Final Exam

1.จงอธิบายการประยุกต์ Machine learning ไปใช้กับงานตาม 11 หัวข้อดังนี้

- **Understand Business Requirement**

ในปัจจุบันตลาดประกันสุขภาพ (Health Insurance) มีมูลค่าสูงถึง 23% ของตลาดประกัน ภัยในปัจจุบัน และมีอัตราการเติบโตต่อปีเฉลี่ย 5% และในช่วงที่ผ่านมาเป็นกลุ่มที่มีแนวโน้มโตเร็วที่สุด ซึ่งในปัจจุบันนั้นได้มีการหลากหลายบริษัทได้นำเทคโนโลยีมาร่วมประยุกต์ใช้กับการซื้อขายประกันภัย โดยแนวโน้มที่เห็นได้มากยิ่งขึ้น คือ Personalize insurane หรือประกันภัยที่ออกแบบมาเฉพาะแต่ละบุคคล ซึ่งการที่จะเช่นนั้นได้เราจำเป็นต้องมีการนำเทคโนโลยี Machine learning มาช่วยในการสร้างเทคโนโลยีตัวนี้ขึ้นมา

ดังนั้นทางกลุ่มจึงเล็งเห็นถึงความสำคัญ และความเป็นไปได้ในอนาคตที่จะมีแนวทางนี้เกิดขึ้นอย่างแพร่หลายในประเทศไทย จึงเลือกที่จะนำข้อมูลด้านประกันสุขภาพมาใช้ในการทำรายงานในครั้งนี้ โดยเราตั้งวัตถุประสงค์ของการทำในครั้งนี้คือ

1.เพื่อออกแบบระบบที่สามารถคำนวณเบี้ยประกันที่เหมาะสมให้แก่ลูกค้าในแต่ละคนได้

2.เพื่อระบุปัจจัยที่ส่งผลต่อการคำนวณเบี้ยประกันของลูกค้าในบริษัท

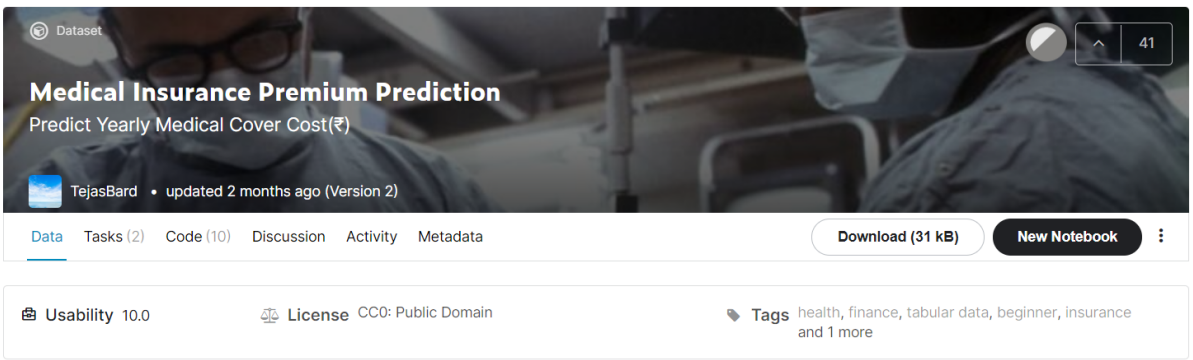
3.เพื่อวิเคราะห์ข้อมูล และหาความสัมพันธ์ของชุดข้อมูล จนสามารถนำไปสู่การวางแผนแนวทางบริษัทในอนาคตได้

- **Data Acquistion (การรวบรวมข้อมูล)**

ในการจัดทำรายงานในครั้งนี้ ทางกลุ่มได้ทำการสืบค้น และทำการคัดเลือกข้อมูลที่สามารถตอบโจทย์ความต้องการของธุรกิจประกันภัยสุขภาพจากข้อที่ 1 ผ่าน Public Dataset บนเว็บไซต์ Kaggle ซึ่งข้อมูลที่ใช้สำหรับการทำรายงานในครั้งนี้ คือ

Medical Insurance Premium Prediction : ซึ่งเป็นข้อมูลลูกค้าจำนวน 1000 รายจากบริษัทประกันภัยแห่งหนึ่ง เนื่องจากข้อมูลดังกล่าวเป็นข้อมูลที่ไม่สามารถระบุตัวตนได้ ทางกลุ่มจึงคาดว่าข้อมูลดังกล่าวสามารถนำมาใช้ในการศึกษาครั้งนี้ได้

ที่มา : https://www.kaggle.com/tejashvi14/medical-insurance-premium-prediction



ภาพที่ 1 : ภาพหน้าตาเว็บไซต์ข้อมูล Data set ที่นำมาใช้

- **Data Preparation (การจัดเตรียมข้อมูล)**

Medical Premium Dataset			
Attributes : 11 | Instance : 986 | Sum of weight : 986			
No	Column name	Type	missing value
1	Age	Numeric	0%
2	Diabetes	Nominal	0%
3	BloodPressureProblems	Nominal	0%
4	AnyTransplants	Nominal	0%
5	AnyChronicDiseases	Nominal	0%
6	Height	Numeric	0%
7	Weight	Numeric	0%
8	KnowAllergies	Nominal	0%
9	HistoryofCancerInFamily	Nominal	0%
10	NumberofMajorSurgeries	Numeric	0%
11	PremiumPrice	Numeric	0%

Age	Diabetes	BloodPressureProblems	AnyTransplants	AnyChronicDiseases	Height	Weight	KnownAllergies	HistoryOfCancerInFamily	NumberOfMajorSurgeries	PremiumPrice
45	0	0	0	0	155	57	0	0	0	25000
60	1	0	0	0	180	73	0	0	0	29000
36	1	1	0	0	158	59	0	0	1	23000
52	1	1	0	1	183	93	0	0	2	28000
38	0	0	0	1	166	88	0	0	1	23000
30	0	0	0	0	160	69	1	0	1	23000
33	0	0	0	0	150	54	0	0	0	21000
23	0	0	0	0	181	79	1	0	0	15000
48	1	0	0	0	169	74	1	0	0	23000
38	0	0	0	0	182	93	0	0	0	23000
60	0	1	0	0	175	74	0	0	2	28000
66	1	0	0	0	186	67	0	0	0	25000
24	0	0	0	0	178	57	1	0	1	15000
46	0	1	0	0	184	97	0	0	0	35000
18	0	0	1	0	150	76	0	0	1	15000
38	0	0	0	0	160	68	1	0	1	23000
42	0	0	0	1	149	67	0	0	0	30000
38	1	0	0	0	154	82	0	0	0	23000
57	1	0	0	0	156	61	0	0	0	25000
21	0	1	0	0	186	97	0	0	0	15000
49	1	0	0	0	160	97	0	0	2	28000
20	1	0	0	0	181	81	0	0	0	15000
35	0	0	0	0	163	92	0	0	1	32000
35	0	1	0	0	175	83	0	0	1	23000
53	0	1	0	0	151	97	0	1	1	35000
31	0	0	0	0	172	57	0	0	0	21000
22	0	0	1	0	151	97	0	0	0	15000
60	0	1	0	0	151	88	0	0	2	28000
30	0	0	0	1	162	73	1	0	0	23000

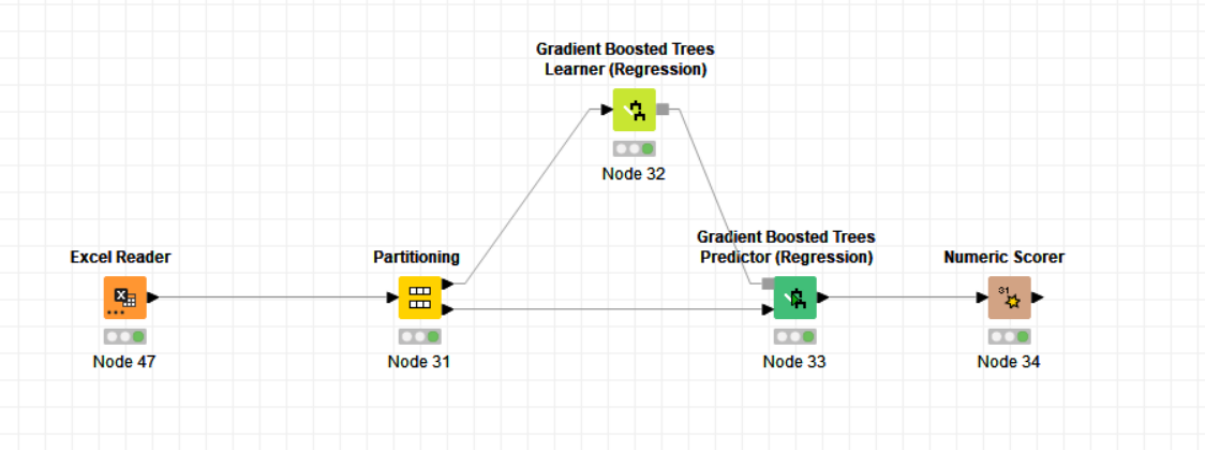
ภาพที่ 2 : ตัวอย่างภาพชุดข้อมูลที่ใช้ในการเทรนข้อมูล

- **Exploratory Data Analysis (การสำรวจข้อมูลเบื้องต้น)**
  - Age : อายุของลูกค้า
  - Height : ส่วนสูง
  - Weight : น้ำหนัก
  - Diabetes : ประวัติการเป็นโรคเบาหวาน
  - BloodPressureProblems : ประวัติปัญหาเกี่ยวความดันโลหิต
  - AnyTransplants : ประวัติการปลูกถ่ายอวัยวะ
  - AnyChronicDiseases : ประวัติการเป็นโรคเรื้อรัง
  - KnowAllergies : ประวัติการรับรู้อาการแพ้
  - HistoryofCancerInFamily : ประวัติผู้ป่วยเป็นโรคมะเร็งในครอบครัว
  - NumberofMajorSurgeries : จำนวนครั้งในการผ่าตัดใหญ่ที่ผ่านมา
  - PremiumPrice : ราคาประกันภัย (รายปี)

2.จากข้อที่ 1 จงแสดงผลการ run ด้วย Knime ในการทำ model อย่างน้อย 3 algorithms ที่แตกต่างกัน และเลือกที่ดีสุดมาทำต่อในขั้นตอนต่อไป

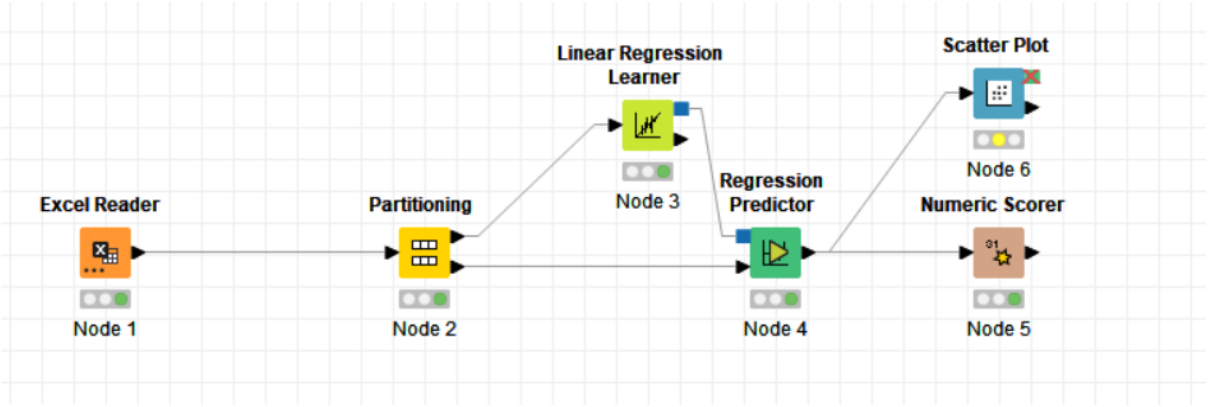
- **Modeling & Evaluation**

1. **Gradient Boosted Trees (Regression) algorithms**



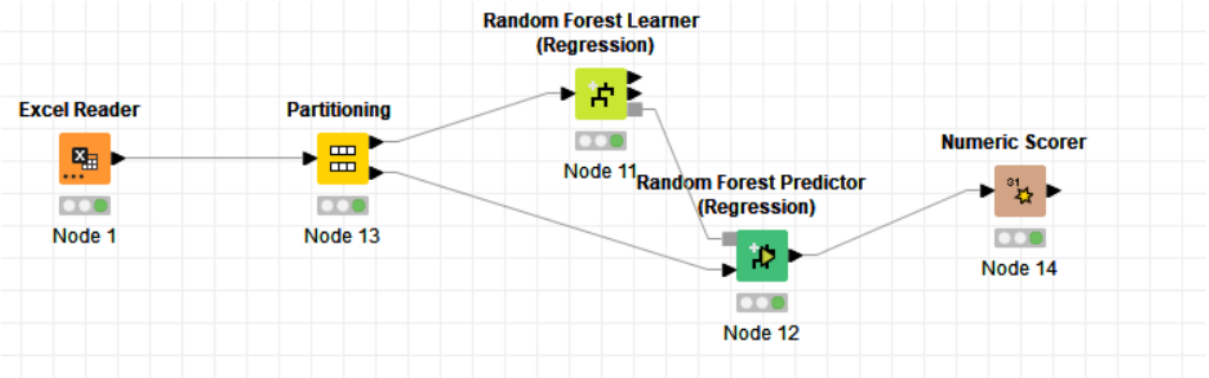
ภาพที่ 3 : Workflow Gradient Boosted Trees (Regression) algorithms

2. **Linear regression algorithms**



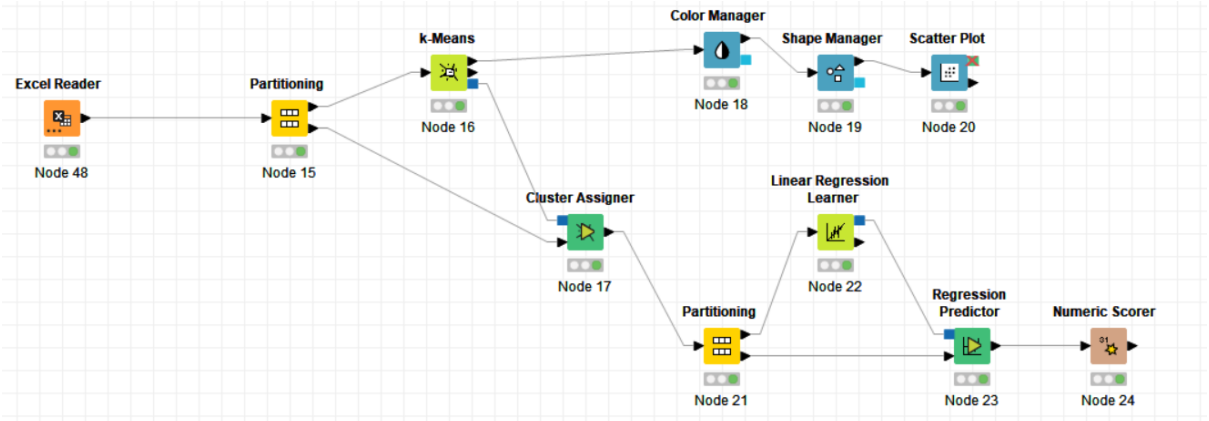
ภาพที่ 4 : Workflow Linear regression algorithms

3. **Random Forest learner (Regression) algorithms**



ภาพที่ 5 : Workflow Random Forest learner (Regression) algorithms

4. **K-Means clustering & linear regression algorithms**



ภาพที่ 6 : K-Means clustering & linear regression algorithms

- **Review Result**

Algorithms	R-Squared	Mean absolute error	Mean squred error
Gradient Boosted Trees (Regression) algorithms	0.803	1,073.921	7,067,584.475
Linear Regression	0.66	2,882.87	14,923,033.855
Random Forest learner (Regression) algorithms	0.781	1,662.3	7,429,765.088
K-Means clustering & linear regression algorithms	0.916	1,241.967	2,837,020.905

จากตารางข้างต้น ทางกลุ่มจะให้ความสำคัญในส่วนของค่า R-Squrared เป็นอย่างมาก เนื่องจากผลลัพธ์ที่ทางกลุ่มต้องการจากการทำโมเดลในครั้งนี้ คือ ความสามารถในการทำนายราคาประกันภัยของลูกค้าในแต่ละคนได้อย่างแม่นยำ ดังนั้น algorithms K-Means clustering & linear regression จึงเป็นโมเดลที่เหมาะสมแก่การใช้งานมากที่สุด ซึ่งทางกลุ่มได้นำโมเดลตัวนี้ไปใช้ในการสร้าง Dashboard สำหรับใช้ภายในองค์กร ดังภาพในโจทย์ข้อที่ 3 หัวข้อ Data visulization

- **Model Deployment**

โดยในครั้งนี้ทางกลุ่มได้มีการ Deploymodel เพื่อนำไปใช้ในตัว Line Chatbot โดยตัวของโมเดลเราได้ทำการเลือกรูปแบบของอัลกอริทึมที่ได้ผล 3 อันดับแรกในโปรแกรม KNIME จากข้อข้างบนมาทำการ Coding ด้วยภาษา Python และนำ Model ที่ได้จากการเขียน Python ไป Deploy ผ่าน Heroku โดยโมเดลที่สามารถทำการ Deploy ได้คือ Random Forest learner (Regression) algorithms เนื่องจากขนาดของไฟล์โมเดลที่สามารถ Deploy ได้มีขนาดที่จำกัด
ดังนั้น จึงนำอัลกอริทึมดังกล่าวไปใช้บน Chatbot

ภาพที่ 7 : ภาพเครื่องมือสำหรับการ Deploy model

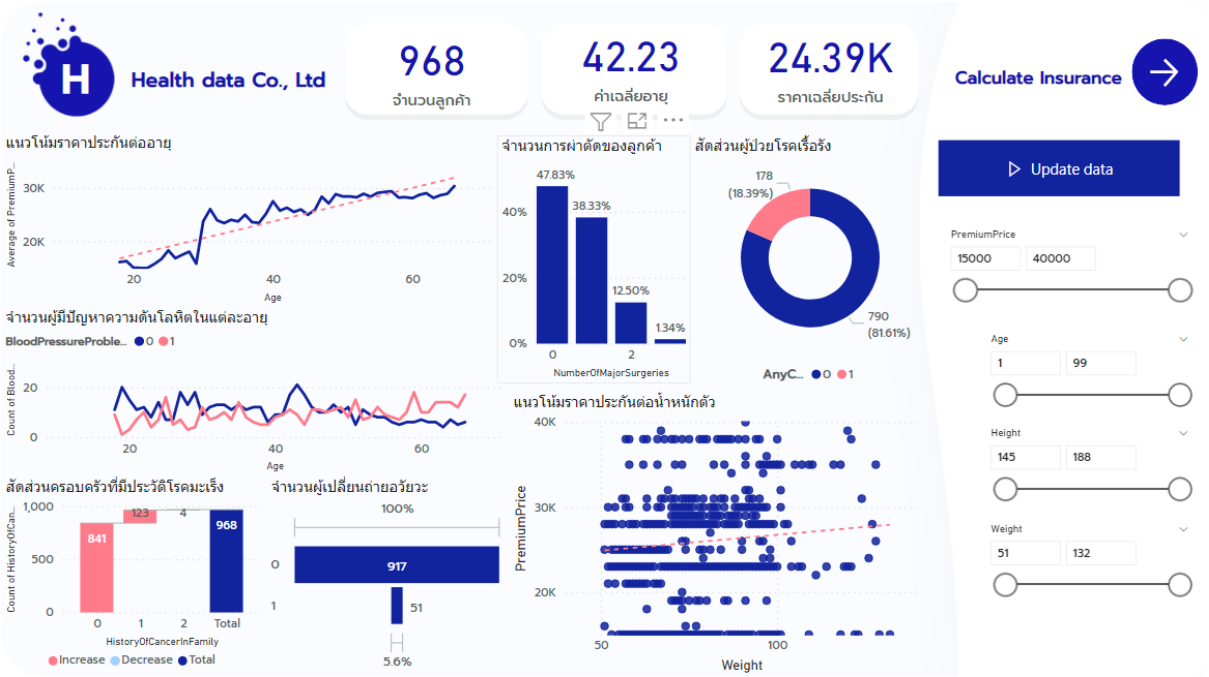
3.จากข้อที่ 1 แสดงผลการทำ Data visualization

Link Digital Dashboard :

https://app.powerbi.com/view?r=eyJrIjoiMzY0NDU2NjEtNjE3OS00ZmZhLWIwNzAtMWU1YzViYTk2ZWM1IiwidCI6IjZmNDQzMmRjLTIwZDItNDQxZC1iMWRiLWFjMzM4MGJhNjMzZCIsImMiOjEwfQ%3D%3D&pageName=ReportSection

ภาพที่ 8 : ภาพ QR Code สำหรับการเข้าใช้งาน Digital Dashboard

- **Real world Testing**



ภาพที่ 9 : ภาพหน้าตา Overview ของ Digital Dashboard

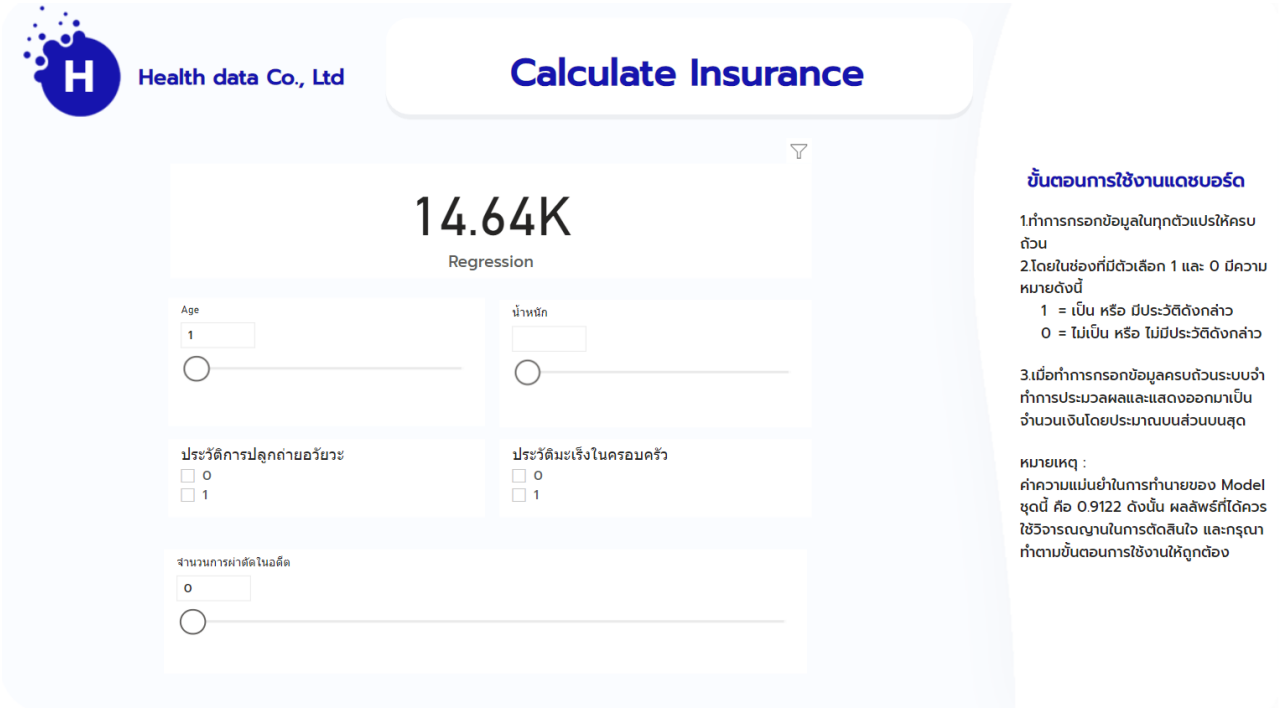
ในส่วนของข้อมูลที่ได้รับมา ทางกลุ่มก็ได้นำไปสร้างเป็น Dashboard สำหรับการ สรุปผลข้อมูลให้สามารถเข้าใจได้มากยิ่งขึ้น ได้แสดงให้เห็นถึงความสัมพันธ์ของข้อมูลในเชิงสถิติพรรณา (Descriptive statistics) ทำให้ผู้ใช้งาน Dashboard สามารถนำไปช่วยในการตัดสินใจในการดำเนินธุรกิจ โดย Dashboard ที่จัดทำได้มีการแบ่งออกเป็น 2 หน้าดังนี้

1.หน้า Overview เป็นหน้าสำหรับการแสดงผลข้อมูลโดยรวม และแสดงปัจจัยหรือข้อมูลที่มีความสัมพันธ์ให้เห็นได้ชัดเจนมากยิ่งขึ้น

2.หน้าประเมินราคาประกันลูกค้า ซึ่งส่วนนี้จะมีการนำสมการที่ได้จาก Model K-Means clustering & linear regression มาใช้ในการประเมินบนหน้า Dashboard ในหน้าที่ 2 เพื่อให้ทางบริษัทสามารถนำไปใช้ในการตัดสินใจในการจัดตั้งราคาเบี้ยประกัน หรืออื่นๆ

- **Business Alignment**

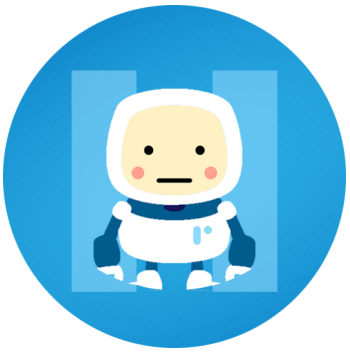
ซึ่งในครั้งนี้เราได้จัดทำ Dashboard ที่มานำตัว Model Machine learning มาใช้ในการทำนายข้อมูลในหน้าแดชบอร์ดเช่นกัน แต่เนื่องจากความแม่นยำของข้อมูลที่ประมาณ 91.22% เท่านั้น ดังนั้นเราจึงจัดตำแหน่งให้ตัว Dashboard ของเราเป็นตัวผู้ช่วยสำหรับการตัดสินใจในเบื้องต้นเพียงเท่านั้น หากต้องทำการตัดสินใจ ควรใช้วิจารณญาณและประสบการณ์ ความเชี่ยวชาญ ในการสนับสนุนการตัดสินใจในครั้งนั้นด้วย



ภาพที่ 10 : ภาพหน้าตา Calculate Insurance ของ Digital Dashboard

4.จากข้อที่ 1 ลองออกแบบ Chatbot ที่สามารถใช้ในงานได้

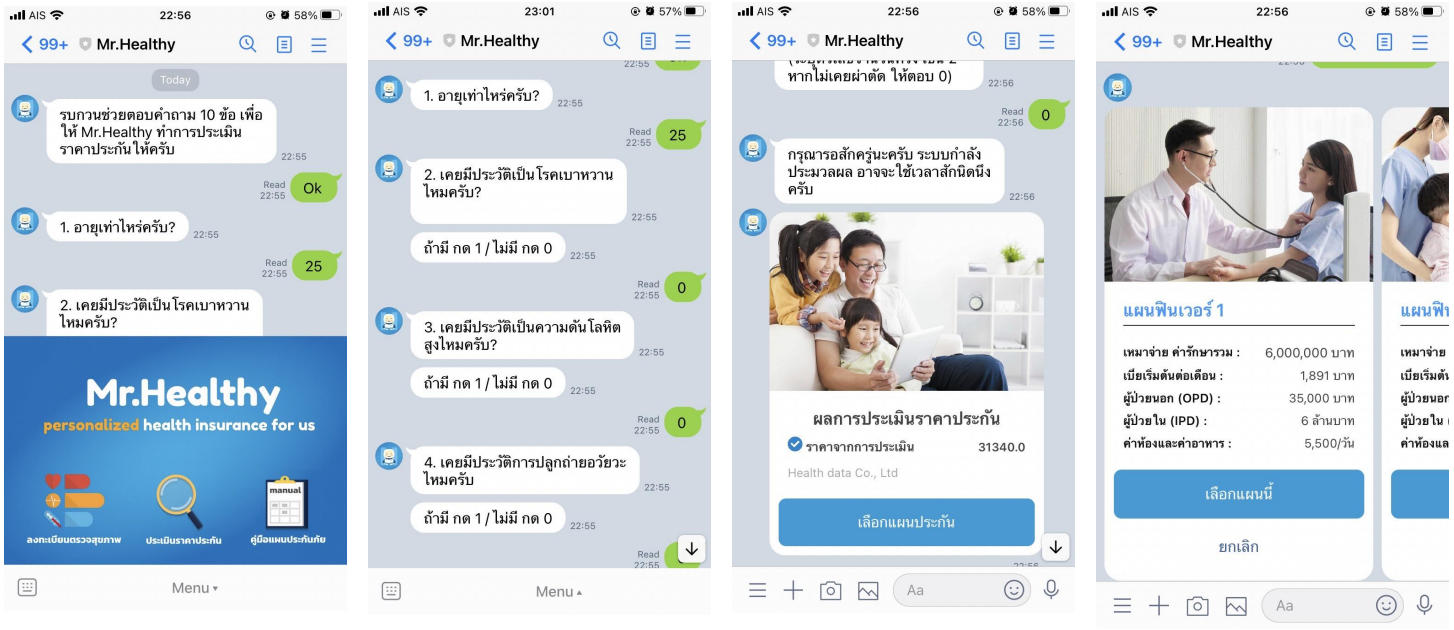
- **Operationalize**



ชื่อแชทบอท : Mr.healthy

ลักษณะของแชทบอท : เป็นแชทบอทที่ผู้ใช้สามารถเข้ามาเพื่อประเมินราคาประกันภัยที่เหมาะสมกับแต่ละบุคคล โดยแชทบอทตัวนี้จะมีการเชื่อมต่อกับ Machine learning ที่เชื่อมต่อผ่าน API เมื่อผู้ใช้งานกรอกข้อมูลครบถ้วนจะได้ผลของประกันภัย และราคาที่เหมาะสมกับคุณขึ้นมา

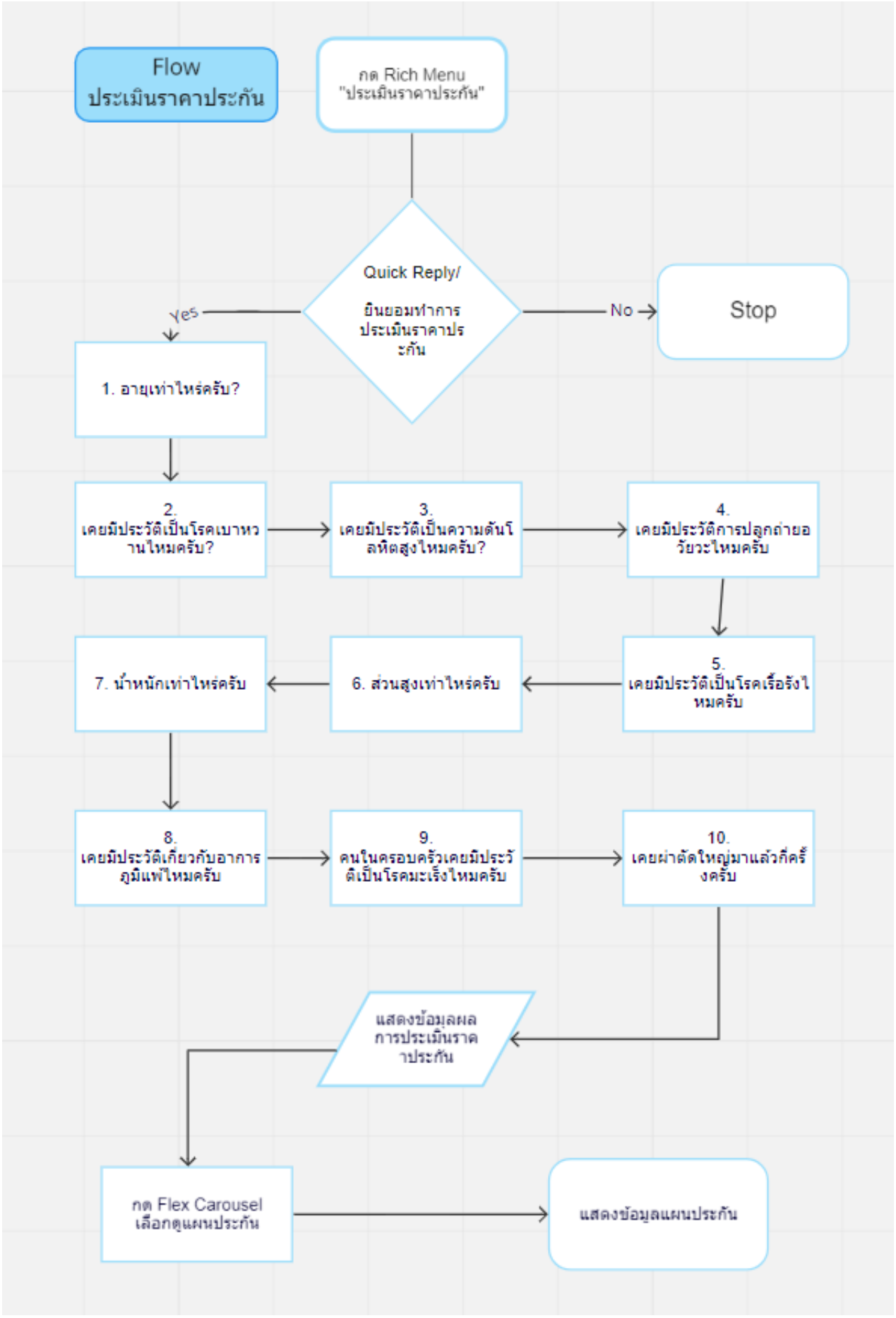
ภาพที่ 11 : ภาพหน้าตาChatbot Mr.healthy



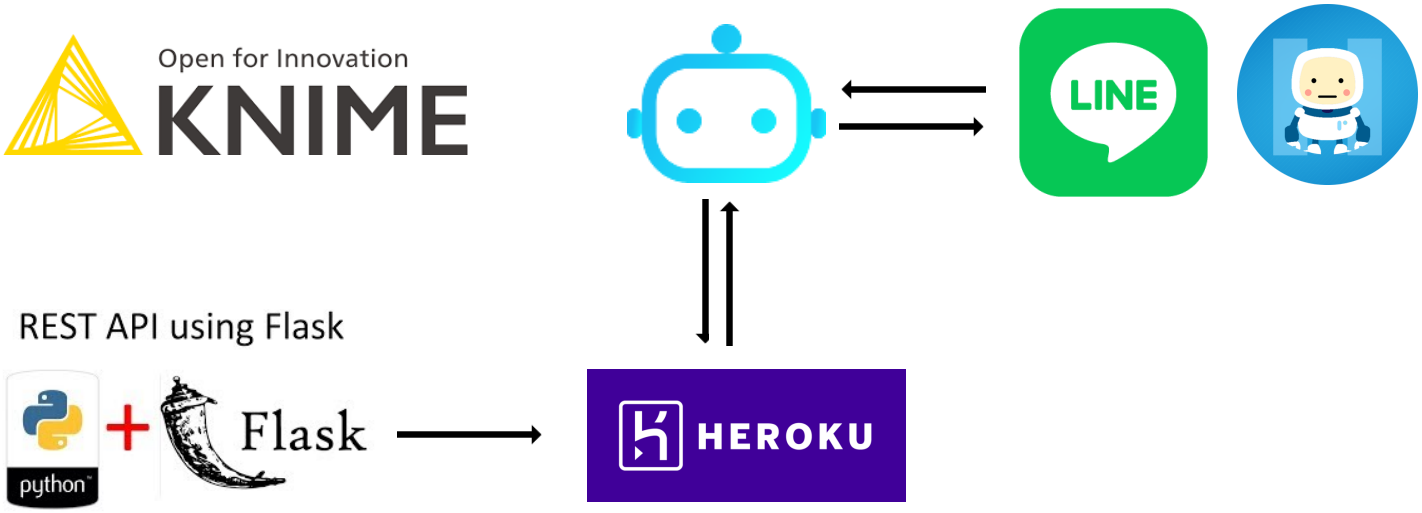
@417yfytt

แสกนเพื่อแอดไลน์ Mr.healthy

แผนผังการทำงานของระบบประเมินราคาประกันสุขภาพ



แผนผังโครงสร้าง Chatbot : Mr.healthy

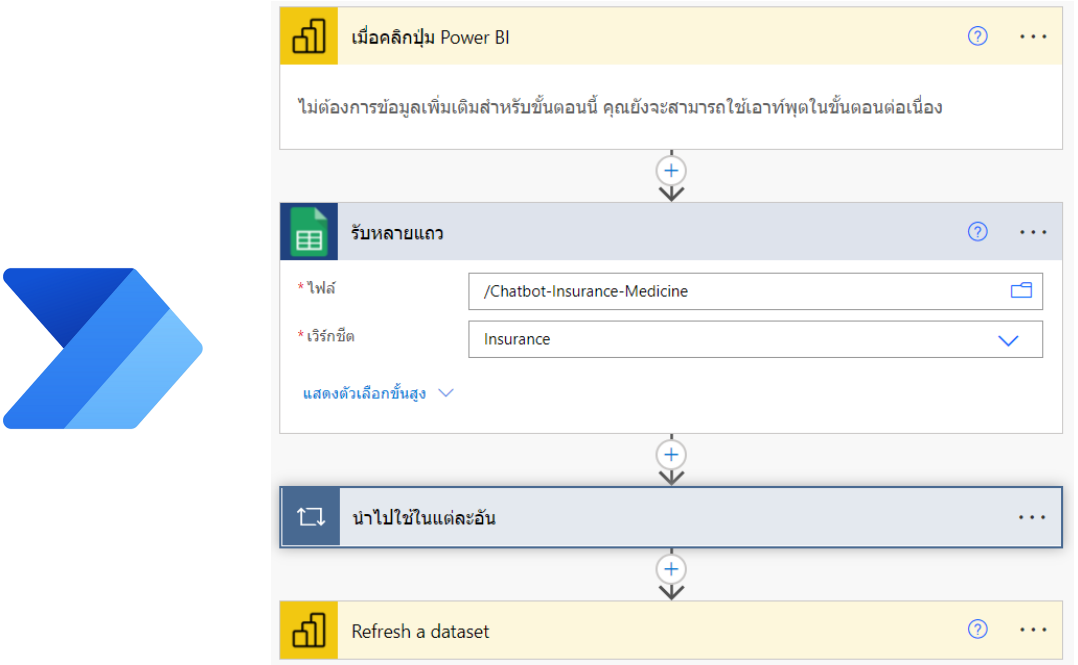


ภาพที่ 12 : แผนผังโครงสร้าง Chatbot : Mr.healthy

5.จากข้อที่ 1 ลองออกแบบ RPA ที่สามารถใช้ในงานได้

- **Learn & optimize**

ในการทำงานในครั้งนี้ทางกลุ่มคิดว่าเราจำเป็นต้องมีการปรับปรุง และพัฒนาโมเดลอย่างต่อเนื่อง เนื่องจากตัวชุดข้อมูลเริ่มต้นในการนำมา Train Model ยังมีในปริมาณที่น้อยและยังไม่มีลักษณะของข้อมูลที่หลากหลาย ทำให้มีผลต่อความแม่นยำในการทำนายของตัว Model ดังนั้น ทางกลุ่มจึงออกแบบวิธีการเก็บข้อมูลหลังจากมีการนำไปทดลองใช้งานกับ User โดยในเบื้องต้นทางกลุ่มได้นำหลักการ RPA มาช่วยในการดึงข้อมูลที่ได้จากตัว Chatbot ให้มาเก็บยังส่วน Sharepoint ใน microsoft office 365 แบบอัตโนมัติเมื่อมีการกดปุ่ม อัพเดทข้อมูลบนหน้า Digital Dashboard โดยในเบื้องต้นที่ทำในลักษณะนี้ เนื่องจากเราต้องการให้ข้อมูลดังกล่าวสามารถเข้าไปสู่หน้าตา Dashboard ได้โดยตรง เพื่อให้เราสามารถตรวจเช็คข้อมูลภาพรวมได้ง่าย และเห็นการอัพเดทที่เป็นปัจจุบันที่สุด ดังนั้นเราจึงตัดสินใจเลือกที่จะทำ Flow RPA ในโปรแกรม Power Automate ที่เป็นส่วนหนึ่งของ office365 เพื่อให้ง่ายต่อการเชื่อมต่อข้อมูลกัน โดยตัว Flow การทำงาน RPA มีลักษณะดังนี้



ภาพที่ 8 ภาพการแสดง Flow การทำงานบน Power automate

ขั้นตอนการทำงานของ Flow RPA

1.เมื่อ Chatbot ทำการส่งข้อมูลการประเมินราคาจากตัวผู้ใช้งานผ่าน API ลง Google sheet ข้อมูลจะถูกบันทึกลงอัตโนมัติ

2.เมื่อบริษัททำการกดปุ่ม Update data ที่อยู่บนหน้า Dashboard ของ Power Bi ระบบจะเริ่มทำงานขึ้นมาในทันที

3.ระบบจะทำการดึงข้อมูลจาก Google sheet ที่เก็บค่าข้อมูลที่ได้รับจาก User มา

4.ระบบจะนำค่าในแต่ละคอลัมป์ที่แตกต่างกันไปเติมในไฟล์ที่ฝากถูกจัดเก็บเอาไว้บน Sharepoint โดยจะเติมต่าจากแถวสุดท้ายของชุดข้อมูลตั้งต้นต่อไปเรื่อย ๆ จนกว่าจะครบจำนวนทั้งหมดของข้อมูลที่ได้รับจาก User

5. Power Bi จะทำการอัพเดทและดึงค่าข้อมูลจาก share point ลงบนหน้าตา Digital Dashboard

รายชื่อสมาชิกภายในกลุ่ม

1.นายไตรรัตน์ อารมฤทธิ์

2.นายทัศนากร รัตนบุรี

3.นายชินพัฒน์ อ้อนประเสริฐ