National College of Ireland

Prediction of Resource Allocation in cloud computing using Machine Learning

MSc Research Project
Data Analytics

Aruna Saravanapandian
Student ID: x22182349

School of Computing
National College of Ireland

Supervisor: Prof. Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Aruna Saravanapandian
Student ID:	X22182349
Programme:	Data Analytics **Year:** 2024
Module:	MSc Research project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	12/12/2024
Project Title:	Prediction of Resource allocation in cloud computing using Machine learning
Word Count:	5077 **Page Count** 18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aruna Saravanapandian
Date:	03/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Resource Allocation in Cloud Computing using Machine Learning

Aruna Saravanapandian
X22182349

Abstract

Cloud computing eliminates the use of physical devices as computational resources are made available virtually, also poses a high demand now due to the increase in necessity for the cloud resources. Cloud resources could cause high costs if they are not used wisely hence, the prediction of the resource demand will help us allocate the resources efficiently. The prediction of the resource demand is analyzed with the various machine learning algorithms like Linear regression, Ridge Regression, Lasso regression, Random Forest Regression, Gradient Boosting Regression and fully connected neural networks. The goal is determining the machine learning algorithm that helps in predicting the resource demand with greater accuracy. The evaluation metrics like root mean square error, mean squared error, mean absolute error and r squared are taken into consideration.

1 Introduction

Cloud computing allows the revolutionary approach to computing by emphasizing the execution of programs over several computers like distributed computing. Resources are shared between economies leading to the rise of shared services and integrated infrastructure. Cloud computing utilizes parallel and distributed computing comprising virtual resources. The resources are made available on demand and are cost-effective depending upon the usage. The data centres work as cloud providers and are provided by companies such as Amazon, Google, Microsoft etc. (Chaudhary and Banga, 2024).

The resource provisioning systems, when they are proactive, have a better response time thereby reducing the overall cost. This can be achieved by creating an efficient resource demand prediction model. The future CPU request, when predicted in advance, will help us allocate resources based on demand and the virtual machines and the number of CPU's required can be allocated accordingly.

The goal of the research is to determine the machine learning model that is used for predicting the CPU request that the machine receives. This system will decrease the cost by predicting the resources that would be required thereby avoiding the CPU overload and allocating the resources beforehand. The machine learning models that are Linear regression, Ridge Regression, Lasso regression, Random Forest Regression, Gradient Boosting Regression, Decision Tree Regression and Artificial Neural Networks can be used for analyzing the data and determining the more accurate regression model. The questions that are addressed through this research are,

1. How can machine learning models be used for determining the resource demand based on the data available?
2. What degree of predictability can be achieved with the machine learning models?

The objective of the research is to analyse the data of Google cluster data from 2019 and determine the resource demand depending on the data. The overall goal is to apply the machine learning models to determine the model that is highly accurate and evaluate the effectiveness of the model. The research would help in reducing the cost while using the cloud resources and avoid overloading of the VM's.

The paper is structured as shown below. The research objective and the environment are described in the section. 1. We examine the related literature on the resource demand prediction in Section 2. The research methodology and its phases are described in Section 3. Section 4 includes the design specification for the study. The implementation of the research is included in Section 5. The evaluation and findings are added in Section 6. The conclusion is present in Section 7 along with future research. The references are present at the end of the research.

2 Related Work

In this section, we will understand the various papers that help us understand the various approaches in which resource utilization is predicted. This section is broken down into the following subsections: 2.1 Prediction of resource utilization using machine learning; 2.2 Prediction of resource utilization using deep learning.

2.1 Prediction of resource utilization using machine learning

The research (Mehmood, Latif and Malik, 2018) recommends the creation of the resource allocation strategy that allows the resource allocator to allocate the resources based on the need thereby it doesn't allow the over utilization of the resource. An Ensemble model implements the mechanism Over Produce and Choose (OPC). The model includes a stacked K-Nearest Neighbour and Decision Tree at the base level followed by the Decision Tree at the first level. This technique is found to have increased the prediction accuracy by 2 percent. The resource utilization was studied using various machine learning models like Linear regression, Lasso regression, ridge regression, random forest regression, gradient boosting regression, decision tree regression. Another research (Mashhadi Moghaddam et al., 2020) consolidated the Virtual Machine thereby predicting the resource utilization using the various machine learning models like Support vector regression, decision tree regression, boosting decision tree regression. The Virtual Machine with overutilization is predicted using the machine learning models and the over utilized hosts are migrated. The study (Ma et al., 2023) focuses on predicting the server's resource utilization where the pending tasks are calculated along with the number of tasks. The Random Forest regression is used for building the model. The research (Baig et al., 2019) allows the accurate estimates of the resource's utilization that allow the user to efficiently allocate Virtual Machines (VM) as well as other resources to the workloads which is performed using various machine learning models like Linear Regression (LR), Support Vector Machines (SVM), Gradient Boosting Tree (GBT) and Gaussian Process also known as Krigin (KR). The best model for prediction is found to be Gradient Boosting Tree with lowest Root Mean Square Error of 9.74 and the Mean Absolute Error of 2.85.

2.2 Prediction of resource utilization using deep learning

The ability to predict is enhanced by neural networks which improves the scope of the responses from the different possibilities and applying the potential answers. The inaccuracy due to local optimum is avoided as opposed to the gradient-based methods. Convoluted Neural Networks (CNN), Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM) neural networks are the deep learning methods used for this purpose as studied (Simaiya et al., 2024). The research (Hong et al., 2023) talks about the nerd for the prediction of resource allocation that in turn improves the efficiency of energy. The accuracy is achieved using the cloud resource agent the acts depending on the state. The Prediction accuracy of the model is 0.645, 0.624, and 0.460, respectively, in the order of CNN, LSTM, and DNN, and it is confirmed that the model applying the CNN neural network has a higher prediction accuracy average by 3.36% than the model of LSTM and 40.2% than the model of DNN. The RNN based study (Purohit, Srivastava and Pandey, 2023) uses the hidden layer output from the previous step to effectively process the data. The historical data is obtained and pre-processed, the RNN model is trained, and the performance is evaluated. The dataset used here contains 40,000 entries and will be used to train the RNN. The RNN algorithm is chosen for its ability to learn and improve with more data. The values obtained after evaluation are RNN with MSE of 4.529*10-09 and RMSE of 6.73009*10-05, followed by MLP with MSE of 3.925*10-06 and RSME of 0.0019, LSTM with MSE of 4.031*10-08 and RSME of 0.00020 where RNN has the lowest MSE and RSME.

3 Research Methodology

The research methodology follows the KDD (Knowledge Discovery from Databases) methodology Figure 1. It includes the six stages like data selection, preprocessing of data, transformation of data, Data mining, Interpretation and evaluation of data.

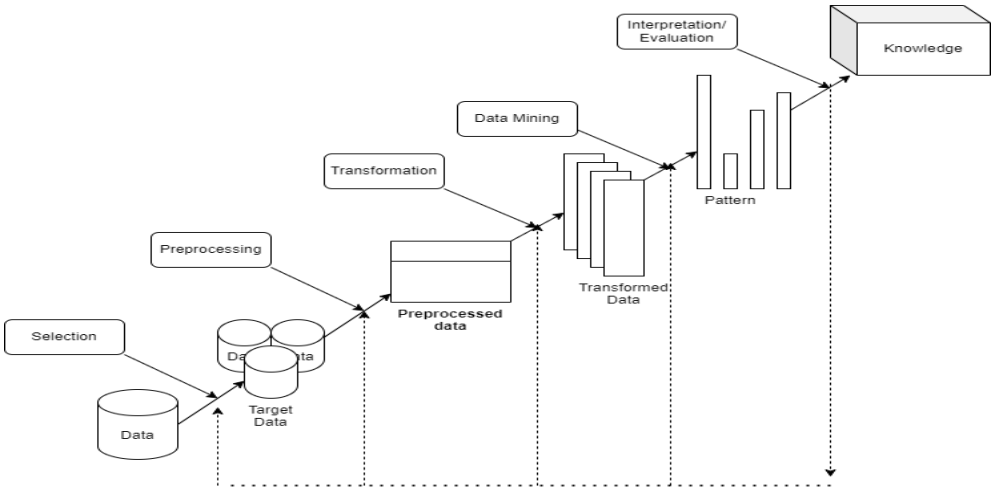

Figure 1: KDD Methodology

3.1 Data Selection

The dataset is obtained from Kaggle * (Google 2019 Cluster sample). The data is a trace of workload running on eight Google Borg compute clusters for the month of May 2019. The trace describes every job submission, scheduling decision and resource usage data for the jobs that has been running in the cluster. The dataset has 405894 observations and 31 feature variables. The variable resource_request_cpu is the target variable and the other variables are the predictor variables.

3.2 Exploratory Data Analysis

Exploratory Data Analysis refers to statistical methods of analysing the data sets to summarize their main characteristics. It is carried out using the statistical graphics and other data visualization techniques. The Figure 2 shows the distribution of the numeric variables in the data frame. The variables like instance_events_type, scheduling_class, priority, vertical scaling, scheduler, sample_rate, cluster are distributed evenly across the graph while the variable assigned_memory, page_cache_memory are skewed towards zero which suggests the minimal usage.

The missing values are to be determined to evaluate the data and process them. If there are lots of missing values in the dataset then there would be uncertainty in the output that is produced. The missing value is visualised in the Figure 3, the values with the large number of missing values like cycles_per_instruction, memory_accesses_per_instruction are removed from the dataset and the remaining data are used for further analysis. If the missing data is greater than 75 percent than the column with the high amount of missing data can be removed.

3.3 Data cleaning and preprocessing

Data pre-processing of the data allows the data to be cleaned and prepared and the target variable and the feature required in predicting the target variable are determined. In this case, the target variable is determined to be the resource_request_cpu, but this value is present as the part of the resource_request variable along with the memory. The memory and the CPU values are present together with the following variables, average_usage, random_sample_usage, maximum_usage and resource_request. These variables are analyzed and the data is split to the average_usage_cpu, average_usage_memory, random_sample_usage_cpu, random_sample_usage_memory, maximum_usage_cpu, maximum_usage_memory and resource_request_cpu and resource_request_memory are obtained and are available for further processing.

The outliers are determined for the target variable as shown in Figure 4, it is essential to remove outliers in the target variable as the outliers cause a skew in the result. The outliers may be due to measurement errors, processing errors or poor sampling. The filter is set up to remove the resource_request_cpu value greater than 0.025.

* https://www.kaggle.com/datasets/derrickmwiti/google-2019-cluster-sample/data

Figure 2 : Histogram the distribution of the numeric variables in the data frame

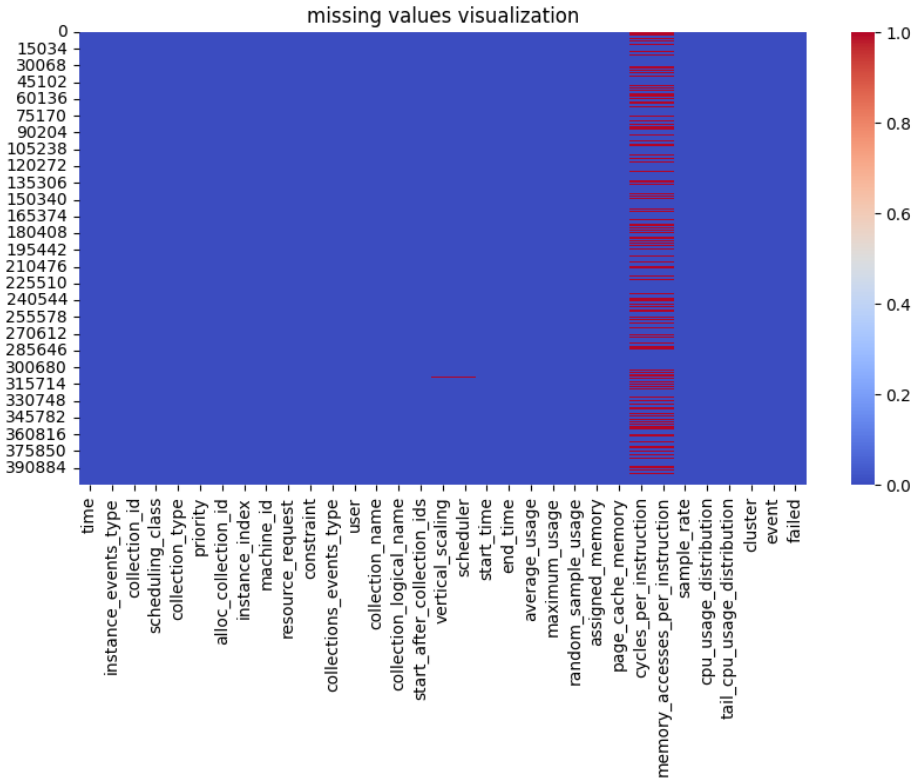

Figure 3 : Visualization of Missing values

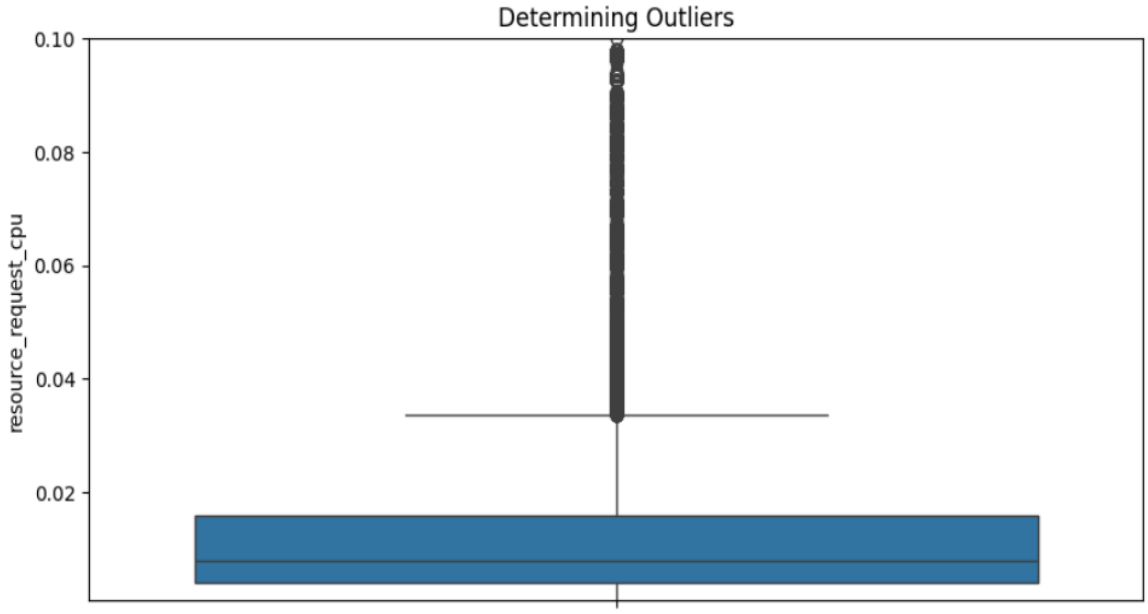

Figure 4 : Determining the outliers

3.4 Data transformation

Data transformation involves transforming the data into a form suitable for analysis. The cleaning, restructuring and the missing values are handled in this process. The values are handled using the minmaxscaler, the standardization allows the normalization of the values in the preprocessed data. It allows the standardization of the data. The missing values are determined and the columns with more missing values are eliminated. The categorical columns like scheduling_class, priority, machine_id, user are encoded with one hot encoder. The train size of the data is 80 percent while the test size of the data is 20 percent for the model.

3.5 Data Modelling

The machine learning algorithm that is used for predicting the resource demand are as follows,

Linear Regression:

Linear regression is used for predicting the outcome of the variable based on the multiple variables and it is also called multiple regression. The dependent variable is the variable to be predicted, and the explanatory variables are used for predicting the dependent variable. The methodology allows us to determine the model variation and the contribution of the independent variable in the variance.

Ridge regression:

Ridge regression is used for the analysis of multicollinearity in multiple linear regression. The data contains the predictor variables and the number of observations. Ridge regression reduces the standard error by adding bias in the regression estimation. This in turn significantly increases the reliability of estimates.

Lasso regression:

Lasso regression is a type of linear regression that uses L1 regularization to reduce the size of the coefficients in regression model. The coefficients become exactly zero, thereby selecting the subset of features reducing overfitting.

Random Forest Regression:

Random Forest Regression combines the predictions of multiple decision trees to reduce overfitting and improve accuracy. It combines prediction from multiple models to create accurate and stable predictions.

Gradient Boosting Regression:

Gradient Boosting Regression helps in minimizing the bias error of the model. It can handle complex relationships in data protecting against overfitting. The gradient boosting regression works by adding models sequentially with every addition trying to resolve the errors made by previous ones.

Decision Tree Regression:

Decision Tree Regression is a supervised, tree-based machine learning technique used for predicting target variables in a nonlinear manner and is used for predicting continuous target variables. The feature variable with a better split based model is chosen.

Artificial Neural Networks:

The Artificial Neural Network has an input layer as the first layer and an input is received from the external sources and it is moved to the hidden layer being the second layer. The hidden layer, the input from the previous layer is computed as weighted sum and sent to the neuron of next layer. Artificial Neural Network produces a higher r squared values and lower mean squared error.

3.6 Model Evaluation

The metrics that can be used for evaluating the regression models are as follows,

R-Squared:

The R-Squared is known as the coefficient of determination and is a statistical measure that is used for determining the amount of variance in the dependent variable that can be explained by the independent variable. The higher the value of the R-Squared greater the variation is explained by the model.

$$R^2 = \frac{Variance\ Explained\ by\ model}{Total\ variance} \quad (1)$$

Root Mean Square Error (RMSE):

Root Mean Square Error is the calculation of the average distance between the predicted values from the model and the actual values in the dataset. The lower the value of root mean square error higher the fit of the model would be.

$$RMSE = \frac{\sqrt{\Sigma\left(y_i^{\wedge} - y_i\ \right)^2}}{n} \quad (2)$$

Where, Σ is a symbol that means "sum"
$y_i^{\wedge}$ is the predicted value for the ith observation
y_i is the observed value for the ith observation
n is the sample size

Mean Squared Error (MSE):

The Mean Squared Error is the average difference between the predicted values and the actual values in the dataset. The value is calculated by taking the average of the squared residuals where the residual is the difference between the predicted value and the actual values.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - y_i^{\wedge})^2 \quad (3)$$

Where, Σ is a symbol that means "sum"
$y_i^{\wedge}$ is the predicted value for the ith observation
y_i is the observed value for the ith observation
n is the sample size

Mean Absolute Error (MAE):

Mean Absolute Error is the mean absolute difference between the predicted and actual values in the dataset. If the value of mean absolute error is the lowest then the model fits the dataset better than the other models.

$$MAE = \frac{1}{n} \cdot \Sigma|y_i - y_i^{\wedge}| \quad (4)$$

where, Σ is a symbol that means "sum"
y_i is the observed value for the ith observation
$y_i^{\wedge}$ is the predicted value for the ith observation
n is the sample size

4 Design Specification

The initial step of the research involves formulating the hypothesis. The research hypothesis must be followed by the data acquisition process, preprocessing of the data where the data is checked for duplicate values and null values are eliminated. The essential features are identified, and the target variables are determined. Finally, the models are trained and tested with the data using the training and the testing data sets that are about 80 percent of the train data and 20 percent of the test data. Machine learning algorithms such as Linear regression, Lasso regression, Ridge regression, Random Forest regression, Decision tree regression and deep learning algorithm like Artificial neural network.

5 Implementation

The dataset for the research is obtained from the Kaggle an online open-source repository. The csv file borg_traces_data.csv is the source of the data. Python is the programming language used for the implementation of machine learning models. Jupyter notebook is the application that uses the python for the implementation, it is an open-source application that supports the code, narrative text and visualization and a great tool for data analysis. Python supports machine learning algorithms with the extensive library like sckit-learn which formulates machine learning algorithm like Linear regression, Ridge regression, Lasson regression, Decision tree regression and Random Forest regression and the deep learning algorithm utilizes the library TensorFlow.

The dataset that is used for the analysis is present in csv format. The initial step of the analysis includes the data preprocessing steps like data cleaning, eliminating the missing values, and the determination of the outliers. The categorical values are identified, and they are encoded with one hot encoding. The normalization of the target variable is done as it is condensed around a certain value, hence the log transformation is done to ensure a get the best outcomes and prevent the skewing of the data. The total number of variables in the dataset is 33 columns out of which 12 features are selected for the prediction of the resource_request_cpu value. The relationship between the features is determined. The categorical columns are encoded and scaled using MinMaxScaler to disperse the values normalized. Finally, the test and train data are obtained in the manner where 80 percent of the data is used for training and 20 percent is used for testing and evaluated using the machine learning algorithms. The evaluation metrics that are used for determining the best fit model are Root Mean Square Error, R-Squared, Mean Absolute Error, Mean Square Error.

6 Evaluation

Several evaluation metrics such as Root Mean Square Error, R-Squared, Mean Absolute Error, Mean Square Error are used for the comparison of the regression machine learning models. The evaluation metrics are essential to determine the working of the model. The evaluation metrics applied to the research and the findings are detailed below.

6.1 Experiment / Case Study 1: Model performance with all the features

In the first experiment, all the selected features are used to train a base model for all the regressor models. The evaluation was performed after the feature selection and the training and testing of the dataset.

	Model	**R2**	**MSE**	**MAE**	**RMSE**	**Execution Time(sec)**
1	Linear Regression	0.547167	0.000016	0.002853	0.003992	0.90
2	Ridge Regression	0.547167	0.000016	0.002852	0.003992	0.36
3	Lasso Regression	0.425282	0.000020	0.003438	0.004497	1.02
4	**Random Forest Regression**	**0.921258**	**0.000003**	**0.000553**	**0.001665**	**44.35**
5	**Decision Tree Regression**	**0.900158**	**0.000004**	**0.000581**	**0.001874**	**6.40**
6	Gradient Boosting Regression	0.493325	0.000018	0.003274	0.004222	14.60
7	Artificial Neural Network	0.827655	0.000006	0.001404	0.002463	253.92

6.1.1 Linear Regression:

The linear regression model is used for establishing linear relationships between predictor and the dependent variables, the R2 value is 0.547167 indicating the features like scheduling_class, priority, cluster significantly influence the predictor variable resource_request. It has a Mean Squared Error of 0.000016 and an execution time of 0.90 seconds with the model performing average.

6.1.2 Ridge Regression:

When the independent variables are highly correlated, it could lead to imprecise estimates of the regression coefficients. The R2 value of Ridge Regression is 0.547167 and

Mean Squared Error is 0.000016 and the low execution time of 0.36 seconds which is same as the performance of the Linear regression but with a slightly lower execution time.

6.1.3 Lasso Regression:

The overfitting of the variables is avoided by using the Lasso regression technique, shrinking the coefficients for those input variables that do not contribute much to the prediction task. The R2 value of the Lasso regression is 0.425282 and Mean Squared Error of 0.000020 and an execution time of 1.02 seconds with a relatively lower performance.

6.1.4 Random Forest Regression:

Random Forest Regression is used for predicting numerical values and the collective intelligence of multiple decision trees to improve accuracy and reduce overfitting. This ensemble method combines the predictions of several decision trees, each trained on different subsets of the data, to produce a more stable and accurate prediction. The R2 value is 0.920811 and Mean Squared Error of 0.000003 but with an execution time of 44.35 seconds which is the highest performing model.

6.1.5 Decision Tree Regression:

The Decision Tree Regression allows a tree-like model with the node representing a decision based on the feature, the branch represents the outcome while the node represents the predicted value. It is highly useful in predicting continuous values. The R2 value of Decision Tree Regression is 0.900158 and Mean Squared Error is 0.000004 with the execution time of 6.40 seconds.

6.1.6 Gradient Boosting Regression:

Gradient Boosting Regression is a combination of several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error. The R2 value is 0.493325 and the Mean squared error is 0.000018 with the execution time of 14.60 seconds.

6.1.7 Artificial Neural Network:

Artificial Neural Network is a deep learning algorithm that allows the data to be passed through the input layers to pass through the output layers. The r2 value is 0.817700 and Mean Squared Error is 0.000006 but has the highest processing time and the highest execution time of 253.92 seconds. The model has a good performance but has a higher execution time than the other models.

6.2 Experiment / Case Study 2: Model performance with feature elimination

	Model	R2	MSE	MAE	RMSE	Execution Time
1	Linear Regression	0.530794	0.000017	0.002961	0.004063	0.87
2	Ridge Regression	0.530481	0.000017	0.002961	0.004065	0.34
3	Lasso Regression	0.416527	0.000020	0.003438	0.004497	1.43
4	**Random Forest Regression**	**0.920811**	**0.000003**	**0.000555**	**0.001669**	**40.39**
5	**Decision Tree Regression**	**0.900637**	**0.000003**	**0.000580**	**0.001870**	**5.96**
6	Gradient Boosting Regression	0.492490	0.000018	0.003292	0.004226	13.36
7	Artificial Neural Network	0.817700	0.000006	0.001485	0.002533	247.55

6.2.1 Linear Regression:

The linear regression model is used for establishing linear relationships between predictor and the dependent variables, the R2 value is 0.530794 indicating the features like scheduling_class, priority, cluster significantly influence the predictor variable resource_request_cpu. It has a Mean Squared Error of 0.000017 and an execution time of 0.87 seconds with the model performing average.

6.2.2 Ridge Regression:

When the independent variables are highly correlated, it could lead to imprecise estimates of the regression coefficients. The R2 value of Ridge Regression is 0.530481 and Mean Squared Error is 0.000017 and the low execution time of 0.34 seconds which is same as the performance of the Linear regression but with a slightly lower execution time.

6.2.3 Lasso Regression:

The overfitting of the variables is avoided by using the Lasso regression technique, shrinking the coefficients for those input variables that do not contribute much to the prediction task. The R2 value of the Lasso regression is 0.072922 and Mean Squared Error of 0.000003 and an execution time of 0.51 seconds with a relatively lower performance.

6.2.4 Random Forest Regression:

Random Forest Regression is used for predicting numerical values and the collective intelligence of multiple decision trees to improve accuracy and reduce overfitting. This ensemble method combines the predictions of several decision trees, each trained on different subsets of the data, to produce a more stable and accurate prediction. The R2 value is 0.920811 and Mean Squared Error of 0.000003 but with an execution time of 40.39 seconds which is the highest performing model.

6.2.5 Decision Tree Regression:

The Decision Tree Regression allows a tree-like model with the node representing a decision based on the feature, the branch represents the outcome while the node represents the predicted value. It is highly useful in predicting continuous values. The R2 value of Decision Tree Regression is 0.900637 and Mean Squared Error is 0.000003 with the execution time of 5.96 seconds.

6.2.6 Gradient Boosting Regression:

Gradient Boosting Regression is a combination of several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error. The R2 value is 0.492490 and the Mean squared error is 0.000018 with the execution time of 13.36 seconds.

6.2.7 Artificial Neural Network:

Artificial Neural Network is a deep learning algorithm that allows the data to be passed through the input layers to pass through the output layers. The r2 value is 0.817700 and Mean Squared Error is 0.000006 but has the highest processing time and the highest execution time of 247.55 seconds. The model has a good performance but has a higher execution time than the other models.

6.3 Discussion

The research aims to provide a detailed view of methodology, implementation and evaluation strategies for predicting the resource request in the cloud computing system. The experiment has been divided into two where the two case studies are evaluated. The first case study involves the evaluation of model performance with all the features and the second case study involves the evaluation of Model performance with all the features removed. The case studies are applied with different machine learning algorithms like linear regression, ridge regression, lasso regression, random forest regression, decision tree regression, gradient boosting regression, artificial neural network.

In the first case study, evaluation of model performance with all the features has the random forest regression which is the highest performing model with the r squared value of 0.921258, MSE with 0.000003, MAE of 0.000553 and RSME of 0.001665 and finally with the execution time of 44.35 seconds. The second highest performing model is decision tree regression with r squared value of 0.900158, MSE of 0.000004, MAE with 0.000581, RSME of 0.001874 and execution time of 6.40 seconds as seen in Figure 5.

In the second case study, evaluation of model performance with the features like scheduling class removed. The random forest regression which is the highest performing model with the r squared value of 0.920811, MSE with 0.000003, MAE of 0.000555 and RSME of 0.001669 and finally with the execution time of 40.39 seconds. The second highest performing model is decision tree regression with r squared value of 0.900637, MSE of 0.000003, MAE with 0.000580, RSME of 0.001870 and execution time of 5.96 seconds as seen in Figure 6.

From the above studies we can find that the Random Forest regression is the best performing algorithm with the highest r squared value and the lowest of errors and a low execution time followed by the Decision tree regression with slightly less value of r squared, and a higher error rate but low execution time. Another algorithm that has better r squared is Artificial neural network, but the execution time is high making it not suitable for the evaluation.

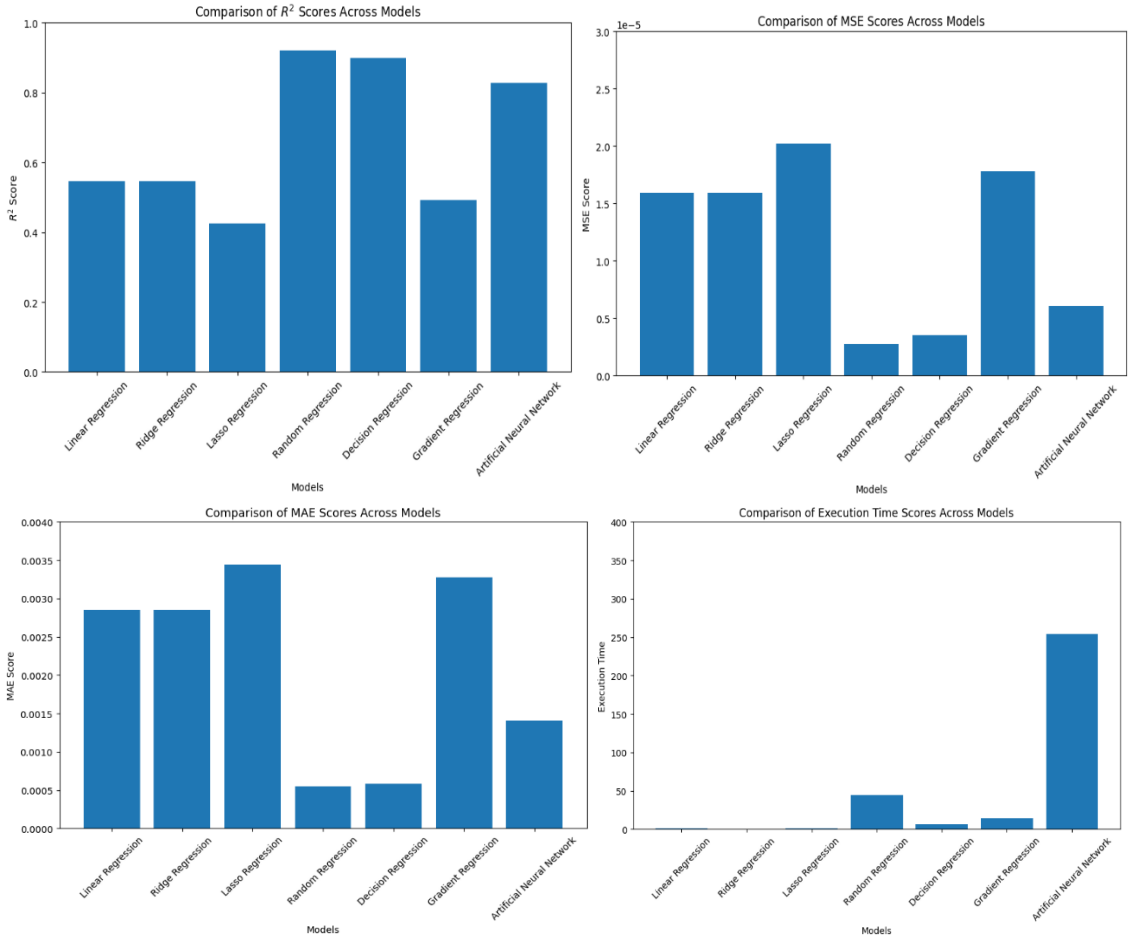

Figure 5: Evaluation metrics of Model Performance

7 Conclusion and Future Work

The prediction of the resource demand is important in the cloud computing environment as it reduces the overhead costs and allows us to allocate the necessary resources on time. From this research it can be found that the machine learning algorithms are very helpful in predicting the resource like CPU. The data cleaning, preprocessing are essential processes that can enhance the performance of the model, whereas the normalization prevents the data from being skewed.

The various machine learning algorithms like linear regression, ridge regression, lasso regression, random forest regression, decision tree regression, gradient boosting regression, artificial neural network was used for evaluating the model performance. The random forest regression is the best performing model in the research. However, the research could be extended to the prediction of the resource memory request, and the hyper performance tuning

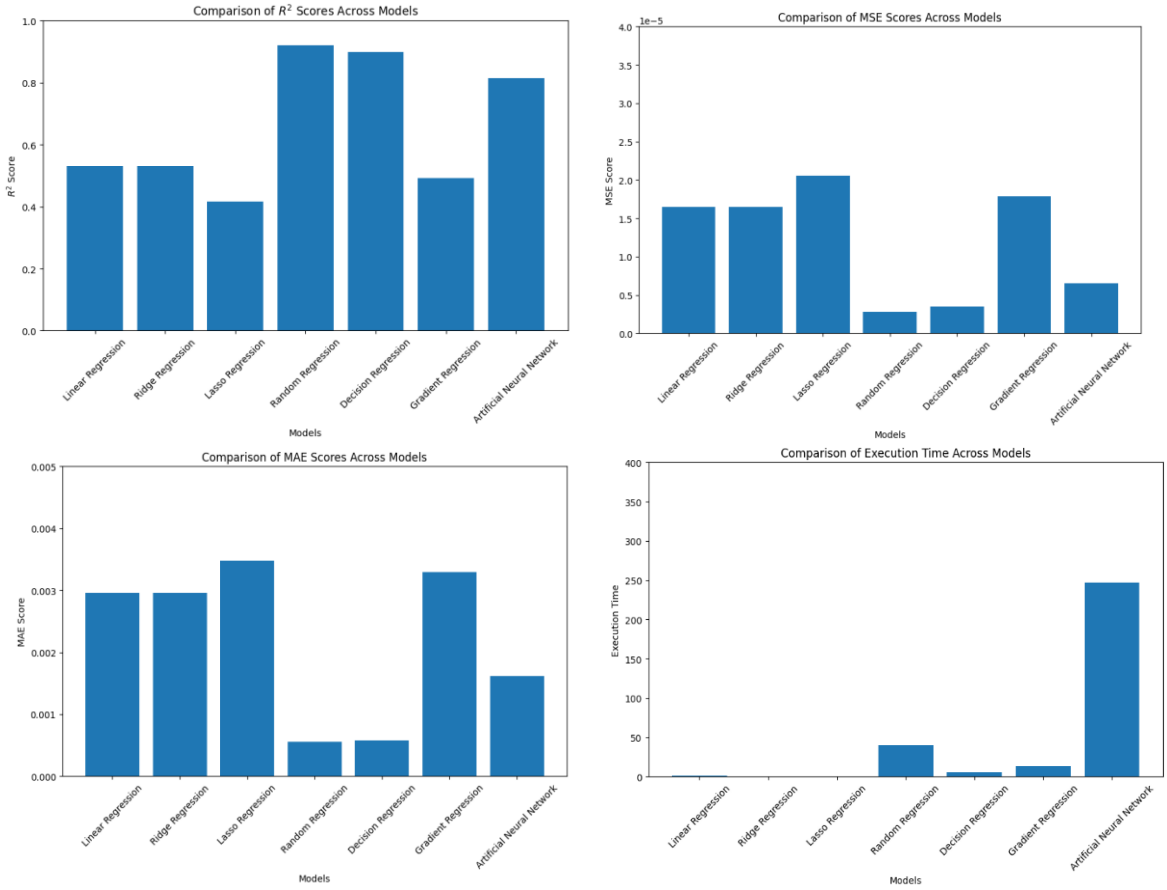

Figure 6: Evaluation metrics of Model Performance with feature scheduling_class removed

of the variables for a better performance. The hybrid machine learning models like the combination of two machine learning models can be studied to produce even better performing models.

Acknowledgements

I would like to express my deepest gratitude to Prof. Mohammed Hasanuzzaman for his constant support and guidance with the submission of this research.

References

Baig, S.-U.-R., Iqbal, W., Berral, J.L., Erradi, A. and Carrera, D. (2019). Adaptive Prediction Models for Data Center Resources Utilization Estimation. *IEEE Transactions on Network and Service Management*, [online] 16(4), pp.1681–1693. doi:https://doi.org/10.1109/TNSM.2019.2932840.

Chaudhary, M. and Banga, P. (2024). Survey of Cloud Computing with Role of Machine Learning. doi:https://doi.org/10.1109/iccica60014.2024.10585116.

Hong, D., Kim, D., Min, O.J. and Shin, Y. (2023). Resource Allocation Reinforcement Learning for Quality of Service Maintenance in Cloud-Based Services. doi:https://doi.org/10.1109/icoin56518.2023.10048905.

Ma, D., Li, Y., Xu, H., Chen, M., Liang, Q. and Li, H. (2023). Optimizing Data Analysis Tasks Scheduling Based on Resource Utilization Prediction. *Lecture notes in networks and systems*, pp.812–823. doi:https://doi.org/10.1007/978-3-031-35314-7_67.

Mashhadi Moghaddam, S., O'Sullivan, M., Walker, C., Fotuhi Piraghaj, S. and Unsworth, C.P. (2020). Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers. *Future Generation Computer Systems*, 106, pp.221–233. doi:https://doi.org/10.1016/j.future.2020.01.008.

Mehmood, T., Latif, S. and Malik, S. (2018). Prediction Of Cloud Computing Resource Utilization. *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*. doi:https://doi.org/10.1109/honet.2018.8551339.

Purohit, N., Srivastava, P. and Pandey, P. (2023). An RNN-Based Resource Allocation Strategy in Cloud Computing. 18, pp.1–6. doi:https://doi.org/10.1109/icicat57735.2023.10263712.

Simaiya, S., Umesh Kumar Lilhore, Yogesh Kumar Sharma, Rao, B., Rao, M., Anupam Baliyan, Anchit Bijalwan and Roobaea Alroobaea (2024). A hybrid cloud load balancing and host utilization prediction method using deep learning and optimization techniques. *Scientific Reports*, 14(1). doi:https://doi.org/10.1038/s41598-024-51466-0.