POLITECNICO DI TORINO

Master of Science: ICT for Smart Societies

Master's Degree Thesis

ML Optimization of Cell-Range Overshooting Detection in Real LTE Networks

Supervisor:

Prof. Tiziano Bianchi

Candidate:

Francesca Fanelli

Co-Supervisors:

Prof. Luis Mendo Tomás
Prof. Luis Alfonso Hernández Gómez
Prof. Zoraida Frías Barroso
Prof. Mateo José Cámara Largo

A.Y. 2023/2024

Abstract

The dynamic evolution of mobile access networks, marked by increasing size and complexity, necessitates innovative approaches to automatic anomaly detection and resolution. Traditional manual methods prove insufficient in handling the complexities of modern networks. Consequently, network optimizers must shift focus towards developing algorithmic solutions that enable automation.

This thesis emerges from the *AI/ML Optimization Program 2024-2026*, a collaborative research initiative between Telefónica's Radio Access Network Optimization teams and the Universidad Politécnica de Madrid (UPM). The primary objective is to develop an automated system for managing and monitoring access networks, leveraging data analysis and Machine Learning (ML) techniques to optimize network performance.

This project concentrates in studying the behavior of commonly used network Key Performance Indicators (KPIs) under various typical issues, referred to as use cases, with the goal of automatically suggesting appropriate adjustments. Specifically, the thesis focuses on the development of a detection system for cell-range overshoot phenomenon in LTE networks. Cell-range overshoot occurs when the coverage area of a cell extends beyond its intended boundaries, leading to inefficient network resource usage, degraded signal quality, and disrupted handover procedures.

This research, after reviewing basic LTE access procedures and optimization techniques, introduces the cell-range case study. To this end, a detailed analysis of commonly used Key Performance Indicators (KPIs) is addressed. Specifically, for this study a carefully designed database of real KPIs collected from the actual Telefónica's LTE deployment in Spain has been created. This database also includes true labels of cell-range anomalies manually detected by current Telefónica Optimization Teams. Several Machine Learning models have been designed and evaluated to test their capabilities to automatically detect cell-range overshooting.

Key findings include the development of primary classification models capable of detecting problematic cells. To this end, tree ensemble models (namely Random Forest and eXtreme Gradient Boosting) are chosen both for their performance and their ability to express the feature importance analysis on which the algorithms build their classification criterion. The models are constructed and trained performing K-fold cross-validation techniques over the KPIs database.

These models provide around 70% accuracy in detecting cell-range anomalies and they have been deployed in a real testing probe by Telefónica.
This research study highlights the challenging task of automatically detecting specific cellular anomalies, which are notably similar to one another and closely linked to the intrinsic characteristics of access networks.

Future research should focus on incorporating additional information specific to cell-range overshooting to enhance detection mechanisms for this particular use case.

In summary, this thesis contributes to the ongoing efforts to automate and optimize Telefónica's mobile network management, showcasing the potential of ML techniques to improve network performance and efficiency. The detection solutions developed in this thesis are anticipated to extend seamlessly to Telefónica's entire network infrastructure, including future technologies like 5G.

Acknowledgements

This thesis was conducted during my Erasmus year as part of a collaborative project between the Universidad Politécnica de Madrid (UPM) and Telefónica. This project was jointly supervised by Politecnico di Torino, UPM, and Telefónica. I am deeply grateful to all the institutions and individuals involved in making this thesis possible.

I would like to express my sincere gratitude to Professor Tiziano Bianchi of Politecnico di Torino for his guidance and assistance which significantly enriched the quality of this thesis, despite the geographical distance.

Words cannot express my gratitude to all professors from ETSI de Telecomunicación (UPM), for being dedicated and passionate teachers. I am deeply grateful to Professor Luis Mendo Tomás, Professor Luis Alfonso Hernández Gómez, Professor Zoraida Frías Barroso and Professor Mateo José Cámara Largo, for their invaluable patience, support and expertise throughout both the research and writing process of this thesis.

Additionally, I am thankful for the expertise and assistance provided by the team at Telefónica. Their insights, resources, and collaborative efforts have been crucial in the completion of this thesis, and I am profoundly grateful for their support.

A special mention goes to my friend and colleague Jorge, for sharing the highs and lows of the research process, celebrating achievements and overcoming obstacles as a team. His insights and feedback have enriched my own understanding and perspective on the subject matter.

Moreover, particular gratitude to my sibling Mica Fanelli for their professional assistance in designing the graphical illustrations used in this thesis.

Last but certainly not least, I want to express my deepest appreciation to my family and friends. Their unwavering love and encouragement have sustained me throughout my entire academic journey. This achievement is as much theirs as it is mine, and I am profoundly grateful for their presence in my life.

Contents

List of Figures

Image Credits

Graphical illustrations designed by M.Fanelli:

- Figure 1.1
- Figure 2.1
- Figure 2.2
- Figure 2.3
- Figure 2.4
- Figure 2.5

List of Tables

Table Credits

Graphical illustrations designed by M.Fanelli:

- Table 2.1
- Table 2.2

1. Introduction

This work stems from *AI/ML Optimization Program 2024-2026*, a research project born in collaboration between various Radio Access Network Optimization teams within Telefónica and a team from the Universidad Politecnica de Madrid (UPM), which includes professors, researchers, PhD students and Master's grad students, based in the Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT).

At its core lies the recognition that mobile networks generate an abundance of data ripe for exploration through data analysis and Machine Learning (ML) techniques. This exploration offers a multiple opportunities to challenge existing procedures and elevate network optimization.

The dynamic evolution of access networks, expanding both in size and complexity, necessitates a fresh approach. With the integration of a myriad of innovative technologies, including slicing and virtual networks, the detection and resolution of issues demand agile solutions. Traditional manual intervention proves inadequate in facing the such complexity of modern networks.

The role of network optimizers must evolve to focus significant time and expertise on developing and optimizing algorithmic solutions that facilitate automation. While traditional optimization approaches remain relevant, the definition and development of automatic optimization processes and algorithms will become fundamental to daily optimization tasks.

The purpose of the project is to develop a system of optimization and automation of access networks' management and monitoring. This work aims at studying the behavior of commonly-used network's Key Performance Indicators (KPIs) under several typical issues and problematics, called use cases, in order to automatically suggest proper adjustments solutions.

In Chapter 1 a brief description of the framework in which the work of this thesis has been developed will be provided, including an introduction to Telefónica company as well as its project *AI/ML Optimization Program 2024-2026*. Additionally, Chapter 1 offers an overview of the cell-range use case, which is the main focus of the thesis.

Chapter 2 contains detailed background knowledge about general Long Term Evolution (LTE) Access procedures, alongside exhaustive presentation of the use case covered in this thesis, as well as a second use case involved during Testing Phase. Additionally, in Chapter 2 presents a global view of current optimization techniques,

algorithms, and technologies used for random access networks in the telecommunications industry.

The work of the thesis is presented in Chapter 3 and Chapter 4. Specifically, Chapter 3 describes the initial phase of the work. It includes Machine Learning problem framing, data processing and exploratory data analysis. Subsequently, in Chapter 4 the testing phase of the project is provided. It covers several tests carried out on the models defined in the previous chapter.

Finally, in Chapter 5, conclusions are drawn and the results are comprehensively discussed addressing possible interesting future research directions.

1.1 Telefónica

In the dynamic landscape of telecommunications, Telefónica is a pioneer in the industry, leading technological interconnections advancements in Spanish territories as well as worldwide.

Since its inception, Telefónica has been synonymous with innovation. Founded in 1924 in Spain, the company has achieved operating across Europe, Latin America, and beyond. Through strategic investments in research and development, Telefónica has pioneered groundbreaking technologies, such as the digital revolution and the dawn of 5G connectivity.

> *At a time when technology is more present than ever in our lives, we cannot forget that the most important connections are human connections.* [15]

Telefónica's mission is to deliver reliable, high-speed connectivity in bustling urban centers as well as remote rural areas.

As the digital era unfolds, Telefónica continuously seeks to optimize its mobile network infrastructure, addressing challenges and maximizing efficiency at every level. One such challenge lies in the optimization of access networks — a critical factor in ensuring seamless coverage and optimal network performance across diverse areas and environments. By refining algorithms, deploying advanced antenna technologies, and leveraging data analytics, Telefónica is willing to achieve highly-reliable and robust cellular network infrastructures, withstanding the increasing weight of demand.

1.2 AI/ML Optimization Program 2024-2026

The "AI/ML Optimization Program 2024-2026" seeks to address the escalating complexity of access networks by embracing dynamic optimization strategies driven by available data. With access networks becoming increasingly intricate, traditional manual methods for problem detection and correction are quickly becoming inadequate and obsolete. As these networks grow in complexity, there's a need for automated optimization processes to efficiently manage them without a significant

increase in resource allocation. Thus, the program aims to advance dynamic optimization techniques leveraging traditional cellular network's KPIs, representative of both network conditions and user experience, to proactively identify and resolve issues. Through a combination of automated data-driven analysis and corrective actions, the program envisions a future where optimization processes operate with minimal human intervention. The project's short-range (2024) objectives include segmenting and clustering the access network elements based on specific scenarios and issues, and implementing initial automated optimization algorithms across defined use cases. This determination considers various factors such as geographic location, topological features, capacity, traffic and mobility patterns, and spectrum usage. By precisely understanding each cell's nature and its impact area, optimization efforts can be tailored more effectively. Through these objectives, Telefónica aims to advance towards a more automated and efficient optimization framework for its access networks.

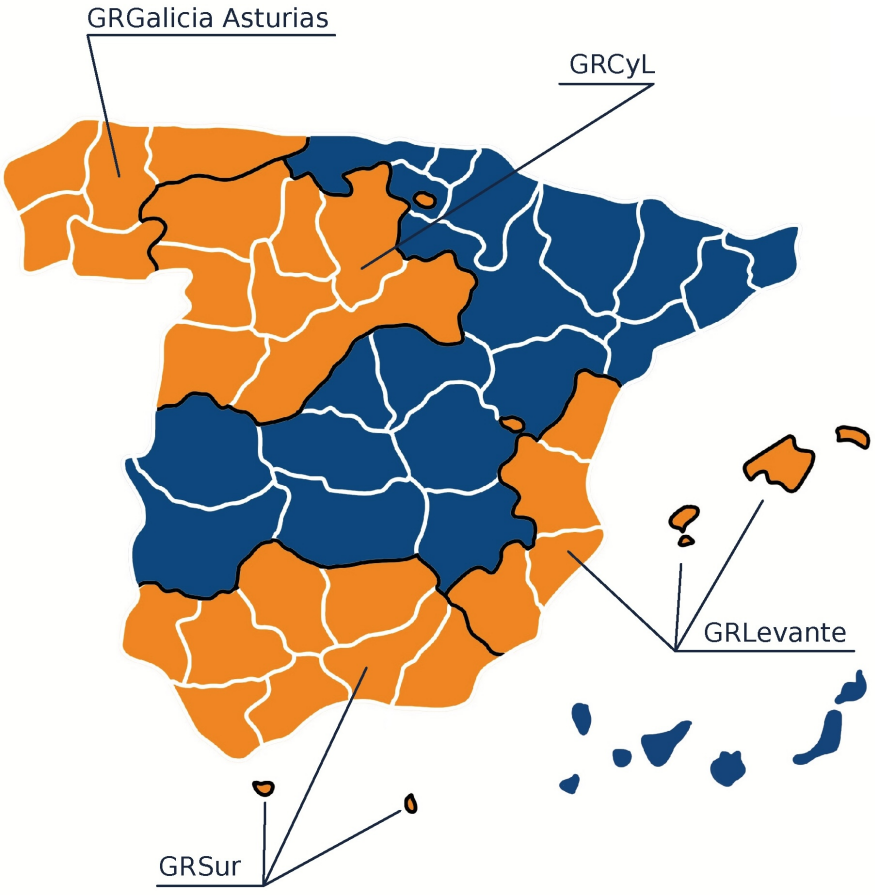


Figure 1.1: Geographical areas covered by data in Spanish territory.

While the comprehensive objective of the program is to encompass Telefónica's whole network infrastructures, including the latest innovative network technologies and the whole Spanish geographical area covered by Telefónica, the scope of this thesis is specifically delimited to LTE networks over four specific regions. In Figure 1.1, the

geographical area covered in this thesis is reported in orange. Hence, four regions are defined: "GRCyL" covering the territory of Castile and León, "GRSur" covering Andalusia, "GRLevante" comprising Murcia, Valencia and Balearic Islands, and finally "GRGaliciaAsturias" including Galicia and Asturias.
The decision to focus on these regions is based on the company's internal policies, which are not pertinent to discuss in this context.

It is important to notice that Telefónica's mobile networks utilize equipment from two different suppliers, each with its own specific KPI formulas. Moreover, in accordance with Telefonica's privacy policies, the names of these two vendors will not be disclosed. For the purposes of this thesis, these vendors will be referred to as Vendor-A and Vendor-B, as this information is considered sensitive.
Therefore, for the sake of data consistency, the work of this thesis focuses exclusively on Vendor-A.

It is presumed that the solutions developed in this thesis can be seamlessly extended to encompass Telefónica's entire network infrastructure in Spain, inclusive of 5G technology and Vendor-B equipment. This adaptability is expected to be put into action in the near future within the program's framework.

1.2.1 Cell-range Use Case

Telefónica's project itself defines several very general use cases to differentiate the most common issues arising and affecting access networks. However, this thesis will focus on one kind of such challenges, the problematic related to the concept of cell-range overshoot, i.e. the situation where the coverage area of a cell extends beyond its intended boundaries. This phenomenon typically occurs due to various factors, including the configuration of cell-range parameters. When configuring the cell-range parameters, network engineers aim to establish boundaries within which random access procedures can reliably be performed to achieve connection of satisfactory signal quality. However, if these parameters are not properly configured or if the signal propagation characteristics are not accurately accounted for, User Equipments (UEs) located beyond the intended coverage area may still attempt to connect to the cell. The overshoot can lead to several issues within the network. For instance, it can cause UEs located beyond the intended coverage area to mistakenly perform random access procedure (more details in Section 2.1), sending preambles to the erroneous base stations. This can result in inefficient use of network resources (and higher power consumption), degraded signal quality and reduced quality of service for users. Additionally, the overshoot phenomenon can complicate handover (HO) procedures and disrupt the seamless transition of connections between cells, leading to dropped calls or data transmission interruptions.

In the forthcoming Chapter 2, specifically within Section 2.2, a comprehensive examination of the potential challenges resulting from the overshooting phenomenon is provided, offering detailed technical insights.

2. Background

As discussed in Section 1.2.1, overshooting phenomenon highly impacts overall network performances. In particular, overshooting is an issue affecting the Random Access Network (RAN), mistakenly reaching UEs located out of the intended cell-range.

As already mentioned in Section 1.2, Telefónica's network equipment is supplied by Vendor-A and Vendor-B, and each supplier's equipment has different - although equivalent - configuration parameters. However, this work focuses on Vendor-A telecommunication equipment.

This Chapter describes, in general terms, the procedures for a terminal to be able to access an LTE-based network.

2.1 LTE Access Procedures

Before an LTE terminal can communicate with an LTE network it has to carry out the following procedures [3]:

1. *cell search*: find and acquire synchronization to a cell within the network;
2. *cell system information*: receive and decode the information needed to communicate properly within the cell;
3. *random access*: request and obtain a connection setup.

The first procedure, *cell search*, is not only necessary at the initial access to the system but rather, to support mobility; LTE terminals need to continuously search for, and synchronize to neighboring cells. The reception quality of such cells is evaluated and compared to the reception quality of the current cell, drawing conclusions about a possible handover (HO) or cell reselection. At the end of this first step, a terminal synchronizes to a cell, acquiring the physical-layer identity of the cell and detecting the cell frame timing.

As second step, the terminal needs to acquire the *cell system information* to be able to access the cell. Such information is enclosed in so-called System-Information Blocks (SIBs) that are repeatedly broadcast by the network. The system information (in particular, the SIB-2) includes, among other things, detailed parameters related to random-access transmission. To this end, the main mechanism used in LTE

networks is by transmitting the SIBs through the Downlink Shared Channel (DL-SCH).

Lastly, the terminals request a connection setup to the cell, commonly referred to as *random access*. The main objective of this last procedure is to eventually establish connection with the base station. Acquisition of uplink syncronization, i.e. timing advance, is essential to this purpose. Either a contention-based or a contention-free (dedicated) scheme can be used, depending on the purpose. As the names suggest, contention-based procedure comes into play when multiple UEs attempt to access the network simultaneously, while contention-free random access is initiated by the network, specifically during a UE HO scenario.

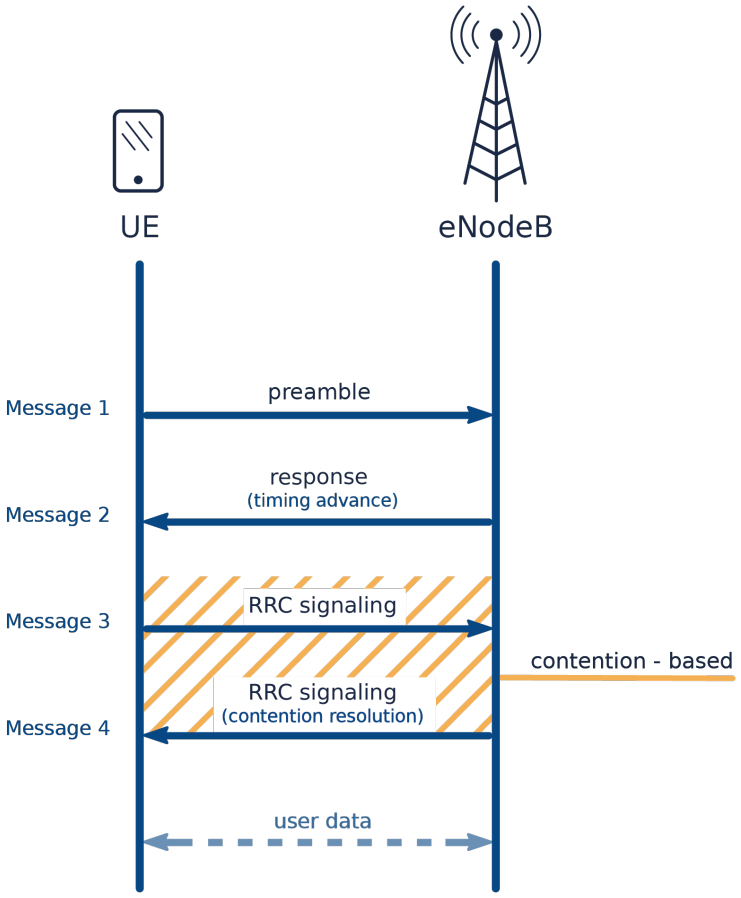


Figure 2.1: Random Access step-procedure.

As shown in Figure 2.1, the random access is basically a four-step procedure, which is initiated by the terminal sending a preamble (also known as message-1) through the Physical Random Access Channel (PRACH). As second step, the evolved Node-B (eNB) responds with message-2 which includes a timing advance command to adjust synchronization and assigns uplink resources to terminal. Then, in case of contention-based connection, the terminal transmits message-3 (RRC signalling) sharing its unique identity with the network and requesting connection. As final step,

the eNB concludes the procedure with message-4, a contention-resolution message. Except for the first step which uses physical-layer processing dedicated to random access, the subsequent messages are sent through DL-SCH and UL-SCH, normally used for data transmission. Although the procedure above is started by the terminal, also the network could initiate a random access, using RRC signalling or so-called PDCCH order (primarly intended for re-establishing uplink synchronization).

2.1.1 PRACH

[1] The Physical Random Access Channel is the time-frequency resource used to transmit the random access preambles, whose main purpose is to indicate to the base station the presence of random-access attempt. The PRACH resource information is broadcast by the network in SIB-2.

In the frequency domain, the PRACH resource has a bandwidth of 1.08MHz. In the time domain, the duration depends on the configured preamble format.

The general structure of a random access preamble is illustrated in the Figure 2.2.

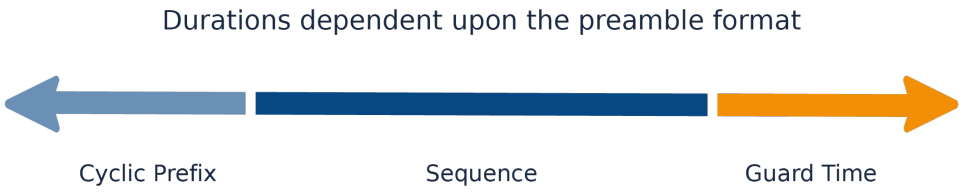


Figure 2.2: Structure of random access preamble.

As the Figure 2.2 shows, the preamble consists of two parts: a cyclic prefix and a preamble sequence. In addition, a guard period is used to handle timing uncertainty. At the beginning of random access procedure, the terminal has already acquired downlink synchronization through the cell search procedure, but uplink synchronization has not been established yet. Hence, there is uncertainty in the uplink timing due to unknown location of terminal within the cell.

The cyclic prefix is required to account for two aspects: the maximum delay spread, i.e. the last delay spread component should arrive within the cyclic prefix period of the first preamble component, as well as the maximum delay, i.e. addressing the case of terminals at the edge whose transmission is affected by the cell range distance. In the latter, the maximum delay is equal to the round trip time: one-way delay accumulated at reception of prior PRACH and one-way delay due to transmission of PRACH.

Figure 2.3 illustrates transmission of PRACH preamble received by an eNB from two different UEs: one user located at short distance from the node, while the other user at cell edge. It is possible to notice that the length of the cyclic prefix is approximately equal to the length of the guard period, which should be enough

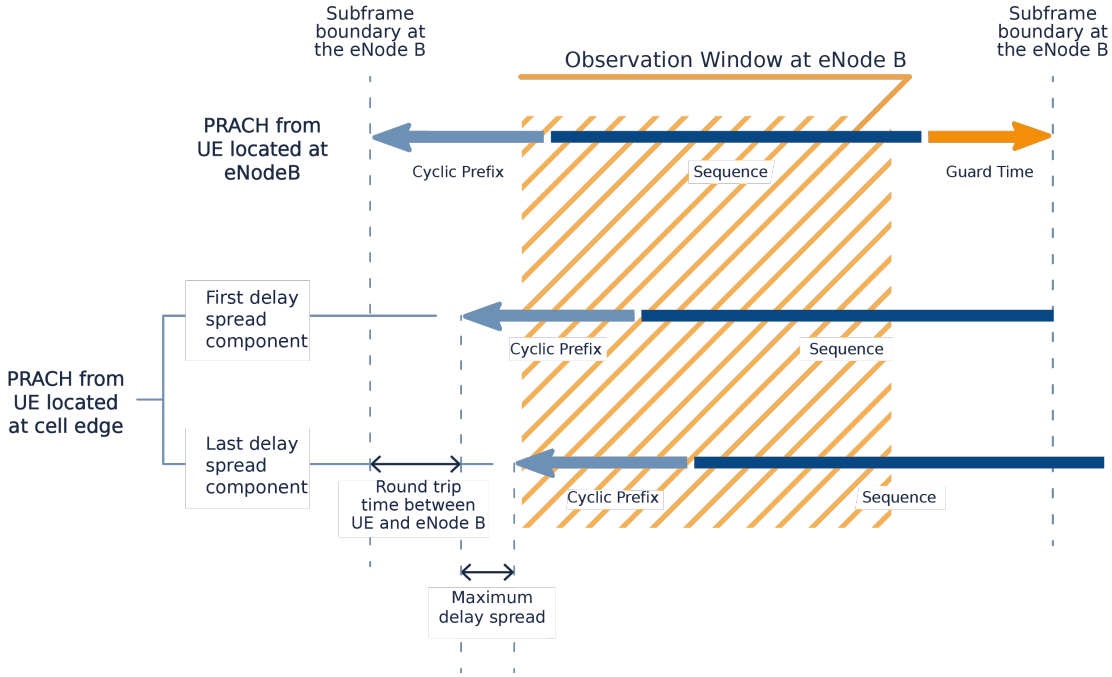


Figure 2.3: Cyclic Prefix allowing for both the maximum round trip time and maximum delay spread.

to accommodate at least the round trip time. This limits any overlap into the subsequent subframe i.e. only delay spread components overlap into the subsequent frame. This is deemed to be acceptable because the overlap is into the cyclic prefix region of the subsequent subframe.

According to Third Generation Partnership Project (3GPP) standards, there are 5 preamble formats presented in the Table 2.1. These are all based on the general structure illustrated in the Figure 2.2, but have different durations of cyclic prefix and guard time. Formats 0 to 3 can be used by either FDD and TDD.

In general, the radio network planner selects the appropriate format based upon the cell range. Formats with longer cyclic prefixes and longer guard times are more

Preamble Format	Application	Cyclic Prefix Duration	Sequence	Guard Time	Total Lenght	Typical Max Cell Range	Vendor - A Configurations for:
0	FDD & TDD	103.13µs	800µs	96.88µs	1ms	14.5km	LTE - 1800
1	FDD & TDD	684.38µs	800µs	515.63µs	2ms	77.3km	LTE - 800
2	FDD & TDD	203.13µs	1600µs	196.88µs	2ms	29.5km	
3	FDD & TDD	684.38µs	1600µs	715.63µs	3ms	100.2km	
4	TDD only	14.58µs	133µs	9.38µs	0.16ms	1.4km	

Table 2.1: PRACH format parameters.

suitable for larger cell ranges, which tend to experience larger delays spreads and have longer propagation channel round trip times. The drawback associated with using longer PRACH duration is an increased PRACH overhead, i.e. an increased number of resource blocks are allocated to the PRACH transmission.

Practically, the preamble format to be used within a specific cell is indicated using the PRACH Configuration Index (broadcast in SIB2). This parameter also defines the subframes during which the PRACH preambles can be transmitted.

In each cell, there are 64 preamble sequences available. These preambles allow multiple UE to share the same Root Sequence Index (RSI) values. Two main sets are defined to dedicate a larger number of preambles to contention-based purposes and a smaller number of them for contention-free purposes.

Moreover, in case of contention-based attempt, two sub-subsets of sequences are defined called group A and group B. Depending on the uplink data quantity to send and the experienced coverage quality, the group to use is identified. Within that group, the terminal selects at random one sequence. Instead, in case of contention-free attempt, for example for handover procedure, the eNB explicitly indicates the preamble sequence to use in order to avoid any collision.

3GPP standard specifies that preambles sequences are generated from a set of 838 root sequences (RSI), the Zadoff-Chu sequences, i.e. particular sequences characterized by great uncorrelation. Each preamble sequence is generated from its root sequence (RSI) by applying a cyclic shift. The zero-correlation (ZC) zone parameter determines the size of the cyclic shift and the number of preamble sequences that can be generated. A small zero-correlation zone means a small cyclic shift is used and therefore a larger number of preambles sequences can be generated from a unique root sequence. This would be suitable in case of a small cell range. The Table 2.2 shows the cell range as a function of the ZC zone.

It is beneficial to generate as many preamble sequences as possible from the same root sequence because such procedure would ensure orthogonality among one sequence and another. On the contrary, preamble sequences generated from different root sequences are not orthogonal, which implies intra-cell air-interface. Furthermore, generating a large number of preambles from the same root sequence also means that each cell would require fewer root sequences to construct the set of 64 preambles. Hence, this allows greater re-utilization of the set of 838 root sequences and make it easier to ensure that the sets of root sequences assigned to neighboring cells are mutually exclusive. However, the important drawback of generating large numbers of preamble sequences from one single root sequence is that the maximum supported cell range significantly decreases. The reason behind this lies in the challenges of distinguishing a frequency offset from distance-dependent delay.

Consequently, the extent of the cell range is delineated by two constraining factors. Primarily, a physical constraint arises from the guard time inherent in the selected preamble format. A secondary constraint emerges from the level of reuse of root sequences, indicated in the configuration of PRACH Cyclic Shift (prachCS) parameter. Additionally, the radio coverage of the cell must be taken into consideration, deter-

Zero Correlation Zone (Cyclic Shift, Ncs)		Preamble Sequences per Root Sequence	Root Sequences Required per Cell	Root Sequences Re-use Pattern	Cell Range	
Signalled Value	Actual Value					
1	13	64	1	838	0.76km	
2	15	55	2	419	1.04km	
3	18	46	2	419	1.47km	
4	22	38	2	419	2.04km	
5	26	32	2	419	2.62km	
6	32	26	3	279	3.47km	
7	38	22	3	279	4.33km	
8	46	18	4	209	5.48km	
9	59	14	5	167	7.34km	
10	76	11	6	139	9.77km	Vendor - A Configurations for:
11	93	9	8	104	12.20km	
12	119	7	10	83	15.92km	LTE - 1800
13	167	5	13	64	22.78km	LTE - 800
14	279	3	22	38	38.80km	
15	419	2	32	26	58.83km	
0	0	1	64	13	118.90km	

PrachCS

Cyclic Shift of Root Sequence

Table 2.2: Cell range as a function of the Zero-Correlation Zone.

mining the actual signal level provided by the cell. Hence, the ultimate maximum cell range is determined as the more stringent limit among such constraints.

Within Vendor-A's infrastructure, the default configuration defines the use of Preamble format `0` (max. cell range of $14.5km$) for LTE-1800 frequency band and format `1` (max. cell range of $77.3km$) for LTE-800 frequency band. However, the current Vendor-A configuration of prachCS stands at 12 (max. cell range of $15.92km$) for the LTE-1800 and 13 (max. cell range of $22.78km$) for the LTE-800. It is very notable that in the latter scenario, upon referencing Table 2.1 and Table 2.2, the cell range dictated by the prachCS is significantly constrained compared to the threshold delineated by the guard-time parameter of the relative preamble format.

2.2 Potential Overshoot Scenarios

This Section builds upon the foundational technical concepts established in the previous Section, providing further elaboration of the possible consequent problem scenarios arising due to the overshooting phenomenon [12].

Before delving into the possible scenarios, it is important to define two key concepts which are often utilized in the following paragraphs. First, the term *cell range* is determined by the configuration of prachCS parameter and represents the maximum distance at which a UE should be able to connect to the cell. Hence, the cell range is usually referred to as the "logical" coverage area of the node. On the

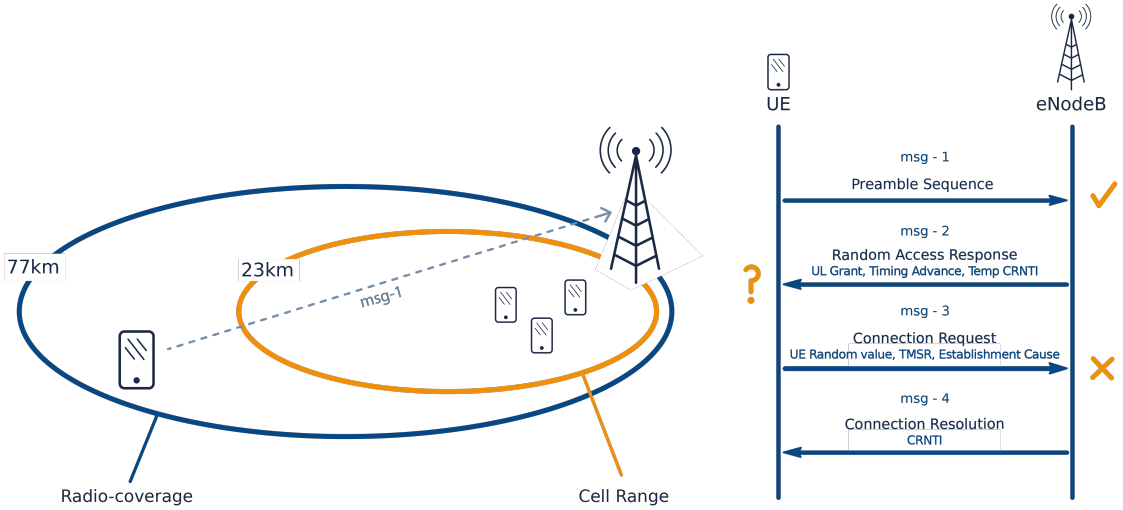


Figure 2.4: Overshooting phenomenon example.

other hand, the term *radio coverage* defines the geographical area within which the cell's antenna must provide reliable radio signal to users. This latter concept, determined by physical factors such as transmission power, antenna characteristics, and surrounding environment, is often referred to as the "physical" coverage of the node.

The overshooting phenomenon is illustrated in Figure 2.4. In the reported example, a LTE-800 cell has a maximum cell-range of 23 km, defined by its configuration of prachCS parameter. Anyway the actual radio coverage is 77 km. For this reason, although messages of type 1 sent by an overshooting UE are received by the cell, the significant time shift causes the preambles to be mistaken with another sequence among the possible time shifts allowed for that RSI. In this scenario, the eNB will respond with an incorrect Random Access Response (message-2), causing the RRC Connection Request (message-3) not to be received by the eNB. As a result, the connection attempt will fail, affecting the overall network performance.

In order to overcome overshooting issues, the idea is to force the cell to have matching configured values of cell-range and a physical coverage. Hence, two possible solutions are commonly implemented to meet such constraint. The first solution implies (physically) adjusting the antenna tilt, i.e. the angle at which the antenna is positioned relative to the ground, which alters the cell's coverage area. By decreasing the tilt angle the antenna's main radiation lobe is directed more towards the ground, reducing coverage in distant areas and potentially mitigating overshoot. This adjustment can help optimize coverage and minimize interference with neighboring cells. On the other hand, a second solution consists in a "logical" adjustment of the PRACH parameters, leading to an extended cell-range. Increasing the cell-range allows UEs located further away to successfully access the network without causing overshooting. However, this approach requires careful planning to ensure optimal performance and minimal interference with neighboring cells.
Both of these approaches aim to optimize the coverage area of the cell and mitigate overshoot effectively.

2.3 Root Sequence Index Collisions

This Section expands on the fundamental technical concepts introduced in the present Chapter (2), offering a general overview of the topic of RSI collisions that may arise in the network [5].

As explained in Section 2.1.1, the parameter prachCS determines the level of reuse of a RSI by specifying the number of cyclic shifts applied to the root sequence, thereby determining the number of preambles created from that RSI.

Regardless of the prachCS configuration, each cell requires 64 preambles, leading to the assignment of a set of possible RSIs for each cell.

The distribution of RSIs must consider several factors:

- the maximum cell range, as defined in Section 2.1.1;
- the number of RSIs needed per cell, depending on on the value of cyclic prefix (and thus on prachCS);
- the set of 838 RSIs is divided into three subsets, based on different usage scenarios: permanent network assignments, integration of new cells, and femtocells.

Currently, the distribution of RSIs does not follow a standard procedure. However it is important to note that in areas close to national borders the number of assigned RSIs may be reduced, resulting in some RSIs being left unused.

In Section 2.1, it is discussed that UE must undergo the LTE random access procedure to connect to a LTE network, establish or re-establish a service connection, perform HO, synchronize for uplink and downlink data transfers. The LTE random access procedure offers two distinct approaches: contention-based and contention-free (also known as dedicated). In the contention-free scenario, the eNB explicitly assigns a preamble sequence to the UE to prevent collisions. Conversely, in contention-based attempts, the UE randomly selects a preamble from the available set and transmits it to the node.

When neighbouring cells operate in the same frequency band and share the same RSI parameter, connected UEs may calculate the same preambles, resulting in increased occurrences of preamble collisions, commonly referred to as “RSI Collisions”. An example is depicted in Figure 2.5.
This issue can lead to failed service establishments or re-establishments, as well as failed handovers.

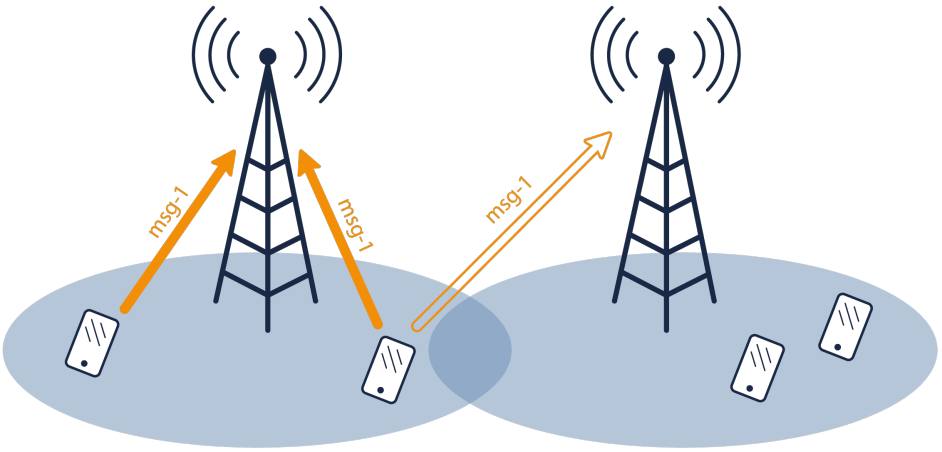


Figure 2.5: RSI Collision phenomenon example.

2.4 State-of-the-art Analysis

As mobile networks expand and the demand for seamless connectivity continues to surge, the optimization and automation of these networks have emerged as critical tasks. To this end, anomaly detection, performance forecasting, and self-healing capabilities are indispensable components of network management strategies. Leveraging advanced techniques such as machine learning (ML) and data-driven optimization, research is directed towards exploring innovative approaches to address the complexities inherent in modern mobile networks.

Recent studies have focused on the feasibility of deploying adaptive ensemble-method frameworks for modeling cell behavior, using KPIs to determine cell-performance status. The abundance of operational data within mobile networks presents an opportunity to detect anomalies and predict performance accurately. Advanced ML techniques applied to aggregated data from multiple sources enable the identification of anomalous behavior and the forecasting of network performance. Novel frameworks propose the aggregation of heterogeneous datasets and apply ML algorithms for diagnosing network issues. Pattern identification and time-series forecasting automatic algorithms efficiently detect spatial-temporal anomalies, predicting customer impact accurately.

The paper *On the feasibility of deploying cell anomaly detection in operational cellular networks* [2] introduces the concept of "self-healing" within the Self-Organizing Network (SON) framework. By utilizing KPIs, the adaptive ensemble-method framework demonstrates a practical solution of automation of the detection of cell anomalies with minimal computational overhead and detection delay.

Similarly, authors in paper *Automatic Root Cause Analysis for LTE Networks Based on Unsupervised Techniques* [4] propose an automatic diagnosis system for LTE networks, leveraging unsupervised techniques such as Self-Organizing Maps (SOMs).

In the paper *Big Data-driven Automated Anomaly Detection and Performance Forecasting in Mobile Networks* [7], leveraging the vast amount of data available in operational mobile networks, the proposed framework aggregates data from diverse sources, including configuration, performance, inventory, locations, and user speeds, applying ML algorithms for diagnosis and impact analysis.

Finally, the paper *A canonical correlation-based framework for performance analysis of radio access networks* [8] introduces a framework based on Canonical Correlation Analysis (CCA), which is a highly effective method for not only dimensionality reduction but also for analyzing relationships across different sets of multivariate data. It provides a case study on energy-saving through cell shutdown in LTE networks, demonstrating the effectiveness of CCA to analyze the impact of capacity cell shutdown on the KPIs of coverage cell in the same sector.

2.4.1 Current Operation Methodologies and Technologies

At Telefónica, the procedure for identifying overshooting issues relies on analyzing the out-of-range signal reception rate of individual cells. The latter can serve as an indicator of the severity of the overshooting occurrence.

Initially, a local filtering process is conducted within specific geographical regions of Spain using a partially automated system implemented with the PowerBI[1] tool (in Figure 3.1). This system allows for the selection of telecommunication equipment between Vendor-A or Vendor-B (as specified in Section 1.2) and the input of the current "expected cell size", which is defined by the current configuration of PRACH parameters. Subsequently, the tool generates a list of cells along with their significant characteristics such as current prachCS, out-of-range signal reception rate (%), calculated-prachCS, and the corresponding calculated expected cell size. Note the term "calculated", which refers to the recommendations for parameter adjustments that the tool provides based on current configuration data. PowerBI users have the ability to filter the information geographically and by setting thresholds for the out-of-range signal reception rate. The PowerBI tool is configured to update its results daily, ensuring the integration of the latest available data into the analysis process. However, it operates with a temporal depth of 14 days, i.e. insights and visualizations presented are based on 2-weeks time span of historical data.

Once overshooting issues are identified, each case is individually assessed, and decision-making is carried out manually by optimization engineers. This process takes into consideration various factors including the topological characteristics of the cell (e.g., location, surrounding environment, network neighborhood). Additionally, consulting appropriate KPIs, optimizators assess the actual impact of these issues on user experience and network performance. One such KPI is the successful message-3/message-1 accesses rate (%), which reflects the ratio of the number of success-

[1] *Power BI* is a business analytics tool created by Microsoft, aiming to facilitate the visualization and analysis of organizational data. It enables users to connect with diverse data sources, such as Excel spreadsheets, databases, and cloud services, consolidating information from disparate origins into a unified and coherent interface.

ful RRC signaling (message-3), transmitted consequently to the reception of UE's preambles, over the total number of preambles (message-1). Comparability between these two quantities is crucial for representing cells that are functioning optimally.

Conversely, Telefónica's Radio Optimization department is currently transitioning certain operations to Databricks[2], an analytics platform built on Apache Spark.

This project marks one of the initial steps in such technological transition towards the field of data analysis.

Databricks, in comparison to the former Telefónica's platform, offers greater potential in scalability, advanced analytics, and real-time processing. Additionally, it provides higher programming flexibility and integration, enabling the utilization of big data databases. Its distributed architecture enables to effectively handle massive data volumes, making it ideal for organizations dealing with big data or complex data processing tasks. Additionally, Databricks support multiple programming languages (e.g. Python, R, SQL) and libraries commonly used in data science and ML.

[2] *Databricks* is a unified, open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. The Databricks Data Intelligence Platform integrates with cloud storage and security in your cloud account, and manages and deploys cloud infrastructure on your behalf.

3. ML problem framing, Data processing and Exploratory Data Analysis

This Chapter delves into the initial phase of a project, which is dedicated to ML problem framing, data processing and Exploratory Data Analysis (EDA).

Mobile network data ingestion, cleaning, and aggregation pose significant challenges in terms of volume, diversity, and reliability. More in details, mobile network management data include Performance Management (PM), Inventory Management (IM) and Configuration Management (CM) data. IM data primarily consists of static equipment details, such as base station specifications and infrastructure inventory. Within IM set, the subset of CM data encompasses crucial radio access network parameters like eNB/cell ID, frequency bands, and neighbor relations. Conversely, PM data includes a multitude of dynamic counters, generating raw measurements across the network.

This phase involves ingestion of mobile network data, in particular PM data, from different sources, followed then by EDA process. The ultimate goal is to understand and identify relevant characteristics, which serve as indices of connection quality, network performance and user experience, hence indicative of overshoot problems.

For this purpose, as anticipated in Section 2.4.1, the initial step involves extracting information from Telefónica's Power BI. Specifically, this tool has been previously designed by Telefónica's Radio Optimization department to identify cells marked as cell-range problematic for specific dates.

Successively, this information enables for further query and extract relevant information for these instances.

The work of this thesis explores the utilization of Azure Databricks, a managed version of the Databricks platform within the Azure cloud environment. Specifically, it leverages Spark[1] sessions within Azure Databricks and the integration of PyS-

[1] *Spark* is an open-source distributed computing system for programming entire clusters with implicit data parallelism and fault tolerance. It's designed for big data processing and analytics, offering in-memory computation and supporting various programming languages like Scala, Java, Python, and R.

park[2] and Kusto Query Language[3] (KQL) queries for efficient data manipulation and analysis. Further details are provided in Appendix D.

3.1 ML Problem Framing

In a ML life cycle, the fist step is the so-called "ML problem framing". It consists in reframing the business problem under study into a ML problem. This first stage involves the articulation of the problem statement, the identification of target variable, pertinent features, context and limitations of the problem domain. The ML problem must be framed taking into account the business objective, the theoretical framework, and the current state-of-the-art.

As presented in Section 1.2.1, the work of this thesis mainly focuses on cell-range overshooting use case. First, the objective is to implement a tool which, given some KPIs as input variables, is able to identify the problematic occurrences related to overshooting among the network.

With this under consideration, the business problem can be designed as a binary classification problem. Classification techniques enable the categorization of instances based on their characteristics. This implies that cells will be classified according to the values of their KPIs, aligning precisely with the objective.

Firstly. it must be clarified that, although a cell may operate on multiple frequencies in real networks, in KQL queries each cell's unique identifier represents a cell associated to a specific frequency. Hence, for nomenclature reasons, in the entire thesis the term "cell" will refers to a cell operating in one single, particular frequency.

Therefore, the dataset will comprises of a cell (sample) in each entry described by a set of KPIs (features) and a binary label (`1` for overshooting issue or `0` for no overshooting issue). Each entry will correspond to a specific time frame during which the KPI values are extracted for a cell operating at a specific frequency.

Accordingly, the *problematic* class (label `1`) represents the cells experiencing cell-range overshooting issues. On the other hand, the *random* class (label `0`) identifies the cells not experiencing cell-range overshooting issues. The term "random" aims to highlight the uncertainty about the cell status. The reason behind this is to create as comparison class a realistic network model comprehensive of cells well-behaving as well as cells experiencing other types (and at different severity levels) of problems. In this way, it is ensured robust training of the ultimate system to distinguish overshoot issues within real - hence heterogeneous - network environment.

[2] *PySpark* is the Python API for Apache Spark, enabling to leverage Spark's distributed computing capabilities using Python.

[3] *Kusto Query Language* (KQL), developed by Microsoft, is a query language used to interact with Azure Data Explorer for analyzing large volumes of data.

3.2 Identification of Problematic Cells

Introducing the Data Processing stage, the initial step is the Data Acquisition which focus on the identification of cells experiencing overshoot problems. To this end, all network cells belonging to Vendor-A's equipment (in the territory of Spain) are monitored.

For the purpose, a Power-BI tool, previously developed in Telefónica, is leveraged, which enables the detection of cells exhibiting an out-of-range signal reception rate higher than a given threshold. In Figure 3.1, the Power-BI tool provides a geographical visualization of selected cells and among other information, the actual prachCS and expected cell size (expected cell range corresponding to current prachCS) as well as the calculated cell-size and PRACH parameters proposed configuration in order to better serve the detected overshooting terminals. Details on the Power-BI tool can be found in Section 2.4.1.

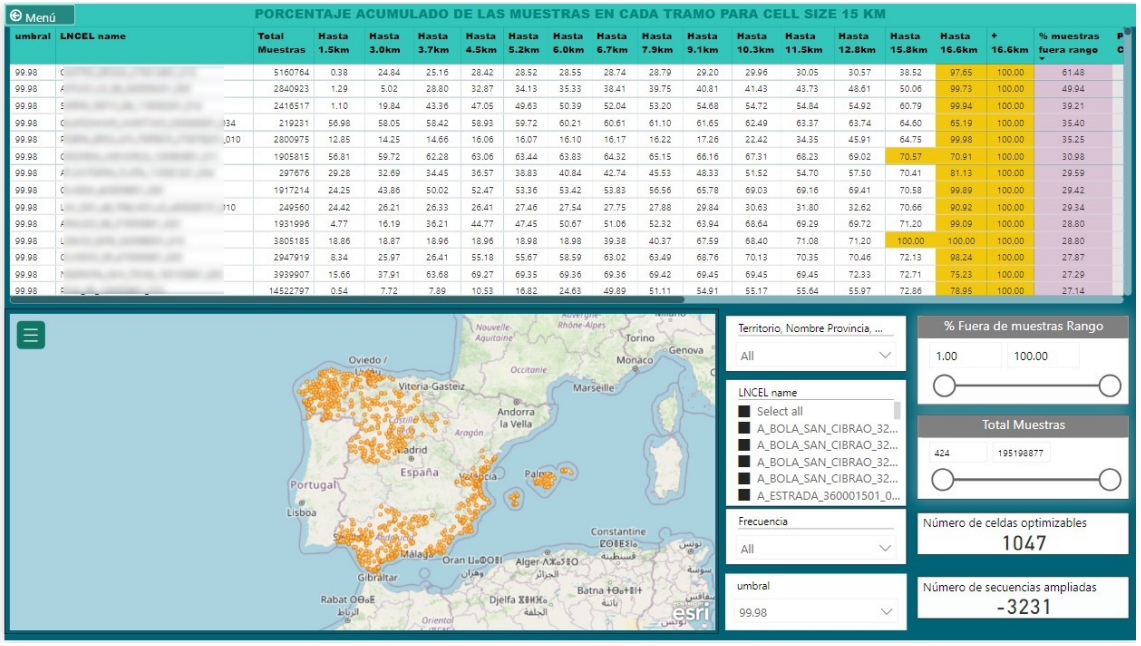


Figure 3.1: Power BI tool screen capture example.

During the stage of Data Processing, this Power-BI tool is exploited for the identification of cells experiencing (in a particular date) problems related to overshooting.

Hence, filtering the result from such Power-BI tool using a minimum threshold of 2% for the out-of-range signal reception rate, a set of 1559 records is constructed, serving as the *ground truth* [4] for the *problematic* class. Each record represents one cell's conditions in a given period of time during which it was experiencing overshooting problems. The data collection spans the time window from March 12th

[4] *Ground truth* is defined as the labels associated with the data points that indicate whether the data represents well-known problematic cell or not.

to March 14th 2024, including the endpoint dates. Furthermore, the 2% filter allows to exclude occurrences of overshooting with very low severity.

In contrast, the *random* class is composed by approximately 1540 cells, which are randomly selected among the remaining cells over the entire Spanish territory covered by Vendor-A. Also for this class, the same time window is used for data collection.

In Figure 3.2 the geographical distribution of *Training* dataset is depicted. Figure 3.2 provides a capture of an interactive visualization (HTML file) generated using `plotly` open-source library in Python. For sake of privacy considerations for sensible information involved, only the captures of such visual outputs will be presented in this thesis.

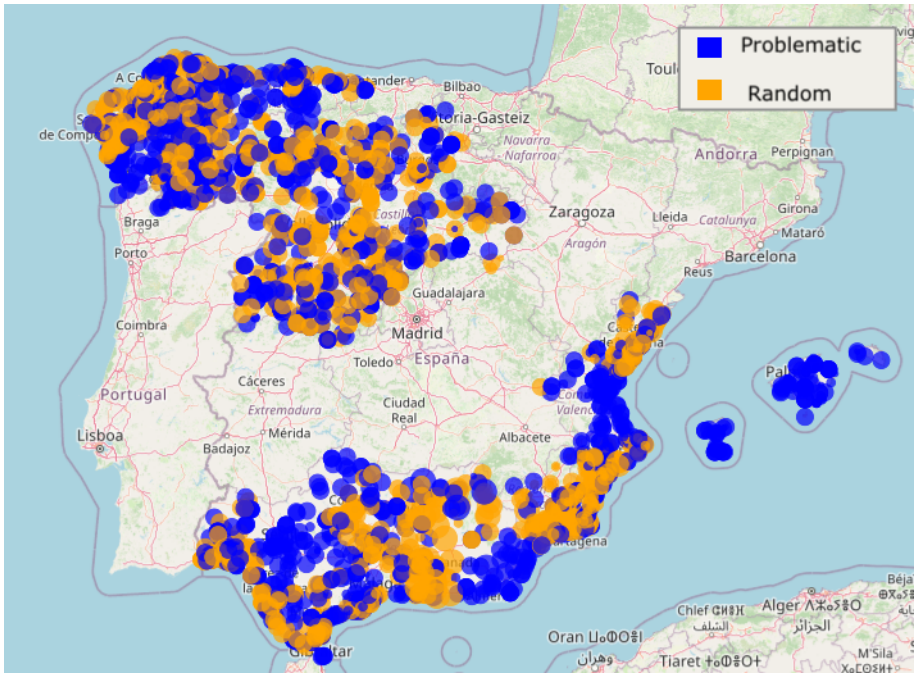


Figure 3.2: *Training* dataset geographical distribution.

It is important to clarify that the same cell observed in a different time window could be in a completely different status and experiencing different conditions. Therefore, the timing of the observation is fundamental. Beside that, during the Testing Phase (Chapter 4), cells that have already been seen in the training dataset are excluded from the testing datasets.

Finally, it is essential to specify that mobile networks exhibit a stringent dependency on the movements and behavioral trends of individuals. Variations in seasons or temporal periods significantly influence the distribution of population densities and mobility patterns, thereby exerting notable fluctuations on network traffic dynamics. Consequently, the operational performance of these networks is intricately tied to temporal factors.

3.3 Key Performance Indicators

The full set of Key Performance Indicators (KPIs) encompasses various aspects of LTE-based access network, including Integrity, Accessibility, Retainability, and Usage. For example, they can quantify the success rate of connection establishment, the failure rate of Handover (or Handoff) (HO) procedures, or the drop rate of Voice over LTE (VoLTE) connections. Traditionally and presently, these KPIs constitute the primary metrics examined to assess network health and cell status.

The KPIs are defined by official formulas, which are functions of several variables known as counters. The KPIs capture diverse network metrics, often with high granularity (e.g., every 15 minutes). This results in vast volumes of time-series data collected at various levels such as cell and neighbor relations. For instance, counters track metrics like cell throughput at the cell level and handover statistics per neighbor relation.

Considering the purpose of the project, prior knowledge and under the supervision of optimizers from Telefónica, a restricted list of 49 KPIs is extracted.

These KPIs, relative to individual cells and 4G technology, span various aspects of network operation, including the RACH procedure, access failures, drop ratios, throughput-related integrity metrics, and HO procedures. In Appendix A a full descriptive table of considered KPIs is provided.

3.3.1 KPIs Time Aggregation

As previously explained, the observed KPIs are very dynamic parameters that efficiently represent the instantaneous (15 minutes window) condition of a network cell.

In order to have comprehensive overview of network cell status, a wider time window observation is more representative of the overall behavior of each KPI. In this direction, taking into account the fact that overshooting issues are currently resolved after roughly a week from the detection (since it is partially manual procedure), it is reasonable to consider the KPIs as time-aggregated values of three days over the period in which the cell was experiencing the issues.

Aggregation is automatically performed in Databricks by PySpark computational resources. The time aggregation consists in the computation of the KPI's official formula using as input variables the counters accounting for the entire given time window.

Moreover, three-day aggregation approach addresses also the issue of NaN values which sometimes are returned for some KPI in a specific 15-min output. Among other reasons, this is due to the dynamic network management systems which frequently involves cell reconfiguration or cell resizing, which could translate in null values for small time windows. However, aggregation over a larger window of three days resolves such challenge.

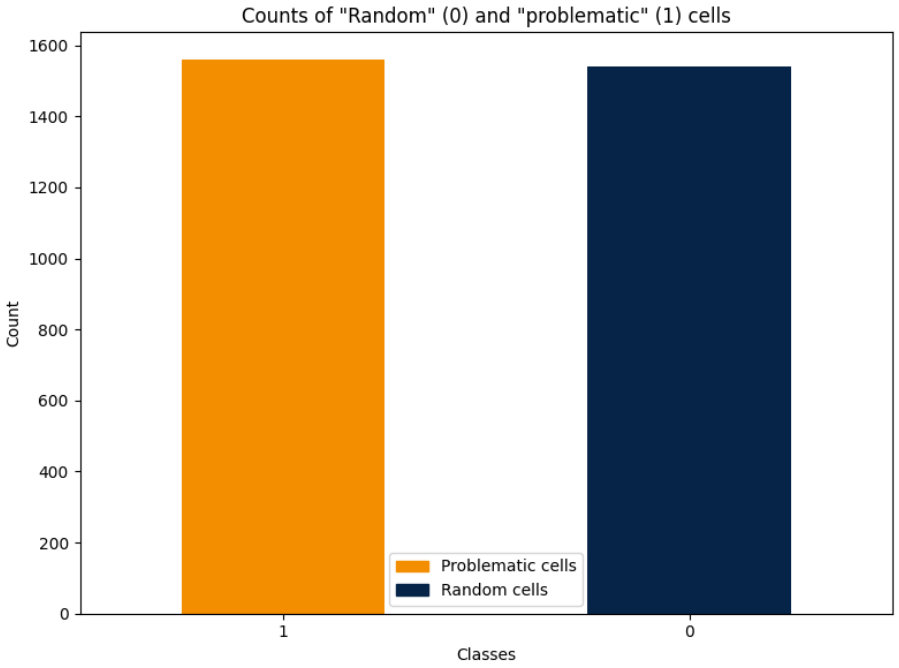


Figure 3.3: Dataset division in *Problematic* and *Random* classes.

3.4 *Training* Dataset

In recapitulation, the dataset consists of approximately 3100 samples, roughly equally divided into *problematic* and *random* classes (Figure 3.3) and 49 features, which are three-day aggregated values of KPIs in the time window from March 9th to March 12th, 2024.

For sake of clarity and consistency, the dataset in question is henceforth referred to as the *Training* dataset. This nomenclature will be consistently employed throughout the remainder of this thesis to denote its role as the primary dataset for training purposes.

3.5 Exploratory Data Analysis

The primary objective of the EDA process is to meticulously examine the dataset in depth, analyzing its various attributes, patterns, and distributions, thereby gaining comprehensive insights and understanding into the underlying structure and characteristics of the data.

A well-balanced dataset holds significant importance for classification algorithms, as it guarantees the effective learning of predictive patterns across all classes, mitigating biases toward dominant ones. Conversely, in datasets skewed towards one class, the model risks favoring the majority class, thereby compromising its ability to effectively discern minority classes.

However, beyond mere class balance, it is crucial to delve into the homogeneity of data records. Specifically, in this case, each record corresponds to a cell, making it

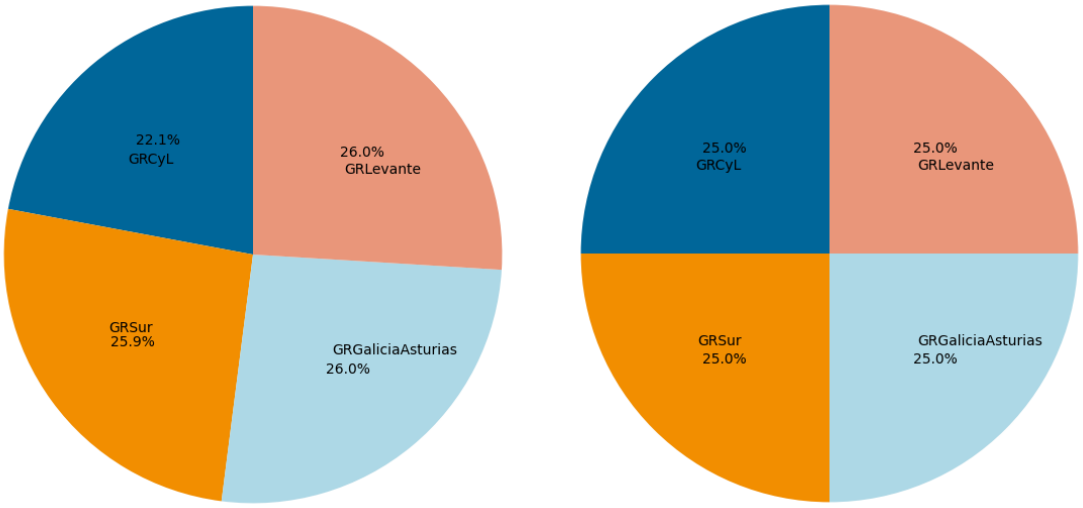
imperative to study not only the overall balance but also how different "types" of cells are represented within the *problematic* and *random* classes.

Therefore, the EDA process is divided into a first data examination related to the network's topology and physical aspects, and a second further statistical analysis of the KPI measurements.

3.5.1 EDA of Cell Nature

The preliminary analysis of data primarily centers around examining various network-related topological and physical factors.

Although the classification model depends on KPI features for pattern recognition and issue identification, the observation of network aspects adds depth and context to the problem under examination. The analysis of network-related information has the potential to reveal hidden correlations, anomalies, or trends not emerging from just KPI data. Furthermore, the inclusion of network data facilitates a cleansing process, enhancing the overall integrity and balance of the dataset.



(a) Problematic class (b) Random class

Figure 3.4: Intra-class distribution of cells across geographical regions.

To begin with, it is crucial for the dataset to be representative and homogeneous across the entire geographical territory. The objective is to prevent situations where, for instance, problematic cells are concentrated in specific areas, which could introduce bias in classifying their behavior within those regions rather than focusing on their performance issues.

In Figure 3.4, it is reported, within each class, the distribution of samples across different regions. The dataset exhibits near-equal distribution across the Spain-Vendor-A territory.

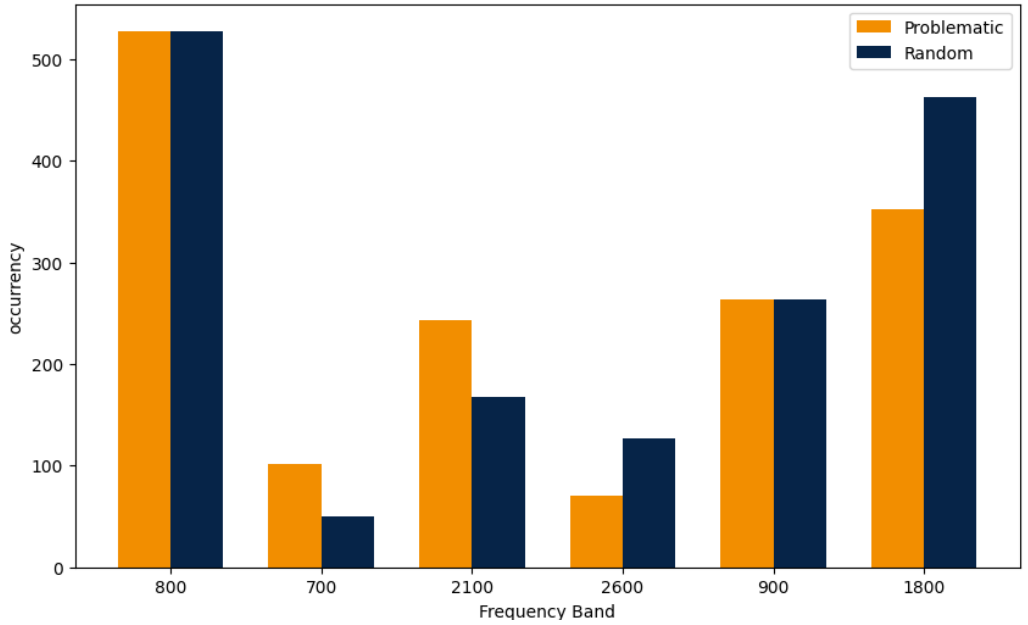


Figure 3.5: Comparison intra-class distribution of cells in terms of frequency bands.

Also, Figure 3.5 demonstrates almost symmetrical allocation of records across the possible frequency bands, meaning that for each cell belonging to *problematic* class there is (approximately) an instance of the same frequency band in the *random* class.

However, it must be recognised that the dataset is not equally divided across the bands, exhibiting better representation within bands 800 MHz and 1800 MHz. This aspect is further discussed and studied in Section 4.2.

Subsequent investigations were conducted to analyze the frequency band distribution of problematic cells relative to the frequency bands employed by co-located cells. The initial unverified hypothesis posited that overshooting incidents predominantly occur in "external" cells, while occurrences in "internal" cells, co-sited with multiple higher frequency bands, are rare. However, as illustrated in Figure 3.6, the findings indicate no discernible correlation between the occurrence of overshooting incidents and the frequency bands utilized within the cell site. Notably, a similar pattern is observed when comparing the two cell classes.

Furthermore, similar observation is carried out considering the size of cells included in the dataset. In particular, the prachCS parameter is exploited to calculate the expected cell range (with respect to Table 2.2). As shown in Figures 3.7a 3.7b, is mainly composed by records of cells with expected cell size of 30km and consistent number of 10km/15km. Nonetheless, the dataset exhibits an overall acceptable inter-class symmetry.

In conclusion, although the dataset presents a comparable composition within each class without regional disparities, cells with varying characteristics are represented non-uniformly.

With the aim of acknowledging these imbalances, during the Testing Phase (in

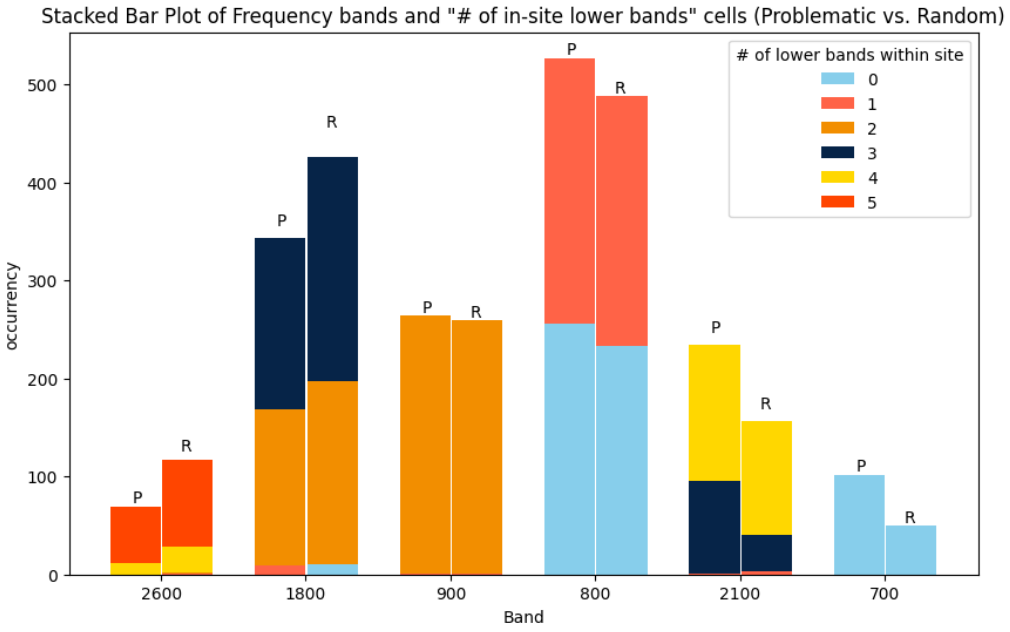
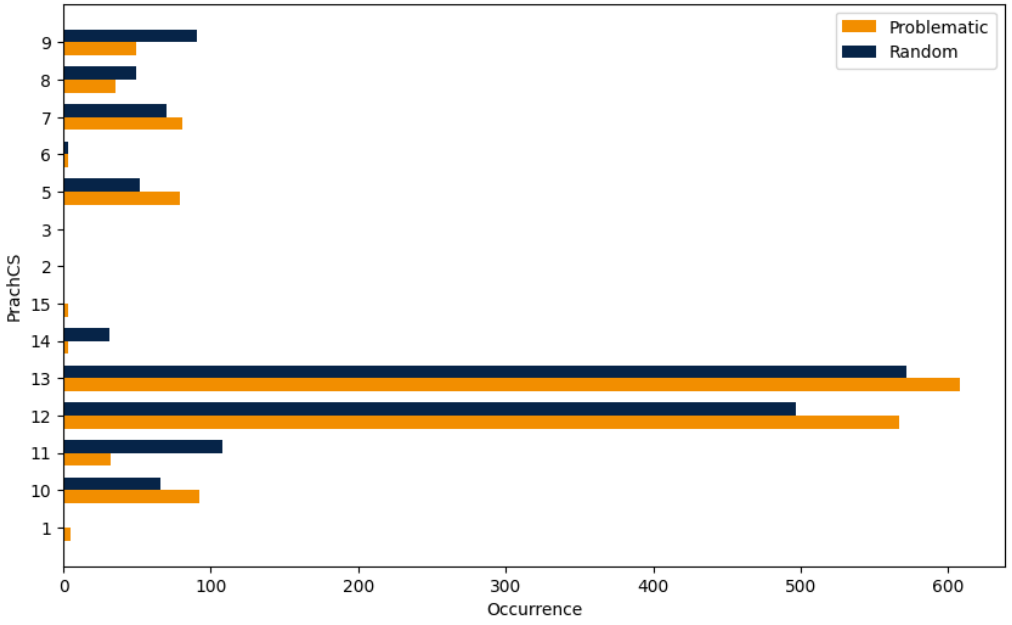


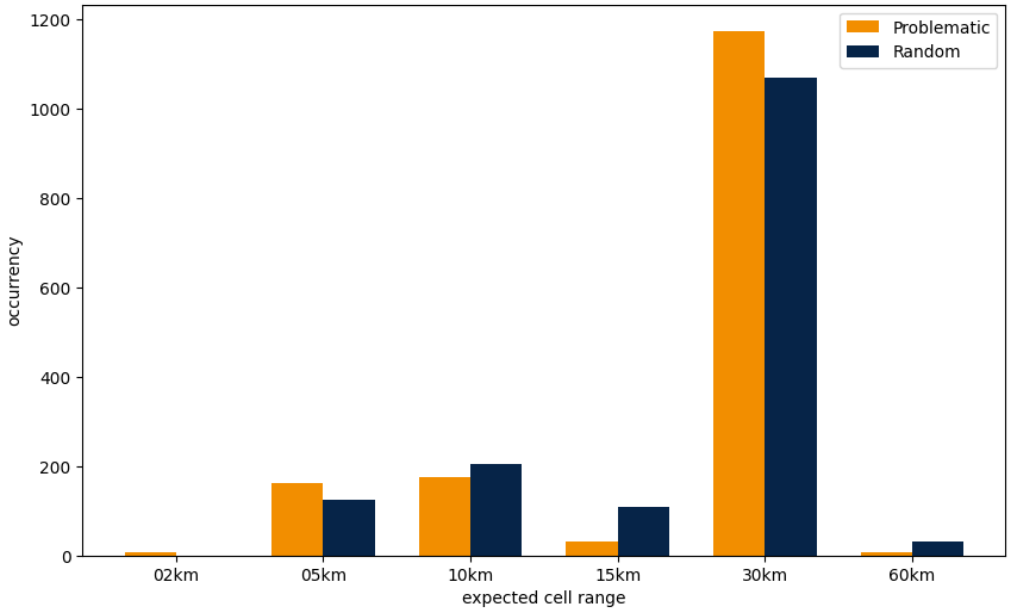
Figure 3.6: Frequency bands distribution per different levels of internal placement with respect to co-site cells. Comparison of *problematic* (P) and *random* (R) classes.

Chapter 4), results are analyzed for cells with different characteristics to detect possible biases in classification performance.

This comprehensive approach enhances the robustness of the classification model, enabling it to provide reliable insights into the network performance across diverse parameters.



(a) PrachCS



(b) Expected cell-range

Figure 3.7: Distribution of cells in terms of prachCS and expected cell-range.

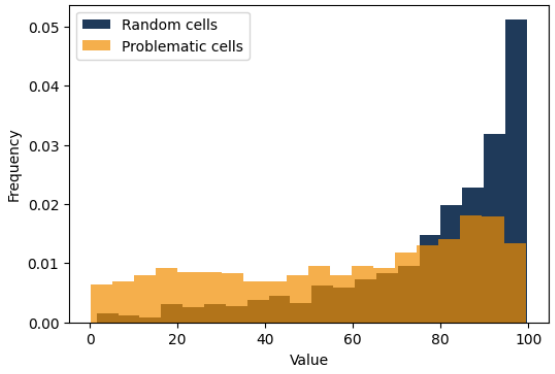
3.5.2 KPI EDA

The preliminary data analysis continues focusing on the KPIs included in the dataset.

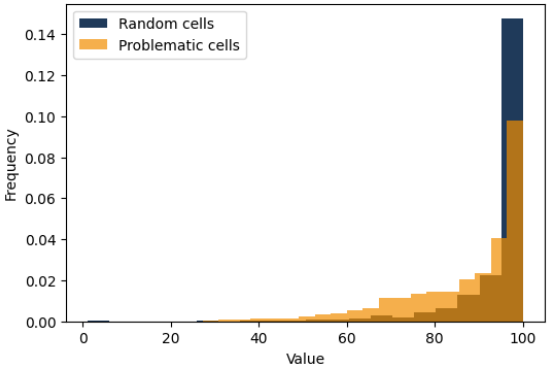
Although the initial selected list comprises of 49 KPIs, not all these features contribute equally to the classification process. Further analysis can allow identify the most valuable KPIs, thus reducing both memory and computational costs.

For the purpose of gaining insights, an examination of the distribution of the KPIs under study is conducted.

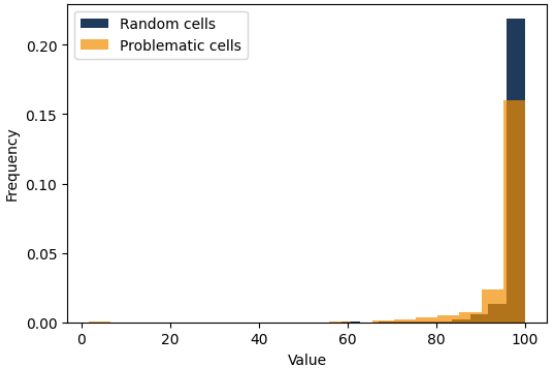
The histograms presented in Figure 3.8 facilitate a comparative analysis of the distribution of a specific KPI across the two classes. Figures 3.8a and 3.8b illustrate the impact of the overshoot phenomenon on the ratios of message-3 to message-1 and message-2 to message-1, respectively. These findings are consistent with theoretical expectations, as overshooting often results in a higher frequency of connection establishment attempts that remain incomplete. Similarly, in Figure 3.8c and 3.8d, it is noticeable the random cells' better performance, which translates in higher success rate of HO execution phase and higher average downlink throughput.



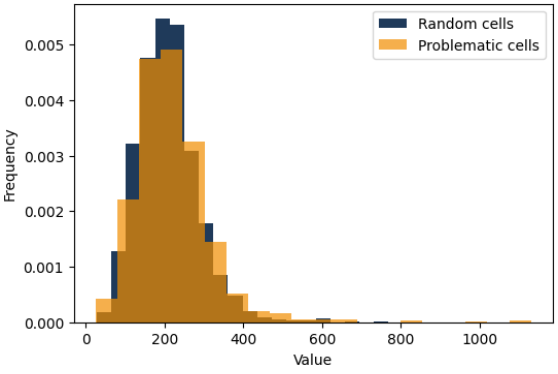
(a) Message-3 / Message-1 Success Rate



(b) Message-2 / Message-1 Success Rate



(c) HO Execution Phase Success Rate



(d) Average IP Scheduled DL Throughput

Figure 3.8: Histograms of Distribution of four KPIs.

In line with this, the Mann-Whitney U Test[5] [9] performed on each KPI distribution returns very low p-values which confirm the significant statistical difference between

[5] *Mann-Whitney U Test* is a non-parametric statistical tests used to assess whether two indepen-

the *random* and *problematic* populations. In Appendix B a table reports the results of Mann-Whitney U tests for the full list of considered KPIs.

In this regard, Feature Importance analysis (next Section 3.6) enables the ranking of KPIs based on their level of information regarding overshoot problems.

3.6 Feature Importance Analysis

The Feature Importance Analysis procedure employs ML techniques to efficiently determine the relevance of each KPI. Its aim is to quantitatively evaluate the impact of each KPI on the predictive capability of our models. This method not only corresponds with our current understanding but also offers a data-driven validation of our hypotheses.

Feature Importance analysis process is carried out implementing two Tree-based classification algorithms, namely Random Forest (RF) and Extreme Gradient Boosting (XGB). Therefore, prior to exploring Feature Importance analysis and its results, such family of algorithms is introduced for sake of better clarification.

3.6.1 Tree-based Classification

Tree-based classification algorithms are ML algorithms that use a decision tree structure to classify instances into different classes based on their features. Decision Trees (DTs) recursively split the feature space into regions, with each split based on thresholding one selected feature [11].

Figure 3.9 offers a visual representation of the Decision Tree (DT) classifier applied to the dataset used in this study. The output displays the structure of the decision tree, including nodes and branches. By setting a maximum depth to a single branch, one can better appreciate the logical criteria employed in the tree-based decision-making process. Each node represents the split criteria, including the KPI used for splitting and the threshold value, while the branches indicate the possible outcomes of that decision. The leaf nodes of the tree represent assigned class labels. While DTs are advantageously easy to interpret and visualize, a side effect is their proneness to overfitting. Thereby, they may not perform well on high-dimensional data. This is one of the reasons why more complex and highly-accurate tree-based classifiers are implemented in this study.

dent samples originate from the same population or have different population medians. It yields the U statistic value, which measures the extent of difference between the two samples, and calculates the p-value. Typically, a low p-value (usually below 0.05) indicates a significant difference between the populations, while a high p-value suggests little difference.

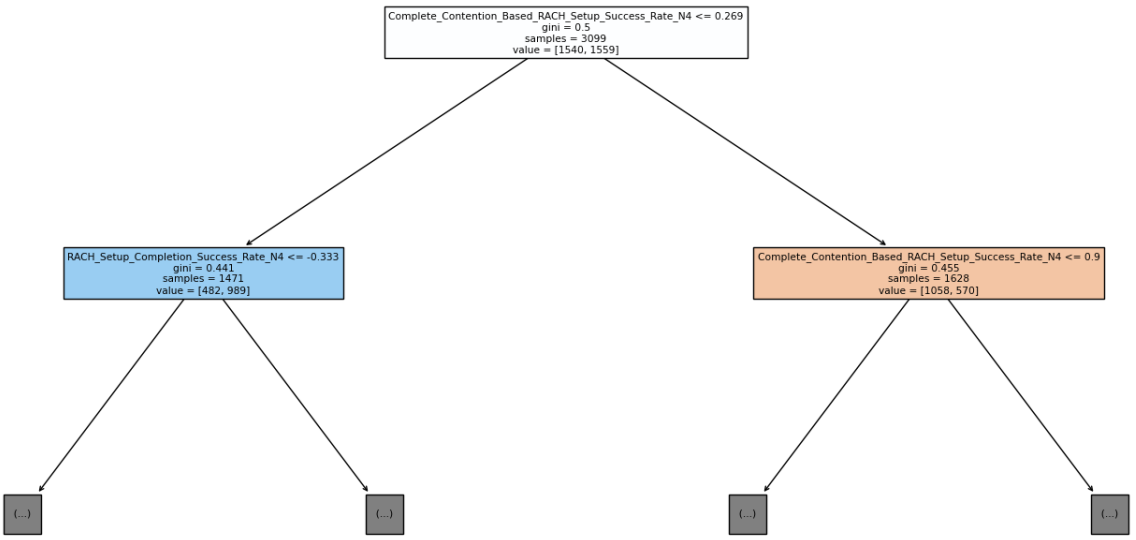


Figure 3.9: Decision Tree Classification.

Random Forest (RF) and Extreme Gradient Boosting (XGB) are widely used machine learning algorithms for classification tasks. Their ability to effectively manage complex patterns and high-dimensional data makes them such robust algorithms against overfitting and noise [13].

Random Forest algorithm is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data and features.

On the other hand, Extreme Gradient Boosting is an optimized implementation of gradient boosting (another ensemble learning technique that builds a sequence of decision trees) that includes additional features such as regularization, sparsity awareness, and parallel computing.

RF and XGB offer distinct advantages over other classification algorithms like Logistic Regression (LR) and Support Vector Machine (SVM). For instance, LR relies on the assumption of a linear relationship between features and the target variable, which might not always hold true. SVM's performance can be influenced by the choice of kernel function, often necessitating meticulous hyperparameter tuning.

3.6.2 Feature Importance Analysis

With the aim of analysing the impact of each KPI in the classification, Feature Importance methods allow to assign scores to KPIs, which are measures of their contribution to the predictive power of the model.

In order to build the models, the two classification algorithms, Random Forest (from

`scikit-learn` Python library) and XGB (from `xgboost`) are trained, cross-validated and tested on the dataset previously presented in Section 3.4.

Both classification algorithms are trained performing K-fold cross-validation with three folds (K = 3) on the whole dataset. Successively, Feature Importance analysis is carried out.

In case of Random Forest, the attribute `feature_importances_` provides a measure based on the Gini impurity[6] [6]. The importance score for a KPI is calculated by averaging the decreases in Gini impurity across all decision sub-trees in the forest.

On the other hand, the XGB offers an equivalent attribute (so-called `feature_importances_`) which provides similar information, although not directly using Gini impurity. It calculates the gain importance of each KPI, i.e. the average gain across all the sub-trees in the model, where the gain is defined as the reduction loss achieved by splitting on that feature.

In Figures 3.10 and 3.11, the graphs show sorted KPI importance scores for RF and XGB. Whereas the sorting order differs, the most important KPIs are highly placed in both results. It must be noticed that both XGB and RF consider the two most important KPI as significantly more informative, while the subsequent KPIs are deemed to have comparable (lower) relevance among themselves.

Finally, the results obtained by Random Forest and XGboost are combined to retrieve a unique list of features. In Table 3.1 the resulting list of KPIs, combined respecting the descending importance order for both RF and XGB, is reported.

KPI
Complete_Contention_Based_RACH_Setup_Success_Rate_N4
RACH_Setup_Completion_Success_Rate_N4
Exitos_de_la_fase_de_ejecucion_del_handover_N4
S1_Initial_Context_Setup_Attempts_N4
Initial_E_RAB_Accessibility_N4
RRC_Connection_Setup_FR_N4
E_RAB_Setup_Attempt_N4
Number_of_successful_Intra_eNB_Handover_completions_per_neighbor_cell_relationship_N4
Number_of_successful_Inter_eNB_Handover_completions_per_neighbor_cell_relationship_N4
HO_Success_Ratio_intra_eNB_N4
RRC_Connection_Setup_Attempts_N4
Average_MCS_used_for_TB_transmission_using_Spatial_Multiplexing_transmission_N4
RRC_Connection_Re_establishment_Attempts_HO_fail_N4
Exitos_de_handover_N4
Percentage_of_PDSCH_transmissions_using_Low_MCS_Codes_MCS_9_N4
PRB_usage_per_TTI_DL_N4

[6] *Gini Impurity* is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution in the dataset.

KPI
Inter_Frequency_HO_Success_Ratio_N4
Average_PDCP_layer_active_cell_throughput_DL_kbps_N4
Average_RSSI_for_PUSCH_dBm_N4
Averaged_IP_scheduled_Throughput_in_DL_QCI1_kbps_N4
RACH_Setup_Attempts_N4
PRB_usage_per_TTI_UL_N4
E_RAB_Setup_SR_N4
HO_Success_Ratio_intra_eNB_N4
Intra_Frequency_HO_Success_Ratio_N4
Maximum_Active_UEs_with_data_in_the_buffer_per_cell_UL_N4
Percentage_of_PDSCH_transmissions_using_High_MCS_Codes_MCS_20_N4
RRC_Connection_Re_establishment_rejection_ratio_N4
Maximum_Active_UEs_with_data_in_the_buffer_per_cell_DL_N4
Initial_Context_Setup_Failure_Ratio_due_to_Failed_Radio_N4
Number_of_Inter_eNB_Handover_attempts_per_neighbor_cell_relationship_N4
Radio_Bearer_Drop_Ratio_N4
Averaged_IP_scheduled_Throughput_in_UL_QCI1_kbps_N4
E_RAB_Drop_Ratio_User_Perspective_N4
Total_E_UTRAN_RRC_Connection_Re_establishment_Failure_Ratio_N4
Average_RSSI_for_PUCCH_dBm_N4
MAX_PRB_usage_per_TTI_UL_N4
HO_Preparation_Success_Ratio_intra_eNB_N4
Exitos_de_la_fase_de_preparacion_del_handover_N4
Number_of_Inter_eNB_Handover_failures_per_cause_per_neighbor_cell_relationship_N4
MAX_PRB_usage_per_TTI_DL_N4
Number_of_Inter_eNB_Handover_preparations_per_cause_per_neighbor_cell_relationship_N4
E_RAB_active_drop_ratio_with_data_in_the_buffer_due_to_RNL_Radio_Connection_with_UE_Lost_N4
Number_of_failed_Inter_eNB_Handover_preparations_per_cause_per_neighbor_cell_relationship_N4
Average_PDCP_layer_active_cell_throughput_UL_kbps_N4
Number_of_failed_Inter_eNB_Handover_preparations_per_neighbor_cell_relationship_due_to_failures_in [...]

Table 3.1: Ordered list of most important KPIs resulting from RF and XGB.

Successively, the resulting output list of KPIs is exploited to build the classification models.

The cross-validation methodology is employed to further conduct Feature Importance Analysis (more details in Section 3.6.2) and identify the optimal set of features, i.e. the smallest subset of most important KPIs allowing to achieve the highest accuracy scores.

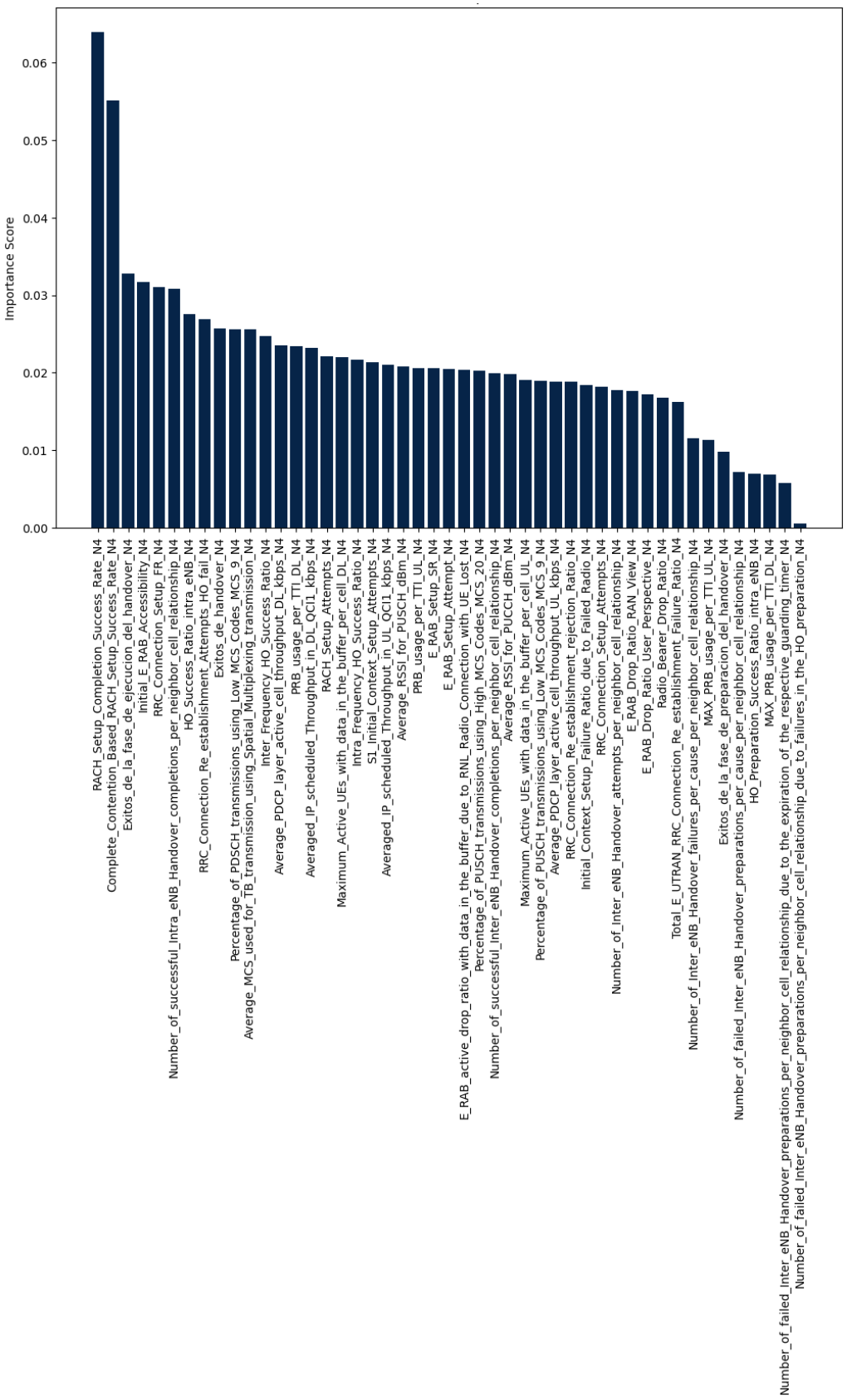


Figure 3.10: Random-Forest Feature Importance sorted scores.

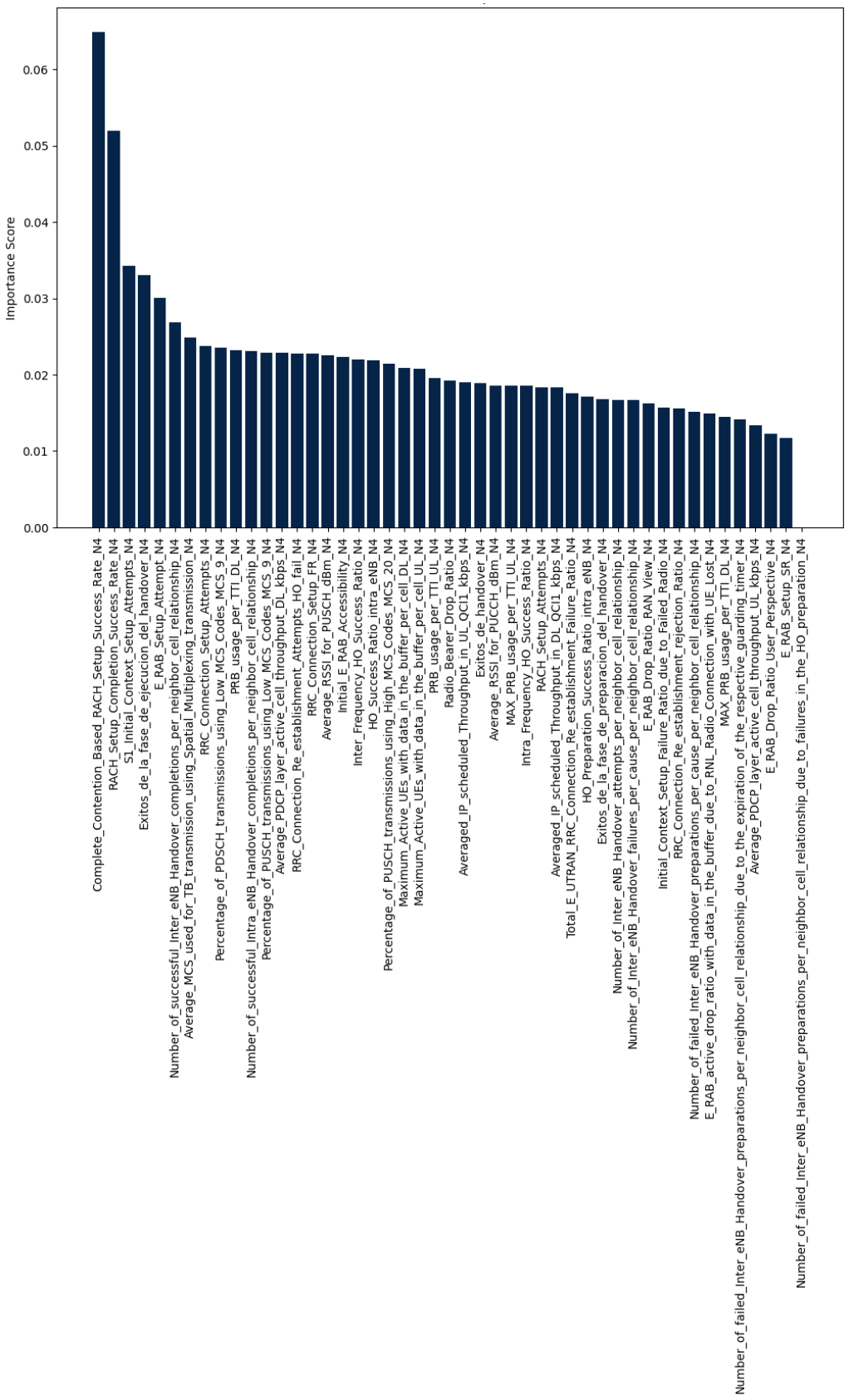
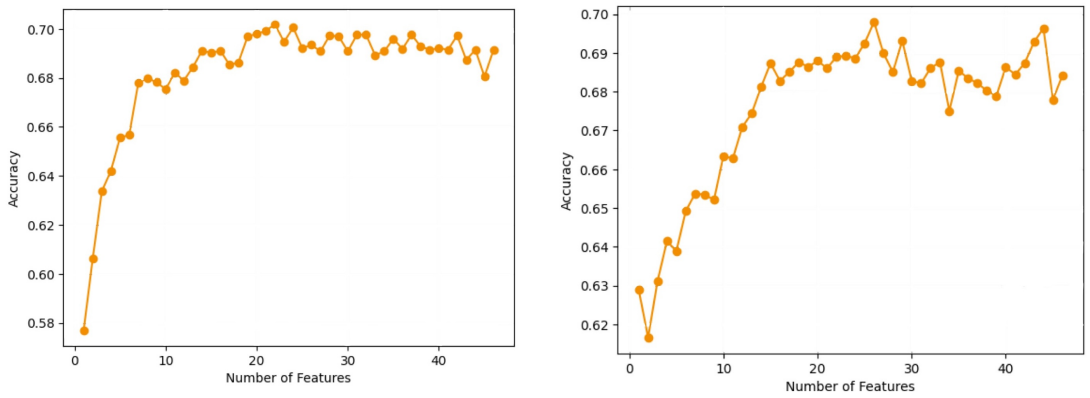


Figure 3.11: XGBoost Feature Importance sorted scores.



(a) Random Forest (b) XGBoost

Figure 3.12: Accuracy vs. number of most important KPIs used.

Cross-validation technique (`cross_val_score` from `scikit-learn` Python library) is implemented in both cases on the whole dataset, in order to evaluate how well the model generalizes to unseen data and can detect issues such as overfitting. This technique involves dividing the dataset into 5 subsets, known as folds, and iteratively training the model on a subset of the data while using the remaining subsets for validation.

As results, Figure 3.12 provides the behavior of accuracy of the two classifiers against the number of features considered (sorted in descending order of importance). It is visible that the accuracy drastically increases using only the first most important KPIs. Therefore, in Table 3.1, the optimal subsets of KPIs (i.e. the subsets allowing to reach the highest accuracy) respectively for the models are highlighted with colors: the KPIs considered by RF are in orange (23 out of 49); the selection of XGB is in blue (25 out of 49); the KPIs common to both are indicated in bronze brown.

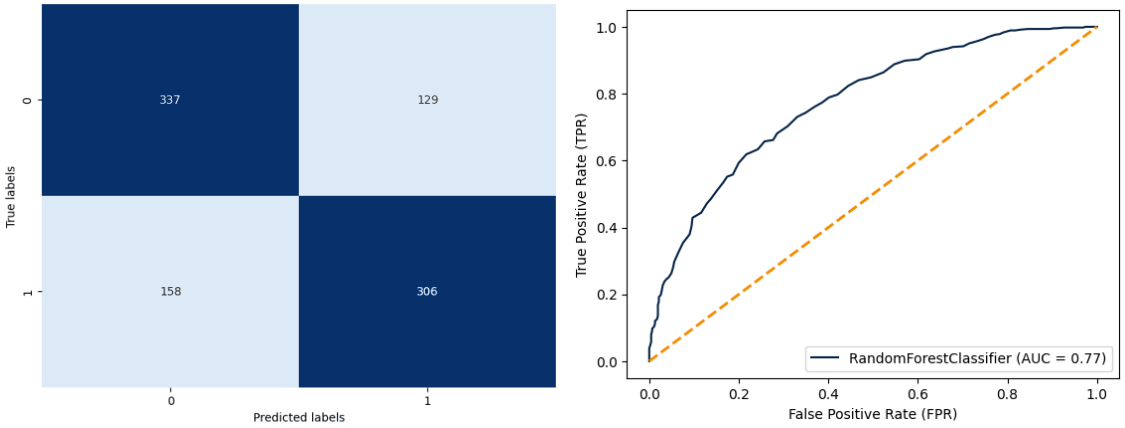
3.6.3 Models definition

After the completion of Feature Importance Analysis, the optimal subset of KPIs is then considered to build the final best models of Random Forest (RF) and XGB classifiers.

It is important to note that a standardized version of *Training* dataset is considered to train the models, as well as conducting Feature Importance Analysis. Specifically, z-score normalization is employed to ensure that the dataset has a mean of 0 and a standard deviation of 1.

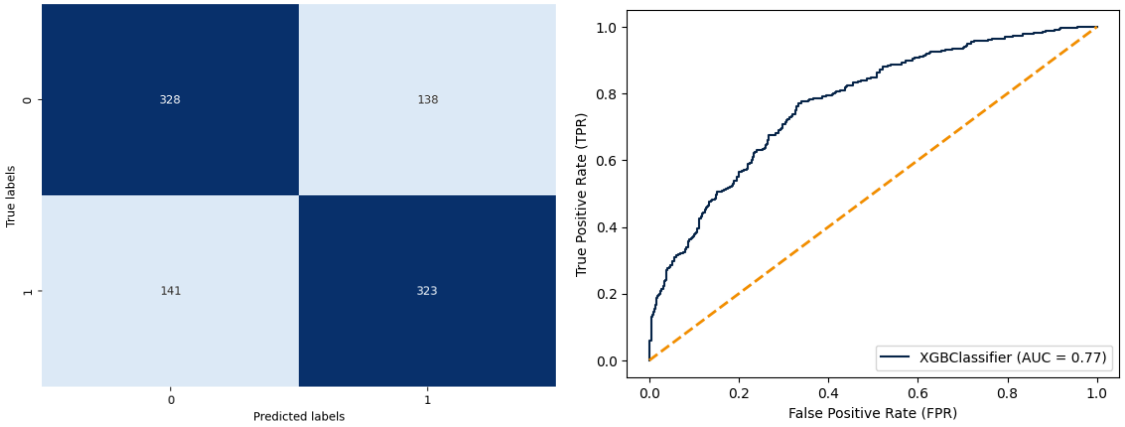
Both the models are trained (performing K-fold cross-validation) again on the dataset considering only the optimal set of most important KPIs.

In Figures 3.14 and 3.13, for each classifier it is provided the relative estimated confusion matrix along with the estimated Receiver operating characteristic (ROC) curve.



(a) RF Estimated Confusion Matrix (b) RF Estimated ROC curve

Figure 3.14: Estimated Confusion Matrices and ROC curves for Random Forest (performing cross-validation).



(a) XGB Estimated Confusion Matrix (b) XGB Estimated ROC curve

Figure 3.13: Estimated Confusion Matrices and ROC curves for XGBoost (performing cross-validation).

Finally, as it results, both algorithms allow to reach acceptable predicting accuracy (around 70%). Considering the average low level of severity of the False Negative prediction, it can be concluded that the models are reliable tools to use for further data analysis.

4. Testing Phase

In this Chapter the previously mentioned tree-based classification models, namely Random Forest and Extreme Gradient Boosting, are further exploited and different experiments are presented.

Mobile network performance is influenced by various factors inherent to the network cells and the environments in which they are situated. Among these, traffic patterns significantly impact network behavior. Similarly, cells operating on different frequency bands exhibit distinct behaviors due to being subject to different stimuli. Additionally, the surrounding environment, which is closely related to traffic and frequency factors, presents diverse challenges to each specific cell.

Building upon the insights obtained from the EDA of Cell Nature conducted in Section 3.5.1, which uncovers the heterogeneous nature of the dataset, attention is now directed towards examining how these differences impact classification predictions, identifying any biases or disparities due to these imbalances.

To accomplish the above, in the following diverse tests are presented to individually examine each of these critical network aspects and assess their impact on classification predictions.

Finally, the last test described in this Chapter aims to provide a deeper understanding of the learnt targets developed by the two classification models, RF and XGB, in their "Traffic-free" versions.

Specifically, the test presented in Section 4.4 challenges the classification models with a particular dataset that includes both *random* instances and occurrences of cells encountering another common network issue: the RSI collisions, as introduced in Section 2.3.

4.1 Traffic impact on KPIs

As discussed in Section 3.2, mobile networks are particularly sensitive to seasonal or temporal variations, and as a result, they experience fluctuations in traffic patterns. Therefore, alterations in traffic dynamics could impact the networks, leading to changes in certain KPIs.

Moreover, considering that KPIs are usually percentage measures, their values obtained in higher-traffic circumstances might be more precise.

Consequently, in pursuit of building generic classification models that do not capture specific traffic patterns, a study is conducted on KPIs believed to be particularly influenced by traffic dynamics.

Thank to the expertise of Telefónica's optimizers and prior knowledge, a subset of traffic-related KPIs is identified among the full set considered in Section 3.3.

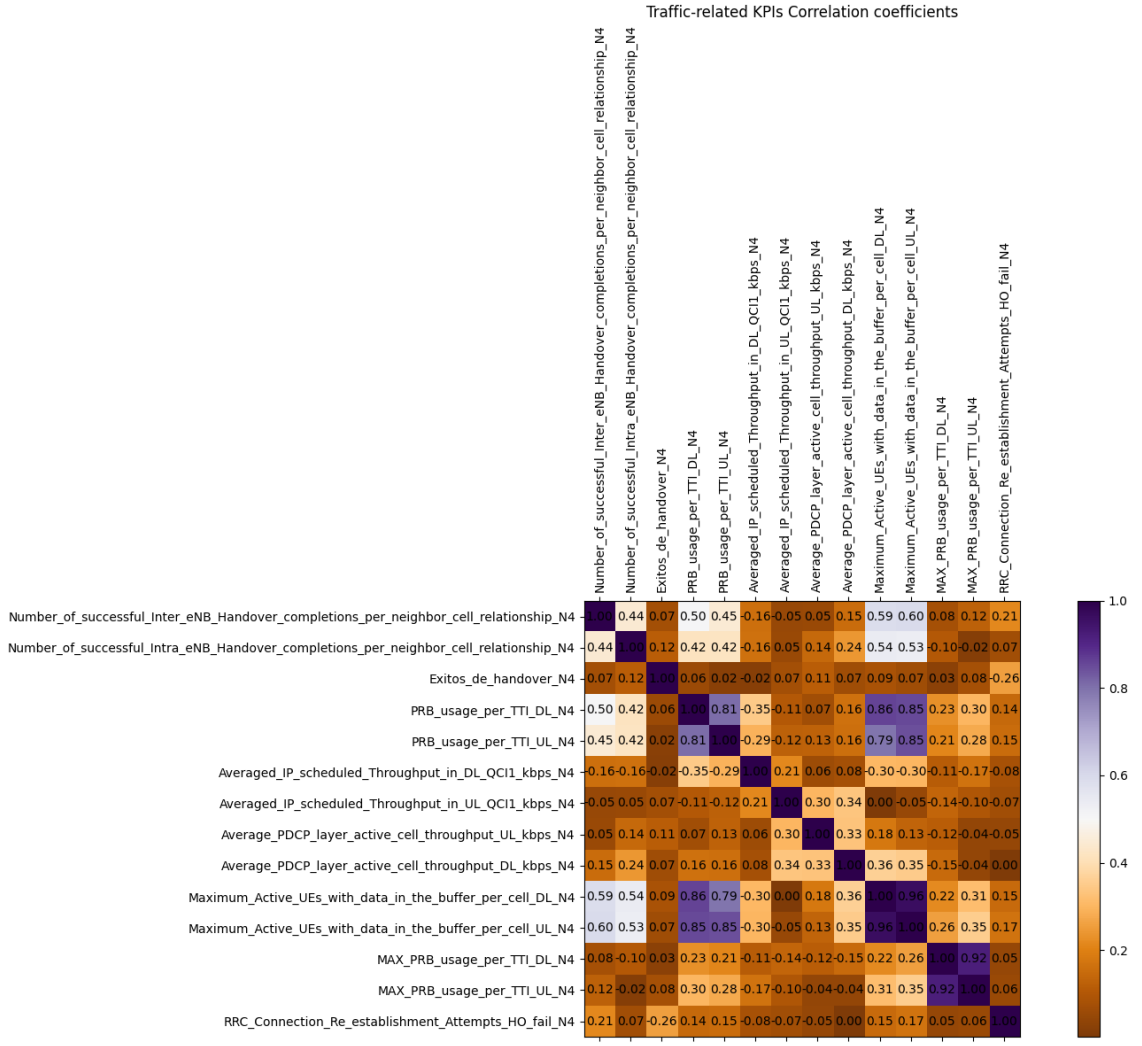


Figure 4.1: Correlation matrix between traffic-related KPIs.

In Figure 4.1, such list of traffic-related features can be appreciated alongside the correlation matrix of these variables. The Pearson correlation coefficients[1] are computed for all selected variables, revealing varying degrees of correlation, including instances of very low correlation. Notably, the strongest correlation observed is among four KPIs: "PRB_usage_per_TTI_UL_N4", "PRB_usage_per_TTI_DL_N4", "Maximum_Active_UEs_with_data_in_the_buffer_per_cell_UL_N4", "Maximum_Active_UEs_-

[1] *Pearson correlation coefficient* is a statistic measure of linear correlation between two variables, with values ranging from −1 to 1. A value of 1 indicates a perfect positive linear relationship, −1 a perfect negative linear relationship, while 0 indicates no linear relationship.

with_data_in_the_buffer_per_cell_DL_N4". Furthermore, these KPIs are highly correlated with "Number_of_successful_Inter_eNB_Handover_completions_per_neighbor_cell_relationship_N4" and "Number_of_successful_Intra_eNB_Handover_completions_per_neighbor_cell_relationship_N4".

Overall, the correlation matrix reveals varying directions of correlations and cases of low correlation, indicating that while the selected KPIs capture traffic-related aspects, they represent diverse information.

4.1.1 Traffic Index and *Traffic-free* Models

Undoubtedly, considering the full set of traffic-related KPIs would be optimal in order to have an holistic view of traffic aspects, which impact on mobile network performance. However, for the sake of comprehensiveness in our project, it could be advantageous to develop a unified "traffic index" to consolidate the major traffic information.

Successively, to achieve the above, Principal Component Analysis (PCA) is conducted on the set of traffic-related KPIs.
PCA [14] is a widely used statistical technique employed for dimensionality reduction and data visualization. By transforming original variables into a new set of uncorrelated variables known as principal components, PCA simplifies complex datasets while preserving as much information as possible. It identifies the directions along which the data varies the most, termed principal components, which are ordered by the amount of variance they explain.

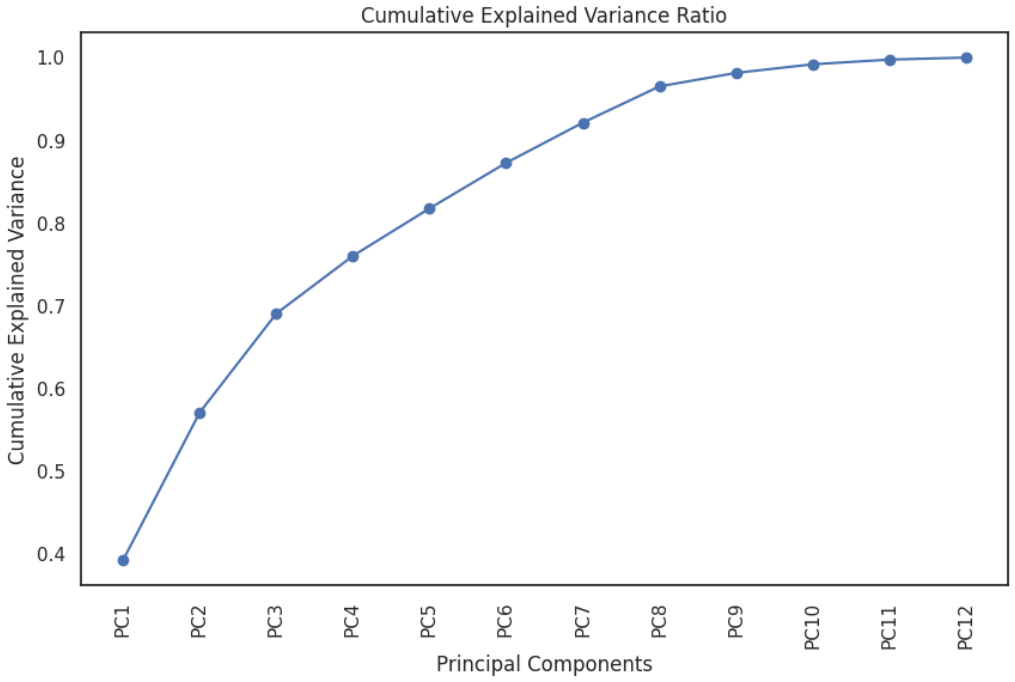


Figure 4.2: PCA Cumulative Explained Variance Ratio.

After having applied z-score normalization to set *Training* dataset to zero-mean and standard deviation of 1, PCA is employed on the set of traffic-related KPIs.
The PCA cumulative explained variance ratio[2], in Figure 4.2, shows that the first principal component (PC1) captures the 40% of total variance, while the second (PC2) adds around the 20%.

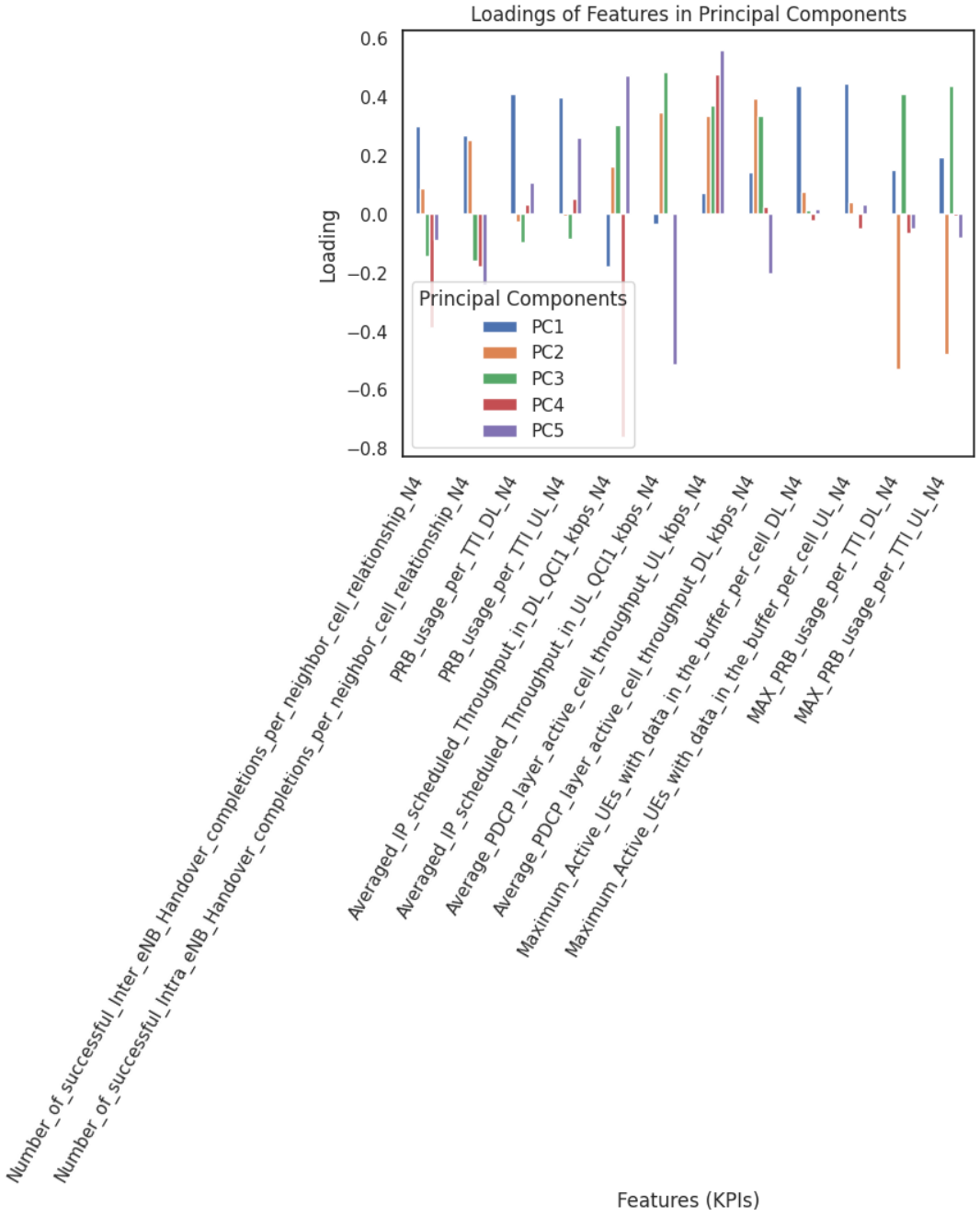


Figure 4.3: PCA Loadings with respect to KPIs.

[2] *Cumulative Explained Variance Ratio* is a metric used to assess how much of the total variance in the dataset is explained by each principal component, as well as by a combination of multiple principal components.

Additionally, in Figure 4.3, the PCA loadings depict the correlation coefficients between the original features (KPIs) and the principal components obtained through PCA. The barplot illustrates how the correlation matrix is mirrored in the PCA loadings. Specifically, the strongest correlation among the primary four KPIs mentioned earlier is evident in PC1, whose prominent loadings originate from these KPIs, followed by the other two KPIs referenced above.

In conclusion, the PCA confirm that PC1 effectively encapsulates the primary traffic information. While further analysis of the other principal components, capturing minor variances of traffic information, could be intriguing, PC1 is currently designated as the "traffic index".

Subsequently, the Random Forest and Extreme Gradient Boosting models are retrained following the entire procedure outlined in Section 3.6.3, with the exception that the traffic-related KPIs are omitted from the feature set. For the sake of clarity in nomenclature, this new version of the models is defined as *Traffic-free* throughout this thesis.

Although the performance of the models do not seem to be significantly affected by the removal of traffic-related KPIs, the final decision is to utilize the *Traffic-free* models in the subsequent tests in order to ensure robustness against traffic dynamics.

Finally, the "traffic index" attribute is leveraged in the following Section to delve deeper into the impact of traffic on the classification models.

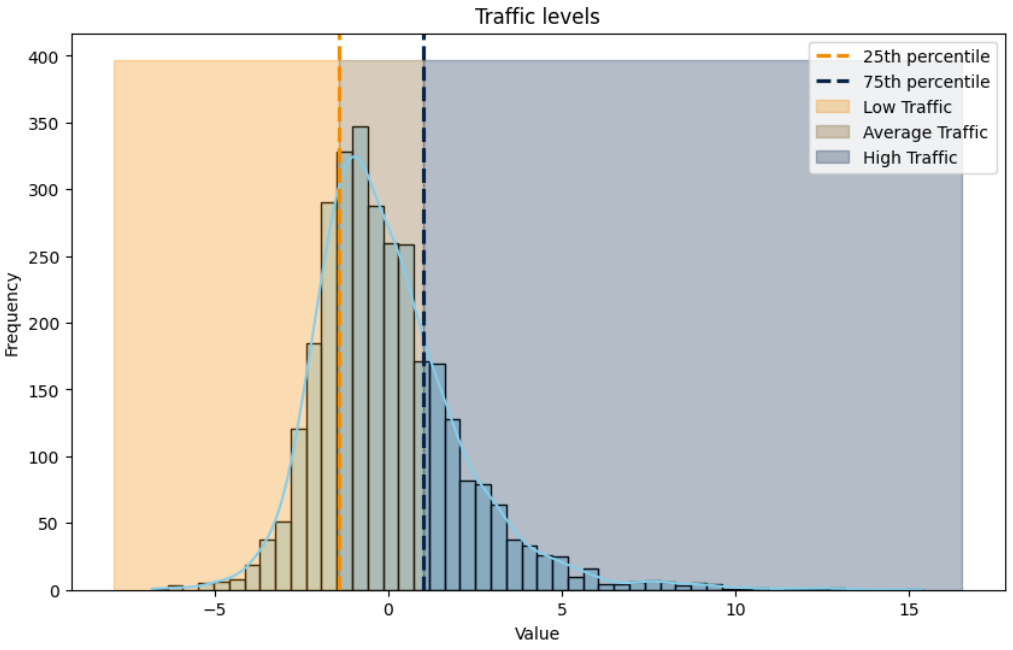


Figure 4.4: Three traffic levels defined by the 25th and 75th percentiles of the *Training* dataset's "traffic index" distribution.

Hence, based on the distribution of the "traffic index" in the *Training* dataset, three different traffic levels are defined, as shown in Figure 4.4.

Therefore, in Figure 4.5 shows the *Training* dataset's instances grouped by traffic level, with differentiation among the classes. Notably, within each traffic level, the class ratio is similar, with the percentage of each class being around 50%. Moreover a higher number of occurrences is observed in the High-Traffic level.

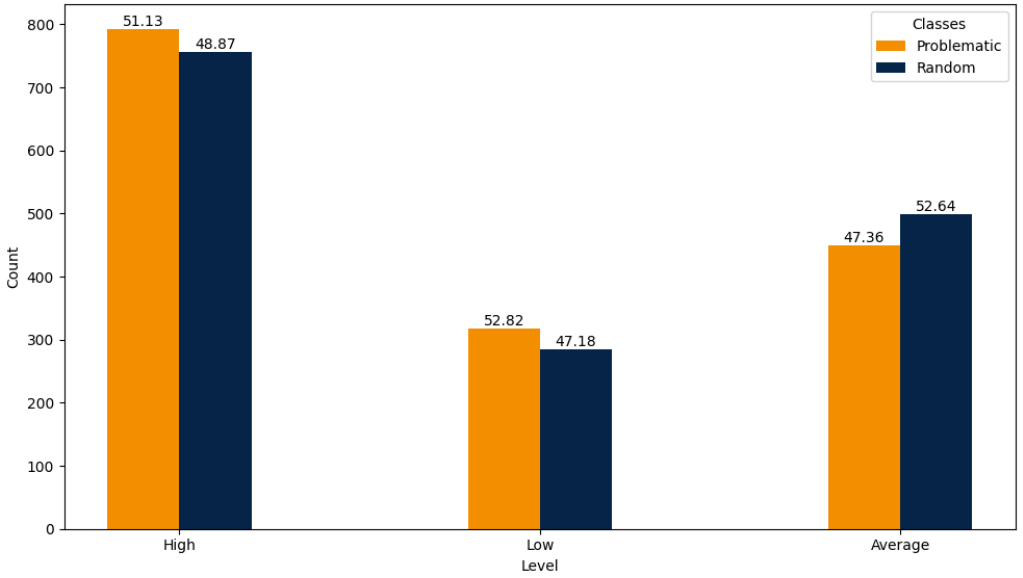


Figure 4.5: Classes Distribution of *Training* dataset by Traffic level.

4.1.2 Testing on *Semana Santa* dataset

The performances of two classification models, namely Random Forest and Extreme Gradient Boosting - *Traffic-free* version (Section 4.1.1) - , are further assessed using a specialized dataset comprising overshooting incidents recorded in date April 2nd.

Semana Santa Traffic patterns

The chosen time window falls in the Easter period of the year: the holiday of *Semana Santa* is widely celebrated in Spain, therefore the dataset is expected to exhibit unique traffic patterns. In fact, the dataset is estimated to be particularly representative of the traffic dynamics and behaviors, since that the temporal variation in traffic patterns during this week poses a considerable challenge to network performance.

Therefore, to validate the hypothesis of traffic changes during Semana Santa period, we analyze the probability density functions associated with both *Training* and *Semana Santa* datasets. As shown in Figure 4.6, comparing the distributions reveals a slight difference between the datasets.

Additionally, the difference is confirmed also by performing the Mann-Whitney U Test with resulting p-value of $1.528e-11$ (more details about the applied test in Appendix B).

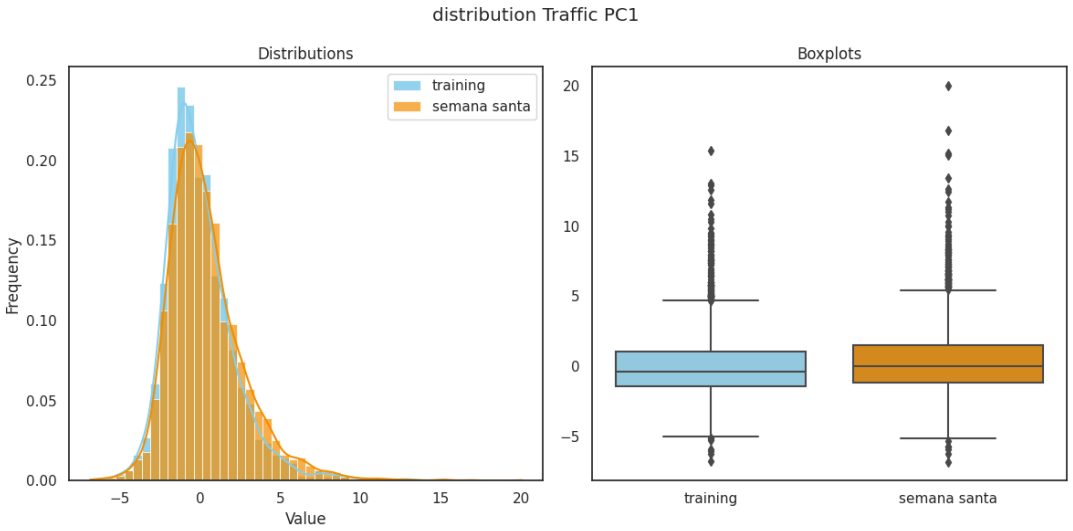


Figure 4.6: PC1 Probability Density Function Comparison *Training* versus *Semana Santa* datasets.

Although the traffic actually changes during this particular period, the anticipated expectations are of more substantial traffic dynamics. Given the highly heterogeneous circumstances in which cells are situated, this modest change may result from the specific selection of cells included in the dataset. Future testing may involve larger datasets to better represent the territory.

However, despite the slight traffic change, it reaffirms the effectiveness of the PC1 feature as the "traffic index" once again.

Semana Santa Test

Hence, in the following, the *Traffic-free* version of the RF and XGB models, trained on the *Training* dataset presented in Section 3.4, are tested on different never-seen samples from *Semana Santa* dataset.

Indeed, testing the models on new dataset, z-score normalization (more details in Section 3.6.3) is consistently applied on *Semana Santa* dataset using the mean and standard deviation derived from the *Training* dataset prior to model testing.

Respecting the format of the *training dataset*, the *Semana Santa* dataset encompasses 3-days aggregation (from March 31st to April 2nd) values of all the used KPIs and comprises 3020 network cells, belonging to *problematic* (1641) and *random* (1379) classes. The problematic cells are identified through the same PBI tool mentioned in Section 2.4.1, following the same procedural steps employed for the construction of the training dataset.

Figure 4.7 presents the visual output generated using `plotly` library in Python, depicting the geographical distribution of *Semana Santa* dataset.

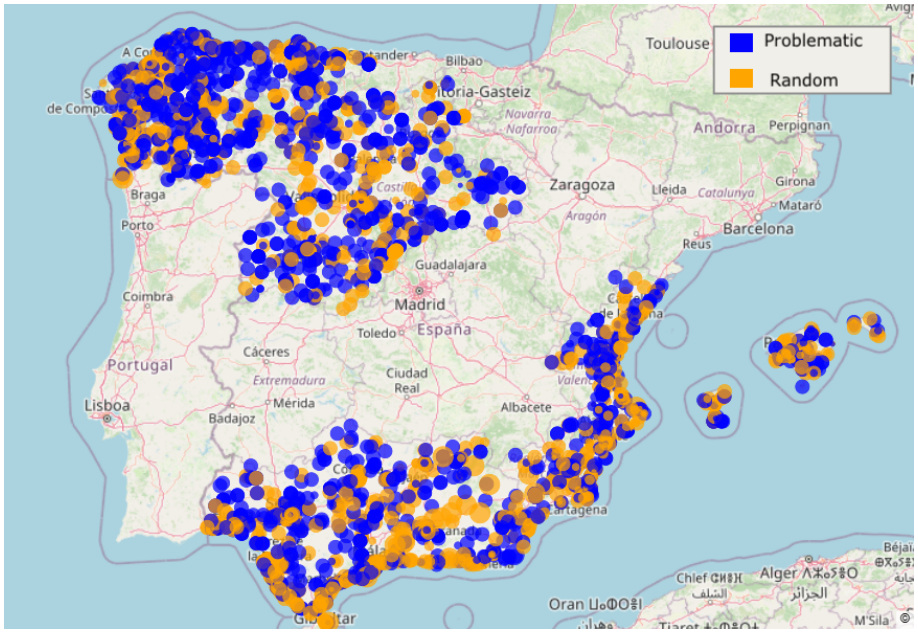
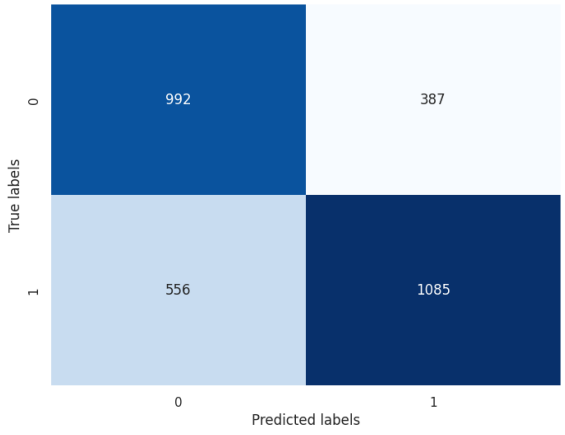


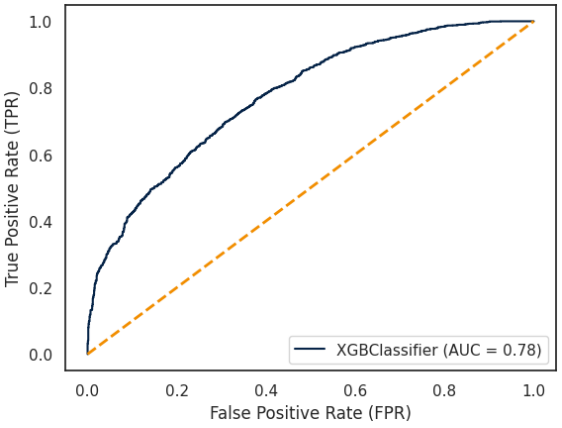
Figure 4.7: Geographical distribution of *Semana Santa* dataset.

The Random Forest classifier achieved an accuracy score of 69%, while the XGBoost classifier yielded 67%. These scores are slightly lower but consistent with their accuracy relative to the original dataset.

In Figures 4.9 and 4.8, the confusion matrices and the estimated ROC curves are presented, showing the predictions of RF and XGB classifiers on the *Semana Santa* dataset. By observing both confusion matrices, in Figures 4.9a and 4.8a, it is evident that there is a higher number of False-Negative (i.e. *problematic* samples predicted as *random*) compared to the False-Positive (i.e. *random* cells predicted as *problematic*).

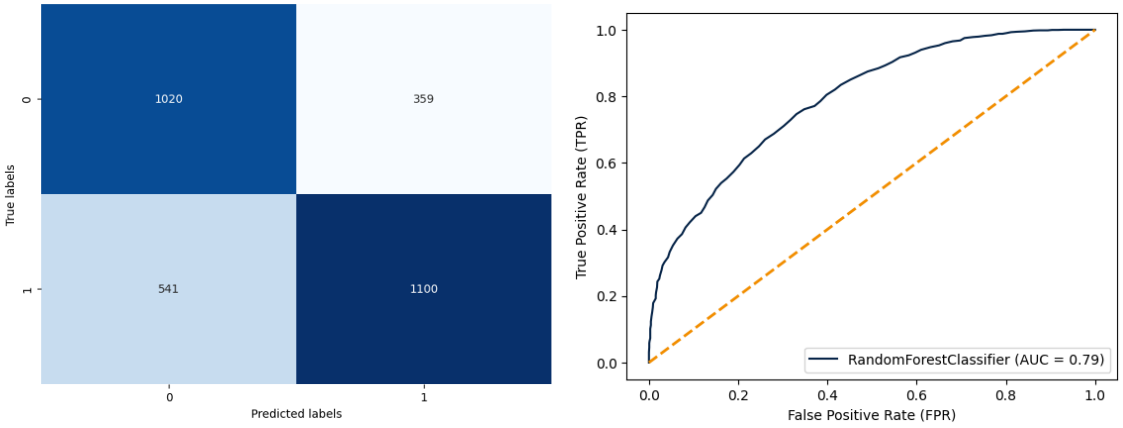


(a) XGB confusion matrix



(b) XGB Estimated ROC curve

Figure 4.8: Confusion Matrices and ROC curves of XGB predictions on *Semana Santa* dataset.

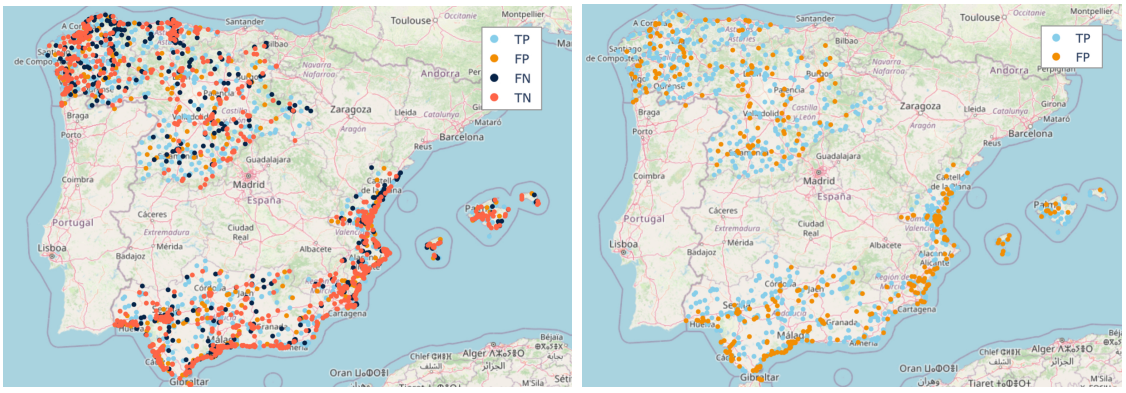


(a) RF confusion matrix (b) RF Estimated ROC curve

Figure 4.9: Confusion Matrices and ROC curves of RF predictions on *Semana Santa* dataset.

One of the requirements set by Telefónica is to maintain a low number of False-Positives, as these are considered significantly more problematic than False-Negatives. False-Positives imply incorrectly identifying cells without actual failures as *problematic*, leading to unnecessary corrections and to a consequent waste of resources.
On the other hand, False-Negatives, while still undesirable, typically correspond to low-severity issues that might not immediately impact users and can be addressed later if they escalate. Therefore, having a low number of False-Positives is advantageous as numerous false alarms can have detrimental consequences on network management.

In Figures 4.10 a geographical distribution of RF predictions is provided. In Figure 4.10b, it can be noticed that False-Positives appear closer to urban areas, while True-Positives are well spread, even in rural zones.

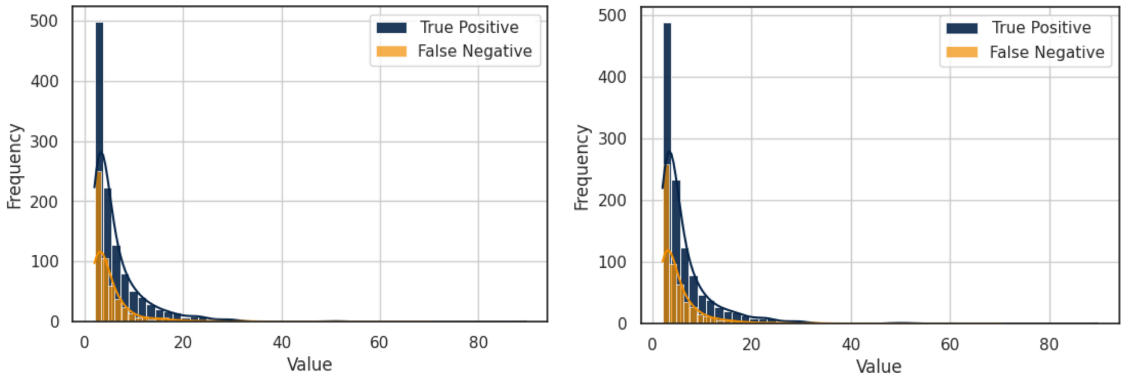


(a) (b)

Figure 4.10: Geographical distribution of RF predictions.

In order to address the high number of False-Negatives, the distribution of the percentage of out-of-range access attempts, as retrieved from the initial PBI tool,

are further examined. Figure 4.11 presents a comparison between the distributions relative to False-Negatives and True-Positives. Although the distributions appear very similar, False-Negatives exhibit a consistently lower average value of out-of-range access attempts rate, indicative of a minimal severity of overshooting instances.



(a) Random Forest (b) XGBoost

Figure 4.11: Distribution of the percentage of out-of-range access attempts of False-Negatives and True-Positives.

Moreover, the Mann-Whitney U Test (more details in Appendix B) is performed on both pairs of distributions and returns very low p-values (Table 4.1) as confirmation of the significant statistical difference between False-Negative and False-Positive populations.
These results are favorable since very low severity levels of overshooting issues likely have no significant impact on user experience and can therefore be disregarded.

	p	**statistic**
RF	2.699e-06	366448
XGB	2.046e-4	342788

Table 4.1: Mann-Whitney U test results comparing distribution of out-of-range access attempts of False-Negatives and False-Positives.

Additionally, considering the results from RF classification model (however similar results can be obtained from XGB model), the predictions on *Semana Santa* dataset are examined in terms of traffic levels in Figure 4.12.
At first sight, it can be noticed that the Average-traffic level is the most represented in the case of *Semana Santa*.
Moreover, it is visible that similar proportions of confusion matrix elements are repeated among the traffic levels. Likewise, it can be observed that the within-level class ratio is similar to that of *Training* dataset (Figure 4.5), with each class representing approximately 50%.

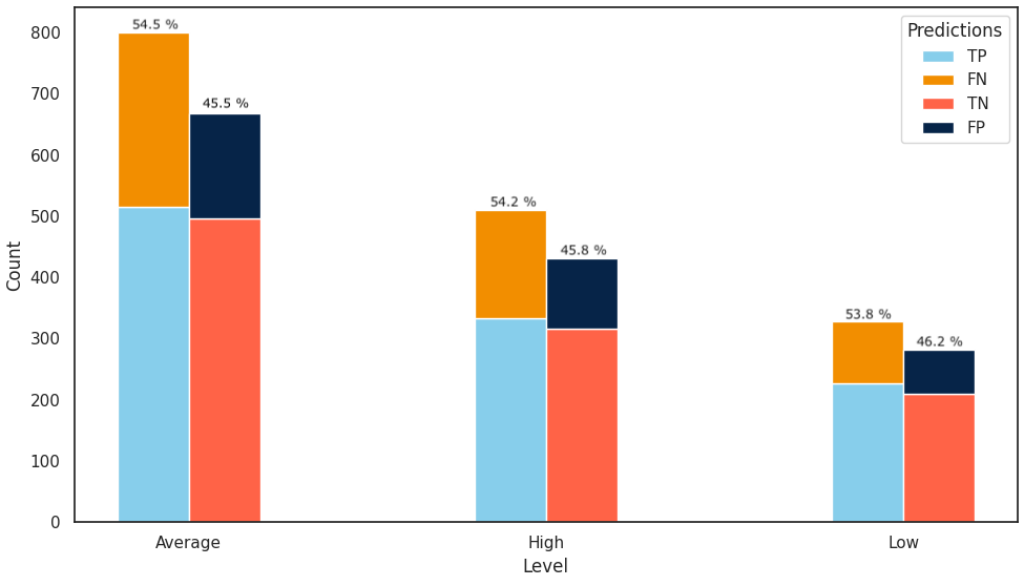


Figure 4.12: Traffic levels distribution of *Semana Santa* dataset (grouped by False-Negatives, False-Positives, True-Negatives, True-Positives from RF).

Furthermore, a closer examination is proposed in Table 4.2, which reports the values of precision, recall and specificity per traffic level, alongside the classes' percentages and total counts. Detailed explanations of these performance metrics can be found in Appendix C.

Traffic Level	**Precision**	**Recall**	**Specificity**	**Problematic (#)**	**Random (#)**
Average	0.64	0.75	0.74	801	668
High	0.65	0.74	0.73	511	432
Low	0.69	0.76	0.74	329	282

Table 4.2: Precision, Recall and Specificity per traffic level (RF model)

Considering the results reported in Table 4.2, the values of precision, recall and specificity are very similar across each traffic level, reflecting the results visualized in Figure 4.12. This indicates that, traffic dynamics do not significantly affect classification performance.

Finally, although traffic patterns slightly influencing the traffic-related KPIs, they do not appear to significantly affect the overall set of KPIs. Hence, the *Traffic-free* version demonstrates greater robustness against traffic dynamics.

4.2 Within-band Classification

To gain a deeper understanding of the effects of inherent network aspects on classification performance, the following Section presents a study on frequency bands within which the cells operate.

As anticipated in the introduction of this Chapter, network cells are exposed to a wide range of stimuli, which vary depending on the specific frequency band in which they operate.

Specifically, higher frequency bands are selected when connection quality is high, while the primary mechanism employed by networks in response to a decrease in quality is switching to the lower band (if feasible at that site).

Indeed, network cells performing in the lowest frequency bands, i.e. the lower end of the frequency spectrum offered in that site, also referred to as "border" frequency, usually operate in highly critical situations, which implies registering numerous and diverse issues. This is often the case of band 800MHz, which is the most common border band.

Semana Santa Test

To begin with, the results from classification test carried on *Semana Santa* dataset (Section 4.1.2) are further examined in terms of frequency bands.

Thus, Figure 4.13 examines the distribution of *Semana Santa* dataset in terms of frequency bands, with a focus on the predictive results of the RF classifier. However, similar results can be obtained from XGBoost. At first sight, it is noteworthy that the majority of samples belong to 1800MHz and 800MHz bands. This aligns with the frequency band distribution of the *Training* dataset, as illustrated in Figure 3.5, where these bands encompass a larger number of instances.

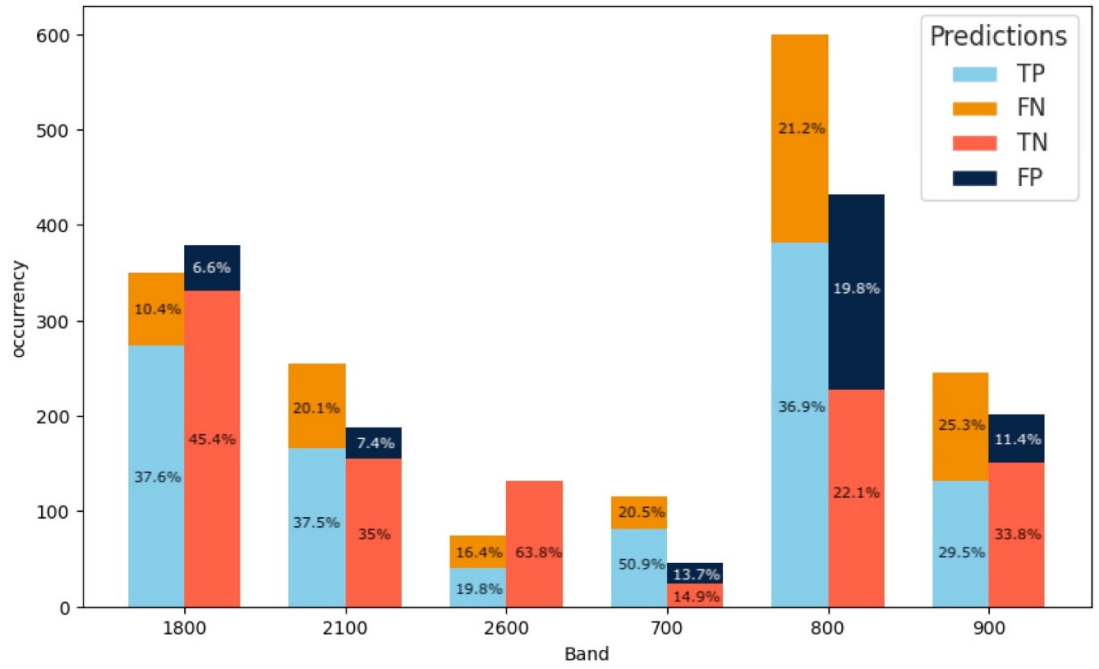


Figure 4.13: Frequency Bands distribution of *Semana Santa* dataset (grouped by False-Negatives, False-Positives, True-Negatives, True-Positives from RF).

Furthermore, a closer examination is proposed in Table 4.3, which reports the values of precision, recall and specificity per band.
Notably, in band 2600MHz, highlighted in blue, anomalous values of precision and specificity (both equal to 1) are observed, due to the absence of True-Positives. This scarcity stems from the limited representation of the 2600MHz band in the dataset, reflecting the sparse network coverage in that frequency range.

Band (MHz)	**Recall**	**Precision**	**Specificity**	**Problematic (#)**	**Random (#)**
800	0.63	0.60	0.53	599	432
700	0.71	0.79	0.52	117	46
2100	0.65	0.83	0.82	255	188
2600	0.55	1	1	75	132
900	0.54	0.72	0.75	245	202
1800	0.78	0.85	0.87	350	379

Table 4.3: Precision, Recall and Specificity per band (RF classifier)

Additionally, attention is drawn to the notably low values observed for the 800MHz band, highlighted in orange, contrasting with the high performance of the 1800MHz band.
Specifically, within the 800MHz band, False-Positives account for approximately 20%, while True-Negatives comprise 22%, resulting in a specificity of 0.53 and True-Positives constitute the 37%, resulting in a precision of of 0.65. Similarly, the low recall value (0.63) is attributed to False-Negatives, which constitute around 21% of the observations.

Thereby, although the band 800MHz is well represented in the *Training* dataset, it exhibits the poorest predictive performance in terms of precision, recall and specificity.
Anyway, this poor performance stems from inherent characteristics of the mobile network itself. The band 800MHz typically operates as "border" frequency, i.e. the lower end of the frequency spectrum offered in that site, making cells operating in this band particularly critical and prone to various issues.

Consequently, the classifiers detect anomalous KPI behaviours, which, however, do not necessarily stem from cell range problems.

Finally, in Figure 4.14 the distribution of out-of-range access attempts rates for False-Negatives and True-Positives within the band 800MHz is presented.
Unsurprisingly, these distributions closely resemble those shown in Figure 4.11, which depicts overall distributions, as band 800MHz is the major contributor to both False-Negatives and True-Positives.

Further details regarding this aspect are provided in the subsequent Section, where it is extensively evaluated within the context of the within-band classification performance study.

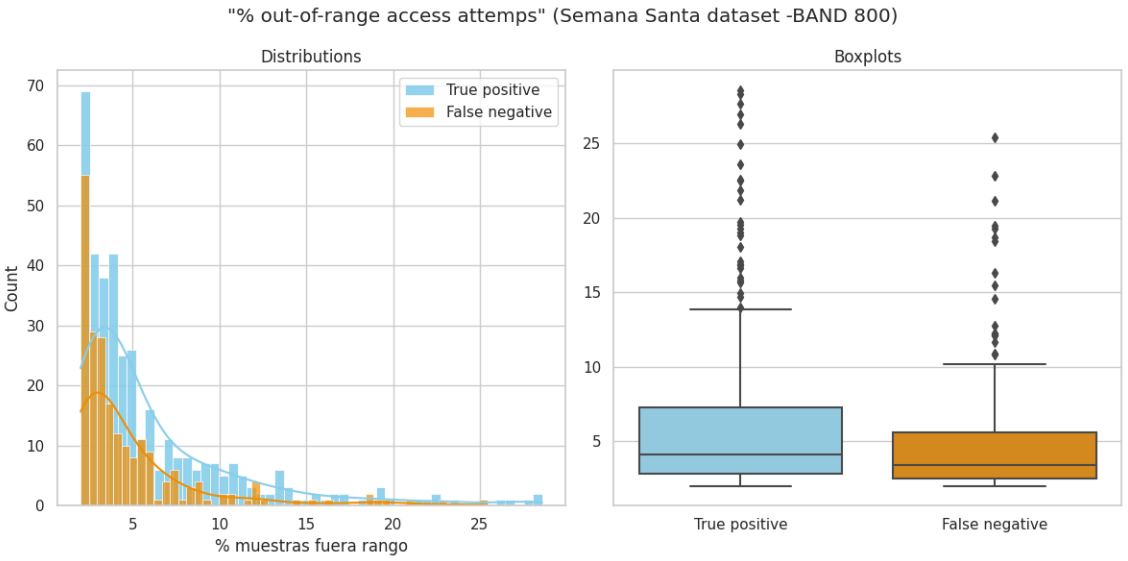


Figure 4.14: Distribution of out-of-range access attempts rate of False-Negatives vs. True-Positives of *Semana Santa* dataset (only in band $800Mz$).

In conclusion, the prediction performances of both models on *Semana Santa* dataset confirm our expectations. While the models are influenced by the characteristics of the mobile network they are still capable of accurately identifying instances of medium and high severity overshooting occurrences.

4.2.1 Low-Band versus High-Band Models

In pursuit of constructing robust and unbiased classification models, the following experiment involves dividing the data into Low and High Bands to build band-specialized classification models.

Consequently, the original "Traffic-free" models, encompassing all bands, are compared to these band-specialized classification models. The objective of this study is to determine whether the original models exhibit bias towards the most represented frequency bands.

Therefore, the Low-Band case focuses on band 800MHz, which is the best represented band. The choice of removing bands 700MHz and 900MHz is motivated by the relative scarcity of data.
Conversely, the High-Band case comprehends bands 1800MHz and 2100MHz. Similarly to low bands, the band 2600MHz is not considered for lack of enough data.

Hence, employing the same training procedure utilized in constructing the "Traffic-free" RF and XGB models (more details in Sections 4.1.1 and 3.6.3), the Low-Band RF and High-Band RF models are built. The decision to exclusively work with RF stems from the notable similarity in results between RF and XGB. Thus, for this phase of observation, it is sufficient to conduct experiments purely with RF.

Initially, the Feature Importance Analysis is obtained using the same process outlined in Section 3.6.2. In Figure 4.15, the normalized results from Low-Band and

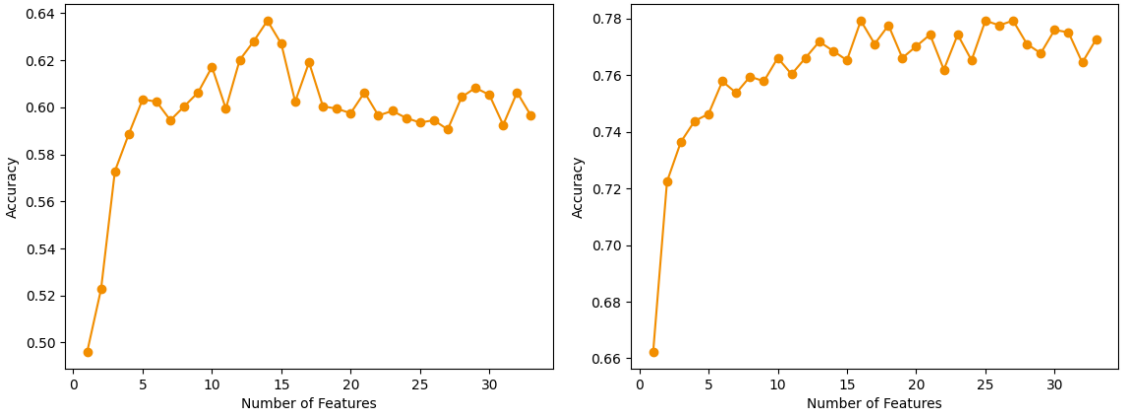
High-Band models are compared. The KPIs are arranged according to the order of importance in the Low-Band model.

Notably, the two models only concur on the first and second most important KPIs, which are message-2/message-1 success ratio and message-3/message-1, relatively. However, the Feature Importance Analysis results reveal remarkably different patterns.

On one hand, the Low-Band model presents an importance order that closely mirrors that of the original RF model (Figure 3.10). Moreover, it assigns similar importance values to the majority of KPIs. On the other hand, the High-Band models assign notably high importance scores to the first two most significant features, creating a substantial gap with the remaining KPIs.

Subsequently, always accordingly to the model-construction procedure of Section 3.6.2, the cross-validation methodology is employed to identify the highest achievable accuracy and the relative optimal subset of most important KPIs.

Thus, the evolution of the accuracy of the two band-specialized classifiers as function of the number of most important KPIs is provided in Figure 4.16.



(a) Low-Band RF model (b) High-Band RF model

Figure 4.16: Accuracy vs. number of most important KPIs used.

In Figure 4.16, it is important to highlight the significant differences in performance between the Low-Band and High-Band models. Firstly, the Low-Band model achieves a lower accuracy (64%) compared to the High-Band model (78%). Moreover, the Low-Band model requires the first 14 most important KPIs to achieve its best score, while the High-Band model needs to select the first 25 features.

Furthermore, considering both the Feature Importance (Figure 4.15) and the accuracy evolution (Figure 4.16), it becomes evident that the Low-Band model faces significant challenges posed by the instances. Consequently, despite assigning comparable importance scores to the majority of KPIs, it fails to enhance its predictive performance. In contrast, the High-Band model succeeds in attaining a notable accuracy score, even with only the initial few most important KPIs, and continues to marginally improve its performance by incorporating the less important features.

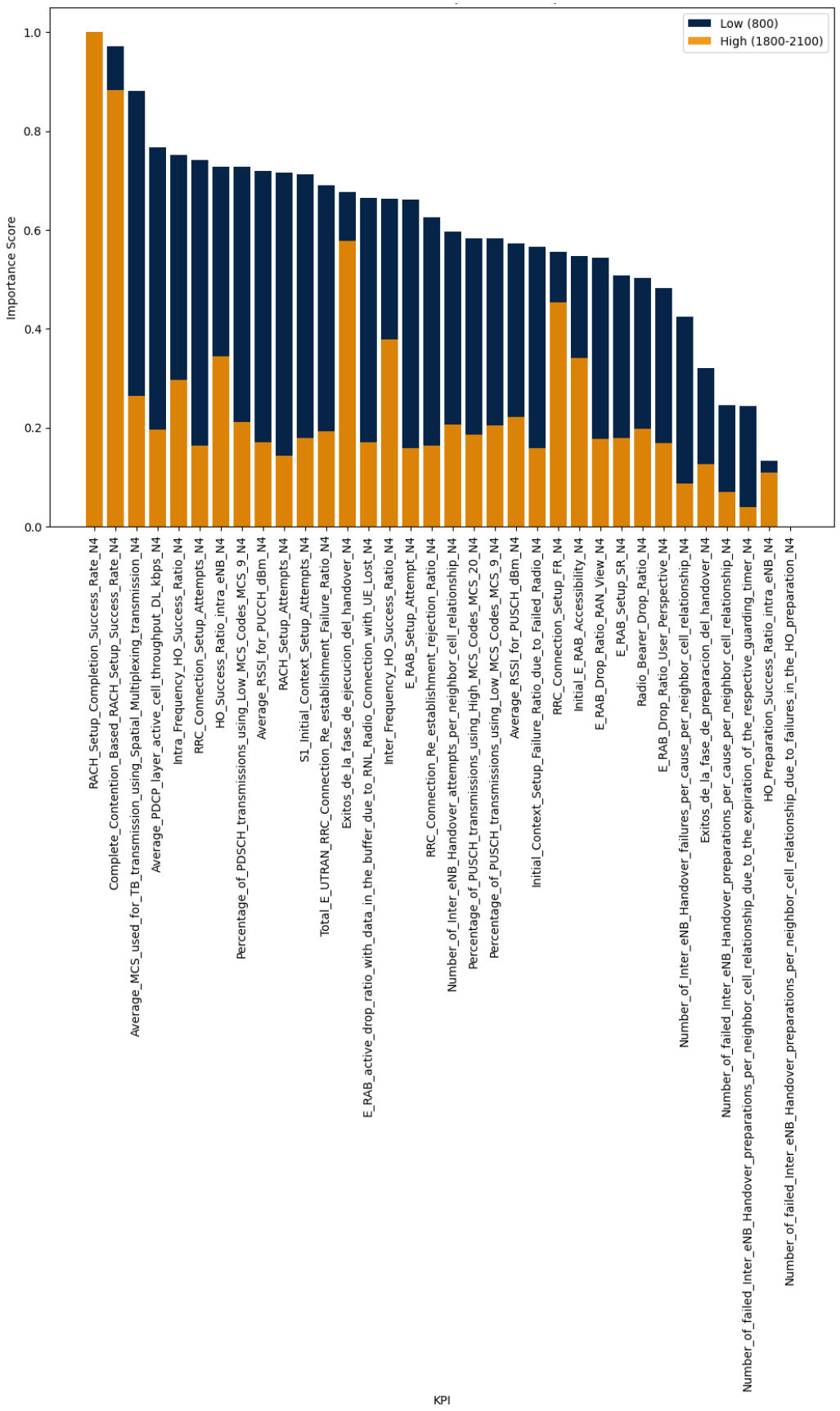


Figure 4.15: Comparison Low-Band versus High-Band RF Feature Importance.

Finally, the two band-specialized models are tested on *Semana Santa* dataset and the performances are compared, including also the original "Traffic-free" RF model.

In Figure 4.17, the Low-Band, High-Band and original RF models are compared in terms of precision versus recall.

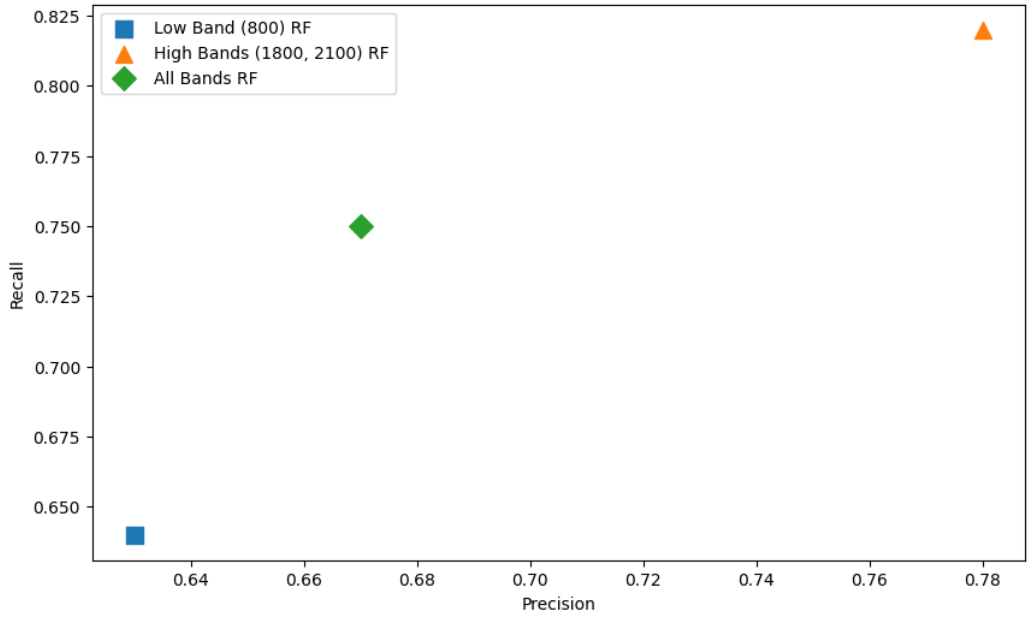


Figure 4.17: Model Comparison in terms of Precision and Recall (Low-Band versus High-Band versus Original RF).

Based on these results, the High-Band specialized model appears to be an improvement over the original RF model, which encompasses all frequency bands.

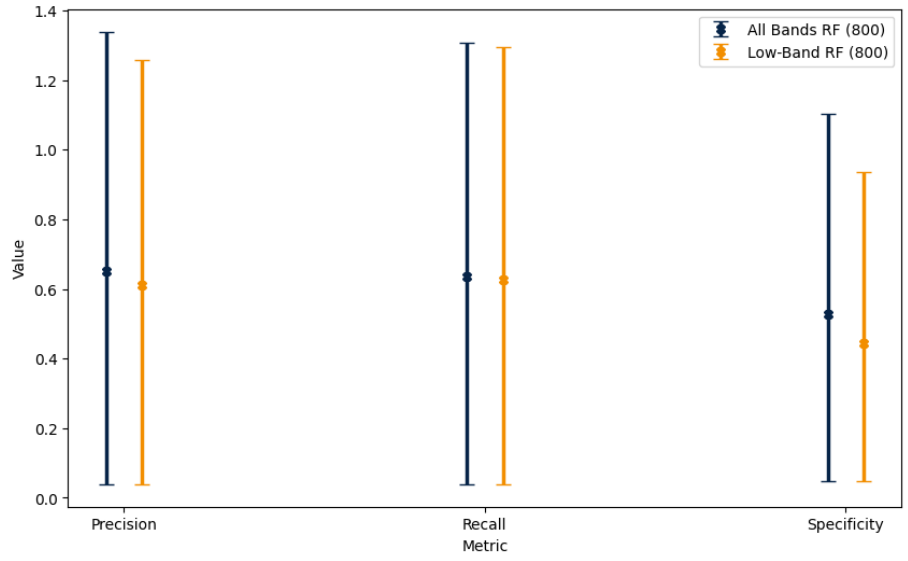
However, a more detailed examination of the band-specialized models is conducted by comparing the performance of each model to that of the original model within the corresponding frequency band.

Therefore, in Table 4.4, the within-band performances of the different models are reported in terms of precision, recall and specificity.

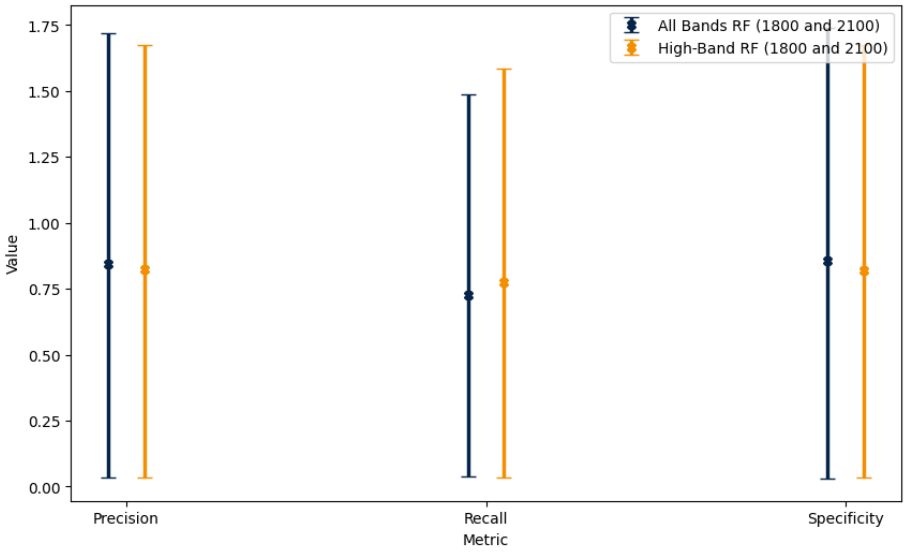
Model	Band (MHz)	Precision	Recall	Specificity
Low Band	800	0.63	0.61	0.44
High Band	1800	0.8	0.86	0.88
	2100	0.69	0.78	0.73
All	800	0.63	0.60	0.53
	1800	0.8	0.86	0.88
	2100	0.62	0.82	0.82

Table 4.4: Precision, Recall, Specificity of Low-Band RF, High-Band RF and original RF models

Additionally, Figures 4.18 show the metrics and the relative confidence intervals of the band-specialized models, in comparison to the original "Traffic-free" RF model.



(a) Low-Band RF vs. original RF



(b) High-Band RF vs. original RF

Figure 4.18: Confidence intervals of Precision, Recall, Specificity of Low-Band RF, High-Band RF versus original RF.

In conclusion, in both cases, the band-specialized models perform very similarly to the original RF model. These results indicate that the original RF model is not biased towards any specific band, and its performance accurately reflects the true predictive potential within each band.

Thus, the original "Traffic-free" RF model is demonstrated to be functioning properly and at its highest potential. Consequently, the instances from the 800 MHz band are shown to be inherently challenging.

Continuing in this line of research, an interesting study could involve including band information as an additional feature in the dataset used by the model. However, this approach is not included in this thesis due to the current unavailability of balanced datasets across bands, as previously shown in Figures 3.5 and 4.13).

4.3 Rural versus Urban Classification

To further delve into the behavior of mobile networks and understand the effects of inherent network aspects on classification performance, this section presents a third study examining the surrounding environment in which the cells operate.

Mobile network performance changes significantly based on the characteristics of the territory in which the network is placed. Thereby, another factor that particularly affects cells is the surrounding environment. This aspect is closely related to traffic dynamics and frequency bands, as, for example, an open area is expected to have lower levels of traffic and few, sparsely placed cells operating at low frequency bands (large cell range).
However, in this study, this factor is treated in isolation, and its impact on network cells is measured in terms of the predictive performance of ground-specialized classification models, as in the previous experiment.

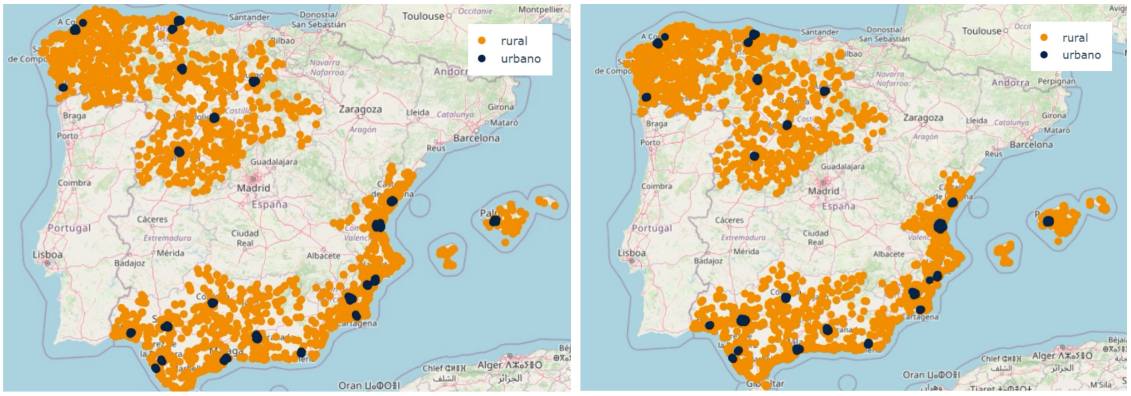
In this study, the data is split into "rural" and "urban" categories, based on the characteristics of the area.

Specifically, to define the terms "rural" and "urban", the Spanish electoral census partitioning is considered.
In Spain, each electoral census section contains at most 2000 residents, representing small towns or parts of larger cities. To determine whether a cell falls within a "rural" area or an "urban" one, the number of overlapping sections within a geographical cell-range of 5km is taken into account. While the exact population within these sections is not known, a higher number of overlapping sections suggests a more urban environment.
Hence, by establishing an arbitrary threshold of 10 overlaps, the term "rural" defines the areas below this threshold while "urban" areas contains at least 11 overlapping electoral census sections.

For the sake of clarity, these definitions will be consistently applied throughout this Section, evaluating the influence of the surrounding environment on network performance.

Therefore, the *Training* dataset as well as the *Semana Santa* dataset are split accordingly into "rural" and "urban" categories.

Figures 4.19 provides a capture of an interactive visualization (HTML file) generated using the Python `plotly` open-source library.



(a) *Training* dataset (b) *Semana Santa* dataset dataset

Figure 4.19: Geographical distribution of "rural" and "urban" cells.

As previously mentioned in the earlier Chapters, for sake of privacy considerations for sensible information involved, only the captures of such visual outputs are presented in this thesis.

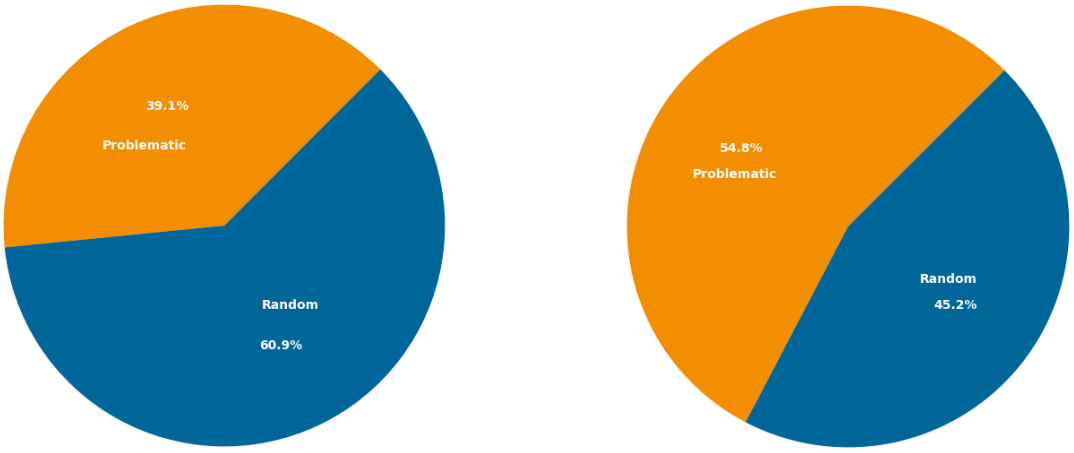
Additionally, considering both *Training* and *Semana Santa* datasets, Table 4.5 reports the number of instances in "rural" and "urban" categories per class. The smaller number of instances in the "urban" category is noticeable and should be considered when discussing the final results of the Urban-specialized model.

Area type	Dataset	Problematic	Random
Rural	*Training*	1384	1258
	Semana Santa	1431	1061
Urban	*Training*	172	279
	Semana Santa	209	315

Table 4.5: Total counts of instances in *Training* and *Semana Santa* datasets for "rural" and "urban" categories per class.

Accordingly, Figure 4.20 provides class distribution within "rural" and "urban" categories, respectively. It shows an almost balanced class distribution in "rural" category, while "urban" category comprises of an higher number of *random* instances than *problematic* ones.

Moreover, Figure 4.21 shows the percentage of instances within each frequency band for both rural and urban categories. It is visible that, as expected, the lowest band 700 MHz is almost absent in "urban" areas while the higher bands are more present in these areas. Moreover, bands 800 MHz and 1800 MHz, being the most used bands, are significantly present in both zones.



(a) "Urban" category (b) "Rural" category

Figure 4.20: Class division of "rural" and "urban" cells.

It is noteworthy that while the concept of rural-urban categorization is closely associated with frequency bands, the distribution of "rural" and "urban" categories within each band, as depicted in Figure 4.21, illustrates that "rural" areas are not exclusively served by low frequency bands, nor do "urban" areas solely consist of high frequency bands. This is because an antenna may be positioned facing either a tall building or an open field, regardless of falling in "urban" or "rural" areas. The distinction lies only in the likelihood of encountering low or high bands within the specific rural or urban category.

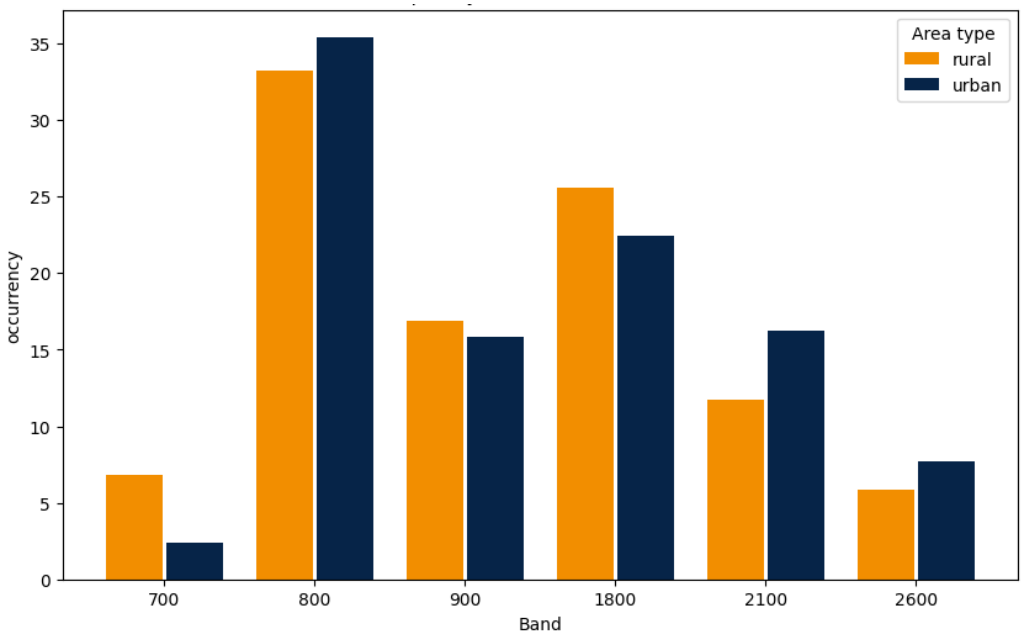


Figure 4.21: Within-band distribution of "rural" and "urban" categories.

4.3.1 Rural versus Urban Models

As in Section 4.2.1, in pursuit of constructing robust and unbiased classification models, the following experiment involves building specialized classification models, focusing on "rural"-"urban" categorization.

Consequently, the original "Traffic-free" RF model, encompassing all bands, is compared to these band-specialized classification models. The objective of this study is to determine whether this categorization leads or not to higher classification performances.

Hence, as in Section 4.2.1, employing the same training procedure utilized in constructing the "Traffic-free" RF model (in Sections 4.1.1 and 3.6.3), the Rural RF and Urban RF specialized models are built.

As first step, Feature Importance Analysis is obtained (same procedure as in Section 3.6.2). In Figure 4.22 the normalized results from Rural RF and Urban RF are compared. The KPIs are arranged according to the order of importance in the Rural RF model.

Notably, the two models differ significantly in assigning importance values to KPIs. However, as highlighted by the semi-transparent orange color in Figure 4.22, it is important to keep in mind that the Urban-specialized model uses a much smaller dataset.

Predictably, the Rural-specialized model presents an importance order that closely mirrors that of the original RF model (Figure 3.10). This outcome aligns with expectations, as the Rural category comprises more than 80% of the total data.

Consistently, the KPIs related to message-2/message-1 and message-3/message-1 success ratios are identified as the most important features, creating a significant gap with the rest of the KPIs.

In contrast, the Urban-specialized model assigns similar importance values to the majority of KPIs, while still placing the two previously mentioned KPIs among the top features.

Subsequently, always accordingly to the model-construction procedure of Section 3.6.2, the cross-validation methodology is employed to identify the highest achievable accuracy and the relative optimal subset of most important KPIs.

Finally, the Urban and Rural specialized models are tested on *Semana Santa* dataset and the performances are compared, including also the original "Traffic-free" RF model.

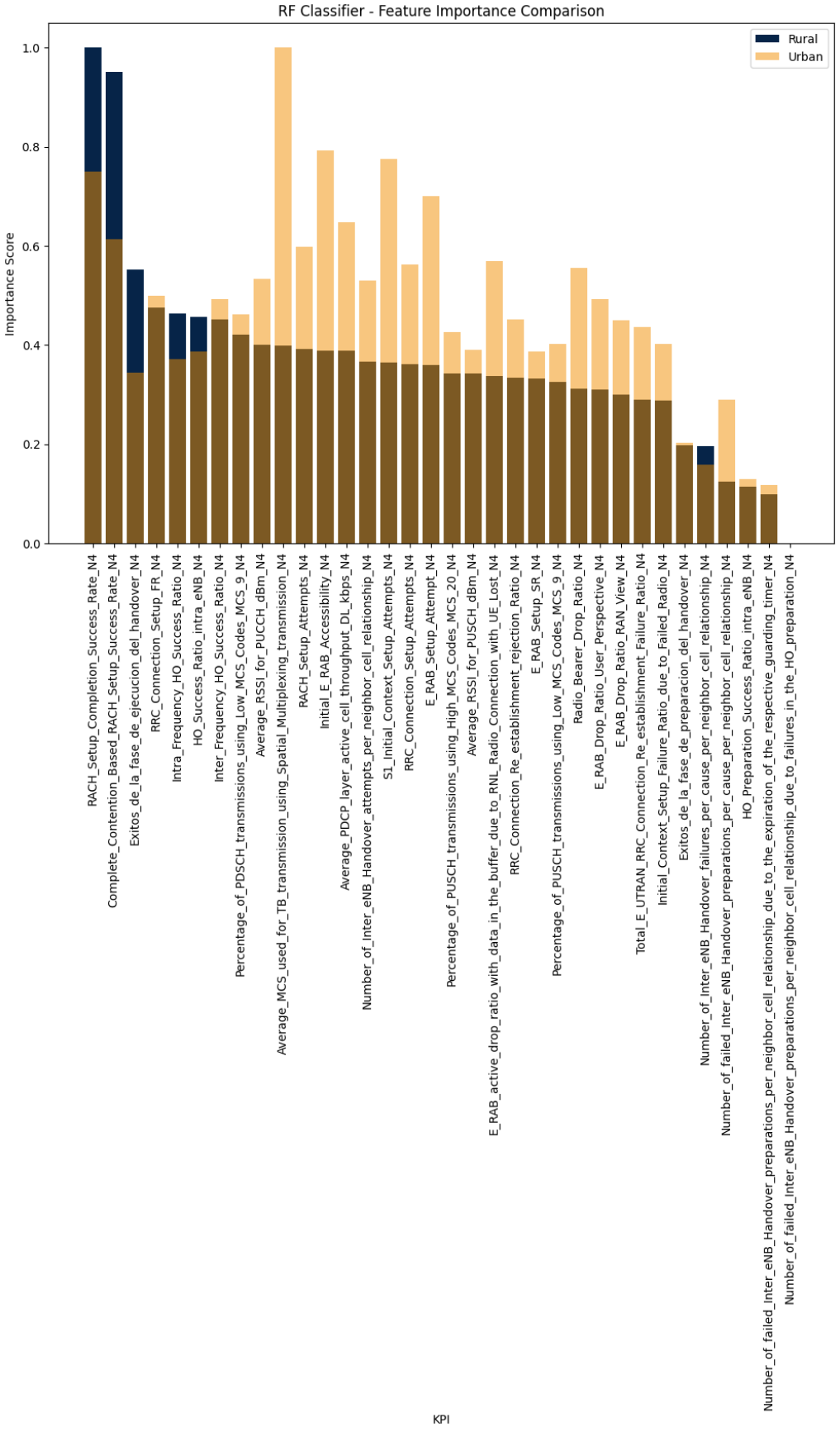


Figure 4.22: Comparison Rural versus Urban RF Feature Importance.

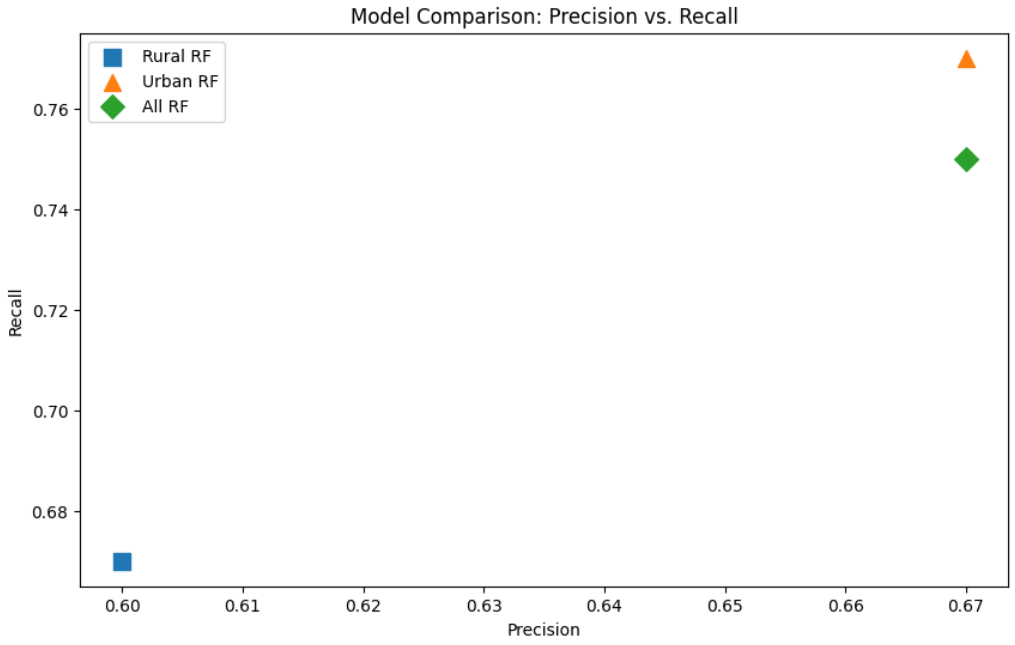


Figure 4.23: Model Comparison in terms of Precision and Recall (Rural versus Urban versus Original RF).

In Figure 4.23, the Rural, Urban and original RF models are compared in terms of precision versus recall.

Based on these results, the Urban-specialized model appears to be improve in terms of recall over the original RF model, which encompasses both categories. Conversely, the Rural-specialized model appears to perform significantly worse compared to the others.

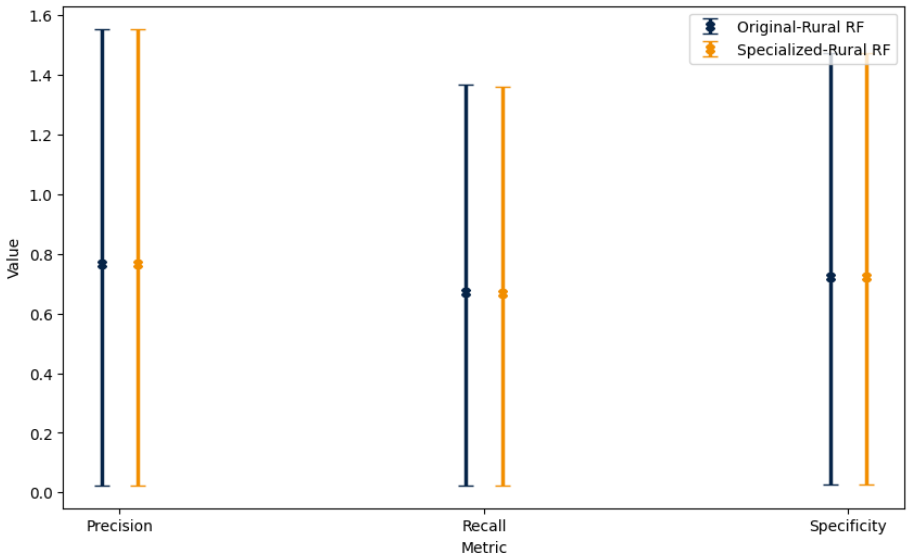
However, a more detailed examination of the Urban and Rural specialized models is conducted by comparing the performance of each model to that of the original model within the corresponding category.

Therefore, in Table 4.6, the within-band performances of the different models are reported in terms of precision, recall and specificity.

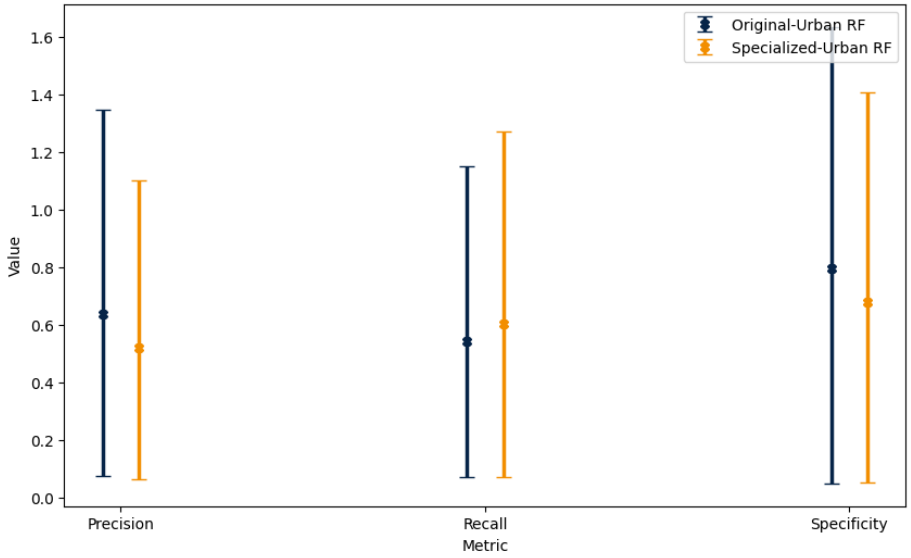
Model		**Precision**	**Recall**	**Specificity**
Rural		0.77	0.67	0.72
Urban		0.52	0.60	0.68
Original	rural	0.77	0.67	0.72
	urban	0.64	0.54	0.80

Table 4.6: Precision, Recall, Specificity of Rural RF, Urban RF and original RF models

Additionally, Figures 4.24 show the metrics and the relative confidence intervals of the Urban and Rural specialized models, in comparison to the original "Traffic-free" RF model, considered within the relative category.



(a) Rural RF vs. original RF



(b) Urban RF vs. original RF

Figure 4.24: Confidence intervals of Precision, Recall, Specificity of Rural RF and Urban RF versus the original RF (within the relative category).

In conclusion, in both cases, the Urban and Rural specialized models perform very similarly to the original RF model. These results suggest that the original RF model is not biased towards any specific category, and its performance accurately reflects the true predictive potential within each band.

However, as anticipated at the beginning of this Section, the significantly smaller number of instances in the "urban" category should be considered when discussing the final results within this category.
Thus, although this study demonstrates that the original "Traffic-free" RF model is functioning properly and at its highest potential, without being influenced by

the specific surrounding environment of the network cells, the same test should be repeated with a larger number of samples from the "urban" category.

Hence, continuing in this line of research, an interesting study could involve including the area category information as an additional feature in the dataset used by the model. However, this approach is not included in this thesis due to the current unavailability of balanced datasets between rural and urban categories, as previously shown in Table 4.5.

4.4 Testing with *RSI* dataset

As anticipated in Section 1.2, Telefónica's overall project aims at targeting a variety of common issues that arise and affect access networks.

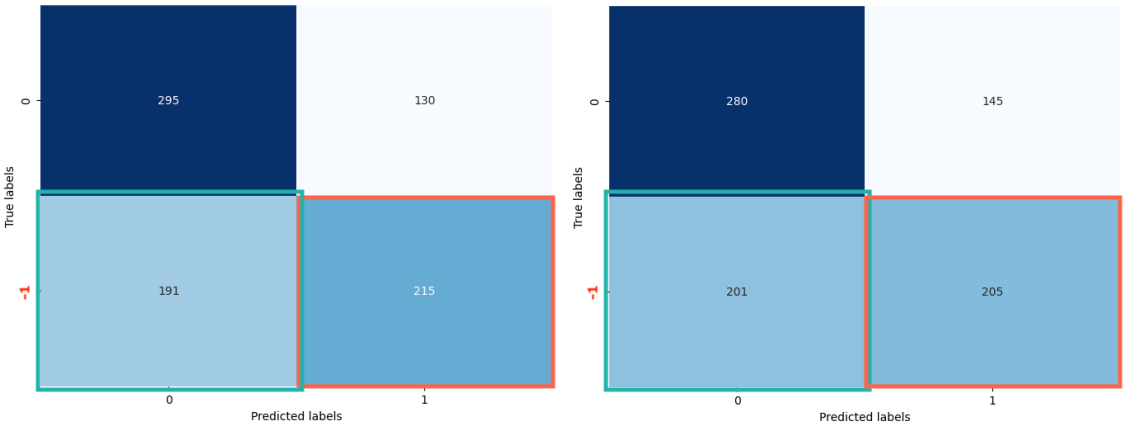
Although this thesis is focused on cell-range overshooting issues, which define the studied *problematic* class, a similar study is conducted in parallel to address another common issue: RSI collisions (detailed in Section 2.3).

As the final study presented in this thesis, a testing experiment is held exploiting a dataset comprising cells experiencing issues related to RSI collisions.

Similarly to the format of the *Training* dataset (Section 3.4), the *RSI* dataset encompasses 3-days aggregation (from 19th to 22nd of February) values of all the used KPIs for each sample.
The chosen *RSI* dataset consists of 831 samples: 406 RSI collision occurrences and 425 non-problematic cells. Thus, regarding the binary classification of *problematic* versus *random* within the study, the entire *RSI* dataset belongs to *random* class.

The output of this final test retrieves that Random Forest and Extreme Gradient Boosting classifiers fail to accurately predict RSI collision occurrences, as approximately half of them are misclassified as *problematic* cells. It is important to clarify that the term *problematic* adheres to the original definition provided in Section 3.2, representing overshooting occurrences. Hence, RSI collision is another type of phenomenon affecting cells.

In the following, the RF model's predictions on RSI collisions instances are further examined. The Figure 4.25 depict the confusion matrices resulting from the predictions of both models. In the true labels, the value "-1" (highlighted in red) denotes the RSI collision occurrences, which constitute a distinct class. Moreover, in the matrices, the coral box highlights the misclassifications (as *problematic* occurrences) while the box in aquamarine emphasizes the correctly classified instances (as *random*).



(a) Random Forest (b) XGBoost

Figure 4.25: Confusion matrix of RF and XGB predictions on *RSI* dataset.

Furthermore, in Figure 4.26 misclassified and correctly-predicted RSI collisions instances are grouped and the percentage per frequency band is depicted. It is evident that there is a substantial disparity between the percentage of misclassification (58%) and correctly-classified instances (31%) in band 800MHz. Also, it is worthy to notice the high performance of the model in band 1800MHz.

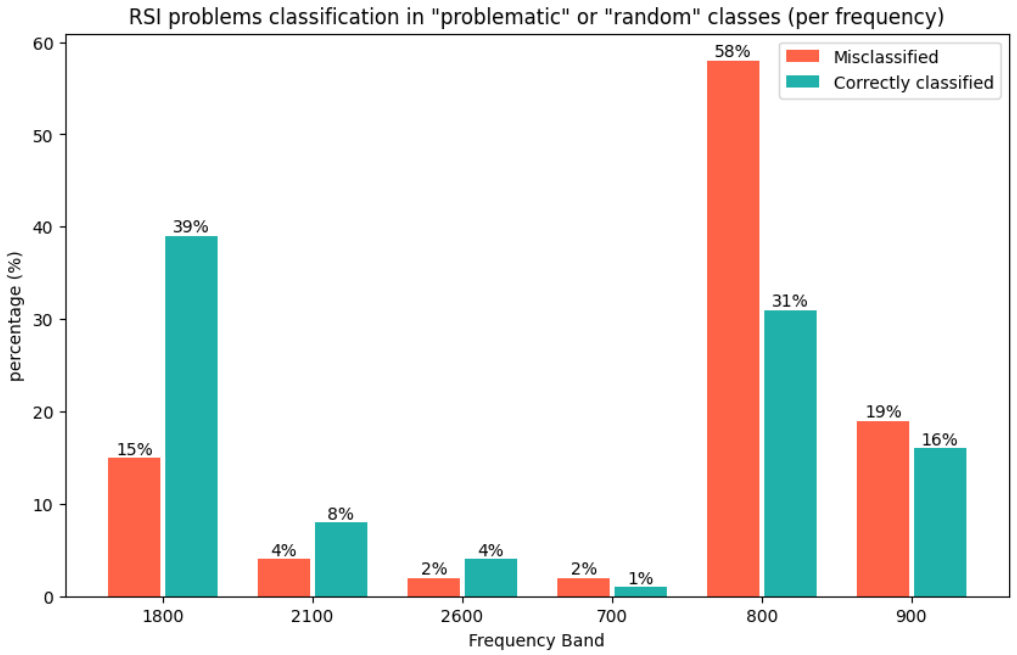


Figure 4.26: Distribution of RF predictions on RSI collisions per frequency band.

Both results confirm, once again, the same conclusions retrieved from testing with *Semana Santa* dataset (Section 4.1.2). On one side, the cells in 800MHz, being critical samples of mobile networks, are difficult to analyze for the models, which also in this case are unable to distinguish the type of KPI anomaly. On the other

hand, the models exhibit high predictive capabilities for the 1800MHz band and demonstrate acceptable performance for the rest. Nevertheless, these bands are less represented in datasets and in the network itself.

The primary goal of this test experiment is to verify whether the RF and XGB models effectively identify the specific issue of overshooting phenomenon. However, it's important to note that RSI collisions are closely associated with the concept of cell-range, which is inherently linked to the overshooting issue.

5. Conclusions and Future Work

In this final chapter, conclusions are drawn by revisiting the tests conducted during the Testing Phase (Chapter 4), with a focus on addressing the associated research limitations. Consequently, an overview of potential research directions for the near future is presented.

5.1 Conclusions

The increasing demand for seamless connectivity, coupled with the enormous expansion of mobile networks, has led to the urgent necessity for optimization and automation of these networks. In response, the world of machine learning (ML) offers powerful and innovative technologies to aid in this endeavor.

In the dynamic evolution of access networks, the traditional manual intervention of optimizers must transition towards the exploration and development of automatic optimization tools for detection, performance forecasting, and self-healing processes.

This work, conducted within Telefónica's research project, the *AI/ML Optimization Program 2024-2026*, focuses on the development of an automatic system for detecting cell-range overshooting in LTE networks.

In the initial phase, the study on available KPIs leads to the identification of a subset that proves to be informative about the status of cells. Consequently, primary classification models are designed and trained. These initial tools yielded promising results, demonstrating a good capability to detect instances of cells experiencing problems.

Subsequently, during the second testing phase, various test experiments are designed to address different aspects of mobile network performance that could influence detection.

Ultimately, the set of KPIs observed in this thesis proves to be informative regarding the status of cells, as the classification models successfully identify the majority of problematic cells. However, as demonstrated in Section 4.4, the developed classification models do not specifically detect overshooting issues; instead, they identify more general problematic behaviors.

Therefore, the primary focus of future research should involve investigating additional information specifically related to cell-range overshooting occurrences. This

effort will facilitate the development of specialized detection mechanisms tailored to identifying *problematic* cells for specific use cases.

5.2 Research Limitations and Future Work

As previously discussed, mobile networks are highly sensible to a diverse range of factors. Among these, traffic patterns, frequency bands and the surrounding environment of cells are examined through specialized testing experiments presented in Chapter 4.

In Section 4.1, the study on traffic impact reveals the influence of traffic on KPIs. The study demonstrates the varying behavior of KPIs in cells based on the level of traffic experienced.

It is important to note that traffic patterns change significantly throughout the year, influenced by user movements and behavioral trends. Traffic dynamics are a direct consequence of population density distribution and mobility patterns, which vary considerably with seasonal or temporal changes.
Consequently, the operational performance of these networks is intricately linked to seasonality.

In this thesis, the data is concentrated within the temporal window between March and April. This aspect must be considered when discussing the obtained results, as it represents a research limitation.

As a primary future direction, the classification models developed in this thesis should be tested over an annual temporal window. This aims to validate the results across a wide variety of seasonal traffic patterns.

Subsequently, leveraging the extensive available data, it may be feasible to create a balanced training dataset by ensuring an equal number of instances within each traffic level.
As consequence, another possibility would be to incorporate the level of traffic as additional information, introducing it as a new feature into the classification model. This approach could enhance the model's ability to directly discern between different traffic levels.

Conversely, in Section 4.2, the testing experiment addressing frequency bands also reveals distinct KPI behaviors within each frequency band.
Specifically, the band 800MHz proves to be very challenging with respect to the others due to its intrinsic characteristics.

Therefore, also in this case, a possibly optimal approach could be introducing the frequency band as explicit information for the classification model. Thereby, similar to the approach taken with traffic levels, leveraging a more extensive dataset and ensuring balance across frequency band categories may facilitate the incorporation of frequency band information as a new feature. This approach would enable the classification model to directly differentiate between cells operating in different fre-

quency bands.
Such effort could assist in addressing the challenge of accurately classifying instances within the 800MHz band, which has been identified as particularly critical.

Finally, in Section 4.3, the concept of network cell surrounding environment is introduced. Within a network, cells behaviors are highly affected by the surrounding environment characteristics, which are strictly related to other factors, e.g. traffic patterns.
Particularly, the study proposed in this thesis defined the "rural" and "urban" categories, dividing Telefónica's Spain-Vendor-A territory accordingly.
Although the results of such study are promising, suggesting a correct behavioral distinction between "rural" cells and "urban" ones, the scarcity of "urban" data must be addressed.

Hence, in the near future, taking advantage of more ample available data, it would be possible to comprehensively analyze the cellular surrounding environment by examining KPI behaviors in both "rural" and "urban" areas in more reliable studies. This methodology may confirm notable distinctions between these two settings. Consequently, similar to the previous considerations, this additional information could also be integrated as a new feature for observation by classification models.

A. List of KPIs

In this Appendix the full list of considered Key Performance Indicators (KPIs) is presented. For each KPI, the first column of the table reports the full original name as well as the simplified name, i.e. the unique `string` used within the datasets during the development and testing phases of the classification models.

KPI	*Description*	Group
This counter provides the average MCS used for TB transmission using Spatial Multiplexing transmission. (#) N4 [Average_MCS_used_for_TB_transmission_using_Spatial_Multiplexing_transmission_N4]	Average Modulation Coding Scheme used for TB transmission using Spatial Multiplexing	Usage
Average PDCP layer active cell throughput DL (kbps) N4 [Average_PDCP_layer_active_cell_throughput_DL_kbps_N4]	Average Packet Data Convergence Protocol throughput in Downlink (kbps)	Usage
Average PDCP layer active cell throughput UL (kbps) N4 [Average_PDCP_layer_active_cell_throughput_UL_kbps_N4]	Average Packet Data Convergence Protocol throughput in Uplink (kbps)	Usage
Average RSSI for PUCCH (dBm) N4 [Average_RSSI_for_PUCCH_dBm_N4]	Average Received Signal Strength Indicator in PUCCH channel (dBm)	Integrity
Average RSSI for PUSCH (dBm) N4 [Average_RSSI_for_PUSCH_dBm_N4]	Average Received Signal Strength Indicator in PUSCH channel (dBm)	Integrity
Averaged IP scheduled Throughput in DL, QCI1 (kbps) N4 [Averaged_IP_scheduled_Throughput_in_DL_QCI1_kbps_N4]	Average scheduled IP throughput in Downlink with QoS Class Identifier 1 (kbps)	Integrity
Averaged IP scheduled Throughput in UL, QCI1 (kbps) N4 [Averaged_IP_scheduled_Throughput_in_UL_QCI1_kbps_N4]	Average scheduled IP throughput in Uplink with QoS Class Identifier 1 (kbps)	Integrity

KPI	*Description*	Group
E-UTRAN Complete Contention Based RACH Setup Success Rate (%) N4 [Complete_Contention_Based_RACH_Setup_ Success_Rate_N4]	Ratio of received message-3 over message-1. Success rate (%) of contention-based connection setup	Accessibility Random Access
E-UTRAN E-RAB active drop ratio with data in the buffer due to RNL Radio Connection with UE Lost (%) N4 [E_RAB_active_drop_ratio_with_data_in_the_buffer_due_to_RNL_Radio_Connection_with_U	Drop rate (in case of data in the buffer) due to the loss of Radio Network Layer connection with UE	Retainability E-RAB
E-UTRAN E-RAB Drop Ratio, RAN View (%) N4 [E_RAB_Drop_Ratio_RAN_View_N4]	Drop Rate (%) of E-RAB connections from network perspective	Retainability E-RAB
E-UTRAN E-RAB Drop Ratio, User Perspective (%) N4 [E_RAB_Drop_Ratio_User_Perspective_N4]	Drop Rate (%) of E-RAB connections from user perspective	Retainability E-RAB
E-UTRAN E-RAB Setup Attempt (#) N4 [E_RAB_Setup_Attempt_N4]	Total number of E-RAB setup attempts	Accessibility E-RAB
E-UTRAN E-RAB Setup SR (%) N4 [E_RAB_Setup_SR_N4]	Success Rate of E-RAB setup attempts	Accessibility E-RAB
Éxitos de handover (%) N4 [Exitos_de_handover_N4]	Success rate (%) of HO procedure	Mobility HO
Éxitos de la fase de ejecución del handover (%) N4 [Exitos_de_la_fase_de_ejecucion_del_handover_N4]	Success rate (%) of HO execution phase (including all HO types)	Mobility HO
Éxitos de la fase de preparación del handover (%) N4 [Exitos_de_la_fase_de_preparacion_del_handover_N4]	Success rate (%) of HO preparation phase (including all HO types)	Mobility HO
E-UTRAN HO Preparation Success Ratio, intra eNB (%) N4 [HO_Preparation_Success_Ratio_intra_eNB_N4]	Success rate (%) of HO preparation phase (specific of intra-eNB)	Mobility HO
E-UTRAN HO Success Ratio, intra eNB (%) N4 [HO_Success_Ratio_intra_eNB_N4]	Success rate (%) of intra-eNB HO procedure	Mobility HO
E-UTRAN Initial Context Setup Failure Ratio due to Failed Radio N4 [Initial_Context_Setup_Failure_Ratio_due_to_Failed_Radio_N4]	Failure rate (%) of Initial Context Setup (due to Failed Radio Interface Procedure)	Accessibility S1
Initial E-RAB Accessibility (%) N4 [Initial_E_RAB_Accessibility_N4]	Success rate (%) of E-RAB connection establishment?	Accessibility E-RAB

KPI	*Description*	Group
E-UTRAN Inter-Frequency HO Success Ratio (%) N4 [Inter_Frequency_HO_Success_Ratio_N4]	Success rate (%) of inter-frequency HO procedure	Mobility HO
E-UTRAN Intra-Frequency HO Success Ratio (%) N4 [Intra_Frequency_HO_Success_Ratio_N4]	Success rate (%) of intra-frequency HO procedure	Mobility HO
MAX PRB usage per TTI DL (%) N4 [MAX_PRB_usage_per_TTI_DL_N4]	Maximum Physical Resource Block usage per TTI DL	Usage
MAX PRB usage per TTI UL (%) N4 [MAX_PRB_usage_per_TTI_UL_N4]	Maximum Physical Resource Block usage per TTI UL	Usage
Maximum Active UEs with data in the buffer per cell DL (#) N4 [Maximum_Active_UEs_with_data_in_the_buffer_per_cell_DL_N4]	Maximum data in the buffer of cell (from active UE in Downlink)	Usage
Maximum Active UEs with data in the buffer per cell UL (#) N4 [Maximum_Active_UEs_with_data_in_the_buffer_per_cell_UL_N4]	Maximum data in the buffer of cell (from active UE in Uplink)	Usage
MIN PRB usage per TTI DL (%) N4 [MIN_PRB_usage_per_TTI_DL_N4]	Minimum Physical Resource Block usage per TTI UL	Usage
MIN PRB usage per TTI UL (%) N4 [MIN_PRB_usage_per_TTI_UL_N4]	Minimum Physical Resource Block usage per TTI UL	Usage
Number of failed Inter eNB Handover preparations per cause per neighbor cell relationship (#) N4 [Number_of_failed_Inter_eNB_Handover_preparations_per_cause_per_neighbor_cell_relationship_N4]	Counter of failed inter-eNB HO preparations (aggregated per cause and per neighbour cell)	Mobility HO
Number of failed Inter eNB Handover preparations per neighbor cell relationship due to failures in the HO preparation (#) N4 [Number_of_failed_Inter_eNB_Handover_preparations_per_neighbor_cell_relationship_due_to_failures_in_the_HO_preparation_N4]	Counter of failed inter-eNB HO preparations due to failure in HO preparation (aggregated per neighbour cell)	Mobility HO

KPI	*Description*	Group
Number of failed Inter eNB Handover preparations per neighbor cell relationship due to the expiration of the respective guarding timer (#) N4 [Number_of_failed_Inter_eNB_Handover_preparations_per_neighbor_cell_relationship_due_to_the_expiration_of_the_respective_guarding_timer_N4]	Counter of failed inter-eNB HO preparations due to expiration of guard-time (aggregated per neighbour cell)	Mobility HO
Number of Inter-eNB Handover attempts per neighbor cell relationship (#) N4 [Number_of_Inter_eNB_Handover_attempts_per_neighbor_cell_relationship_N4]	Counter of inter-eNB HO attempts (aggregated per neighbour cell)	Mobility HO
Number of Inter eNB Handover failures per cause per neighbor cell relationship (#) N4 [Number_of_Inter_eNB_Handover_failures_per_cause_per_neighbor_cell_relationship_N4]	Counter of failed inter-eNB HO procedures (aggregated per cause and per neighbour cell)	Mobility HO
Number of successful Inter-eNB Handover completions per neighbor cell relationship (#) N4 [Number_of_successful_Inter_eNB_Handover_completions_per_neighbor_cell_relationship_N4]	Counter of successful inter-eNB HO completions (aggregated per neighbour cell)	Mobility HO
Number of successful Intra-eNB Handover completions per neighbor cell relationship (#) N4 [Number_of_successful_Intra_eNB_Handover_completions_per_neighbor_cell_relationship_N4]	Counter of successful intra-eNB HO completions (aggregated per neighbour cell)	Mobility HO
Percentage of PDSCH transmissions using Low MCS Codes (MCS¡=9) (%) N4 [Percentage_of_PDSCH_transmissions_using_Low_MCS_Codes_MCS_9_N4]	Percentage (%) of Downlinklink transmissions with Modulation Coding Scheme loweror equal to 9	Usage
E-UTRAN Percentage of PUSCH transmissions using High MCS Codes (MCS¿=20) (%) N4 [Percentage_of_PUSCH_transmissions_using_High_MCS_Codes_MCS_20_N4]	Percentage (%) of Uplink transmissions with Modulation Coding Scheme higher or equal to 20	Usage
E-UTRAN Percentage of PUSCH transmissions using Low MCS Codes (MCS¡=9) (%) N4 [Percentage_of_PUSCH_transmissions_using_Low_MCS_Codes_MCS_9_N4]	Percentage (%) of Uplink transmissions with Modulation Coding Scheme lower or equal to 9	Usage

KPI	*Description*	Group
PRB usage per TTI DL (#) N4 [PRB_usage_per_TTI_DL_N4]	Average Physical Resource Block usage per TTI DL	Usage
PRB usage per TTI DL (#) N4 [PRB_usage_per_TTI_UL_N4]	Average Physical Resource Block usage per TTI UL	Usage
E-UTRAN RACH Setup Attemps (%) N4 [RACH_Setup_Attempts_N4]	Total number of received Random Access preambles (message-1).	Accessibility Random Access
E-UTRAN RACH Setup Completion Success Rate (%) N4 [RACH_Setup_Completion_Success_Rate_N4]	Ratio of received message-2 over message-1. Success rate (%) of transmitted message-2 after reception of message-1	Accessibility Random Access
E-UTRAN Radio Bearer Drop Ratio (%) N4 [Radio_Bearer_Drop_Ratio_N4]	Drop rate (%) of E-RAB connection	Accessibility E-RAB
RRC Connection Re-establishment Attempts, HO fail (#) N4 [RRC_Connection_Re_establishment_Attempts_HO_fail_N4]	Total number of RCC conection re-establishment attempts due to HO fails	Retainability RCC
RRC Connection Re-establishment rejection Ratio (%) N4 [RRC_Connection_Re_establishment_rejection_Ratio_N4]	Rejection rate (%) of RCC connection re-estabishments	Retainability RCC
RRC Connection Setup Attempts (#) N4 [RRC_Connection_Setup_Attempts_N4]	Total number of RCC conection setup attempts	Retainability RCC
E-UTRAN RRC Connection Setup FR (%) N4 [RRC_Connection_Setup_FR_N4]	Failure rate (%) of RCC connection setup	Retainability RCC
E-UTRAN S1 Initial Context Setup Attempts (#) N4 [S1_Initial_Context_Setup_Attempts_N4]	Total number of S1 initial context setup attempts	Accessibility S1
Total E-UTRAN RRC Connection Re-establishment Failure Ratio (%) N4 [Total_E_UTRAN_RRC_Connection_Re_establishment_Failure_Ratio_N4]	Failure rate (%) of RCC connection re-estabishments	Retainability RCC

B. Mann-Whitney U Test

This Appendix introduces the Mann-Whitney U Test, which is employed in Chapter 3 and Chapter 4 of this thesis.
Moreover, in this Appendix the full list of results of the Mann-Whitney U tests of the considered KPIs is reported.

The *Mann-Whitney U Test*[1] is a non-parametric test of the null hypothesis that the distribution underlying one sample x is the same as the distribution underlying on other sample y. It is often used as a test of difference in location between distributions.

Calculation of Mann-Whitney U test

$$U = n_1 * n_2 + \frac{n_2 * (n_2 + 1)}{2} - \sum_{i=n_1+1}^{n_2} R_i \tag{B.1}$$

In the context of the Mann-Whitney U test, the U statistic is a measure of the extent of difference between the two samples. The p-value represents the probability of obtaining a test statistic at least as extreme as the observed value, assuming the null hypothesis is true. Typically, a low p-value (usually below 0.05 to ensure 95% confidence interval) indicates a significant difference between the populations, while a high p-value suggests little difference.

Assumptions of Mann-Whitney U test

Non-parametric tests, also known as distribution-free tests, are utilized when the data within the populations of interest do not adhere to a normal distribution. The Mann-Whitney U-test serves as a non-parametric counterpart to the unpaired Student's T-test[2], which is employed under the assumption that the two populations being compared follow a normal distribution, characterized by their means and standard deviations.

Key assumptions for the Mann-Whitney U Test include the following:

[1] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`

[2] `https://www.technologynetworks.com/informatics/articles/paired-vs-unpaired-t-test-differences-assumptions-and-hypotheses-330826`

- The variable under comparison between the two groups must be continuous. This requirement stems from the test's reliance on ranking observations within each group.
- The data are presumed to follow a non-normal (or skewed) distribution.
- Although the data in both groups are not expected to be normally distributed, it is assumed that the data exhibit a similar shape across the two groups.
- The samples should consist of two randomly selected independent groups, with no interrelation between them.
- A sufficient sample size is necessary for the test's validity, typically requiring more than five observations in each group.

These assumptions ensure the appropriate application and interpretation of the Mann-Whitney U Test.

As shown in the graphic visualization of the distributions in Chapter 3, the considered KPIs prove to be non-normally distributed. Moreover, it is assumed that the two classes, namely *problematic* and *random*, are independent groups, with very low interrelation between them.

In conclusions, the Mann-Whitney U Test is the appropriate to perform the comparison of the *problematic* and *random* populations under consideration in this thesis.

Mann-Whitney U tests of *Problematic* versus *Random* KPI Distributions

The results of the Mann-Whitney U tests for the full list of considered KPIs are reported in the following table.

KPI	*p*	statistic
Average PDCP layer active cell throughput DL (kbps) N4	< 0.001	1351182
Average PDCP layer active cell throughput UL (kbps) N4	< 0.001	1446754
Average RSSI for PUCCH (dBm) N4	0.02856	1254959.5
Average RSSI for PUSCH (dBm) N4	< 0.001	1473436
Averaged IP scheduled Throughput in DL, QCI1 (kbps) N4	0.0046	1129778.5
Averaged IP scheduled Throughput in UL, QCI1 (kbps) N4	< 0.001	1408884
E-UTRAN Complete Contention Based RACH Setup Success Rate (%) N4	< 0.001	1728854

KPI	p	statistic
E-UTRAN E-RAB Drop Ratio, RAN View (%) N4	< 0.001	885074.5
E-UTRAN E-RAB Drop Ratio, User Perspective (%) N4	< 0.001	887257.5
E-UTRAN E-RAB Setup Attempt (#) N4	< 0.001	1487654.5
E-UTRAN E-RAB Setup SR (%) N4	< 0.001	1501339
E-UTRAN E-RAB active drop ratio with data in the buffer due to RNL Radio Connection with UE Lost (%) N4	< 0.001	897039
E-UTRAN HO Preparation Success Ratio, intra eNB (%) N4	< 0.001	1048302.5
E-UTRAN HO Success Ratio, intra eNB (%) N4	< 0.001	1554976
E-UTRAN Initial Context Setup Failure Ratio due to Failed Radio N4	< 0.001	915922
E-UTRAN Inter-Frequency HO Success Ratio (%) N4	< 0.001	1555893
E-UTRAN Intra-Frequency HO Success Ratio (%) N4	< 0.001	1547623.5
E-UTRAN Percentage of PUSCH transmissions using High MCS Codes ($MCS >= 20$) (%) N4	< 0.001	1462345
E-UTRAN Percentage of PUSCH transmissions using Low MCS Codes ($MCS <= 9$) (%) N4	< 0.001	956962
E-UTRAN RACH Setup Attemps (%) N4	0.2691	1227954.5
E-UTRAN RACH Setup Completion Success Rate (%) N4	< 0.001	1715613
E-UTRAN RRC Connection Setup FR (%) N4	< 0.001	802276
E-UTRAN Radio Bearer Drop Ratio (%) N4	< 0.001	874858
E-UTRAN S1 Initial Context Setup Attempts (#) N4	< 0.001	1484997.5
Initial E-RAB Accessibility (%) N4	< 0.001	1575624.5
MAX PRB usage per TTI DL (%) N4	0.9920	1200635.5
MAX PRB usage per TTI UL (%) N4	0.0886	1239639
Maximum Active UEs with data in the buffer per cell DL (#) N4	< 0.001	1416103.5
Maximum Active UEs with data in the buffer per cell UL (#) N4	< 0.001	1385457.5
Number of Inter eNB Handover failures per cause per neighbor cell relationship (#) N4	< 0.001	1004520.5
Number of Inter-eNB Handover attempts per neighbor cell relationship (#) N4	< 0.001	1365211.5

KPI	*p*	statistic
Number of failed Inter eNB Handover preparations per cause per neighbor cell relationship (#) N4	0.0710	1242077.5
Number of failed Inter eNB Handover preparations per neighbor cell relationship due to failures in the HO preparation (#) N4	0.0031	1219361
Number of failed Inter eNB Handover preparations per neighbor cell relationship due to the expiration of the respective guarding timer (#) N4	0.3772	1218302
Number of successful Inter-eNB Handover completions per neighbor cell relationship (#) N4	< 0.001	1373572.5
Number of successful Intra-eNB Handover completions per neighbor cell relationship (#) N4	< 0.001	1510015
PRB usage per TTI DL (#) N4	< 0.001	1362877.5
PRB usage per TTI UL (#) N4	< 0.001	1289764
Percentage of PDSCH transmissions using Low MCS Codes ($MCS <= 9$) (%) N4	0.7094	1209711
RRC Connection Re-establishment Attempts, HO fail (#) N4	< 0.001	901396.5
RRC Connection Re-establishment rejection Ratio (%) N4	< 0.001	1289654
RRC Connection Setup Attempts (#) N4	< 0.001	1473430.5
This counter provides the average MCS used for TB transmission using Spatial Multiplexing transmission. (#) N4	0.0057	1131641.5
Total E-UTRAN RRC Connection Re-establishment Failure Ratio (%) N4	< 0.001	1107305.5
Éxitos de handover (%) N4	< 0.001	1530418.5
Éxitos de la fase de ejecución del handover (%) N4	< 0.001	1595462.5
Éxitos de la fase de preparación del handover (%) N4	0.0617	1155605.5

C. Performance Metrics and Confidence Intervals

In this Appendix, the methodology used to compute the performance metrics - precision, recall and specificity - are described.

Furthermore, the bootstrap resampling technique utilized to estimate the confidence intervals relative to the previously mentioned performance metrics is presented in details.

Performance Metrics

In this thesis three different performance metrics are exploited during the testing phase of the project, presented in Chapter 4.

Specifically, precision, recall and specificity are utilized to measure and compare the classification performances.

Precision

Precision measures the accuracy of positive predictions. It is the fraction of true positive results among all positive results predicted by the model.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{C.1}$$

Recall

Recall measures the model's ability to identify all relevant instances. It is the fraction of true positive results among all actual positive instances.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{C.2}$$

Specificity

Specificity measures the model's ability to identify only the relevant negative instances. It is the fraction of true negative results among all actual negative instances.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN))} + \text{False Positives (FP)}} \tag{C.3}$$

Bootstrap Method for Confidence Intervals

[10] To estimate the 95% confidence intervals for the previously mentioned performance metrics, the bootstrap resampling technique is employed.

The bootstrap method provides a reliable means of estimating confidence intervals for performance metrics without assuming a specific distribution. By leveraging this non-parametric method, the robustness of interval estimates is ensured.

Bootstrap Method Steps

1. **Resampling**: generate a large number of *bootstrap samples* (specifically, $n = 10000$) from the original dataset by random sampling with replacement.
2. **Metric Calculation**: for each *bootstrap sample*, compute the desired performance metric.
3. **Confidence Interval Estimation**: determine the confidence interval by finding the appropriate percentiles from the distribution of the bootstrapped metric values. For a 95% confidence interval, the 2.5th and the 97.5th percentiles are used.

D. Key Technologies

This Appendix provides a comprehensive overview of the key technologies employed in this thesis.

As Telefónica's Radio Optimization department transitions certain operations to Databricks, this project represents an initial step in this technological shift towards advanced data analysis.

The work of this thesis utilizes Azure Databricks, leveraging PySpark and Kusto Query Language (KQL).

Azure Databricks

Azure Databricks is a managed version of the Databricks platform within the Azure cloud environment. *Databricks* is a unified, open analytics platform built on Apache Spark, designed for creating, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. The Databricks Data Intelligence Platform integrates with cloud storage and security in the user's cloud account, managing and deploying cloud infrastructure on their behalf.

Databricks offers scalability, advanced analytics, and real-time processing capabilities. It provides high programming flexibility and integration, enabling the use of big data databases. Its distributed architecture effectively handles massive data volumes, making it ideal for organizations dealing with big data or complex data processing tasks. Additionally, Databricks supports multiple programming languages (e.g., Python, R, SQL) and libraries commonly used in data science and ML.

Specifically, for the purpose of this thesis, among others, Python `sklearn`[1] library and `xgboost`[2] package are exploited to develop the classification models, alongside `sciPy`[3] library which is employed to perform the statistical Mann-Whitney U test.

The following sections detail the data accessible from the Databricks platform at Telefónica.

Firstly, topological and physical information about network cells can be extracted, filtering by supplier (e.g., Vendor-A or Vendor-B) and geographical region or tech-

[1] `https://scikit-learn.org`
[2] `https://xgboost.readthedocs.io/en/stable/python/python_intro.html`
[3] `https://scipy.org`

nology employed (e.g., 3G, 4G/LTE or 5G). Each record represents a cell, and each field provides specific information about that cell. It is recommended to filter by registration date to ensure accurate information at a specific time.

Additionally, a different database can be consulted to provide information about specific Key Performance Indicator (KPI)s. Each record includes the KPI name, the formula defined to calculate it, and the counters involved.

As previously explained, KPIs are functions of several counters captured every 15 minutes. The KPI computation time window can be expanded by specifying the time aggregation parameter. Time aggregation involves computing the official formula using the sum of all counters related to the same metric over the entire specified time window.

Finally, a third database can be queried to obtain cells' KPIs for a specific date. This is the primary database used in this thesis to query records for various datasets. A particular function implemented in Python takes the list of requested cell names, the desired KPIs, the query date, and the time aggregation as inputs. It processes this information and returns a PySpark DataFrame with entries representing KPI values for specific cells operating in a particular frequency band.

Spark and PySpark

This thesis leverages Spark sessions within Azure Databricks and integrates PySpark.

Spark is an open-source distributed computing system for programming entire clusters with implicit data parallelism and fault tolerance. It is designed for big data processing and analytics, offering in-memory computation and supporting various programming languages like Scala, Java, Python, and R.

PySpark is the Python API for Apache Spark, enabling users to leverage Spark's distributed computing capabilities using Python.

Kusto Query Language

To access Telefónica's data sources, Kusto Query Language (KQL) is used for defining queries. KQL queries are defined and integrated with PySpark commands to retrieve and store the results in a PySpark DataFrame, facilitating further data processing and analysis within the Azure Databricks environment.

Kusto Query Language (KQL), developed by Microsoft, is a query language used to interact with Azure Data Explorer for analyzing large volumes of data.

KQL provides a rich set of operators and functions for filtering, aggregating, and transforming data. It supports complex queries involving joins, unions, and subqueries.

In KQL, operators are sequenced by a — (pipe), and the data is filtered or manipulated at each step before being fed into the following step. This sequential piping

of information makes the order of query operators important, which can affect both results and performance.

An example KQL query might look like the following:

```
cells_new
| where tech_id == '4' and registration_date >= datetime('2024-02-02')
| project cell_name, supplier, region, province, site_name, site_id,
band, transmit_power, longitude, latitude, azimuth, registration_date
| summarize arg_max(registration_date, *) by cell_name
```

This example query filters records within a specific registration date and technology of LTE (4G), returning a selection of attributes aggregates the data by category, and sorts the results in descending order based on the total value.

Acronyms

3GPP Third Generation Partnership Project. 8, 9

CCA Canonical Correlation Analysis. 14

CM Configuration Management. 17

DL-SCH Downlink Shared Channel. 6, 7

DT Decision Tree. 28

EDA Exploratory Data Analysis. 17, 22, 23, 37

eNB evolved Node-B. 6, 7, 9, 11, 12

FDD Frequency Division Duplexing. 8

HO Handover (or Handoff). 4–6, 12, 21, 27

IM Inventory Management. 17

KPI Key Performance Indicator. v, xviii, 1, 4, 13, 14, 18, 21–23, 27–31, 34, 37–41, 43, 47, 49, 51, 58, 62, 65–67

KQL Kusto Query Language. xvii, xviii, 18

LR Logistic Regression. 29

LTE Long Term Evolution. 1, 3, 5, 12–14, 65

ML Machine Learning. xvii, 2, 15, 17, 18, 28

PCA Principal Component Analysis. 39–41

PDCCH Physical Downlink Control Channel. 7

PM Performance Management. 17

PRACH Physical Random Access Channel. 6, 7, 9, 11, 14, 19

prachCS PRACH Cyclic Shift. 9–12, 14, 19, 24

RACH Random Access Channel. 21

RAN Random Access Network. 5

RF Random Forest. 28–30, 34, 37, 41–46, 48, 50, 51, 53, 54, 58, 60–62, 64

ROC Receiver operating characteristic. 34

RRC Radio Resource Control. 6, 7, 11, 15

RSI Root Sequence Index. 9, 11, 12, 37, 62, 64

SIB System-Information Block. 5–7

SOM Self-Organizing Map. 13

SON Self-Organizing Network. 13

SQL Structured Query Language. xvii, 15

SVM Support Vector Machine. 29

TDD Time Division Duplexing. 8

UE User Equipment. 4–7, 9–12, 15

UL-SCH Uplink Shared Channel. 7

XGB Extreme Gradient Boosting. 28–30, 34, 37, 41–44, 46, 50, 62, 64

Bibliography

[1] Chris Johnson. *Long Term Evolution in Bullets*. 1st edition. Createspace Independent Pub, 2010. ISBN: 978-14528346419.

[2] Gabriela Ciocarlie et al. "On the feasibility of deploying cell anomaly detection in operational cellular networks". In: *2014 IEEE Network Operations and Management Symposium (NOMS)*. 2014.

[3] Stefan Parkvall Erik Dahlman and Johan Sköld. *4G: LTE/LTE-Advanced for Mobile Broadband*. 2nd edition. Chapters 7, 14. Academic Press, 2014. ISBN: 978-0124199859.

[4] Ana Gómez-Andrades et al. "Automatic Root Cause Analysis for LTE Networks Based on Unsupervised Techniques". In: *IEEE Transactions on Vehicular Technology* 65.4 (2016), pp. 2369–2386.

[5] R. Veríssimo et al. "PCI and RSI conflict detection in a real LTE network using supervised learning". In: *URSI Radio Science Bulletin* 2018.364 (2018), pp. 11–19.

[6] V. Zhou. "A Simple Explanation of Gini Impurity". In: *Victor Zhou* (2019). URL: `https://victorzhou.com/blog/gini-impurity/` (visited on 04/28/2024).

[7] Jessica Moysen et al. "Big Data-driven Automated Anomaly Detection and Performance Forecasting in Mobile Networks". In: *2020 IEEE Globecom Workshops (GC Wkshps)*. 2020.

[8] Furqan Ahmed, Muhammad Zeeshan Asghar, and Jyri Hämäläinen. "A canonical correlation-based framework for performance analysis of radio access networks". In: *2022 IEEE Globecom Workshops (GC Wkshps)* (2022). URL: `https://api.semanticscholar.org/CorpusID:252596005`.

[9] Elliot McClenaghan. "Mann-Whitney U Test: Assumptions and Example". In: *Technology Networks logo* (2022). URL: `https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425` (visited on 06/01/2024).

[10] Samantha Lomuscio. "Bootstrap Estimates of Confidence Intervals". In: *University of Virginia Library* (2023). URL: `https://library.virginia.edu/data/articles/bootstrap-estimates-of-confidence-intervals` (visited on 06/09/2024).

[11] P. Pandey. "Decision Tree Ensemble Model". In: *Medium* (2023). URL: `https://medium.com/@pankaj_pandey/decision-tree-ensemble-model-f422ba280fa8`.

[12] Sonia Trirahmi, Siska Aulia, and Dikky Chandra. "Analisa Pengaruh Issue Overshooting Cell Terhadap Coverage dan Quality Jaringan 4G LTE Kelurahan Jati Baru Kota Padang". In: *JURNAL AMPLIFIER : JURNAL ILMIAH BIDANG TEKNIK ELEKTRO DAN KOMPUTER* (2023). URL: `https://api.semanticscholar.org/CorpusID:265548755`.

[13] B. Wohlwend. "Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning". In: *Medium* (2023). URL: `https://`

medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948.

[14] Sadrach Pierre and Brennan Whitfield. "A Step-by-Step Explanation of Principal Component Analysis (PCA)". In: *Builtin* (2024). URL: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (visited on 06/16/2024).

[15] S.A. Telefonica. *Mission*. 2024. URL: https://www.telefonica.com/en/about-us/mission/ (visited on 05/31/2024).