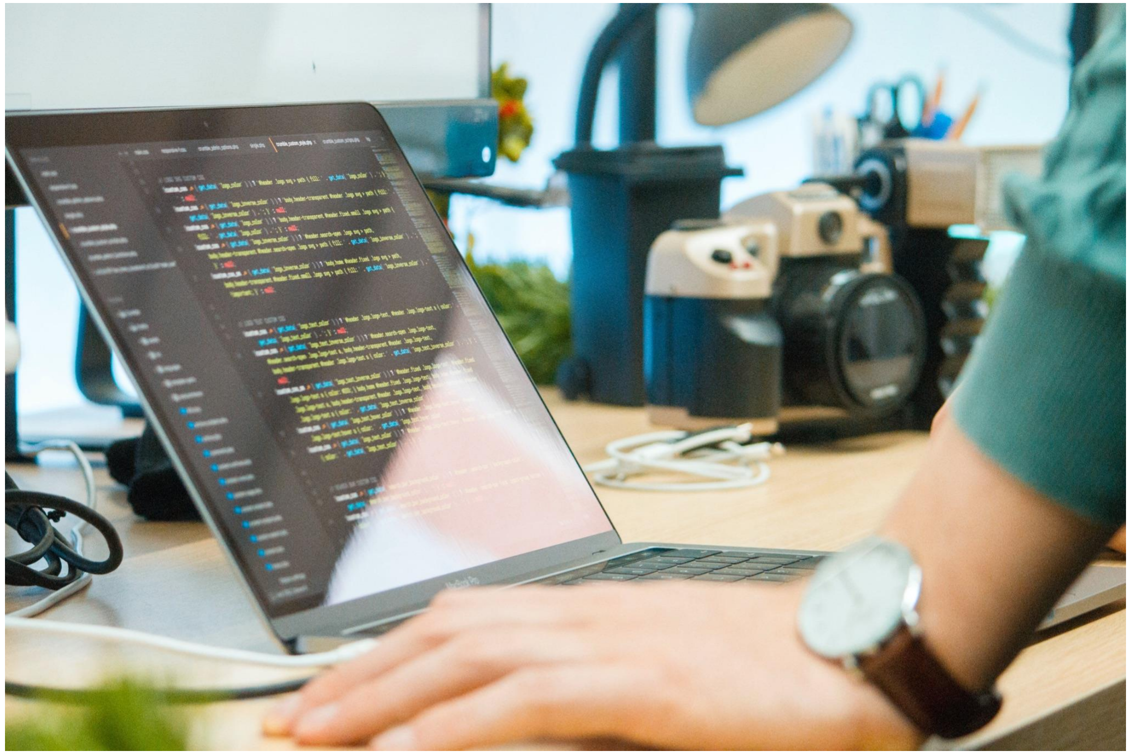
Indonesia AI
AI for Everyone, AI for Indonesia



Random Forest

Indonesia AI
AI for Everyone, AI for Indonesia

Objective & Outline

Objective & Outline

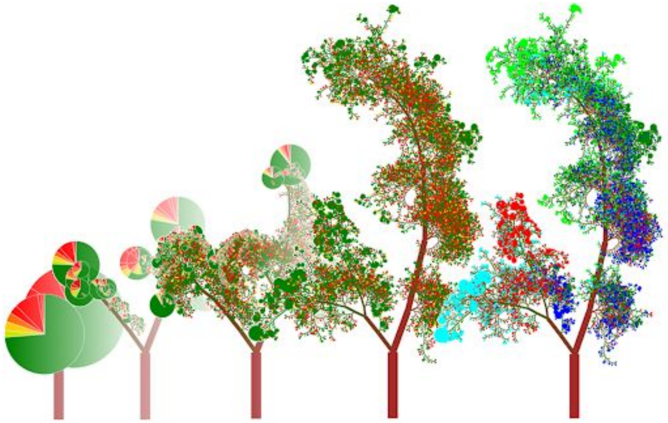
Indonesia AI
AI for Everyone, AI for Indonesia

- **Objective**

  Understanding one of the top five supervised algorithms which is Random Forest algorithm

- **Outline**

  ❖ Overview the history of Random Forest
  ❖ Basic Concept of Random Forest
  ❖ Model Evaluation
  ❖ Application of Random Forest

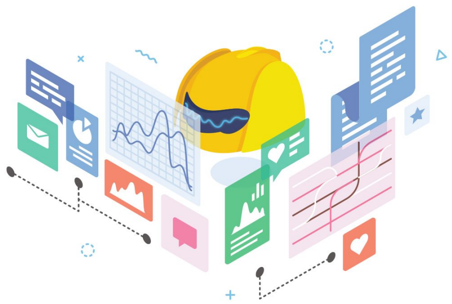


Overview the history of Random Forest

Indonesia AI
AI for Everyone, AI for Indonesia

Overview the history of Random Forest

- The general method of random decision forest was first proposed by Ho in 1995, Tin Kam Ho
- It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions
- The random forest model combines the predictions of the estimators to produce a more accurate prediction



Basic concept of Random Forest

Basic concept of Random Forest (Decision Tree)

❖ Decision Tree is the key concept of Random Forest

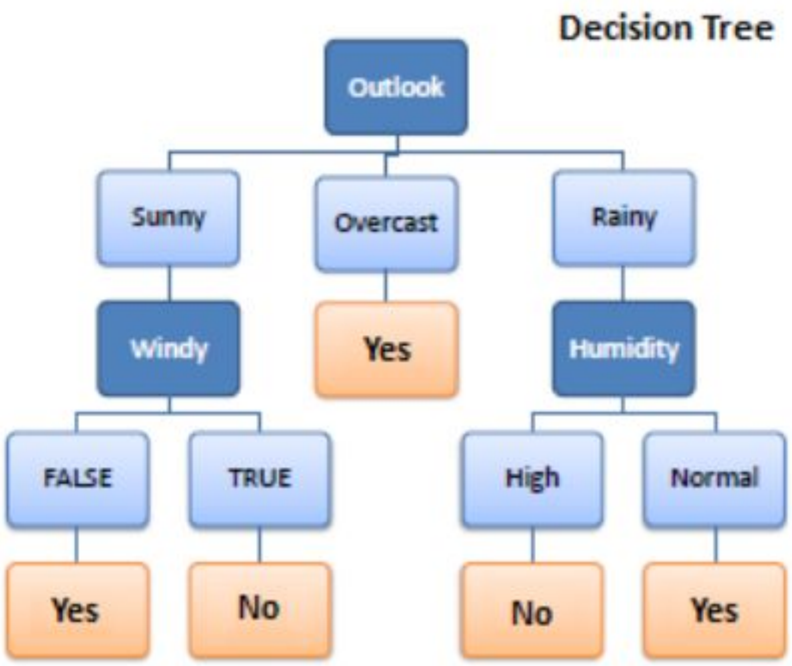
❖ Consider this example: What are the factors which decide if we ar to play golf?

➢ Outlook? (Sunny, Overcast, Rainy)

➢ Temperature? (Hot, Mild, Cool)

➢ Humidity? (High, Normal)

➢ Windy? (False, True)



❖ Label: Play Golf? (Yes / No)

Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia

Predictors				Target
Outlook	**Temp.**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Basic concept of Random Forest (Decision Tree)

We are going to use some points deducted from **information theory**.

To measure the randomness or uncertainty of a variable X (feature) is defined by **Entropy**.

Find the entropy of the target feature:

a. If all examples are positive or all are negative then entropy = 0

b. If all examples are equally divided then entropy = 1

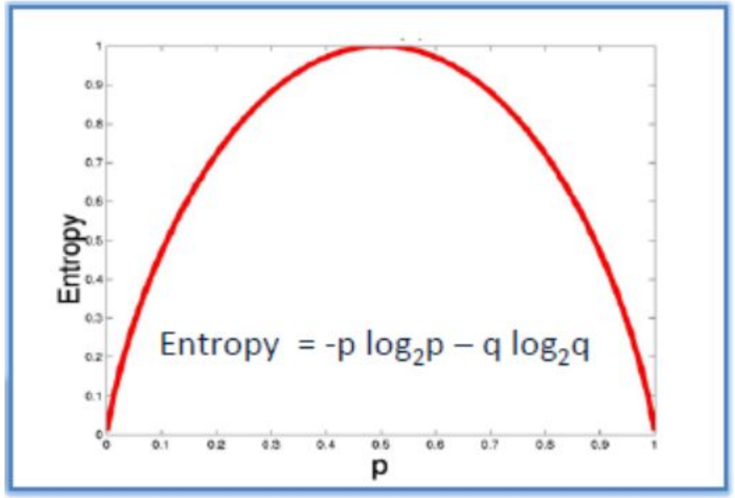
Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)

$= \text{Entropy}\ (0.36, 0.64)$

$= -\ (0.36 \log_2 0.36) - (0.64 \log_2 0.64)$

$= 0.94$



Basic concept of Random Forest (Decision Tree)

Find the entropy of each feature towards the target.

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$\mathbf{E}(\text{PlayGolf, Outlook}) = \mathbf{P}(\text{Sunny})*\mathbf{E}(3,2) + \mathbf{P}(\text{Overcast})*\mathbf{E}(4,0) + \mathbf{P}(\text{Rainy})*\mathbf{E}(2,3)$$

$$= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971$$

$$= 0.693$$

Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia

- Find the information gain of each feature used to predict the target.

- **Information gain**: Decrease of entropy after the dataset is split on an attribute
  - Creating decision tree classification is about finding attribute that return the highest Information Gain

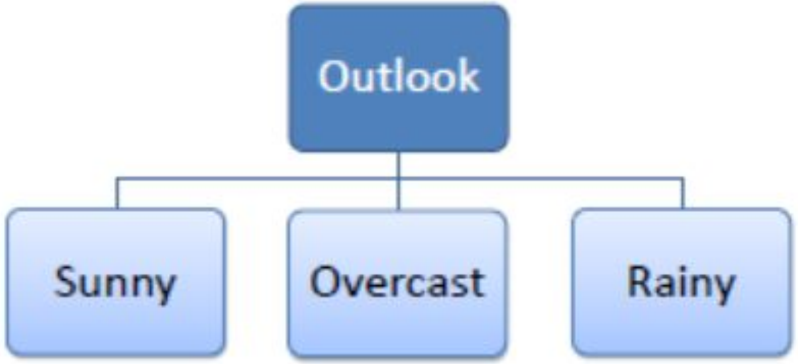
$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

$$\begin{aligned}\mathbf{G}(\text{PlayGolf, Outlook}) &= \mathbf{E}(\text{PlayGolf}) - \mathbf{E}(\text{PlayGolf, Outlook}) \\ &= 0.940 - 0.693 = 0.247\end{aligned}$$

Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia

Select feature with the highest Information Gain as the root node.



		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

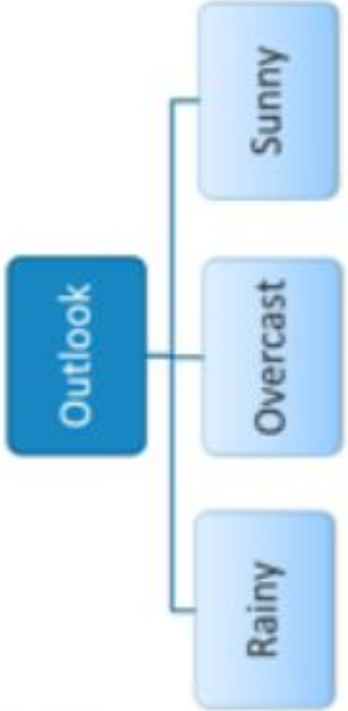
		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia



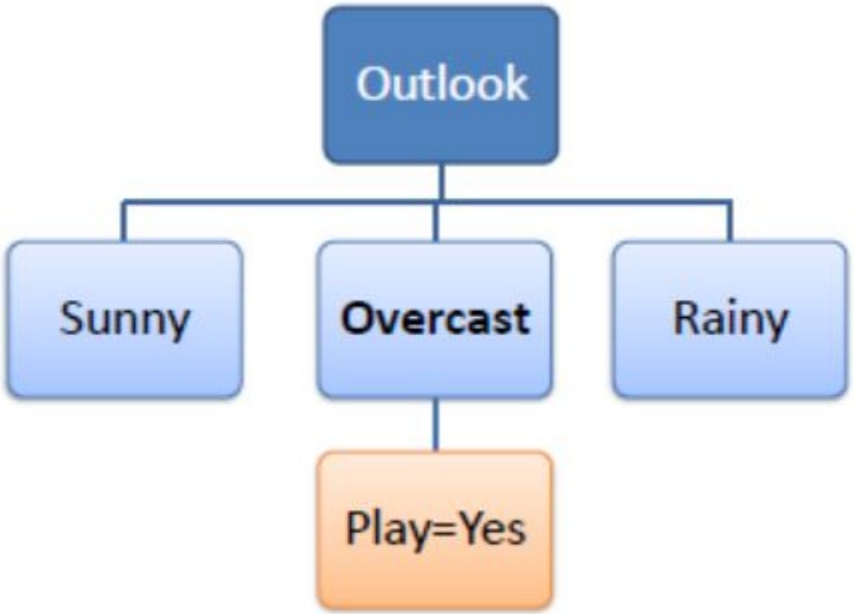
Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Basic concept of Random Forest (Decision Tree)

If all examples are positive (yes) or all are negative (no) then entropy will be zero, in this case it is overcast.

Branch with entropy is 0 will be a terminal node (leaf).

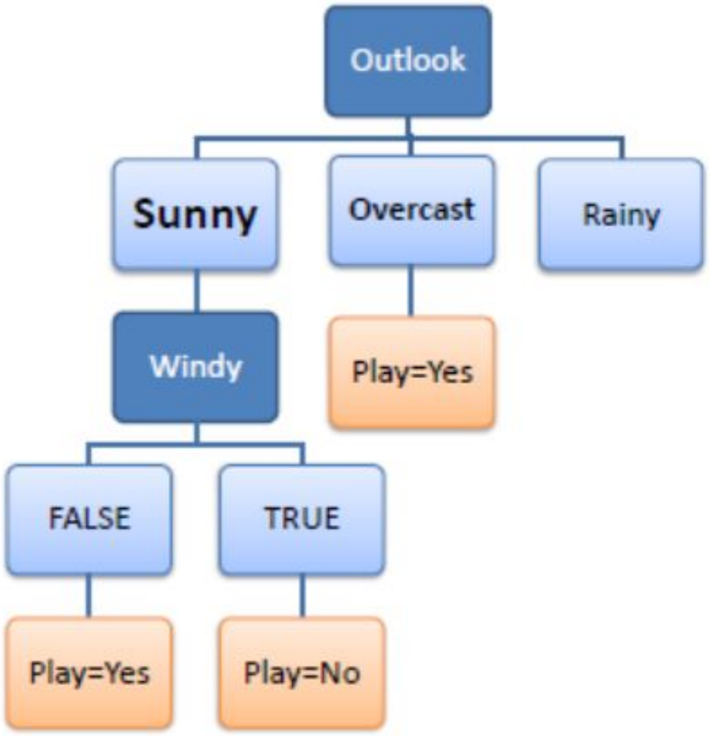
Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes



Basic concept of Random Forest (Decision Tree)

Branch with entropy more than 0 needs further splitting.

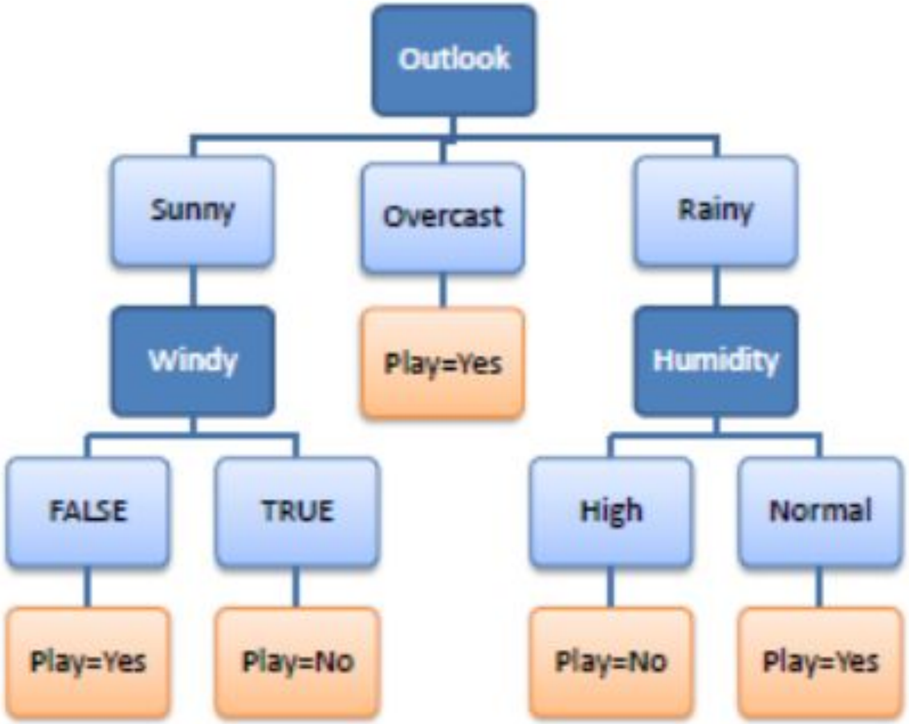
Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia

Decision tree algorithm is run recursively on the non-leaf branches, until all data is classified.

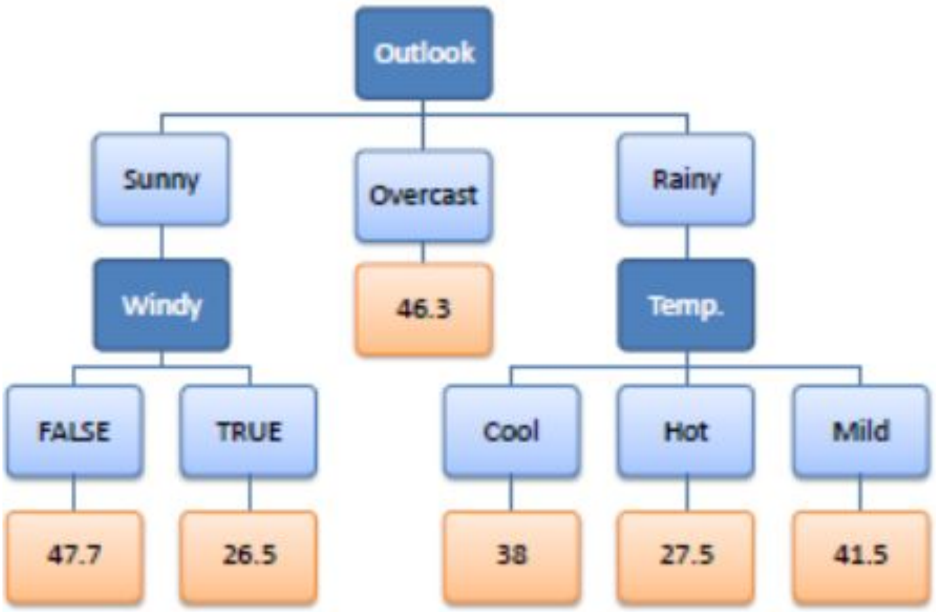


Basic concept of Random Forest (Decision Tree)

Indonesia AI
AI for Everyone, AI for Indonesia

While using information Gain as a criterion, we assume target attributes to be categorical, and for gini index, target attributes are assumed to be continuous.

Predictors				Target
Outlook	**Temp.**	**Humidity**	**Windy**	**Hours Played**
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



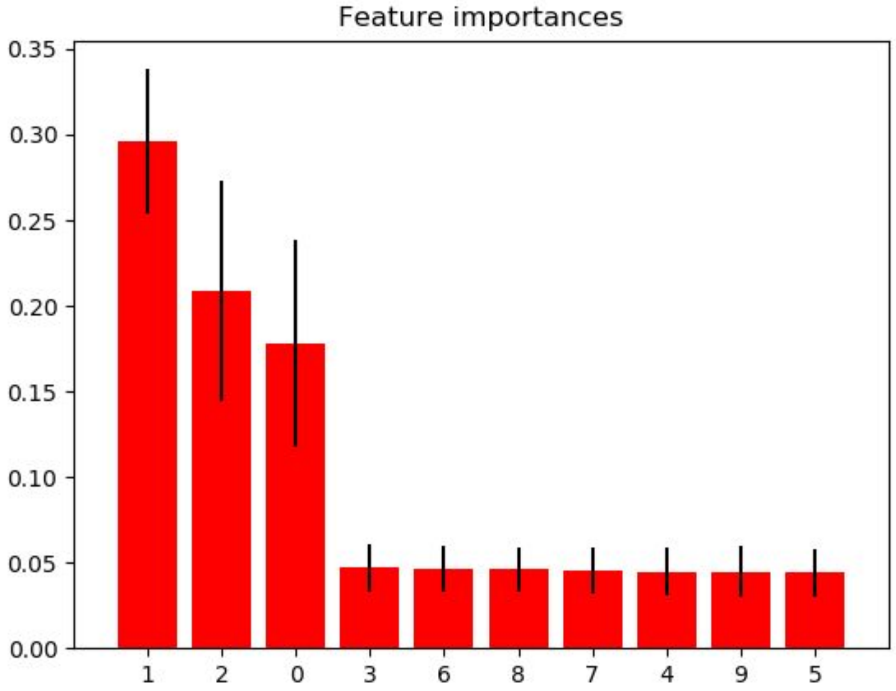
Basic concept of Random Forest

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by **constructing multiple decision tree**
  - The classification forest chooses the result with most votes overall trees (max voting)
  - The regression forest average outputs of different trees (average voting)

- **Ensemble learning** is a popular machine learning technique that combines several models to improve overall result (in this case, combines several trees)

Basic concept of Random Forest

- Shows the use of forests of trees to evaluate the importance of features on an artificial classification task. The red bars are the feature importances of the forest, along with their inter-trees variability.
- As expected, the plot suggests that 3 features are informative, while the remaining are not.



Basic concept of Random Forest

- Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples

```
Feature ranking:
1. feature 1 (0.295902)
2. feature 2 (0.208351)
3. feature 0 (0.177632)
4. feature 3 (0.047121)
5. feature 6 (0.046303)
6. feature 8 (0.046013)
7. feature 7 (0.045575)
8. feature 4 (0.044614)
9. feature 9 (0.044577)
10. feature 5 (0.043912)
```

Quiz Session

Quiz

Can Random Forest Algorithm be used both for Continuous and Categorical Target Variables?

Why we should use Random Forest?

Why we should use Random Forest?

Indonesia AI
AI for Everyone, AI for Indonesia

- Random forest algorithm is suitable for both classifications and regression task
- It gives robust accuracy, meaning that it is not so fragile that will drop when tested with testing data
- Random forest classifier can handle the missing values and maintain the accuracy of a large proportion of data
- Random Forest will be used as baseline model for any kind of project in Industry

Indonesia AI
AI for Everyone, AI for Indonesia

Model
Evaluation

Model Evaluation (Confusion Matrix)

Membangun model *machine learning* saja tidaklah cukup, kita perlu mengetahui seberapa baik model kita bekerja. Tentunya, dengan sebuah ukuran (atau istilah yang seringkali digunakan adalah *metric*).

1. **True Negative (TN):** Model memprediksi data ada di kelas **Negatif** dan yang sebenarnya data memang ada di kelas **Negatif.**
2. **True Postive (TP):** Model memprediksi data ada di kelas **Positif** dan yang sebenarnya data memang ada di kelas **Positif.**
3. **False Negative (FN):** Model memprediksi data ada di kelas **Negatif,** namun yang sebenarnya data ada di kelas **Positif.**
4. **False Positive (FP):** Model memprediksi data ada di kelas **Positif,** namun yang sebenarnya data ada di kelas **Negatif.**

		Predicted 0	1
Actual	0	TN	FP
	1	FN	TP

Accuracy
(TP + TN) / (TP+TN+FP+FN)

Model Evaluation (Precision & Recall)

Secara definisi, precision adalah perbandingan antara True Positive (TP) dengan banyaknya data yang diprediksi positif. Atau bisa juga dituliskan secara matemetis:

$$precision = \frac{TP}{TP + FP}$$

Sedangkan untuk Recall, secara definisi adalah perbandingan antara True Positive (TP) dengan banyaknya data yang sebenarnya positif. Dan dapat dituliskan secara matematis seperti ini:

$$recall = \frac{TP}{TP + FN}$$

Model Evaluation (F1 Score)

Indonesia AI
AI for Everyone, AI for Indonesia

Secara definisi, F1-Score adalah harmonic mean dari precision dan recall. Yang secara matematik dapat ditulis begini:

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

Nilai terbaik F1-Score adalah 1.0 dan nilai terburuknya adalah 0. Secara representasi, jika F1-Score punya skor yang baik mengindikasikan bahwa model klasifikasi kita punya precision dan recall yang baik.

Quiz Session

Quiz

Why the name is Confusion Matrix?

Indonesia AI
AI for Everyone, AI for Indonesia

Terima Kasih!

[Indonesia AI | AI for Everyone, AI for Indonesia]

contact@aiforindonesia.org