



# Interpretable Machine Learning

Hendra Bunyamin    Maranatha Christian University    October 2, 2023

# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# Prerequisites

Pemahaman mengenai

- Perbedaan masalah regresi & klasifikasi.
- Model *machine learning* seperti linear regression.
- Statistika  $\Rightarrow$  distribusi marginal.

# Github repository

## The repository

# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# What is Interpretability?

- Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019).
- Interpretability is the degree to which a human can consistently predict the model's result (Kim et al., 2016).

*The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made.*

- Christoph Molnar (a statistician, a machine learner)

*A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model.*

- Christoph Molnar



# Taksonomi Teknik Interpretability

- Berbagai taksonomi teknik *Interpretability* dapat dibaca di Molnar (2022).
- Fokus kepada taksonomi berdasarkan **model-specific** atau **model-agnostic**?

# Teknik Interpretasi Model yang Spesifik (*Not Limited*)<sup>†</sup>

Algorithm	Linear	Interaction	Task
Linear regression	✓	✗	regr
Logistic regression	✗	✗	class
Decision trees	✗	✓	class, regr
RuleFit	✓	✓	class, regr
Naïve-Bayes	✗	✗	class
<i>k</i> -nearest neighbors	✗	✗	class, regr

# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# Metode Model-Agnostic

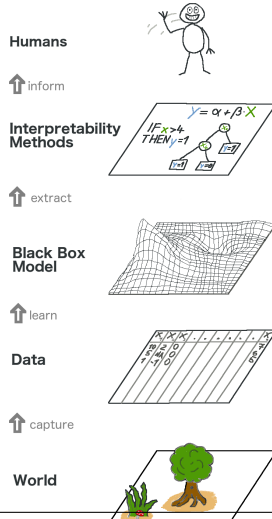
- Memisahkan penjelasan dari model machine learning mempunyai beberapa keuntungan (Ribeiro et al., 2016).
- Keuntungan terbesar metode ini adalah **fleksibilitasnya**.
- Pengembang model machine learning bebas menggunakan model machine learning apa saja.

# Aspek yang Diinginkan<sup>†</sup>

Aspek yang diinginkan dari penjelasan model-agnostic (Ribeiro et al., 2016) adalah

- Model flexibility,
- Explanation flexibility, and
- Representation flexibility.

# High Level Look (Molnar, 2022)<sup>†</sup>



# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# Global Methods



# Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.

# Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.
- Global methods  $\approx$  **expected values** based on the distribution of the data.

# Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.
- Global methods  $\approx$  **expected values** based on the distribution of the data.
- Contoh:  $\hat{f}(x_1, x_2, x_3)$  = fungsi prediksi dengan 3 fitur.  
Untuk melihat efek  $x_1$  pada fungsi prediksi, maka

$$\hat{g}(x_1) = \sum_{x_2} \sum_{x_3} \hat{f}(x_1, x_2, x_3).$$

# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# Partial Dependence Plot (PDP)<sup>†</sup>

- PDP menunjukkan **efek marginal satu atau dua fitur** pada hasil prediksi sebuah model machine learning (Friedman, 2001).
- PDP dapat menunjukkan hubungan antara target dan fitur apakah linier, monotonik atau lebih kompleks.

# Definisi Fungsi Partial Dependence<sup>†</sup>

Bila

$x_S$  = fitur-fitur yang akan diplot oleh fungsi partial dependence,

$X_C$  = fitur-fitur lainnya dalam model machine learning  $\hat{f}$ , maka

$$\hat{f}_S(x_S) = E_{X_C} [\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C).$$

# Estimasi Fungsi Partial Dependence<sup>†</sup>

Fungsi partial  $\hat{f}_S$  diestimasi dengan menghitung rata-rata di *train set* (metode Monte Carlo):

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)}).$$

# Estimasi Fungsi Partial Dependence<sup>†</sup>

Fungsi partial  $\hat{f}_S$  diestimasi dengan menghitung rata-rata di *train set* (metode Monte Carlo):

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)}).$$

**Asumsi:** fitur di C tidak berkorelasi dengan fitur di S.



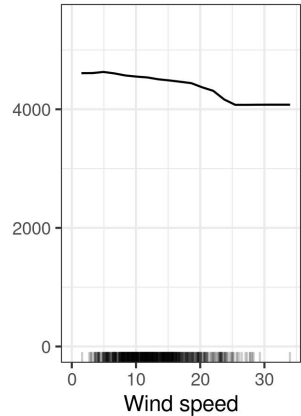
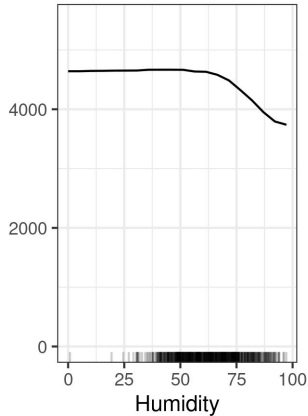
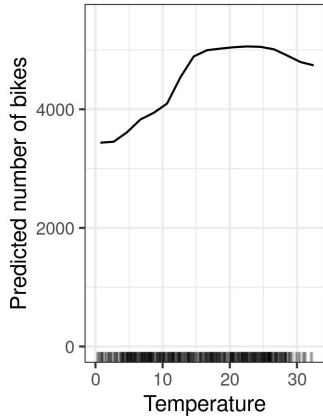
# Bagaimana dengan fitur kategorikal?<sup>†</sup>

- Untuk setiap nilai kategori, kita hitung nilai PDP dengan "memaksa" semua instance data mempunyai nilai kategori yang sama.
- Hitung rata-rata dari semua nilai PDP yang sudah diperoleh.

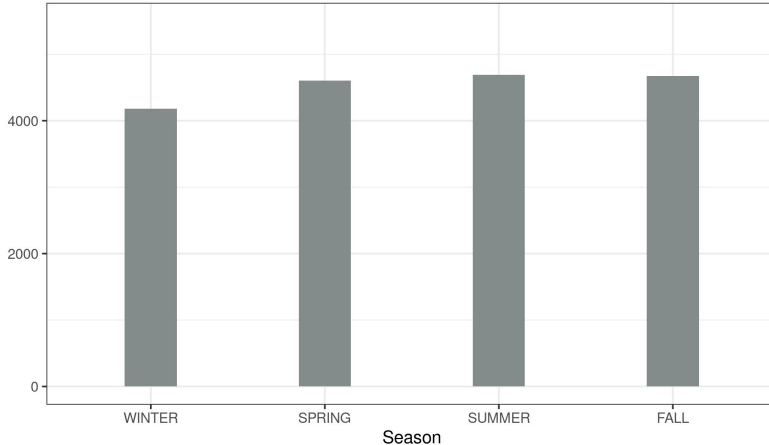
## Contoh: #Sepeda yang dipinjam (1/3)

- Model machine learning, *random forest* dilatih.
- PDP plot  $\Rightarrow$  visualisasi relationship yang model sudah pelajari.

## Contoh: #Sepeda yang dipinjam (2/3)<sup>†</sup>



## Contoh: #Sepeda yang dipinjam (3/3)<sup>†</sup>



# Contoh: Prediksi Lead Menjadi Customer<sup>1</sup>

- Perusahaan edukasi (X Education) menjual online courses ke profesional industri.
- Perusahaan memasarkan course-course pada beberapa website dan search engines like Google.

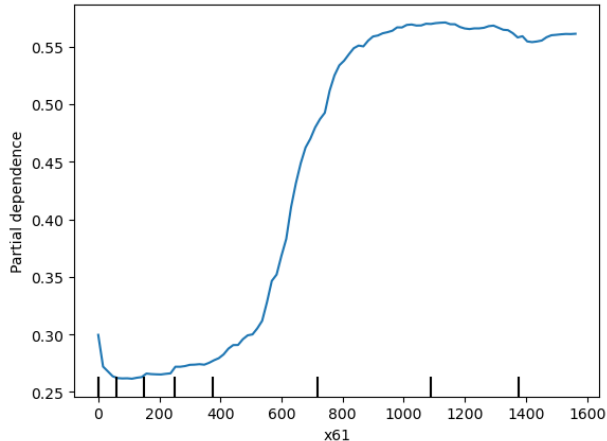
---

<sup>1</sup>link dataset

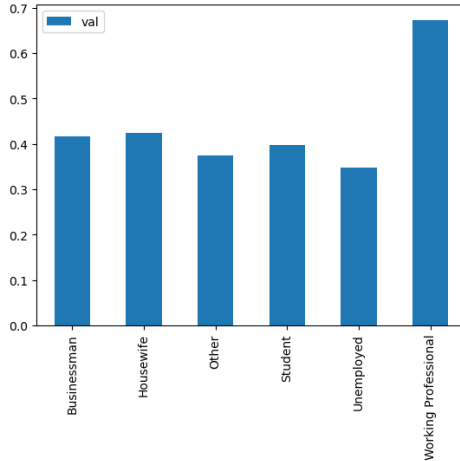
# Contoh: Prediksi Lead Menjadi Customer (Tidak Semua)

- Lead origin
- Lead source
- Do Not Email
- Do Not Call
- Converted
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Last activity
- Country
- Specialization
- How did you hear about X Education
- What is your current occupation
- What matters most to you in choosing this course
- Search

# PDP for Model Prediksi Converted & Total Time Spent on Website



# PDP for Model Prediksi Converted & What is Your Occupation

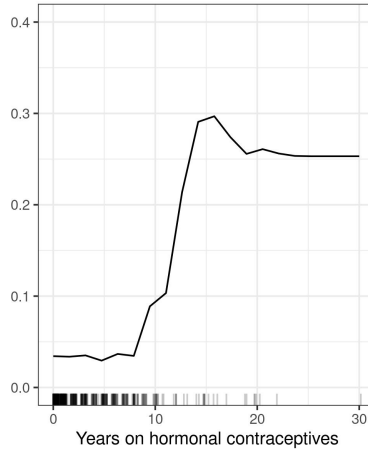
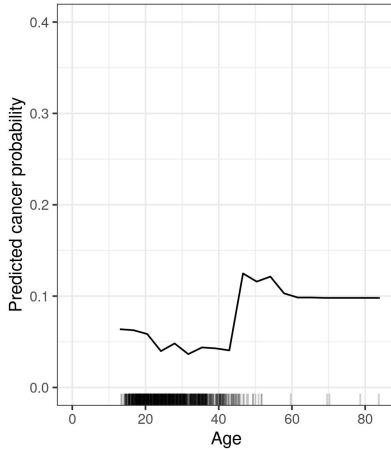




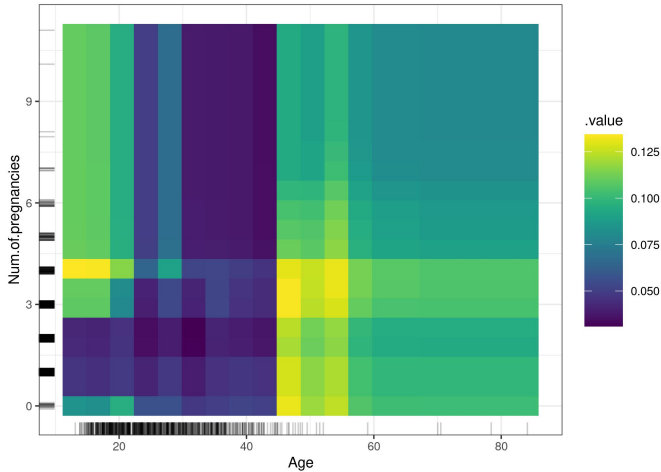
# Contoh: Kanker Serviks (Fernandes and Fernandes, 2017)

- Age in years
- Number of sexual partners
- First sexual intercourse (age in years)
- Number of pregnancies
- Smoking yes or no
- Smoking (in years)
- Hormonal contraceptives yes or no
- Hormonal contraceptives (in years)
- Intrauterine device yes or no (IUD)
- Number of years with an intrauterine device (IUD)
- Has patient ever had a sexually transmitted disease (STD) yes or no
- Number of STD diagnoses
- Time since first STD diagnosis
- Time since last STD diagnosis
- The biopsy results : "Healthy" or "Cancer".  
Target outcome.

# Contoh: Kanker Serviks<sup>†</sup> (1/2)



## Contoh: Kanker Serviks<sup>†</sup> (2/2)



# Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot
5. Advantages & Disadvantages

# Advantages<sup>†</sup>

- Perhitungan PDP *intuitif*.
- Dalam kasus tidak ada korelasi, interpretasi PDP jelas.
- PDP mudah untuk diimplementasi.
- Perhitungan PDPs mempunyai interpretasi *causal* (Zhao and Hastie, 2021).

# Disadvantages<sup>†</sup>

- Jumlah maksimum fitur yang realistik dalam PDPs = 2.
- Beberapa PD plots tidak menampilkan distribusi dari fitur.
- Asumsi *independence* adalah masalah terbesar dengan PD plots  $\Rightarrow$  *Accumulated Local Effect* (ALE) plots.
- Efek heterogeneous mungkin dapat tersembunyi  $\Rightarrow$  kurva Individual Conditional Expectation (ICE).

# Softwares

- R programming language: package `iml`, `pdp`, atau `DALEX`.
- Python programming language: `scikit-learn` atau library `PDPBox`

# References I

- Fernandes, Kelwin, C. J. and Fernandes, J. (2017). Cervical cancer (Risk Factors). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5Z310>.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.
- Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*.
- Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.