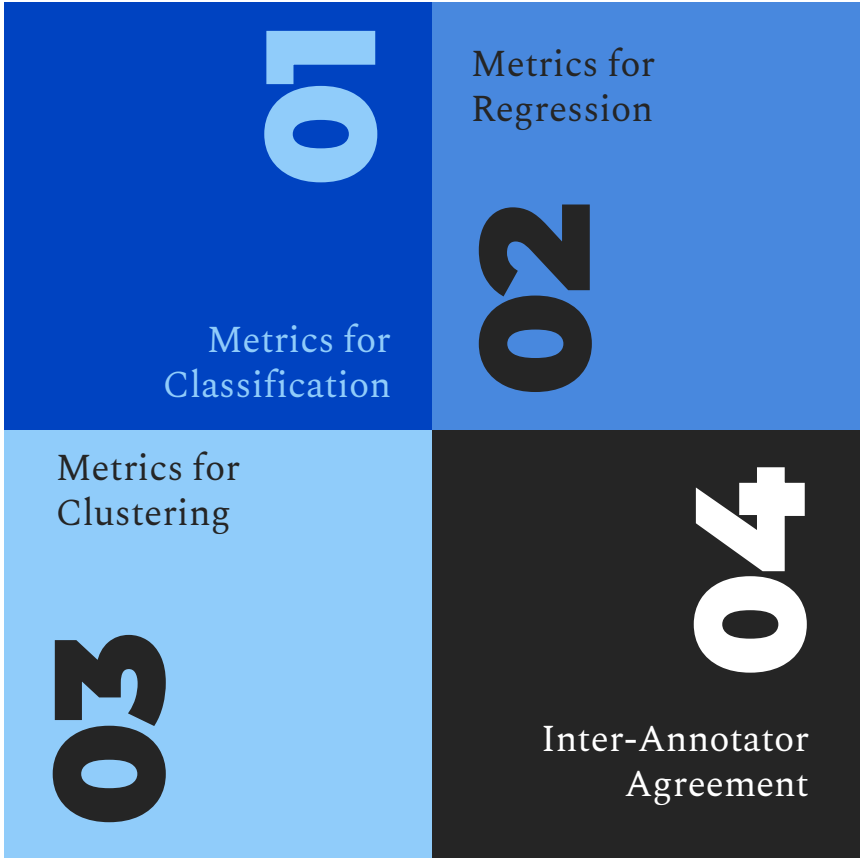
Model Evaluation

Metrics For Evaluation Machine Learning Model

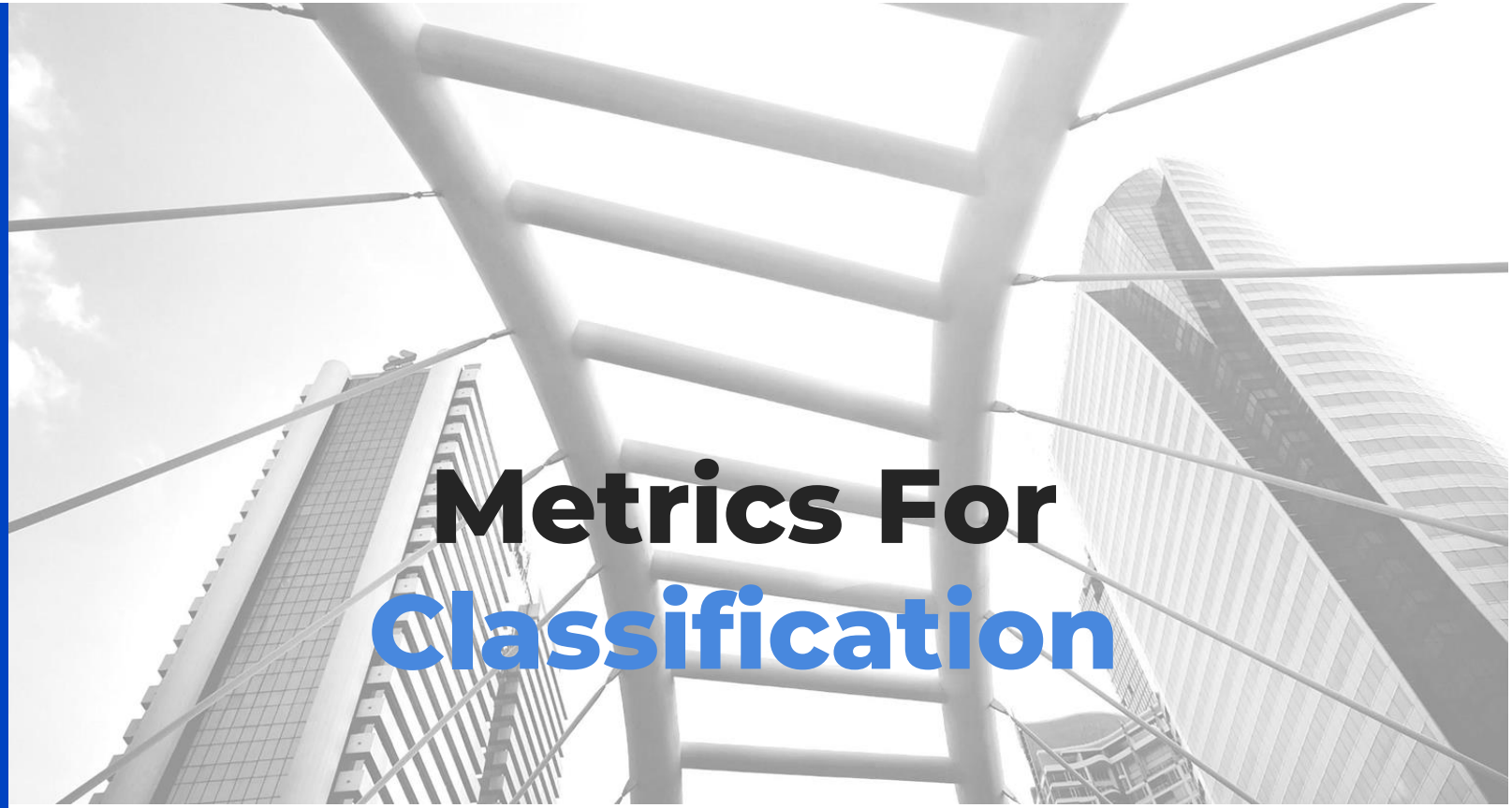
Nama Pengajar

- Dian Ade Kurnia, M.Kom
- Rusnanda Farhan
- Rusnandi Fikri
- Rika Sahriana


01
Metrics for Classification
Metrics for Regression
02
Metrics for Clustering
03
04
Inter-Annotator Agreement

Agenda

01



Metrics For Classification

Confusion Matrix

The confusion matrix/error matrix is a visualization table of the predicted label vs the original label (groundtruth label).

		True/Actual	
		Positive ()	Negative
Predicted	Positive ()	7	0
	Negative	0	3

Confusion Matrix

Name	Predicted	Groundtruth
True Positive	+	+
False Positive	+	-
True Negative	-	-
False Negative	-	+

		True/Actual	
		Positive ()	Negative
Predicted	Positive ()	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

Accuracy

Accuracy is defined as the ratio of the correct prediction to the total of all predictions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision

Question : what proportion of predicted Positives is truly Positive?

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

Question : what proportion of actual Positives is correctly classified?

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Recall also known as **True Positive Rate** (**TPR**) or **Sensitivity**

F1-Score

One popular metric which combines precision and recall is called F1-score, which is the harmonic mean of precision and recall defined as:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

Use Beta = 1 to calculate F1-Score

Specificity

Question : what proportion of actual Negative is correctly classified?

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Specificity also known as **True Negative Rate** (**TNR**) as the opposite of Sensitivity. Specificity and Sensitivity are two other popular metrics mostly used in medical and biology related fields.

How About MultiClass?

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

How About MultiClass?

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

How About
MultiClass?

↔ : Precision

↕ : Recall

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

02

Metrics For Regression

Mean Squared Error

"Mean squared error" is perhaps the most popular metric used for regression problems. It essentially finds the average squared error between the predicted and actual values.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error

As its name, RMSE is essentially the square root of MSE. RMSE shows what is the average deviation in your model predicted values from the target values

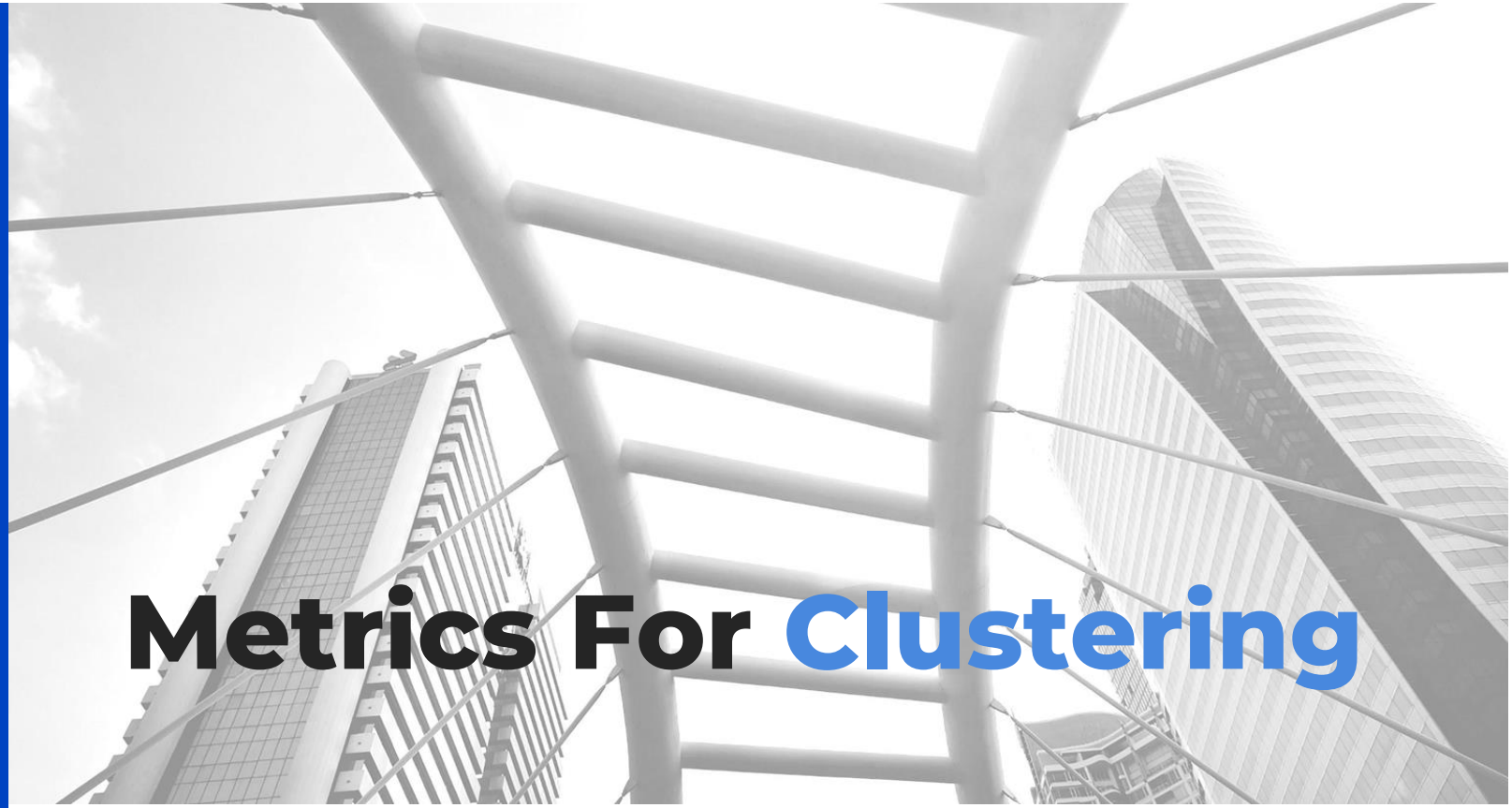
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error

Mean absolute error (or mean absolute deviation) is another metric which finds the average absolute distance between the predicted and target values.
MAE is known to be more robust to the outliers than MSE. The main reason being that in MSE by squaring the errors, the outliers (which usually have higher errors than other samples) get more attention and dominance in the final error and impacting the model parameters.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

03



Metrics For Clustering

Silhouette **Score**

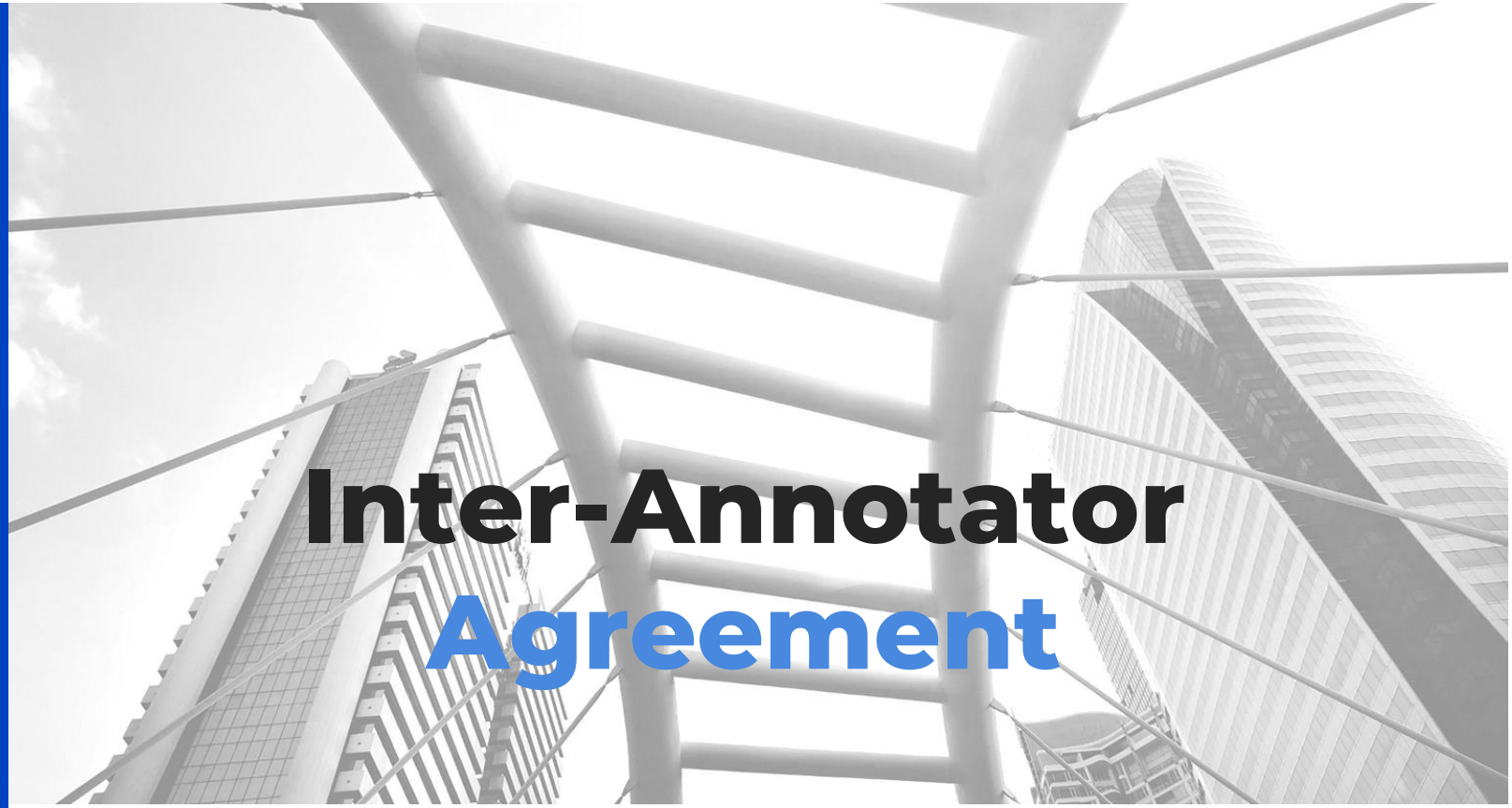
The silhouette score is a metric to evaluate a clustering algorithm. The silhouette score is calculated from two scores, a and b.

- a is the mean distance between a sample and all other points in the same cluster
- b is the mean distance between a sample and all other points in the next nearest cluster.

The silhouette score can range from -1 to +1. A silhouette score of -1 means incorrect clustering and +1 means correct and highly dense clustering. A silhouette score of 0 means the clusters are overlapping.

$$s = \frac{b - a}{max(a, b)}$$

04



Inter-Annotator Agreement

Inter-Annotator Agreement

Inter-Annotator Agreement (IAA), is a measure of how well multiple annotators can make the same annotation decision for a certain category. IAA shows you how clear your annotation guidelines are, how uniformly your annotators understood it, and how reproducible the annotation task is. It is a vital part of both the validation and reproducibility of classification results.

Inter-Annotator Agreement

best practices for annotations and explores the IAA metrics for qualitative annotations: Cohen and Fleiss' kappa. Cohen kappa is calculated between a pair of annotators and Fleiss' kappa over a group of multiple annotators.

$$kappa(\kappa) = \frac{P_o - P_e}{1 - P_e}$$

Cohen Kappa

Cohen's kappa coefficient (κ) is a statistic to measure the reliability between annotators for qualitative (categorical) items. It is a more robust measure than simple percent agreement calculations, as κ takes into account the possibility of the agreement occurring by chance. It is a pairwise reliability measure between two annotators. Cohen's kappa statistic is the agreement between two raters where Po is the relative observed agreement among raters (identical to the accuracy), and Pe is the hypothetical probability of chance agreement.

Fleiss' Kappa

Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to several items or classifying items. It is a generalization of Scott's pi (π) evaluation metric for two annotators extended to multiple annotators. Whereas Scott's pi and Cohen's kappa work for only two raters, Fleiss' kappa works for any number of raters giving categorical ratings, to a fixed number of items. In addition to that, not all raters are required to annotate all items.

Conclusion

- Knowing and choosing the right metric is crucial while evaluating machine learning (ML) models.
- It often happens that clusters are manually and qualitatively inspected to determine whether the results are meaningful.
- In classification, accuracy is not everything. Use accuracy in conjunction with other metrics to better ensure model performance.
- In the real case, sometimes the data doesn't have labels and requires us to annotate the labels. The main annotators who are able to annotate are experts, but we as non-experts can also annotate using the IAA.

References

- https://towardsdatascience.com/
- https://en.wikipedia.org/