบุผู้เข้าเยี่ยมชมเว็บไซต์ด้วย IP Address
2	Time stamp	วันเวลาที่ผู้ใช้เรียกดูเว็บไซต์
3	Request	คำขอโดยผู้ใช้
4	Status code	รหัสที่ส่งโดยเซิร์ฟเวอร์หลังจากผู้ใช้แต่ละคนขอ
5	Bytes	ปริมาณของเนื้อหาเอกสาร
6	User agents	บราวเซอร์ที่ผู้ใช้ ใช้ส่งคำขอ
7	Request type	วิธีที่ผู้ใช้ ใช้ส่งคำขอ (GET, POST)

ตารางที่ 2.3 ตารางแสดงข้อมูล Server status codes ที่พบเจอบ่อย ๆ

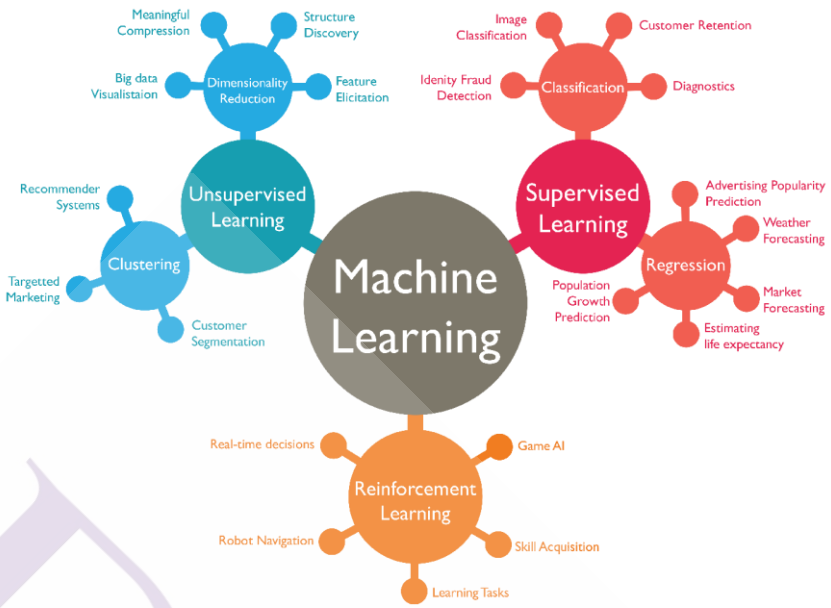
Success Code	
2xx Success (สำเร็จ)	
200	OK : สำเร็จ
201	Created : ข้อมูลถูกเพิ่มเรียบร้อย
202	Accepted : ได้รับคำขอและประมวลผลอยู่ แต่ยังดำเนินการไม่สำเร็จ
204	No Content : ประมวลผลคำขอเสร็จแล้ว แต่ไม่มีการตอบ Response body กลับไป
3xx Redirection (เปลี่ยนเส้นทาง)	
301	Multiple Choices : มีตัวเลือกหลายทางในทรัพยากรที่ทาง client สามารถเลือกได้
302	Moved Permanently : คำขอในปัจจุบันและอนาคตจะถูกพาไปยังตำแหน่ง URL ที่กำหนด
303	Found : ทรัพยากรที่ร้องขอถูกย้ายไป URL อื่นชั่วคราว
304	Not Modified : เว็บไซต์ที่ร้องขอเข้าใช้ไม่ได้มีการอัปเดตตั้งแต่ครั้งล่าสุดที่คุณเข้าถึง โดยปกติบราวเซอร์จะบันทึก (หรือแคช Cache) หน้าเว็บเพื่อไม่ให้ต้องดาวน์โหลดข้อมูลเดิมซ้ำ ๆ
Failure Code	
4xx Client Error (Client ผิดพลาด)	
400	Bad Request : เซิร์ฟเวอร์ไม่เข้าใจสิ่งที่ client ร้องขอ อาจเกิดจากคำขอผิดรูปแบบ

ตารางที่ 2.3 (ต่อ)

401	Unauthorized : client ไม่ได้ทำการ authenticate มาก่อน เซิร์ฟเวอร์จึงไม่สามารถให้ request นี้ทำงานได้
403	Forbidden : คล้ายกับ 401 แต่เซิร์ฟเวอร์รู้ว่า client เป็นใคร แต่ client ไม่มีสิทธิเข้าถึงข้อมูล
404	Not Found : ที่อยู่ URL ผิด หรือเว็บไซต์นั้นไม่มีอยู่จริง
5xx Server Error (เซิร์ฟเวอร์ผิดพลาด)	
500	Internal Server Error : พบข้อผิดพลาดภายในเซิร์ฟเวอร์
502	Bad Gateway : เซิรฟ์เวอร์ที่ทำเป็น Gateway หรือ Poxy ได้รับการตอบสนองที่ไม่ถูกต้องจากเซิร์ฟเวอร์ต้นทาง
503	Service Unavailable : เซิร์ฟเวอร์ไม่สามารถดำเนินการตามคำขอได้ อาจมาจากภาระการทำงานหนักเกินกว่าจะรับไหว หรืออยู่ในช่วงปรับปรุงเซิร์ฟเวอร์
504	Gateway Timeout : เซิรฟ์เวอร์ที่ทำเป็น Gateway หรือ Poxy ไม่ได้รับการตอบสนองภายในระยะเวลาที่กำหนดจากเซิร์ฟเวอร์ต้นทาง

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง (Machine Learning) คือ การสอนให้ระบบคอมพิวเตอร์การเรียนรู้ด้วยตัวเองโดยการใช้ข้อมูล การเรียนรู้ของเครื่องนั้นเป็นได้สองรูปแบบใหญ่ ๆ คือ Supervised Learning คือการที่คอมพิวเตอร์เรียนรู้ด้วยการที่มีข้อมูลมาสอน และ Unsupervised Learning คือการที่คอมพิวเตอร์เรียนรู้โดยที่ไม่ต้องมีข้อมูลมาสอน นอกจากนี้ยังมี Machine Learning สำหรับงานเฉพาะด้าน เช่น Reinforcement Learning คือการที่เครื่องเรียนรู้และเปลี่ยนไปตามสภาพแวดล้อม

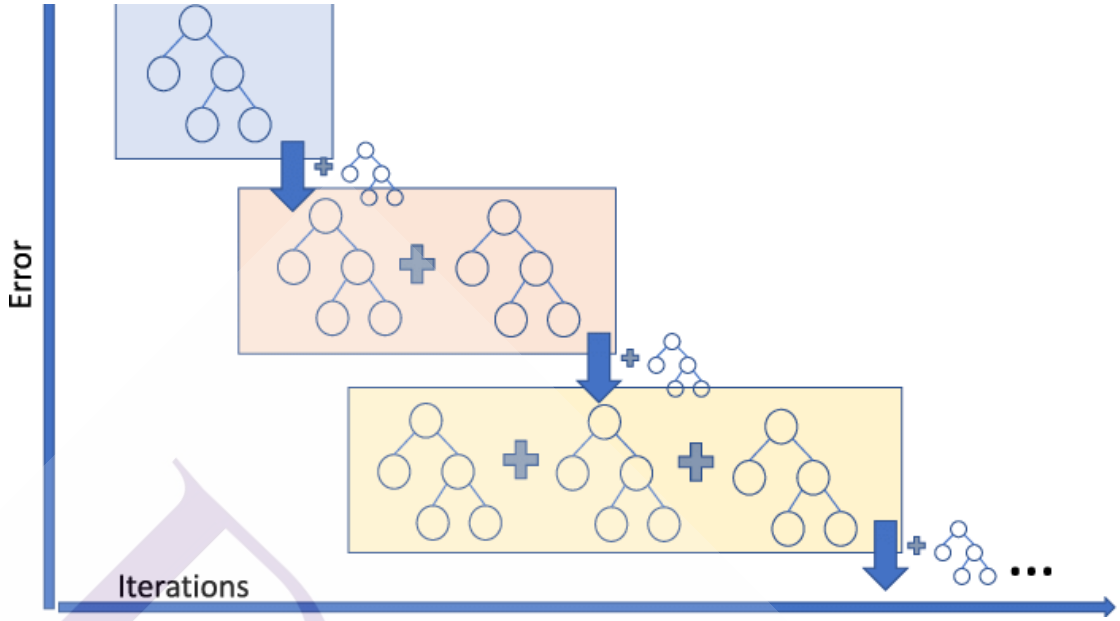


ภาพที่ 2.2 ประเภทหลัก ๆ ของ Machine Learning

ที่มา: https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

2.3 Gradient Boosted Trees

Gradient Boosted model เป็นชุดของแบบจำลองการถดถอย (Regression) และการจำแนกประเภท (Classification) ทั้งสองวิธีเป็นวิธีการเรียนรู้แบบ Forward-learning มีพื้นฐานมาจาก Decision tree ซึ่งจะเป็นการสร้าง Decision tree ต่อกันหลายแบบ โดยที่แต่ละ Decision tree จะเรียนรู้จาก Error ของแบบจำลองก่อนหน้า ซึ่งเป็นการปรับปรุงประสิทธิภาพของแบบจำลองให้สูงขึ้น และประเมินผลแต่ละแบบจนกว่าจะได้ Decision tree ที่สมบูรณ์



ภาพที่ 2.3 ลักษณะการทำงานของ Gradient Boosted Trees

ที่มา: https://pub.towardsai.net/gradient-boosting-technique-b3dbb7069b74

2.4 Autoregressive Integrated Moving Average (ARIMA)

เทคนิค ARIMA เป็นหนึ่งในการวิเคราะห์ข้อมูลแบบอนุกรมเวลา คือการอาศัยข้อมูลในอดีตมากำหนดรูปแบบของข้อมูลและการพยากรณ์ข้อมูลในอนาคต องค์ประกอบของแบบจำลอง ARIMA ประกอบด้วย 3 ส่วน ได้แก่ Autoregression process หรือ AR(p), Integrated (d) และ Moving average process หรือ MA(q)

Autoregression process หรือ AR(p) เป็นกระบวนการอธิบายตัวแปร Y ด้วยค่าของตัวแปร Y(Yt-1, Yt-2, Yt-3,…, Yt-p) ค่า Y ที่ใช้สามารถมีได้หลาย order, Integrated (d) d คือจำนวนความแตกต่างที่จำเป็นในการทำให้อนุกรมเวลาคงที่ และ Moving average process หรือ MA(q) คือการดูโมเดลว่าพึ่งพาระหว่างการสังเกตและข้อผิดพลาดที่เหลือจากแบบจำลองค่าเฉลี่ยเคลื่อนที่ที่ใช้กับการสังเกตที่ยังไง

รูปแบบ	สมการของ (Y)
ARIMA(1,1,0)/AR(1)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \varepsilon_t$
ARIMA(2,1,0)/AR(2)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + \varepsilon_t$
ARIMA(0,1,1)/MA(1)	$\Delta Y_t = \alpha + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
ARIMA(0,1,2)/MA(2)	$\Delta Y_t = \alpha + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$
ARMA(1,1,1)	$\Delta Y_t = \alpha + \phi_1 \Delta Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

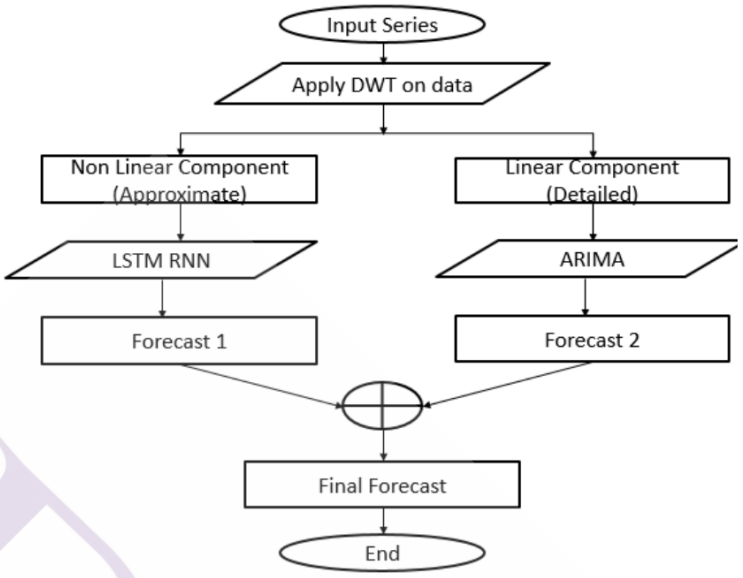
ภาพที่ 2.4 ตัวอย่างสมการพยากรณ์ ARIMA model

ที่มา: https://nutdnuy.medium.com/การพยากรณ์ข้อมูลอนุกรมเวลาด้วยเทคนิค-arima-ด้วย-python-44809eb8e990

2.5 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลและพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ ที่ผู้วิจัยได้ศึกษาและสรุปได้ดังนี้

Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020). Web Traffic Time Series Forecasting using ARIMA and LSTM RNN. ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในการศึกษาปริมาณการเข้าใช้งานเว็บไซต์ โดยมีการเปรียบเทียบประสิทธิภาพของโมเดลทั้งหมด 2 แบบได้แก่ ARIMA model และ LSTM RNN

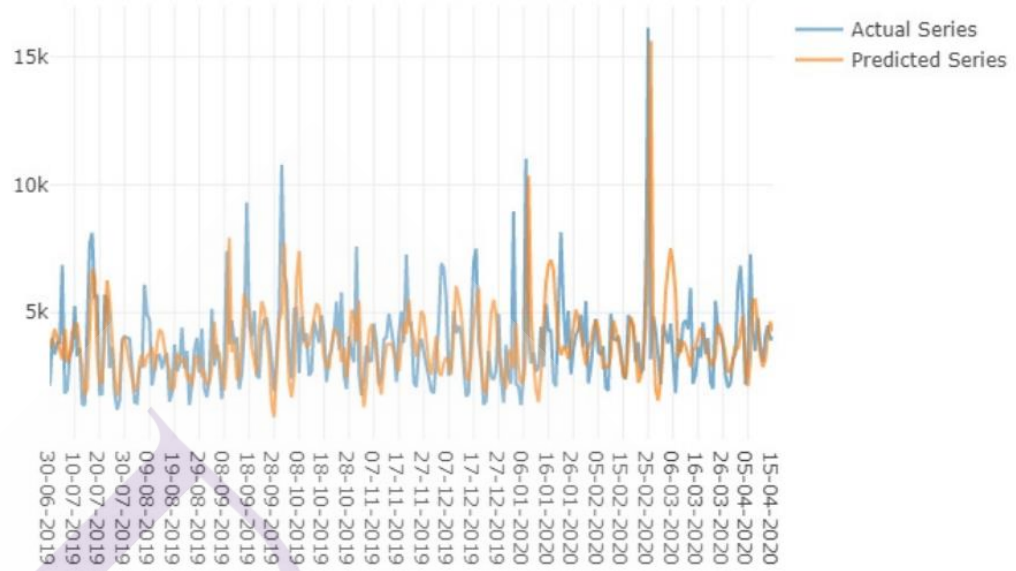


ภาพที่ 2.5 แผนผังของระบบ

ที่มา: Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020).

ภาพที่ 2.6 ผลการพยากรณ์ปริมาณการเข้าใช้งานโดยใช้โมเดล ARIMA

ที่มา: Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020).

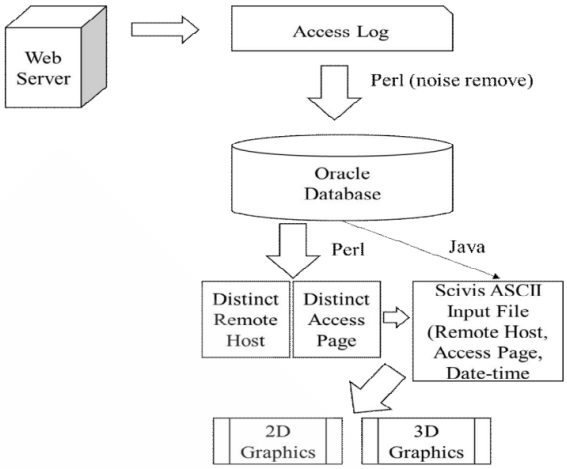


ภาพที่ 2.7 ผลการพยากรณ์ปริมาณการเข้าใช้งานโดยใช้โมเดล LSTM-RNN

ที่มา: Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020).

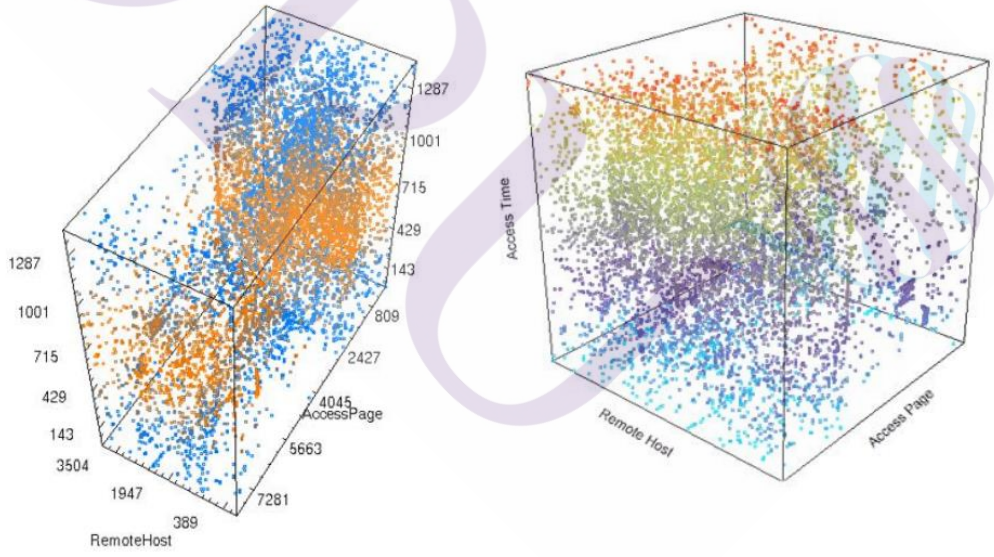
จากผลการทดลองจะเห็นได้ว่าการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ ที่เป็นอนุกรมเวลาโดยใช้ LSTM-RNN มีประสิทธิภาพและแม่นยำมากกว่า ARIMA model

Jungkee Kim, (2018). Web Server Log Visualization. งานวิจัยนี้ได้นำเสนอข้อมูลของ Web log และการวิเคราะห์บันทึกการใช้เว็บไซต์ของ NPAC โดยการใช้ Shell script ในการบันทึกข้อมูลและใช้ Perl ในการสร้างข้อมูลรายวัน, รายสัปดาห์ และ โดเมน และบันทึกข้อมูลลงในฐานข้อมูล Oracle สำหรับขั้นตอน Pre-process ข้อมูลจะใช้ข้อมูลเพียงสามฟิลด์จากที่บันทึกไว้ในฐานข้อมูลได้แก่ Remote Host, Access Page และ Date-time โดยนำข้อมูลเหล่านี้เข้าโปรแกรม Java เพื่อแสดงผลข้อมูลการวิเคราะห์ออกมาเป็นกราฟ



ภาพที่ 2.8 แผนผังการเตรียมข้อมูลในการวิเคราะห์บันทึกการใช้งานเว็บไซต์

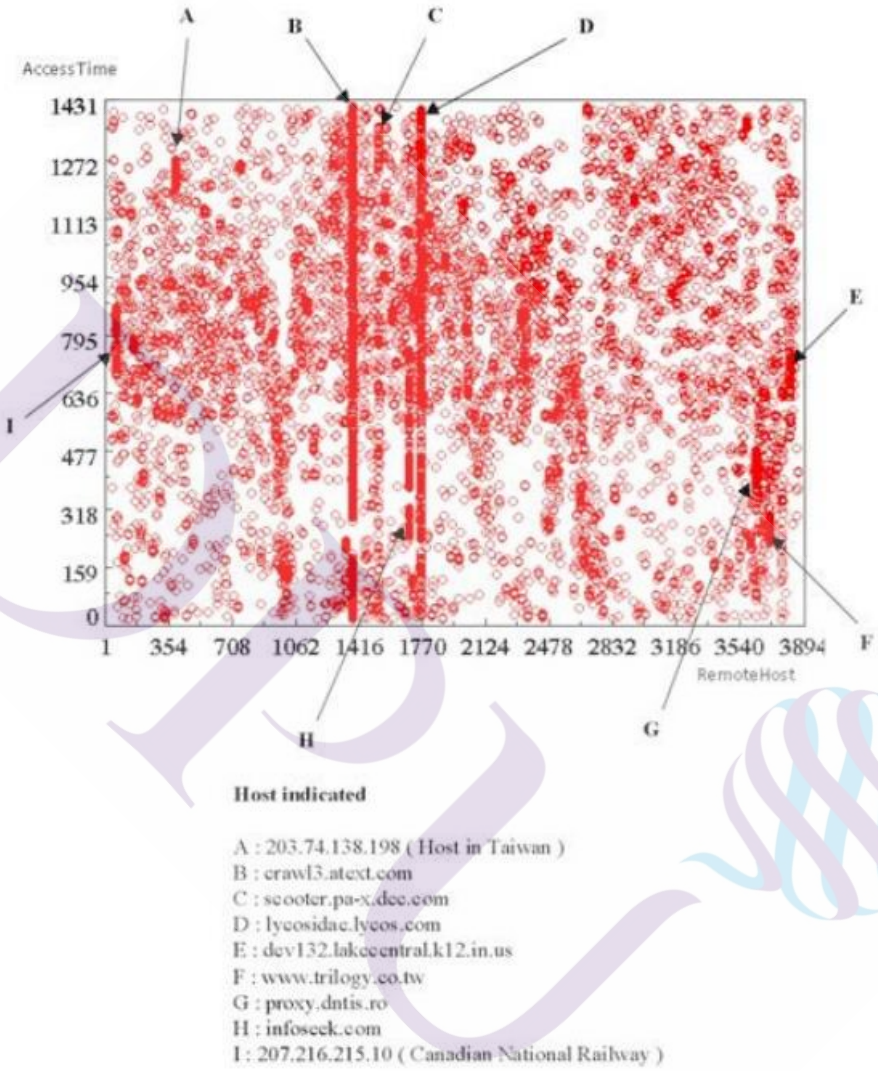
ที่มา: Jungkee Kim, (2018).



ภาพที่ 2.9 กราฟสามมิติของข้อมูลการใช้งานรายวัน

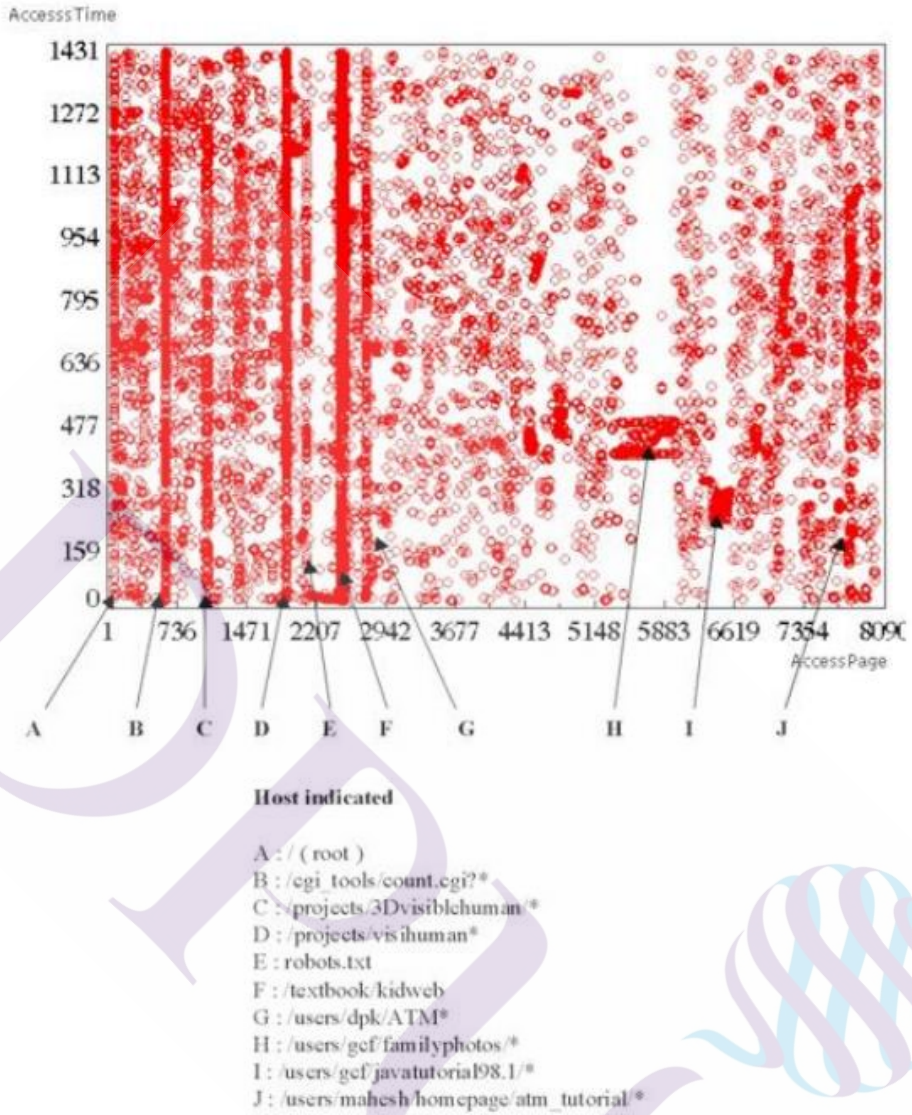
ที่มา: Jungkee Kim, (2018).

ความสัมพันธ์ระหว่าง Access time และ Remote host ในภาพที่ 2.9 เส้นยาวแสดงให้เห็นถึงเว็บโรบอท ส่วนเส้นหนาสั้น ๆ แสดงให้เห็นถึงการเข้าใช้งานของ Remote Host หลายครั้งภายในระยะเวลาสั้น ๆ



ภาพที่ 2.10 กราฟสองมิติของข้อมูลระหว่าง Access time และ Remote host

ที่มา: Jungkee Kim, (2018).



ภาพที่ 2.11 กราฟสองมิติของข้อมูลระหว่าง Access time และ Access page

ที่มา: Jungkee Kim, (2018).

บทที่ 3
วิธีวิจัย

งานวิจัยนี้เป็นการนำเสนอเครื่องมือที่ช่วยพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ และเครื่องมือการแสดงผลการวิเคราะห์ข้อมูลแบบพื้นฐาน โดยมีขั้นตอนและแนวทางการวิจัยดังนี้

ภาพที่ 3.1 ขั้นตอนการทำงานของระบบตั้งแต่เริ่มต้น (Input) จนถึงการแสดงผลข้อมูลบน dashboard (Output)

3.1 การเก็บข้อมูล

งานวิจัยนี้ได้ใช้ข้อมูลการบันทึกของเว็บเซิร์ฟเวอร์ที่บันทึกการเข้าใช้งานของเว็บไซต์ www.zanbil.ir จาก Kaggle ซึ่งเป็นข้อมูลการเข้าใช้งานเว็บไซต์อีคอมเมิร์ซของประเทศอิหร่าน ตั้งแต่วันที่ 22 มกราคม 2562 ถึง 26 มกราคม 2562

```
40.77.167.129 - - [22/Jan/2019:03:56:19 +0330] "GET /image/6229/productModel/
150x150 HTTP/1.1" 200 2739 "-" "Mozilla/5.0 (compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)" "-"

207.46.13.136 - - [22/Jan/2019:03:56:19 +0330] "GET /product/14926
HTTP/1.1" 404 33617 "-" "Mozilla/5.0 (compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)" "-"

40.77.167.129 - - [22/Jan/2019:03:56:19 +0330] "GET /image/6248/productModel/
150x150 HTTP/1.1" 200 2788 "-" "Mozilla/5.0 (compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)" "-"

40.77.167.129 - - [22/Jan/2019:03:56:20 +0330] "GET /image/64815/productModel/
150x150 HTTP/1.1" 200 3481 "-" "Mozilla/5.0 (compatible; bingbot/2.0;
+http://www.bing.com/bingbot.htm)" "-"
```

ภาพที่ 3.2 ตัวอย่างข้อมูล Web access log บน Kaggle ที่นำมาใช้งาน

3.2 การเตรียมข้อมูล

3.2.1 นำข้อมูลเข้าฐานข้อมูล

นำข้อมูล Access log เข้าฐานข้อมูลโดยใช้ภาษา PHP ทำคำสั่ง loop ข้อมูลขึ้นฐานข้อมูลทั้งหมด

```php
$dsn = 'mysql:dbname=dpu_is;host=127.0.0.1';
$user = 'root';
$password = '';

try {
    $dbh = new PDO($dsn, $user, $password);
} catch (PDOException $e) {
    echo 'Connection failed: ' . $e->getMessage();
}

$file= new SplFileObject('access_.log');

while(!$file->eof())
{
    $line=$file->fgets();
    list($log)=explode(PHP_EOL,$line);

    $sth = $dbh->prepare('INSERT INTO `access_log` (`log`) VALUES (?);');
    $sth->bindValue(1, $log, PDO::PARAM_STR);
    $sth->execute();
}
```

ภาพที่ 3.3 ตัวอย่าง code สำหรับขั้นตอนการ loop ข้อมูล Web access log ขึ้นฐานข้อมูล

id	log
1	54.36.149.41 - - [22/Jan/2019:03:56:14 +0330] "GET...
2	31.56.96.51 - - [22/Jan/2019:03:56:16 +0330] "GET ...
3	31.56.96.51 - - [22/Jan/2019:03:56:16 +0330] "GET ...
4	40.77.167.129 - - [22/Jan/2019:03:56:17 +0330] "GE...
5	91.99.72.15 - - [22/Jan/2019:03:56:17 +0330] "GET ...
6	40.77.167.129 - - [22/Jan/2019:03:56:17 +0330] "GE...
7	40.77.167.129 - - [22/Jan/2019:03:56:18 +0330] "GE...
8	40.77.167.129 - - [22/Jan/2019:03:56:18 +0330] "GE...
9	66.249.66.194 - - [22/Jan/2019:03:56:18 +0330] "GE...
10	40.77.167.129 - - [22/Jan/2019:03:56:18 +0330] "GE...

ภาพที่ 3.4 ตัวอย่างข้อมูล Web access log ที่ถูกบันทึกลงฐานข้อมูลหลังจากใช้ PHP loop

3.2.2 Feature Extraction

สำหรับข้อมูลจากฐานข้อมูลจะอยู่ในลักษณะ Semi-structure จะต้องทำการแบ่งข้อมูลออกมาเป็นส่วนต่าง ๆ และทำการเปลี่ยนชื่อแอตทริบิวต์เพื่อให้ง่ายต่อการใช้งาน และทำการเปลี่ยนแอตทริบิวต์วันที่จาก Norminal ให้เป็น Date ก่อนที่จะนำไปใช้งาน

timezone	referer	user_agent	ip_address	method	request_url	protocal	request_sta...	size	datetime_old
+0330	-	Mozilla/5.0 (c...	54.36.149.41	GET	/filter/27|13%...	HTTP/1.1	200	30577	22/Jan/2019 03:5...
+0330	https://www.z...	Mozilla/5.0 (Li...	31.56.96.51	GET	/image/6084...	HTTP/1.1	200	5667	22/Jan/2019 03:5...
+0330	https://www.z...	Mozilla/5.0 (Li...	31.56.96.51	GET	/image/6147...	HTTP/1.1	200	5379	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (c...	40.77.167.129	GET	/image/1492...	HTTP/1.1	200	1696	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (W...	91.99.72.15	GET	/product/3189...	HTTP/1.1	200	41483	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (c...	40.77.167.129	GET	/image/2348...	HTTP/1.1	200	2654	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (c...	40.77.167.129	GET	/image/4543...	HTTP/1.1	200	3688	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (c...	40.77.167.129	GET	/image/576/a...	HTTP/1.1	200	14776	22/Jan/2019 03:5...
+0330	-	Mozilla/5.0 (c...	66.249.66.194	GET	/filter/b41,b66...	HTTP/1.1	200	34277	22/Jan/2019 03:5...

ภาพที่ 3.5 ตัวอย่างข้อมูลหลังจากทำการแบ่งและเปลี่ยนชื่อแอตทริบิวต์

3.3 การวิเคราะห์ข้อมูลพื้นฐานและบันทึกค่าลงฐานข้อมูล

หลังจากเสร็จสิ้นกระบวนการเตรียมข้อมูลแล้ว ข้อมูลจะถูกนำมาวิเคราะห์หาผลลัพธ์ในหลายๆด้าน เพื่อนำไปบันทึกลงฐานข้อมูลและแสดงผลบน dashboard

3.3.1 Count data group by request status

ทำการหาค่าผลรวมจำนวนของแต่ละ request status

request_status	count(request_status)
200	9579760
206	3
301	67553
302	199835
304	340228
400	586
401	323

ภาพที่ 3.6 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของแต่ละ request status

3.3.2 Count data group by request url

ทำการหาค่าผลรวมจำนวนของแต่ละ request url

request_url	count(request_url)
/	47580
/%20	3
/%2Fstatic%2Fima...	23
/%D8%A6/big-kitch...	1
/%D8%A7%D8%B...	2
/%D8%AA%D8%B...	4

ภาพที่ 3.7 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของแต่ละ request url

3.3.3 Count IP address group by IP address and datetime

ทำการหาค่าผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์โดยอิงข้อมูลตาม IP address และ วันเวลา (รายนาที)

ip_address	datetime	count(ip_ad...
2.178.180.33	26/01/2019 19:04	149
2.178.180.33	26/01/2019 19:05	31
2.178.181.20	26/01/2019 16:22	15
2.178.181.20	26/01/2019 16:23	20
2.178.181.240	25/01/2019 16:11	27
2.178.181.240	25/01/2019 16:12	26

ภาพที่ 3.8 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์

3.3.4 Web access traffic

ทำการหาค่าผลรวมจำนวนของ request โดยอิงข้อมูลตามวันเวลา

datetime	count(ip_address)
22/01/2019 07:26	242
22/01/2019 07:27	287
22/01/2019 07:28	331
22/01/2019 07:29	425
22/01/2019 07:30	246
22/01/2019 07:31	245

ภาพที่ 3.9 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ request โดยอิงข้อมูลตามวันเวลา

3.3.5 Most active IP address

ทำการหาค่าผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์โดยอิงค่าตาม IP address

ip_address	count(ip_address)
10.1.48.115	125
10.1.52.71	147
10.1.68.25	446
10.10.56.92	282
10.103.61.149	56
10.104.208.32	96

ภาพที่ 3.10 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ IP address ที่มีการ request เข้าเว็บไซต์โดยอิงค่าตาม IP address

3.3.6 Count request status group by request status and datetime

Map ค่า request status ให้อยู่ในรูปแบบเดียวกัน เช่น (request status 200, 201, 202 จะถูก map ค่าให้เป็น 2XX) และทำการหาค่าผลรวมจำนวนของ request status เข้าเว็บไซต์โดยอิงค่าตาม request status และ วันเวลา

request_status	datetime	count
2XX	22/01/2019 07:26	230
2XX	22/01/2019 07:27	273
2XX	22/01/2019 07:28	299
2XX	22/01/2019 07:29	399
2XX	22/01/2019 07:30	220
2XX	22/01/2019 07:31	222

ภาพที่ 3.11 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ request ที่ผ่านการ map ค่าให้อยู่ในรูปแบบเดียวกันโดยอิงข้อมูลตามรายวันเวลา (นาที)

3.3.7 Count active url by path

Split ค่า url ออก โดยใช้ / (Slash) เป็นตัวแบ่งข้อมูลออกจากกัน ก่อนนำข้อมูลของ path ไปหาค่าผลรวม

request_url_2	count(request_url_2) ↓
image	5682852
static	2000256
m	569326
settings	352047
filter	341412
site	240432

ภาพที่ 3.12 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ path ที่ผู้ใช้เรียกใช้งาน

3.3.8 Count request by user agent

ทำการหาค่าผลรวมจำนวนของ user agent ที่มีการ request เข้าเว็บไซต์โดยอิงค่าตาม user agent

agent	count ↓
Mozilla/5.0 (Windows NT 10.0; Win64; x64) Apple...	746572
Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleW...	702672
Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...	636897
Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:64.0...	551966
Mozilla/5.0 (Windows NT 6.1; rv:64.0) Gecko/2010...	454961

ภาพที่ 3.13 ตัวอย่างผลลัพธ์การหาผลรวมจำนวนของ user agent ที่มีการเข้าใช้งานบนเว็บไซต์

3.3.9 Map device user agent

แปลงค่า user agent ให้กลายเป็น device (desktop, tablet, smartphone) โดยใช้ curl API ส่งค่า user agent ไปยังเว็บผู้ให้บริการข้อมูล user agent และนำค่าที่ return กลับมาบันทึกลงฐานข้อมูล

```php
<?php
$conn = new mysqli("localhost","root","","dpu_is");
if ($conn -> connect_errno) {
  echo "Failed to connect to MySQL: " . $mysqli -> connect_error;
  exit();
}

$sql = "SELECT agent FROM user_agent";
$result = $conn->query($sql);
$agent_array = array();
if ($result->num_rows > 0) {
    while ($row = mysqli_fetch_array($result)) {
        $agent = str_replace(' ', '%20', $row["agent"]);
        $string = $row["agent"];
        $agent_array[] = $agent;
    }
}

foreach ($agent_array as $value) {
  $string = str_replace('%20', ' ', $value);
  $url = "https://api.apicagent.com/?ua=".$value;

  $ch = curl_init();
  curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
  curl_setopt($ch, CURLOPT_URL, $url);
  $result = curl_exec($ch);
  $sub = explode(',',$result);

  if (isset($sub[8])) {
    $sub_2 = explode('"',$sub[8]);
    if($sub_2 != 'unknown'){
    $sql = "UPDATE user_agent SET type = '$sub_2[3]' WHERE agent = '$string'";
    $result = $conn->query($sql);
    }
  }
}
```

ภาพที่ 3.14 ตัวอย่าง code สำหรับ curl API ส่งค่า user agent เพื่อ map ค่า device

agent	type	count ▼ 1
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWeb...	desktop	746572
Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebK...	desktop	702672
Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (K...	desktop	636897
Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:64.0)...	desktop	551966
Mozilla/5.0 (Windows NT 6.1; rv:64.0) Gecko/201001...	desktop	454961
Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/...	unknown	450555
Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:64.0) ...	desktop	340970
Mozilla/5.0 (compatible; bingbot/2.0; +http://www....	unknown	197769
Mozilla/5.0 (compatible; Googlebot/2.1; +http://ww...	unknown	191450
Mozilla/5.0 (iPhone; CPU iPhone OS 12_1_2 like Mac...	smartphone	180005
Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebK...	desktop	175913
Googlebot-Image/1.0	unknown	159118
Mozilla/5.0 (Windows NT 6.3; Win64; x64; rv:64.0) ...	desktop	94603
Mozilla/5.0 (Windows NT 5.1; rv:52.0) Gecko/201001...	desktop	82152
Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (K...	desktop	68960

ภาพที่ 3.15 ตัวอย่างผลลัพธ์การ map ค่า device ของ user agent

3.3.10 Map IP address geolocation

แปลงค่า IP address ให้กลายเป็นประเทศโดยใช้ curl API ส่งค่า IP address ไปยังเว็บผู้ให้บริการข้อมูล IP address geolocation และนำค่าที่ return กลับมาบันทึกลงฐานข้อมูล

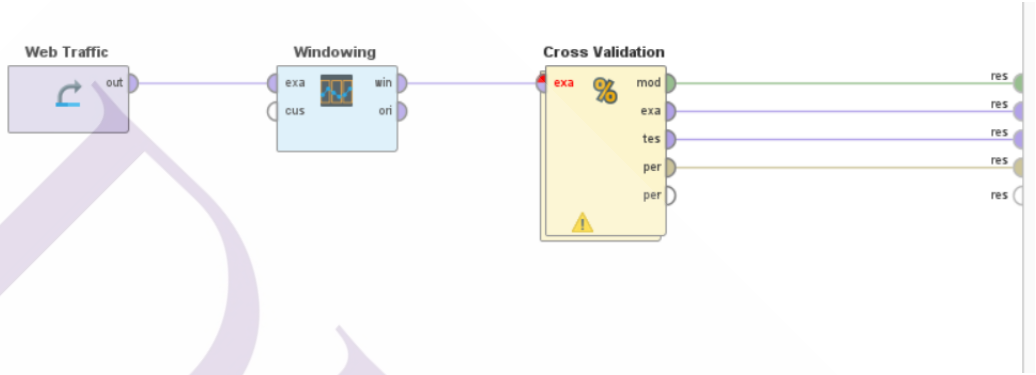
ip_address	code	country	capital
54.36.149.41	FRA	France	Paris
31.56.96.51	IRN	Iran	Tehran
40.77.167.129	USA	United States	Washington, D.C.
91.99.72.15	IRN	Iran	Tehran
66.249.66.194	USA	United States	Washington, D.C.
207.46.13.136	USA	United States	Washington, D.C.
178.253.33.51	USA	United States	Washington, D.C.
66.249.66.91	USA	United States	Washington, D.C.
5.78.198.52	DEU	Germany	Berlin
34.247.132.53	IRL	Ireland	Dublin
54.36.149.70	FRA	France	Paris
2.177.12.140	IRN	Iran	Tehran
89.199.193.251	IRN	Iran	Tehran
66.111.54.249	USA	United States	Washington, D.C.

ภาพที่ 3.16 ตัวอย่างผลลัพธ์การ map ค่า IP address ให้เป็นประเทศ

3.4 การสร้างแบบจำลอง

หลังจากได้ข้อมูลปริมาณการเข้าใช้งานเว็บไซต์แล้ว ก็นำไปสร้างแบบจำลองโดยงานวิจัยนี้ได้ทดลองกับแบบจำลอง Gradient Boosted Trees และ ARIMA model

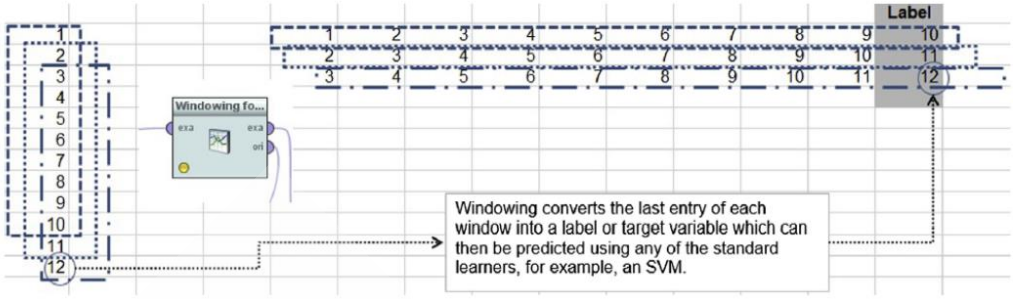
3.4.1 แบบจำลอง Gradient Boosted Trees



ภาพที่ 3.17 ตัวอย่างขั้นตอนการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ของแบบจำลอง Gradient Boosted Trees

โดยวิธีการเริ่มจากนำค่าของ Web access traffic มาสร้างตัวแปรปริมาณการเข้าใช้งานเว็บไซต์ย้อนหลัง โดยงานวิจัยนี้ใช้โอเปอร์เรเตอร์ Windowing ใน Rapidminer ในการแปลงข้อมูล Web access traffic ที่เป็นข้อมูลอนุกรมเวลาให้เป็นข้อมูลที่เหมาะสมต่อการใช้งาน โดยได้มีการปรับค่าพารามิเตอร์ของโอเปอร์เรเตอร์ Windowing เป็นแบบ Time based และได้มีการกำหนด Window size time ที่ต่างกัน

ในส่วนของ Cross Validation กำหนดพารามิเตอร์ Number of folds ที่ 10 folds และ Sampling type เป็น Automatic โดยค่าพารามิเตอร์ที่ใช้ในการสร้างแบบจำลอง Gradient Boosted Trees ได้แก่ Number of trees ที่ 100, maximal depth ที่ 5 และ Number of bins ที่ 20

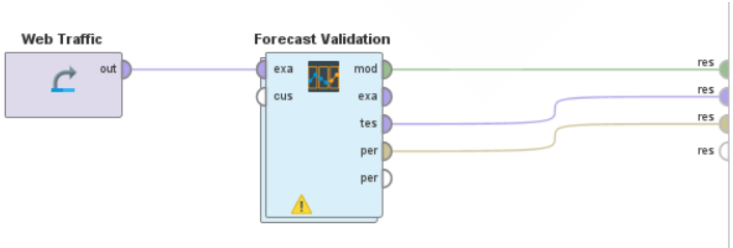


ภาพที่ 3.18 การทำ Windowing transformation สำหรับสร้างตัวแปรปริมาณการเข้าใช้งานย้อนหลัง

Last dat...	count + 1 (h...	prediction(c...	count - 719	count - 718	count - 717	count - 716	count - 715	count - 714	count - 713	count - 712	cou
Jan 22, 2019 ...	2098	2655.637	242	287	331	425	246	245	166	168	197
Jan 22, 2019 ...	1723	2135.007	218	206	209	190	211	175	520	327	214
Jan 22, 2019 ...	2108	1849.418	245	279	258	259	270	339	257	225	282
Jan 22, 2019 ...	1645	1910.574	455	457	211	386	420	358	543	568	599
Jan 22, 2019 ...	1634	1790.801	1055	1032	466	898	1236	1384	1246	1697	1432
Jan 23, 2019 ...	2064	1833.535	1760	2535	1890	2346	2024	2200	1626	2476	2058
Jan 23, 2019 ...	1597	1563.889	2070	2341	2218	1855	2567	1991	2626	2267	2024
Jan 23, 2019 ...	1880	1852.288	3355	3207	3510	2722	2149	2710	2079	2041	2704
Jan 23, 2019 ...	1653	1771.840	1816	2541	2741	2688	1067	2404	1709	3552	2684
Jan 23, 2019 ...	1151	1248.935	2614	2985	2751	2978	3441	2469	2184	2885	2451
Jan 23, 2019 ...	400	653.848	3573	3172	3164	3037	2307	2622	2961	2204	1971
Jan 23, 2019 ...	377	490.330	2023	1614	1840	2946	2866	2459	2366	2496	2134
Jan 23, 2019 ...	398	341.491	2098	1812	1784	2446	2421	2901	2221	2152	2115
Jan 23, 2019 ...	218	295.446	1723	1937	1962	2257	2010	1865	1741	2161	2338
Jan 23, 2019 ...	307	395.408	2108	2122	1849	1838	2368	1367	1869	2371	1926
Jan 23, 2019 ...	528	474.175	1645	1687	1693	1609	1646	2183	2145	2110	2244

ภาพที่ 3.19 ตัวอย่างผลลัพธ์การพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์โดยใช้แบบจำลอง Gradient Boosted Trees

3.4.2 แบบจำลอง ARIMA



ภาพที่ 3.20 ตัวอย่างขั้นตอนการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ของแบบจำลอง ARIMA

โดยวิธีการเริ่มจากนำค่าของ Web access traffic มาสร้างตัวแปรปริมาณการเข้าใช้งานเว็บไซต์ย้อนหลัง ก่อนนำเข้า ARIMA โดยใช้โอเปอร์เรเตอร์ Forecast Validation ใน Rapidminer ในการสร้างข้อมูลย้อนหลังแบบ Time based และได้มีการกำหนด Window size time ที่ต่างกันแบบเดียวกับแบบจำลอง Gradient Boosted Trees

ในส่วนของแบบจำลอง ARIMA กำหนดค่าพารามิเตอร์ p หรือ Autoregressive process เท่ากับ 2 , ค่า d หรือ Integrated เท่ากับ 0 และค่า q หรือ Moving average process เท่ากับ 1

datetime ↑	count	forecast of count	forecast position	Last datetime in window
Jan 23, 2019 1:26:00 AM ICT	1597	1794.981	1	Jan 23, 2019 1:25:00 AM ICT
Jan 23, 2019 2:26:00 AM ICT	1880	1867.502	1	Jan 23, 2019 2:25:00 AM ICT
Jan 23, 2019 3:26:00 AM ICT	1653	1600.779	1	Jan 23, 2019 3:25:00 AM ICT
Jan 23, 2019 4:26:00 AM ICT	1151	1298.478	1	Jan 23, 2019 4:25:00 AM ICT
Jan 23, 2019 5:26:00 AM ICT	400	791.054	1	Jan 23, 2019 5:25:00 AM ICT
Jan 23, 2019 6:26:00 AM ICT	377	411.580	1	Jan 23, 2019 6:25:00 AM ICT
Jan 23, 2019 7:26:00 AM ICT	398	265.494	1	Jan 23, 2019 7:25:00 AM ICT
Jan 23, 2019 8:26:00 AM ICT	218	282.438	1	Jan 23, 2019 8:25:00 AM ICT
Jan 23, 2019 9:26:00 AM ICT	307	283.251	1	Jan 23, 2019 9:25:00 AM ICT

ภาพที่ 3.21 ตัวอย่างผลลัพธ์การพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์โดยใช้แบบจำลอง ARIMA

3.5 เครื่องมือที่ใช้ในงานวิจัย

3.7.1 RapidMiner เป็นซอฟต์แวร์ ที่ใช้ทำงานด้าน Data Science ใช้ในการเตรียมข้อมูล การทำ Data Mining และ Machine learning รวมไปถึงการแปลงข้อมูลและการวิเคราะห์ข้อมูลได้โดยที่ไม่ต้องเขียนโค้ด เพราะตัวซอฟต์แวร์ออกแบบมาให้ใช้งานง่ายเพียงแค่ทำการ Drag and drop ให้ผู้ใช้งานได้ออกแบบ Workflow ในการวิเคราะห์ข้อมูลในหน้า Design view

3.7.2 XAMPP เป็นโปรแกรมสำหรับจำลองเครื่องคอมพิวเตอร์ส่วนบุคคลให้ทำงานในลักษณะของเว็บเซิร์ฟเวอร์ ซึ่งประกอบด้วย Apache, PHP, MySQL, PHP MyAdmin, Perl ซึ่งเป็นโปรแกรมพื้นฐานที่รองรับการทำงาน CMS สำหรับออกแบบเว็บไซต์

3.7.3 Laravel เป็น PHP Framework ที่ใช้ในการพัฒนาเว็บแอปพลิเคชันในรูปแบบ MVC (Model View Controller) มีจุดเด่นและข้อดีคือ ทำให้การเขียนโค้ดของเรานั้นดูสะอาดง่ายต่อการอ่านและแก้ไข และยังสามารถดาวน์โหลดมาใช้งานได้ฟรี

3.7.4 Bootstrap เป็น Frontend Framework แบบหนึ่งที่ช่วยให้เราสามารถสร้างหน้าเว็บให้ตรงตามแบบที่ต้องการได้ง่ายขึ้น เพราะมีทั้งระบบ Grid ที่ช่วยเรื่องการจัดวาง Layout ที่รองรับในการทำเว็บแอปพลิเคชันในรูปแบบ Responsive หรือให้เหมาะสมกับการแสดงผลบนมือถือและแท็บเล็ต และมี Component สำเร็จรูปให้ใช้งาน โดยนำส่วนของ HTML, CSS , JS มาพัฒนา

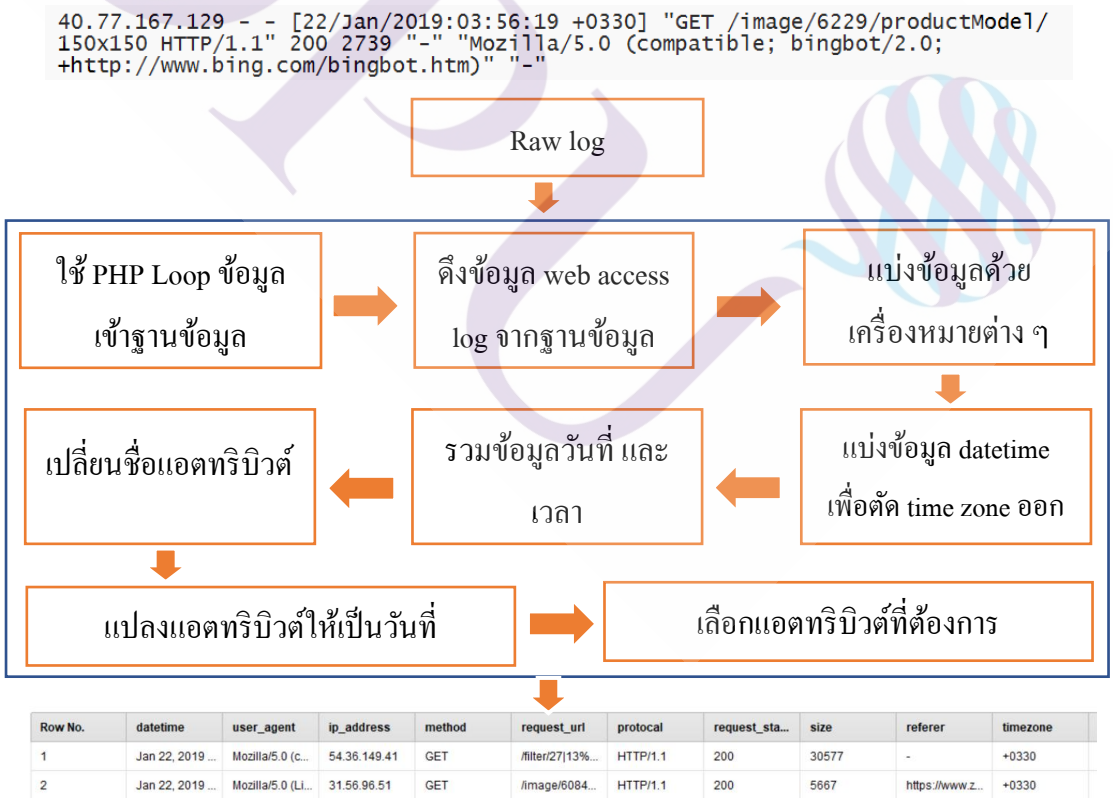
3.7.5 VS Code หรือ Visual Studio Code เป็นโปรแกรมประเภท Editor ที่ใช้ในการแก้ไขโค้ด สามารถใช้งานได้โดยไม่มีค่าใช้จ่าย รองรับหลายภาษาและสามารถเชื่อมต่อกับ Git ได้ง่ายและไม่ซับซ้อน มี Tools และ Extension ให้เลือกติดตั้งมากมาย รองรับการใช้งานภาษาอื่น ๆ ทั้ง ภาษา C++, Java, Python หรือ PHP

บทที่ 4

ผลการศึกษา

งานวิจัยนี้เป็นการวิเคราะห์ Web access log ผ่านเว็บแอปพลิเคชันในรูปแบบของ Dashboard เพื่อให้มองเห็นภาพรวมและพฤติกรรมของผู้ใช้งานเว็บไซต์ และนำไปสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) สำหรับพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ โดยมีรายละเอียดของผลการศึกษาดังต่อไปนี้

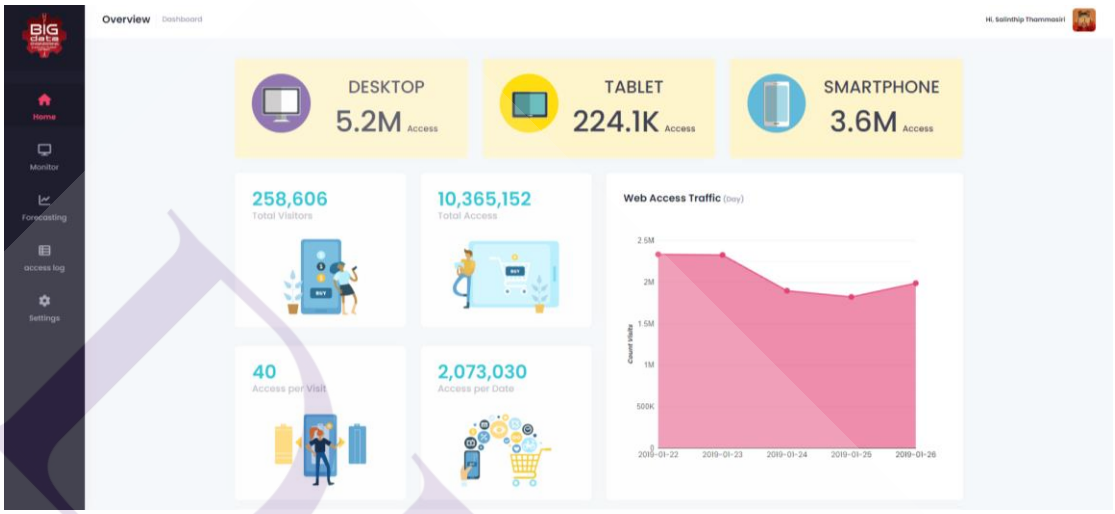
4.1 ผลการเตรียมข้อมูลก่อนทำการวิเคราะห์ข้อมูลพื้นฐานและสร้างแบบจำลอง



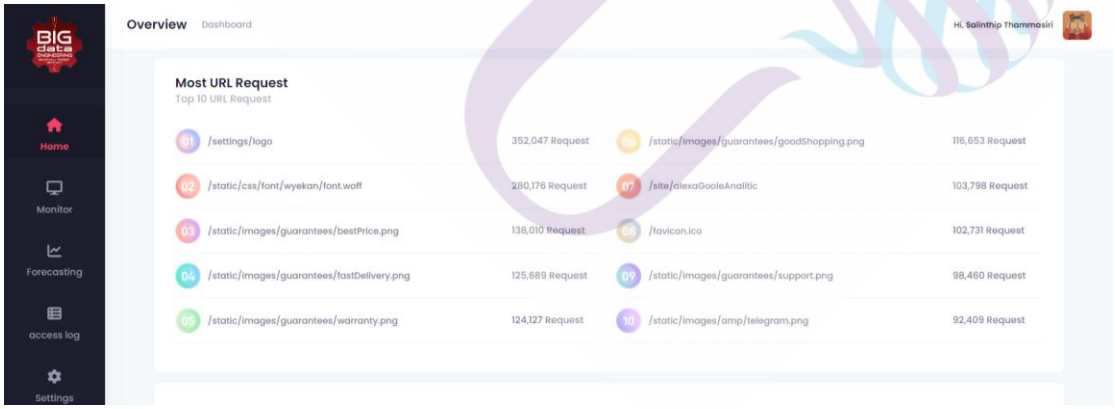
ภาพที่ 4.1 ผลการเตรียมข้อมูล

4.2 ผลการดำเนินงานในส่วนของการวิเคราะห์ข้อมูลพื้นฐาน

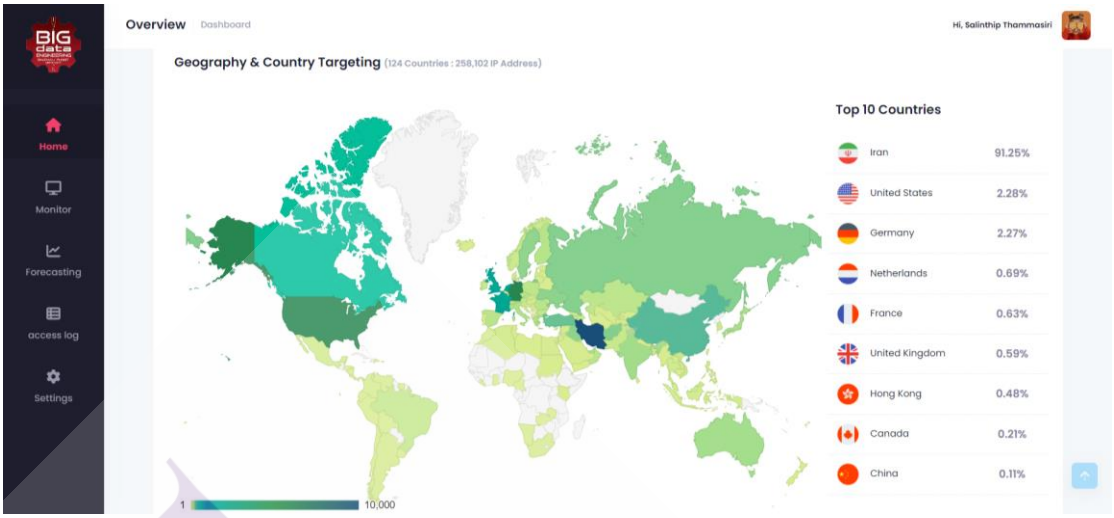
ผลลัพธ์ของการวิเคราะห์ข้อมูลพื้นฐาน จะถูกนำมาใช้ในการแสดงผลบน Dashboard ในรูปแบบกราฟและตาราง



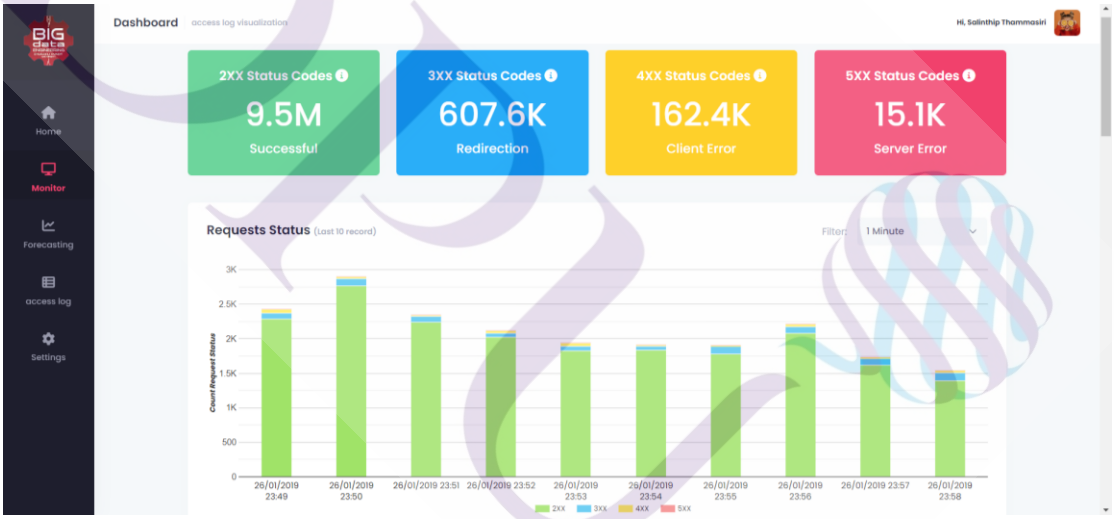
ภาพที่ 4.2 ตัวอย่างการแสดงผล User agent ที่ผ่านการ map ค่า device แล้วและกราฟแสดงข้อมูล traffic รายวัน



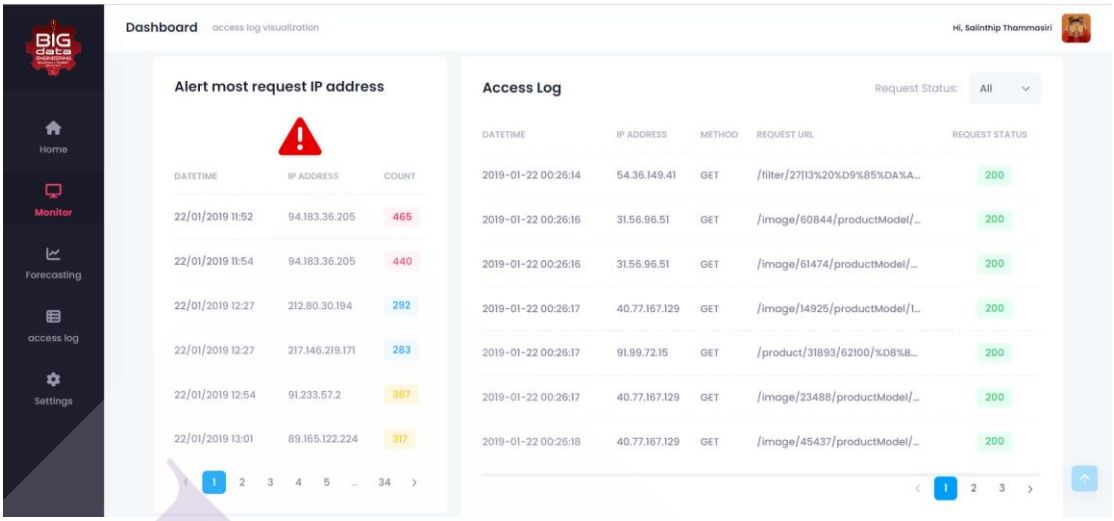
ภาพที่ 4.3 ตัวอย่างการแสดงผล 10 อันดับ URL ที่มีการ request มากที่สุด



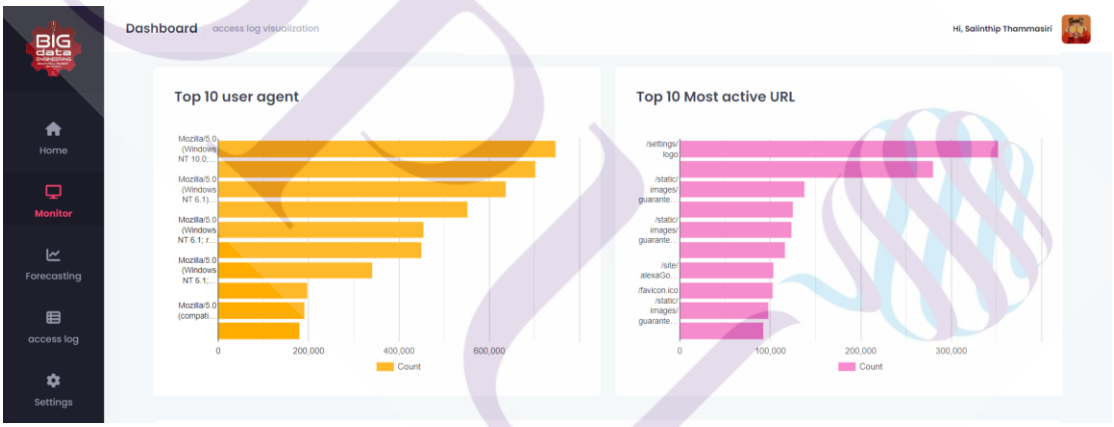
ภาพที่ 4.4 ตัวอย่างการแสดงผลข้อมูล 10 อันดับสูงสุดของประเทศที่เข้าใช้งานเว็บไซต์



ภาพที่ 4.5 ตัวอย่างการแสดงผลของ Request status บน Dashboard



ภาพที่ 4.6 ตัวอย่างการแสดงผลรวมของ IP Address ที่มีการ request เข้ามารายนาที และตารางแสดงข้อมูล access log



ภาพที่ 4.7 ตัวอย่างการแสดงผลข้อมูล 10 อันดับสูงสุดของ User agent และ URL ที่ผู้ใช้งานเรียกใช้มากที่สุด

ภาพที่ 4.8 ตัวอย่างการแสดงผลข้อมูล Traffic รายนาที และ dropdown สำหรับ filter ระยะเวลา

4.3 ผลการดำเนินงานในส่วนของการวัดประสิทธิภาพความถูกต้องของโมเดล

งานวิจัยนี้ได้ทำการทดลองสร้างแบบจำลอง Gradient Boosted Trees และ ARIMA โดยการใช้ข้อมูลหลังจากการเตรียมข้อมูลจำนวน 6,754 แถว โดยปรับค่าพารามิเตอร์ของ Windowing เพื่อใช้สำหรับเปรียบเทียบประสิทธิภาพ ค่า RMSE ค่า Absolute Error ค่า Relative Error ได้ผลดังตารางที่ 4.1

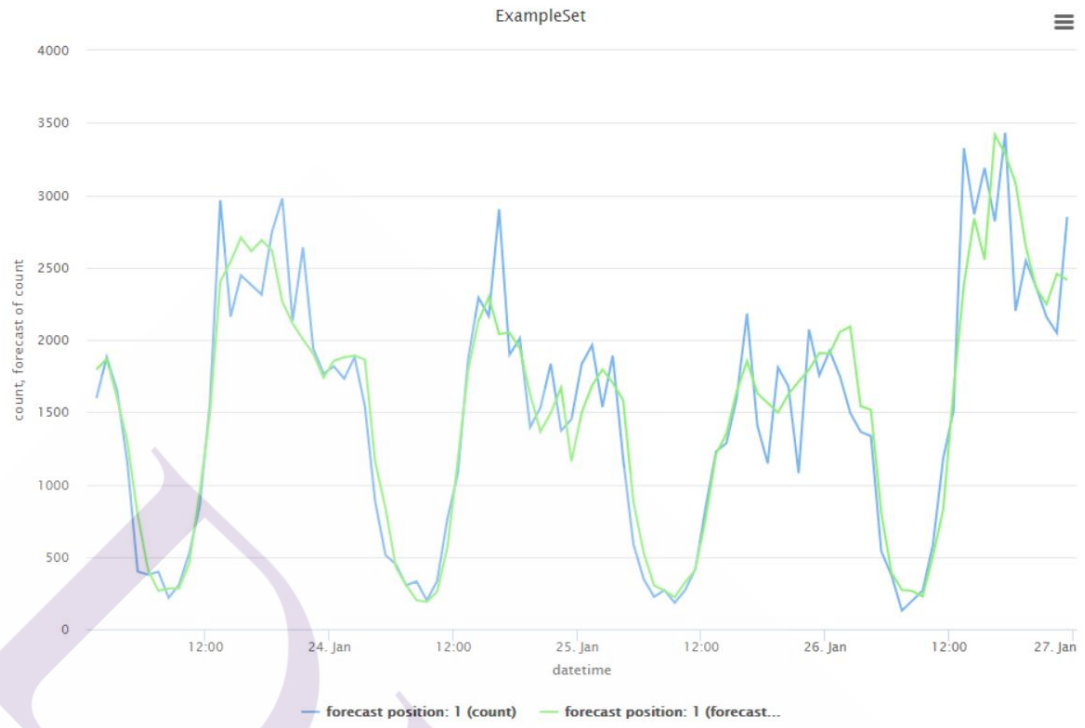
ตารางที่ 4.1 ตารางแสดงผลการทดสอบประสิทธิภาพของโมเดล Gradient Boosted Trees

Windowing (size time)			Performance		
Window	**Step**	**Horizon**	**RMSE**	**Absolute Error**	**Relative Error**
3 Hours	1 Hours	1 Minutes	366.210	277.051	21.24%
6 Hours	1 Hours	1 Minutes	395.264	291.497	21.36%
9 Hours	1 Hours	1 Minutes	373.654	275.233	20.24%
12 Hours	1 Hours	1 Minutes	335.589	257.569	20.93%
15 Hours	1 Hours	1 Minutes	314.679	244.693	21.54%
18 Hours	1 Hours	1 Minutes	340.993	259.125	21.47%
21 Hours	1 Hours	1 Minutes	343.225	256.483	22.85%
24 Hours	1 Hours	1 Minutes	350.484	263.864	20.60%

ตารางที่ 4.2 ตารางแสดงผลการทดสอบประสิทธิภาพของโมเดล ARIMA

Windowing (size time)			Performance		
Window	**Step**	**Horizon**	**RMSE**	**Absolute Error**	**Relative Error**
3 Hours	1 Hours	1 Minutes	239.639	239.639	17.90%
6 Hours	1 Hours	1 Minutes	249.171	249.171	18.64%
9 Hours	1 Hours	1 Minutes	221.621	221.621	17.46%
12 Hours	1 Hours	1 Minutes	221.478	221.478	17.91%
15 Hours	1 Hours	1 Minutes	220.450	220.450	18.29%
18 Hours	1 Hours	1 Minutes	215.373	215.373	18.04%
21 Hours	1 Hours	1 Minutes	217.468	217.468	18.33%
24 Hours	1 Hours	1 Minutes	222.689	222.689	17.72%

จากตารางผลการทดสอบของโมเดล Gradient Boosted Trees และ ARIMA พบว่าโมเดล ARIMA ที่ใช้ตัวแปร Window size time เป็น 18 Hours มีประสิทธิภาพในการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ที่ดีที่สุด โดยให้ค่า RMSE อยู่ที่ 215.373 ค่า Absolute Error อยู่ที่ 215. 373 และค่า Relative Error อยู่ที่ 18.04%



ภาพที่ 4.9 กราฟแสดงผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ของแบบจำลอง ARIMA โดยตั้งค่าตัวแปร Window size time เป็น 18 Hours

ภาพที่ 4.10 ตัวอย่างการแสดงผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์บน Dashboard

4.4 ผลการวัดความพึงพอใจของผู้ใช้งาน

การวัดผลความพึงพอใจของผู้ใช้งาน ทำการวัดผลจากทีมงานที่ดูแล Web server รวมถึงทีมงานอื่น ๆ ที่เกี่ยวข้อง โดยทำการวัดผลคำถามจำนวน 5 ข้อและข้อเสนอแนะอีก 1 ข้อ ซึ่งมีรายละเอียดดังนี้

4.4.1 คำถาม (Questionnaire) วัดความพึงพอใจ

1. โดยรวมแล้ว ท่านชอบเครื่องมือการแสดงผลการวิเคราะห์และการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ในรูปแบบ dashboard หรือไม่
2. เครื่องมือการแสดงผลการวิเคราะห์และการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ในรูปแบบ dashboard ตรงกับความต้องการในการทำงานของท่านหรือไม่
3. เครื่องมือนี้ทำให้การทำงานหรือตรวจสอบข้อมูลมีประสิทธิภาพมากขึ้นกว่าวิธีการเดิมของท่านหรือไม่
4. เครื่องมือนี้มีขั้นตอนการใช้งานที่สะดวกและง่ายหรือไม่
5. ท่านคิดว่า ท่านจะแนะนำ ให้ทีมงานที่ทำงานเดียวกันกับท่านต่อหรือไม่

4.4.2 กลุ่มเป้าหมาย (Target Users) ทั้งหมด 10 คนแบ่งเป็น

1. ทีมงานฝ่ายดูแล Web server จำนวน 5 ท่าน
2. ทีมงานฝ่ายพัฒนาเว็บไซต์ จำนวน 2 ท่าน
3. ทีมงานฝ่ายการตลาด จำนวน 3 ท่าน

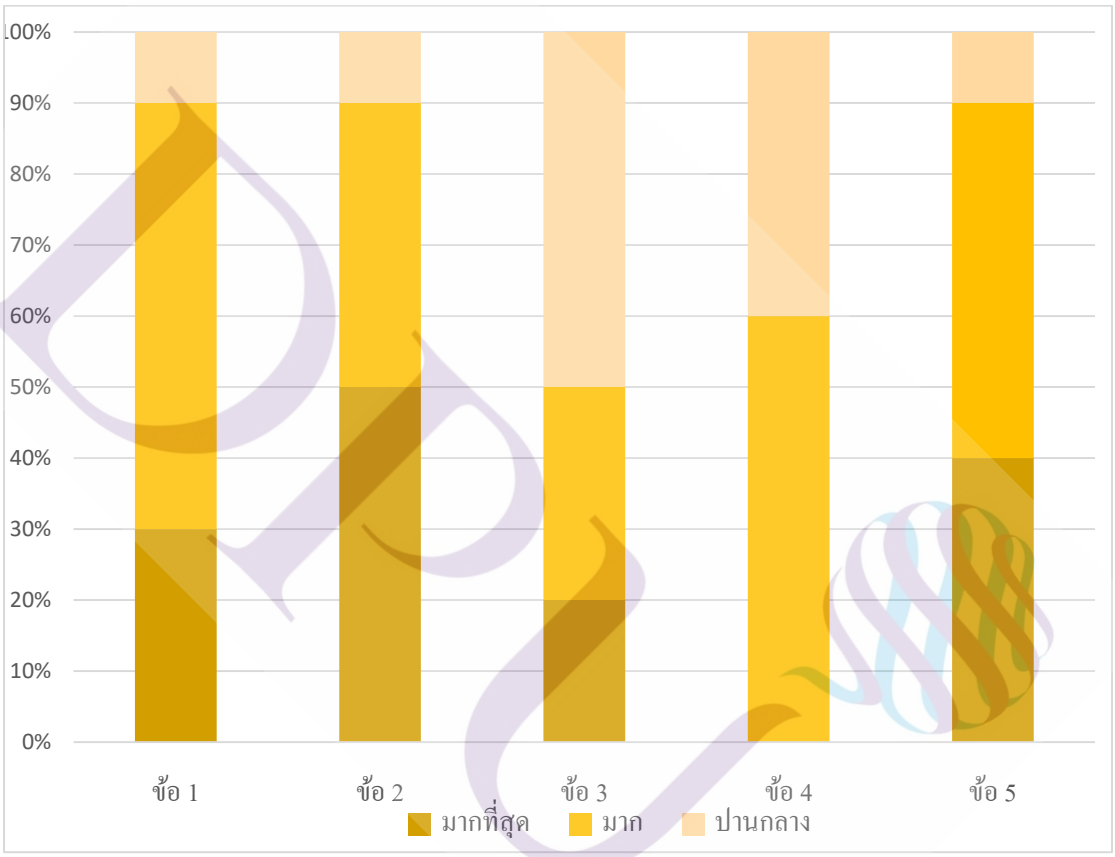
4.4.3 สรุปความคิดเห็นและข้อเสนอแนะสำหรับผลิตภัณฑ์ ทีมงานที่ดูและ Web server รวมถึงทีมงานอื่น ๆ ที่เกี่ยวข้อง จำนวน 10 ท่าน

1. ในส่วนของการกรองข้อมูลอยากให้มีการเลือกวันที่ได้ด้วย
2. อยากให้เพิ่มคำอธิบายของการแสดงแต่ละส่วน สำหรับคนที่ไม่รู้ใช้งานจะได้อ่านคำอธิบายแล้วเข้าใจ

4.4.4 ผลการประเมินความพึงพอใจ

จากผลการสำรวจความพึงพอใจของผู้ใช้งาน พบว่าโดยรวมแล้ว ผู้ใช้งานชอบเครื่องมือการแสดงผลการวิเคราะห์และการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์มากที่สุด 30% ชอบมาก 60% และชอบปานกลาง 10% และเครื่องมือนี้ตรงกับความต้องการของผู้ใช้งานมากที่สุด 50% มาก 40% และตรงกับความต้องการของผู้ใช้งานปานกลาง 10% นอกจากนี้ผู้ใช้งานให้คะแนนเรื่อง

เครื่องมือนี้ทำให้การทำงานมีประสิทธิภาพมากขึ้นหรือไม่ เห็นด้วยมากที่สุด 20% มาก 20% และปานกลาง 50% ในส่วนของมุมมองการใช้งานเครื่องมือ พบว่าการใช้งานเครื่องมือที่ง่ายมาก 60% และใช้งานง่ายปานกลาง 40% ในข้อสุดท้ายสำหรับการแนะนำให้กับผู้อื่นพบว่า ผู้ใช้งานจะแนะนำต่อแน่นอน มากที่สุด 40% มาก 50% และปานกลาง 10% โดยคะแนนในภาพรวมจะแสดงดังภาพที่ 4.11



ภาพที่ 4.11 กราฟแสดงผลการวัดความพึงพอใจของผู้ใช้งาน

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเกี่ยวกับการสร้างเครื่องมือที่แสดงผลการวิเคราะห์ข้อมูลในหลายมุมมอง เข้าใจง่ายในรูปแบบของ Dashboard และเพื่อพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยสามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการศึกษา

5.1.1 จากการดึงข้อมูลจำนวน 10,365,152 แถว จากฐานข้อมูลเข้าโปรแกรม Rapidminer เพื่อทำการวิเคราะห์ข้อมูล พบว่าใช้เวลาในการดึงข้อมูลนานและการดึงข้อมูลบางครั้งทำให้โปรแกรมค้าง ทางผู้วิจัยจึงแก้ปัญหาด้วยการดึงข้อมูลมาทำการ Feature extraction และ Cleansing data ให้ข้อมูลอยู่ในรูปแบบที่พร้อมใช้งานและทำการบันทึกข้อมูลที่ผ่านการ Clean แล้วเข้า Repository สำหรับนำมาใช้วิเคราะห์ข้อมูลต่อ เพื่อลดขั้นตอนและลดระยะเวลาในการดึงข้อมูลจากฐานข้อมูลและกระบวนการเตรียมข้อมูล

5.1.2 ข้อมูลแบบรายวันมีปริมาณเพียง 5 วัน ซึ่งเป็นปริมาณที่น้อยเกินไป ทำให้ไม่สามารถนำข้อมูลใช้กับแบบจำลองที่เป็น deep learning ได้ ทางผู้วิจัยจึงได้ใช้แบบจำลอง ARIMA และ Gradient Boosted Trees ในการสร้างแบบจำลองและเปรียบเทียบประสิทธิภาพ เนื่องจากโดยทั่วไปแล้วแบบจำลอง ARIMA จะใช้สำหรับการคาดการณ์อนุกรมเวลา จึงเหมาะที่จะนำมาใช้งานกับชุดข้อมูลของงานวิจัยนี้ที่เป็นข้อมูลแบบอนุกรมเวลา และในส่วนของแบบจำลอง Gradient Boosted Trees ก็เป็นอีกหนึ่งแบบจำลองที่สามารถนำมาใช้งานกับข้อมูลที่เป็นแบบอนุกรมเวลาได้ด้วย

5.1.3 ได้พัฒนาแบบจำลอง สำหรับพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ โอเปอร์เรเตอร์ Forecast Validation ที่ปรับค่าพารามิเตอร์ Window size time ที่ 18 Hours, Step ที่ 1 Hours และ Horizon ที่ 1 Minutes และโมเดล ARIMA ที่ปรับค่าพารามิเตอร์ p หรือ Autoregressive process เท่ากับ 2 , ค่า d หรือ Integrated เท่ากับ 0 และค่า q หรือ Moving average process เท่ากับ 1 ซึ่งให้ค่า RMAE อยู่ที่ 215.373 ค่า Absolute Error อยู่ที่ 215.373 และค่า Relative Error อยู่ที่ 18.04%

5.1.4 ได้พัฒนาเครื่องมือที่แสดงข้อมูลการวิเคราะห์พื้นฐานและข้อมูลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ในรูปแบบของ Dashboard บนเว็บแอปพลิเคชัน

5.2 ข้อสังเกต

จากผลการทดสอบพบว่าข้อมูลที่นำมาใช้งานกับแบบจำลองมีค่าต่ำสุด (min value) และค่าสูงสุด (max value) ของข้อมูลห่างกันเกินไป ส่งผลให้ค่า RMSE, Absolute error และ Relative Error ที่ใช้สำหรับวัดประสิทธิภาพของแบบจำลองสูง ซึ่งทั้งสามค่าที่ใช้วัดประสิทธิภาพของแบบจำลองนี้ยิ่งค่าน้อยจะยิ่งดี จากการทดลองกับแบบจำลองเดียวกันด้วยชุดข้อมูลอื่นที่มีค่าต่ำสุด (min value) และค่าสูงสุด (max value) ของข้อมูลที่ไม่ห่างกันมาก ส่งผลให้ตัววัดประสิทธิภาพของแบบจำลอง RMSE, Absolute error และ Relative error มีค่าที่ต่ำ

5.3 ข้อเสนอแนะ

5.3.1 ข้อมูลที่นำมาใช้งานเมื่อทำการ Cleaning data แล้วข้อมูลในรูปแบบวันที่มีปริมาณที่น้อย จึงทำให้เกิดความยากในการใช้งาน

5.3.2 ข้อมูล Access log ที่มีข้อมูลจำนวนมาก การดึงข้อมูลจากฐานข้อมูลปกติและการนำข้อมูลมาวิเคราะห์อาจใช้เวลานานและโปรแกรมอาจรันไม่ไหว อาจต้องเปลี่ยนจากการเก็บข้อมูลในฐานข้อมูลไปเก็บที่ Hadoop แทน

5.3.3 ถ้าข้อมูล Access log ที่นำมาใช้งานมีข้อมูลของ Destination IP ด้วยจะสามารถนำมาต่อยอดการวิเคราะห์อื่น ๆ ได้อีก เช่น Anomaly Detection, Social Network Analysis

5.3.4 พัฒนาประสิทธิภาพของโมเดล หรือทดลองทำโมเดลอื่นเพิ่มเติม เพื่อให้ได้ผลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์ที่แม่นยำมากขึ้น

บรรณานุกรม

บรรณานุกรม

เอกสิทธิ์ พัชรวงศ์ศักดา. (2563). *Practical Data Mining with RapidMiner Studio 9.* บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด

DiTC : ไขข้อข้องใจ Log File คืออะไร? ทำไมร้านกาแฟต้องจัดเก็บข้อมูล ตาม พ.ร.บ. คอมฯ. สืบค้น 19 พฤษภาคม 2565, จาก https://ditc.co.th/knowledge/log-file/

Frank Andrade(2021) : How To Easily Scrape Multiple Pages of a Website Using Python. สืบค้น 6 พฤษภาคม 2565, จาก https://betterprogramming.pub/how-to-easily-scrape-multiple-pages-of-a-website-using-python-73e85bd06f8c

Junyan Shao(2019) : Web Traffic Time Series Prediction Using ARIMA & LSTM. สืบค้น 23 พฤษภาคม 2565, จาก https://medium.com/@jyshao53/web-traffic-time-series-prediction-using-arima-lstm-7ef3911845ae

Jungkee Kim, (2018) : Web Server Log Visualization.

L.K. Joshila Grace1, V.Maheswari and Dhinaharan Nagamalai (2015) : Web Log Data Analysis and Mining

MDN Web Docs : HTTP response status codes สืบค้น 20 พฤษภาคม 2565, จาก https://developer.mozilla.org/en-US/docs/Web/HTTP/Status

Nuthdanai Wangpratham(2020) : การพยากรข้อมูลอนุกรมเวลาด้วยเทคนิค ARIMA ด้วย Python สืบค้น 10 มิถุนายน 2565, จาก https://nutdnuy.medium.com/การพยากรณ์ข้อมูลอนุกรมเวลาด้วยเทคนิค-arima-ด้วย-python-44809eb8e990

Tawfiq A. Al-asadi1 and Ahmed J. Obaid (2016) : Discovering similar user navigation behavior in Web log data

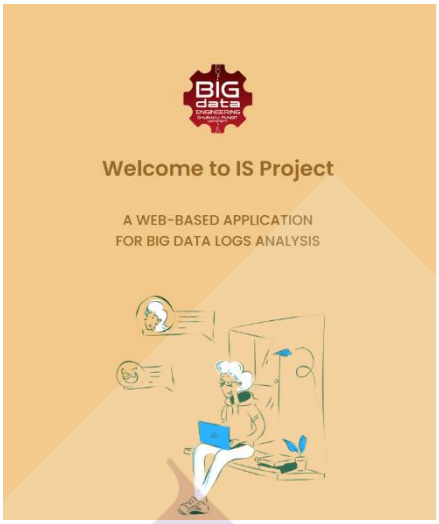
TECHSAUCE(2020) : ทำความรู้จัก Dashboard คืออะไร มีความสำคัญอย่างไร ทำไมควรทำ? สืบค้น 19 พฤษภาคม 2565, จาก https://techsauce.co/tech-and-biz/what-is-dashboard

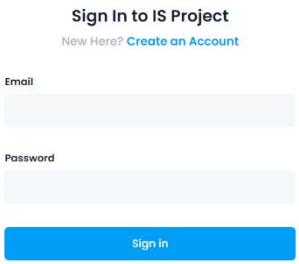
Tejas Shelatkar, Stephen Tondale, Swaraj Yadav and Sheetal Ahir, (2020) : Web Traffic Time Series Forecasting using ARIMA and LSTM RNN.

ภาคผนวก

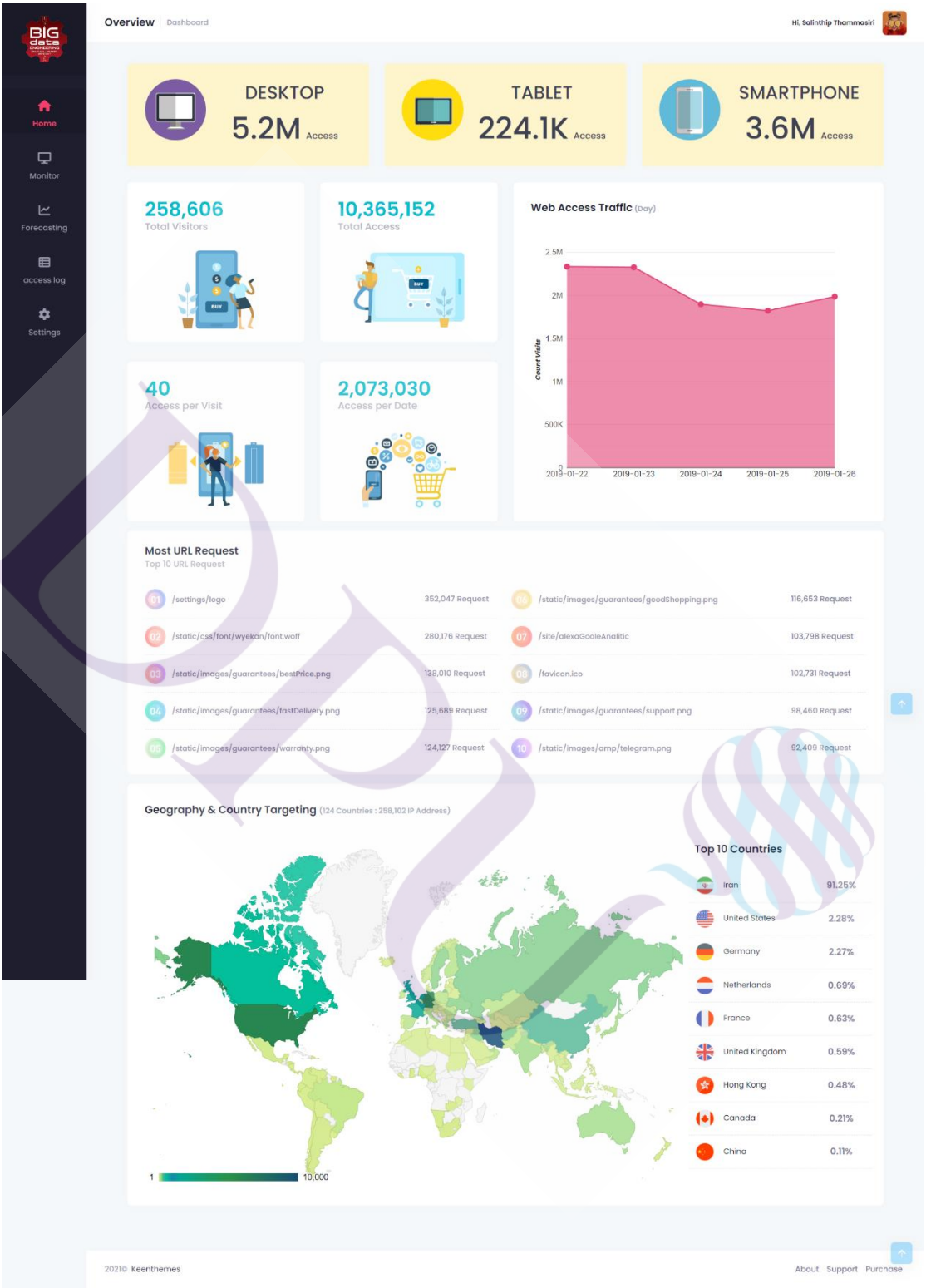
ภาคผนวก ก

เว็บแอปพลิเคชันสำหรับแสดงผลข้อมูลการวิเคราะห์ (Dashboard)

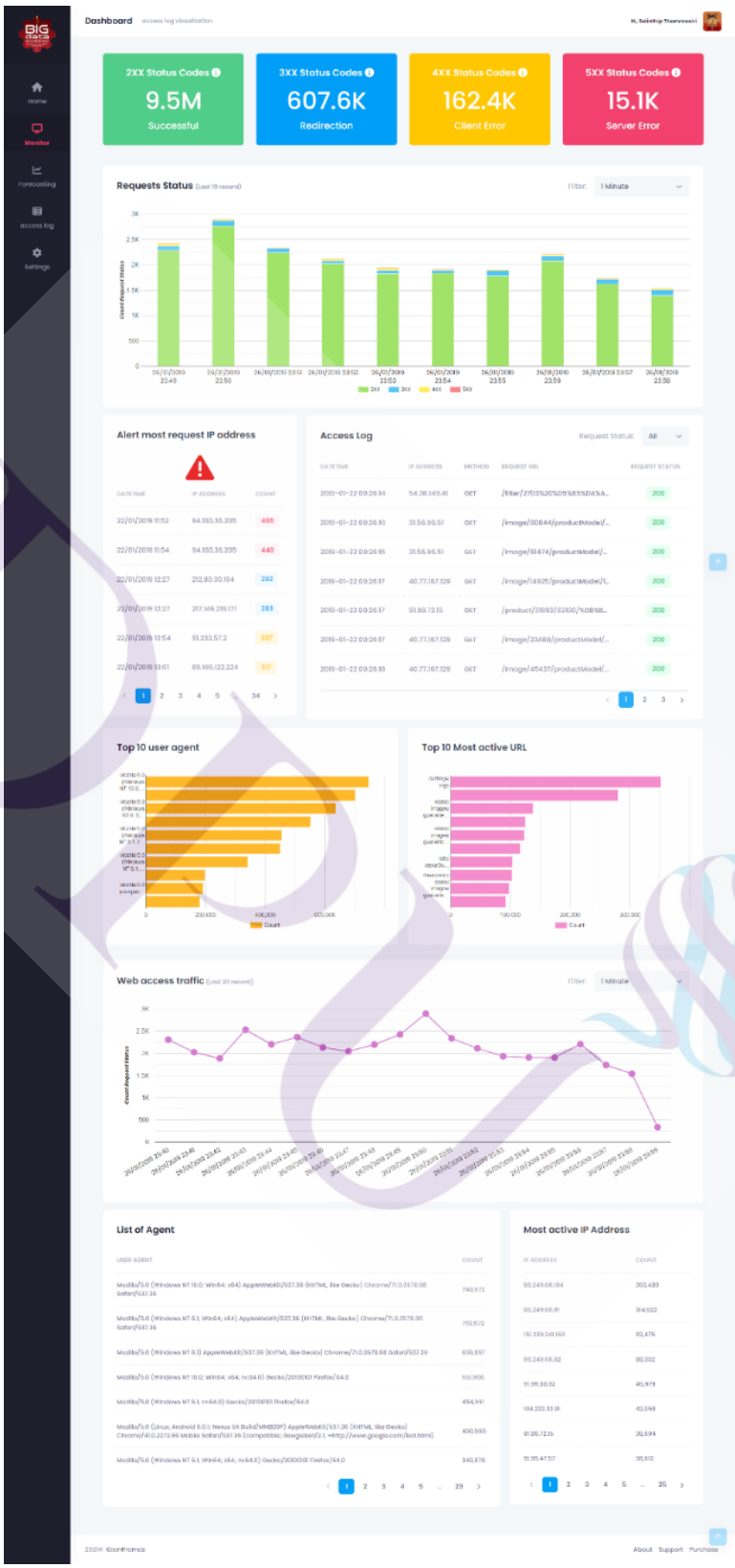




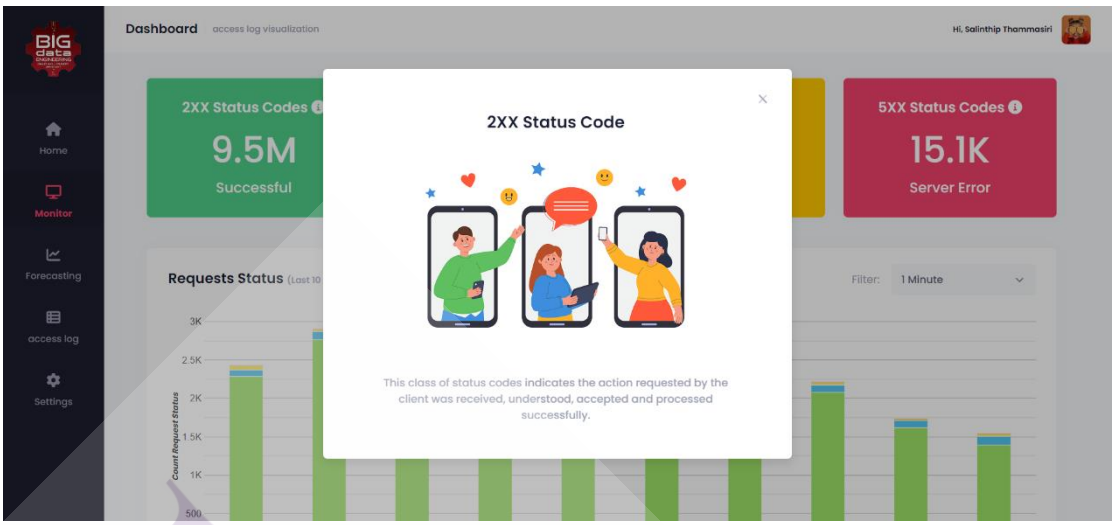
ภาพที่ 1 หน้าเว็บสำหรับ Login เข้าสู่ระบบ



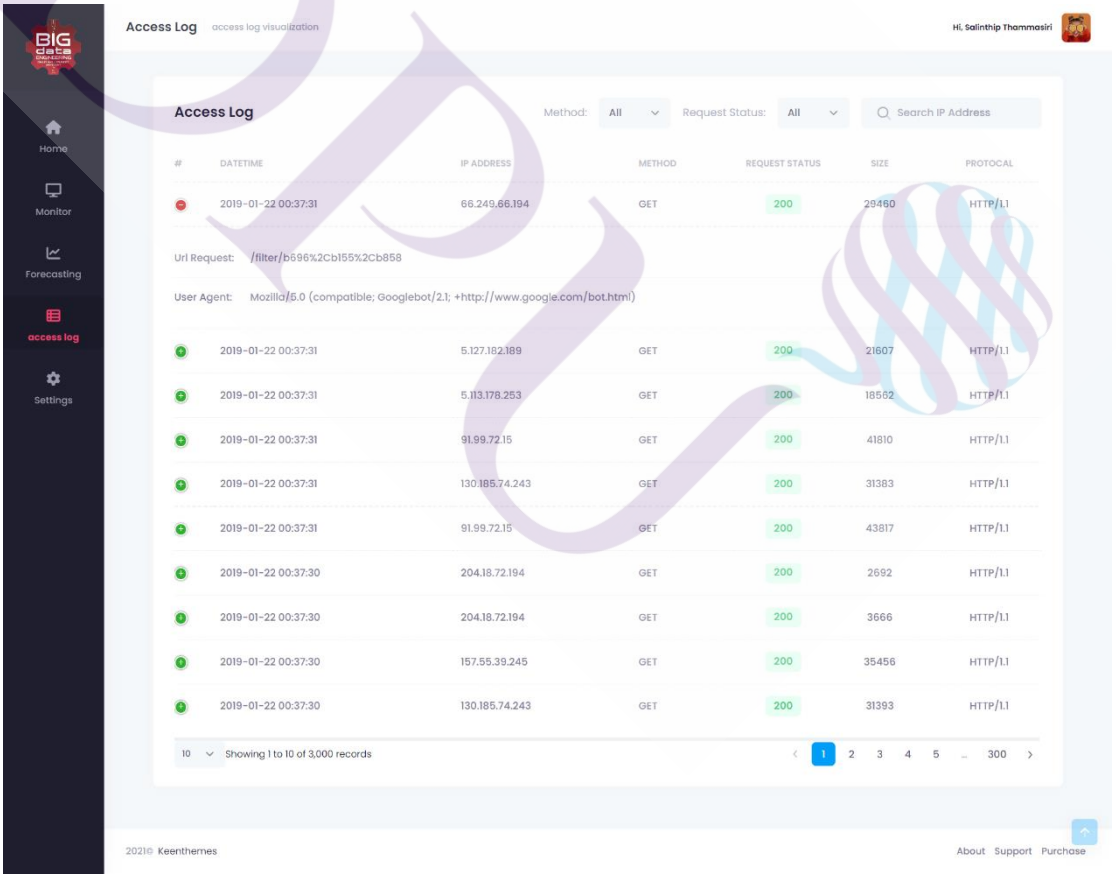
ภาพที่ 2 หน้า Overview dashboard แสดงภาพรวมของข้อมูล



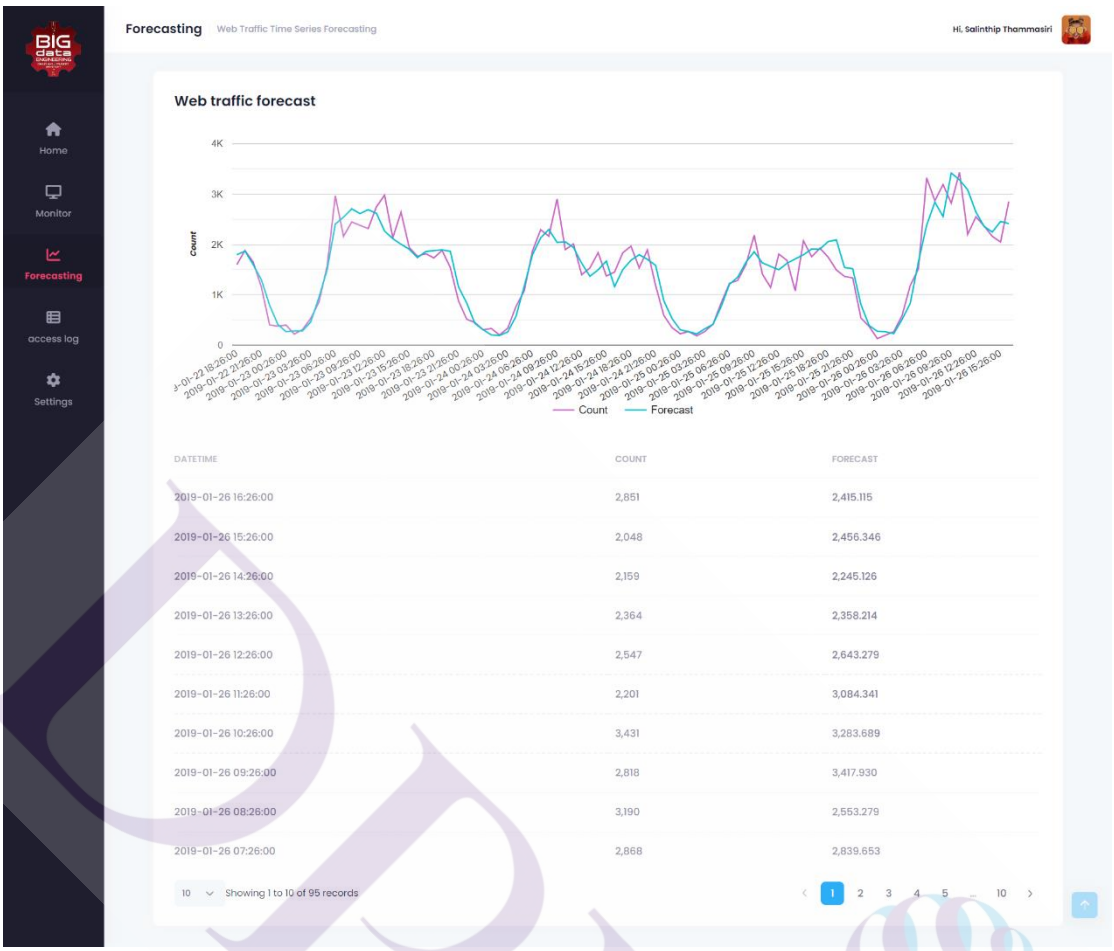
ภาพที่ 3 หน้า Monitor dashboard แสดงข้อมูลการวิเคราะห์ของ Access log



ภาพที่ 4 Modal popup แสดงข้อมูลคำอธิบายของ Request status code



ภาพที่ 5 ตารางแสดงข้อมูล Access log



ภาพที่ 6 ตารางแสดงผลข้อมูลการพยากรณ์ปริมาณการเข้าใช้งานเว็บไซต์

ประวัติผู้เขียน

ชื่อ-นามสกุล	สลิลทิพย์ ธรรมศิริ
ประวัติการศึกษา	บริหารธุรกิจบัณฑิต สาขาระบบสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์ ปีการศึกษา 2558
ตำแหน่งและสถานที่ทำงานปัจจุบัน	Programmer GOFX Thailand Co., Ltd.