

Ensemble methods

Getting the best possible solution
by combining multiple models
(many heads are better than one)



Industry focus - healthcare

Data is noisy, inaccuracy can be fatal.



Instead of trying to learn one super accurate model, train a large number of low accuracy models, combine the predictions of weak models to gain a high accuracy 'meta-model'

The ensemble learning paradigm



What are the known weaknesses of machine learning algorithms?

- I. Bias
- II. Assumptions
- III. Variance/fluctuation against different data sets
- IV. Continues to learn incrementally
- V. Small sample limitation
- VI. Rubbish in, rubbish out
- VII. Minor changes can have big impact

All models are wrong, but some are useful

Parallel ensemble methods

Bagging : improve unstable estimations- eg RF model

Boosting: iteratively training the same weak model

Bootstrapping : drawing samples with replacement

Bootstrap Aggregating trains multiple models in parallel

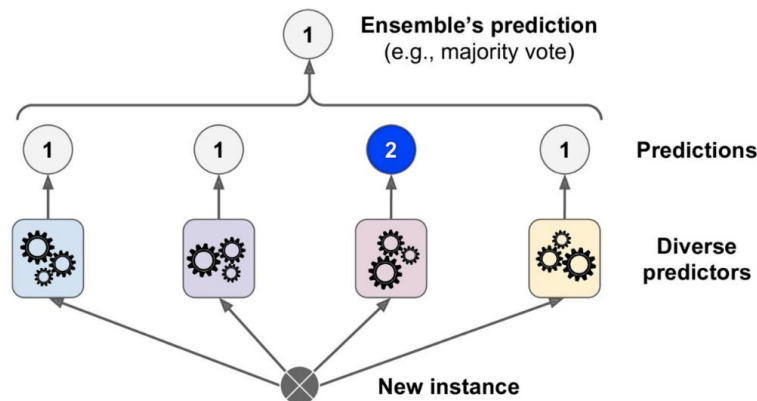


Figure 7-2. Hard voting classifier predictions



Parallel ensemble methods

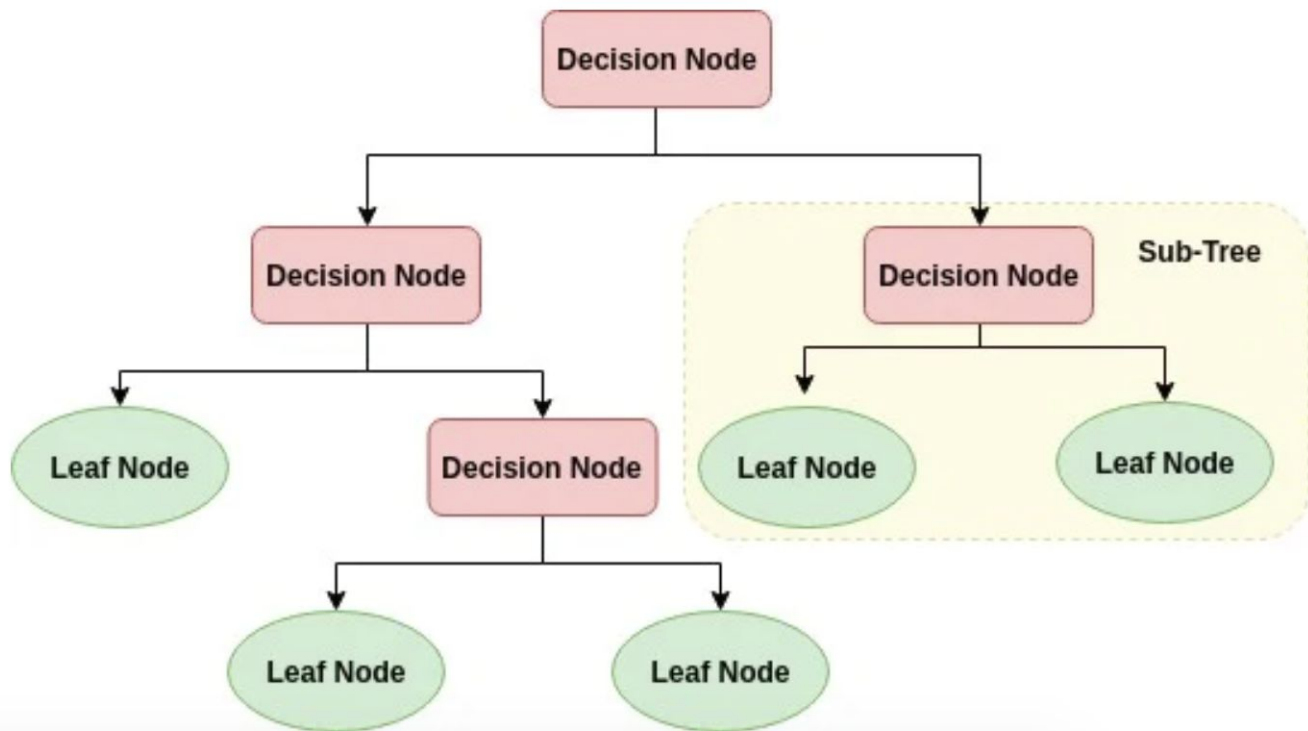
Bootstrap Aggregation Example : Random Forest

Widely used algorithm. For the given training set we create random samples and build decision tree models using each sample as the training set.

At each split in the learning process a random subset of the features is taken. This avoids correlation of the trees and the influence of strong features.

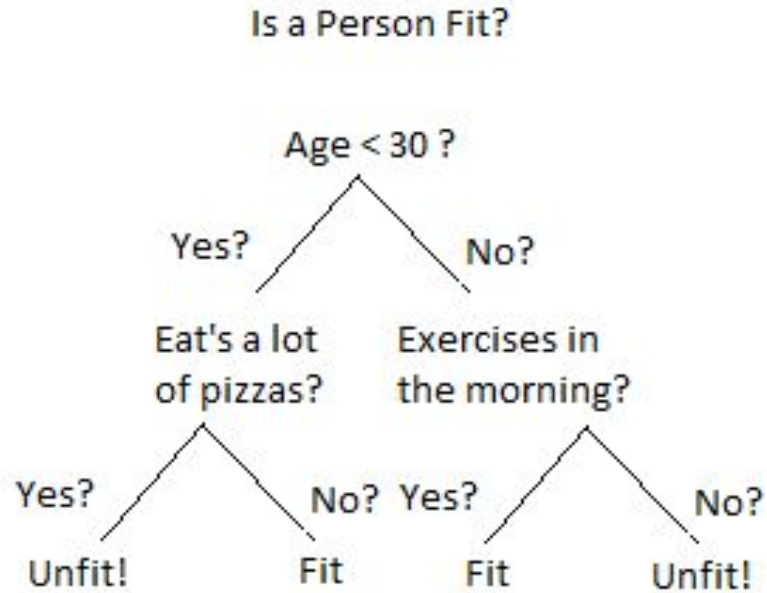
Using multiple samples of the same dataset reduces the variance of the final model, and thus reduces overfit.

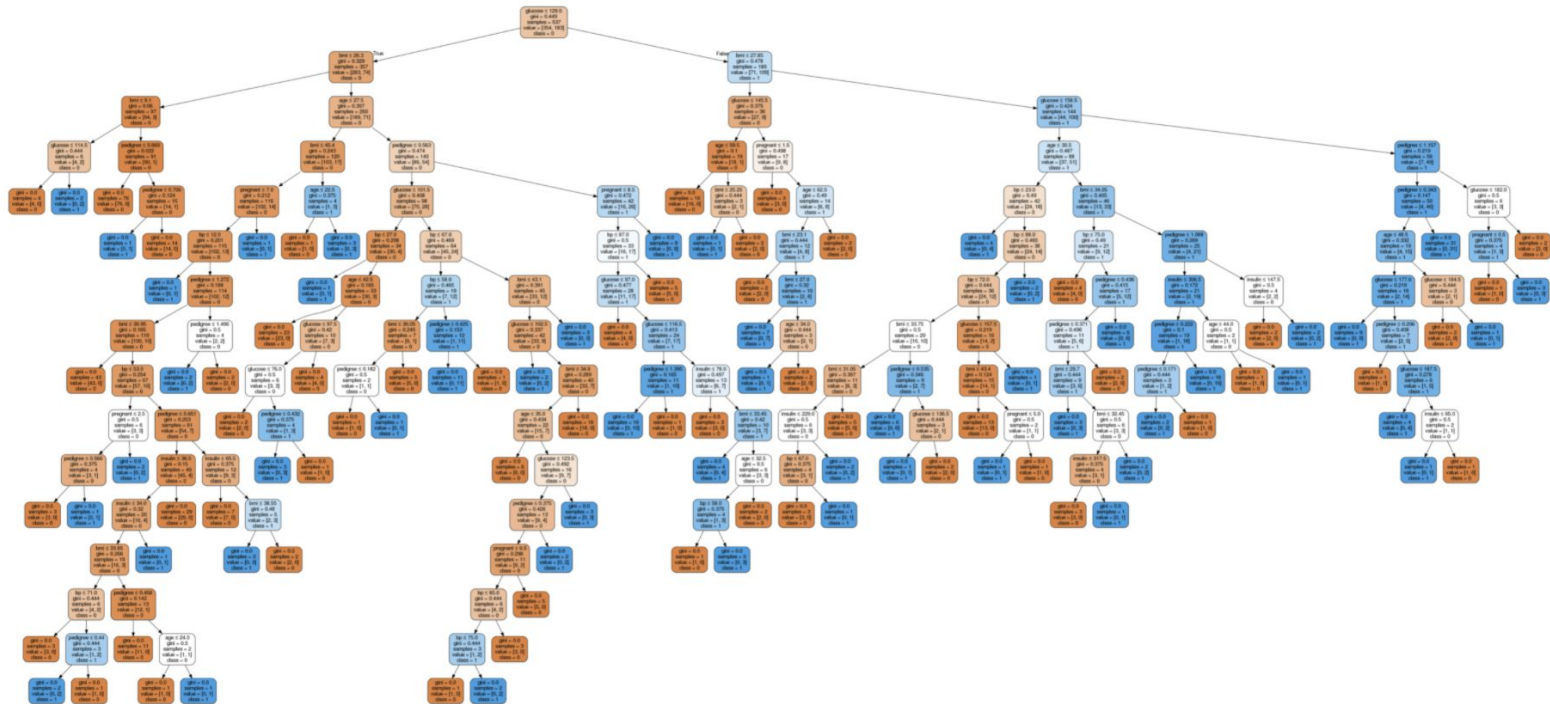
Decision tree- structure



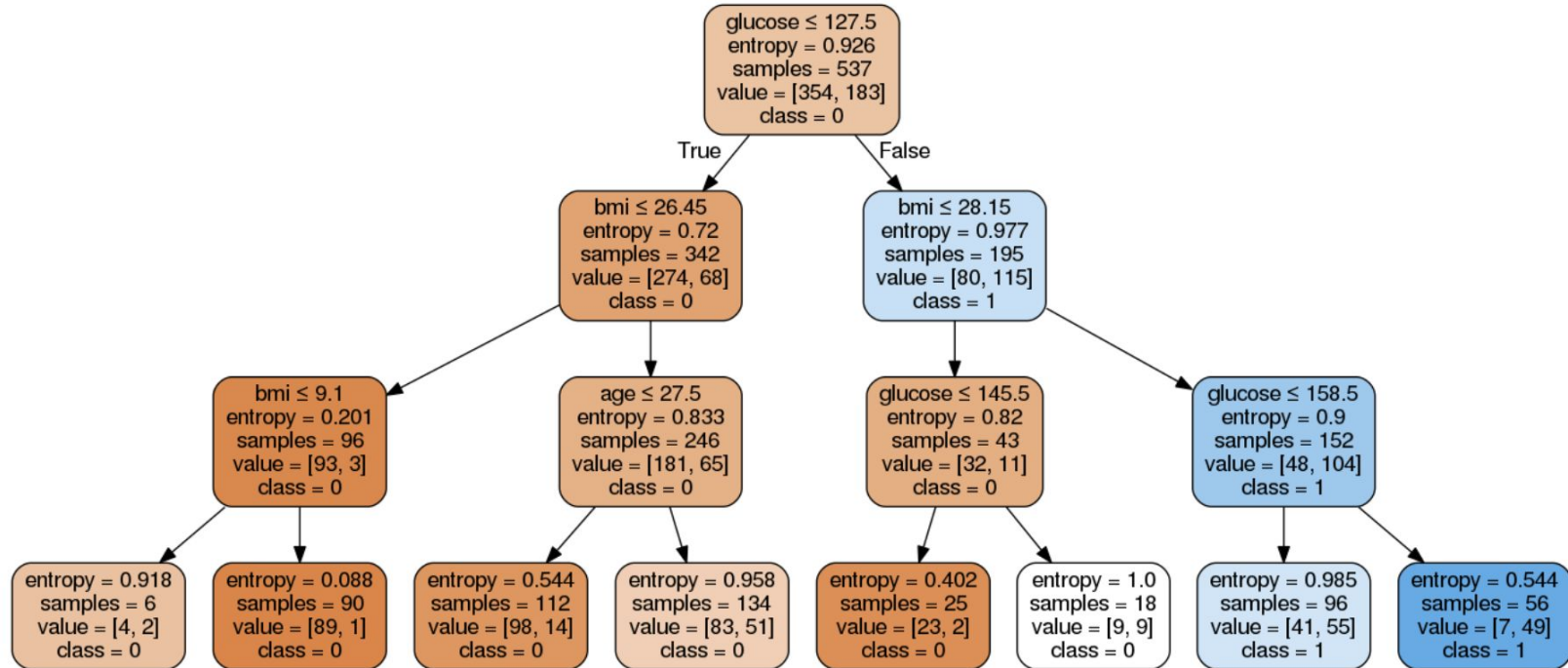


Decision tree- reminder



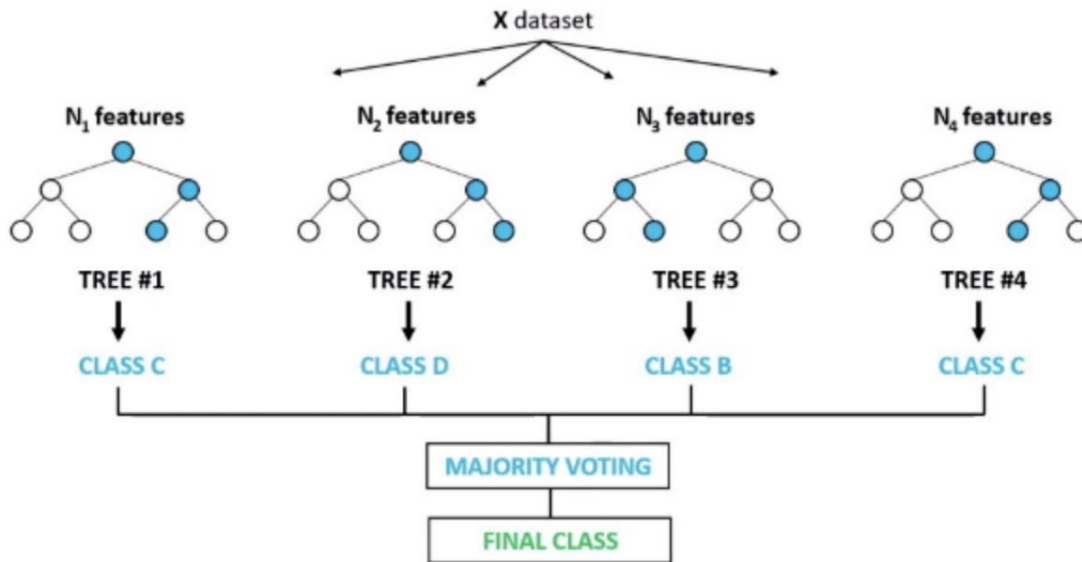


Decision tree- pruning



Random Forests- majority vote

Random Forest Classifier



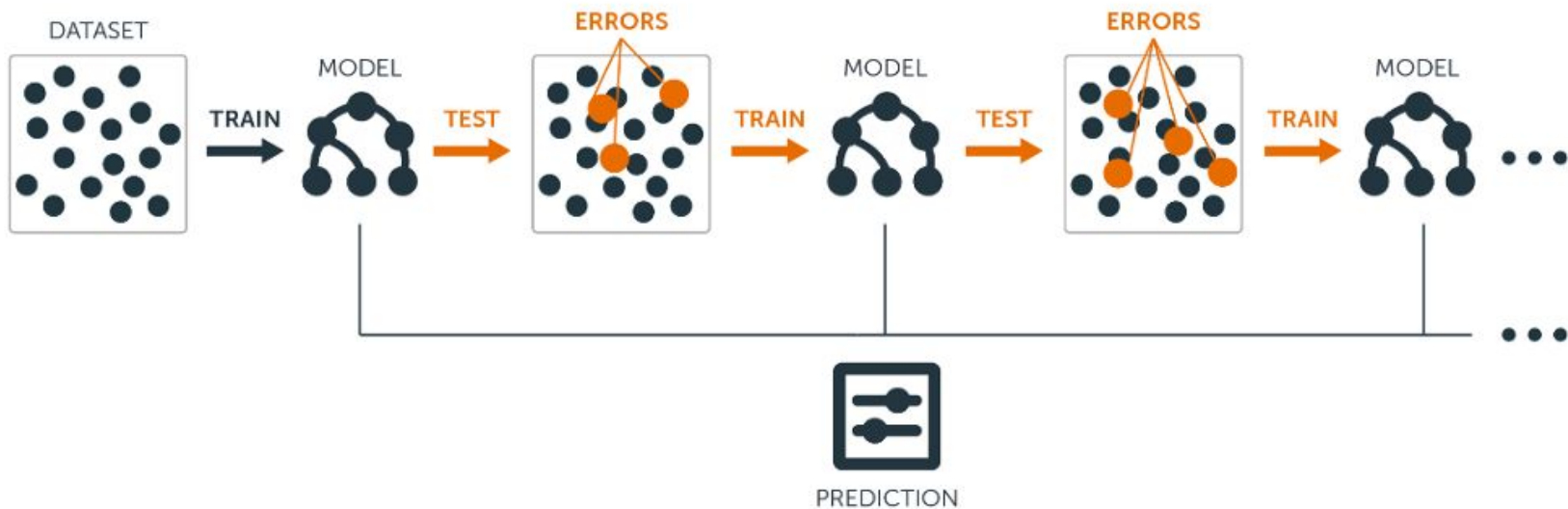


RF Examples from Medical journals

- Critical care study - ICU mortality with RF
- Covid 19 outcomes predictor
- Bed occupancy in ICU
- Predicting heart disease

Sequential ensemble methods

Boosting : sequentially reduce bias because each new model learns from the mistakes of the previous



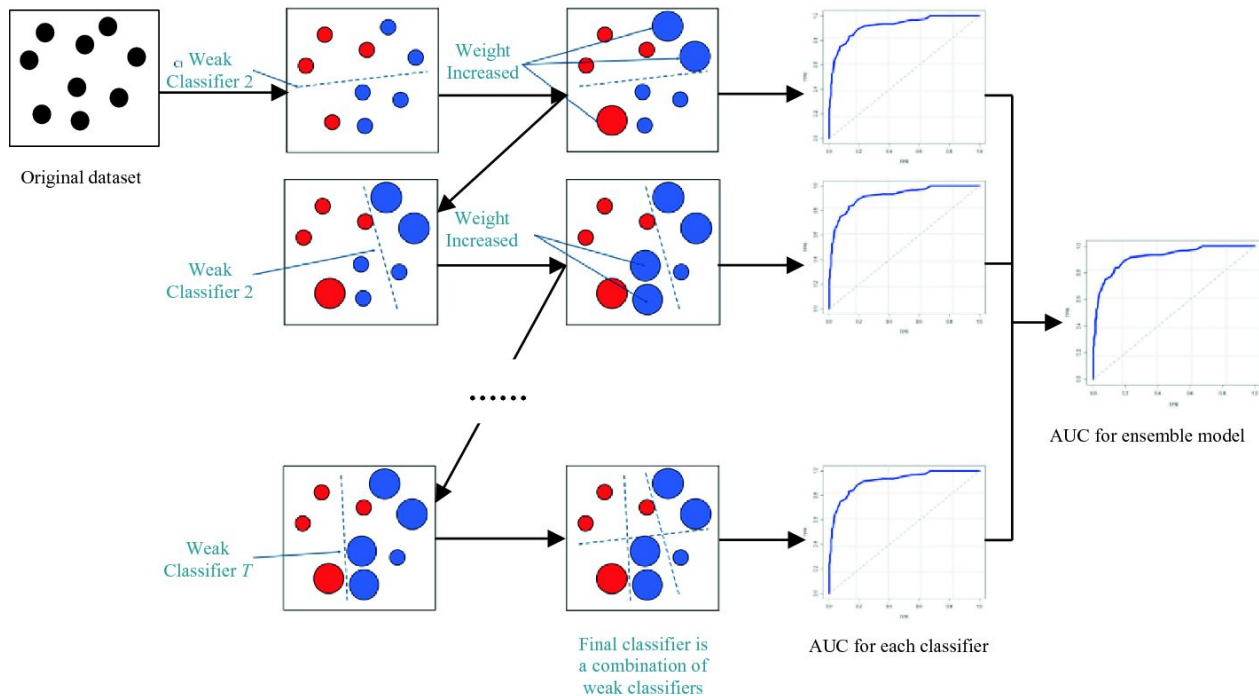


Sequential ensemble methods

Adaptive boosting : training made upon misclassified observations. Increase the weight of sample observations in the training data set, where observations are hard to classify. Final prediction is based on majority vote, weighted by individual accuracy per tree.

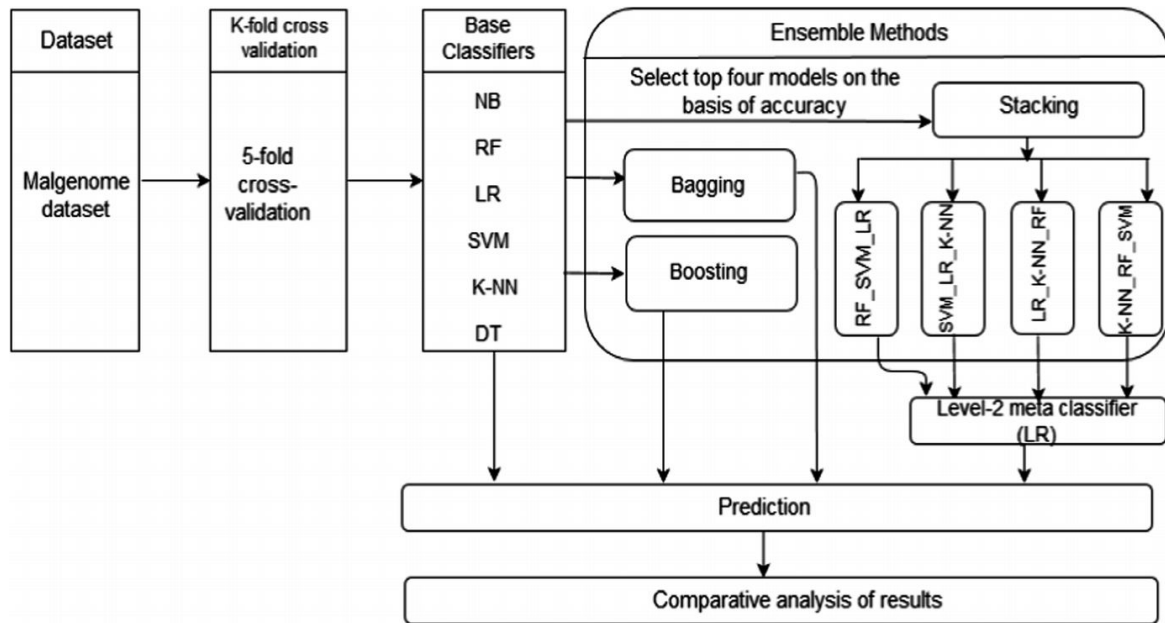
Gradient boosting : trains to minimise loss using residual error, rather than update the weights of data points. Decision trees based on purity scores are added to the model, against the prior tree residual error. The final prediction is the sum of all tree predictions.

Sequential ensemble methods





Stacking



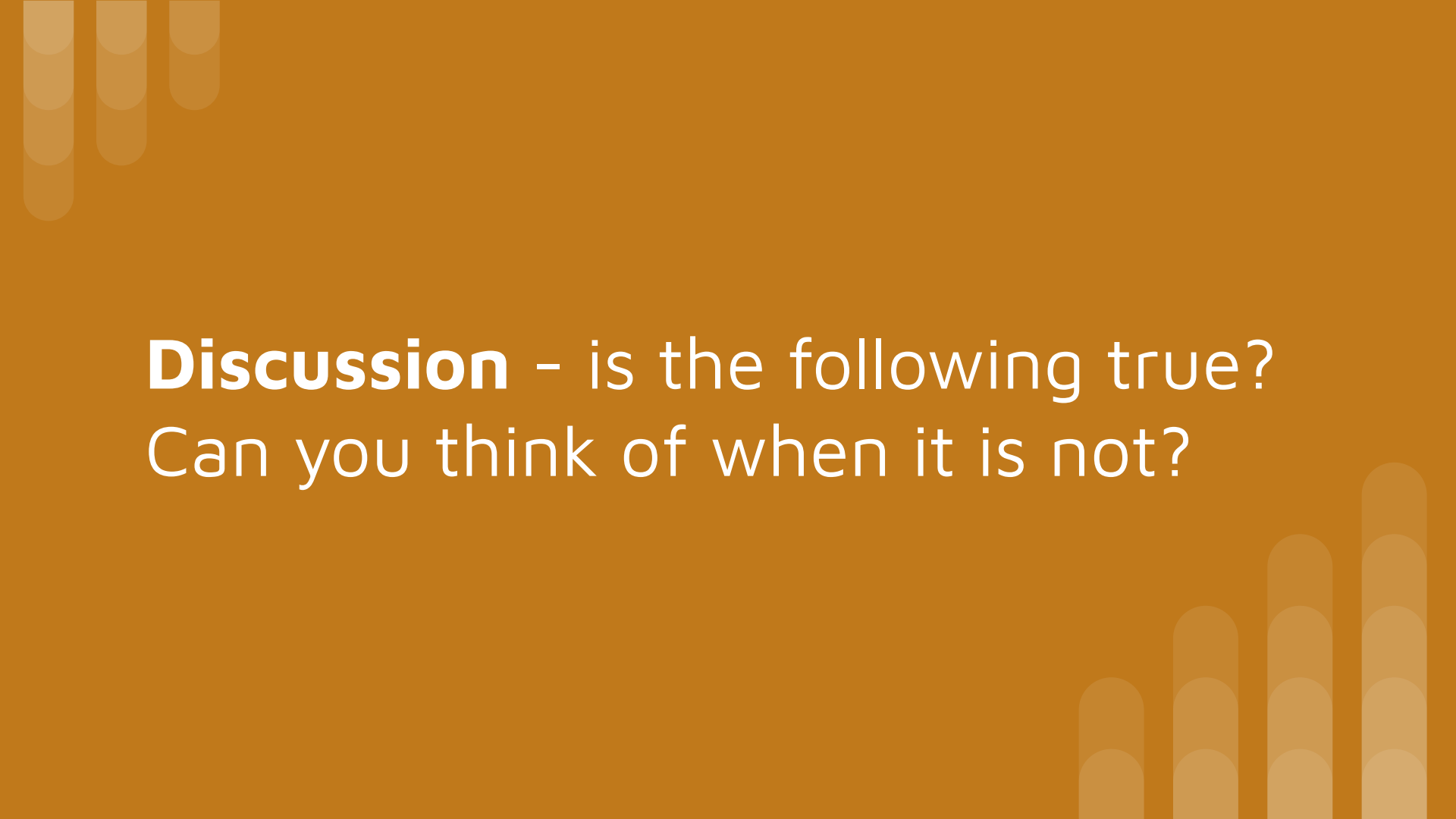
Stacked generalisation:

Given multiple models that are skillful in different ways, and the observed errors are uncorrelated, how to choose which to trust?

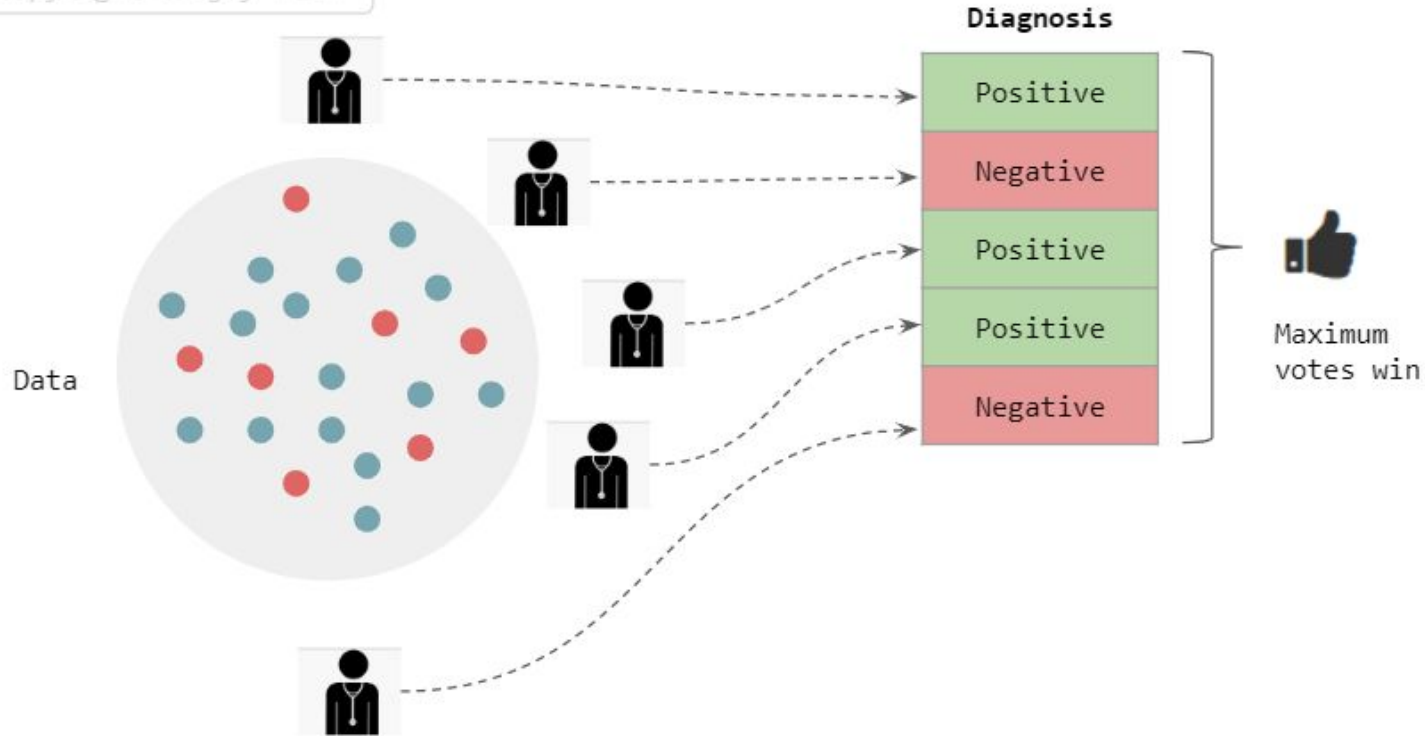
The new super learner model **learns** how to best combine 2+ predictions of base machine learning algorithms

The base models chosen are typically **different** (unlike bagging)

A **single simple meta** model is then used to combine eg linear regression

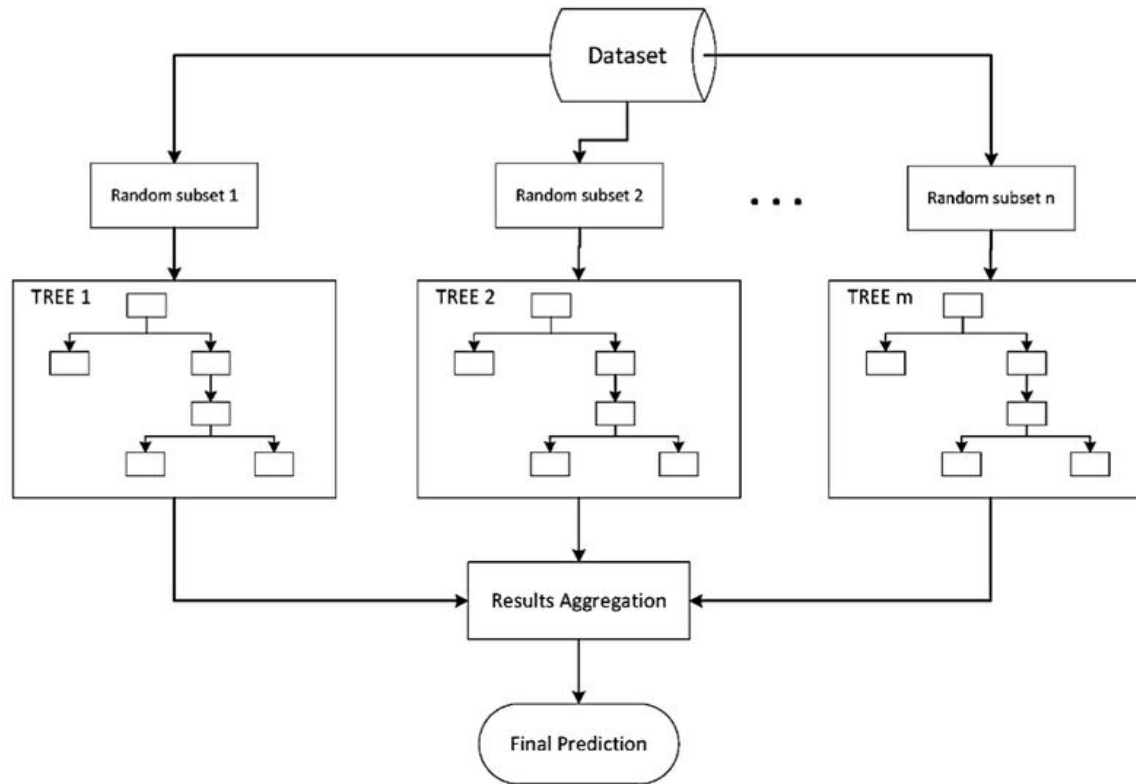


Discussion - is the following true?
Can you think of when it is not?



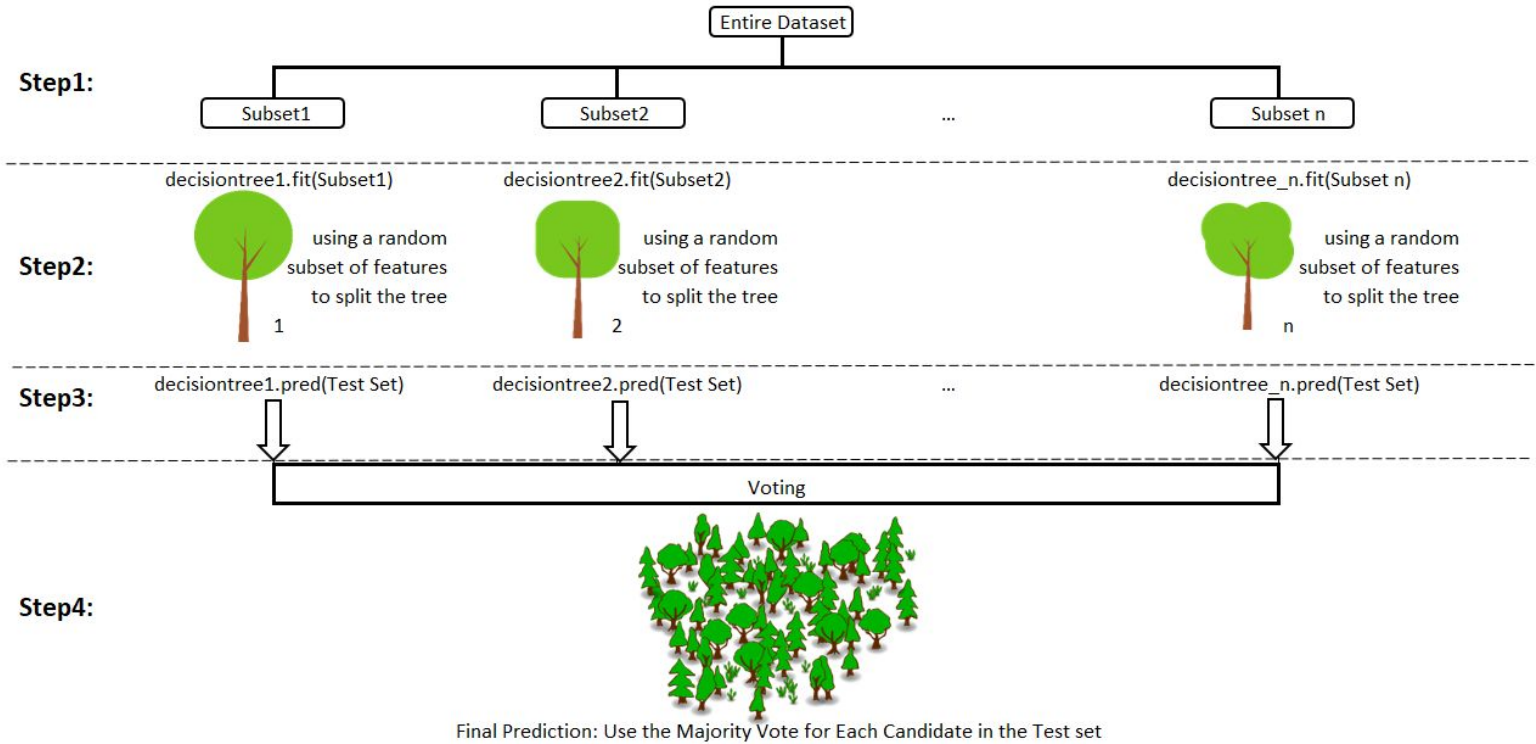
Theory 1 : taking the average will get us a more reliable result.

Counter argument :

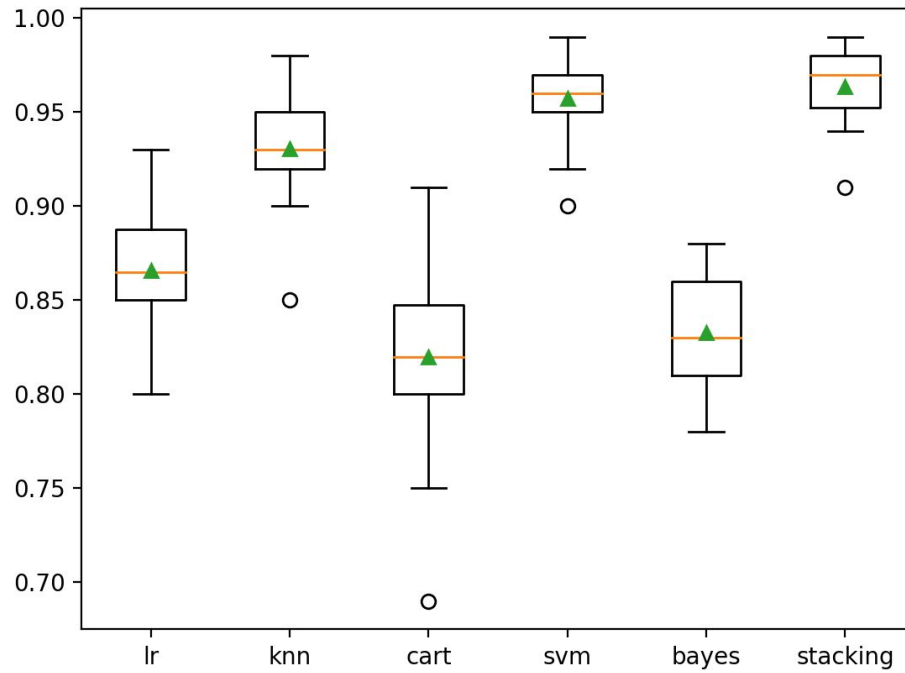


Theory 2 : Aggregating results and applying an algorithm will find the most accurate predictor.

Counter argument :



Theory 3 : override the effect of multiple appearances of strongly predictive variables via random selection
Counter argument :



Theory 4 : stacking will give us a better performance than any base model
Counter argument :



Case studies with ensemble learning

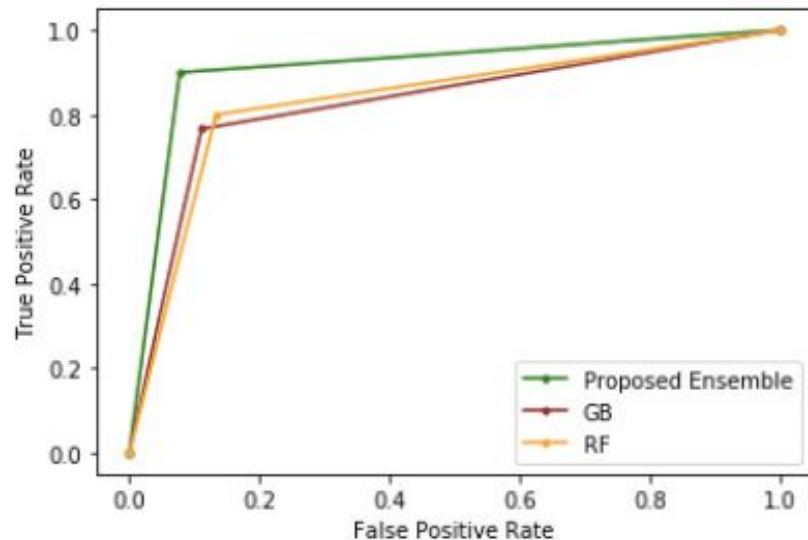




Case study - heart disease predictor

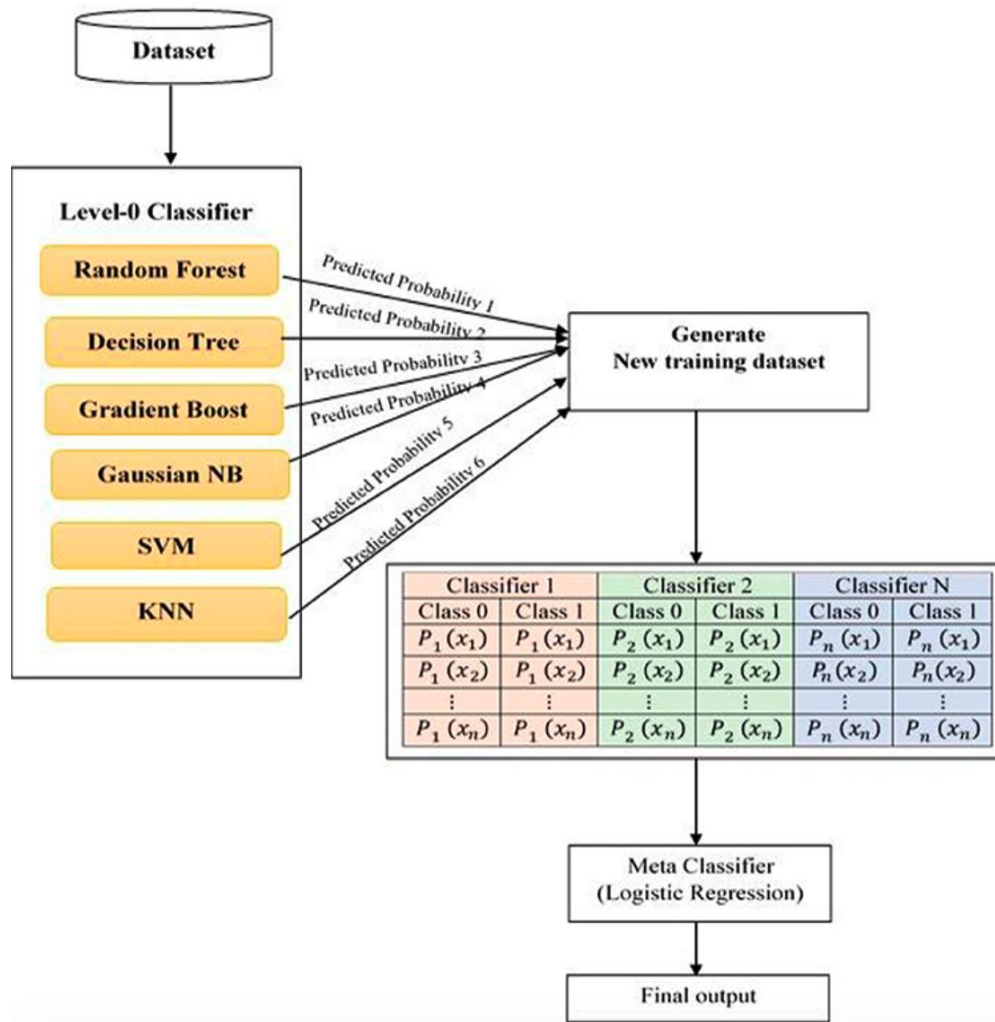
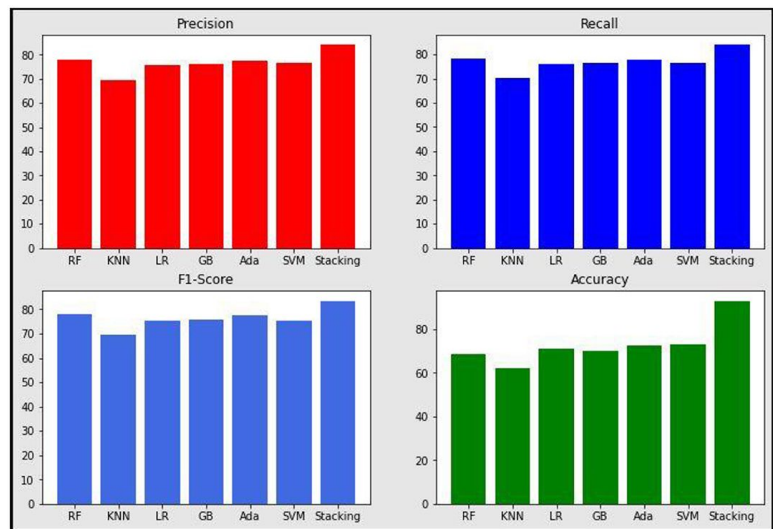
Heart disease prediction across two datasets, randomly partitioning the data into smaller subsets using a mean based split, followed by an accuracy based weighted classifier ensemble, achieves 93% accuracy.

Author(s)	Method	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score
Latha and Jeeva [5]	Majority vote with NB, BN, RF, and MLP	85.48	–	–	–
Ali et al. [6]	L_1 Linear SVM + L_2 Linear & RBF SVM	92.22	–	82.92	–
Mohan et al. [24]	HRFLM	88.4	90.1	92.8	90
Repaka et al. [25]	NB and AES	89.77	–	–	–
Samuel et al. [26]	ANN and Fuzzy_AHP	91	–	–	–
Proposed method	Randomized decision tree ensemble	93	96	91	93



Case study - diabetes predictor

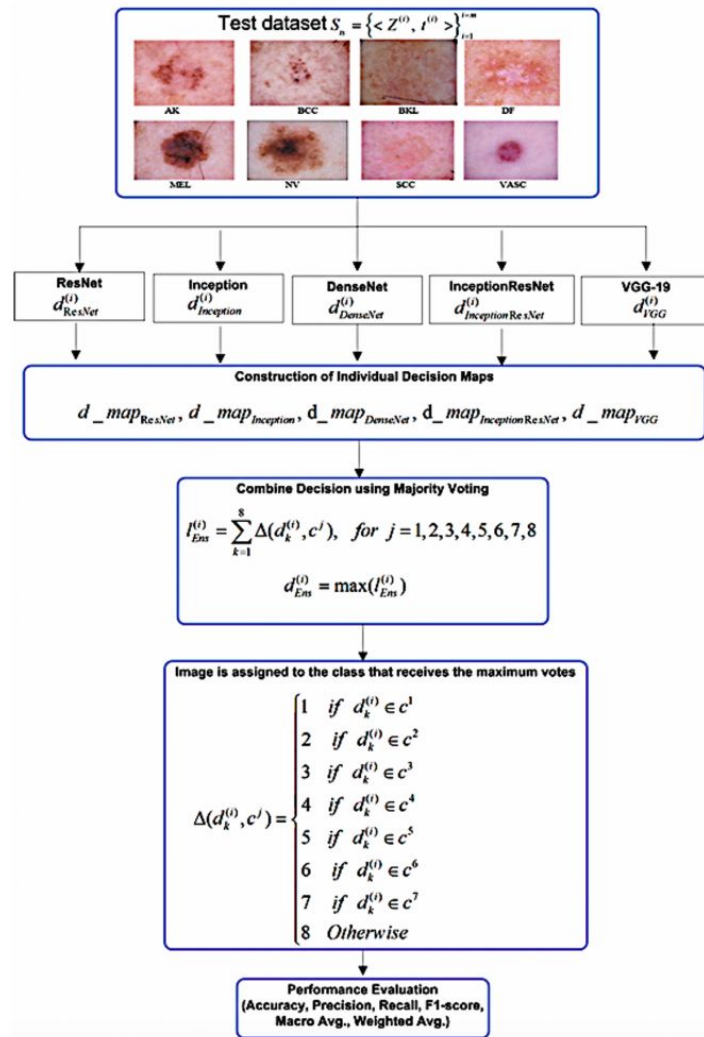
Early prediction of diabetic positives; achieves 93% accuracy using stacked ensemble learning, in two combined layers



Case study - skin cancer classification

Deep learning ensemble method for multi class skin cancer classification, plus majority and weighted majority voting to find 98% accuracy

■ Accuracy (%)





Your SciKit Learn Toolkit



Python calls for ensemble methods

- From sklearn.metrics import
- From sklearn.tree import DecisionTreeRegressor
- From sklearn.ensemble import
 - a. RandomForestRegressor
 - b. AdaBoostRegressor
 - c. GradientBoostingRegressor
 - d. StackingRegressor

Example

- Baseline LR model (includes cleaning)
- RF for feature selection
- LR model v2
- Ensemble method for better accuracy
 - Stacking regressor
- (bonus) competition





Competition time



Who?

CookieGo is a startup which distributes cookies

Problem?

They are losing money

Why ?

Too many different brands and recipes, leading to stock management issues



Competition time



What's the plan?

Optimise their supply chain - identify which cookies will sell better based on quality by predicting the quality of the cookies. Highest accuracy wins.

Data source

TRAIN - 5198 rows, 16 columns with cookie measurements and quality label

TEST - 779 rows, 16 columns with cookie measurements, NO quality label



Competition time



Assessment

VALIDATE - 779 rows, 16 columns with cookie measurements, and quality label.

Submit your predictions as an array via slack (*length must be 779*).

You can submit up to 5 prediction arrays per team - you will get your RMSE score back within minutes from the teachers

Lowest RMSE score through validation wins all the cookies!!



Tips!

- Split the work
- (github) Pull before you push
- Learn from each other
- Automate to save time

Kahoot quiz!

