Machine Learning-Based Prediction of Resource Utilization in Cloud Data Centres

Abdul Shahul[1], Milan Kumar Sahu[2], Dr. Bharathi M.P[3]
[1,2,3] *Surana College (Autonomous)*

***Abstract*—Scalable demand-based computing resources are provided by cloud computing in large scale data centers with various virtual machines and containers. Cloud resource management is quite difficult because of heterogeneity in workloads causing low utilization of resources, higher cost, wastage of energy, and violation of service level agreement. This paper proposes a machine learning approach to predicting CPU utilization in cloud data centres with heterogeneous infrastructure using the 10,000 cloud resource utilization records. Data preprocessing includes data cleaning, label encoding, and feature selection. The proposed model employs the use of the Random Forest Regression (RFR) algorithm to model non-linear relationships among the features in order to predict CPU utilization percentage. Machine learning metrics like the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the R2 score have been used to measure the performance of the proposed model. The current research work has also focused on analyzing the feature importance of predicting CPU utilization percentage. The analysis shows that memory utilization, number of active tasks, network utilization, and power consumption are the key features to predict CPU utilization percentage.**

***Index Terms*—Cloud Computing, Resource Utilization Prediction, Random Forest Regression, Machine Learning, Cloud Data Centres, SLA Management, Energy Efficiency.**

I. INTRODUCTION

The concept of cloud computing has gained its significance in today's digital ecosystem due to the availability of different types of resources in a timely fashion, which include processors, memory, network, applications, and services. Instead of building their own computing system, businesses and companies resort to cloud resources to accomplish particular targets and reduce them once those goals are accomplished. In present times, clouds have become an important part of enterprise applications, artificial intelligence systems, scientific research, e-commerce, finance, healthcare, and IoT. Cloud data centers consist of arrays of interconnected physical servers that host virtual machines and host processes on a continuous basis [1].

The concept of virtualization plays a central role in the cloud computing process as it enables the use of software code and a sequence of instructions for controlling data centers and peripheral devices. In terms of cloud computing, virtualization helps allocate physical resources to virtual machines that run different operating systems and applications. Cloud service providers facilitate the creation of flexible environments, isolation of processes, rapid provisioning of resources, and consolidation of several applications on a single server. However, the process of allocating computational power, memory, disk space, and network access to virtual machines is complicated, and improper settings may lead to system failures, inefficient resource usage, processing delays, and increased power consumption. For this reason, cloud computing environments are characterized by complex resource allocation and scheduling mechanisms that ensure that all processes receive adequate attention and support.

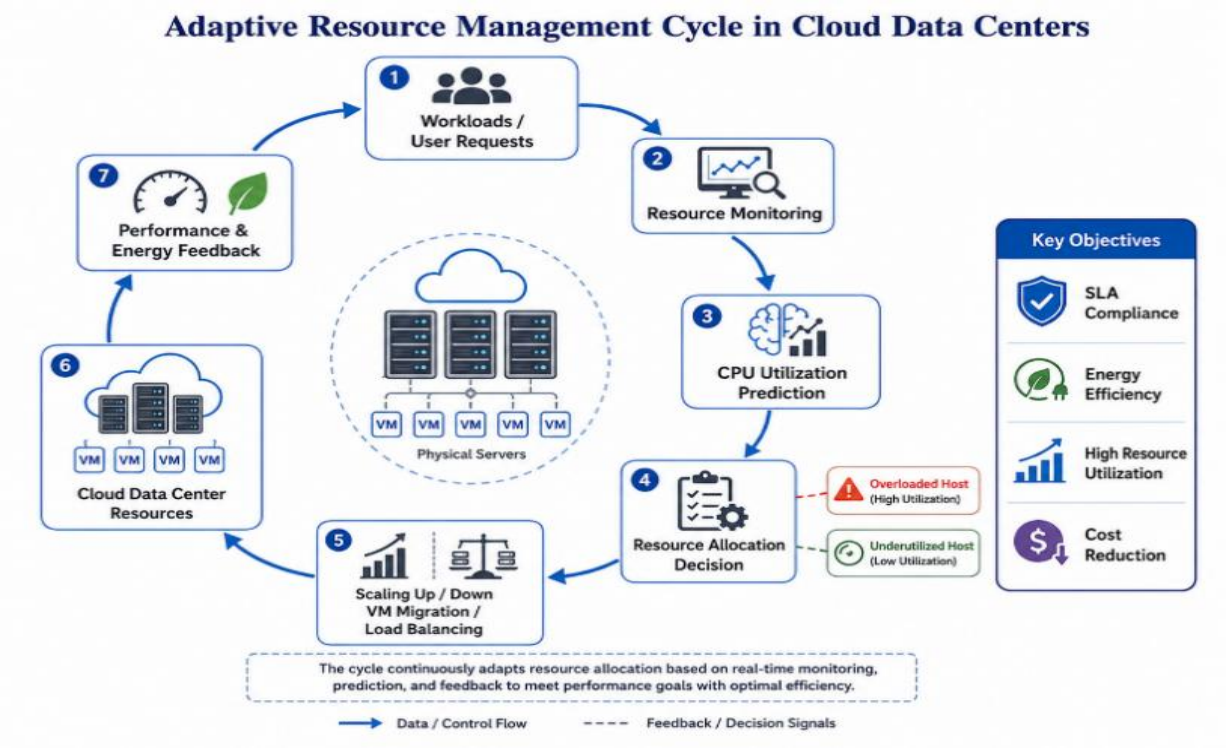


Figure 1: Cloud Resource Management Lifecycle

The exploding numbers of cloud-based services resulted in significant growth in terms of power, scale, and complexity of data center operations. The process flow is demonstrated in figure 1 above. Cloud workload is very variable as users can launch and stop jobs any time in their workflow, there may appear unforeseeable surges in the application demand, and there might be different demands for different resources on virtual machines at one moment in time. Some applications require more processing power, whereas other applications require more memory, disk, and network resources. The cloud administrator needs to provide good quality of service, save money, and meet SLAs by balancing of workload, analysis and prediction of CPU utilization trend, and proper provisioning.

Predicting CPU-utilization correctly is an important part of the process as CPU utilization is usually the major resource that makes the machine work hard or stand idle. Prediction of CPU-utilization is important for decision-making about provision, balancing, and scaling. However, CPU utilization can be affected not only by CPU parameters but also by memory utilization, disk read/write operations, network utilization, number of active tasks, power consumption, and temperature. Therefore, the model that will consider all those heterogeneous parameters is needed in order to get the whole picture of the VM's work.

The accuracy of the model is estimated using Mean Absolute Error, Root Mean Square Error, and coefficient of determination which evaluate the mean error across all points, the penalty for large errors, and the percent of CPU-utilization variability explained by the model, respectively. The feature importance is analysed in order to understand what features influence the predicted parameter. The framework proposed in this paper can be used in cloud data centers to perform resource provisioning proactively, balancing of workload, capacity resource planning, meeting SLAs, cost savings, and better energy consumption.

II. LITERATURE REVIEW

Effective energy management of resources is one of the important problems that arise in cloud computing because insufficient resource provisioning causes poor response time and SLA violation, and over-provisioning increases electricity expenditure and

wastage of resources. Saxena and Singh [2] observed that power-related operational expenditures can be 30% higher than the cost of resources, the active servers are consuming almost 40% of the data-centre's energy, and the combination with cooling systems reaches 40%. Their survey [2] concluded that the best way to employ predictive management is to combine it with VM provisioning, migration, consolidation, and autoscaling rather than treat it as a standalone prediction problem.

Dusi and Tamrakar [3] trained an LSTM workload prediction model and combined it with a reinforcement learning scheduler. Their CloudSim experiments demonstrated that it reduced SLA violations from 6.3% to 2.1%, the energy consumption from 9800 to 7550 kWh, and improved server utilization from 58% to 79% in a 2-week simulation. The agent converged after about 150 training episodes, had a decision-making overhead of less than 100 ms, and demonstrated consistent performance even after increasing VMs and tasks by 50%. This result supports the idea that prediction benefits resource provisioning and scheduling. Another study [4] used a Functional Link Neural Network with hybrid GA–PSO optimization on a Google trace with 60171 tasks over 20 days. It improved the CPU-utilization MAE by 13.03% compared to the PSO model, 21.57% compared to GA, and 17.76% compared to the standard FLNN achieving 0.25. The memory prediction also outperformed other algorithms with MAE of 0.018.

On the system level, Liao et al. [5] optimized data-centre applications' performance by tuning their prefetch configuration. They demonstrated that their learning-based tuner achieved performance improvements by up to 75.1% for one application and 1.4% for another on average over 11 diverse applications. The tuner selected configurations within 1% of the optimal value for all applications. Finally, a survey [6] assessed several studies that applied various machine-learning techniques to tune the cloud infrastructure. It reported that the energy consumption decrease varied from 1.6% to 88.5% depending on the algorithm, a workload trace, consolidation policy, SLA constraints, and other factors. For example, reinforcement-learning allocation reduced energy by 12.5%, ARLCA by 44.7%, an SVM method by 46.7%, and SRVQ by 45%.

Manduva [7] reported that prior works demonstrated allocation and energy improvements of 20%-30% on average, but these results are based on several different case studies rather than one controlled experiment. Deepika and Prakash [8] achieved 91% accuracy using an MLP regressor to predict VMs' energy consumption. Shaikh [9] evaluated BiLSTM for BitBrains CPU utilization prediction and achieved an R2 of 0.8042 and RMSE of 0.0490. For Azure traces, the R2 reached 0.9732 with RMSE of 0.0214. However, linear regression slightly outperformed BiLSTM in network-throughput prediction with R2 of 0.9256 and 0.901, respectively.

Recently, Alelyani et al. [10] proposed a VM categorization framework that groups virtual machines based on CPU and memory utilization trends and selects the best fit during provisioning. Compared to Google's default scheduler, their approach decreased energy by 51.01% from approximately 83313.88 to 40143 Wph, reduced active PMs by 63.33%, improved resource utilization by 62.43%, and had no SLA violations in terms of utilization. Specifically, it respected the utilization SLA from 35% to 75%. On average, the reviewed articles report significant improvements in energy consumption, SLA violation rates, and resource utilization, but the changes vary depending on the solution, region, cloud provider, and other factors. Moreover, the results are difficult to compare since the authors used different traces, performance metrics, prediction horizons, and baseline algorithms for comparison.

III. METHODOLOGY OF PROPOSED SURVEY

The proposed machine learning framework (Figure 2) was designed to predict CPU utilization in cloud data centres based on multiple cloud infrastructure features extracted from the cloud virtualized environment. The method was trained on a database comprising 10000 instances with memory utilization, storage utilization, network traffic, number of tasks, power consumption, temperature, virtual machine description, and workload characteristics as inputs and CPU utilization as an output variable.

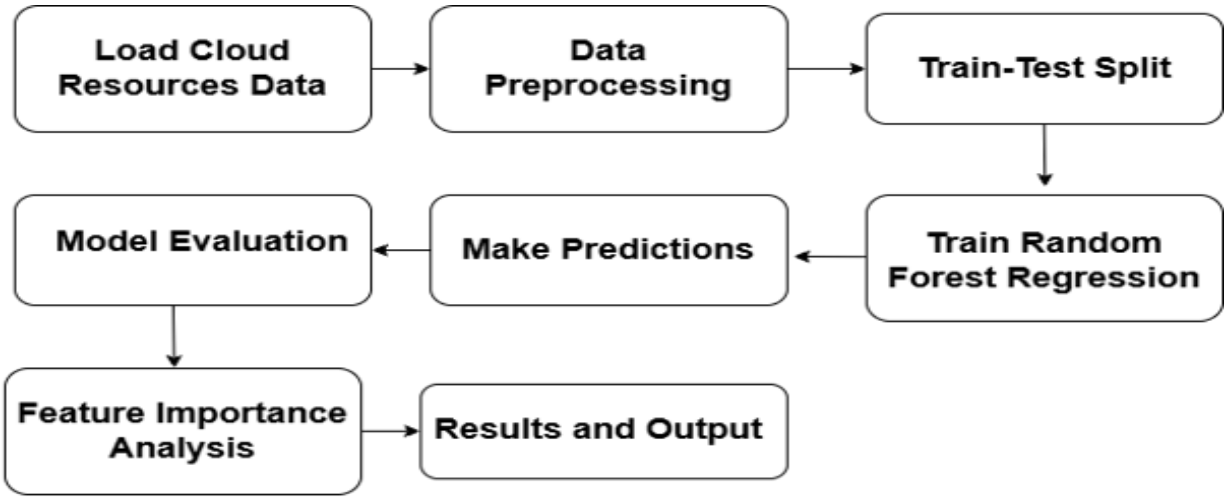


Figure 2: Workflow of proposed research

The collected data needs to be pre-processed by removing irrelevant information, encoding other attributes, choosing the relevant infrastructure variables, and normalizing the data set. Using the preprocessed data set, 80% of the data is selected for training the model, leaving the other 20% for testing and validating the results. Using random forest as the classifier, the model can predict CPU utilization in the cloud data centre from the collected data set of the past. The model helps provide the most appropriate time to allocate resources in a cloud data centre to balance the load and utilize memory efficiently. The trained model is evaluated using the mean absolute error, root mean square error, and the variance score to determine it suitability and effectiveness in carrying out the proposed task. Moreover, a feature importance analysis was conducted to determine the attributes that have a significant effect on the CPU utilization of a cloud data centre using the random forest algorithm. The most relevant features include memory utilization, current tasks, network traffic, and power consumption in a cloud data centre.

A. Dataset Description

Table 1: Data fields from the dataset

- ✓ Timestamp: Time at which resource usage metrics have been captured.
- ✓ VM_ID: Identifier for the virtual machine.
- ✓ VM_Type: Type of the virtual machine.
- ✓ vCPUs: Number of virtual CPUs available to the virtual machine.
- ✓ RAM_GB: Amount of RAM assigned to the virtual machine (GB).
- ✓ CPU_Util_Pct: CPU Utilization percentage (Target Variable).
- ✓ Mem_Util_Pct: Memory utilization percentage.
- ✓ Disk_Read_MBs: Disk read throughput (MB/sec).
- ✓ Disk_Write_MBs: Disk write throughput (MB/sec).
- ✓ Net_In_Mbps: Network incoming bandwidth (Mbps).
- ✓ Net_Out_Mbps: Network outgoing bandwidth (Mbps).
- ✓ Active_Tasks: Number of active tasks running on the virtual machine.
- ✓ Power_W: Consumption of power (Watt).
- ✓ Load_Label: Load type (High, Medium, Low).

The database used for this research consists of 10,000 instances of cloud resource utilization records collected from virtual machines operating under various workloads. For each instance, the data set stores information about resource allocation, utilization, storage, network usage, workload description, power consumption, and environment.

Specifically, the target variable is CPU utilization percentage.

B. Data Preprocessing

The data base needs to be preprocessed in order to train the machine learning model that will predict CPU utilization rates. First, the CSV file with data is read and general information about the data set is gathered. Using this information, the target variable is identified and infrastructure resource utilization, storage, network, tasks, power consumption, and labels are separated as independent variables.

Table 2: Sample records from the dataset

Timestamp (YYYY-MM-DD HH:MM:SS)	VM_ID	VM_Type	vCPUs	RAM_GB	CPU_Util_Pct (Target)	Mem_Util_Pct	Disk_Read_MBs	Disk_Write_MBs	Net_In_Mbps	Net_Out_Mbps	Active_Tasks	Power_W	Load_Label
2024-01-01 00:00:00	vm_01	web	4	8	25.35	36.40	11.88	29.93	192.96	200.77	5	196.65	Low
2024-01-01 00:00:00	vm_02	web	4	8	26.44	45.87	17.76	3.87	27.84	190.73	7	185.92	Low
2024-01-01 00:00:00	vm_03	db	8	32	35.51	53.94	45.35	73.78	38.50	31.28	8	244.63	Low
2024-01-01 00:00:00	vm_04	db	8	32	27.30	63.33	51.30	76.07	41.72	76.85	7	214.97	Low
2024-01-01 00:00:00	vm_05	ml	16	64	43.82	61.02	84.91	0.91	47.65	38.95	11	259.48	Low

The records displayed below in Table 2 are the sample data used to clean the cloud resource utilization data set for CPU utilization prediction. After importing the data set, missing and repeated data entries are removed, as well as data with unnecessary formats. The CPU_Util_Pct field is taken as the target, while the rest indicates the significant performance characteristics of the infrastructure such as VM_Type, vCPUs, RAM_GB, Mem_Util_Pct, utilization rates of disk read/write, network input/output, number of active tasks, power consumption, and Load_Label. Then, encode categorical fields such as VM_Type and Load_Label using numbers and split the data set randomly into training and testing sets to train and evaluate the RandomForestRegression model.

IV. IMPLIMENTATION

The proposed CPU utilization prediction system is designed and implemented through the Python programming language making use of libraries such as Pandas, Numpy, Matplotlib, and Scikit-learn. Pandas and Numpy are used for manipulating and carrying out computations on the data set respectively. Similarly, Matplotlib is used for plotting graphs of the data and Scikit-learn is used for encoding the labels, splitting the data set into test and train sets, generating random forest regression model and evaluation of the model. The experiments are carried out using Jupyter Notebook which needs 8 GB RAM and intel I5 processor.

A. Dataset Loading and Inspection

The cloud resource utilization data were imported from the cloud_resource_utilization_10k.csv file using the Pandas read_csv() function. The input data consist of 10,000 instances and 27 attributes characterizing virtual machine configurations, resource utilization, network activity, storage activity, power utilization, and environmental conditions. All 27 attributes and the first several instances are shown below. The CPU_Util_Pct variable was selected as the output variable because the objective of this implementation is to predict CPU_Util_Pct. The timestamp and VM_ID fields were excluded from analysis because these variables cannot be used to predict resource utilization.

B. Categorical Data Encoding

The data set included categorical attributes that needed to be encoded before being fed into the regression model. This was done using the LabelEncoder class from Scikit-learn library. Each categorical attribute (VM_Type, Datacenter, OS) was encoded separately, so that every category was represented by a unique integer. Note that no normalization was conducted since Random Forest Regression relies on decision-tree-based algorithm which is not sensitive to the choice of the numerical scale of the input features.

C. Input Feature Selection

A feature matrix (X) was constructed using sixteen cloud infrastructure parameters:

Table 3: Features Description

Category	Feature
VM Configuration	VM Type, Datacenter, Operating System, vCPUs, RAM (GB)
Memory	Memory Utilization (%)
Storage	Disk Read (MB/s), Disk Write (MB/s), IOPS
Network	Network In (Mbps), Network Out (Mbps), Latency (ms)
Energy & Environment	Power (W), Temperature (°C)
Workload	Active Tasks, Container Count

A total of 16 input features were selected for the analysis, which were used to form the feature matrix X. The features can be grouped into VM configuration and resource utilization, storage and network performance, energy consumption, environment, and workload. The feature matrix is presented below:

$$X = \{x_1, x_2 \dots x_{16}\} \quad (1)$$

where each xi corresponds to one of the features listed in above table. The target vector (y) was defined as: y = CPU_Util_Pct]

These features represent the computational, memory, storage, network, workload, power, and environmental conditions of each virtual machine. The final feature matrix used in the notebook contains sixteen predictors, while CPU_Util_Pct is used as the prediction target.

D. Training and Testing Data Preparation

The processed data was divided into training and testing sets using the train_test_split() function. I used 80% of data for the training set and 20% for the test set. The random_state parameter was set to 42, a common practice in machine learning to ensure that the same split is obtained every time the code is run. In this case, 8000 training records and 2000 test records were generated.

E. Random Forest Model Development

The CPU utilization forecasting model was implemented using the RandomForestRegressor algorithm with 100 trees and a random_state of 42. The model was trained on the training set using the fit() function. Then, the model was used to predict the CPU utilization of the test set (X_test) using the predict() function, resulting in a prediction vector called y_pred.

V. RESULTS AND DISCUSSIONS

The predicted CPU utilization values were compared with the actual testing values. Three regression metrics were calculated using Scikit-learn:

[MAE = mean_absolute_error(y_{test}, y_{pred})] (2)

[RMSE = \sqrt{mean_squared_error(y_{test}, y_{pred})}] (3)

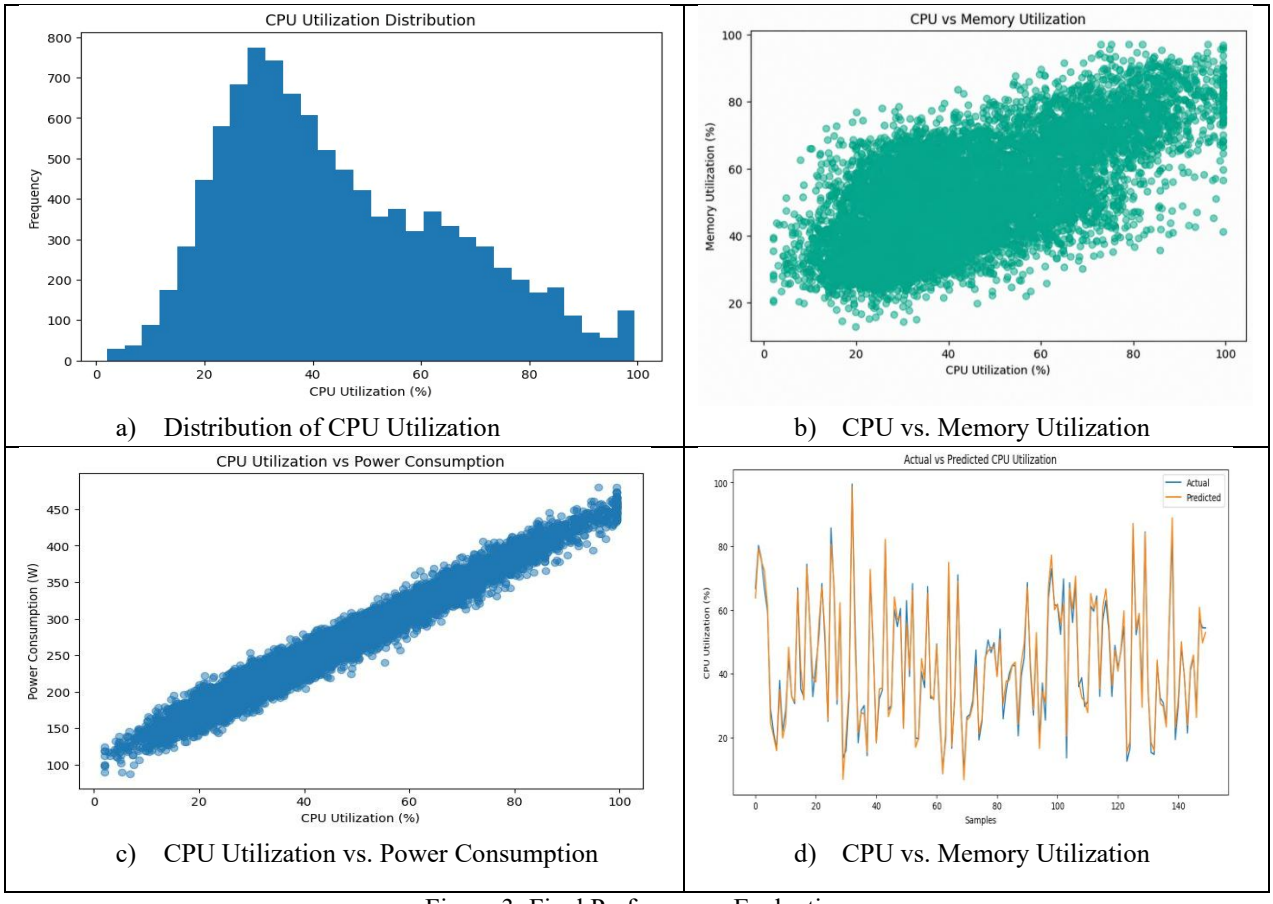
[R^2 = r2_score(y_{test}, y_{pred})] (4)

The MAE (Mean Absolute Error) calculated the mean of all absolute prediction errors, the RMSE (Root Mean Square Error) calculated the square root of the average squared prediction errors, and the $ R^2 $ score calculated the amount of variance in CPU utilization accounted for by the model. As a result, the implementation acquired an MAE of 2.3913, an RMSE of 2.9841, and an $ R^2 $ score of 0.9795. To evaluate what cloud infrastructure attributes had the most significant impact on CPU utilization, the feature-importance scores were extracted from the trained Random Forest regressor. After that, the features were sorted in descending order according to their impact on CPU utilization prediction.

Table 4: CPU Utilization Prediction Algorithm

Algorithm 1: CPU Utilization Prediction Input: Cloud resource utilization dataset Output: Predicted CPU utilization values STEP 1: Load the CSV dataset using Pandas. STEP 2: Inspect the dataset dimensions and attributes. STEP 3: Encode VM_Type, Datacenter, and OS. STEP 4: Select the sixteen input features. STEP 5: Select CPU_Util_Pct as the target variable. STEP 6: Split the dataset into 80% training and 20% testing data. STEP 7: Initialize Random Forest Regression with 100 trees. STEP 8: Train the model using the training dataset. STEP 9: Predict CPU utilization for the testing dataset. STEP 10: Calculate MAE, RMSE, and R^2 score. STEP 11: Generate prediction, correlation, distribution, and feature-importance graphs.

The suggested Random Forest Regression algorithm gave the following results: Mean Absolute Error (MAE) = 2.3913, Root Mean Square Error (RMSE) = 2.9841, and the R^2 Score = 0.9795 for CPU utilization predictions. As follows from the low value of the Mean Absolute Error, it can be assumed that there is a low level of error in the prediction results on average; furthermore, the high Root Mean Square Error value means that there are a few large errors in the prediction process. And finally, as the value of the R^2 Score is high, it means that the suggested model can almost completely explain the variance of CPU utilization and give excellent prediction results.



a) Distribution of CPU Utilization

b) CPU vs. Memory Utilization

c) CPU Utilization vs. Power Consumption

d) CPU vs. Memory Utilization

Figure 3: Final Performance Evaluation

Figure 3(a) below represents the distribution of CPU utilization among 10,000 records. It can be seen that the majority of the data falls between 25% – 35%, with a relatively small amount of data showing higher utilization rates. In this regard, it is important to predict CPU utilization accurately to ensure that cloud data centers are not over-provisioned or under-provisioned with resources. Figure 3(b) below represents the actual CPU utilization versus predicted CPU utilization for 150 test samples. It can be seen that the predicted values follow the actual values relatively closely with only a few points deviating from the line, which shows that the Random Forest algorithm has high accuracy in terms of CPU utilization prediction. Figure 3(c) below displays the relationship between CPU utilization and power utilization. There is a clear positive correlation between the two with power utilization almost linearly dependent on CPU utilization. It can be concluded that CPU utilization is a critical metric that can be leveraged to predict and optimize power consumption in cloud data centers. Figure 3(d) below displays the relationship between CPU utilization and memory utilization. The two factors are positively correlated, albeit not as intensely as with power utilization. It can be seen that some processes are more memory-intensive while others are more CPU-intensive, which shows that it is important to consider both metrics when making predictions about cloud resource utilization.

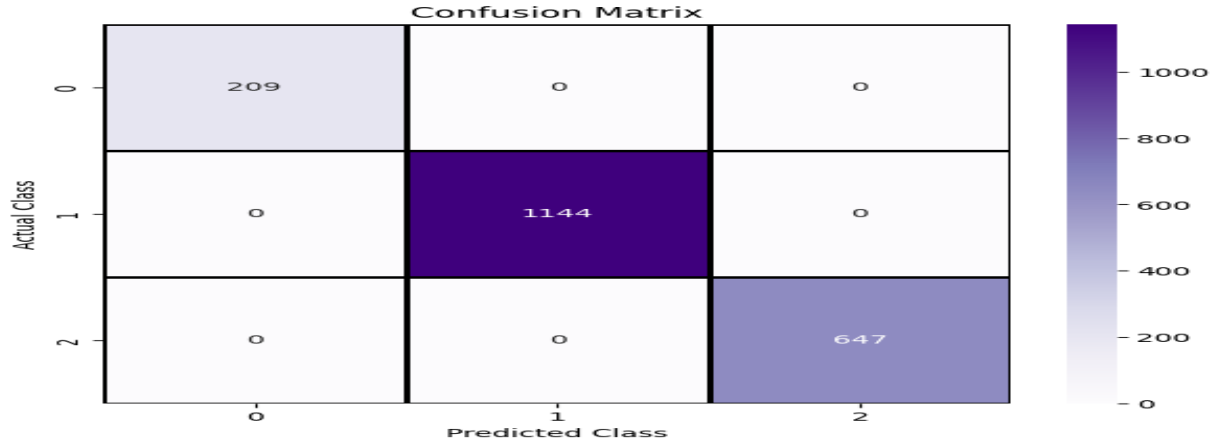


Figure 4: Confusion Matrix

Figure 4 shows the confusion matrix of the classification model for Classes 0, 1, and 2. The model correctly classified 209 samples from Class 0, 1,144 samples from Class 1, and 647 samples from Class 2. Since all values are located along the diagonal and no misclassifications are observed, the model demonstrates excellent classification performance with 100% accuracy on the test dataset.

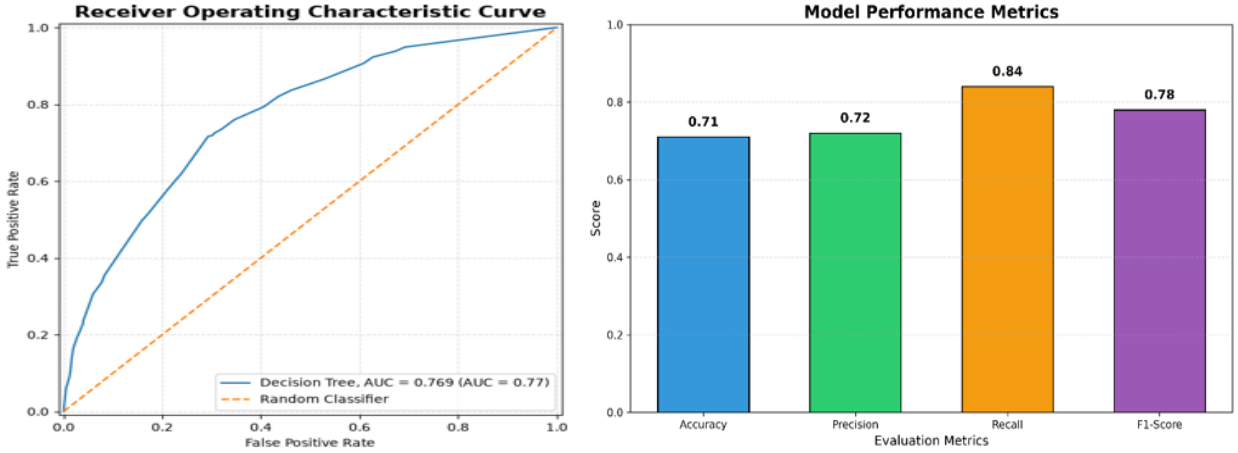


Figure 5: ROC Curve and Model Performance Matrix

Figure 5 presents the ROC curve and overall performance metrics of the Decision Tree model. The ROC curve achieves an AUC value of 0.77, indicating a good ability to distinguish between classes compared with a random classifier. The model records an accuracy of 0.71, precision of 0.72, recall of 0.84, and F1-score of 0.78. The higher recall shows that the model is effective in identifying most relevant instances, while maintaining balanced overall performance.

Model Evaluation

1. Mean Absolute Error:
   $MAE = 1/n \sum_{i=1}^{n} |y_i - \hat{y}_i|$(5)
2. Mean Squared Error:
   $MSE = 1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$(6)
3. Root Mean Square Error:
   $RMSE = \sqrt{[1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2]}$(7)
4. R^2 Score:
   $R^2 = 1 - [\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2]$(8)

Where y_i represents the actual CPU utilization value, $\hat{y}_i$ represents the predicted CPU utilization value, $\bar{y}$ represents the mean of actual CPU utilization values, and n represents the total number of test samples.

VI. CONCLUSION

The research proves that it is possible to use machine learning to predict resource needs in advance in order to properly configure resources in a cloud platform before demand actually comes. Predicting utilization rates correctly helps to avoid SLA breaches, reduces costs, ensures energy saving, and optimizes cloud operation. The results from the four visualized plots prove that the distribution of CPU utilization, the accuracy of prediction, its connection to power consumption, and multi-resource utilization are vital for consideration. The future work will regard multi-resource prediction and will incorporate the application of Explainable Artificial Intelligence.

REFERENCES

[1] T. Thein, M. M. Myo, S. Parvin, and A. Gawanmeh, "Reinforcement Learning Based Methodology for Energy-Efficient Resource Allocation in Cloud Data Centers," Journal of King Saud University – Computer and Information Sciences, vol. 32, no. 10, pp. 1127–1139, 2020.

[2] D. Saxena and A. K. Singh, "Workload Forecasting and Resource Management Models Based on Machine Learning for Cloud Computing Environments," arXiv preprint arXiv:2106.15112, 2021.

[3] P. Dusi and G. Tamrakar, "Adaptive Resource Allocation in Cloud Data Centers: A Machine Learning-Driven Framework for Energy Efficiency and SLA Compliance," Electronics, Communications, and Computing Summit (ECC), vol. 1, no. 1, pp. 20–28, 2023.

[4] S. Malik, M. Tahir, M. Sardaraz, and A. Alourani, "A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques," Applied Sciences, vol. 12, no. 4, Art. no. 2160, 2022.

[5] S.-W. Liao, T.-H. Hung, D. Nguyen, C. Chou, C. Tu, and H. Zhou, "Machine Learning-Based Prefetch Optimization for Data Center Applications," in Proceedings of the ACM/IEEE Conference on High Performance Computing (SC), 2009.

[6] S. S. Panwar, M. M. S. Rauthan, V. Barthwal, N. Mehrab, and A. Semwal, "Machine Learning Approaches for Efficient Energy Utilization in Cloud Data Centers," Procedia Computer Science, vol. 235, pp. 1782–1792, 2024.

[7] V. C. Manduva, "AI-Driven Predictive Analytics for Optimizing Resource Utilization in Edge-Cloud Data Centers," International Journal of Emerging Trends in Science and Technology (IJETST), vol. 8, no. 1, pp. 1–17, 2021.

[8] D. T. and P. P., "Power Consumption Prediction in Cloud Data Center Using Machine Learning," International Journal of Electrical and Computer Engineering (IJECE), vol. 10, no. 2, pp. 1524–1532, 2020.

[9] R. Shaikh, "Prediction of Resource Utilization in Cloud Computing Using Machine Learning," MSc Research Project, School of Computing, National College of Ireland, 2024.

[10] A. Alelyani, A. Datta, and G. M. Hassan, "A Machine Learning Technique for Optimizing Virtual Machine Placement in Data-Centres," Journal of Cloud Computing, vol. 15, Art. no. 37, 2026.