Water Quality Prediction using Statistical, Machine Learning and hybrid models

Shriya B
Vikram V
Vyshali S

SSN College of Engineering, Chennai

October 26, 2021

Introduction

- Water- Essential to sustain life.
- India- only 4 percentage of freshwater resources.
- Tamilnadu- only 2.5 percent out of the 4
- Water contamination is getting alarming- oil spills, industrial wastes, chemical fertilizers, pesticides, etc.
- Water quality analysis- done before using ML,DL, Auto-ML, Auto-DL,etc
- Our objective- Overcome shortcomings
- Statistical models, Machine Learning models and a hybrid model

Problem Statement

- The proposed system will use a combination of statistical and Machine learning models.
- Real-world data is generally incomplete, inconsistent and noisy.
- The statistical model- pre-processes the data set
- Then, the Machine Learning model-predicts the water quality of the water sample.
- The results of the conventional statistical, machine learning models and the hybrid model is compared.

Justification for problem statement

- ML, DL, Auto-ML and Auto-DL have previously been used to analyse water quality.
- Statistical models- not found much
- Our approach: Analyse the results of an ML model, a statistical model, and a hybrid of the two.
- Hybrid: Data pre-processing done using statistical models like min-max, Z-score, etc and the prediction to be done using Random Forest.
- The exact statistical model is yet to be decided because not much of research is available. Need to test before deciding the best suit for the project

Literature Survey and Feasibility Study

Paper title	Methodology	Limitations
Statistics-Based Data Preprocessing Methods and Machine Learning Algorithms for Big Data Analysis	compared the pre-processing methods decimal scaling, min-max normalization and z-score normalization statistical techniques. It was found that z-score technique worked the best as the size of the dataset increases. The research also compared machine learning models such as HMCM(Hidden Markov chain model), Box-Cox Transformation and Linear Transformation out of which HMCM(Hidden Markov chain model) has the highest accuracy rate, prediction ability and utility. It shows that HMCM is the best statistical machine learning algorithm to use for big data analysis	A high dimensional data analysis technique needs to be integrated along with this model to find why this model performed the best
Pre-processing and audit of power consumption data based on composite mathematical statistics model	designed a pre-processing model based on mathematical statistics methods using quartile detection and Z-score method. The original data obtained from electricity consumption in Shanghai is abnormal and has many errors. The statistical methods extract required data, expel abnormal data and structure the dataset thus increasing the quality of the dataset.	The model can be developed more by summarizing more features so that data classification in pre processing stage is easier and precise.

Literature Survey and Feasibility Study

Paper title	Methodology	Limitations
Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model	developed two machine learning models - the long short-term memory(LSTM) and Convolutional Neural Network(CNN) along with two deep learning models - support vector regression(SVR) and decision tree(DT). A coupled CNN-LSTM model was also developed. According to the statistical metrics LSTM outperformed CNN for DO prediction and DL models yielded similar results for Cha-a prediction. The hybrid model - integration of LSTM and CNN models outperformed both ML and DL standalone models in DO and Cha-a prediction. The dataset was from Small Prespa Lake, Greece. Statistical metrics such as correlation coefficient, root mean square error(RMSE), mean absolute error(MSE) and others were used to assess the performance of the models. The hybrid model captured both low and high levels of water quality variables for DO concentrations.	NO limitations cited
Prediction on water quality of a lake in Chennai, India using machine learning algorithms	used different machine learning models such as decision tree, random forest, logistic regression, support vector machine and navie bayesian network for binary and multi class classification. Random forest was the beat performing model among them with a accuracy of 96 percent. The dataset was acquired from Korattur Lake, Tamil Nadu.	The machine learning models handle less data than what a hybrid model or a deep learning model would use. Increase in the dataset will lead to better results. So, there is a need for deep learning or hybrid models.

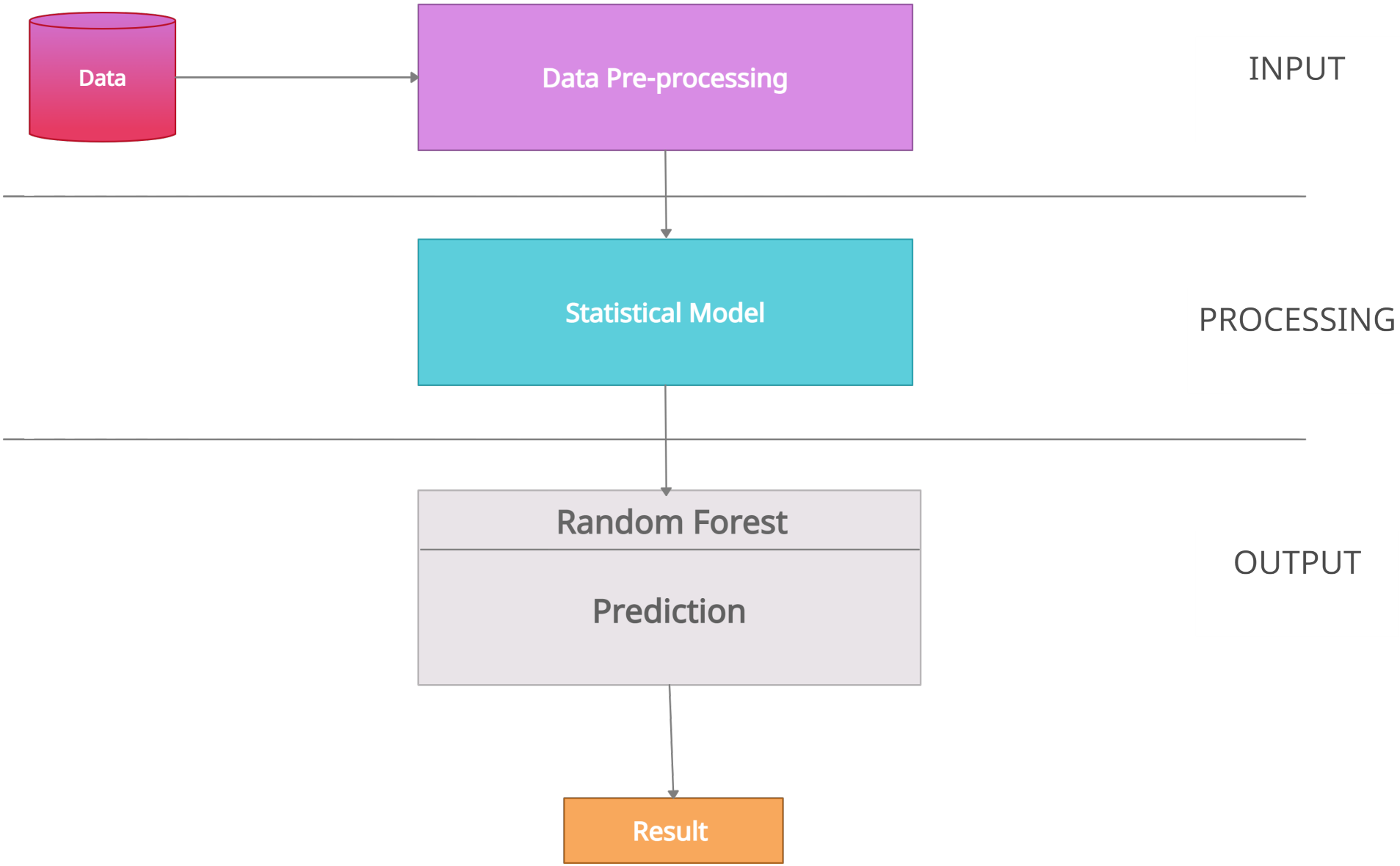
Literature Survey and Feasibility Study

Paper title	Methodology	Limitations
Improving water quality index prediction in Parak River basin Malaysia through a combination of multiple neural networks	Proposed multiple neural networks. Proved to be more accurate in classifying data set into its WQC(water quality class) with an accuracy of 85 percent. FANN(Feed Forward Artificial Neural Network) provided non-robust nature of prediction while MNN(multiple neural network) introduced robustness into the picture. The values of the coefficient of determination and MSE prove that polynomial regression and gradient boosting helped increase accuracy.	The metrics only showed slight changes between the models of FANN and MNN
Machine learning approaches for anomaly detection of water quality on a real-world data set	used machine learning and deep learning techniques such as SVM(Support Vector Machine), LDA(Linear Discriminant Analysis), LSTM(Long Short-Term Memory), DNN(Deep Neural Network), ANN(Artificial Neural Network), RNN(Recurrent Neural Network). They predicted water quality by proposing a new model least squares support vector machine(LS-SVM). The dataset was from Liuxi river, Guangzhou from which eight features were used.	This paper has shown that SVM does not work well on time series dataset. But after scaling the dataset, the prediction of SVM classifier has improved. It shows that there is a need to find better models for imbalanced datasets since all real world data are imbalanced

Proposed System

- Initially, results obtained from Water quality analysis
- Using: machine learning model (Random Forest) and a statistical model (such as PCA, CA, DA etc)
- Later, we device a hybrid system combining both the models
- Data pre-processing unit processes the data set based on several parameters using a statistical model
- Models to be tried: Z-score normalization, min-max normalization, decimal scaling and quartile detection method
- Later we use a machine learning model (Random Forest) that would predict the water quality.
- Compare the results of all the three and see if the hybrid model has good performance instead of individual models.

Hybrid model Architecture



Random Forest

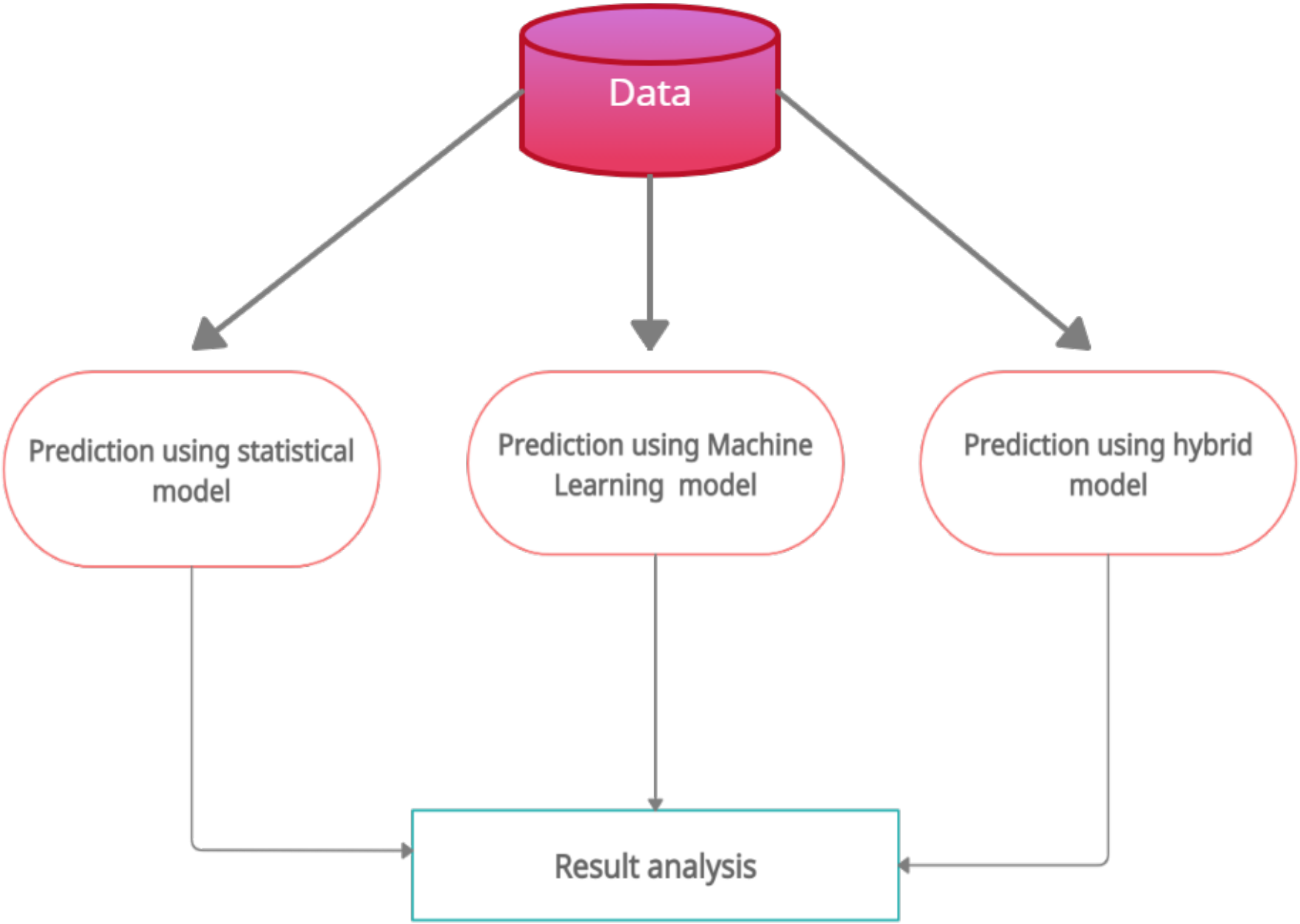
- Random Forest - Supervised Machine Learning Algorithm
- Solving problems on classification and regression
- The majority vote after building decision trees on different samples is taken for classification.
- can handle categorical values and performs better for classification rather than regression.
- example for bagging type of ensemble technique
- creates a different training subset from sample training data with replacement the final output is based on majority voting.
- Main limitation - large number of trees can make the algorithm too slow - ineffective for real-time predictions.

PCA,CA,DA

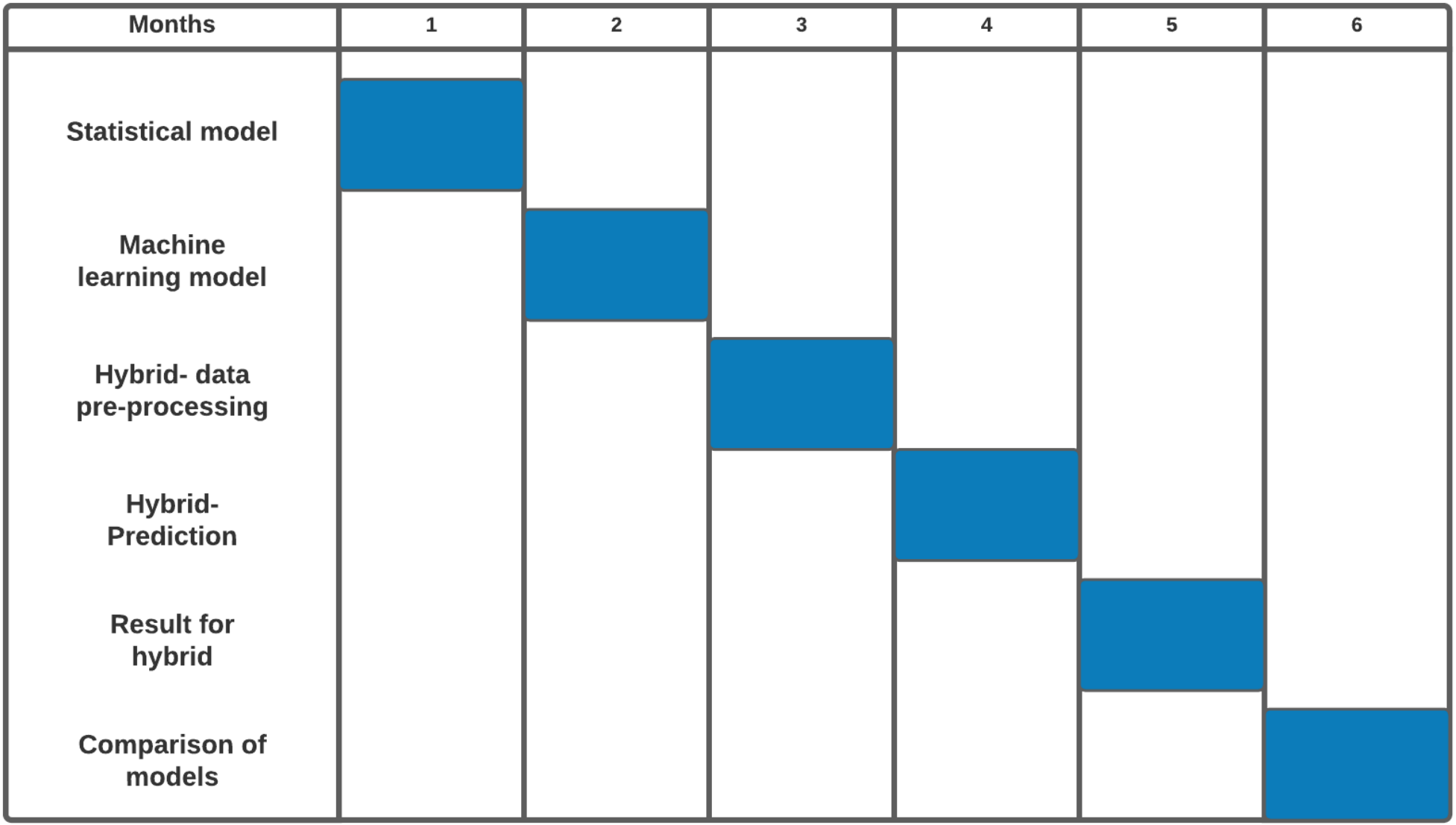
- Principal component analysis (PCA) - technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.
- Important use of PCA - represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers.
- Cluster analysis (CA) - task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- classify different objects into groups in such a way that the similarity between two objects is maximal if they belong to the same group and minimal otherwise.
- Discriminant analysis (DA) - method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events.
- Performs a multivariate test of differences between groups

Modules split-up

- Our project comprise of four major modules:
  1. ML module
  2. Statistical module
  3. Hybrid module
  4. Result analysis


Data
Prediction using statistical model
Prediction using Machine Learning model
Prediction using hybrid model
Result analysis

Project Timeline



Thank You