IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Resource Planning Analytics of HPCC Systems

[1]Dr.R.DilliBabu, [2]Gorty Prasanna Srinivasan, [3]Dr.L.Sudha

[1]Professor, [2]Student, [3]Assistant Professor
[1]Department of Industrial Engineering,College of Engineering Guindy,Anna University,
[2]Department of Industrial Engineering,College of Engineering Guindy,Anna University,
[3]VISTAS,Chennai,

Abstract : For ONGC, estimating how much an HPCC system will be used is quite important. The purpose of this project is to determine the variables affecting HPCC System Utilization and to locate an appropriate data analytics approach to forecast Utilization. The project was created in order to meet the HPCC System's requirement for Resource Planning analytics. With the aid of the computer, various information was noted, including user utilization, idle utilization, and so forth. And utilizing this, a scatter plot between the two parameters was created. We make an effort to use machine learning algorithms to investigate the link between the two variables.

***IndexTerms* - Linear Regression, Correlation, Data cleaning**

INTRODUCTION

A strategic method for ensuring that resources are used efficiently across a single project or a portfolio of tasks is resource planning. When implemented correctly, organizations employ resources as efficiently and optimally as possible without under- or overusing any one resource. Additionally, they gain visibility into ongoing projects, upcoming resource needs and shortages, prospective project bottlenecks, and capacity allocations.Organizations can respond to changing markets and initiatives with greater flexibility by using resource planning. The capacity of firms to change direction quickly becomes crucial when new disruptive technologies hit the market at ever-increasing rates. Business objectives that were significant yesterday might not be as valuable tomorrow.Companies must use every effort to develop the adaptability required to change course when objectives and strategies evolve.

Analyzing data collections to identify trends and make judgments about the information they contain is known as data analytics (DA). Data analytics is increasingly carried out with the use of specialist hardware and software. In order to help businesses make better business decisions, data analytics technologies and methodologies are widely used in the commercial sector. In order to support or refute scientific models, theories, and hypotheses, scientists also use analytics techniques.

METHODOLOGY

The methodology, which is followed for the Predictive analysis comprises:
(i) Collection, sorting and classification of sample data
(ii) Detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
(iii) Understanding of the sample data by computing descriptive statistics.
(iv) Training and testing the dataset using Machine Learning Algorithms.
(v) Creating a mathematical model for the given dataset and using the model to predict the Utilization.

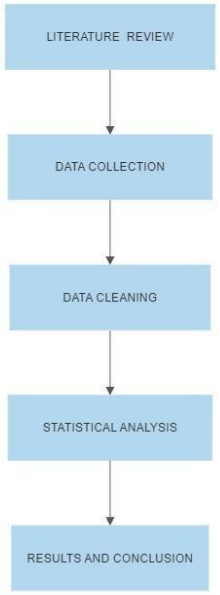


FIGURE 1: PICTORIAL REPRESENTATION OF METHODOLOGY

The above is the flowchart of the methodology that's being taken place here.

METHODS USED

Now it is the time to articulate the research work with ideas gathered in above steps by adopting any of below suitable approaches:

A.Data Collection

The ONGC statistics were utilised in my project. The graphic shows several characteristics such as user Utilization and idle Utilization.In order to maximize the value of user usage, our goal is to identify the variables that depend on user utilization.The information is available in its given raw format.

B.Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. It helps the user to easily understand the data. Steps Involved in Data Preprocessing: Data cleaning, data integration, data reduction.

C. Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate,or incomplete data within a dataset. First step is Removing unwanted/irrelevant observations. We can Handle the missing data with the help of Data cleaning Process.

D. Correlation matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

E.Scatterplot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates valuesfor an individual data point. Scatter plots are used to observe relationships
between variables. Here we draw a scatterplot between User Utilization and Idle Utilization to find out the relationship between the two attributes.

F. Predictive analysis

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events.

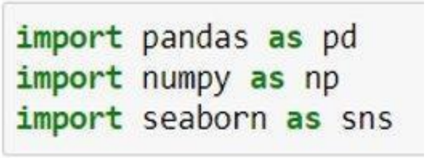
G. Linear regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. I used a Programming language called Python to perform linear regression to perform predictive analysis on Utilization.

PYTHON LIBRARIES USED

Seaborn is python library that are used for data visualization.They have inbuilt modules for plotting different graphs
Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool.

NumPy is a general-purpose array-processing package. It provides a high- performance multidimensional array object, and tools for working with these arrays.
Scikit-learn is a free software machine learning library for the Python programming language.



```
import pandas as pd
import numpy as np
import seaborn as sns
```

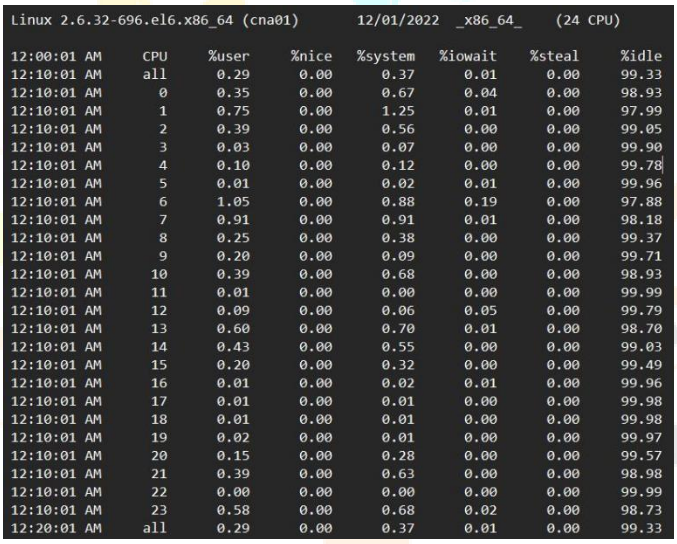
FIGURE 2: SCREENSHOT OF THE LIBRARIES USED

DATA PREPROCESSING

Any type of processing done on raw data to get it ready for another data processing operation is referred to as data preprocessing, which is a part of data preparation. It has historically been a crucial first stage in the data mining process.

5.1 Data collection and cleaning

For this project, we have used the data of ONGC. Different variables like user Utilization, idle Utilization can be seen in the image.Our aim is to find the variables that depend on the user utilization so as to maximise the value of user utilization.
The data is present in raw format as given.



```
Linux 2.6.32-696.el6.x86_64 (cna01)       12/01/2022  _x86_64_    (24 CPU)

12:00:01 AM     CPU     %user     %nice   %system   %iowait    %steal     %idle
12:10:01 AM     all      0.29      0.00      0.37      0.01      0.00     99.33
12:10:01 AM       0      0.35      0.00      0.67      0.04      0.00     98.93
12:10:01 AM       1      0.75      0.00      1.25      0.01      0.00     97.99
12:10:01 AM       2      0.39      0.00      0.56      0.00      0.00     99.05
12:10:01 AM       3      0.03      0.00      0.07      0.00      0.00     99.90
12:10:01 AM       4      0.10      0.00      0.12      0.00      0.00     99.78
12:10:01 AM       5      0.01      0.00      0.02      0.01      0.00     99.96
12:10:01 AM       6      1.05      0.00      0.88      0.19      0.00     97.88
12:10:01 AM       7      0.91      0.00      0.91      0.01      0.00     98.18
12:10:01 AM       8      0.25      0.00      0.38      0.00      0.00     99.37
12:10:01 AM       9      0.20      0.00      0.09      0.00      0.00     99.71
12:10:01 AM      10      0.39      0.00      0.68      0.00      0.00     98.93
12:10:01 AM      11      0.01      0.00      0.00      0.00      0.00     99.99
12:10:01 AM      12      0.09      0.00      0.06      0.05      0.00     99.79
12:10:01 AM      13      0.60      0.00      0.70      0.01      0.00     98.70
12:10:01 AM      14      0.43      0.00      0.55      0.00      0.00     99.03
12:10:01 AM      15      0.20      0.00      0.32      0.00      0.00     99.49
12:10:01 AM      16      0.01      0.00      0.02      0.01      0.00     99.96
12:10:01 AM      17      0.01      0.00      0.01      0.00      0.00     99.98
12:10:01 AM      18      0.01      0.00      0.01      0.00      0.00     99.98
12:10:01 AM      19      0.02      0.00      0.01      0.00      0.00     99.97
12:10:01 AM      20      0.15      0.00      0.28      0.00      0.00     99.57
12:10:01 AM      21      0.39      0.00      0.63      0.00      0.00     98.98
12:10:01 AM      22      0.00      0.00      0.00      0.00      0.00     99.99
12:10:01 AM      23      0.58      0.00      0.68      0.02      0.00     98.73
12:20:01 AM     all      0.29      0.00      0.37      0.01      0.00     99.33
```

FIGURE 3: RAW DATA SET

5.2 Definition about the data

We can see there are **24 CPUs** from the raw data format.
Steal time is the percentage of time the virtual machine process is waiting on physical the CPU for its CPU time.
I/O wait (iowait) is the percentage of time that the CPU (or CPUs) were idle during which the system had pending disk I/O requests.
User Utilization is the percentage of time that the User is Utilizing the system.
Idle Utilization is the percentage of time the system is in Idle state.
Nice Utilization/Agent utilization is a workforce management metric that indicates agent productivity.

5.3 Data Preprocessing

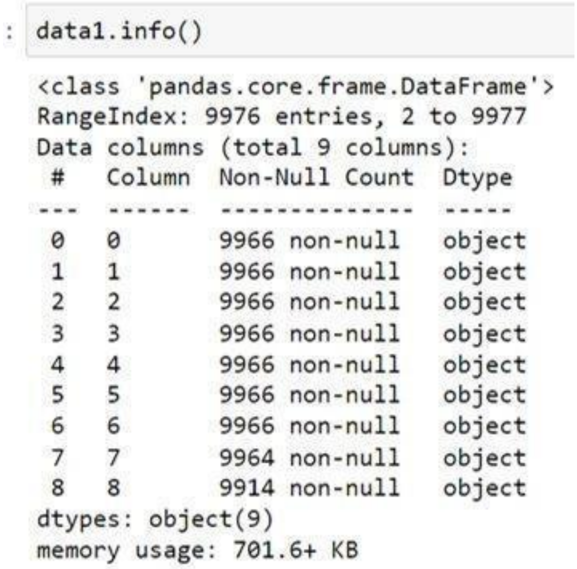
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. It helps the user to easily understand the data. Steps Involved in Data Preprocessing: Data cleaning, data integration, data reduction, and data transformation. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. First step is Removing unwanted/irrelevant observations .Before Data cleaning the columns had empty values in some of the columns so we fill those values with the help of Mean of the data.

DATA CLEANING

The practise of correcting or deleting inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are numerous ways for data to be duplicated or incorrectly labelled when merging multiple data sources.

6.1 Before Data Cleaning

We can see that the range of the Index is 9976 entries but just 9964 entries are therein the table. Data.info is the command which can be used to get the information about the data . Hence Data cleaning is performed and we fill the empty columns
with the help of the command as shown in the figure below.After filling the data we can seethere are 9976 entries as shown in the figure.

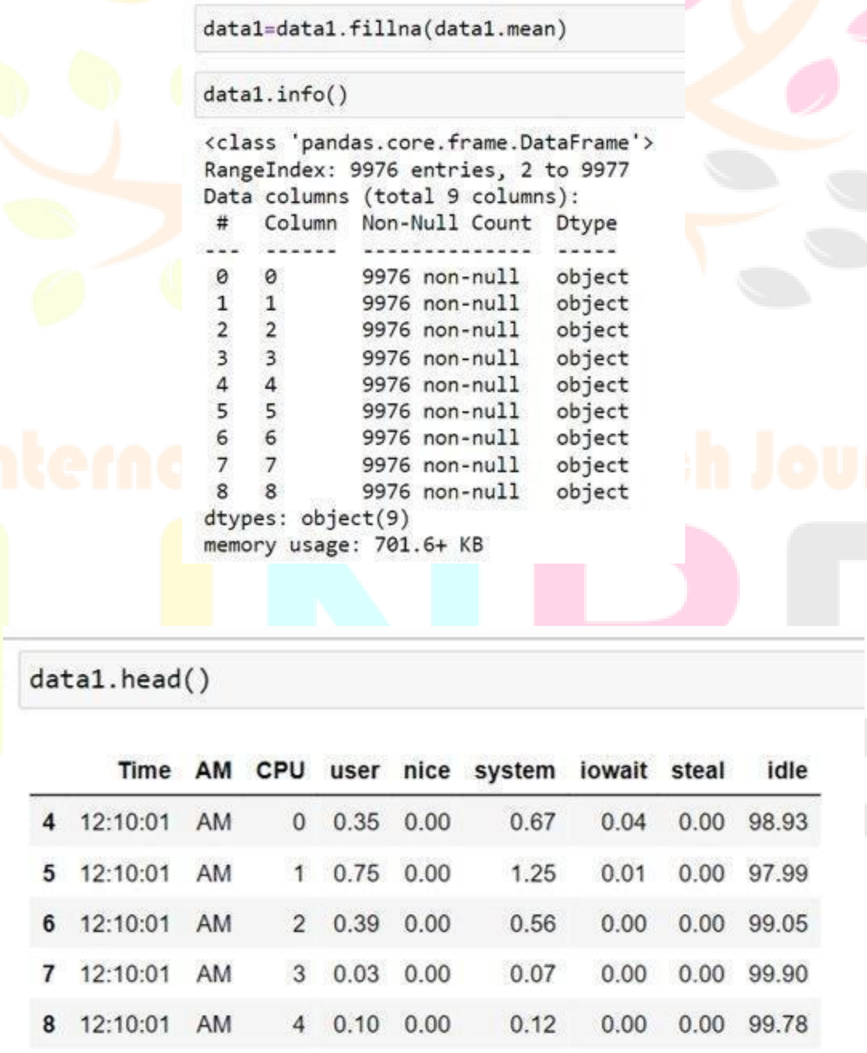


```
:  data1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9976 entries, 2 to 9977
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   0       9966 non-null   object
 1   1       9966 non-null   object
 2   2       9966 non-null   object
 3   3       9966 non-null   object
 4   4       9966 non-null   object
 5   5       9966 non-null   object
 6   6       9966 non-null   object
 7   7       9964 non-null   object
 8   8       9914 non-null   object
dtypes: object(9)
memory usage: 701.6+ KB
```

FIGURE 4: INFORMATION ABOUT THE DATA BEFORE DATA CLEANING

6.2 After Data Cleaning

After filling the data we can see there are 9976 entries as shown in the figure. After the data cleaning process we get the updated dataset which looks like as shown below. We know that the raw data set is complex and hard to understand , hence we convert it to a simple and understandable dataset as shown below.



```
data1=data1.fillna(data1.mean)

data1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9976 entries, 2 to 9977
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   0       9976 non-null   object
 1   1       9976 non-null   object
 2   2       9976 non-null   object
 3   3       9976 non-null   object
 4   4       9976 non-null   object
 5   5       9976 non-null   object
 6   6       9976 non-null   object
 7   7       9976 non-null   object
 8   8       9976 non-null   object
dtypes: object(9)
memory usage: 701.6+ KB

data1.head()
```

	Time	AM	CPU	user	nice	system	iowait	steal	idle
4	12:10:01	AM	0	0.35	0.00	0.67	0.04	0.00	98.93
5	12:10:01	AM	1	0.75	0.00	1.25	0.01	0.00	97.99
6	12:10:01	AM	2	0.39	0.00	0.56	0.00	0.00	99.05
7	12:10:01	AM	3	0.03	0.00	0.07	0.00	0.00	99.90
8	12:10:01	AM	4	0.10	0.00	0.12	0.00	0.00	99.78

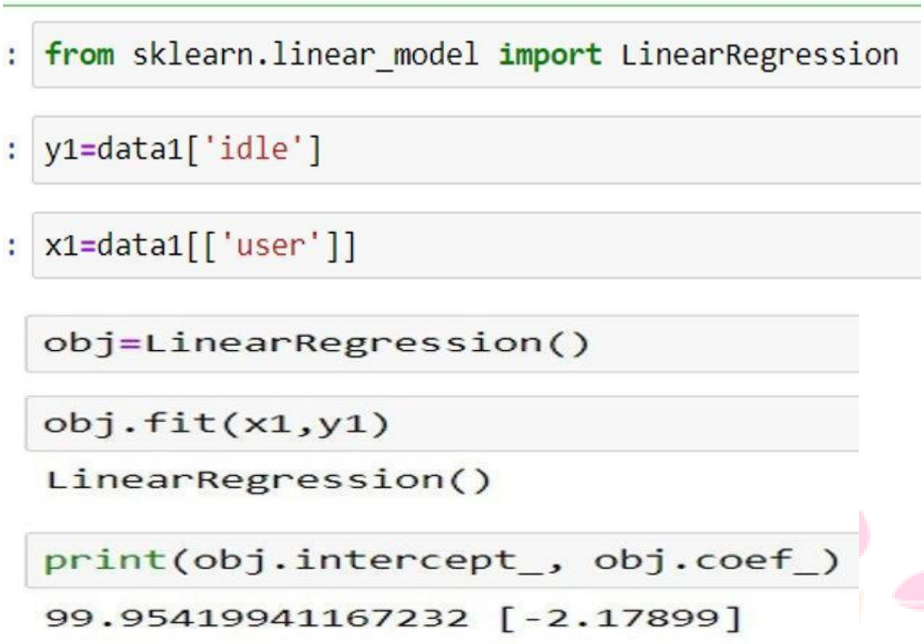
FIGURE 5: INFORMATION ABOUT THE DATA AFTER DATA CLEANING AND UPDATED DATA SET

RESULTS

The Results of this paper are as given below.

7.1 Predictive Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.



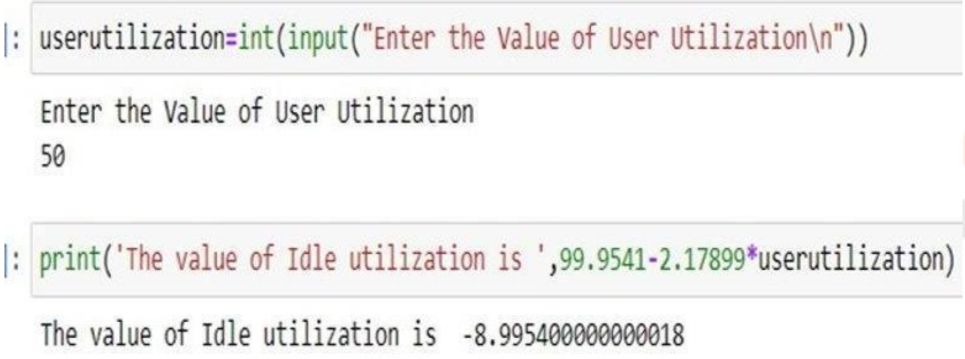
```
: from sklearn.linear_model import LinearRegression

: y1=data1['idle']

: x1=data1[['user']]

obj=LinearRegression()

obj.fit(x1,y1)
LinearRegression()

print(obj.intercept_, obj.coef_)
99.95419941167232 [-2.17899]
```

FIGURE 6: PICTORIAL REPRESENTATION OF REGRESSION OBJECT

Hence the equation between User Utilization and Idle Utilization is **Y= 99.9541-2.17899x**

7.2 Predicting Utilization

We use Python to predict the value of Idle Utilization by building a User Interface to get the values of User Utilization. We try to Input the values ofthe User Utilization with the help of which we try to find the output of IdleUtilization. The image below shows the predicted value.



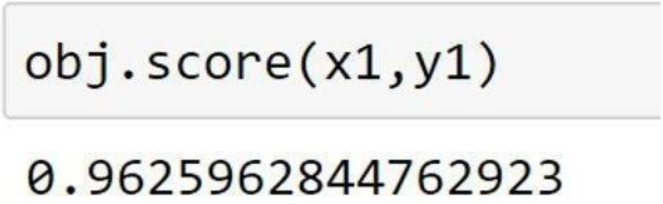
```
|: userutilization=int(input("Enter the Value of User Utilization\n"))

Enter the Value of User Utilization
50

|: print('The value of Idle utilization is ',99.9541-2.17899*userutilization)

The value of Idle utilization is  -8.995400000000018
```

FIGURE 7: PICTORIAL REPRESENTATION OF PREDICTING UTILIZATION

In this example, we predicted the Idle Utilization by inputting the value of User Utilization from the user. And then after that we calculate the Value of Coefficient of Correlation at last.

7.3 Predicting Coefficient Of Correlation

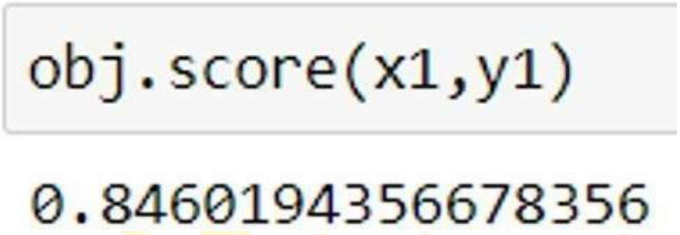
The correlation coefficient is a statistical measure of the strength of a linearrelationship between two variables. Its values can range from -1 to 1.



```
obj.score(x1,y1)

0.9625962844762923
```

FIGURE 8: PICTORIAL REPRESENTATION OF CORRELATION COEFFICIENT

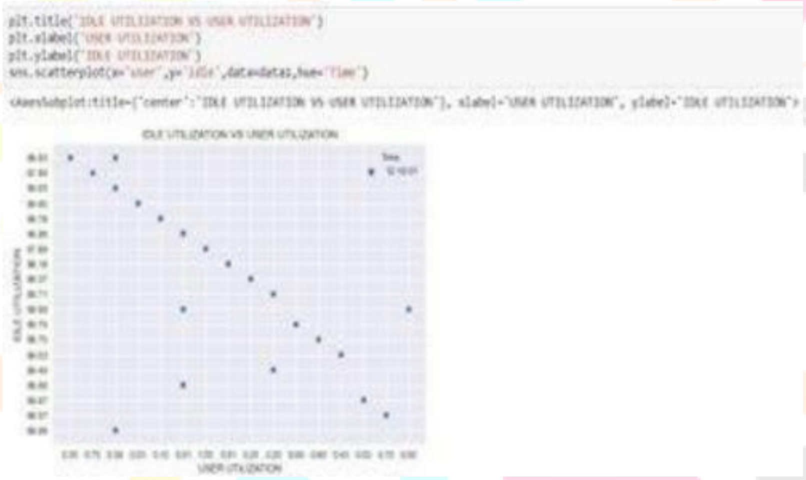
From the figure we can see that there is highly Negative Linear Relationship between User Utilization and Idle Utilization. Hence If we increase the Value of Idle Utilization the User Utilization will get decreased and Vice Versa.



```
obj.score(x1,y1)

0.8460194356678356
```

FIGURE 9: PICTORIAL REPRESENTATION OF CORRELATION COEFFICIENT

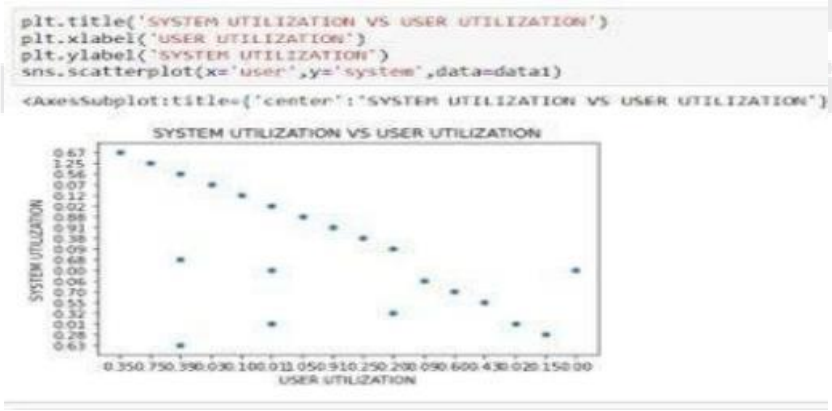
From the figure we can see that there is highly Negative Linear Relationship between User Utilization and System Utilization. Hence If we increase the Value of System Utilization the User Utilization will get decreased and Vice Versa.



```
plt.title('IDLE UTILIZATION VS USER UTILIZATION')
plt.xlabel('USER UTILIZATION')
plt.ylabel('IDLE UTILIZATION')
sns.scatterplot(x='user',y='idle',data=data1,hue='Time')

<AxesSubplot:title={'center':'IDLE UTILIZATION VS USER UTILIZATION'}, xlabel='USER UTILIZATION', ylabel='IDLE UTILIZATION'>
```

FIGURE 10: PICTORIAL REPRESENTATION OF SCATTERPLOT BETWEEN IDLE AND USER UTILIZATION

The above Scatterplot plot shows us that both the variables are Negatively Correlated with each other.



```
plt.title('SYSTEM UTILIZATION VS USER UTILIZATION')
plt.xlabel('USER UTILIZATION')
plt.ylabel('SYSTEM UTILIZATION')
sns.scatterplot(x='user',y='system',data=data1)

<AxesSubplot:title={'center':'SYSTEM UTILIZATION VS USER UTILIZATION'}
```

FIGURE 11: PICTORIAL REPRESENTATION OF SCATTERPLOT BETWEEN SYSTEM AND USER UTILIZATION

The above Scatterplot plot shows us that both the variables are Negatively Correlated with each other.

CONCLUSION

We discover that user utilization influences system and idle utilization more so than the other parameters that are available. Therefore, both of these are important elements that affect user utilization.in contrast to the other parameters, we discover that both system and idle utilization have a substantially negative linear relationship with user utilization. But compared to other elements or variables, idle utilization has a more detrimental effect on user utilization; as a result, it is a crucial or significant factor. We therefore draw the conclusion that we can improve the value of user utilization if we lower the values of idle utilization and system utilization.

With a correlation coefficient of -0.96, there is a linear relationship between user utilization and idle utilization. By doing so, we may use User Utilization to forecast the value of the measure Idle Utilization. Also created was a scatter plot between these two variables. With a coefficient of -0.84, there is a linear relationship between user utilization and the measure of system utilization. With the help of this, we were able to forecast the value of System Utilization using UserUtilization. Also created was a scatter plot between these two variables. Using System and IdleUtilization as the independent variables and User Utilization as the dependent variable, a multiple regression line was also created.

erm.

REFERENCES

[1] Barnaghi, P, Amit, S, and Cory, H. "From Data to Actionable Knowledge: Big Data Challenges in the Web of Things [Guest Editors' Introduction]." IEEE Intelligent

[2] Systems 28.6 (2013): 6-11. Chaudhary, R., Pandey, J. R., & Pandey, P. (2015, October) Business model innovation through big data. In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on (pp. 259-263). IEEE.

[3] Chen, H, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188. Davenport, H., and Jill, D. "Big data in big companies." International Institute for Analytics (2013).

[4] Jiang, W. and Chai, H., 2016, July. Research on big data in business model innovation based on GA-BP model. In Service Operations and Logistics, and Informatics (SOLI), 2016 IEEE International Conference on (pp. 174-177). IEEE

[5] Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., et al. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062.

[6] Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., et al. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062.

[7] Russom, P. "Managing big data." TDWI Best Practices Report, TDWI Research (2013)11. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint =

[8] The INE webpage provides historical data on employment. https://www.ine.es. Accessed 20 Nov 202 Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), 457–466 (2018)