Water Quality Prediction using Statistical,Ensemble and Hybrid models

Vyshali S
Vikram V
Shriya B

SSN College of Engineering, Chennai

March 31, 2022

Problem Statement

- Objective- analyse the data and predict the water quality of the resources by building a model with better prediction ability.
- Proposed Hybrid system- combination of statistical and ensemble learning models.
- Statistical model- pre-processes the data set in order to resolve the shortcomings of real world data. Statistical techniques to be studied- Linear Regression, Classification, and Unsupervised Learning Algorithms.

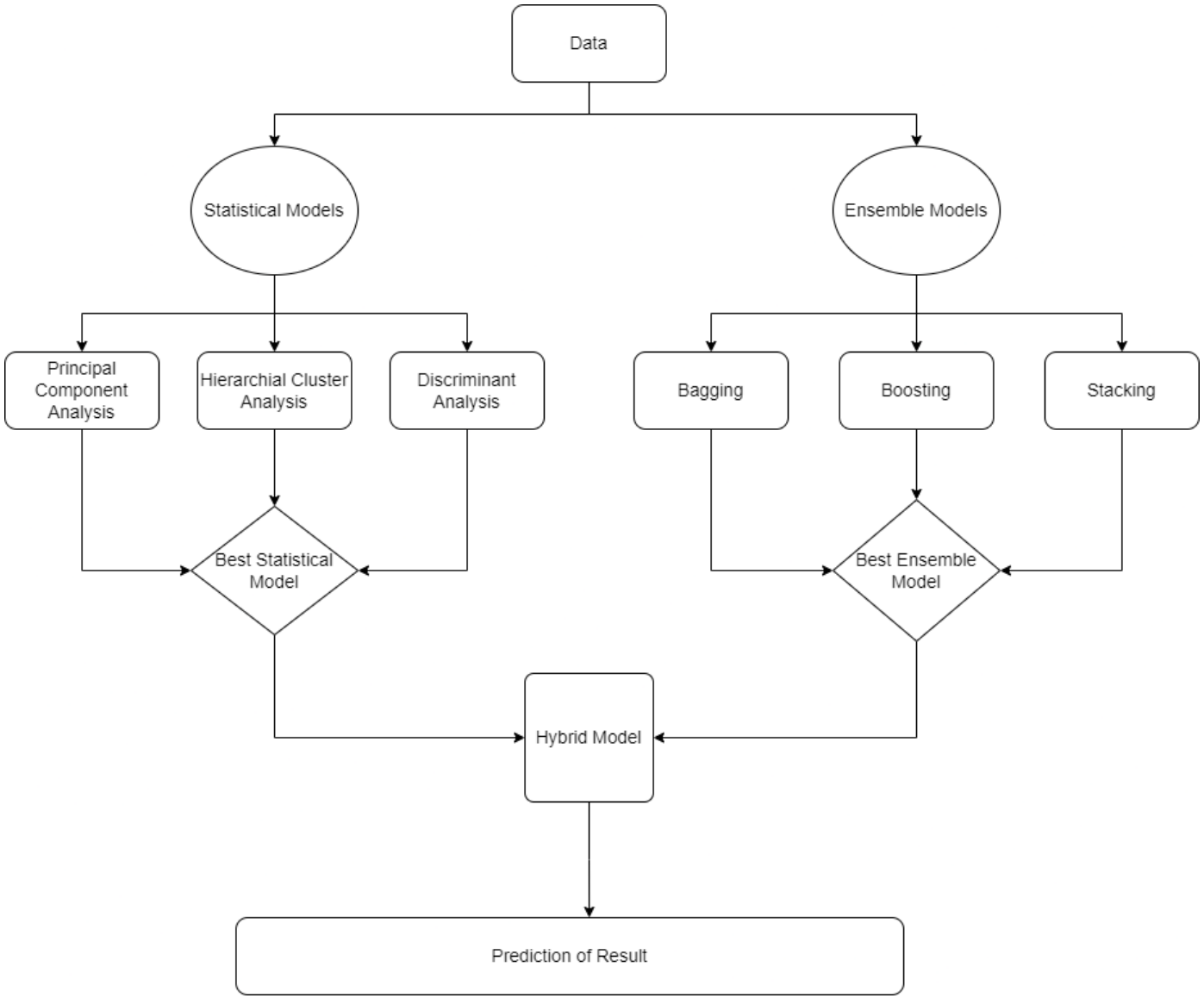
Follow up of previous review

- Initially proposed to use and compare base models in Machine Learning and Statistics, then combine the best model from both to propose a hybrid model.
- Received a feedback to try out ensembling the proposed models
- Advised to bring out better predictability for the proposed hybrid model, to provide a positive thesis

Overview

- Ensemble- create multiple models and combine them to produce better results.
- Ensemble methods used: Bagging, Boosting and Stacking.
- Base models: Decision tree classifier, random forest, XGboost, K neighbours and logistic regression.

Proposed system Architecture



Models and techniques used

Base models :

- Decision tree :- tree like model ; branches-decision rules ; leaf nodes-outcomes
- Random Forest: combines outcomes of several small decision trees
- XGBoost :- parallel tree boosting ; uses loss function to identify shortcomings of weak learners
- K-neighbours - likelihood of belonging to one group based on proximity
- Logistic regression :- predict a dependent categorical target variable ; only applicable to binary classification

Models and techniques used

Ensemble Techniques :

- Bagging :- bootstrap aggregating ; decrease the variance in the prediction
- Boosting :- combines a set of weak learners into a strong learner to minimize training errors
- Stacking :- ensemble multiple classifications or regression models ; produce one optimal predictive model ; better performance than the base learners taken alone

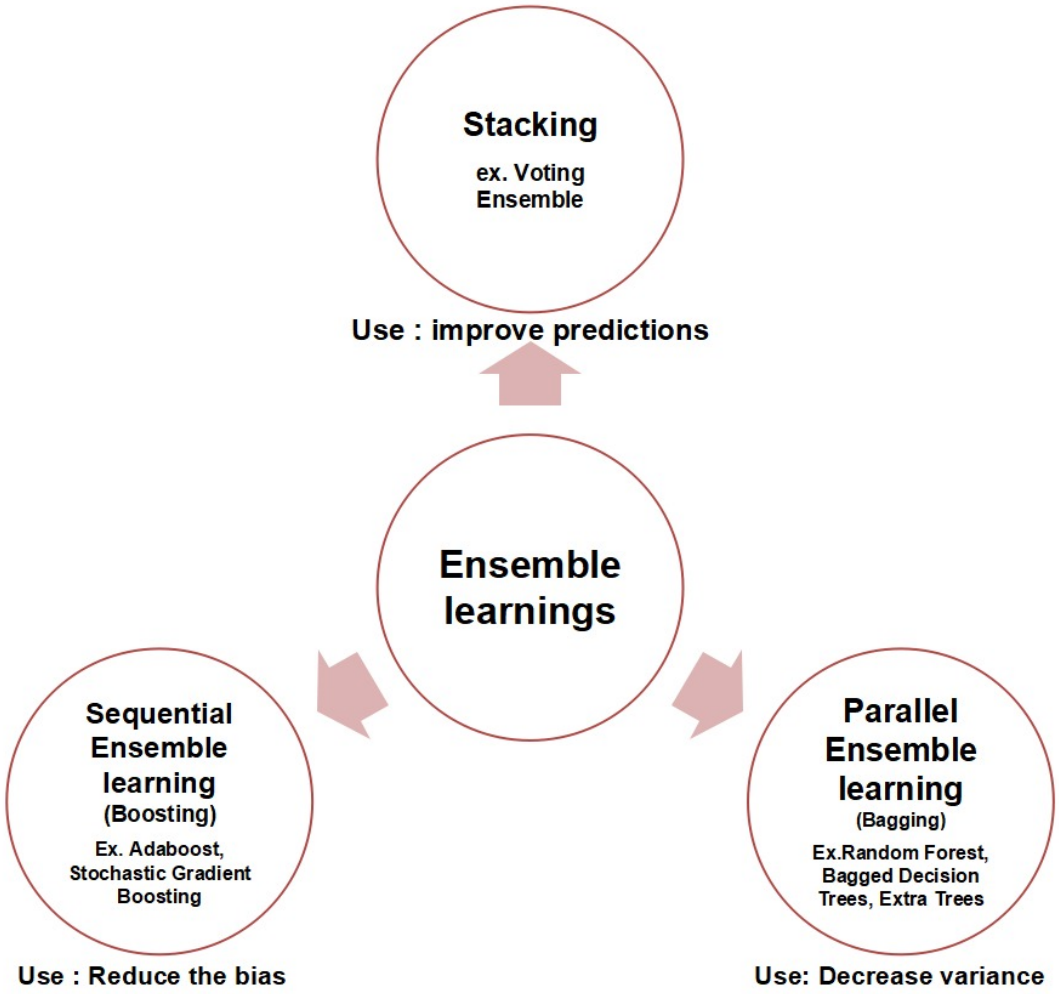
Milestones Achieved

- Selected ensemble models to explore
- Selected base models for Bagging, Boosting and Stacking
- Built and trained the models with both binary and multi class data and found the best models
- Binary class data - Adaboost[decision tree classifier] boosting model
- Multi class data - Decision tree classifier based bagging model

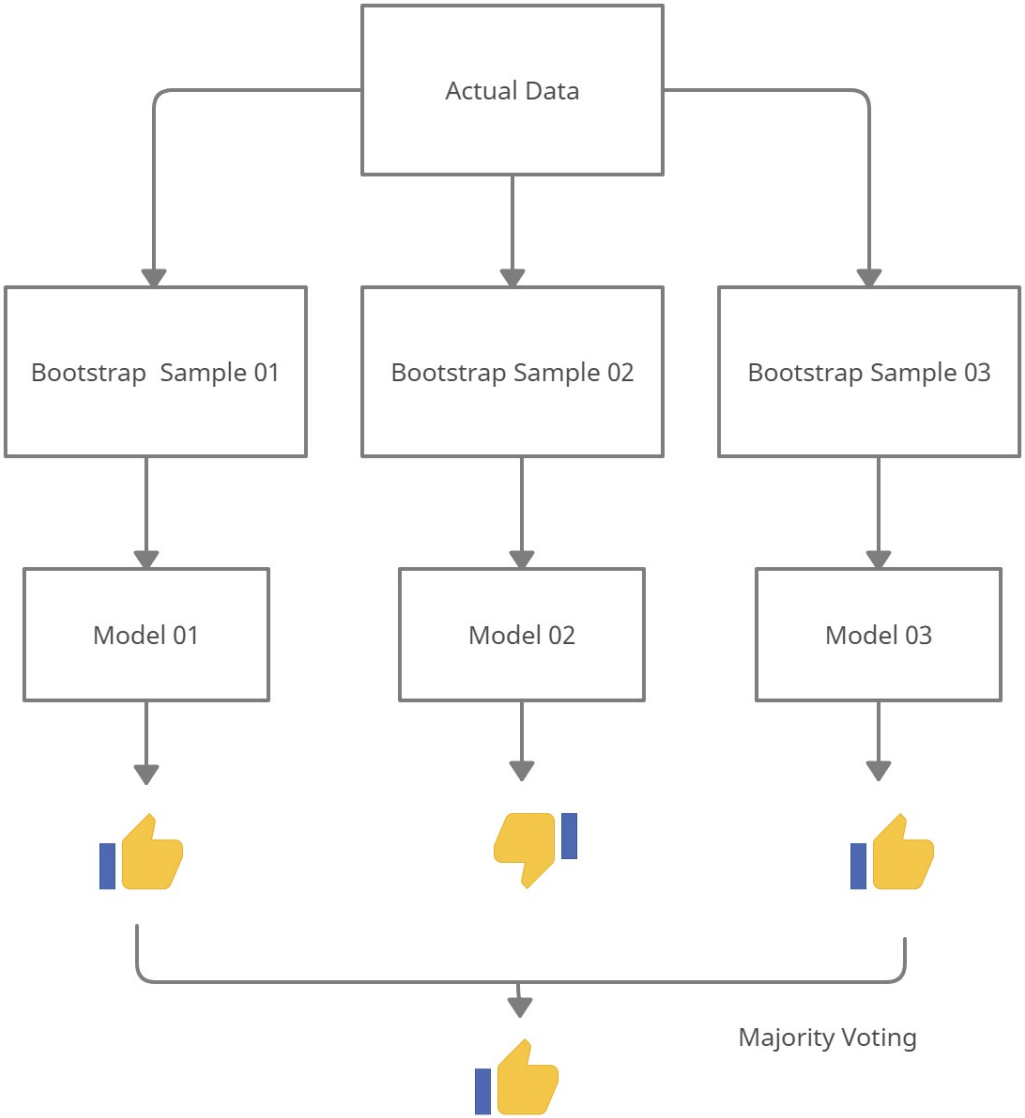
Ensemble Learning

- Ensemble learning is a machine learning paradigm where multiple models (often called "weak learners") are trained to solve the same problem and combined to get more accurate and/or robust models.
- A low bias and a low variance, are the two most fundamental features expected for a model. This is the called bias-variance trade-off.
- The idea of ensemble methods is to try reducing bias and/or variance of such weak learners by combining several of them together in order to create a strong learner that achieves better performances.

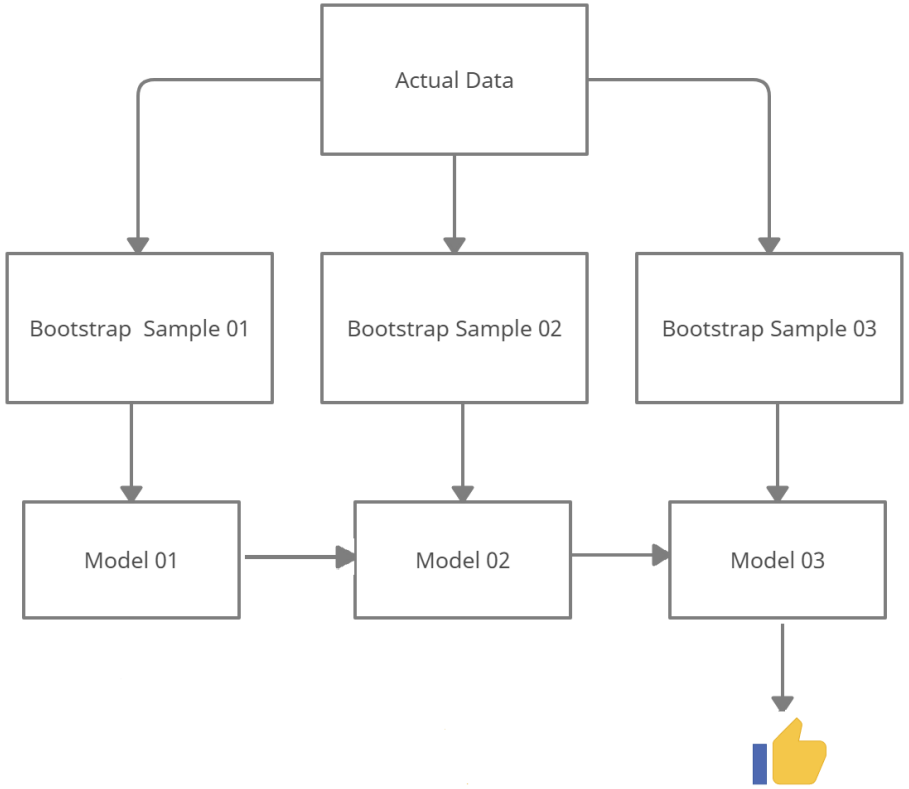
Ensemble Learning



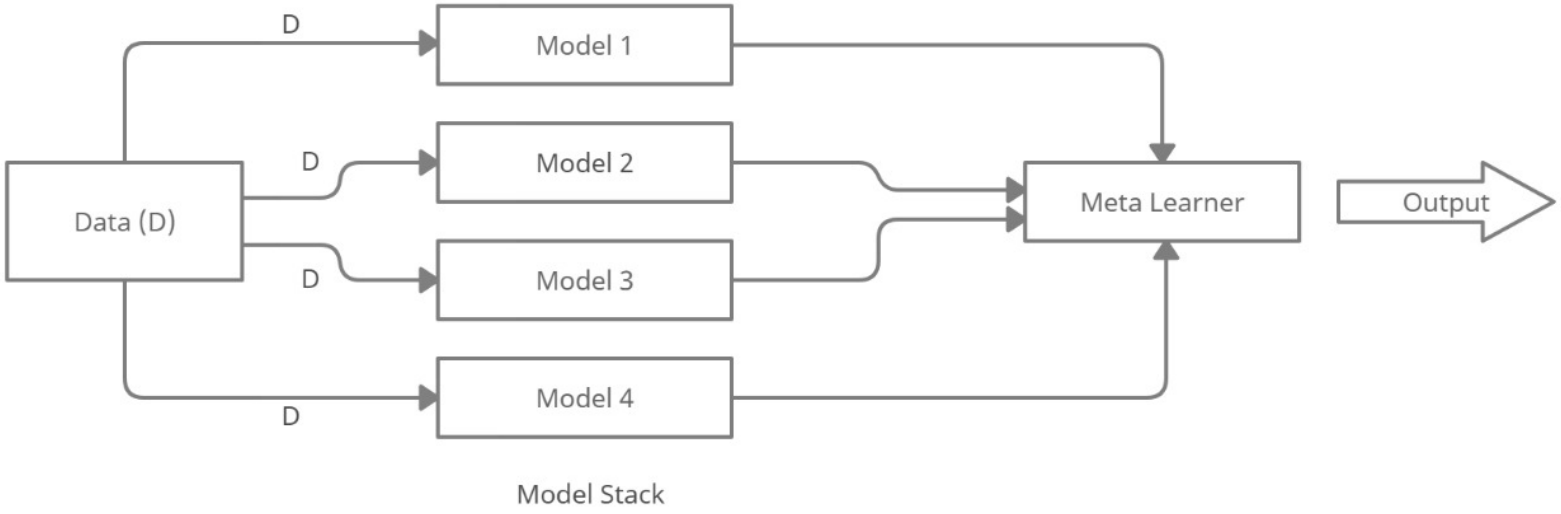
Bagging


Actual Data
Bootstrap Sample 01
Bootstrap Sample 02
Bootstrap Sample 03
Model 01
Model 02
Model 03
Majority Voting

Boosting


Actual Data
Bootstrap Sample 01
Bootstrap Sample 02
Bootstrap Sample 03
Model 01
Model 02
Model 03

Stacking


D
Model 1
D
Model 2
Data (D)
Meta Learner
Output
D
Model 3
D
Model 4
Model Stack

Accuracy of base models for binary and multi class data

	ACCURACY	
CLASSIFIER	BINARY CLASS DATA	MULTI CLASS DATA
DECISION TREE	0.9992	1
RANDOM FOREST	0.9992	0.999704142
XG BOOST	0.926	0.9998027613
K NEIGHBOURS	0.9998	0.9271130177

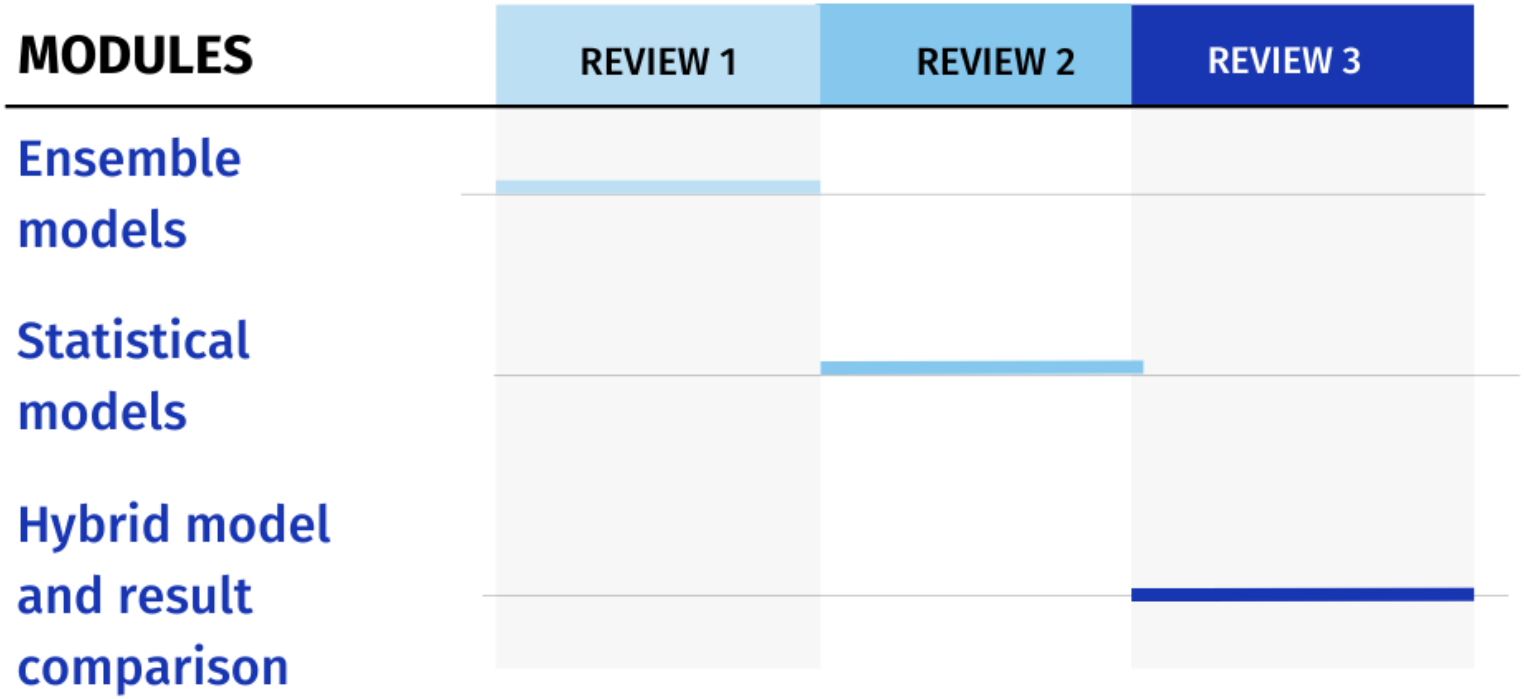
Outcomes of Binary class data

	BINARY CLASS DATA					
	ACCURACY	PRECISION	RECALL	F1 SCORE	CROSS VALIDATION	MODEL TRAINING TIME
BAGGING	0.9988888889	1	0.9915966387	0.9957805907	0.9996	0.9419000149
ADABOOST	1	1	1	1	0.9998	0.05208396912
STACKING	0.9988888889	1	0.9910714286	0.9955156951	0.9994	3.159189701

Outcomes for Multi class data

	MULTI CLASS DATA					
	ACCURACY	PRECISION	RECALL	F1 SCORE	CROSS VALIDATION	MODEL TRAINING TIME
BAGGING	1	1	1	1	1	0.7528047562
ADABOOST	1	1	1	1	0.9998027613	0.09389972687
STACKING	1	1	1	1	1	8.813961267

Project Timeline



Proof of weekly meeting with Supervisor

Sl No.	Date	Work done during the week	Supervisor signature
1	07/03/2022	Exploring, data and different base models for bagging ensemble technique	
2	14/03/2022	Exploring adaboost ~~and~~ and XGBoost for boosting ensemble technique	
3	21/03/2022	Exploring Decision tree, random forest, XG boost, K-N-neighbours & Logistic Regression for Stacking ensemble technique	
4	28/03/2022	Report and Presentation drafting for Review 1	

Thank You