**LINEAR REGRESSION MODEL**

**4. Methodology**

**4.1 Theoretical**

**LINEAR REGRESSION MODEL**

**Linear regression model is a supervised learning technique in machine learning algorithm. Normally, regression model is used to predict some continues values depending on some independent values, as example forecasting a future value depending on past values. In statistics, Linear regression is often used to find the correlation between input and output variables. Linear regression model is very easy to represent in machine learning field where a distinct set of input numbers(X) is used to predict the output numbers(Y), where both are numerical. Linear regression analysis can be separated into two parts, simple linear regression model and multiple variable linear regression model. This paper will mainly analyze the multiple carrier linear regression model, where there will be two or more independent variables and one dependent variable.**

**The formula of linear regression model is,**

**Y=Bo+B1X1+……BnXn+e**

* **Y=the predicted value**
* **Bo=the y-intercept**
* **B1X1= the regression coefficient (B1) and the first independent variable(X1)**
* **BnXn= regression coefficient of last independent variable**
* **e= model error**

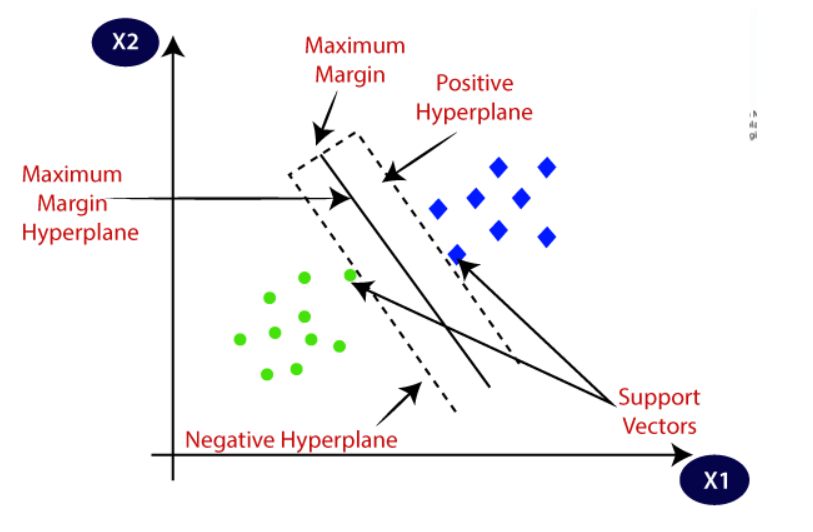
**To find the best fit line for independent variables, multiple variable linear regression mainly calculates following three things,**

* **the regression coefficients**
* **t-statistics of the model**
* **Associated p-value**

**Support vector Machine**

**Support vector machine or SVM is one the popular supervised learning algorithm which can be used for both classification and regression problems. The main objective is to create a best line or decision boundary that can precisely segregate the n-dimensional spaces between different classes so that we can put the new data in correct category in future. The decision boundary is named as hyperplane.**

**SVM normally choses the extreme points or we can call vectors to create the hyperplane. These points are called support vectors and the algorithm is named as support vector machine. Consider below diagram where two different categories are classified using hyperplane or decision boundary.**



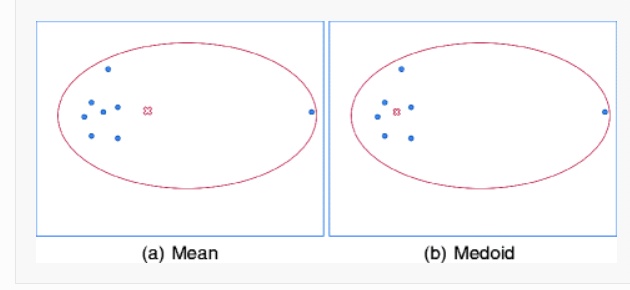
**Fig 1: Support Vector Machine**

**K-Means Clustering**

**K-Means clustering is one of the simple machine learning algorithms to solve any clustering problems.it divides the dataset into different clusters (assume K clusters). The main idea of this model is defining the K-center of each cluster. These centers should be identified in a intelligent way because of different locations cause different results. So, its better to place the centers as far away from each other. The next step is to collect each point belonging a given data and place it to the nearest center. After covering all points, the first step is done and we need find the new centroids and new calculation has to be done with the data set like first step. So, in this way a loop has been created and k-centers change their location step by step until no changes are done.**

***K*-medoids Clustering**

**K-medoids clustering is a variant of K-means clustering which is more robust to noises and outliers. In this clustering method it doesn’t use mean points like k-means clustering rather than it uses actual points in the cluster to represent it. Medoid is the most centrally located object of the cluster with very minimum sum of distance compare to other points. In below figure we can see the difference between means and medoid method.**



**Fig 2: K-means and K-Medoids Clustering**

**Lime Library**

**The Main purpose of LIME, Local Interpretable Model-agnostic Explanations is to explain what machine learning are doing. LIME can explain tabular models, test classifiers and also image classifiers. Lime normally explains what happens in our predictions when give variations of data in our model. This is kind of interpretable models that is used to explain the black box of machine learning models.**

**4.2 System Model**

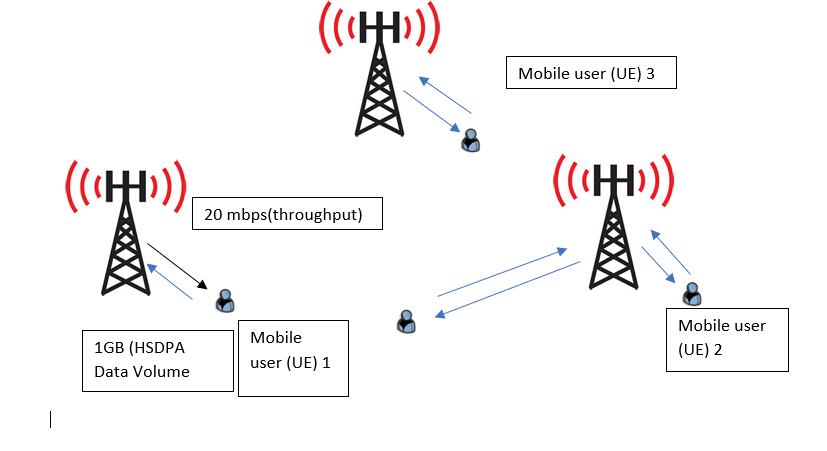
**In the system model, we consider KPI (key performance indicator) data of a UMTS (Universal Mobile Telecommunications System) network in a specific geographical area. In this KPI data we have some term which need to be cleared before we proceed to next step.**

**HSDPA\_Data\_Volume\_MB (MB): High speed data packet access is a packet based mobile telephony protocol mainly used in 3G UMTS network. To increase data speed and data capacity HSDPA is a popular protocol in 3G network. It evolved in WCDMA network which provides five-time faster data speed than previous version. So HSDPA data volume is simply the data consumed by a user in a telecom network by using HSDPA protocol.**

**MAX and Mean HSDPA UE: MAX and Mean HSDPA UE means maximum and average number of users taking traffic or service in a certain telecom zone under a node(nodeb). Nodeb is a telecommunication node which is serving in a telecommunication network where users are using its resource for voice and data service.**

**Cell Availability: Cell Availability is the success rate of a radio access network availability in specific zone or cluster. It helps the network optimization team to find various issue in telecom network during network access and roaming.**

**HSDPA.MeanChThroughput: Throughput is the actual amount of data successfully sent or received between a link or between a user and a mobile tower. It is normally measured as kbps,mbps,gbps.**



**Our main objective in this paper is to predict Data volume in a cellular network by multiple variable linear regression model and worst cell clustering by K-means clustering algorithm. From our KPI data Max HSDPA UE, Mean HSDPA UE and Cell availability are the independent variables where HSDPA means high speed data packet access. HSDPA\_Data\_Volume\_MB is only the dependent variable in our model. We will predict this HSDPA data volume based on the independent variables.**

**The procedure to predict the data volume in a cellular network from max, mean users and cell availability data is as flow chart in fig 1.**

Get the 3G network KPI data from Wireless OSS terminal

Plot the graph and show output result

Met the expectation

Predict the Data Volume

Analyze the KPI data by using Multiple carrier linear regression model and Support vector machine

Plot the graph and show output result

Predict the Data Volume

Analyze the KPI data by using Multiple carrier linear regression model and Support vector machine

Yes

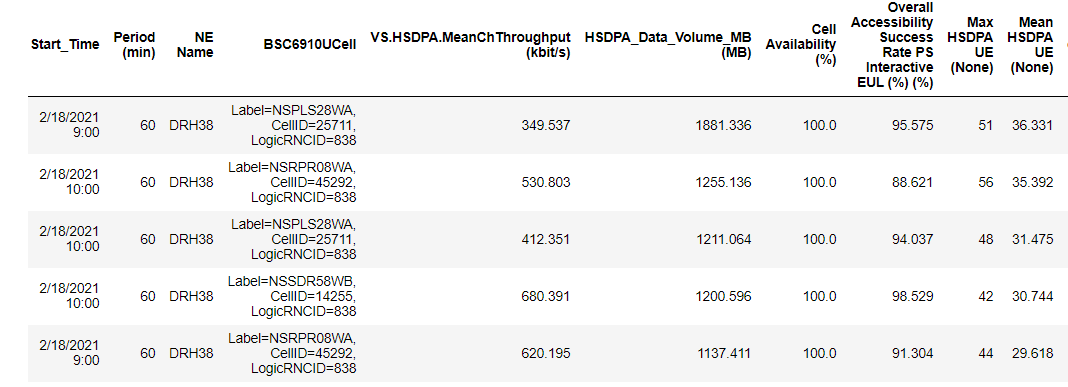
No

Met the expectation

**Fig 3: Flow chart for data volume prediction**

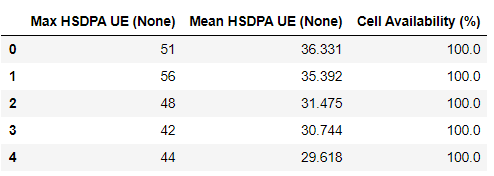
**5. Result Analysis**

**Let’s Explain our data set. We have imported 24 hours KPI data of a UMTS cellular network where we have Time periods, start time, cell name,** **HSDPA.MeanChThroughput,** **HSDPA\_Data\_Volume\_MB,** **Cell Availability,** **Overall Accessibility Success Rate PS Interactive EUL,** **Max HSDPA UE,** **Mean HSDPA UE. We can see our dataset distribution from fig 4.**

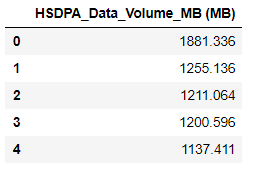


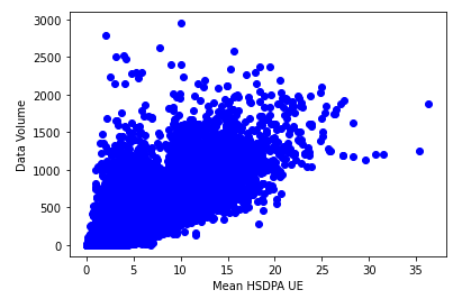
**First of all, we split the data set into training and test data set to fit in the model. We are considering 80% data for our training purpose and rest 30% for our test data. The length of independent and dependent data set is 10938 and test data set is 2735.After data split is completed, we will now fit our data into a multiple variable linear regression model and predict the output. In the same way we will fit our data into a SVM (support vector machine) for comparison purpose. So, after analyzing both model we observed following output from our both models.**

**Independent variables**



**Dependent Variable**



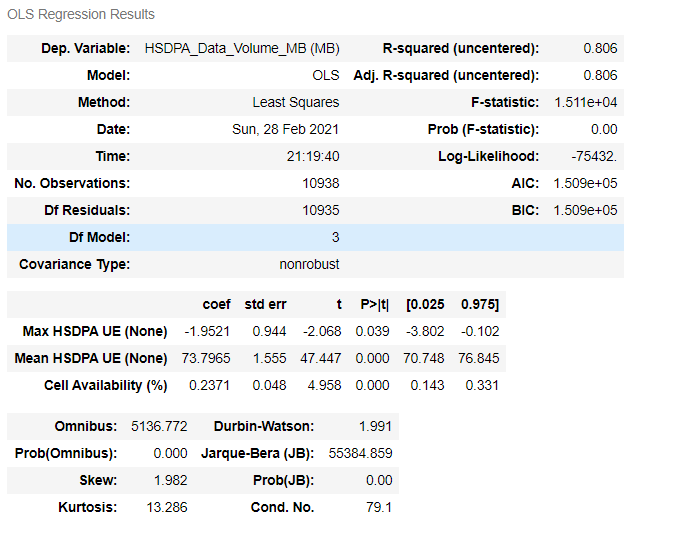


**Fig 4: Data set distribution**

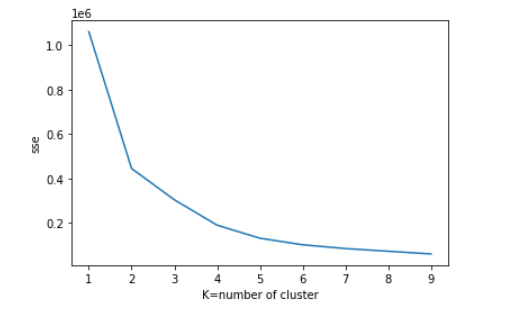
**Model Performance**



**So, here we can see our model accuracy is almost 62% for both model, that means our model can predict data volume 62% correctly from given data sets. For more analysis, we applied the OLS (ordinary least square method) for better understanding our model performance. From this analysis, we will able to find coefficients, R-squared values (how well the data points fit the regression line), P-value, t-value, number of observations. Normally, OLS is statistical method where it can analyze the relationship between one or more independent variables.**



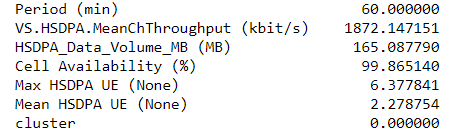
**Form this analysis, we can observe, or R-squared value is .806 which a better output. Also, we can see the coefficients, standard error and t-values from our analysis. “Mean HSDPA UE” is covering 73% coefficient data to find the data volume in our dataset. The next step is to find the number of clusters to determine the worst cells from our KPI data. We applied an Elbow method to find the number of clusters.**



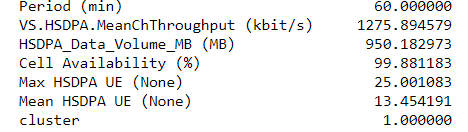
**Fig 5: Elbow method to find the number of clusters**

**From the result, we can fix three clusters for our analysis. Now, we will apply K-Means clustering algorithm to find the worst cells from our dataset. After analyzing, we have found three different clustering output from our model.**

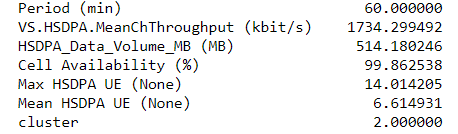
**Cluster 0:**



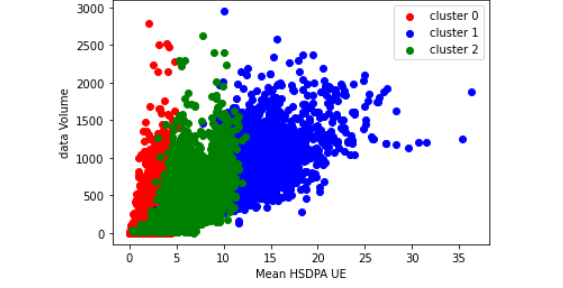
**Cluster 1:**



**Cluster 2:**

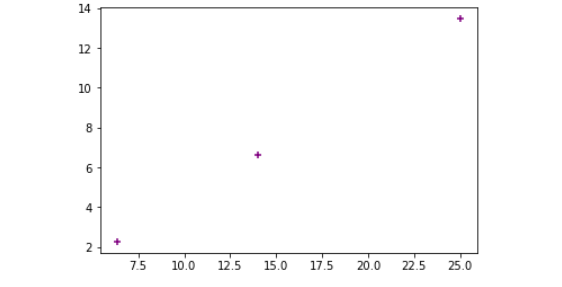


**So, we have divided our data set into three clusters where the minimum data users found in cluster 0 where mean HSDPA UE is only 2.278. In cluster 1, mean HSDPA UE is 13.454 which is better than cluster 2 where mean HSDPA user is only 6.614. So, if we plot the data, we can have a clear idea of our worst cells clustering from our analysis.**



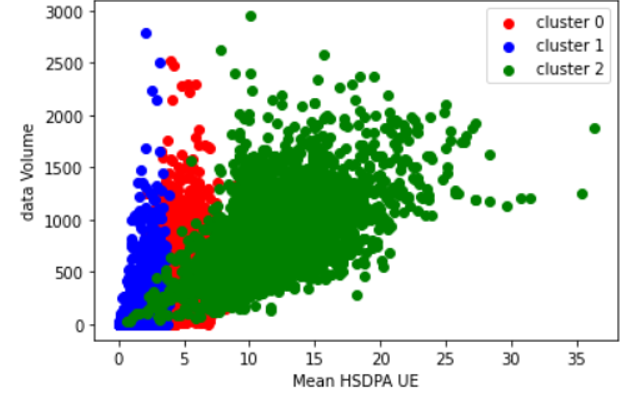
**Fig 6: Worst cell clustering by K-means algorithm**

**From below figure we can observe the centroids points determined by our k-means clustering algorithm.**



**Fig 7: Centroids points in K means clustering**

**We have also applied a K-Medoids and found almost a similar output.**



**Fig 8: Worst cell clustering in K-Medoids algorithm**

# 5.1 Model Interpretation with Lime Library

# We have applied lime library in our model for a better model interpretation. As if we don’t know what is going inside our model, we will not be able to improve it, that’s why model interpretation is very useful method in machine leaning models. so our analysis shows the below prediction

# 

# Fig 9: Model Interpretation by Lime Library.

# From our analysis we can see, the most positive impact from our independent variables is mean HSDPA rather than Cell Availability and Mean HSDPA UE to determine the data volume in a cellular network.