**Data Center Performance Prediction**

**1.1 INTRODUCTION:**

Data centers are simply centralized locations where computing and networking equipment is concentrated for the purpose of collecting, storing, processing, distributing or allowing access to large amounts of data. They have existed in one form or another since the advent of computers.

In the days of the room-sized behemoths that were our early computers, a data center might have had one supercomputer. As equipment got smaller and cheaper, and data processing needs began to increase and they have increased exponentially we started networking multiple servers (the industrial counterparts to our home computers) together to increase processing power. We connect them to communication networks so that people can access them, or the information on them, remotely. Large numbers of these clustered servers and related equipment can be housed in a room, an entire building or groups of buildings. Today's data center is likely to have thousands of very powerful and very small servers running 24/7.

**1.2.OBJECTIVES OF RESEARCH:**

There are four objectives in the design of any high performance data center:

1. Security

2. Availability

3. Scalability

4. Manageability

Today’s business and competitive environment requires that each of these objectives be considered from a complete end-to-end perspective. Clients or users connecting to the data center measure performance by timely access to the desired application data, whether the connection is point to point or via an Internet connection. The user must perceive reasonable response and connection time. Access is no longer measured solely on an arbitrary measure of “network availability” or ping times. This fact does not lesson the goal of developing networks that provide 99.999 percent availability. However, the implication is that the designers of the traditional networking elements must also become “content aware".

**1.3.PROBLEM STATEMENT:**

Data center known to consume lot of electric power. Power generation, we know more harmful effect on environment. So there is urgent need to make data center energy efficient. Finally by this project we expect all electronic gadgetry to be ‘energy efficient’ to possibly achievable limits.

**2.LITERATURE REVIEW:**

Monitoring the cloud performance is an important issue for both the cloud users and the cloud provider. Monitoring facilitates analysis of both the real time and historical data for performance analysis of the cloud. Existing cloud monitoring tools are either provided by the cloud provider itself (such as cloud watch by Amazon) or user can also use these tools provided by the third parties, e.g. the tolls like reveal uptime and cloud status.

Disadvantages:

1. Possibility of Unpleasant Surprises

2. Data Security

### 3.Loss of Control

### 4.Vendor Lock-In

**3.DATA COLLECTION:**

The dataset is originally from the Green data center. The objective of the dataset is to predict whether efficiency of the power consumption, based on certain input measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database.

The datasets consists of several power predictor variables and one target variable, Outcome. Predictor variables includes the Input volts AC, Current in Amps , power factor, Input power in Watts, output power in Watts, time, TP, PDR, Delay . Efficiency in Percentage this is the column that you predict yourself!

**4.METHODOLOGY:**

The datacenter believes that for the industry to make real progress any data centre efficiency metric will need to be part of a measurement methodology designed to calculate a reasonable and fair approximation of the total power efficiency of the service provision from the data centre.

Brief Description of Algorithm Used:

Multiple linear regression:

Machine learning, more specifically the field of predictive modelling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explain ability. In applied machine learning we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

## Linear Regression Model Representation:

[Linear regression](https://en.wikipedia.org/wiki/Linear_regression) is an attractive model because the representation is so simple.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 \* x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Now that we understand the representation used for a linear regression model, let’s review some ways that we can learn this representation from data.

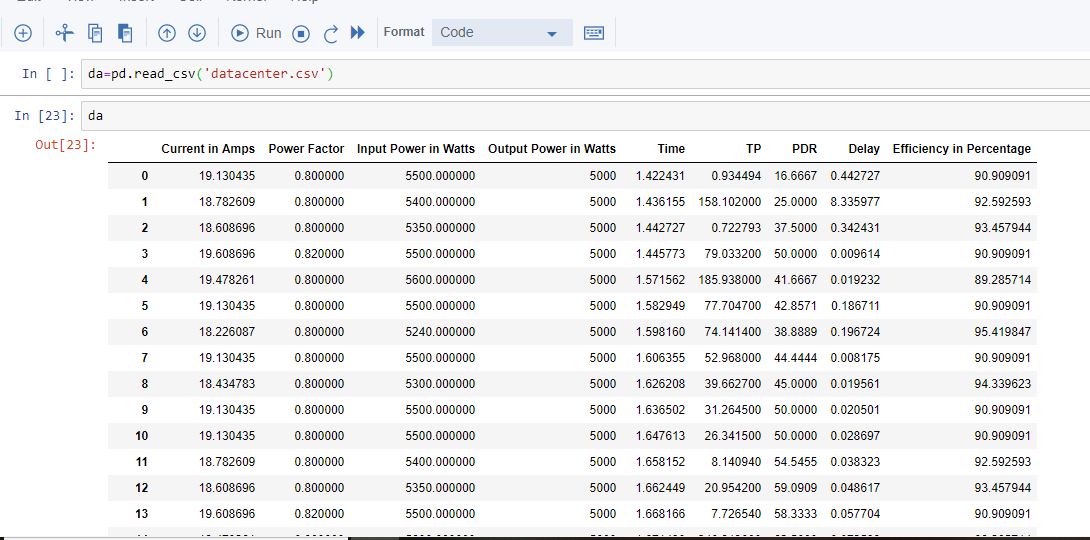
Accuracy Measures:

Multi Linear Regression algorithm is used in this research work. Experiment is performed by using r2\_score.

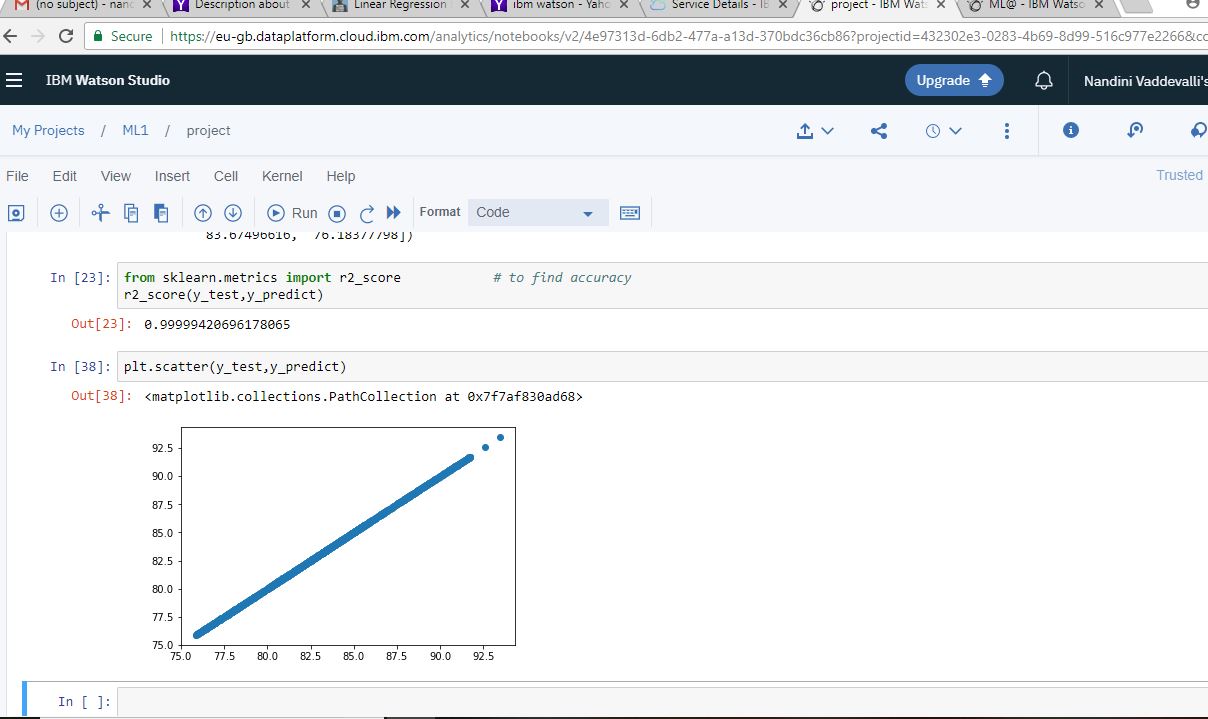
**4.1 Exploratory data analysis:**

4.1.1 Figures and tables:

Dataset for Data Center:



Multi Linear Regression:

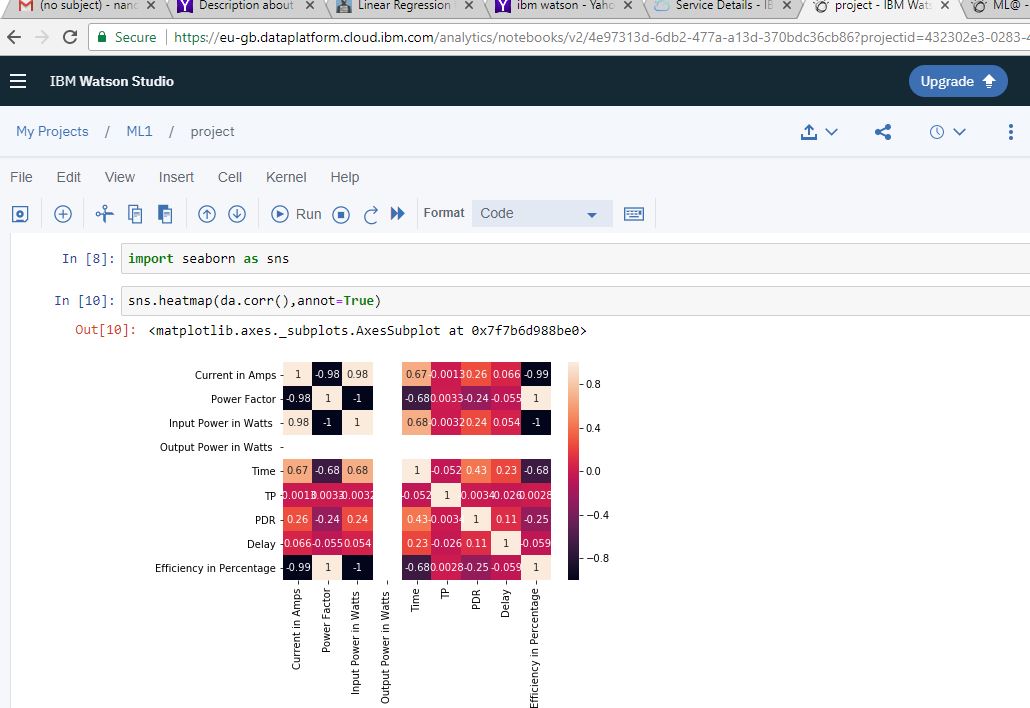


4.2 Statistical techniques and data visualization:

By importing matplotlib.pyplot library we have drawn graphs. We have visualized the percentage.

And we have used the co-relation function to demonstrate the impact of every factor on each other.

Correlation graph:

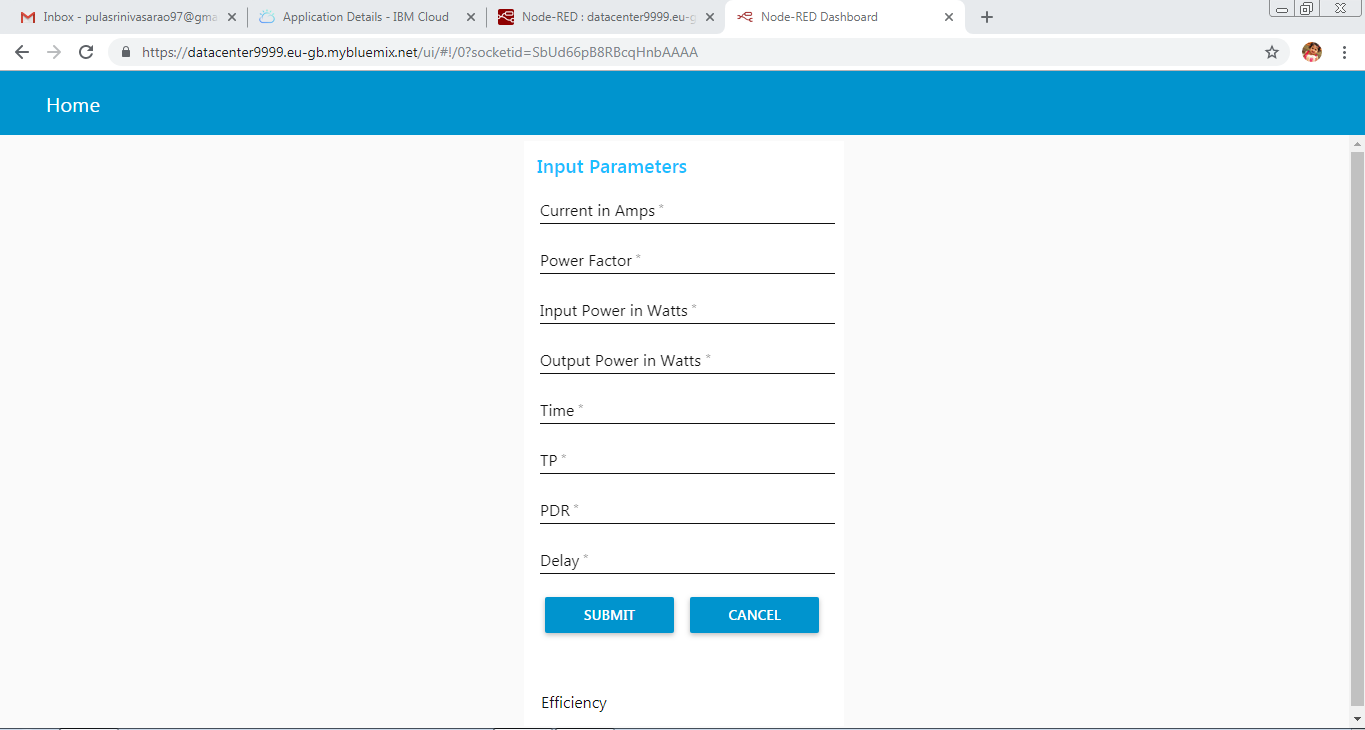


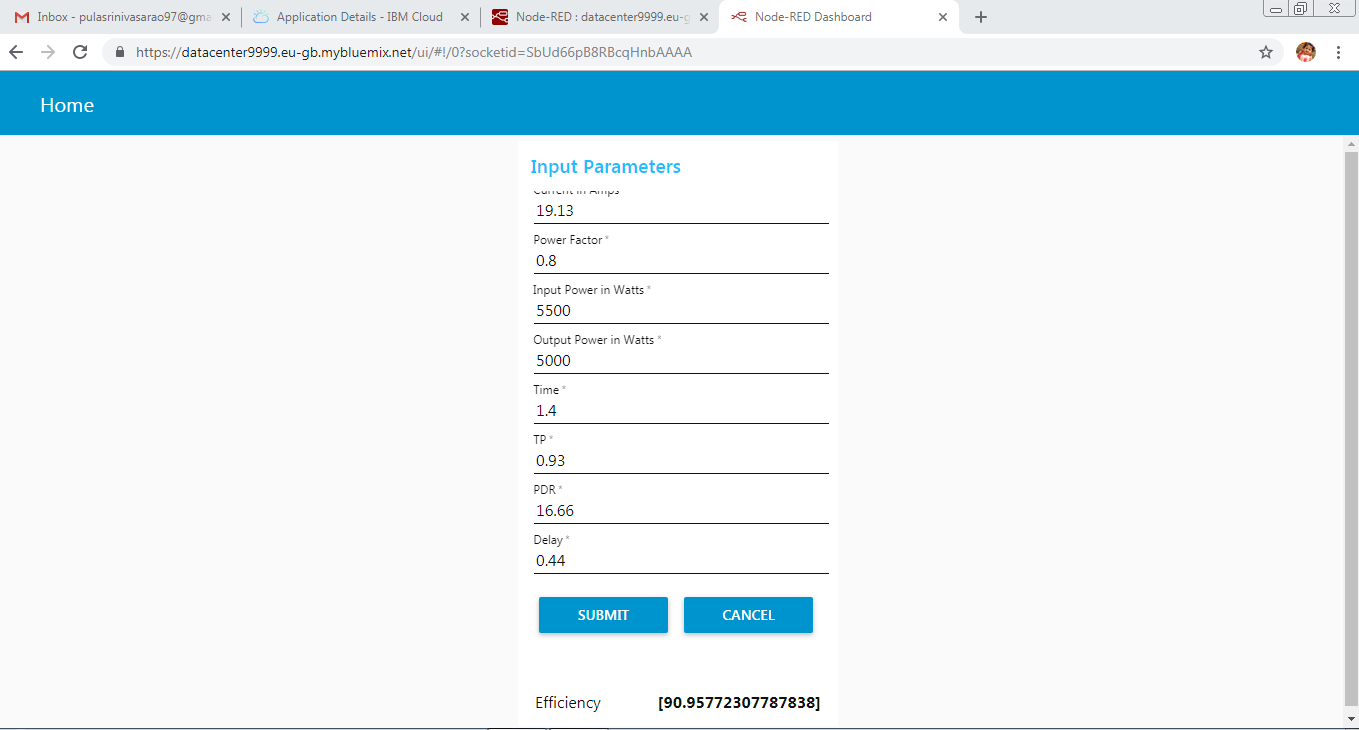
4.3 Data modeling using supervised learning algorithms:

In general we have two types of learning algorithms, supervised and unsupervised learning algorithms and in detail it consists of different techniques like, Multi Linear Regression

Since our model comes under supervised learning algorithm we applied every technique, of the algorithm and based on the accuracy values we obtained we chosen Multi Linear Regression and our model has dependent variable which is Efficiency means the output can be the performance of data center.

**5. FINDINGS AND SUGGESTIONS:**





**6. CONCLUSION:**

One of the important real-world problems is the detection of power consumption. In this study, systematic efforts are made in designing a system which results in the prediction of Efficiency. Experiment is performed on Data center Dataset. Experimental results determine the adequacy of the designed system with an achieved accuracy of 0.98 using the Multi Linear Regression algorithm. In future, the designed system with the used machine learning algorithms can be used to predict. The work can be extended and improved for the Efficiency of data center.