APPENDICITIS PREDICTION

METHODOLOGY: CRISP-DM

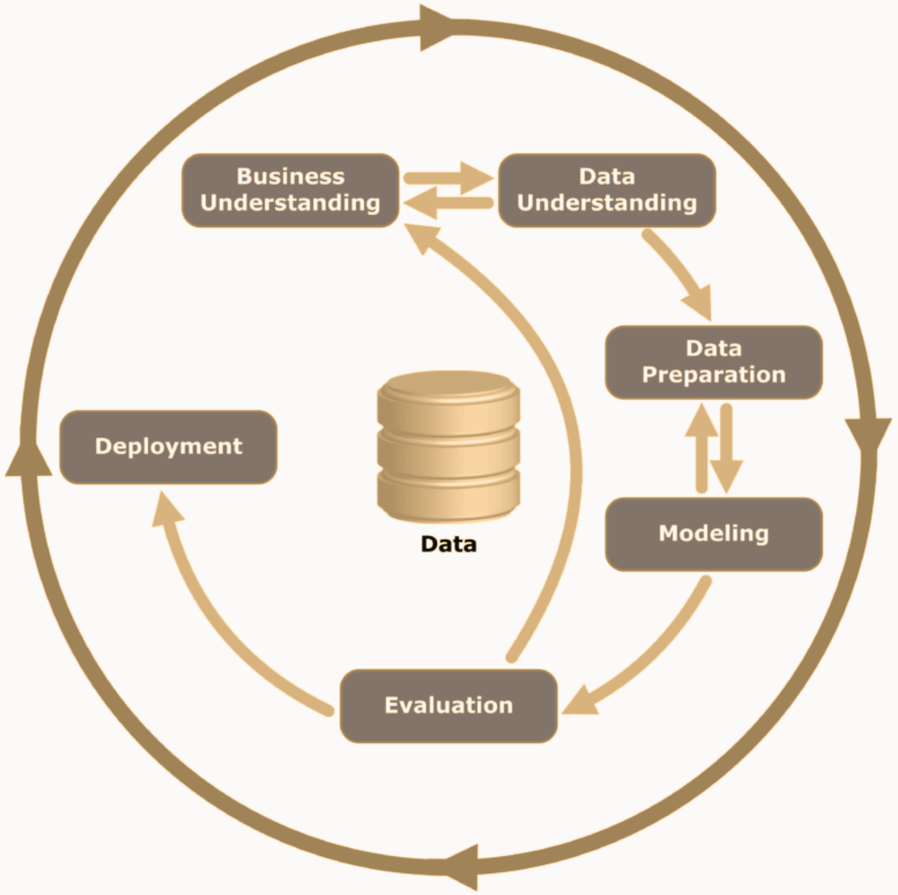
Pisit Janthawee (Data Scientist)

Bachelor of science in Computer Science, Bangkok University

Date: 29 May 2023

METHODOLOGY

CRISP-DM
(CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING)



BUSINESS UNDERSTANDING

HARD TO DIAGNOSE!
VARIOUS LABS TO DO
HIGH COST

BUSINESS TYPE

Healthcare, Medical

PROBLEM?

Appendicitis can present with a wide range of symptoms, making it **difficult** to accurately **diagnose** without additional medical tests.

GOAL

- Prediction Tool for appendicitis suspecting
- Decision support system
- Improve healthcare outcomes

FAST DIAGNOSE
FAST TREATMENT

- Medical personnel
  - Doctor diagnoses faster
  - Reduce diagnosis time
  - Diagnosed at the last delayed or missed, which can lead to complications
- Patient
  - Reduce the chance of getting into complications.
  - Helps reduce costs for patients
  - Blood, urine, X-ray
  - Save time
  - Reduce unnecessary steps

DATA UNDERSTANDING

01

VERIFY DATA QUALITY

How clean/dirty/missing is the data? Document any quality issues.

02

DATA

Info: 500 rows x 23 columns
Target: Disease = 1, No disease = 0
Contains: Demographic, Laboratory, CC (Chief Complaint)
Nation: Thai population

03

ANALYSIS

- What is the factor of disease?
- The relationship between disease A and disease B leading to Main disease?
- Recurrence?

04

TARGET

Probability of disease

Disease	No disease
90%	10%

DATA PREPARATION

SELECT DATA

Appendicitis is among the commonest **childhood diseases**, between 10 and 19 years of age.

Age 1 → 40

CLEAN DATA

- **Remove** laboratory test results from erroneous values.
- Data Imputation

CONSTRUCT DATA

Create related **conditions** from Text data (Chief complaint)

- Migration of pain
- Rebound tenderness
- Cough tenderness
- Dysuria
- Stool
- Tenderness in right lower quadrant (RLQ)
- Nausea/vomiting
- Anorexia
- AS / PAS Score

MODELING

01 DIMENSIONALITY REDUCTION

What is the factor of disease

02 OVER-SAMPLING

SMOTE (Synthetic Minority Over-sampling Technique)

Solved: Class imbalance

After performing SMOTE oversampling, Will the correlation change?

03 SELECT MODELING

- Ensemble Methods:
  - Random Forest
  - Gradient Boosting Classifier
  - AdaBoost Classifier
- Linear Models:
  - Logistic Regression
  - GaussianNB (though it's technically a Naive Bayes classifier, it falls under this category due to its linear nature)
- Tree-based Models:
  - Decision Tree Classifier
- Nearest Neighbors:
  - KNeighbors Classifier
- LightGBM and XGBoost:
  - LGBM Classifier
  - XGB Classifier
- Dummy Classifier:
  - Dummy Classifier (most frequent strategy)

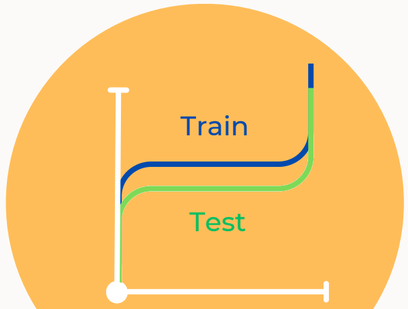
EVALUATION

N

Number of features

Factors

Helps reduce costs for patients



Learning curve

Generalization ability

Can model able to predict on Unseen data?

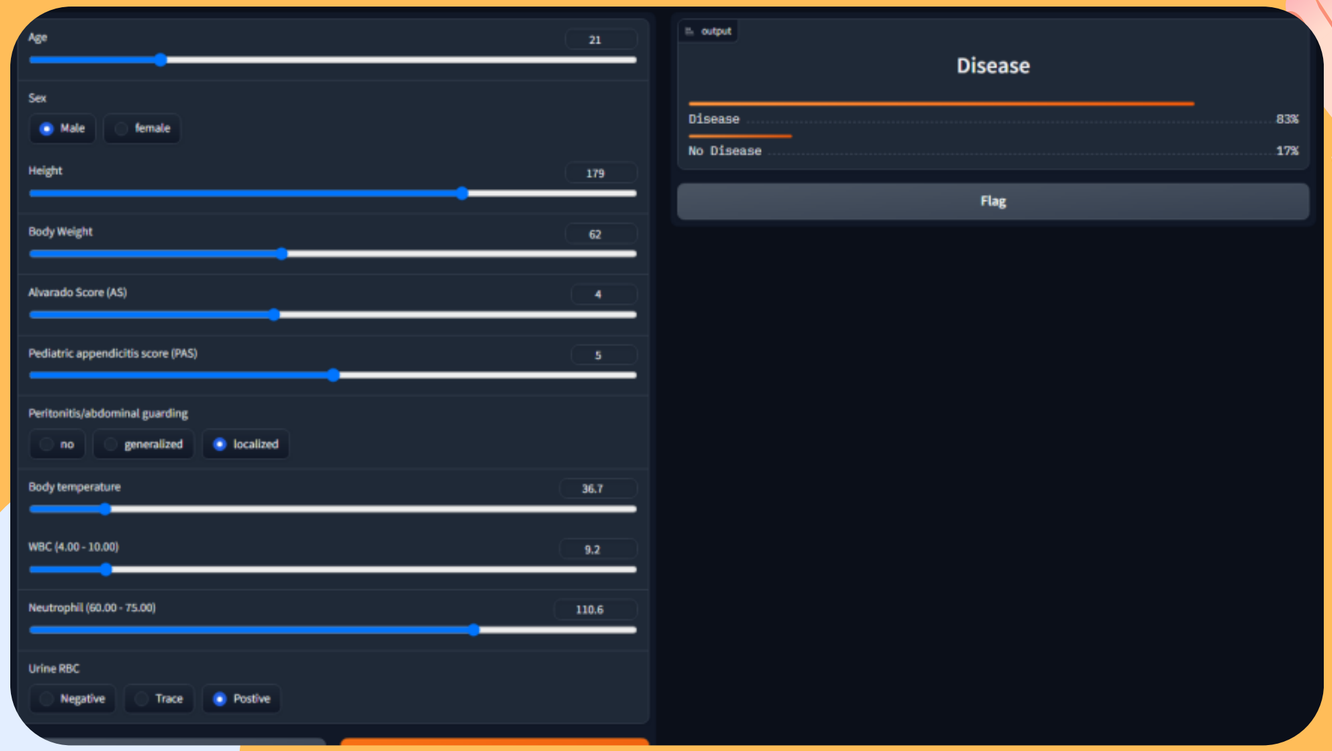
Sensitivities, Specificities, Accuracy, AUROC

Ensure the reliability

Diagnostic accuracy, patient safety, treatment decisions

DEPLOYMENT

User interface design Entering patient information is easy to use.



EXPECTED RESULT

93%

7%

Disease

No disease

Thai population

Treatment decisions

Antibiotics / Surgical

THANKS YOU

CONTACT

099-465-5516

pisit.jant@gmail.com

Bangkok, Bangkok City,
Thailand

- **Portfolio**
  https://pisit-janthawee.mystrikingly.com/
- **LinkedIn**
  https://www.linkedin.com/in/pisit-janthawee/
- **Github**
  https://github.com/Pisit-Janthawee

Pisit Janthawee (Data Scientist)
Bachelor of Science in Computer Science,
Bangkok University