WATER QUALITY PREDICTION USING STATISTICAL, ENSEMBLE AND HYBRID MODELS

A PROJECT REPORT

Submitted By

VYSHALI S. 185001202

VIKRAM V. 185001194

SHRIYA B. 185001149

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

Rajiv Gandhi Salai (OMR), Kalavakkam - 603110

May 2022

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this project report titled **"WATER QUALITY PREDICTION USING STATISTICAL, ENSEMBLE AND HYBRID MODELS"** is the *bonafide* work of "**VYSHALI. S (185001202)**, **VIKRAM. V (185001202)**, and **SHRIYA. B (185001149)**" who carried out the project work under my supervision.

Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

DR. T.T. MIRNALINEE
HEAD OF THE DEPARTMENT
Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

DR. D. VENKATA VARA PRASAD
SUPERVISOR
Associate Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

Place:
Date:

Submitted for the examination held on. . . . . .

Internal Examiner **External Examiner**

ACKNOWLEDGEMENTS

I would like to thank and deep sense of gratitude to my guide **DR. D. VENKATA VARA PRASAD**, Professor, Department of Computer Science and Engineering, for his valuable advice and suggestions as well as his continued guidance, patience and support that helped me to shape and refine my work.

My sincere thanks to **Dr. T.T. MIRNALINEE**, Professor and Head of the Department of Computer Science and Engineering, for her words of advice and encouragement and I would like to thank our project Coordinator **Dr.B. BHARATHI**, Associate Professor, Department of Computer Science and Engineering for her valuable suggestions throughout this project.

I express my deep respect to the founder **Dr. SHIV NADAR**, Chairman, SSN Institutions. I also express my appreciation to our **Dr. V. E. ANNAMALAI**, Principal, for all the help he has rendered during this course of study.

I would like to extend my sincere thanks to all the teaching and non-teaching staffs of our department who have contributed directly and indirectly during the course of my project work. Finally, I would like to thank my parents and friends for their patience, cooperation and moral support throughout my life.

VYSHALI S. **VIKRAM V.** **SHRIYA B.**

ABSTRACT

Water is an essential elixir for several living organisms to function and survive. But it gets contaminated through several sources such as industrial wastes, oil spills, marine dumping, etc. With a growing population, availability of good quality water is of grave importance. This has become the motivation to probe into analysis of water quality from the outcomes of Statistical and Ensemble methods and to find the best working models from both methods. Research has been done to predict water quality analysis using standalone statistical and ensemble models. So, this research focuses on obtaining the best Statistical and Ensemble model separately among the models tried. The statistical models implemented for comparison are Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA). The Ensemble models used are Bagging, Boosting and Stacking. The models are then combined to build a Hybrid model to observe the comparisons between the three. The performance metrics used are Confusion Matrix, Accuracy, Precision, Recall, F1-score and ROC curve. While comparing the models, it is observed that Hybrid model produces the most accurate results, hence proving that the combination of Statistical and Ensemble model is efficient.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BPNN- Backpropagation Neural Network
ANFIS- adaptive neuro-fuzzy inference system or adaptive network-based fuzzy inference system
SVR- support vector regression
MLR- Multiple Linear Regression
WQI- water quality index
LSTM- Long short-term memory
ELM- Extreme Learning Machines
GRNN-Generalized regression neural network
SAE-sparse autoencoder
BPNN- Back Propagation Neural Network
NSE- national stock exchange
RMSE-root-mean-square error
MAE-mean absolute error
MSE- mean square error
SMOTE-Synthetic Minority Oversampling Technique
ARIMA-autoregressive integrated moving average
ANN- artificial neural network
CART- Classification And Regression Tree
MLR- multiple linear regression
COD- Chemical Oxygen Demand
BOD- Biochemical Oxygen Demand
SWAT-Soil and Water Assessment Tool
TPOT- Tree-Based Pipeline Optimization Tool
SHAP- SHapley Additive exPlanations
MLP- MultiLayer Perceptron
AUC- Area Under the Curve
ROC-Receiver Operating Characteristic curve

CHAPTER 1

INTRODUCTION

Water is a fundamental resource for life to survive. India contributes to 4 percent of the world's total freshwater resources. Out of the 4 percent Tamil Nadu contributes only 2.5 percentage. The ratio of population versus the freshwater availability is concerning. At this point, water contamination is not to be taken lightly. With so many sources of water contamination in hand, it is necessary to analyse water quality as efficiently and accurately as possible. Water quality of resources has been analysed by several researchers before. Several models using Machine Learning, Deep Learning, Auto-ML(Machine Learning) and Auto-DL(Deep Learning) have been proposed before for the same. The objective of this research is to analyse the data and predict the water quality of the resources by building a model with better prediction ability. In order to reduce dimensionality and noise from a real world data set, statistical models are applied. Statistical models are mathematical techniques and statistical assumptions that generate sample data and make predictions. It usually is a collection of probability distributions on a set of all possible outcomes of an experiment. The models used in this research are Principal Component Analysis, Hierarchical Clustering Analysis, Quadratic Discriminant Analysis and Linear Discriminant Analysis. Principal Component Analysis is a dimensionality reduction technique that reduces the dimension of a large data set while preserving the important information. Hierarchical Clustering Analysis technique clusters points that are more closely related to each other. Linear Discriminant Analysis used to find a linear combination of attributes that separates several classes of objects or events.

Quadratic Discriminant Analysis is another version of LDA in which a separate covariance matrix is assumed for every class of outcomes. PCA, HCA QDA and LDA are compared and the best model is used for the data pre-processing.

Once the pre-processing of the data is done using the best performing statistical model, it is fed into the water quality prediction model. In order to determine the water quality, ensemble learning methods are used. Ensemble methods create multiple models and combine them to produce better results.They usually produce solutions that are higher in accuracy than a single model. Bagging, Boosting and Stacking are the methods used in this research. These methods are implemented using decision tree classifier, random forest, XGboost, K neighbours and logistic regression as base models. Bagging is generally used to reduce variance in a data set that contains noise. Boosting is a technique that creates a strong classifier from several weak classifiers. Stacking is a method that combines predictions of multiple models to create an optimal model. All three methods are fed with both binary and multi class data. Finally, Bagging, Boosting and Stacking models are compared and the best model is selected to use for further stages.

The statistical and ensemble learning models are combined to form a hybrid model. The outcome of the hybrid model is compared with ensemble learning and statistics based systems in order to analyse the performance of the hybrid model.

1.1 Problem statement

Water quality analysis has always been a popular topic of study. A combination of statistical and Ensemble models will be used in the proposed system. In most cases, real-world data is incomplete, inconsistent, and noisy. The statistical model pre-processes the data set to eliminate the flaws that exist in real-world data. The Ensemble model then forecasts the water quality of the sample. The Ensemble model predicts the class feature based on these features (water quality). In order to assess the hybrid model's performance, the hybrid model's output is compared to ensemble learning and statistics-based models.

1.2 MOTIVATION

Water bodies have played a critical role in personal and industrial uses. Polluted water bodies can lead to series of infection and sometimes could even lead to death. Several industries like fabrication, food, agriculture, automotive and so on, require water quality prediction tools with great prediction ability in order to manufacture products with good quality. This research aims to analyse the water quality to make sure that good quality water is available to drink and use for all purposes. Traditional methods like lab analysis have several shortcomings and could consume a lot of time. This research adopts innovative techniques to solve such shortcomings. This research helps to get an insight on how combining two best working models can produce better results than standalone models.

1.3 ORGANIZATION OF CHAPTERS

Chapter 1 gave an introduction about the need of water quality analysis in various aspects and the amount of contamination water is subjected to.

Chapter 2 will be a survey of all the literature readings done for the research in the areas of Ensemble Learning and Statistical techniques.

Chapter 3 will provide a detailed insight on the Data sets used for the research along with the architecture of the Hybrid model proposed.

Chapter 4 contains a thorough study with visual representations about the various Ensemble and Statistical techniques implemented in this research.

Chapter 5 presents the results after using various performance metrics and compares the results of the standalone models and Hybrid model produced using several visual representations.

Chapter 6 concludes the research and mentions future work that is to be done.

CHAPTER 2

LITERATURE SURVEY

Sani Isah Abba et al., [1] implemented data intelligence models along with ensemble methods in order to predict water quality index. The algorithms used were BPNN, ANFIS, SVR and MLR to predict the WQI of three stations- Nizamuddin, Palla, and Udi along the river Yamuna. Determination coefficient, root mean square error, and correlation coefficient were used for assessing the results. The results indicated that NNE was the best approach for the prediction of water quality index. It was also concluded that the above method can be proved effective for rivers across different regions. Nevertheless, it was agreed that the model needed further more optimization by coupling with advanced optimization algorithms for higher levels of accuracy.

Nguyen Thi Thuy Linh et al., [2] proposed a system that applied AI based models like LSTM, ELM, GRNN and HW along with ensemble models such SAE and WAE (linear) and BPNN-E and HW-E (non-linear) and a hybrid random forest ensemble for prediction of dissolved oxygen in water. The performance metrics used wre NSE, WI, RMSE, MAE, CC and MSE. Furthermore, sensitivity analysis was conducted using non linear input variables selection methods. The results indicated that HW(M3) was the best model for prediction. GRNN (M4) was considered to be the second best followed by ELM(M1) and LSTM(M3). Among ensemble models BPNN-E was found to be the most effective compared to the other three models. All the hybrid models showed great results but HW-RF ensemble seemed to be the best.

Rodelyn Avila et al., [5] performance of statistical model in water quality prediction. The statistical models used were naive model, multiple linear regression, dynamic regression, regressiontree, Markov chain, classification tree, random forests, multinomial logistic regression, discriminant analysis and Bayesian network. The data was collected as weekly data over the summer months fro 2006 to 2014 on the Oreti river in Wallacetown, New Zealand. Results indicated that Bayesian network had the best performance compared to all other models that were researched upon. They also mentioned that they were going to research on other sites using the same tool.

Rahim Barzegar et al., [6] researched on multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. They used ELM and WA-ELM to forecast 1, 2 and 3 months ahead EC and employed an integrated method to use the advantages of both by using boosting ensembles. EC values from a study site in Iran over a time period of 26 years were used for training and testing. The performance metrics used were root mean squared error, coefficient of determination, and Nash-Sutcliffe model efficiency coefficient. ELM and ANFIS didnt show great results. However, making hybrid WA-ELM and WA-ANFIS models indicated that the hybrid version had greater performance than the individual models. Between the two hybrids, WA-ELM had better performance.

Xingguo Chen et al., [7] identified suitable model for water quality prediction among traditional, ensemble, cost-sensitive, outlier detection learning models and sampling algorithms. The dataset was obtained from Genetic and Evolutionary Computation conference. The Traditional models used were decision tree, Logistic regression, K-nearest neighbour, and support vector machine. The ensemble techniques chosen were RF, DCF, and gradient boostingdecision tree.

Outlier detection models used were ODPCA, ODLOF, ODMCD and ODHBOS. Sampling algorithms used were SMOTE+TLTE and SMOTE+ENNTE. Cost sensitive learning models were LR, SVM, DT, RF and AdaCost. The performance metrics chosen were precision and recall and F1 score. DCF was found to be best performing among all the chosen models for water quality prediction. Cost sensitive RF and AdaCost were also observed to produce excellent results.

Gozen Elkiran et al., [10] researched on Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. The AI models used were BPNN, ANFIS, SVM and a traditional linear model ARIMA along with three other ensemble techniques to enhance their performances.These were applied on single and multi-step ahead modelling of dissolved oxygen in the Yamuna river. The performance metrics used were Determination coefficient and root mean square error. For SL1, ANFIS has better results, for SL2, ANFIS had better performance too. However for SL3, SVM had better results compared to others. To increase performance, ensembling was used. SAE AND WAE (linear) and NNE (nonlinear) were used. It was recorded that NNE was proved the most effective of all other techniques. The research also indicated thaat introducing more algorithms with many combinations of ensemble techniques could produce better performance and reliability.

Ming Chen et al., [14] attempted to design a data pre-processing model based on statistical detection methods n order to help power grid works improve efficiency by optimizing the data set quality. The statistical methods chosen were quartile detection method and Z-score method. The Z-score method was used to filter the result produced by the quartile detection method. The intersection of the two screening results were used as the final screening result, thus producing a

composite mathematical statistics screening model. The main functionality was based on quartile detection method, but its screening result was corrected by the Z-score method. Finally, the produced result after screening was processed into a matrix format.They also indicated that the model can further be optimised by adding ore data features so that more meaningful tags can be added which will in turn produce greater performance.

Y. Khan et al., [15] proposed an ensemble of ANN and ANFIS for water quality prediction and analysis. The data consisted of the measurements of water quality parameters with 30-minute time interval from the year of 2015. The research used a hybrid of ANN and ANFIS in order to record the prediction accuracy. The compared the results with ANN and ANFIS individually. The performance metric used were Correlation coefficient, Mean squared error, and root mean squared error. While comparing he ensemble and the unique models, it was proved that the ANN-ANFIS model was the most accurate with greater prediction accuracies. They also indicated that more hybrid models need to be devised in order to improve the prediction ability furthermore.

Ozgur Kisi et al., [16] proposed a new ensemble model called Bayesian model averaging to predict dissolved oxygen lecvels in water. The data used was a time series data from Link River below Keno Canal and Klamath River above Keno Dam near Keno operated by USGS. The proposed model was compared with extreme learning machine, artificial neural network, adaptive neuro-fuzzy inference system, classification and regression tree and multilinear regression.The performance metrics used were root mean square errors (RMSE), Nash-Sutcliffe efficiency, and determination coefficient. The results indicated that the proposed method performed better than the ELM, ANN, ANFIS, CART, and MLR.

Li-ming (Lee) he et al., [13] developed an ANN model to predict TC, FC and EN in two popular beaches of San diego, California. The parameters used were water temperature, conductivity, turbidity, rainfall and time lapse from last rainfall. The performance metrics were root mean squared error and false positive/negative rates. The model was found to have great performance with fast prediction ability. They also indicated that the model is ready to be employed to other coastal beaches too.

A. Najah et al., [20] used models such as multi-layer perceptron neural networks, ensemble neural networks and support vector machine. The dataset used was from Johor river in Johor state, Malaysia. The parameters upon which the dataset was based were COD, Dissolved oxygen, and BOD. The performance metrics used were correlation coefficient, mean square error, and correlation of efficiency. In the research, it was cited that SVM was found to overcome all the drawbacks of the other models, showing greater prediction ability with minimal computation.

Navideh Noori et al., [21] researched on water quality prediction using SWAT-ANN coupled approach. The research indicated that hybrid models can be more robust and accurate than standalone models. ycombined a process-based watershed model and an artificial neural network. The dataset used was from the watersheds in the Atlanta Metropoliton area, USA. The performance metric used was mean square error. The model was found to work very well, so much that it outperformed the SWAT models at each site.

Sanghyun Park et al., [24] researched on variable update strategy to improve water quality forecast accuracy in multivariate data assimilation using ensemble kalman filter. The dataset used was from Yeongsan river in the southwest part of Korea. The Hydrologic Simulation Program-Fortran (HSPF) model and the

Environmental Fluid Dynamics Code (EFDC) model were employed. The performance metric used was root mean square error. Case IV, which had CHL, PO4 3–P, NH4+ -N, and DO as parameters was found to be the best performing case.

Ali Omran Al-sulthani et al., [4] proposed ensemble data-intelligence models for surface water quality prediction. The ensemble models used were Quantile regression forest, random forest, radial support vector machine, stochastic gradient boosting and gradient boosting machines. There were used to predict the biochemical oxyggen demand values of the Euphrates River Iraq. The performance metrics used were determination coefficient, root mean square error, mean absolute error, Nash-Sutcliffe model efficiency coefficient, Wlimott Index and percent bias. The results indicated that the QRF model seemed to have the maximum performance. It was shown that the PCA-QRF integrated model had greater performance than standalone models. They mentioned that using metaheuristic optimization algorithm can further improve performance.

Leizhi Wand et al., [28] researched on improving robustness of Beach water quality modeling using ensemble machine learning approach. This research proposed an ensemble machine learning technique called model stacking. The outcome of five common individual machine learning models, that is- multiple linear regression, partial least square, sparse partial least square, random forest, and Bayesian network were taken as input for another model that gave the final prediction. The dataset used was from three beaches along eastern Lake Eric, New York, USA. The performance etrics used were Cross validation, MSE and accuracy. It was observed that different models performed well for different

samples. It was cited that model stacking can improve the effectiveness of beach water quality modeling.

Abobakr Saeed Abobakr Yahya et al., [3] developed a model using Support vector machine to predict the quality of water in the Langat river Basin. The parameters in the dataset were pH, suspended solids, dissolved oxygen, ammonia nitrogen, cod, and bod. The performance metrics used were mean squared error and correlation coefficient. The model was found to be very efficient and robust. The maximum error in prediction was found to be only 1 percent. However, the research also indicated that there is a need to improve the performance by deploying optimal kernel parameters and select Nu-RBF as the optimal model.

D. Venkata Vara Prasad et al, [8] explored different types of machine learning algorithms to predict the water quality index and the water quality class. The dataset was collected from Korattur Lake in Chennai city and tested for its necessary hydro-chemical parameters. The ML models used were support vector machine, decision tree, logistic regression, random forest, and naive Bayesian. The performance metrics were accuracy and precision. Among all the algorithms, the random forest algorithm produced an accuracy of 95% which was the highest with least execution time. Their scope for future work was mentioned to be including some more classes to the data set and training using hybrid models of machine learning and deep learning.

Venkata Vara Prasad D et al, [9] explored many deep learning algorithms to predict the Water Quality Index and the Water Quality Class. The dataset was collected from Korattur Lake in the Chennai city, Tamilnadu. Artificial Neural Networks, Recurrent Neural Networks, and Long-Short Term Memory were the deep learning models used. Accuracy, precision, and execution time were the

performance criteria used. The results showed that LSTM had the highest accuracy of 94 percent while also taking the shortest time to execute.

D. Venkata Vara Prasad et al. [25] compared AutoML with an expert architecture established by the authors to evaluate the Water Quality Index and the Water Quality Class using Machine Learning techniques. The findings revealed that the AutoML and TPOT were 1.4 percent more accurate than traditional machine learning algorithms.For water data of the binary class AutoML was used in the situation of multi-class water data. Traditional ML approaches have a 0.5 percent higher TPOT and a 0.6 percent higher TPOT.

ShuangyinLiu et al, [19] presented a hybrid model called real-value genetic algorithm support vector regression (RGA–SVR), which searched for the optimal SVR parameters using real-value genetic algorithms, to further construct the SVR models. This was used to predict the aquaculture water quality data collected from the aquatic factories of YiXing, in China. The results showed that RGA–SVR performed better than the traditional SVR and back-propagation (BP) neural network models based on the root mean square error (RMSE) and mean absolute percentage error (MAPE).

Yunrong Xiang et al, [29] dealt with water quality prediction model through application of LS-SVM in Liuxi River in Guangzhou. Least squares support vector machine (LS-SVM) along with particle swarm optimization (PSO) was used for time series prediction. Testing the model showed high efficiency in predicting the water quality of the Liuxi River.

Chenguang Song et al,[27] proposed a hybrid model based on the ensemble learning method that combined the entire ensemble empirical mode

decomposition with adaptive noise (CEEMDAN) and improved LSTM to predict the water quality parameters. The results showed that the proposed model had greater accuracy than models.

ZilinLi et al, [18] presented a new stacking ensemble model for detection of water quality using multiple parameters. The model consisted of a number of machine learning base predictors and a meta-predictor, and was trained using cross-validation for water quality prediction. The stacking method had a higher true positive rate, lower false positive rate and higher F1 score.

Park, Jungsu [22] developed an ensemble machine learning model to predict Suspended sediment concentration using the XGBoost (XGB) algorithm. The discharge (Q) and SSC in a couple fields monitoring stations were used to construct the model. The variables were divided in two groups with low and high ranges of Q employing the k-means clustering algorithm. The RSR were 0.51 and 0.57 in the two monitoring stations for Model 2, respectively, while the model performance improved to RSR 0.46 and 0.55, respectively, for Model 1.

JungsuPark, [23] developed an XGBoost ensemble machine learning (ML) model from 18 input variables to predict Chl-a concentration.The performance metrics used were root mean square error (RMSE), RMSE-observation standard deviation ratio, and Nash-Sutcliffe efficiency. This study successfully demonstrated a good example of XAI application to improve the machine learning model performance in predicting water quality.

LingboLi et al, [17] evaluated five tree-based models- classification tree, random forest, CatBoost, XGBoost, and LightGBM, and employed an explanation method SHAP to explain the models used. The results suggested that the combination

of LightGBM and SHAP had good potential to develop interpretable models for predicting microbial water quality in freshwater lakes.

Farid Hassanbaki Garabaghi et al, [12] analysed the performance of four machine learning algorithms with ensemble learning approach and proposea a classification model (classifier) with highest performance. Three feature selection methods employing machine learning were applied.As a result XGBoost classifier was suggested as the best classifier with the maximum accuracy of 95.606%.

Arshia Fathima et al, [11] the present study focuses on devising a prediction model for BOD using ensemble techniques in data mining. A correlation coefficient of 0.9541 and a root mean-squared error of 0.4679 were obtained for the proposed BOD prediction model on river water quality data. Bagging for the river data showed good results without over-fitting.

Rosaida Rosly et al, [25] compared ensemble classifiers based on 10-fold cross validation on water quality for datasets of Kinta River, Perak. The ensemble methods used were boosting, bagging, and stacking. The result showed that the stacking method with MLP algorithm achieved higher accuracy of 96.39%.

Dipankar Ruidas et al, [26] researched on three important machine learning algorithms- bagging, random forest (RF), and an ensemble of bagging and RF were employed to assess the HHRM. Performance was analysed using statistical validating methods such as AUC-ROC, sensitivity, specificity, accuracy, precision, F-score, kappa, and Taylor diagram. The results showed that ensemble technique was best performing.

CHAPTER 3

SYSTEM DESIGN

3.1 SYSTEM MODULES

The datasets to be used in this research are obtained through the following processes - dataset collection, pre-processing and manipulation. The data is then fed into the statistical and ensemble models in parallel. Then these models are individually built and trained. The results are then analysed and the best performing statistical and ensemble models are then chosen for building the hybrid model. The architecture diagram of the proposed system is shown in Figure 3.1.

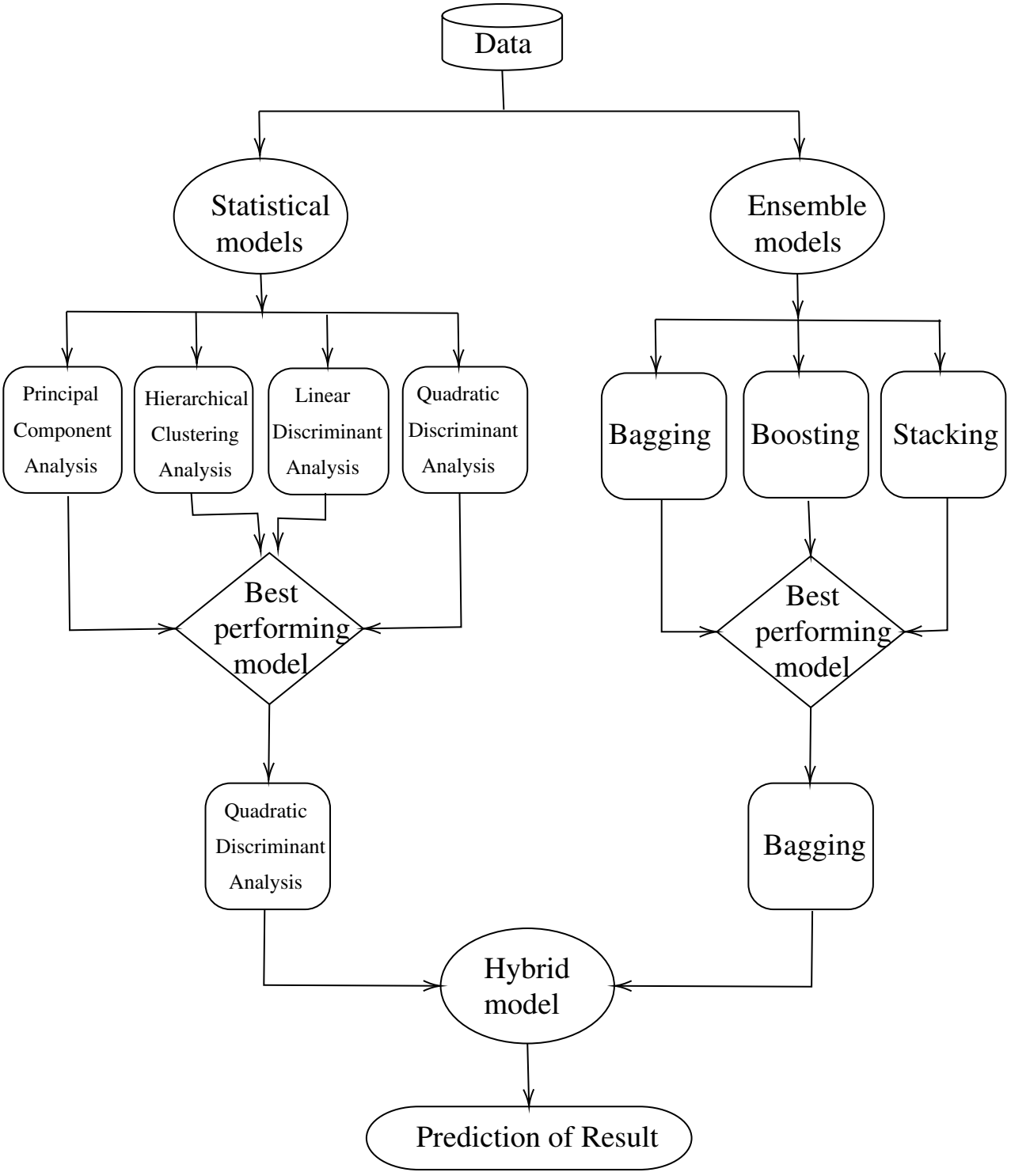

FIGURE 3.1: Architecture diagram of Proposed System

3.2 DATA COLLECTION AND EXPLORATORY DATA ANALYSIS

3.2.1 Data Collection

This research compares the outcomes of the statistical, ensemble and the hybrid models using four datasets out of which two are binary class and two are multiclass dataset. The datasets sourced from Korattur Lake as mentioned in Table 3.1, have parameters which were collected for over 10 consecutive years (2010 to 2019). The Korattur Lake is one of the biggest lakes in the Chennai city. The dataset obtained from Kaggle contains water quality metrics obtained from 3276 different water bodies.

Dataset	Dataset Shape	No of Classes
Korattur Lake	5001 x 10	two(0 and 1)
Kaggle	8000 x 21	two(0 and 1)
Korattur Lake	10140 x 10	three(0, 1 and 2)
Korattur Lake	5100 x 10	five(0, 1, 2, 3 and 4)

TABLE 3.1: Dataset Description

3.2.2 Exploratory Data Analysis

In the pipeline of execution, the data is first analysed to understand the inherent distribution and to discover if pre-processing is required.

3.2.3 Binary Class Dataset from Korattur Lake

The rows in the dataset are classified into two classes 0 and 1 based on 9 attributes. The 9 attributes are - pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD(mg/L), Chlorine and Sodium. The size of the dataset is 5001 x 10, that is it has 5001 rows and 10 columns including the class label.

Figure: 3.2 shows the class distribution of Korattur Lake Binary Class Dataset with classes 0 and 1.

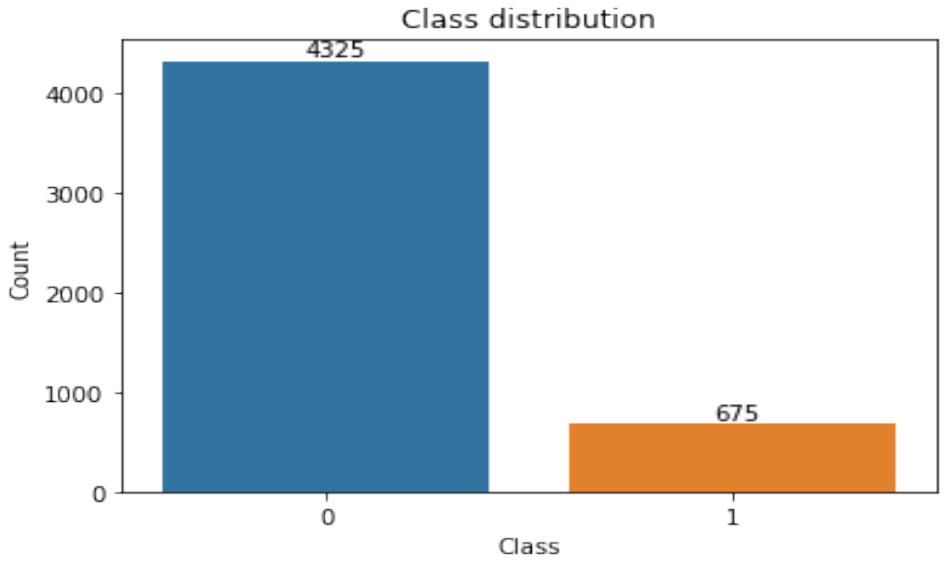

FIGURE 3.2: Class distribution of Korattur Lake Binary Class Dataset

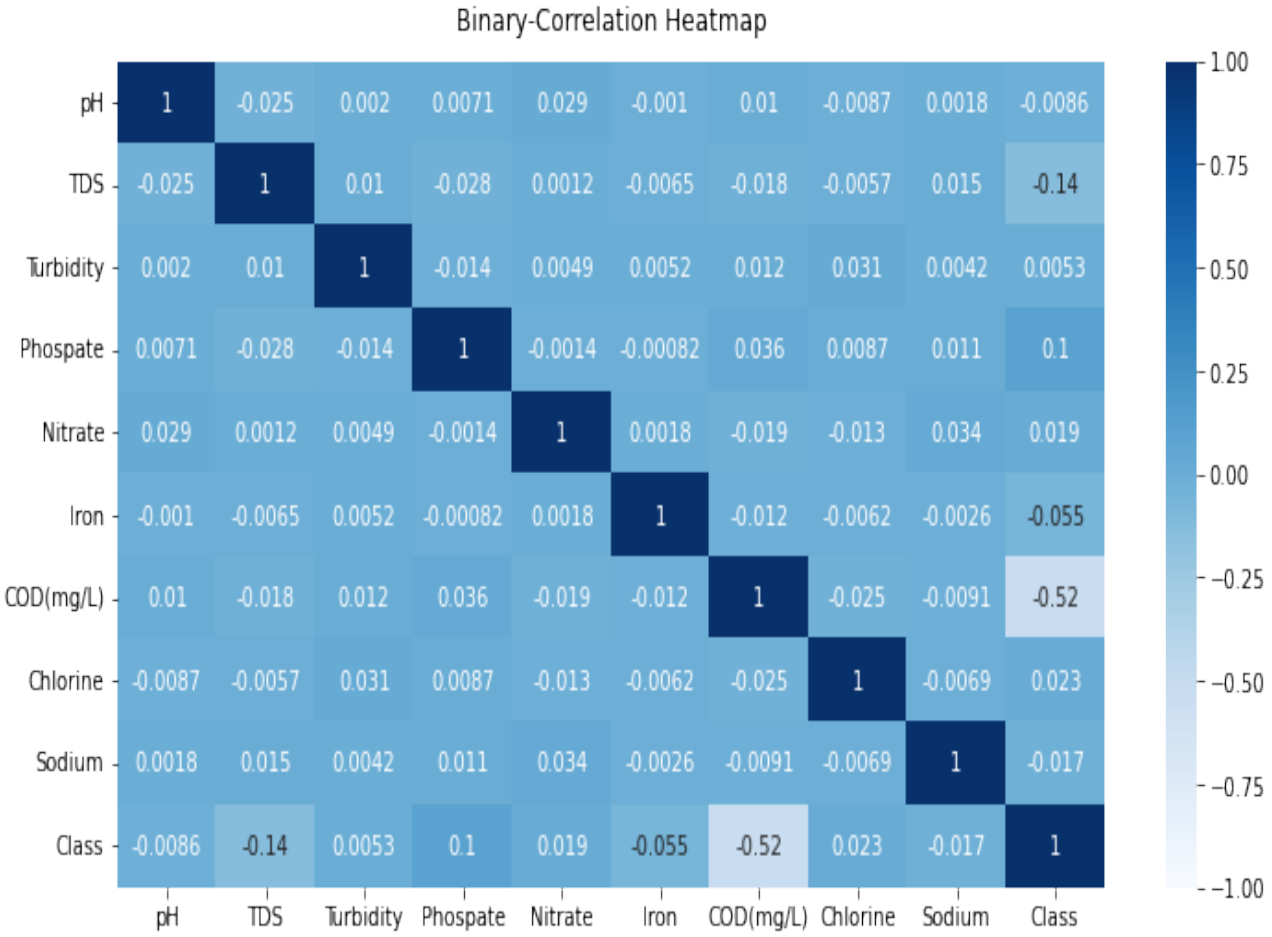

FIGURE 3.3: Heatmap of Korattur Lake Binary Class Dataset

Figure: 3.3 shows the correlation between different parameters in the Binary Class Dataset sourced from Korattur Lake using a Heatmap.

3.2.4 Binary Class Dataset from Kaggle

The rows in the dataset are classified into two classes 0 and 1 based on 20 attributes. The 20 attributes are - Aluminium, Ammonia, Arsenic, Bar- ium, Cadmium, Cloramine, Chromium, Copper, Fluoride, Bacteria, Lead, Ni- trates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver and Uranium. The size

of the dataset is 8000 x 21, that is it has 8000 rows and 21 columns including the class label.

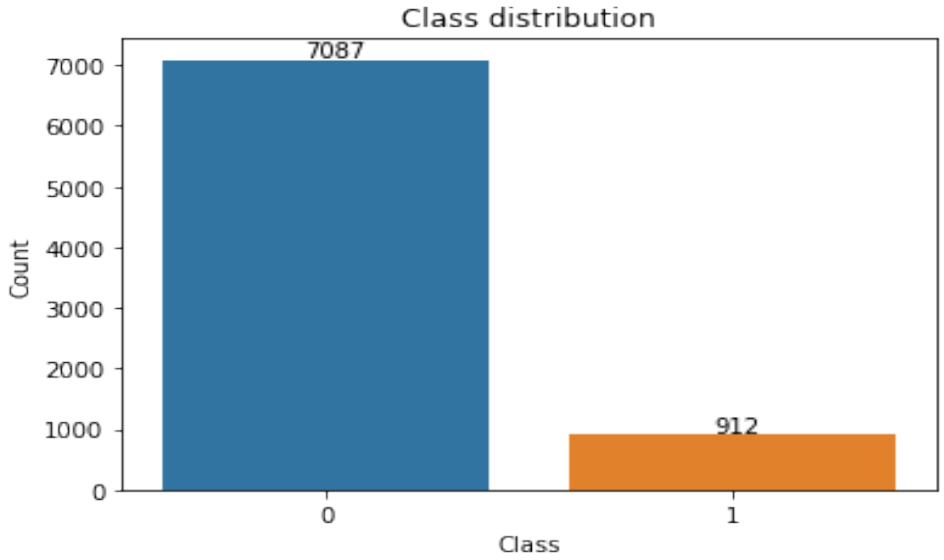

FIGURE 3.4: Class distribution of Kaggle Binary Class Dataset

Figure 3.4 shows the class distribution of classes 0 and 1 in the Binary Class Dataset sourced from Kaggle.

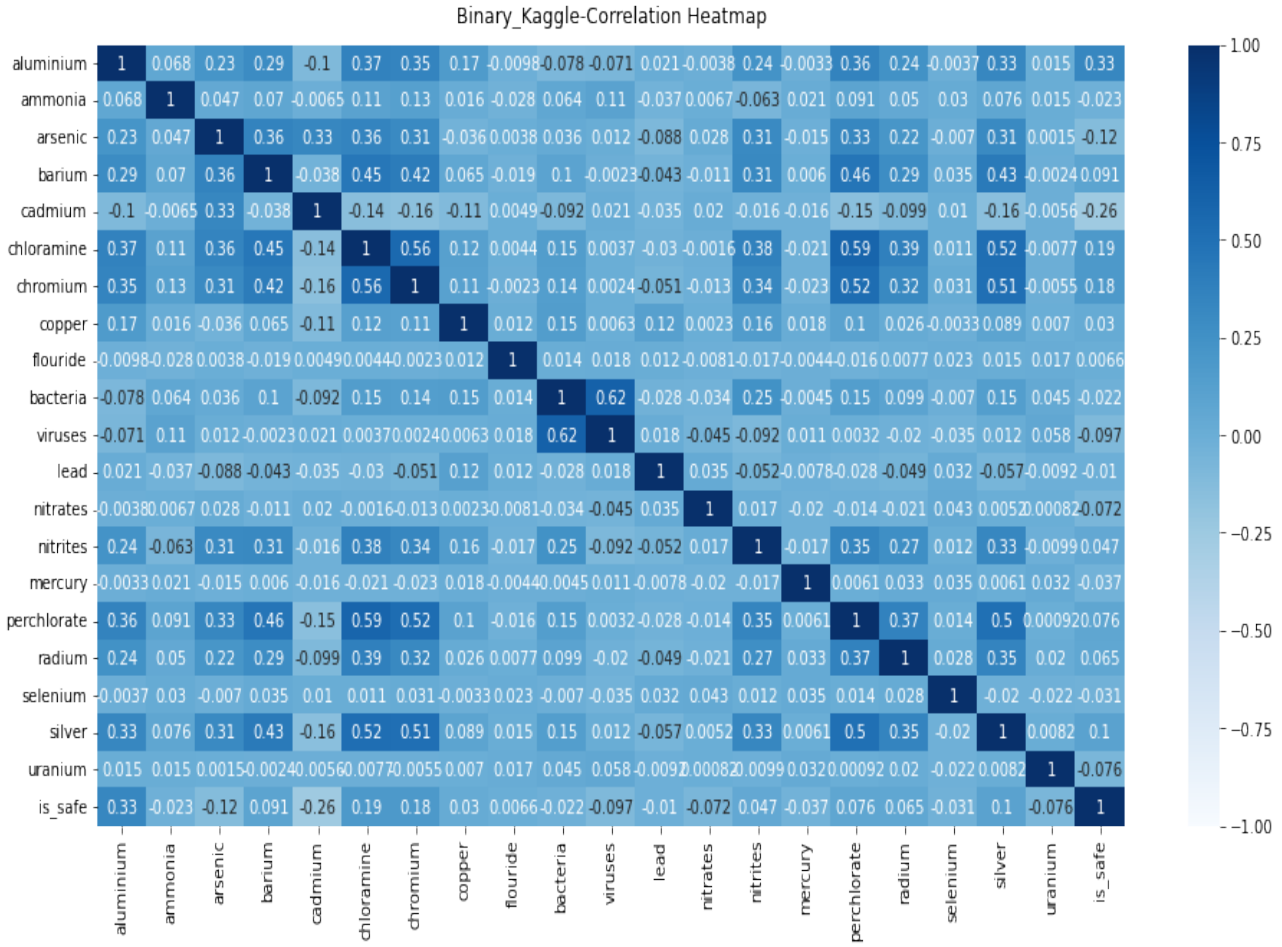

FIGURE 3.5: Heatmap of Kaggle Binary Class Dataset

Figure 3.5 shows the correlation between different parameters in the Binary Class Dataset sourced from Kaggle using a Heatmap.

3.2.5 Three Class Dataset from Korattur Lake

The dataset has three classes namely 0, 1 and 2 where 0 indicates that the quality of water is excellent, 1 indicates that the water is good, and 2 indicates that the water quality is poor. The classification is done based on 9 attributes. They are - pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD(mg/L), Chlorine and Sodium.

The size of the dataset is 10140 x 10, that is it has 10140 rows and 10 columns including the class label.

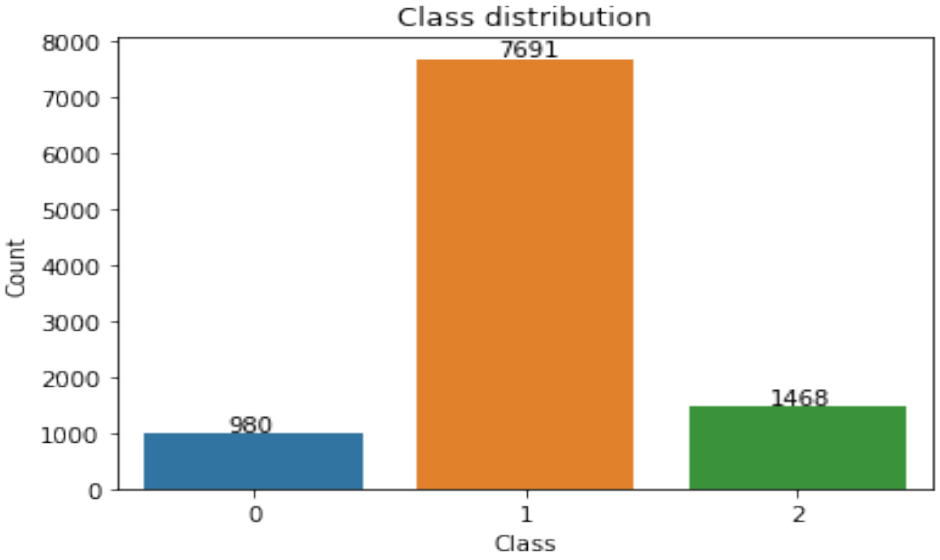

FIGURE 3.6: Class distribution of Korattur Lake Three Class Dataset

Figure 3.6 shows the class distribution of classes 0, 1 and 2 in the Three Class Dataset sourced from Korattur Lake.

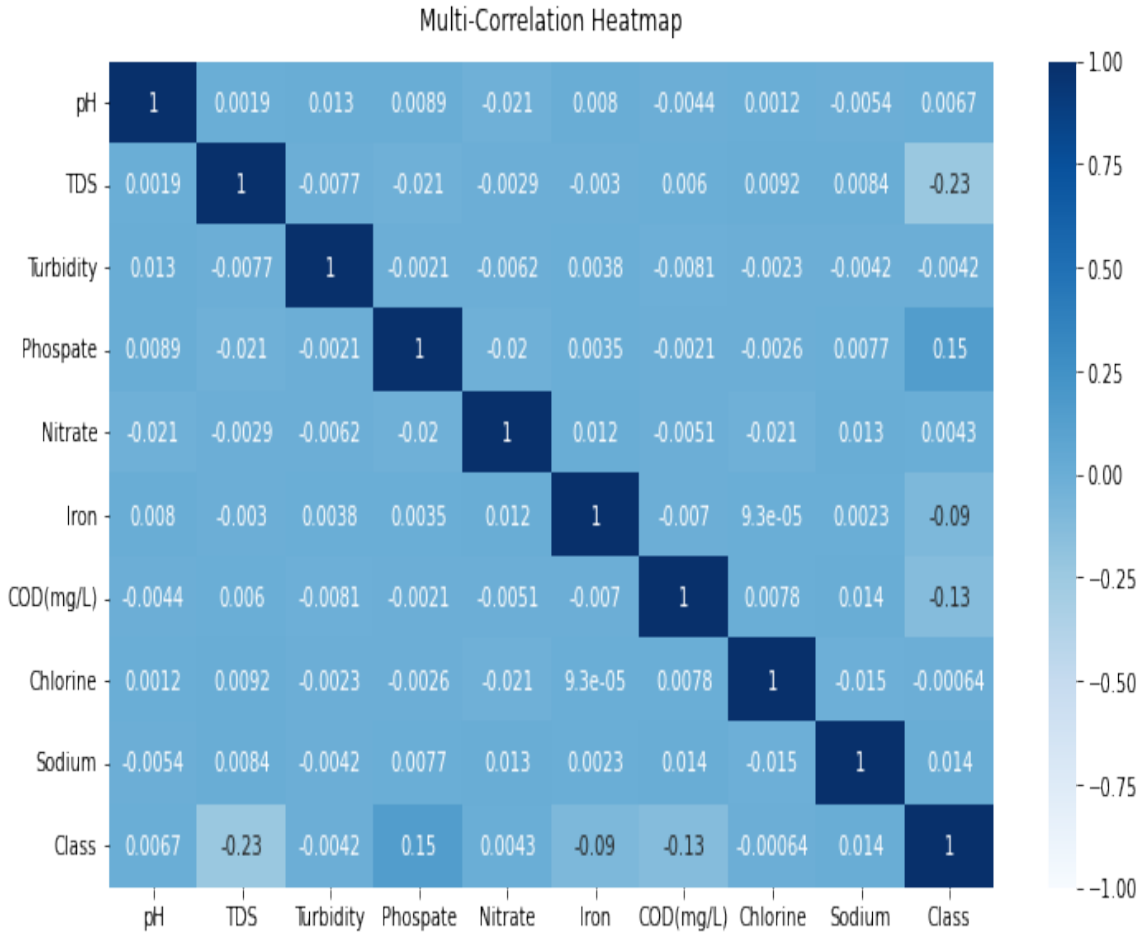

FIGURE 3.7: Heatmap of Korattur Lake Three Class Dataset

Figure 3.7 above shows the correlation between different parameters in the Three Class Dataset sourced from Korattur Lake using a Heatmap.

3.2.6 Five Class Dataset from Korattur Lake

The dataset consists of five classes where 0 stands for excellent, 1 stands for good, 2 stands for average, 3 stands for bad and 4 for poor water quality. The classification is done based on 9 attributes. They are - pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD(mg/L), Chlorine and Sodium. The size of the

dataset is 5100 x 10, that is it has 5100 rows and 10 columns including the class label.

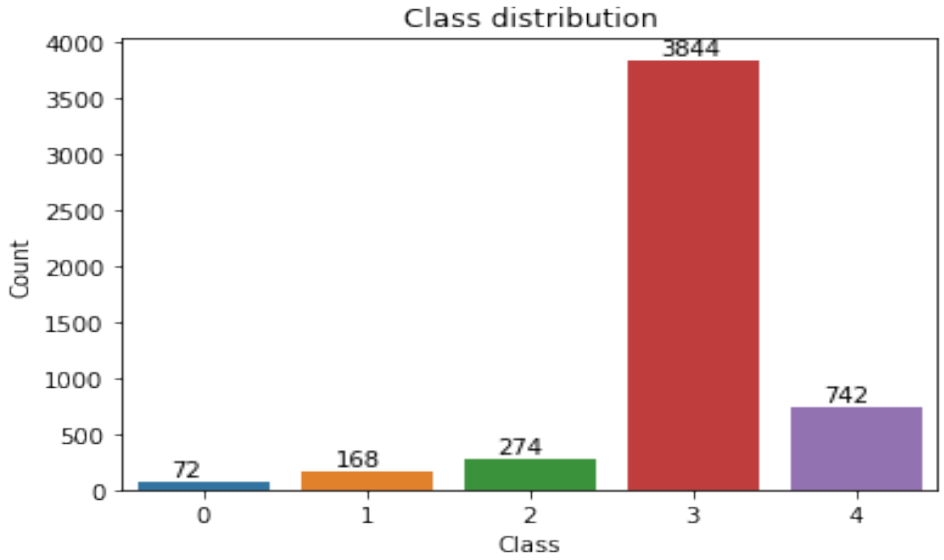

FIGURE 3.8: Class distribution of Korattur Lake Three Class Dataset

Figure 3.8 above shows the class distribution of classes 0, 1, 2, 3 and 4 in the Five Class Dataset sourced from Korattur Lake.

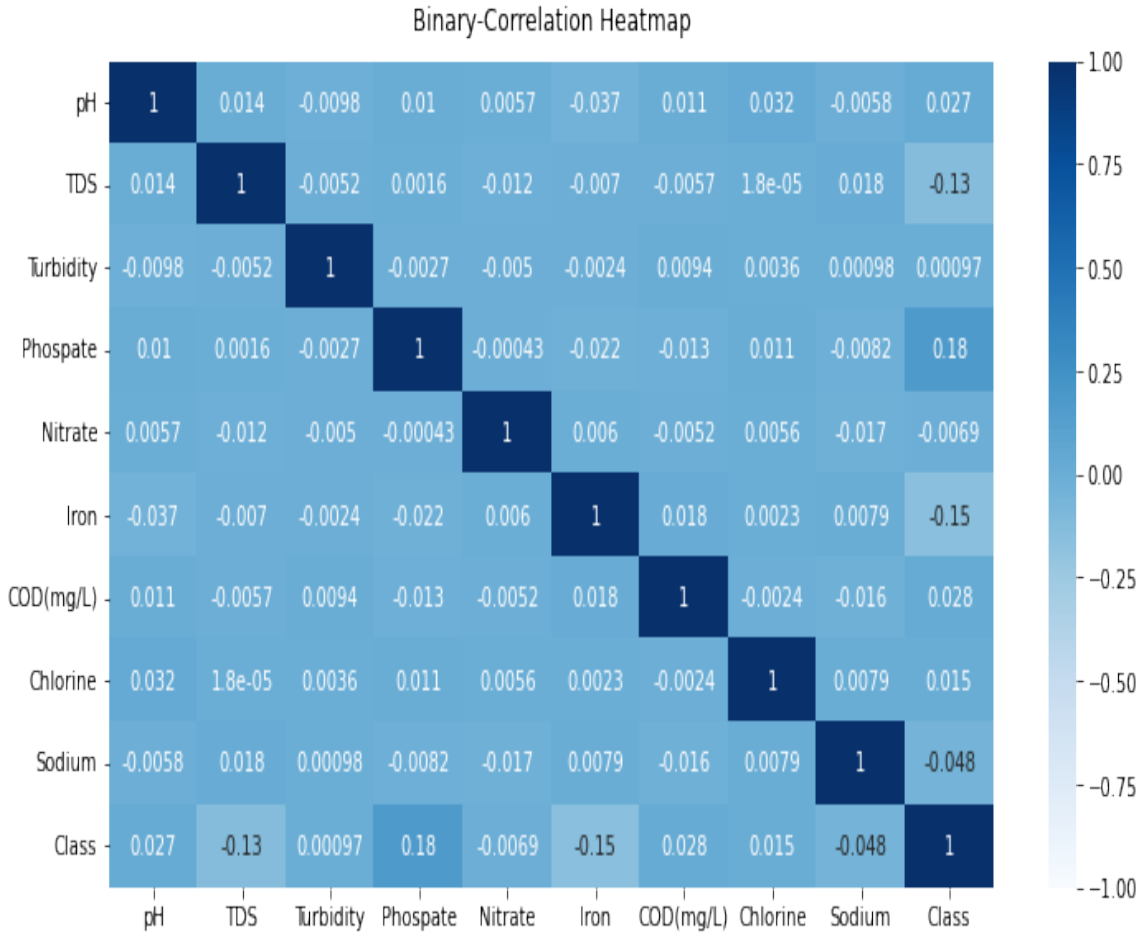

FIGURE 3.9: Heatmap of Korattur Lake Five Class Dataset

Figure 3.9 shows the correlation between different parameters in the Five Class Dataset sourced from Korattur Lake using a Heatmap.

3.3 MODEL TRAINING

The data was trained on various Statistical and Ensemble Models to decide on the best Statistical and Ensemble Model to be used in the Hybrid Model. The Statistical models which are implemented are the PCA[Principal Component Analysis], HCA[Hierarchial Clustering Analysis], LDA[Linear Discriminant

Analysis] and QDA[Quadratic Discriminant Analysis]. Out of the ensemble models, Bagging, Boosting and Stacking were implemented. Bagging was implemented using the Decision Tree classifier. Boosting was implemented using the AdaBoost classifier. The Stacking model was implemented using the base models such as decision trees, random forest, XGBoost, K neighbours and Linear regression as the final estimator. These various Statistical and Ensemble algorithms were tested to obtain parallel results.

3.4 MODELS COMPARISON AND ANALYSIS

The Statistical models were compared and the results were analysed to check which Statistical Algorithm works efficiently. The Ensemble models were also compared and the results were then analysed to find the most efficient algorithm. The conclusions were drawn based on the data in hand in context of the case being studied.

3.5 SELECTION FOR HYBRID MODEL

The best performing Statistical Model as well as the most efficient Ensemble Model are combined to form the Hybrid model.The Statistical, Ensemble and the Hybrid models are compared and their results are analysed to observe if the Hybrid model is more efficient than the independent Statistical and Ensemble models.

3.6 HYBRID MODEL

The architecture diagram for the hybrid model is shown in Figure 3.10 (where 'M' represents a ML model) The flow of the data is seen to first go to the selected Statistical model, then to the Ensemble model. The data is trained and tested through these two models which combine to form the hybrid model. The Hybrid model is then used to predict the result.

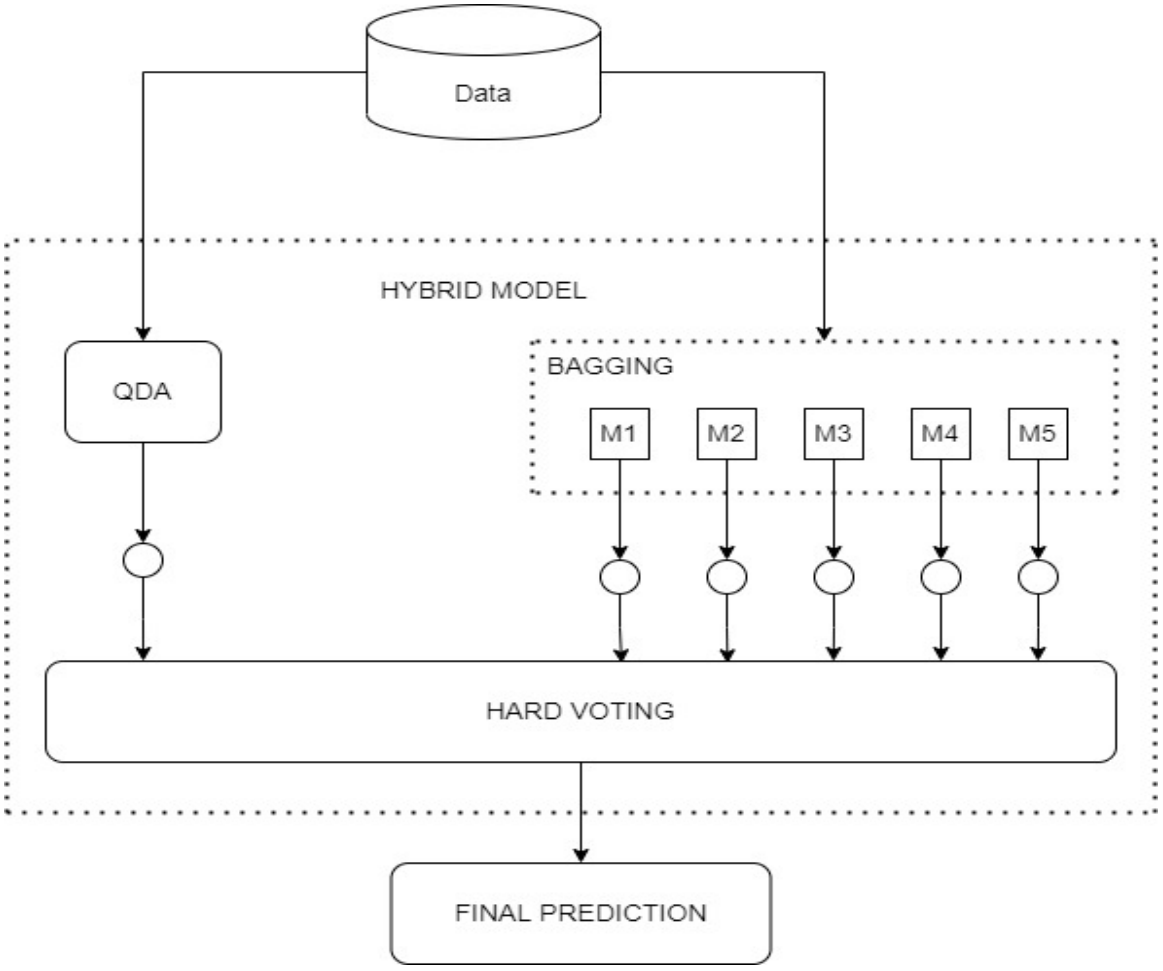

FIGURE 3.10: Architecture of the Hybrid Model

CHAPTER 4

IMPLEMENTATION

This research engages two parallel techniques to apply to the data. Various Machine and Deep Learning techniques are combined to construct the Ensemble models. Both the Statistical and Ensemble models are built in parallel. They are trained, and the results are obtained. Using the results we draw conclusions and select the best performing Statistical and Ensemble model to build the Hybrid Model.

4.1 STATISTICAL TECHNIQUES

4.1.1 Principal Component Analysis

PCA is a common model used for **dimensionality reduction**, that is,reducing the feature space by removing several features from a real world dataset that is noisy and unclean. By removing these features, the dataset is made much **easier to visualise, analyse and interpret**.PCA also **determines correlations between the features**. It is commonly used in the areas of pattern recognition and signal processing. There are two classes that come under PCA, namely - Feature Elimination and Feature Extraction. In this research, we use feature extraction. Figure 4.1 shows the steps in Principal Component Analysis.

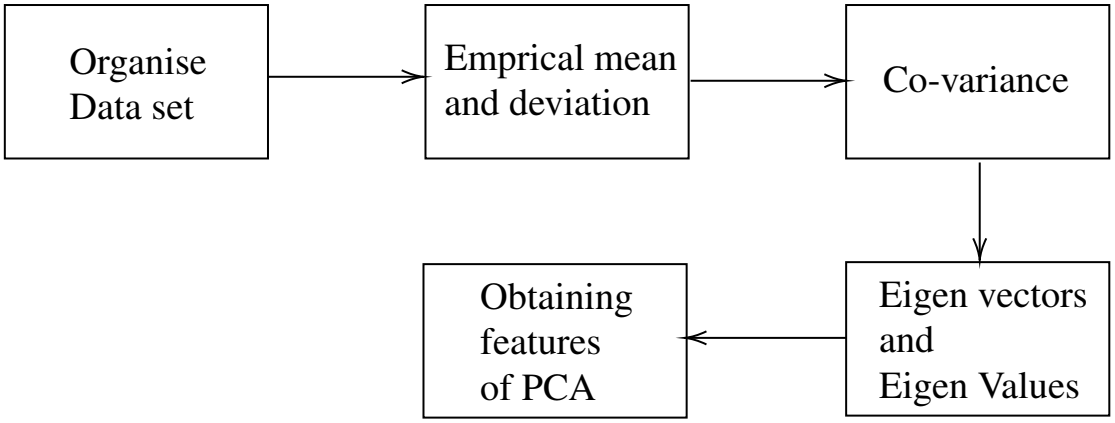

FIGURE 4.1: PCA

4.1.2 Hierarchical clustering Analysis

Hierarchical clustering analysis is a common model in which the objective is to **group several features/data points in such a way that they are close to one another**. The fundamental technique is to repeatedly calculate the distance between the features and further calculate the distances between the clusters once the features/ data points start forming clusters, as shown in Figure 4.2. The **outputs** are usually **represented** as a **dendrogram**. The two methods that fall under HCA are Divisive methods and Agglomerative methods.

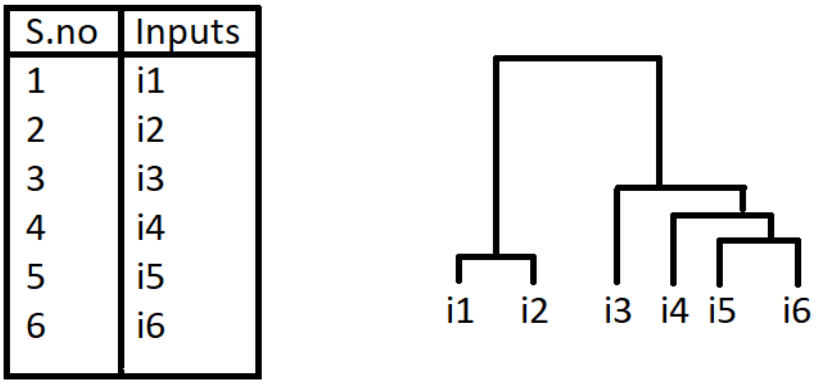

FIGURE 4.2: HCA

4.1.3 Linear Discriminant Analysis

Linear Discriminant analysis is also a commonly used **dimensionality reduction method** that is usually used in **supervised** classification problems. It is more precisely used to **model the differences between the groups/classes**. The higher dimension space is projected into the lower dimension space. Quadratic Discriminant analysis, flexible discriminant analysis and regularised discriminant analysis are the extensions to linear discriminant analysis. LDA is commonly used in the areas of medicine, face recognition, customer identification and so on. The flow of LDA is shown in Figure 4.3.

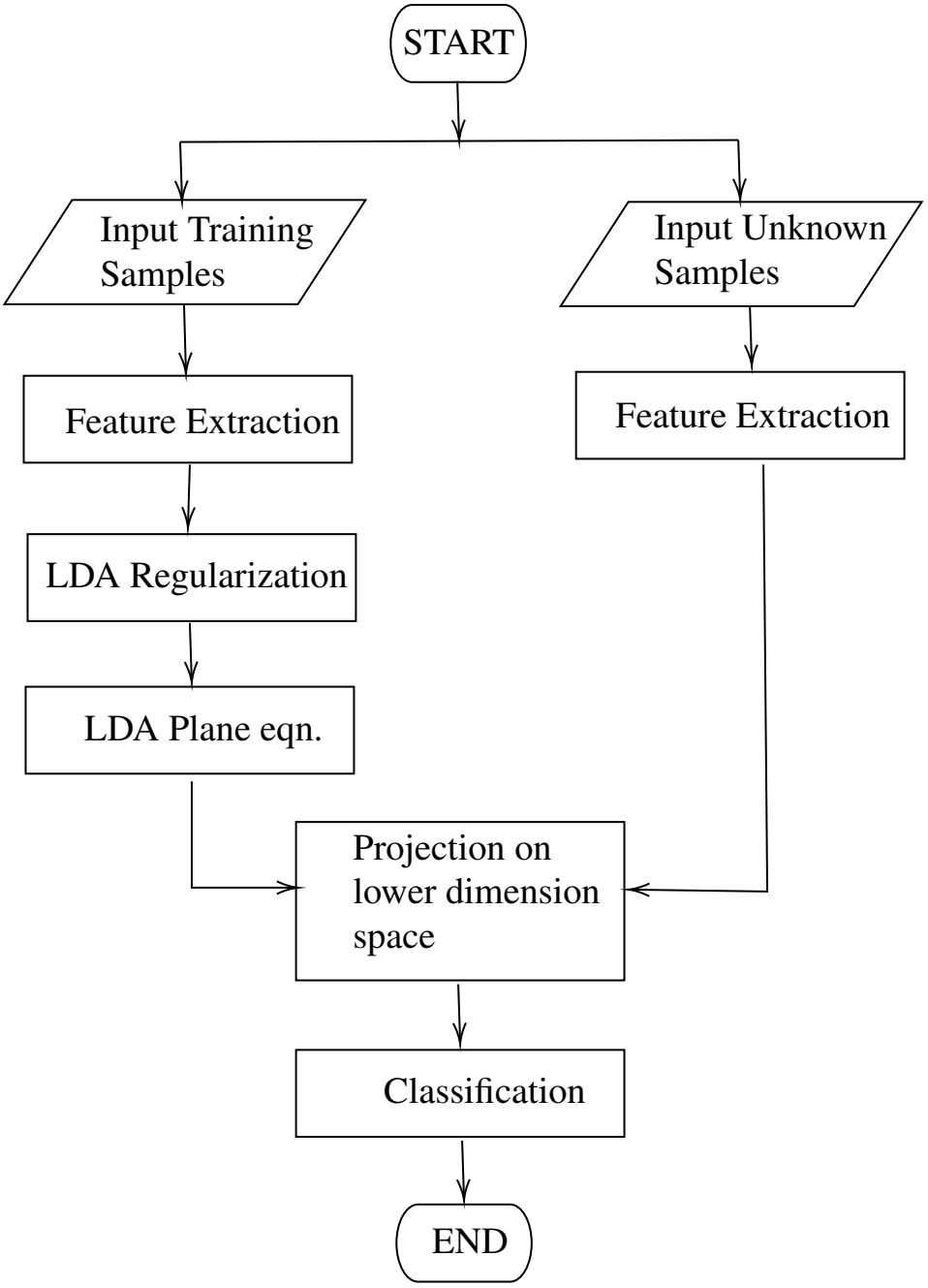

FIGURE 4.3: LDA

4.1.4 Quadratic Discriminant Analysis

QDA is quite related to linear discriminant analysis (LDA). It is assumed that the **measurements are normally distributed**. Unlike LDA, in QDA there is no assumption that the covariance of each of the classes is the same. QDA is a

generative model and it assumes that every class follows a Gaussian distribution.One aspect in which the two differ is that LDA assumes the feature covariance matrices of both classes are identical, which leads to a linear decision boundary. However, QDA is less stringent. The comparison of LDA and QDA is shown in Figure 4.4.

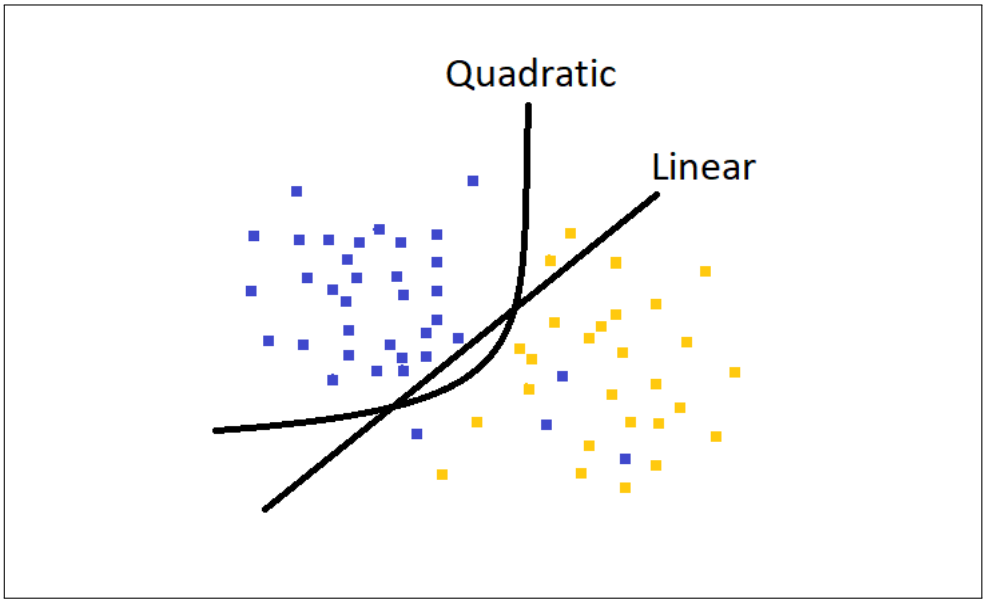

FIGURE 4.4: QDA vs LDA

4.2 BASE MACHINE LEARNING MODELS

4.2.1 Decision tree

A decision tree is a decision support tool that uses a tree-like model of **decisions and their possible consequences**, including chance event outcomes, resource costs, and utility. Their internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.

4.2.2 Random Forest

Random forest is an ensemble method that is made up of a large number of **small decision trees**, called **estimators**, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

4.2.3 XGBoost

XGBoost, which stands for **Extreme Gradient Boosting**, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides **parallel tree boosting** and is designed to be highly efficient, flexible and portable.It implements Machine Learning algorithms under the Gradient Boosting framework.

4.2.4 K-Nearest Neighbours

The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the **likelihood** that a data point will become a **member of one group** or another **based on** what group the data points **nearest** to it belong to.It is a **lazy learning** and **non-parametric algorithm**.

4.2.5 Logistic regression

Logistic regression is used to **predict a dependent categorical target variable**.The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

4.3 ENSEMBLE TECHNIQUES

4.3.1 Bagging

In parallel methods we fit the different considered learners independently from each other and, so, it is possible to train them concurrently. The most famous such approach is “bagging” (standing for “bootstrap aggregating”) that aims at producing an ensemble model that is more robust than the individual models composing it. Figure 4.5 shows the process of Bagging Ensemble.

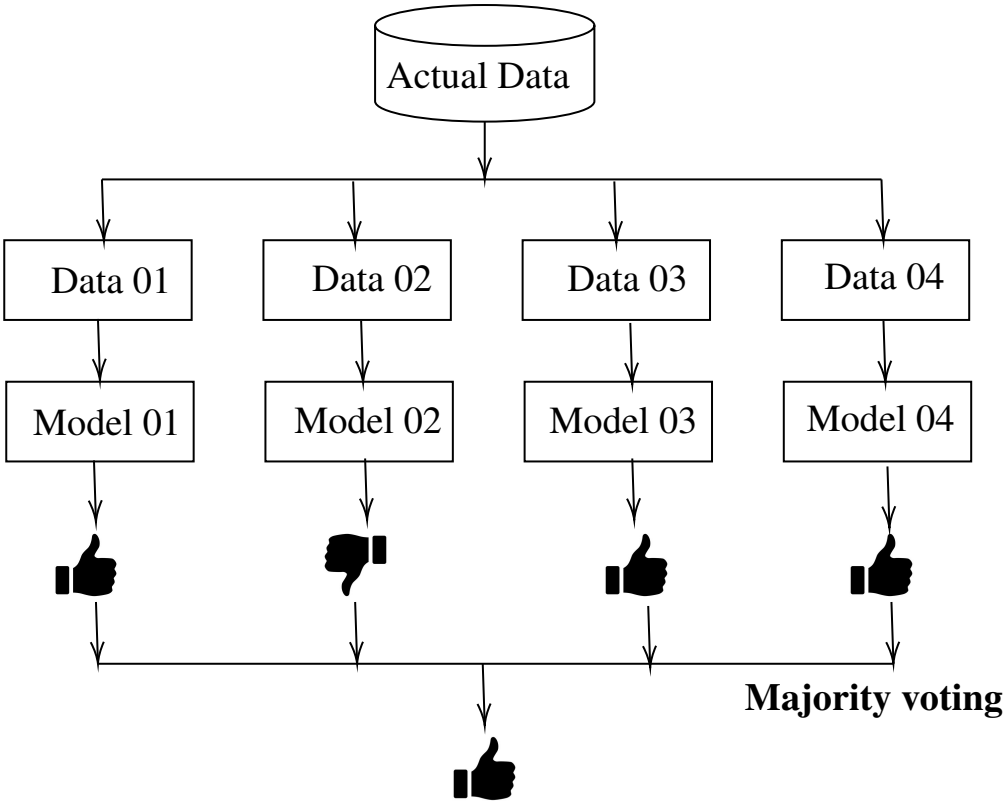

FIGURE 4.5: Bagging

4.3.2 Boosting

The various combined weak models are not fitted independently from each other in sequential methods. The process is to fit models repetitively such that the training of models at a given point relies on the models fitted at the previous points. "Boosting" is the most popular of these approaches and it presents an ensemble model that is much less biased than the weak learners that compose it. Figure 4.6 shows the process of Boosting Ensemble.

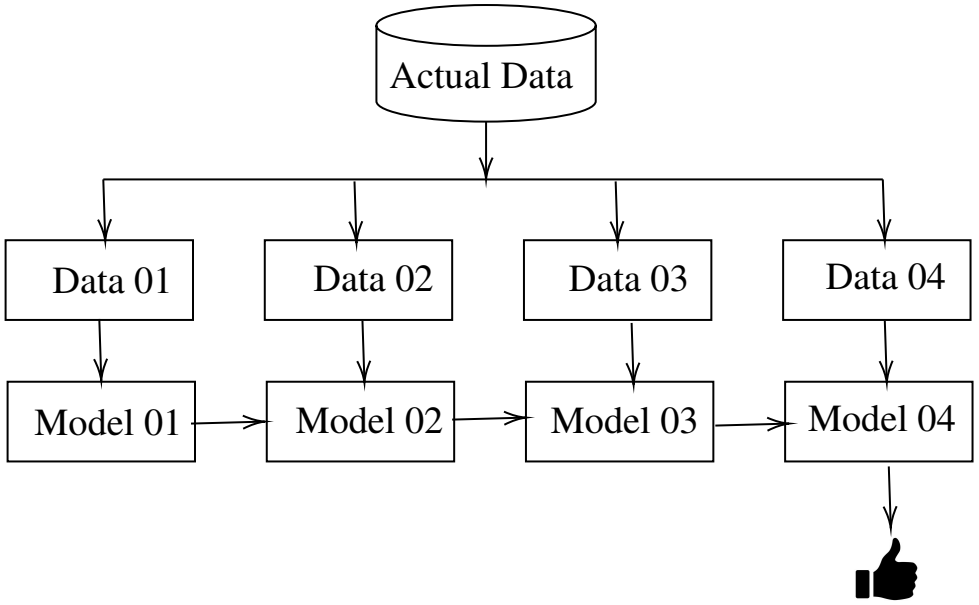

FIGURE 4.6: Boosting

4.3.3 Stacking

Learning several independent weak learners and combining them by training a meta-model to output predictions based on the multiple predictions produced by these weak models is the process of stacking. So, in order to develop our stacking model, we need to define two things: the L learners we want to suit and the meta-model that mixes them. The Stacking Ensemble procedure is depicted in Figure 4.7.

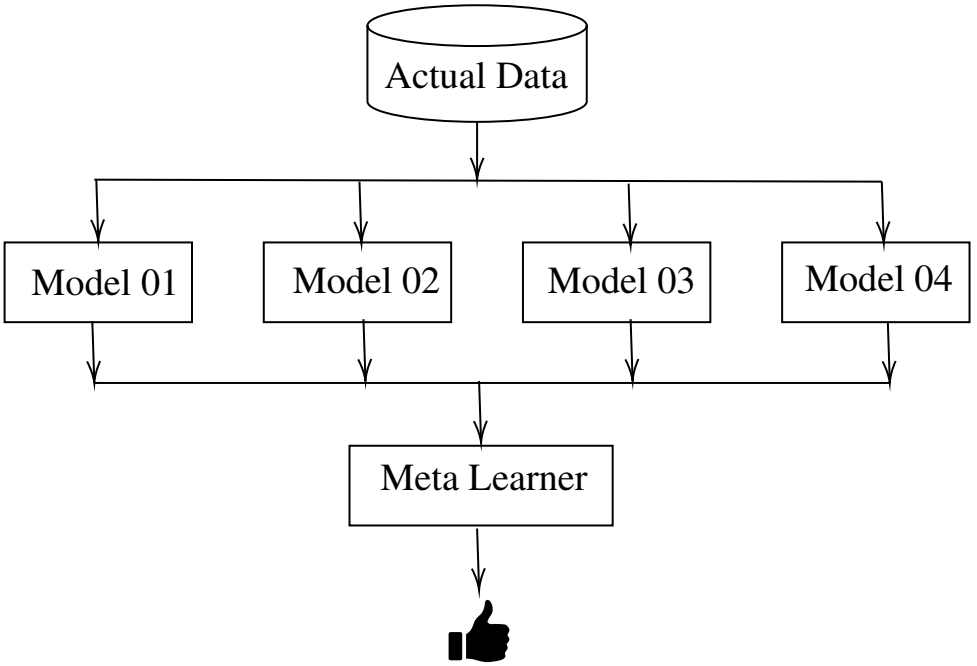

FIGURE 4.7: Stacking

CHAPTER 5

RESULTS AND CONCLUSION

The result will be consolidated as per the following terms. The overall results with respect to the approach used along with feature wise contribution to the prediction will enable better explainability. This consolidation provides a better holistic approach as well as a proper recognize the independent contribution.

5.1 MODEL EVALUATION METRICS

5.1.1 Confusion Matrix

The Confusion Matrix is a matrix used for making out the performance of a classification model. The confusion matrix cam=n be plotted only if the data is already labelled. It is also known as Error matrix since it shows the errors of a model in the form of a matrix. The confusion matrix as shown in Figure 5.1, compares the actual values with the predicted values by the model.

True Positive(TP): Positive samples are correctly predicted as positive. Eg: Drinkable water is predicted as drinkable.

True Negative(TN): Negative samples are correctly predicted as negative. Eg: Non drinkable water is predicted as non drinkable.

False Positive(FP): Negative samples are incorrectly predicted as positive. Eg: Non drinkable water is predicted as drinkable.

False Negative(FN): Positive samples are incorrectly predicted as negative. EG: Drinkable water is predicted as non drinkable.

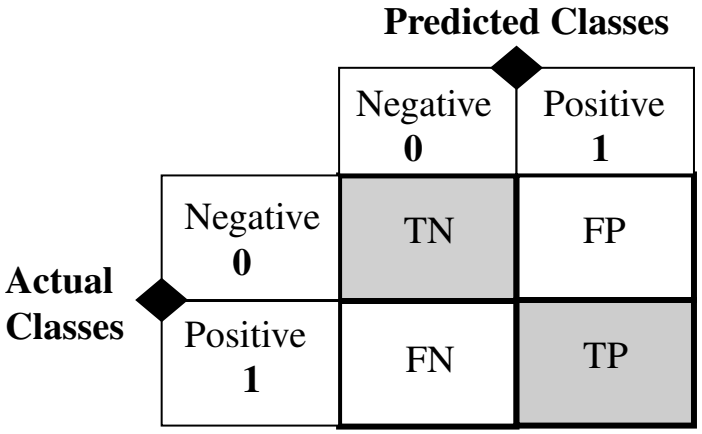

FIGURE 5.1: Confusion Matrix

5.1.2 Accuracy

Informally the definition of accuracy goes as follows: It is the fraction of predictions the model built got right. Formally, Accuracy is the ratio of number of correct predictions to the total number of predictions. The closer the Accuracy of a model is to 1, the better the model.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{5.1}$$

In mathematical form, it may also be represented as follows:

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{5.2}$$

5.1.3 Precision

Precision attempts to find the proportion of positive identifications that was actually correct. It is defined as the ratio of number of true positives to the total number of positive predictions. Eg: It refers to the rate of number of samples

correctly predicted as drinkable out of all the samples classified as drinkable by the model.

$$Precision = \frac{TP}{(TP + FP)} \tag{5.3}$$

5.1.4 Recall

Recall refers to the proportion of actual positives that was identified correctly. It is defined as the ratio of true positives to the sum of true positive and false negative, as shown in Figure 5.2. Eg: It refers to the number of samples correctly predicted as drinkable out of all the samples that are actually drinkable.

$$Recall = \frac{TP}{(TP + FN)} \tag{5.4}$$

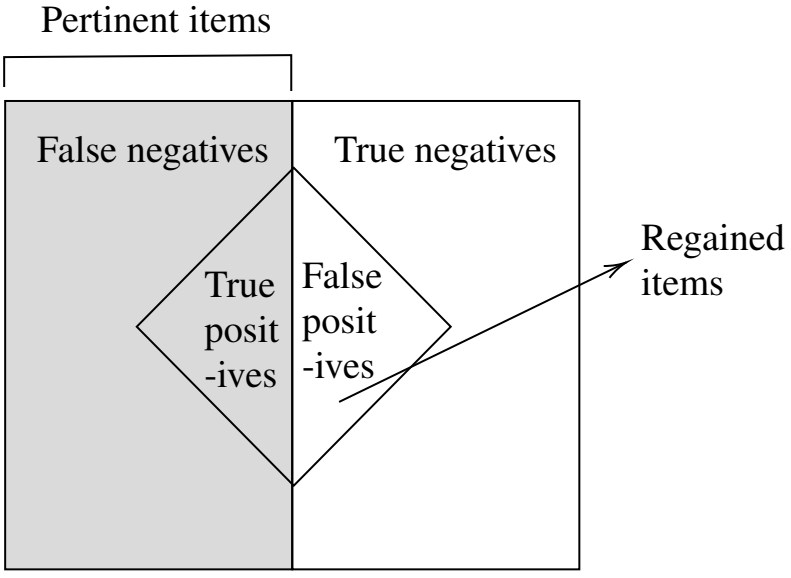

FIGURE 5.2: Precision and Recall

5.1.5 F1-Score

F1-score is the harmonic mean of precision and recall. It combines precision and recall into a single number.The higher the precision and recall, the higher the F1-score. F1-score ranges between 0 and 1. The closer it is to 1, the better the model.

$$F1 - score = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{5.5}$$

5.1.6 ROC curve

ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

1.True Positive Rate(TPR)

$$TPR = \frac{TP}{(TP + FN)} \tag{5.6}$$

2.False Positive Rate(FPR)

$$FPR = \frac{FP}{(FP + TN)} \tag{5.7}$$

AUC stands for ”Area under the ROC Curve.” AUC represents the probability that a random positive example is positioned to the right of a random negative example.

5.2 TABULATION AND INFERENCES

The Statistical, Ensemble and Hybrid Models are all evaluated in the scale of the above mentioned performance metrics. The classification is in four ways one for each dataset used. There are two binary class datasets and 2 multi class datasets. The classes are distributed based on the parameters and it refers to the purity of the water analysed.

5.3 RESULTS: STATISTICAL MODELS

5.3.1 Binary Classification

5.3.1.1 Korattur Lake Dataset

Table 5.1 displays the accuracies of the Statistical models - PCA, HCA, LDA and QDA when they were trained with the Korattur Lake Binary Classification Dataset. The classes were either 'drinkable' or 'non-drinkable'.

Statistical Algorithm	**Accuracy**
Principal Component Analysis	0.87
Hierarchical Clustering Analysis	0.53
Linear Discriminant Analysis	0.91
Quadratic Discriminant Analysis	0.95

TABLE 5.1: Binary Class Korattur Lake Dataset Classification using Statistical Models

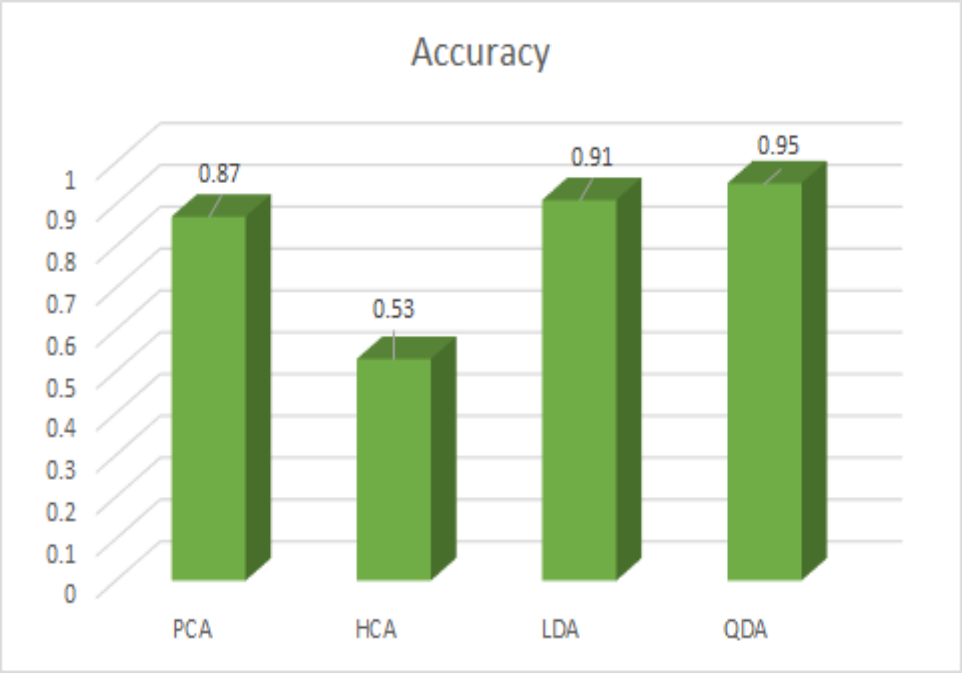

FIGURE 5.3: Accuracy bar plot of Statistical models using Binary Korattur Lake Dataset

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 95% as shown in Figure 5.3.

5.3.1.2 Kaggle Dataset

Table 5.2 below displays the results of the Statistical models - PCA, HCA, LDA and QDA when they were trained with the Kaggle Binary Classification Dataset. The classes were either 'drinkable' or 'non-drinkable'.

Statistical Algorithm	**Accuracy**
Principal Component Analysis	0.88
Hierarchical Clustering Analysis	0.57
Linear Discriminant Analysis	0.88
Quadratic Discriminant Analysis	0.87

TABLE 5.2: Binary Class Kaggle Dataset Classification using Statistical Models

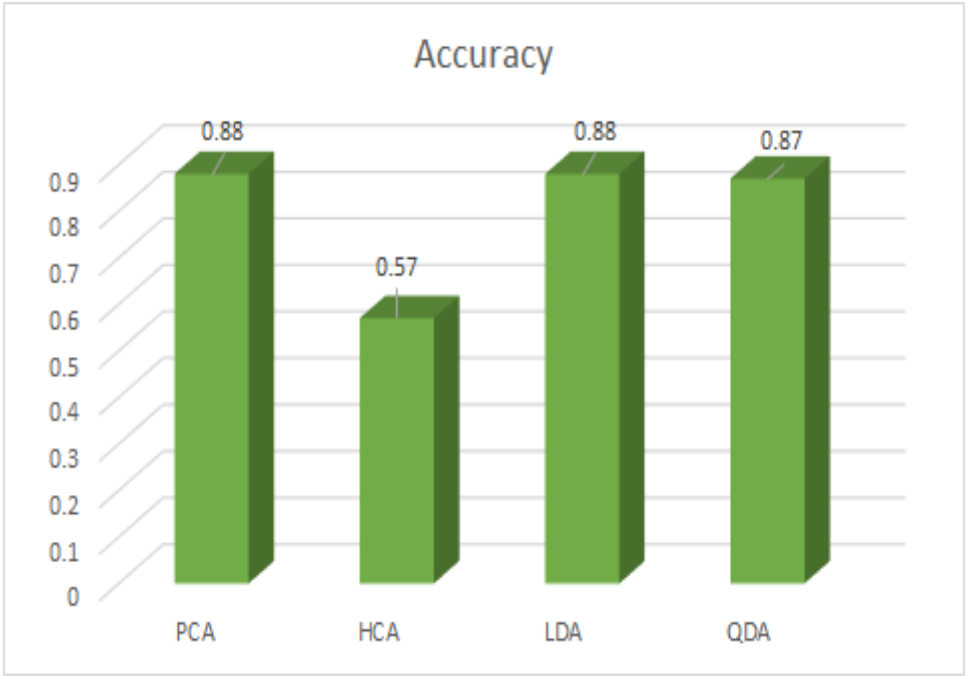

FIGURE 5.4: Accuracy bar plot of Statistical models using Binary Kaggle Dataset

Here, Linear Discriminant Analysis (LDA) is the best algorithm with an accuracy of 88% as shown in Figure 5.4.

5.3.2 Multi Class Classification

5.3.2.1 3 Class Korattur Lake Dataset

Table 5.3 displays the results of the Statistical models - PCA, HCA, LDA and QDA when they were trained with the Korattur Lake Multi Class Classification Dataset with 3 classes. The classes were 'excellent', 'good' or 'poor'.

Statistical Algorithm	Accuracy
Principal Component Analysis	0.75
Hierarchical Clustering Analysis	0.38
Linear Discriminant Analysis	0.92
Quadratic Discriminant Analysis	0.94

TABLE 5.3: 3 Class Korattur Lake Dataset Classification using Statistical Models

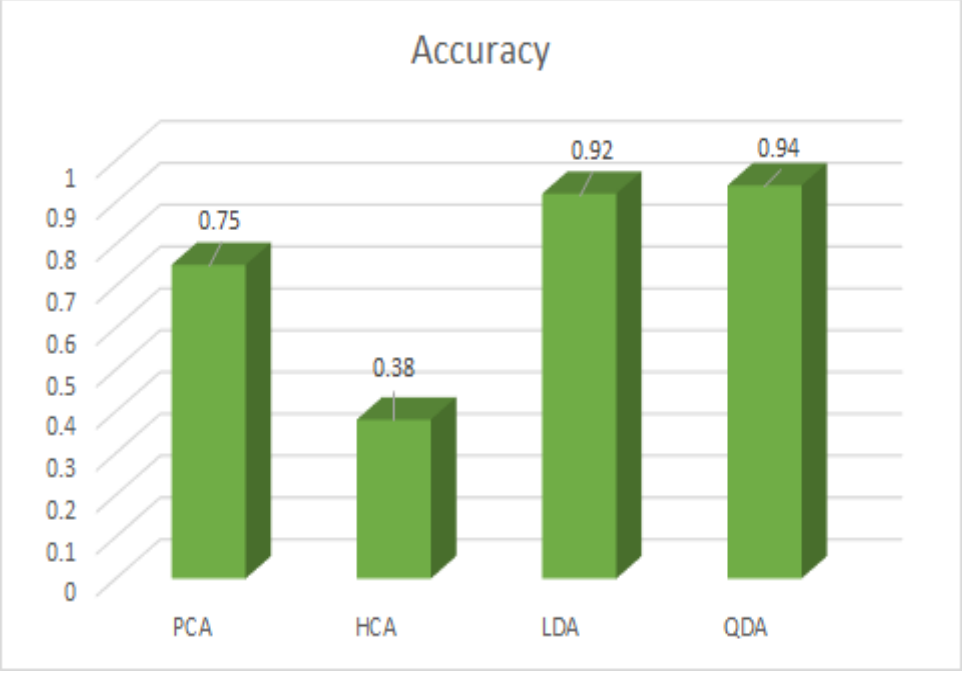

FIGURE 5.5: Accuracy bar plot of Statistical models using 3 Class Korattur Lake Dataset

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 94%, as shown in Figure 5.5.

5.3.2.2 5 Class Korattur Lake Dataset

Table 5.4 displays the results of the Statistical models - PCA, HCA, LDA and QDA when they were trained with the Korattur Lake Multi Class Classification Dataset with 5 classes. The classes were 'excellent', 'good' . 'average', 'bad' or 'poor'.

Statistical Algorithm	**Accuracy**
Principal Component Analysis	0.75
Hierarchical Clustering Analysis	0.17
Linear Discriminant Analysis	0.94
Quadratic Discriminant Analysis	0.97

TABLE 5.4: 5 Class Korattur Lake Dataset Classification using Statistical Models

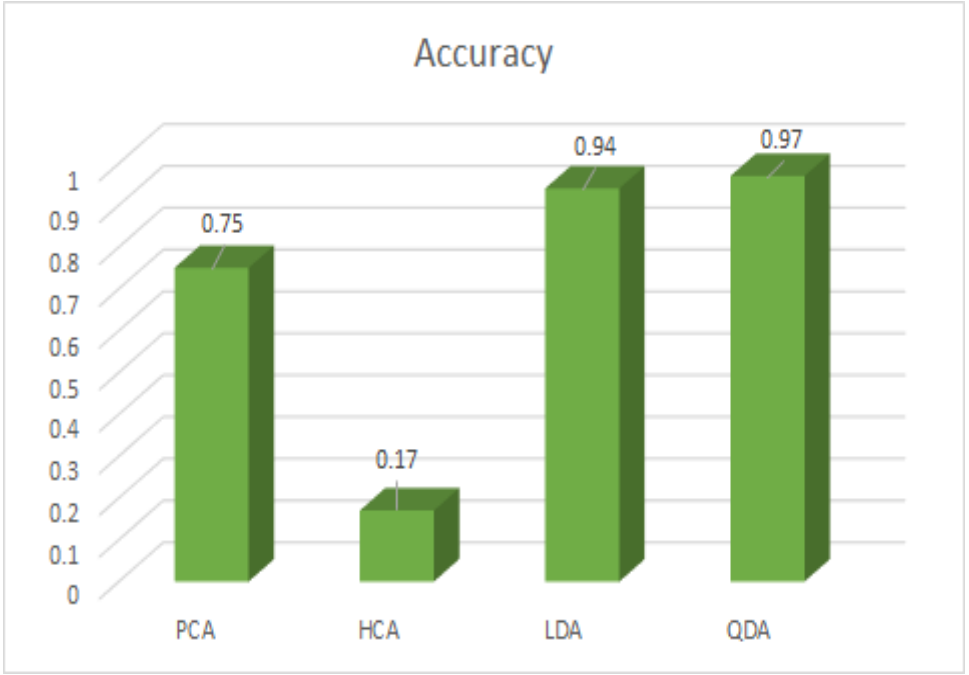

FIGURE 5.6: Accuracy bar plot of Statistical models using 3 Class Korattur Lake Dataset

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 97% as shown in Figure 5.6. Overall, we can see that QDA is the most efficient algorithm from all the Statistical models. So it is chosen as the Statistical Algorithm to be used in the Hybrid Model.

5.4 RESULTS: ENSEMBLE MODELS

5.4.1 Binary Classification

5.4.1.1 Korattur Lake Dataset

Table 5.5 displays the accuracies of the Ensemble models - Bagging, Boosting and Stacking when they were trained with the Korattur Lake Binary Classification Dataset. The classes were either 'drinkable' or 'non-drinkable'.

Ensemble Algorithm	Precision	Recall	F1-Score
Bagging	1.0	1.0	1.0
Boosting	1.0	1.0	1.0
Stacking	1.0	1.0	1.0

TABLE 5.5: Precision, Recall and F1-Score for binary Korattur Lake dataset using Ensemble Models

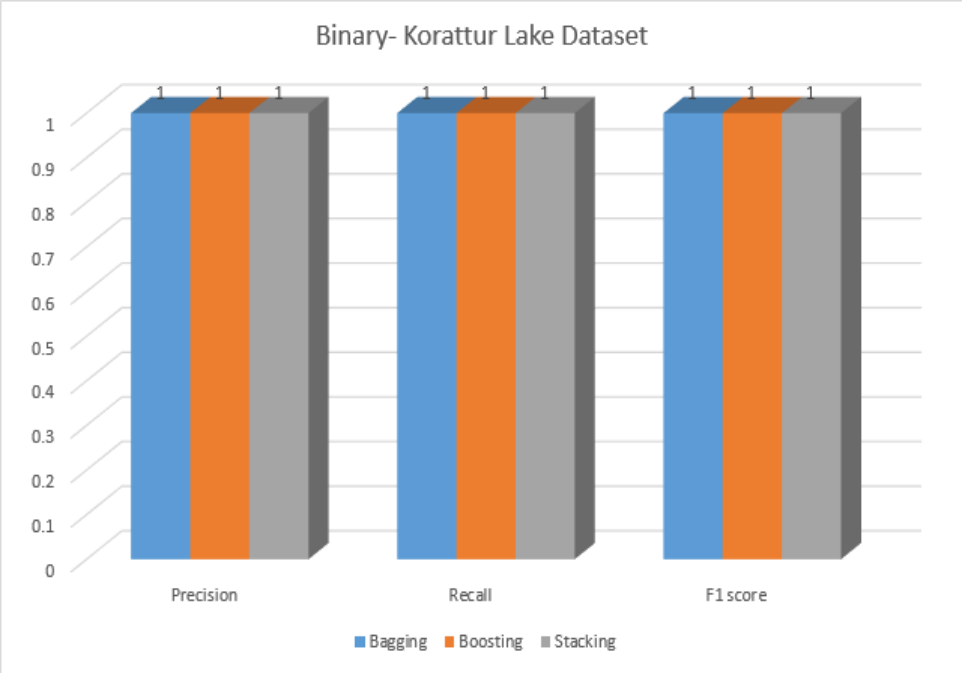

FIGURE 5.7: Visual representation of the above table

Ensemble Algorithm	**Accuracy**	**Time(s)**
Bagging	1.0	0.464
Boosting	1.0	0.081
Stacking	1.0	2.916

TABLE 5.6: Accuracy and time for binary Korattur Lake dataset using Ensemble Models

Here, all three algorithms work best with an accuracy of 100%, as shown in Figure 5.8.

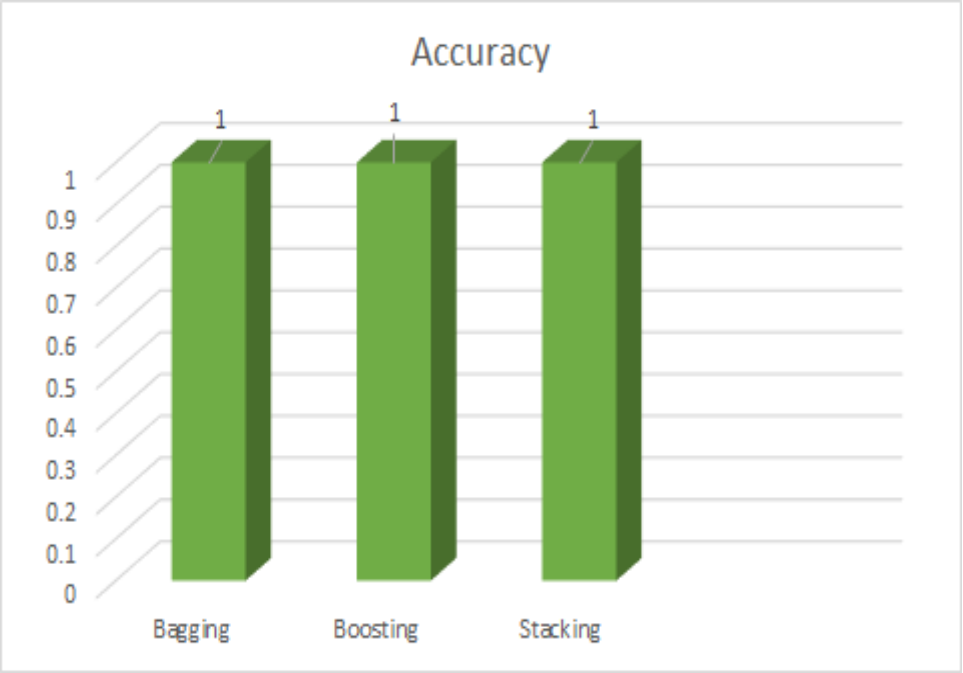

FIGURE 5.8: Accuracy bar plot of Statistical models using Binary Korattur Lake Dataset

5.4.1.2 Kaggle Dataset

Table 5.7 displays the accuracies of the Ensemble models - Bagging, Boosting and Stacking when they were trained with the Kaggle Binary Classification Dataset. The classes were either 'drinkable' or 'non-drinkable'.

Ensemble Algorithm	Precision	Recall	F1-Score
Bagging	0.898	0.785	0.837
Boosting	0.7	0.047	0.089
Stacking	0.916	0.8	0.854

TABLE 5.7: Precision, Recall and F1-Score for binary Kaggle dataset using Ensemble Models

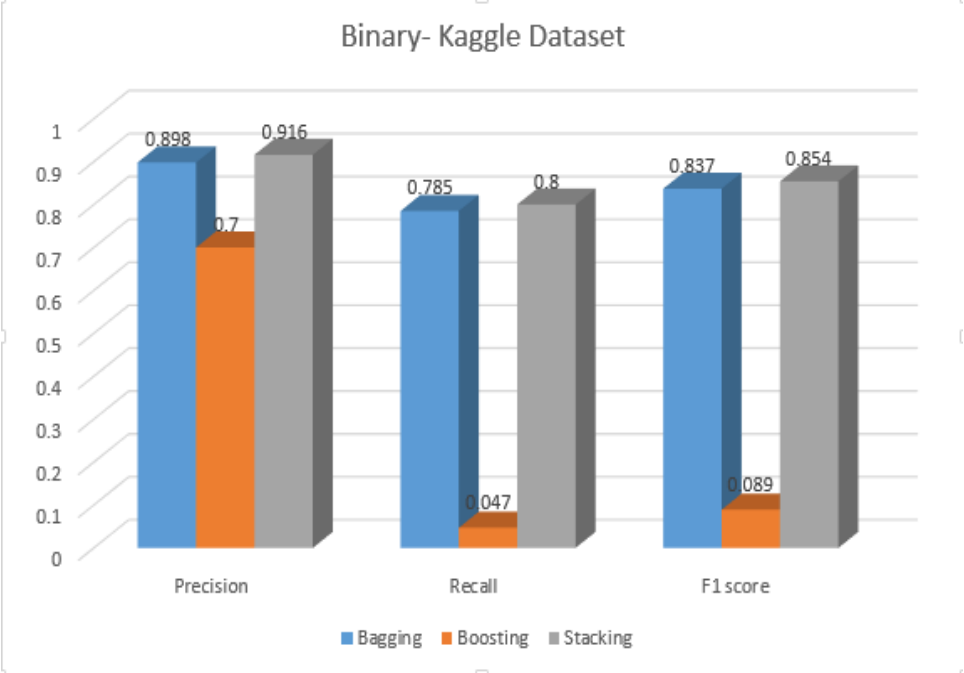

FIGURE 5.9: Visual representation of the above table

Ensemble Algorithm	**Accuracy**	**Time(s)**
Bagging	0.967	6.321
Boosting	0.88	0.114
Stacking	0.96	11.324

TABLE 5.8: Accuracy and time for binary Kaggle dataset using Ensemble Models

Here, the Bagging Algorithm works best with an accuracy of almost 97%, as shown in Figure 5.10.

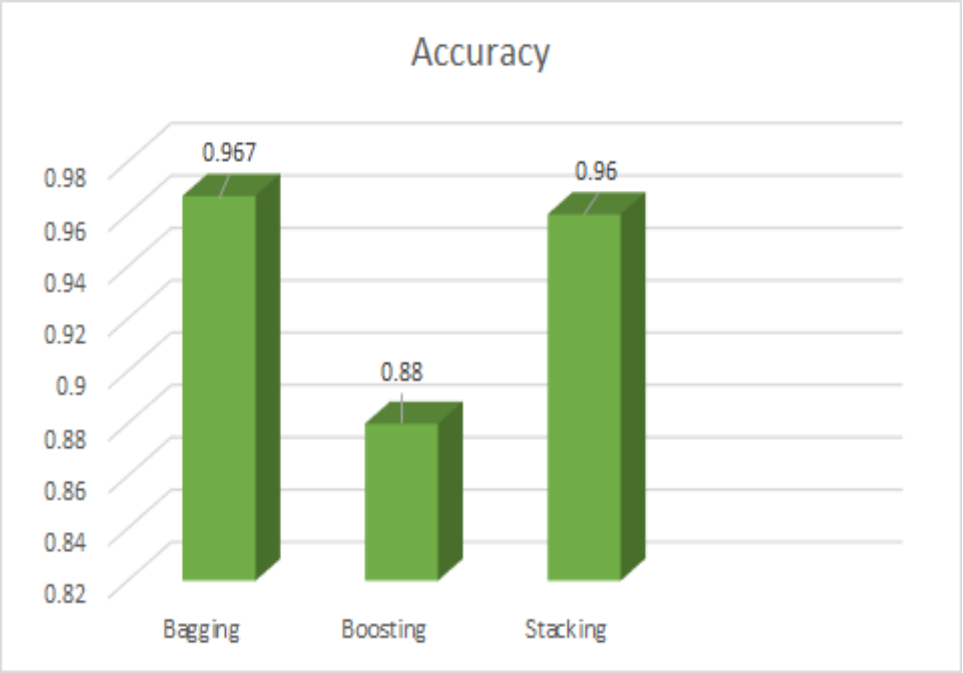

FIGURE 5.10: Accuracy bar plot of Statistical models using Binary Kaggle Dataset

5.4.2 Multi Class Classification

5.4.2.1 3 Class Korattur Lake Dataset

Table 5.9 displays the results of the Ensemble models - Bagging, Boosting and Stacking when they were trained with the Korattur Lake Multi Class Classification Dataset with 3 classes. The classes were 'excellent', 'good' or 'poor'.

Ensemble Algorithm	Precision	Recall	F1-Score
Bagging	1.0	1.0	1.0
Boosting	0.99	0.99	0.99
Stacking	0.99	0.99	0.99

TABLE 5.9: Precision, Recall and F1-Score for 3 class Korattur Lake dataset using Ensemble Models

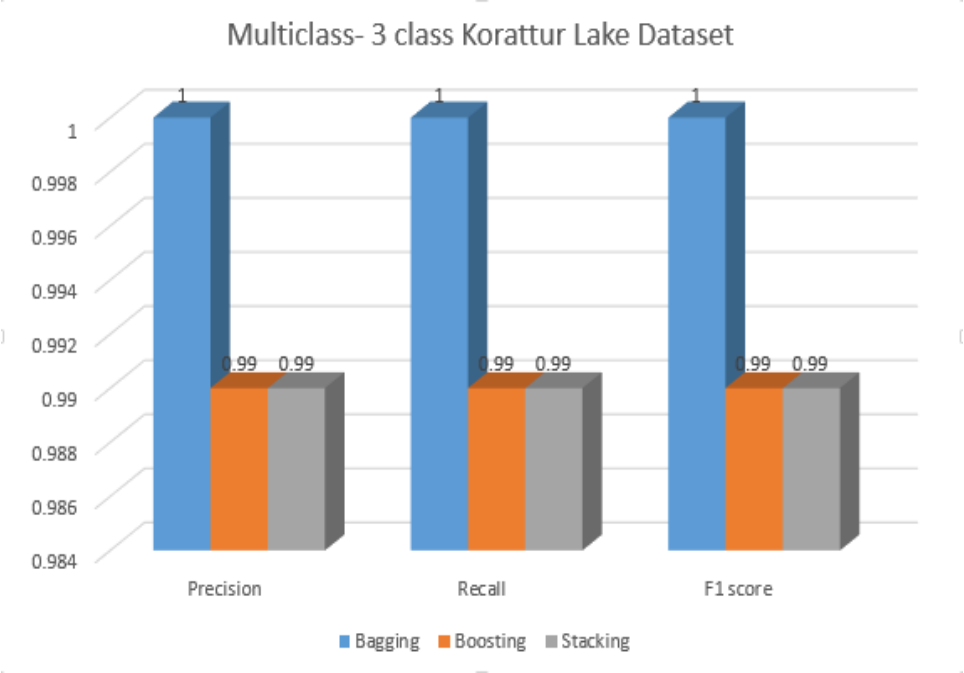

FIGURE 5.11: Visual representation of the above table

Ensemble Algorithm	**Accuracy**	**Time(s)**
Bagging	1.0	2.482
Boosting	0.99	0.101
Stacking	0.99	7.746

TABLE 5.10: Accuracy and time for 3 class Korattur Lake dataset using Ensemble Models

Here, the Bagging Algorithm works best with an accuracy of 100%, as shown in Figure 5.12.

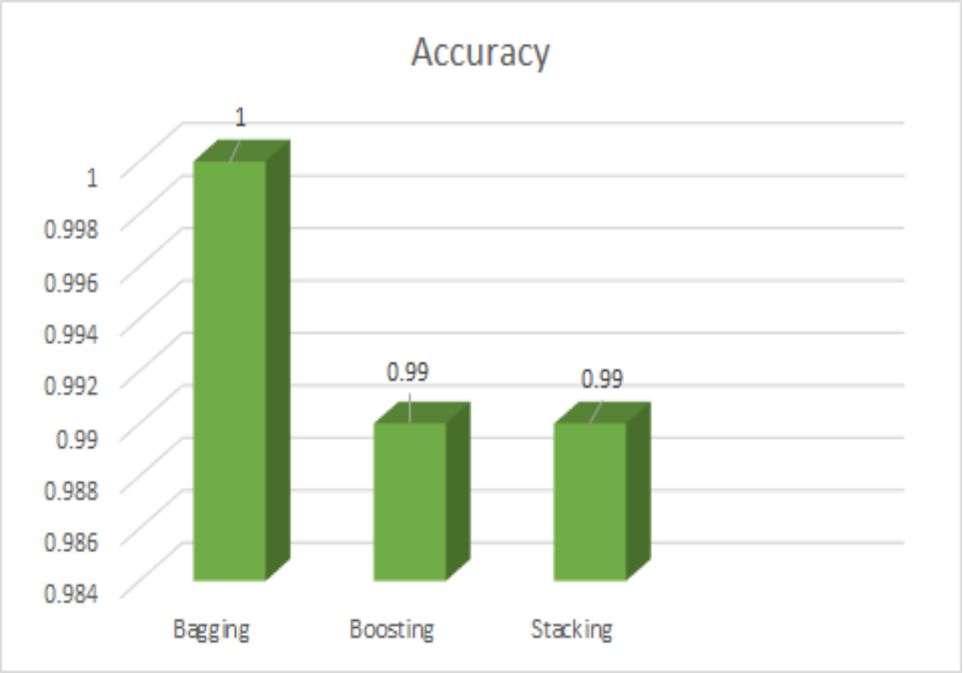

FIGURE 5.12: Accuracy bar plot of Statistical models using 3 Class Korattur Lake Dataset

5.4.2.2 5 Class Korattur Lake Dataset

Table 5.11 displays the results of the Ensemble models - Bagging, Boosting and Stacking when they were trained with the Korattur Lake Multi Class Classification Dataset with 5 classes. The classes were 'excellent', 'good' . 'average', 'bad' or 'poor'.

Ensemble Algorithm	Precision	Recall	F1-Score
Bagging	1.0	1.0	1.0
Boosting	1.0	1.0	1.0
Stacking	0.99	0.99	0.99

TABLE 5.11: Precision, Recall and F1-Score for 5 class Korattur Lake dataset using Ensemble Models

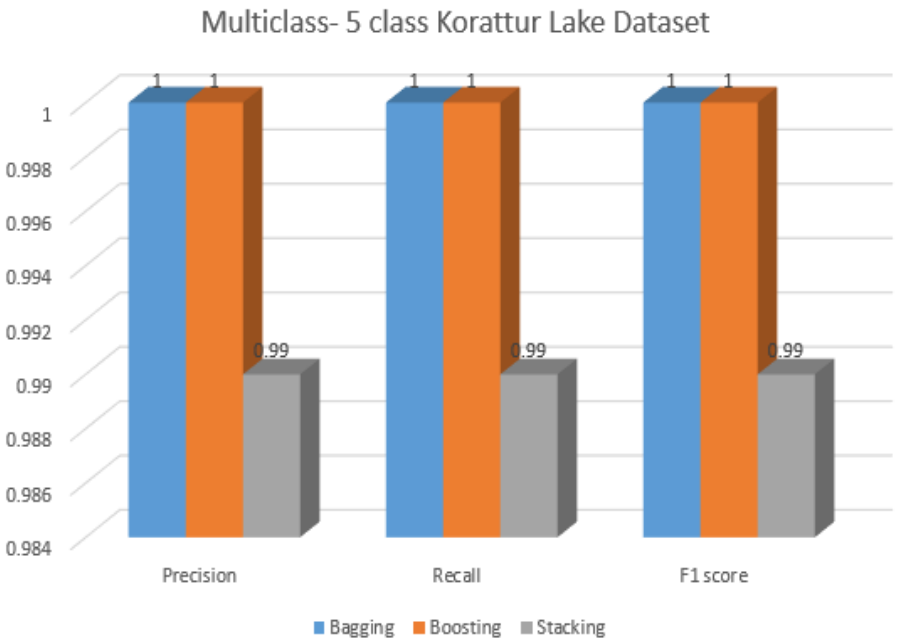

FIGURE 5.13: Visual representation of the above table

Ensemble Algorithm	Accuracy	Time(s)
Bagging	1.0	2.468
Boosting	1.0	0.165
Stacking	0.99	10.444

TABLE 5.12: Accuracy and time for 5 class Korattur Lake dataset using Ensemble Models

Here, both the Bagging and Boosting algorithms work best with an accuracy of 100%,as shown in Figure 5.14.

Overall, Bagging model of the Ensemble algorithms is the most stable when trained with different datasets. Also, it has the best performance when the performance metrics of all the ensemble models are compared. Hence, Bagging is chosen to be used in the Hybrid Model.

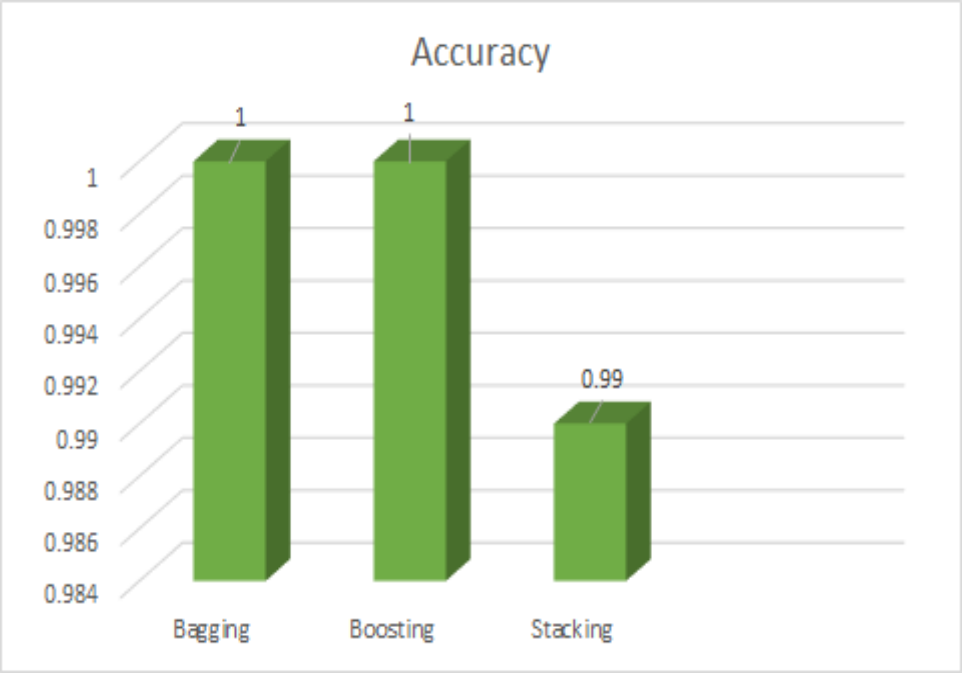

FIGURE 5.14: Accuracy bar plot of Statistical models using 5 Class Korattur Lake Dataset

5.5 RESULTS: HYBRID MODEL

The Hybrid Model is the combination of both the best Statistical as well as Ensemble methods. From the Statistical methods implemented, QDA(Quadratic Discriminant Analysis) was found to be the best performing model. Bagging was found to be the best choice among the implemented Ensemble models. Thus, both the QDA and Bagging models were combined to form the Hybrid model.

The combination of both QDA and Bagging was done 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. There are two types of aggregation in voting classifier - Hard voting and soft voting.

1.Hard Voting is when voting is evaluated based on predicted output class. 2.Soft

Voting is when voting is evaluated based on predicted probability of the output class. In the proposed Hybrid model, hard voting is used.The below diagram describes how the voting ensemble model based on hard voting predicts the output class.

The comparison of results produced by the Hybrid model is shown in Figure 5.16.

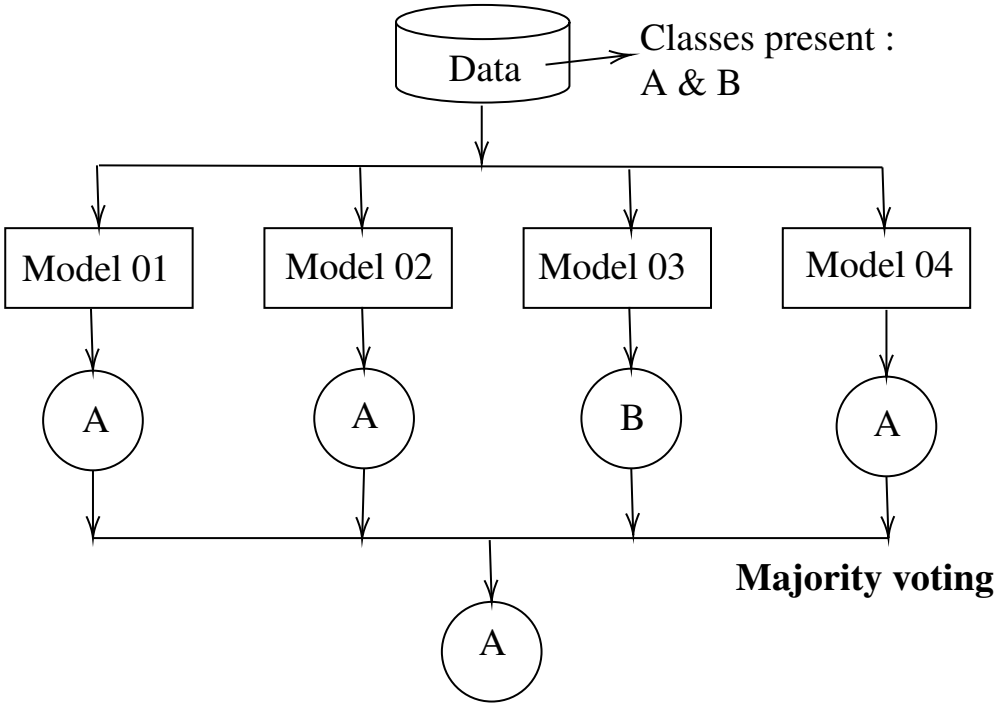

FIGURE 5.15: Working of Hybrid model

5.5.1 Binary Classification

5.5.1.1 Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 99 percent on the Binary Class Korattur Lake Dataset.

5.5.1.2 Kaggle Dataset

The Hybrid Model performs with an accuracy of 96 percent on the Binary Class Kaggle Dataset.

5.5.2 Multi Class Classification

5.5.2.1 3 Class Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 100 percent on the 3 Class Korattur Lake Dataset.

5.5.2.2 5 Class Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 100 percent on the 5 Class Korattur Lake Dataset.

5.6 COMPARISON

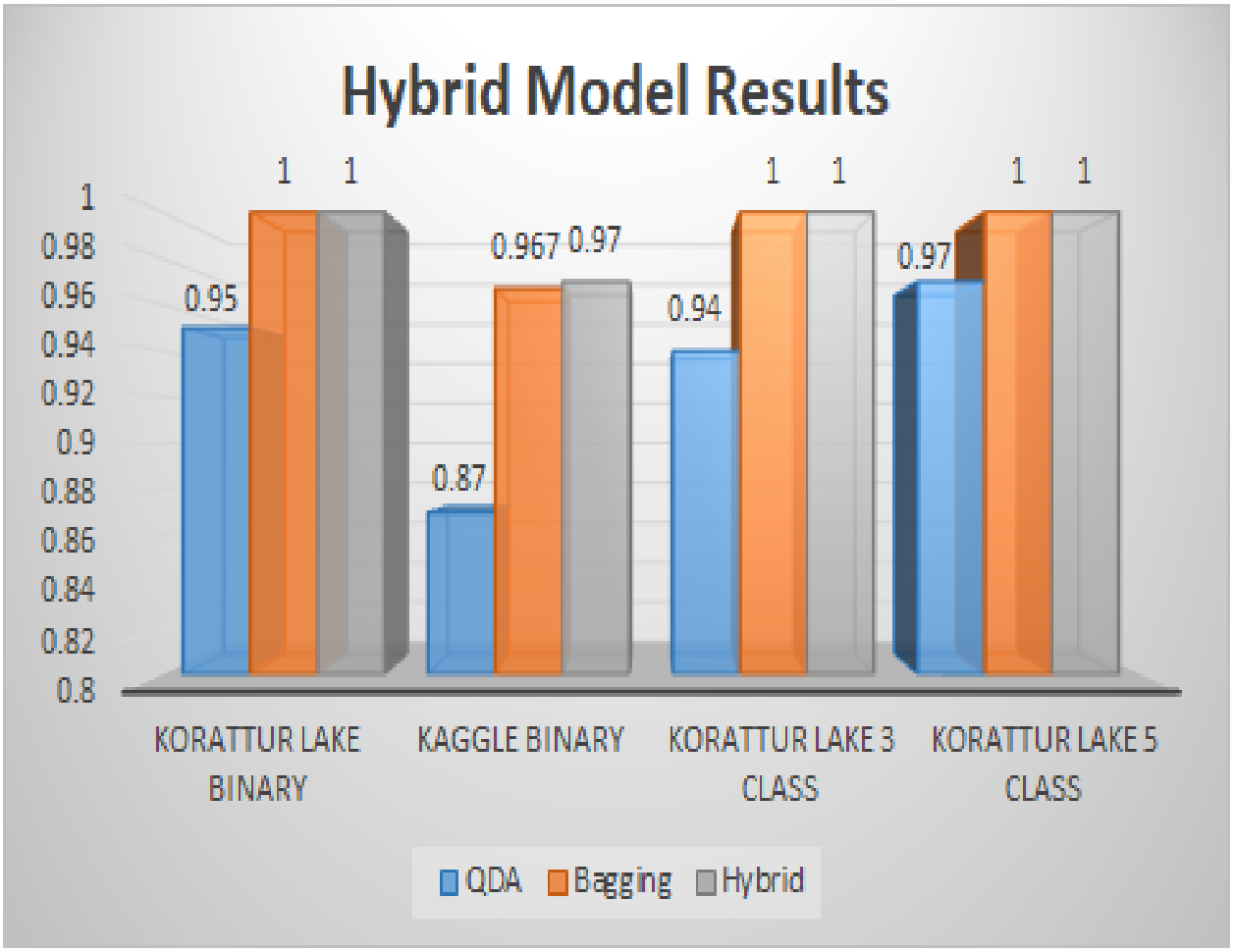

FIGURE 5.16: Accuracy comparison of the Ensemble,Statistical and Hybrid models across all the dastasets

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

To conclude, as discussed in this report, the evaluation and performance comparison of Statistical and Ensemble models were done and a Hybrid model was implemented, combining both. The results of the Hybrid model was compared with both the best performing Statistical and Ensemble models. All the models were trained with the four datasets mentioned in this report.
According to the results of the comparison of the performance of the three models, the Hybrid model is observed to be performing the best, irrespective of the dataset used. Thus, this report has achieved its objective and has built an Hybrid model, with the best results.

As a part of the future work, this research intends to expand its scope by introducing timestamp into the Binary, 3-Class and 5-Class Korattur lake datasets, because the rows of the dataset are ordered by time.As explained earlier in the Exploratory Data Analysis, the Korattur lake datasets are real-world datasets, where the rows are inputed based on real-time values as observed at that particular instance. This research intends to expand the scope of the project by implementing a time-series models such as : MA(Moving Average), ARIMA(Auto Regressive Integrated Moving Average), SARIMA(Seasonal Auto Regressive Integrated Moving Average) using the above timestamped datasets.

REFERENCES

1. Pham Q.B. Saini G. et al Abba, S.I. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environ Sci Pollut Res, 27:41524–41539, 2020.

2. Sani Isah Abba, Nguyen Thi Thuy Linh, Jazuli Abdullahi, Shaban Ismael Albrka Ali, Quoc Bao Pham, Rabiu Aliyu Abdulkadir, Romulus Costache, Van Thai Nam, and Duong Tran Anh. Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration. IEEE Access, 8:157218–157237, 2020.

3. Abobakr Saeed Abobakr Yahya, Ali Najah Ahmed, Faridah Binti Othman, Rusul Khaleel Ibrahim, Haitham Abdulmohsin Afan, Amr El-Shafie, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, and Ahmed Elshafie. Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. Water, 11(6), 2019.

4. Ali Omran Al-Sulttani, Mustafa Al-Mukhtar, Ali B. Roomi, Aitazaz Ahsan Farooque, Khaled Mohamed Khedher, and Zaher Mundher Yaseen. Proposition of new ensemble data-intelligence models for surface water quality prediction. IEEE Access, 9:108527–108541, 2021.

5. Rodelyn Avila, Beverley Horn, Elaine Moriarty, Roger Hodson, and Elena Moltchanova. Evaluating statistical model performance in water quality prediction. Journal of Environmental Management, 206:910–919, 2018.

6. Rahim Barzegar, Asghar Asghari Moghaddam, Jan Adamowski, and Bogdan Ozga-Zielinski. Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Stochastic environmental research and risk assessment, 32(3):799–813, 2018.

7. Xingguo Chen, Houtao Liu, Xiuying Xu, Luoyuan Zhang, Tianchi Lin, Min Zuo, Yichao Huang, Ruqin Shen, Da Chen, and Yongfeng Deng. Identification of suitable technologies for drinking water quality prediction: A comparative study of traditional, ensemble, cost-sensitive, outlier detection learning models and sampling algorithms. ACS ES&T Water, 1(8):1676–1685, 2021.

8. Venkata Vara Prasad D, Lokeswari Y Venkataramana, P. Senthil Kumar, Prasannamedha G, Soumya K., and Poornema A.J. Water quality analysis in a lake using deep learning methodology: prediction and validation. International Journal of Environmental Analytical Chemistry, 0(0):1–16, 2020.

9. Venkata Vara Prasad D, Lokeswari Y Venkataramana, P. Senthil Kumar, Prasannamedha G, Soumya K., and Poornema A.J. Prediction on water quality of a lake in chennai, india using deep learning algorithms. Desalination and Water Treatment, 218:44–51, 2021.

10. Gozen Elkiran, Vahid Nourani, and S.I. Abba. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. Journal of Hydrology, 577:123962, 2019.

11. Arshia Fathima, J Alamelu Mangai, and Bharat B Gulyani. An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques. International journal of river basin management, 12(4):357–366, 2014.

12. FARID HASSANBAKI GARABAGHI, Semra Benzer, and Recep Benzer. Performance evaluation of machine learning models with ensemble learning approach in classification of water quality indices based on different subset of features. 2021.

13. Li-Ming (Lee) He and Zhen-Li He. Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern california, usa. Water Research, 42(10):2563–2573, 2008.

14. Zengrui Huang, Wei Mao, Ming Chen, Qiang Wu, Boyue Xiong, and Wei Xu. An intelligent operation and maintenance system for power consumption based on deep learning. IOP Conference Series: Materials Science and Engineering, 486:012107, 07 2019.

15. Y Khan and SS Chai. Ensemble of ann and anfis for water quality prediction and analysis-a data driven approach. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-9):117–122, 2017.

16. Ozgur Kisi, Meysam Alizamir, and AliReza Docheshmeh Gorgij. Dissolved oxygen prediction using a new ensemble method. Environmental Science and Pollution Research, 27(9):9589–9603, 2020.

17. Lingbo Li, Jundong Qiao, Guan Yu, Leizhi Wang, Hong-Yi Li, Chen Liao, and Zhenduo Zhu. Interpretable tree-based ensemble model for predicting beach water quality. Water Research, 211:118078, 2022.

18. Zilin Li, Chi Zhang, Haixing Liu, Chao Zhang, Mengke Zhao, Qiang Gong, and Guangtao Fu. Developing stacking ensemble models for multivariate contamination detection in water distribution systems. Science of The Total Environment, 828:154284, 2022.

19. Shuangyin Liu, Haijiang Tai, Qisheng Ding, Daoliang Li, Longqin Xu, and Yaoguang Wei. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling, 58(3):458–465, 2013. Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture 2012.

20. Al-Mahfoodh Najah, Ahmed El-Shafie, Othman A Karim, Othman Jaafar, and Amr El-Shafie. An application of different artificial intelligences techniques for water quality prediction. International Journal of Physical Sciences, 6, 10 2011.

21. Navideh Noori, Latif Kalin, and Sabahattin Isik. Water quality prediction using swat-ann coupled approach. Journal of Hydrology, 590:125220, 2020.

22. Jungsu Park. "the effect of input variables clustering on the characteristics of ensemble machine learning model for water quality prediction.". Journal of Korean Society on Water Environment, 37(5):335–43, September 30, 2021.

23. Jungsu Park, Woo Hyoung Lee, Keug Tae Kim, Cheol Young Park, Sanghun Lee, and Tae-Young Heo. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of The Total Environment, 832:155070, 2022.

24. Sanghyun Park, Kyunghyun Kim, Changmin Shin, Joong-Hyuk Min, Eun Hye Na, and Lan Joo Park. Variable update strategy to improve water quality forecast accuracy in multivariate data assimilation using the ensemble kalman filter. Water Research, 176:115711, 2020.

25. Rosaida Rosly, Mokhairi Makhtar, Mohd Khalid Awang, and Nordin Abdul. Comparison of ensemble classifiers for water quality dataset.

26. Dipankar Ruidas, Subodh Chandra Pal, Towfiqul Islam, Abu Reza Md, and Asish Saha. Hydrogeochemical evaluation of groundwater aquifers and associated health hazard risk mapping using ensemble data driven model in a water scares plateau region of eastern india. Exposure and Health, pages 1–19, 2022.

27. Chenguang Song and Leihua Yao. A hybrid model for water quality parameter prediction based on ceemdan-ialo-lstm ensemble learning. Environmental Earth Sciences, 81(9):1–14, 2022.

28. Leizhi Wang, Zhenduo Zhu, Lauren Sassoubre, Guan Yu, Chen Liao, Qingfang Hu, and Yintang Wang. Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Science of The Total Environment, 765:142760, 2021.

29. Yunrong Xiang and Liangzhong Jiang. Water quality prediction using ls-svm and particle swarm optimization. pages 900–904, 2009.