머글들을 위한
Machine Learning 4
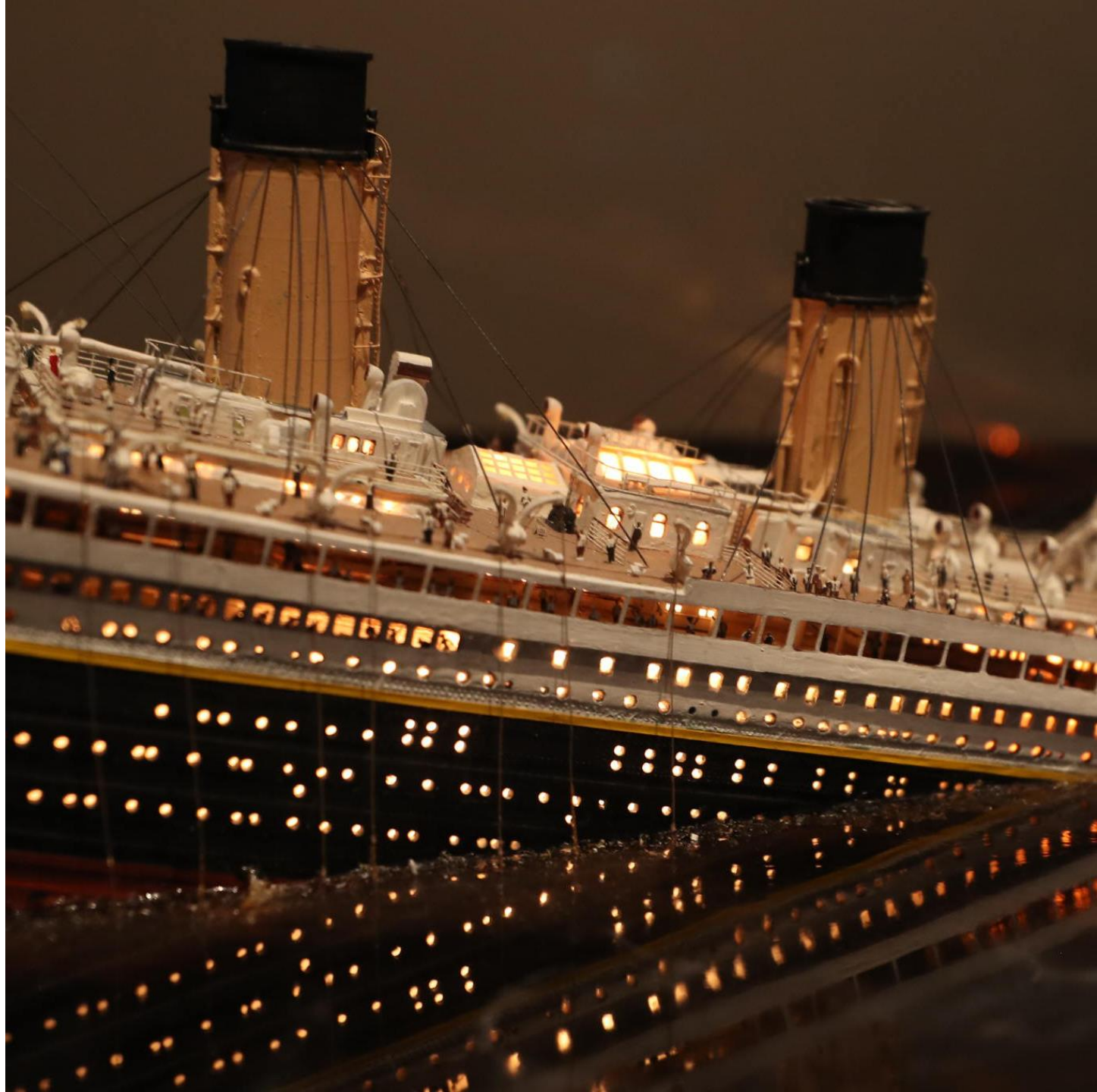
김영욱
부장/ PS / Microsoft

youngwook@outlook.com
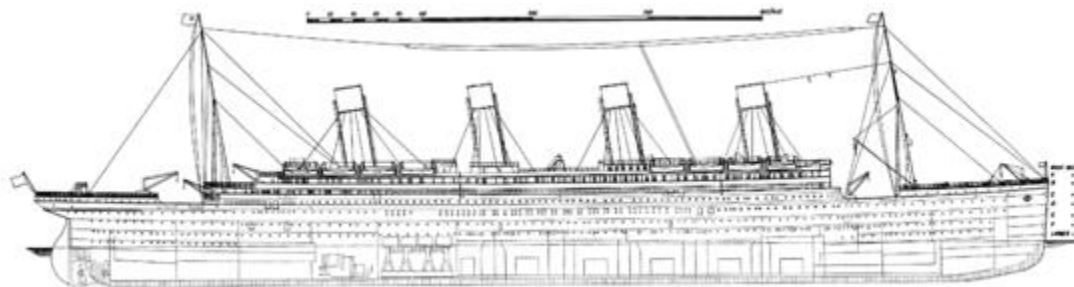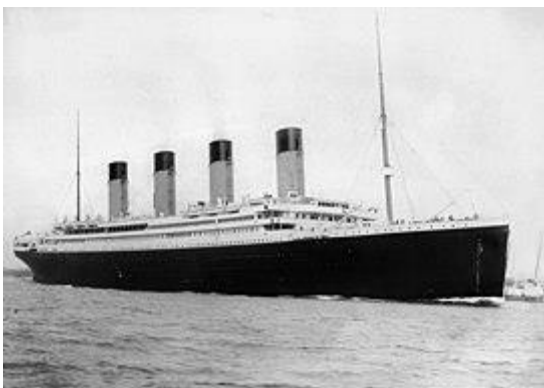Blog: Youngwook.com

Microsoft

Titanic

문제 정의

어떤 모델을 만들 것인가?

1514명 사망

탑승자 종류 ▲	총원 ⬥	구조율 ⬥	사망율 ⬥	구조인원 ⬥	사망인원 ⬥
1등실, 남성	175	33%	67%	57	118
1등실, 어린이	6	83%	17%	5	1
1등실, 여성	144	97%	3%	140	4
2등실, 남성	168	8%	92%	14	154
2등실, 어린이	24	100%	0%	24	0
2등실, 여성	93	86%	14%	80	13
3등실, 남성	462	16%	84%	75	387
3등실, 어린이	79	34%	66%	27	52
3등실, 여성	165	46%	54%	76	89
남성 총합	1690	20%	80%	338	1352
승무원, 남성	885	22%	78%	192	693
승무원, 여성	23	87%	13%	20	3
어린이 총합	109	51%	49%	56	53
여성 총합	425	74%	26%	316	109
전체 총합	**2224**	**32%**	**68%**	**710**	**1514**

Feature

Label

이름	나이	성별	선실 등급	티켓넘버	티켓 요금	부모 자식	형제 자매	키	출항지	생존 여부
Rose	25	여	A	EA-1039	300	2명	1명	167	런던	Y
Jack	20	남	C	GE-3059	29	-	-	178	도버	N
Mark	57	남	B	BA-2031	89	4명	3명	167	뉴포트	N
Andy	48	남	B	NN-3928	102	5명	7명	182	런던	Y

https://cyc1am3n.github.io/2018/10/09/my-first-kaggle-competition_titanic.html

https://www.kaggle.com/c/titanic/data

Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 10,820 teams · Ongoing

Overview   **Data**   Notebooks   Discussion   Leaderboard   Rules   **Join Competition**

Data Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth
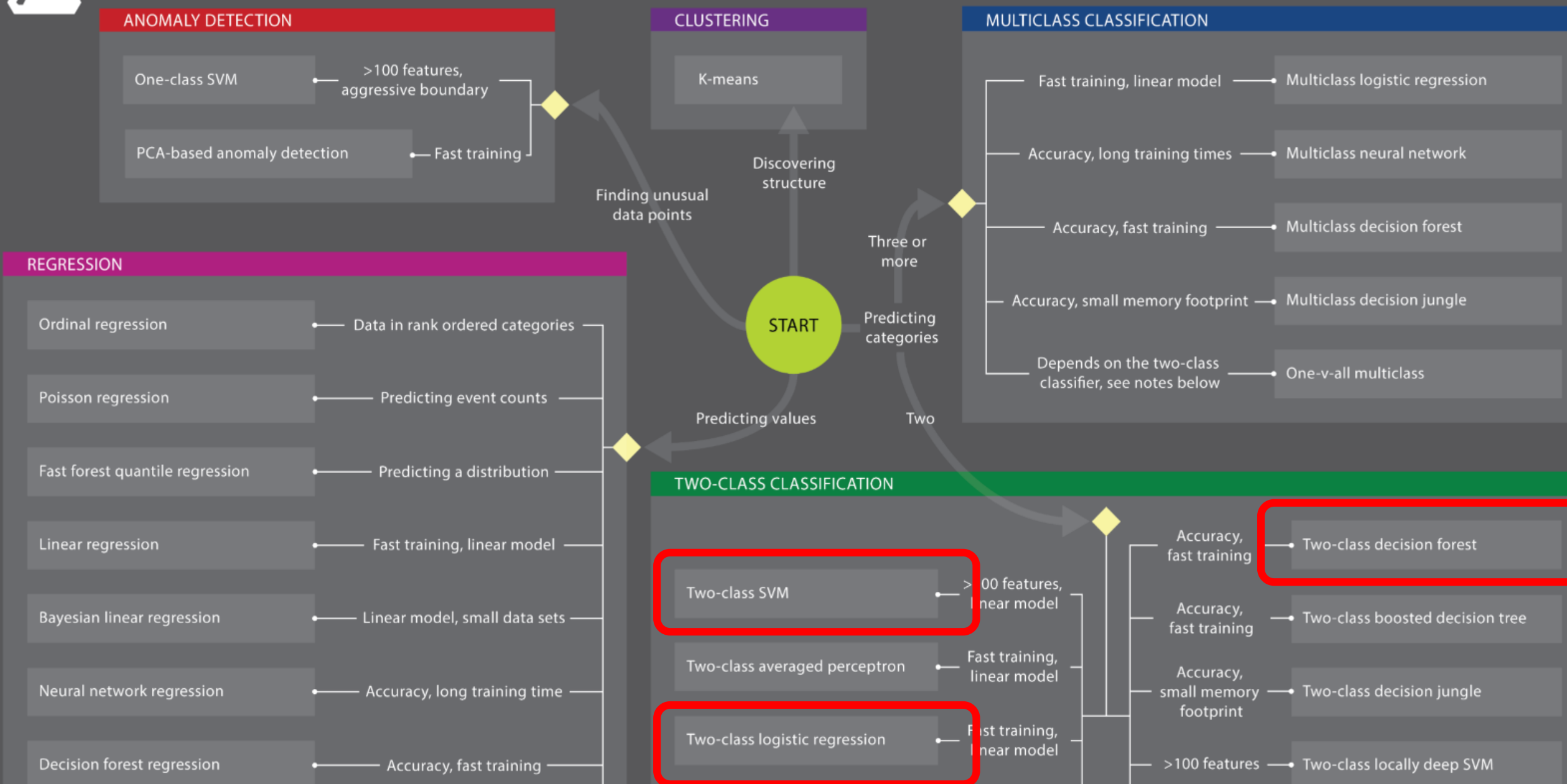
실습1:
Titanic Data Preprocessing

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

ANOMALY DETECTION

One-class SVM —— >100 features, aggressive boundary

PCA-based anomaly detection —— Fast training

Finding unusual data points

CLUSTERING

K-means

Discovering structure

MULTICLASS CLASSIFICATION

Fast training, linear model —— Multiclass logistic regression

Accuracy, long training times —— Multiclass neural network

Accuracy, fast training —— Multiclass decision forest

Accuracy, small memory footprint —— Multiclass decision jungle

Depends on the two-class classifier, see notes below —— One-v-all multiclass

Three or more

Predicting categories

START

REGRESSION

Ordinal regression —— Data in rank ordered categories

Poisson regression —— Predicting event counts

Fast forest quantile regression —— Predicting a distribution

Linear regression —— Fast training, linear model

Bayesian linear regression —— Linear model, small data sets

Neural network regression —— Accuracy, long training time

Decision forest regression —— Accuracy, fast training

Predicting values

Two

TWO-CLASS CLASSIFICATION

Two-class SVM —— >100 features, linear model

Two-class averaged perceptron —— Fast training, linear model

Two-class logistic regression —— Fast training, linear model

Accuracy, fast training —— Two-class decision forest

Accuracy, fast training —— Two-class boosted decision tree

Accuracy, small memory footprint —— Two-class decision jungle

>100 features —— Two-class locally deep SVM

실습1:
Titanic Machine Learning

https://docs.microsoft.com/ko-kr/azure/machine-learning/studio/algorithm-choice

감사합니다

Thank you~!