Prediction of Workload Performance of Data Center Using Machine Learning Techniques

[1]Arul Natarajan, [2]A.Sai Preetham, [3] K.Ramakrishna, [4]Vidya Rajasekaran, [5]Amjan Sheik, [6]J.Ramesh Babu

[1, 2, 3, 5, 6,] Department of Computer Science and Engineering, St. Peter's Engineering College, Medchal, Hyderabad

[4,] B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai

***Abstract*—The workload performance in a data center depends on the available resources and the workload. If the workload is too low whereas there are many computational and network resources, then those resources are not utilized to their maximum capacity because of low workload. Likewise, a high workload with low resources is not also advisable as the resources will not be able to meet up the demand. This project aims at predicting the performance by analyzing a data set, consisting of the above mentioned properties by using the Random Forest Classifier, Gradient Booster Algorithm, Logistic Regression, ANN (Artificial Neural Networks) of Sklearn's Ensemble Module and the results of these algorithms are further tallied by executing Quadratic Discriminator Analysis on the same dataset.**

***Index Terms*— Machine Learning, Quadratic Discriminator Analysis, Ensemble Method**

I. INTRODUCTION

The workload performance in a datacenter depends on the available resources and the workload. If the workload is too low whereas there are many computational and network resources, then those resources are not utilized to their maximum capacity because of low workload. Likewise, a high workload with low resources is not also advisable as the resources will not be able to meet up the demand. The workload performance depends on various network properties like minimum and maximum bitrates switch ports, channel medium and line encoding. Also other properties like Auto Negotiation, Topology and sub layers play a role in getting good performance.

This project aims at predicting the performance by analyze a data set, consisting of the above mentioned properties by using the Random Forest classifier, Gradient Boost algorithm, logistic regression, Artificial Neural Network of Sklearn's Ensemble Module and The results of these algorithms are further tallied by executing Quadratic Discriminator Analysis with the same dataset.

In the existing system sufficient measures are not taken to predict the datacenter workload performance, as a result of which people using data center services are facing network issues at high network traffic times like office opening hours. The proposed project suggests analyze the datacenter workload performance using the above mentioned machine learning techniques, so that additional measures can be planned to enhance the performance.

II. LITERATURE SURVEY

In order to optimize resource allocation, increase energy efficiency, and lower operating costs, workload prediction is a critical component of datacenter management. Due to their capacity to increase forecast accuracy by merging numerous models, ensemble learning techniques have recently grown in prominence in the field of workload prediction.

A summary of the state of the art in workload performance prediction for data centre is given by authors C. Lain et al [1]. The authors look at a number of methods, including data mining, machine learning, and statistical modeling. The obstacles and potential directions in this field are also covered in the study.

Machine learning techniques for performance prediction in cloud computing, including linear regression, decision trees, neural networks, and support vector regression. The author M. A. Khan et al [2] provide a detailed comparison of these techniques based on their prediction accuracy and computational complexity.

The performance of numerous workload prediction techniques, such as auto regressive integrated moving average (ARIMA), support vector regression (SVR), and artificial neural networks (ANNs), is compared by author F. Wang et al [3]. The authors compare these techniques using actual data from a cloud data centre and find that ANNs provide more accurate predictions than ARIMA and SVR.

A method using deep learning for multi-step workload forecasting in cloud data centre is proposed [4]. Using actual data, the authors assess their suggested strategy and contrast it with alternative methodologies like ARIMA and SVR.

The results show that the deep learning approaches out performs other methods in terms of prediction accuracy. The performance prediction for cloud computing, including workload prediction and resource allocation prediction is processed. The authors examine various techniques, including statistical modeling, machine learning, and artificial intelligence. The paper by the author X. Liu et al [5] also discusses the challenges and future directions in this area.

The above studies demonstrate that there are various techniques available for workload performance prediction in the data center, including statistical modeling, machine learning, and deep learning. These techniques have been shown to improve prediction accuracy compared to traditional methods. As data centers become more complex and dynamic, the importance of accurate performance prediction is likely to increase, and new techniques will continue to be developed to meet this need [6] [7].

Machine learning algorithms have been in great use for various industries in the present world. Furthermore, they have been capable of not just analyzing but also in providing respective levels of accuracy leaving us with a wide spectrum of choice to select the suitable algorithm in order to produce the most effective outcomes [8] [9] [10].

In this experiment, we have determined to work with machine learning algorithms and neural networks to propose the most effective one in order to have a successful detection of tempo romandibular disease. This experiment deals with machine learning algorithms such as, Logistic Regression, Artificial Neural Networks, Gradient Boost, Quadratic Discriminate Analysis, and Random Forest.

To summarize the experiment, we have proposed three algorithms Artificial Neural Networks, Gradient Boost Algorithm and Logistic Regression to predict workload performance in a datacenter with highest accuracy and successful outcomes compared to all the algorithms dealt in the project.

III. PROPOSED SYSTEM

This flowchart in Fig 1 depicts the project's procedure, which begins with data collection, pre-processing, feature selection, algorithm training and testing, and concludes with our final analysis.

Quadratic Discriminate Analysis

Using the classification technique Quadratic Discriminate Analysis (QDA), data points are classified into multiple classes based on their properties. It is a statistical strategy that relies on the fact that each class's covariance matrices are unique and the data points are normally distributed.

QDA is a powerful tool for dealing with non-linear decision boundaries and complex data. When working with large datasets, if the number of features is excessive in comparison to the total number of data points, it can be computationally expensive and prone to over fitting.

Gradient Boosting

For classification and regression tasks, a powerful machine learning technique known as gradient boosting is used. It is an ensemble boosting strategy that combines a number of weak learners, such as decision trees, to build a strong learner capable of making precise predictions.

Gradient Boosting, on the other hand, is computationally expensive and requires careful tuning of hyper parameters such as the learning rate, the number of trees, and the depth of each tree. It may also have uneven data, in which one class has far fewer samples than the others. To remedy this issue, it may be required to employ additional procedures such as class weighting or re- sampling.

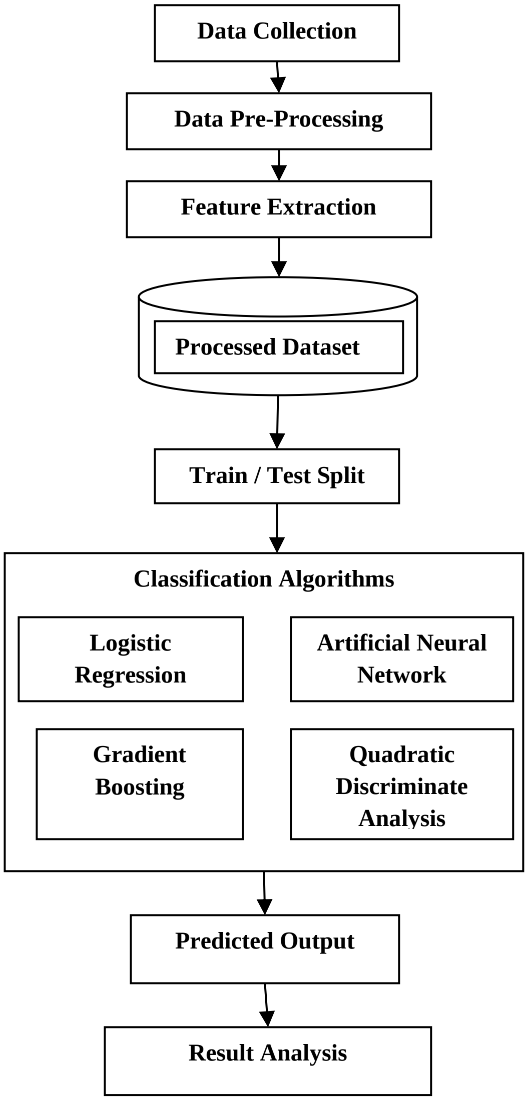


Fig 1. Architectural Design

Random Forest

Random Forest is an excellent machine learning technique that is utilized for both classification and regression issues. In this type of ensemble learning technique, a number of decision trees are joined to build a single powerful learner capable of making accurate predictions.

Random Forest, on the other hand, may necessitate careful tuning of hyper parameters such as the number of trees and the maximum depth of each tree because it can be computationally expensive, particularly when working with large datasets. It may also have imbalanced data, in which one class has a disproportionately large number of samples in comparison to the others. To solve this issue, extra methods such as class weighting or re sampling may be required.

Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models inspired by the structure and operation of the human brain. This method can be used to perform a range of machine learning tasks such as classification, regression, and unsupervised learning.

ANNs, on the other hand, can be computationally expensive, especially when working with deep networks with multiple layers or large datasets. They may also suffer from over fitting, which occurs when a network is extremely complex and matches the training data too closely, resulting in poor performance on new data.

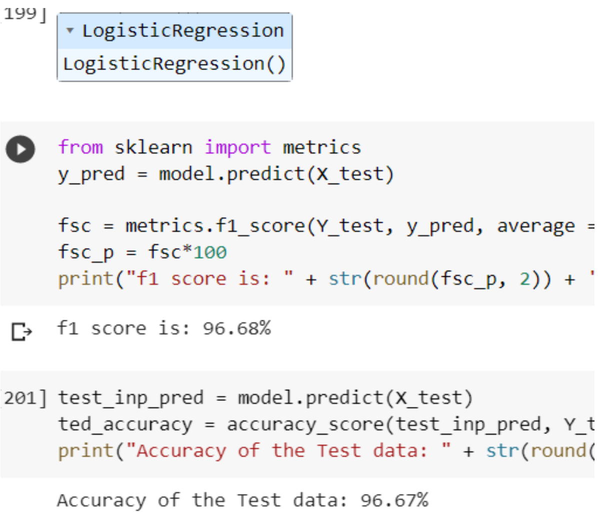
Logistic Regression

For situations involving binary classification, a statistical technique known as logistic regression is utilized. The goal is to predict a binary result (such as yes or no, true or false) based on one or more input variables. It is a type of supervised learning technique that uses a logistic function to assess the likelihood of a binary result.

However, intricate datasets with non-linear correlations between the input variables and the output may not respond well to logistic regression. The input variables are also assumed to be independent of one another, which may not be the case for all datasets.

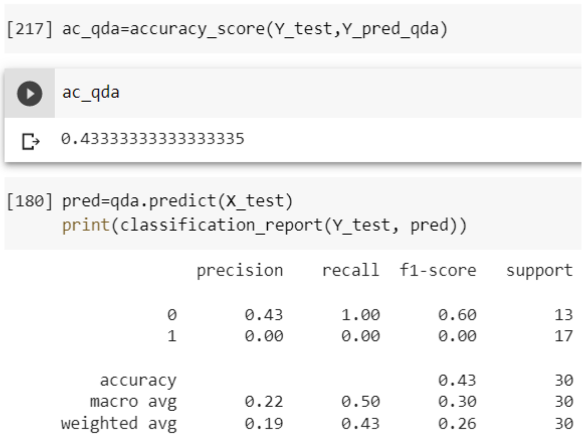
IV. RESULTS

The best model is defined in terms of accuracy. The accuracy of the classification algorithms are evaluated using the metrics. The results of the accuracy are depicted in the below figures. Figure 2 indicates the accuracy of Logistic Regression with 96.67% which outstands all other machine learning models with maximum accuracy, Figure 3 represents the accuracy of QDA with 43%, and Figure 4, 5 and 6 represents an accuracy of 96% using Gradient Booster, Random Forest and Artificial Neural Network.



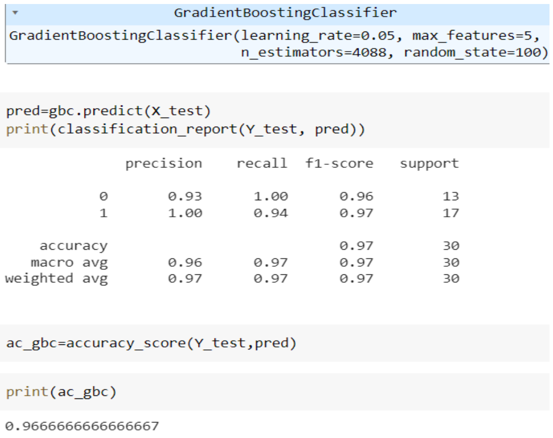
```
199]  ▾ LogisticRegression
      LogisticRegression()

from sklearn import metrics
y_pred = model.predict(X_test)

fsc = metrics.f1_score(Y_test, y_pred, average =
fsc_p = fsc*100
print("f1 score is: " + str(round(fsc_p, 2)) + '

f1 score is: 96.68%

201] test_inp_pred = model.predict(X_test)
     ted_accuracy = accuracy_score(test_inp_pred, Y_t
     print("Accuracy of the Test data: " + str(round(

     Accuracy of the Test data: 96.67%
```

Fig 2. Architectural Design



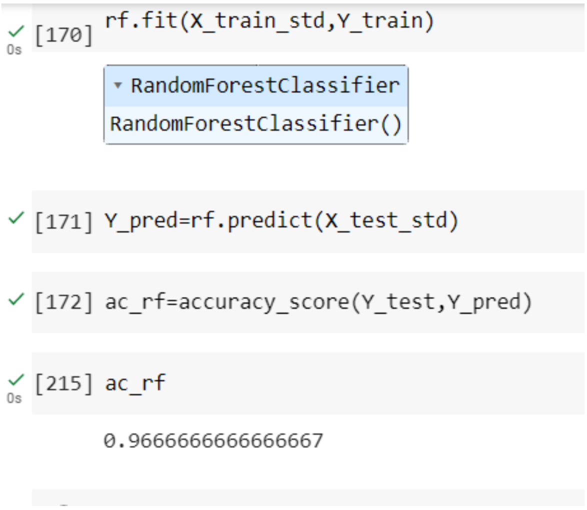
```
[217] ac_qda=accuracy_score(Y_test,Y_pred_qda)

ac_qda

0.43333333333333335

[180] pred=qda.predict(X_test)
      print(classification_report(Y_test, pred))

              precision    recall  f1-score   support

           0       0.43      1.00      0.60        13
           1       0.00      0.00      0.00        17

    accuracy                           0.43        30
   macro avg       0.22      0.50      0.30        30
weighted avg       0.19      0.43      0.26        30
```

Fig 3. Quadratic Discriminate Analysis



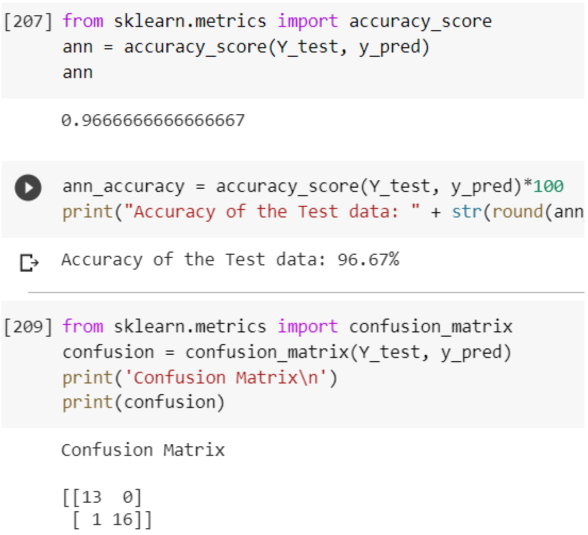
```
▾                 GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.05, max_features=5,
                           n_estimators=4088, random_state=100)

pred=gbc.predict(X_test)
print(classification_report(Y_test, pred))

              precision    recall  f1-score   support

           0       0.93      1.00      0.96        13
           1       1.00      0.94      0.97        17

    accuracy                           0.97        30
   macro avg       0.96      0.97      0.97        30
weighted avg       0.97      0.97      0.97        30

] ac_gbc=accuracy_score(Y_test,pred)

] print(ac_gbc)

0.9666666666666667
```

Fig 4. Gradient Boosting



```
✓ [170] rf.fit(X_train_std,Y_train)
0s
        ▾ RandomForestClassifier
        RandomForestClassifier()

✓ [171] Y_pred=rf.predict(X_test_std)

✓ [172] ac_rf=accuracy_score(Y_test,Y_pred)

✓ [215] ac_rf
0s
        0.9666666666666667
```

Fig 5. Random Forest



```
[207] from sklearn.metrics import accuracy_score
      ann = accuracy_score(Y_test, y_pred)
      ann

      0.9666666666666667

ann_accuracy = accuracy_score(Y_test, y_pred)*100
print("Accuracy of the Test data: " + str(round(ann

Accuracy of the Test data: 96.67%

[209] from sklearn.metrics import confusion_matrix
      confusion = confusion_matrix(Y_test, y_pred)
      print('Confusion Matrix\n')
      print(confusion)

      Confusion Matrix

      [[13  0]
       [ 1 16]]
```

Fig 6. Artificial Neural Network

V. CONCLUSION

The project is useful in understanding the factors leading to improving the datacenter workload performance. In this project accurately predicting workload performance in a data centers a vital activity that could greatly affect the datacenter effectiveness and productivity. Despite the difficulties involved in this undertaking, there are a number of potential methods that may be applied to the creation of an efficient workload prediction system. The initiative helps data analysts learn more about the employed algorithms. Finally, thanks to this initiative, datacenter services will improve.

Currently, the accuracy of the Quadratic Discriminate Analysis Classification Random Forest analysis, Logistic Regression, ANN, and Gradient Boost Algorithm is calculated in order to carry out the project. By taking into account additional classifiers like Gaussian Naive Bayes and Gaussian Process Classifiers, the project can be further expanded implementing it with real time data set.

REFERENCE

[1] C.Lianetal:"Workloadperformancepredictionfordatacenters: Asurvey"(2018), IEEEAccess.

[2] M.A.Khanetal:"Areviewofmachinelearningtechniquesforperformancepredictionincloud computing"(2018), IEEE Access.

[3] F.Wangeta l:"A comparison of workload prediction methods in a cloud data center"(2017), IEEE Access.

[4] S.Wangetal."Multi-step- head work load prediction using deep learning for cloud datacenters" (2019), IEEE Access.

[5] X.Liuetal:"Performance prediction for cloud computing: A review of current research trends" (2020) ,IEEE Access.

[6] Johnson, E. E. (2009). Performance envelope of broadband HF data waveforms. *MILCOM 2009 - 2009 IEEE Military Communications Conference*. https://doi.org/10.1109/milcom.2009.5379871

[7] Leung, K. K. (2002). Load-dependent service queues with application to congestion control in broadband networks. *Performance Evaluation*, *50*(1), 27-40. https://doi.org/10.1016/s0166-5316(02)00045-7

[8] Load-dependent relationships between frontal fNIRS activity and performance: A data-driven PLS approach. (n.d.). https://doi.org/10.37473/dac/10.1101/2020.08.21.261438

[9] Middleton, C. A., & Ellison, J. (2021). Understanding internet usage among broadband households: A study of household internet use survey data. https://doi.org/10.32920/ryerson.14639490

[10] Rani K. R., U., Ravishankar, S., & Mahesh, H. M. (2011). Analysis of noise and load effects on broadband performance over residential power lines employing VDSL2. *2011 Annual IEEE India Conference*. https://doi.org/10.1109/indcon.2011.6139421