ITC 1/51
Information Technology and Control
Vol. 51 / No. 1 / 2022
pp. 59-77
DOI 10.5755/j01.itc.51.1.29467

Data Analytics and Reporting API – A Reliable Tool for Data Visualization and Predictive Analysis

Received 2021/07/16 | Accepted after revision 2022/01/13

http://dx.doi.org/10.5755/j01.itc.51.1.29467

HOW TO CITE: Ignatius, J. L. P., Selvakumar, S., JSN, S., Govindarajan, S. (2022). Data Analytics and Reporting API – A Reliable Tool for Data Visualization and Predictive Analysis. *Information Technology and Control*, 51(1), 59-77. https://doi.org/10.5755/j01.itc.51.1.29467

Data Analytics and Reporting API – A Reliable Tool for Data Visualization and Predictive Analysis

Joe Louis Paul Ignatius, Sasirekha Selvakumar
Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering,
Rajiv Gandhi Salai (OMR), Kalavakkam-603110, Chennai, Tamil Nadu, India

Spandana JSN, Subasri Govindarajan
Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering,
Rajiv Gandhi Salai (OMR), Kalavakkam-603110, Chennai, Tamil Nadu, India

Corresponding author: joelouisi@ssn.edu.in

Data analytics, the science of analyzing raw data is widely being used in a broad range of applications to make effective, efficient and timely decisions in the real-time prediction problems. Analytics requires the user to have an in-depth knowledge about the various steps involved in the process for arriving at the conclusions. Hence, to make it easier for the naïve users, Data Analytics and Reporting API (DARAPI) provides Analytics as a Service in the form of an Application Programming Interface (API) has been proposed. DARAPI is aimed to ease the process of analyzing the data by the users without requiring their expertise in the field. It accepts the input file from the user and performs pre-processing techniques at various stages including imputation of missing data, detection and replacement of outliers, encoding of categorical variables, and etc. Furthermore, feature engineering is performed, based on which DARAPI will execute the different classification/regression models and finally delivers the model which provides the best accuracy for future predictions. This entire system is rendered to the user in the form of an API that can be called from any device that is Internet enabled. DARAPI stands unique with its embedded feedback mechanism which generates constructive input for future predictions. This feature enhances the performance of the system in comparison with the existing tools. DARAPI proves beneficial not only to the naïve users but also to the experts by saving their time and efforts needed in understanding the data.

KEYWORDS: DARAPI, data analytics, feature engineering, machine learning, reporting API.

1. Introduction

Over the last few decades, data analytics and machine learning have become buzzwords in the IT world. The majority of businesses have refocused their efforts and are investing heavily on machine learning and data analytics. Although technology is rapidly advancing, it is difficult to master and can produce inaccurate results if it is being utilized incorrectly [24]. Intelligent software and devices, on the other hand, will gradually take over the technology sector as time and technology mature. Analytics as a Service is offered to assist developers in integrating machine learning and data analytics into their products.

Big data analytics is a growing research area in computer science and a variety of other fields around the globe. It has had a lot of success in a wide range of applications. This includes topics such as social media, economy, finance, healthcare, agriculture, and etc. To deliver big data predictive analytics solutions, several machine learning approaches are used [1]. A reporting API is a general reporting method where web apps use to make reports available in a consistent manner based on numerous platform features and to make appropriate decisions based on the prediction [20].

The science of getting computers to learn and act based on data, without being programmed explicitly is known as machine learning. Data analytics refers to the process of using prior data to gain insights into data [27-28].

The primary goal of DARAPI is to provide data analytics interface for the amateur users to work with ease. Though, there are various tools available for the process of analysing and visualizing data, they require prior knowledge of the user to operate the tool and to decipher the results. In addition, the rigidity in usage of the existing tools poses a difficulty. As a result, to aid developers to integrate machine learning and data analytics into their products, a tailor-made Data Analytics and Reporting API (DARAPI) is proposed with the goal of efficiently utilising the time and resources. Through its intuitive usability and precise results, this aids naïve and expert users in analysing and acquiring insights about the data.

From this research, the main contributions are as follows,

1. Most of the business organizations require skilled professionals for data analytics to make decisions. Hence, to enhance and improve the productivity of the business organizations, a Data Analytics tool (DARAPI) is proposed that does not demand skilled professionals.
2. Though some data analytics tools provide easy user interface for the users, this is limited by the technical support price. Thus, to enable amateur users to master the functionalities of the tool, a simple, cost-free and easy-to-understand user interface is designed.
3. The users are restricted from customizing the functioning of the data analytic tool for their specific needs and they are not given flexibility in choosing the target variable. Hence, the system with a number of data pre-processing tools, which allow the users to modify and rearrange the data to meet their needs and to give freedom to the users in selecting the target variable and the features of their choice to build the model, is designed.

2. Related Works

Data analytics is rapidly becoming a vital tool in numerous areas, including healthcare, robotics, and many others. There are four forms of data analytics: descriptive, diagnostic, predictive, and prescriptive analysis. Prescriptive analysis is used in health care to achieve the best results and make better judgments. Muneeswaran et al. [19] discussed that big data enables data analysts to acquire and retain data from resources more efficiently. Sarker [30] suggested that the digital world has a huge amount of data in this age of the Fourth Industrial Revolution (4IR or Industry 4.0), such as Internet of Things (IoT) data, cyber security data, mobile data, business data, social media data, health data, and so on. Pathak et al. [23] discussed both the principles and the tools employed in the field of data science and data analytics.

Wieringa et al. [35] suggested that data can be viewed as the new lifeline for the economy, yet privacy issues limit its usage, leading to a common perception that data analytics and privacy are contradictory. Moorthi et al. [18] studied that the data's characteristics are growing and changing day by day. Delen and Demirkan [4] discussed that many firms are turning to service-oriented decision support systems (DSS on the cloud) to help them become more agile where Service Oriented Architecture (SOA) helps to commoditize business processes, architectures and infrastructures.

In the era of e-learning platforms, data analytics techniques in education give unrivalled potential to help students. In this context, Elfeky and Elbyaly [7] proposed a learning management system (LMS) as a platform allowing the learners to record their activities in order to investigate data types that are specific to the academic environment and improve their learning experience. In various fields of data application, data analytics has shown its value in knowledge discovery and decision assistance. Zulkernine et al. [36] proposed a conceptual architecture for Cloud-based Analytics-as-a-Service (CLAaaS), a big data analytics service provisioning platform in the cloud, based on the taxonomy and an analysis of existing analytic software and systems. A machine learning algorithm (MLA) is a method or tool for assisting with big data analytics (BDA) of applications. Rahul et al. [26] examined the various challenges that machine learning tools and technologies face, as well as the present state of industry adoption. Kumar et al. [16] focused on the use of several parameters in machine learning approaches like k-nearest neighbours (KNN), Nave Bayes, support vector machine (SVM), decision trees, and random forest.

Chong et al. [3] proposed machine learning models to find the best combination of feature subsets and p1rediction algorithms for predicting activity class from hip-based raw acceleration data. Hodge and Austin [10] discussed the importance of detecting and dealing with the outliers and other applications in which outlier detection plays a significant role in providing insights about the multiple technologies existing for handling the outliers. Jijo and Abdulazeez [12] discussed a detailed approach based on supervised machine learning algorithms that use externally supplied examples to form broad hypotheses and then make predictions about future cases.

Khalid et al. [13] investigated certain commonly used feature selection and feature extraction approaches to see how well they can be utilised to improve the performance of learning algorithms and, as a result, the predicted accuracy of classifiers. There are various ways for assessing the vast stream of data, as well as the whole life cycle of big data analysis and many practical applications for obtaining, processing, and analysing this massive data. Balusamy et al. [2] suggested many sorts of data analysis techniques that can be used in data analytics and explained the methodologies used in big data analysis, such as quantitative and qualitative analysis, as well as different forms of statistical analysis including A/B testing, correlation, and regression.

Sumbaly et al. [31] discussed the performance of decision tree data mining technique for the breast cancer diagnosis problem using the Winconsin Breast Cancer Dataset (WBCD). The findings demonstrate that J48 classifiers with feature selection are a superior approach for breast cancer detection with accuracy of 94.5637 %. In another research on the WBC dataset, Ojha and Goel [21] investigated the performance of several clustering and classification techniques. The results indicate that classification algorithms outperform clustering methods in predicting the outcomes. The decision tree (C5.0) and SVM were the best predictors on the holdout sample, with 81 % accuracy whereas c-means had the lowest accuracy rate 37% among the algorithms.

Salama et al. [29] discussed the fusion of Multilayer Perceptron (MLP) and Decision Tree J48 classifiers with features selection using WEKA data analytics tool for the WBC datasets which outperforms the other classifiers for predictive analysis. Uzut and Buyrukoglu [34] proposed the prediction of real estate prices using data mining algorithms and compare multiple data mining algorithms for real estate price prediction, such as linear regression, random forest, and gradient boosting. The results indicate that gradient boosting algorithm provided the highest performance for the real estate price prediction.

Dutta et al. [6] discussed the cost prediction of health insurance for the general public using different regression methods such as Decision Tree Regression, Random Forest, Linear Regression, and Polynomial Regression. By comparing all these results, the best model was found to be Random Forest regression with 0.862533 Coefficient of determination (R2_score), 2.12 Mean Squared Error (MSE) score and 4604.86 Root Mean Squared Error (RMSE) scores. According to the previous works, the required features for the training phase are not chosen based on the impact percentage of the features on the data set. Table 1 discusses the strengths and limitations of some of the related works.

Table 1
Strengths and limitations of some of the related works

Reference	Methodology	Strengths	Limitations
Sumbaly et al. [31]	– Decision Tree Data Mining Technique (J 48)	– Breast cancer detection with accuracy of 94.5637%. – PCA is used to transform possible correlated variables into uncorrelated variables.	– All the data sets utilised here are conventional, other raw data sets can be used. – Deciding an appropriate number of principal components to represent feature space is critical.
Ojha and Goel [21]	– Clustering algorithms (K-means, EM, PAM and Fuzzy c-means) & Classification algorithms (SVM, C5.0 Decision Tree, KNN and Naïve Bayes)	– Classification algorithms are better predictor than clustering algorithms. – The classification algorithms, C5.0 and SVM have shown 81% accuracy (better than clustering algorithms).	– A feature selection strategy such as Correlation-based Feature Selector can be used to minimize the computation overhead of huge data. – Classification models and clustering models evaluated on different datasets.
Salama et al. [29]	– Decision tree (J48), Multi-Layer Perceptron (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) with PCA	– The fusion of MLP, J48, SMO and IBK is superior to the other classifiers in WPBC dataset. – Avoided the "curse of dimensionality," by using the optimal feature set while reducing the redundancy of feature space.	– For all datasets, no one multi-classifier fusion level or combination is superior. – Utilising more layers of fusion classifiers does not imply greater accuracy. – For any breast cancer dataset, substituting mean or median values for instances with missing values may achieve improved accuracies rather than eliminating it.
Buyrukoglu [34]	– Random Forest, Linear Regression and Gradient Boosting	– Gradient boosting method delivered the best performance for real estate price prediction. – Various metrics such as MAE, RMSE, MSE are used for comparison.	– Predicting the most efficient property pricing for real estate customers in terms of their budgets is critical. – Future prices have to be projected by examining recent market trends and price ranges, as well as forthcoming changes.
Dutta et al. [6]	– Decision tree, random forest, polynomial regression, and linear regression.	– Lasso Regularization has been used to reduce model complexity and prevent over-fitting. – Random Forest Regressor was found to be the best model was found to bewith 0.862533 R2_score, 2.12 MSE score and 4604.86 RMSE scores.	– Dataset can be extended and Deep learning can be used. – Difficult to anticipate a patient's health-care costs solely based on clinical data. Previous health-care costs are the strongest predictor of future costs.

Thus, the focus of this proposed research is to test the generalizability of various datasets, such as the Wisconsin breast cancer dataset, Mobile price range dataset, Health insurance dataset, and Boston house pricing dataset, on the Data analytics and reporting API which uses exclusive feature selection and data visualisation techniques for analysing the predictive performance for various datasets.

3. Survey of Data Analytics Tools and Reporting API

The Reporting API allows you to access report data in Data Analytics through programmatic ways. It's used to create custom dashboards that show Analytics data as well as assisting in the automation of time-consuming reporting tasks. In recent years, demand for machine learning expertise has grown faster than supply. To close this gap, Machine Learning-as-a-Service (MLaaS) acts as an umbrella term encompassing a variety of cloud-based platforms that addresses a wide range of infrastructure concerns, including data pre-processing, model training, and model assessment, as well as prediction [17]. Four notable cloud MLaaS services that enable quick model training and deployment are Amazon Machine Learning, Azure Machine Learning, Google AI Platform, and IBM Watson Machine Learning [33].

Amazon Web Services (AWS) provides easy to use interface for the users with the array of analytics products. AWS offers a comprehensive set of managed services to help us rapidly and easily create, protect, and scale end-to-end big data applications. AWS offers an ecosystem of analytical tools that are particularly built to manage the increasing amount of data and provide insight into one's organisation. However, AWS is limited by the technical support fee to be paid and security issues [32].

Azure Machine learning is a platform that offers MLaaS developed by Microsoft. It gives a lot of choices for creating data-driven machine learning models. Azure Machine Learning, as a Microsoft MLaaS platform, allows for simple model deployment via a cloud web service. However, it necessitates previous understanding of machine learning, making it only accessible to advanced users rather than beginners [14]. Furthermore, the unimposing user interface makes it tough for people to use.

Google Analytics is a free online statistic monitoring tool that delivers a vast amount of data about a website's performance. It allows website writers and marketers to collect an excessive quantity of data about their website, their users, and the advertising they perform, and then alter that data in predefined ways by those who created the reporting component [11].

On IBM Cloud Pak® for Data, IBM Watson® Studio allows data scientists, developers, and analysts to create, execute, and manage Artificial Intelligence (AI) models, as well as enhance decisions. Watson Analytics entry into the realm of analytical software products gives customers access to new features not seen in many other applications [22].

The majority of these digital tools provide application programming interfaces (APIs) or allow delimited file export. As a result, a team may combine a lot of the data in Statistical Analysis System (SAS) Visual Analytics for high-level analysis and reporting [15]. According to the literature, the Google Analytics Reporting API v4 allows users to create custom dashboards to show Google Analytics data using the Google Analytics Reporting API, which saves time by automating complicated reporting processes and also connects user's Google Analytics data to other software applications. This API is used to retrieve the actual Google Analytics data to generate reports in different contexts, as the name implies. This is the most basic and important API [25].

Many open-source data mining tools and applications, such as, Waikato Environment for Knowledge Analysis (WEKA), R-Programming, Konstanz Information Miner (KNIME), Orange, and etc, are now accessible for usage. These software and applications provide a collection of methodologies and algorithms that aid in data analytics. Data visualization, decision trees, cluster analysis, predictive analytics, regression analysis, text mining, and other tasks are made easier using these technologies. For data analytics, these free tools are also extensively used [8-9]. These tools, on the other hand, have drawbacks in that they don't allow the user to choose the target variable. The development of this proposed work was inspired by the implementation approaches and the shortcomings of the above tools.

4. System Design

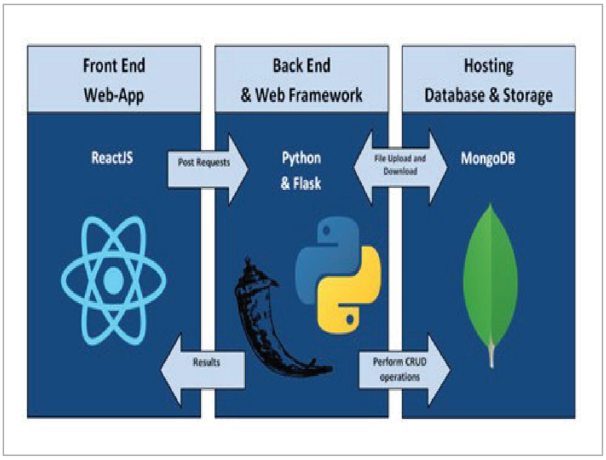
The proposed system has been meticulously designed to serve all types of users. A variety of technologies have been used to provide the best user experience with quickest response time.

4.1. System Architecture

Figure 1 describes the architecture of the proposed system.

Figure 1

System architecture of the proposed DARAPI



The system architecture has three components. They are as follows:

1 Front End Web App

2 Back End and Web Framework

3 Hosting Database and Storage.

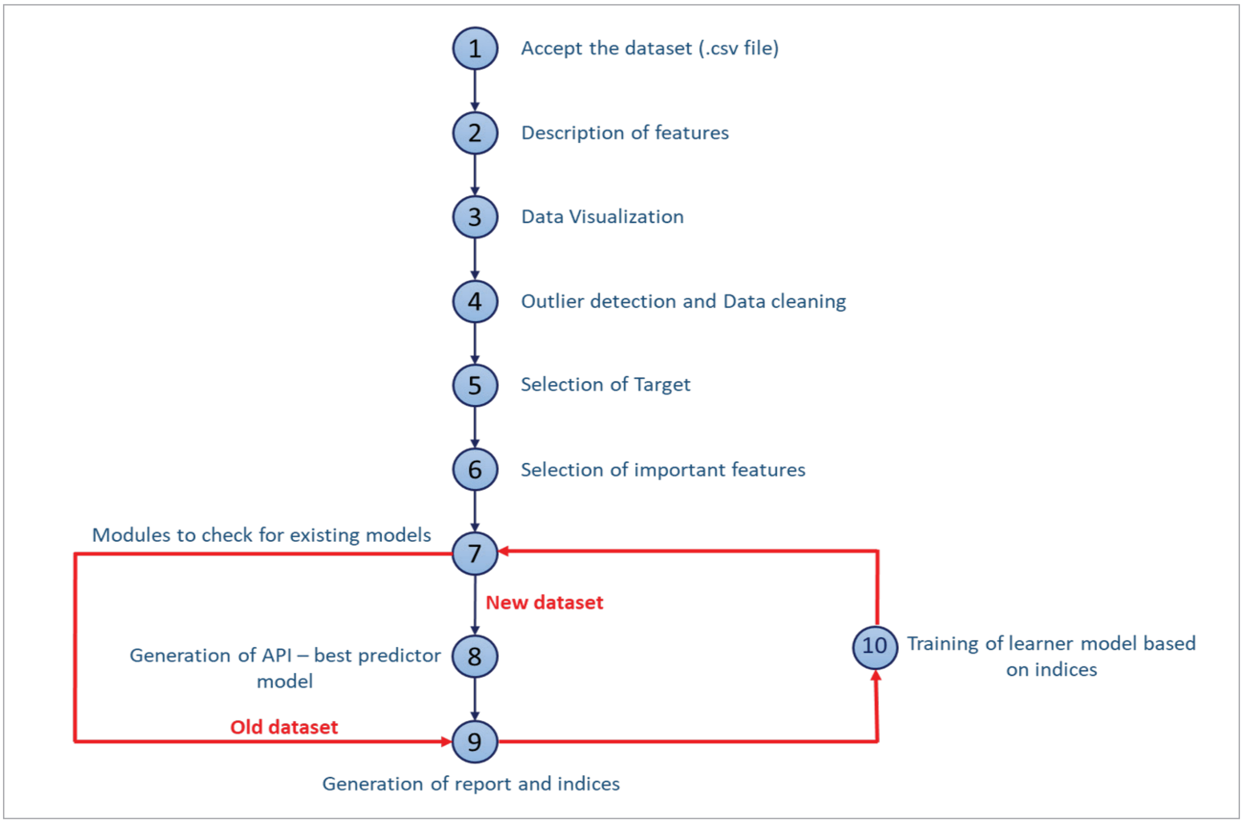
React.js is used for developing the User Interface of the system. It accepts input from the user and forwards it to flask through axios. Flask, a Python-based web framework, accepts the data from the API and performs the required operations on the MongoDB. If it is a GET API, the data is fetched from the MongoDB in Python and it is passed to the React.js code in JSON format. The JSON data is parsed to retrieve the required data. Similarly, for a POST API, the data is sent in the JSON format to flask in Python and required acknowledgement is provided back.

4.2. Work Flow

The proposed system begins with data collection and then moves on to data visualization, preprocessing, feature selection, training and finally ends with final predictions. This work flow is explained in Figure 2.

Figure 2

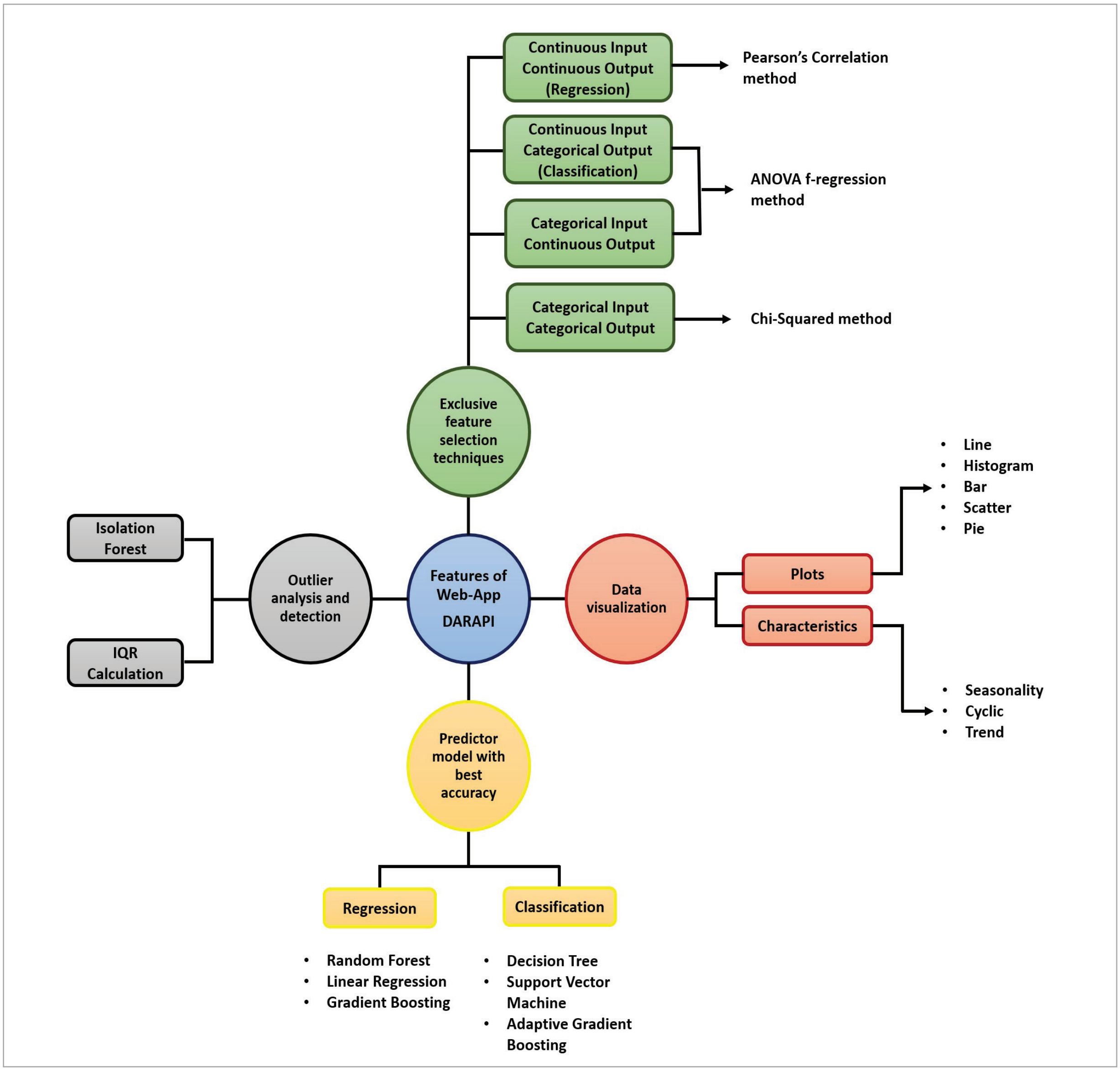
Work flow of the proposed DARAPI


1 Accept the dataset (.csv file)
2 Description of features
3 Data Visualization
4 Outlier detection and Data cleaning
5 Selection of Target
6 Selection of important features
Modules to check for existing models
7
New dataset
Generation of API – best predictor model
8
10 Training of learner model based on indices
Old dataset
9
Generation of report and indices

There are four features of the Web-App. It includes Exclusive feature selection techniques, Data visualization, Outlier analysis and detection and Predictor model with best accuracy. Figure 3 describes the features of the Web-App and its components.

Figure 3

Modules of the proposed system

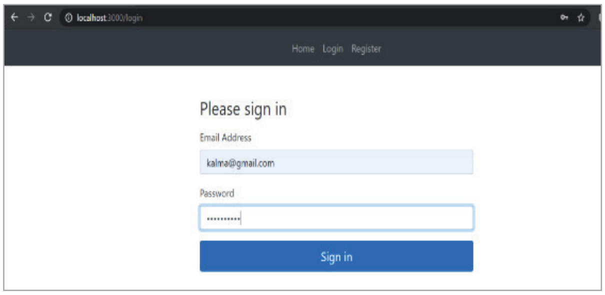

Continuous Input Continuous Output (Regression)
Pearson's Correlation method
Continuous Input Categorical Output (Classification)
Categorical Input Continuous Output
ANOVA f-regression method
Categorical Input Categorical Output
Chi-Squared method
Exclusive feature selection techniques
Isolation Forest
IQR Calculation
Outlier analysis and detection
Features of Web-App DARAPI
Data visualization
Plots
Characteristics
Line
Histogram
Bar
Scatter
Pie
Seasonality
Cyclic
Trend
Predictor model with best accuracy
Regression
Classification
Random Forest
Linear Regression
Gradient Boosting
Decision Tree
Support Vector Machine
Adaptive Gradient Boosting

5. Methodology

Step 1: The user first logs-in to the system as shown in Figure 4. When the user enters the username and password into the login page, a request is made from React.js to flask, which starts a local development server. The server provides access to the user if the pair exists in the NoSQL database i.e., the MongoDB and also displays the profile page to the user as depicted in Figure 5.

Figure 4

Login page



localhost:3000/login

Please sign in

Email Address

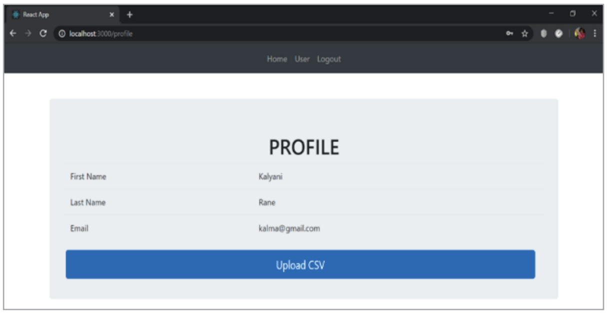
kalma@gmail.com

Password

••••••••••

Figure 5

Profile page



localhost:3000/profile

PROFILE

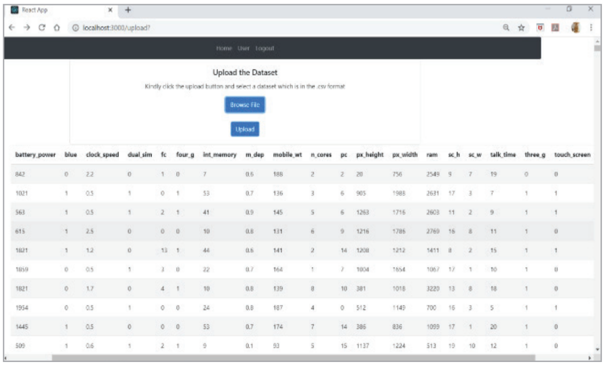
First Name	Kalyani
Last Name	Rane
Email	kalma@gmail.com

Step 2: A Python code for preprocessing is called through flask internally. The dataset is checked for missing data. The missing data is imputed using an Iterative Imputer. Further, the categorical features are label encoded for making them machine understandable.

Step 3: The user is prompted to browse for the dataset. Upon browsing and selecting the required file, the records of the dataset are displayed on the screen as shown in Figure 6. The user has to press the upload button to store the dataset in the MongoDB with the same name as detected while browsing. Currently, the system supports only Comma Separated Values (CSV) files.

Figure 6

File upload page



localhost:3000/upload?

Upload the Dataset

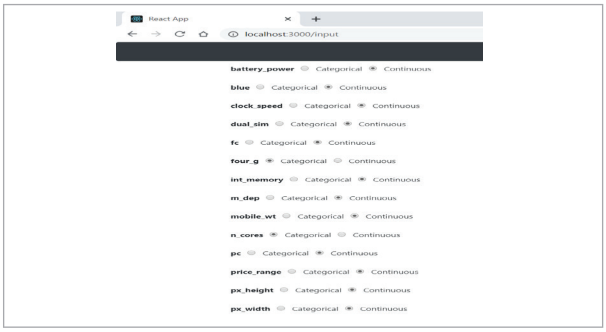
Kindly click the upload button and select a dataset which is in the .csv format

battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen
842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0
1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1
563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1
615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0
1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1
1859	0	0.5	1	3	0	22	0.7	164	1	7	1004	1654	1067	17	1	10	1	0
1821	0	1.7	0	4	1	10	0.8	139	8	10	381	1018	3220	13	8	18	1	0
1954	0	0.5	1	0	0	24	0.8	187	4	0	512	1149	700	16	3	5	1	1
1445	1	0.5	0	0	0	53	0.7	174	7	14	386	836	1099	17	1	20	1	0
509	1	0.6	1	2	1	9	0.1	93	5	15	1137	1224	513	19	10	12	1	0

Step 4: Then, the user is directed to a description page, where the details of the features should be provided as shown in Figure 7.

Figure 7

Feature description page



localhost:3000/input

battery_power ○ Categorical ● Continuous
blue ○ Categorical ● Continuous
clock_speed ○ Categorical ● Continuous
dual_sim ○ Categorical ● Continuous
fc ○ Categorical ● Continuous
four_g ● Categorical ○ Continuous
int_memory ○ Categorical ● Continuous
m_dep ○ Categorical ● Continuous
mobile_wt ○ Categorical ● Continuous
n_cores ● Categorical ○ Continuous
pc ○ Categorical ● Continuous
price_range ○ Categorical ● Continuous
px_height ○ Categorical ● Continuous
px_width ○ Categorical ● Continuous

For each feature, the user has to select if it is continuous or categorical in nature. Categorical variables are data types that may be split into categories based on some qualitative feature. Genders, ages, educational levels, and blood types are examples of categorical data that emerge regularly in real-life applications. Continuous variables can have an infinite number of values between the lowest and highest points of measurement. Continuous data such as height, weight, and temperature are common examples.

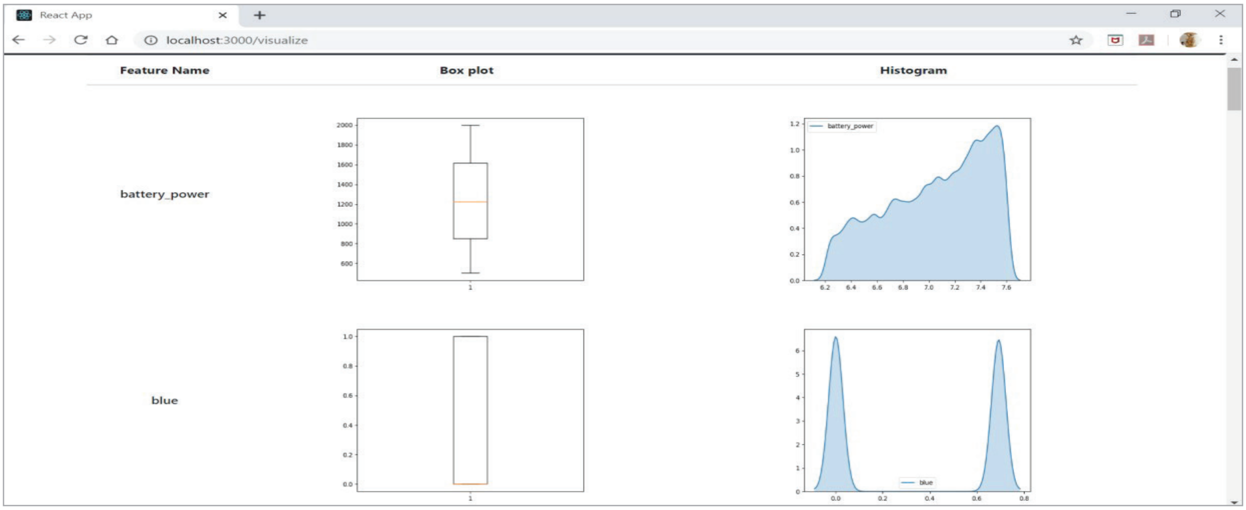
Based on the attribute's relation to the dataset, the user has the freedom to choose whether the attribute is categorical or continuous in nature. There are no limitations for choosing the number of categorical and continuous features [5].

Step 5: The graphical depiction of information and data is known as data visualization. Data visualization tools make it easy to examine and comprehend trends, outliers, and patterns in data by employing visual elements like graphs and chart. Data visualization tools and technologies are critical in the Big Data analytics for analyzing enormous volumes of data and making decisions.

Different visualizations, such as the box plots and the histogram plots of all the features, are displayed in the Dataset Visualization page as illustrated in Figure 8,

Figure 8

Data visualization page



which follows the description page. The plots are generated through the Seaborn Sand Matplotlib libraries in Python, which is called via flask. This outlier visualization through box plot and histogram describes the distribution of data in the dataset.

Step 6: An outlier is a value in a random sampling from a population that deviates abnormally from other values. It helps data analysts to distinguish between normal and abnormal observations. The mode, median, and mean values are replaced for the extreme values in outlier replacement.

In addition to the visualizations, the system provides the list of features containing outliers which are calculated by taking the intersection of outlier records produced by Inter-Quartile Range (IQR) method and isolation forest method as given in Figure 9.

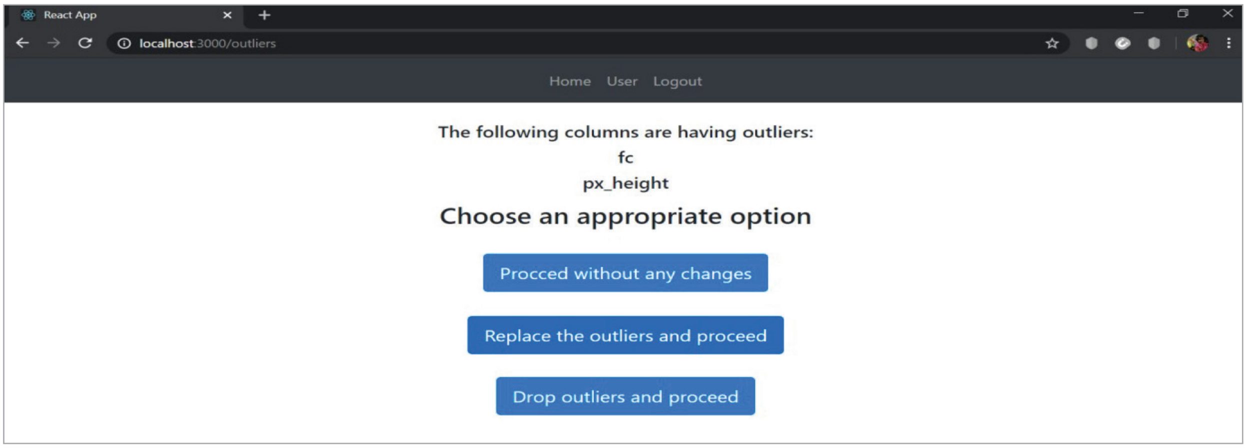
For handling the outliers, the user is provided with the following options:

Proceed without any changes

1. Replace the outliers:
   a. Replace with mean
   b. Replace with nearest quartile
2. Drop outliers

Figure 9

Outlier analysis page


React App
localhost:3000/outliers
Home
User
Logout

The following columns are having outliers:

fc

px_height

Choose an appropriate option

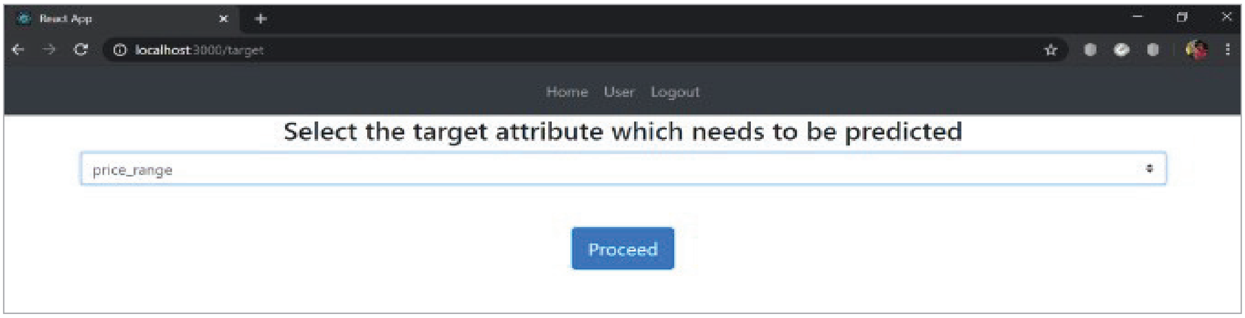
Procced without any changes
Replace the outliers and proceed
Drop outliers and proceed

Step 7: Datasets, after executing the respective outlier handling technique, are generated for future processing.

Step 8: Further, the user is prompted to select the target feature out of all the attributes displayed as shown in Figure 10.

Figure 10

Target selection page



Select the target attribute which needs to be predicted

price_range

Step 9: The importance of each feature, in predicting the target, is calculated based on the nature of the input features and the target feature as shown in Table 2. The Python code, which contains the calculations and final results, is called through flask while loading the page.

Table 2

Selection of feature importance algorithm based on the nature of the features

Sl. No.	Classification	Regression
1.	Linear Support Vector Machine	Linear Regression
2.	Random Forest Classifier	Random Forest Regressor
3.	Decision Tree	Gradient Boosting
4.	Adaptive Gradient Boosting	---

Step 10: The features, along with their importance scores, are displayed in order to allow the user to select the required features for the training phase as depicted in Figure 11. The importance of each feature over the target is shown in percentage in the feature selection page. The features of a dataset are selected based on the impact percentage of the features on the dataset.

Step 11: After feature selection, the user needs to click on the 'Train the model' button to train different models with the refined dataset. Classification algorithms are used if the target feature is categorical in nature and Regression algorithms are used if the target feature is continuous in nature. A combination of primitive and ensemble algorithms has been chosen with the goal of developing a prototype. The different types of algorithms invoked are discussed in Table 3.

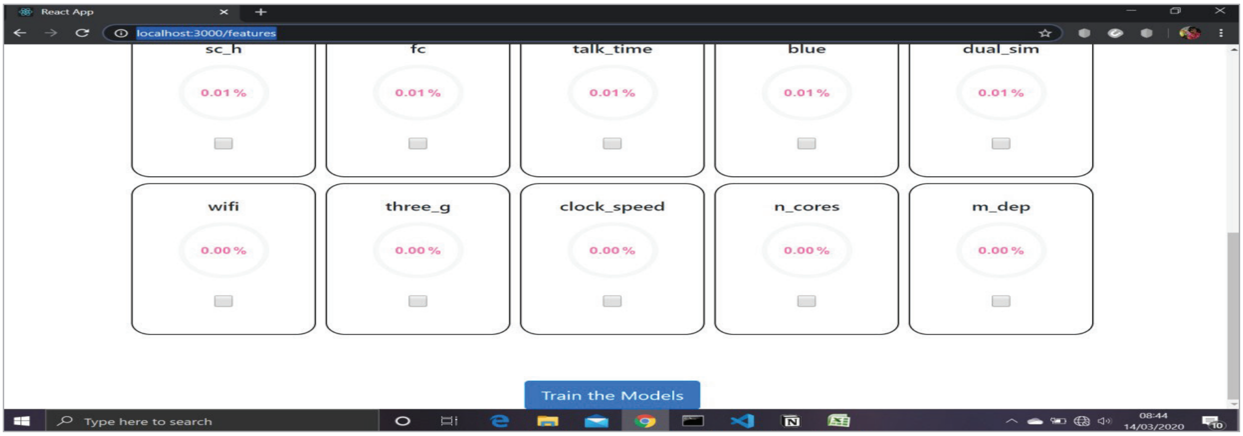
Table 3

Different types of algorithms used in the proposed system

Sl. No.	Input	Target	Method
1.	Continuous	Continuous	Pearson's Correlation (p-test)
2.	Continuous	Categorical	ANOVA f-regression (f-test)
3.	Categorical	Continuous	ANOVA f-regression (f-test)
4.	Categorical	Categorical	Chi-Squared
5.	Continuous and Categorical (Mixed)	Continuous/ Categorical	Combination of above methods

Figure 11

Important feature selection page

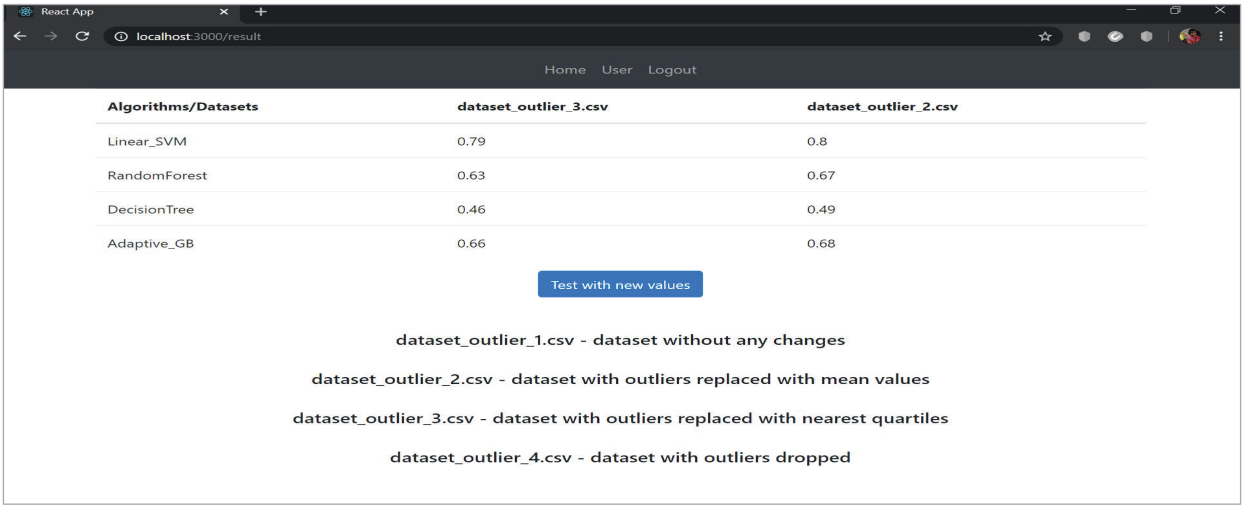


sc_h	fc	talk_time	blue	dual_sim
0.01 %	0.01 %	0.01 %	0.01 %	0.01 %
☐	☐	☐	☐	☐

wifi	three_g	clock_speed	n_cores	m_dep
0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
☐	☐	☐	☐	☐

Figure 12

Machine learning test score page



Algorithms/Datasets	dataset_outlier_3.csv	dataset_outlier_2.csv
Linear_SVM	0.79	0.8
RandomForest	0.63	0.67
DecisionTree	0.46	0.49
Adaptive_GB	0.66	0.68

dataset_outlier_1.csv - dataset without any changes

dataset_outlier_2.csv - dataset with outliers replaced with mean values

dataset_outlier_3.csv - dataset with outliers replaced with nearest quartiles

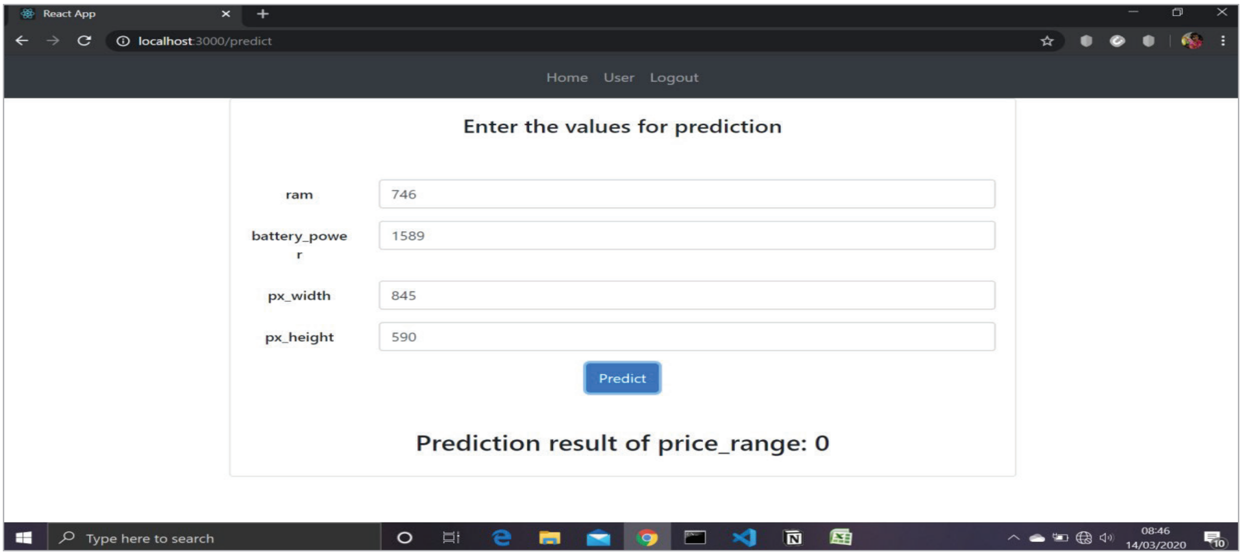
dataset_outlier_4.csv - dataset with outliers dropped

The dataset is first scaled using a Standard Scaler. Then, it is divided into a train set and test set using 70:30 ratios respectively. Further, the Scikit-Learn machine learning models are imported and trained using the train dataset. The trained model is tested using the test dataset and the corresponding accuracy scores are stored in a data frame. These scores are parsed in the flask and then, forwarded to React.js for display.

Step 12: The test score of each algorithm is displayed and the model with the best accuracy is used for future predictions as shown in Figure 12.

Step 13: The classification or regression estimator yields the prediction value. Estimator is an object that fits a model based on input data (i.e., training data) and performs specific calculations on new, unseen data. Further, the user can enter the values for the selected set of features for prediction as depicted in Figure 13. On clicking the predict button, a Python code is called through flask. It executes the predictor which has the best accuracy, for the given set of input values of the features and finally generates the prediction output. This output is carried forward from flask to React.js for display purpose.

Figure 13
Prediction page



Enter the values for prediction

ram: 746

battery_power: 1589

px_width: 845

px_height: 590

Prediction result of price_range: 0

Step 14: Throughout the process, all the valuable details such as choice of feature selection algorithm, choice of handling of outliers, number of categorical features, number of continuous features, types of all the features, best machine learning algorithm chosen, etc. are stored in the MongoDB collection for further use.

If a dataset with similar characteristics is provided as an input in future, the previously stored information is used for automating the decisions straight away instead of processing all the possible options. This way, the execution time is reduced and the system efficiency is improved. It shows the embedded feedback mechanism of the system.

6. Results and Discussions

The proposed system has been tested with different types of datasets, which includes the following:

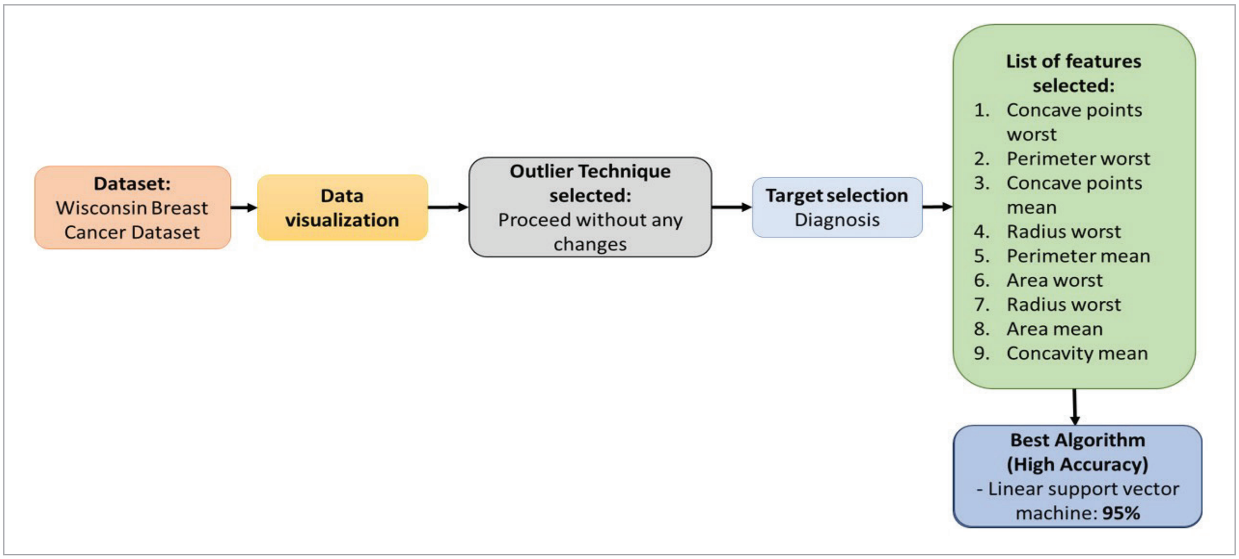
6.1. Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer Dataset is used for predicting if the cancer is malignant or benign in the patients based on various biomarkers like radius, perimeter of cancer cells, etc. It has 32 attributes in total, out of which, 31 input attributes are continuous in nature and one target attribute is categorical in nature. The attributes which are present in the dataset are: Id, Radius mean, Texture mean, Perimeter mean, Area mean, Smoothness mean, Compactness mean, Concavity mean, Concave points mean, Fractal dimension mean, Radius se, Texture se, Perimeter se, Area se, Smoothness se, Compactness se, Concavity se, Concave points se, Symmetry se, Fractal dimension se, Texture worst, Perimeter worst, Area worst, Smoothness worst, Compactness worst, Concavity worst, Concave points worst, Symmetry worst, Fractal dimension worst, Symmetry mean, Radius worst and Diagnosis (Target).

As shown in Figure 14, in DARAPI, Wisconsin breast cancer dataset is uploaded, after choosing the nature of the features (categorical or continuous), the data visualization is displayed through box plot and histogram. For handling the outliers, 'Proceed without any changes' is been selected in the available list of options. After executing the respective outlier handling technique, it is generated for further processing. Diagnosis is selected as the target in the target selection page, followed by selection of required features for the training phase, which is chosen based on the impact percentage of the features on the dataset. Classification algorithms that are used as the target feature is categorical in nature. The Wisconsin Breast Cancer dataset is a binary classification problem with a high degree of correlation. As the correlation increases, so does the accuracy. The test score of each algorithm is displayed and Linear support vector machine is the model with highest accuracy of 95%.

Figure 14

Performance of DARAPI Wisconsin Breast Cancer Dataset



6.2. Mobile Price Range Dataset

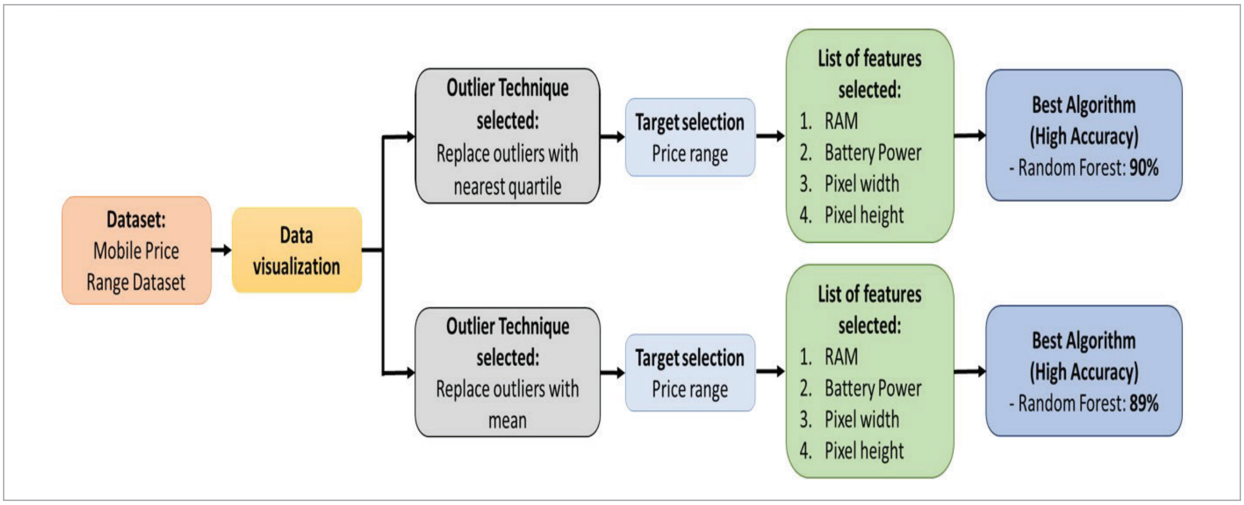
The Mobile Price Range Dataset is used for predicting the price range within which a mobile phone price is likely to fall based on various parameters such as RAM size, 3G, 4G, Bluetooth facilities, etc. The dataset has 21 features in total, out of which, 6 input features and 1 target feature are categorical in nature, and 14 input features are continuous in nature.

The attributes which are present in the dataset are: Battery power, Blue, Clock speed, Dual sim, Fc, Four g, Int memory, M deep, Mobile wt, N cores, pc, Pixel height, Pixel width, RAM, Screen height, Screen weight, Talk time, Three g, Touch screen, WIFI and Price range (Target).

As shown in Figure 15, in DARAPI, Mobile price range dataset is uploaded, after choosing the nature of the

Figure 15

Performance of DARAPI using Mobile Price Range Dataset


Dataset:
Mobile Price Range Dataset
Data visualization
Outlier Technique selected:
Replace outliers with nearest quartile
Target selection
Price range
List of features selected:
1. RAM
2. Battery Power
3. Pixel width
4. Pixel height
Best Algorithm (High Accuracy)
- Random Forest: 90%
Outlier Technique selected:
Replace outliers with mean
Target selection
Price range
List of features selected:
1. RAM
2. Battery Power
3. Pixel width
4. Pixel height
Best Algorithm (High Accuracy)
- Random Forest: 89%

features (categorical or continuous), the data visualization is displayed through box plot and histogram.

For handling the outliers, 'Replace the outliers' is selected, where two outlier handling technique has been proceeded. Price range is selected as the target in the target selection page, followed by selection of required features for the training phase, which is chosen based on the impact percentage of the features on the dataset.

Classification algorithms are used as the target feature that is categorical in nature. For the 'Replace with nearest quartile' outlier handling technique, the test score of each algorithm is displayed and Random Forest is the model with highest accuracy of 90%.

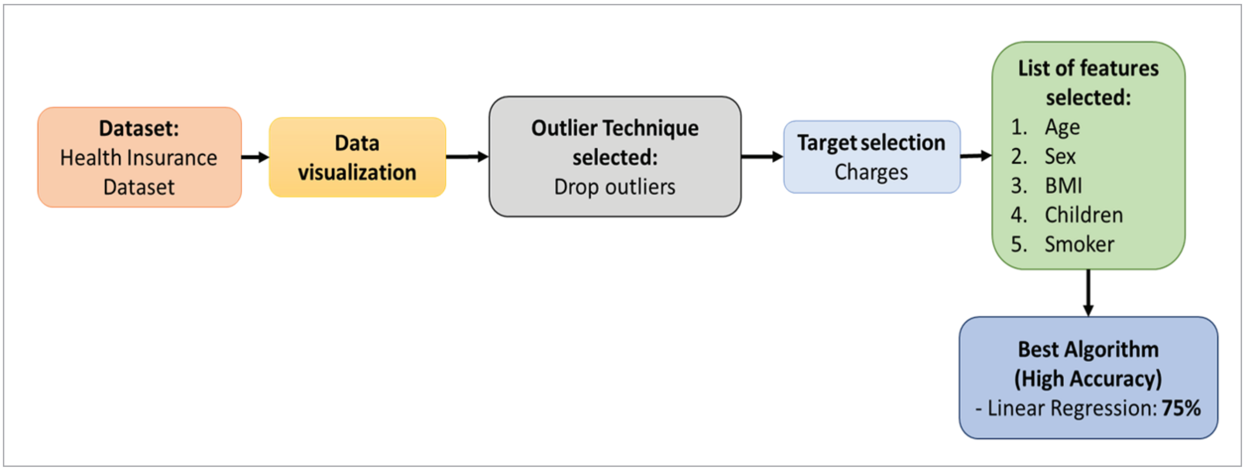
For 'Replace with mean' outlier handling technique, the test score of each algorithm is displayed and Random Forest is the model with highest accuracy of 89%.

6.3. Health Insurance Dataset

The Health Insurance Dataset is used for predicting the insurance charges based on various details of the client. The dataset has 7 attributes in total, out of which, 2 input features and 1 target feature are continuous in nature, and remaining 4 input features are categorical in nature. The attributes which are present in the dataset are: Age, Sex, BMI, Children, Smoker, region and Charges (Target).

As shown in Figure 16, in DARAPI, Health Insurance Dataset is uploaded, after choosing the nature of the features (categorical or continuous), the data visualization is displayed through box plot and histogram. For handling outliers, the option 'Drop outliers' is selected from the list of options. It is generated for further processing after the corresponding outlier handling technique has been executed. Charges is selected as the target in the target selection page and the selection of required features for the training phase is chosen based on the impact percentage of the features on the dataset. Regression algorithms are used as the target feature is continuous in nature. The test score of each algorithm is displayed and Linear regression is the model with highest accuracy of 75%.

Figure 16
Performance of DARAPI using Health Insurance Dataset



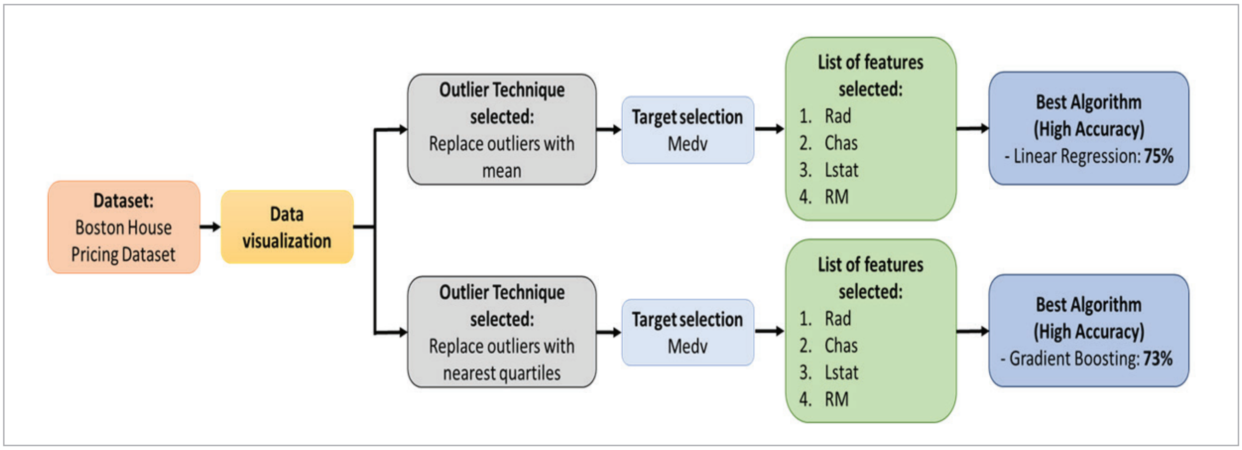
6.4. Boston House Pricing Dataset

The Boston House Pricing Dataset is used for predicting the medv of a house based on parameters like area, tax, crime rate, etc. The dataset has 14 features in total, out of which, 2 input features are categorical in nature and the remaining input features and the target features are continuousare continuous in nature. The attributes which are present in the dataset are: Crim, Zn, Indus, Chas, Nox. Rm, Age, Dis, Rad, Chax, Ptratio, B, Lstat and Medv (Target).

As shown in Figure 17, in DARAPI, Boston House Pricing Dataset is uploaded, after choosing the nature

Figure 17

Performance of DARAPI using Boston House Pricing Dataset


Dataset:
Boston House Pricing Dataset
Data visualization
Outlier Technique selected:
Replace outliers with mean
Target selection
Medv
List of features selected:
1. Rad
2. Chas
3. Lstat
4. RM
Best Algorithm (High Accuracy)
- Linear Regression: 75%
Outlier Technique selected:
Replace outliers with nearest quartiles
Target selection
Medv
List of features selected:
1. Rad
2. Chas
3. Lstat
4. RM
Best Algorithm (High Accuracy)
- Gradient Boosting: 73%

of the features (categorical or continuous), the data visualization is displayed through box plot and histogram. For handling the outliers, 'Replace the outliers' is selected, where two outlier handling technique has been proceeded. Medv is selected as the target in the target selection page, followed by selection of required features for the training phase, which is chosen based on the impact percentage of the features on the dataset. Classification algorithms are used as the target feature that is categorical in nature. For the 'Replace with mean' outlier handling technique, the test score of each algorithm is displayed and Linear Regression is the model with highest accuracy of 75%.

For 'Replace with nearest quartile' outlier handling technique, the test score of each algorithm is displayed and Gradient Boosting is the model with highest accuracy of 73%.

The size of the Boston House Pricing dataset is very small and it is a regression problem with the data points being completely numerical and diverse. As the diversity of the data increases, the correlation decreases, thus reducing the accuracy.

Table 4 shows the performance of DARAPI in terms of accuracy for different datasets, including the problem type of the processed dataset (classification or regression) and the outlier handling technique used for each

Table 4

DARAPI performance in terms of accuracy for different datasets

Sl. No.	Dataset	Problem Type	Outlier Technique	Features Selected	Best Algorithm	Accuracy (in %)
1.	Wisconsin Breast Cancer	Classification	Proceed without any changes	Concave points worst, perimeter worst, concave points mean, radius worst, perimeter mean, area worst, radius mean, area mean, concavity mean	Linear SVM	**95**
2.	Mobile Price Range	Classification	Replace with nearest quartile	RAM, Battery power, Pixel width, Pixel height	Random Forest	**90**
3.	Mobile Price Range	Classification	Replace with mean	RAM, Battery power, Pixel width, Pixel height	Random Forest	**89**
4.	Health Insurance	Regression	Drop outliers	Age, BMI, Children, Sex, Smoker	Linear Regression	**75**
5.	Boston House Pricing	Regression	Replace with mean	Rad, Chas, Lstat, RM	Linear Regression	**75**
6.	Boston House Pricing	Regression	Replace with nearest quartile	Rad, Chas, Lstat, RM	Gradient Boosting	**73**

Table 5
Comparison of open-source data analytics tools with the proposed work

Sl. No.	Open-source Tool	Features	Limitations
1.	WEKA	– Machine learning techniques for data mining – Data pre-processing tools – Experiment scripting flexibility – Attribute selection and visualization – Independent and portable	– Memory constraints – Lack of proper documentation – Slightly slower execution – Poor database connectivity – Poor parameter optimization – weaker in classical statistics
2.	R	– ML operations – Cross interoperability – Data wrangling – Impressive reports – Wide variety of packages – Provides graphical facilities for data analysis	– Difficult for novice users to comprehend – Inadequate documentation – Poor memory management – System is slow and insecure
3.	KNIME	– intuitive user interface – Data mining, text mining – Business Intelligence – Works with Java runtime – Includes extensions for big data and bio informatics – Supports all major operating systems	– Slower execution – Lower accuracy – Knowledge of R/Python is necessary – Visualization has to be improved – Memory utilization can be a concern
4.	Orange	– Suitable for novice and experts – Shortest scripts – Easiest to learn – Add-ons for bio-informatics and text mining	– Suitable for smaller datasets – It is weak in classical statistics – No automatic parameter optimization of machine learning
5.	DARAPI (Our proposed tool)	– Simple user interface – Suitable for novice and experts – Data pre-processing tools – Data visualization – Attribute and target selection – Options for handling the outliers – Combination of primitive and ensemble algorithms for best accuracy – Embedded feedback mechanism	– Currently, the system supports only CSV files. – Does not support deep learning algorithms and hybrid models – Multiple datasets cannot be uploaded – Video and audio files are not supported

dataset, features chosen for the training phase, and the best algorithm with its accuracy for each problem type. Table 5 summarizes the comparison of the features and the limitations of top few open-source data analytics tools with our proposed tool.

7. Discussion of the Findings

The results obtained from testing DARAPI for four diverse datasets highlight the following features of the system that has enhanced the performance:

1 Feature selection technique is one of the most striking features of the system that adds immense value to the outcome. It displays the contribution of various features towards the selected target variable based on their nature (categorical or continuous). In the Mobile Price Range dataset, this feature selection technique is very apparent where the ‘ram’ adds over 98 percent to the final prediction, and thus provides a fairly high accuracy of 90%. Also, the feature selection mechanism gives the flexibility to the user to include features which are important by virtue.

2 Handling of outliers has proven to play a pivotal role in improving the performance of the prediction models. This is evident from the observations made while analyzing the Boston House Pricing dataset, where the accuracy increased from 73% to 75% upon changing the outlier handling technique. A similar trend can be seen in performances of Mobile Price Range dataset.

3 Machine Learning models are used for arriving at the best accuracy. This kind of an extensive analysis is useful in identifying which algorithm works best for a given type of dataset. Thus, it enables the data science researchers to further study and enhance the relation between data and machine learning algorithms.

8. Assumptions and Limitations

- DARAPI is equipped to accept the datasets only in CSV format as of now. Hence, all datasets are required to be converted into CSV format by the user manually before uploading.
- DARAPI is restricted to basic machine learning algorithms. It does not support deep learning algorithms and hybrid models.
- The API accepts only one dataset at a time. Multiple datasets cannot be uploaded at the same time to perform cross join operations. Hence, the user is required to merge the datasets and format them accordingly before uploading it, in case there are more than one data files.
- DARAPI is capable to work with text and numerical data only. Video and audio files are not supported.

9. Conclusion

The main focus of DARAPI is to save the user's time and effort in analyzing the data by providing them with an easy-to-use analysis tool. This is ensured by the following features:

1 The DARAPI has been tested on diverse datasets which shows the generalizability of the proposed system.

2 From the results that are obtained, it is very clear that the system uses only that model which provides the best accuracy for the process of prediction. This improves the overall performance of the system.

3 In addition, the system is equipped with a variety of data pre-processing options, using which, the user can refine and restructure the data to fit his/her needs.

4 Moreover, the embedded feedback mechanism of the system captures the characteristics of the data and the choices made to reach the conclusion. Further, this information is used for taking direct decisions for similar datasets, therefore reducing the overall time taken for processing.

10. Future Scope

DARAPI has an exceptional feature that displays the correlation of the attributes present in the dataset to the target variable which allows the user to select or ignore a particular feature for building the model. The combination of these notable elements in DARAPI makes it suitable to be used for innumerable applications. Some of the applications are listed below:

1 Sales Prediction– The sales of various products sold in the mart across its stores in different locations is collected along with the attributes associated with each product. This data when fed to DARAPI, it can be used to find out the properties of a product, and store, which impacts the sales of a product

2 Bank Credit Analysis – The huge volumes of credit data held by the banks is uploaded to the server. The model is built by the server and the built model can then be used by the bank to find if a customer is potentially liable to fraud in the future or not. This endpoint can also be used in the bank's website, so that users can do a self-check of their eligibility.

3 Weather Prediction– Since the system can also predict continuous values (Regression), it can also be used in the prediction of weather conditions.

In future, this system can be augmented with hybrid models which are built by combining the conventional algorithms. Deep learning techniques can also be integrated to improve the efficiency of the system. In order to improve the accuracy, data augmentation can be done as an extension to the current application. Further, the system can be enhanced to support image and video mining techniques and various other kinds of pre-processing approaches. This system is available as an open-source tool and can be accessed from its GitHub repository at any time from the link as given below. https://github.com/sspppaaa123/DARAPI/tree/e6ffebe108bf8b19f951d7124205810854f91444

References

1. Athmaja, S., Hanumanthappa, M., Kavitha, V. A Survey of Machine Learning Algorithms for Big Data Analytics. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, 1-4. https://doi.org/10.1109/ICIIECS.2017.8276028

2. Balusamy, B., Abirami, R. N., Kadry, S., Gandomi, A. H. Big Data Analytics. In Big Data: Concepts, Technology, and Architecture, Wiley, 2021, 161-186. https://doi.org/10.1002/9781119701859.ch6

3. Chong, J., Tjurin, P., Niemelä, M., Jämsä, T., Farrahi, V. Machine-learning Models for Activity Class Prediction: A Comparative Study of Feature Selection and Classification algorithms. Gait & Posture, 2021, 89, 45-53. https://doi.org/10.1016/j.gaitpost.2021.06.017

4. Delen, D., Demirkan, H. Data, Information and Analytics as Services. Decision Support Systems, 2013, 55(1), 359-363. https://doi.org/10.1016/j.dss.2012.05.044

5. Dinh, D-T., Huynh, V.-N., Sriboonchitta, S. Clustering Mixed Numerical and Categorical Data with Missing Values. Information Sciences, 2021, 571, 418-442. https://doi.org/10.1016/j.ins.2021.04.076

6. Dutta, K., Chandra, S., Gourisaria, M. K., GM, H. A Data Mining Based Target Regression-Oriented Approach to Modelling of Health Insurance Claims. Proceedings of IEEE 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, 1168-1175. https://doi.org/10.1109/ICCMC51019.2021.9418038

7. Elfeky, A. I. M., Elbyaly, M. Y. H. The Use of Data Analytics Technique in Learning Management System to Develop Fashion Design Skills and Technology Acceptance. Interactive Learning Environments, 2021. https://doi.org/10.1080/10494820.2021.1943688

8. Fu, I.-K., Chaves, M., Fagan, A., Hazen, J. Bringing Google Analytics, Facebook, and Twitter Data to SAS® Visual Analytics. Proceedings of SAS Conference: SAS Global Forum, 2016, 1-12. https://www.lexjansen.com/wuss/2016/153_Final_Paper_PDF.pdf

9. Google. Google Analytics Reporting API. Available from https://developers.google.com/analytics/devguides/reporting/core/v4. Accessed on 2021 June 5.

10. Hodge, V., Austin, J. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 2004, 22(2), 85-126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

11. Hoyt, R. E., Snider, D., Thompson, C. J., Mantravadi, S. IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. JMIR Public Health and Surveillance, 2016, 2(2), 1-12. https://doi.org/10.2196/publichealth.5810

12. Jijo, B. T., Abdulazeez, A. M. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2021, 2(1), 20-28. https://doi.org/10.38094/jastt20165

13. Khalid, S., Khalil, T., Nasreen, S. A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. Proceedings of Science and Information Conference, 2014, 372-378. https://doi.org/10.1109/SAI.2014.6918213

14. Klein, S. Azure Machine Learning. In: IoT Solutions in Microsoft's Azure IoT Suite. Apress, Berkeley, CA, 2017, 227-252. https://doi.org/10.1007/978-1-4842-2143-3_14

15. Kawade, B., Deoskar, A. Comparative Study of Data Analytics Open Source Tools for Educational Data Analytics. Journal of Emerging Technologies and Innovative Research (JETIR), 2019, 6(2), 299-301. http://www.jetir.org/papers/JETIRAE06072.pdf

16. Kumar K. U. P., Gandhi, O., Reddy, M. V., Srinivasu, S. V. N. Usage of KNN, Decision Tree and Random Forest Algorithms in Machine Learning and Performance Analysis with a Comparative Measure. In: Bhattacharyya D., Thirupathi Rao N. (eds), Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, 1280, Springer, Singapore, 2021, 473-479. https://doi.org/10.1007/978-981-15-9516-5_39

17. Lee, Y.-S. Analysis on Trends of Machine Learning-as-a-Service. International Journal of Advanced Culture Technology, 2018, 6(4), 303-308. https://doi.org/10.17703//IJACT2018.6.4.303

18. Moorthi, K., Dhiman, G., Arulprakash, P., Suresh, C., Srihari, K. A Survey on Impact of Data Analytics Techniques in E-commerce. Materials Today: Proceedings, 2021, 1-8. https://doi.org/10.1016/j.matpr.2020.10.867

19. Muneeswaran, V., Nagaraj, P., Dhannushree, U., Lakshmi, S. I., Aishwarya, R., Sunethra, B. A Framework for Data Analytics-Based Healthcare Systems. In: Raj J. S., Iliyasu A.M., Bestak R., Baig Z.A. (eds), Innovative Data Communication Technologies and Application, 59, Springer, Singapore, 2021, 83-96. https://doi.org/10.1007/978-981-15-9651-3_7

20. Nandigramwar, H., Mittal, A., Bhatnagar, A., Rashid, M. A Distributed and Unified API Service for Machine Learning Models. Proceedings of IEEE 2nd International Conference on Intelligent Engineering and

Management (ICIEM), 2021, 480-485. https://doi.org/10.1109/ICIEM51511.2021.9445348

21. Ojha, U., Goel, S. A Study on Prediction of Breast Cancer Recurrence Using Data Mining Techniques. Proceedings of IEEE 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017, 527-530. https://doi.org/10.1109/CONFLUENCE.2017.7943207
22. Pandian, I., Krishnamoorthy, M., Meenalaxmi, Abinaya. A Comparative Study on Three Big Data Analytics Tools. International Journal of Advanced Scientific Research and Management, 2018, 3(9), 192-194. http://ijasrm.com/wp-content/uploads/2018/09/IJASRM_V3S9_831_192_194.pdf
23. Pathak, P., Iyengar, S. P., Abhyankar, M. A Survey on Tools for Data Analytics and Data Science. Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics, IGI Global, 2021, 28-49. https://doi.org/10.4018/978-1-7998-3053-5.ch003
24. Qin, S. J., Chiang, L. H. Advances and Opportunities in Machine Learning for Process Data Analytics. Computers & Chemical Engineering, 2019, 126, 465-473. https://doi.org/10.1016/j.compchemeng.2019.04.003
25. Rajeswari, C., Basu, D., Maurya, N. Comparative Study of Big data Analytics Tools: R and Tableau. IOP Conference Series: Materials Science and Engineering, 2017, 263(4), 1-9. https://doi.org/10.1088/1757-899X/263/4/042052
26. Rahul, K., Banyal, R. K., Goswami, P., Kumar, V. Machine Learning Algorithms for Big Data Analytics. In: Singh V., Asari V., Kumar S., Patel R. (eds), Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing, 1227, Springer, Singapore, 2021, 359-367. https://doi.org/10.1007/978-981-15-6876-3_27
27. Rashid, M., Hamid, A., Parah, S. A. Analysis of Streaming Data Using Big Data and Hybrid Machine Learning Approach. In: Singh A., Mohan A. (eds), Handbook of Multimedia Information Security: Techniques and Applications, Springer, Cham, 2019, 629-643. https://doi.org/10.1007/978-3-030-15887-3_30
28. Sagiroglu, S., Sinanc, D. Big Data: A Review. Proceedings of International Conference on Collaboration Technologies and Systems (CTS), 2013, 42-47. https://doi.org/10.1109/CTS.2013.6567202
29. Salama, G. I., Abdelhalim, M. B., Zeid, M. A. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of Computer and Information Technology, 2012, 1(1), 36-43. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.9343&rep=rep1&type=pdf
30. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2021, 2, 1-21. https://https://doi.org/10.1007/s42979-021-00592-x
31. Sumbaly, R., Vishnusri, N., Jeyalatha, S. Diagnosis of Breast Cancer Using Decision Tree Data Mining Technique. International Journal of Computer Applications, 2014, 98(10), 16-24.https://doi.org/10.5120/17219-7456
32. Swensson, E., Dame, E., Kenghe, S. Big Data Analytics Options on AWS. Technical Report, Dec. 2018. Available from: https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf. Accessed on 2021 June 5.
33. Tadejko, P. Cloud Cognitive Services Based on Machine Learning Methods in Architecture of Modern Knowledge Management Solutions. In: Poniszewska-Marańda A., Kryvinska N., Jarząbek S., Madeyski L. (eds), Data-Centric Business and Application, Lecture Notes on Data Engineering and Communications Technologies, 40, Springer, Cham, 2019, 169-190. https://doi.org/10.1007/978-3-030-34706-2_9
34. Uzut, O. G., Buyrukoglu, S. Prediction of Real Estate Prices with Data Mining Algorithms. Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences. 2020, 7(9), 77-84. https://www.euroasiajournal.org/Makaleler/1026312952_11%2081%2056%2077-84.pdf
35. Wieringa, J., Kannan, P. K., Ma, X., Reutterer, T., Risselada, H., Skiera, B. Data Analytics in a Privacy-Concerned World. Journal of Business Research, 2021, 122, 915-925. https://doi.org/10.1016/j.jbusres.2019.05.005
36. Zulkernine, F., Martin, P., Zou, Y., Bauer, M., Gwadry-Sridhar, F., Aboulnaga, F. Towards Cloud-Based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud. Proceedings of IEEE International Congress on Big Data, 2013, 62-69. https://doi.org/10.1109/BigData.Congress.2013.18