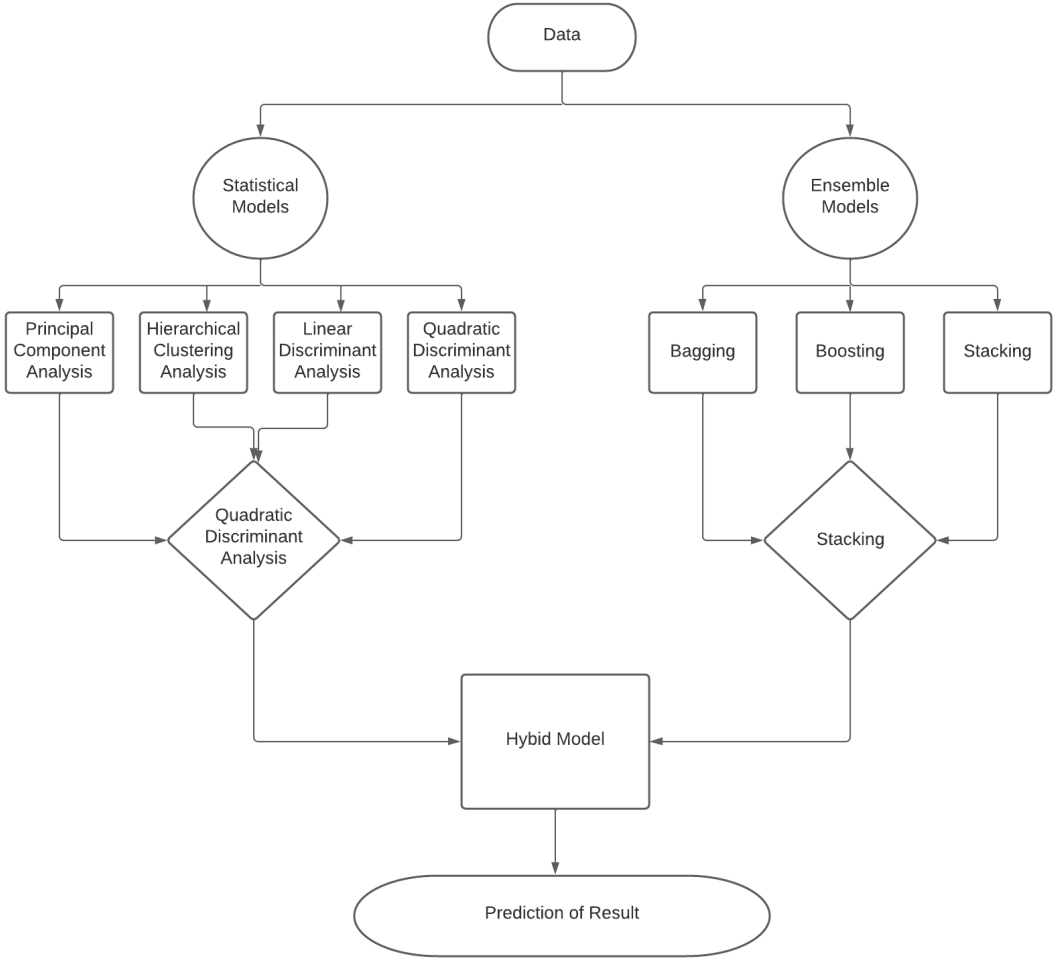
Water Quality Prediction using Statistical,Ensemble and Hybrid models

Vyshali S
Vikram V
Shriya B

SSN College of Engineering, Chennai

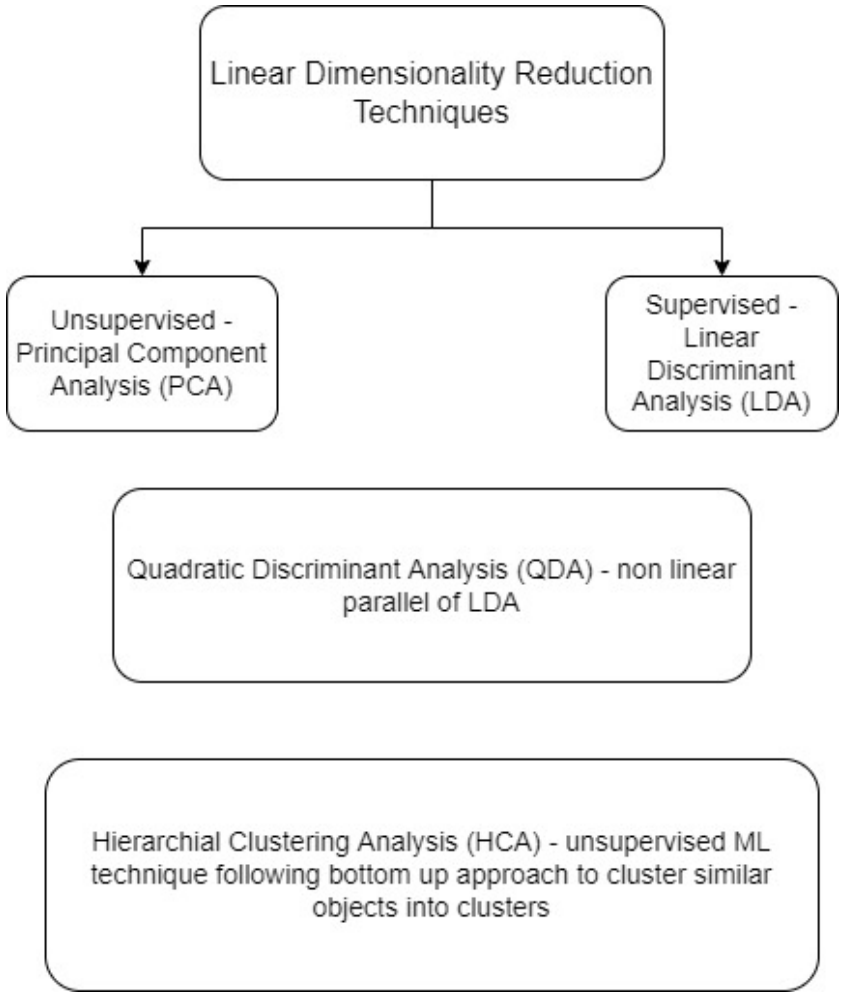
March 31, 2022

System Architecture



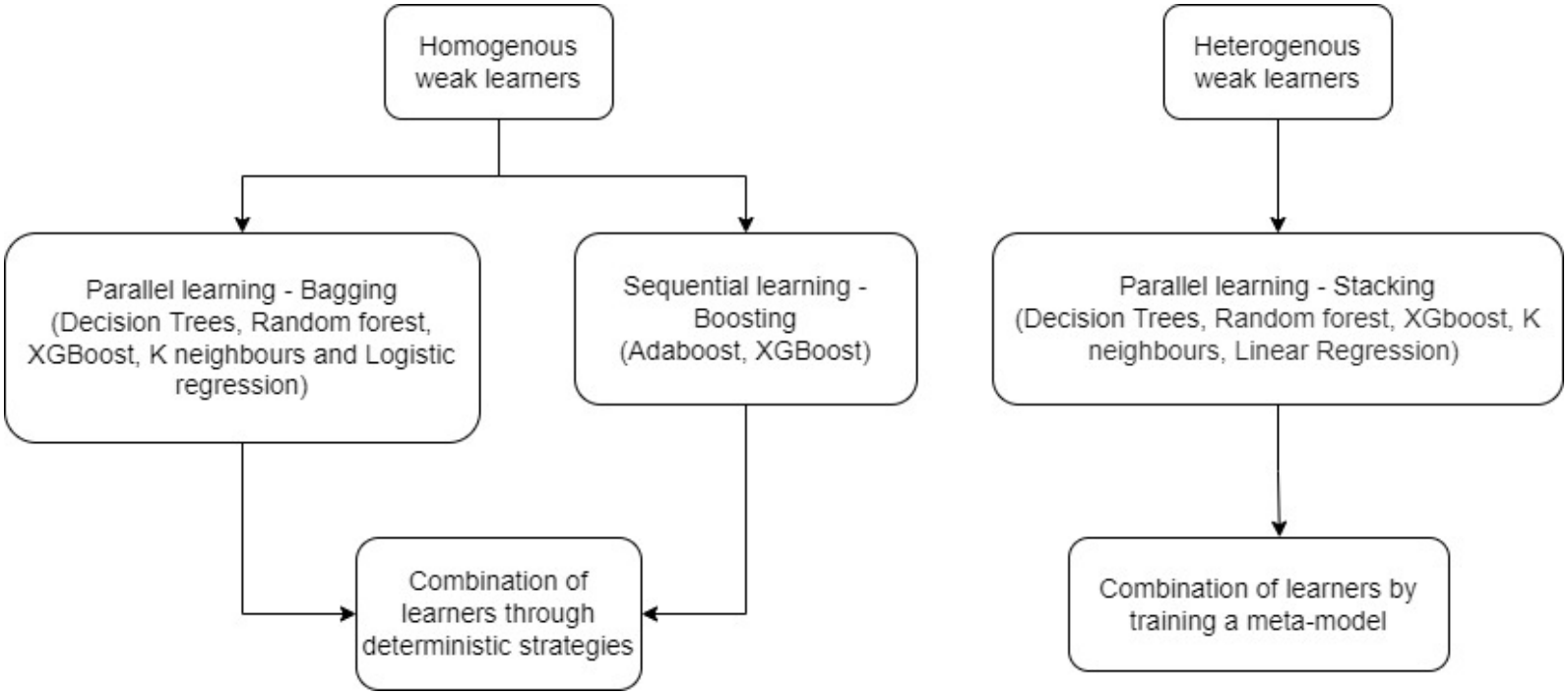
System Design - Statistical Models

Analyse the data set in order to resolve the shortcomings of real world data.



System Design - Ensemble Models

Combine several base models in order to produce one optimal predictive model.



Datasets used

Binary Class Korattur Lake Dataset –> 5000 x 10

	pH	TDS	Turbidity	Phospate	Nitrate	Iron	COD(mg/L)	Chlorine	Sodium	Class
0	7.6	877	3.59	0.026136	8	0.378500	397	4	8	1
1	7.6	729	1.75	0.020622	6	0.333759	397	3	9	1
2	7.5	622	3.44	0.004071	7	0.382368	394	4	16	0
3	7.6	864	2.80	0.022071	7	0.313915	403	2	5	0
4	7.6	656	1.81	0.004031	7	0.333226	421	4	11	0

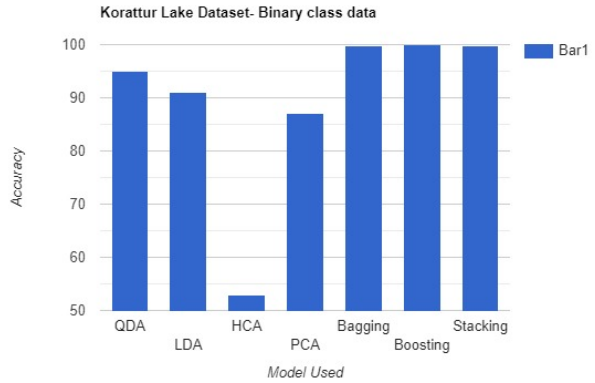
Binary Class Kaggle Dataset : –> 7999 x 21

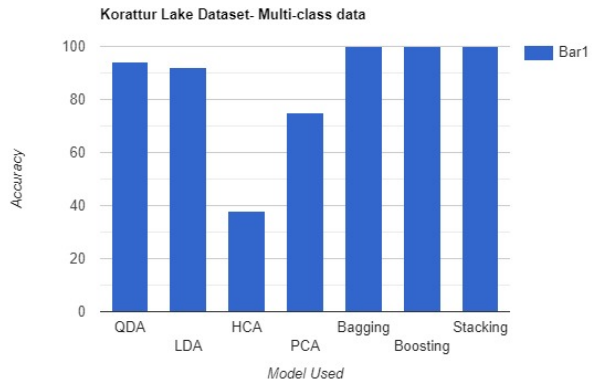
luminium	Ammonia	Arsenic	Barium	Cadmium	Cloramine	Chromium	Copper	Flouride	Bacteria	...	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	Class
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.20	...	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.90	0.65	...	0.100	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	...	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	...	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57	0.61	0.13	...	0.117	6.74	1.11	0.003	16.90	2.41	0.02	0.06	0.02	1

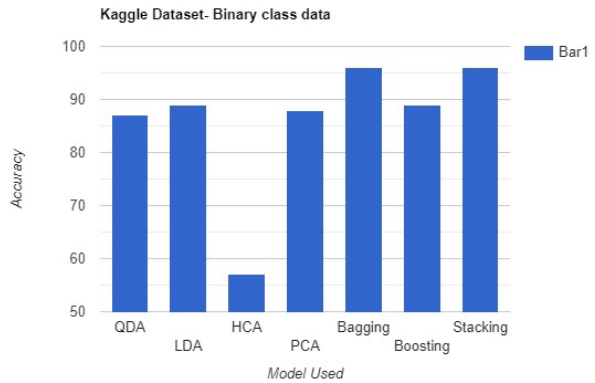
Multi Class Korattur Lake Dataset : –> 10139 x 10

	pH	TDS	Turbidity	Phospate	Nitrate	Iron	COD(mg/L)	Chlorine	Sodium	Class
0	7.6	973	0.16	0.012967	8	0.328568	422	2	12	1
1	7.6	975	3.17	0.016066	7	0.332097	427	4	2	1
2	7.5	755	2.53	0.019433	7	0.338934	398	2	16	2
3	7.6	686	4.15	0.018559	7	0.303969	409	4	12	1
4	7.6	858	3.90	0.002456	8	0.383476	390	4	17	0

Results


Korattur Lake Dataset- Binary class data
Bar1
100
90
80
70
60
50
Accuracy
QDA
LDA
HCA
PCA
Bagging
Boosting
Stacking
Model Used


Korattur Lake Dataset- Multi-class data
Bar1
100
80
60
40
20
0
Accuracy
QDA
LDA
HCA
PCA
Bagging
Boosting
Stacking
Model Used


Kaggle Dataset- Binary class data
Bar1
100
90
80
70
60
50
Accuracy
QDA
LDA
HCA
PCA
Bagging
Boosting
Stacking
Model Used

Next Review

- Building a hybrid model using the best statistical and ensemble models, that is Quadratic Discriminant Analysis(QDA) and Stacking.
- Try Implementing time series forecasting models such as Moving-Average, Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA).

Thank You