

# DSC3108: Big Data Mining and Analytics

## Lecture 09 (BSCS\_3:1)

Topic: *Ensemble Learning Techniques*



Dr. Daphne Nyachaki Bitalo  
Department of Computing & Technology  
Faculty of Engineering, Design & Technology

# COURSE OVERVIEW



## Big Data

### Fundamentals and Mining

Architectures, storage  
of data

Data Mining Techniques

### Data Analytics

Descriptive analytics

Predictive analytics



# Lecture Objectives and Learning outcomes

The Objectives of this lecture are to:

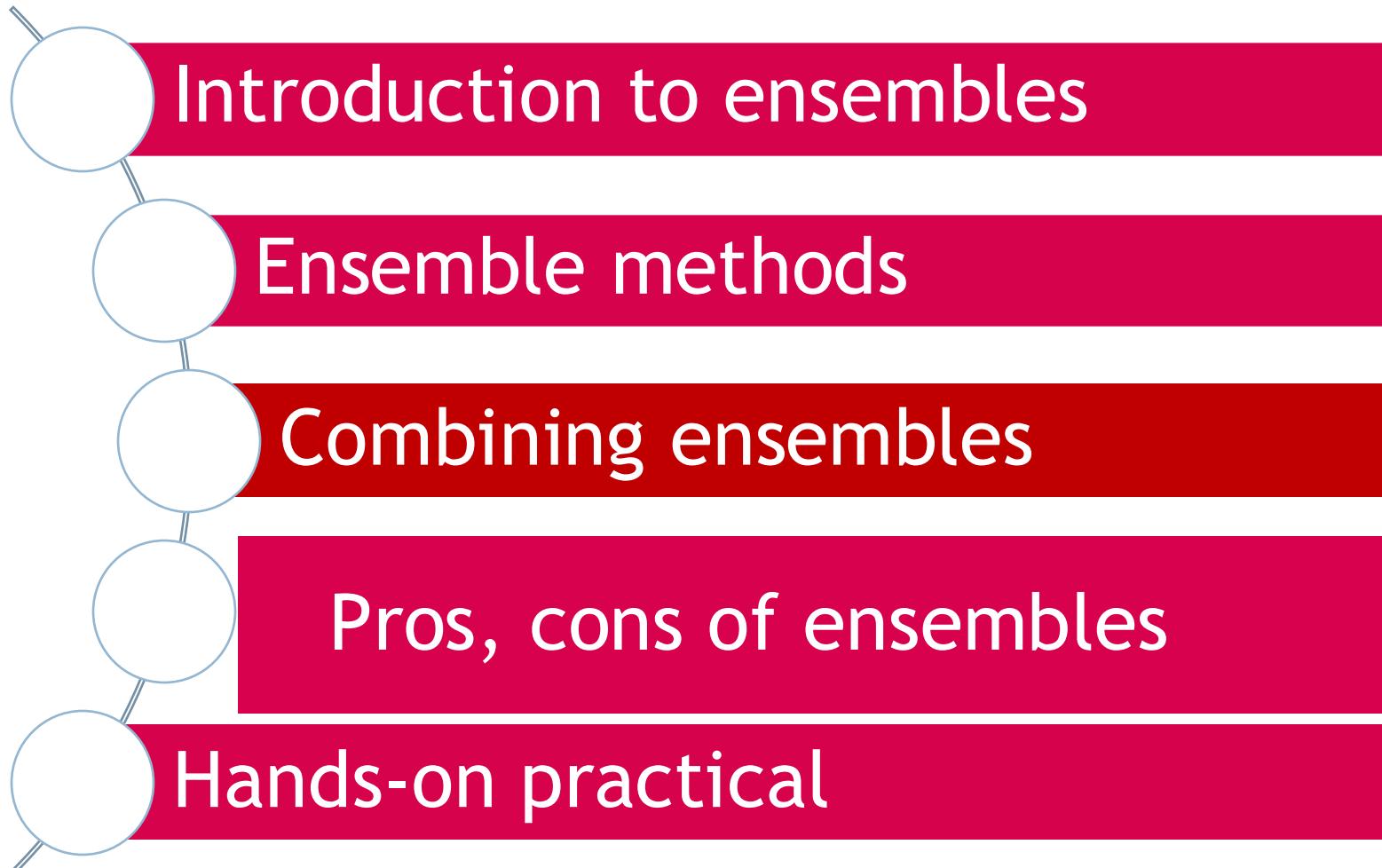
- ❑ Learn how to improve the predictive performance of an ML model
- ❑ Learn about the different ensemble methods and how they are trained and combined

By the end of this lecture, students should be able to:

- ❑ Get practical experience working with ensemble techniques to improve classification and regression predictive models



# Lecture Overview



# Why use ensembles?

- ML/ predictive models need constant updates as new data becomes available for accurate and reliable predictions. Therefore, ensemble techniques aid in boosting this accuracy.

Concept	Explanation	Analogy
Definition	Ensemble learning is a meta-algorithm that combines the predictions from multiple base models (called weak learners or base estimators) to produce a single, superior prediction.	Asking a panel of experts (diverse perspectives) for an opinion rather than relying on a single individual. The panel's average/majority decision is often better.
Goal	To improve predictive performance (accuracy, robustness) and reduce common issues like overfitting and high variance or high bias that a single model might exhibit.	
Weak Learner	A model (e.g., a simple Decision Tree) that performs slightly better than random chance. Ensembles make a "forest" from these "trees."	



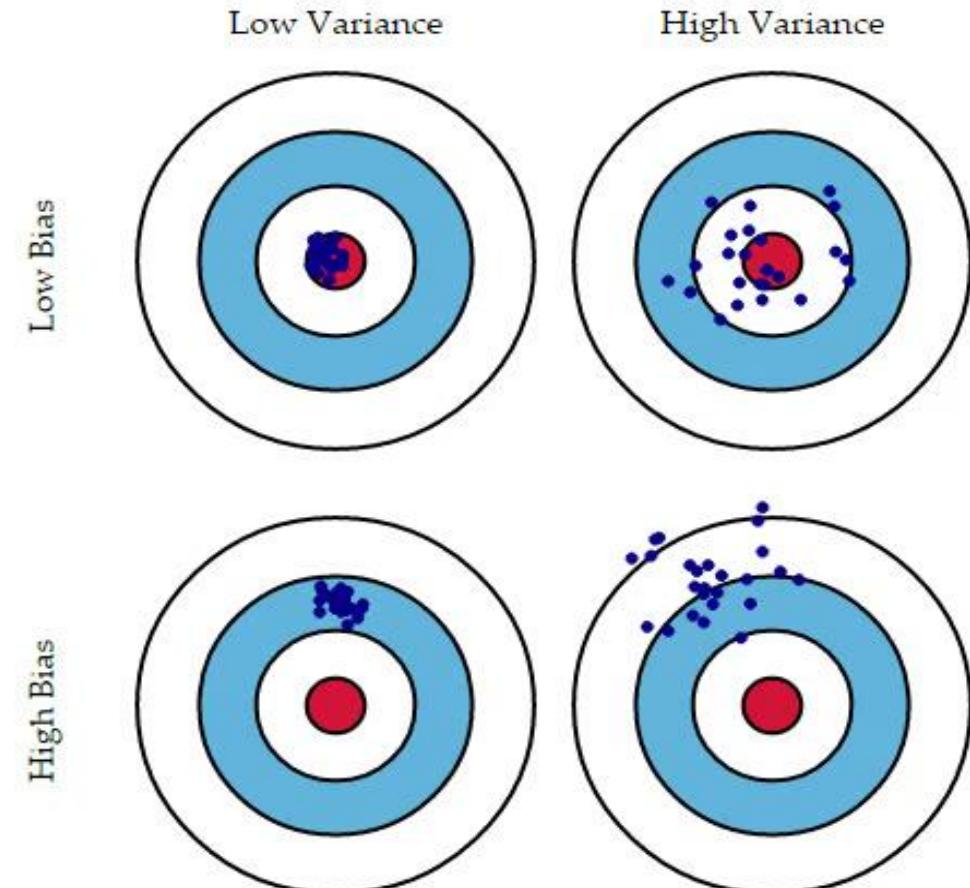
# Model errors and ensemble correction

- Errors emerging from ML models can be broken down into three components mathematically: **Bias + Variance + Irreducible error**
- Bias: quantifies how much, on an average, the predicted values are different from the actual value. A high bias error means we have an underperforming model that keeps missing essential trends.
- Variance: quantifies how the predictions made on the same observation differ. A high variance model will over-fit on your training population and perform poorly on any observation beyond training.



# Model errors and ensemble correction

- As model complexity increases, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model, and hence your model will start suffering from the high variance.



# Ensemble methods

- Ensemble methods are categorised based on how the base models are trained and combined.
- The methods are also categorised for their purpose (either to reduce variance or to reduce bias)
- Averaging/Parallel ensemble methods reduce variance
- Boosting/Sequential ensemble methods reduce bias



# Ensemble methods: Averaging techniques

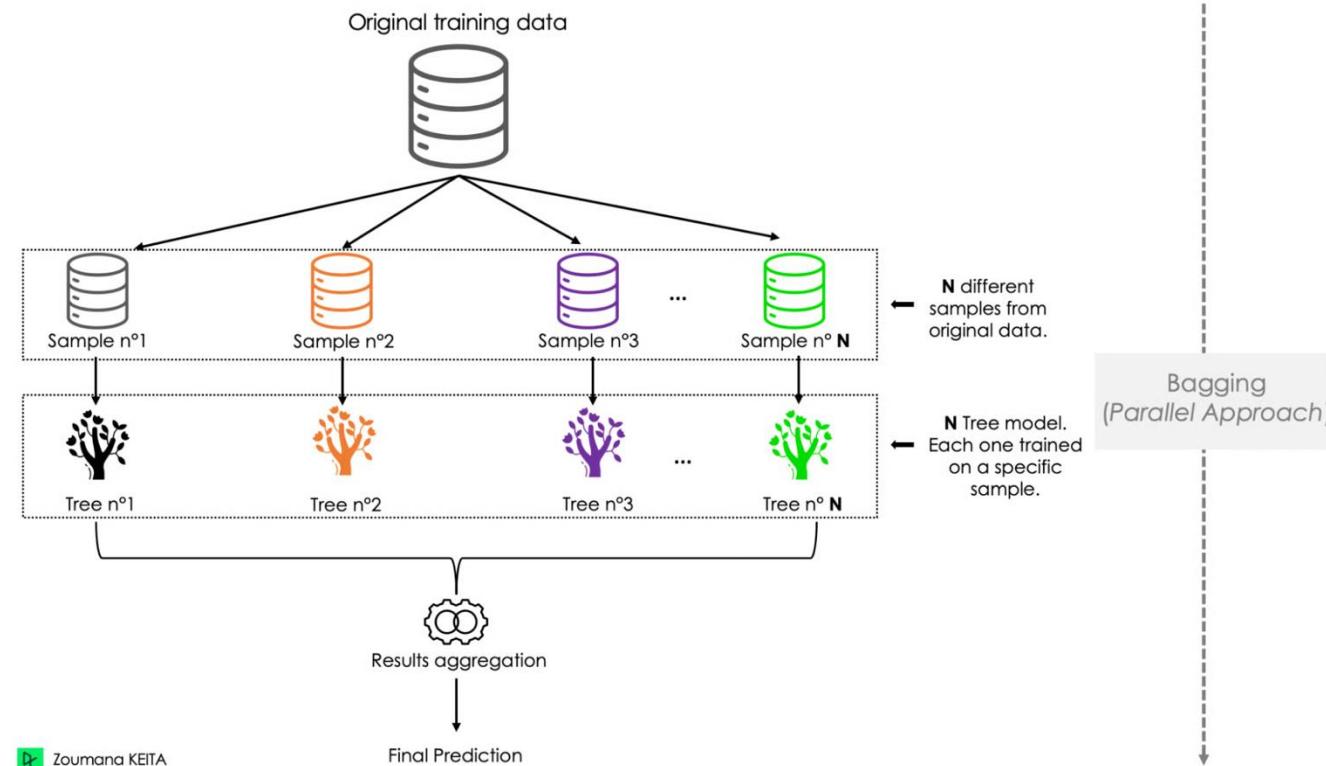
- These methods train the base models/learners independently and in parallel. They reduce variance and overfitting in the final model.

Method	Training Strategy	Combination	Key Benefit
1. Bagging (Bootstrap Aggregating)	Each weak learner is trained on a different, random bootstrap sample (sampling with replacement) of the original training data.	Averaging (for regression) or Majority Voting (for classification).	Reduces variance (overfitting) by training models on slightly different datasets, thereby decorrelating their errors.
2. Random Forest (RF)	An extension of Bagging. Each tree is built on a bootstrap sample, AND at each node split, only a random subset of features is considered.	Majority Voting/ Averaging.	Further reduces correlation between trees, making the ensemble much more robust than standard Bagging. High performance and feature importance readily available.



# Example bootstrapping/Bagging

- In bagging, a random sample of data from the training set is selected with replacement, which enables the duplication of sample instances in a set. Below are the main steps involved in bagging:
  - Generation of multiple bootstrap resamples.
  - Running an algorithm on each resample to make predictions.
  - Combining the predictions by taking the average of the predictions or taking the majority vote (for classification).



Zoumana KEITA



# Ensemble methods: Boosting techniques

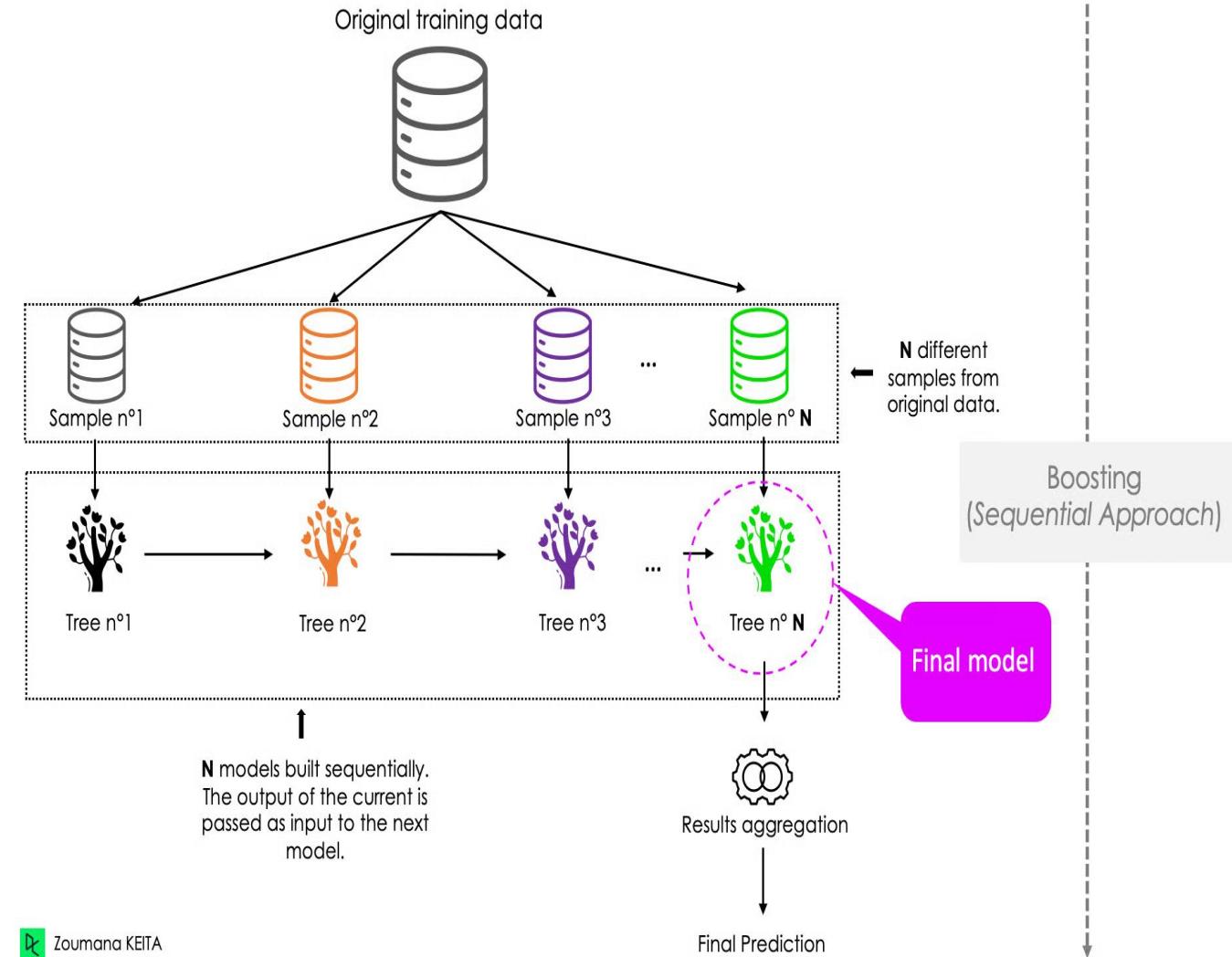
- These methods train the base models/learners sequentially (in series).  
 Each new learner attempts to correct the errors of the previous learners.  
 They reduce bias and model predictive accuracy.

Method	Training Strategy	Combination	Key Benefit
1. AdaBoost (Adaptive Boosting)	Each successive learner is trained on the same data, but the data points that were misclassified by the previous model are given higher weights.	Weighted Majority Voting, where better-performing models (lower error) are given higher weight in the final vote.	Focuses training iteratively on hard examples (reducing bias).
2. Gradient Boosting Machines (GBM)	Each successive learner is trained to predict the residual error (or gradient) of the previous ensemble of trees.	Predictions are summed up: $Y^{\wedge} = \sum_{k=1}^K f_k(X)$ .	Highly effective for complex problems. State-of-the-art performance in many tabular data tasks.
3. XGBoost, LightGBM, CatBoost	Highly optimized, scalable versions of GBM (e.g., parallel processing for tree construction, handling missing values).	Summation of weighted tree predictions.	Computational efficiency and superior predictive power on massive datasets.



# Example Boosting techniques

- The prediction of the current model is transferred to the next one. Each model iteratively focuses attention on the observations that are misclassified by its predecessors.



# Ensemble combination approaches

## Combination:

- This is the approach in which the predictions made by base models are aggregated by the ensemble technique:
  - Majority Voting (Classification): The final class is the one predicted by the majority of the individual models.
  - Averaging (Regression): The final prediction is the arithmetic mean of the individual models' predictions.
  - Stacking (Stacked Generalization): Stacking (Stacked Generalization): Uses heterogenous base learners. The predictions from the base learners are stacked together and are used as the input to train the meta learner to produce more robust predictions. The meta learner is then used to make final predictions



# Pros and Cons of employing ensembles

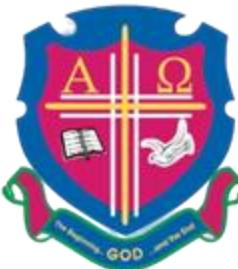
Aspect	Pros	Cons
Performance	Typically achieves higher accuracy than any single base model, especially Boosting methods (XGBoost).	Computationally expensive and time-consuming to train, especially for large datasets.
Robustness	Reduces overfitting (Bagging/RF) and handles noisy data better due to the collective decision-making.	Loss of interpretability. A single Decision Tree is easy to explain; a Random Forest of 500 trees is a black box (confusing).
Ease of Use	Modern implementations (e.g., in scikit-learn or R's caret) are easy to implement with few lines of code.	Hyperparameter complexity: More parameters to tune (e.g., number of estimators, learning rate, tree depth).



# Reading Assignment

- ❖ Sources of errors in predictive/ML models
- ❖ Advantages and Disadvantages of different ensemble techniques
- ❖ Advantages and Disadvantages of ensemble combination approaches
- ❖ Ensemble techniques for clustering mining/unsupervised Machine Learning
- ❖ Applications of various ensemble techniques





## Uganda Christian University

P.O. Box 4 Mukono, Uganda

Tel: 256-312-350800

<https://ucu.ac.ug/> Email: [info@ucu.ac.ug](mailto:info@ucu.ac.ug).



@ugandachristianuniversity



@UCUniversity



@UgandaChristianUniversity



## Department of Computing & Technology FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY

Tel: +256 (0) 312 350 863 | WhatsApp: +256 (0) 708 114 300



@ucucomputeng



@ucu\_ComputEng



<https://cse.ucu.ac.ug/> Email: [dct-info@ucu.ac.ug](mailto:dct-info@ucu.ac.ug)

A Complete Education for A Complete Person

P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: [info@ucu.ac.ug](mailto:info@ucu.ac.ug) Web: <https://ucu.ac.ug>  
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda