MSc Project Report
School of Engineering and Computer Science
University of Hertfordshire

COMMUNICATION NETWORKS TRAFFIC PREDICTION USING MACHINE LEARNING

Report by
Aslihan Reyhanoglu

Supervisor
Iosif Mporas

Date
05/09/2019

DECLARATION STATEMENT

I certify that the work submitted is my own and that any material derived or quoted from the published or unpublished work of other persons has been duly acknowledged (ref. UPR AS/C/6.1, Appendix I, Section 2 – Section on cheating and plagiarism)

Student Full Name: Aslihan Reyhanoglu

Student Registration Number: 17019396

Signed: ………………………………………………….
Date: 05/09/2019

ABSTRACT

When machine learning is used effectively and accurately for traffic modelling and forecasting, it will be inevitable for future large data-driven intelligent cellular networks as it can assist in providing control and management of the autonomous network. Moreover, the demand for better radio sources has increased, especially with the increase in cellular data traffic. The Radio Access Network (RAN), which is an element of the communication network and is also effective in resource management, uses Machine Learning, a remarkable olution to develop forecast resource allocation techniques. In addition, the Telecom Italia Big Data Challenge dataset which is publicly open will be utilized in order to apply this project. It includes real-life traffic data of the network in the city of Milan. The dataset involves several types of information named as such as the call detail records (CDRs). SMS in/out, call in/out, and internet traffic activity can be given as examples for CDRs. This Project purposes to examine thoroughly the structure of communication networks and to introduce a data-driven architecture for the practical applications of machine learning techniques to predict internet traffic activity of a network. Regression algorithms in supervised learning have been used based on past research to predict the internet traffic. Two of these algorithms are Linear Regression (LR) and Decision Tree (DT), which are easy to use, and the other two are Support Machine Vector and Feedforward Neural Network to cope with nonlinearity that may result from the temporal and spatial data set. RMSE and MAE are used to evaluate performances of algorithms.

ACKNOWLEDGEMENTS

I would like to express gratitude to my supervisor Iosif Mporas for useful comments and remarks throughthe learning process of this Msc project, and also thank to my family for their support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES AND GRAPHS

GLOSSARY

ABBREVIATIONS

[LR Linear Regression
FNN Feedforward Neural Network
NN Neural Network
BS Base Stations
RMSE Root Mean Square Error
MAE Mean Absolute Error
SVM Support Vector Machine
DT Decision Tree
CDMA Code-Division Multiple Access
FDMA Frequency Division Multiple Access
TDMA Time Division Multiple Access
LSTM Long Short-Term Memory
VNF Virtual Network Functions
SDN Software Defined Network
ML Machine Learning
RAN Radio Access Networks
RRU Remote Radio Unit
]

Chapter 1

Introduction

The development of smartphones with the acceleration of technological changes in recent decades has led to an increase in the production of big data and even a burst. Mobile traffic which is one of the data sources is said to show internet traffic 20 percent by 2021. Data traffic generated by smartphones, in particular, will exceed 86 percent of the entire traffic of mobile data with the rise of different mobile apps. Virtual reality, Internet of vehicles and live streaming can be given as examples for these applications. (Chuai et al,2019)

The increase in traffic of the mobile network will cause dissatisfied of costumers due to traffic overloading. It is known that something needs to be done to increase customer satisfaction by using available radio resources efficiently. It is possible to make better the quality of service supplied by mobile operators while allocating radio resources by analyzing the mobile network and the behaviour of customers. While this analysis is made, it is utilized from the data traffic in radio access network. For instance, forecasting of the use of spectrum can be used to help considerably to share spectrum(Jin,2016)

Machine learning suggests a more intelligent and higher-level control and management of networks to fulfil various requirements of costumers. Therefore, the studies about applying algorithms of machine learning to mobile networks have been increased in recent years. It has been known that next-generation systems such as large-scale MIMOs, device-to-device (D2D) networks, heterogeneous networks constituted by femtocells and small cells, and so on will meet with some technical challenges in the future. Machine learning is one of the most effective artificial intelligence tools, which will be able to become a solution to these technical challenges(Jiang et al., 2016).

Among information obtained from the communication network, one of the most used data sources in the researches is the call record details(CDR) data. For implementations of analysis of CDRs, call patterns, urban computing, and criminal investigation can be given as examples. A CDR, including temporal and spatial information, is a form in which lays practicable data regarding a given action of telephone containing a mobile user of a frame of a communication network. For this reason, CDRs have a major significance while making analysis and optimization in a communication network. Past CDR information analysed and modelled can be used to forecast the upcoming trend of CDRs, in other words, it takes an active role in allocating resources and providing load balance in the network. (Zhang,2019)

The Telecom Italia Big Data Challenge dataset which is publicly open was utilized in order while applying this project. The dataset involves data of call/sms/internet which are named the call detail records(CDRs) of mobile users for each area in Milan, measured during November and December 2013. Objectives of this study are to examine thoroughly the structure of communication networks and to introduce a data-driven architecture for the practical applications of machine learning techniques to predict data traffic of the network of Milan. In this project, the algorithms utilized for internet activity traffic prediction were linear regression(LR), support vector regression (SVR), decision tree(DT) and neural network(feedforward).

Chapter 2

Methodology

This project focussed on which machine learning algorithms are able to be applied to traffic prediction for communication networks to ensure a requirement of a better quality of service, and how this process was followed. Therefore, this project accepted the method of qualitative research including the use of secondary data and resources like literature materials. This data was utilized to increase the deep comprehension of the developments to reinforce the suggestions of successful machine learning methods concentrated on increasing performance of communication networks and service delivery.

The first step of this study was to investigate the related work made on this topic in recent years. Secondly, after obtaining enough information about communication networks, radio access network(RAN), machine learning techniques and big data processing and analysing recent studies on this area, the dataset utilized was chosen and Matlab is a program which will be used during this project. Moreover, taking consideration of the existing machine learning implementations in communication networks, it was decided to use different supervised machine learning algorithms such as linear regression and support vector regression. Another step was to process the dataset to make them easy to use. Following this, the necessary features of the processed information were extracted in order to use for machine learning. Then forecasting was made based on the results of algorithms applied to data. Additionally, root mean square error (RMSE) and mean absolute error (MAE) values were calculated in order to evaluate accuracy and outcomes were compared. Finally, it was identified which is better among machine learning algorithms implemented in this project.

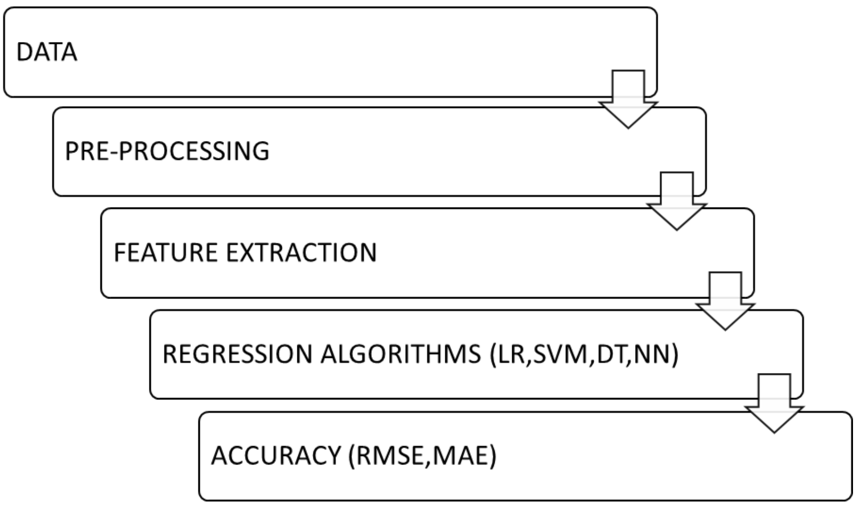


Figure 2.1 Flowchart for methodology

Chapter 3

Theoretical Consideration And Background

3.1 Literature Review

Jiang et al. (2017) studied machine learning paradigms for next-generation wireless networks. They addressed that the next- generation wireless networks are assumed to be supported higher data rates and entire recent implementations working in the paradigm of wireless radio technology. Moreover, it was forecasted that the future smart terminals of 5G are predicted to autonomously reach high spectral bandwidths with the help of the sophisticated spectral efficiency inference and learning(Jiang et al. ,2017). However, machine learning could be a solution in modelling different technic challenges of next-generation systems like heterogeneous networks, the device to device networks and large-scale MIMOs (Chin et al., 2014). In addition, some researches considered that the machine learning models and data mining are so significant for optimization and efficient performance of smart grid systems being excellent applications of next-generation networks. This view supports the study of Jiang et al. (2017). Klaine et al.(2017) presented a few effective machine learning implementations in Self-Organizing Networks (SONs).Moreover, they debated the positive and negative of various machine learning methods and address directions of future study in this field.

Chen et al.(2015) investigated traffic offloading methods in wireless networks and offered reinforcement learning from machine learning algorithms. Additionally, Zhang et al. believed that this view can emerge a novel study direction toward embedding machine learning towards greening cellular networks. Baştuğ et al. (2014) studied the role of proactive caching in 5G wireless networks. They addressed that proactive caching indicate important gains through backhaul savings and growth in service delivery between many consumers and also raised storage capacity. 5G networks showed peak data traffic requests can be decreased by serving the user requests proactively. The research specifies that supervised machine learning algorithms can be combined with a proactive caching procedure to increase the potential of utilizing the predictive capabilities of 5G networks. The predictive abilities of 5G networks provided aid to the large part of the mobile operators to reduce the problems data which continues to strain current networks. On the other side, Pérez et al. (2017) investigated machine learning aided cognitive RAT selection for 5G heterogeneous networks. Furthermore, The Heterogeneous Networks strategy play a role as the key Radio Access Network architecture for future 5G networks. This can cause critical problems to mechanisms of the current user association utilized in cellular networks. Therefore, they proposed the reinforcement learning method over supervised machine learning because of its flexibility and low computational complexity for learning the policy of user association. Alawe et al.(2018) suggested a new method to scale resources of 5G core network by predicting traffic load changes with the use of machine learning algorithms. They applied it by utilizing and training a Neural Network. Additionally, the outcomes indicated that the prediction-based scalability process makes better the threshold-based solutions such as latency to react to traffic change, and delay to have novel resources which are ready to be utilized by the VNF to react to traffic growth. Le et al.(2018) firstly presented a construction to gather and process the big data in the network so that they investigated the connection among the key progress indicators and the traffic patterns. An algorithm using machine learning for traffic prediction was also introduced to cope with wide cells which generate voice traffic and data traffic. In addition, this study utilized the Bayesian network (BN) to investigate the probabilistic relationships between variables. According to this study, autoregressive and neural network methods showed alike performances. However, the gaussian process indicated the best performance as dealing with fast alterations in traffic(Le et al., 2018).It has been developed a process and created a tool for intelligent and fruitful network planning. This tool utilizes and learns from the data based on working history collected over a network anywhere and at any time. In this research, they gathered measurements of user equipment (UE) in accordance with 3GPP MDT functionality to construct quality of service forecast depending on past information. Methods of regression analysis were implemented to predict the Physical Resource Blocks (PRB) per transmitted Mb. According to research results, when the entire regression algorithms such as k-NN, bagged-SVM and NN were applied, they showed great accuracy (Moysen et al.2017). However, Moysen et al.(2017) believe that the approach of bagged -SVM learning is more appropriate for the requirements of this study. Bagged -SVM learning presents more precise forecasts. Polese et al.(2019) studied a data-driven association methods among the Next Generation Node Base(gNBs) and the RAN controllers. Moreover, they implemented different methods of machine learning adapted for forecasts such as the Bayesian Ridge Regressor (BRR) for the local-based prediction, and the Gaussian Process Regressor (GPR) and Random Forest Regressor (RFR) for both the local and the cluster-based forecasts. The clustering solution defined in this article has restricted the interplays between various controllers to decrease the requirement for synchronization of inter-controller and minimize the latency of control plane. In other research, the mobile traffic prediction challenge was handled the call detail records(CDRs)in the city of Milan. The assessed architectures of deep learning with multitask learning improve the accuracy of the forecasts by implementing the convolutional neural network (CNN) and the recurrent neural network (RNN) to capture features of spatial and temporal successfully(Huang et al., 2017). Furthermore, Oliveira et al. (cited in Huang et al.) investigated machine learning algorithms used artificial neural networks for Internet traffic prediction. They believe that RNN performed better and it is more suitable for time series network traffic prediction.

Chen et al(2018) pointed out that it is not easy to predict and characterize the patterns of traffic because the traffic pattern of each base station is active at diverse locations and times. Moreover, they suggested a structure of C-RAN optimization using deep learning to solve the difficulties. By utilizing the multivariate Long Short Term Memory (MuLSTM) model they aimed to understand the spatial correlation and the temporal dependence between traffic models of base stations and at the same time to forecast the traffic more accurately for the future time. In addition, Telecom Italia Dataset which is data of a real-world network was used in this study. It was noted in the article that the outcomes display that the model used significantly enhances capacity utilization and decreases the total cost of deployment of conventional RAN structure. Chuai et al(2019) studied on a model which integrates analysis of real-world telecommunication data and multivariate forecast system, fitting for communication networks in urban areas. Furthermore, they claim that causality analysis operated in practice first time while analysing telecommunication data which was taken by utilizing from the spatial-temporal model. Using the multivariate LSTM algorithm, future traffic data were also forecasted with the data obtained from the analysis. Based on the outcomes of the experiment, they emphasize that casualty analysis can increase the performance of forecasting of multivariate time series. Zhang et al (2019) investigated Spatial-Temporal Cross-domain neural network(STCNET) which is one of the new structure of deep learning to get the complicated and confidential patterns in mobile traffic data efficiently. They argue that STCNet, one of the components of convolutional long-short term memory, is good for overcoming problems associated with spatial-temporal dependence when forecasting traffic. Furthermore, three different domain datasets were used to capture the outer elements influencing the creation of traffic on this research. Since there are differences and similarities between mobile traffic data in different regions of cities, the inter-cluster transfer learning strategy was designed to reuse information by using the clustering model to separate parts of the city into diverse categories. In addition to the STCNet model, performances of Linear Regression (LR), Support Vector Machine (SVM), LSTM networks, and Spatial-Temporaly Densely Connected CNN (DenseNet) algorithms were calculated and compared to prove STCNet's superiority in predicting traffic in mobile communication networks. This study emphasizes that transfer learning with STCNet improves performance even further.

3.2 Background

In this project, traffic prediction using machine learning in communication networks was studied. Unstructured data used in this study comes from Radio Access Network (RAN) where is responsible for resource allocation in a communication network. Machine Learning methods have been applied to unstructured big data in RAN to ensure patterns and to gather information about the status of network and make better resource allocation. In this part, the main topics mentioned are wireless cellular technologies, radio access network, some recent communication networks, machine learning and the algorithms used for this project, data processing, accuracy and programming.

3.2.1 Wireless Cellular Technologies

The technologies of mobile wireless have begun to create, improve and make revolution since the beginning of the 1970s. Over the last few decades, 4-5 generations of technology revolution have taken place in technologies of mobile communication from 0G to 4G(Bhalla et al., 2010). In addition, it is expected to switch to 5G in early 2020. Some features of mobile communication generations mentioned are below:

1G: Cellular concept was mentioned in 1G that had analog cellular technology for the first time. The data bandwidth of 1G was 2 Kbps and it provided basic voice service. In addition, it used Advanced mobile phone service(AMPS) as a standard and FDMA as a multiplexing technique. After, 2G was developed with more superior features than 1G, which incorporates digital communication technologies, taking into account the negative characteristics of 1G technology such as low capacity, bad voice connectivity and untrustworthy handoffs(Bhalla et al., 2010).

2G: As mentioned above, when the second generation (2G) wireless mobile systems are compared with the first generation systems, it will be seen that contrary to 1G, the second-generation wireless systems used digital modulation methods such as time division multiple access (TDMA) and code division multiple access (CDMA). Moreover, Moreover, 2G provides SMS messaging in addition to the call service as primary services, and the most commonly utilized 2G standard is GSM. One of the weaknesses of 2G whose data bandwidth is 64 Kbps, is rates of limited data.

3G: The technology of 3G intended to ensure 144kbps - 384kbps data rates for large coverage regions and 2Mbps for local coverage regions. Internet-based, E-mail, video teleconferencing and services of multimedia occurring of mixed audio and data streams are some conceivable implementations of 3G. 3G wireless communication standards come from the IMT2000 family and were advanced by ITU-R. The basic standards of IMT-2000 are below:

- **UMTS / WCDMA**: The Universal Mobile Telecommunications System (UMTS), which utilizes Wideband Code Division Multiple Access (WCDMA), has been widely used in the rapidly expanding European Union system.
- **CDMA2000**: A US system, the first standard distributed utilizing CDMA techniques.
- **TDS-CDMA**: The system, which emerged in China and accepted many facts of GSM / UMTS, was optimized for Time Division Duplex (TDD).

4G: 4G technology is a fully IP based network system. Integration is a word that defines the attributes of 4G. In other words, it has the ability to integrate present and technologies of the future wireless network (e.g OFDM) to provide freedom of mobility and uninterrupted roaming(Li et al., 2009).

5G: It, the next-generation communication system, will allow many new applications to emerge and be realized. 5G technology is expected to have many advantages such as high speed, tremendous capacity, IoT ability and low latency. In addition, 5G networks have already begun to appear in some countries and are said to be spread around the world by 2020.

Main Factors Determining Demand for 5G

The main goal for 5G mobile networks is to provide distinct services according to customers' needs in an efficient way. In other words, the quality of service is the basic request.

Some of these requests are as follows:

-The increment in the number of devices requiring connection

- Need for higher data rates than currently offered data rates

- Near-zero latency

- Consumption of Energy

Key Attributes for 5G

- **Large Scale Antennas:** Base stations(BSs) utilizing large scale antennas ensure more spatial multiplexing, which allows supporting a lot of customers(multi-access) per time-frequency signalling resource. Large scale antennas used can provide an increase in the gain of transmitter antenna(downlink) or the gain of the receiver antenna(uplink) in the formula of the Friis transmission, where causes higher SNR.
- **Milimeter Wave Communications:** It is one of the favourite technologies of 5G, which is still under development. It enables the usage of signals having high-frequency in the unlicensed spectrum.
- **Flexibility:** The capability of a network to comply quickly to an altering need.

Furthermore, other features contain cell size and density of base station,networks of a device to device communication with multiple radio access technology.

3.2.2 Radio Access Network

A radio access network (RAN) which links single devices to other network components by establishing a radio connection is an important element of the communication system. RAN consists of a base station and antennas with a varying coverage area relying on their capacities. Moreover, a RAN makes radio resource management possible. When a machine is wirelessly linked to the core network, the RAN conveys machine's signal to diverse endpoints, and this signal moves along with the traffic of other networks.

RAN, which has been used since the early days of mobile communication technology, has developed among generations of cellular communications. For example; GRAN, GERAN, UTRAN and E-UTRAN(Ries et al., 2007).

- **Generic Radio Access Network (GRAN**): Its components are base transmission stations and controllers. It is in charge of radio connections of circuit-switched and packet-switched core networks.
- **GSM Edge Radio Access Network (GERAN):** It is available for Real-time packet data.
- **UMTS Terrestrial Radio Access Network (UTRAN):** It is used for both circuit-switched and packet-switched services
- **Evolved Universal Terrestrial Radio Access Network (E-UTRAN):** E-UTRAN is only used on packet-switched services. It ensures low latency and high data rates.

RANs of the next generations of cellular communication requires to be more complicated to be able to support technologies of software-defined networking (SDN) and network functions virtualization (NFV) and millimetre wave (mmWave). Therefore, radio access networks as well will benefit from virtualization and cloud technologies. Virtualization (vRAN) and Cloud (C-RAN) are examples of RANs in the future.

3.2.2.1 E-UTRAN in LTE Network

The evolved Universal Terrestrial Radio Access Network (E-UTRAN), with evolved base stations called eNodeB or eNB, manages radio links between mobile and evolved packet cores.

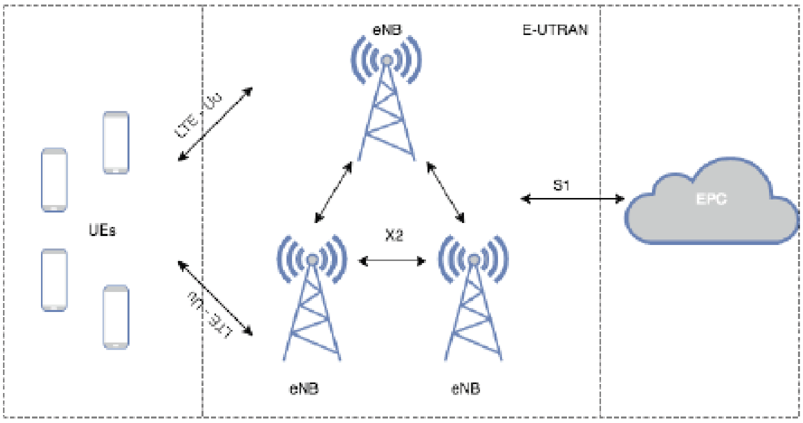


Figure 3.1 LTE Network Architecture(Jin,2016)

E-UTRAN whose basic elements are defined below has known as RAN of LTE network.

- **UE:** The user equipment (UE). For instance; a mobile phone, a computer or any remotely controlled device.

- **eNB:** An evolved B (NB) is a base station in LTE. The eNB transmits and receives radio signals to entire mobile devices utilizing the analogue and digital functions of signal processing of the LTE air interface. The eNB is in charge of the low-level operation of the entire mobile devices, transmitting to them signalling messages like handover commands.

- **X2 interface**: It is the interface among the eNBs ensuring links.

- **EPC**: the evolved packet core (EPC) which is the core network of the LTE system is IP-based network.

- **S1 interface**: It is the interface between the evolved Universal Terrestrial Radio Access Network (E-UTRAN) and the evolved packet core (EPC)

The network architecture of LTE is indicated in the figure(3.1). It can be seen that E-UTRAN occurs of eNBs which supplies a direct link to the UE. In the E-UTRAN, eNBs are linked via X2 interface which is mainly utilized for signalling and packet forwarding throughout handover. The link among E-UTRAN and EPC is achieved by the S1 interface.

Radio Resources Management in LTE

Radio Resources Management (RRM) is a basic application-level function of the eNB in LTE network, providing the effective usage of the present radio resources(Das, 2012). Some of them are defined below:

- Radio Admission Control(RAC): It accepts or refuses demands of installation for new radio bearers. The aim of RAC is to provide usage of high radio resource by admitting demands of radio bearer as long as radio resources are present. It concurrently provides suitable QoS for in-progress sessions, refusing radio bearer demands when they may not be ensured.

- Radio Carrier Control (RBC): RBC includes installation, maintenance and release of radio carriers for mobile users. Additionally, It is related to the maintenance of radio carriers of ongoing sessions in the event that the radio source changes. For instance, because of mobility. RBC takes a role in the release of radio resources related to radio carriers containing at session ending and handover.

- Dynamic Resource Allocation(DRA): It is in charge of allocating and de-allocating resources which involves resources of buffer and processing. It also allocates resource blocks in uplink and downlink. The other name of DRA is Packet Scheduling(PS).

- Inter cell Interference co-ordination(ICIC): It is in charge of for radio blocks to control intercell interference depending on feedback from multiple cells.

- Load Balancing(LB): It is in charge of dealing with the uneven traffic distribution load over multiple cells.

Cognitive Radio Network

Cognitive Radio Network is a type of network of wireless communication which intelligently perceives the convenient channels in the spectrum by using a transceiver that takes part in itself so that it can utilize the best channels in its surrounding. Cognitive Radio aims to provide sharing of radio resources in the best way and to reduce interference arising due to reusing the spectrum.

In addition, many kinds of research have been made by methods of machine learning and data mining in the area of CR. Implementations of machine learning are so popular for CR since it developed from software-defined radio(SDR).

3.2.2.2 C-RAN in 5G Networks

Cloud Radio Area Networks(C-RAN) which is presented by China Mobile to handle the arising issues due to extreme growth in the cellular network is a new structure based on the idea of virtualization. It is capable of dealing with such a number of base stations the network needs thanks to virtualization. Centralization and sharing have a positive impact on more dynamic traffic management and deployments of base stations, which means better use of resources. Moreover, such structure will have a possibility to reduce the cost of the expenses since base stations are virtualized rather than physically positioned in diverse regions. It xdecreases the use of energy and power when it is compared with conventional networks by the reason of the fact that base stations would be positioned on the identical physical machine(Salman, 2016).

A design of C-RAN occurs of three main components called the Baseband Unit (BBU) pool, the Remote Radio Unit (RRU) networks, and the perimeter network. Features of the main components of C-RAN are summarized below(Huang et al., 2014):

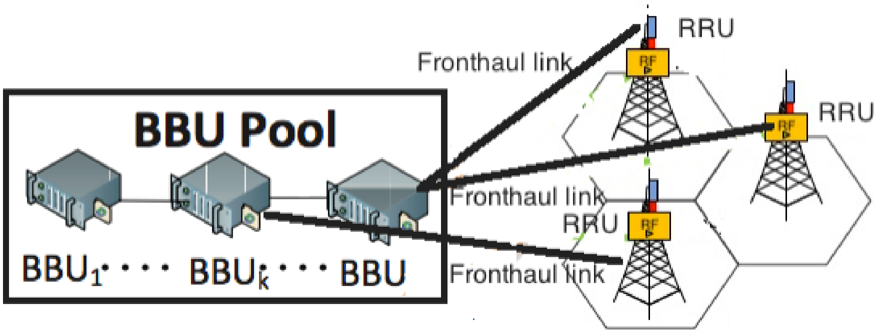


Figure 3.2 C-RAN Architecture(Salman,2016)

- **Baseband Unit (BBU) pool:** a BBU pool which is found at a centralized location such as a cloud or information centres occurs of multiple nodes of BBU that have excellent abilities of computation and storage. In addition, BBUs process resources and dynamically allocate them to Remote Radio Units (RRU) depending on the present network requires.
- **Remote Radio Unit (RRU)** networks: It is a wireless network which links wireless devices. RRU network works in a similar way to access points or towers in conventional mobile networks.
- **Fronthaul or transport network:** Fronthaul is the link-layer among a BBU and a series of RRUs that ensure high bandwidth connections to meet the needs of multiple RRUs. Moreover, Fronthauls could be implemented by utilizing various technologies which contain optical fiber communication, cellular communication or millimetre-wave communication. It ensures the highest bandwidth needs. Therefore, optical fiber communication is contemplated to be ideal in C-RAN. Nevertheless, this

implementation causes a high expense because it won't be flexible. On the other side, cellular communication or millimetre-wave communication is low-priced and easier to use than optical fiber communication. In addition, However, cellular or millimeter-wave communication results in less bandwidth and higher latency costs.

In addition, old virtualization methods that are used in IT clouds won't meet requirements of C-RAN. Therefore, more complicated technologies have been suggested finding a solution such as Network function virtualization(NFV) and Software defined network(SDN) to problems that will arise.

Network Function Virtualization

The purpose of Network Function Virtualization(NFV) is to handle the issue of virtualization by utilizing standard virtualization methods to deliver a variety of network hardware devices. These devices of network cover high-price routers and switches which are modified by software applied on high-volume servers, switches in the clouds, and elements of storage taking a place in the clouds. The advantages of this technology include reduced cost, the power and location of this equipment, and the ability to easily adapt devices. Another advantage is that operators can have resilient services depends on geographic position and privileges of the user. Furthermore, it will be more easily to be able to share resources with other operators on the identical server.

In addition, NFV technology, which displays significant advantages, is most likely to influence C-RAN implementation in practical application in the near future because it provides ease of use of the software as hardware elements supported by diverse suppliers.

Software-Defined Network

Software-defined network(SDN) at 5G or cellular network of next generation enables network management to be cost-effective by offering novel abilities and solutions for programmable, centrally controlled, flexible and high bandwidth implementations. Actually, the basic network infrastructure in SDN is completely isolated from implementations and the intelligence of networking is centrally controlled(Kitindi et al., 2017).In other words, This technology, which develops greatly with researches made in networking, allocates the control plan from the data and centralizes the control. As mentioned above, this technology can also be used with C-RAN(Salman,2016). The main advantages of SDN-enabled C-RAN structure are adaptable optimization and reconfiguration capabilities. Moreover, the state of RAN can modify with time and mobility of customers. The SDN controller of this structure is able to take the information of RAN in real-time, adapt the links between RRHs and BBUs. Moreover, It can rearrange them and redeliver resources among the baseband units(He et al., 2016).

3.3 Machine Learning

Arthur Samuel introduced that machine learning is a field giving computers the capability to learn with no being obviously programmed in 1959. Machine learning contains algorithms which forecast and learn based on data (Carbonell et al., 1983). These algorithms are used to create a model to perform data-driven forecasts or decisions instead of using exactly static program instructions.

In the last years, algorithms of machine learning to apply autonomous operations in networks has been investigated(Polese, 2019). Optimization of video flows (Zorzi,2017) energy efficient networks(Li, 2015) and resource allocation (Chinchali, 2018) can be given as examples for the studies made. The study (Jiang cited in Polese et al) surveyed the importance of machine learning for next-generation 5G networks. It gives information about the appropriate methods of machine learning.

Algorithms of machine learning are examined in three various groups. The features of these groups have explained below. The common aim among algorithms of supervised, unsupervised, and reinforcement learning is to predict the class conditional distribution by using the training data. The crucial distinction between methods of these different machine learning is the presence of information in the vector of feature showing which pattern class created a vector of the feature. Each vector of feature within the training data includes clear data concerning which class of model created that vector of features in supervised learning algorithms. When it comes to unsupervised learning, data respecting which pattern class created a special vector of feature is not obtainable within the training or the test data set. In addition, reinforcement learning ensures ‘clues’ regarding which class of pattern that is formed a special vector of the feature either periodically or aperiodically along with the operation of learning(Sutton and Barto, 1998)

Supervised Learning: Dependent variables and independent variables are introduced to the computer, and the aim of these algorithms is to acquire knowledge of a common function which maps inputs to desired outputs. When the algorithm reaches the wanted level of performance on the training data, the process stops. Regression, Decision Tree, Linear Regression, Support Vector Machine, Random Forest, KNN, Logistic Regression, Neural Network can be given as examples of supervised learning. In this study, supervised algorithms were used to make traffic prediction.

Unsupervised Learning: Algorithms of unsupervised learning doesn't utilize from target or outcome variable while making a prediction. The purpose is to explore hidden pattern in inputs by itself. Unsupervised algorithms usually work when the specialist doesn't understand what to seek in the dataset. For example; K-means.

Reinforcement Learning: Reinforcement Learning is a branch of AI, while it is a Machine Learning method. A computer attempts to capture the best feasible information to make the right decisions by using its experiences. This technique proposes to utilize observations collected from the interplay with the environment in order to increase the reward to the maximum or decrease the risk to the minimum. For instance; Markov Decision Process.

The fundamentals of the algorithms used during this project are explained simply as follows.

Linear Regression

Linear regression(LR) which is one of the algorithms of supervised learning mentions a set of methods for investigating the linear intercourse among two variables. A model of linear regression that contains a single independent variable is called simple linear regression in general.

In the equation (3.1), the regression factors of intercept β0 and slope β1 were predicted by using linear regression.

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (3.1)$$

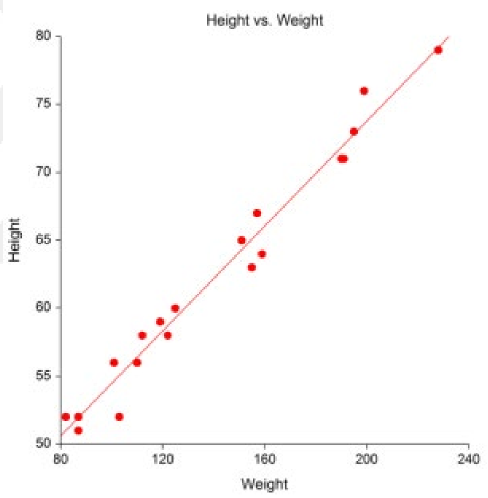


Figure 3.3 Linear Regression

It is presumed that errors have unknown variance and mean zero by calculating confidence intervals and hypothesis tests. Furthermore, it is accepted that errors do not correlate. In other words, the values of the errors do not affect each other.

It is expedient for x to be seen in the manner controlled by the expert and measured by a negligible error, whereas y is random. The probability distribution for y is at each value that is suitable for x.This distribution mean is

$$E(y \mid x) = \beta_0 + \beta_1 x \qquad (3.2)$$

$$Var(y \mid x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \qquad (3.3)$$

Despite the variance of y does not rely on the variable of x, it can be said that the mean of y is a linear function. In addition, errors do not correlate, and it is valid for the responses as well. β0 and β1 named regression coefficients have a straightforward and frequently beneficial interpretation. β1 is an alter in a mean of a distribution of y when altering unit in x. if data range on x does not contain 0, β0 does not convenient interpretation. If range of x contains 0, β0 is a mean of a distribution of y when x = 0(Montgomery et al., 2006).

Support Vector Machine

Support Vector Machines(SVM) is supervised learning algorithms for classification and regression issues. They can handle linear and non-linear issues and work well for many implementations. While SVM is utilized for classification problems, the dots of data are set apart by using a maximum margin decision boundary. Simple SVM depends on the linear classifier, which contains datum with linear functions which are as well named hyperplane.

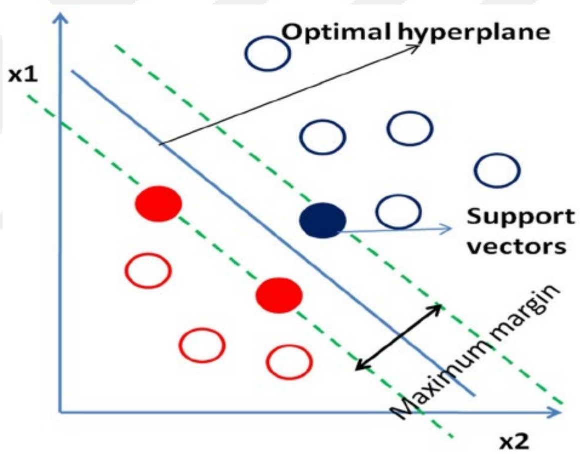


Figure 3.4 Support Vector Machine(Rout et all, 2014)

In Figure 3.4, The straight line is a hyperplane which splits the blue points and red points. As can be seen, support vectors are data dots which are closest to the hyperplane. A margin is a space among the two vectors on the nearest class dots, which is computed as the vertical distance from the hyperplane to these vectors or nearest dots. If the margin is major, it is called as a good margin.

As mentioned above, SVM can as well be utilized as a technique of regression, by keeping the entire important features characterized the method(maximum margin). The main characteristics of SVM for classification and support vector regression (SVR) are almost only the same in other words there are only some small segregations between them. It is so hard to forecast the data, which has limitless possibilities since the output is an actual number.

Tolerance (epsilon) margin in regression is decided to proximity to the SVM, which could have already asked from the difficulty. The main difference is that SVR is used when working with continuous values and making a prediction.

Additionally, SVM can utilize from math functions which have been known as the kernel. These functions, which have many different types, are used to convert input data from the current form to another form. Linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid can be given as examples for the kernel functions. The kernel functions are responsible for turning the inner product among two points in an appropriate space of feature. By utilizing kernel functions, an algorithm is also applied in a high dimensional space. Another important attribute of kernels is modularity. In other words, steps of using kernel functions are to select an algorithm which utilizes only inner products among inputs and to use this algorithm with a kernel function that computes inner products among input images in a feature space.

Decision Tree

Decision Tree Analysis(DT) is an algorithm of general and predictive learning, which is used for implementations in diverse fields, and also built over an algorithmic approximation which defines methods to divide a dataset based on varied situations. Moreover, Decision Tree, commonly used and beneficial method, a non-parametric supervised learning technique utilized for classification and regression problems. Decision tree regression is utilized if the dependent variable is continuous, and classification trees are utilized if it is categorical. The aim of DT is to form a model that forecasts the value of targets, learning basic decision rules deduced of the attributes of data.

In addition, it is a flowchart-like architecture which every inner node indicates a test on a feature . Each leaf of DT represents one class. The structure of a decision tree depends on the information entropy.

$$\boldsymbol{Entropy = -p * \log(p) - (1-p) * \log(1-p)} \qquad \textbf{(3.4)}$$

In equation (3.5), p shows the percent of positive representatives on the present node of DT. The value of entropy is wanted to be minor because if entropy is bigger, the sample space of the node will be more complicated. The node is split to obtain a smaller entropy. For instance; The node at hand is separated into left and right nodes. The gain provided by this division is named information gain(Quinlan,1986).

$$\boldsymbol{InformationGain = Entropy(parent) - \{Prob(left) * Entropy(left) + Prob(right) * Entropy(right)\}} \qquad \textbf{(3.5)}$$

In the equations, while Prob(left) indicates the percent of samples separated into a left node within the former sample space, Prob(right) represents the percent of samples split into a right node within the former sample space.

Structure of Decision trees:

The Decision Tree begins with the root node to indicate the entire train dataset(Alharan,2017).

1.First of all, when the training lists give identical results, the node is leafed and also will be labelled with the class.

2. If not, the tree chooses the best information feature to split into the set. It labels the node with the title of the feature.

3. Repeat the steps and it is stopped when the entire samples possess the identical class or there are not any other samples or there are no new features.

4.Decision Tree concludes.

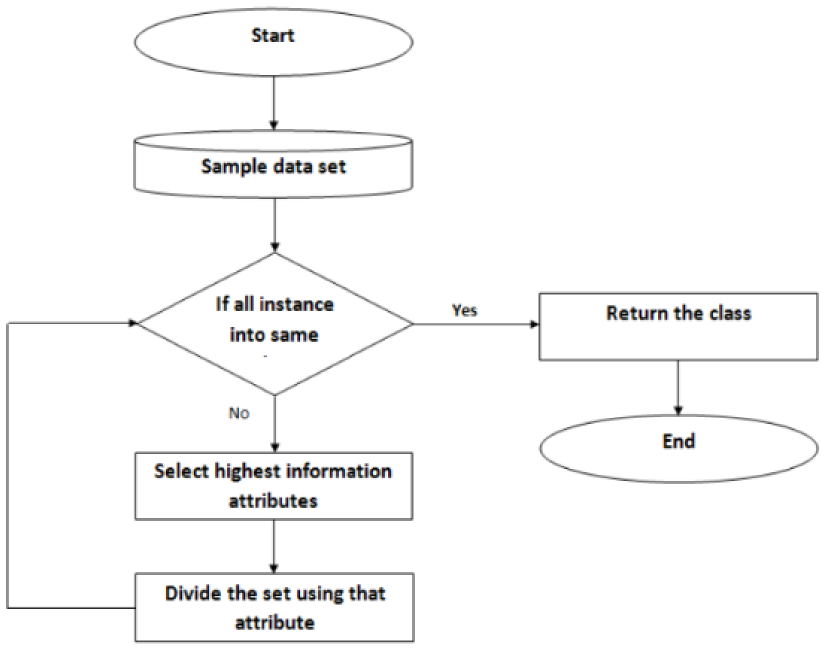


Figure 3.5 Decision trees construction

Feedforward Neural Network

The links among the units of Feedforward neural networks, a kind of artificial neural network, do not create a cycle. Feedforward neural networks are also the initial sort of artificial neural networks, but they are more straightforward than their equivalent, recurrent neural networks. They are named as feedforward since data only moves forwards on the network. The data follows the path of the input nodes, then the hidden nodes (if any) and finally the output nodes respectively. In other words, every neuron in one layer binds to every neuron in the subsequent layer and there is no link between perceptrons on the same layer.

Multi-layered Network(MLN) is formed of sigmoid neurons which are competent in dealing with the nonlinearly separable information. The layers named as hidden layers which are utilized to overcome the complicated nonlinearly detachable relations among input and the output take part between the input and output layers.

Feedforward neural networks which are a type of supervised learning are utilized if the information that will be learned is not sequential or time-dependent. Feedforward neural network calculates a function of f on constant-magnitude input of x, and f(x) is about equal to y for training pairs of (x, y).

$$f(x) \approx y \qquad (3.7)$$

Moreover, if it will be mentioned recurrent neural networks, they learn sequential information by calculating g at the input of X_k={x_1,..., x_k}. g(X_k) is about equal to y_k for training pairs (X_n, Y_n) for 1≤k≤n.

$$g(X_k) \approx y_k \qquad (3.8)$$

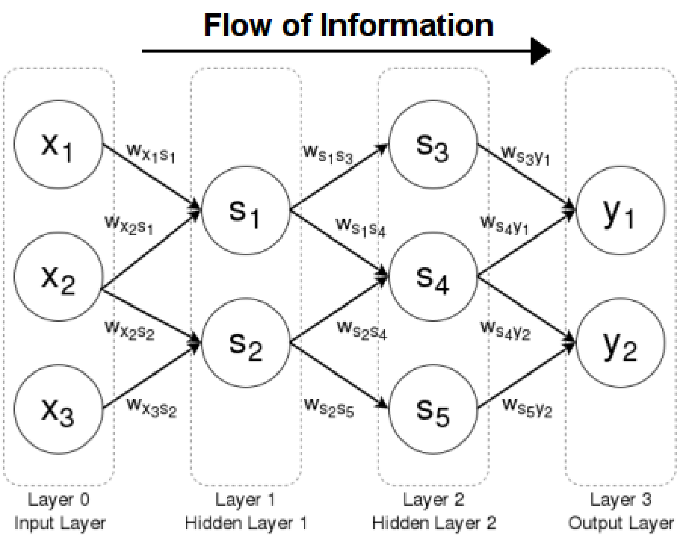


Figure3.6 A feedforward neural network with information flowing left to right

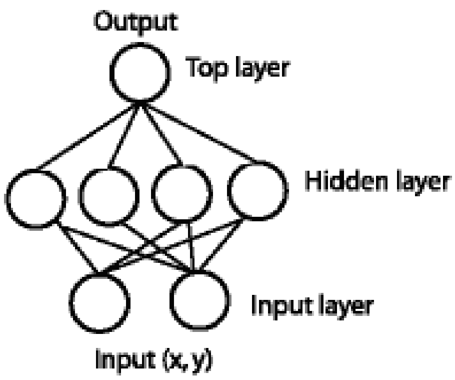


Figure 3.7 A feed-forward network with one hidden layer(Gupta,2013)

(x, y) is supplied to the network via neurons in the input layer. There are 4 independent neurons inside a hidden layer. The point is divided into 4 sets of linear detachable areas. Every set has an original line which divides the area(Gupta,2013).

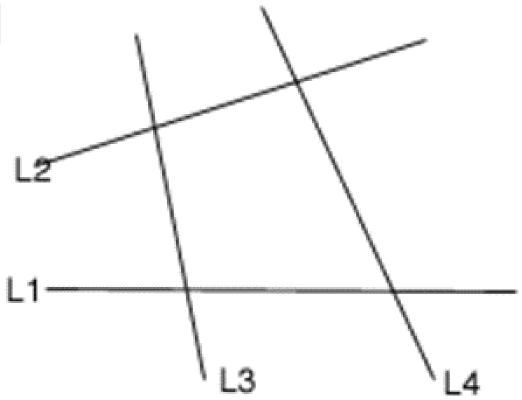


Figure 3.8 Lines each dividing the plane into 2 linearly separable regions(Gupta,2013)

The top perceptron in the hidden layer makes logical operations such as AND on perception's outputs in order to categorize points of input in 2 areas which cannot be separated linearly. When the AND operator is used in these four outputs, it comes to the intersection of the 4 areas which make up a central area.

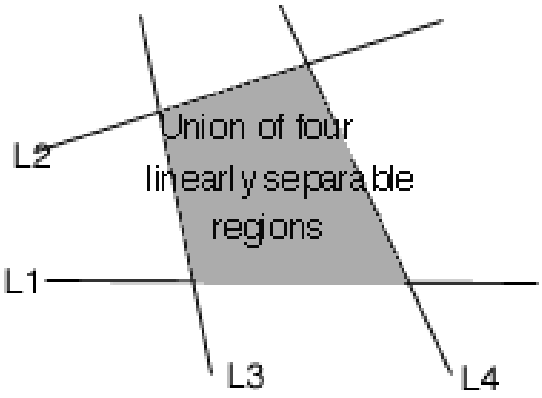


Figure 3.9 Intersection of 4 linearly separable regions forms the center region(Gupta,2013)

Changing the number of nodes in the hidden layer, the number of layers and the number of input and output nodes take a role in classification. It is possible to classify points of arbitrary size in accordance with an arbitrary number of groups. Therefore, feed-forward networks are more widely utilized for classification problems(Gupta,2013).

3.4 Data Preprocessing

3,4,1 Feature Engineering

The aim of Feature engineering, which is the operation of converting raw data into features by utilizing domain information of data, is to generate features to help ensure the best performance of ML algorithms. The performance of simple machine learning methods will also improve when the features are good. Feature engineering is divided into 3 sections as feature extraction, feature construction and feature selection.

- **Feature Selection:** Feature Selection is an operation of choosing the attributes that will be beneficial the most to the forecasting variable or output.
- **Feature construction:** It is the process of creating novel features from the present set of attributes. These new features can be utilized for forecasting.
- **Feature extraction:** It is the dimensionality reduction process in which the first raw information set is reduced to more controllable sets to handle.

3.4.2 Feature Vector

Vectors of features are utilized to indicate numeric or symbolic features of an item mathematically and based on analyze in machine learning. The equivalent of variables vectors which are utilized in a statistical operation like linear regression is also feature vectors. Due to the effectiveness and practicality of exemplifying items numerically to assist with types of analysis, vectors of features are widely to utilize for machine learning as well. The Euclidean distance which is one of the easy methods is used to compare the feature vectors related to two different items. Feature vectors can be exploited for k-nearest neighbours, artificial neural networks, and classification problems(Pang et al).

3.5 Missing Values

Each real-life datasets nearly is formed of incomplete values. It is important to take exclusive consideration for missing values to analyse the data.

Some methods of missing data are like below.

1. Discarding the tuples: If a missing value happens on any of the parts in the data, remove all group of observation.

2. A method of filling the incomplete value manually: It is time-wasting and not conceivable for tremendous data set with a lot of missing values.

3. Utilizing a constant: Modifying entire missing feature values by an appropriate value or label. For instance; "unknown" or " - ∞ ".

4. Using the attribute mean value: Replacing missing values by making estimations. For example; filling with a measurement of central tendency, mean, midrange, median and mode.

One of the solutions for deficient data issues is the usage of linear interpolation which is an imputation technique used to predict incomplete values. Furthermore, Linear Interpolation was the technique used in this study.

Linear Interpolation: Linear Interpolation, one of the basic variety of interpolation, purposes to construct a straight line among two discrete data points. Utilizing the notions of algebra, the formula of the function of linear interpolation is:

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \quad (3.9)$$

Where x, x_0 and x_1 are the independent factors, x_0 and x_1 are also known values. In addition, $f(x)$ is the dependent factor to the value of x.

Moreover, $\frac{f(x_1) - f(x_0)}{x_1 - x_0}$ is an approach of finite-split-difference of the first derivative, indicating the slope of the line constructing the values. The small interval among the data values generally shows the better the estimation. For this reason, while the distance reduces, a continuous function can be better approximated by a solid line(Noor et al., 2008).

3.6 Accuracy

In this study, two different metrics were acknowledged for evaluating exhaustively of distinct forecasting methods.

One of them is Root Mean Square Error (RMSE) utilized to measure the distinction among values forecasted by an algorithm and the values of actual observation.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} |A_t - F_t|^2}{n}} \quad (3.10)$$

The other one was Mean Absolute Error (MAE) which finds the absolute difference among the predicted values and actual observations and then calculates the average.

$$MAE = \sqrt{\frac{\sum_{t=1}^{n} |A_t - F_t|}{n}} \quad (3.11)$$

Small values of RMSE and MAE are an indication of the better the performance(Zhang et al., 2019).

3.7 Programming

In this Project MATLAB which was the software programme used to implement machine learning algorithms. Statistics and Machine Learning Toolbox™ in Matlab includes functions and applications to define, evaluate, and pattern data. Regression and classification algorithms allow taking deductions by using data and create predictive patterns. Regression algorithms portray the connection among an answer(output) variable and one or more predictor (input) variables. (MATLAB)

3.8 Dataset

In this study, it is used the dataset supported by "Open Big Data" project (Dandelion), containing varied datums gathered from November 2013 to December 2013 by some service provider of distinct fields such as energy suppliers, Internet Service Providers. Additionally, it is also publicly available with Database Open License (ODbL v1.0). This dataset includes the traffic information of call/sms/internet for every squared area of the Milan city, which ensures the precise location of the base station(BS) of the four apex cells of grids of the city as well.

According to the WGS84 standard (World Geodetic System 1984):

- Cell 1= [9.011490619692509, 45.356685994655464]

- Cell 100 = [9.311521155996243, 45.356261753717845]
- Cell 9901 = [9.011533669936474, 45.56821407553667]
- Cell 10000 = [9.312688264185276, 45.56778671132765]

which permits of localizing of various areas as can be seen in the figures (3.10) and (3.11) using the cell size information.

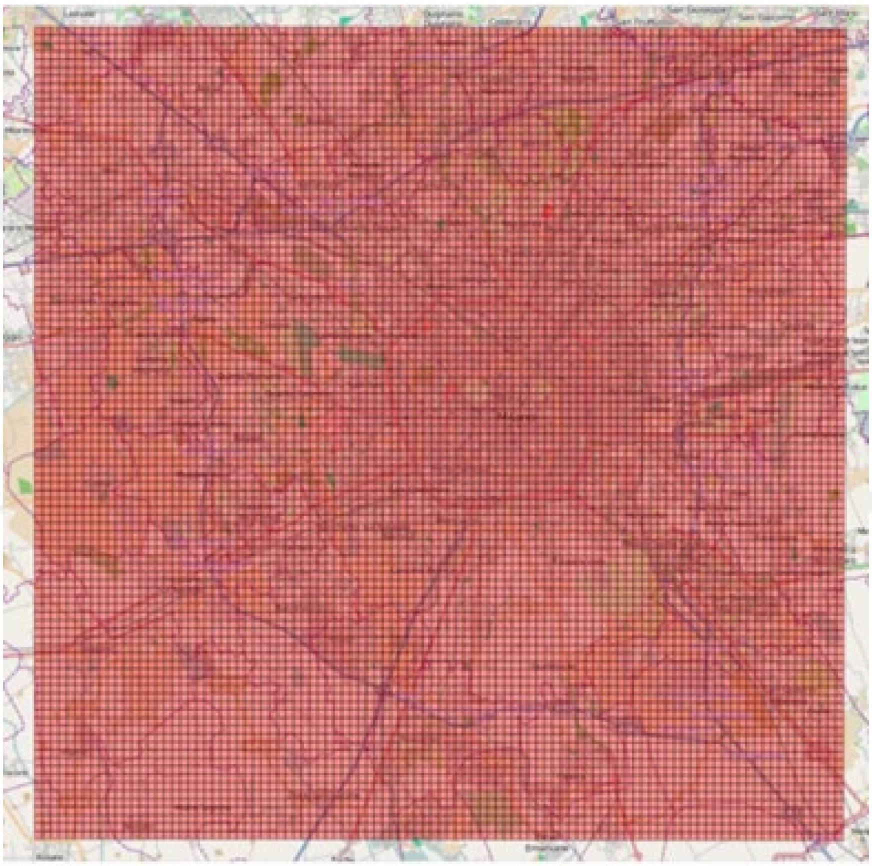


Figure 3.10 Milan Map(Barlachi et al., 2015)

Milan city consists of the grid layer more than 1,000 (areas with a dimension of about 235×235 meters).

9901	9902	...	9999	10000
9801	...	...	9899	9900
...	...	...	...	...
101	102	...	...	200
1	2	3	...	100

Figure 3.11 Milan Grid(Barlachi et al., 2015)

Although there are so many different sorts of CDRs, It has considered that telecom Italia used the activities below when creating this dataset:

- Received SMS: a CDR created every time when a customer gets an SMS

- Sent SMS: a CDR created every time when a customer sends out an SMS

- Incoming Calls: a CDR created every time when a customer receives a phone call

- Outgoing Calls: a CDR created every time when a customer a phone call

- Internet: a CDR created each time

•when a customer starts a connection

•when a customer ends a connection

•when one of the upper limits below is attained during the same connection:

o 15 minutes from the latest created CDR

o 5 MB from the latest created CDR

By combining these records, a data set containing information about SMS /call/internet activity was created. This data set also indicates the level of interaction of the customer with the cellular network. For example, when a higher number of sms sent, sms sent activity increases.

Logs of Call Detail Records (CDRs) which created by the mobile network of Telecom Italia of Milan indicate the action of customers and are used to bill customers, and handle the network issues. The first column of the table below refers to the cell ids of all 10000 cells, while the second column indicates the timestamp which can be converted to Unix Epoch Format. The other one is a code of a country which the connection becomes. Others are to show indicate traffic activities of the network. In addition, the CDRs are stored every ten minutes, creating 144 records for each country code in a day.

Square id	Time interval	Country code	SMS-in activity	SMS-out activity	Call-in activity	Call-out activity	Internet activity
1	13835196	39	0.02613742	0.03087508	0.026137424	0.0552251	9.2601138
1	13835208	39	0.02792473	0.02792473	0.001787310	0.0546009	8.6690075
…							
40	13835196	39	0.04044804	0.01862799	0.001596025	0.0186279	4.5503902
40	13835196	49					0.0015960
…							
10000	13835196	39	0.02613742	0.03087508	0.026137424	0.0552251	9.2601138

Table 3.1 The architecture of the dataset of Milan

The data schema(Barlacchi, 2015):

• **Square ID:** The cell identification of the square taking part in the Milan GRID, which is a type of numeric.

• **Time interval:** The starting of the time interval is described as milliseconds passed from the Unix Epoch on January 1st, 1970 at UTC. If the end of the time interval is to be obtained, 600000 milliseconds (10 minutes) must be added to this value. It is the type of numeric.

• **Country code:** It represents the phone country code of a nation, which is a type of numeric.

• **SMS-in activity:** It shows a received sms activity belonging to a given square id over a given time interval and known from where it was sent by country code. Its type is numeric.

• **SMS-out activity:** It indicates a transmitted sms activity belonging to a given square id over a given time interval and known from where it was sent by country code. Its type is numeric.

• **Call-in activity:** Activity proportional to the number of received calls belonging to a given square id over a given time interval. It has been known from where it was issued by country code. Its type is numeric.

• **Call-out activity:** Activity proportional to the number of issued calls belonging to a given square id over a given time interval. It has been known from where it was received by country code. Its type is numeric.

• **Internet traffic activity:** The number of Internet traffic activities that belong to a particular square id over a given time period. Internet connection starts from the country indicated by the country code.

Some notes: Despite the fact that the data studied only contains information for two months data (November 2013, December 2013, and the first day of January 2014), the files are pretty much big. The files are in tsv format, and also includes 62 files which are a total text size of 5 GB(Dandelion).

Call Detail Record: When a customer achieves a telecommunications interplay, a Radio Base Station (RBS) is allocated by the service provider and makes tranmission over the network. After that, a novel CDR is generated, which logs the interplay time and the RBS that managed it. Moreover, it is feasible to get an indicator of the customer's geographic position by making use of the coverage maps Cmap in relation to each RBS in the area it operates(AKA coverage area).

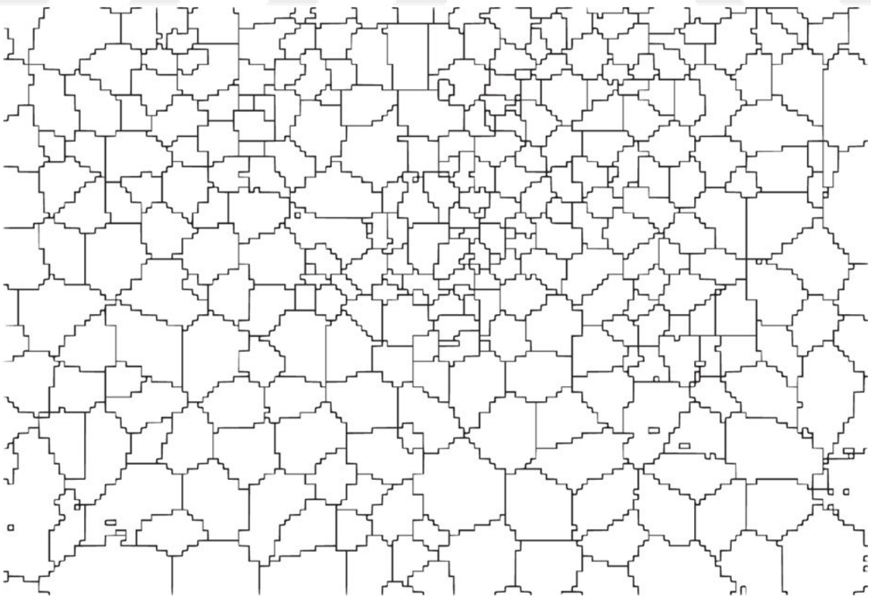


Figure 3.12 An example of coverage map of Milan(Barlachi et al., 2015)

Every interaction is related to the coverage region v of the RBS which handled it with the aim of spatially clustering the CDRs in the grids. Therefore, the amount of records si(t) in area i at time t is calculated as below:

$$S_i(t) = \sum_{\mathrm{v \in Cmap}} R_v(t) \frac{A_{v \cap i}}{A_v} \qquad (3.12)$$

where Rv(t) represents the number of records within the coverage region v at time t.While Av is the surface of the coverage area v, Av∩i indicates the spatial intersection surface between the coverage area v and the square i.

Chapter 4

Data Preprocessing and Experimental Settings

The mobile network is one of the most prevalent technologies worldwide at the present time. The network is separated to micro or macro regions named as cells and, a base station is located at fix place in each cell. In communication system, the signal coming from a mobile device is gathered by a base station that is near to a place where mobile user locates, and thus a connection is established. Most of the adjacent ones overlap, even if little, with each other to ensure a sustained link at the network while mobile devices move. Many neighbour cells are grouped in regions defined by a local area code (LAC) called Cell ID. Service suppliers make a record of information called call detail records of mobile equipment in use. Service suppliers make a record of information called call detail records of mobile equipment in use. As has mentioned at dataset part of the background, CDRs usually contains time-stamp, cell number, time type and internet traffic activity etc(Zhang et al., 2019).

As already mentioned in this study, the cellular traffic data set studied comes from Telecom Italy, a major operator in Europe as part of the Big Data Challenge. This record, which was held for 8 weeks in Milan, contains temporal and spatial data. Traffic in each area changes in real-time and the trend of change is different. Therefore, the system can be considered as dynamic.

4.1 Data preprocessing

In this study, internet traffic activity data of Milan city were analyzed.

Data processing section is as follows.

• Internet traffic activity data of 100 areas from different parts of Milan were extracted according to spatial information (according to their Square ID).

• After the Internet Traffic Activity data was extracted, the data in the first seven weeks were used to generate the training dataset, and the data in the last week were utilised to create the test dataset to test the accuracy of the different forecasting algorithms.

• Some cells' traffic data values may be missing in a specific time interval due to a data storage error or improper transmission. Missing data must be completed before proceeding to the forecasting process (Zhang et al., 2019). Since this dataset also contains missing values, the missing values in the data of each area were completed using the linear interpolation method, which provides good performance for datasets containing low levels of missing data (Noor et al, 2008).

• The previous values of internet traffic activity are used to estimate future values in each area. In other words, N-previous value was used as a feature vector.

• Regression algorithms, which is a type of supervised learning, are widely used in forecasting. Therefore, the data set was adapted to supervised learning using v_enframe, one of the voice box algorithms developed for Matlab. V_enframe frames the vector pretending to be (overlapping) the signal. Thus, predictor values and target values to be used to feed algorithm models were obtained. This process was repeated for each area.

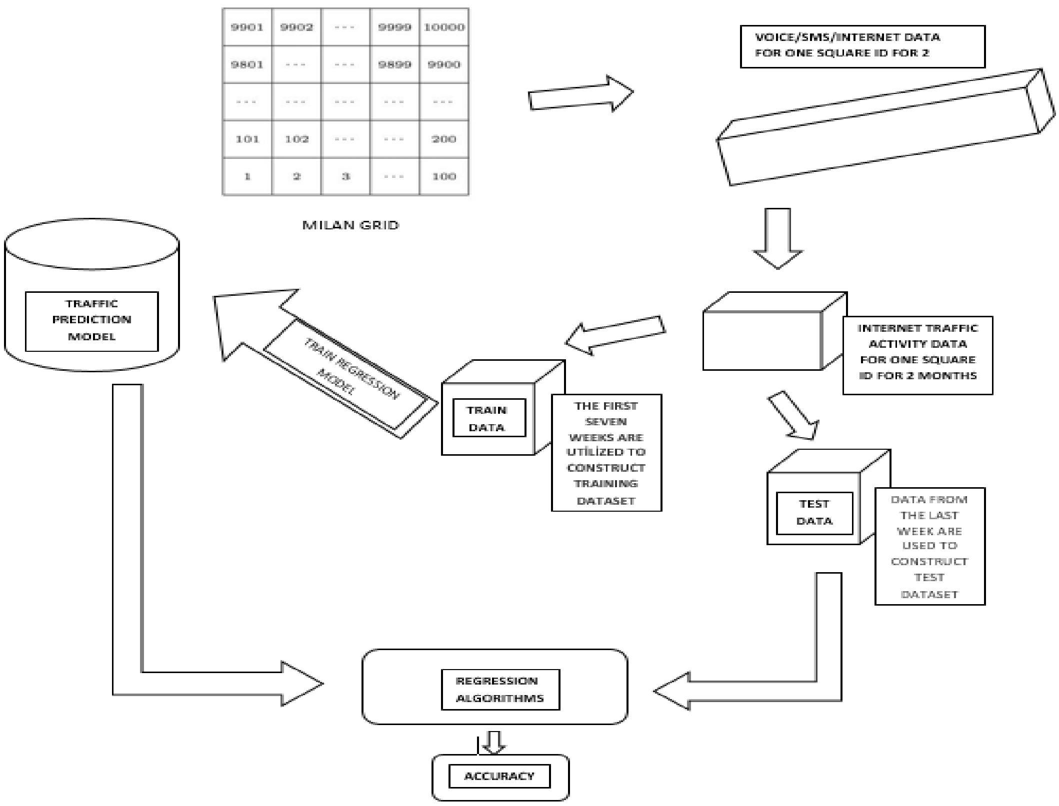


Figure 4.1 A Flowchart for data processing and implementations of algorithms

4.2 Prediction

Naturally, mobile trafic forecastic can be considered as a time series issue. In accordance with ways of solving, previous researches can be splitted into two parts where are named as statistical-based machine learning-based. The intense increase in cellular traffic and improvements in data-driven machine learning forecasting and ai methods have made them a powerful opponent toward classical statistical models. It is a remarkable research topic in the field of wireless communication.

4.2.1 Implementation of Algorithms and Results

In this thesis, supervised learning algorithm, linear regression, support vector machine, decision tree and neural network are used. The most common metric systems, mae and rmse, were used to measure accuracy. Both the MAE and RMSE can take a value from 0 to ∞. The following algorithms and metric systems are used separately for each field.

Linear Regression

Linear regression(LR) is one of classical methods which are commonly utilized for time series forecasting. In previous research made, It has seemed that LR was extensively used for traffic prediction (Zhang et al., 2019). This algorithm was selected to use depends on this information.

Linear model was generated by using predictor values and target values. Moreover, prediction was made to calculate response values of the linear regression model by using test data. After calculating errors as well, accuracy was measured.

Support Vector Machine

The main reason for using Support Machine Vector(SVMs) in this dissertation is the capability of this methodology to exactly predict time series data if nonlinear, non-stationary and not defined a-priori are commonly used as the processes of basic system. Sapankevych et al.(2009) has claimed that SVMs have also been shown better performance than different non-linear methods contaning neural-network based non-linear forecasting methods such as multi-layer perceptrons.

When Support Machine Vector was created, it was utilized input arguments which are kernel function, gaussian and standardize true, in addition to predictor values and test values. Kernel function was utilized to calculate the Gram matrix and model of the SVM regression which takes advantage of the Gaussian kernel was used in this study. Gaussian shows better performance than the one utilizing the linear kernel. In addition, every column of information of predictor is centered and weighted using the average of weighted column and standard deviation(MATLAB).After modelling, responses were forecasted for model of Gaussian kernel regression and, errors were computed. Finally, The values of MAE and RMSE that were calculated.

Algorithms/Accuracy	RMSE	MAE
Linear Regression	160.78	55.18
Support Vector Machine	105.65	47.16
Desion Tree	158.39	61.87
Neural Network(Feedforward)	108.67	53.61

Table 4.1 Accuracy Table Using RMSE and MAE

Decision Tree

The reasons to utilize from the Decision Tree in this experiment are that it is easy to use, flexible and versatile. A regression tree are obtained by depending on the input variables (predictor data)and the output (response data). The obtained tree is a binary tree where every branching node is divided depends on the values of predictor data column.

Decion Tree regression model was formed by using predictor values and target values. After, prediction was made to calculate response values of the Decision Tree model by using test data. After calculating errors as well, accuracy was measured. Accuracy results were calculated.

Neural Network (Feedforward)

Feedfoward neural network is able to handle complicated(non-linear) functions(Chavoya et al., 2013). Based on this information and literature research, it was decided that feedforward neural network would be the last algorithm that will be used in this study. As the model of network was generated, input argumentts used are as below.

- Number of Neuron: 50
- Hidden Layer: 1
- TrainFcn: trainscg

After modelling, prediction was made and, errors and accuracy values were computed by using MAE and RMSE.

While forming of feedforward neural network with matlab code used (fitnet), the network gives a function suitable neural network as a result with a hidden layer size of hiddenSizes stated as a row vector and function of training, identified by trainFcn(Matlab). Trainscg's purpose is to update values of bias and weight accordance with the scaled conjugate gradient method(MATLAB).

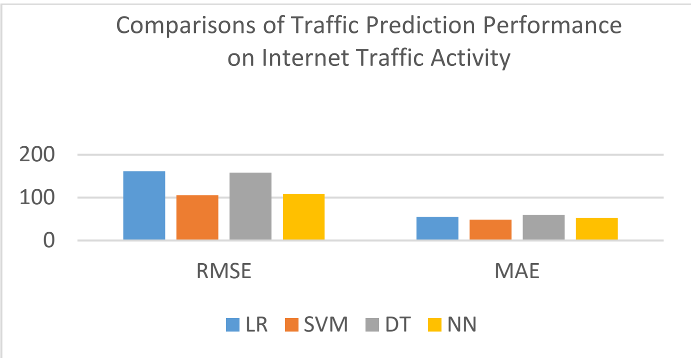
In addition,training continues until at least one of the following occurs.

- The highest number of epochs (repetitions)
- Passing maximum amount of time.
- Performance is reduced to the aim.
- The gradient performance become lower than min_grad.

4.3 Analysis of Results

After the MAE and RMSE values of 100 square ids are calculated separately for each algorithm, the mean values of the MAE and RMSE of each algorithm are calculated. Recent accurcy values are above in Table .4.1.

While Linear regression shows the worst performance, Support Vector Machine has the best one as can be seen in the Graph (4.1).below. There is a big difference between the values obtained of linear regression and support vector machine. The reason to get better the for Support Vector Machine regression is that it can handle the nonlinearity of mobile traffic. Linear Regression has not enough parameter space to deal with the complicated dynamics of mobile traffic. The improved parameter capacity is likely to assist to better catch the spatial and temporal dependencies of mobile traffic created by geographically allocated base stations. Although the accuracy results of Feedforward Neural Network(FNN) are worse than SVM regression, the difference between them is so small, which shows that the values are close to each other. FNN is to give the second-best result since it can deal with nonlinearity. The performance values of DT are not as good as LR, or even very close to each other.



Graph 4.1Traffic Prediction Performance on Internet Traffic Activity

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In a communication network that is a complicated system, network optimization is a significant solution to allocate resources fairly. A precise forecast is one of the foundations of network optimization. Due to the increase in mobile traffic, radio resources in Radio Access Network should be used more efficiently. This would be possible when using more intelligent forecasting methods of resource allocation. In this project, the main goal was to present a data-driven predictive approach using machine learning in communication networks. Firstly, It was investigated the present implementations of machine learning in communication networks. After the literature review, the process that will be followed was specified. Next step is to make data ready for algorithms. Therefore, the methods of data processing were used.100 areas were extracted from different regions of Milan city according to cell ids. As the first seven weeks of internet activity traffic of the data extracted was used for training, the last week was for testing. The missing values were completed by linear interpolation. The data was framed and while historical values in the frame were used as predictor values, another vector in the frame was called as target values. The reason for this process was to make data suitable for supervised learning. The processes mentioned above were made for each square id extracted. After completing the data process, models of the algorithms decided to use were generated. Predictions for each area extracted was made for coming internet activity values and MAE and RMSE measured to able to evaluate performances. The mean of all MAE values and the mean of all RMSE values coming from 100 square ids were calculated.

It can be seen that while the results of Support Vector Machine(SVM) and Feedforward Neural Network(FNN) are close, Linear Regression(LR)is the worst one. SVM is better than all four algorithms used. The reason for that it can deal with nonlinearity that is an outcome of the spatiotemporal dataset

5.2 Future Work

The mobile traffic can be affected by external factors. Therefore, this would be extended by using external factors such as the number of base stations and social activity. When examining the previous research made related to this topic, it was mentioned the power of RNN algorithm. It has thought that RNN can be suitable for this Project to obtain better performance.

References

1) Montgomery, D. C.; Peck, E. A. & Vining, G. G. (2006), *Introduction to Linear Regression Analysis (4th ed.)* , Wiley & Sons . Available from: https://www.academia.edu/32079757/Introduction_to_Linear_Regression_Analysis_5th_ed._Douglas_C._Montgomery_Elizabeth_A._Peck_and_G._.pdf [Accessed 4th September 2019]

2) Kitindi, E. J., Hu,S., Jia, Y., Kabir,A., Wang,Y.(2017). "Wireless Network Virtualization With SDN and C-RAN for 5G Networks: Requirements, Opportunities, and Challenges". *IEEE Access* vol. 5, pp. 19099-19115. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8025644&tag=1 [Accessed 4th September 2019]

3) Gupta, K. D., Joshi,J.,Gupta,P.,Gupta,S.(2013).' Linear Neural Network Structural Design For Discrete Regions'. *International Journal of Engineering Technology and Computer Research (IJETCR)* Available from: https://www.researchgate.net/publication/331275577_LINEAR_NEURAL_NETWORK_STRUCTURAL_DESIGN_FOR_DISCRETE_REGIONS [Accessed 4th September 2019]

4) Chuai et al (2019).' A new method for traffic forecasting in urban wireless communication network'*. EURASIP Journal on Wireless Communications and Networking*. Available from: https://www.researchgate.net/publication/331794945_A_new_method_for_traffic_forecasting_in_urban_wireless_communication_network [Accessed 4th September 2019]

5) Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.C. and Hanzo, L. (2017). Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, 24(2), pp.98-105. Available from: http://www.eng.usf.edu/chen/pdf/2%20IEEE%20WC%202017-4.pdf [Accessed 4th September 2019]

6) Zhang, C.,Patras, P., Haddadi,H. (2018). "Deep Learning in Mobile and Wireless Networking". *A Survey. IEEE Communications Surveys & Tutorials.* Available from: https://arxiv.org/pdf/1803.04311.pdf [Accessed 4th September 2019]

7) Jin J. Traffic Burst Prediction in Radio Access Network with Machine Learning [Internet] [Dissertation]. 2016. (TRITA-EE). Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-197206

8) Chen, L., Yang, D., Zhang, D., Wang, C., Li, J.,Thi, M.,Trang, N. (2018). "Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization". *Journal of Network and Computer Applications,* vol.121, pp. 59-69. Available from: https://www.sciencedirect.com/science/article/pii/S1084804518302455 [Accessed 4th September 2019]

9) Zhang, C. , Zhang, H., Qiao, J. , Yuan, D. and Zhang, M.(2009) "Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389-1401. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8667446&isnumber=8715825 [Accessed 4th September 2019]

10) Li, X., Gani, A. , Salleh, R. and Zakaria, O. "The Future of Mobile Wireless Communication Networks," (2009) *International Conference on Communication Software and Networks*, Macau, pp. 554-557. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5076913&isnumber=5076790[Accessed 4th September 2019]

11) Bahalla, M.,R. And Bahalla ,A.,V.(2010)" Generations of Mobile Wireless Technology:A Survey". *International Journal of Computer Applications (0975 – 8887)*,vol.5, no.4. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.5216&rep=rep1&type=pdf [Accessed 4th September 2019]

12) Ries, M., Rupp, M. (2007). "Performance Evaluation of Mobile Video Delivery Technologies". Available from: https://www.researchgate.net/publication/237475752_Performance_Evaluation_of_Mobile_Video_Delivery_Technologies [Accessed 4th September 2019]

13) Das,M.K.(2012)" Radio Resource Management - Connection Mobility Control eNodeB Source Handover Strategies"*Tata Consultancy Services.* Available from:

https://pdfs.semanticscholar.org/e307/4fa1bca6858a1a0893be8beeb50f03f6b5a9.pdf [Accessed 4th September 2019]

14) Salman,T.(2016)" A Survey of C-RAN Basics, Virtualization, Resource Allocation, and Challenges" Available from: https://www.cse.wustl.edu/~jain/cse574-16/ftp/cloudran/index.html [Accessed 4th September 2019]

15) Huang ,C. I J. , Duan ,R. , Cui ,C. , Jiang, J. and Li, L.(2014) "Recent Progress on C-RAN Centralization and Cloudification," in *IEEE Access*, vol. 2, pp. 1030-1039. Available from: https://ieeexplore.ieee.org/document/6882182 [Accessed 4th September 2019]

16) He ,M., Alba, A. ,M., Basta A., Blenk, A. and Kellerer ,W.(2019) "Flexibility in Softwarized Networks: Classifications and Research Challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2600-2636. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8611160&isnumber=8809933 [Accessed 4th September 2019]

17) Quinlan, J. R. (1986). "Induction of Decision Trees". 1(1), pp.81-106. Available from: https://dl.acm.org/citation.cfm?id=637969 [Accessed 4th September 2019]

18) Noor et al." Using The Linear Interpolation Technıqueto Estimate Missing Valuesfor Air Pollution Data". Available from: https://www.academia.edu/208141/Using_the_Linear_Interpolation_Technique_to_Estimate_Missing_values_for_Air_Pollution_Data [Accessed 4th September 2019]

19) Dandelion: Available from: https://dandelion.eu/datagems/SpazioDati/telecom-sms-call-internet-mi/description/ [Accessed 4th September 2019]

20) Sapankevych, I. N. & Sankar, R. (2009). Time Series Prediction Using Support Vector Machines: A Survey. *Computational Intelligence Magazine, IEEE*. 4. 24 - 38. Available from: https://www.researchgate.net/publication/224408260_Time_Series_Prediction_Using_Support_Vector_Machines_A_Survey [Accessed 4th September 2019]

21) López-Martín ,C., Chavoya, A. and Meda-Campaña ,M. E.(2013) "Use of a Feedforward Neural Network for Predicting the Development Duration of Software

Projects." *12th International Conference on Machine Learning and Applications*, Miami, FL, pp. 156-159. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6786099&isnumber=6786067 [Accessed 4th September 2019]

22) Klaine P., Imran M., Onireti O., and Souza D. (2017).A survey of machine learning techniques applied to self organizing cellular networks. *IEEE Communications Surveys and Tutorials.* Available from: https://ieeexplore.ieee.org/document/7982603 [Accessed 4th September 2019]

23) Chen X., Wu J., Cai Y., Zhang Y., and Chen T.(2015).Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(4):627–640. Available from: https://ieeexplore.ieee.org/abstract/document/7012044 [Accessed 4th September 2019]

24) Zhang et al. (2018). Deep Learning in Mobile and Wireless Networking. *A Survey. IEEE Communications Surveys & Tutorials.* Available from: https://arxiv.org/pdf/1803.04311.pdf [Accessed 4th September 2019]

25) Panwar et al. (2016). A Survey on 5G: The Next Generation of Mobile Communication. Physical Communication. Available from: https://www.sciencedirect.com/science/article/pii/S1874490715000531 [Accessed 4th September 2019]

26) Baştuğ, E., Bennis, M. and Debbah, M.(2014). Living on the edge: The role of proactive caching in 5G wireless networks. Available from: https://ieeexplore.ieee.org/document/6871674 [Accessed 4th September 2019]

27) Pérez, J.S., Jayaweera, S.K. and Lane, S.(2017). Machine learning aided cognitive RAT selection for 5G heterogeneous networks. *IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Istanbul, Turkey*.

Available from: https://ieeexplore.ieee.org/document/8277675 [Accessed 4th September 2019]

28) Alawe I., Ksentini A., Hadjadj-Aoul Y., Bertin P.(2018). Improving traffic forecasting for 5G core network scalability: A Machine Learning approach. *IEEE Network Magazine,* pp.1-10. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6786099&isnumber=6786067 [Accessed 4th September 2019]

29) Stankovic, J.A., (2014). Research directions for the internet of things. *IEEE Internet of Things Journal*, *1*(1), pp.3-9. Available from: https://ieeexplore.ieee.org/document/6774858 [Accessed 4th September 2019]

30) Zorzi M., Zanella A., Testolin A., Grazia M. D. F. D.(2015).Cognition-based networks: A new perspective on network optimization using learning and distributed intelligence. *IEEE Access*, vol. 3, pp. 1512–1530. Available from: https://www.researchgate.net/publication/281768893_Cognition-Based_Networks_A_New_Perspective_on_Network_Optimization_Using_Learning_and_Distributed_Intelligence [Accessed 4th September 2019]

31) Li R., Zhao Z., Zhao X., Ding G., Chen Y., Wang Z., and Zhang H.(2017). Intelligent 5G: When Cellular Networks Meet Artificial Intelligence*. IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183. Available from: http://www.rongpeng.info/files/Paper_wcm2016.pdf [Accessed 4th September 2019]

32) Chinchali S., Hu P., Chu T., Sharma M., Bansal M., Misra R., Pavone M., and Sachin K.(2018).Cellular network traffic scheduling with deep reinforcement learning.*National Conference on Artificial Intelligence (AAAI)*. Available from: http://asl.stanford.edu/wp-content/papercite-data/pdf/Chinchali.ea.AAAI18.pdf [Accessed 4th September 2019]

33) Polese, M., Jana, R., Kounev, V., Zhang, K., Deb, S., & Zorzi, M. (2018). Machine Learning at the Edge: A Data-Driven Architecture with Applications to 5G Cellular Networks. *CoRR*. Available from: https://arxiv.org/abs/1808.07647 [Accessed 4th September 2019]

34) Calabrese F. D., Wang L., Ghadimi E., Peters G., Hanzo L. and Soldati P.(2018).Learning Radio Resource Management in RANs: Framework, Opportunities, and Challenges. *IEEE Communications Magazine*, vol. 56, no. 9, pp. 138-145. Available from: https://ieeexplore.ieee.org/document/8466370 [Accessed 4th September 2019]

35) Le L., Sinh D., Tung L. and Lin B. P.(2018).A practical model for traffic forecasting based on big data, machine-learning, and network KPIs.*15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1-4. Available from: https://ieeexplore.ieee.org/document/8319255 [Accessed 4th September 2019]

36) Huang C., Chiang C. and Li Q.(2017)A study of deep learning networks on mobile traffic forecasting.*IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1-6. Available from: https://ieeexplore.ieee.org/document/8292737 [Accessed 4th September 2019]

37) Telecom Italia, *Harvard Dataverse*

38) Carbonell, Jaime & S. Michalski, Ryszard & M. Mitchell, Tom & Carbonell, Jaime. (2002). Machine Learning: A Historical And Methodological Analysis. *Artificial Intelligence Magazine. 4*. Available from: https://pdfs.semanticscholar.org/3522/d171d0af99fb1e0a06f8d31734987967870a.pdf [Accessed 4th September 2019]

39) Chin, W. H., Fan, Z., & Haines, R. J. (2014). Emerging Technologies and Research Challenges for 5G Wireless Networks. *IEEE Wireless Communications*, 21(2), 106-112. Available from:

https://www.researchgate.net/publication/260393079_Emerging_Technologies_and_Research_Challenges_for_5G_Wireless_Networks [Accessed 4th September 2019]

40) Moysen, Jessica & Giupponi, L & Mangues-Bafalluy, Josep. (2017). A Mobile Network Planning Tool Based on Data Analytics. *Mobile Information Systems* ,pp 1-16. Available from: https://www.researchgate.net/publication/313372790_A_Mobile_Network_Planning_Tool_Based_on_Data_Analytics [Accessed 4th September 2019]

41) Barlacchi et al.(2015). A Multi-Source of Urban Life in the City of Milan and Province of trentino. *Sci Data.* Available from: https://www.nature.com/articles/sdata201555 [Accessed 4th September 2019]

42) Alharan,A.,Alsgheer,R.,Al-habobi,A.(2017)"Popular Decision Tree Algorithms ofData Mining Techniques:A Review"*International Journal of Computer Science and Mobile Computing,vol.6, pp.133-142.* Available from: https://www.researchgate.net/publication/317731072_Popular_Decision_Tree_Algorithms_of_Data_Mining_Techniques_A_Review [Accessed 4th September 2019]

Some useful links:

http://www.artizanetworks.com/resources/tutorials/cran.html

https://in.mathworks.com/help/stats/index.html

https://www.sdxcentral.com/5g/definitions/radio-access-network/

Appendices

MODELS

```
clear all


load('DataTrn\InternetTrn')
Trn=fillmiss(InternetTrn);

N=3;

ftrain = v_enframe(Trn, N+1, 1);%VoiceBox Toolbox enfame.m
X = ftrain(:,1:N);
Xresponse = ftrain(:,N+1);

load('DataTrn\InternetTst')

Tst=fillmiss(InternetTst);

ftest = v_enframe(Tst, N+1, 1);%VoiceBox Toolbox enfame.m
Y = ftest(:,1:N);
Yresponse = ftest(:,N+1);

clear   callin_a callout_a smsin_a smsout_a sqid_a timestamp_a
countrycode_a

%%%%%%% LR %%%%%%%%%

% X1=X(1:500,:);
% X1response=Xresponse(1:500,:);

LM=fitrlinear(X,Xresponse);
ypred = predict(LM,Y);%returns the predicted response values of  LM
to the points in Z

[Yresponse, ypred];
errorL = Yresponse- ypred;

RMSELinear = sqrt(mean(errorL .* errorL, 'omitnan')) %Root Mean
Squared Error
MAELinear = mean(abs(errorL),'omitnan') %Mean Absolute Error


 %%%%%%%% SVM%%%%%%%%%

Mdl =
fitrsvm(X,Xresponse,'KernelFunction','gaussian','KernelScale','auto',
...
    'Standardize',true);
YFit = predict(Mdl,Y);

[Yresponse, YFit];
errorSVM = Yresponse - YFit;

RMSESVM = sqrt(mean((errorSVM).^2)) %Root Mean Squared Error
```

```
MAESVM = mean(abs(errorSVM)) %Mean Absolute Error



%%%%%%%%Decision Tree%%%%%%%%%%%%

tree = fitrtree(X,Xresponse);
Zfit = predict(tree,Y);

[Yresponse, Zfit];

errorDT =Yresponse - Zfit;

RMSEDT = sqrt(mean((errorDT).^2)) %Root Mean Squared Error
MAEDT = mean(abs(errorDT)) %Mean Absolute Error


%%%%%%%%Neural Network(feedforwardnet)%%%%%%%%%%%%%%

TransposeX=X';
TransposeY=Xresponse';
TransposeZ=Y';
TransposeH=Yresponse';

net = feedforwardnet(250);

net.trainFcn='trainscg';
net = train(net,TransposeX,TransposeY);

%view(net)
y = net(TransposeZ);

perf = perform(net,y,TransposeZ);

%scatter(TransposeH,y)

% hold on
e=gsubtract(TransposeH,y);
% e   = TransposeH-y;

RMSENN = sqrt(mean((e).^2)) %Root Mean Squared Error
MAENN = mean(abs(e)) %Mean Absolute Error
```

Linear Interpolation

```
%  Missing values are calculated as weighted sum of linear
%  interpolations from nearest available points.
%  Altogether 5 estimates from column-wise and 5 for row-wise
%  1-d linear interpolation are calculated.
%  Weights are such that for the best case (isolated missing
%  points away from the boundary) the interpolation is equvalent
%  to average of 4-point Lagrangian polynomial interpolations
%  from nearest points in a row and a column.

%  Kirill K. Pankratov,  kirill@plume.mit.edu
%  12/19/94

 % Handle input ..........................................
 if nargin==0, help fillmiss, return, end

 % If no missing values, output=input, quick exit.
 % If all values are missing, give up  (output=input)
 % and prevent endless recursion ....................
if all(all(isnan(M)))|~any(any(isnan(M)))
   Mf = M;
return
end

 % Interpolation coefficients ......................
 % [left right left-left left-right right-right]
coef =  [eps eps 1/2 4/3 1/2];

 % Sizes ..................
n_miss = length(find(isnan(M)));
szM = size(M);
szMf2 = szM(2)+4;
is_vec = szM==1;

 % Auxillary ...............
v4 = -1:2;
o4 = ones(1,4);
o5 = ones(5,1);
o2 = ones(2,1);
omiss = ones(n_miss,1);
wsum = zeros(n_miss,2);
a = wsum;

 % Interpolate from both rows and columns, if possible
for jj = find(~is_vec)

  bM = zeros(2,szM(3-jj));    % Make margins
  Mf = M; if jj==2, Mf = M'; end
  Mf = [bM; isnan(Mf); bM];

  Mf = Mf(:);
  miss = find(Mf==1);        % Missing ##
  exis = find(~Mf);          % Available ##
  Mf = cumsum(~Mf);
  i_m = Mf(miss);

  Mf = M; if jj==2, Mf = M'; end
```

```
  Mf = [bM*nan; Mf; bM*nan];

  % Indices ................
  I = i_m(:,o4)+v4(omiss,:);
  I = reshape(exis(I),n_miss,4);     % Quartets of neib. pts for
                                     % each missing pts.

  % Make 5 estimates ...................................
  W = miss(:,o4)-I;
  A = zeros(n_miss,5);
  A(:,1:2) = reshape(Mf(I(:,2:3)),n_miss,2);
  A(:,3:5) = (W(:,1:3)-W(:,2:4));
  A(:,3:5) = reshape(Mf(I(:,2:4)),n_miss,3).*W(:,1:3);
  A(:,3:5) = A(:,3:5)-reshape(Mf(I(:,1:3)),n_miss,3).*W(:,2:4);
  A(:,3:5) = A(:,3:5)./(W(:,1:3)-W(:,2:4));

  % Calculate weights ......
  W = [abs(W(:,2:3)) abs(W(:,1:3))+abs(W(:,2:4))];
  W = (~isnan(A))./W;

  W = W.*coef(omiss,:);
  wsum(:,jj) = sum(W')';
  wsum(:,jj) = wsum(:,jj)-(wsum(:,jj)==0);
  W = W./wsum(:,jj*ones(1,5));
  i_m = find(isnan(A));
  A(i_m) = zeros(size(i_m));
  A = A.*W;
  a(:,jj) = A*o5;
end

 % Correspondence between row and column numbering .....
exis = ceil(miss/szMf2);
exis = exis+(miss-(exis-1)*szMf2)*szMf2;
[exis,i_m] = sort(exis);
wsum(:,1) = wsum(i_m,1);
a(i_m,1) = a(:,1);

 % Combine estimates from rows and columns .............
wsum = wsum+(wsum==-1);
exis = wsum*o2;
i_m = exis==0;
exis(i_m) = exis(i_m)+nan;
wsum = wsum./exis(:,o2);
exis = (a.*wsum)*o2;

 % Insert interpolated values into Mf ..................
Mf(miss) = exis;

  % Remove NaNs at the margins
Mf = Mf(3:szM(jj)+2,:);
if jj==2, Mf = Mf'; end

 % If there are still missing pts, repeat the procedure
if any(any(isnan(Mf))), Mf = fillmiss(Mf); end
```

V_enframe/VOICEBOX

function [f,t,w]=v_enframe(x,win,hop,m,fs)

%V_ENFRAME split signal up into (overlapping) frames: one per row. [F,T]=(X,WIN,HOP)

%

% Usage: (1) f=v_enframe(x,n) % split into frames of length n

% (2) f=v_enframe(x,hamming(n,'periodic'),n/4) % use a 75% overlapped Hamming window of length n

% (3) calculate spectrogram in units of power per Hz

%

% W=hamming(NW); % analysis window (NW = fft length)

% P=v_enframe(S,W,HOP,'sdp',FS); % computer first half of PSD (HOP = frame increment in samples)

%

% (3) frequency domain frame-based processing:

%

% S=...; % input signal

% OV=2; % overlap factor of 2 (4 is also often used)

% NW=160; % DFT window length

% W=sqrt(hamming(NW,'periodic')); % omit sqrt if OV=4

% [F,T,WS]=v_enframe(S,W,1/OV,'fa'); % do STFT: one row per time frame, +ve frequencies only

% ... process frames ...

% X=v_overlapadd(v_irfft(F,NW,2),WS,HOP); % reconstitute the time waveform with scaled window (omit "X=" to plot waveform)

%

% Inputs: x input signal

% win window or window length in samples

% hop frame increment or hop in samples or fraction of window [window length]

% m mode input:

% 'z' zero pad to fill up final frame

% 'r' reflect last few samples for final frame

% 'A' calculate the t output as the centre of mass

```
%             'E'  calculate the t output as the centre of energy
%             'f'  perform a 1-sided dft on each frame (like v_rfft)
%             'F'  perform a 2-sided dft on each frame using fft
%             'p'  calculate the 1-sided power/energy spectrum of each frame
%             'P'  calculate the 2-sided power/energy spectrum of each frame
%             'a'  scale window to give unity gain with overlap-add
%             's'  scale window so that power is preserved:
sum(mean(v_enframe(x,win,hop,'sp'),1))=mean(x.^2)
%             'S'  scale window so that total energy is preserved:
sum(sum(v_enframe(x,win,hop,'Sp')))=sum(x.^2)
%             'd'  make options 's' and 'S' give power/energy per Hz:
sum(mean(v_enframe(x,win,hop,'sp'),1))*fs/length(win)=mean(x.^2)
%         fs   sample frequency (only needed for 'd' option) [1]
%
% Outputs:  f   enframed data - one frame per row
%          t    fractional time in samples at the centre of each frame
%              with the first sample being 1.
%          w    window function used
%
% By default, the number of frames will be rounded down to the nearest
% integer and the last few samples of x() will be ignored unless its length
% is lw more than a multiple of hop. If the 'z' or 'r' options are given,
% the number of frame will instead be rounded up and no samples will be ignored.
%


% Bugs/Suggestions:
%  (1) Possible additional mode options:
%       'u'  modify window for first and last few frames to ensure WOLA
%       'a'  normalize window to give a mean of unity after overlaps
%       'e'  normalize window to give an energy of unity after overlaps
%       'wm' use Hamming window
%       'wn' use Hanning window
```

```
%      'x'  hoplude all frames that hoplude any of the x samples


%           Copyright (C) Mike Brookes 1997-2014
%    Version: $Id: v_enframe.m 10865 2018-09-21 17:22:45Z dmb $
%
%  VOICEBOX is a MATLAB toolbox for speech processing.
%  Home page: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  This program is free software; you can redistribute it and/or modify
%  it under the terms of the GNU General Public License as published by
%  the Free Software Foundation; either version 2 of the License, or
%  (at your option) any later version.
%
%  This program is distributed in the hope that it will be useful,
%  but WITHOUT ANY WARRANTY; without even the implied warranty of
%  MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
%  GNU General Public License for more details.
%
%  You can obtain a copy of the GNU General Public License from
%  http://www.gnu.org/copyleft/gpl.html or by writing to
%  Free Software Foundation, Inc.,675 Mass Ave, Cambridge, MA 02139, USA.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


nx=length(x(:));
if nargin<2 || isempty(win)
   win=nx;
end
if nargin<4 || isempty(m)
   m='';
```

```
end
nwin=length(win);
if nwin == 1
   lw = win;
   w = ones(1,lw);
else
   lw = nwin;
   w = win(:).';
end
if (nargin < 3) || isempty(hop)
   hop = lw; % if no hop given, make non-overlapping
elseif hop<1
   hop=lw*hop;
end
if any(m=='a')
   w=w*sqrt(hop/sum(w.^2)); % scale to give unity gain for overlap-add
elseif any(m=='s')
   w=w/sqrt(w*w'*lw);
elseif any(m=='S')
   w=w/sqrt(w*w'*lw/hop);
end
if any(m=='d') % scale to give power/energy densities
   if nargin<5 || isempty(fs)
      w=w*sqrt(lw);
   else
      w=w*sqrt(lw/fs);
   end
end
nli=nx-lw+hop;
nf = max(fix(nli/hop),0);   % number of full frames
na=nli-hop*nf+(nf==0)*(lw-hop);     % number of samples left over
```

```
fx=nargin>3 && (any(m=='z') || any(m=='r')) && na>0; % need an extra row
f=zeros(nf+fx,lw);
indf= hop*(0:(nf-1)).';
inds = (1:lw);
if fx
  f(1:nf,:) = x(indf(:,ones(1,lw))+inds(ones(nf,1),:));
  if any(m=='r')
    ix=1+mod(nf*hop:nf*hop+lw-1,2*nx);
    f(nf+1,:)=x(ix+(ix>nx).*(2*nx+1-2*ix));
  else
    f(nf+1,1:nx-nf*hop)=x(1+nf*hop:nx);
  end
  nf=size(f,1);
else
  f(:) = x(indf(:,ones(1,lw))+inds(ones(nf,1),:));
end
if (nwin > 1)  % if we have a non-unity window
  f = f .* w(ones(nf,1),:);
end
if any(lower(m)=='p') % 'pP' = calculate the power spectrum
  f=fft(f,[],2);
  f=real(f.*conj(f));
  if any(m=='p')
    imx=fix((lw+1)/2); % highest replicated frequency
    f(:,2:imx)=f(:,2:imx)+f(:,lw:-1:lw-imx+2);
    f=f(:,1:fix(lw/2)+1);
  end
elseif any(lower(m)=='f') % 'fF' = take the DFT
  f=fft(f,[],2);
  if any(m=='f')
    f=f(:,1:fix(lw/2)+1);
```

```
   end
end
if nargout>1
   if any(m=='E')
      t0=sum((1:lw).*w.^2)/sum(w.^2);
   elseif any(m=='A')
      t0=sum((1:lw).*w)/sum(w);
   else
      t0=(1+lw)/2;
   end
   t=t0+hop*(0:(nf-1)).';
end
```