Supervised Learning - Classification

Mhd. Arsya Fikri | 191402066

Apa yang akan kita bahas?

What is Supervised Learning?

Data Imbalance

Modelling (Classification)

Model Evaluation

.

.

01

Supervised Learning - Classification

What is it?

What is Supervised Learning?

There are 3 major groups of machine learning types:

- **Supervised Learning**: Input data is called training data and has a known label or result (such as spam/not-spam or a stock price at a time).
- **Unsupervised Learning**: Input data is not labeled and does not have a known result.
- **Reinforcement Learning**: A special type of Machine Learning where the model learns from each action taken. The model is rewarded for any correct decision made and penalized for any wrong decision.

What is Supervised Learning?

Supervised Learning

- Making predictions with a rule/often called as a model
- Has input data and labels

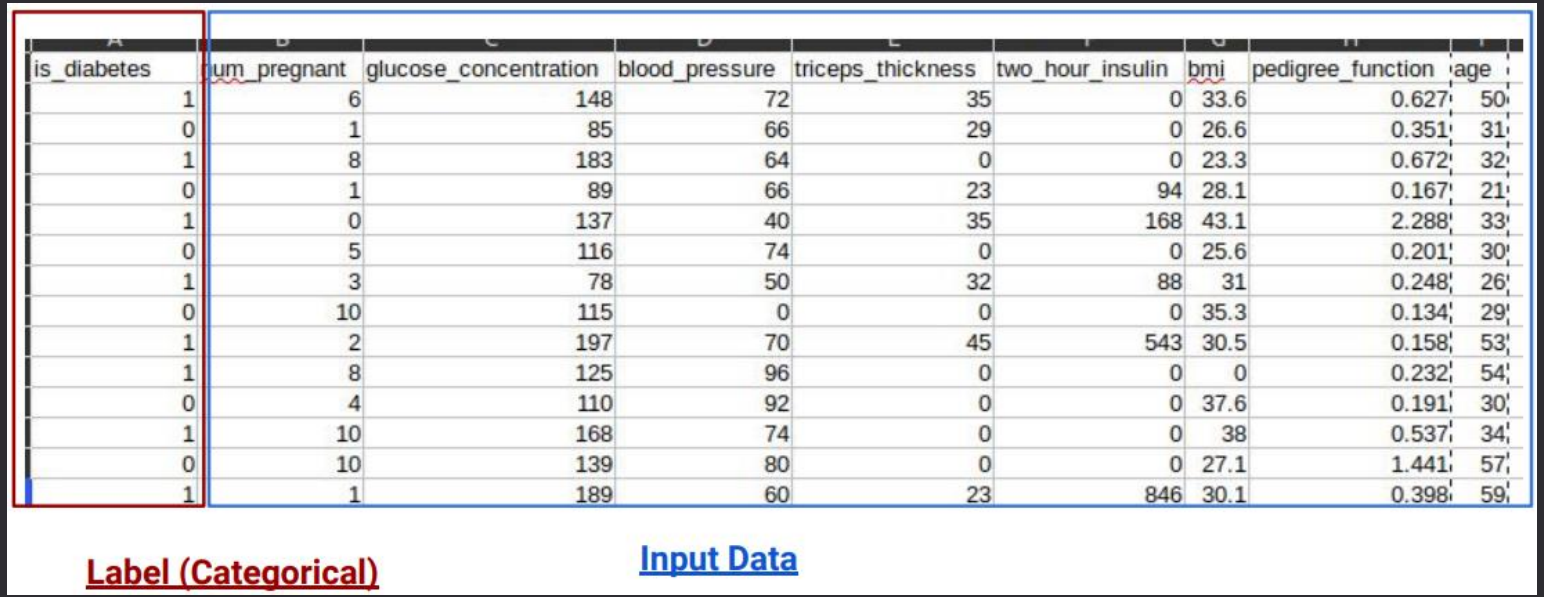
What is Supervised Learning?

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
221900	3	1	1180	5650	1	0	0	3	7	1180
538000	3	2.25	2570	7242	2	0	0	3	7	2170
180000	2	1	770	10000	1	0	0	3	6	770
604000	4	3	1960	5000	1	0	0	5	7	1050
510000	3	2	1680	8080	1	0	0	3	8	1680
1225000	4	4.5	5420	101930	1	0	0	3	11	3890
257500	3	2.25	1715	6819	2	0	0	3	7	1715
291850	3	1.5	1060	9711	1	0	0	3	7	1060
229500	3	1	1780	7470	1	0	0	3	7	1050
323000	3	2.5	1890	6560	2	0	0	3	7	1890
662500	3	2.5	3560	9796	1	0	0	3	8	1860
468000	2	1	1160	6000	1	0	0	4	7	860

Label (Numerical)

Input Data

What is Supervised Learning?

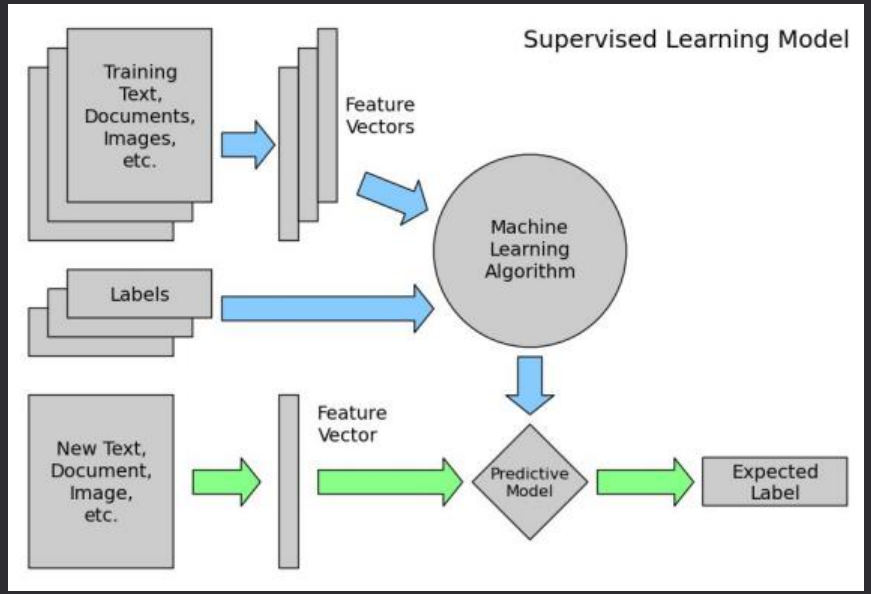


is_diabetes	um_pregnant	glucose_concentration	blood_pressure	triceps_thickness	two_hour_insulin	bmi	pedigree_function	age
1	6	148	72	35	0	33.6	0.627	50
0	1	85	66	29	0	26.6	0.351	31
1	8	183	64	0	0	23.3	0.672	32
0	1	89	66	23	94	28.1	0.167	21
1	0	137	40	35	168	43.1	2.288	33
0	5	116	74	0	0	25.6	0.201	30
1	3	78	50	32	88	31	0.248	26
0	10	115	0	0	0	35.3	0.134	29
1	2	197	70	45	543	30.5	0.158	53
1	8	125	96	0	0	0	0.232	54
0	4	110	92	0	0	37.6	0.191	30
1	10	168	74	0	0	38	0.537	34
0	10	139	80	0	0	27.1	1.441	57
1	1	189	60	23	846	30.1	0.398	59

Label (Categorical)

Input Data

How Supervised Learning Works?



02

Supervised Learning - Classification

Data Imbalance

Imbalance Problem

Imbalanced Classification: A classification predictive modeling problem where **the distribution of examples across the classes (label) is not equal**. Examples:

- Identification of rare diseases like cancer; tumours etc.
- Fraudulent transactions in banks
- Identify customer churn rate (what fraction of customers continue using a service)
- Natural Disasters like Earthquakes
- Spam emails, etc.

Imbalance Problem: Majority and Minority

Majority & Minority

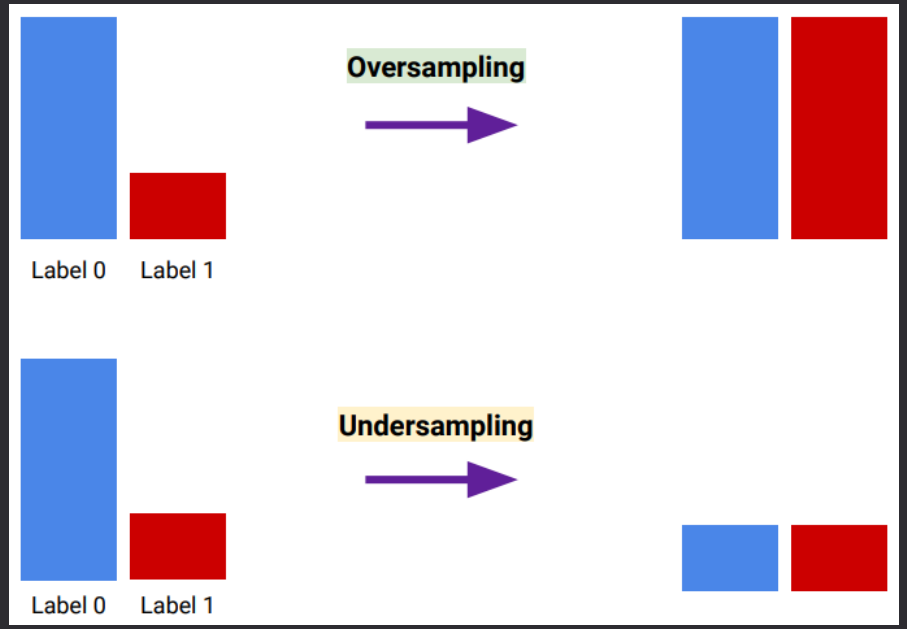
- Majority Class: The class (or classes) in an imbalanced classification predictive modeling problem that has many examples.
- Minority Class: The class in an imbalanced classification predictive modeling problem that has few examples.

When working with an imbalanced classification problem, **the minority class is typically of the most interest**. This means that a model's skill in correctly predicting the class label or probability for the minority class is more important than the majority class or classes. However, the minority class is harder to predict because there are few examples of this class. And most machine learning algorithms are designed on problems that assume **an equal distribution of classes**.

Imbalance Problem: How to Handle?

- **Resampling** (Oversampling & Undersampling)
- **Choosing Proper Evaluation Metrics** (will be explained later together with other evaluation metrics for classification)

Imbalance Problem: Resampling



03

Supervised Learning - Classification

Modelling (Classification)

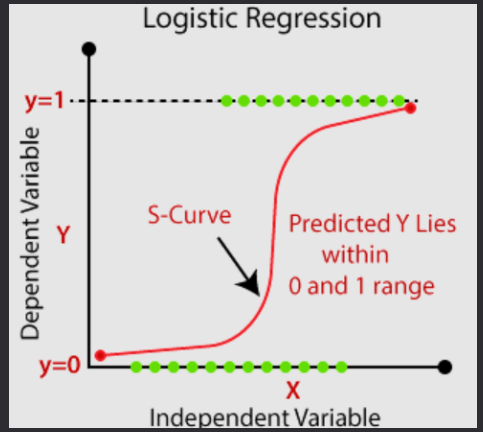
Modelling

- The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data.
- There are so many algorithm we can use as a model.

Modelling

- **Logistic Regression**

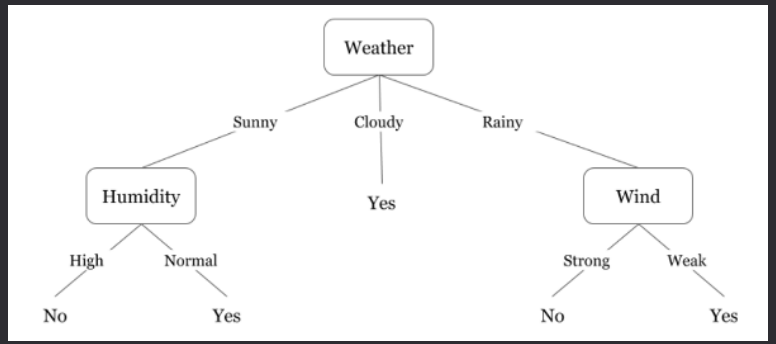
Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no



Modelling

- **Decision Tree**

A decision tree is a tree-like structure that represents a series of decisions and their possible consequences. It is used in machine learning for classification and regression tasks. An example of a decision tree is a flowchart that helps a person decide what to wear based on the weather conditions.



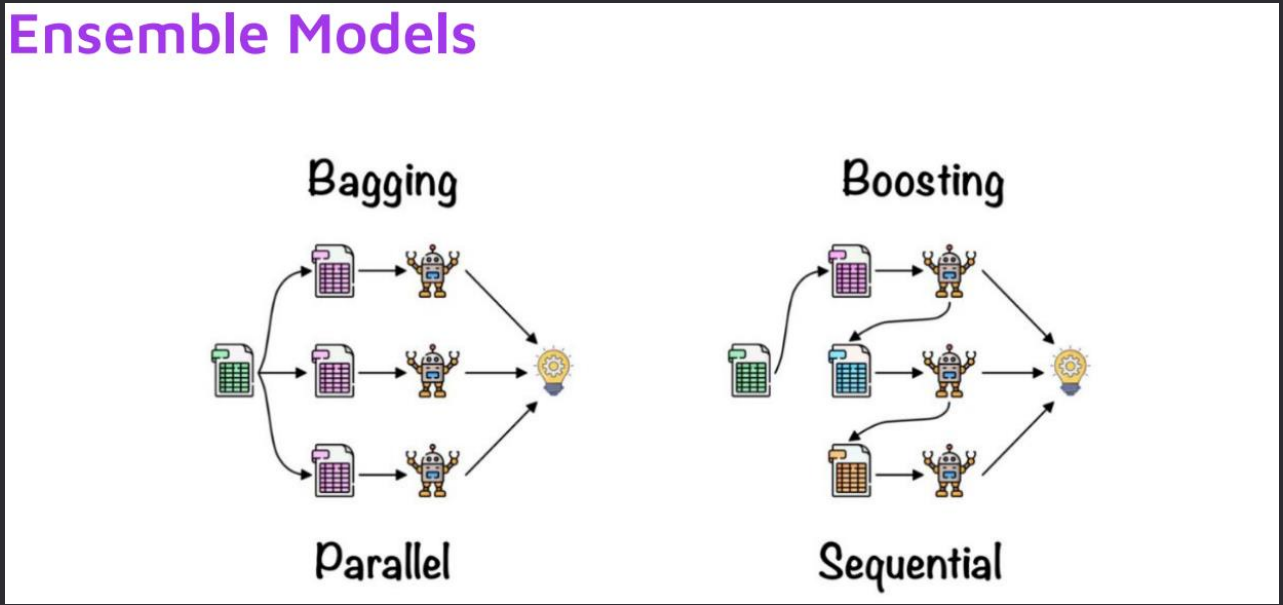
Modelling

- **Random Forest**

Random forests is an ensemble learning method. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.

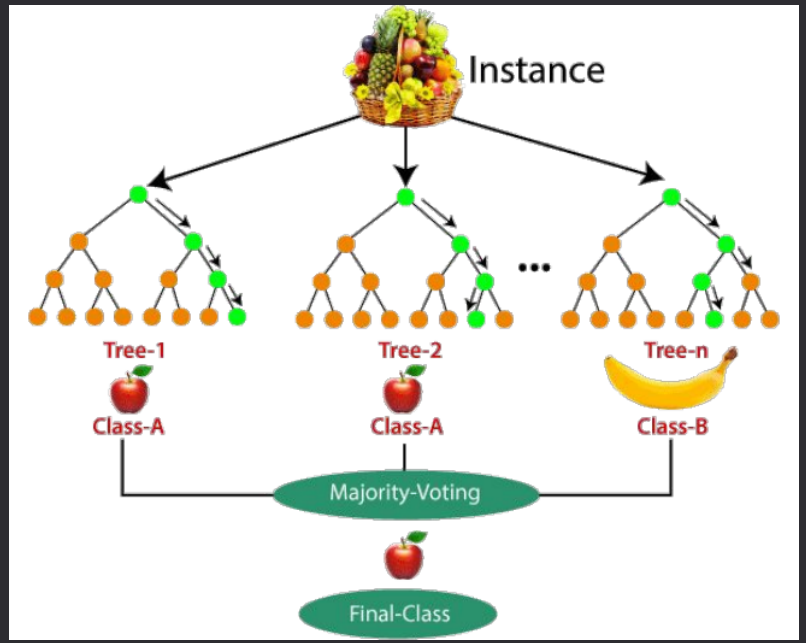
Modelling

- **Random Forest**



Modelling

- **Random Forest:**



04

Supervised Learning - Regression

Model Evaluation (Classification)

Model Evaluation

Confusion Matrix in Classification

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 🕵)
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉ ✉ 320	FALSE NEGATIVE (FN) ✉ 🕵 43
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉ ✉ 20	TRUE NEGATIVE (TN) ✉ 🕵 538

Model Evaluation

Accuracy

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 🕵)
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉ ✉ 320	FALSE NEGATIVE (FN) ✉ 🕵 43
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉ ✉ 20	TRUE NEGATIVE (TN) ✉ 🕵 538

Accuracy = (320+538)/(320+538+43+20)

= 0.93

Model Evaluation

- Is measuring the accuracy of a model enough for evaluation?
- No, we have to match it with the case. Accuracy is usually only suitable when the comparison of the number of data labels is actually relatively the same (balance)

Model Evaluation

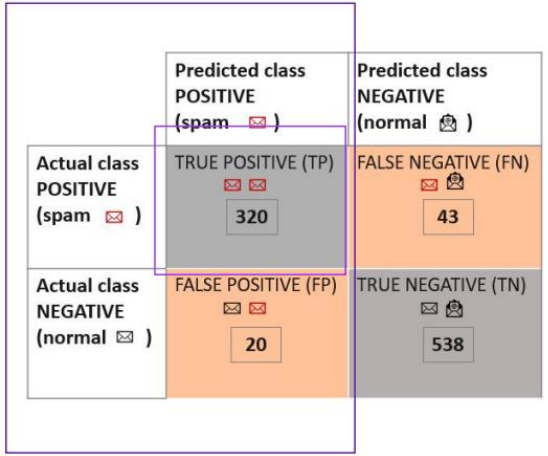
- **Precision**

Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. Precission menjawab pertanyaan “Berapa persen pelanggan yang benar churn dari keseluruhan pelanggan yang diprediksi churn?”

Model Evaluation

Precision

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$



Precision = (320)/(320+20)
= 0.91

Model Evaluation

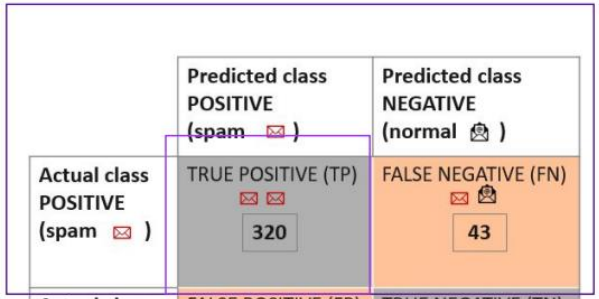
- **Recall**

Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Recall menjawab pertanyaan “Berapa persen pelanggan yang diprediksi churn dibandingkan keseluruhan pelanggan yang sebenarnya churn”.

Model Evaluation

Recall (True Positive Rate)

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$



Recall = (320)/(320+43)
= 0.88

Model Evaluation

F1 Score

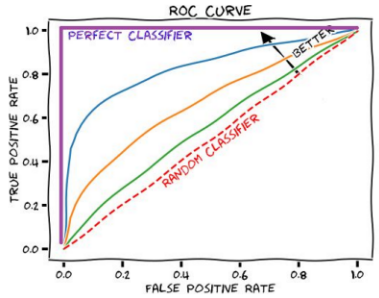
F1 Score merupakan perbandingan rata-rata presisi dan recall yang dibobotkan.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Model Evaluation

Area Under ROC Curve (AUC)

ROC is a probability curve and AUC represents the degree or measure of separability. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.



Model Evaluation

In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

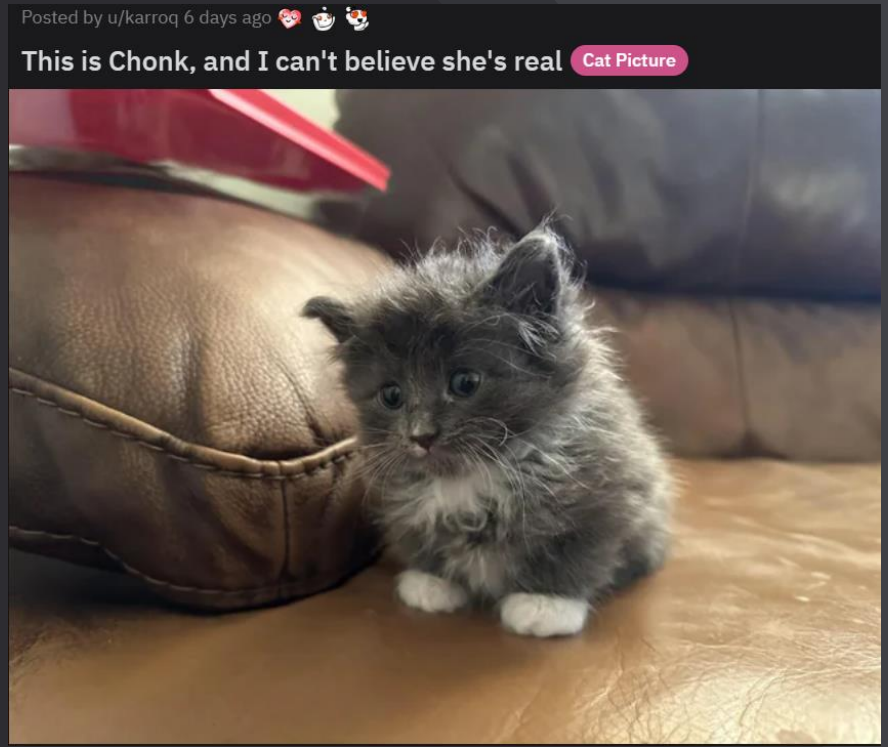
Model Evaluation

The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but **not ideal for the imbalanced class problem**.

For an imbalanced class dataset, **AUC and F1** score is a more appropriate metric.

Thanks!

Do you have any questions?



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**