ENSEMBLE LEARNING

a road to have super models

Bagus Sartono

Department of Statistics – IPB University

in collaboration with: Gerry Alfa Dito and Fransdana Nadeak

Presented in the 9th Basic Science International Conference
Brawijaya University – Malang
20-21 March 2019

what is a
super model?

definitely not these ones!

- what is the super model?
- how can ensemble help us?
- some empirical results of super learner methodology
- conclusion

outline

super predictive-model

has excellent accuracy

be tough in handling
ill-conditioned datasets

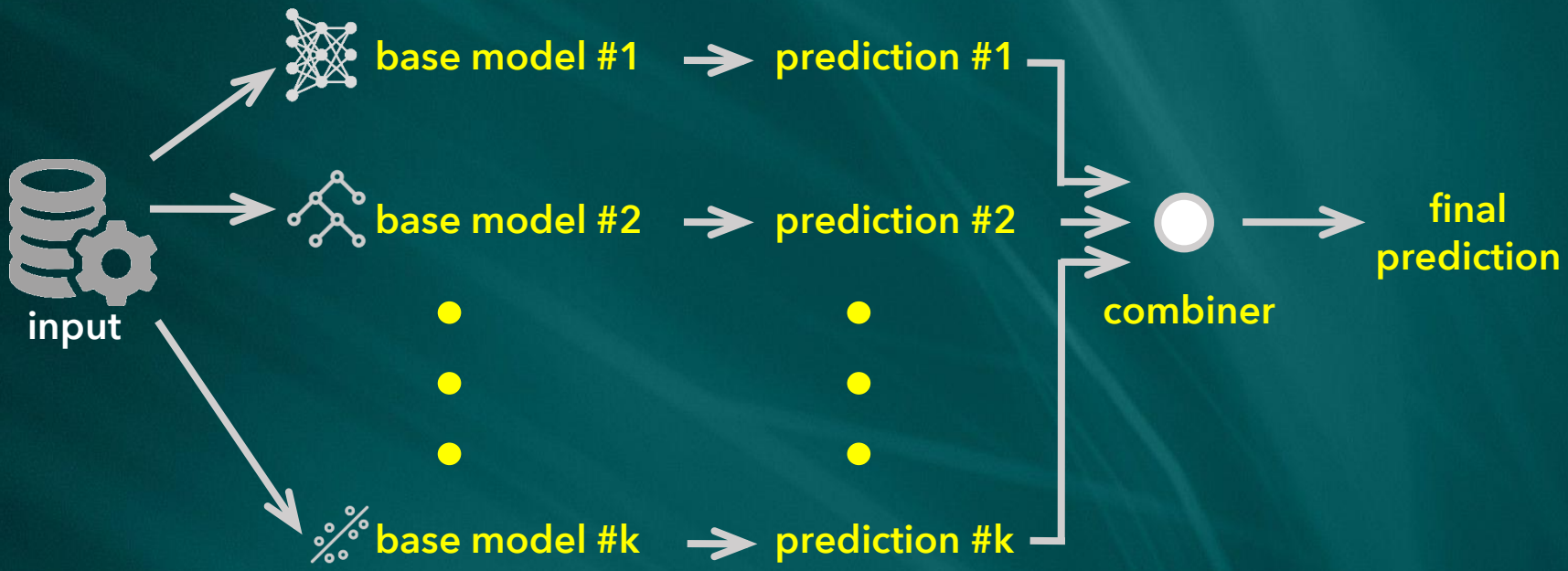
demands on predictive models

- business
  - propensity
  - risk
  - behavior
- bio-science
  - risk of disease
  - content of substances
- image analysis
  - predicting land-cover type from satellite image
- education
  - predicting student performance

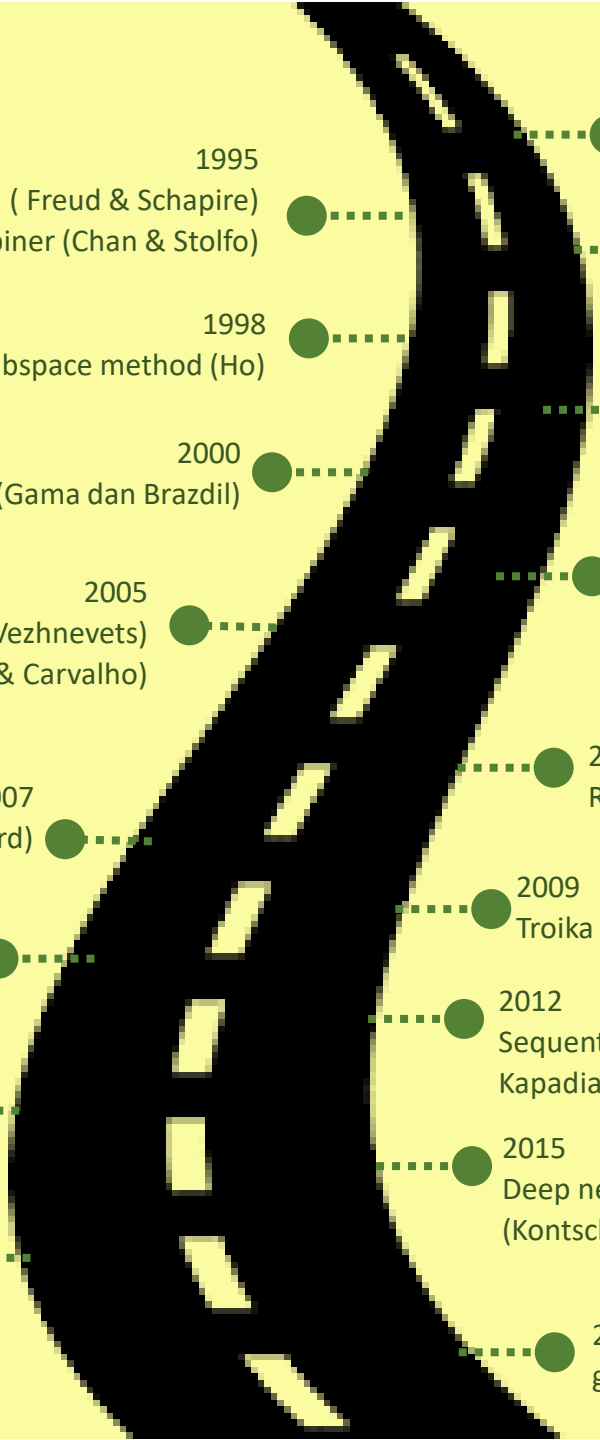
ensembling model could increase the model performance

fact or fake?

what is the ensemble approach?



milestones



1992 Stacking (Wolpert)

1995
Adaboost (Freud & Schapire)
Combiner (Chan & Stolfo)

1996
Bagging (Breiman)

1998
random subspace method (Ho)

1999
Wagging/Weight
aggregating(Bauer & Kohav)

2000
Cascade Generalization(Gama dan Brazdil)

2001
Random Forest (Breiman)
Gradient Boosting Machine
(Friedman)

2005
Modest AdaBoost (Vezhnevets & Vezhnevets)
Stacked sequential learning (Cohen & Carvalho)

2006
Rotation Forest (Rodriguez *et al*.)

2007
Super Learner (van der Laan, Polley, Hubbard)

2009
Troika (Menahem *et al*.)

2010
GA-stacking (Ledezma *et al*.)

2012
Sequential Stacking (Trivedi and
Kapadia)

2014
Random projection random discretization
ensembles (Ahmad & Brown)

2015
Deep neural decision forests
(Kontschieder *et al*.)

2016
XGBoost - Extreme Gradient Boosting
(Chen & Guestrin)

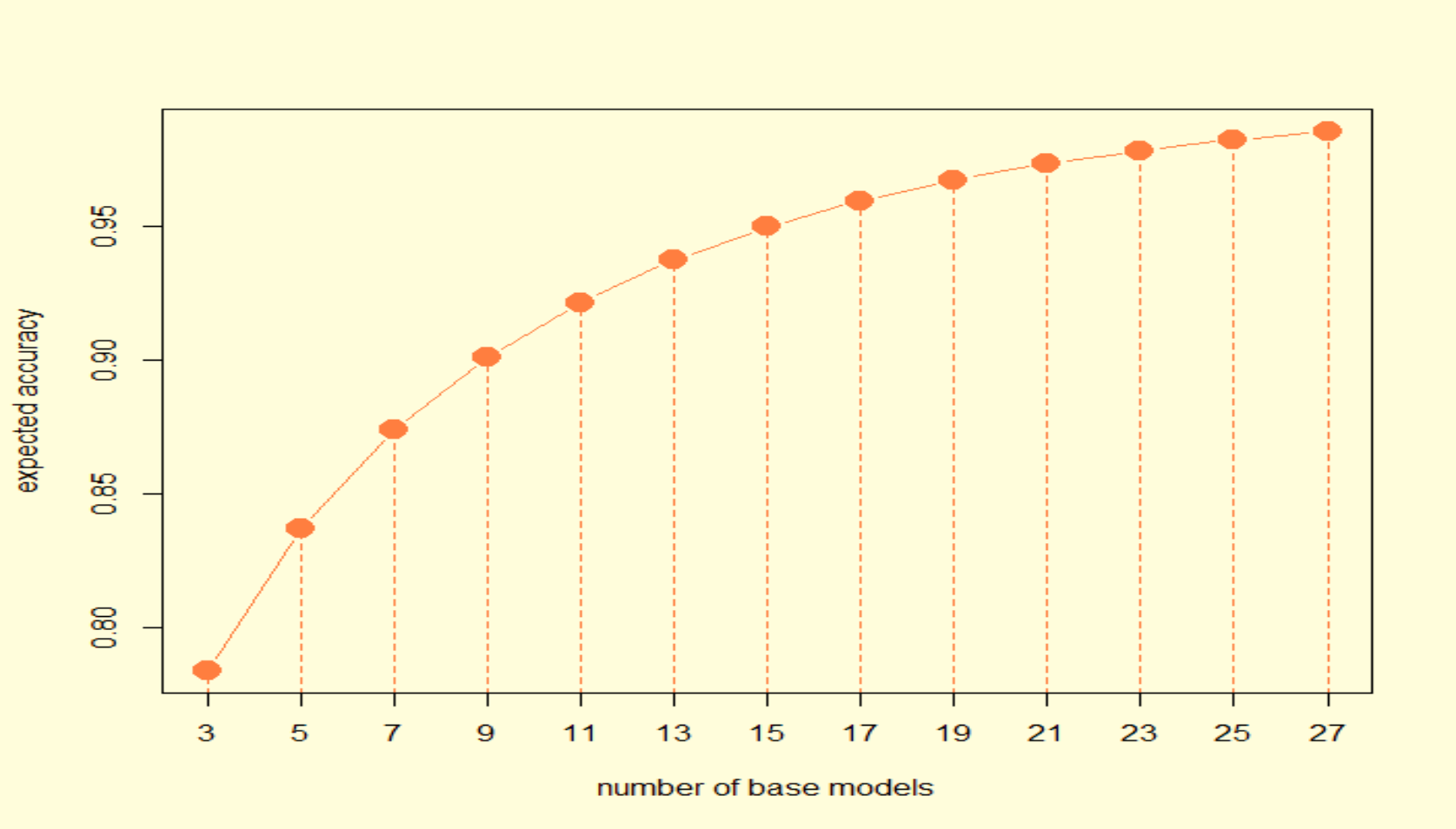
2017
gcForest (Zhou & Feng)

accuracy of the ensemble: an illustration

- 5 base models, independent
- accuracy level @70%
- ensemble prediction → majority vote
- correctly predict if at least 3 base models do so
- accuracy of the ensemble:

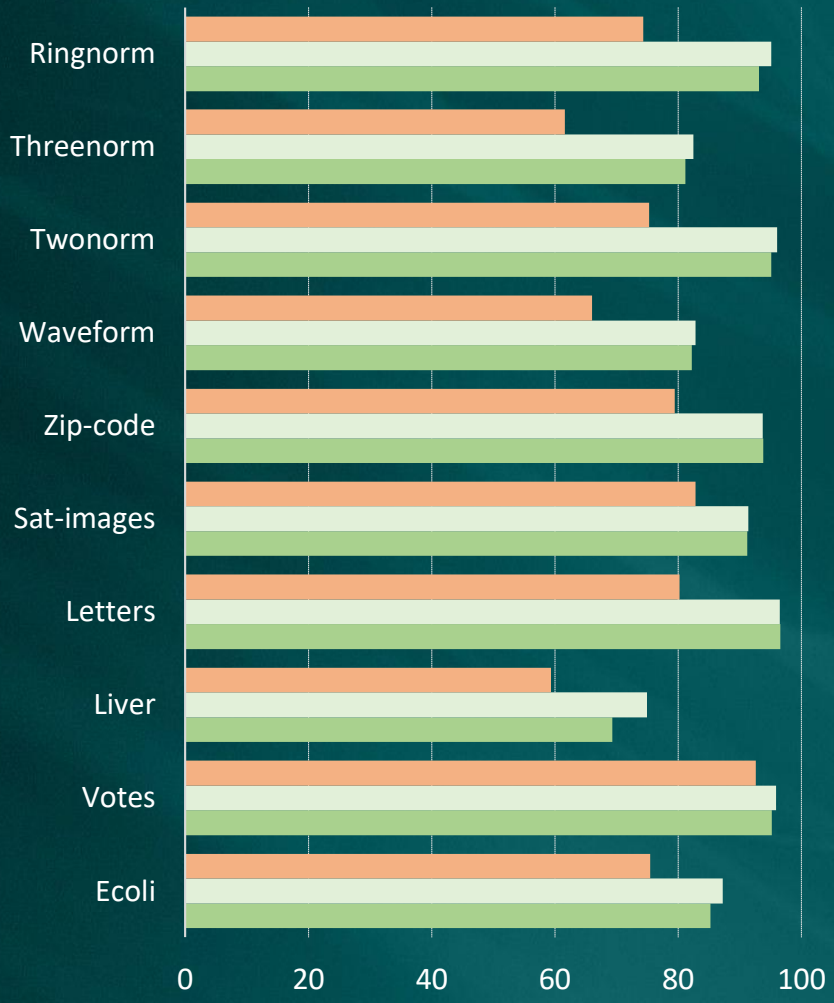
$$\sum_{k=3}^{5} \binom{5}{k} \times 0.7^{k} \times 0.3^{5-k} = 83.69\%$$

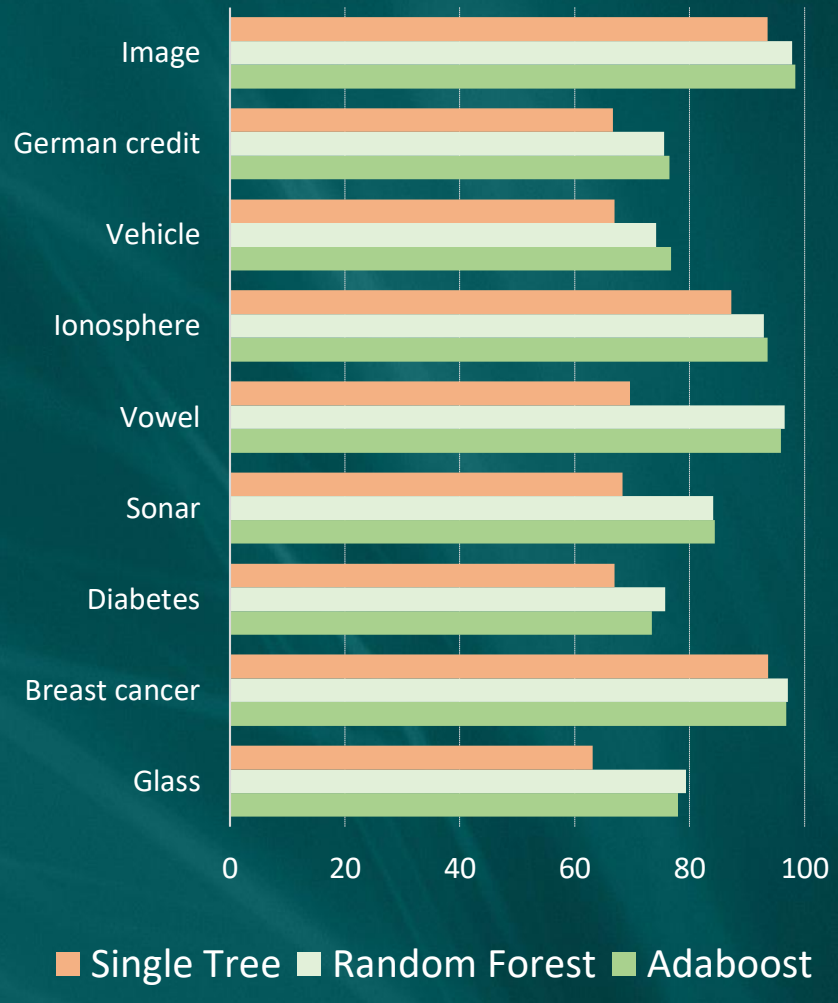
accuracy of the ensemble: an illustration



success story

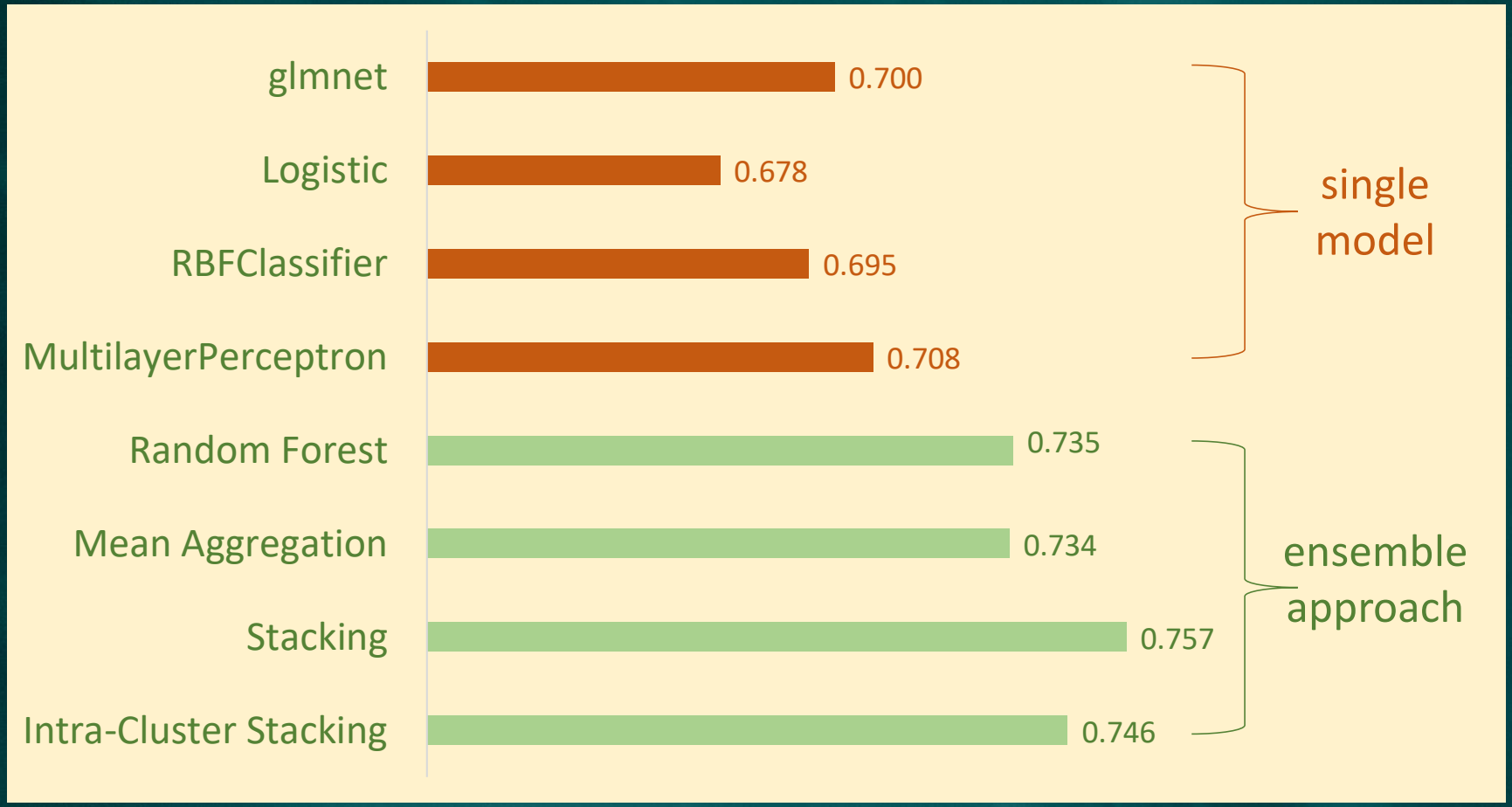
*Breiman L (2001). "Random Forests". Machine Learning. **45** (1): 5–32.*





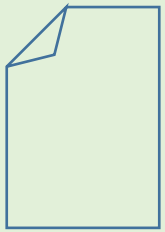
success story

Sean Whalen, Gaurav Pandey. 2013. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. ICDM 2013: 807-816



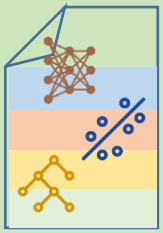
super learner methodology

how to create the combiner?



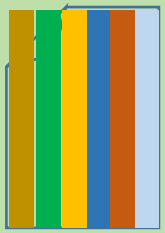
training dataset

Level-0 Data



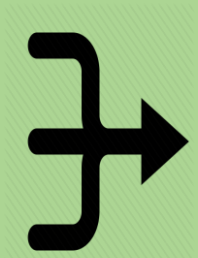
Partition data into k parts and run m base-models k times using obs from k-1 parts

Base Models



Predict the other fold, as in k-fold CV predictions

Level-1 Data



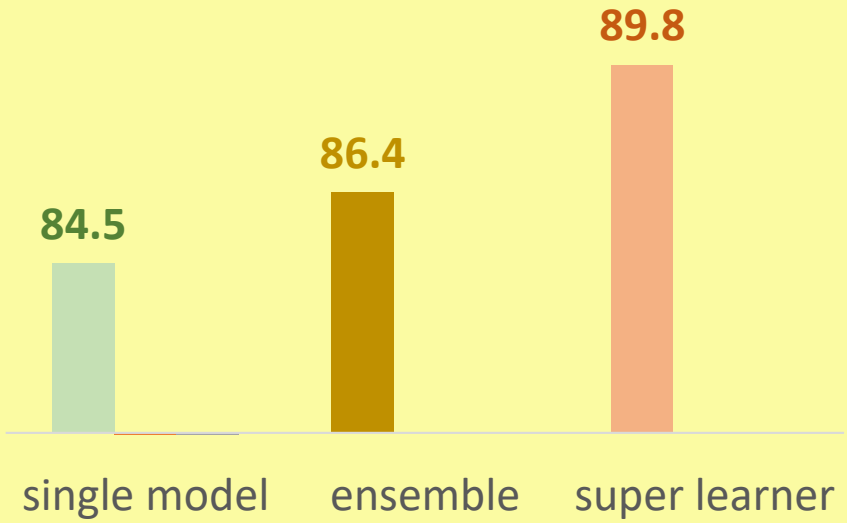
Run a meta-learner algorithm to combine using predictions as predictors

Meta Learner

some results

Gerry Alfa Dito & Bagus Sartono

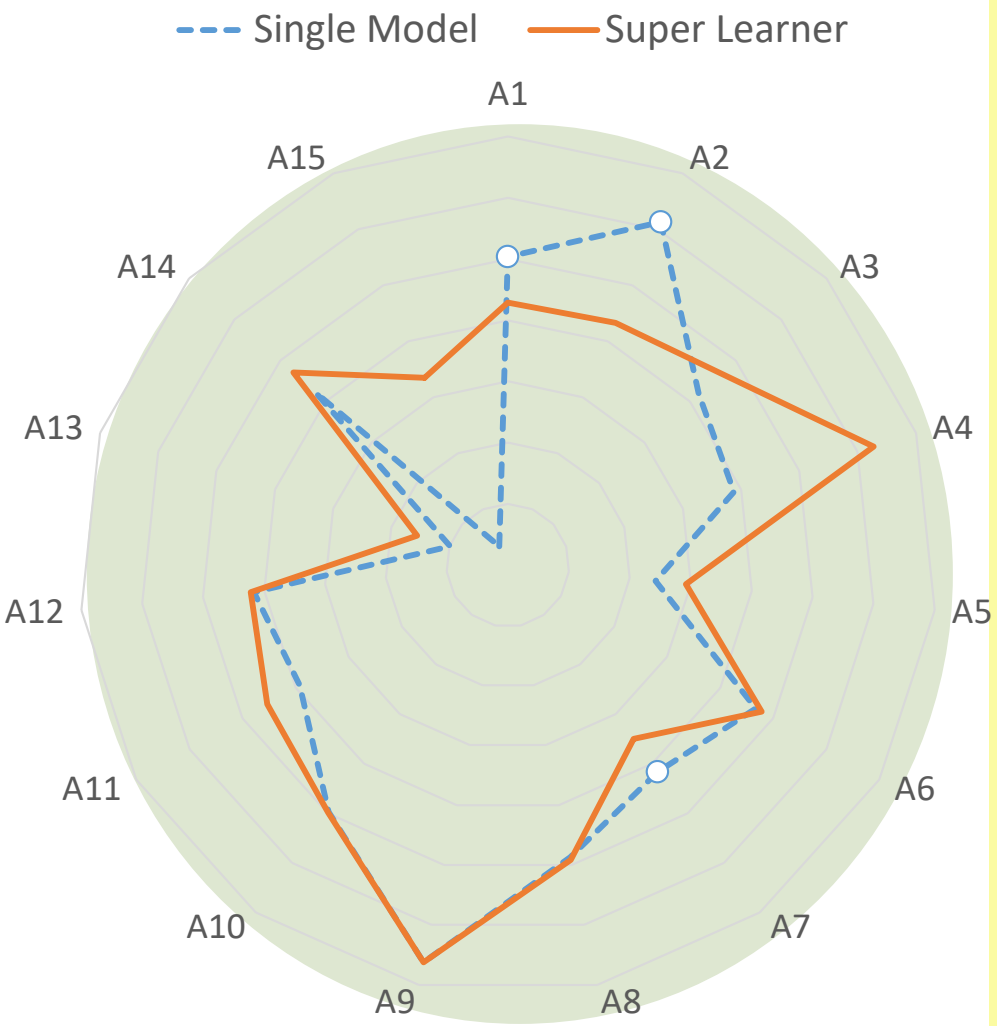
Data Source:
UCI Machine Learning Repository



Datasets	Disc	NN	SVM	CART	Bagging	Boost	RF	Cascade	Super Learner
Australian	85.9	84.2	86.9	85.9	87.9	87.4	87.2	86.8	89.9
Adult	78.1	80.6	-	86.1	85.5	-	85.9	86.3	87.2
Heart Disease	83.9	89.0	82.5	79.3	81.9	82.6	81.8	84.9	93.2
German	78.0	75.0	75.0	72.1	77.2	77.1	73.8	77.0	78.0
Hepatitis	78.4	81.9	85.4	80.5	83.5	86.2	85.2	84.7	87.1
Votes	96.4	96.1	96.8	96.7	96.4	95.6	95.5	96.7	93.1
Breast Cancer	94.5	98.1	96.2	95.1	96.8	97.0	97.4	97.4	99.3
Diabetes	77.3	71.5	76.6	73.5	78.2	77.8	75.8	77.6	79.1
Ionosphere	89.5	96.2	87.1	92.0	94.1	94.1	92.6	93.4	95.7
Sonar	74.7	90.4	75.9	71.6	82.2	87.0	82.1	80.8	95.7

some results

Fransdana Nadeak, Bagus Sartono & Anwar Fitrianto



Data Source:
Students' theses, 2011 - 2018

Dataset	Researcher	Year	Single Model
A1	Resty Indah Sari	2011	Stepwise Discriminant
A2	Gitania N Rahisti	2015	CART
A3	Rindy Pertiwi (2)	2013	CHAID
A4	Dimas Adiangga	2015	C5.0
A5	Nur Fitriani	2015	QUEST
A6	Nita Nurgenita	2015	Logistic Regression
A7	Adi Nugraha	2015	Logistic Regression
A8	Rossi A Barro	2013	Logistic Regression
A9	Shafa R Surbakti	2015	Logistic Regression
A10	Meita A Rubiati	2014	Logistic Regression
A11	Alfin Khairi	2014	Logistic Regression
A12	Rakhmawati	2011	Ridge Logistic Regression
A13	Devi Adrian	2018	K-Nearest Neighbor
A14	Dairul Fuhron	2018	Multivariate GLMM
A15	Dairul Fuhron(2)	2018	Multivariate GLMM

conclusion

need a super model?

think ensemble!

thank you

matur nuwun

contact:
bagusco@gmail.com
bagusco@ipb.ac.id