Universidad del País Vasco Euskal Herriko Unibertsitatea

FACULTY OF ENGINEERING BILBAO
UNIVERSITY OF THE BASQUE COUNTRY

MASTER IN TELECOMMUNICATION ENGINEERING

MASTER'S DEGREE FINAL PROJECT

QOX ENHANCEMENT IN CAMPUS WI-FI NETWORKS: A MACHINE LEARNING APPROACH

Student: Casado-O'Mara Corral, Itziar

Supervisor: Ibarrola Armendariz, Ana Eva

Academic year: 2022-2023

Date: Bilbao, 18, September, 2023

BLANK PAGE

RESUMEN

El desarrollo de nuevas tecnologías, junto con las crecientes exigencias de los usuarios de éstas, ha puesto en el punto de mira la gestión de la calidad de servicio (QoS) de las redes de telecomunicaciones. Por este motivo, el grupo de investigación Networking Quality and Security (NQaS) de la Universidad del País Vasco (UPV/EHU) ha desarrollado un modelo para su gestión global conocido como QoXfera.

Este Trabajo Fin de Máster (TFM) tiene como objetivo la validación de la metodología desarrollada para la implementación del modelo QoXfera en entornos reales. El trabajo se centra en la primera parte de esta metodología, que se basa en el uso de herramientas de aprendizaje automático para la identificación de diferentes perfiles de usuarios. Para ello, se ha realizado un estudio de los posibles algoritmos a utilizar y se han recogido datos reales del servicio Wi-Fi de la UPV/EHU para aplicar dichos algoritmos e identificar los perfiles de usuarios. Además, para profundizar en el proceso de validación, se han contrastado los resultados obtenidos con aquellos de un estudio similar realizado en otro campus universitario.

Palabras clave: Calidad de servicio (QoS), Aprendizaje automático, Wi-Fi

LABURPENA

Teknologia berrien garapenak, haien erabiltzaileen eskakizun gorakorrekin batera, telekomnikazio sareen zerbitzuaren kalitatearen (QoSaren) kudeaketa jomugan jartzea ekarri du. Hori dela eta, Euskal Herriko Unibertsitateko (UPV/EHU) Networking Quality and Security (NQaS) ikerketa taldeak QoSaren kudeaketarako eredu globala garatu du, QoXfera izenekoa.

Master Amaierako Lan (MAL) honek QoXfera eredua ingurune errealetan ezartzeko garatutako metodologia balioztatzea du helburu. Lan hau metodologia horren lehenengo atalean oinarritzen da, non ikaskuntza automatikoa erabiliz erabiltzaile profil desberdinak identifikatzen saiatuko den. Horretarako, erabili daitekeen algoritmoak aztertu dira eta UPV/EHUko Wi-Fi zerbitzuaren benetako datuak bildu dira algoritmo horiek aplikatzeko eta erabiltzaileen profilak identifikatzeko. Gainera, balioztatze prozesuan sakontzeko, aurkitutako erantzunak beste unibertsitate campus batean aurkitutakoekin alderatu egin dira.

Hitz gakoak: Zerbitzuaren kalitatea (QoS), Ikaskuntza automatikoa, Wi-Fi

ABSTRACT

The development of new technologies, together with the growing demands of their users, has brought the management of the quality of service (QoS) of telecommunications networks to the forefront. For this reason, the Networking Quality and Security (NQaS) research group of the University of the Basque Country (UPV/EHU) has developed a model for the QoS' global management known as QoXphere.

This Master's Degree Final Project aims to validate the methodology developed for the implementation of the QoXfera model in real environments. The work focuses on the first part of this methodology, which is based on the use of automatic learning tools for the identification of different user profiles. To this end, a study of the possible algorithms to be used has been carried out and real data has been collected from the UPV/EHU Wi-Fi service in order to apply these algorithms and identify user profiles. Furthermore, in order to deepen the validation process, the results obtained were compared with those of a similar study carried out on another university campus.

Key words: Quality of Service (QoS), Machine Learning, Wi-Fi

List of acronyms and abbreviations

ACK: Acknowledgement

AC: Access Category

ADV: Advertisement

AI: Artificial Intelligence

AIFS: Arbitration Interframe Space

ANN: Artificial Neural Network

AP: Access Point

ARPU: Average Revenue Per User

BSS: Basic Server Set

CHURN: Attrition rate

CLARA: Clustering Large Application

CLARANS: Clustering Large Applications based on Randomized Search

CLICK: Cluster Identification via Connectivity Kernels

CLIQUE: Clustering in Quest

CoS: Class of Service

CSMA/CA: Carrier Sense Multiple Access with Collision Avoidance

CP: Contention Period

CFP: Contention Free Period

CW: Contention Window

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DCF: Distributed Coordination Function

DE: Differential Evolution

DENCLUE: Density-based Clustering

DHCP: Dynamic Host Configuration Protocol

DIFS: DCF Interframe Space

EAP: Extensible Authentication Protocol

EDCA: Enhanced Distributed Channel Access

EM: Expectation-Maximization

ENCLUS: Entropy-Based Subspace Clustering

EXP: Expectation

GA: Genetic Algorithm

GoS: Grade of Service

HCCA: HCF-Controlled Channel Access

HCF: Hybrid Coordination Function

IBSS: Independent Basic Service Set

IEEE: Institute of Electrical and Electronics Engineers

IID: Institute Intelligence and Data

IoT: Internet of Things

IT: Information Technology

ITU: International Telecommunication Union

ITU-T: ITU Telecommunication Standardization Sector

KIS: Department of Communications Engineering

KPI: Key Performance Indicator

MAC: Medium Access Control

MAL: Master Amaierako Lana

ML: Machine Learning

NaN: Not a Number

NP: Network Performance

NQaS: Networking, Quality and Security

OPTICS: Ordering Points to Identify the Clustering Structure

PC: Point Coordinator

PCF: Point Coordination Function

PIFS: PCF Interframe Space

PROCLUS: Projected Clustering

PSO: Particle Swarm Optimization

QAP: QoS-enhanced access point

QBSS: QoS supporting Basic Service Set

QoBiz: Quality of Service of Business

QoE: Quality of Experience

QoP: Quality of Service Perceived

QoR: Quality of Resilience

QoS: Quality of Service

QoSD: Quality of Service delivered/achieved by service provider

QoSP: Quality of Service perceived by service provider

QoSO: Quality of Service offered/planned by service provider

QoSR: Quality of Service requirements of user/customer

QSTA: QoS-enhanced station

RSSI: Received Signal Strength Indicator

RTS: Request to Send

SAT: Satisfaction

SDG: Sustainable Development Goals (SDGs)

SIFS: Short Interframe Space

SNR: Signal-to-Noise Ratio

SOM: Self-Organizing Maps

SSID: Service Set Identifier

STA: Station

TFM: Trabajo Fin de Máster

TU Dublin: Technological University Dublin

TXOP: Transmission Opportunity

UN: United Nations

UPS: User's selection

UPV/EHU: University of the Basque Country

VLAN: Virtual local area network

WCSS: Within-Cluster Sum of Squares

WECA: Wireless Ethernet Compatibility Alliance

WLAN: Wireless Local Area Networks

WNIC: Wireless Network Interface Controller

WPA: Wi-Fi Protected Access

INDEX

List of figures

List of tables

1 INTRODUCTION

Few sectors have undergone such rapid evolution as the telecommunications sector. Starting with the first telegraph line deployed in the early nineteenth century, in just under 200 years, telecommunications have evolved in such a way that they are currently capable of connecting 5.3 billion users [1] around the world in just a matter of seconds.

Many are the reasons for this exponential evolution, some being the privatization of the sector that led to increased investment and the enhancement of infrastructure development, the major technological advancements obtained in the last few years and the ever-growing user demand that has driven the development of new services and applications. Another important factor that has led to the increase in the number of Internet users is wireless communications. It must be considered that initially, wireless communications were under government control and not open for public use. It was not until 1985, when the US Federal Communications Commission decided to open three of their controlled bands for unlicensed use, that companies started to develop proprietary wireless technology. Alongside this, in 1997, the Institute of Electrical and Electronics Engineers (IEEE) published the first wireless networking standard, IEEE 802.11, today known as Wi-Fi. This very same year, various technological companies established the Wireless Ethernet Compatibility Alliance (WECA), currently known as the Wi-Fi Alliance, with the goal of ensuring interoperability between various types of devices. The fact that people were no longer tied to a specific location to access the Internet meant that the number of Internet users increased dramatically.

As technology evolved, various QoS mechanisms were developed. These mechanisms have continued to gain importance to the extent of becoming an essential part of today's technologies. Wi-Fi, which is the technology that will be analyzed in the present study, is no exception. There are various mechanisms, such as Transmission Opportunity (TXOP), Enhanced Distributed Channel Access (EDCA) or HCF-Controlled Channel Access (HCCA), whose goal is to provide a better QoS. However, QoS is a complex matter, as there are many factors, both objective and subjective, that must be taken into consideration. To handle this challenge, various standardization bodies have dedicated specialized teams to QoS analysis and there are also several independent research groups working on the topic. These groups continuously work on developing new mechanisms and methodologies to improve QoS. One of the most recent methodologies, which is the central point of this study, is the application of Machine Learning (ML) techniques to QoS analysis. In recent decades, the use of ML has increased considerably, as it has been applied in many different fields. It is believed that the application of ML for QoS analysis could be of great use in this field too, as it

could offer the possibility of discovering underlying relationships amongst the different objective and subjective factors affecting QoS.

2 CONTEXT

This research study is carried out within the framework of a project promoted by the Networking, Quality and Security (NQaS) research group [2]. The NQaS is a research group of the Department of Communications Engineering (DIC/KIS) of the Faculty of Engineering of Bilbao belonging to the University of the Basque Country (UPV/EHU) [3]. Its main area of research is telematics, which is an interdisciplinary field that combines computing and communications technology for data transmission. The topics the research group focuses on are QoS analysis and management, security in data networks and the study of new technologies and transmission systems. This study is related to the first topic, that is, QoS analysis and management, as the project's goal is to determine the validity of the methodology proposed by the research group in the QoXphere model. QoXphere is a user-centric QoS model proposed by the NQaS research group whose objective is to consider all QoS-related aspects in order to perform an effective analysis and management of the service quality. This model will be further discussed in section 5.1.

The present project is also a continuation of a study previously conducted by the same author for her Bachelor's Degree Final Project under the Ikasiker collaboration grant. The study named "Machine Learning-based methodology for improving QoS in Wi-Fi networks" also sought to determine the validity of the methodology proposed in the QoXphere model. As the model is large and complex, the validation of its methodology is not performed all at once but in sections. The Bachelor's Degree Final Project focused on the initial step of this methodology, Context extraction. The purpose of this initial step is to identify different scenarios and user profiles within a certain context. Therefore, the project's goal was to find the most appropriate ML algorithm for performing this task. The context, that is, the real case scenario that was studied, was the Wi-Fi service of the university campus of the Technological University Dublin (TU Dublin) [4] and the data was obtained from the project's director's research stay at this same university. After comparing the results obtained by using several different ML algorithms, it was determined that the one that offered the best results was K-Means. It was with this algorithm that the initially set-out goal was achieved, that is, that the validation of the methodology was obtained, as the K-Means algorithm was able to identify four types of users.

A closer look at the obtained results will now be performed. As was mentioned earlier, the data was collected at the Technological University Dublin. To be more precise, it was collected at the library of the university and at one of its academic buildings. That is, a building containing classrooms, laboratories and offices. The data was therefore divided into these two groups when performing the data analysis, as can be

seen in Figure 1. Images on the first row show data belonging to the library and images on the second row show data of the academic building. Additionally, data was divided into three time frames: morning, afternoon and night. Images on the first, second and third column show data belonging to those time frames respectively. However, the night scenario is not considered when talking about user types as there are no active users to be identified at this time. The variables used for this analysis were transmission rate and total number of devices. By looking at the morning and afternoon time frames of the library scenario it can be seen that two clusters were identified vertically, while in the academic building scenario the clusters were identified horizontally. The clusters obtained in the library scenario were understood by linking them to the different study habits of the library students:

- **Library scenario user profile 1 (Cluster 1):** Users that do not use the network while studying, hence the lower transmission rates. This lack of network usage could be due to various reasons, such as, alternative use of books or notes.
- **Library scenario user profile 2 (Cluster 2):** Users that actively use the network while studying, resulting in higher transmission rates.

As for the clusters identified in the academic building scenario, it was concluded that each one represented data from a different area of the building.

- **Academic building scenario user profile 1 (Cluster 1):** Office and laboratory users. The lower total devices value is explained by the fact that the number of people in this area is expected to be lower.
- **Academic building scenario user profile 2 (Cluster 2):** Classroom users. The higher total devices value represents a larger number of users in this area.

It is in this context, that is, in the QoXphere validation context, that the present work is set. Having said that, in this project, a more in-depth validation of the methodology will be performed. To do so, a larger and more varied dataset will be used, as data will be gathered from various campuses of the University of the Basque Country. In addition to this a more extensive study and analysis of ML fields and unsupervised learning techniques will be performed to determine the most appropriate algorithms to use for the validation process. Lastly, it must be underlined that the previous study used OptiCube Wi-Fi probes that were capable of capturing Wi-Fi data and processing it to extract further information. However, these probes were not available for the data collection of this project. Instead, data was gathered directly from the APs installed in the university's facilities. These APs are not capable of processing data as was the case with the OptiCube

probes. Therefore, the collected data will be exclusively composed of real Wi-Fi traffic. This implies that a larger effort will have to be made in the data preprocessing step as this lacking information will have to be made up for by means of a more exhaustive analysis.

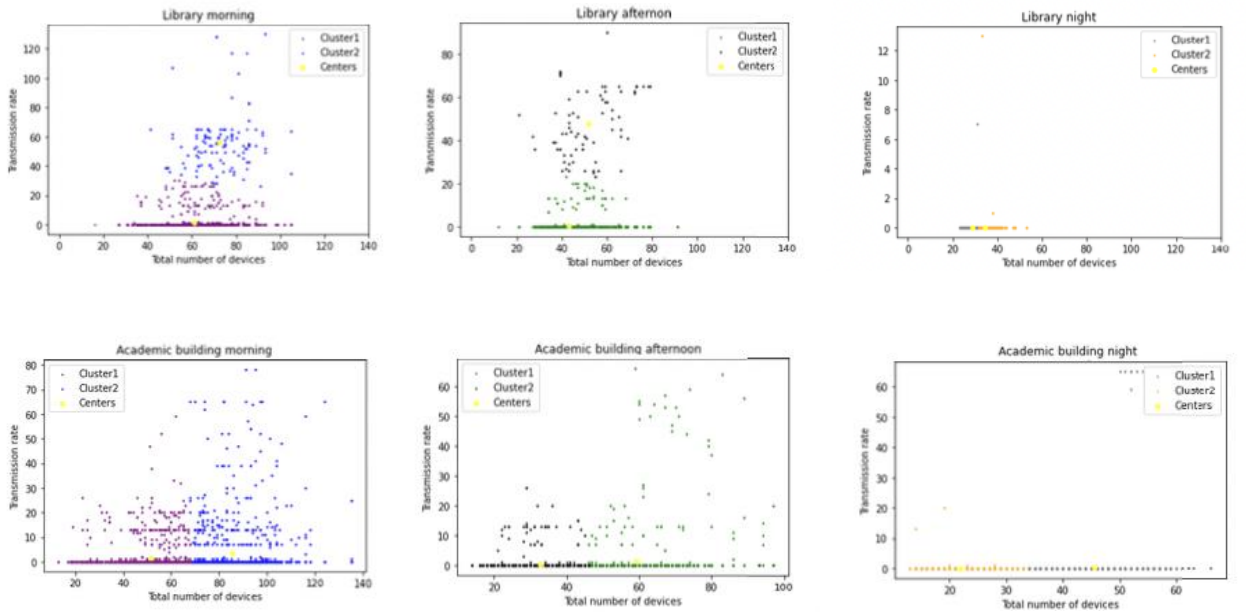

Figure 1: Results obtained in Bachelor's Degree Final Project

3 OBJECTIVES

As was previously mentioned, one of the main goals of the present research is to validate the methodology of the QoXphere model in a real case scenario. That is, to prove that the proposed methodology can be implemented in real life settings. This research will focus on the methodology's Context extraction step. That is, its goal will be to apply ML techniques to a given dataset in order to identify different users and scenarios based on their QoS requirements. To prove the validity of this methodology, a large amount of data has been collected from a real networking scenario and different ML algorithms will be used to analyze it.

On the other hand, the second objective of this project is to compare the results obtained in this study with those of the study carried out in the final degree project. The data that was analyzed for that study was collected at the Technological University Dublin, whereas for the present study, it was acquired from the University of the Basque Country. Given that the data was collected in two different scenarios, this study's objective is to analyze the influence of contextual aspects in user's QoS requirements by comparing the results obtained in these two different settings.

The third and last objective of this project is to work towards the Sustainable Development Goals (SDGs) proposed by the United Nations (UN) in the Agenda 2030 for Sustainable Development, whose ultimate goal is to achieve development in the economic, social and environmental domains by 2030. The Agenda 2030 is being adopted by governments, organizations and companies in many countries. An example of this is the definition of EHUAgenda 2023 by the University of the Basque Country. Among its proposals, it includes that students' Bachelor's Degree Final Project and Master's Degree Final Project should adopt a perspective in line with the agenda and that they should try to focus their efforts on improving some of the aspects included in the SDGs. This project concentrates on two SDGs: Quality and education and Industry, innovation and infrastructure. The first SDG aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" [5]. This project achieves this by not only allowing the researchers to grow professionally but also by offering future researchers and students the possibility to feed off the analysis made. The second SDG's focus is to "build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation" [5]. If the methodology in this project were to be validated, this would mean an improvement in terms of sustainability, as networks could be parametrized according to user needs leading to a reduction of the employed resources.

4 BENEFITS

As with any project, it is important to specify and understand the benefits the project will produce in order to justify its execution. This section covers the main technological, economic, social and scientific benefits that this research will provide.

4.1 Technological

This project's goal is to perform an analysis aimed at identifying different user types based on their QoS requirements. The identification of the different user types and their main requirements will provide extra knowledge which can then be used to adjust network parameters accordingly. This will allow for more efficiently and securely run networks.

4.2 Economic

The technological benefits highlighted how this study would help run networks more efficiently thanks to the network parameter adjustment. However, this would not only have a technological impact, but it would also provide financial benefits mainly from the provider's perspective. With the collected information, providers will be able to adjust parameters and employ their resources in a more efficient manner, which will in turn lead to cost reduction. An example of this would be that a more efficient use of bandwidth could be performed and therefore, there would be no need to invest in a project to increase it. Another benefit for providers would be that by acquiring more knowledge on their users' profiles and requirements they would be able to offer better QoS and reduce the attrition rate. The attrition rate, also known as CHURN, is the pace at which people leave a company. Therefore, reducing CHURN implies reducing company losses. Alternatively, users could also benefit from this study. Cost reduction can help providers offer more competitive pricing which could grant users lower prices for the same services.

4.3 Social

As was previously said, knowing what users' needs are will help adjust parameters to achieve such requirements. This will help improve users' Quality of Experience (QoE) when using different services or applications, which will in turn increase their overall satisfaction.

4.4 Scientific

The present study seeks to validate one of the steps of the QoXphere model. This model has been produced in accordance with many recommendations and standards published by different regulatory bodies. Therefore, if the results of this study are conclusive it could be useful for future research regarding QoS. Additionally, as the model's goal is to establish a framework for the analysis of QoS, if it were proven that it can be put into practice, it could be adopted as a new framework for QoS analysis or it could be used as a starting point to develop a new one.

5 STATE-OF-THE-ART KNOWLEDGE

5.1 QoXphere

As was mentioned previously, QoXphere is a user-centric QoS model proposed by the NQaS research group. Its goal is to consider all QoS-related aspects in order to perform an effective analysis and management of network QoS. In an effort for the reader to properly understand the intricacies of this model, this section will provide a thorough explanation of the most relevant concepts related to QoS. Once these concepts have been presented and understood, the QoXphere model will be introduced.

5.1.1 Quality of Service (QoS) in telecommunications

In recent years there has been a true technological revolution in which the development of cutting-edge technology has led to the creation of many new services. Several of these services involve real-time and multimedia applications which are very sensitive in terms of delay and jitter. These services demand high quality standards and therefore impose large requirements on the network. All of this has created an urgent need to develop mechanisms to ensure QoS and is the reason why there are currently many standardization bodies and research groups focusing on this field. However, the sudden need to develop QoS frameworks and mechanisms, alongside the lack of a well-defined QoS terminology, has led to the misuse of many QoS terms or to the creation of new QoS concepts that are often overlapping. In this section, a review of the most relevant QoS concepts will be performed by analyzing the QoS terminology proposed and adopted by the International Telecommunication Union (ITU) and the model presented in [6].

The first QoS-related ITU framework was proposed in the ITU Telecommunication Standardization Sector (ITU-T) E.800 Recommendation [7] published in 1994. This recommendation gave a user-centric definition of QoS defining it as the “collective effect of service performance which determines the degree of satisfaction of a user of the service” [7]. However, the framework presented in this recommendation lacked precision in many aspects and was therefore too vague to be used. That is why, in 2001, the ITU-T G.1000 Recommendation [8] was published in an effort to establish a more consistent framework for QoS. This framework was much more precise and it defined four viewpoints that should be considered when analyzing QoS by contemplating the user’s and provider’s perspective. A few years later, in 2008, the ITU-T E.800

Recommendation was updated [9] in order to introduce a more consistent set of terms and definitions. The four viewpoints proposed in ITU-T G.1000 were adopted and defined in the following manner (Figure 2):

- **QoS requirements of user/customer (QoSR):** "A statement of QoS requirements by a customer/user or segment/s of customer/user population with unique performance requirements or needs" [9].
- **QoS offered/planned by provider (QoSO):** "A statement of the level of quality planned and therefore offered to the customer by the service provider" [9].
- **QoS delivered/achieved by provider (QoSD):** "A statement of the level of QoS achieved or delivered to the customer" [9].
- **QoS perceived by user/ customer (QoSP):** "A statement expressing the level of quality that customers/users believe they have experienced" [9]. Therefore, it "may also be considered as QoSD received and interpreted by a user with the pertinent qualitative factors influencing his/her perception of the service" [9].

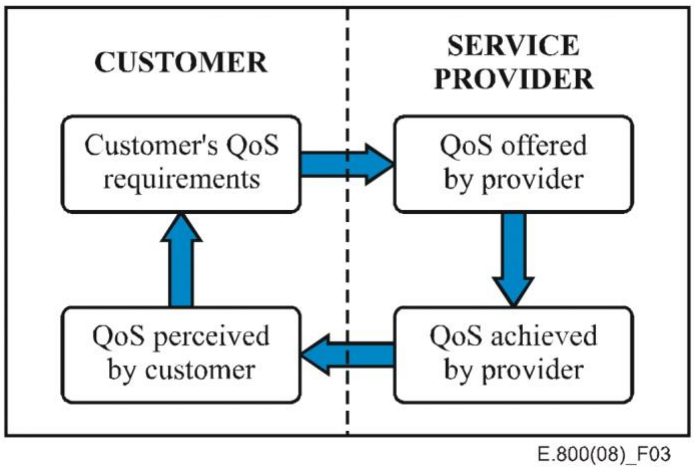

Figure 2: Four viewpoints of QoS [9]

In the updated version of the ITU-T E.800 Recommendation a new definition of QoS was also given: "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [9]. This definition of QoS is different from the previous one in the sense that it highlights the influence that all the components of the communication system have on QoS, including the end user. That is why it is also called end-to-end QoS (Figure 3).

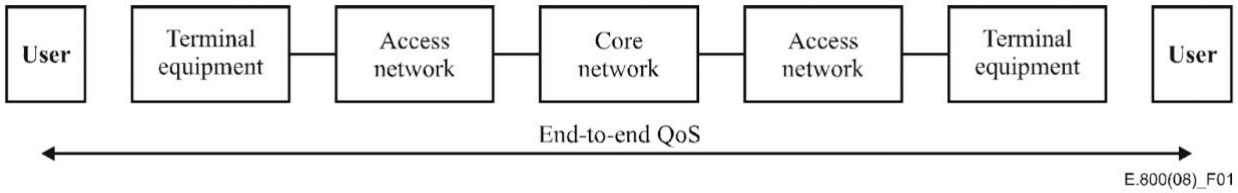

Figure 3: End-to-end QoS [9]

QoS is a concept that is determined by objective and subjective components (Figure 4). The ITU-T E.800 Recommendation defines objective or quantitative parameters as "parameters that are measurable (with instruments or observation*s)"* [9]. These quantitative parameters depend on the network infrastructure and that is why they are collectively referred to as Network Performance (NP). Some examples of NP parameters are bit error rate or latency [9]. On the other hand, subjective or qualitative parameters are defined as "parameters that can be expressed using human judgment and understanding" [9]. These parameters "can be influenced by user expectations, ambient conditions, psychological factors, application context, etc" [9]. As these parameters do not rely on the network infrastructure itself, they are collectively referred to as non-network related performance. Examples of non-NP parameters are repair time, rate offer range or the complaints resolution time offered by the provider [9].

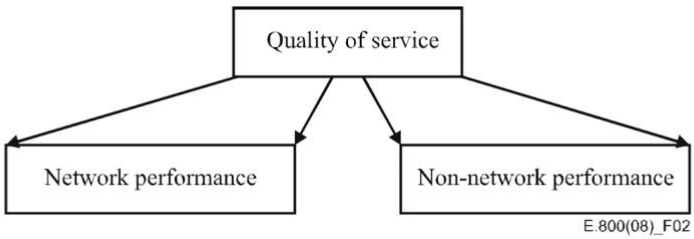

Figure 4: Quality of service components [9]

Lastly there is the ITU-T P.10/G.100 Recommendation [10] which was published in 2017. The most important contribution of this recommendation was the introduction of the QoE concept, which was defined as the "the degree of delight or annoyance of the user of an application or service" [10]. The level of user satisfaction depends on "the application or service, context of use, the user's expectations with respect to the application or service and their fulfilment, the user's cultural background, socio-economic issues, psychological profiles, emotional state of the user" [10] amongst other factors.

On the other hand, there is the model proposed in [6] which proposes a different view of QoS from the one proposed by the ITU. According to this model QoS should be considered from three different perspectives:

- **Intrinsic QoS**: It depends on the technical aspects of the network, such as the transport network design, provisioning and protocol and parameter selection. It is evaluated by comparing expected and measured performance values [11].
- **Perceived QoS**: It reflects the customer's overall satisfaction when using a service. That is, it is the user's evaluation of the intrinsic quality of service based on his or her personal expectations. Therefore, it can be said to be subjective as it will vary from one user to another [11].
- **Assessed QoS**: It is the user's decision as to whether to continue using a service or not. It depends on many factors such as the customer service offered by the service provider [11].

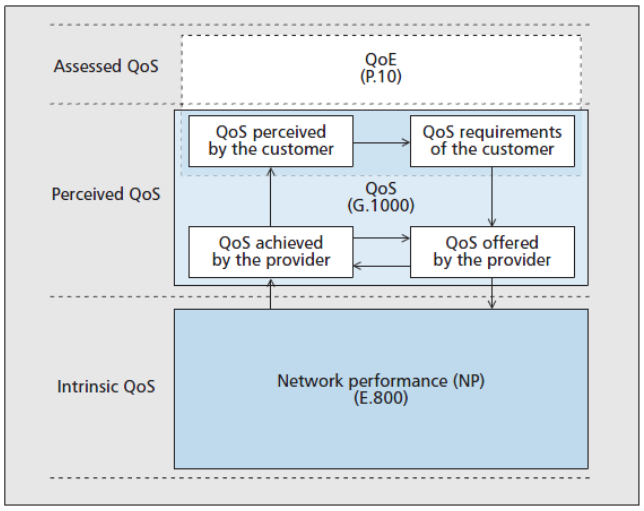

Figure 5: ITU-T terminology and standards in relation to the general QoS model [12]

The frameworks proposed by ITU and the model described in [6] have some concepts in common and others that differ, Figure 5 shows the relationships between them. As can be seen, intrinsic QoS is what in the ITU framework is known as NP. As for perceived QoS, this is represented in the ITU framework by the four viewpoints mentioned previously: QoSR, QoSO, QoSD and QoSP. Lastly, there is assessed QoS. In terms of the ITU terminology, this would be the ensemble of QoE, QoSR and QoSP.

5.1.2 QoX and QoXphere

The flourishing of all these QoS-related terms drew attention to the need of creating a new concept that would gather them all. This concept is what is known as QoX, where X stands for anything. Along with the creation of QoX came the need to create a suitable framework for it. This has been one of the main research points of the NQaS research group, who propose the QoXphere named framework. QoXphere is a "user-centric and business-oriented multi-layer model that takes into account most of the different concepts and aspects defined in the current QoS regulation and standards" [12]. In order to consider all the different QoS-related aspects, it is divided into four layers, as can be seen in Figure 6.

The first three layers coincide with the layers proposed in [6], that is, intrinsic, perceived and assessed QoS. Additionally, a fourth layer is considered in the QoXphere framework, the business QoS layer. As can be seen in Figure 6, there are specific QoS aspects that need to be taken into account in each of these layers. These aspects are intra-related within the same layer but are also interconnected with aspects of the layers directly above or below, which means that each layer influences its adjacent layers.

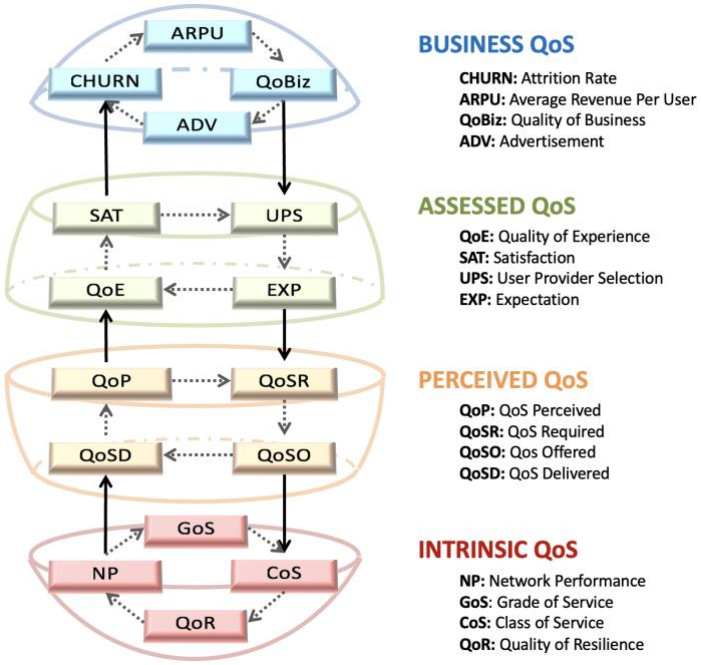

Figure 6: QoXphere layer specification [12]

- **Intrinsic QoS layer**: It "reflects the service features stemming from the technical aspects" [12]. It is the lower layer of the QoXphere model and it contains a total of four QoS aspect to be analyzed:

- **Network performance (NP):** "Ability of a network to provide the functions related to communications between users" [12]. It has a direct effect on the QoS Delivered (QoSD) of the perceived QoS layer.
- **Grade of Service (GoS): "**It is the categorization of services with respect to requirements that can be verified through NP" [12].
- **Class of Service (CoS):** "Any of the network-oriented designations that can distinguish between various services" [12]. CoS is used to divide similar types of traffic into groups to handle each group with a different level of priority. It is directly affected by the QoS Offered (QoSO) of the perceived QoS layer.
- **Quality of Resilience (QoR): "**Describes network survivability" [12]. That is, it refers to the probability of having the service available.

- **Perceived QoS:** It "reflects the customer's experience of using a particular service" [12]. It is the second layer of the model and it also contains four QoS-related aspects. Those aspects are based on the four viewpoints presented in the ITU-T G.1000 Recommendation:
  - **QoS Required (QoSR)**: "A statement of QoS requirements by a customer/user or segment/s of customer/user" [12]. It is directly affected by the Expectation (EXP) aspect of the assessed QoS layer.
  - **QoS Offered (QoSO)**: "A statement of the level of quality planned and therefore offered to the customer" [12].
  - **QoS Delivered (QoSD)**: "A statement of the level of QoS achieved or delivered to the customer" [12].
  - **QoS Perceived (QoP)**: "A statement expressing the level of quality that customers believe they have perceived" [12]. It has a direct effect on the upper layer as it is interconnected to the QoE aspect of the assessed QoS layer.
- **Assessed QoS:** It "reflects the user's satisfaction and decision of remaining with the provider or not" [12]**.**
  - **Quality of Experience (QoE): "**The overall acceptability of an application or service, as perceived by the user" [12].

- **Satisfaction (SAT): "**Global customer's satisfaction with the service" [12]. It has a direct effect on the CHURN, an aspect contemplated on the last layer of the model.
- **User's selection (UPS): "**Provider selection made by the user/customer" [12]. The Quality of Business (QoBiz) aspect of the Business QoS layer has a direct effect on this aspect.
- **Expectation (EXP): "**User expectations concerning the quality of the service" [12].

- **Business QoS:** It "reflects the provider business stage" [12] and its goal is to ensure that operator's achieve profitability, which is achieved by guaranteeing user loyalty. As user loyalty depends on the objective and subjective parameters mentioned in previous layers, this layer is placed on the upper part of the QoXphere model. Once again, it contains a total of four QoS aspect to be analyzed:
  - **Attrition rate (CHURN):** "Measure of customers moving out of a collective over a specific period of time" [12].
  - **Revenue (ARPU):** "Average Revenue Per User (services provided/the number of users buying the services)" [12].
  - **QoS of Business (QoBiz):** "QoS metric that quantifies the business return of a service provider (profit/revenue)" [12].
  - **Advertisement (ADV):** "Providers media and publicity policy" [12].

5.2 QoS in IEEE 802.11

The data used in this study has been collected from the eduroam network of the University of the Basque Country. Eduroam [13] is an international Wi-Fi service for students, researchers and personnel of higher and further education institutions. As was explained in section 1, Wi-Fi is the common name given to the IEEE 802.11 family of standards. Therefore, as data relying on these standards will be used, it is necessary to have a good comprehension of them. More specifically, as this study is focused on QoS, it is important to understand the mechanisms that these standards implement for providing QoS. Therefore, in this section, an overview of the IEEE 802.11 standards will be given, followed by a more detailed look at the QoS enhancement mechanisms.

5.2.1 IEEE 802.11 overview

IEEE 802.11 is a set of standards that define the implementation of the physical and data link layers for Wireless Local Area Networks (WLAN). The first standard was published in 1997 and all the standards that have followed have provided some type of improvement regarding their predecessors. In order to guarantee the correct functioning of Wi-Fi, all the new standards must be compatible with the previously published ones. Therefore, many of the standards share certain characteristics such as, medium access protocols or frame structure [14]. This subsection will perform a brief review of the IEEE 802.11 standards by focusing on the architectures used and the functioning of the data link or medium access layer (MAC) layer.

There are two components in WLAN communications, stations (STAs) and access points (APs). The former are the devices trying to exchange information, which could be mobile phones, computers or Internet of Things (IoT) devices. As for the latter, this is a base station that coordinates the information exchange amongst the STAs and that provides the connection towards the Internet. Based on how these components are organized there are two main architectures (Figure 7):

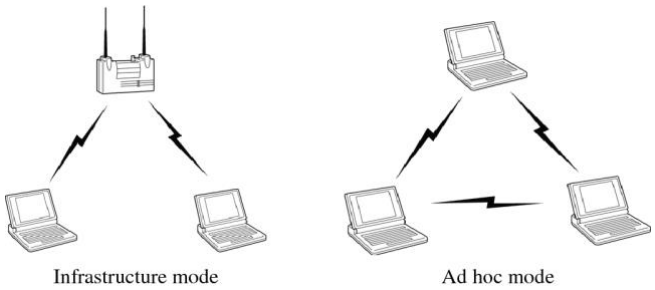

Figure 7: IEEE 802.11 architectures [15]

- **Infrastructure mode**: In infrastructure mode, STAs communicate through an AP that is connected to a router that provides access to the Internet. This means that in order to communicate amongst STAs or to connect to an outside network all information must pass through the AP. The ensemble of the AP and STAs make up what is known as the Basic Server Set (BSS).
- **Ad-hoc mode**: In this mode, two or more STAs communicate amongst one another in an independent manner, that is, without the need of an AP or any other network device. Ad hoc networks are especially useful when a network is only required temporarily or when wiring cannot be laid down. If in addition to information exchange between STAs Internet access were required, at least one of

the STAs would have to be connected to it. The ensemble of STAs in ad-hoc mode is known as Independent Basic Service Set (IBSS).

IEEE 802.11 defines the functioning of the physical and data link layers. To comprehend the main QoS mechanisms adopted by these standards it is important to have a good understanding of how the data link layer works. One of the most important aspects controlled by this layer is medium access control. It is essential because in Wi-Fi, many STAs try to transmit on the same channel at the same time. If this were to happen, a data collision would occur and the data would not be transmitted correctly. In the first IEE 802.11 standard two medium access control techniques were defined: Distributed Coordination Function (DCF) and Point Coordination Function (PCF).

DCF is a technique based on the access protocol known as carrier sense multiple access with collision avoidance (CSMA/CA). This type of protocol is based on the idea that in order to transmit, STAs must first sense the channel to determine if it is free. As the name states, DCF is distributed technique as the decision of accessing the medium is performed by each STA. This technique works in the following manner:

1. **Carrier sense**: First, the station senses if the medium is free for a time interval known as DCF Interframe Space (DIFS).
2. **Backoff**: Once the first step is completed, that is, once the channel is free for a DIFS interval, a random time interval known as backoff time is selected. A backoff time is a waiting time that is randomly chosen from a range of values spanning form 0 to a value known as contention window (CW). Initially, all STAs begin with the same CW value (CWmin). Once the value has been selected, the STA starts counting down from this value (this countdown is only performed while the channel remains idle, otherwise the countdown is stopped). Once the backoff value reaches 0, the data is transmitted.
3. **Acknowledgement (ACK) transmission:** As was mentioned previously DFC is based on the CSMA/CA protocol. As this is a collision avoidance protocol it does not guarantee that collisions will not occur. Therefore, it is necessary to ensure that the transmissions are carried out successfully. The way to confirm this is by means of ACKs. Upon correct reception, the receiver will send an ACK to the sender using the previously described steps. The only difference is that, when doing so, instead of waiting for the channel to be idle for a DIFS interval it will wait for a shorter interval known as short interframe space (SIFS). This will give ACKs transmission priority over regular data frames. In the event that an ACK is not received, it is supposed that the data was not transmitted correctly and

therefore the transmission is repeated. However, this time, the CW value of the sending STA is increased. The reason for increasing it is that the bigger the value, the larger the interval for selecting the backoff time will be and thus, the less likely it will be for STAs to select the same value and for a collision to occur.

In contrast to DCF, PCF is not decentralized but centralized. As its name indicates, there is one point known as the Point Coordinator (PC), usually the AP, in charge of coordinating the access to the medium. In this mode two periods are defined, the contention period (CP) and the contention free period (CFP). In the former, the DFC technique is used for accessing the medium whereas in the second PCF is employed. PCF works in the following manner:

1. **Carrier sense**: In this mode, the PC senses if the medium is free for a time interval known as PCF Interframe Space (PIFS), which is shorter than DIFS but longer than SIFS. This will give it priority over the transmission of stations using DCF but not over the transmission of ACKs.
2. **Initiating the CFP**: Once the medium has been free for a PIFS interval, in order to indicate the beginning of the CFP, the PC sends a beacon frame to all STAs.
3. **Polling**: The PC then polls the STAs one by one asking them to send data. If they do not have any data to be transmitted, they will simply respond with a frame with no payload and an ACK indicating they have received the poll. Additionally, if the PC has data destined for the STA it has polled, it will send it along with the poll.
4. **Ending the CFP**: The polling process is repeated until all stations have been polled or until the CFP expires. At that point, a contention free-end (CF-end) frame is sent to indicate the end of the CFP.

5.2.2 QoS limitations and the IEEE 802.11e standard

From the previously presented explanation of DCF and PCF it can be determined that these techniques do not offer QoS mechanisms. In the case of DCF, all the STAs have the same priority when competing to access the medium. That is, DCF only provides best effort service, it does not give priority to services with more stringent requirements such as real-time multimedia applications. PCF was designed to overcome this problem, that is, to offer support for real-time multimedia applications. However, it still faces strong QoS limitations. On the one hand, PCF uses a single-class round robin system when polling and therefore it cannot offer priority access based on the different types of traffic. On the other hand, it cannot calculate how long

an STA will be transmitting once it has been polled. The difficulty for calculating this duration relies mainly on the fact that the length of the data frame to be transmitted is not known and also that the data rate will vary depending on the medium's condition [16].

In an effort to offer the lacking QoS support, the IEEE 802.11e standard was published in 2005. In terms of architecture, the main difference in this standard is the name given to the APs and STAs that implement it, which are now referred to as QoS-enhanced APs and (QAPs) and QoS-enhanced STAs (QSTAs). The BSS that these components make up is known as a QoS supporting BSS (QBSS). As for the data link layer, the following mechanisms are introduced [17]:

- **Transmission Opportunity (TXOP):** It is a time interval assigned to a QSTA during which it has the opportunity to send data frames, once this time is up the QSTA must stop transmitting. This mechanism solves the aforementioned problem of not knowing how long an STA will transmit for [16].
- **Hybrid Coordination Function (HCF):** In the IEEE 802.11e standard a new medium access technique is proposed. This access technique is known as the hybrid coordination function (HCF). It is called hybrid because it combines the use of contention periods and contention-free periods. For the CP, a method named enhanced distributed channel access (EDCA) is used. Whereas for the CFP, the HCF-controlled channel access (HCCA) method is employed. Thanks to these new medium access techniques, it is no longer just a best effort service that is offered.

  In EDCA, traffic is classified into one of the four following access categories (AC): AC_VO (voice), AC_VI (video), AC_BE (best effort) and AC_BK (background). The functioning of EDCA is based on the fact that each STA will have four queues (one per AC), known as backoff entities, competing to access the medium. Each backoff entity will have its own maximum and minimum CW values and TXOP duration. Additionally, in order to access the medium, instead of waiting for the medium to be idle for a DIFS, QSTAs will have to wait for an arbitration interframe space (AIFS). Each AC has a different AIFS duration with higher priority ACs having a lower AIFS duration. In this way, when accessing the medium, preference is given to the ACs with the highest priority. It could also happen that the backoff time of two or more of the backoff entities finishes simultaneously. If this were to occur, only the AC with highest priority would be allowed to transmit.
  As for the HCCA technique, it is considered more flexible than PCF. The reason for this is that in PCF polling is only permitted in the CFP, whereas in HCCA a QAP can initiate HCCA at any time given the

channel has been free for an interval of PIFS. The PIFS time interval is shorter than DIFS and AIFS, therefore HCCA will have priority over EDCA.

- **Block Acknowledgements:** This standard introduces block acknowledgement. This mechanism enables multiple consecutive frames to be sent without the need to wait for each frame to be acknowledged. Thanks to this, the frames are acknowledged in a single ACK frame instead thereby improving throughput efficiency.
- **Direct Link Setup:** When using infrastructure mode, each time STAs want to communicate amongst one another they must do it through the AP. However, when this communication take place only between two STAs channel resources are wasted. A solution to optimize this is to set up a direct link between the two communicating STAs.

5.3 Machine Learning

Machine learning (ML) is a field focused on the development of computer systems that are capable of learning to analyze and infer patterns from data without being explicitly programmed to do so. It is important to note that ML is considered a subfield of Artificial Intelligence (AI) and not the opposite way around. AI is a broad term that refers to systems or machines that try to mimic human intelligence. Therefore, although all ML is considered AI, not all AI is ML. The 'machine learning' term was introduced in the 1950s, therefore, it can be said that it is quite a recent field of study. However, the increasing amount of data available today has allowed this field to evolve very rapidly and that is why today many different algorithms and types of ML are used. In this section, a brief explanation of the main fields of ML will be carried out. The three most important ones are:

- **Supervised Learning**: This type of learning uses labeled data, which is data to which a label has been added to assign a specific class or output value. In supervised learning, the labeled data is used to train a model in order to make it capable of predicting the class or value of non-labeled. There are two main types or supervised learning: if the predicted value is a class, that is, a discrete value, the problem is known as classification, whereas if the output is a continuous value, it is called regression.

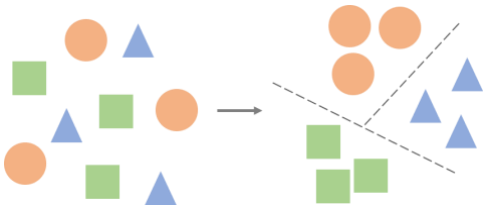

Figure 8: Supervised learning [18]

- **Unsupervised Learning**: On the other hand, in unsupervised learning, non-labeled data is used. That is, in this type of learning the data samples used have no output value assigned. That is why, in this type of learning, the goal is to find the underlying relation in the data. The most important unsupervised algorithm is clustering, where data is separated into groups (clusters) according to common characteristics.

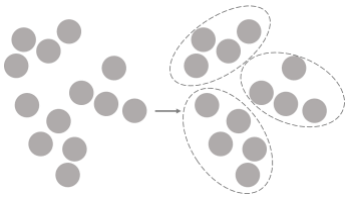

Figure 9: Unsupervised learning [18]

- **Reinforcement Learning**: In reinforcement learning, an agent autonomously makes decisions to achieve a final goal. As this type of learning is based on the use of a reward system, for each decision made, the agent will either receive a reward or a penalty depending on whether the action taken was beneficial or not to reach the final goal. This way, the best set of actions needed to obtain the final goal are determined.

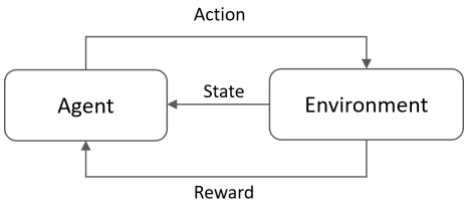

Figure 10: Reinforcement learning [18]

Although these are the three main fields of ML, many subfields of ML have emerged. As labeled data is usually hard and expensive to obtain different solutions have been offered in order to exploit the use of unlabeled data which is much more accessible. Some of these methods are:

- **Semi-supervised Learning:** In this type of learning, a large amount of unlabeled data and a smaller amount of labeled data is used. In order to label the unlabeled data, first a classifier is trained on the labeled data. The most reliable unlabeled points, along with their predicted labels, are added to the training set. The classifier is then retrained with the new data and the procedure is iteratively repeated [19].

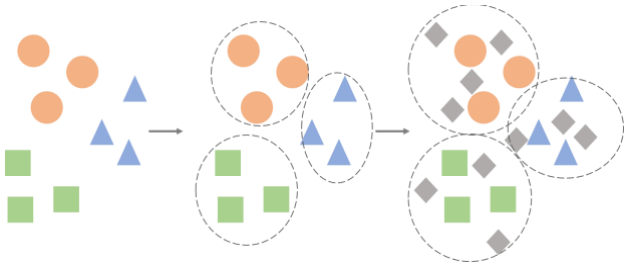

Figure 11: Semi-supervised learning [18]

- **Self-supervised Learning**: For this type of learning an unlabeled dataset is initially available. This data is autonomously labeled by using unsupervised learning which is capable of identifying underlying relations between samples. This means that data is labeled without human supervision. Once the data has been labeled, it is used to train a supervised learning model in order to perform the main task.

Figure 12: Self-supervised learning [18]

- **Self-taught Learning**: The difference between this type of learning and self-supervised learning is that, for self-taught learning, it is not assumed that the unlabeled data can be labeled with the labels required for the supervised learning task. For example, consider a task of classifying images of cats and dogs. With self-taught learning, random Internet images (not necessarily of cats or dogs) would be used to learn a representation. Then this representation would be used to classify the images of cats and dogs. This type of learning presents two main advantages. The first, that once the representation is learned it can be applied to different classification tasks, not only to cats and dogs.

Second, data obtention is much simpler as it is easier to obtain random images than images of cats and dogs [20].

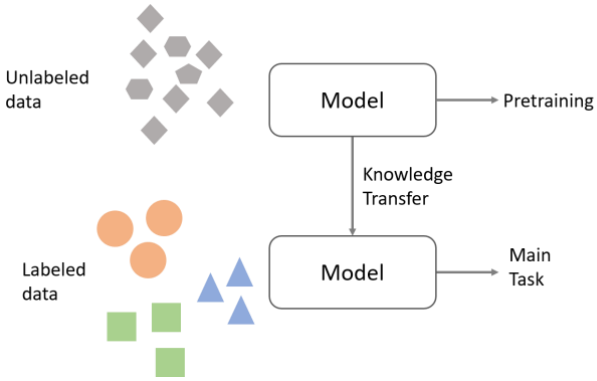

Figure 13: Self-taught learning [18]

- **Active Learning**: In active learning, the most important or relevant data instances are selected from a set of unlabeled data. Once these samples have been selected, a human expert with domain knowledge is asked to help in the task of labeling them. However, this human labeling interaction should be kept to a minimum, as active learning's goal is to obtain as good a model as possible without having to label more data than necessary. Finally, once the data has been labeled, a supervised learning model is trained to achieve the final task [21].

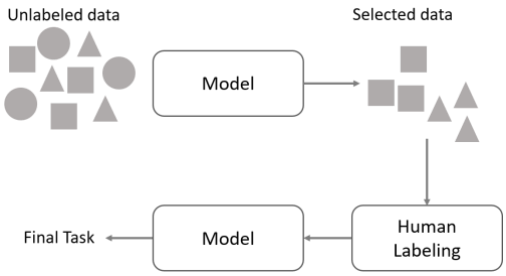

Figure 14: Active learning [18]

As well as the previously mentioned hybrid approaches where labeled and unlabeled data is used, there are some other ML subfields that are worth mentioning.

- **Multi-task Learning**: In this type of learning various related tasks are solved simultaneously by means of a single model.

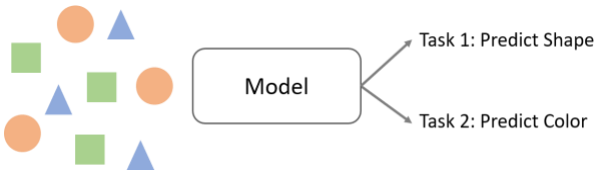

Figure 15: Multi-task learning [18]

- **Online Learning:** The main difference between online learning and other subfields of ML is that in online learning data is available in a sequential order. That is, in contrast with other approaches, it is not all available from the beginning. This is an interesting approach to use when the model must dynamically adapt to new data.

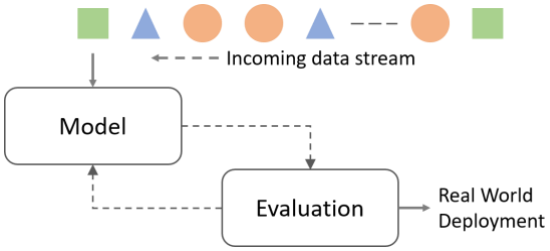

Figure 16: Online learning [18]

- **Transfer Learning:** In transfer learning a model is obtained by training or fine-tuning another model that has already been trained and used for a similar task. It gets its name from the idea that knowledge is transferred from one task to the other.

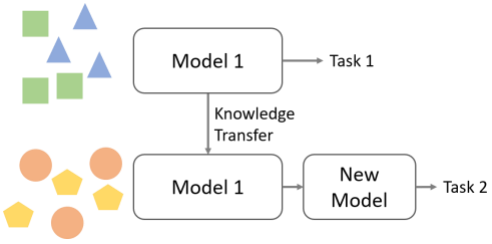

Figure 17: Transfer learning [18]

- **Federated Learning:** The data that is used to train ML models is usually centralized. However, sometimes, due to privacy restriction data cannot be shared and gather in a unique database. Federated learning is a ML approach that uses distributed data. In this case, models are locally trained

on the distributed data and instead of sharing the data, it is the model parameters that are sent to a central server in order to train a global model.

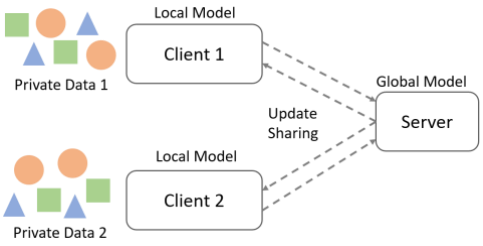

Figure 18: Federated learning [18]

- **Ensemble Learning:** The rationale for this kind of learning lies in improving the predictive performance of a single model through the training of several models and the combination of their predictions.

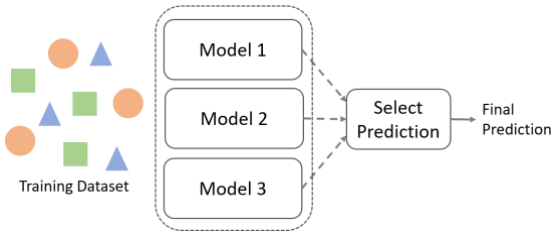

Figure 19: Ensemble learning [18]

- **Adversarial Learning**: In adversarial learning deceptive data samples, also known as adversarial examples, are added to the training data. This way the model is trained as to not be fooled by these examples.

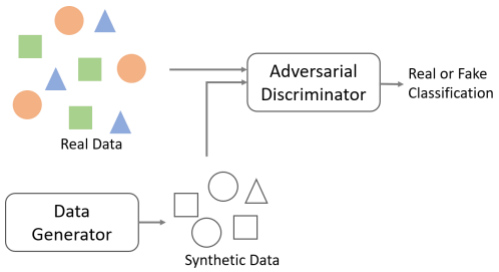

Figure 20: Adversarial learning [18]

- **Deep Learning:** This type of learning imitates the neural networks of the human brain by means of artificial neural networks (ANNs). Deep references the number of layers in the ANN. There are three types: input layer (receives data), output layer (outputs the prediction) and hidden layer or layers (data representations are learned with multiple levels of abstraction) [18].

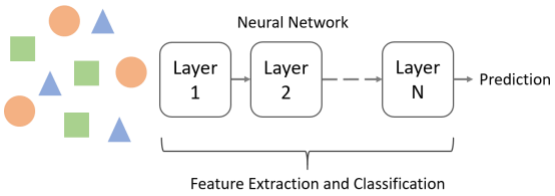

Figure 21: Deep learning [18]

6 ANALYSIS OF ALTERNATIVES

6.1 Clustering methods

In the previous section a review of different ML techniques was carried out. Given that the data that will be used in this project is unlabeled data, a good option for this study is to employ unsupervised learning techniques. This means that clustering algorithms will be used in this data analysis.

However, there are several different types of clustering methods (Figure 22). Hence, an analysis of alternatives will be performed to determine the most appropriate one for this particular case. According to recent studies [22], [23], there are two main types of clustering: hierarchical clustering and partitional clustering. Therefore, the first important decision is to determine which option is the most appropriate for the present study. An alternative matrix has not been used for this decision as it depends mainly on the characteristics of the dataset and not on the advantages or disadvantages of the clustering method.

- **Hierarchical clustering**: In this type of clustering, a hierarchical structure known as dendrogram is created to depict the relationships between the data points. Hierarchical clustering is a good option for cases where understanding the hierarchy and relations between clusters is of great importance, when the number of clusters is not predetermined and there is an interest in exploring the different levels of granularity or when the dataset is relatively small, as the calculation of the dendrogram is computationally expensive.
- **Partitional clustering**: In this type of clustering, data is divided into disjoint clusters. Partitional clustering can be useful when the data does not seem to follow a hierarchical structure, when the number of clusters is predefined or when the employed dataset is large.

Given that no hierarchical structure is expected amongst the data (the instances offer information of independent eduroam sessions), the partitional clustering method has been determined more adequate. Additionally, even though there is no specified number of clusters, this value is expected to be between a given range and therefore, it has not been considered pertinent to analyze the different levels of granularity that a hierarchical clustering method could offer.

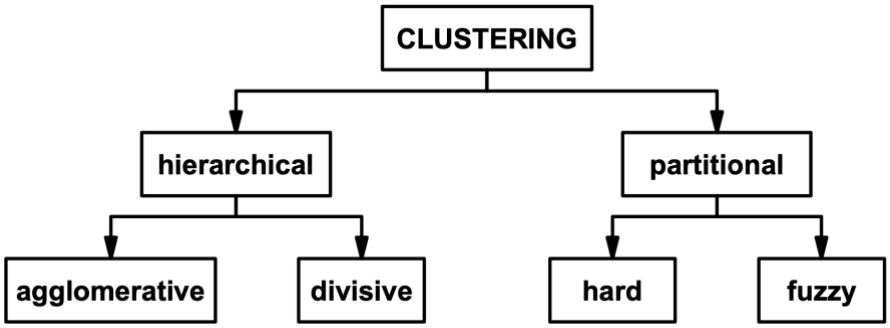

Figure 22: Taxonomy of clustering methods [24]

As can be seen in Figure 22, partitional clustering can further be divided into two subcategories: hard clustering or fuzzy clustering. Therefore, as was done for hierarchical and partitional clustering, it must be decided which of the two methods is the most appropriate for the current project's case.

- **Hard clustering**: In this type of clustering each data point is assigned to a unique cluster. This type of clustering is easily interpretable but is sensitive to outliers and cannot handle ambiguous memberships to clusters.
- **Fuzzy clustering**: In this type of clustering data can belong to multiple clusters. A membership value is assigned to each data point for each cluster, indicating the degree to which the data point belongs to that cluster. Fuzzy clustering is more flexible as it can handle data ambiguities by means of the membership assignments but is not as easily interpretable.

Once again, this decision does not require the use of a decision matrix, it can be determined simply by analyzing the type of dataset and the projects end goal. Given that the goal of the project is to analyze and obtain different user profiles, it is expected that the clusters that will be obtained will not be overlapping. Therefore, a hard clustering approach will be used.

Now that it has been determined that a hard partitional clustering method will be used, the specific method must be selected. Once again, as was the case with the previous methods, hard clustering can be divided into subcategories [22], [23]: search-based methods, graph-theoretic methods, density-based methods, model-based methods, subspace methods and centroid-based methods. In order to select the most appropriate clustering method an alternatives matrix containing the following selection criteria will be used. Additionally, a scale from 1 to 5 (1 being the lowest score and 5 the highest) will be employed to evaluate each method on the criteria below.

- **Handling of irregular cluster shapes:** Many algorithms assume that the clusters will have a certain shape. This assumption can lead to an incorrect clustering as the clusters do not necessarily have a defined format. Given that this aspect can have a significant impact in the clustering process, especially if the clusters are of different or irregular shapes, this criterion has been given a weight of 20% for the score calculating process.
- **Robustness to outliers:** It is important for the algorithms to be able to deal with data points, known as outliers, that deviate significantly from the rest of the data. If these outliers are not detected and correctly managed, they can change the result of the clustering significantly. This criterion is given the same weight as the previous one, 20%, as it is considered to have a similar effect and importance.
- **Parameter sensitivity:** The initially selected values of the parameters can make the end result vary, as the algorithms may converge to different answers based on them. Parameter selection also affects the quality and reliability of the result. Therefore, the same weight, 20%, has been given to this criterion.
- **Complexity and implementation:** This will depend on two main aspects; the learning curve involved in grasping the theoretical and mathematical concepts behind the functioning of the algorithm and the difficulty involved in the calculation of the initial parameters. The weight of this criterion is 40%.

6.1.1 Search-based clustering

One of the main problems with partitional clustering methods is the need of prior information to determine the number of clusters that the data should be divided into. To overcome this problem, search-based clustering, also known as automatic data clustering, offers the possibility of determining the number of clusters without prior knowledge. It does so by treating the clustering problem as an optimization problem, where the goal is to maximize the similarity within a cluster and maximize the dissimilarity between them by applying metaheuristic techniques inspired on natural processes [25]. Some well-known search-based algorithms are Genetic Algorithm (GA), Differential Evolution (DE) and Particle Swarm Optimization (PSO) [22].

- **Handling of irregular cluster shapes:** The accurate detection of irregular shaped clusters will depend on the employed search strategy. As this varies from one type of search-based algorithm to another this criterion has been given a 3.

- **Robustness to outliers:** As with the previous criteria, the ability to detect outliers varies from one search strategy to another. Therefore, this criterion has also been awarded a 3.
- **Parameter Sensitivity:** The strong point of this method is the fact that it does not need to rely on prior knowledge for determining the number of clusters, which simplifies the parameter tuning process. This is why a score of 5 has been given.
- **Complexity and implementation:** The complexity of the algorithms that are classified within this type of method is quite high, as they are based on sophisticated natural processes, meaning that their implementation is also difficult. Given this, the complexity and implementation aspect of search-based clustering has been given a 1.

6.1.2 Graph-theoretic clustering

Graphs are structures made up of a set of nodes or vertices and a set of edges that connect these nodes. These nodes or vertices represent the data points of a dataset and the edges represent the similarities between them. Graph-theoretic clustering is the process of grouping the data points or nodes into clusters. This should be done in such a way that the formed clusters have the maximum number of edges within them and the minimum number of edges between different clusters [26]. An example of this type of clustering is Cluster Identification via Connectivity Kernels (CLICK) [22].

- **Handling of irregular cluster shapes:** Graph-theoretic clustering is considered a flexible approach as it can model irregular cluster shapes as nodes and edges. As it is considered adequate when dealing with irregularly shaped clusters, a 5 has been awarded for this particular criterion.
- **Robustness to outliers:** As with the search-based method, the capacity to deal with outliers depends on the algorithm. Following the same logic, a 3 has been given.
- **Parameter Sensitivity:** Depending on the algorithm, a different set of parameters must be adjusted, all of which will have an impact on the end result. Therefore, as with the previous criterion, a 3 has been given.
- **Complexity and implementation:** The two main problems related to this method are the tuning of the parameters and the construction of the graph, this second one being the most intricate. This is why this criterion has been scored with a 1.

6.1.3 Density-based clustering

Density-based clustering relies on the idea that clusters are continuous regions of high data points density separated from other clusters by continuous low data point density regions. In order to determine what is considered as a high- or low-density region a threshold value is defined [27]. Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS) and Density-based Clustering (DENCLUE) are some examples of density-based clustering algorithms [22].

- **Handling of irregular cluster shapes:** One of the main advantages of this method is that it does not assume that the clusters have a specific shape. This is due to the fact that it simply searches for contiguous high-density regions that could have any type of form. As this is one of the main advantages of this type of method, it has been awarded a 5.
- **Robustness to outliers:** Another important advantage of this method is that once again, due to the fact that it is density-based, separated data points, i.e., outliers, will be considered as noise points. In this case, a 4 has been granted.
- **Parameter Sensitivity:** It is quite sensitive to parameters such as the distance between data points that is required to be considered a high-density region. As this is one of the disadvantages of the model a 2 has been given.
- **Complexity and implementation:** Due to the difficulty of determining the correct values of the parameters a 2 has also been given.

6.1.4 Model-based clustering

Model-based clustering, also known as mixture model clustering, considers that the dataset is generated from a combination of mathematical models, where each model represents a different cluster. Therefore, in order to identify the clusters, the goal is to obtain the parameters of these different models [28]. Examples of these types of clustering include Expectation-Maximization (EM) and Self-Organizing Maps (SOM) [22].

- **Handling of irregular cluster shapes:** This method does not assume that the cluster will have a certain shape. It is capable of handling irregular cluster shapes effectively by fitting distributions or other probabilistic models to the data. As this is the main advantage of the method a 5 has been awarded as the score of this criterion.

- **Robustness to outliers:** Most model-based clustering methods assume that the dataset is outlier free, therefore they are not capable of dealing with them. This is why a score of 2 has been given.
- **Parameter Sensitivity:** This model type is very sensitive to the parameter choice, such as the number of components of the model. Therefore, a 1 has been given.
- **Complexity and implementation:** Due to the difficulty of selecting the model's parameters a 2 has been selected as the score of this criterion.

6.1.5 Subspace clustering

Subspace clustering is a type of clustering based on the idea of feature selection. Many datasets include features that are irrelevant or that offer little to no additional information. In this type of clustering, the goal is to utilize the most relevant features in order to define low-dimensionality subspaces where clusters can be better identified [29]. Clustering in Quest (CLIQUE), Entropy-Based Subspace Clustering (ENCLUS) and Projected Clustering (PROCLUS) are all examples of subspace clustering algorithms [22].

- **Handling of irregular cluster shapes:** As with some of the other methods, the ability to deal with irregular clusters will depend on the specific algorithm. That is why, as in the other cases, a 3 has been given.
- **Robustness to outliers:** The fact that several subspaces are used reduces the effect that outliers can have. This is why this criterion has been scored with a 4.
- **Parameter Sensitivity:** The algorithms classified within this group are generally sensitive to the tuning parameters that are required. Choosing these parameters, such as the features to be used for the different sets of subdimensions, is a non-intuitive task. Due to all of this, a 2 has been awarded.
- **Complexity and implementation:** This method can be more complex to implement, particularly for high-dimensional data, and might require more computational resources. That is why it is given a 1.

6.1.6 Centroid-based clustering

Centroid-based clustering is a clustering method where a random central position (centroid) is selected for each cluster. Then, the distances between the centroids and the data points are calculated and the data points are assigned to the closest centroid, forming a cluster. Then the centroids positions are updated by recalculating the center of the newly formed clusters. This is done iteratively until the centroids' position

does not change significantly from one iteration to the next. In order to calculate the distance between the centroids and the data points, different distance measures can be used. The most well-known centroid-based clustering algorithms include K-means, K-medoids and Clustering Large Application (CLARA) [22].

- **Handling of irregular cluster shapes:** This method assumes that clusters have a spherical shape. This means that it is not appropriate when dealing with irregular shaped cluster. As this is one of the main disadvantages of the method a 1 has been given.
- **Robustness to outliers:** Outliers can have a big impact on the final result as they can influence the calculation of the centroid. Therefore, a score of 2 has been given.
- **Parameter Sensitivity:** The initial position of the centroids can cause the algorithm to get stuck in a local minimum. Although there are methods to potentially avoid this, it is still an issue and therefore, a 3 has been given.
- **Complexity and implementation:** Out of all the methods, centroid-based clustering is the simplest to implement. As this is the strong point of the method, a 5 has been granted.

An alternative matrix has been constructed in order to sum up the previously given scores (Table 1). These scores, together with the assigned weights, have been used for the calculation of the total score of each method. As can be seen, the highest scored method is centroid-based clustering.

Table 1: Clustering method evaluation matrix

	Clustering methods						
Criteria	Search-based	Graph-theoretic	Density-based	Model-based	Subspace	Centroid-based	**Weights**
Handling irregular cluster	3	5	5	5	3	1	20%
Robustness to outliers	3	3	4	2	4	2	20%
Parameter sensitivity	5	3	2	1	2	3	20%
Complexity and implementation	1	1	2	2	1	5	40%
Score	2.6	2.6	3	2.4	2.2	**3.2**	

6.2 Centroid-based clustering algorithms

Now that the method has been chosen, it is important to select the specific algorithm to be used. Within the centroid-based clustering algorithms there are four that stand out [30]: K-Means, K-Medoids, Clustering for Large Application (CLARA) and Clustering Large Applications based on Randomized Search (CLARANS). In this section, a comparison of these algorithms will be made, based on the following criteria, in order to determine the most appropriate one. Three parameters have been chosen to evaluate the algorithms, two of which were already used in the clustering method analysis.

- **Robustness to outliers:** It is important for the algorithms to be able to deal with data points, known as outliers, that deviate significantly from the rest of the data. If these outliers are not detected and correctly managed, they can change the result of the clustering significantly. Given that this aspect can have a significant impact in the clustering process this criterion has been given a weight of 30% for the score calculating process.
- **Complexity and implementation**: This will depend on two main aspects; the learning curve involved in grasping the theoretical and mathematical concepts behind the functioning of the algorithm and the difficulty involved in the calculation of the initial parameters. The weight of this criterion is 35%.
- **Scalability**: It refers to the algorithm's ability to work with large data sets. Many algorithms are very computationally demanding and therefore do not work well with large data sets. This criterion has been assigned the same weight as the previous one, i.e., 35%.

6.2.1 K-Means

K-Means is a popular centroid-based algorithm developed by James Macqueen [31]. This algorithm requires the number of clusters to be previously specified. Once the number of clusters has been determined, a centroid is randomly initialized for each cluster. Then, as long as a stopping criterion is not met, the following two operations are performed repeatedly. First, the distance from each data points to each centroid is calculated, and then the data points are assigned to the nearest centroid. The data points assigned to the same centroid form what is known as a cluster. Second, the position of the centroids is recalculated by calculating the mean of the data points in each newly formed cluster.

- **Robustness to outliers:** It is not very robust to outliers as these significantly affect the calculation of the centroids, leading to the miscalculation of the clusters. As this is the main disadvantage of the algorithm it has been given a 2.
- **Complexity and implementation:** K-Means is a simple and easy to understand algorithm and in terms of parameter tunning it is also quite simple. Due to this, a 5 has been given.
- **Scalability:** It is a relatively not a very demanding algorithm in terms of computational capacity. It is most appropriate for medium sized datasets and that is why a 3 has been awarded.

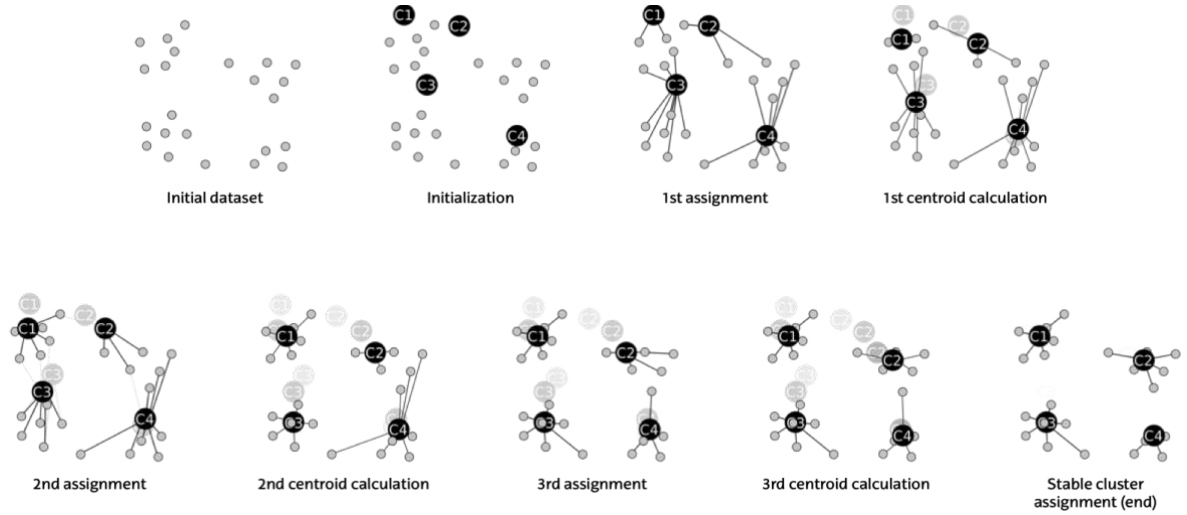

Figure 23: K-Means algorithm [32]

6.2.2 K-Medoids

This model created by Leonard Kaufman, and Peter J. Rousseeuw [33], was developed to improve some of the limitations of K-Means, such as, sensitivity to outliers and presumption of spherical shaped clusters. As with K-Means, in K-Medoids, it is also required for the number of clusters to be defined previously. However, in contrast with K-Means, K-Medoids selects actual data points from the dataset to act as centroids. Once this is done, the distance from each data point to each centroid is calculated, and then each data point is assigned to the nearest cluster. The next step is to calculate the cost of each centroid to its assigned data points. Then, each non-centroid of a cluster is imagined to be the centroid and the cost is recalculated. Finally, the data point with the lowest cost is considered as the new centroid and the process is repeated until a stopping criterion is satisfied.

- **Robustness to outliers:** The fact that data points are used as centroids reduces the algorithms sensibility to outliers. As this is considered a slight improvement when compared to K-Means a 3 has been awarded.
- **Complexity and implementation:** K-Medoids is also a relatively simple and easy algorithm to understand. It is a bit more complex than K-Means to implement as actual data points are used as medoids, that is why a score of 4 has been chosen.
- **Scalability:** It is not the best algorithm for large datasets as the swapping of all the data points of a cluster as centroids requires many calculations. Due to this large number of calculations, a 2 has been given.

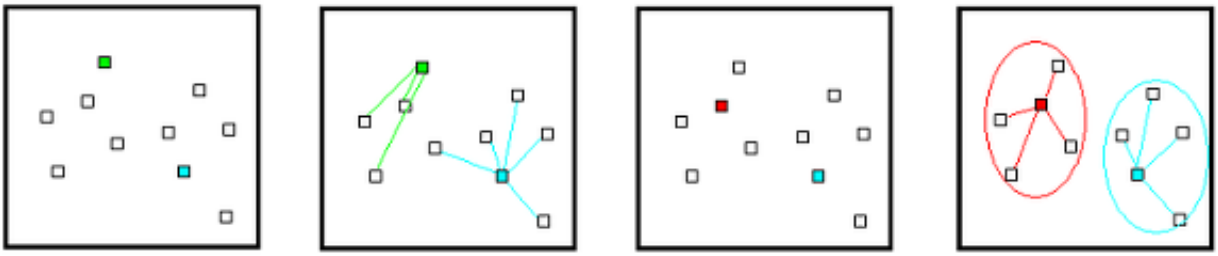

Figure 24: K-Medoids [34]

6.2.3 Clustering for Large Application (CLARA)

Clustering for Large Application, commonly referred to as CLARA, was also developed by Leonard Kaufman, and Peter J. Rousseeuw [35]. It is an extension of their K-Medoids algorithm that tries to solve the problems K-Medoids faces when dealing with large datasets. The CLARA algorithm selects a number of samples from the dataset to create a subset of the data. Then the K-Medoids algorithm is used with the subset and the obtained centroids are applied to the whole dataset for the clustering of all the data points. Then the cost of the clusters is calculated and compared to the cost of the previous iteration. If the cost is lower, then the centroid positions are updated. Then a new subset of the dataset is selected and the process is repeated iteratively until a stopping criterion is satisfied [36].

- **Robustness to outliers:** As in this algorithm internally uses K-Medoids, its sensibility towards outliers is the same. Also, the fact that a subset is selected could introduce bias to the calculations. As the

selected samples could be outliers and therefore not food representatives of the data set. Given this, a lower score, 2, than for the case of K-Medoids is given.

- **Complexity and implementation:** CLARA is also an algorithm that is easy to understand. It is a bit more complex than K-Medoids to implement as subsets must be created. That is why a lower score, 3, has been given.
- **Scalability:** It is very adequate for large datasets as the number of calculations is significantly reduced when compared to K-Medoids. Because of this, a higher score of 4 is given.

6.2.4 Clustering Large Applications based on Randomized Search (CLARANS)

Clustering Large Applications based on Randomized Search (CLARANS) was developed by Ng and Han [37]. It is an extension of CLARA and therefore of K-Medoids. The difference between CLARA and CLARANS is that CLARANS tries to solve CLARA's bias problem produced from sampling. This is why CLARANS does not use sampled subsets. Instead, it employs a dynamic exploration of the dataset that consists of swaps between data points to find cluster medoids. These swaps are performed taking into consideration a optimization function to minimize the clustering cost.

- **Robustness to outliers:** As the K-Medoids method is also used in CLARANS, its sensibility towards outliers is similar. However, CLARANS uses a randomized search strategy which offers a slightly more robust approach than the one proposed in CLARA. That is why it achieves a higher score of 3 points.
- **Complexity and implementation:** CLARANS is also an algorithm that is easy to understand and it is at a par with CLARA in terms of implementation difficulty. Therefore, the same score, 3, is given.
- **Scalability:** It is very adequate for large datasets. As the number of calculations is also significantly reduced when compared to K-Medoids. Because of this, a higher score of 4 is given.

An evaluation matrix has been constructed in order to sum up the previously given scores (Table 2). These scores, together with the assigned weights, have been used for the calculation of the total score of each algorithm. As can be seen, the highest scoring method is K-Means.

Table 2: Centroid-based clustering algorithms evaluation matrix

	Centroid-based clustering methods				
Criteria	K-Means	K-Medoids	CLARA	CLARANS	**Weights**
Robustness to outliers	2	3	2	3	30%
Complexity and implementation	5	4	3	3	35%
Computational cost	3	2	4	4	35%
Score	**3.4**	3	3.05	3.35	

6.3 Programming languages

Over the last decade, ML and data science have gained great importance. That is why, in 2019, GitHub performed a study [38] to determine the most popular programming languages for these fields. Four languages from that list have been selected (Python, C++, JavaScript, and R) and will be evaluated using the following criteria.

- **Ease of use and syntax:** This criterion evaluates if the language is easy to learn and use and if its syntax is clear and easy to comprehend.
- **Library and framework ecosystem:** This criterion evaluates if the programming language has ML libraries available and if these offer a wide range of algorithms and tools that are ready to be implemented.
- **Performance and speed:** This evaluates the efficiency of the algorithm, especially in terms of speed.

6.3.1 Python

Python is an open-source programming language created by Guido van Rossum in 1991. To be more specific, it is a high-level interpreted programming language. A high-level programming language is a language that is written in a similar way to natural speech, meaning that the low-level and hardware

operations are hidden from the user to provide a greater level of simplicity when coding. On the other hand, interpreted programming means that another software known as the interpreter is needed to run the code.

- **Ease of use and syntax:** Due to its high-level nature, Python is one of the easiest languages to start learning from scratch. Additionally, users do not have to face the tedious task of creating methods and algorithms themselves as these are already available in libraries. Python is the most popular language for ML and data science programming, meaning that there is a lot of information and documentation available regarding the use and implementation. This eases the task of learning how to program in Python. Considering all of this, a 5 has been awarded.
- **Library and Framework Ecosystem:** Python offers a large number of libraries with ready-to-implement ML algorithms and tools. Given this, another 5 has been awarded.
- **Performance and Speed:** As Python is an interpreted languages it is not considered as efficient in terms of speed when compared with other compiled languages. As this is the main disadvantage of Python, a 2 has been given.

6.3.2 C++

C++ is a programming language created by Bjarne Stroustrup in 1979. It is often referred to as a middle-level language as it shares some aspects both with low-level and high-level programming languages. Contrary to Python, C++ is a compiled programming language, meaning that the code is directly converted into machine code for the processor to execute.

- **Ease of use and syntax:** The learning curve associated to C++ is much steeper than the previously mentioned Python. C++ is considered to be a verbose language that requires manual memory management and explicit typing, which makes it much harder to learn and understand.
- **Library and framework ecosystem:** This programming language also offers a vast range of libraries. That is why, as in the previous case, a 5 has been awarded.
- **Performance and speed:** As it is a compiled programming language it is more efficient in terms of speed. This is one of the main advantages of C++ for which a score of 5 has been given.

6.3.3 JavaScript

JavaScript was developed in 1995 by Brendan Eich. It is a high-level and interpreted language whose use is very extended nowadays, especially in web development. However, as it offers many high-level tools and libraries it is also used in ML.

- **Ease of use and syntax:** As it is a high-level programming language, JavaScript is an easy language to learn. However, it is considered to be more complex than Python. Therefore, a slightly lower score of 4 has been given.
- **Library and framework ecosystem:** JavaScript also offers a wide selection of libraries. However, when compared to other languages, such as Python, JavaScript is slightly behind. This is the reason for which the score, 4, is also a bit lower.
- **Performance and speed:** As was already mentioned, JavaScript is an interpreted language, meaning that it is not considered as efficient in terms of speed when compared to other compiled languages. It is given the same score as for the case of Python, i.e., 2.

6.3.4 R

R was developed by Robert Gentleman and Ross Ihaka in 1993. It is also a high-level and interpreted language with a more statistical approach, making it very popular in various research fields. It is also used in ML thanks to the user defined packages or libraries it offers.

- **Ease of use and syntax:** Once again, due to it being a high-level programming language, R is considered an easy language to learn. Even with little to no previous programming knowledge, it is still feasible to learn how to program quickly in R. That is why, a 5 has been awarded.
- **Library and framework ecosystem:** As was previously mentioned, R offers the opportunity for users to create and share libraries. Meaning that many academics working on research can share their developed models, pathing the way for other users. This also means that a vast number of libraries are available. However, a lot of the academic applications are used for more statistical purposes and not that much for ML. Therefore, a score of 4 is awarded.
- **Performance and speed:** As explained earlier, being an interpreted language, its efficiency in terms of speed is not as good as in the case of a compiled language. This is why a punctuation of 2 is given.

An evaluation matrix has been constructed in order to sum up the previously given scores (Table 3). These scores, together with the assigned weights, have been used for the calculation of the total score of each programming language. As can be seen, the highest scoring language is Python.

Table 3: Programming-language evaluation matrix

	Programming languages				
Criteria	Python	C++	JavaScript	R	**Weights**
Ease of Use and Syntax	5	2	4	4	40%
Library and Framework Ecosystem	5	5	4	5	40%
Performance and Speed	2	5	2	2	20%
Score	**4.4**	3.8	3.6	4	

7 RISK ANALYSIS

Risks are events that affect project performance in a negative manner. Therefore, performing a risk analysis is one of the crucial steps in project management. It is considered a proactive effort that aims at identifying and managing potential problems that might occur during the implementation of the project. That is, the goal is to determine all the possible risks, reduce their probability of occurring and their impact and setup a contingency plan to manage those events that do materialize.

When developing a risk analysis, the following three steps are performed. The first step is risk identification. In this step the complete project is analyzed in order to determine risk sources. Second is risk assessment. In this phase the likelihood, controllability and impact of the risk event are calculated. The likelihood of each risk is evaluated on a scale from 1 to 5, where 1 is “very unlikely” and 5 is “very likely”. For the impact, a five-level scale is also used, where 1 is “negligible” and 5 is “severe”. Risk level is then calculated by multiplying the likelihood by the impact. This can be seen in the risk matrix (Table 1) where the different risk levels are visualized. The last step consists of developing strategies to prevent risks from materializing and creating a contingency plan for each risk in case it becomes a reality.

Table 4: Risk event evaluation matrix

		Impact				
		1 - Negligible	*2 - Minor*	*3 - Moderate*	*4 - Significant*	*5 - Severe*
Likelihood	*5 - Very likely*	Medium - 5	Medium - 10	High - 15	Very high - 20	Very high - 25
	4 - Likely	Low - 4	Medium - 8	Medium - 12	High - 16	Very high - 20
	3 - Possible	Low - 3	Medium - 6	Medium - 9	Medium - 12	High - 15
	2 - Unlikely	Very low - 2	Low - 4	Medium - 6	Medium - 8	Medium - 10
	1 - Very unlikely	Very low - 1	Very low - 2	Low - 3	Low - 4	Medium - 5

Risk analysis typically includes the following four risk categories: technical, external, internal and project management risks [39]. Technical risks are associated with the complications that appear when executing the different project tasks. External risks include those situations that are out of the project team’s control, while internal risks are those problems that arise within the organization. Lastly, management risks embrace all problems related to project control.

Bearing these four categories in mind, as well as the three steps mentioned previously, the subsequent risks have been identified and a set of preventive measures and contingency plans have been developed.

7.1 Technical risks

7.1.1 Bad data quality

Big Data analysis are highly dependent on the quality of the used data. If there is not a sufficiently large amount of data, the collected data is not accurate enough or the structure of the collected data varies greatly, the resulting analysis will most certainly not be appropriate for drawing conclusions. The reason for this is that, on the one hand, if there is not enough data, trained models will not be general enough to use in other situations with different data. On the other hand, if the data is not accurate enough, that is, if there are values missing or if the collected values are erroneous, the extracted conclusions will be incorrect.

This risk has been given the following punctuation:

- **Likelihood**: Possible (3)
- **Impact**: Severe (5)
- **Result**: High (15)

To avoid this from happening, Cisco probes have been used to collect data at the University of the Basque Country. This way, data integrity is guaranteed, and it is also ensured that all the collected data will have the same structure.

7.1.2 Inaccurate data analysis

When performing a Big Data analysis there are many decisions that the analyst must make. That is, the analyst must use the information obtained as well as his or her own judgement for model selection or parameter tuning amongst others. These decisions can lead to biased results and therefore, to an inaccurate data analysis.

This risk has been given the following punctuation:

- **Likelihood**: Unlikely (2)

- **Impact**: Severe (5)
- **Result**: Medium (10)

One way of reducing this risk is by having a good comprehension of the different models and parameters that must be employed. As well as this, different models can be used to help determine the optimal parameters and prevent the analyst from having to choose them.

7.1.3 Information loss

All projects are exposed to information loss, that is to the loss of data or documentation. The impact of this risk materializing is very high as, depending on the quantity of information lost, it could lead to starting the project from scratch again. Nowadays, as most information is stored on information technology (IT) equipment, information loss is usually caused by hardware or software malfunction.

This risk has been given the following punctuation:

- **Likelihood**: Very unlikely (1)
- **Impact**: Significant (4)
- **Result**: Low (4)

To prevent this from happening, a copy of all information is stored in a different storage unit or most likely in the Cloud. This way if information is lost, there is always a backup copy.

7.2 External risks

7.2.1 Data provisioning interruption

As was previously mentioned, the data used in this project is being collected at the University of the Basque Country under an agreement between the university and the NQaS research group. This data collection is carried out continuously in order to have a sufficiently large amount of data as well as the most up-to-date version of it. As this information is required for the different analysis that are being performed, the interruption of the agreement between both entities would greatly impact the project. This breach of contract could occur if the university felt that the data was not being treated in a way that guaranteed its security or privacy.

This risk has been given the following punctuation:

- **Likelihood**: Very unlikely (1)
- **Impact**: Severe (5)
- **Result**: Medium (5)

To avoid this from happening, the contract between both parties clearly states how the data should be treated. In addition, a person will be assigned to supervise throughout the project's execution that the terms established in the contract are being complied with.

7.2.2 Regulatory changes

The QoXphere model is based on various standards from different regulatory bodies. It could be the case that changes would be made to some of those standards as research regarding these topics is still ongoing. For example, a change in KPI values could be established. However, most of the changes would not impact the functioning of the model.

This risk has been given the following punctuation:

- **Likelihood**: Possible (3)
- **Impact**: Negligible (2)
- **Result**: Medium (6)

To reduce the impact of this type of risk it is sufficient with checking and keeping up to date with the new standards and regulations and making sure that the model continues to comply with them.

7.3 Internal risks

7.3.1 Delays

It is often the case that the initially established deadline of a project is not met. This can sometimes be due to the fact that a poor estimation of task duration is made or that a set of unexpected events occur.

This risk has been given the following punctuation:

- **Likelihood**: Likely (4)
- **Impact**: Significant (4)
- **Result**: High (16)

To ensure that the time estimations made are as accurate as possible, it is often a good idea to gather information from similar projects that were previously executed. Another measure that is often taken is to allocate extra time to tasks in order to ensure that there is some margin if an unexpected event occurs. However, these are all supplementary measures that can be put into practice. The key to avoid delays is to have a well-though-out and detailed planning as the one available in section 10.2.

7.4 Project management risks

7.4.1 Lack of project control

Many projects fail due to a lack of project control. Control is based on contrasting performance and plan to detect alterations and find a solution to get the project back on track. Lack of control can have many consequences such as failure to meet outlined scope, cost, schedule or quality expectations which can result in project failure.

This risk has been given the following punctuation:

- **Likelihood**: Very unlikely (1)
- **Impact**: Moderate (3)
- **Result**: Low (3)

The best way to ensure that control is implemented is to set a baseline plan in terms of scope, cost, schedule and quality. That is determining the objectives, the budget, the timeline and the requirements of the project. Then, while the project is underway, the progress and performance must be measured and compared against the previously outlined plan. Lastly, if it is detected that the project is not meeting some of the requirements, corrective measures should be taken.

In Table 2, the result of the risk analysis is shown. That is, each of the previously mentioned risks are displayed according to the risk level they were assigned:

Table 5: Result of risk analysis

		Impact				
		1 - Negligible	*2 - Minor*	*3 - Moderate*	*4 - Significant*	*5 - Severe*
Likelihood	*5 - Very likely*					
	4 - Likely				Delays	
	3 - Possible		Regulatory changes			Bad data quality
	2 - Unlikely					Inaccurate data analysis
	1 - Very unlikely			Lack of project control	Information loss	Data provisioning interruption

8 METHODOLOGY

This section provides a description of the methodology used for the data analysis. Typically, data analysis studies follow five main phases:

1. **Problem understanding**: First, the context of the problem must be understood in order to set out the goals and objectives of the study.
2. **Data collection**: Second, the data to be used in the analysis must be collected. This might mean contacting different entities to collect data from them and requesting consent for privacy issues.
3. **Data preparation:** Once the data has been collected, it must be processed in order to adapt it to the specific needs of the study. This could include selecting the appropriate features, eliminating outliers and normalizing the data, amongst other steps that must be taken.
4. **Data analysis**: After the data has been prepared and preprocessed, the ML techniques and algorithms can be applied.
5. **Result analysis**: Finally, once the algorithms and techniques have been applied and results have been obtained, these are analyzed in order to extract further conclusions.

8.1 Problem understanding

As with all data analysis, the first step is to perform thorough research regarding the main topics of the investigation. This will provide good background knowledge that will offer a good context of the problem and enable a correct study of the data.

This project is focused on validating a QoS framework, known as QoXphere. Therefore, it was important on the one hand to understand the most used QoS terminology and frameworks. As well as this, a study of the main QoS and QoE factors was performed, as these would be determining in identifying the different user types. Then, the intricacies of the QoXphere framework had to be understood to have a correct approach towards the validation process. Additionally, as this data analysis uses data from a Wi-Fi network, it was considered important to perform a study of the Wi-Fi standard, also known as IEEE 802.11. This research involved understanding the architecture of Wi-Fi, the functioning of the MAC layer of this standard and the QoS mechanisms that this technology uses. Lastly, as mentioned in previous sections, one of the goals of this project is to validate the Context extraction step of the QoXphere framework. This step suggests using ML in

order to identify different users and scenarios based on their QoS requirements. Therefore, it was also considered relevant to study the different ML fields to determine which one the present study most appropriately fell into. By doing so, it was identified that the study at hand was an unsupervised learning or clustering problem.

Once this background research had been carried out, more research had to be conducted to perform the analysis of alternatives. Firstly, the type of clustering method had to be determined. Once it was decided that centroid-based clustering was the method that would best suit the given problem, a study of the main centroid-based clustering algorithms was conducted to determine the most appropriate one. Lastly, the programming language to be used had to be decided upon, which is why a study regarding the most relevant programming languages in ML was performed.

8.2 Data collection

As was mentioned previously, the data used in this study was collected from the University of the Basque Country. This data was collected from the three different campuses that make up this university, i.e., the Campus of Araba, the Campus of Biscay and the Campus of Gipuzkoa (Figure 25). As can be seen, some of these campuses have multiple locations. That is the case of the Campus of Biscay, which can be found in the localities of Bilbao, Portugalete and Leioa-Erandio, and of the Campus of Gipuzkoa, which is divided between Eibar and Ibaieta (Donostia-San Sebastian). On the contrary, the Campus of Araba is entirely located in Vitoria-Gasteiz. In order to have a large and varied dataset, data samples were collected in the university buildings of each one of these localities.

In addition to diversity in the location of data collection, there is also diversity in the period of data collection. This data collection process was carried out at different points in the year so that data was collected at different stages of the academic year (start of the academic year, middle of the school term, exam period and holiday period). This is interesting and important, as it may reflect different user habits in terms of network use depending on the stage of the academic year.

Figure 25: Location of UPV/EHU centers [40]

Lastly, as was mentioned in the analysis of risks, the data used in this project is being collected at the University of the Basque Country under an agreement between the university and the NQaS research group. One of the conditions that allow the data collection is that the data is treated secure and privately, so that no personal information of the users is shared or made public. In order to do so, all the data was anonymized by deleting the users' username information in a way that the information could not be traced back to them.

8.3 Data preparation

The data preparation step consists of selecting the most relevant data to use in the data analysis step. As many datasets were available for this project, the first step was to decide which dataset was the most appropriate to use. It was finally decided to use the `Clientes_Eduroam_20221027_235912_075` dataset for the following reasons:

1. **Sufficient data**: A good dataset is one that contains a sufficiently large number of instances. The reason behind this is that the larger the dataset is, the easier it will be to detect more subtle patterns that would probably go unnoticed in a smaller dataset. As can be seen in Table 6, the selected dataset has a total of 26104 instances, which is a sufficiently large amount.

Table 6: Dataset instance quantity

Dataset	Instances
Sessiones_cliente_20220610_130640_730	341401
WIFI_clientes_Junio2022_TablaDinamica_anonimo	69795
Clientes_Eduroam_20221027_235912_075	26104
Clientes_Eduroam_20220908_235909_338	14928
Clientes_Eduroam_20220720_235906_333	12149
Client_Session_Ingenieria_20230511	3856
Datos_clientes_20230109	1622
ExportDevice_EHU_20230109	1622
Client_Session_Educacion_20230511	1327
ExportDevice_BIZ_Leioa_20230109	159
ExportDevice_BIZ_Bilbao_20230109	122
Clientes_U024650_Magisterio_20221027_235902_627	16
Clientes_U022594yU033595_IngenieroP4_20221027_235916_236	3

2. **Good features**: It is important that the selected dataset contains features that are useful in the data analysis. For this, a domain expert's knowledge is usually required. It was considered that the selected dataset offered a good set of features as it contained categorical features offering information on user location and connection time, as well as non-categorical features providing information on the amount of data transferred and speed.
3. **Representative data**: The chosen dataset contains data collected on October 27th, 2022. This day happens to be a Thursday, and therefore a school day, of the first semester of the school year. This dataset was considered more useful than other datasets collected on months spanning May to August, where students have already finished their courses, or those collected at the beginning of January or September, where students have just returned from holidays and have not had time to pick up their usual study habits.

The next step was to organize the data in the dataset. As was mentioned in the data collection step, this data was gathered in a number of different campuses. However, due to the familiarity of the researchers with the Campus of Biscay, it was decided to only use the data selected in this campus, as it still offered a

large amount of varied information. This was done by means of the `Map Location` feature of the dataset. To be more specific it was decided to use the data from four different locations: Faculty of Science, Faculty of Engineering, Faculty of Business and Economics and the university Library. The selection of these locations was not arbitrary, it was done according to the following criteria:

1. **Sufficient data:** This first criteria, is common to the one of the previous selection process. It is important to have a large dataset to find underlying relations. shows the top 6 faculties in terms of instance number and as can be seen all four selected faculties appear in the table.

Table 7: Faculty instances quantity

Faculty	Instances
Faculty of Science (Leioa)	3000
Faculty of Medicine (Leioa)	2360
Faculty of Engineering (ETSI building, Bilbao)	2184
Faculty of Engineering (EUITIMOP building, Bilbao)	2150
Faculty of Business and Economics (Sarriko, Bilbao)	1283
Library (Leioa)	1198

2. **Different study domains**: In order to have a heterogeneous dataset, it was decided to select faculties from different study domains. The reason behind this is that, theoretically, this could influence the user profile, as in one study domain a higher use of the network might be made. That is why, to eliminate similar profiles, the Faculty of Medicine was discarded and the Faculty of Science was chosen for the study, as it had a larger number of instances. The similarities mentioned refer to related study areas (biology, chemistry, etc.) and therefore similar study spaces, such as laboratories. The same happens with the ETSI and EUITIMOP buildings of the Faculty of Engineering. In this case, the ETSI was also selected due to its higher number of instances. Lastly, the Faculty of Business and Economics and the Library were chosen, due, once again, to the different user profile they could offer.

It was also decided to only use the data corresponding to the time gap between 7 am and 6 pm. The reason for this is that these are the busiest campus times and when most students and personnel are there. Choosing

to use later hours could bias the results and it has therefore been considered best to omit them. This is done by using the `Association Time` feature.

The last step in the data preparation phase is to pre-process the data for the application of the ML techniques. In this case, the pre-processing consisted of various steps:

1. **Eliminating constant columns**: Constant columns are those columns whose value is the same for each instance. As the value is the same for every instance, no additional information is offered and therefore they can be eliminated from the dataset. In this case, the constant columns were `Device Name`, `SSID`, `Profile`, `VLAN ID`, `Policy Type`, `Client Type`, `Device IP`, `Controller Port`, `Anchor Controller`, `Authentication`, `Encryption Cipher`, `Authentication Algorithm`, `Web Security`, `RTS Retries` and `Mobility Status`. This meant that the number of features was reduced from 44 to 29.
2. **Eliminating columns with NaN values**: A NaN (Not a Number) value indicates that no data was collected for the given feature and instance. When all the instances of a feature are NaNs, the feature is not offering additional information and it can therefore be removed. The features with NaN values are `Host Name`, `Speed` and `Network Access Id`. This meant that the number of features was reduced from 29 to 26.
3. **Eliminating outliers**: Outliers are data points that deviate significantly from the rest of the data. This is an important step, as outliers can make cluster centers get stuck on local minima.
4. **Eliminating Root Area instances**: In the dataset there where some instances whose `Map Location` value did not follow the format established for this feature, which was `Campus > Faculty > Floor`. Instead, the value was Root Area. By observing the `AP Name` parameter, it was seen that all these instances had APs whose name started with AP-Derecho, meaning that they had been collected at the Faculty of Law. As this could bias the result, these instances were eliminated from the dataset.

To sum up, at the end of the data preparation step, four groups or scenarios were extracted from the dataset: Faculty of Science, Faculty of Engineering, Faculty of Business and Economics and Library. Each of these contained data belonging to the previously mentioned faculties, collected between 7 am and 8 pm, with all features containing constant and NaN values and outlier instances removed.

8.4 Data analysis

Once the data had been correctly prepared and the non-useful data had been filtered out the data analysis could be started. This analysis was composed of three main steps: feature selection, parameter tunning and algorithm application.

8.4.1 Feature selection

As was explained before, the data used in this study was collected at the University of the Basque Country. To be more specific, it was directly collected from the APs set up in these campuses. When proceeding with feature selection, it is important to first understand the meaning of each of the variables. That is why, a summary of the features' meaning and relevance was made:

- **Client Username:** It is the identification used by a person with access to the eduroam network and is therefore useful for identifying each user in the network. It is important to note that a username may appear multiple times simultaneously as each user can have more than one device connected to the eduroam network.
- **Client IP Address:** It is the IP address of the user (device) of the eduroam network. An IP address is a unique address that identifies a device on the Internet or on a local network. All clients receive a private IP address assigned by the network by means of the Dynamic Host Configuration Protocol (DHCP). This address is represented by 4 decimal numbers separated by dots, whose values range from 0 to 255 and is useful when identifying the different users (devices) on the network.
- **Client MAC Address:** It is the MAC address of the user (device) of the eduroam network. A MAC address is a unique identifier of a computer's device or network interface, which is assigned by the manufacturer of the device. It is represented as a series of 12 hexadecimal digits grouped in pairs and is very useful as it allows each device or network interface to be identified unequivocally.
- **Association Time:** It indicates the date and time at which the user (device) connected to the AP. The format used is: `Abbreviated Day | Abbreviated Month | DD | hh:mm:ss CEST YYYY`. An example of this would be: `Wed Jul 20 09:18:28 CEST 2022`. This parameter is useful for studying the patterns between the time of the day and the user's connections.
- **Vendor:** This parameter offers information regarding the manufacturer of the device. This can be useful when studying the type of device the users of the network employ.

- **AP Name:** It is the name given by the network manager in order to unequivocally identify the APs within the network's centralized management system. A good AP naming system is crucial for correctly identifying each AP and for the correct management of the network.
- **Radio Type:** This parameter specifies the different IEEE 802.11 standards the AP supports. It is an interesting parameter to consider as it can affect other transmission factors such as throughput.
- **Device name:** It indicates the name of the controller, whose function is to manage and coordinate the functioning of the network.
- **Map Location:** This variable offers information as to where the AP is physically located. That is, the campus, faculty and floor where it is situated. The format it uses is the following: `Campus > Faculty > Floor`. An example of this is `Alava > Filologia > Planta Primera`. This variable is very useful for sorting information according to the different buildings and floors and to study the different user behavior from one campus to another.
- **SSID:** The Service Set Identifier (SSID) is the technical term for a Wi-Fi name. It is used so that a network can be distinguished from the other networks in the area. In this case the network that will be studied is the eduroam network. Therefore, this variable is important as it will allow to filter out all data not belonging to the eduroam network.
- **Profile:** This parameter indicates the profile assigned to the AP. A profile is a set of configurations and policies that are applied to APs to manage their functioning within the network. This is practical in order to optimize the network's functioning.
- **VLAN ID:** Virtual local area network **(**VLAN) is a networking technology that allows the creation of independent logical networks within the same physical network. Since there can be several VLANs on the same physical network, a system of tags (VLAN IDs) is used to indicate the VLAN of which a network device should be a member. This column shows the tag of the VLAN to which each user (device) belongs to. This information can be interesting to analyze when studying network segmentation and routing.
- **Protocol:** It indicates the IEEE 802.11 standard and the band (2.4 GHz or 5 GHz) that has been used for the communication. As with Radio Type, this is also an interesting parameter as it can influence other transmission factors, such as the previously mentioned throughput.
- **Session Duration:** The session duration indicates the length of time the user (device) has been connected to the specific AP mentioned in AP Name. It is a good variable for studying user types

depending on how long they are connected to the AP. The format used for indicating the session duration is `X hrs X min X sec`. For example, `1 hrs 40 min 20 sec`.

- **Policy Type:** In order to secure Wi-Fi networks, the Wi-Fi Alliance developed three different types of security certification programs known as Wi-Fi Protected Access (WPA). The three types are WPA, WPA2 and WPA3. This indicates the WPA that is being used.
- **Avg. Session Throughput (Kbps):** It is the mean transmission rate expressed in kilobytes per second. It is a very useful in this context of QoS.
- **Host Name:** It is a parameter used for human-readable identification of the AP. Although similar to `AP Name`, the difference between these two parameters is that `AP Name` is usually named following some type of convention established by the management system, while `Host Name` is normally a meaningful name that helps humans easily identify the AP.
- **Client Type:** This parameter is used to categorize clients according to their characteristics and roles. This is practical in order to configure the network to act in a specific manner with each of the client types.
- **Speed :** It indicates the maximum velocity at which a client can communicate with the AP. It is useful as it is an important parameter when studying QoS.
- **AP MAC Address:** It is the MAC address of the AP the user (device) is connected to. As with the Client MAC Address, it allows each AP to be identified unequivocally.
- **AP IP Address:** It is the IP address of the AP the user (device) is connected to. Like in the case of the Client IP Address, it is useful when identifying the APs on the network.
- **Device IP:** It is the IP address of the network's controller. This information is essential for a good functioning of the network, as the controller is in charge of the network's control and coordination functions.
- **Controller Port:** It indicates the port used by the network controller to communicate with the APs**.** It is useful for network control and management operations.
- **Anchor Controller:** It indicates the Anchor controller the client is associated to. An anchor controller is a controller that manages client sessions in scenarios that involve guest access and mobility between network segments, such as roaming between APs. It is also useful for network control and management operations.

- **Association ID:** It is used to unequivocally identify each association between a client and the AP. It is important for the AP's traffic and QoS management.
- **Disassociation Time:** It indicates the date and time at which the user (device) disconnected from the AP. The format used is: `Abbreviated Day | Abbreviated Month | DD | hh:mm:ss CEST YYYY`. An example of this would be: `Wed Jul 20 15:09:53 CEST 2022`. This parameter is useful for studying the patterns between the time of the day and the user's disconnections.
- **Authentication:** It indicates if authentication mechanisms were employed in the association process between the client and the AP. It is a useful when studying security aspects of the network.
- **Encryption Cipher:** It specifies the cipher suite used to secure the communications between the client and the AP. As with the previous parameter, it is useful from a security point of view.
- **EAP Type:** Extensible Authentication Protocol (EAP) is a framework that offers a safe way to exchange identifying information for network authentication processes. There are several types of EAP methods and this variable is used to specify the method used in each connection. It is useful when studying security aspects of the network.
- **Authentication Algorithm:** It specifies the algorithm employed for authenticating the clients trying to access the network. It is important for securing the network correctly.
- **Web Security:** This parameter indicates whether or not a set of configurations for the security of web-based communications between client and AP has been enabled. Once again, it is a security-related parameter.
- **Bytes Sent:** It is used to quantify the number of bytes sent by the user (device) in the present connection.
- **Bytes Received:** It is used to quantify the number of bytes received by the user (device) in the present connection.
- **Packets Sent:** Similar to the bytes sent variable, instead of quantifying the received bytes, this variable counts the number of packets received by the user (device).
- **Packets Received:** Like the bytes received variable, instead of quantifying the sent bytes, this variable counts the number of packets sent by the user (device).

- **SNR (dB):** The Signal-to-Noise Ratio (SNR) indicates quality and strength of the signal with respect to the background noise. It is expressed in decibels and is an important parameter as it can have great effect on the QoS.
- **RSSI (dBm):** The Received Signal Strength Indicator (RSSI) expresses the power of the signal upon reception at the AP. It is expressed in decibels-milliwatts (dBm) and is an important parameter to study as it can influence the perceived QoS.
- **Status**: It indicates whether the client is associated or disassociated from the AP. It is important for dealing with connectivity issues and client movement tracking, as well as for studying the network utilization.
- **Reason**: It indicates the reason behind the disassociation of the user (device) from the AP. There are three possible reasons: `new association detected`, `no longer seen from controller` or `disassociation detected`.
- **Data Retries:** It indicates the number of times a packet has been transmitted unsuccessfully. It is an important parameter for QoS as it shows the quality or reliability of the communication.
- **RTS Retries:** The Request to Send (RTS) is a mechanism used in the CSMA/CA protocol to manage channel access. This parameter indicates the number of times a RTS frame has been transmitted unsuccessfully. As with the previous parameter it is important I for QoS aspects.
- **Mobility Status:** It indicates if the client is capable of maintaining connectivity while moving within the network, that it, when switching from one AP to another. This is another important parameter for QoS, as seamless mobility will provide a better service experience for the client.
- **Network Access Id:** It is a standard format used to identify the clients that are trying to access a network. It is important for authentication and roaming operations.
- **Session ID:** It is used to reference a specific connection between client and AP. It is important in terms of network management.

Now that the features' meaning is understood, a reasoned selection of them can be made. In order to apply the ML algorithms, all the data must have numerical values. However, many features have non-numerical values. Some examples are the `Map Location` feature, used for selecting the data belonging to a certain campus, or the `Association Time` feature, used for only selecting the instances within the specified time gap. However, there are many more categorical features that do not offer additional

information and can therefore be removed. By doing this, the feature number is reduced from 26 to 9. These 9 features are: `Session Duration`, `Avg. Session Throughput (Kbps)`, `Bytes Sent`, `Bytes Received`, `Packets Sent`, `Packets Received`, `SNR (dB)`, `RSSI (dBm)` and `Data Retries`.

To better understand the relations between these nine features, a correlation matrix was computed (Figure 26). Correlation matrixes are square matrixes used to comprehend the relations and dependencies between a set of variables. These matrixes are made up of cells that indicate the correlation coefficient between two variables. The values of these cells can range from -1 to 1.

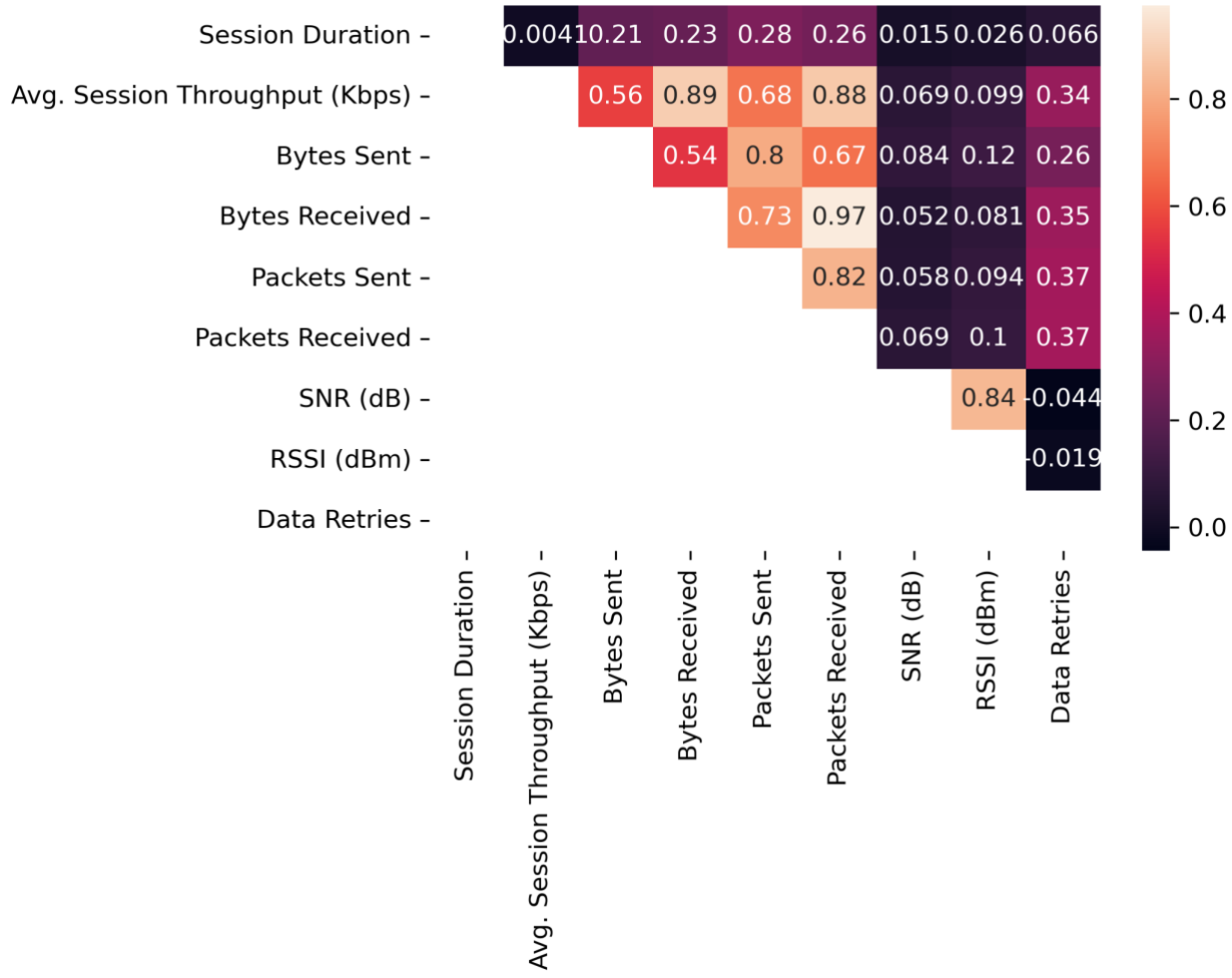

Figure 26: Correlation matrix

- **No correlation**: A value of zero implies that there is no linear dependance between the two variables.

- **Positive correlation**: Positive values indicate that as one variable increases, the other tends to increase as well. Values close to zero indicate low correlation, while those close to one indicate high correlation.
- **Negative correlation**: Negative values indicate that as one variable increases, the other tends to decrease. As with positive correlation, values close to zero indicate low correlation, while those close to minus one indicate high correlation.

To determine which features were most appropriate for the clustering, the next two steps, parameter tunning and algorithm application, were applied using different feature combinations. By doing this, it was observed that clusters were distinguished when using the `Packets Sent – Packets Received` feature combination. Therefore, in the next two sections the results and graphics obtained when using these parameters will be shown.

8.4.2 Parameter tunning

It was previously explained that when applying the K-Means clustering algorithm, the number of clusters must be specified. To determine what number of clusters offers the best results, different parameter tunning mechanisms can be used. One of the most widely used mechanisms is the elbow method. This method consists of applying the K-Means algorithm several times but using a different number of clusters in each iteration. Each time the algorithm is applied with a different number of clusters, the within-cluster sum of squares (WCSS) is calculated. The WCSS is used to measure the proximity between data points in a cluster. The lower the WCSS value is, the tighter the clusters will be, and therefore it could be thought that the better the clustering will be. However, choosing a very low value of WCSS, and hence a high value of k, is not optimal either. The goal is to obtain a good tradeoff between the value of k and WCSS. This is what is known as looking for the “elbow” of the curve, hence the name of the method. The elbow point is the spot of the graph at which the rate of decrease in WCSS changes significantly from one iteration to another.

This method was applied to each of the previously mentioned scenarios. By looking at the figures in Figure 27 it can be observed that the most abrupt changes in the curve occur for k=2. Therefore, this was the value that was finally used.

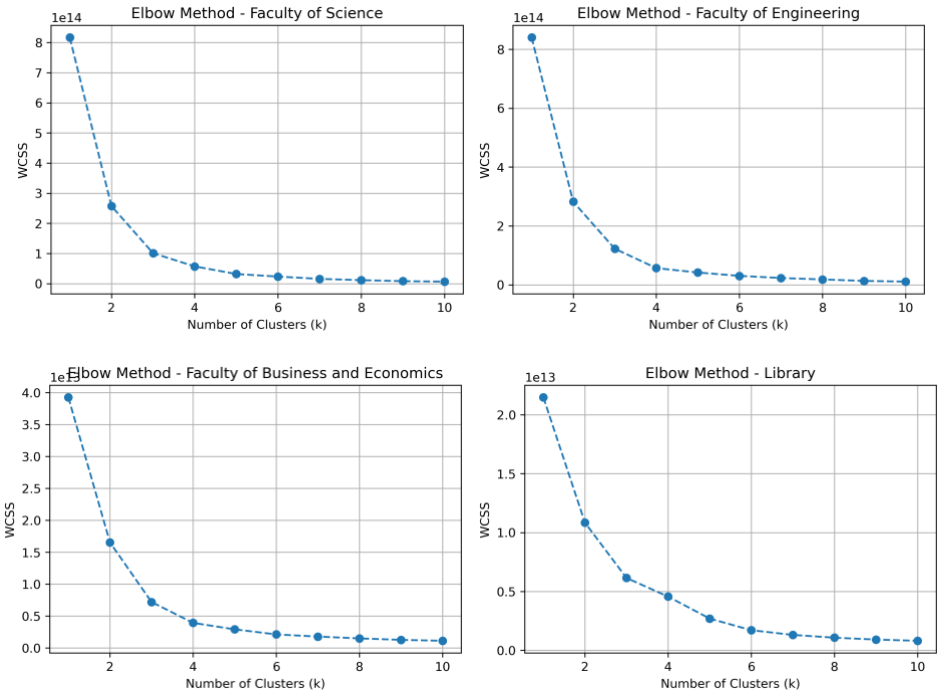

Figure 27: Elbow method results

8.4.3 Algorithm application

Now that the features were selected and the parameters tunned, the K-Means algorithm could be applied. As was explained in section 6.2.1, K-Means is a centroid-based algorithm. This means that a set of random central positions (centroids) are selected for each cluster and that the data points are assigned to the cluster that is closest according to a distance measure.

As was explained earlier the algorithm was applied to several combinations of features. However, in this section, in order to reduce the number of graphics, only the results of the `Packets Sent - Packets Received` feature combination are showed, as this is the combination that produced significant results.

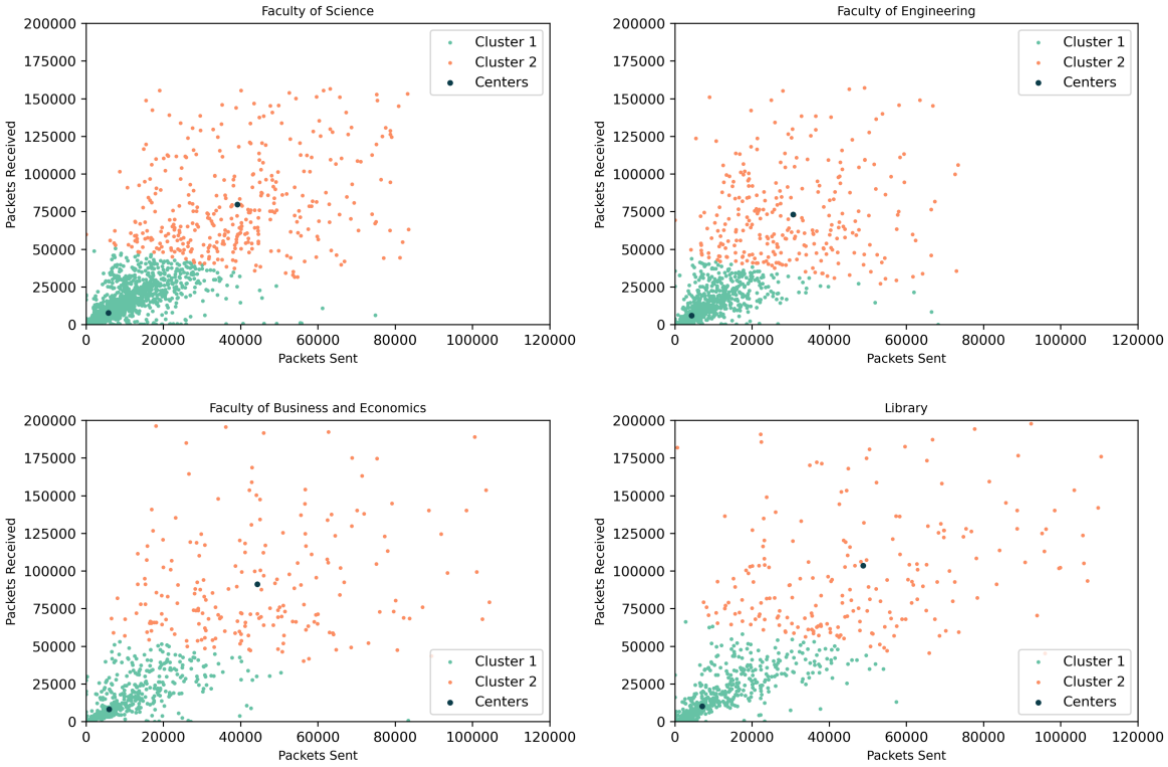

Figure 28: K-Means results

As can be seen in Figure 28, two clear clusters are obtained when applying the algorithm. Cluster 1 gathers those instances with low packet reception and transmission and Cluster 2, gathers those with higher values. These two groups are repeatedly found in all four scenarios but by observing the graph it can also be seen that they present some differences. These will be studied in the following section in order to try grasp a better understanding of these user profiles.

9 RESULTS AND DISCUSSIONS

At this point, the data has been collected, the features selected, the parameters defined and clusters have been obtained. The next step is to try different approaches to analyze the data in order to assign user profiles to each one of them.

9.1 Device analysis

The first analysis is known as device profiling, a process that consists of identifying the most commonly used devices in Cluster 1 and Cluster 2. Possessing knowledge of the types of devices connected to a network is beneficial in terms of management and security. In this case, by identifying the type of devices the clients use, a better understanding of their network usage can be achieved, which will aid in the task of defining the users' profiles.

In order to perform the device profiling, the `Vendor` feature will be used. Although this feature does not directly indicate if the client's device is a mobile phone or laptop, it does offer information on the manufacturer of the client device's Wireless Network Interface Controller (WNIC). A WNIC is a hardware component that enables the connection of a device to a WLAN and an AP is capable of determining the manufacturer by inspecting the WNIC's MAC address. As was explained in section 8.4.1, MAC addresses are made up of a series of 12 hexadecimal digits grouped in pairs. As these addresses must be unique, a certain convention is followed when assigning the MAC addresses to the components. The convention is regulated by the IEEE who is in charge of assigning a certain combination of the most significant hexadecimal pairs, that is the leftmost pairs, of the MAC address to each manufacturer. Depending on the manufacturer's needs, the IEEE assigns a different number of pairs or multiple different combinations of pairs. Then, the rest of the numbers are assigned by the manufacturer. Therefore, as the IEEE keeps a list indicating the pair combinations assigned to each manufacturer, the manufacturer of a WNIC can easily be identified by looking at its MAC address and the aforementioned list. As was stated before, even though this feature does not give information about the type of device the client is using, by knowing the manufacturer of the WNIC assumptions of the equipment type can be made, as some manufacturers specialize on a certain type of devices. It is also important to keep in mind that the manufacturer of the WNIC, is not necessarily the same as the manufacturer of the device. For example, a laptop manufactured by Dell can use a WNIC or other network components manufactured by Intel.

When performing the analysis, the most common values of the `Vendor` feature were calculated for each of the clusters in all four scenarios. Then those manufacturers for which more than ten devices were counted were plotted in the pie charts. The values in these pie charts were then expressed as percentages in order to be able to make good comparisons between them. Figure 30 shows the results obtained for the two clusters in the four contexts, where the top row graphs correspond to Cluster 1 and the bottom row ones to Cluster 2. The first, second, third and fourth columns correspond to the Faculty of Science, Faculty of Engineering, Faculty of Business and Economics and Library namely.

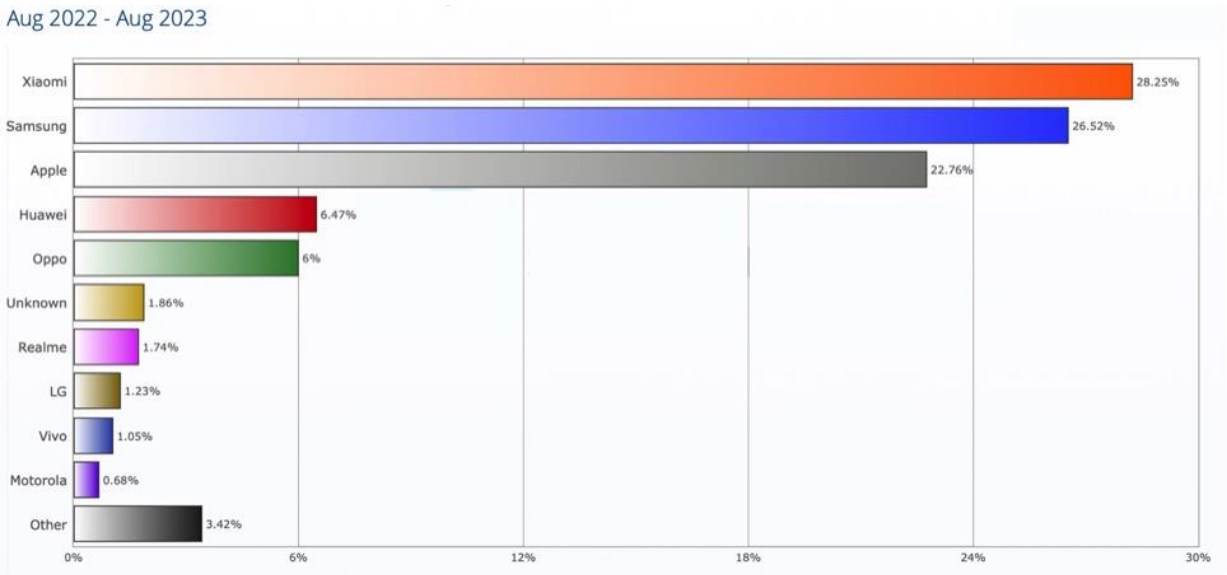

Figure 29: Mobile vendor market share (Spain) [41]

The graphics on the first row, that is, the graphics belonging to Cluster 1 all include Intel, Apple and LiteOn values. Additionally, three out of four of the graphs contain Huawei, Xiaomi, AzureWave and Samsung values. Those manufacturers that are common for all four scenarios, that is, Intel, Apple and LiteOn, do not offer a lot of information about the devices used, as all three produce WNICs that are installed both in mobile phones and laptops. However, in contrast, some of the other manufacturers mentioned, such as Huawei, Xiaomi, and Samsung, specialize on mobile phones. In fact, as can be seen in Figure 29, these manufacturers are some of the leading mobile phone manufacturers in Spain. Therefore, it could be assumed with a significant degree of certainty that the majority of the values represent mobile phone devices. On the other hand, when observing the second row which contains the graphics from Cluster 2, a significant difference is noticed. While the first row contains a wide variety of manufacturers, the second row only has one, two or

at a maximum four. Here, the main manufacturers are Intel, Apple, LiteOn and AzureWave. From these results the following conclusions were made and two possible user profiles were defined:

- **Passive users:** Users connected to the network but that do not make an active use of it.

The users classified as Passive users are those whose `Packets Sent` and `Packets Received` values are relatively low. That is, users belonging to Cluster 1. It was previously explained that after performing the device analysis, a high number of mobile phones was identified amongst the devices of these users. This is coherent with the proposed user profile, as the network use on mobile phones is not expected to be very high in an academic context. Normally, in this type of environment, the more tedious study tasks are carried out with a computer. The reason behind this being that computers are better equipped for the different tasks students have to perform, such as research, assignment composition or use of specific software. However, it must be pointed out too that some of the devices of this user profile can also be computers. These could belong to students who are in class and use their laptop to view classroom slides or to take notes, and who therefore do not use the network actively.

- **Active users:** Users connected to the network who do make an active use of it.

The users classified as Active users are those whose `Packets Sent` and `Packets Received` values are relatively high. That is, users belonging to Cluster 2. In this user type, none of the previously mentioned mobile phone manufacturers were observed. This, once again, supports the proposed profiles as the more network-active users in an academic scenario are expected to work on their computer.

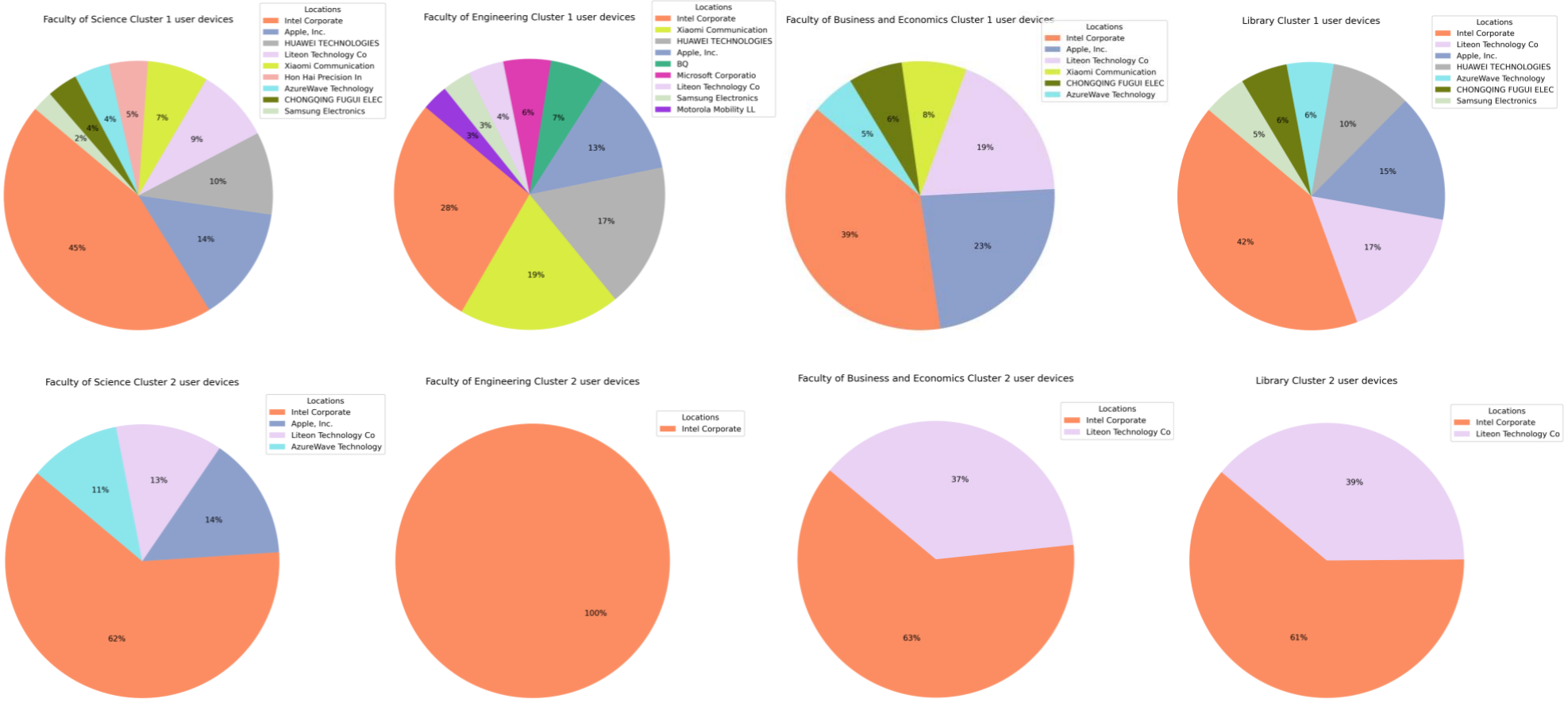

Figure 30: Device analysis results

9.2 Location analysis

The second analysis consists of identifying the spaces that Cluster 1 and Cluster 2 users are more prone to occupy within the university facilities. Possessing knowledge of the locations of the users within a network is beneficial in terms of network dimensioning and management. In this case, by identifying the whereabouts of the clients and understanding the environment they are surrounded by a better understanding of their network usage can be achieved. This in turn, will facilitate the identification of the user profiles.

To perform the location analysis, the `Map Location` feature will be used. As was explained in section 8.4.1, `Map Location` is a variable that offers information as to where the AP is physically located. That is, the campus, faculty and floor where it is situated. This will be used as it offers intuitive information of the users' position. When performing the analysis with the `Map Location` feature, the most common values of this variable were calculated for each of the clusters in all four scenarios. The results were then plotted in bar charts, where the results of both clusters appeared side by side. Figure 31 shows the results obtained in the four contexts, where the first, second, third and fourth rows correspond to the Faculty of Science, Faculty of Engineering, Faculty of Business and Economics and Library namely. The left column shows the number of users for each space of the building and the right column indicates the same values but in percentages. It was decided to plot both the count and the percentages as the first helps visualize the difference between the number of Cluster 1 and Cluster 2 users and the second gives a better understanding of the distribution of these user types in the different spaces of the building.

The first thing that is noticed when observing the graphs on the left is that in all four cases the number of Cluster 1 users is much higher. This makes sense, as each Cluster 2 student, that is, each student that is intensely using the network most probably on a laptop, is likely to have a mobile phone device connected to the network that he or she is making little to no use of. This mobile phone device will count as a Cluster 1 user, meaning that for each Cluster 2 user there is probably also a Cluster 1 user. In addition to this, all other students that are either in a classroom where they are not actively using the network, in a hallway, in a cafeteria or in some other non-working environment must also be considered. These students will significantly increase the number of Cluster 1 users.

For the study of the distribution of the user types in the different spaces of each building, that is for the study of the graphs of the right column, the results of each scenario must be studied independently, as the disposition of the buildings varies from one to another.

- **Faculty of Science**: There is no peak in the distribution of students in Cluster 2, as is seen in the case of the Faculty of Engineering and the Faculty of Economics and Business. However, the basement floors (S1 and S2) have significantly lower values than the rest of the floors.
- **Faculty of Engineering**: In this case the biggest percentage of the Cluster 2 users are on the ground floor (Planta Baja)
- **Faculty of Business and Economics:** Most of the Cluster 2 users are found in the basement (Sótano A).
- **Library:** As for the library, the situation is a bit different. There is no peak in the distribution of Cluster 2 students as was the case for the Faculty of Engineering and the Faculty of Business and Economics. However, floors 2, 3 and 6 have significantly lower values than the rest of the floors.

After analyzing the areas that contain a larger percentage of Cluster 2 users, it has been discovered that most of these areas include spaces dedicated to studying. For the Faculty of Engineering, the area where most Cluster 2 students were found, i.e. the ground floor (Planta Baja), is where the two study rooms of the faculty are located. The same happens for the Faculty of Business and Economics. The two study rooms are found in the basement, which happens to be the area where most Cluster 2 users are gathered. In the Library, the same phenomenon is repeated. Those floors where no study areas are found, floors 2 (reception), 3 (newspaper archive) and 6 (this floor is not open for students), are the ones that show a lower percentage of Cluster 2 users. The only exception is the case of the Faculty of Science, where there is no specific are for the students to study which explains that no peak is observed amongst the Cluster 2 percentage data. Therefore, it can be concluded that the results obtained reinforce the proposed user profiles. Cluster 2 users, that is, active users, will be found in areas that are dedicated to studying. The reason being that these areas are specially designated for students to carry out their study activities which, in some cases, require a higher use of the network. As for Cluster 1 users, that is passive users, these can be found in different places of the building. Some will be roaming from one place of the building to another while making barely no network use, others will be found in study areas, either using notes for studying and therefore making no use of the network or acting at the same time as passive and active users (active, while using their computer and network for study and research activities and passive while their mobile device is simultaneously connected to the network).

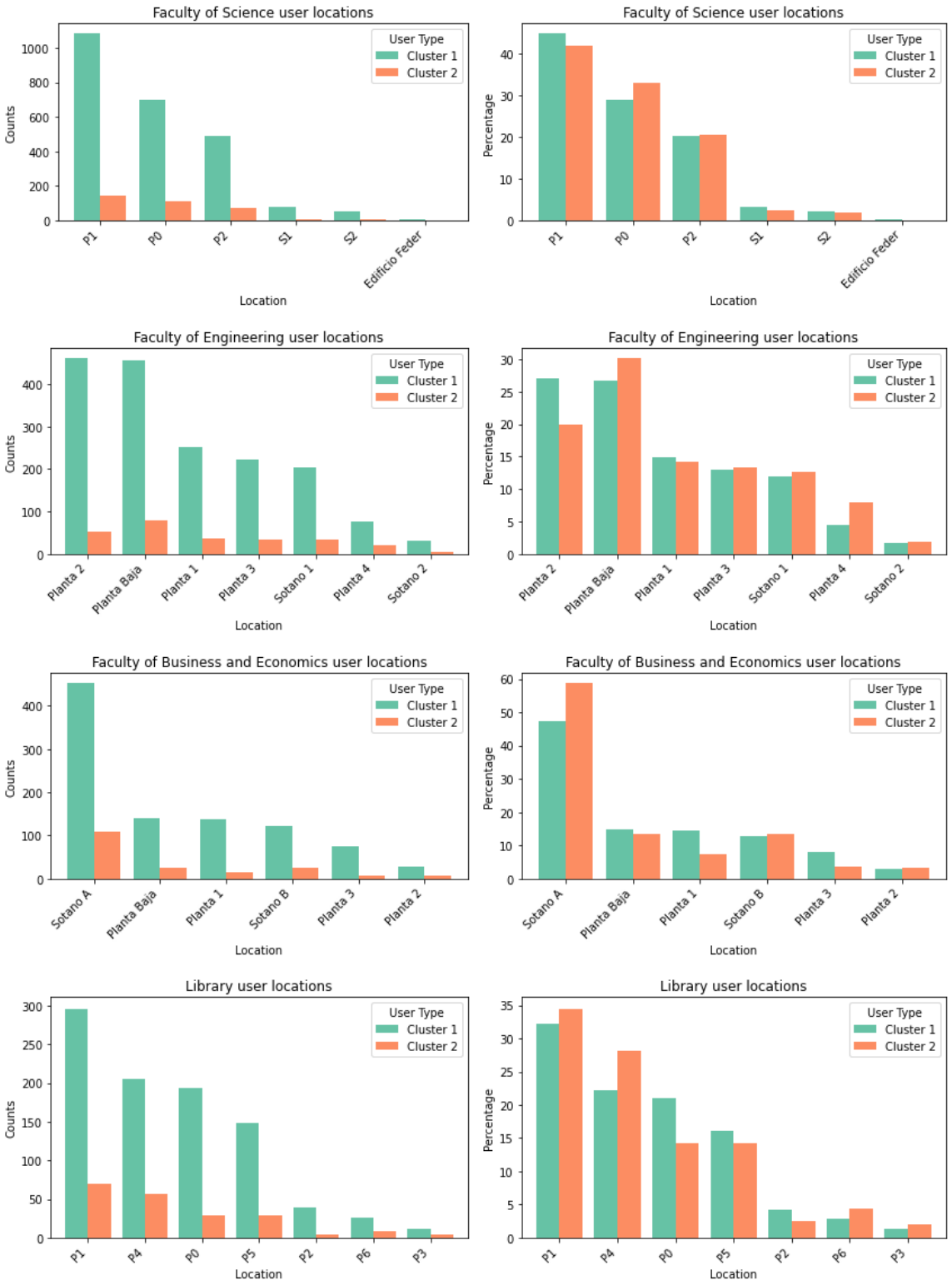

Figure 31: Location analysis results

9.3 Case comparison: University of the Basque Country and Technological University Dublin

This section aims to perform an analysis and comparison of the results obtained in the two real case scenarios where the QoXphere methodology has been attempted to be validated. These two scenarios correspond to two universities: Technological University Dublin and University of the Basque Country.

The first validation attempt was performed in Dublin, where data was collected form the library and a certain faculty building. As explained in Section 2, the following user profiles where identified:

- **Library scenario user profile 1:** Users that do not use the network while studying, observed by means of the lower transmission rates. This lack of network usage could be due to various reasons, such as, alternative use of books or notes for studying.
- **Library scenario user profile 2:** Users that actively use the network while studying, resulting in higher transmission rates.
- **Academic building scenario user profile 1:** Office and laboratory users. The lower total devices value is explained by the fact that the number of people in this area is expected to be lower.
- **Academic building scenario user profile 2:** Classroom users. The higher total devices value represents a larger number of users in this area.

The second validation attempt was carried out in the University of the Basque Country, where data was collected form the Faculties of Science, Engineering, Business and Economics and from the Library. Here, the two previously mentioned profiles were identified:

- **Active users**: Users connected to the network and who make an active use of it. These users usually occupy spaces dedicated to studying, such as study rooms, teamwork rooms or the library, as the reason behind their high network usage is probably related to study-tasks.
- **Passive users**: Users connected to the network and who do not make an active use of it. As the users are not expected to carry out any specific study tasks they can be found in many different areas of the campus.

Although the proposed profiles are different in each scenario, they do share some similarities. In fact, the profiling performed in the second study can cover the profiles from the first study. For example, the Library scenario user profile 1 and 2 have a direct relation with the Passive and Active user profiles

respectively. The Library scenario user profile 1 is defined as "*Users that do not use the network while studying*" which is similar to the Passive users' definition of "*Users connected to the network and who do not make an active use of it".* On the other hand, the Library scenario user profile 2 is defined as "*Users that actively use the network*", which coincides with the Active users' profile definition "*Users connected to the network and who make an active use of it*". As for the Academic building scenario user profiles, it would depend on the case. Office users would probably be included in the Passive users' group. Even though these users do make use of the network for emailing, researching, etc., they usually do it on computers provided by the university. These computers are connected to the university's LAN and therefore do not use the WLAN which is what is being studied. As for the students in the lab, this will probably depend on their study area. Those belonging to fields like engineering, might have to use specific programs that they have installed on their laptops for convenience. Therefore, their use of the network will probably be higher than a chemistry student's that is performing experiments and is not allowed to have the computer nearby. Lastly, for the classroom students, the situation can also vary. Depending on whether students use their laptop and network while listening to the teacher's explanation, they might either be classified as Active of Passive users.

10 DESCRIPTION OF THE RESEARCH TEAM AND WORK PLANNING

10.1 Work team

The work team for this project is composed by the following two members:

- **Ana Eva Ibarrola Armendariz**: Professor of the Communications Engineering Department at the University of the Basque Country's Faculty of Engineering in Bilbao. She received her Ph.D. degree in telecommunications engineering in 2010 and was honored with the Best Thesis Award in Management, Economy and Telecommunications Regulation. She has worked with several standardization organizations and currently participates in the NQaS research group of the University of the Basque Country where her research focuses on the development of user centric QoS management models and frameworks. In this study she plays the role of project director, where she will be responsible of establishing the project's objectives, performing continuous supervision of the project's development and performance and controlling the quality of the produced results and documentation.
- **Itziar Casado-O'Mara Corral**: Student of the Master in Telecommunication Engineering at the University of the Basque Country, specialized in Telematics Engineering. She has worked as a collaborator of the NQaS research group (UPV/EHU), where her researched has focused on the analysis and management of QoS in telecommunication networks through the application of ML techniques. She has also worked in the Institute Intelligence and Data (IID) [42] at Université Laval (Canada) [43], investigating the use of federated learning models. She is the project planner and, in this capacity, must design the schedule and cost baseline of the project, perform the data analysis in order to achieve the objectives laid out by the project director, extract conclusions from the obtained results and prepare the documentation.

The following table summarizes the members of the work team by indicating their name, position and role:

Table 8: Summary of work team members

Name and surnames	Position	Role
Ana Eva Ibarrola Armendariz	Senior engineer	Project director
Itziar Casado-O'Mara Corral	Junior engineer	Project planner

10.2 Project planning

To ensure project success and completion it is important to develop a detailed plan of the project. The project plan lists all the tasks (T) that must be performed and indicates their duration. Similar tasks are grouped together forming what are known as work packages (WP) which are used as control points to check if the project is on track. At the completion of a work package a deliverable (D) should be produced and the project's progress in terms of scope, schedule, budget and quality requirements should be assessed.

All projects go through the following five phases:

1. **Scope definition**: In this stage the specifications of the project are established. That is, the objectives of the project are defined and the benefits of conducting the project are specified.
2. **Planning**: In this stage a more detailed outlining of the project is made by defining the scope, schedule, budget and quality requirements. To be able to do this, a set of analysis must be carried out, such as, analysis of alternatives and risk analysis. Additionally, the project team must be formed and inquiries must be made about available resources.
3. **Execution**: In this stage, all efforts are put into producing the product of the project which in this case is the data analysis.
4. **Closeout**: This stage includes delivering the product of the project, finalizing documentation and drawing conclusion based on the project's performance to gather knowledge for future projects.
5. **Control**: Stages 1 to 4, are performed sequentially, that is, one after the other. However, the control stage spans the whole project. The goal is to carry out project surveillance to ensure that the project is on time, on budget and meeting the specified requirements.

In Table 9, the different work packages are shown alongside the tasks that compose each one of them and the deliverable expected after their completion.

Table 9: Project plan

ID	Name and description	Human resources	Duration (hours)
WP1	**Scope definition**		
T101	Project definition: Proposition of the project by the project director to the project planner and definition of the scope by establishing the goals to be achieved.	Junior engineer Senior engineer	8
G1	**Project proposal**		
WP2	**Planning**		
T201	Analysis of alternatives: Research on machine learning algorithms in order to determine the best option.	Junior engineer	30
T202	Risk analysis: Identification and evaluation of possible risks and development of preventive measurements and contingency plans.	Junior engineer	5
T203	Task planning: Breakdown of the tasks to be carried out to achieve the project objectives. Each task is assigned duration and resources.	Junior engineer	8
T204	Schedule planning: Activity scheduling by using the duration times established in the task planning.	Junior engineer	5
T205	Budget planning: Budget estimation.	Junior engineer	4
G2	**Project plan**		
WP3	**Execution**		
T301	Research: Research on IEEE 802.11 standards and ML models.	Junior engineer	120
T302	Data collection: Understanding of the entity-relationship model used when gathering and measuring the Wi-Fi data.	Junior engineer	15
T303	Data preprocessing: Preparation of data to ensure that it is reliable and robust for data processing.	Junior engineer	50

T304	Data analysis: Application of ML algorithms for user-type identification.	Junior engineer	215
T305	Result discussion: Analysis of the obtained results to determine if the methodology is validated and, if so, interpretation of the data to identify the different user-types' characteristics.	Junior engineer Senior engineer	110 22
G3	**Data analysis results**		
WP4	**Closeout**		
T401	Product and documentation handover: Preparation of the documentation to be delivered. The documentation includes a description of the process carried out during the big data analysis as well as the description of the results achieved. The information presented in the project proposal and the project plan are also included.	Junior engineer	30
G4	**Documentation completion**		
WP5	**Control**		
T501	Scope and risk management: Monitoring by the project manager to ensure that the project is on track to achieve the objectives, so that if it is not, she can intervene. Documentation revision.	Senior engineer	30
G5	**Documentation completion**		

10.3 Gantt diagram

TASK	START	DAYS	END
WP1 - Scope definition	**15-03-2023**	**3**	**17-03-2023**
T101 - Project definition	15-03-2023	3	17-03-2023
G1 - Project proposal			
WP2 - Planning	**18-03-2023**	**24**	**10-04-2023**
T201 - Analysis of alternatives	18-03-2023	12	29-03-2023
T202 - Risk analysis	30-03-2023	2	31-03-2023
T203 - Task planning	01-04-2023	5	05-04-2023
T204 - Schedule planning	06-04-2023	2	07-04-2023
T205 - Budget planning	08-04-2023	3	10-04-2023
G2 - Project plan			
WP3 - Execution	**11-04-2023**	**140**	**28-08-2023**
T301 - Research	11-04-2023	35	15-05-2023
T302 - Data collection	16-05-2023	5	20-05-2023
T303 - Data preprocessing	21-05-2023	15	04-06-2023
T304 - Data analysis	05-06-2023	55	29-07-2023
T305 - Result discussion	30-07-2023	30	28-08-2023
G3 - Data analysis results			
WP4 - Closeout	**29-08-2023**	**21**	**18-09-2023**
T401 - Product and documentation handover	29-08-2023	21	18-09-2023
G4 - Documentation completion			
WP5 - Control	**15-03-2023**	**188**	**18-09-2023**
T501 - Scope and risk management	15-03-2023	188	18-09-2023
G5 - Documentation completion			

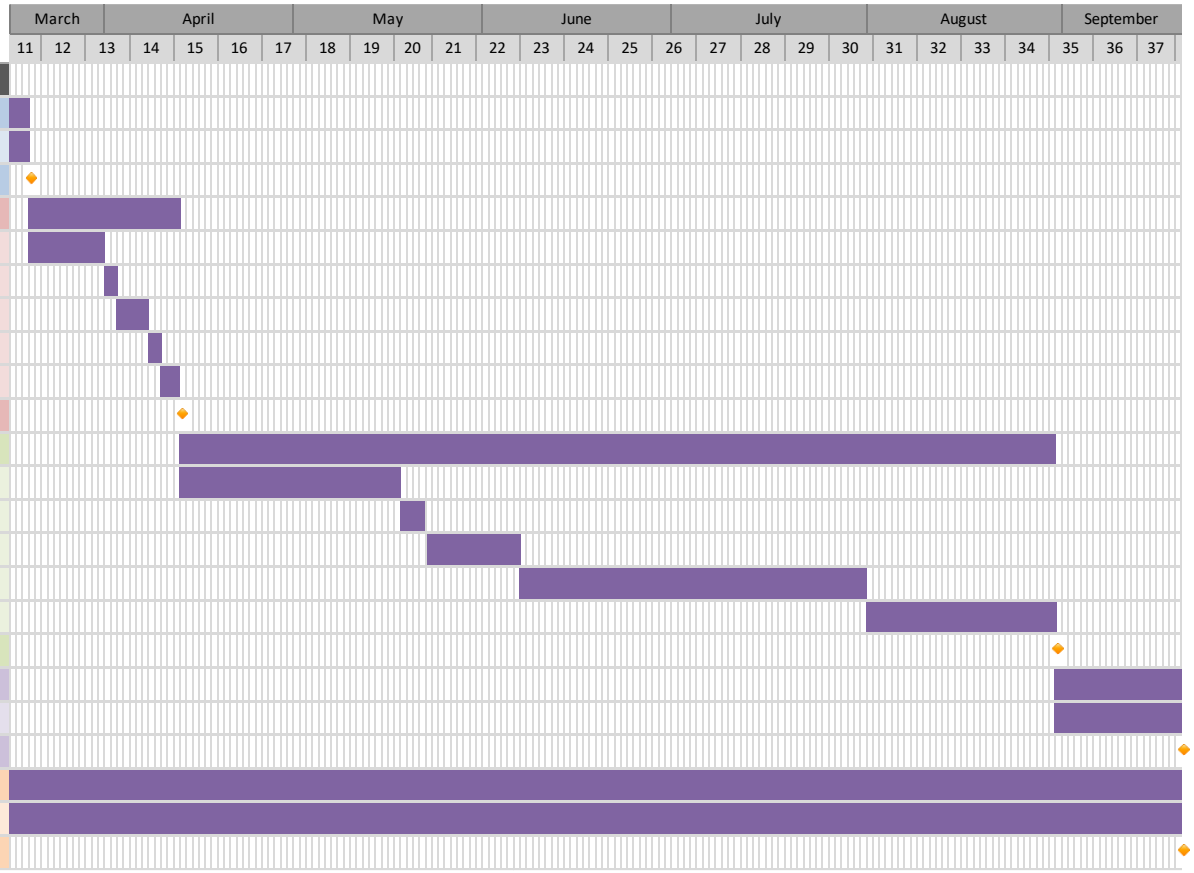

Figure 32: Gantt diagram

11 BUDGET

This section details the costs of the present project. There are two types of costs that are usually considered when preparing the project's budget: direct costs and indirect costs [39]. Direct costs are those that can be clearly assigned to one of the project's tasks. Common direct costs are those associated to labor, materials, equipment or sometimes subcontractors. Indirect costs, on the other hand, represent overhead costs that cannot be assigned to a specific project task. Typical indirect costs are rent, utilities and general and administrative expenses.

11.1 Direct costs

The direct costs that have been considered are the following:

- **Labor cost**: Sum paid to cover the salary and benefits of an employee. In this case it refers to the amount paid to the project members based on their hourly wage and the number of hours they have worked on the project (Table 10).
- **(Direct) material cost**: Cost of materials employed in the project to develop the product or service that can be associated with a particular task. As this is a data analysis project where the data was obtained by means of a university collaboration, no money was spent on the purchase of materials.
- **Equipment cost**: Sum of the items' purchase price and other costs related to acquisition. As free software was used to avoid the purchase of licenses or subscriptions, the only equipment that was used was a MacBook Air (M2, 2022) in order to perform the data analysis and produce the required documentation (Table 11).

Table 10: Budget items under labor cost

Concept	Hourly cost (€/h)	Hours (h)	Cost (€)
Project director	60.00	60	3600.00
Project planner	15.00	600	9000.00
Total			**12600.00**

Table 11: Budget items under equipment cost

Concept	Cost (€)
MacBook Air (M2, 2022)	1500.00
Total	**1500.00**

11.2 Indirect costs

The indirect costs that have been considered are:

- **Amortizations**: Recognition of the depreciation of an asset according to its contribution to generate income for the company (Table 12).
- **(Indirect) material cost:** Cost of materials employed in the project that cannot be associated with a specific task (Table 13).

Table 12: Budget items under amortizations

Concept	Initial cost (€)	Lifespan (months)	Monthly cost (€/month)	Use (months)	Cost €)
Project director's computer	2100.00	60	35.00	4	140.00
Project planner's computer	1500.00	60	25.00	4	100.00
Total					**240.00**

Table 13: Budget items under (indirect) material costs

Concept	Cost (€)
Stationery	50.00
Utilities	200.00
Total	**250.00**

11.3 Total costs

Once direct and indirect costs have been calculated, these are summed together to calculate the total cost of the project. As can be seen in the following table (Table 14), the total cost amounts to 14590.00 €.

Table 14: Total project costs

Type of cost	Concept	Cost (€)
Direct	Labor cost	12600
	Equipment cost	1500.00
Indirect	Amortizations	240.00
	(Indirect) material costs	250.00
Total		**14590.00**

12 CONCLUSIONS

This section aims to briefly present the conclusions of the project in a clear way in order to get an overview of what has been accomplished.

First, the main goal of this project was achieved, that is, a specific part of the QoXphere methodology (Context extraction) was validated for its application in real case scenarios. This was done by studying the real case scenario of the corporate Wi-Fi network of the University of the Basque Country. As the methodology indicates, by applying ML techniques and taking into consideration the QoS requirements, two different user profiles, Active and Passive users were identified. The main findings related to these users are the following:

- **Active users**: Users connected to the network and who make an active use of it. These users usually occupy spaces dedicated to studying, such as study rooms, teamwork rooms or areas of the library, as the reason behind their high network usage is probably related to study-tasks.
- **Passive users**: Users connected to the network and who do not make an active use of it. As the users are not expected to carry out any specific study tasks they can be found in many different areas of the campus.

In fact, due to the satisfactory outcome of this study, the author of this project has participated in the elaboration of a paper entitled “QoX Management for Modern Networks” where the project’s results have been included. This paper has been sent and its abstract has already been registered for the special issue “Advancements in QoS/QoE for Future Networks and Their Applications” of the "Electrical, Electronics and Communications Engineering" section of the Applied Sciences journal (ISSN 2076-3417).

Second, the results obtained in the present project have been compared to the ones obtained in the Bachelor’s Final Degree project, in order to analyze the influence of contextual aspects in users QoS requirements for the two different settings. Given that both settings are academic scenarios it has been seen how there are many similarities in the obtained results and how the classification made in the present project can also be applied to the results obtained in the other study.

In conclusion, it may be affirmed that this study has successfully achieved the initially established objectives. However, there are still many steps to be validated in the QoXphere methodology, therefore, some future research areas are envisaged. As it was discovered that the user profiles identified in the

University of the Basque Country could also be applied to the users of the Technological University of Dublin, it would be interesting to expand this research to other universities to determine if this user profiling could be generalized for university campuses. On the other hand, it would also be interesting to study completely different scenarios to identify the differences amongst them and to better identify the characteristics of each one.

13 BIBLIOGRAPHY

[1] "Measuring digital development Facts and Figures 2022," Geneva, 2022.

[2] "Networking, Quality and Security Research Group," *http://det.bi.ehu.eus/NQAS/?home*.

[3] "University of the Basque Country," *https://www.ehu.eus/en/en-home*.

[4] "Technological University Dublin," *https://www.tudublin.ie/*.

[5] "THE 17 GOALS," *https://sdgs.un.org/goals*.

[6] W. Hardy, "QoS: Measurement and Evaluation of Telecommunications Quality of Service," 2002, pp. i–xvi. doi: 10.1002/0470845910.fmatter_indsub.

[7] ITU-T Recommendation E.800, "Terms and definitions related to the quality of telecommunication services." 1994.

[8] ITU-T Recommendation G.1000, "Communications quality of service: A framework and definitions." 2001.

[9] ITU-T Recommendation E.800, "Definitions of terms related to quality of service." 2008.

[10] ITU-T Recommendation P.10/G.100, "Vocabulary for performance, quality of service and quality of experience." 2017.

[11] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, "Quality of service terminology in IP networks," *IEEE Communications Magazine*, vol. 41, no. 3, pp. 153–159, 2003, doi: 10.1109/MCOM.2003.1186560.

[12] E. Ibarrola, E. Saiz, J. Xiao, L. Zabala, and L. Cristobo, "QOXPHERE: A new QoS framework for future networks," in *2013 Proceedings of ITU Kaleidoscope: Building Sustainable Communities*, 2013, pp. 1–7.

[13] "Eduroam," *https://eduroam.org*.

[14] J. Kurose and K. Ross, *Computer networking: A top-down Approach*, 7th Edition. Madrid: Pearson, 2017.

[15] Y. El khatib, “WiFi ad-hoc message propagation over GPRS networks,” 2006.

[16] Q. Ni, “Performance analysis and enhancements for IEEE 802.11e wireless networks,” *IEEE Netw*, vol. 19, no. 4, pp. 21–27, 2005, doi: 10.1109/MNET.2005.1470679.

[17] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke, “Analysis of IEEE 802.11e for QoS support in wireless LANs,” *IEEE Wirel Commun*, vol. 10, no. 6, pp. 40–50, 2003, doi: 10.1109/MWC.2003.1265851.

[18] S. Sah, “Machine Learning: A Review of Learning Types.” May 2020. doi: 10.20944/preprints202007.0230.v1.

[19] X. Zhu, J. Lafferty, and R. Rosenfeld, “Semi-Supervised Learning with Graphs,” USA, 2005.

[20] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-Taught Learning: Transfer Learning from Unlabeled Data,” in *Proceedings of the 24th International Conference on Machine Learning*, in ICML ’07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 759–766. doi: 10.1145/1273496.1273592.

[21] F. Olsson, “A literature survey of active machine learning in the context of natural language processing,” May 2009.

[22] A. E. Ezugwu *et al.*, “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” *Eng Appl Artif Intell*, vol. 110, p. 104743, 2022, doi: https://doi.org/10.1016/j.engappai.2022.104743.

[23] S. Pitafi, T. Anwar, and Z. Sharif, “A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms,” *Applied Sciences*, vol. 13, no. 6, 2023, doi: 10.3390/app13063529.

[24] H. G. Wilson, B. Boots, and A. A. Millward, “A comparison of hierarchical and partitional clustering techniques for multispectral image classification,” *IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, pp. 1624–1626 vol.3, 2002, [Online]. Available: https://api.semanticscholar.org/CorpusID:53793183

[25] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature," *Neural Comput Appl*, vol. 33, no. 11, pp. 6247–6306, 2021, doi: 10.1007/s00521-020-05395-4.

[26] S. E. Schaeffer, "Graph clustering," *Comput Sci Rev*, vol. 1, no. 1, pp. 27–64, 2007, doi: https://doi.org/10.1016/j.cosrev.2007.05.001.

[27] J. A. dos Santos, T. I. Syed, M. C. Naldi, R. J. G. B. Campello, and J. Sander, "Hierarchical Density-Based Clustering Using MapReduce," *IEEE Trans Big Data*, vol. 7, no. 1, pp. 102–114, 2021, doi: 10.1109/TBDATA.2019.2907624.

[28] H. Banerjee Arindam and Shan, "Model-Based Clustering," in *Encyclopedia of Machine Learning*, G. I. Sammut Claude and Webb, Ed., Boston, MA: Springer US, 2010, pp. 686–689. doi: 10.1007/978-0-387-30164-8_554.

[29] L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, Jun. 2004, doi: 10.1145/1007730.1007731.

[30] S. Pandya and S. Saket, "An overview of partitioning algorithms in clustering techniques," *International Journal of Electrical and Computer Engineering*, vol. 5, Aug. 2020.

[31] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, 1967, pp. 281–297.

[32] "File:K-means.png," *https://commons.wikimedia.org/wiki/File:K-means.png*.

[33] L. Kaufmann and P. Rousseeuw, "Clustering by Means of Medoids," *Data Analysis based on the L1-Norm and Related Methods*, pp. 405–416, Aug. 1987.

[34] https://sens.tistory.com/297, "k-medoids".

[35] L. Kaufman and P. Rousseeuw, "Clustering Large Data Sets," 1986, pp. 425–437. doi: 10.1016/B978-0-444-87877-9.50039-X.

[36] L. Kuang and L. Zhang, "A scheduling algorithm based on Clara clustering," *AIP Conf Proc*, vol. 1864, no. 1, p. 020016, Aug. 2017, doi: 10.1063/1.4992833.

[37] R. T. Ng and J. Han, "CLARANS: a method for clustering objects for spatial data mining," *IEEE Trans Knowl Data Eng*, vol. 14, no. 5, pp. 1003–1016, 2002, doi: 10.1109/TKDE.2002.1033770.

[38] Thomas Elliott, "The State of the Octoverse: machine learning," *https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/*.

[39] E. W. Larson, C. F. Gray, and C. F. Gray, *Project management : the managerial process*.

[40] "Localización de los centros UPV/EHU," *https://www.ehu.eus/es/ikastegien-kokapena-bizkaia*.

[41] "Mobile Vendor Market Share Spain," *https://gs.statcounter.com/vendor-market-share/mobile/spain/#monthly-202208-202308-bar*.

[42] "Institut intelligence et données," *https://iid.ulaval.ca*.

[43] "Université Laval," *https://www.ulaval.ca*.