NETWORK TRAFFIC ANALYSIS USING MACHINE LEARNING TECHNIQUES

[1]B.Sarada, [2]D.Sowmya, [3]G.Santhoshi, [4]K.Prashanthi, [5]K.Srithraya
[1]Assitsant Professor, [2,3,4,5]UG Students, Dept. Computer Science and Engineering-Data Science, Mallareddy Engineering college for Women, Hyderabad, India.

ABSTRACT
Recent development in smart devices has leads us to an explosion in data generation and heterogeneity, which requires new network solutions for better analyzing and understanding traffic. These solutions should be intelligent and scalable in order to handle the huge amount of data automatically. These solutions should be intelligent and scalable in order to handle the huge amount of data automatically. With the progress of high-performance computing (HPC), it becomes feasible easily to deploy machine learning (ML) to solve complex problems and its efficiency has been validated in several domains (e.g., healthcare or computer vision).
In this project, we have focused on analyzing network data with the objective of defining network slices according to traffic flow behaviors. For dimensionality reduction, the feature selection has been applied to select the most relevant features from a real dataset (80 Apps) of more than 3 million instances. Then, a unsupervised K-Means clustering is applied to better understand and distinguish behaviors of traffic. The results demonstrated a good correlation among instances in the same cluster generated by the unsupervised learning. A supervised ML technique regression is used to analyze the traffic flow of a network in the form of bytes. This solution can be further integrated in a real environment using network function virtualization.

INTRODUCTION
Under the evolution of smart devices, the networks become increasingly heterogeneous, dynamic, and this has pushed network operators to search for new concepts of management. Consequently, designing network architecture that can handle the heterogeneity and maximize resource utilization efficiency is a major challenge. In this context, several solutions have been proposed including Machine Learning (ML). Applying these concepts can provide an autonomic and intelligent network resources management as well as can yield performance optimization on a large-scale environment, which will accommodate 5G quality of service (QoS) requirements. In fact, ML is deployed to solve complex problems without explicit programming where the algorithm can model and learn the underlying behavior using a training dataset/environment. Its efficiency is validated and it has achieved promising results in several domains and network management is one of them. To efficiently understand the traffic and automatically generate, ML offers these benefits since it can inspect a large quantity of data and find a useful pattern from this data in a reasonable time. It allows the system to come up with rules for automating tasks. The most important advantage of ML is its capability to deal with complex problems. Moreover, analysis of these vast datasets using the machine learning approach promises novel discoveries. Within the field of ML, a broad distinction could be made between supervised and unsupervised learning. The main objective behind supervised learning is to identify a mapping from the input features to an output class, which requires a fully labeled dataset. On the other hand, the objective of unsupervised learning is to find a structure (i.e., pattern) in the inputs without the need of an output class. In practice, the learning ability of ML models is only as good as the given data and features. The increase in the data dimensionality may decrease the performance of an algorithm as well as cause an extra computational cost (e.g., storage and processing) and curse of dimensionality problem. Consequently, to make the raw data suitable for analysis,

preprocessing steps should be applied and feature selection is one of them. In this study, an ML- based slice-defining solution is introduced. The proposed architecture uses network statistics and an offline process for understanding network traffic patterns with a clustering algorithm.

Therefore, the main contributions they are:

Feature selection: We include additional experiments involving features selection. Since the networks is a time-critical system and ML algorithms are as good as the quality of data, we show that the set of features needed to distinguish between applications can be reduced using feature reduction techniques, which can improve the performance of learning algorithms (i.e., classification or clustering) and decreases the overhead both for data collection and model computation.

Traffic clustering: We apply an unsupervised-learning method to find different clusters of the traffic using the previously selected features. These clusters are constructed from traffic with similar behavior.

Cluster analysis: We focus on the analysis of each cluster (i.e., future slice definition) and explain its behavior according to the property of selected features.

PROBLEM STATEMENT

Recent development in smart devices has lead us to an explosion in data generation and heterogeneity, which requires new network solutions for better analysing and understanding traffic. These solutions should be intelligent and scalable in order to handle the huge amount of data automatically.

LITERATURE SURVEY

Specific algorithms based on Machine Learning are being proposed and implemented for network slicing and analyzing the network .

NETWORK SLICING USING MACHINE LEARNING:AN UNSUPERVISED APPROACH TO UNDERSTAND AND SLICE YOUR NETWORK

AUTHOR : Ons Aouedi, KandarajPiamrat, Salima Hamma, JkMenuka Perera

Network Slicing provides the Network AS A Service(NAAS) for different use cases .It partitions the network infrastructure into isolated network slices.

The methodology used here is K-Means Algorithm. Conclusion that can be drawn isclustering is done using ML Algorithm k-means.

NETWORK SLICING AND SOFTWARIZATION: A SURVEY ON PRINCIPLES, TECHNOLOGY AND SOLUTIONS

AUTHOR : L.Afolabi, Tarik Taleb, K.Samdanis

Network Slicing Has been identified as the backbone of the rapidly evolving 5G technology. This paper elaborates network slicing from an end-to end perspective detailing its historical heritage, principal concepts , enabling technologies and solutions .

NETWORK SLICING BASED ON NEXT GENERATION WIRELESS NETWORK

AUTHOR :Saleh Yousefi

Software defined networking (SDN) introduced a programmable and scalable networking solution that enables Machine Learning (ML) applications to automate networks. Network data gathered by the SDN controller will allow data analytics methods to analyze and apply machine learning models to customize the network management.

NETWORKING SLICING FOR 5G: CHALLENGES AND OPPORTUNITIES

AUTHOR : Mustafa Haider, Hisham

Network slicing for 5G is significant attention from the telecommunications industry as a means to provide network as a service (NaaS) for different use cases. 5G mobile networks will carry a large mobile devices and provide faster network connection speed.

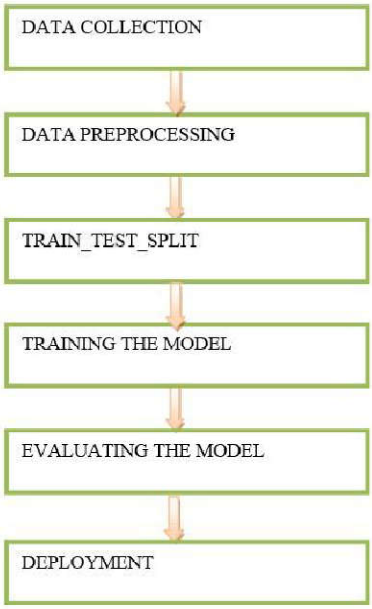
APPLYING BIG DATA, MACHINE LEARNING, SDN/NFV TO 5G TRAFFIC CLUSTERING

AUTHOR :Luong-Vy Le, DoSinh, Li-ping Tung
Traffic clustering, forecasting, and management play a crucial role in improving network efficiency, network quality, load balancing (LB), and energy saving of mobile networks. Especially, in 5G networks,a dense heterogeneous architecture of various types of cells make traffic management become more complicated.
COMPREHENSIVE SURVEY ON MACHINE LEARNING FOR NETWORKING
AUTHOR : Nour Limam, SaraAyuobi
Machine Learning has been enjoying an unprecedented surge in applications that solve problems and enable automation in diverse domains. Primarily, this is due to the explosion in the availability of data, significant improvements in ML techniques, and advancement in computing capabilities.

EXISTING SYSTEM
Feature Selection method and K-Means are used for feature selection and clustering respectively.
The existing solution or system consists of four steps .In the first step the data set is cleaned and prepared for the second step, which is the feature selection. Then an unsupervised ML algorithm is applied to cluster the applications that have similar traffic behavior. Finally network data is analyzed to define network slices according to traffic of each cluster.
The main contributions are:
• Feature selection
• Traffic clustering.
The parameters that are considered are:
• Application related
• Time related

PROPOSED SYSTEM
To overcome the main limitation of the existing system we are going to detect the traffic flow network by implementing linear regression.
This is done by using ML algorithm K-Means and by using LR the data that is taken is divided or clustered using the k-means clustering and then the linear regression is applied to the train and test data set for analyzing the traffic flow.
The proposed system has the following parameters,
Application related
Time related
Bandwidth related

SYSTEM ARCHITECTURE



METHODOLOGY

To implement this project, we have used following modules:

Step 1-Data Collection : We have used a dataset which contains information about different networks.

Step 2-Data preprocessing :

Preprocessing is the step where we prepare the data for Modeling.

High cardinality columns

Converting data types

Handling null values

Standardization

Encoding

Step 3-Train_test_split : The main purpose of this step is for model evaluation.

Step 4-Training the Model : Algorithms used:

1.K-Means

2.Linear Regression.

Step 5-Evaluating the Model :

To check for the Minimum error we used three methods:

RMSE-Root Mean Square Error

MSE-Mean Square Error

MAE-Mean Absolute Error

Step 6-Deployment

o Comparing the Actual output and the predicted output to see how much accurate our model is.

IMPLEMENTATION

In this project we implemented two algorithms K-Means Clustering and Linear Regression.

6.1 K-Means

K-Means Clustering is a type of unsupervised learning algorithm that is used to address clustering issues in data science or machine learning.

It is an iterative algorithm that separates the unlabeled dataset into k distinct clusters, each of which contains only one dataset and shares a set of characteristics.
The algorithm starts with an unlabeled dataset as its input, divides it into k clusters, and then repeats the process until it runs out of clusters to use. In this algorithm, the value of k should be predetermined.
The two main functions of the k-means clustering algorithm are:
Through an iterative process, chooses the best value for K centre points or centroids .Each data point is assigned to the nearest k-center.
A cluster is formed by the data points that are close to a specific k-center.
Elbow Method:
 One of them ost well-liked techniques for determining the ideal number of clusters is the Elbow method.

 The WCSS value concept is used in this technique.

 Theterm"totalvariationswithinacluster"isabbreviatedas"WCSS,"whichstandsforWithin ClusterSumofSquares.

 ThefollowingformulacanbeusedtodeterminethevalueofWCSS(for3clusters)

 WCSS= ΣPiinCluster1 distance(PiC1)2 +ΣPi inCluster2distance(PiC2)2+ΣPi inCLuster3distance(PiC3)2

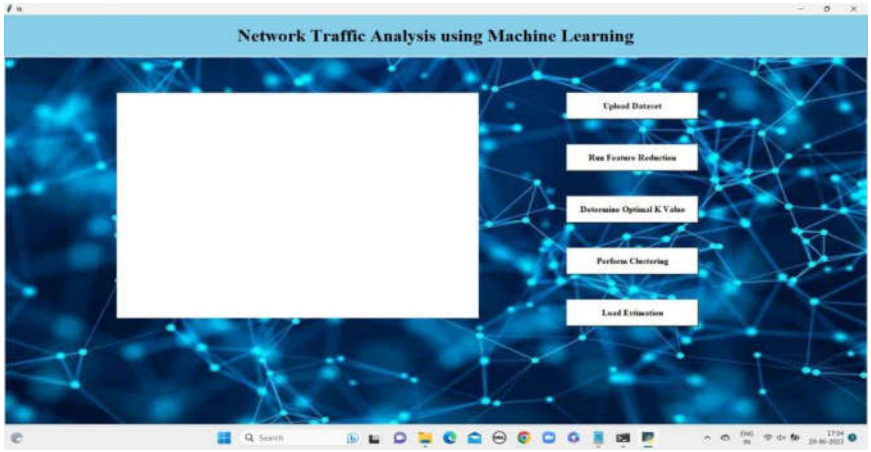
 In the above formula of WCSS,

 ΣPi in Cluster1 distance(PiC1)2: It is the sum of the square of the distances betweeneach data point and its centroid within a cluster1 and the same for the other two terms.
To find the optimal value of clusters ,the elbow method follows the below steps:
It executes the K-means clustering on a given dataset for different K values (rangesfrom1-10).
For each value of K, calculates the WCSS value.

Plots a curve between calculated WCSS values and the number of clusters K.

The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.
Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow Method

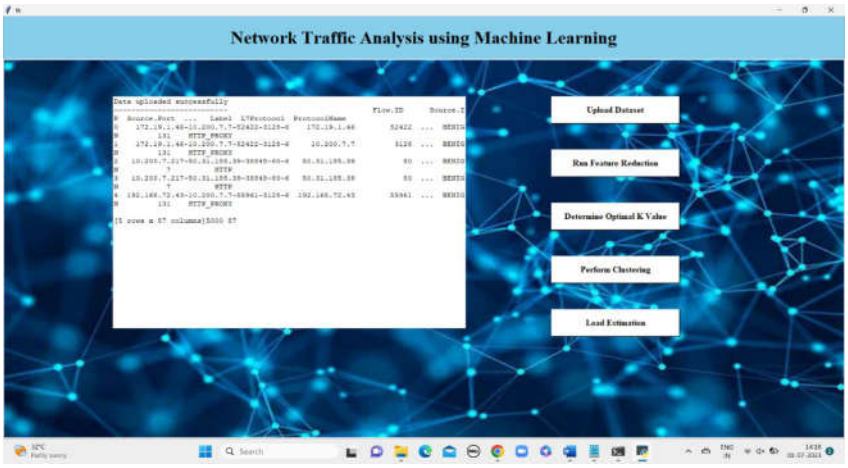
RESULTS



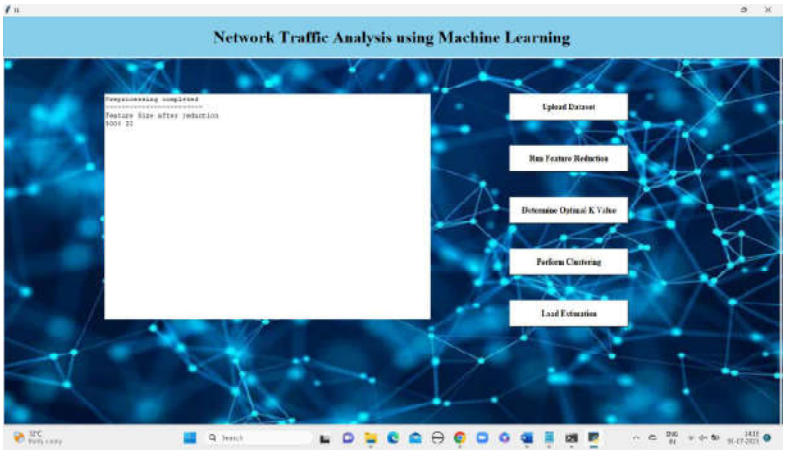
This is the GUI of our model created by using Tkinter . It consists of five buttons. They are :
Upload Dataset ,
Run Feature Reduction,
Determine Optimal K value,

Perform Clustering,

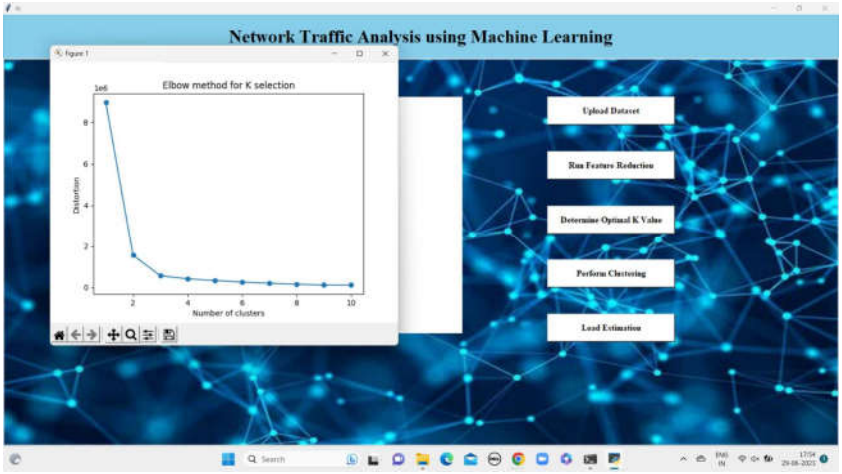
Load Estimation.



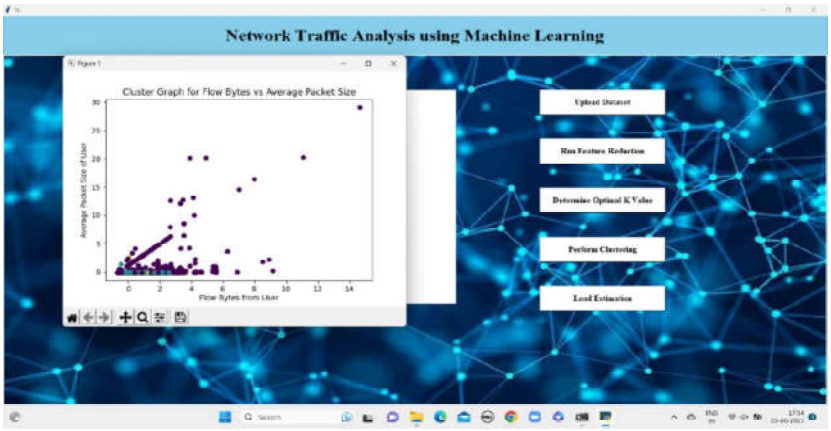
First step is uploading dataset. When we upload dataset the dataset get's uploaded. Upload dataset button displays top five row of data set.



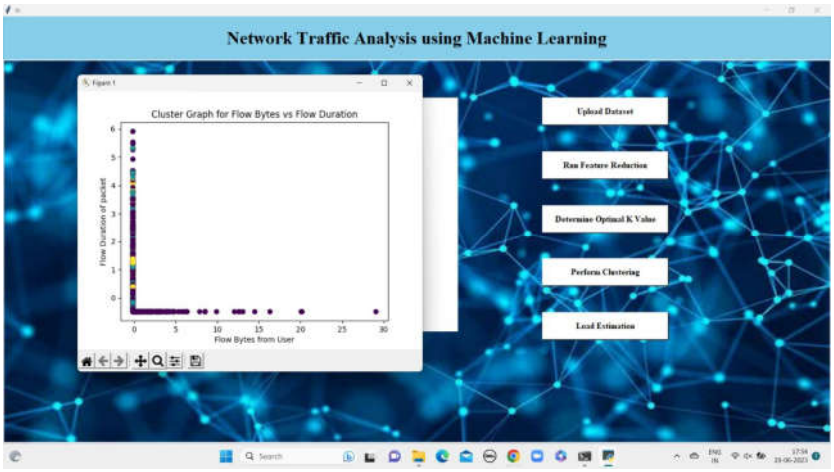
Second step is performing preprocessing and feature reduction. Run feature reduction button displays how many samples and features we have taken in our project



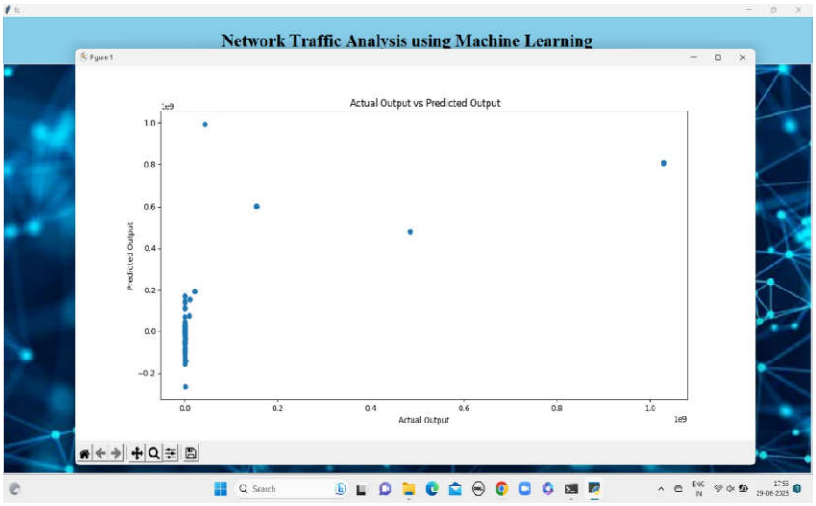
Next step is determining optimized k value. By using elbow method we have determined optimal value of k



The above graph is showing the number of clusters



The above graph is showing the traffic flow between the network in the form of clusters



Network traffic prediction using machine learning

CONCLUSION

In this project, K-Means clustering (Unsupervised machine learning algorithm) is used to cluster on the given dataset which is considered after performing feature selection by selecting relevant features. The features are classified according to application related, time related and bandwidth related features. The network traffic flow is analyzed by using Linear Regression method(Supervised machine learning algorithm) and the behavior is shown.

FUTURE SCOPE

For future work, other clustering algorithms such as DBSCAN (Density based spatial clustering of applications with noise) or hierarchical clustering should also be applied to the dataset in order to compare the performance to the K- mean clustering used in this paper. Implementation of network slicing infrastructure should also be done in order to bench-marking the signaling and overhead generated by this solution

REFERENCES

1. X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "Aiassisted network-slicing based next- generation wireless networks," IEEE Open Journal of Vehicular Technology, vol. 1, pp. 45–66, 2020.
2. R. Fantacci and B. Picano, "When network slicing meets prospect theory: A service provider revenue maximization framework," IEEE Transactions on Vehicular Technology, vol. 69, no. 3, pp. 3179–3189, 2020.
3. R. Boutaba, M. A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. M. Caicedo, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," Journal of Internet Services and Applications, vol. 9, no. 1, pp. 1–99, 2018.
4. X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network slicing for 5g: Challenges and opportunities," IEEE Internet Computing, vol. 21, no. 5, pp. 20–27, 2017.
5. M. H. Abidi, H. Alkhalefah, K. Moiduddin, M. Alazab, M. K. Mohammed, W. Ameen, and T. R. Gadekallu, "Optimal 5g network slicing using machine learning and deep learning concepts," Computer Standards & Interfaces, p. 103518, 2021.