DATA ANALYTICS AND VISUALIZATION: Alibaba cluster trace GPU 2020.

Abstract

The research is to find the appropriate dataset for preparing an environment-friendly and sustainable system with the operation management of an organization. Thus this paper develops the concept of the application of machine learning and other technologies to maintain SDG goals with increasing the sustainability of the organization. Data analytics is important in managing the data set and directly helps in the optimization of the performance of the business. Implementing a proper business model helps the different companies help to reduce the costs of the different companies and also helps in identifying the different ways in doing effective business and creating a maximum store of data in the case of larger amounts.

Table of Contents

1. Introduction

1.1 Background

Increasing sustainability is the new normal in different organizations and it is applicable to data scientists for increasing ***"sustainability policies"*** within the organization. Moreover, developing environmental data as per the operational criteria of an organization is essential for increasing its sustainability while fulfilling the ***"sustainable development goals"*** of the UN. Hence, the research has considered the dataset of Alibaba for discussing smart systems through rigorous investigation to build the visualization through data analysis. Looked at the hybrid cluster's runtime state and provided some key information concerning cloud imbalance. The difficulty and complexity of managing cloud resources are exacerbated by such imbalance. Several contemporary type of strategies are built on the premise of a perfect cluster environment.(Lu et al 2017).

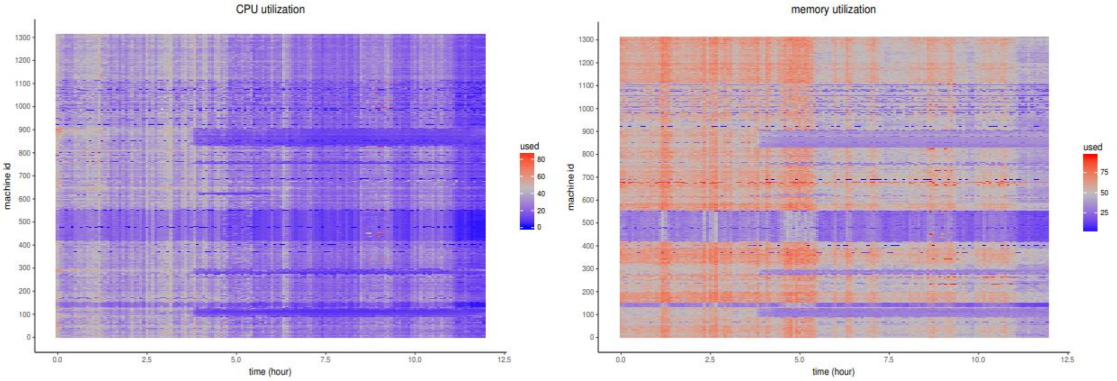


Figure 1. heat maps show the cluster's computer systems' CPU and memory use. The white area denotes the trace's absence of data. Blue hue denotes low consumption, whereas red denotes excessive utilization. (Source Lu et al 2017).

1.2 Research aim and objectives

The research aims to develop a ***"Sustainable and Smart System"*** for generating sustainability-related policies with easier decision-making policies with the application of ***"data science and visualization"***.

Research objectives are

Level 1 : To perform a descriptive analysis of the dataset in order to find the mean, median, and mode.

Level 2: To execute a correlation in inferential statistics based on the “Alibaba Cluster Trace GPU 2020” Dataset.

Level 3: To develop linear relationship between a single dependent continuous variable and more than one independent variable, and also develop multilinear regression for predicting data.

1.3 Data analytics with Random Forest Classifier for regression and classification.

Level 4: To predict data accuracy model using CNN (convolutional neural network).

- To explore the environmental data for creating a sustainable operating system
- To develop a dataset for maintaining the ***“sustainable development goals”*** of the UN in Alibaba
- To examine the importance of data science for supporting datasets for maintaining the ***“sustainable development goals”*** of the UN in Alibaba

1.3 Research questions

- What is the importance of environmental data for creating a sustainable operating system?
- What is the process of preparing the dataset for maintaining the ***“sustainable development goals”*** of the UN in Alibaba?
- What is the importance of data science for supporting datasets for maintaining the ***“sustainable.***
- ***development goals”*** of the UN in Alibaba?

1.4 Rational of research

Researchers are motivated to research this aspect due to the increasing importance of sustainable development around the globe and its required dataset to prepare the strategies of development. The sustainable development process can be enhanced based on the dataset and the appropriate operating system can be used to maintain the function of the data analysis process. Concerning method has been used to gather data for meeting the goal based on the data analysis process

(Bohidar *et al.* 2022). The importance of the sustainable development process can be described for finding the strategies for developing the forecasted outcomes by using the proper dataset.

1.5 Contribution

The paper develops the concept of data science and its application in real organizations for increasing competitive advantage with the implementation of SDG obligations within the organization (Joshi *et al.* 2021).

1.6 Organization of the report

The report develops through different sectional analyses starting from the introduction with the research aim and objectives up to drawing the concluding statement along with the future implementation.

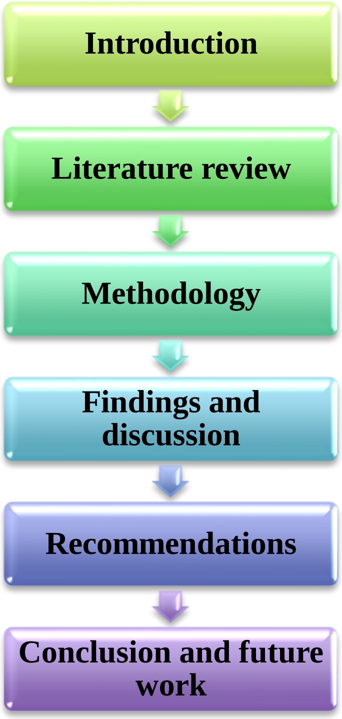


Figure 1: Report structure

(Source: Self-created)

2. Literature review

2.1 Important of data analytics of workload

Data analytics is important in managing the data set and directly helps in the optimization of the performance of the business. Implementing a proper business model helps the different companies help to reduce costs of the different companies and also helps in identifying the different ways in doing effective business and creating a maximum store of data in the case of larger amounts (Palkar *et al.* 2018). Using the data analysis of collecting different data set is effective in doing a better business decision-making process and help to maintain customer satisfaction and provide proper customer trends in their better product and services. Doing the data analytics process in the business first must have to collect the data set and separate it by doing demographic, gender, and another process. The collected data must be organized and analyzed by using a spreadsheet and using software to take the statistical data. In this data analytics process, it must have to evaluate different factors such as regression analysis, T-test, correlation, mean and median mode. Data analytics is important to understand the trends of the pattern of massive amounts of data being collected (Jindal *et al.* 2019). Analysis of this dataset is effective to understand the audience, forecasting future results, and reducing the cost of the business process. In order to work should measure the business process that helps to measure the ROI and monitor the performance metrics that directly impact the business, maximum time of delivery, and execution of the cost in the allocated budget. In this data analytics process of the workload, it must have to adopt the cloud decoupling that has a legacy of the IT environments and it has required limited pool resources.

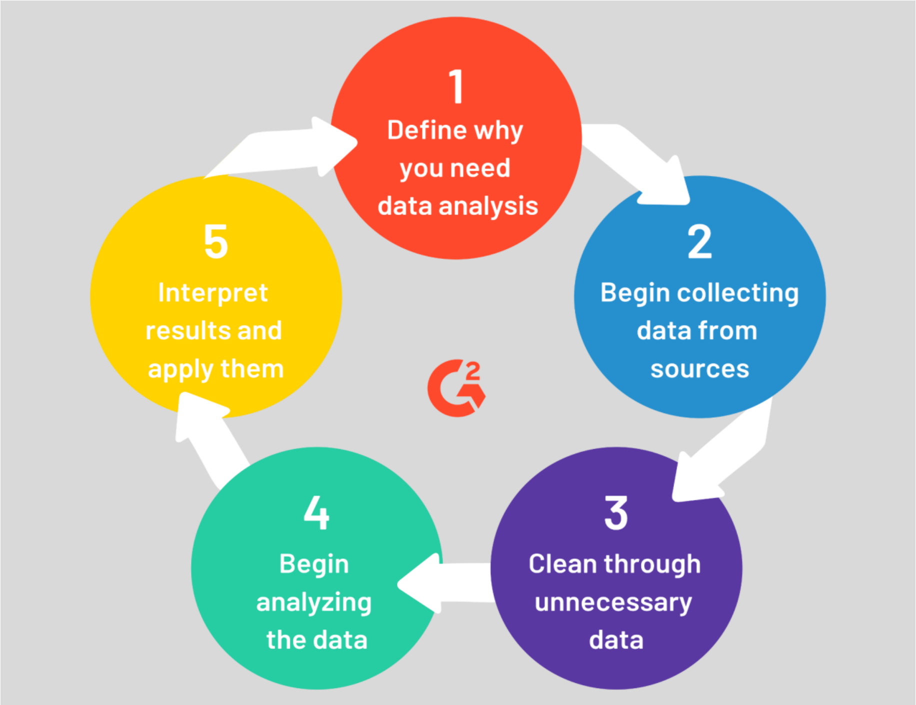

1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

Figure 2: Data analytics process

(Source: Jindal *et al.* 2019)

2.2 Scheduling in Large-Scale Heterogeneous GPU Clusters

Sustain technological advantages of the machine learning technology is effective to do the massive availability of the dataset and properly compare with the tech companies in large scale of the ML services in clouds. Thus the adoption of these technological advantages is effective to provide the heterogeneous GPUs and host ML applications (Alam *et al*. 2022). In order to run the ML workloads and creation of the heterogeneous GPU system process it must have to raise a cluster number of challenges. On the other hand, the collection of the two months of the workload trace from the production of the "MLaaS cluster" has about 6000 GPUs of challenges. this has explained the overall challenges posed in the cluster scheduling process that consists of low GPU utilization that has the presence of hard schedule tasks in a high GPU system process that imbalance load across the heterogeneous machine (Makrani *et al*. 2018). Adoption of the GPU clusters schedulers in the long-running and intensive resources. It heavily makes powerful and expensive accelerators in the GPU computation system process. Thus it is crucial to effectively manage the cluster GPU system and ensure the different job packs for allocated resources. Adoption of the machine learning is effective that helping to drive the force of automation and cutting the time of the human workload. On the other hand, the creation of the automation process in the scheduling and training process is effective to reduce the complex algorithms and helps in the creation of reliable and efficient work (Szárnyas *et al.* 2018). Further Ml job predictable has generated the memory locations and computed irritations of predictable. Moreover, ML jobs across the GPU have created proper weights and it creation of the overall network that helps to make the scheduling decisions and provide a proper resource allocation system in dicu8ss of the different dimensions.

2.3 Advantages and disadvantages of using Python in the data analytics process

Using the software of Python in analytics of the dataset is effective to properly progress in this research and evaluate effective data in order to understand the machine load in the progress of the work. There are several advantages and disadvantages to using Python in this data analytics process (Tallent *et al.* 2018). Further another advantage of using this software is that it can easily import different libraries that easily help to do the code and include it in different projects.The main libraries that can be reached through Python are Pandas, Numpy, Pyglet, Scrapy, and so on.

Besides the several advantages of this Python, there are certain disadvantages to using this software.

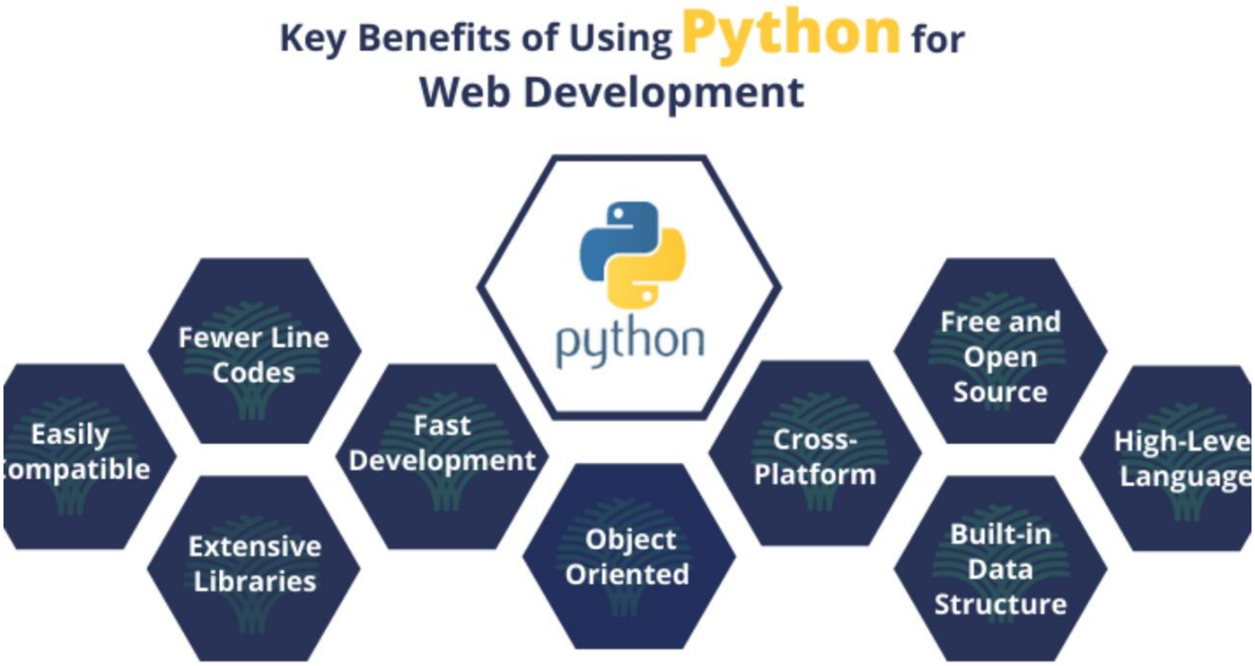


Figure 3: Effectiveness of this using Python

(Source: Al-Zubaidi *et al.* 2020)

2.4 Effectiveness of Artificial intelligence in data analytics

The adoption of Artificial Intelligence in the business process and data analytics process has helped in the high performances of the software and provides real-time performances in this data analytics process. Using advanced AI technology help to make possible improve accuracy and not feasible in the manual process. Using artificial intelligence and machine learning has been effective in predicting personalization and improving experiences (Yeung *et al*. 2021). Using AI-based software helps to automatically analyze the data and deliver valuable insights. On the other hand, it is effective to customize the data and help in the development of the product. Adoption of the AI technology in the data analytics process of business has created automated processes and helped to forecast the best data trend process. Further AI is effective in the diagnosis and analysis process that helps to reduce the crisis in understanding the service provide concerning the past activity. In another hand, it has faced a complex volumetric nature in the process of big data and explores artificial intelligence in some of the troublesome. The main advantage of using this AI tool is that it is effective to manage big data and forward-thinking in solutions of the in improve the efficiency of data analytics (Fathi *et al.* 2021)..

3. Methodology

3.1 Data mining motive

The data mining process has been effective in order to show the number of increased workability and cluster of production in the from the past decade. In the most commonly established process of data mining, the field is effective to discover the knowledge of databases in KDD. In order to monitor and discover the common techniques it has used the DM in CRISP-DM extended of versions of design the KDD process. Moreover, the macro methodology is the process that helps to produce entire the NOE and components of the different parts in means of the single model. Besides this, in the exploration of the modified model, it has mostly developed the SAS for doing similar steps of KDD for designing the purpose of the dataset of Alibaba UN (Roettger *et al.* 2019). In order to do the data mining process, it has mainly developed the pre-KDD value and posted it as similar of the KDD. Besides this, using the python google collaborate of the DM strategies is effective to consider the field of the data analysis process. Among the simple steps of the KDD selection and transformation of the data mining process is effective to customize the no-deployment value and do the post-KDD process involved with the business. In order to target the quality data of the hybrid training of Alibaba the ML algorithms have helped to collect the large production cluster by using the Alibaba PAI. Thus in this data mining process, it leads to monitoring the data from the given data set. The selected dataset for this process of the data analytics of the Alibaba workload it has consisted of seven rows and 7034 columns in the dataset. Thus this amount of big data must have to do the pre experimentations by doing the irrelevant data omitted.

3.2 Data preparation

After collecting the data in order to of the world and production rate of Alibaba it has measured the different analyzation factors after sequentially modified in the data set. First, it must have to initialize all the datasets and then omit the missing value in the dataset. Thus after omitting the mission value from the dataset it has the final seven rows in this dataset process. The creation of this data is effective in properly analyzing the mean median mode by doing the descriptive analysis and the correction of the T-test using ANOVA (Dai, H.N., Wang *et al.* 2020). On the other hand, after doing the T-test process in this analysis of the data by doing the linear regression using the machine learning process. Besides this, it must have to create the neural network considering the quantity of the missing values that can easily be imputed later on.

3.3 Micro methodology

Analyzation of the micro methodology has been done by the design dataset in four different levels of analysis. In order, to reach the objectives there must have to do 3 to 54 different types of algorithms to be properly applied to each level of analysis. In order to do the descriptive statistical analysis it has calculated the different algorithms such as mean, median, standard deviation must have to be done. This algorithm process it has consisted of two samples of the T-test in the application of the significant differences of the workload capacity.

3.4 Use of tools and techniques

Data visualization and the data analytical technique has been implemented for determining the challenges of the analysis process and the python programming language helps to describe the analysis process. In this project, describe(), as well as various types of statistical analysis processes can be applied for collecting data based on the dataset analysis process. An effective software tool has been applied for describing the details of the dataset to meet the problem of the analysis method.

3.5 CRISPDM (Cross Industry Standard Process for Data Mining):

Three "veterans" of the developing DM sector came up with the idea for the Cross Industry Standard Process for Data Mining (CRISP-DM) initiative in 1996. This methodology is used for predictive analytics and data mining across several industries (Rivo et al 2012).

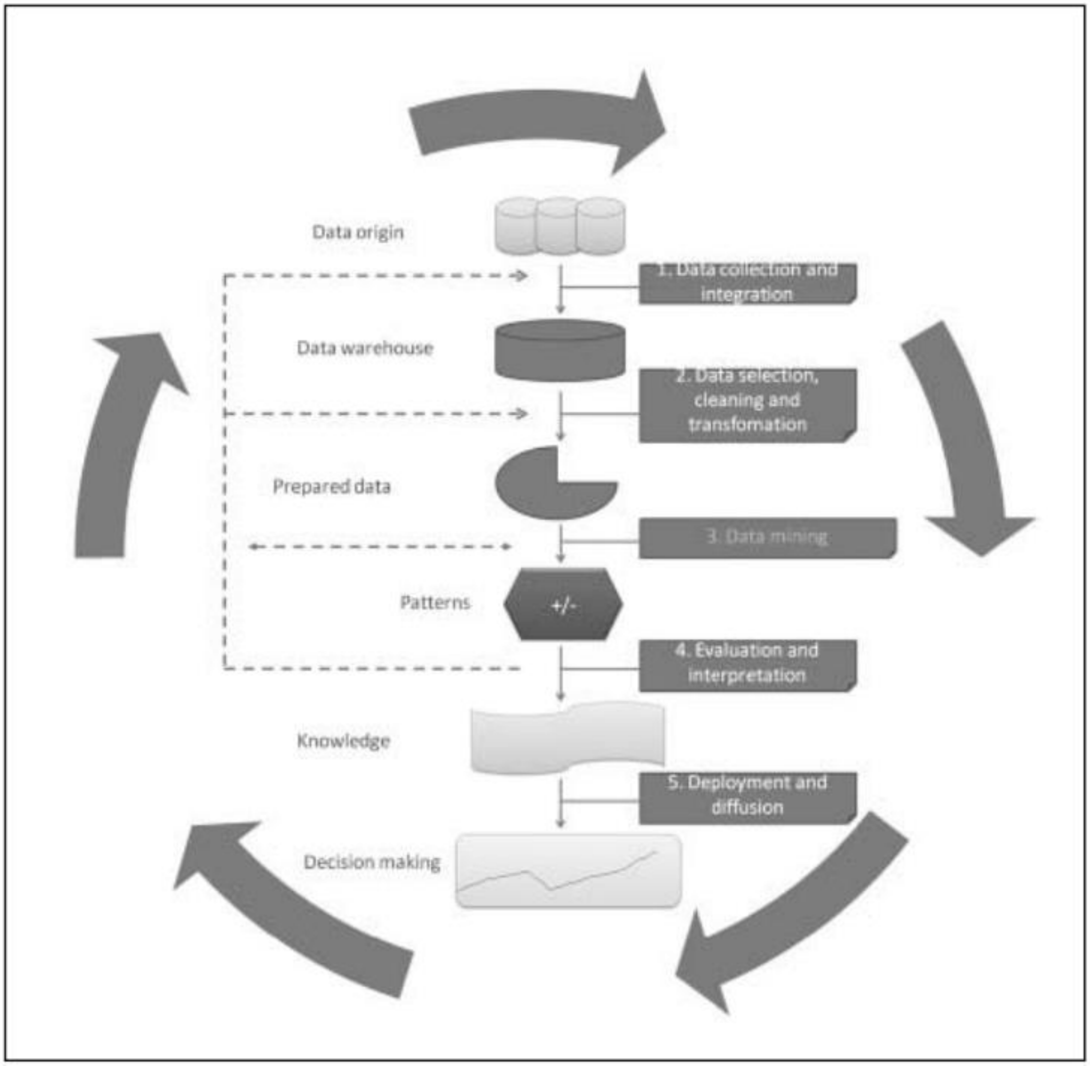

Data origin
1. Data collection and integration
Data warehouse
2. Data selection, cleaning and transfomation
Prepared data
3. Data mining
Patterns
+/-
4. Evaluation and interpretation
Knowledge
5. Deployment and diffusion
Decision making

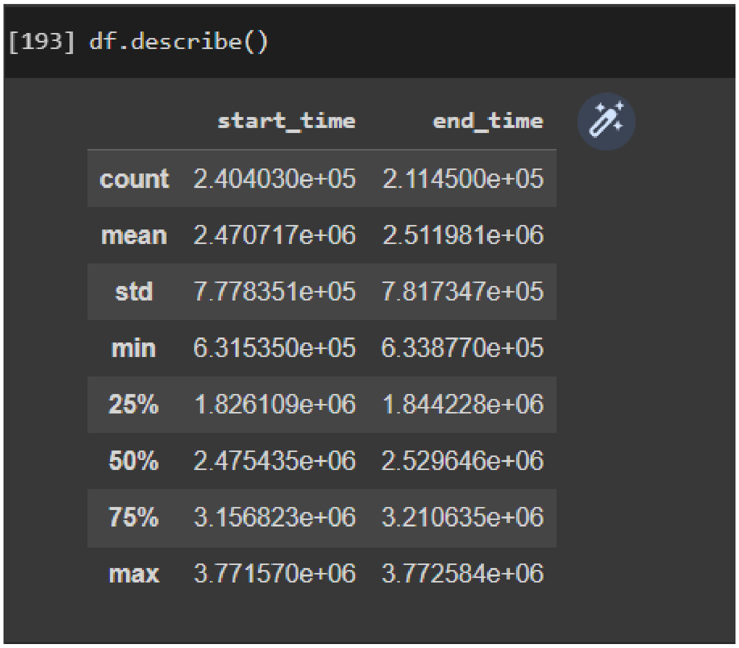
Figure 4. One of the DM cycle's main stages. The outside arrows stand for the cycle of DM.(Source Rivo et al 2012).

4. Findings and discussions

Python is a good program used for data analytics. In this project of “Data analytics and visualization” python programming language is used. The data set used in this project is used for visualization. This data set has many null values and empty data points removing this is very important for this project (Chen *et al.* 2022). The null values can create problems in the visualization. So removing the null value is important. The null value present is removed in the early stage as this remove all null value from the dataset. The data set has a few columns that are not useful for this visualization. This unnecessary column requires extra processing power and increases the processing time. So the column that is not important for the project is removed in the early stage. This makes the process efficient. GPU cluster is a process of processing graphics and the GPU is at every node. This is used to make visualization of the data set. The data set used in the project is first checked if there is any null value in the data set and any unnecessary value present in the dataset. First, the values that are not necessary for the project is removed and then the data is used for analysis and visualization. This improves the accuracy of the project and makes the analysis and visualization more accurate (Arifuzzaman *et al.* 2019). The analysis is done on google collab. This makes the analysis more effective.

4.1 Descriptive statistics analysis

This analysis is crucial for mentioning the essential features of the selected data for running the simulation program in the ***"Python software".*** Thus, this part discusses the importance of the selected data for describing the requirement of data science in the sustainability measurements of an organization (Gosling-Goldsmiths, 2018). The selected data is from Alibaba and its previous operational status. Thus, researchers have analyzed these data to analyze the performance of the organization and the possibilities of implementation of different technologies within the organization. This shows the type of data and importance of those data in the dataset selected for the analysis part of this research. Thus, researchers have developed the concept of the application of modern technological tools for improving the sustainable performance of the organization. According to Amrhein *et al.* (2019), increasing sustainability is the new strategy of global organizations for improving their overall performance, Thus, the data selected is based on the overall performance and sustainability measurements of the selected organization.



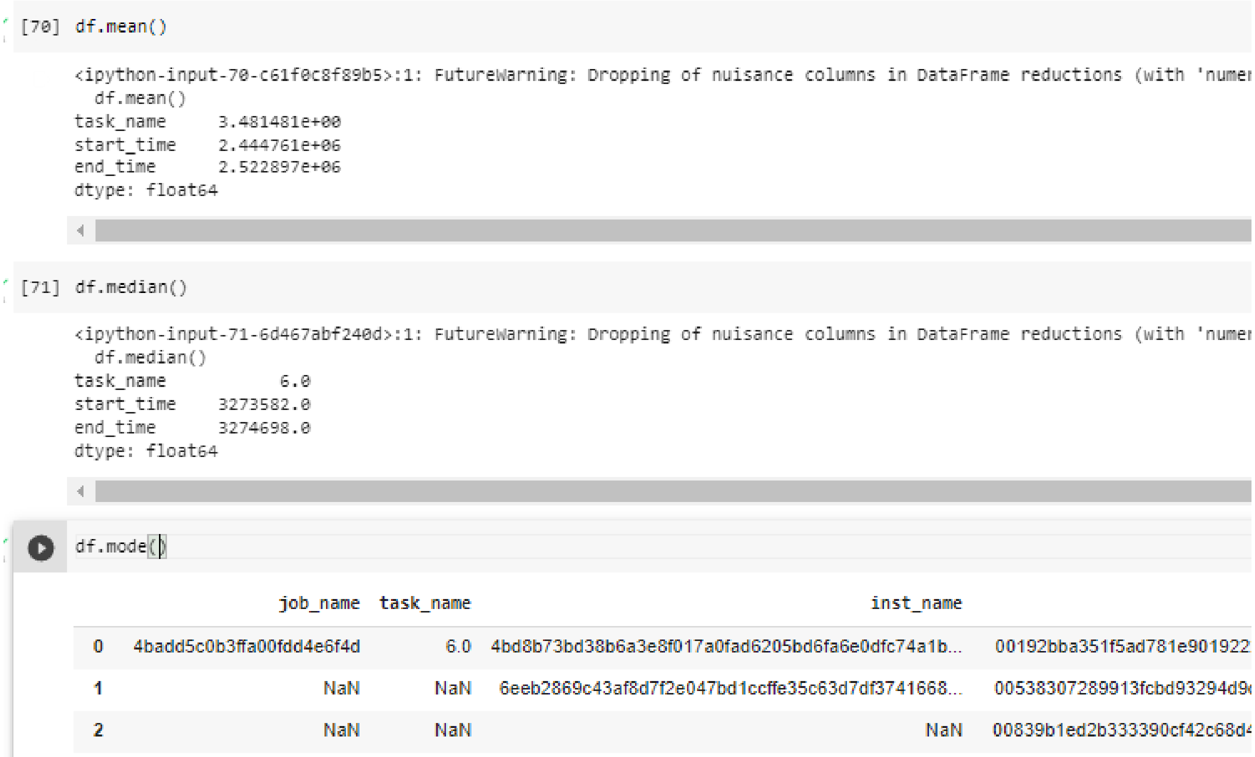
```
[193] df.describe()
```

	start_time	end_time
count	2.404030e+05	2.114500e+05
mean	2.470717e+06	2.511981e+06
std	7.778351e+05	7.817347e+05
min	6.315350e+05	6.338770e+05
25%	1.826109e+06	1.844228e+06
50%	2.475435e+06	2.529646e+06
75%	3.156823e+06	3.210635e+06
max	3.771570e+06	3.772584e+06

Figure 5: Descriptive analysis

(Source: Created by the learner on google collab)

The above picture shows the important library that is imported. The pandas and NumPy are the two important libraries imported into the project.



```
[70] df.mean()

<ipython-input-70-c61f0c8f89b5>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numer
  df.mean()
task_name     3.481481e+00
start_time    2.444761e+06
end_time      2.522897e+06
dtype: float64
```

```
[71] df.median()

<ipython-input-71-6d467abf240d>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numer
  df.median()
task_name           6.0
start_time    3273582.0
end_time      3274698.0
dtype: float64
```

```
df.mode()
```

	job_name	task_name	inst_name	
0	4badd5c0b3ffa00fdd4e6f4d	6.0	4bd8b73bd38b6a3e8f017a0fad6205bd6fa6e0dfc74a1b...	00192bba351f5ad781e901922
1	NaN	NaN	6eeb2869c43af8d7f2e047bd1ccffe35c63d7df3741668...	00538307289913fcbd93294d9
2	NaN	NaN	NaN	00839b1ed2b333390cf42c68d4

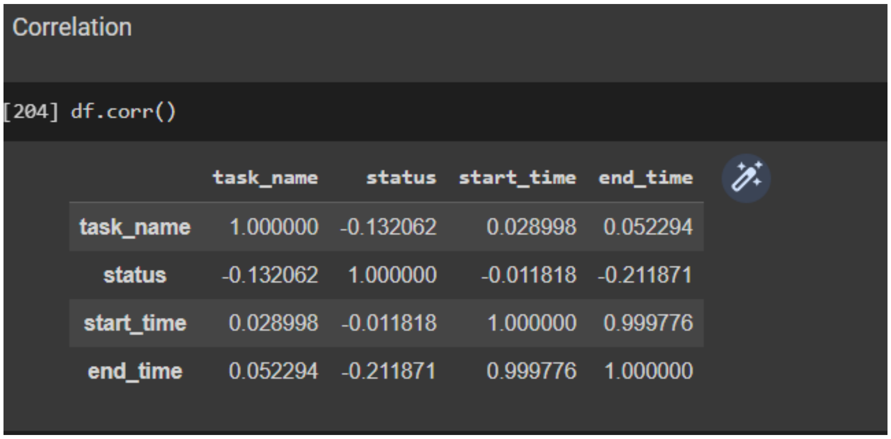
Figure 6: Calculate mean, median, mode

(Source: Created by the learner on google collab)

The pandas work as a wrapper over NumPy and matplotlib, and combine them. This help to access the command of this library in a single click. The library NumPy gives an array that is faster than a normal python list.

4.2 Inferential statistics analysis

The statistic is generated for measuring the sample taken based on the subject organization to compare these data with the treatment group for making generalizations from those data. Researchers have defined the population in the dataset and their planning procedure as per the requirement of the analysis. According to Gareiou *et al.* (2018), this process involves the analysis of the selected data for achieving the expected goal of the paper as per the generation of the outcome of the data science application. Researchers have developed inner relationships between different elements and components in the dataset for analyzing the data and creating the ***"heat map"*** for the related analysis (Bzdok *et al.* 2019). Thus, they have created the heat map for describing the required improvement in the overall system with the ***"regression analysis"*** from the dataset. Researchers have proved the relationship between the application of data science and organizational development as per the SDG goals through this analysis and map generations.



Correlation

```
[204] df.corr()
```

	task_name	status	start_time	end_time
task_name	1.000000	-0.132062	0.028998	0.052294
status	-0.132062	1.000000	-0.011818	-0.211871
start_time	0.028998	-0.011818	1.000000	0.999776
end_time	0.052294	-0.211871	0.999776	1.000000

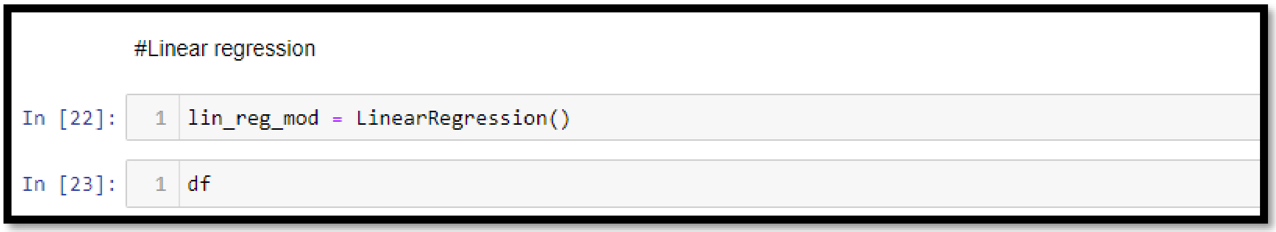
Figure 7: Correlation analysis

(Source: Created by the learner on google collab)

The above picture shows the data set visually. The Correlation analysis is assigned to the data set and the df command shows the data set.

4.3 Machine learning

This is a process of analyzing the automation of the data analysis model involved in the paper. Researchers have developed the concept of ***“artificial intelligence”*** for developing models of machine learning in data analysis. This offers a minimum amount of human reference for describing the developed model during the designing of the data science (De Felice *et al.* 2020). ***“Machine learning”*** is applied in different sections of organizations and the data analysis processes as an interactive process for exploring new data sets to develop organizational performance. Thus, this section developed an algorithm for finding the techniques of ***“machine learning”*** with the research as per the requirement of the paper. According to Berg *et al.* (2019), the application of modern technologies increases around the globe, and along with this the application of different data science techniques also increasing. This algorithm of the researchers developed the security system for the data science of the organization by increasing capabilities of preparing the dataset as per the requirement of the research.



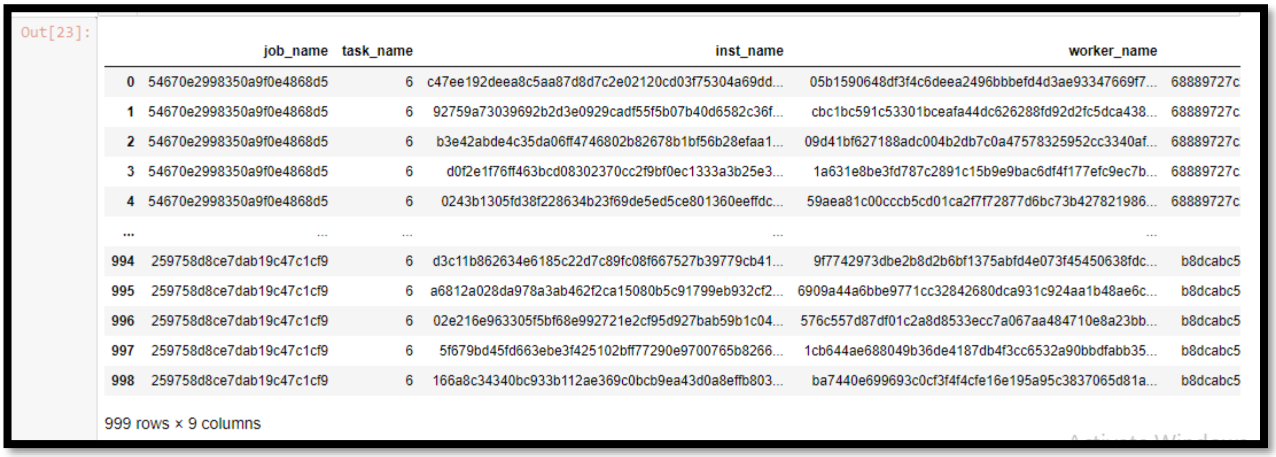
#Linear regression

```
In [22]:  lin_reg_mod = LinearRegression()

In [23]:  df
```

Figure 8: Code for linear regression

(Source: Created by the learner on google collab)

The above figure demonstrates Python code for linear regression. The Linear regression method in the machine Learning Algorithms helps in prediction analysis. It has helped in modeling a target prediction value based on the independent variables.



Out[23]:

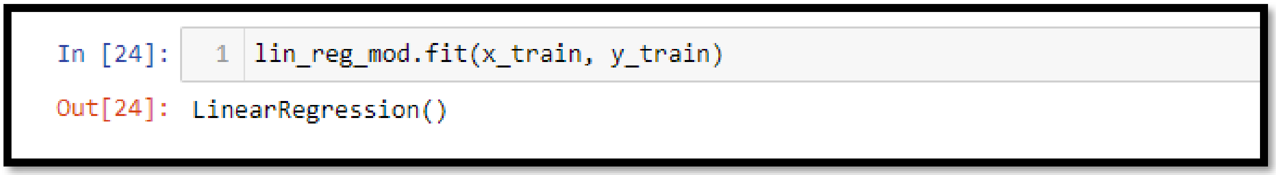
	job_name	task_name	inst_name	worker_name	
0	54670e2998350a9f0e4868d5	6	c47ee192deea8c5aa87d8d7c2e02120cd03f75304a69dd...	05b1590648df3f4c6deea2496bbbefd4d3ae93347669f7...	68889727c
1	54670e2998350a9f0e4868d5	6	92759a73039692b2d3e0929cadf55f5b07b40d6582c36f...	cbc1bc591c53301bceafa44dc626288fd92d2fc5dca438...	68889727c
2	54670e2998350a9f0e4868d5	6	b3e42abde4c35da06ff4746802b82678b1bf56b28efaa1...	09d41bf627188adc004b2db7c0a47578325952cc3340af...	68889727c
3	54670e2998350a9f0e4868d5	6	d0f2e1f76ff463bcd08302370cc2f9bf0ec1333a3b25e3...	1a631e8be3fd787c2891c15b9e9bac6df4f177efc9ec7b...	68889727c
4	54670e2998350a9f0e4868d5	6	0243b1305fd38f228634b23f69de5ed5ce801360eeffdc...	59aea81c00cccb5cd01ca2f7f72877d6bc73b427821986...	68889727c
...	...	...	...	...	
994	259758d8ce7dab19c47c1cf9	6	d3c11b862634e6185c22d7c89fc08f667527b39779cb41...	9f7742973dbe2b8d2b6bf1375abfd4e073f45450638fdc...	b8dcabc5
995	259758d8ce7dab19c47c1cf9	6	a6812a028da978a3ab462f2ca15080b5c91799eb932cf2...	6909a44a6bbe9771cc32842680dca931c924aa1b48ae6c...	b8dcabc5
996	259758d8ce7dab19c47c1cf9	6	02e216e963305f5bf68e992721e2cf95d927bab59b1c04...	576c557d87df01c2a8d8533ecc7a067aa484710e8a23bb...	b8dcabc5
997	259758d8ce7dab19c47c1cf9	6	5f679bd45fd663ebe3f425102bff77290e9700765b8266...	1cb644ae688049b36de4187db4f3cc6532a90bbdfabb35...	b8dcabc5
998	259758d8ce7dab19c47c1cf9	6	166a8c34340bc933b112ae369c0bcb9ea43d0a8effb803...	ba7440e699693c0cf3f4f4cfe16e195a95c3837065d81a...	b8dcabc5

999 rows × 9 columns

Figure 9: Output of Linear Regression

(Source: Created by the learner on google collab)

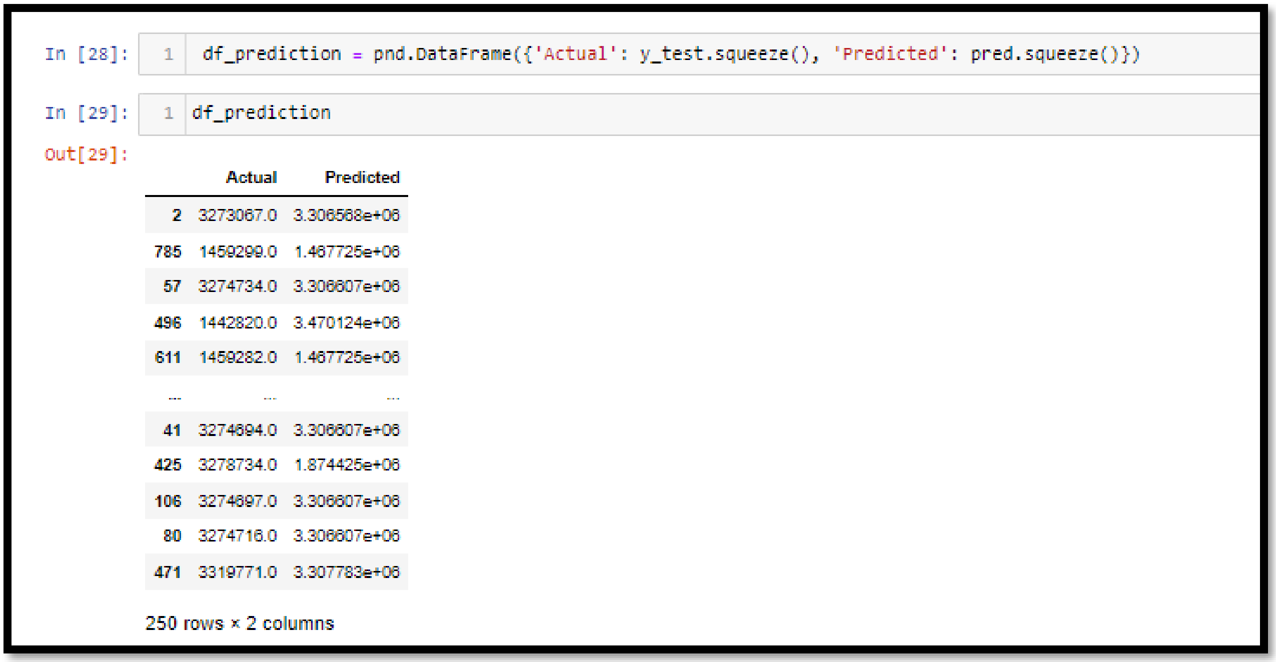
The figure shows the output of the linear regression method. Job_name, task_name, inst_name, and worker_name are the variables here that are displayed in the output.



```
In [24]:  lin_reg_mod.fit(x_train, y_train)
Out[24]: LinearRegression()
```

Figure 10: Code for linear regression

(Source: Created by the learner on google collab)

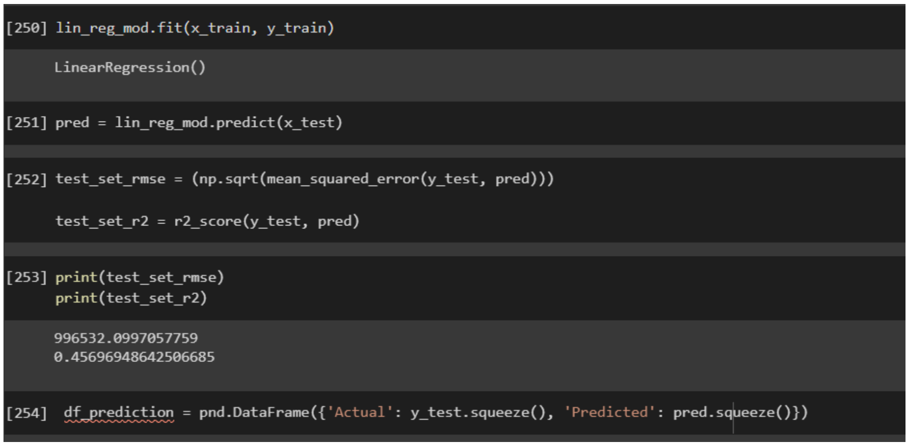


```
In [28]:  df_prediction = pnd.DataFrame({'Actual': y_test.squeeze(), 'Predicted': pred.squeeze()})

In [29]:  df_prediction
Out[29]:
          Actual     Predicted
  2    3273067.0  3.306568e+06
785    1459299.0  1.467725e+06
 57    3274734.0  3.306607e+06
496    1442820.0  3.470124e+06
611    1459282.0  1.467725e+06
...          ...           ...
 41    3274694.0  3.306607e+06
425    3278734.0  1.874425e+06
106    3274697.0  3.306607e+06
 80    3274716.0  3.306607e+06
471    3319771.0  3.307783e+06

250 rows × 2 columns
```

Figure 11: Data frame

(Source: Created by the learner on google collab)

The above figure shows the data frame used for the linear regression model. The fundamental cause of the linear regression model is simplicity and accuracy.

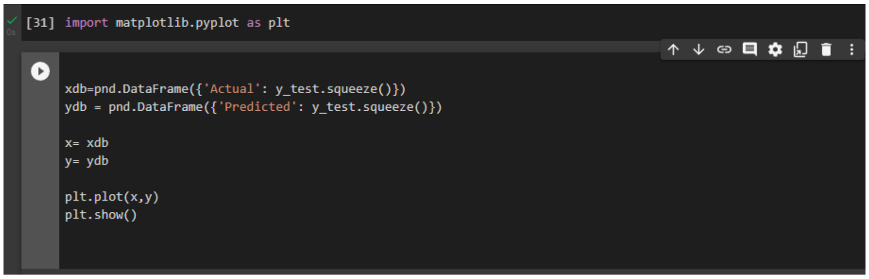


```
[250] lin_reg_mod.fit(x_train, y_train)

      LinearRegression()

[251] pred = lin_reg_mod.predict(x_test)

[252] test_set_rmse = (np.sqrt(mean_squared_error(y_test, pred)))

      test_set_r2 = r2_score(y_test, pred)

[253] print(test_set_rmse)
      print(test_set_r2)

      996532.0997057759
      0.45696948642506685

[254]  df_prediction = pnd.DataFrame({'Actual': y_test.squeeze(), 'Predicted': pred.squeeze()})
```

Figure 12: Multilinear regression model

(Source: Created by the learner on google collab)

The above picture shows the description of the data set. The Multilinear regression model command shows the description of the dataset.



```
[31] import matplotlib.pyplot as plt

xdb=pnd.DataFrame({'Actual': y_test.squeeze()})
ydb = pnd.DataFrame({'Predicted': y_test.squeeze()})

x= xdb
y= ydb

plt.plot(x,y)
plt.show()
```

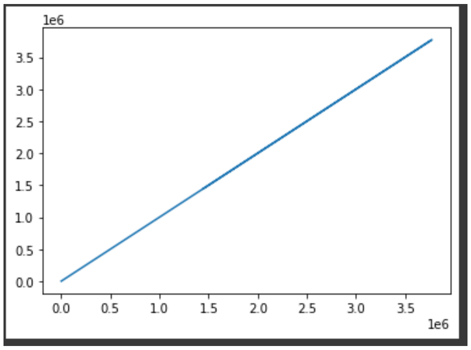


Figure 13: Multilinear regression LinePlot

(Source: Created by the learner on google collab)

The above picture shows the data type and the command multilinear regression LinePlot is used to check the data type in the data set.

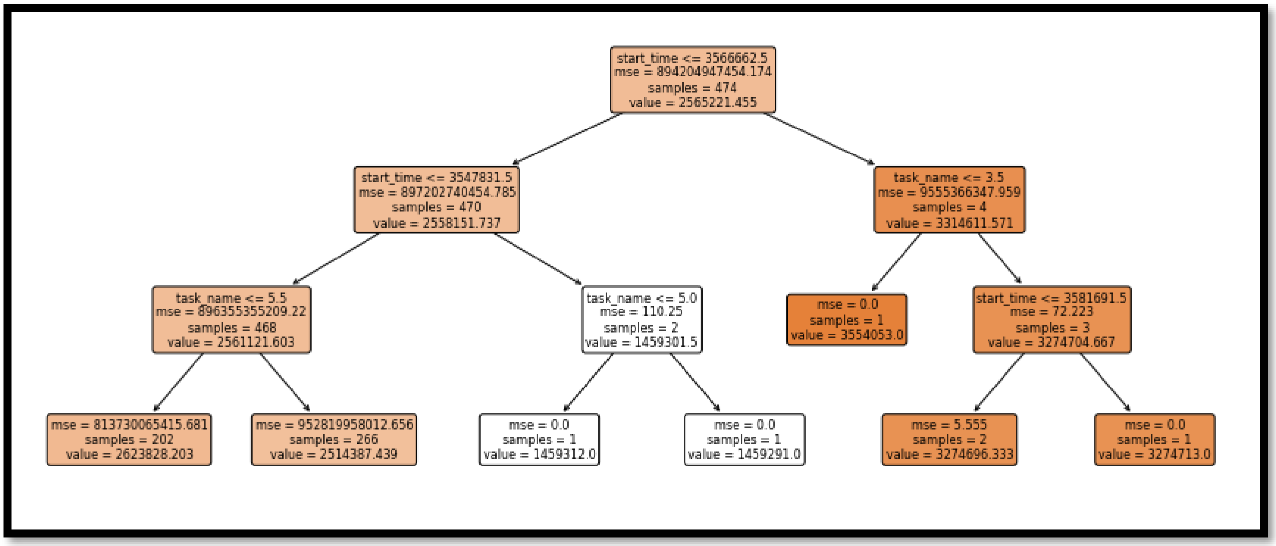


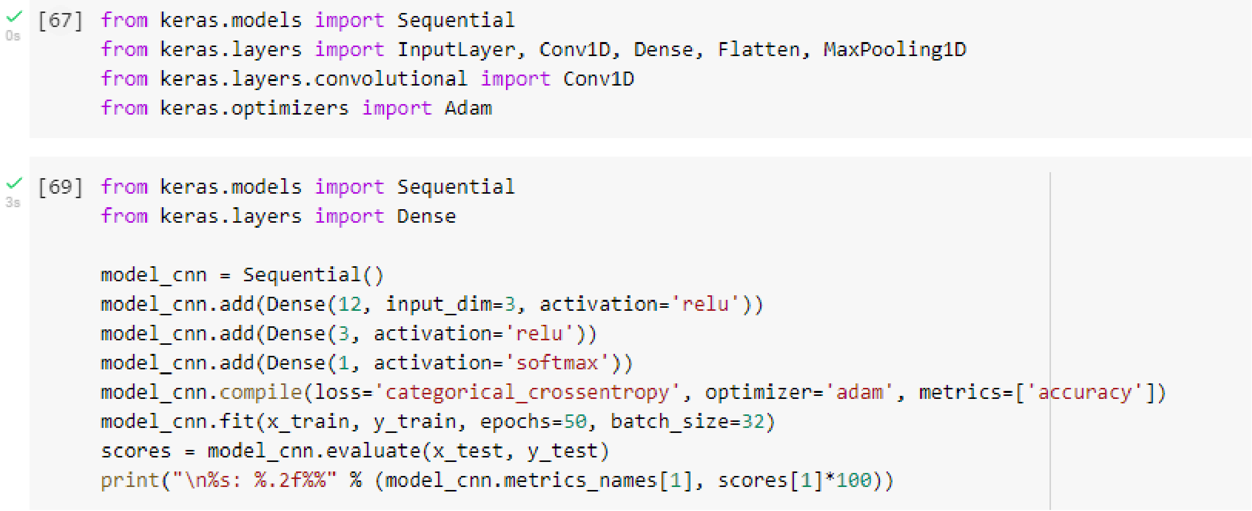
Figure 14: Random forest

(Source: Created by the learner on google collab)

Random Forest is a machine learning algorithm that combines all the output of different decision-tress to reach to a single point. It is used for classification and regression problems. The random forest model uses ransom subsets of the data. With the help of the random subsets of data, it becomes possible for developing a forest of various decision trees. The above figure demonstrated the random forest method for classification and regression.

4.4 Deep learning :

This is an integral part of the application of data science for predicting the model of implementation within an organization. Researchers have analyzed the aspects of defining the dataset as per the given information of Alibaba to prepare the analytical model of data science. This is a process of ***"automated predictive analysis"*** in the paper as per the requirement of the research and its data set (Yadav and Vishwakarma, *et al.* 2020). Researchers have prepared the required algorithm for developing the concept of deep learning along with defining the parameters involved in the predictive model. At the same time, they have discussed the importance of variables involved in the algorithm of the ***"machine learning process"*** so that it can be applied within the system efficiently without any barrier. This algorithm analysis increases the complexity of the system but predicts the success rate of the application within the organization (Roy *et al.* 2019). Hence, researchers have discussed the algorithm development in this section as per their developed model with the system design. Computer programs have been generated by researchers for describing the application of deep learning within the organization along with the implementation of data science techniques in the corresponding organization.



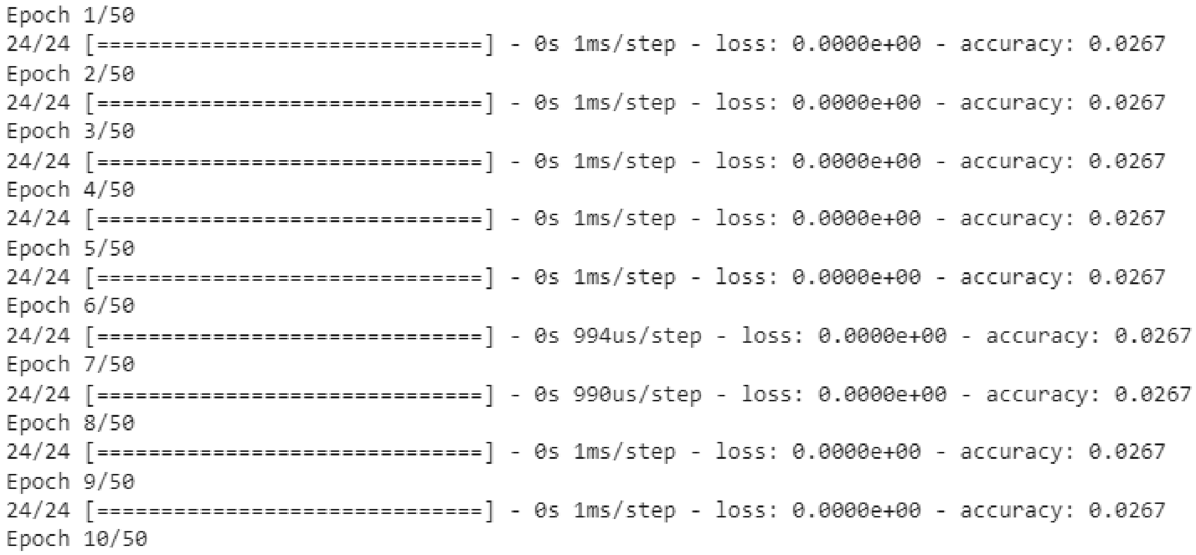
```
[67] from keras.models import Sequential
     from keras.layers import InputLayer, Conv1D, Dense, Flatten, MaxPooling1D
     from keras.layers.convolutional import Conv1D
     from keras.optimizers import Adam
```

```
[69] from keras.models import Sequential
     from keras.layers import Dense

     model_cnn = Sequential()
     model_cnn.add(Dense(12, input_dim=3, activation='relu'))
     model_cnn.add(Dense(3, activation='relu'))
     model_cnn.add(Dense(1, activation='softmax'))
     model_cnn.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
     model_cnn.fit(x_train, y_train, epochs=50, batch_size=32)
     scores = model_cnn.evaluate(x_test, y_test)
     print("\n%s: %.2f%%" % (model_cnn.metrics_names[1], scores[1]*100))
```

Figure 15: CNN

(Source: Created by the learner on google collab)

The CNN shows the missing value in the data set. The above picture shows the missing value present in the dataset.

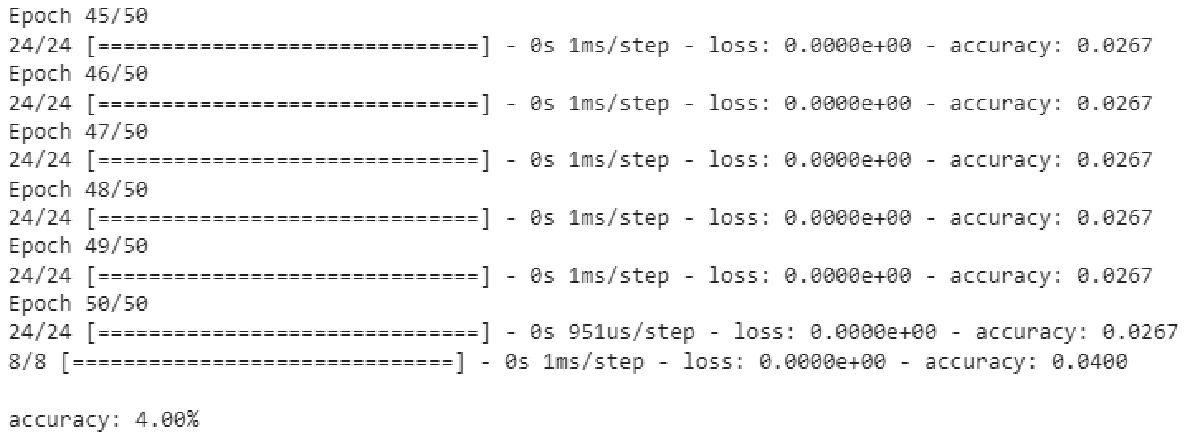


```
Epoch 1/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 2/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 3/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 4/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 5/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 6/50
24/24 [==============================] - 0s 994us/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 7/50
24/24 [==============================] - 0s 990us/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 8/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 9/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 10/50
```

Figure 16: Epoch calculation

(Source: Created by the learner on google collab)

The CNN has been formulated with the above code.



```
Epoch 45/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 46/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 47/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 48/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 49/50
24/24 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0267
Epoch 50/50
24/24 [==============================] - 0s 951us/step - loss: 0.0000e+00 - accuracy: 0.0267
8/8 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.0400

accuracy: 4.00%
```

Figure 17: CNN

(Source: Created by the learner on google collab)

The accuracy score has been evaluated for describing the accuracy score based on the deep learning technique. The CNN model has been applied for evaluating the value of the model's overall accuracy based on the dataset.

5. Recommendations

5.1 Recommendations for the management team

The management team of the organization is an integral part of this application as they are responsible for managing the organizational activities during the application of data science and different modern technologies. Thus, they should follow some important recommendations for increasing the benefits of the application of different technological advancements with the analysis of data science within the organization.

- There are a total of 17 goals in the ***"global compact policies of the UN"*** and the managers should adopt the most relevant ones from all these to apply within the organization through the analysis of data science (Gutberlet, 2021).
- Thus, managers can increase organizational profitability with the application of datasets so that they can manage the workforce and workload as per the requirement of the organization (Sweileh, 2020).
- Managers should develop the dataset for analysis of the IT information and acceptability of the world of their taken policies or strategies in the market for evaluating the efficiency of the application in the organization.

- Moreover, the management team of the organization should increase their expertise by decreasing the communication gap for increasing organizational benefits from the application of big data and different new innovative technologies (Giuliani *et al.* 2021).

5.2 Recommendations for other stakeholders

Other stakeholders are also essential parts of the strategic planning of an organization as they are responsible for maintaining the change within the organization. Such stakeholders are ***"employees, shareholders, customers, government bodies, and regulatory authorities".*** Thus, they are recommended to find the relevant update over the online portal of the organization and assist the managers to implement the change so that the organization can meet obligations of the SDG policies through its application of data science and big data policies. The concept of ***"big data for social good"*** should be incorporated within the working policies for providing key insights of the organization to society with the implementation of visualization (World Health Organization, 2021). The machine learning approach can be used for determining the implementation process by using the data analysis method. In this case, descriptive analysis has been used for gaining information about the dataset and the various factors have been determined for fulfilling the entire process (Goloshchapova *et al.* 2019). The forecasted model can be implemented to predict the marketing strategies for improving the performance of the market.

6. Conclusions and future work

6.1 Future work

"Sustainable development goals" are referred to as the blueprint for becoming successful in global competitiveness by an organization. This allows the organization to invest in the removal of poverty, fighting against inequalities, and reducing climate impact from their internal operation. All these activities by the organization require the development of a proper dataset for preparing the data science application strategy within the organization. The application of data science has a great impact on the creation of ***"disaggregated indicators"*** for ensuring the modernization of the business.

Thus, the overall work will apply to increasing the utilization of modern technologies and machine learning techniques within the organization. Moreover, the ninth SDG Gopal is about ***"sustainable industrialization"*** which is focused in this research on developing sustainable industries as per the required quality assessment for Alibaba. Furthermore, the future work on this aspect is increasing with the increasing concept of industrialization and modern technologies such as***"machine***

learning, IoT, data science, big data, social media, and others". Thus, the application of different technologies will be easier with these findings as it has developed the data analysis model with the data set of Alibaba to develop strategies for establishing SDG policies within the organization.

Machine learning techniques can be applied to describe the evaluation process based on the sustainability analysis. Several kinds of technologies can be implemented to describe the different parameters based on the data approach, as well as the analysis technique has been used for evaluating the performance of the data analysis process (Sanchez-Segura *et al*. 2022). The deep learning technique is an advanced data analysis technique that allows for gathering information about the analysis process for describing the forecasted data based on the market analysis process.

6.2 Conclusions

The overall research funds the implementation of data science policies within an organization by referring to the dataset of Alibaba. This is a multinational technology-based organization that implements several IT technologies within the organization for increasing its overall performance. Hence, the research is relevant to the application of data science and its policies within this organization. Researchers have developed different analysis processes for describing the process as per the given scenario of the organization along with its dataset. Thus, this paper has developed the concept of IT and other smart systems for developing sustainable goals within the organization to increase its competitive advantage around the globe.

Moreover, the overall paper has been developed based on the data available on Alibaba and its performance as per the application of different software processes for creating a secure system within the organization. Thus, this paper can be implemented within the organizational context as per the essential capacity requirement within the organization. The research paper has developed several recommendations for the management team of this organization for implementation of the data science in the organization. This can influence profitability increment within the organization based on its current performance analysis as power the dataset provided.

Machine learning method has been implemented for determining the performance, and accuracy, as well as evaluating the various parameters of the analysis process. The learning process allows for describing the entire statistics, as well as developing a prediction model to predict the outcomes of the analysis process (Huang *et al*. 2022). Concerning method has the ability to obtain the effectiveness of the analysis process and the descriptive statistics allow for meeting the requirements of the analytical method. Machine learning methods can evaluate the different factors of the sustainability of the industrialization process that helps to collect information about the industrialization process. The deep learning method can also use for describing the details of the industry and the appropriate model is required to describe the forecasting analysis. The forecasting analysis technique has been implemented for gathering information about the industrialization process.

Bibliography

Alam, M.M., Torgo, L. and Bifet, A., (2022). A survey on spatio-temporal data analytics systems. *ACM Computing Surveys*, *54*(10s), pp.1-38.

Al-Zubaidi, W.H.A., Thongtanunam, P., Dam, H.K., Tantithamthavorn, C. and Ghose, A., (2020), November. Workload-aware reviewer recommendation using a multi-objective search-based approach. In *Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering* (pp. 21-30).

Amrhein, V., Trafimow, D. and Greenland, S.,(2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don’t expect replication. The American Statistician, 73(sup1), pp.262-270.

Arifuzzaman, S. and Pandey, B., (2019). Scalable mining, analysis and visualisation of protein-protein interaction networks. *International Journal of Big Data Intelligence*, *6*(3-4), pp.176-187.

Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M. and Eren, K., (2019). Ilastik: interactive machine learning for (bio) image analysis. Nature methods, 16(12), pp.1226-1232.

Bohidar, S.K., Surakasi, R., Karuna, M.S., Janardhana, K., Praveen, S.L. and Narayan, P., (2022). Sustainable manufacturing for turning of Inconel 718 using uncoated carbide inserts. *Materials Today: Proceedings*.

Bzdok, D. and Ioannidis, J.P., (2019). Exploration, inference, and prediction in neuroscience and biomedicine. Trends in neurosciences, 42(4), pp.251-262.

Chen, M., Abdul-Rahman, A., Archambault, D., Dykes, J., Ritsos, P.D., Slingsby, A., Torsney-Weir, T., Turkay, C., Bach, B., Borgo, R. and Brett, A., (2022). RAMPVIS: Answering the challenges of building visualisation capabilities for large-scale emergency responses. *Epidemics*, *39*, p.100569.

Dai, H.N., Wang, H., Xu, G., Wan, J. and Imran, M., (2020). Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, *14*(9-10), pp.1279-1303.

De Felice, F. and Polimeni, A., (2020). Coronavirus disease (COVID-19): a machine learning bibliometric analysis. in vivo, 34(3 suppl), pp.1613-1617.

Fathi, M., Haghi Kashani, M., Jameii, S.M. and Mahdipour, E., (2021). Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering*, pp.1-29.

Gareiou, Z. and Zervas, E., (2018). Analysis of Environmental References in the Texts of the Four Gospels Using Descriptive Statistics. Religions, 9(9), p.266.

Giuliani, G., Mazzetti, P., Santoro, M., Nativi, S., Van Bemmelen, J., Colangeli, G. and Lehmann, A., (2020). Knowledge generation using satellite earth observations to support sustainable development goals (SDG): A use case on Land degradation. International Journal of Applied Earth Observation and Geoinformation, 88, p.102068.

Goloshchapova, I., Poon, S.H., Pritchard, M. and Reed, P., (2019). Corporate social responsibility reports: topic analysis and big data approach. *The European Journal of Finance*, *25*(17), pp.1637-1654.

Gossling-Goidsmiths, J., (2018). Sustainable development goals and uncertainty visualization. Unpublished master's thesis]. University of Twente, p.55.

Gutberlet, J., (2021). Grassroots waste picker organizations addressing the UN sustainable development goals. World Development, 138, p.105195.

Huang, J., Tao, Y., Shi, M. and Wu, J., (20220. Empirical Study on Design Trend of Taiwan (1960s–2020): The Evolution of Theme, Diversity and Sustainability. *Sustainability*, *14*(19), p.12578.

Jindal, A., Patel, H., Roy, A., Qiao, S., Yin, Z., Sen, R. and Krishnan, S., (2019), November. Peregrine: Workload optimization for cloud query engines. In *Proceedings of the ACM Symposium on Cloud Computing* (pp. 416-427).

Joshi, A., Morales, L.G., Klarman, S., Stellato, A., Helton, A., Lovell, S. and Haczek, A., (2021), June. A knowledge organization system for the united nations sustainable development goals. In European Semantic Web Conference (pp. 548-564). Springer, Cham.

Lu, C., Ye, K., Xu, G., Xu, C.Z. and Bai, T., (2017), December. Imbalance in the cloud: An analysis on alibaba cluster trace. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2884-2892). IEEE.

Makrani, H.M., Sayadi, H., Dinakarra, S.M.P., Rafatirad, S. and Homayoun, H., (2018), October. A comprehensive memory analysis of data intensive workloads on server class architecture. In *Proceedings of the International Symposium on Memory Systems* (pp. 19-30).

Palkar, S., Thomas, J., Narayanan, D., Thaker, P., Palamuttam, R., Negi, P., Shanbhag, A., Schwarzkopf, M., Pirk, H., Amarasinghe, S. and Madden, S., (2018). Evaluating end-to-end optimization for data analytics applications in weld. *Proceedings of the VLDB Endowment*, *11*(9), pp.1002-1015.

Rivo, E., de la Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M.Á. and Gil, P., (2012). Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology*, *14*(1), pp.73-79.

Roettger, T.B., Winter, B. and Baayen, H., (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, pp.1-7.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H. and Faubert, J., (2019). Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering, 16(5), p.051001.

Sanchez-Segura, M.I., González-Cruz, R., Medina-Dominguez, F. and Dugarte-Peña, G.L., (2022). Valuable Business Knowledge Asset Discovery by Processing Unstructured Data. *Sustainability*, *14*(20), p.12971.

Sweileh, W.M., (2020). Bibliometric analysis of scientific publications on “sustainable development goals” with emphasis on “good health and well-being” goal (2015–2019). Globalization and health, 16(1), pp.1-13.

Szárnyas, G., Prat-Pérez, A., Averbuch, A., Marton, J., Paradies, M., Kaufmann, M., Erling, O., Boncz, P., Haprian, V. and Antal, J.B., (20180, June. An early look at the LDBC social network benchmark's business intelligence workload. In *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)* (pp. 1-11).

Tallent, N.R., Gawande, N.A., Siegel, C., Vishnu, A. and Hoisie, A., (2018). Evaluating on-node gpu interconnects for deep learning workloads. In *International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems* (pp. 3-21). Springer, Cham.

World Health Organization, (2021). Stronger collaboration for an equitable and resilient recovery towards the health-related sustainable development goals: (2021) progress report on the global action plan for healthy lives and well-being for all.

Yadav, A. and Vishwakarma, D.K., (2020.) Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), pp.4335-4385.

Yeung, G., Borowiec, D., Yang, R., Friday, A., Harper, R. and Garraghan, P., (2021). Horus: Interference-aware and prediction-based scheduling in deep learning systems. *IEEE Transactions on Parallel and Distributed Systems*, *33*(1), pp.88-100.

Appendices:

Appendix 1: Code of Python with Google colab

```
import pandas as pnd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt

df=pnd.read_csv("pai_instance_table.csv")
df
df.describe()
df.head()
df.isna().sum()
df.info()
from sklearn import preprocessing
label_encoder=preprocessing.LabelEncoder()
df['task_name']=label_encoder.fit_transform(df['task_name'])
df['task_name'].unique()
df['status']=label_encoder.fit_transform(df['status'])
df
df['status'].unique()
df.corr()
df.bfill()
```