Water Quality Analysis and Prediction using Statistical,Ensemble and Hybrid Models

Dr. D. Venkata Vara Prasad S. Vyshali V. Vikram B. Shriya

Department of CSE, Sri Sivasubramaniya Nadar college of Engineering, Chennai, India

Highlights of Proposed Model

To develop an new efficient model that

- Predicts the water quality class with very good accuracy.
- Performs better than models proposed by previous papers.
- Performs better than the implemented statistical and ensemble models.
- Is an combination of the best statistical and ensemble models

Dataset Description

- One binary, 3 and 5 class datasets sourced from Korattur Lake, taken over a period of 10 years
- A binary class dataset sourced from Kaggle
- From the new augmented dataset 70% of total is taken for training, 20% for validation and 10% for testing

Proposed Architecture Model

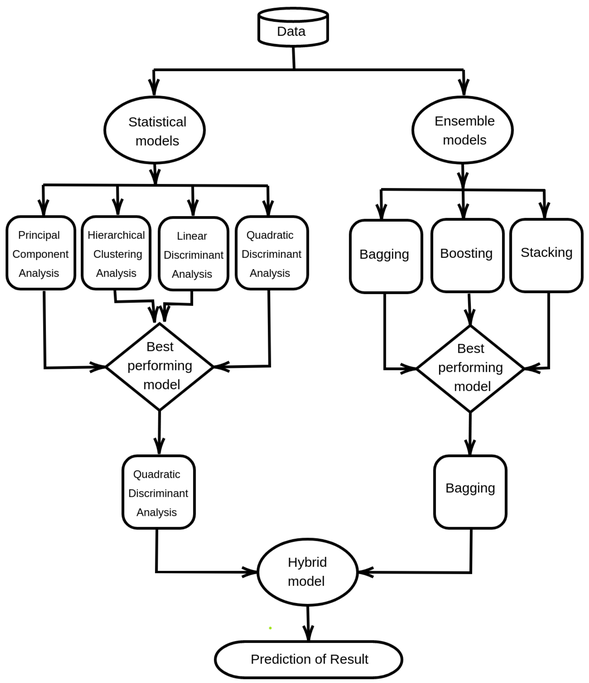

Figure 1. Architecture of Proposed Model

Functional Modules

- Data Pre-processing
  - Data cleansing
- Implementing Statistical Models
  - Principal Component Analysis (PCA)
  - Hierarchical Clustering Analysis (HCA)
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
- Implementing Ensemble Models
  - Bagging
  - Boosting
  - Stacking
- Building the Hybrid Model
  - Choosing best statistical model
  - Choosing best ensemble model
  - Ensembling both to build hybrid model

Implementing Statistical Models

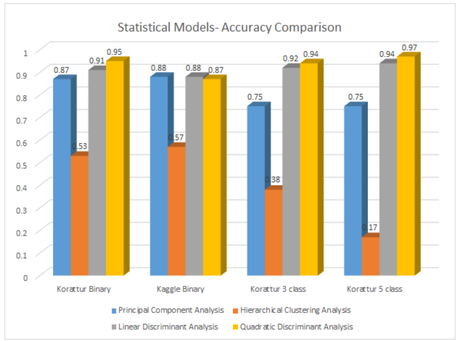

Figure 2. Quadratic Discriminant Analysis works best among all Statistical models

Implementing Ensemble Models

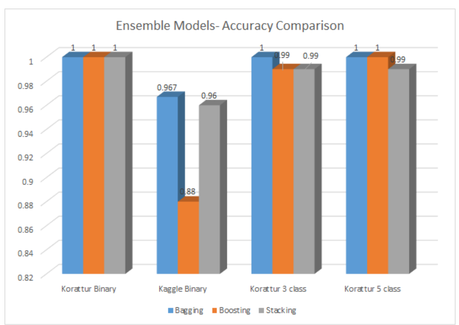

Figure 3. Bagging works best among all Ensemble models

Hybrid Model Implementation

Combining models using Voting Classifier Voting Classifier

- QDA(Quadratic Discriminant Analysis) was found to be the best performing Statistical model
- Bagging using Decision tree classifier was the most efficient Ensemble Model
- Both QDA and Bagging are combined using the Voting Classifier.
- The voting classifier is an ensemble classifier algorithm which trains various base models / estimators.
- The prediction is then done based on the combination of the findings of each base estimator.
- The aggregating criteria can be combined decision of voting for each estimator output. It is of 2 types - hard and soft voting.
- Hard Voting is when voting is evaluated based on predicted output class.In the proposed Hybrid model, hard voting is used.

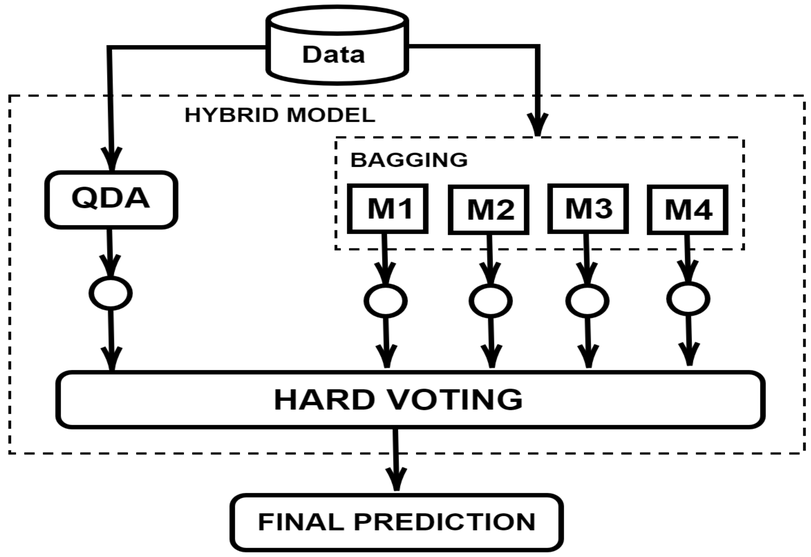

Figure 4. Hybrid Model Architecture

Performance Analysis and Inferences

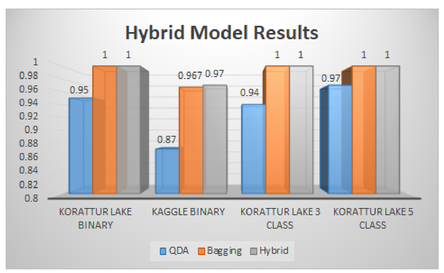

Figure 5. Comparison of QDA, Bagging and Hybrid Model

Inferences

- The results of the Hybrid model was compared with both the best performing Statistical model - QDA and Ensemble models - Bagging for all the datasets.
- Hybrid model performs the best, irrespective of the dataset.