UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

PREDICCIÓN DE COMPORTAMIENTO EN TRÁFICO DE RED LTE Y AJUSTE DE PARAMETRIZACIÓN PARA MAXIMIZAR PERFORMANCE DE RED

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

CRISTOFER JESÚS BENAVIDES RIVERA

PROFESOR GUÍA:
JAVIER SANTA ANA ÁLVAREZ

MIEMBROS DE LA COMISIÓN:
CESAR AZURDIA MEZA
PATRICIO VALENZUELA CANO

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA
POR: CRISTOFER JESÚS BENAVIDES RIVERA
FECHA: 26/03/2021
PROF. GUÍA: JAVIER SANTA ANA ALVAREZ

PREDICCIÓN DE COMPORTAMIENTO EN TRÁFICO DE RED LTE Y AJUSTE DE PARAMETRIZACIÓN PARA MAXIMIZAR PERFORMANCE DE RED

Debido al actual crecimiento y aumento en la complejidad de las redes de telefonía móvil, cada vez se hace más difícil el monitoreo y optimización de estas para mantener una calidad de servicio acorde al mercado. Es por esto que empresas como Entel están en busca de soluciones para automatizar procesos y dejar de lado el análisis descriptivo, reemplazándolo por análisis predictivo. Esto es posible utilizando nuevas tecnologías desarrolladas actualmente basadas en inteligencia artificial, específicamente aprendizaje de máquinas.

Se implementó un modelo predictivo compuesto por tres bloques, con el objetivo principal de detectar celdas LTE con mal rendimiento de throughtput, utilizando un umbral de 3.3 Mbps, esto para su posterior optimización utilizando balanceo de carga, es decir, traspaso de usuarios de una celda crítica a una de buen estado. El primer bloque es un algoritmo clasificador que diferencia entre celdas con buen y mal rendimiento, mientras que el segundo bloque selecciona las mejores celdas para el balanceo de carga, para finalmente aplicar un algoritmo regresivo a cada una en el tercer bloque.

El modelo desarrollado cumplió los objetivos estipulados, obteniendo altas tasas de detección de celdas con mal rendimiento y un bajo error en las curvas obtenidas con la regresión. Finalmente gracias a los buenos resultados el modelo esta siendo implementado por el equipo de optimización de Entel.

A mi familia por todo el apoyo desde siempre.

Tabla de Contenido

Índice de Tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación

Con la entrada del 5G se espera que las celdas de telefonía móvil se hagan más pequeñas y en consecuencia aumenten su densidad por área. Para garantizar una buena calidad de servicio es necesario optimizar de forma individual parámetros en cada celda, debido a la naturaleza del tráfico de datos pues es distinto según la zona. Actualmente este procedimiento se realiza manualmente, pero las implicancias de 5G y el aumento de la densidad del tráfico de datos en la población apuntan a la automatización en la configuración de redes, concepto conocido como SON (Self Operating Networks).

Actualmente en red LTE Entel existen más de 30000 celdas distribuidas a nivel nacional, esta gran cantidad de celdas en conjunto con el alto número de parámetros manejados por equipo de Optimización de red genera una dificultad mayor al momento de optimizar y parametrizar las celdas de una manera eficiente y con la celeridad que la tecnología y clientes necesitan. Es así como la empresa desea implementar un sistema de automatización, en base a predicciones del comportamiento futuro de parámetros en sitios de telefonía móvil, para así maximizar el rendimiento de la red y satisfacción de los clientes.

1.2. Objetivos y Alcances

1.2.1. Objetivo general

El objetivo de esta memoria consiste en la implementación de una prueba de concepto para el control preventivo, que determine el estado de celdas LTE según el tráfico de red para luego proponer su uso práctico. El foco esta en la predicción del comportamiento de red a partir de los KPI, para luego aplicar el control preventivo. La importancia de este trabajo radica en la factibilidad de su aplicación, vale decir evaluar positivamente el rendimiento de la herramienta diseñada, y los procedimientos llevados a cabo.

1.2.2. Objetivos específicos

- Predecir el comportamiento del throughput en base al historial de indicadores de rendimiento.
- Identificar con alta precisión las celdas con buen y mal rendimiento de throughtput utilizando el umbral mínimo aceptado por Entel de 3.3 Mbps.
- Permitir aplicar cambios de parámetro para llevar a cabo el proceso de optimización de la red en base al traspaso de usuarios entre celdas de bajo a buen rendimiento. .

1.2.3. Alcances y Limitaciones

- El Estudio y análisis solo se realizará en la tecnología LTE FDD, pues aun no hay instalaciones comerciales 5G en Chile.
- Análisis considerará solo en RM, en especial se escogerá uno o más polígonos de RM, que cumplan mejor con las condiciones iniciales para el trabajo.
- Entel entregará toda la información necesaria para la realización de la memoria.

Capítulo 2

Marco Teórico

2.1. Antecedentes

La telefonía móvil se ha caracterizado por su rápida evolución durante las ultimas tres décadas, partiendo con voluminosos teléfonos que solo permitían realizar llamadas, hasta los pequeños smartphones, cuyo potencial posibilita actividades que requieren una muy buena calidad de servicio, como por ejemplo streaming. Hoy en día la mejora del servicio se basa principalmente en el aumento de las tasas de transmisión y la disminución de latencia, aspectos que en la actualidad motivan la implementación del 5G.

A continuación se abordarán las últimas etapas de esta evolución, partiendo con una introducción de la tercera generación para luego dar paso a LTE, cuyas características serán el tema principal. En particular se estudiará la arquitectura de red y los conceptos de capa física asociados, para así llegar a los parámetros e indicadores de rendimiento relacionados a una red 4G, finalmente se mencionan aspectos generales de 5G.

2.2. Tercera generación

El trabajo de determinar las especificaciones que definen la tercera generación fue llevado a cabo por la 3GPP, entidad conformada por organizaciones pertenecientes al rubro de las telecomunicaciones. Esto es implementado en una serie de documentos denominados *releases* o lanzamientos, en particular se da inicio a la tercera generación con el release 99 o R99, publicado en el año 2000.

En dicho release se definió el sistema UMTS (Universal Mobile Telecommunications System) como el estándar sucesor de la tecnología 2G. Como fue mencionado, la primera etapa de la tercera generación que se denominó R99, fue creada con el objetivo de satisfacer los requisitos necesarios para navegación en internet, fundamentalmente servicios multimedia como transmisión de imágenes en tiempo real o videollamadas. Para permitir dichos servicios las redes implementadas con este estándar permitieron tasas de hasta 384 kbps para usuarios en movimiento y hasta 2 Mbps en el caso de usuarios estáticos [1].

Antes de definir la estructura especifica de la tercera generación es necesario mencionar la arquitectura general definida por la 3GPP, pues esta se ha mantenido hasta la actualidad. Se basa en 3 grandes bloques que todo estándar ha seguido [2], como se muestra en la figura 2.1.

- Equipo de usuario (UE): Dispositivo destinado al usuario final, permite acceder a los servicios ofrecidos en la red. Se compone del dispositivo físico mas una tarjeta inteligente (UICC o SIM), cuya utilidad es dar un identificador único con capacidad de acceso a la red.
- Red de acceso (RAN): Componente intermedia entre el equipo de usuario y la red CORE, posibilita la transmisión por radio entre estas dos. Proporciona los llamados *radio bearer* o RB, que corresponde al enlace donde ocurre el flujo de información. En otras palabras, la red de acceso organiza los recursos de radio disponibles según lo ordenado por la red troncal.
- Red central o CORE (CN): Bloque final de la red, entre sus funciones esta el control de los servicios aplicados en la red de acceso, gestión de la información de usuarios, gestión de movilidad y la administración de enrutamiento.

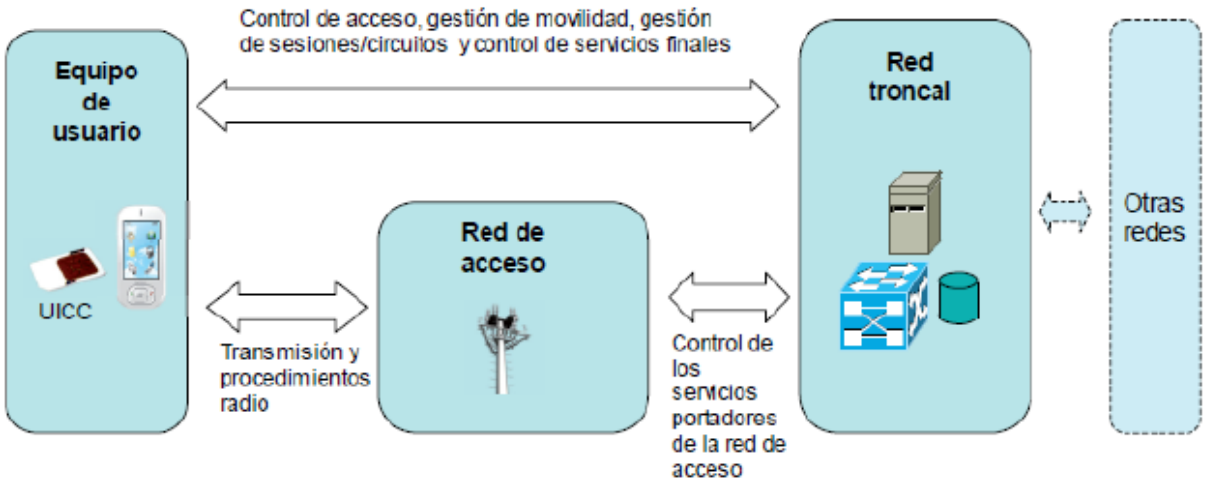


Figura 2.1: Arquitectura de red simplificada.

En el caso especifico del estándar R99, la red CORE mantiene el mismo nombre y su arquitectura es heredada de las redes 2G, por otro lado la red de acceso recibió el nombre UTRAN, es en esta componente donde se aplican los principales cambios especificados por el R99. En el caso del equipo de usuario, mantiene la misma lógica descrita anteriormente. En la figura 2.2 se muestra un esquema detallado de la arquitectura definida en el R99 junto a sus interfaces.

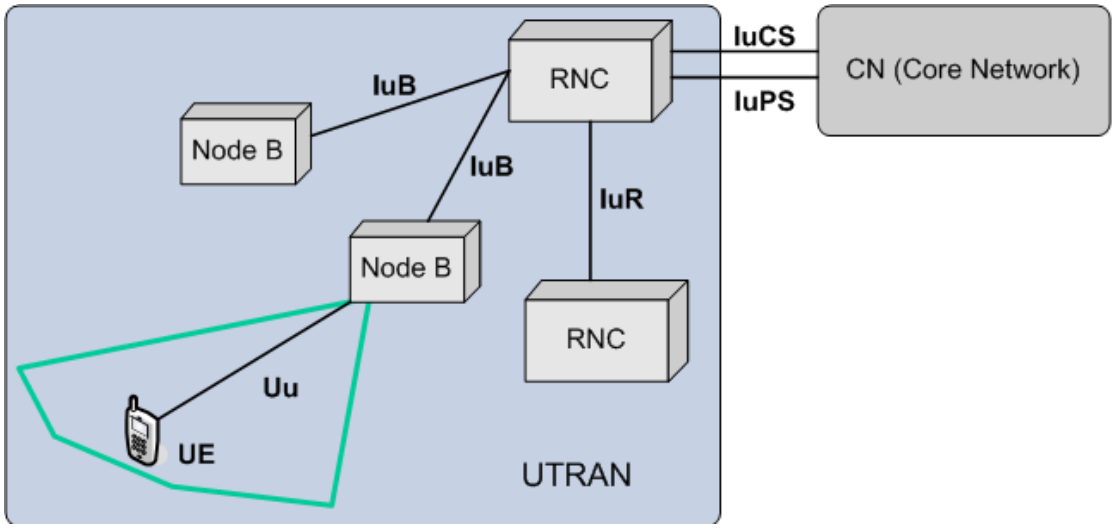


Figura 2.2: Esquema simplificado de una red 3G.

2.2.1. Red de acceso: UTRAN

Bloque encargado de permitir la conexión entre el terminal y la red CORE. Se compone de dos sistemas, la estación base o NodeB y el controlador o RNC. Entre sus principales funciones destacan las siguientes:

- Funcionalidades de capa física como codificación, modulación y acceso al medio.
- Aplicaciones de capa de enlace como control de errores y compresión de paquetes, como el protocolo de reenvío de mensajes ARQ (Automatic Repeat reQuest).
- Gestión de los recursos de radio y traspaso de usuarios entre estaciones base, mecanismo denominado *handover*.
- Cifrado de paquetes y protección de identidad.

UTRAN se divide en múltiples subsistemas llamados RNS (Radio Network Subsystem), los cuales son constituidos por 2 elementos, el NodeB y el RNC. específicamente, el RNS es conformado por solo un controlador y múltiples NodeB, a continuación se describen estos elementos, además de sus funciones respectivas.

- Estación base (NodeB): Es el nodo que maneja la transmisión y recepción de señales en un área física, definida como celda. En consecuencia, es el responsable de todas las funcionalidades de capa física necesarias para transmitir/recibir por la interfaz radio, como las ya mencionadas modulación y codificación, además de gestión de potencia. Como ejemplo particular, el NodeB recibe reportes de los terminales y en función de estos puede disminuir la potencia de sus antenas o notificar al terminal para que haga lo mismo, de esta forma busca evitar la interferencia.
- Controlador (RNC): Es el nodo que administra los recursos de radio del NodeB, como fue mencionado un RNC puede estar controlando múltiples NodeB a la vez, en general realiza las siguientes tareas:
  - Establecer y administrar la conexión de radio.
  - Elegir las propiedades de la transmisión, como por ejemplo tipo de modulación, en base a la capacidad del NodeB y requerimientos de calidad del servicio.
  - Administración de movilidad, dicho de otra forma gestión del handover.
  - Gestionar la sobrecarga de recursos de radio, es decir limitar el establecimiento de nuevas conexiones cuando hay saturación en la red.

2.2.2. Interfaces de UTRAN

- Iub: Interfaz interna de la red de acceso que conecta el NodeB y el RNC.
- IuR: Interfaz que permite la conexión entre distintos RNC cuya finalidad es brindar soporte a la aplicación de handovers entre NodeB de distintos RNS.
- IuPS: Interfaz entre la red de acceso y el dominio PS de la red CORE, conecta el RNC y SGSN.
- IuCS: Interfaz entre la red de acceso y el dominio CS de la red CORE, conecta el RNC y el MSC.

2.2.3. Red CORE

Parte central de la red que se encarga de proveer los servicios al usuario, estos pueden clasificarse en dos grandes grupos, los cuales son el dominio de conmutación de circuitos (CS) y el dominio conmutación de paquetes (PS). Los servicios basados en CS son las llamadas telefónicas o videollamadas, vale decir, servicios cuya naturaleza requiere de una conexión dedicada entre los usuarios. Mientras que en PS están aquellos cuyo transporte de información es realizado con paquetes IP, aquí se encuentran los servicios ofrecidos en internet.

La red CORE se compone de los siguientes elementos, en la figura 2.5 se muestran en detalle las conexiones entre los elementos.:

- Elementos que forman parte del dominio CS:
  - MSC (Mobile Switching Center): Lleva a cabo la conmutación de circuitos y el control de llamadas, además administra la movilidad de usuarios.
  - GMSC (Gateway Mobile Switching Center): Elemento dedicado a direccionar llamadas telefónicas en redes externas, es la puerta de salida del dominio CS.
  - HSS (Home Subscriber Server): Base de datos que administra información del usuario en la red local, como por ejemplo ubicación o detalles de los servicios .
  - VLR (Visitor Location Register): Base de datos que administra información del usuario para servicios de roaming, realiza funciones similares al HSS pero en redes externas.
- Elementos que forman parte del dominio PS:
  - SGSN (Serving GPRS Support Node) [1] : Elemento encargado de interconectar la red de acceso con la red CORE, junto al GGSN administra rutas de tráfico.
  - GGSN (Gateway GPRS Support Node): Análogo al GMSC, pues es dedicado al direccionamiento de tráfico con redes externas (internet) pero en el dominio PS, realiza el enrutamiento de paquetes.

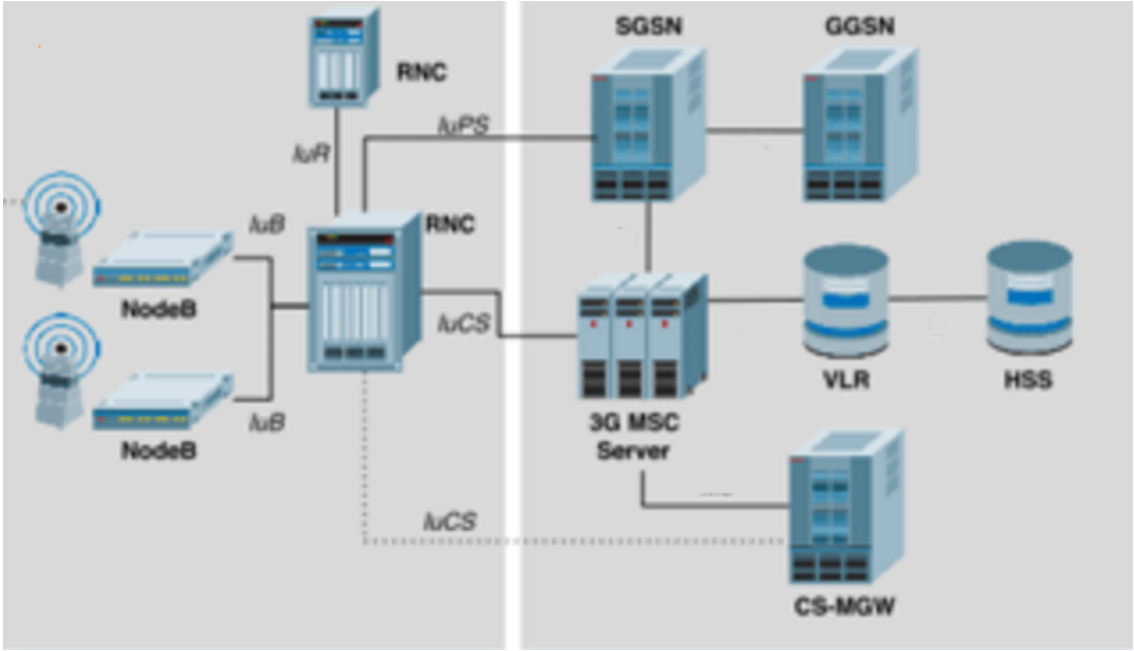


Figura 2.3: Elementos de UTRAN y red CORE.

[1] GPRS es el acrónimo de General Packet Radio Service o en español, servicio general de paquetes vía radio.

2.2.4. Interfaz radio

También se conoce como interfaz de aire, pues es la interfaz inalámbrica que utiliza la UTRAN para conectarse con los terminales. En el caso de 3G, la especificación indica que el acceso al medio se realiza con multiplexación por código o WCDMA, a diferencia de redes anteriores que usaban multiplexación por tiempo o frecuencia.

La información transmitida por esta interfaz se divide en dos grandes grupos, estos son el plano de usuario y de control. En el primero fluye toda la información relacionada a servicios finales dedicados al usuario, mientras que en el segundo se transmite la señalización entre la red y el terminal para establecer la comunicación [2].

Para clasificar los datos transmitidos se define el concepto de canal, el cual corresponde a estructuras de datos pertenecientes a las primeras dos capas (capa física y enlace), existen 3 tipos de canales: lógico, transporte y físico, los cuales funcionan de forma jerárquica, ya que el canal de transporte encapsula los datos del canal lógico, que a su vez es encapsulado por el canal físico, a continuación son descritos cada uno de los canales.

- Canal lógico: Corresponde a los datos donde se define el tipo de información, es decir si corresponde a señalización o datos de usuario.

  Los canales lógicos en UMTS son los siguientes:

  - Broadcast Control Channel (BCCH): Canal DL dedicado al broadcast con el que una estación base notifica su información.
  - Paging Control Channel (PCCH): Canal DL dedicado al control del paging [2].
  - Dedicated Y Common Control Channel (DCCH y CCCH): Canales UL/DL dedicados a datos de control.
  - Dedicated Y Common Traffic Channel (DTCH y CTCH): Canales UL/DL dedicados a datos de usuario.

- Canal de transporte: Datos que definen el cómo debe ser transportada la información dada por el canal lógico.

[2] Paging es el procedimiento realizado por la estación base para conectar un terminal no conectado a la red

Los canales de transporte en UMTS son los siguientes:

- Random Access Channel (RACH) UL
- Forward Access Channel (FACH) DL
- Uplink Common Packet Channel (CPCH) UL
- Downlink Shared Packet Channel (DSCH) DL
- Dedicated Transport Channel (DCH): UL/DL
- Broadcast Channel (BCH) DL
- Paging Channel (PCH) DL

Tanto el canal lógico como el de transporte pertenecen a la capa de enlace, el primero forma parte de la subcapa RLC (Radio Link Control) mientras que el segundo pertenece a la subcapa MAC (Medium Access Control). En la figura 2.4 se muestra la totalidad de los los canales definidos para UMTS, además de la jerarquía que los relaciona.

- Canal físico: Son los datos que finalmente son transmitidos, tanto la información del canal lógico y de transporte son mapeados aquí.

Los canales físicos en UMTS son los siguientes:

- Dedicated Physical Control Channel (DPCCH)
- Primary Common Control Physical Channel (P-CCPCH)
- Physical Random Cccess Channel (PRACH)
- Dedicated Physical Data Channel (DPDCH)
- Dedicated Physical Control Channel (DPCCH)
- Physical Downlink Shared Channel (PDSCH)
- Physical Common Packet Channel (PCPCH)
- Synchronization Channel (SCH)
- Common Pilot Channel (CPICH)
- Adquisition Indicator Channel (AICH)
- Paging Indicator Channel (PICH)
- CPCH Status Indicator Channel (CSICH)
- Colission Detection/Channel Assignment Indicator Channel (CD/-CAICH)

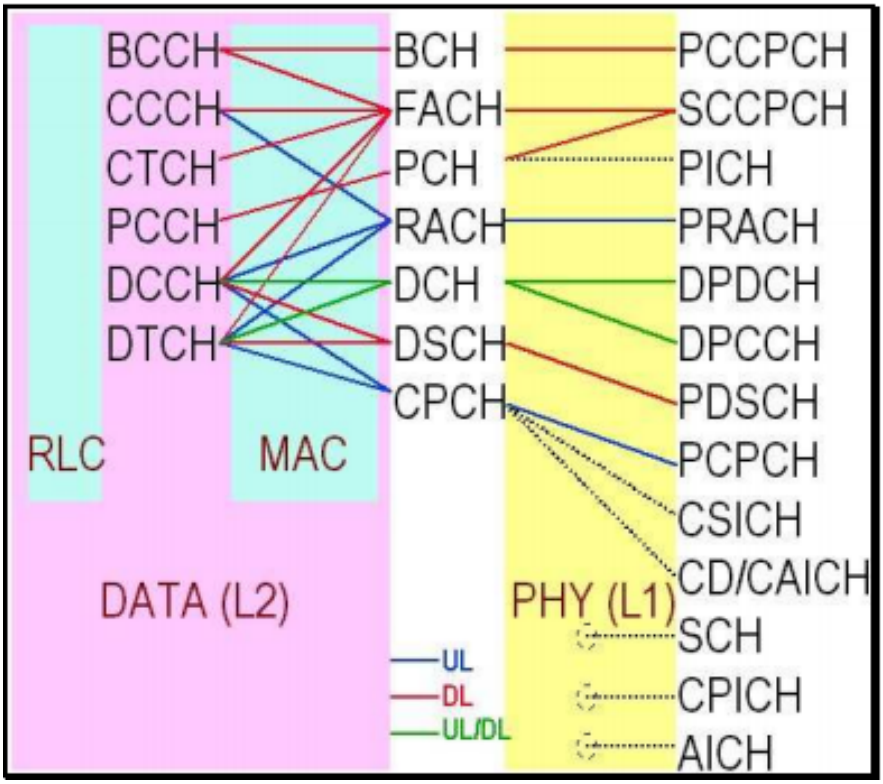


Figura 2.4: Canales UMTS

2.2.5. HSPA y HSPA+

En el año 2002 la 3GPP se encargó de potenciar el rendimiento de la red, esto con el objetivo de traer mejoras en aspectos como tasa de datos, latencia y optimización en los protocolos de transmisión. Como consecuencia, fue introducida la tecnología HSDPA (High Speed Downlink Packet Access) en el release 5, posteriormente se continuó con mejoras hasta 2007, llegando a HSPA+ [3].

HSPA fue una mejora realizada en dos pasos, siendo la primera HSDPA (High Speed Downlink Packet Access) con la finalidad de mejorar la tasa de transmision en el canal de descarga. Luego el segundo paso fue HSUPA (High Speed Uplink Packet Access), donde se buscó la mejora del canal de subida. Finalmente en el release 7 se introduce HSPA+ con aun más mejoras, como la incorporación de técnicas de transmisión con múltiples antenas.

2.2.5.1. HSDPA

El objetivo de HSDPA fue aumentar la tasa de descarga y reducir la latencia, pero en el marco de que los cambios deben ser compatibles con la red, por lo que estos debían ser mínimos en cuanto a cambios de arquitectura de red. Fue así que el foco de esta tecnología se basó en mejoras de la asignación de recursos, minimizando el tiempo de planificación y usando al máximo los recurso de radio.

Para optimizar el tiempo de respuesta en la asignación de usuarios, adaptación a condiciones del canal y reducción de latencia, se agregaron al NodeB funcionalidades como *scheduling* y adaptación de enlace, que antes eran llevadas a cabo por el RNC, así se busca que sean aplicadas más cerca de la interfaz radio.

- Scheduling: Mecanismo donde el NodeB selecciona qué terminal accede a los recursos radio para cada intervalo fijo de tiempo, denominado TTI (Transmission Time Interval), la duración del TTI cambia de 10 ms (con R99) a 2 ms.
- Adaptación de enlace: El NodeB selecciona la codificación y modulación, en el caso de modulación se trabaja con QPSK y 16QAM.

Entre otros cambios, se crea un nuevo canal de transporte compartido llamado High Speed-DL Shared Channel (HS-DSCH), el cual es encapsulado en el nuevo canal físico High Speed Physical DL Shared Channel (HS-PDSCH). Además se mejoró el uso de recursos radio de forma que el usuario con las mejores condiciones de enlace puede acceder a todos los recursos durante un TTI. Es de esta forma que con HSDPA se mejoró la tasa de descarga hasta un máximo teórico de 14.4 Mbps.

2.2.5.2. HSUPA

El objetivo principal de HSUPA se basó en aumentar la transmisión de datos en el canal de subida, problema complejo pues el flujo ascendente de los datos es a partir de múltiples fuentes independientes (los distintos UE que se comunican con el eNodeB), lo que provoca dificultades en el uso óptimo de multiplexación por código.

Para ello se implementó un canal de transporte dedicado al tráfico de subida denominado Enhanced Dedicated Channel (E-DCH), cuyos datos son encapsulados por el nuevo canal físico E-DCH-Dedicated Physical Data Channel (E-DPDCH).

Además se aplicó como apoyo al *scheduling* del NodeB un mecanismo de negociación entre el terminal y la estación base, de tal forma que los recursos sean asignados por el *scheduling* eficientemente. En esta negociación se consideran aspectos como nivel de interferencia o potencia con la que transmite el terminal. Con el uso de todas las mejoras mencionadas se aumentó la tasa de subida hasta 5.7 Mbps en el release 6.

2.2.5.3. HSPA+

Finalmente, con HSPA+ se buscó mejorar aun más los canales de subida y bajada, es así que se implementaron las siguientes funcionalidades:

- Modulación de mayor orden, como 16QAM para UL y 64QAM para DL.
- Multiplexación espacial o uso de múltiples antenas, técnica conocida como MIMO (Multiple Input Multiple Output), en HSPA+ se usó exclusivamente la configuración de 2x2, es decir 2 transmisores y 2 receptores.

Las mejoras logradas con HSPA+ fueron 21 Mbps para DL y 11 Mbps UL, usando MIMO 2x2 el canal de descarga puede aumentar hasta 42 Mbps. De esta forma termina la tercera generación, luego es con el release 8 publicado en 2008 que se estipulan las primeras especificaciones de la próxima generación, lo que pasaría a llamarse LTE (Long Term Evolution).

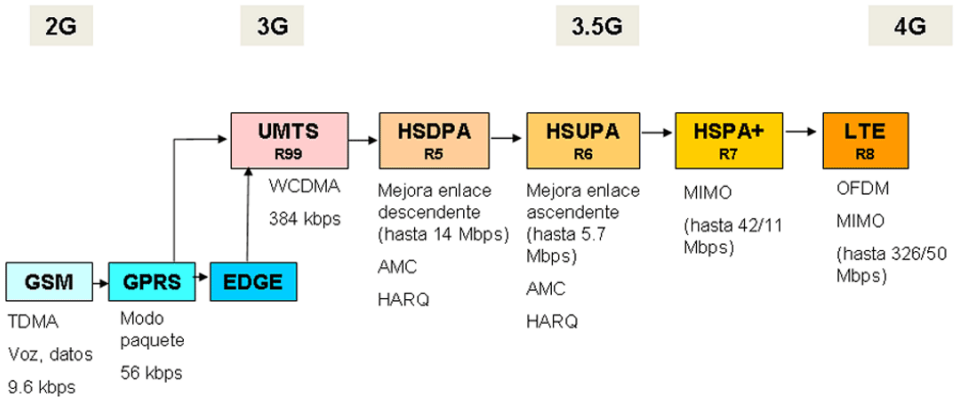


Figura 2.5: Evolución de telefonía móvil hasta 4G.

2.3. Cuarta generación: LTE

Long Term Evolution (LTE) es el nombre que recibe la cuarta generación y consistió en un proyecto llevado a cabo por 3GPP, cuya finalidad fue estudiar tecnologías que permitan cumplir con los nuevos requisitos previstos en la telefonía móvil. Entre los principales objetivos se encuentra la optimización del ancho de banda, reducción de costos y la integración con redes de generaciones anteriores.

Los estudios comenzaron en el año 2004 cuando se creó una comisión con miembros de 3GPP y asociaciones externas. En dicho estudio se definieron los siguientes requisitos para el nuevo estándar:

- Reducir el costo por bit transmitido.
- Aumentar los servicios disponibles junto a mejoras en su calidad.
- Flexibilizar el uso del ancho de banda tanto en las nuevas como actuales bandas de frecuencia.
- Simplificar la arquitectura de red.
- Uso razonable de energía en los UE.

El proyecto LTE consistió en mejoras sobre la red de acceso donde se buscó una evolución del estándar usado para UTRAN, es así como en el nuevo estándar se definió una nueva arquitectura para la red de acceso basada completamente en tráfico paquetizado IP, además se llevo a cabo el desarrollo de una nueva interfaz de radio. Esta nueva red de acceso recibió el nombre de Evolved UTRAN, mientras que la interfaz de radio fue llamada Evolved UTRA [4].

De forma paralela a LTE se estaba desarrollando un proyecto de estandarización sobre la red CORE con el nombre de System Architecture Evolution (SAE), cuya finalidad fue la creación del Evolved Packet Core (EPC) como la nueva red CORE, su principal característica fue migrar exclusivamente a tráfico IP pero brindando interoperabilidad con redes 2G y 3G.

Tanto el proyecto LTE como SAE formaron parte del release 8 llevado a cabo en el año 2008, si bien LTE a nivel de estándar fue solo una parte de la nueva generación hoy en día es aceptado referirse al 4G en su totalidad como sinónimo de LTE. En la figura 2.6 se muestra un esquema general de la nueva red de acceso y CORE.

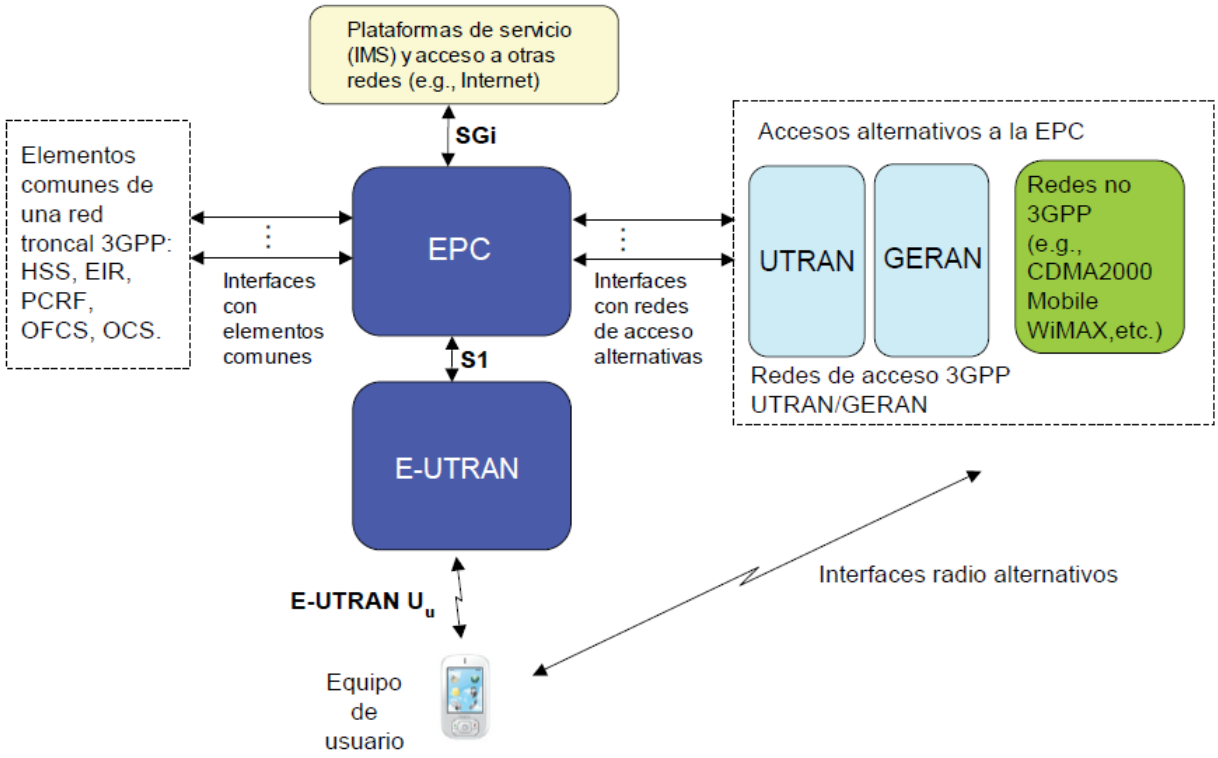


Figura 2.6: Arquitectura general de red LTE.

2.3.1. Red de acceso: E-UTRAN

Entre los principales requisitos planteados en LTE para la red de acceso E-UTRAN y la interfaz de radio están los siguientes [4]:

- Tasas de transmisión de 100 Mbps DL y 50 Mbps UL para bandas de 20 MHz.
- Latencia menor a 10 ms en el plano de usuario.
- Soportar al menos 200 usuarios por celda.
- Brindar conexión para usuarios en movimiento con velocidades de hasta 350 km/h.
- Flexibilidad para operar en distintas bandas de frecuencia y con distinto ancho de banda (1.4, 3, 5, 10, 15 y 20 MHz UL/DL).
- Celdas capaces de brindar cobertura con buena calidad de servicio en un radio de 5 km y de mantener el servicio hasta los 30 km.

Uno de los elementos clave en LTE para lograr los requisitos mencionados es la estación base o eNodeB, sucesor del NodeB en 3G, el cual se encarga de todas las funcionalidades presentes en la red de acceso, a diferencia del NodeB que necesita el apoyo del RNC. Permite la transmisión de datos en paquetes IP entre la red CORE y el UE, asimismo se encarga de operaciones de control; además, mantiene una base de datos donde almacena información sobre usuarios conectados, estado de conexión, enlaces de radio activos y seguridad, elementos que agrupados se denominan *contexto* del UE. En cuanto a arquitectura la red E-UTRAN es conformada solo por los eNodeB, a diferencia de UMTS donde existe la estación base y el controlador.

Sumado a esto, las funcionalidades más importantes que realizan son la gestión de recursos de radio, lo que se puede entender como el problema de asignar dinámicamente los recursos en los canales de subida o bajada, destinar usuarios a otro eNodeB, control de interferencias o balanceo de carga con otras estaciones bases. Así pues, una característica importante es que el eNodeB puede comunicarse con otros eNodeB para llevar a cabo estas funcionalidades, la interfaz entre distintas estaciones base se denomina X2, cuyas características serán descritas más adelante.

Otra de las ventajas y novedades presentadas es la capacidad de poder conectarse a múltiples MME, entidad de conexión con la red CORE, de esta forma se tiene una mayor flexibilidad en el balanceo de carga de señalización pues no hay solo una conexión entre estación base y red CORE, además se minimizan los puntos de falla haciendo que la red sea más robusta a caídas, esta interfaz entre estación base y múltiples MME se conoce como S1.

2.3.2. Interfaces en la red de acceso

Entre las principales interfaces de la red de acceso se encuentran las siguientes:

- Interfaz eNodeB-UE (U_u): Es la interfaz de radio entre el equipo de usuario y el eNodeB, aquí ocurren los siguientes procedimientos:
  - Transferencia de datos entre usuario y estación base, se realiza a partir de los enlaces de radio, denominados *radio bearers* (RB). La transferencia es realizada estrictamente en paquetización IP.
  - Señalización de control a partir del protocolo RRC (Radio Resource Control), con ello se gestiona el establecimiento de los RB y el estado de actividad del usuario, también aplica el cambio de gestión de eNodeB sobre el equipo o handover.
  - Broadcast: señal de control emitida por el eNodeB donde se da conocimiento de los parámetros de la red, como por ejemplo potencia máxima e identidad del operador. Entre otras acciones esta la conexión forzada de un equipo sin conexión a la estación base, lo que se conoce como paging.
- Interfaz eNodeB-eNodeB (X2): Interfaz ubicada entre distintos eNodeB, la transferencia de datos es sin entrega garantizada ni control de errores, cumple un rol importante en el balanceo de carga entre estaciones bases, sus funcionalidades son las siguientes.
  - Soporte al handover: A través de esta interfaz las estaciones bases transfieren el

contexto de un usuario, vale decir, la información necesaria para que el nuevo eNodeB pueda establecer los RB.

- Estado de carga: Notificación del estado de los recursos de radio disponibles, sirve para por gestionar el tráfico en casos de interferencia o sobrecarga.

- Interfaz eNodeB-EPC (S1): Es el nexo entre las redes de acceso y CORE. Es dividido en dos subplanos: el de usuario y el de control, denominados S1-U y S1-MME, respectivamente. Las funcionalidades a las que da soporte son las siguientes [5] :

  - Procedimientos para establecer, mantener y terminar los RB.
  - Procedimientos para establecer handovers.
  - Mensajes de señalización entre UE y UTRAN.
  - Funcionalidades de localización.

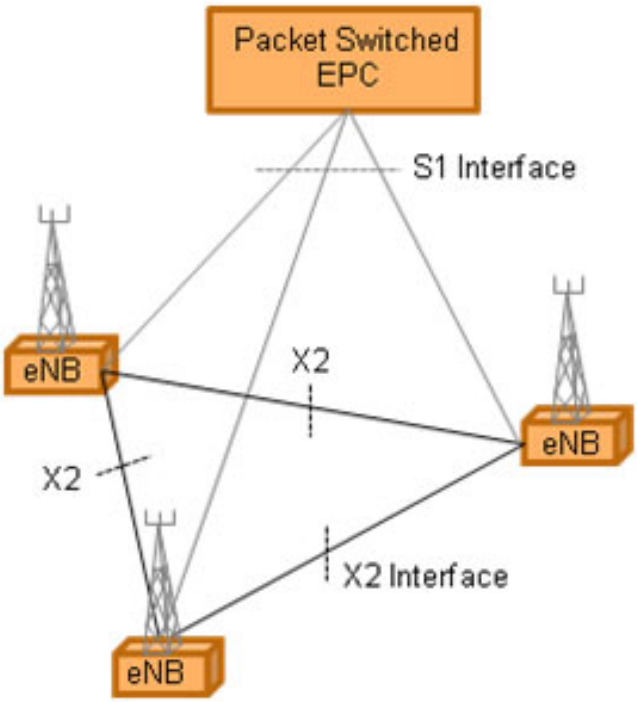


Figura 2.7: Interfaces X2 y S1.

2.3.3. Protocolos E-UTRAN

Los protocolos en capa 2 y 3 de E-UTRAN son similares a sus versiones anteriores de UTRAN, ya que en general las funcionalidades y nombres se mantienen, de todas formas si existen diferencias entre las cuales destacan dos:

- En el caso de E-UTRAN los protocolos son llevados a cabo por exclusivamente por el eNodeB, a diferencia de UTRAN donde estos son repartidos entre el NodeB y el RNC.

- Los protocolos en E-UTRAN son más simples en términos de complejidad y funcionalidades al compararlos con las contrapartes de UTRAN.

2.3.3.1. NAS: Non Access Stratum

Conjunto de protocolos llevados a cabo entre el UE y el MME, por lo que no es estrictamente parte de la red de acceso. Toda señalización llevada a cabo por estos protocolos es transparente a nivel E-UTRAN, es decir que el eNodeB nunca intercepta estos paquetes.

Dos protocolos importantes son el EPS Mobility Management protocol (EMM) y el EPS Sesion Management protocol (ESM).

- EMM: Realiza procedimientos como registro del UE en el MME, autenticación y paging de UE inactivos.
- ESM: Responsable de establecer y administrar radio bearers.

2.3.3.2. RCC: Radio Resource Control

Protocolos responsables de toda la señalización de control entre el UE e eNodeB a nivel capa 3, algunos de los procedimientos que lleva a cabo son:

- Broadcast de la información del sistema: Notificar a los UE con toda la información esencial para entablar una conexión con E-UTRAN, como por ejemplo ancho de banda y lista de celdas vecinas.
- Establecimiento de conexión RCC: Procedimiento que cambia el estado de conexión RCC del usuario de inactivo a activo, es necesario para que el UE pueda recibir/enviar datos a la red.
- Control de handovers: Procedimiento donde se lleva a cabo la señalización de control para realizar handovers entre eNodeB u otras tecnologías de acceso al medio.

2.3.3.3. PDCP: Packet Data Convergence Protocol

Protocolo ubicado entre el UE/E-UTRAN y que es responsable por la compresión de headers en paquetes, entre otros procedimientos como cifrado, integridad y entrega de paquetes desfasados cuando se aplica handovers.

2.3.3.4. RLC: Radio Link Control

Protocolo encargado de controlar el enlace de radio para datos y paquetes de señalización de control. Un UE puede estar conectado a varias entidades RLC simultáneamente donde cada una puede actuar bajo 3 modos: acknowledgement mode (AM), unacknowledgement mode (UM) y transparent mode (TM).

- Acknowledgement mode: Modo donde se garantiza la entrega sin errores de todos los paquetes a la capa superior, esto es realizado con mensajes de confirmación ACK (acknowledgement) y solicitudes de retransmisión NACK.
- Unacknowledgement mode: Modo sin entrega garantizada, pues no se utilizan paquetes ACK ni NACK.
- Transparent mode: Modo que minimiza el uso de headers en los paquetes.

2.3.3.5. MAC: Medium Access Control

La principal funcionalidad de este protocolo es la asignación dinámica o scheduling de de paquetes provenientes de capas superiores, además del manejo de corrección de errores. El algoritmo de asignación elige cómo se van a repartir los recursos de radio entre distintos terminales, algunos de los aspectos tomados en cuenta por el planificador son los siguientes:

- Disponibilidad de los recursos de radio.
- Prioridad según el UE.
- Condiciones de canal para cada UE.
- Cantidad de datos a ser transmitidos.
- Cuanto tiempo lleva un UE sin recibir datos.
- Retransmisiones de paquetes pendientes.

2.3.4. Red CORE: EPC e IMS

El diseño de la red CORE en LTE se diferencia de redes anteriores por brindar exclusivamente conectividad IP, pero con posibilidad de conectarse a redes de otras generaciones, está compuesto mayoritariamente por el EPC y en segundo lugar por el IMS (IP Multimedia Subsystem). El IMS es un elemento secundario en LTE y no es obligación que toda red de cuarta generación lo tenga, a menos que se busque brindar servicios donde es fundamental, como por ejemplo telefonía de voz sobre IP (VoLTE).

En general la red CORE cumple objetivos similares a los vistos en UMTS, como controlar y establecer los servicios que aplica la red de acceso o la conexión con redes exteriores. En cuanto a la arquitectura interna, el EPC se compone de distintos elementos como el MME, S-GW y P-GW, junto a otros ya existentes en generaciones anteriores [6]. La figura 2.8 muestra el esquema general de una red CORE LTE.

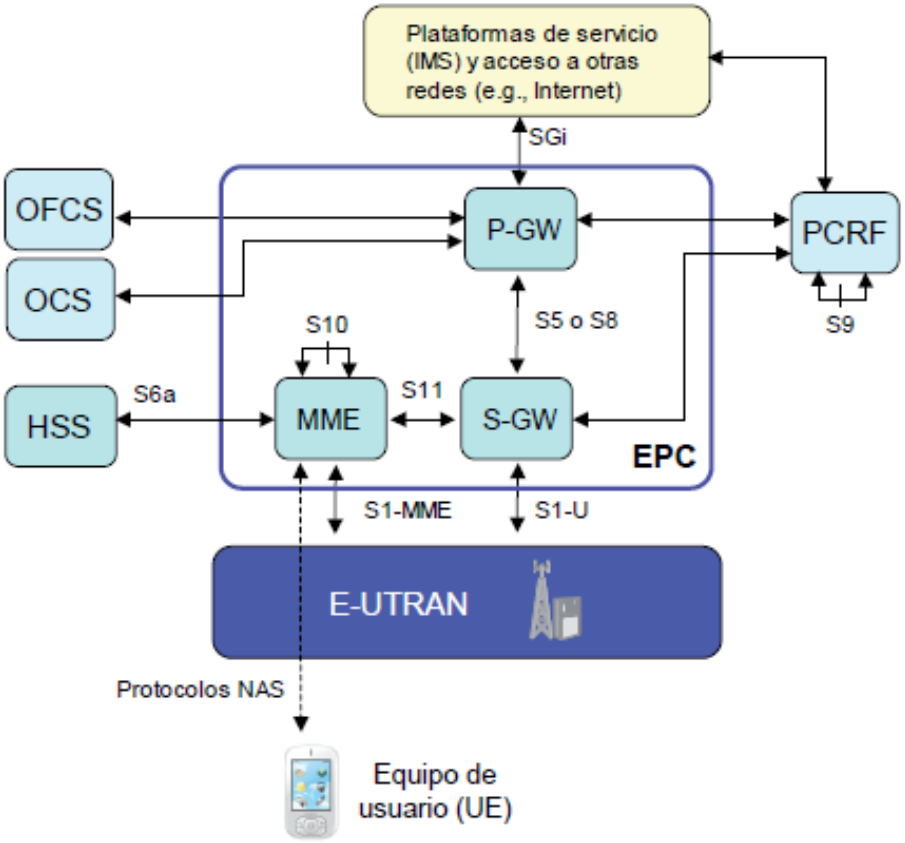


Figura 2.8: Red CORE, compuesta por el IMS y EPC.

A continuación se describen los elementos principales que componen una red EPC:

- MME (Mobility Manage Entity): Elemento principal en plano de control de la red, pues se encarga de administrar la conexión de todos los UE, cada terminal debe tener un identificador o contexto en el MME. Lleva a cabo las siguientes acciones:
  - Gestión de movilidad o localización de usuarios, puede forzar un equipo a conectarse a la red.
  - Autenticación de usuarios.
  - Señalización necesaria para establecer los RB.
- S-GW (Service-Gateway): Elemento dedicado al plano de usuario, es la puerta de enlace donde llegan todos los paquetes IP, cada usuario tiene definido un S-GW y su asignación depende generalmente de la ubicación geográfica. Por otro lado, encamina el tráfico y paquetes de datos por un periodo de tiempo, esto ocurre cuando el usuario se encuentra en estado inactivo.

- P-GW (Packet-Gateway): Elemento dedicado al tráfico con redes externas, además es el punto visible de un UE en internet, razón por lo que la asignación de IP ocurre en este bloque. Por otro lado, cumple funciones de tarificación y reglas de uso de la red, un usuario puede tener varias puertas P-GW asignadas, pero solo un S-GW.

Los elementos mencionados anteriormente forman parte del EPC, pero este no compone en totalidad la red CORE. El resto de bloques aparecen en la figura 2.8 y son comunes con redes 3G, estos son los siguientes:

- HSS (Home Subscriber Server): Base de datos que contiene información del usuario, como estado de subscripción o valores necesarios para establecer conectividad con el terminal. Además almacena la ubicación local, vale decir, el grupo de estaciones base donde se tiene la mayor probabilidad de encontrar al usuario, por lo tanto para establecer un RB el MME consulta al HSS.
- PCRF (Policy and Charging Rules Function): Elemento encargado de los permisos para suplir servicios, es decir, permite el tráfico de datos mientras sea concordante con la subscripción del usuario y los parámetros de calidad de servicio.
- OCS (Online Charging System): Elemento que aplica la tarificación de forma online para los servicios que lo requieran.
- OFCS (Offline Charging System): Elemento que almacena información de tarificación para su posterior cobro.

2.3.5. Interfaces en la red CORE

- Interfaz P-GW/IMS (SGi): Interfaz entre la red y redes IP externas (internet), estas pueden ser redes tanto publicas como privadas, además soportan IPv4 e IPv6.
- Interfaz P-GW/S-GW (S5 y S8): Estas dos interfaces permiten el tráfico de datos entre ambos elementos, por un lado S5 se usa para transferencia entre la misma red, mientras que S8 es cuando S-GW pertenece a una red externa.
- Interfaz MME/S-GW (S11): Interfaz donde ocurren los procedimientos de creación, eliminación o modificación en los RB establecidos en los terminales, por ejemplo, indicación de handover o notificación de tráfico pendiente a ser enviado en el S-GW.
- Interfaz MME/MME (S10): En esta interfaz ocurre el cambio de MME asignado a un terminal, vale decir, el procedimiento donde se envía el contexto del UE al nuevo MME.
- Interfaz HSS/MME (S6a): Interfaz donde transitan datos de localización, autenticación y autorización del terminal.

Si bien gran parte de la red CORE esta compuesta por el EPC, es necesario mencionar las características y funcionalidades llevadas a cabo en el IMS. Este subsistema cumple la tarea de proporcionar los mecanismos de control necesarios para servicios multimedia, como voz o video en tiempo real, pero usando transferencia de paquetes, básicamente reemplaza las funcionalidades del dominio de conmutación de circuitos. Se compone de servidores, bases de datos y gateways dedicados al tráfico IP.

El servicio ofrecido por el IMS actúa sobre las capas de transporte, control y aplicación, por un lado en la capa de transporte la transferencia es llevada a cabo a través del protocolo IP, mientras que en la capa de control utiliza el protocolo SIP, lo que lo diferencia del EPC pues el protocolo SIP esta especializado en establecer y liberar sesiones multimedia.

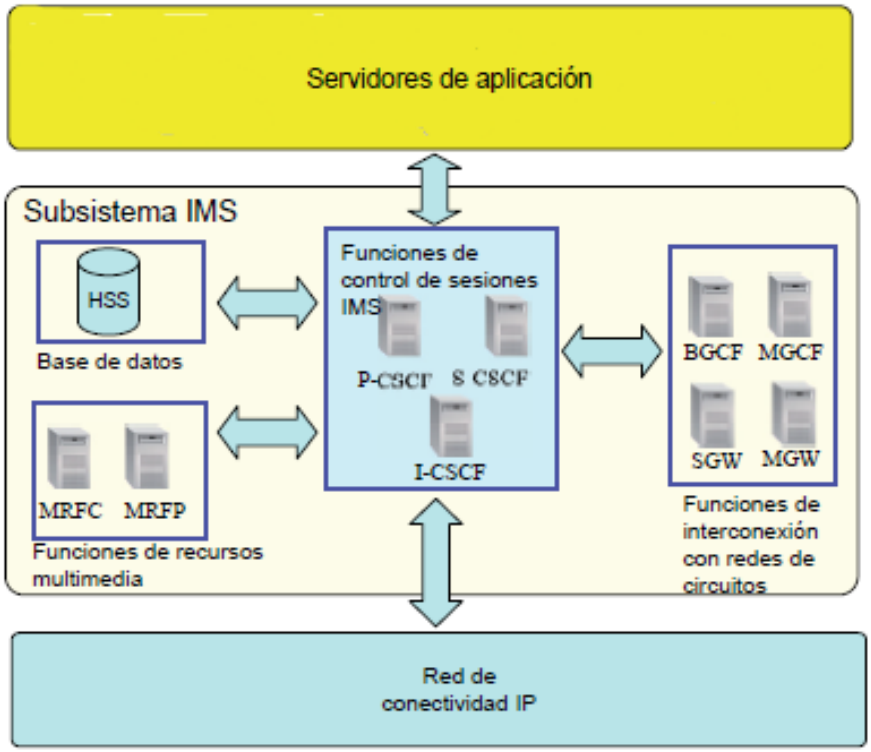


Figura 2.9: Arquitectura del IMS.

La arquitectura del IMS es especificada por la 3GPP [7], donde se pueden identificar 3 elementos principales, estos son : S-CSCF, P-CSCF e I-CSCF. En la figura 2.9 se muestra de forma esquematizada como el IMS está compuesto por estos elementos y con qué bloque del exterior interactúan.

- S-CSCF (Serving CSCF): Servidor de registro SIP que funciona como el nodo central del IMS, cuya función es encaminar señalización SIP a los servidores de aplicación correspondientes.
- P-CSCF (Proxy CSCF): Servidor SIP que actúa como la entrada al IMS desde la red LTE, es por esto que toda la señalización SIP del EPC viaja por este servidor. Entre otras funciones trabaja junto al PCRC para aplicar permisos de conectividad en la capa de transporte.
- I-CSCF (Interrogating CSCF): Servidor SIP que actúa como la entrada al IMS desde redes externas, por lo que la IP de este servidor es el punto visible desde internet y así el I-CSCF se convierte en el primer punto de redireccionamiento de mensajes bajo protocolo SIP.

2.3.6. Canales en LTE

Uno de los objetivos en la estandarización de LTE fue reutilizar soluciones de redes anteriores cuando fuera posible, es así como se hereda la arquitectura de canales de la UTRAN compuesta por los canales lógicos, transporte y físicos. En general se repiten los mismos canales ya mencionados en la sección anterior, se muestra en la figura 2.10 la arquitectura de canales para el caso LTE.

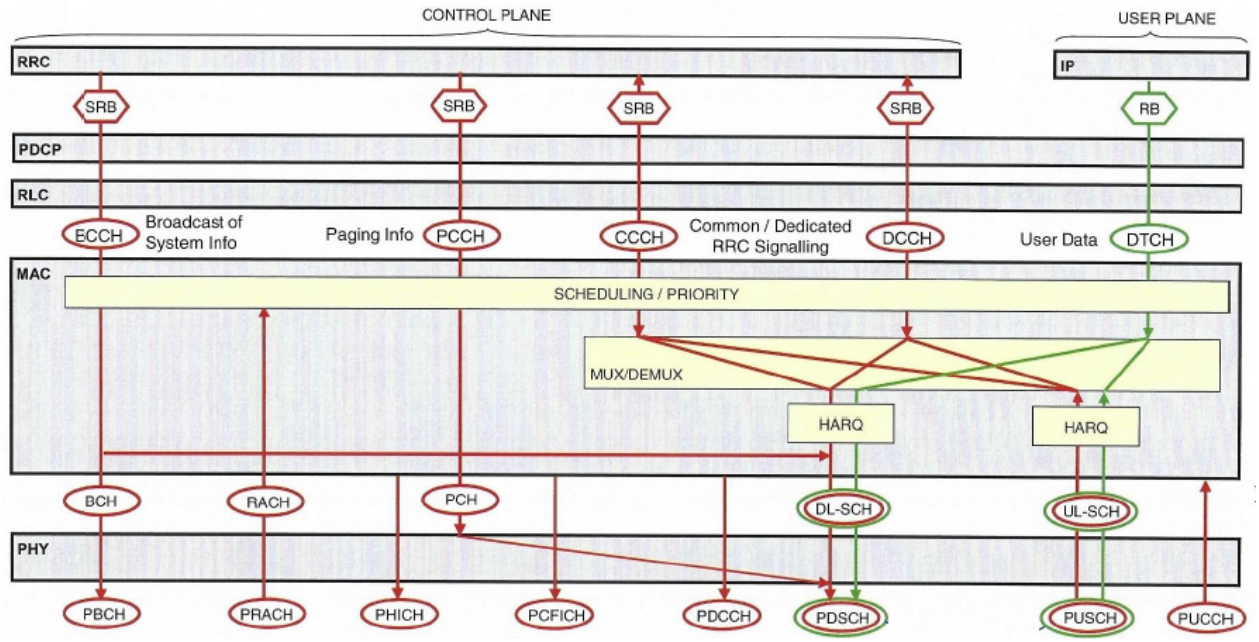


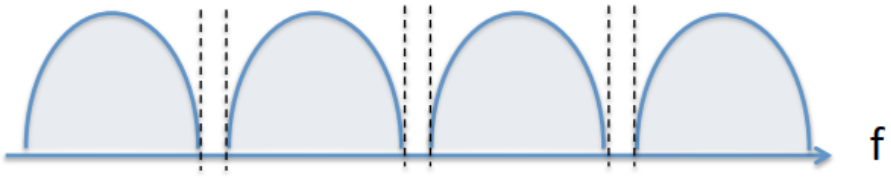
Figura 2.10: Arquitectura de canales lógicos, transporte y físicos en LTE.

2.4. Capa Física en LTE

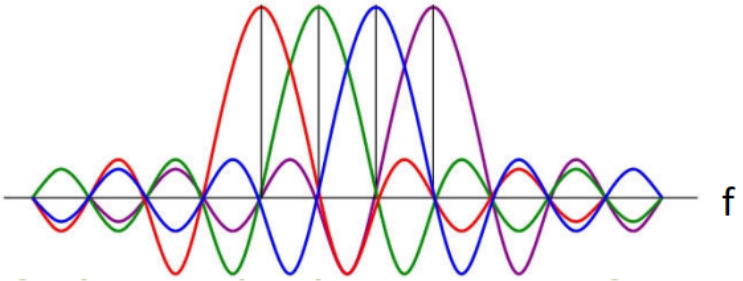
Otra de las principales diferencias de la cuarta generación con sus antecesores se encuentra en las tecnologías de capa física, de modo que permiten alcanzar mejores tasas de transmisión y una gestión eficiente del ancho de banda. En particular la más importante es el uso de multiplexación ortogonal por división en frecuencia (OFDM), que si bien ya existía conceptualmente desde los años 60, recién a partir del año 2000 se pudo dar un uso práctico gracias a la introducción de hardware con mayor poder computacional. De la misma manera se introdujeron esquemas de modulación de mayor orden y el uso MIMO con configuraciones superiores a 2x2.

2.4.1. OFDM: Multiplexación por división de frecuencias ortogonales

La transmisión usando multiplexación ortogonal por división en frecuencia es una técnica que permite ingresar símbolos o agrupaciones de bits en distintas señales subportadoras. Tiene como principal ventaja que estas señales cumplen la propiedad de ser ortogonales entre si, por lo que se pueden transmitir simultáneamente todos los símbolos de cada subportadora sin necesidad de aplicar separaciones entre estas, en otras palabras OFDM permite agrupar en un ancho de banda reducido varias portadoras sin que ocurra interferencia. En la figura 2.11 se muestra un ejemplo de 4 señales ortogonales, es clara la ventaja sobre técnicas como FDM pues no hay necesidad de asignar ancho de banda en separaciones.



(a) Multiplexación en frecuencia.



(b) Multiplexación ortogonal en frecuencia.

Figura 2.11: Comparación entre FDM y OFDM.

2.4.2. Técnicas de Acceso Múltiple

Se entiende como técnicas de acceso múltiple a los métodos o tecnologías que aplica la estación base para llegar a los varios UE, en el caso de LTE se especificó a partir del release 8 el uso de OFDM como tecnología de acceso tanto para el canal de bajada como el de subida. Para el primero se denominó OFDMA o acceso múltiple por división ortogonal en frecuencia, mientras que para el segundo SC-FDMA o acceso múltiple por división de frecuencia de portadora única.

2.4.2.1. OFDMA

Aplicación directa de OFDM ya que es una técnica de acceso múltiple al medio que consiste en asignar bloques temporales de subportadoras a cada usuario, estos bloques se denominan *resource blocks* (PRB) y permiten una mayor flexibilidad para manejar la calidad de servicio. El centro de las portadoras está distanciado en 15 kHz, además estas señales se caracterizan por un factor de cresta (PAPR) bastante alto, a causa de ello la transmisión requiere de amplificadores de alto consumo energético, razón que explica porqué se usa solo para el canal de bajada, pues es la operadora la que puede acceder a equipos de alto consumo energético [8].

Como fue mencionado, facilita la gestión de recursos o *scheduling* debido a que la asignación se convierte en solamente cambiar los símbolos ingresados en los PRB y tiene como ventaja el corto periodo de tiempo con el que se realiza (orden de milisegundos). Para el *scheduling* se considera tanto la calidad del canal como las necesidades del usuario o calidad de servicio, en la figura 2.12 se muestra como se lleva a cabo la asignación de recursos.

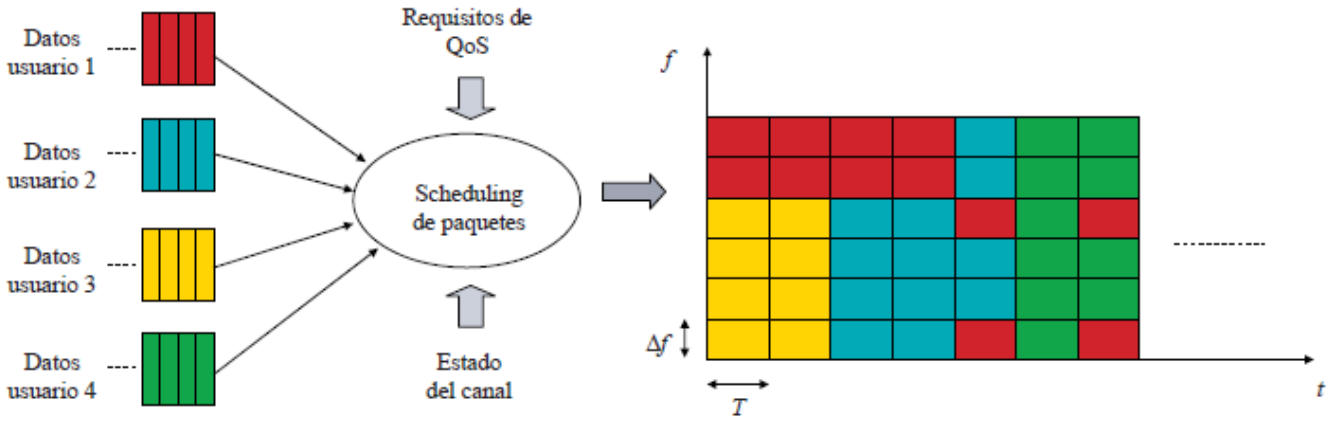


Figura 2.12: Asignación de recursos.

2.4.2.2. SC-FDMA

La aplicación de OFDMA en los terminales del usuario es inviable por alto consumo energético producto de un alto PAPR, es por esto que la 3GPP especificó el uso de SC-FDMA como alternativa para el canal de subida.

Para reducir el consumo energético SC-FDMA trabaja de la siguiente forma: se define el uso de una única portadora como canal de subida, además se aplica una precodificación en los símbolos, la cual consiste en aplicar la transformada discreta de Fourier (DFT) para después volver al dominio del tiempo usando la transformada inversa (FFT), operación que da como resultado un menor PAPR en la señal final. En la figura 2.13 se muestra la comparación entre ambas técnicas, como se puede ver con SC-FDMA cada usuario usa una única portadora, mientras que OFDMA permite una asignación dinámica entre varias [8].

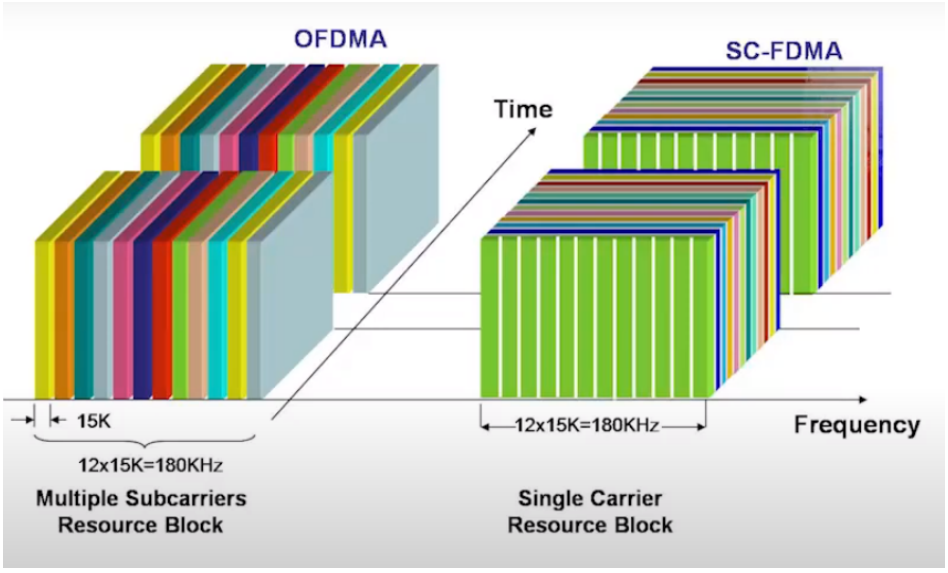


Figura 2.13: Comparación entre OFDMA y SC-FDMA.

2.4.3. Resource Blocks y Estructura de Tramas

El *resource block* o PRB es la unidad mínima de información que un eNodeB puede asignar a un terminal, estos consisten en bloques compuestos por una grilla de dos dimensiones definida por el número de portadoras que contiene y un slot de tiempo. La 3GPP especificó que para LTE un *resource block* se compone de 12 portadoras con un ancho de 15 kHz cada una y una duración de 0.5 ms, como se puede ver en la figura 2.14. En cuanto a capacidad de información cada subportadora puede almacenar 6 o 7 símbolos en un periodo, por lo que un PRB contiene 84 símbolos.

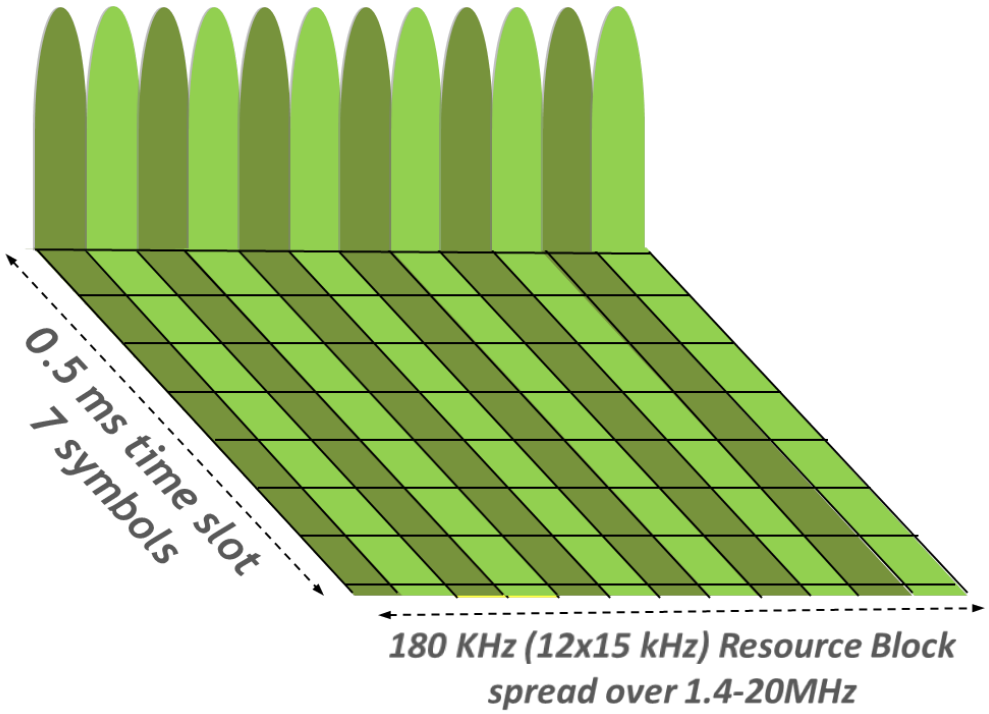


Figura 2.14: Esquema de un resource block.

Como fue mencionado para las técnicas de acceso múltiple, el PRB cumple un rol fundamental en el *scheduling* pues el esquema de bloques permite una gestión dinámica de los recursos de radio. La cantidad de PRB que una estación base puede emitir va a depender del ancho de banda, el mínimo es 1.4 MHz admitiendo 6 PRB simultáneos, mientras que el máximo 20 MHz con 100 PRB simultáneos.

En LTE existen dos formas de estructurar los bloques de recursos en la dimensión temporal y se denominan tramas tipo 1 y 2, la primera esta diseñada para duplexación en frecuencia (FDD), mientras que la segunda para duplexación en tiempo (TDD).

- Estructura de trama tipo 1: Trama con duración temporal de 10 ms, se divide en 20 elementos de 0.5 ms llamadas ranuras temporales o slots. Se define que dos ranuras temporales adyacentes conformaran una subtrama con 1 ms de duración.

  Cada ranura temporal puede albergar 6 o 7 símbolos, en cuanto a gestión de recursos la asignación de los usuarios es determinada dinámicamente por el *scheduler* y es en forma de subtramas, dentro de cada subtrama se asignan dos PRB.

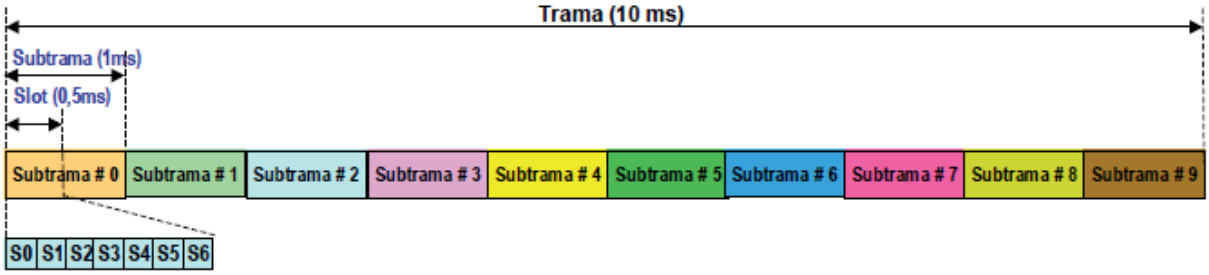


Figura 2.15: Estructura de trama 1.

- Estructura de trama tipo 2: Mantiene la misma estructura que la trama tipo 1, es decir, se compone de tramas con 10 ms de duración donde cada una es integrada por 10 subtramas de 1 ms, la diferencia está en que los canales de subida y bajada son duplexados en tiempo (TDD).

  Como se puede ver en la figura 2.16 la trama tipo 2 tiene campos predefinidos como el DwPTS, que corresponde a la transmisión de bajada y se compone de un símbolo de sincronización seguido de señalización o datos, también esta el UpPTS cuya estructura es la misma pero para el enlace de subida y finalmente se tiene el GP que es un periodo de guarda cuya función es evitar interferencias.

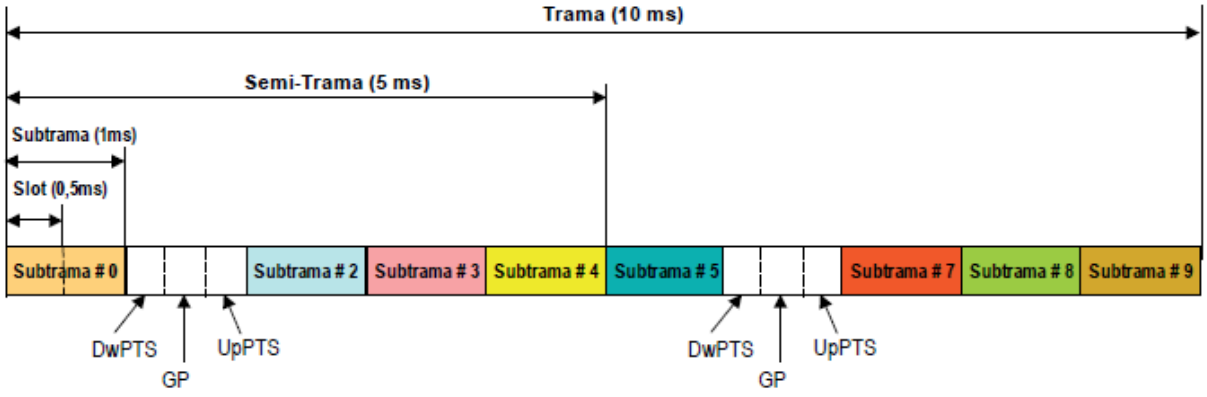


Figura 2.16: Estructura de trama 2.

2.4.4. MIMO: Multiple Input Multiple Output

Se conoce como sistema MIMO a la configuración entre emisor y receptor que utiliza múltiples antenas, la finalidad de usar esta técnica es aumentar las tasas de transmisión sin aumentar el ancho de banda. Configuraciones simétricas de antenas, es decir, misma cantidad de antenas tanto en el receptor como el emisor aumentan la tasa de transmisión de forma lineal, por ejemplo una configuración 2x2 como la que se muestra en la figura 2.17 dobla la tasa de transmisión.

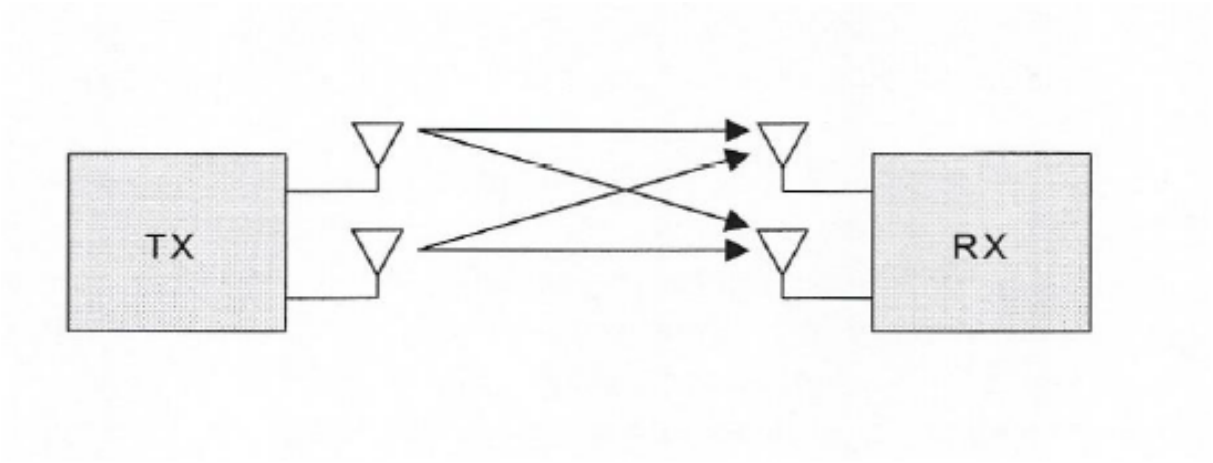


Figura 2.17: Configuración MIMO 2x2.

La forma en la que funciona MIMO es a partir de multiplexación espacial, es decir apuntar las antenas de tal forma que la propagación tome distintos caminos (fenómeno conocido como multipath) y el receptor reciba ambas señales sin interferencia destructiva. Esto es ventajoso pues permite el uso de la misma banda de frecuencia, ademas existe una variación con codificación de canal [4] Space-Time Coding MIMO o STC-MIMO cuya ventaja es reducir los efectos negativos del multipath.

2.5. Quinta generación

A partir del año 2017 la 3GPP comenzó con las especificaciones de la nueva generación 5G en el release 14 y actualmente ya se encuentra completamente estandarizada para el lanzamiento comercial (release 16, año 2020). Si bien el foco de la memoria no es la quinta generación, es necesario mencionar sus principales aspectos, pues en el futuro las redes migrarán a esta nueva tecnología.

Entre los principales objetivos del 5G se encuentran los siguientes [9]:

- Aumento de al menos 20 veces el throughtput experimentado con celdas LTE.
- Aumentar de al menos 10 veces la tasa de descarga experimentada por usuario.
- Muy baja latencia y alta movilidad.
- 1 Millón de terminales por km^2.

En general el 5G está diseñado para soportar 3 tipos de servicios definidos por la unión internacional de telecomunicaciones (ITU), los cuales son: eMBB, uRLLC y mMTC.

- eMBB (enhanced Mobile Broadband) : Consiste en aplicaciones similares a los servicios ofrecidos en 4G, pero mejorando la tasa de datos (orden de Gbps), por ejemplo descargar una película desde el celular en un minutos.
- uRLL (Ultra Reliable Low Latency Control): Consiste en aplicaciones que por naturaleza requieran una alta tasa de transmisión y que sea constante (muy baja latencia), por ejemplo los autos autónomos.
- mMTC (Massive Machine Type Communications): Consiste en aplicaciones relacionadas con el internet de las cosas, no requiere altas tasas de transmisión, pero si la capacidad de manejar muchos terminales a la vez, por ejemplo el control automático de los sensores en una fabrica.

2.6. Indicadores de rendimiento : KPI

Los KPI son como su nombre lo indica indicadores de rendimiento (Key Performance Indicator) y cumplen la función cuantificar la calidad de servicio percibida por el usuario final, además de caracterizar el comportamiento de la red. Algunas de las tareas donde se usan los KPI son:

- Supervisión y optimización de los recursos según la calidad de servicio.
- Notificar oportunamente cuando hay bajo rendimiento en la red.
- Proveer dimensionamiento de la red.
- Detectar las fuentes o causas del bajo rendimiento en la red.

2.6.1. Calidad de servicio: QoS

La ITU definió un modelo general en redes de acceso para cuantificar la calidad de servicio (QoS) percibida por el usuario final, las categorías especificadas son mencionadas a continuación [10].

- Accesibilidad: Capacidad de obtener un servicio dentro de las tolerancias especificadas cuando lo solicite un usuario, algunos ejemplos de KPI presentes en esta clasificación son la tasa de llamadas exitosas o el tiempo de configuración de sesión.
- Retenibilidad: Capacidad de un servicio para continuar prestándose según lo solicitado, un ejemplo de KPI presente en esta clasificación es la velocidad con la que un terminal cierra la sesión de un servicio.
- Disponibilidad: Capacidad de una celda para prestar algún servicio.
- Integridad: Capacidad de que una vez obtenido el servicio la calidad experimentada por el usuario final sea buena, algunos ejemplos de KPI presentes en esta clasificación son la latencia y el packet loss.
- Movilidad: Capacidad de permitir handovers en la red, un ejemplo de KPI presente en esta clasificación es la tasa de handovers exitosos.
- Uso: Descripción del uso de la red en términos de tráfico, gestión de recursos y congestión de la red, algunos ejemplos de KPI presentes en esta clasificación son el throughtput por celda y usuario.

En general los KPI representan información a nivel macro de la red y cuantifican la percepción del servicio desde el usuario final, por otro lado el aspecto micro es cuantificado por los PI (Performance Indicator) y los contadores PM. Un KPI se construye como una función de los parámetros, los cuales pueden estar conformados por aun más contadores, cuya finalidad es definir características a nivel micro de la red, en la figura 2.18 se muestra de forma esquematizada cómo se construyen los KPI.

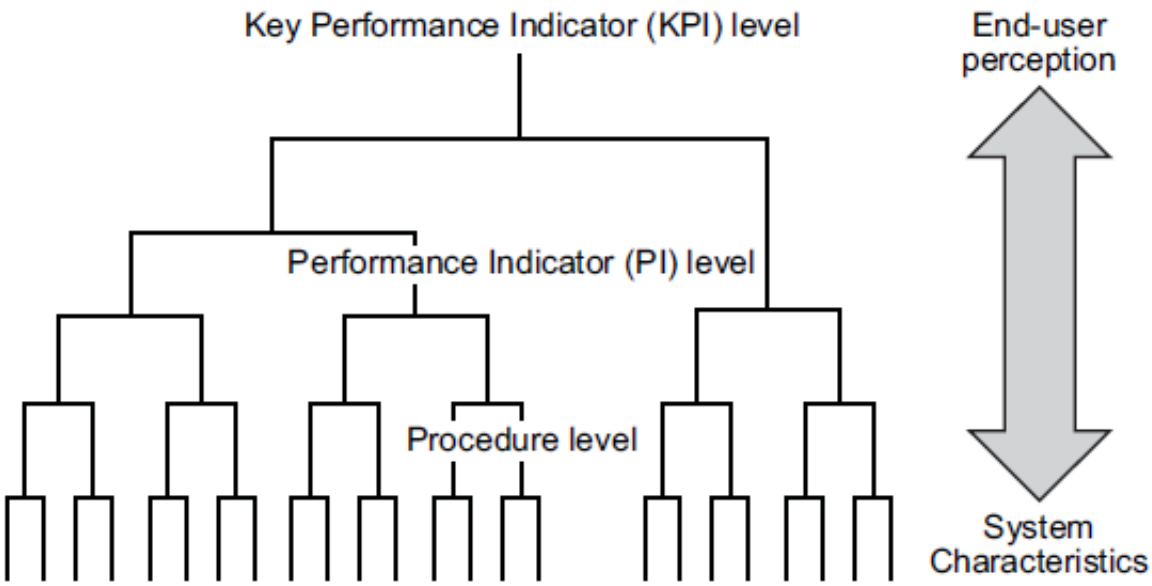


Figura 2.18: Construcción de los KPI

2.7. Machine Learning

Machine learning o aprendizaje de máquinas es una de las ramas de la inteligencia artificial, cuyo concepto consiste en la capacidad de máquinas o computadores para resolver problemas de forma inteligente, pues esta basado en aprendizaje de patrones en grandes volúmenes de datos.

Si bien aprendizaje de máquinas es solo un subconjunto de la inteligencia artificial ambos conceptos tienden a ser confundidos en la práctica, un buen ejemplo para marcar la diferencia entre ambos conceptos es el detector de spam en los correos electrónicos, donde una posible solución es crear un programa que etiquete como spam a correos con palabras claves como promoción o venta, mientras que otra posible solución es entregarle a un programa distintos ejemplos de correos, tanto spam como normales, y en función de ello aprender patrones que diferencien ambos tipos. La primera solución cae en la categoría de inteligencia artificial, y no aprendizaje de máquinas, mientras que la segunda es un claro ejemplo de ello.

2.7.1. Esquema de modelos y datos

La implementación de un modelo de aprendizaje de máquinas no se resume en la aplicación de algún algoritmo, más bien es una secuencia de acciones sobre los datos donde el algoritmo es solamente un bloque. Un modelo pasa por distintas etapas, partiendo por la elección de la base de datos, que pueden ser las mediciones de un sensor en una aplicación industrial, hasta la evaluación de las métricas de rendimiento del modelo, como por ejemplo la tasa de éxito en la detección de fallas de una cadena industrial.

A continuación se mencionan cada uno de los bloques que deben ser considerados en la construcción de un modelo [11]:

- Base de datos: Fuente de los datos crudos a usar en el modelo, deben estar relacionados con el problema a solucionar. Ejemplo: Clasificación de salmones y truchas en una cadena pesquera, la base de datos es un sensor que entrega el largo y ancho de cada pez.
- Limpieza de datos: Todo sensor esta sujeto a un margen de error o a fallas, por ello es necesario aplicar una revisión de datos nulos o fuera del rango esperado. Ejemplo: Eliminar muestras donde el largo o ancho son negativos o excesivamente altos.
- Ingeniería de datos: Manipulación de muestras para ingresar correctamente al algoritmo, ya sea por eficiencia o requisito obligatorio según la naturaleza del algoritmo . Ejemplo: Normalizar datos o transformarlos en una etiqueta binaria de 1 o 0.
- Partición de datos: Los algoritmos aprenden gracias al entrenamiento con datos, esto consiste en la entrega de un gran volumen de muestras al programa para que pueda encontrar y aprender patrones. Para probar el correcto funcionamiento es necesario reservar una parte del conjunto de datos para probar el rendimiento del modelo. Ejemplo: Partir el conjunto de datos en un 80 % entrenamiento y 20 % prueba.

- Algoritmo de ML: Técnica de aprendizaje de máquinas con la que se busca resolver el problema, es el bloque que luego de ser entrenado entrega como salida la respuesta o solución. Ejemplo: Programa que muestra si según el largo y ancho ingresados la muestra corresponde a un salmón o trucha.
- Métricas de rendimiento y evaluación: El algoritmo es sometido a pruebas con la partición de datos reservada, es necesario elegir métricas de rendimiento adecuadas para cuantificar la calidad del modelo. Ejemplo: Tasa de éxito en la detección de salmones o truchas, también conocido como accuracy.

En la figura 2.19 se muestran los pasos a considerar en la construcción de un modelo.

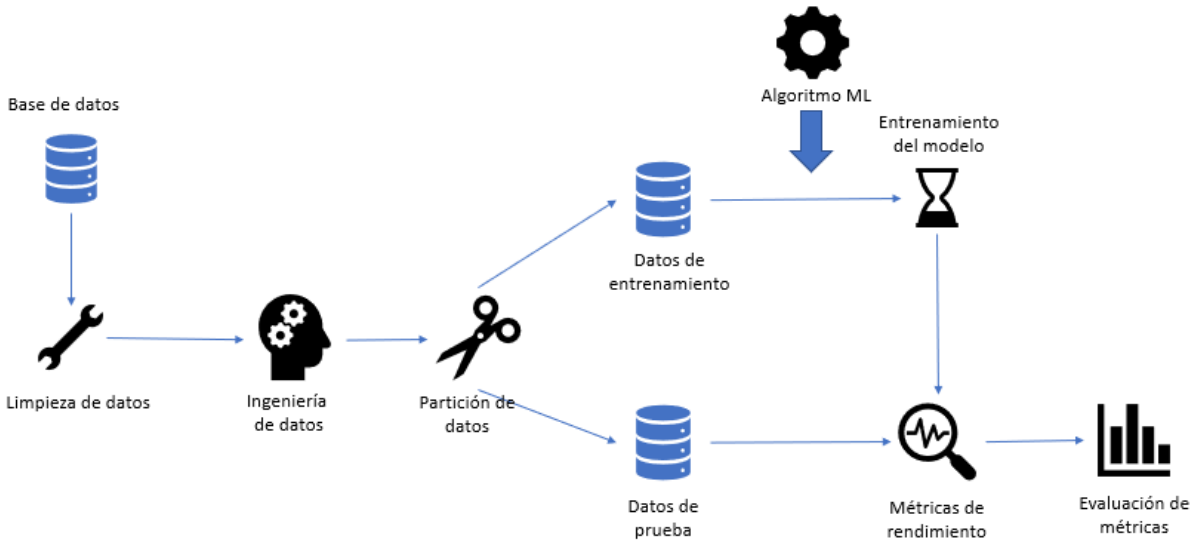


Figura 2.19: Pasos en la construcción de un modelo.

2.7.2. Tipos de aprendizaje de máquinas

El aprendizaje de máquinas se divide en tres grandes grupos, estos son el aprendizaje supervisado, no supervisado y reforzado [11], la principal diferencia es la salida esperada de los algoritmos según el grupo al que pertenezca:

2.7.2.1. Aprendizaje supervisado

Grupo de algoritmos que requieren conocer tanto la entrada como salida de los datos, es decir que necesitan saber la respuesta esperada o etiqueta para aprender y así generalizar la salida para las futuras entradas que ingresen al modelo.

Existen dos grupos de algoritmos supervisados, el primer grupo son las técnicas de clasificación, cuya principal característica es que la respuesta es una variable categórica, un ejemplo de clasificación es la detección de spam en correos electrónicos, pues la salida esperada son solamente dos opciones: clasificar un correo como normal o spam, se muestra el ejemplo en la figura 2.20.

Figura 2.20: Ejemplo de clasificación

El segundo grupo son los algoritmos de regresión, se diferencia de los clasificadores porque en este caso la etiqueta es una valor y por lo tanto que puede en un rango continuo, un ejemplo de regresión es predecir el valor de un automóvil dado su kilometraje, en la figura 2.21 se ilustra este ejemplo donde se espera predecir el valor de un auto dado un kilometraje X.

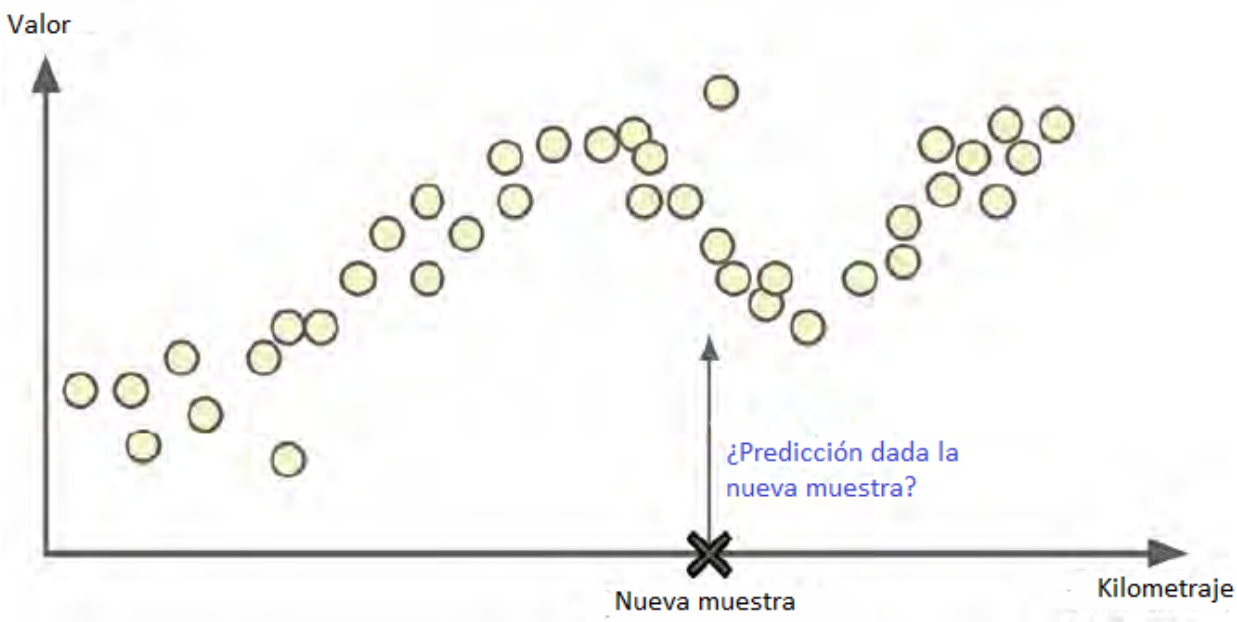


Figura 2.21: Ejemplo de regresión.

2.7.2.2. Aprendizaje no supervisado

Grupo de algoritmos donde los datos no están etiquetados, en consecuencia estas técnicas en general buscan aprender patrones o grupos en los datos, por ejemplo, según las compras en un supermercado es posible segmentar en varios grupos a los clientes y con ello concluir que hay un grupo significativo de personas que lleva vino y carne en la misma compra, usando esta información el supermercado puede definir una estrategia de organización posicionando más cerca las secciones de vinos y carne.

Los tipos de algoritmos no supervisados son los siguientes:

- Clustering: Algoritmos donde se buscan grupos de datos entre las distintas variables, como por ejemplo recomendación de películas en sitios de streaming según el perfil de usuario.
- Visualización y reducción de dimensionalidad: Algoritmos cuya funcionalidad es encontrar relaciones o importancias entre distintas variables y reducir el número de estas, son comúnmente usados en la etapa de ingeniería de datos para simplificar los datos sin perder mucha información.
- Reglas de asociación: Algoritmos que buscan dependencias entre variables, la tendencia estadística de gente que compra vino y carne en el supermercado es un uso de reglas de asociación.

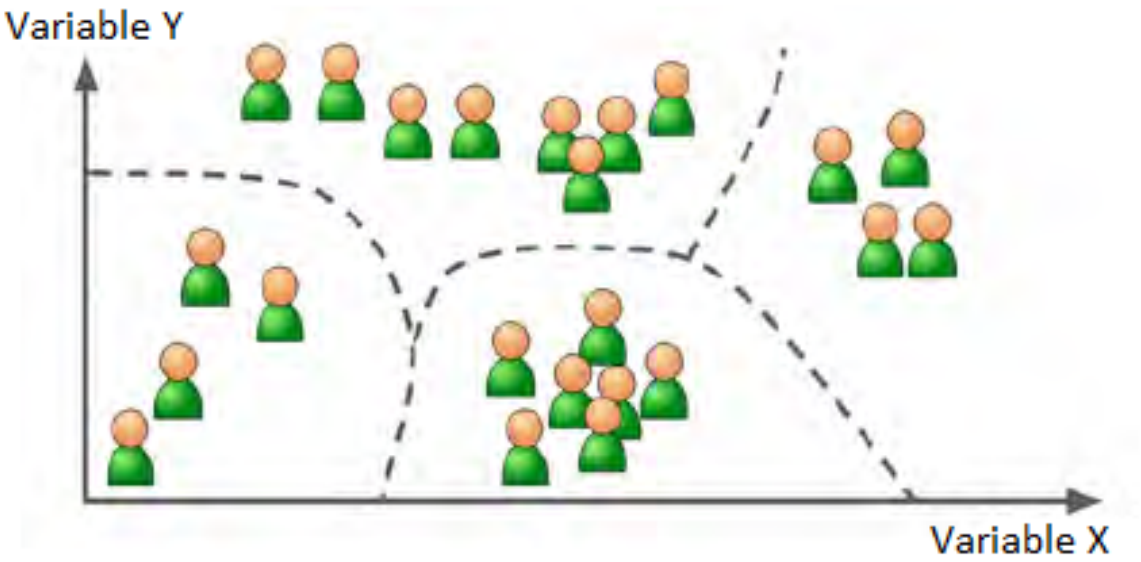


Figura 2.22: Ejemplo de clustering entre dos variables.

2.7.2.3. Aprendizaje reforzado

Tipo de aprendizaje donde un agente aprende utilizando el concepto de premio/castigo, esto lo logra observando el medio ambiente para luego realizar una acción con un objetivo en particular. Se diferencia del aprendizaje supervisado porque no existen etiquetas o respuestas bien definidas del objetivo, más bien se espera que el agente realice acciones y es premiado o castigado según se acerque a su meta u objetivo.

Un uso común es en la robótica, donde se aplican estrategias de aprendizaje reforzado para que los robots aprendan a caminar, de por si no existe una etiqueta que le muestre al robot si camina bien o mal como en el aprendizaje supervisado, más bien tiene que analizar su alrededor y dar pasos, si logra mantenerse en pie es premiado y se incentiva a que mantenga esa forma de caminar, caso contrario es penalizado e intenta de otra manera, en la figura 2.23 se muestra un ejemplo similar donde un robot aprende a apagar incendios.

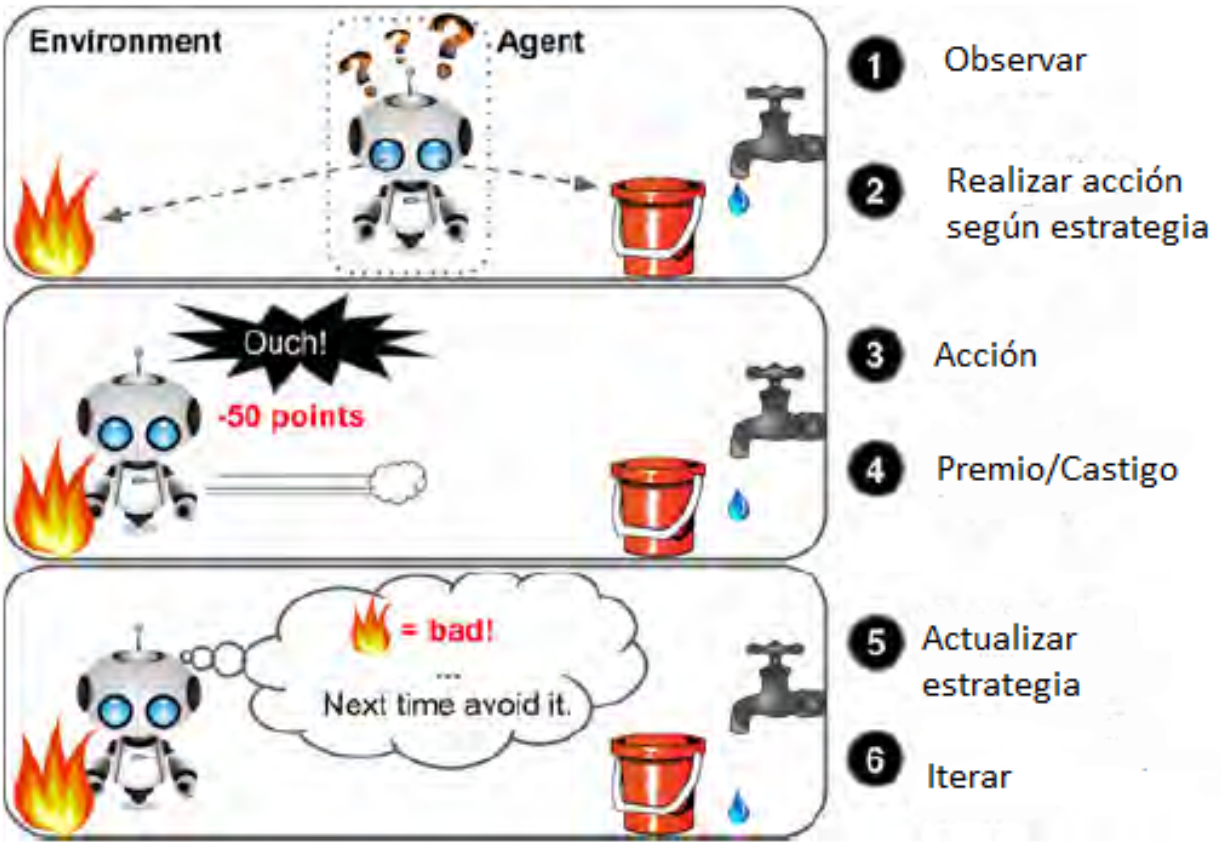


Figura 2.23: Aprendizaje reforzado para que un robot apague incendios.

2.7.3. Técnicas de clasificación

A continuación se describirán los principales algoritmos de clasificación usados [11].

2.7.3.1. KNN: K nearest neighbour

Este algoritmo clasifica cada muestra según los vecinos más cercanos pues utiliza la lógica de que muestras cercanas pertenecen a la misma clase, la optimización de este algoritmo se realiza variando el numero de vecinos K que son necesarios para decidir la clase de la nueva muestra, la decisión final depende de la clase con mayor cantidad de vecinos cerca, esto es ejemplificado en la figura 2.24 donde si K=3 la muestra es clasificada como clase B, mientras que con K=7 es catalogada como clase A.

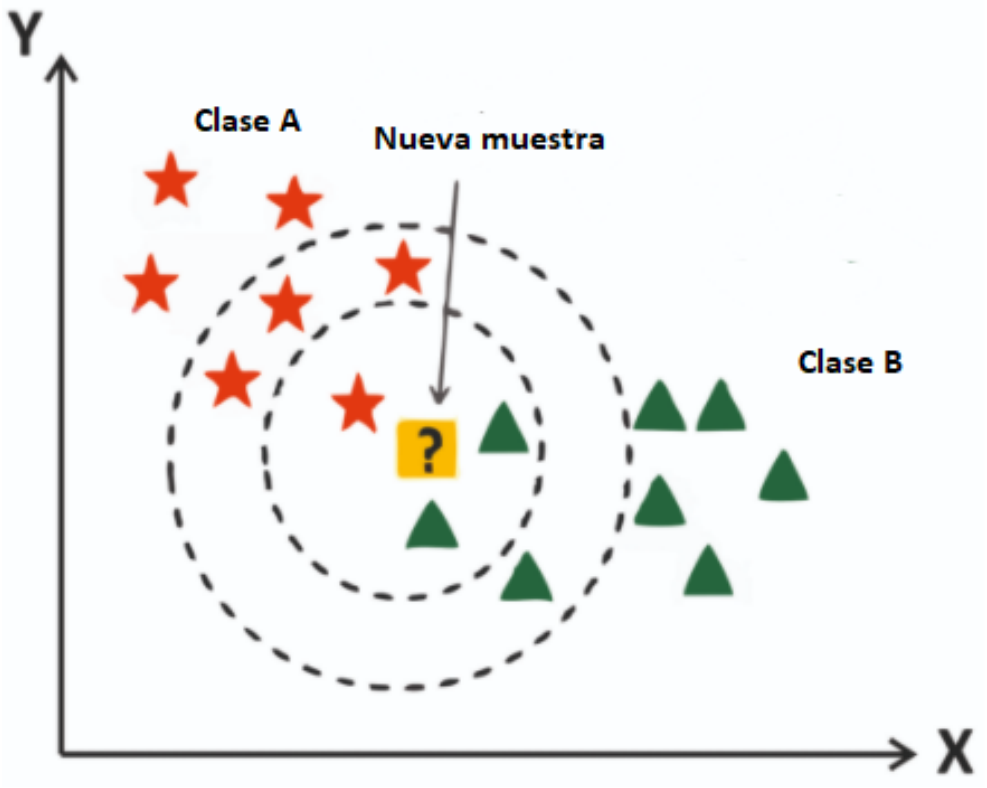


Figura 2.24: Ejemplo de KNN con dos clases.

2.7.3.2. SVM: Support Vector Machine

Los Support Vector Machine o SVM son una familia de clasificadores cuyo funcionamiento consiste en separar las clases utilizando hiperplanos, para efectos de esta memoria se ilustrará la formulación de los SVM lineales y entre solo dos clases. Un hiperplano es una linea recta en el caso de dos dimensiones, como se muestra en la figura 2.25, la idea del algoritmo es buscar el hiperplano óptimo para clasificar ambas clases.

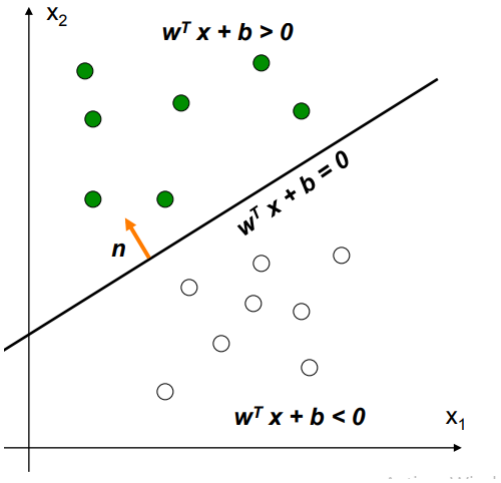


Figura 2.25: Ejemplo de hiperplano en dos dimensiones.

La construcción de un hiperplano es formulada por la ecuación (2.1) donde el vector w y b son los parámetros, la construcción consiste en obtener estos vectores a partir del conjunto de entrenamiento. Cuando la ecuación es igual a cero se obtiene la linea recta, mientras que los casos mayores y menores denotan las dos clases respectivamente como es ilustrado en la figura 2.25.

$$g(x) = w^T x + b \tag{2.1}$$

Ahora bien, la formulación dada anteriormente no entrega un solo hiperplano, si no que infinitos, para solucionar el problema se crea el concepto de margen, que consiste en una zona adyacente al hiperplano donde no existen muestras. El objetivo de esta formulación matemática es buscar el hiperplano cuyo margen sea el máximo, transformando así la búsqueda del margen óptimo en un problema de optimización, como el ejemplo de la figura 2.26.

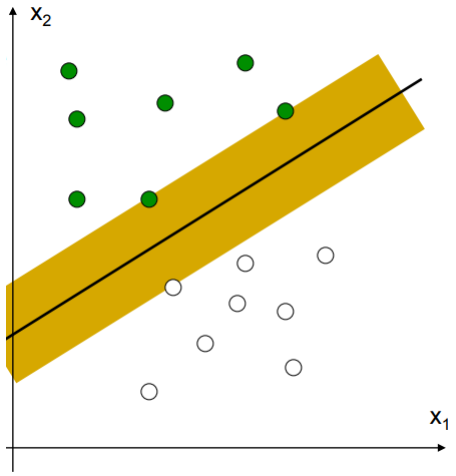


Figura 2.26: Bosquejo de un margen.

La formulación matemática del margen esta en la ecuación (2.2) donde se define como la distancia entre los vectores de soportes, que en este caso son $x+$ y $x-$, multiplicada por el vector normal al hiperplano, esto es mostrado de forma gráfica en la figura 2.27.

$$M = (x^+ - x^-) \cdot n = (x^+ - x^-) \cdot \frac{w}{\|w\|} = \frac{2}{\|w\|} \tag{2.2}$$

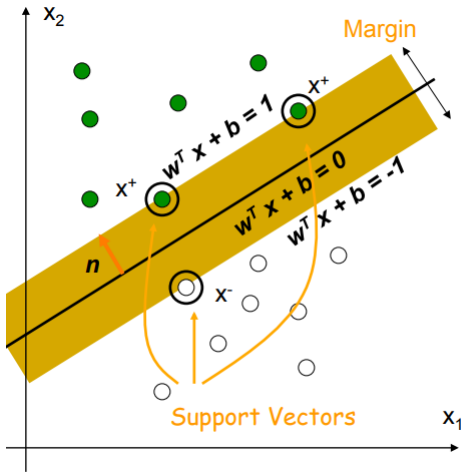


Figura 2.27: Formulaciones gráfica del margen.

En el caso de dos clases definidas como $y_i = +1$ o $y_i = -1$, el problema de optimización se formula de la siguiente manera:

$$\max \quad \frac{2}{\|w\|} \tag{2.3}$$

$$\text{Para } y_i = 1 \; w^T x + b \geq 1 \tag{2.4}$$

$$\text{Para } y_i = -1 \; w^T x + b \leq -1 \tag{2.5}$$

Realizando una transformación vectorial y buscando el mínimo de inverso, el problema queda escrito así:

$$\min \quad \frac{1}{2} \|w\|^2 \tag{2.6}$$

$$\mathrm{y}_i(w^T x + b) \geq 1 \tag{2.7}$$

$$\tag{2.8}$$

Una de las soluciones se obtiene con los multiplicadores de lagrange, una vez el problema se resuelve la decisión de clasificación se entrega con el signo de la ecuación (2.1).

2.7.3.3. Árboles de decisión: Gradient Boosting

Existen muchos tipos de algoritmos que usan árboles de decisión, estos se diferencian de otros clasificadores porque el valor a optimizar no es una distancia paramétrica, como en el caso de SVM, si no que es un diagrama de decisiones binarias en función de los valores de las variables de entrada.

Gradient Boosting es parte de este grupo de clasificadores y consiste en el uso de muchos árboles de decisión de forma iterativa, ya que se crean árboles con subconjuntos de las variables de entrada donde se prueban distintas combinaciones con el objetivo de encontrar la configuración óptima, en la figura 2.28 se muestra de forma resumida como en gradient

boosting se itera sobre distintos subárboles buscando el clasificador óptimo, en la actualidad gradient boosting es usado en gran medida para aplicaciones reales [12].

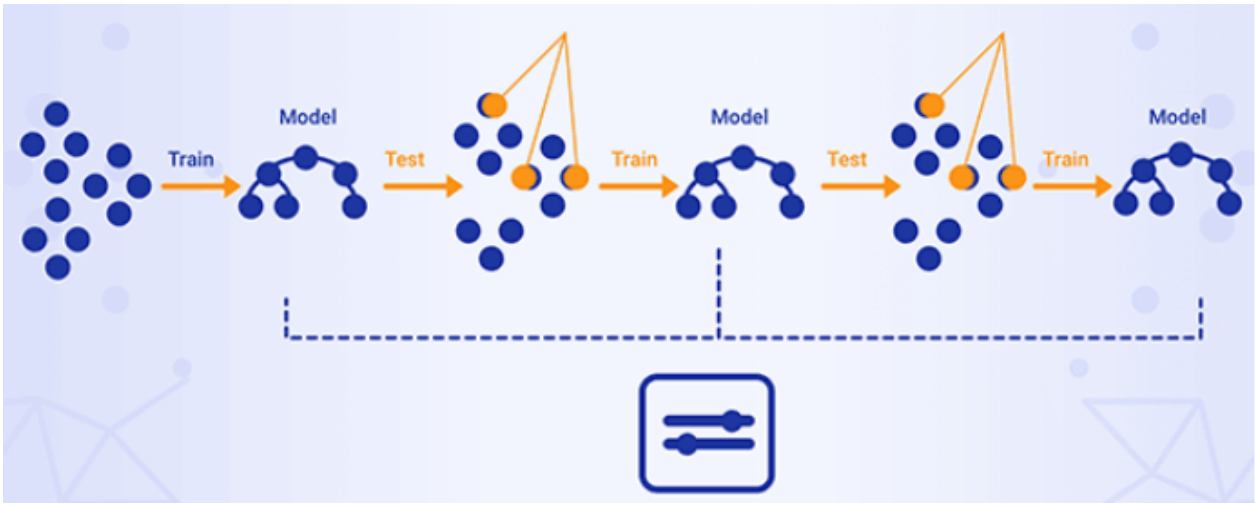


Figura 2.28: Esquema del funcionamiento de gradient boosting, fuente [13].

2.7.4. Técnicas de regresión

A continuación se mencionan algunas de las principales técnicas o algoritmos de regresión [11].

2.7.4.1. Regresión lineal

Es una de las técnicas de regresión más simples y populares, consiste en asumir que las variables de entrada tienen una relación lineal con respecto a la variable de salida, de esta forma se propone que la salida y es una combinación lineal de todas las entradas, como se muestra en la ecuación (2.9) donde cada variable x tiene un peso w, el problema de optimización es encontrar el vector de pesos w que minimice el error entre la curva real y predicha.

$$y(x) = w^T x + c = \sum_{j=1}^{N} w_j x_j + c \tag{2.9}$$

Una de las métricas más comunes a la hora de optimizar el error es el MSE o error cuadrático medio, una vez que se encuentra la curva óptima se obtienen rectas como la de la figura 2.29.

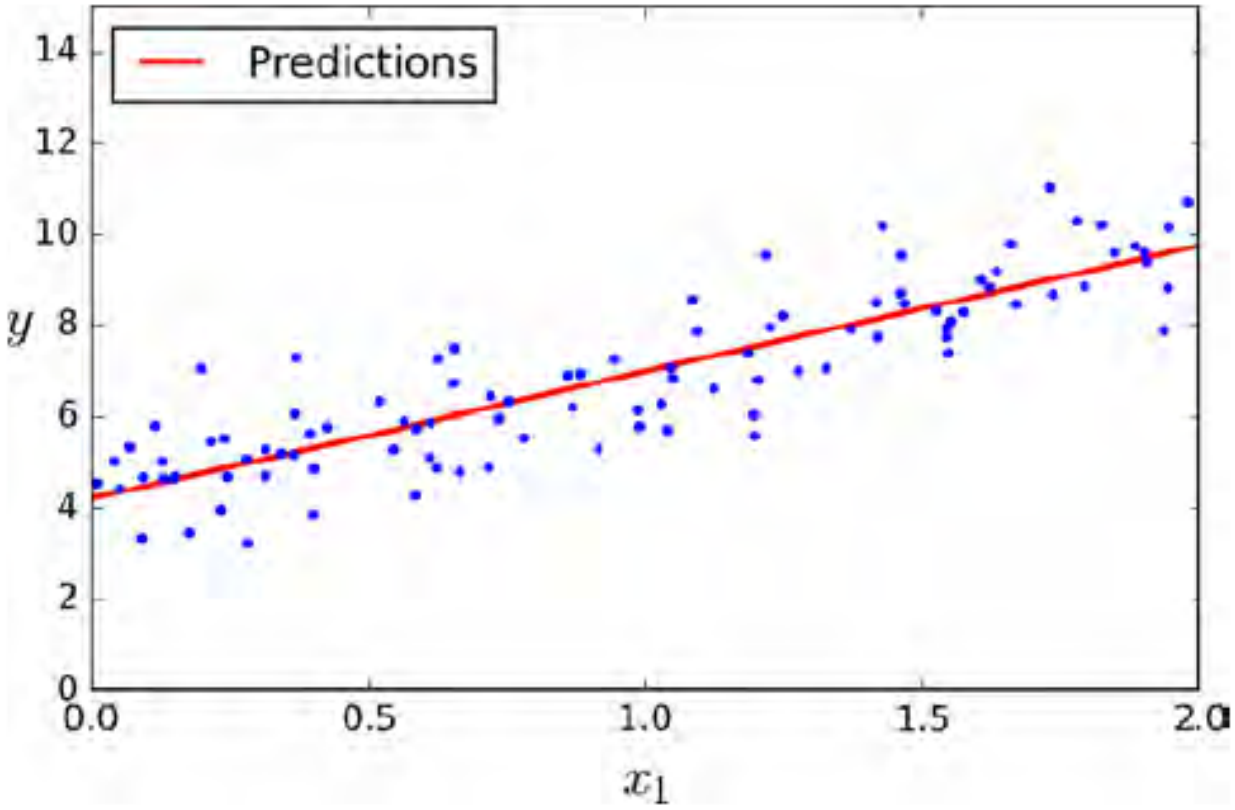


Figura 2.29: Ejemplo de regresión lineal.

2.7.4.2. Redes neuronales: Deep learning

Es necesario definir de forma general el cómo funcionan los algoritmos con redes neuronales, pues se utilizarán distintas configuraciones en la arquitectura de red con la cual se obtendrán las predicciones de series de tiempo.

Las redes neuronales artificiales son una técnica de machine learning supervisado que simula el proceso de aprendizaje biológico a través de nodos artificiales, estos se conocen como perceptrón y su estructura es análoga a la de una neurona biológica, pues tiene muchas entradas y una única salida, simulando las dendritas y axón respectivamente [13]. En figura 2.30 se ilustra un ejemplo de la comparación entre ambas estructuras.

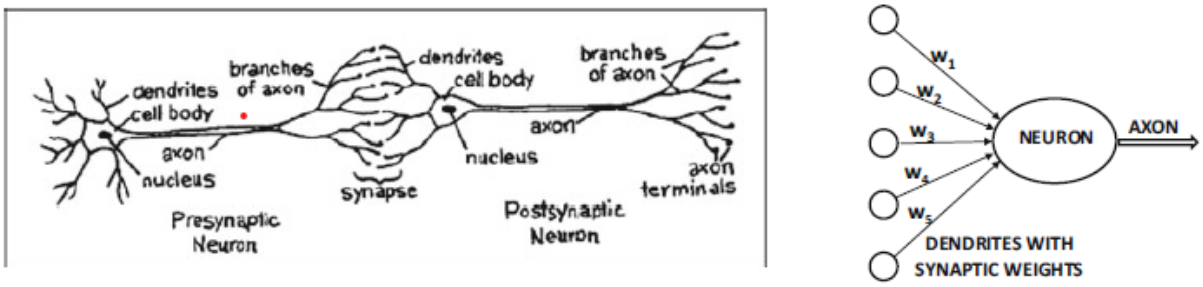


Figura 2.30: Comparación de neuronas biológicas (izquierda) y una artificial (derecha) [13].

La unidad básica de una red neuronal es el perceptrón y se compone de múltiples entradas donde cada una de estas es multiplicada por un escalar llamado peso w_k, el proceso que realiza la neurona es sumar todas las entradas x_k por su peso respectivo para finalmente aplicar sobre el resultado una función especial conocida como función de activación, cuya cualidad es ser no lineal. Dentro de la sumatoria es agregado un termino constante e independiente de las entradas, a este valor se le conoce como bias b_k.

La red más simple consiste en una única neurona, como es esquematizado en la figura 2.31, la forma en la que opera es tratar cada característica como una entrada distinta, luego a cada entrada se le asigna un peso para después sumar todos estos los términos, finalmente el resultado de esta sumatoria es operado con la función de activación entregando la salida u output y_k [14].

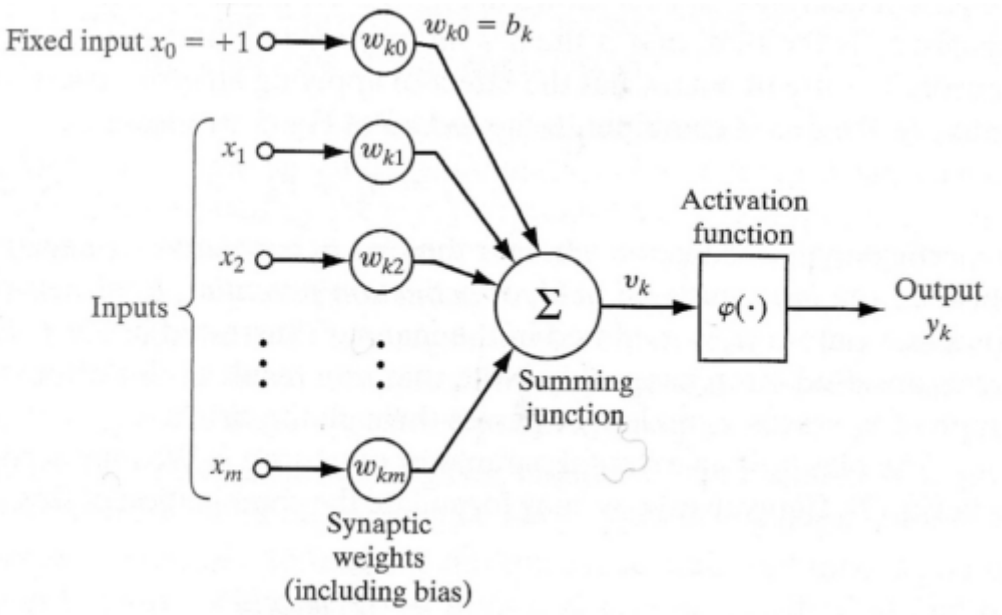


Figura 2.31: Esquema de una red neuronal compuesta por un único perceptrón, fuente [13].

Las dos operaciones que realiza una neurona son descritas por la relación (2.10), donde la entrada se compone de m variables, en cuanto al procedimiento de operación de la neurona lo primero que hace es sumar todos los pesos w_j por su entrada x_j, para después ingresar el resultado en la función de activación φ, finalmente se obtiene la salida como el resultado de la función de activación.

$$\begin{Bmatrix} S = \sum_{j=0}^{m} w_k \cdot x_j \\ y = \varphi(S) \end{Bmatrix} \tag{2.10}$$

El principal elemento de entrenamiento en una red neuronal son los pesos, la particularidad de este algoritmo es que se trata de un aprendizaje secuencial, es decir que por cada muestra o dato ingresado se aplica una corrección en los pesos. Estos pesos son expresados como un vector, en la ecuación (2.11) se muestra como w esta sujeto a cambios según el error (si la salida fue correcta o no), en cuanto al resto de variables se denota el valor real como t_i, mientras que la salida obtenida como y_i, además se usa el parámetro η que recibe el nombre de tasa de aprendizaje y es usado para controlar el error.

$$\Delta\vec{w} = \eta(t_i - y_i) \cdot \vec{x_i} \tag{2.11}$$

Observando la relación (2.10) se puede ver que la sumatoria esta compuesta de solo operaciones lineales, de allí nace la importancia de la función de activación, pues agrega un factor no lineal a la red lo que sirve para interpretar problemas más complejos, como por ejemplo la predicción de variables en el futuro, algunos ejemplos de función de activación se muestran en la figura 2.32.

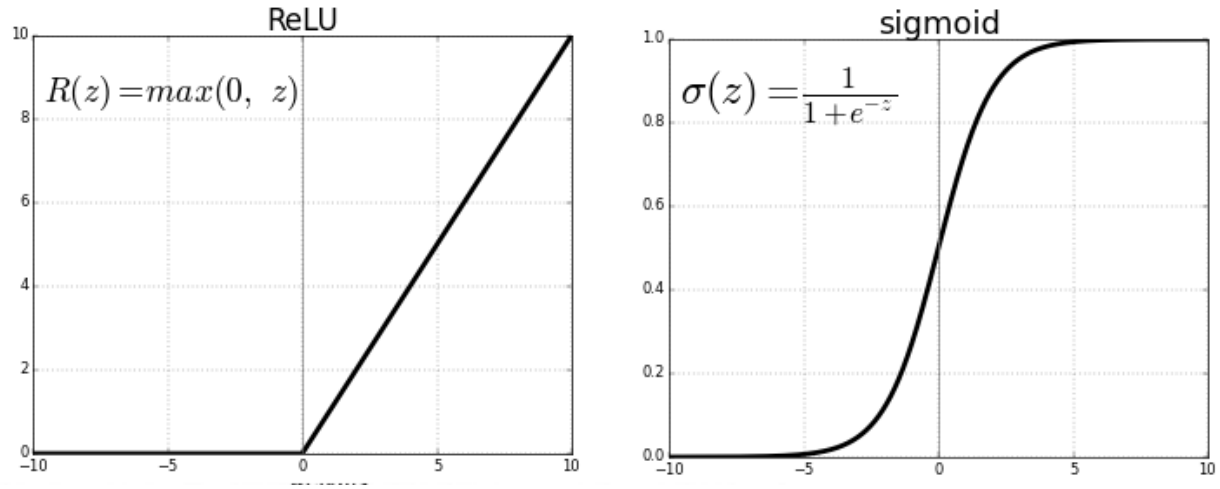


Figura 2.32: Función ReLU y sigmoide.

Hasta el momento se presentó una red compuesta por solo una neurona, la cual es la arquitectura más simple y por lo tanto no permite resolver problemas complejos, para ello se usan otras arquitecturas de red con más neuronas.

Para arquitecturas de multicapa se consideran dos capas fundamentales, estas son las de entrada y salida, la primera esta conectada directamente a las variables o features, mientras que la segunda a las salidas y es por esto que el número de neuronas en estas capas es igual al número de variables y salidas, respectivamente. Es posible agregar más capas intermedias, las cuales reciben el nombre de capas ocultas pues no son visibles para el usuario (caja negra).

En una red de multicapa cada neurona de cada capa está conectada con todas las neuronas de la capa anterior y posterior, lo que se conoce como esquema feedforward, pues el procesamiento es alimentar el output o salidas de forma secuencial a la capa siguiente, en la figura 2.33 se ilustra una arquitectura feedforward con dos capas ocultas [13].

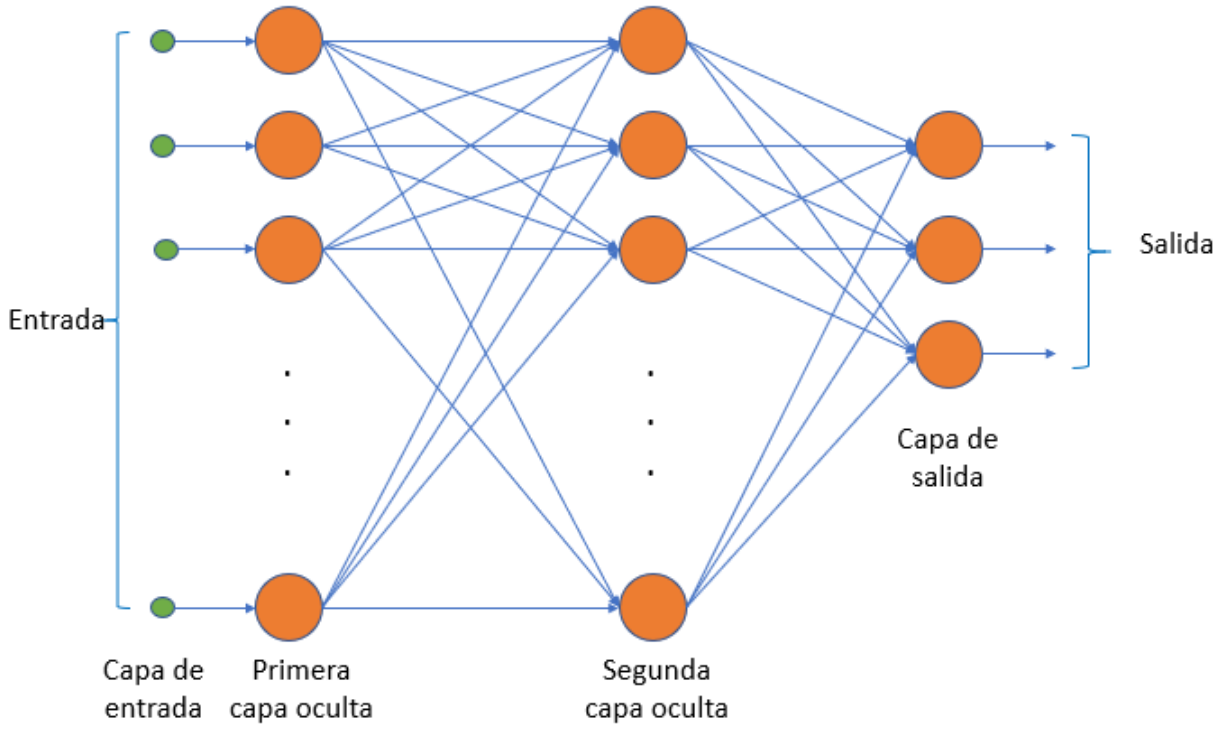


Figura 2.33: Ejemplo de una red multicapa feedforward.

Para el caso de predicción de series de tiempo se utiliza una configuración especifica de redes multicapa conocida como redes recurrentes LSTM (Long Short-Term Memory Network), estas se caracterizan por realimentar la salida a la red nuevamente, por lo que tienen buenas propiedades para aprender y predecir el comportamiento de patrones no lineales en series de tiempo.

2.7.5. Validación de error y métricas

Uno de los pasos más importantes en la construcción de un modelo con aprendizaje de máquinas es el análisis de errores, pues permite validar el comportamiento del modelo para finalmente ser llevado a producción o uso comercial.

Dependiendo el tipo de algoritmo se usan métricas distintas, a continuación se mencionan las principales métricas para algoritmos de clasificación y regresión.

2.7.5.1. Métricas de clasificación

En clasificación se busca etiquetar con una clase a las distintas entradas, es por esto que resulta natural evaluar el comportamiento de un clasificador según qué tan bien o mal acierta al predecir cada clase. Una forma intuitiva de graficar los resultados de un clasificador es comparando las veces que las muestras de cada clase fueron catalogadas correctamente y en el caso de error notificar con que clase fue erróneamente etiquetada, esto es lo que se conoce como matriz de confusión.

La matriz de confusión nos muestra la cantidad de aciertos y errores según la clase real y predicha, en la figura 2.34 se muestra el caso de un clasificador de flores, se puede ver que en diagonal están los casos donde cada flor fue correctamente clasificada, mientras que en el triangulo superior e inferior los errores, en este caso la matriz nos dice que la flor de tipo versicolor fue confundida 6 veces con la flor tipo virginica.

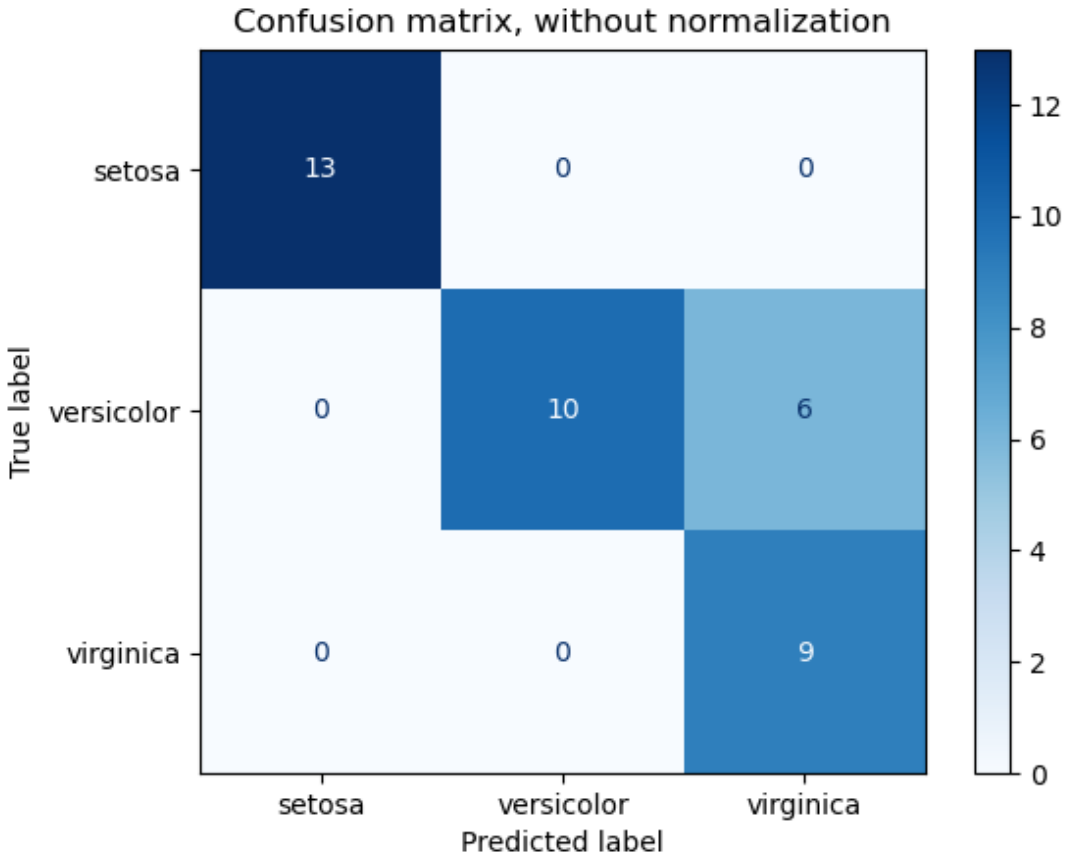


Figura 2.34: Ejemplo de una matriz de confusión para tres clases.

Existen tecnicismos para mencionar si una clase fue catalogada correctamente, para explicar los tecnicismos se usa como referencia una clase teórica denominada A.

- Verdadero positivo (VP): Predicción correcta que ocurre cuando una muestra de clase A fue etiquetada como clase A.
- Verdadero negativo (VN): Predicción correcta que ocurre cuando una muestra distinta a la clase A fue correctamente catalogada con su respectiva clase real.
- Falso positivo (FP): Predicción incorrecta que ocurre cuando una muestra de clase distinta a clase A fue catalogada como clase A.
- Falso negativo (FN): Predicción incorrecta que ocurre cuando una muestra de clase de clase A fue catalogada como una de las clases distintas a la A.

Utilizando los tecnicismos mencionados es posible calcular las tasas de acierto, estas son las siguientes:

- Accuracy: Tasa de aciertos considerando tanto los verdaderos positivos como los negativos sobre el total de las muestras.

$$Acc = (VP + VN)/(VP + VN + FP + FN) \tag{2.12}$$

- Precisión: Tasa de aciertos en las muestras etiquetadas como positivas para una respectiva clase, es decir cuantas de las muestras que fueron catalogadas como positivas realmente lo son, cuantifica la presencia de falsos positivos en el clasificador.

$$Precisión = (VP)/(VP + FP) \tag{2.13}$$

- Recall: Tasa de aciertos sobre todas las muestras de una clase real, cuantifica la presencia de falsos negativos en el clasificador.

$$Recall = (VP)/(VP + FN) \tag{2.14}$$

- F1: Métrica compuesta por la precisión y el recall, es considerada para casos donde existe desbalanceo entre las distintas clases.

$$F1 = 2 * Precision * Recall/(Precisión + Recall) \tag{2.15}$$

2.7.5.2. Métricas de regresión

Para el caso de regresión las métricas usadas están relacionadas con la distancia entre el valor real y el valor predicho, las tres principales métricas son el error promedio absoluto, error cuadrático medio y el coeficiente de determinación o r^2.

Sea y_r el valor real de una muestra e y_p la predicción, las métricas mencionadas se calculan de la siguiente forma:

- Error absoluto (MAE): Distancia entre el valor de la predicción y el valor real, debido a que esta en las unidades de medición de la variable sirve para tener una idea clara del error, por lo tanto es una buena herramienta para visualizar el error.

$$MAE = \frac{\sum_{i=0}^{n} |y_{r_i} - y_{p_i}|}{n} \tag{2.16}$$

- Error cuadrático medio (MSE): Distancia al cuadrado entre la predicción y el valor real, sirve como métrica de optimización en los algoritmos debido a que le da mayor importancia a errores grandes.

$$MSE = \frac{\sum_{i=0}^{n} (y_{r_i} - y_{p_i})^2}{n} \tag{2.17}$$

- Raíz del error cuadrático medio (RMSE): Raíz de la distancia al cuadrado entre la predicción y el valor real, mantiene la unidad de medida y da mayor importancia a errores grandes

$$RMSE = \frac{\sum_{i=0}^{n} \sqrt{(y_{r_i} - y_{p_i})^2}}{n}$$

(2.18)

- Coeficiente de determinación (r^2): Métrica entre 0 y 1 que cuantifica la similitud entre curva de predicción y la curva real, mientras más cercana sea a 1 mejor es la predicción, esta métrica es calculada como la división entre covarianza de los valores reales con la predicción y el producto de las varianzas.

$$r^2 = \frac{\sigma^2_{y_r y_p}}{\sigma^2_{y_r}\sigma^2_{y_p}} \tag{2.19}$$

Capítulo 3

Metodología

Los objetivos de esta sección son dos, el primero consiste en entregar un dimensionamiento acotado de la red LTE de Entel con el cual trabajar ya que existen limitaciones de hardware para el manejo datos para. El Siguiente objetivo es definir la metodología a seguir para la construcción de los algoritmos de predicción, junto a los KPI a usar y los métodos de validación de resultados.

3.1. Dimensionamiento de red LTE Entel

3.1.1. Zona geográfica

Actualmente Entel posee 36 zonas geográficas en la región metropolitana, de los cuales 29 están ubicados en Santiago, estas zonas también reciben el nombre de cluster o polígonos. Es importante destacar la diferencia entre dos elementos que componen un cluster, estos son los sitios y celdas, a continuación se muestra una descripción de ambos conceptos, también se muestra en la figura 3.1 la diferencia entre celda y sitio.:

- Sitio: Estación base donde se emplazan múltiples antenas, es decir, el punto físico que puede alojar múltiples celdas.
- Celda: Área de cobertura para una tecnología específica, por ejemplo en el caso de LTE existen celdas para las bandas de 700, 1900 y 2600 MHz respectivamente.

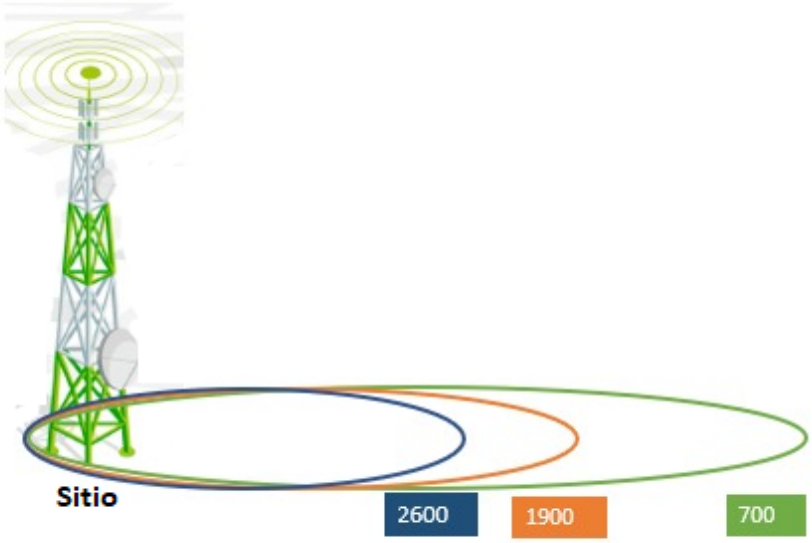


Figura 3.1: Área de cobertura de celdas LTE en un sitio.

En la región metropolitana la red LTE de Entel cuenta a la fecha con 1556 sitios y 10662 celdas, en la figura 3.2 se muestran los sitios ubicados en los distintos clusters de Santiago, por otro lado la cantidad de celdas según la tecnología se muestra en la tabla 3.1

Tabla 3.1: Cantidad de celdas por banda.

Banda	Cantidad de celdas
700 MHz	2406
1900 MHz	4132
2600 MHz	4124

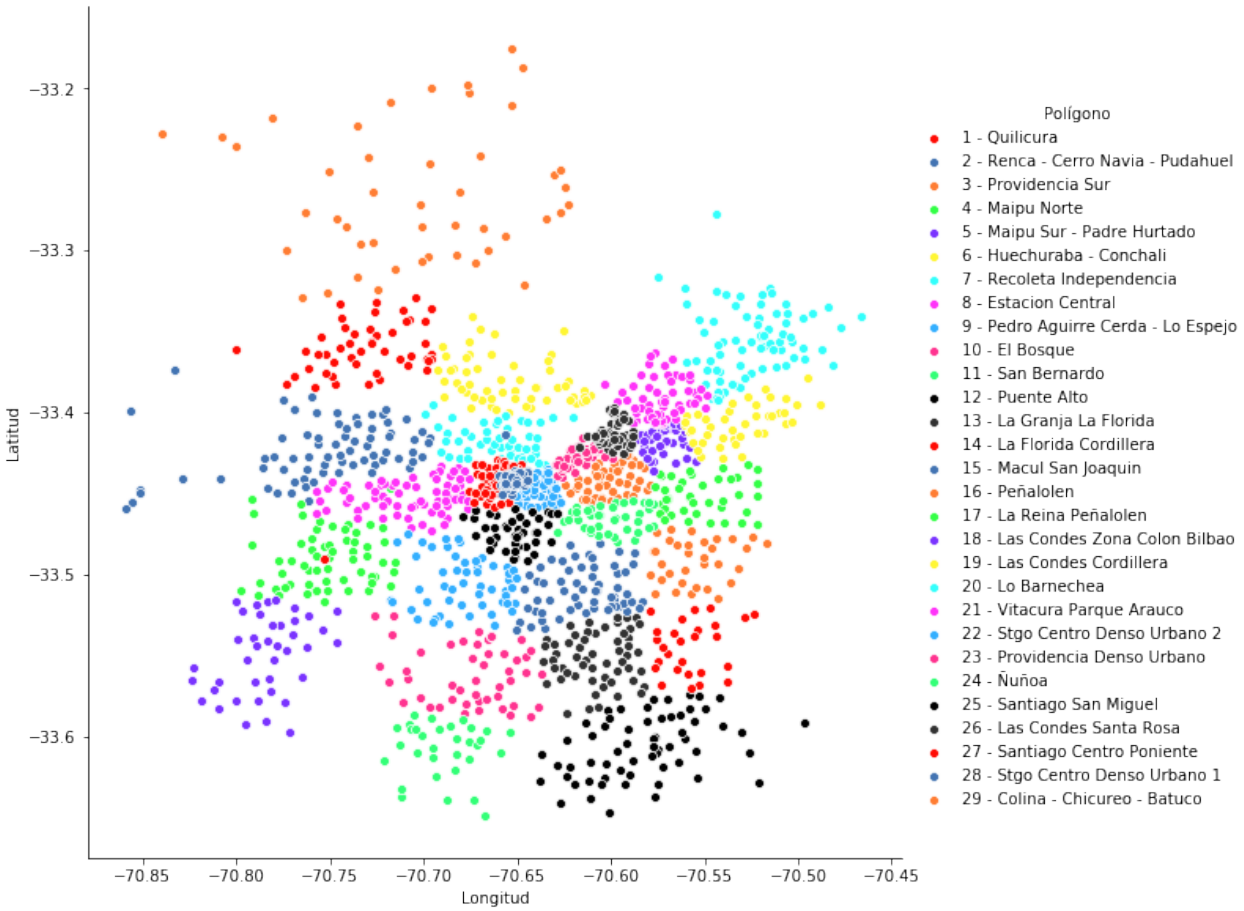


Figura 3.2: Sitios en los distintos clusters de Santiago.

3.1.2. Selección de clusters

Considerando que para efectos de esta memoria se plantea utilizar un muestreo de datos por celda cada 15 minutos, resulta inviable un análisis en toda la región metropolitana, es por esto que es imprescindible acotar el estudio a solo dos clusters. Es necesario que los clusters seleccionados en primer lugar sean vecinos, para que así tengan un comportamiento similar y en segundo lugar que tengan ciertas características de interés para el estudio, como por ejemplo clusters con tendencia a tener mal throughtput para así tener una muestra representativa.

Es así como se realiza un ranking de clusters utilizando los siguientes criterios:

- Volumen de datos
- Perfil de usuario
- Rendimiento del cluster
- Variabilidad en curvas de throughput
- Zona geográfica

3.1.2.1. Datos utilizados

Para la selección de clusters fueron considerados 5 indicadores de rendimiento relacionados al tráfico de datos, estos corresponden a datos de la interfaz de radio como se propone en [15] :

- Vol_DL: Volumen de datos en el canal de descarga, su unidad de medida es en Gb.
- DL_User_Thp: Throughput promedio de descarga experimentado por el usuario, su unidad de medida es en Mbps.
- USO_DLPRBs: Uso porcentual promedio de los PRB, o sea de la cantidad de PRB disponibles cuantos son efectivamente usados.
- USER_DL: Cantidad de usuarios activos por sitio
- LTE_Active_Users_DL: Cantidad de usuarios activos por celda.

Los datos fueron tomados cada 15 minutos a nivel de sitio, mientras que a nivel de celda las mediciones son diarias y en hora peak, estos datos contemplan un periodo de 3 semanas dado entre el 29 de Junio y 19 de Julio de 2020. Una vez seleccionados los clusters finales el análisis a nivel de celda pasará a muestras cada 15 minutos.

3.1.2.2. Criterio de selección: Volumen

Ordenando los cluster según cantidad de datos traficados se obtiene el ranking de la tabla 3.2. En general los polígonos de mayor área son los que más trafican, con el objetivo de normalizar los datos se realiza el mismo ranking en la tabla 3.3 pero considerando el numero de celdas por cluster.

Tabla 3.2: Clusters con mayor volumen de tráfico de datos.

Polígono	Volumen [Gb]
12 - Puente Alto	7.510203e+06
2 - Renca - Cerro Navia - Pudahuel	7.184878e+06
10 - El Bosque	6.262924e+06
4 - Maipú Norte	5.988855e+06
13 - La Granja La Florida	5.929605e+06
8 - Estacion Central	5.628435e+06
9 - Pedro Aguirre Cerda - Lo Espejo	5.215288e+06
5 - Maipú Sur - Padre Hurtado	5.131589e+06
29 - Colina - Chicureo - Batuco	5.089579e+06
15 - Macul San Joaquin	4.719646e+06
20 - Lo Barnechea	4.455000e+06
6 - Huechuraba - Conchali	4.283874e+06

Se observa que al normalizar según cantidad de celdas el ranking cambia considerablemente, pues es liderado por clusters perimetrales y residenciales de la región metropolitana.

A partir del volumen de datos se obtienen dos pares candidatos, estos son:

- Maipú Norte/Maipú Sur - Padre Hurtado (Clusters 4/5).
- Puente Alto/La Granja La Florida (Clusters 12/13).

Tabla 3.3: Clusters con mayor volumen de tráfico de datos por celda.

Polígono	Volumen por celdas [Gb]
Lampa	32226.828994
Buin	24631.665994
Talagante	23707.153169
Peñaflor	21032.672136
El Monte	20250.245788
29 - Colina - Chicureo - Batuco	16797.291369
11 - San Bernardo	16584.604548
5 - Maipú Sur - Padre Hurtado	16447.399301
12 - Puente Alto	16433.704180
10 - El Bosque	15976.845787
16 - Peñalolén	15549.077211
2 - Renca - Cerro Navia - Pudahuel	15385.178151

3.1.2.3. Criterio de selección: Perfil de usuario

Se entiende como perfil de usuario la cantidad de datos que trafica un usuario en promedio según el cluster, los resultados del ranking consideran el periodo de 3 semanas y se muestran en la 3.4, resulta interesante este criterio para detectar polígonos con perfil de alto consumo de datos.

Tabla 3.4: Clusters con mayor tráfico de datos por usuario.

Polígono	Datos por user promedio [Gb]
Buin	3.540263
Lampa	3.277685
13 - La Granja La Florida	3.246643
27 - Santiago Centro Poniente	3.205053
4 - Maipú Norte	3.162720
5 - Maipú Sur - Padre Hurtado	3.146896
El Monte	3.138607
2 - Renca - Cerro Navia - Pudahuel	3.128443
10 - El Bosque	3.088872
3 - Providencia Sur	2.998283
7 - Recoleta Independencia	2.943590
24 - Ñuñoa	2.938516

Nuevamente se puede observar el par de clusters Maipú norte/sur, también se repite la presencia de polígonos residenciales y perimetrales de la región metropolitana, como por ejemplo Buin y Lampa.

3.1.2.4. Criterio de selección: Rendimiento del cluster

El rendimiento del cluster es cuantificado a través del throughtput en el canal de descarga, un sitio o celda tendrá un buen rendimiento si cumple el siguiente requisito:

- DL_User_Thp > 3,3 Mbps en hora peak para celdas.
- DL_User_Thp > 4,6 Mbps en hora peak para sitios.

Considerando este criterio es calculada la cantidad de veces que una celda tuvo un mal rendimiento en el periodo estudiado, los resultados se muestran en la tabla 3.5 y son separados por banda. Destacan los dos pares de clusters vecinos ya mencionados (clusters 4/5 y 12/13), sobre todo Puente Alto con la mayor cantidad de eventos con mal rendimientos, también se observa que la banda de 700 MHz es la que tiene peor rendimiento, seguida por la de 1900 MHz y 2600 MHz.

Tabla 3.5: Cantidad de eventos con mal rendimiento en celdas.

Polígono	Banda_700	Banda_1900	Banda_2600	Total
12 - Puente Alto	714	278	160	1152
10 - El Bosque	454	259	190	903
29 - Colina - Chicureo - Batuco	394	236	156	786
4 - Maipú Norte	369	190	104	663
2 - Renca - Cerro Navia - Pudahuel	384	118	159	661
16 - Peñalolén	267	245	122	634
9 - Pedro Aguirre Cerda - Lo Espejo	314	178	140	632
8 - Estación Central	321	151	153	625
11 - San Bernardo	328	138	145	611
5 - Maipú Sur - Padre Hurtado	254	194	76	524
13 - La Granja La Florida	210	156	116	482

Segmentando los eventos de mal rendimiento por banda se obtiene el gráfico mostrado en la figura 3.3 donde se pueden ver dos patrones distintos: el primero es cuando celda, generalmente de la banda 700 o 1900, tiene muy pocos usuarios conectados y estos experimentan bajo throughtput, mientras que la segunda es cuando la celda esta sobre exigida pues el uso de los PRBs es cercano al 100 %.

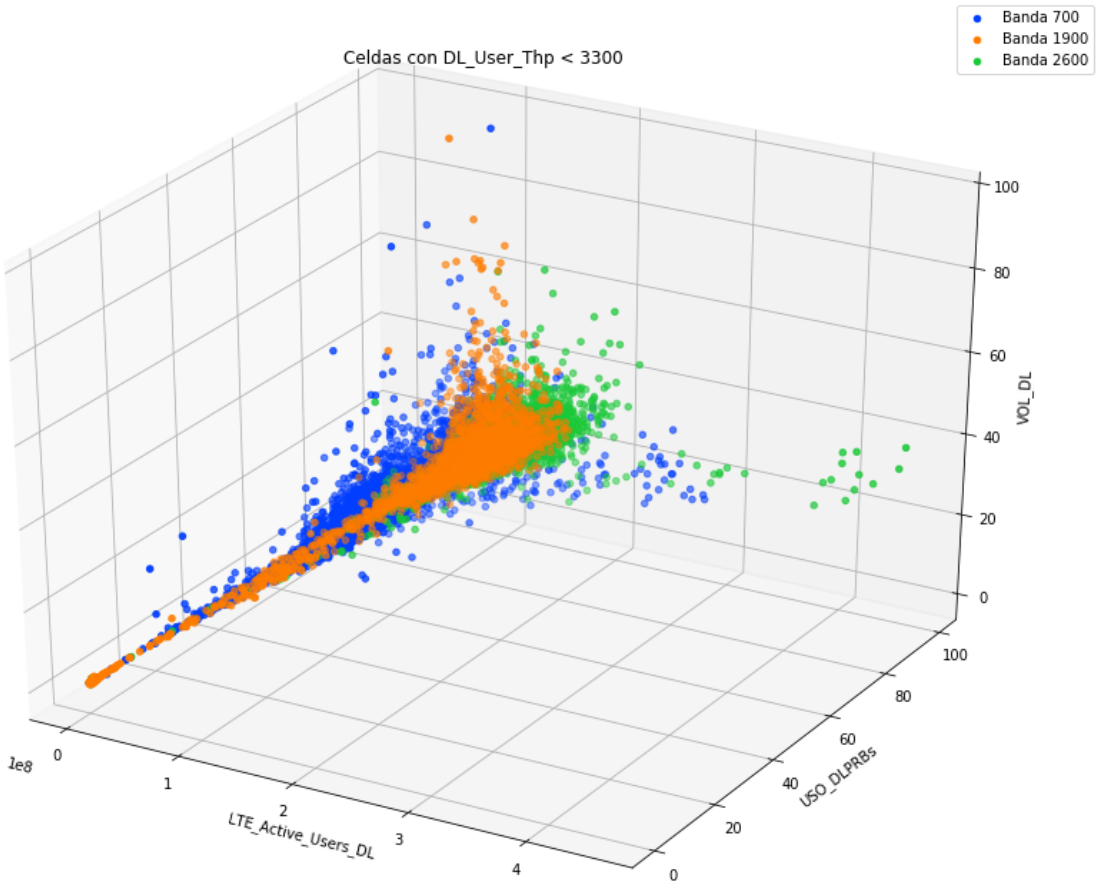


Figura 3.3: Distribución de celdas con mal rendimiento por KPI.

Otro aspecto considerable es cuantificar las celdas únicas que presentan mal rendimiento, ya que la tabla 3.5 contempla eventos repetidos en distintos días por las misma celdas, en la tabla 3.6 se muestra la cantidad de celdas únicas que presentaron mal rendimiento según polígono, donde nuevamente se obtiene que el peor rendimiento es en la banda de 700 MHz, además se puede interpretar que son comunes las celdas con mal rendimiento durante varios días.

Tabla 3.6: Cantidad de celdas únicas que presentaron eventos de mal rendimiento.

Polígono	Banda_700	Banda_1900	Banda_2600
12 - Puente Alto	75	47	35
10 - El Bosque	55	44	31
29 - Colina - Chicureo - Batuco	56	34	24
4 - Maipú Norte	45	36	24
2 - Renca - Cerro Navia - Pudahuel	56	29	37
16 - Peñalolén	41	38	20
9 - Pedro Aguirre Cerda - Lo Espejo	44	36	35
8 - Estación Central	41	38	34
11 - San Bernardo	34	19	25
5 - Maipú Sur - Padre Hurtado	31	30	20
13 - La Granja La Florida	38	27	19

3.1.2.5. Criterio de selección: Variabilidad en curvas de throughput

Este criterio tiene como finalidad encontrar los clusters con mayor variación en la curva de throughtput, esto es de interés pues un punto importante de la memoria es obtener predicciones del tráfico para encontrar posibles celdas con mal rendimiento en el futuro y realizar el mantenimiento preventivo necesario; por lo tanto es necesario considerar polígonos que tengan una variabilidad dinámica en la curva de throughtput, ya que es en celdas con este perfil donde se esperan eventos con mal rendimiento.

La variabilidad se mide con la desviación estándar del throughtput, es decir, se calcula la dispersión de los datos con respecto al promedio, el ranking se muestra en la tabla 3.7.

Tabla 3.7: Desviación estándar del Throughtput a nivel cluster.

Polígono	std_Throughput
Buin	10400.00
Melipilla	10394.11
18 - Las Condes Zona Colon Bilbao	10246.81
Lampa	9837.90
3 - Providencia Sur	9707.47
5 - Maipú Sur - Padre Hurtado	9634.99
1 - Quilicura	9516.33
Aeropuerto Arturo Merino Benítez	9370.39
14 - La Florida Cordillera	9304.61
4 - Maipú Norte	9243.67
21 - Vitacura Parque Arauco	9162.62

3.1.2.6. Criterio de selección: Zona geográfica

Para el análisis por zona geográfica fue considerado el comportamiento de los KPI según el cluster, se encontró que los polígonos siguen dos patrones distintos al graficar las mediciones de los KPI a nivel sitio:

- Primer grupo: Curvas de KPI bien definidas y con único comportamiento, la mayoría de los polígonos sigue este comportamiento, como se puede observar en el ejemplo de la figura 3.4.
- Segundo grupo: Curvas de KPI difusas y con comportamiento doble como se puede observar en el ejemplo de la figura 3.5, generalmente los polígonos del centro de Santiago siguen este comportamiento.

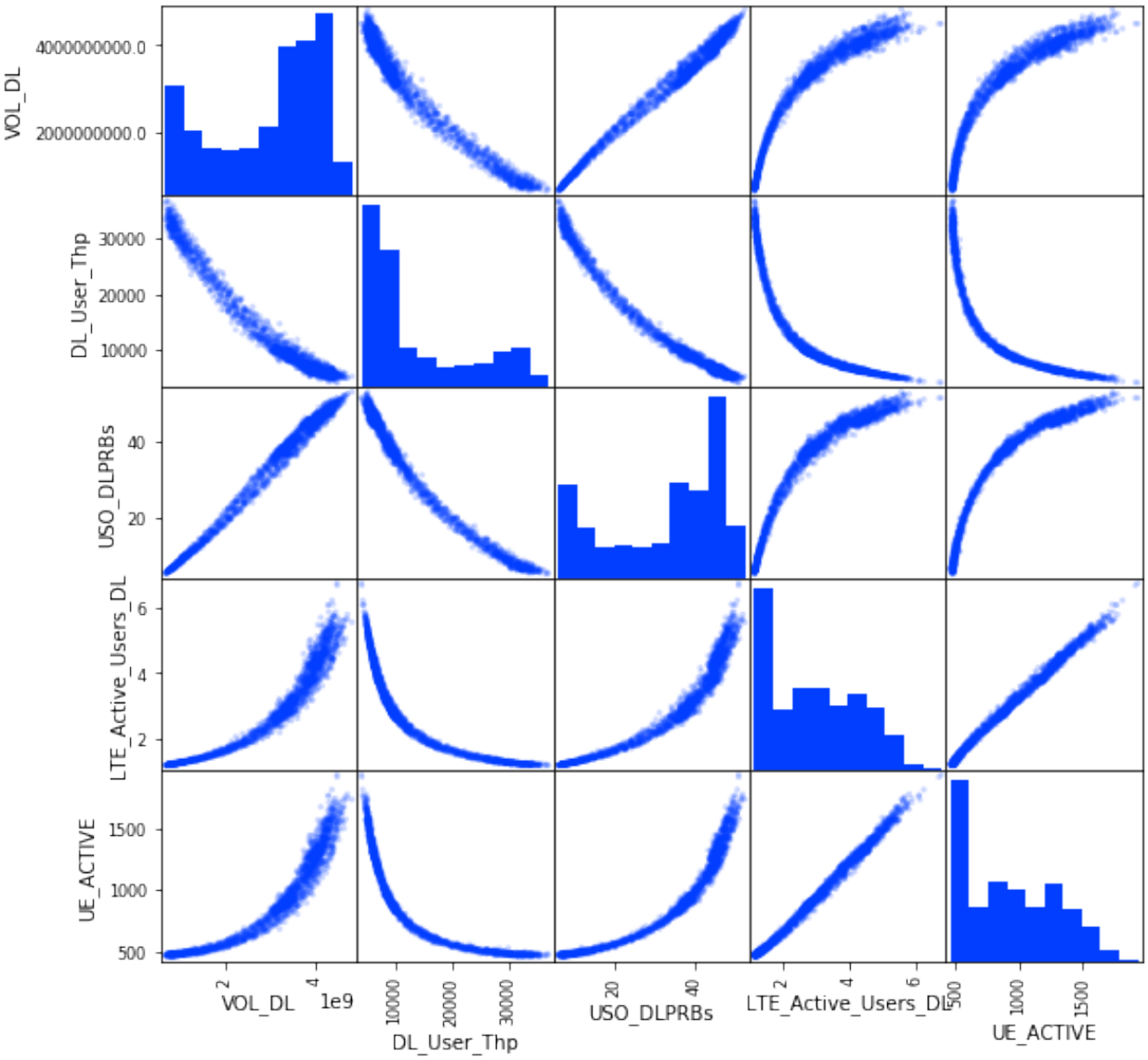


Figura 3.4: Curvas KPI del polígono Maipú norte, sigue el patrón de primer grupo.

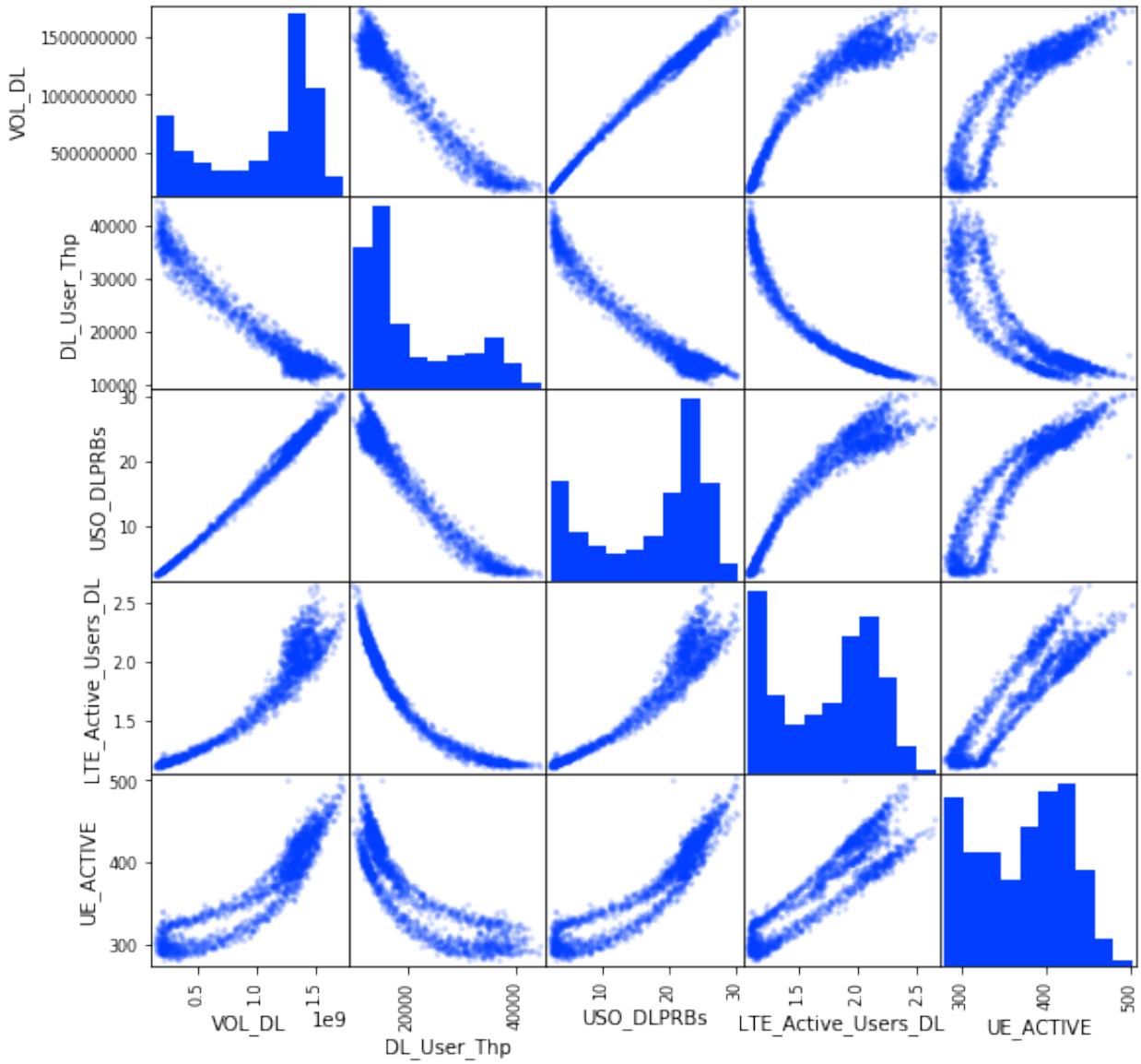


Figura 3.5: Curvas KPI del polígono Providencia, sigue el patrón de segundo grupo.

3.1.2.7. Clusters seleccionados

En función de los resultados observados según cada criterio se obtienen dos pares candidatos, pues aparecen entre los primeros clusters de todos los ranking, además de ser polígonos con una considerable cantidad de eventos con bajo rendimiento lo que motiva el análisis predictivo.

- Maipú Norte/Maipú Sur - Padre Hurtado.
- Puente Alto/La Granja La Florida.

Entre ambos pares no hay grandes diferencias, se optó por seguir el estudio con los polígonos Maipú norte/sur, un cluster es utilizado durante todo el desarrollo de pruebas y cambios mientras que el segundo es reservado para los modelos finales, esto a causa de las limitaciones técnicas de hardware.

3.2. Metodología para aprendizaje de máquinas

Para la construcción de un modelo funcional utilizando aprendizaje de máquinas es necesario considerar distintas alternativas y estrategias tanto en el modelamiento de datos como en los algoritmos a utilizar. La siguiente sección contempla de forma detallada el diagrama de flujo llevado a cabo para obtener el mejor modelo.

El objetivo principal de este trabajo de titulo es optimizar el rendimiento de la red LTE utilizando metodologías predictivas sobre el tráfico, para ello se define como KPI objetivo el Throughtput de descarga a nivel de usuario (DL_User_Throughtput) con el fin de predecir cuando una celda esta bajo el umbral aceptado e identificar cuales son las celdas candidatas para aplicar traspaso de usuarios y así lograr mejoras en el valor de este indicador.

Para ello el modelo se divide bloques principales, como se muestra en el diagrama de la figura 3.6 donde la primera etapa es un algoritmo clasificador de celdas con mal rendimiento, seguido de un filtro de celdas candidatas según KPI para finalmente aplicar un algoritmo de regresión sobre las celdas candidatas y observar en detalle o punto a punto el valor futuro del throughtput, a continuación se detallan los tres puntos:

- Bloque de clasificación: Fase del algoritmo que tiene como objetivo predecir de forma binaria el estado de las celdas con respecto al throughtput, si la predicción de throughtput de una celda esta bajo 3,3 Mbps entonces es etiquetada como celda con mal rendimiento.

  Este modelo toma los datos ya procesados de todas las celdas de un cluster y entrega una tabla con la etiqueta de estado para cada una, es decir la salida son los subconjunto de celdas con con mal y buen rendimiento.

- Bloque de filtrado por KPI: Fase del algoritmo que tiene como objetivo encontrar para cada celda con mal rendimiento el grupo de celdas candidatas a las cuales aplicar el balanceo de usuarios. La idea es que las celdas candidatas tengan un cierto perfil de uso de prb, cantidad de usuarios y características técnicas como pertenecer al mismo sitio, así que para ello se creó una función que realiza un ranking de las mejores celdas según criterio experto, además este filtro asegura que las celdas candidatas sean efectivamente de buen rendimiento según el bloque clasificador.

  La salida de este bloque son las 3 mejores celdas candidatas para cada celda de mal rendimiento.

- Bloque de regresión: Fase del algoritmo que tiene como objetivo entregar una predicción en detalle sobre el valor futuro de las celdas candidatas, tiene como entrada las 3 mejores celdas candidatas para cada celda con mal rendimiento y entrega como salida la curva de throughtput a nivel horario del día de predicción. Con ello se busca dar la mayor cantidad de información al usuario del modelo para la elección óptima en el balanceo de carga.

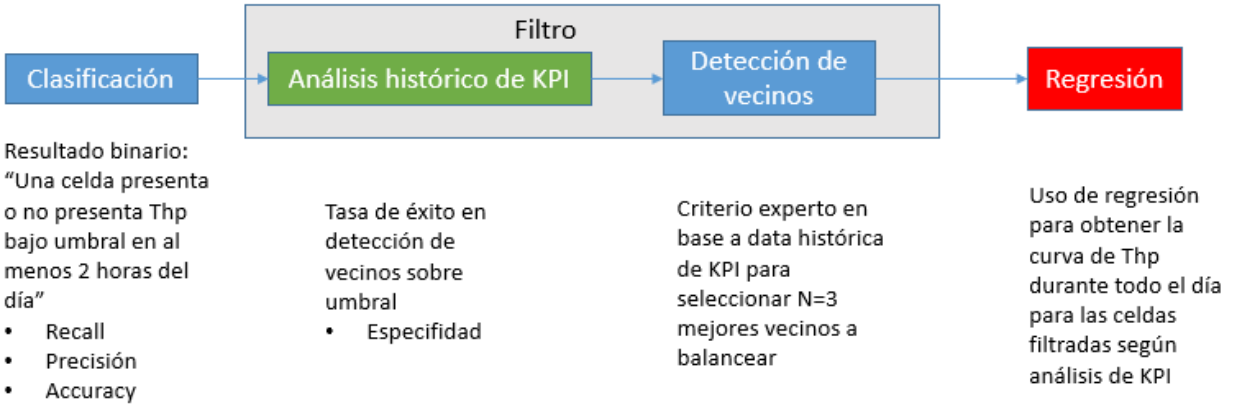


Figura 3.6: Diagrama de flujo del modelo.

Entre los algoritmos utilizados se contempló el uso de arboles de decisión para la etapa de clasificación, como por ejemplo random forest y gradient boosting. Mientras que para el bloque de regresión se optó por el uso de redes neuronales. Existen trabajos similares con redes neuronales en [16], otro punto de vista es utilizado en [17] aplicando métodos lineales como ARIMA.

3.2.1. Modelo: Datos

Previo a la etapa de clasificación es llevado a cabo el preprocesamiento de datos, esto contempla limpieza de valores nulos y selección de los KPI más importantes con respecto al throughtput.

Entel proporcionó una base de datos compuesta por 126 KPI de la red de acceso, además de la altura del sitio y tilt eléctrico (inclinación de la antena). Con el objetivo de simplificar el análisis y optimizar los algoritmos se buscó filtrar la cantidad de KPI a no más de 10, los criterios usados para el filtro son los siguientes:

- Criterio experto: Existen KPI que por naturaleza están más relacionados al throughtput, ademas de KPI redundantes que deben ser omitidos.
- Correlación: La correlación de Pearson permite encontrar relaciones lineales entre el throughtput y cada KPI, por lo que es un buen filtro de variables.
- Feedback de los algoritmos: A partir de ensayo y error se pueden buscar los KPI que mejoran el rendimiento de cada algoritmo, esta estrategia también es conocida como wrapper.

  Los datos son series de tiempo de cada KPI durante un periodo de 3 meses, las pruebas son realizadas con mediciones cada 15 minutos y cada una hora. En cuanto a la estructura de datos el algoritmo considera una muestra como una ventana de tiempo con los cuales predice el día futuro, como se muestra en la figura 3.7, en este trabajo se consideran distintas configuraciones en las ventanas de tiempo.

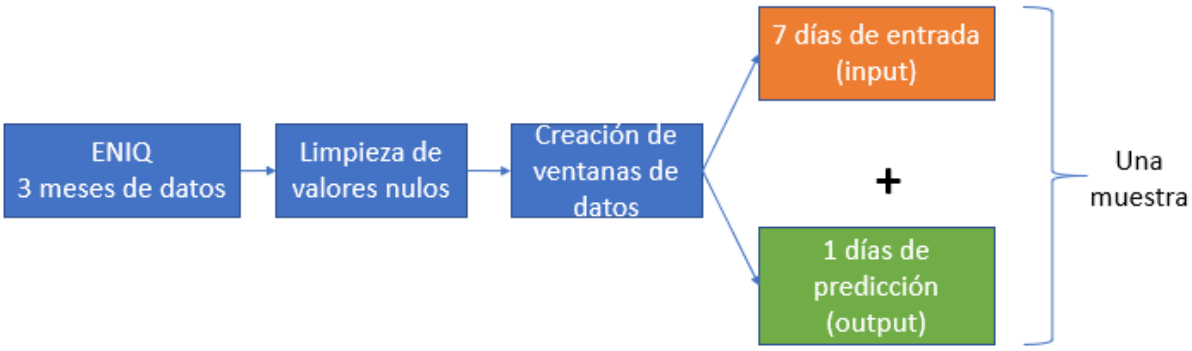


Figura 3.7: Estructura de los datos con un ejemplo de ventana 7+1 días.

3.2.2. Modelo: Clasificación

3.2.2.1. Clasificación: Procesamiento de datos

El bloque de clasificación considera como entrada las series de tiempo de los KPI seleccionados, debido a que el clasificador trabaja con una predicción de un día completo es necesario realizar un nuevo procesamiento a las series de tiempo, pues estas tienen un muestreo de 15 o 60 minutos. Es así que con el objetivo de caracterizar la información de los datos se calculan distintas funciones o estadísticos sobre las series de tiempo a nivel diario, sea $\vec{x}$ una serie de tiempo de un KPI las funciones usadas son las siguientes:

- Máximo: Valor máximo diario en la serie de tiempo.

$$Máximo = max(\vec{x}) \tag{3.1}$$

- Mínimo:Valor mínimo diario en la serie de tiempo.

$$Mínimo = min(\vec{x}) \tag{3.2}$$

- Promedio: Valor promedio diario en la serie de tiempo.

$$Promedio = \frac{\sum_{i=0}^{n} \vec{x_i}}{n} \tag{3.3}$$

- Rango: Rango en el que oscilan los valores de un KPI durante todo el día.

$$Rango = abs(max - min) \tag{3.4}$$

- Varianza: Estadístico de dispersión a nivel horario.

$$Varianza = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_{promedio})^2 \tag{3.5}$$

- Entropía promedio: Estadístico que cuantifica el nivel de irregularidad o incertidumbre en una serie de tiempo.

$$Entropía = -\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(x_i)log(\mathbb{P}(x_i)) \tag{3.6}$$

- Skewness: Estadístico que cuantifica el nivel de asimetría en un conjunto de datos al ser comparados con una distribución gaussiana.
- Kurtosis: Estadístico que cuantifica la frecuencia de datos en los bordes de una distribución guassiana.
- Mean Crossing Rate: Estadístico que cuenta la cantidad de veces que una serie de tiempo oscila con respecto a su valor promedio.

Para el etiquetado de la clase se utiliza la serie de tiempo DL User Throughtput, esta es transformada en un valor binario a nivel diario, cuyo número indica si durante el día estuvo al menos 2 horas bajo el umbral de 3.3 Mbps.

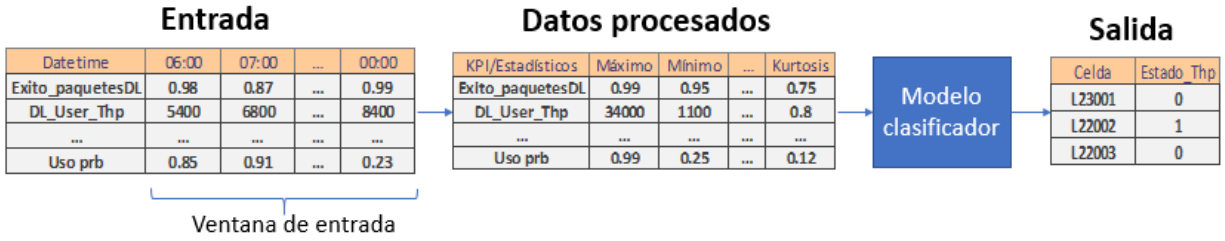


Figura 3.8: Entrada y salida del modelo clasificador.

3.2.2.2. Clasificación: Métricas de evaluación

Se utilizan las métricas típicas para evaluar un algoritmo de clasificación, estas son la matriz de confusión, accuracy, precisión, recall y F1, para efectos de este trabajo se define como clase positiva (clase = 1) la detección de una celda bajo el umbral de throughtput, caso contrario es etiquetada como clase negativa (clase = 0), en cuanto al resto de métricas estas son interpretada de la siguiente forma:

- Accuracy: Tasa de detección tanto en celdas con throughtput sobre y bajo el umbral.
- Precisión: Tasa de detección de celdas detectadas como bajo umbral, pero que realmente estaban sobre el umbral, por lo tanto indica el porcentaje de celdas que serán balanceadas pero que no se deben trabajar.
- Recall: Tasa de detección de las celdas que están bajo el umbral, es decir el porcentaje de celdas detectadas que si se deben balancear.
- Especificidad: Tasa de detección de celdas de celdas sobre umbral y que realmente estaban sobre el umbral, es el porcentaje de celdas candidatas de balanceo que son etiquetadas correctamente.

- F1: Indicador que considera tanto el recall y precisión, es útil para este problema pues las celdas con throughtput sobre el umbral son muchas mas que las de bajo umbral.

3.2.3. Modelo: Filtro de celdas según KPI

El bloque clasificador entrega el estado futuro de una celda, es decir si esta sobre o bajo el umbral de throughtput, el siguiente paso consiste en seleccionar cual es la mejor celda vecina con la que cada celda con throughtput bajo el umbral pueda balancear usuarios. Para eso se diseña una función de puntuación utilizando el historial del uso de prb, cantidad de usuarios, throughtput, intentos de handover y si la celda es cosite; con ello se elabora un ranking de las mejores celdas. La función entrega una nota que pondera de forma distinta cada elemento, a continuación se describe la formula y sus factores cuyos valores fueron decididos según criterio experto:

$$Puntuación = 0.35 \cdot \alpha + 0.45 \cdot \beta + 0.15 \cdot \gamma + 0.05 \cdot \delta \tag{3.7}$$

- Factor α: Factor de cosite cuya funcionalidad es cuantificar si la celda candidata pertenece al mismo sitio que la celda detectada con bajo throughtput, este factor solo puede tomar como valor 0 o 1 (Booleano).
- Factor β: Factor de handover cuya funcionalidad es cuantificar la cantidad de veces que una celda candidata realiza traspaso de usuarios con la celda objetivo, cuenta la cantidad de intentos de handovers realizados en la última semana.
- Factor γ: Factor de uso de prb cuya funcionalidad es cuantificar el uso de recursos radio de la celda candidata, mide el uso de prb diario en hora peak según volumen de datos, el valor usado es el promedio de las últimas dos semanas.
- Factor δ: Factor de throughtput cuya funcionalidad es medir el tráfico de la celda candidata, mide el throughtput diario en hora peak según volumen de datos, el valor usado es el promedio de las últimas dos semanas.

Una vez que las celdas candidatas son rankeadas según la puntuación se aplica un segundo filtrado para que ninguna celda candidata tenga una predicción de throughtput bajo umbral, esto es realizado con el bloque clasificador.

3.2.4. Modelo: Regresión

3.2.4.1. Regresión: Procesamiento de datos

El modelo regresor mantiene la misma estructura de datos utilizada por el clasificador, vale decir la misma cantidad de días en la ventana de entrada y solo un día de predicción. En cuanto al procesamiento de datos se utilizan las series de tiempo de cada KPI punto a punto, es decir se usa la serie de tiempo completa y no un estadístico como en el caso de la clasificación como se muestra en la figura 3.9, otra diferencia es que para el caso de regresión los datos de entrada son normalizados.

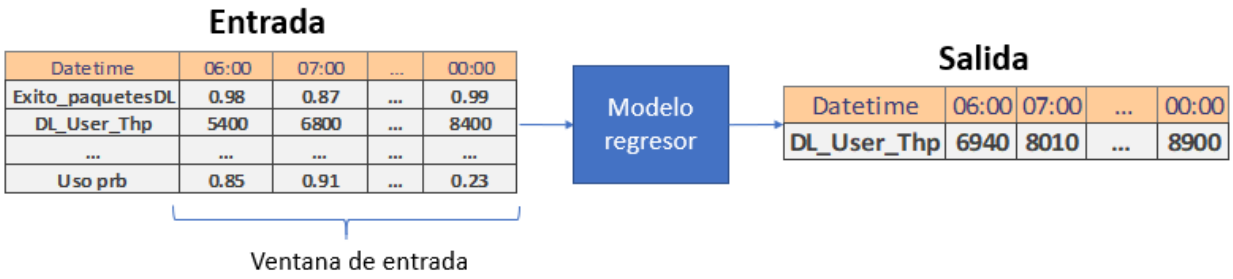


Figura 3.9: Entrada y salida del modelo regresor.

3.2.4.2. Regresión: Métricas de evaluación

La evaluación del rendimiento del algoritmo regresor es realizada de dos formas; la primera utilizando las métricas naturales de error, como por ejemplo el error absoluto y el error cuadrático medio, mientras que la segunda evaluando con tasas de acierto si la curva predicha cumple o no el criterio de bajo throughtput.

- Error: Se utiliza el error medio absoluto como herramienta de visualización, pues esta en la unidad de medida del DL User Throughtput (kbps) y por lo tanto permite dar feedback al modelo según criterio experto. Por otro lado el algoritmo utiliza el error cuadrático medio como función de optimización, ya que castiga en mayor medida el error de la curva y mejora el aprendizaje del modelo.
- Tasas de acierto: Método de evaluación que discretiza en dos estados la curva resultante utilizando el mismo criterio del modelo de clasificación, es decir si en la curva futuro del throughtput existen al menos dos puntos bajo el umbral de 3.3 Mbps, entonces toda la curva es catalogada como celda de bajo rendimiento, mientras que para el caso contrario como de buen rendimiento.

  Es así como para el método de regresión se pueden usar las mismas métricas de clasificación como lo son el accuracy, precisión y recall.

Capítulo 4

Resultados y análisis

A continuación se presentan en orden cronológico las distintas configuraciones utilizadas, partiendo por el algoritmo de regresión con el cual se probaron distintas ventanas de datos y combinaciones de KPI.

4.1. Algoritmo de regresión

4.1.1. Regresión: Primera configuración

Inicialmente se filtraron los KPI espejo, es decir, que cuantifican la misma información, utilizando el criterio experto, es así como en la primera configuración se usaron los 45 indicadores mostrados en la figura 4.1

Acc_RrcConnSetupSuccRate	Ret_ERabRetainabilityRate	DL_User_Thp
Acc_S1SigEstabSuccRate	Ret_ERabDrop	UL_User_Thp
Acc_InitialErabSetupSuccRate	Ret_ERabDropENB	DL_Cell_THROUGHPUT
Init_E_RAB_Estab_SR	Ret_ERabDrosumpMME	UL_Cell_THROUGHPUT
E_RAB_Norm_Act_RR	Active_Users_DL	DL_DRB_THROUGHPUT
SARR	Active_Users_UL	UL_DRB_THROUGHPUT
Avg_Sched_Ue_TTI_DL	Avg_Sched_Ue_TTI_UL	DL_Latency
DL_Packet_Loss_Rate	UL_Packet_Loss_Rate	Int_MacHarqDlSuccRate
Int_MacHarqUlSuccRate	Int_RlcArqDlSuccRate	Int_RlcArqUlSuccRate
Int_DlRadioThroughput	Int_UlRadioThroughput	Int_AverageDlRlcDelay
Int_AverageDlMacDelay	BestCellEvalReport	DL_TRAFFIC_VOLUME_GB
UL_TRAFFIC_VOLUME_GB	uso_prb	Exito_paquetesDL
RLC_DL_BLER	RLC_UL_BLER	MAC_DL_BLER
MAC_UL_BLER	MAC_DL_BLER_QPSK	MAC_DL_BLER_16QAM
MAC_DL_BLER_64QAM	MAC_UL_BLER_QPSK	MAC_UL_BLER_16QAM

Figura 4.1: Indicadores usados en la primera configuración del regresor.

Con ello se construyó el siguiente modelo:

- Red LSTM de una capa.
- Ventana de entrada de 7 días.
- 312 modelos unicos para cada celda del cluster 5 (Maipú sur).
- 45 Series de tiempo muestreadas cada 15 minutos.
- Entrega como salida una serie de tiempo del día 8 con un muestreo de 15 minutos.

Esta primera configuración considera la construcción de un modelo único para cada celda, por lo que el entrenamiento es realizado utilizando solo los datos una celda, como se muestra en la figura 4.2 donde para N celdas se crean N modelos distintos.

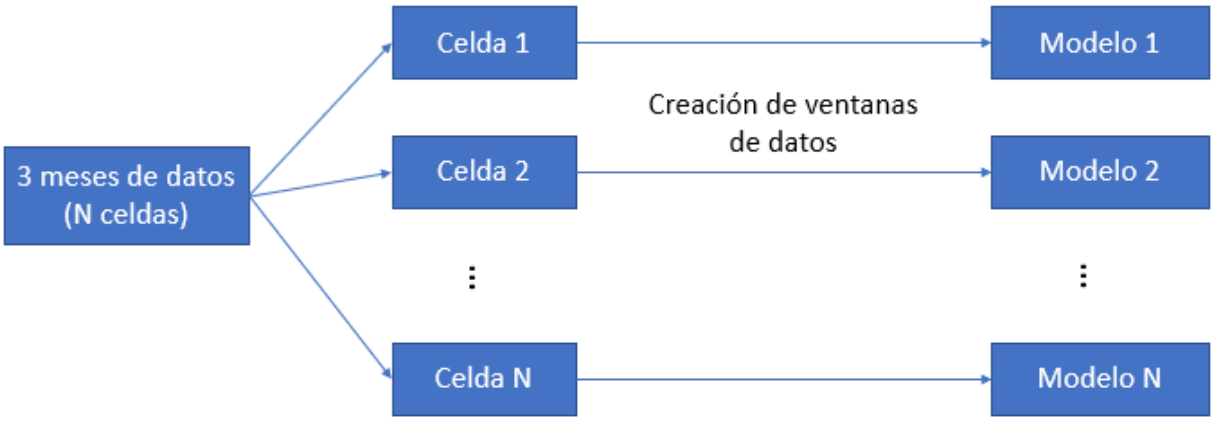


Figura 4.2: Configuración de un modelo por celda.

El entrenamiento de por si contempla una partición fija de los datos, el 60 % de la serie de tiempo se usa para entrenamiento, mientras que el 20 % de validación para evitar sobre-entrenamiento de la red neuronal, mientras que con el 20 % restante se analiza el error del modelo, como se esquematiza en la figura 4.3

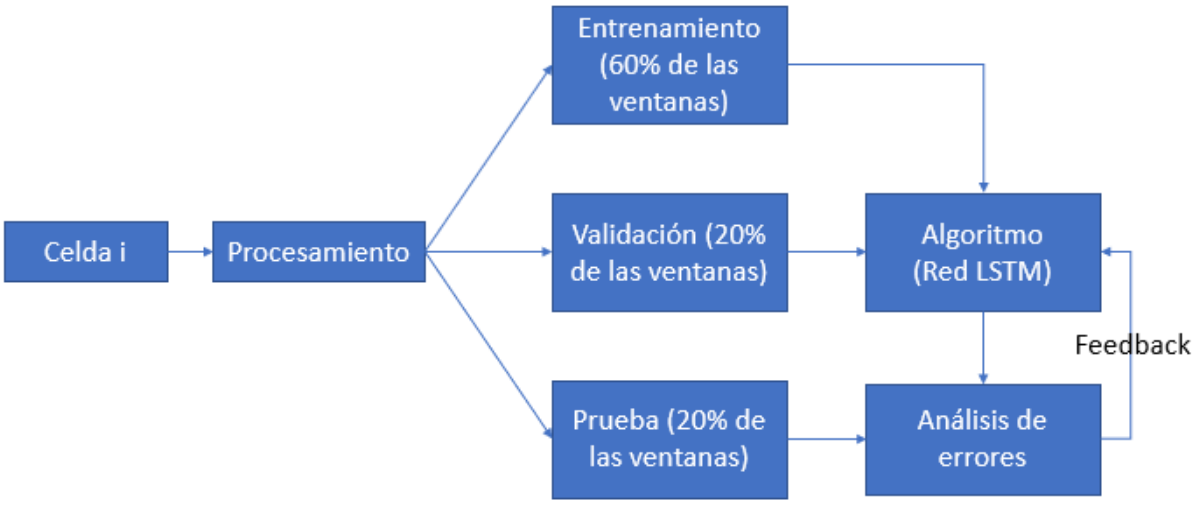


Figura 4.3: Diagrama de entrenamiento de una celda i usando la primera configuración

En la figura 4.4 se muestran de ejemplo las predicciones de un día futuro para dos celdas utilizando esta configuración.

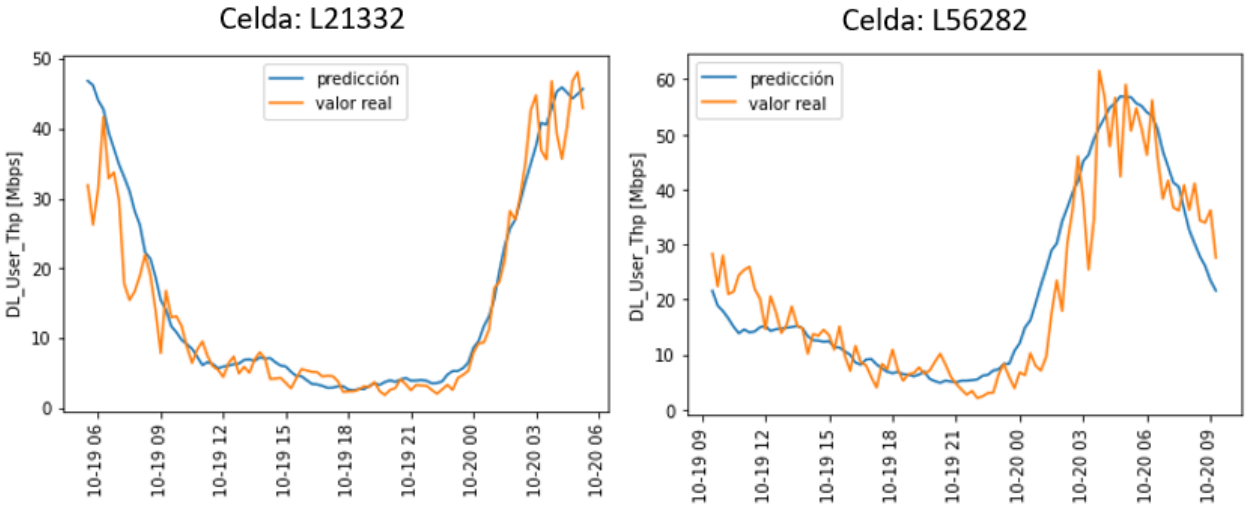


Figura 4.4: Predicción de las siguientes 24 horas para celdas L21332 y L56282.

Con esta configuración se analizó el error de la salida normalizada, de tal forma que este sea un error porcentual pues la normalización distribuye los datos entre 0 y 1; en cuanto a la métrica observada se utilizó el RMSE obteniendo un error promedio de 0.08, en la figura 4.5 se ilustra la distribución del error.

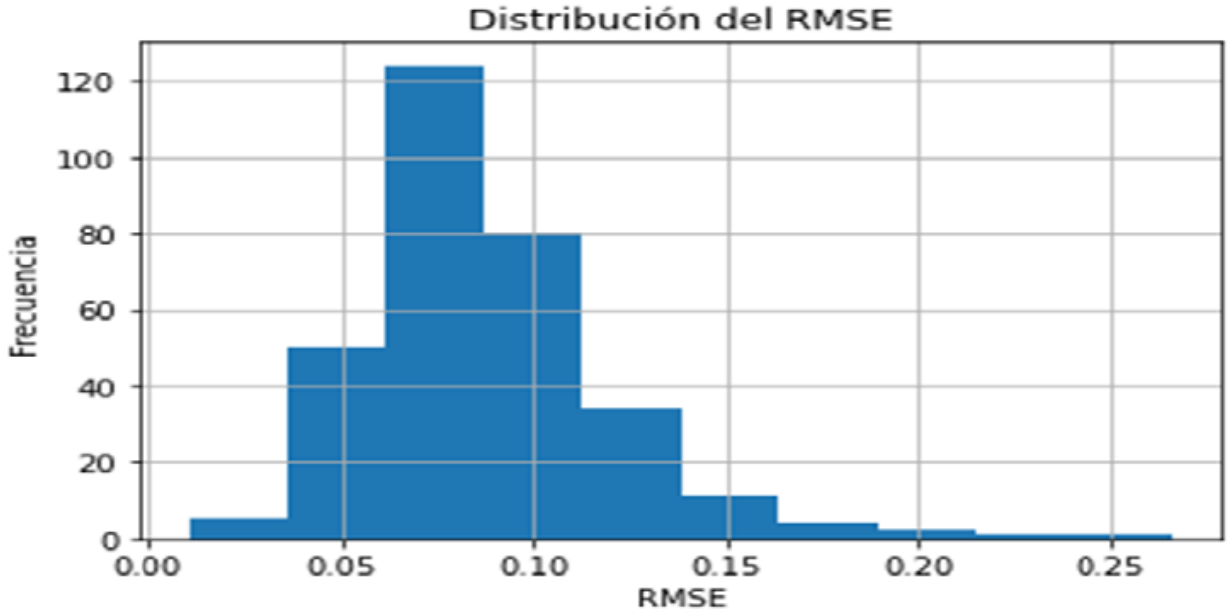
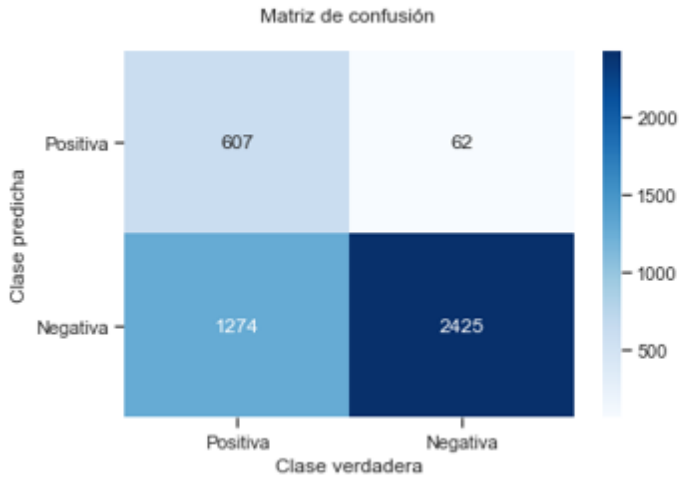


Figura 4.5: RMSE para la primera configuración.

En cuanto a las métricas de clasificación adaptadas a este modelo se obtienen los resultados ilustrados en la figura 4.6 para las 312 celdas del cluster 4 durante un periodo de 2 semanas del conjunto de prueba. Los valores obtenidos no son buenos puesto que la métrica de recall se interpreta como que solo el 32.27 % de las celdas con throughput bajo el umbral son detectadas, si bien la precisión es 90.73 % esto se debe a que el modelo tiene clases desbalanceadas y solo detecta bien las celdas con buen throughtput.



Métrica	Valor
Accuracy	69.41
Precisión	90.73
Recall	32.27

Figura 4.6: Métricas de clasificación adaptadas para la primera configuración.

Aun cuando se obtienen curvas de predicción con error muy bajo, como las mostradas anteriormente, existen casos donde la curva de throughtput futuro empeora notoriamente, como se puede ver en la figura 4.7. Entre las posibles causas esta la inestabilidad de la curva,

lo que dificulta el aprendizaje de un algoritmo ya que son fenómenos azarosos, sobre todo en el horario de la madrugada donde el throughtput llega a oscilar hasta 40 Mbps en solo minutos, también el uso de un muestreo con un periodo corto influye en la variabilidad de este KPI.

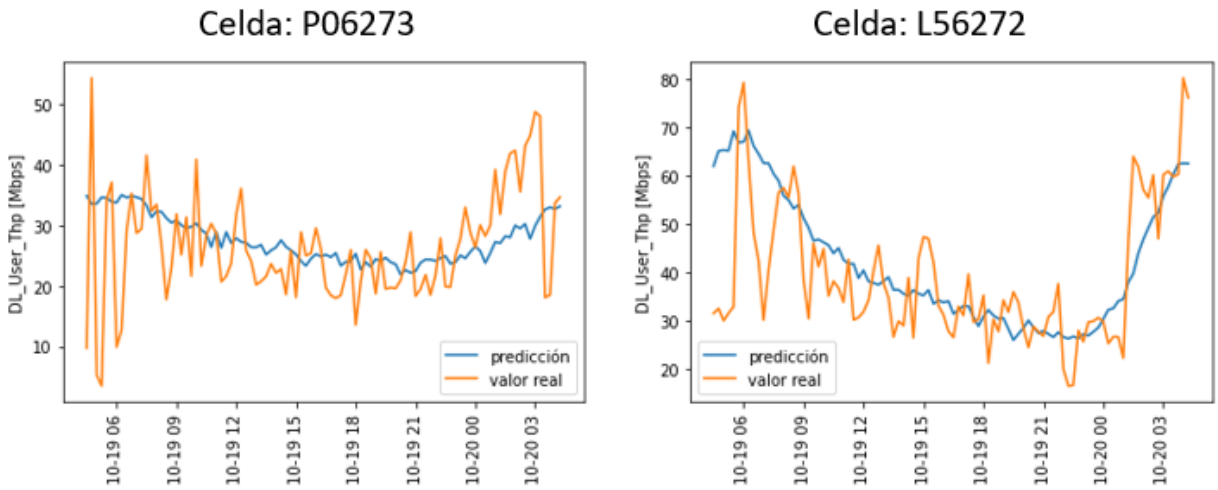


Figura 4.7: Predicción de las siguientes 24 horas para celdas.

Es posible obtener el perfil uso de recursos en función del error, como en la figura 4.8 donde se caracteriza cada predicción con su error RMSE y el coeficiente de determinación contra el uso de prb, es posible identificar como tendencia que a medida que peor es la predicción, es decir, que aumenta el RMSE o disminuye r2, menor es el uso de prb. A partir de ello se infiere que cuando una celda esta con baja carga peor es la predicción, en consecuencia esto puede traducirse en celdas con pocos usuarios conectados, justamente celdas con este perfil son las que más variabilidad tienen en la curva de throughtput, lo que explica el error de los ejemplos de la figura 4.7.

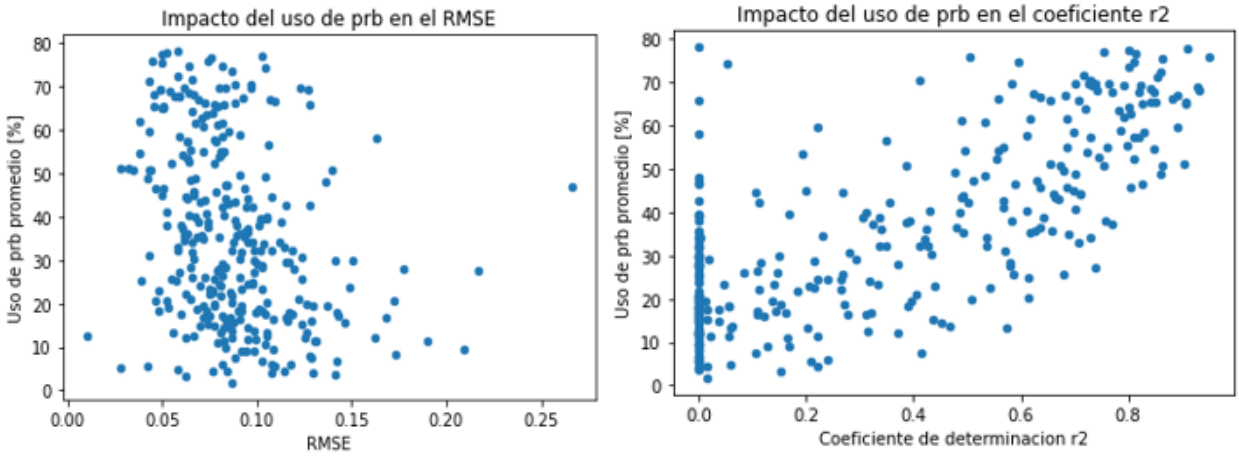


Figura 4.8: Impacto del RMSE y coeficiente r2 en el uso de prb.

Consideraciones de esta configuracion:

- Una desventaja importante es que no permite predecir el throughtput para celdas nuevas, ya que como es un modelo único por celda requiere datos pasados para el entrenamiento, debido a esto es necesario utilizar una configuración que agrupe varias celdas por modelo.
- Considerar un periodo de muestreo más largo para obtener curvas estables y en consecuencia facilitar el aprendizaje del algoritmo.
- No es una buena practica medir el error del throughput normalizado entre 0 y 1, pues la presencia de outliers puede distorsionar la distribución como en el siguiente ejemplo: El 1 % de una ventana con un máximo de 60 Mbps no es igual que el 1 % de una ventana con un máximo de 5 Mbps.

4.1.2. Regresión: Segunda configuración

Para la segunda configuración se usaron inicialmente el mismo conjunto de KPI, pero aplicando como estrategia eliminar KPI y observar si el rendimiento del algoritmo disminuía se filtraron mas indicadores aun (método wrapper), además de un análisis de correlación con el indicador DL User Throughtput. En la figura 4.9 se muestran los KPI finales junto a la correlación con el Throughtput, donde fueron agregados como variables de entrada la hora y el día de la semana.

KPI	Correlación
DL_User_Thp	1
UL_User_Thp	0.328
Active_Users_DL	-0.394
DL_Latency	-0.263
DL_Cell_THROUGHPUT	0.24
Active_Users_UL	-0.361
DL_TRAFFIC_VOLUME_GB	-0.313
UL_TRAFFIC_VOLUME_GB	-0.094
uso_prb	-0.603
Exito_paquetesDL	0.183
día	0.002
hora	-0.388

Figura 4.9: Indicadores usados en la segunda configuración del regresor.

Las características del modelo son las siguientes:

- Red neuronal compuesta por una capa LSTM y dos capas ocultas.

- Ventana de entrada de 7 días.
- Ventana de entrada considera solo datos entre las 06:00 y 00:00.
- Muestreo de los datos por hora y cada 15 minutos (dos pruebas distintas).
- Un modelo único por banda de frecuencia.
- 12 Series de tiempo de entrada.
- Entrega como salida una serie de tiempo del día 8 con un muestreo de 15 minutos o 1 hora según corresponda.

La segunda configuración contempla la construcción de 3 submodelos debido a el análisis llevado a cabo anteriormente, por lo tanto para un conjunto de N celdas se crean 3 modelos, como se muestra en la figura 4.10

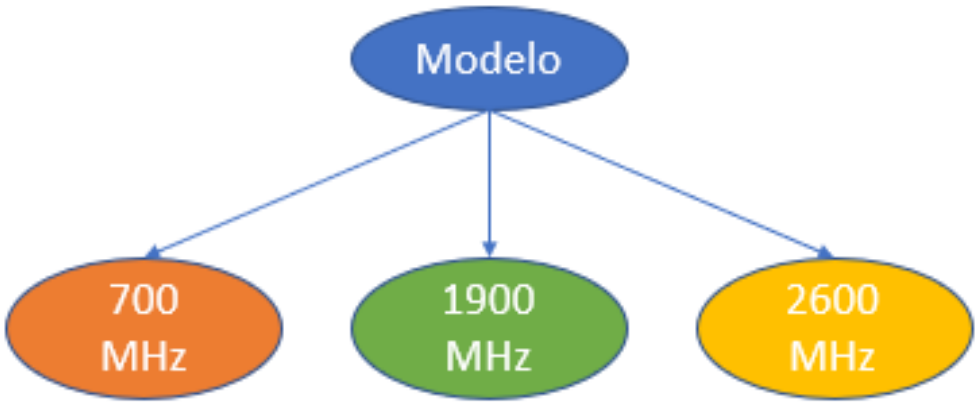


Figura 4.10: Configuración de un modelo por banda.

En cuanto a la partición de entrenamiento y prueba se mantiene la misma configuración, es decir, 60 % de las muestras son utilizadas como entrenamiento, 20 % como validación y el ultimo 20 % como prueba., con ello se obtienen los siguientes resultados:

4.1.2.1. Muestreo cada 15 minutos

Para esta configuración se estudio el RMSE no normalizado (en Mbps) y se obtuvo que el error aumenta a medida que aumenta la frecuencia de la banda, como se puede ver en la figura 4.11, además el promedio del RMSE es cercano a 3 Mbps lo que es alto si se considera que el umbral es 3.3 Mbps, es importante considerar que el error obtenido contempla tanto celdas que bordean 3 Mbps como otras que oscilan alrededor de 50 Mbps, para un análisis más detallado es necesario visualizar aparte el error de las celdas con bajo throughtput.

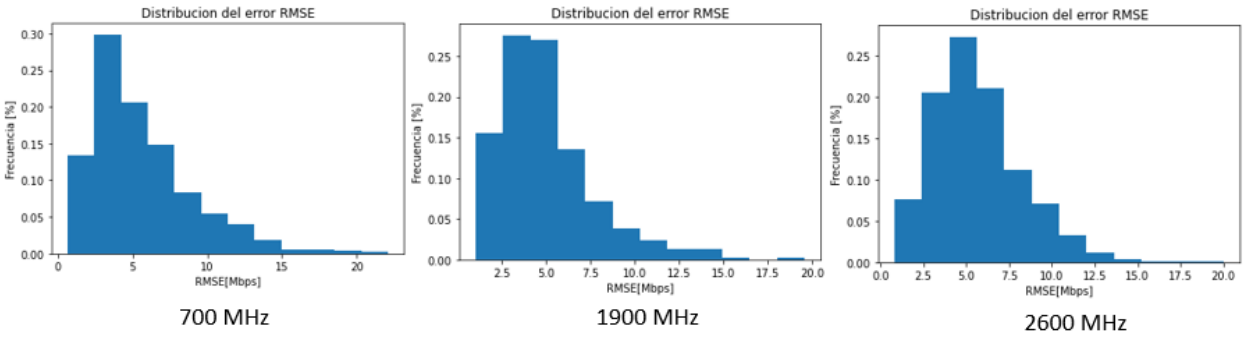


Figura 4.11: RMSE para la segunda configuración.

Por otro lado las métricas de clasificación mejoran notablemente al ser comparadas con la primera configuración, ver figura 4.12, sobre todo el accuracy que llega a valores superiores al 90 %, si bien el recall también aumenta, aun sigue dando valores bajos, puesto que en el peor caso (banda de 2600 MHz) solo logra detectar el 46 % de las celdas con throughtput bajo umbral.

Métrica	Banda 700	Banda 1900	Banda 2600
Accuracy	0.9403	0.9533	0.966
Precisión	0.8229	0.8188	0.8166
Recall	0.6038	0.5288	0.4634

Figura 4.12: Métricas de clasificación adaptadas para la segunda configuración (15 minutos).

En general mejoran los resultados producto de la segmentación por bandas pero siguen sujetos a mejoras, como aumentar el periodo de muestreo a 1 hora.

4.1.2.2. Muestreo cada 1 hora

Considerando los resultados de la configuración anterior se optó por agregar la distribución del error para muestras etiquetadas con throughtput bajo el umbral por separado, como se muestra en las figuras 4.13, 4.14 y 4.15 para cada banda respectiva. Se observa que el error promedio de celdas con bajo throughtput disminuye, puesto que pasa de 3 Mbps a un promedio entre entre 0 y 1 Mbps para las 3 bandas.

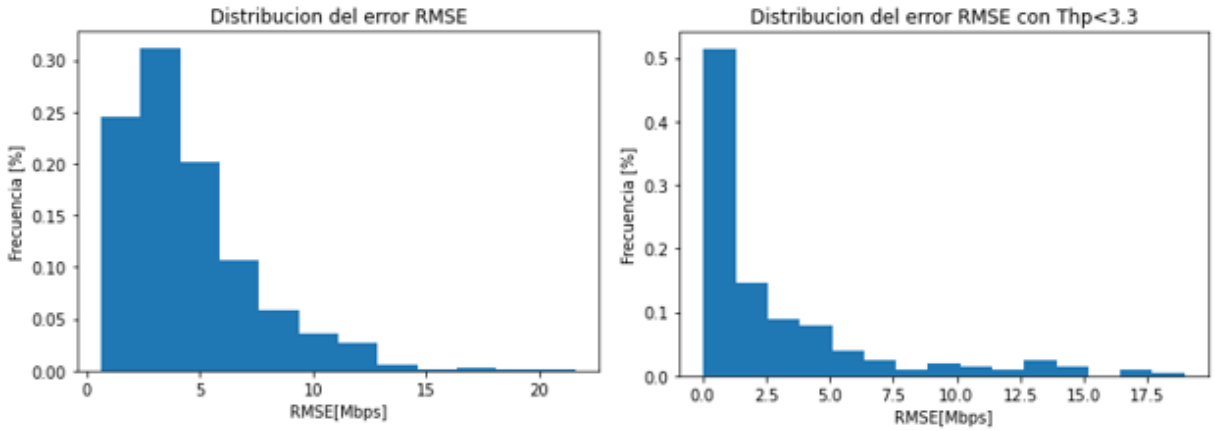


Figura 4.13: RMSE para la segunda configuración (700 MHz).

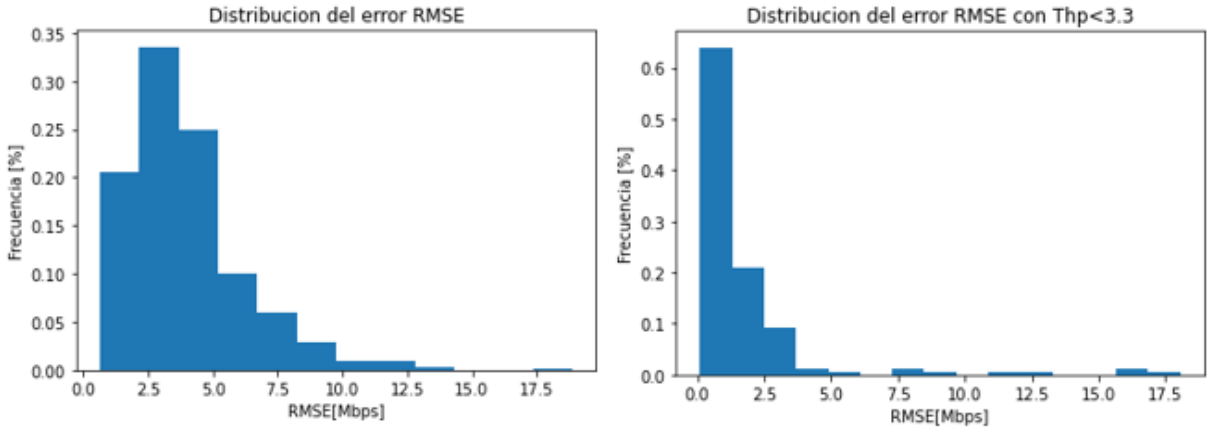


Figura 4.14: RMSE para la segunda configuración (1900 MHz).

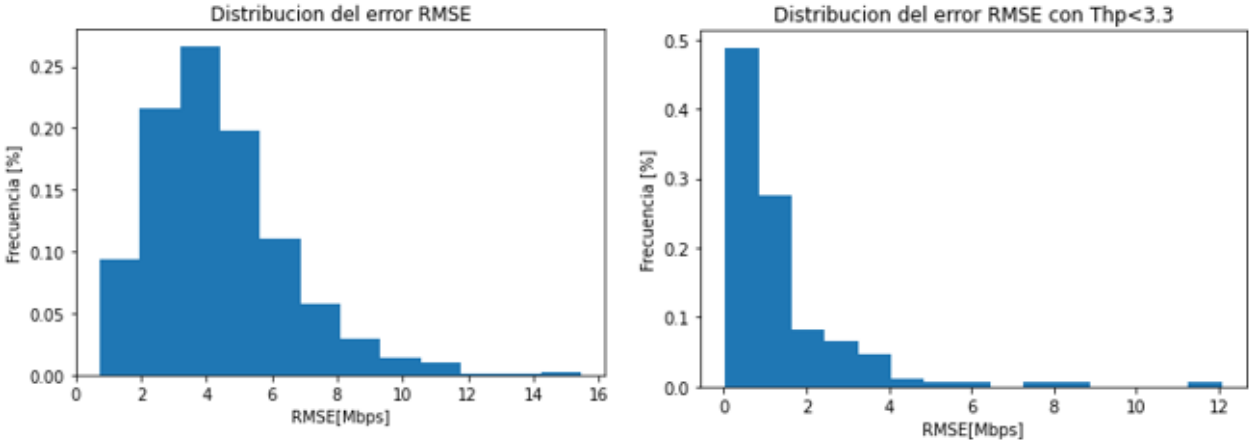


Figura 4.15: RMSE para la segunda configuración (2600 MHz).

Las métricas de clasificación adaptadas se muestran segmentadas por banda en la figura 4.16 donde es visible el efecto de cambiar el muestreo a 1 hora, sobre todo para la banda de 2600 MHz ya que el recall aumenta en un 25 %

Métrica	Banda 700	Banda 1900	Banda 2600
Accuracy	0.9107	0.9486	0.9483
Precisión	0.9191	0.9579	0.8655
Recall	0.6107	0.6993	0.7152

Figura 4.16: Métricas de clasificación adaptadas para la segunda configuración (1 hora).

En cuanto al impacto de la cantidad de usuarios sobre el error se encontró que existe una relación donde a menor cantidad de usuarios mayor es el error y vice versa, como se evidencia en los gráficos segmentados de la figura 4.17.

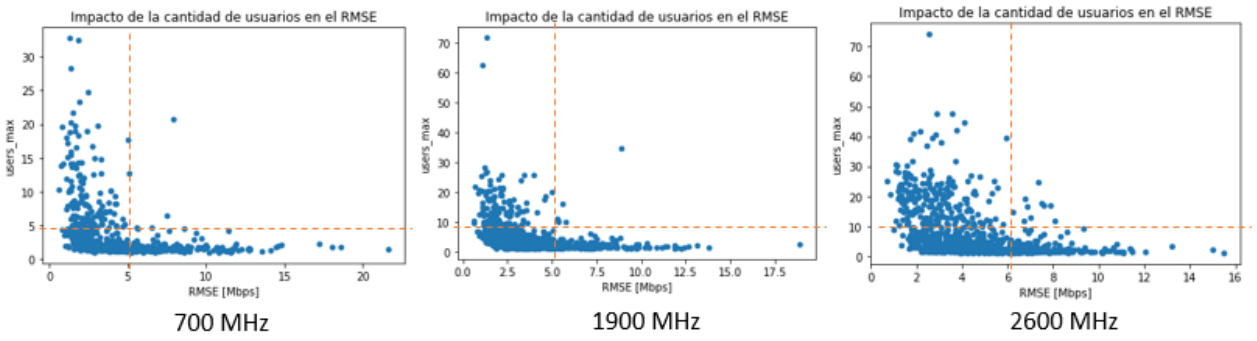


Figura 4.17: Impacto de la cantidad de usuarios en el error (1 hora).

De la misma forma se observó que celdas sobrecargadas (uso de prb cercano al 100 %) tienen menos error que celdas con baja carga, como se muestra en la figura 4.18, tanto este fenómeno como el anterior sugieren como mejora segmentar el modelo no solo por banda, sino que también por uso de recursos.

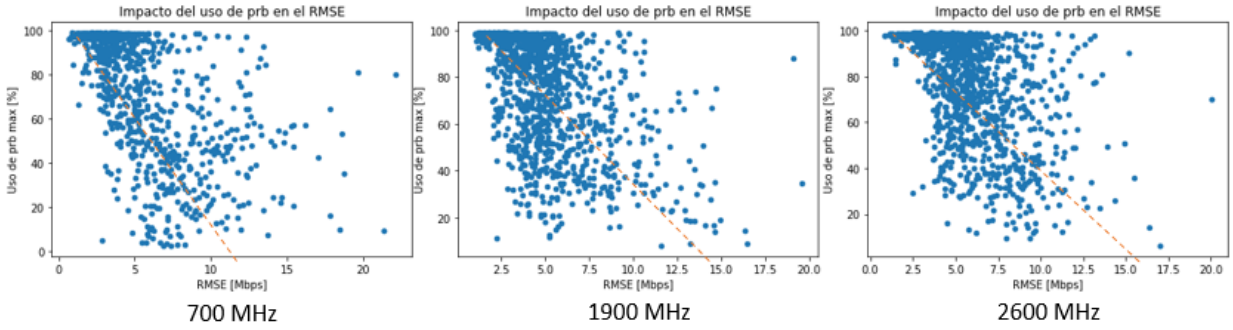


Figura 4.18: Impacto del uso de prb máximo en el error (1 hora).

Entre las observaciones generales de la segunda configuración están las siguientes:

- Muestreo de 1 hora tiene mejores resultados que muestreo de 15 minutos.
- Segmentación por banda mejora considerablemente los resultados, como fue observado en las métricas de clasificación.

- Posible mejora: Segmentar aún más el modelo según uso de prb.

4.1.3. Regresión: Tercera configuración

Para la tercera configuración la principal diferencia es la adaptación de una nueva segmentación, esta vez según uso de prb máximo, obteniendo así 9 submodelos como es esquematizado en la figura 4.19. En cuanto a la estructura de datos se mantiene solo el muestreo cada una hora a razón del análisis llevado a cabo anteriormente, a continuación se muestra un listado con las características:

- Red neuronal compuesta por una capa LSTM y dos capas ocultas.
- Ventana de entrada de 7 días.
- Ventana de entrada considera solo datos entre las 06:00 y 00:00.
- Muestreo de los datos por hora.
- Un modelo único por banda de frecuencia y rango de uso de prb máximo.
- 12 Series de tiempo de entrada, estos se muestran en detalle en la figura 4.20.
- Entrega como salida una serie de tiempo del día 8 con un muestreo de 15 minutos o 1 hora según corresponda.

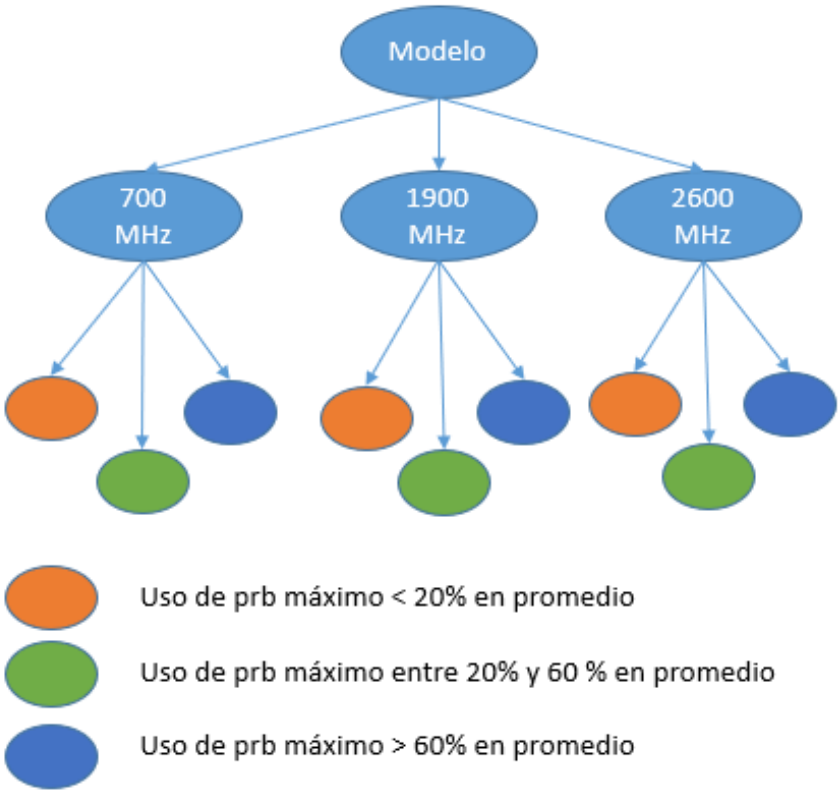


Figura 4.19: Configuración de un modelo por banda y por rango de uso de prb máximo.

DL_User_Thp	UL_User_Thp	Active_Users_UL
Dl_Cell_Throughtput	Ul_Cell_Throughtput	Active_Users_DL
DL_Latency	DL_Traffic_Volume	UL_Traffic_Volume
Uso_prb	Éxito_paquetesDL	hora

Figura 4.20: Indicadores usados en la tercera configuración.

Los resultados de las métricas de clasificación mejoran levemente en comparación a los resultados anteriores, como se muestra en la figura 4.21, además se observa que las tasas de acierto aumentan con la frecuencia de la banda, obteniendo los mejores resultados en la banda de 2600 MHz con un 76.38 % de aciertos para celdas bajo el umbral.

Métrica	Banda 700	Banda 1900	Banda 2600
Accuracy	0.9171	0.9476	0.9465
Precisión	0.9278	0.9426	0.8148
Recall	0.604	0.7055	0.7638

Figura 4.21: Métricas de clasificación adaptadas para la tercera configuración.

Anteriormente se usó el RMSE como métrica de error, para el caso de esta configuración se cambió por el error absoluto con el objetivo de que la métrica sea lineal, recordar que el RMSE pondera cuadráticamente y por lo tanto puntos de la curva con mayor error tienen más peso sobre la métrica.

En las figuras 4.22, 4.23 y 4.24 se muestra la distribución del error absoluto y el valor acumulado en celdas con throughtput bajo el umbral, para las 3 bandas el 50 % de las celdas bajo umbral tienen un error menor a 500 kbps lo que es una mejora considerable con respecto a las dos configuraciones anteriores. Los gráficos son concordantes con las métricas de clasificación porque la banda con menor error es la que presentó el mejor recall (Banda de 2600 MHz).

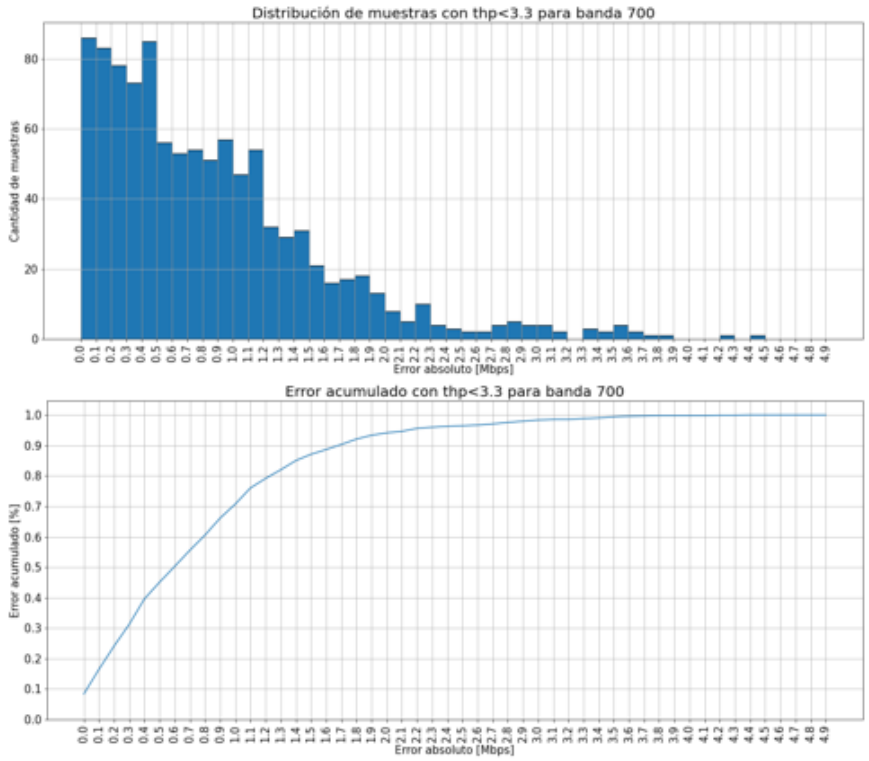


Figura 4.22: Error absoluto de ventanas con Thp $<$ 3.3 Mbps (700 MHz).

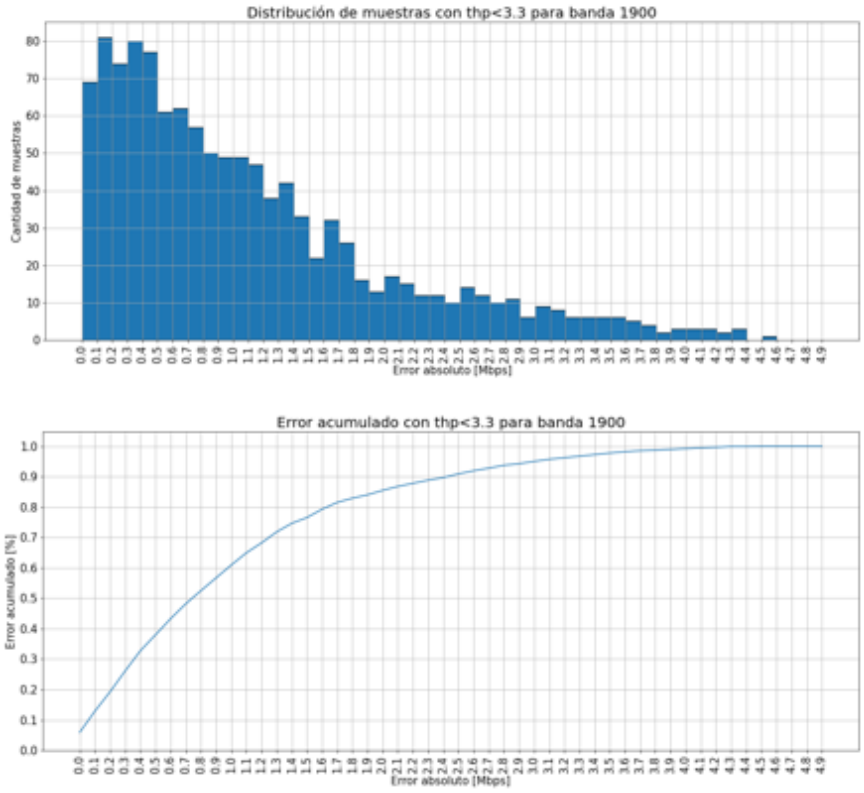


Figura 4.23: Error absoluto de ventanas con Thp $<$ 3.3 Mbps (1900 MHz).

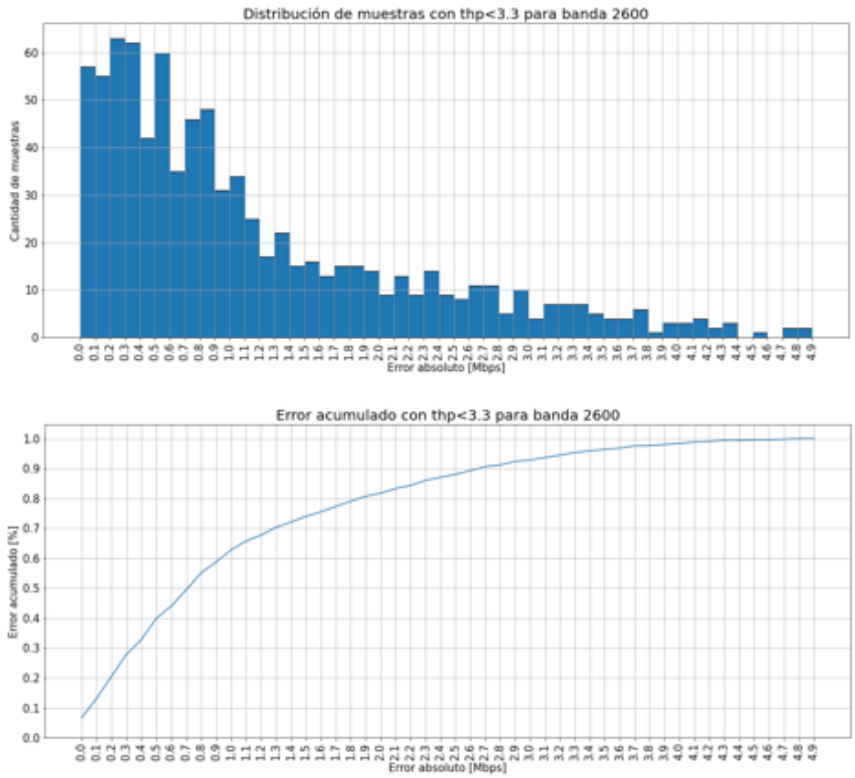


Figura 4.24: Error absoluto de ventanas con Thp $<$ 3.3 Mbps (2600 MHz).

Las celdas con throughtput sobre el umbral fueron analizadas de forma separada y se muestran en la figura 4.25 los resultados, para este caso se utilizó el error relativo para que celdas con distinto throughtput promedio sean comparables, de esta forma se cuantifican equitativamente tanto las celdas que bordean 5 Mbps como las que oscilan en 60 Mbps. Se obtiene que el 50 % de las celdas sobre el umbral tienen un error relativo menor al 15 %.

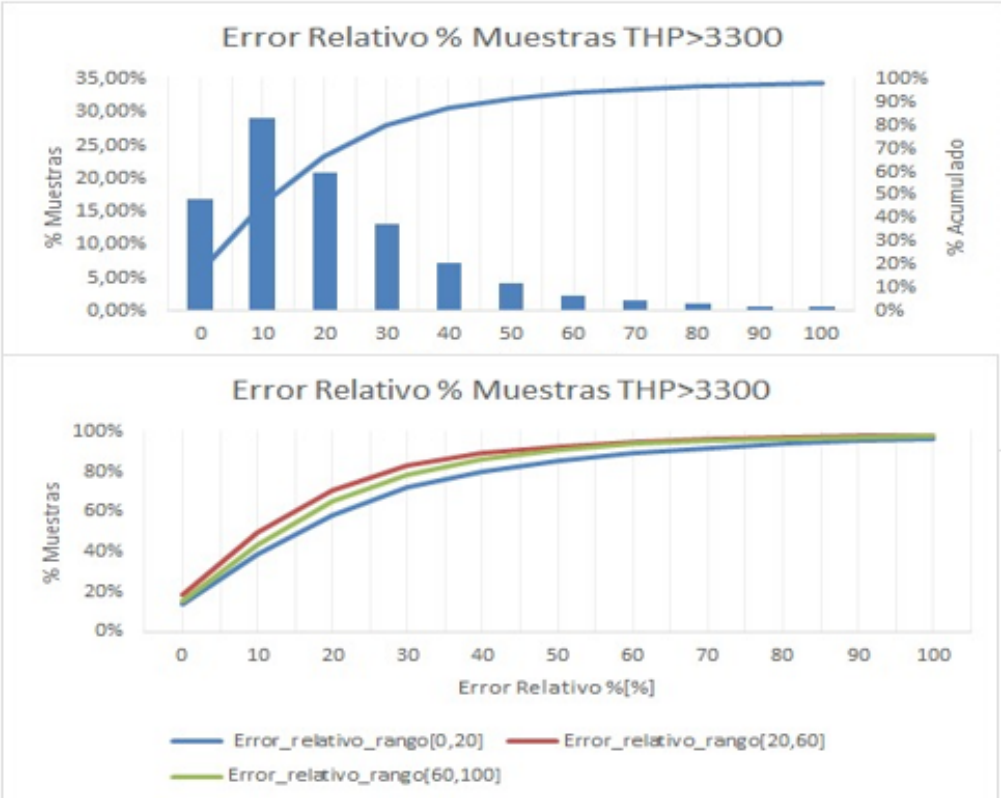


Figura 4.25: Error absoluto en ventanas con Thp $>$ 3.3 Mbps (2600 MHz).

Entre las observaciones generales de la tercera configuración están las siguientes:

- Segmentación por el uso de prb y omitir muestras en horario nocturno mejora el rendimiento del algoritmo.
- Métricas de clasificación aseguran que al menos el 70 % de las celdas bajo el umbral son detectadas de forma efectiva.
- El error de celdas bajo el umbral es en promedio 500 kbps, lo que es aceptable pero no suficiente.
- El error de celdas sobre el umbral es suficiente para asegurar pocos falsos positivos en la predicción.
- Se pueden hacer mejoras al modelo aplicando cambios avanzados en la arquitectura de la red neuronal, pero esto escapa del foco de la memoria.

4.2. Algoritmo de clasificación

Las pruebas del modelo de clasificación consideran dos algoritmos, estos son random forest y gradient boosting, las características en ambos casos son similares al regresor y son las siguientes:

- Ventana de entrada de 7 días considerando datos entre las 06:00 y 00:00.

- La entrada son los estadísticos calculados sobre 12 KPI seleccionados.
- Muestreo de datos por hora.
- Modelo único por banda.
- Entrega salida binaria con el estado de si una celda tiene el throughtput bajo el umbral de 3,3 Mbps.
- Uso de dos algoritmos distintos para comparar resultados, estos son Random Forest y Gradient Boosting.

Para el modelo de clasificación se trabaja con los KPI ya filtrados en las configuraciones anteriores, como ya se mostró en la figura 4.20.

4.2.1. Clasificación: Random Forest

Se crea un modelo preliminar de clasificación considerando una partición fija de 80 % entrenamiento y 20 % prueba del conjunto de muestras, con ello se obtienen la matriz de confusión por banda mostradas en la figura 4.29.

Como se puede ver la cantidad de verdaderos negativos es considerablemente mayor a la de verdaderos positivos, es decir, la cantidad de celdas en buen estado son mayoría, en general en las bandas de 700 MHz y 2600 MHz se detectan la mayoria de las celdas de mal rendimiento, por otro lado en la banda de 1900 MHz la mitad de estos casos no son detectados, siendo asi la de peor rendimiento.

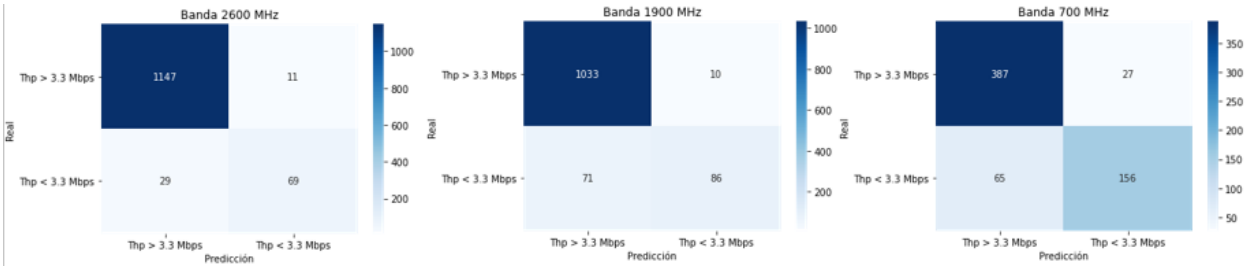


Figura 4.26: Matriz de confusión por banda utilizando random forest.

A partir del resultado anterior se obtiene la tabla de la figura 4.30 con las métricas para la partición fija, se refleja lo visto en las matrices ya que la banda de 1900 tiene un recall de solo 55 %, mientras que el resto de bandas superan el 70 %. Es importante destacar que el uso de una partición fija puede sesgar los resultados ya que el conjunto de prueba esta sujeto a anomalías, por lo tanto el siguiente análisis es verificar el resultado variando la partición con Kfolds.

Banda	F1	Accuracy	Precisión	Recall	Especificidad
2600 MHz	0,77	0,97	0,86	0,70	0,99
1900 MHz	0,67	0,93	0,89	0,55	0,99
700 MHz	0,77	0,86	0,85	0,71	0,93

Figura 4.27: Métricas para partición fija utilizando random forest.

Utilizando Kfolds para validar el error se obtienen las métricas de la figura 4.28, dando así un resultado generalizado, a diferencia de la partición fija. En cuanto a los valores se logra observar que todas las métricas tienen superan 0.8, con especial importancia el recall donde en el mejor de los casos no detecta el 13 % de las celdas bajo el umbral de throughtput. Considerando que las clases son desbalanceadas el indicador F1 evidencia un buen rendimiento del clasificador debido a ser mayor a 0.8 para todas las bandas, lo que no se logró visualizar en la partición fija.

Banda	Accuracy	Precisión	Recall	F1
2600 MHz	0.95	0.91	0.85	0.88
1900 MHz	0.94	0.91	0.87	0.87
700 MHz	0.89	0.9	0.8	0.83

Figura 4.28: Métricas para random forest usando Kfolds = 5.

4.2.2. Clasificación: Gradient Boosting

Al mantener la misma estructura salvo el cambio de random forest por gradient boosting se observa una leve mejora en todas las métricas salvo la precisión para la partición fija, como se ve en la figura 4.30. En este caso la banda de 1900 MHz llega a 76 %, mientras que 700 y 2600 MHz superan el 80 %, el resto de métricas también mejoran, se destaca que en el accuracy ninguna banda esta por debajo del 90 %.

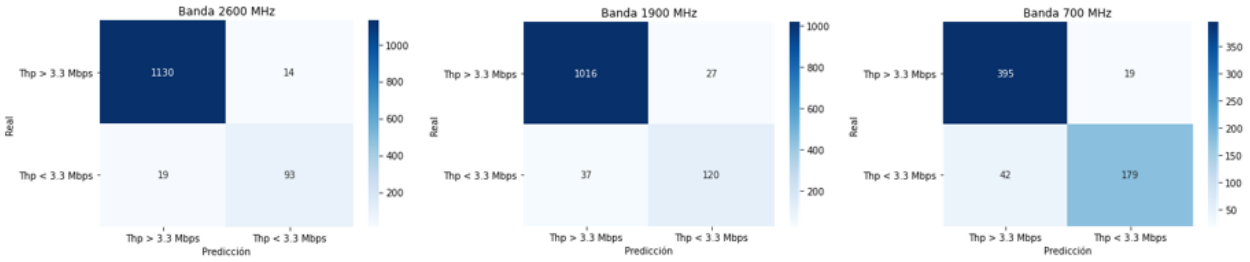


Figura 4.29: Matriz de confusión por banda utilizando gradient boosting.

Banda	F1	Accuracy	Precisión	Recall	Especificidad
2600 MHz	0,85	0,97	0,87	0,83	0,98
1900 MHz	0,79	0,94	0,81	0,76	0,97
700 MHz	0,85	0,90	0,90	0,80	0,95

Figura 4.30: Matriz de confusión por banda utilizando gradient boosting.

A pesar del buen resultado de la partición fija aun es necesario realizar un experimento generalizado, por lo tanto se utiliza nuevamente kfolds con lo que se obtiene la tabla de la figura 4.31. En este caso el recall mejora considerablemente, sobre todo para la banda de

1900 MHz donde no detecta solo el 10 % de los casos con mal rendimiento, también aumenta el F1 en todas las bandas resultando en valores superiores a 0.8 lo que da seguridad de la calidad del clasificador ya que categoriza entre clases desbalanceadas.

Banda	Accuracy	Precisión	Recall	F1
2600 MHz	0.93	0.87	0.85	0.86
1900 MHz	0.95	0.92	0.9	0.91
700 MHz	0.91	0.89	0.85	0.87

Figura 4.31: Métricas para gradient boosting usando Kfolds = 5.

A partir de los experimentos realizados con random forest y gradient boosting se determina que el mejor algoritmo para el uso practico de la herramienta corresponde al segundo, ya que en todas las métricas resulta ser superior, sobre todo en el recall cuya importancia es determinante a la hora de definir si el clasificador funciona bien, pues la detección de celdas con mal rendimiento es clave. Por lo tanto para la prueba final sobre el cluster 4 se opta por el uso de gradient boosting.

4.3. Prueba final

En función de los resultados obtenidos en esta sección se muestran a continuación las pruebas del modelo final sobre el cluster 4 utilizando las configuraciones óptimas tanto en el bloque de regresión como de clasificación. Se muestra paso a paso como se plantea utilizar el modelo una vez pase a fase de producción. La prueba es realizada el dia 18 de Diciembre de 2020 para la totalidad del cluster Maipú norte- Padre Hurtado con un total de 438 celdas LTE.

4.3.1. Primera fase: Clasificación

El primer bloque entrega la clasificación entre ambas clases para toda celda, es decir si corresponde a una predicción con buen o mal rendimiento a partir del criterio de throughtput menor a 3.3 Mbps. El resultado se muestra en la matriz de confusión de la figura 4.35, se observa que de un total de 48 celdas con mal rendimiento se detectaron 42, mientras que 3 celdas fueron etiquetadas con mal rendimiento cuando este no era el caso.

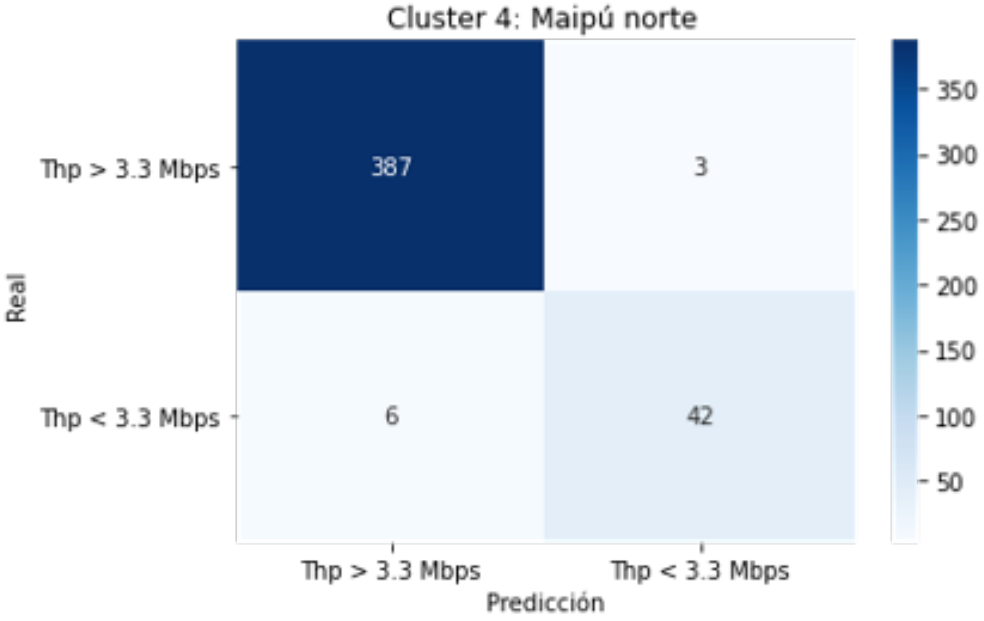


Figura 4.32: Matriz de confusión en el cluster 4 para el día 18 de Diciembre.

En primer lugar como resultado del algoritmo de clasificación se obtiene una tabla para cada celda con el estado del throughput futuro, como se muestra en el ejemplo de la figura **??** con 5 celdas donde se puede ver que la celda L24621 tiene una predicción bajo el umbral de 3.3 Mbps. De la misma forma se muestra el siguiente procedimiento donde para cada celda con predicción positiva (bajo umbral) se calcula el ranking de las celdas más aptas para el balanceo de usuarios.

En forma más detallada las métricas son las que se muestran a continuación:

- Accuracy: 0.98
- Precisión: 0.93
- Recall: 0.88
- F1: 0.90

Se logra reflejar el buen rendimiento del clasificador, solo el 8 % de las celdas con mal rendimiento no se detectaron, mientras que el 98 % del total fue clasificado correctamente.

4.3.2. Segunda fase: Filtro y selección de candidatos

La segunda fase consiste en identificar cuales son las celdas vecinas mas aptas para el traspaso de usuarios, utilizando la función implementada se obtiene el conjunto mostrado en la figura 4.33, donde las celdas rojas son de mal rendimiento y las verdes las celdas vecinas aptas para el balanceo de carga.

Figura 4.33: Cluster 4 con celdas de mal rendimiento (rojo) y celdas candidatas (verde).

Una vez se filtra el top 3 de las mejores celdas vecinas se pasa al modulo de regresión, en la figura 4.35 se ilustra la predicción de throughtput para las 3 mejores celdas a lo largo del día, finalmente queda a criterio del usuario de la aplicación la elección de la mejor celda de balanceo, en este mismo ejemplo es preferente balancear con la celda L58711 pues es la que tiene el throughtput mas alto durante todo el día y por lo tanto no corre el riesgo de caer bajo el umbral tras recibir nuevos usuarios.

4.3.3. Tercera fase: Regresión

La última fase contempla el uso del algoritmo regresor para obtener las curvas de throughtput para las 3 celdas candidatas de cada celda con mal rendimiento, de esta forma se muestra en detalle al usuario de la aplicación cual es las más apta para el traspasado de carga.

Para ejemplificar este bloque se utiliza el caso de la celda L24623 en particular, como se muestra en la figura 4.34 junto a los 3 vecinos más aptos.

Celda	Predicción
L42618	0
L24623	1
...	...
L44083	0
M44085	0

Ranking	Celda bajo umbral	Celda candidata
1	L24623	L24618
2	L24623	L44083
3	L24623	M44085

Figura 4.34: Caso particular de celda L24623.

Los tres vecinos más aptos para la celda L23632 se ordenan según el ranking de la función hecha en el bloque de filtro, finalmente el bloque entrega la regresión del siguiente día para cada una, como se muestra en la figura 4.35. Queda a criterio de usuario la elección de la mejor celda, un posible criterio de selección es la celda con mayor throughtput durante el día, como por ejemplo la celda L42618.

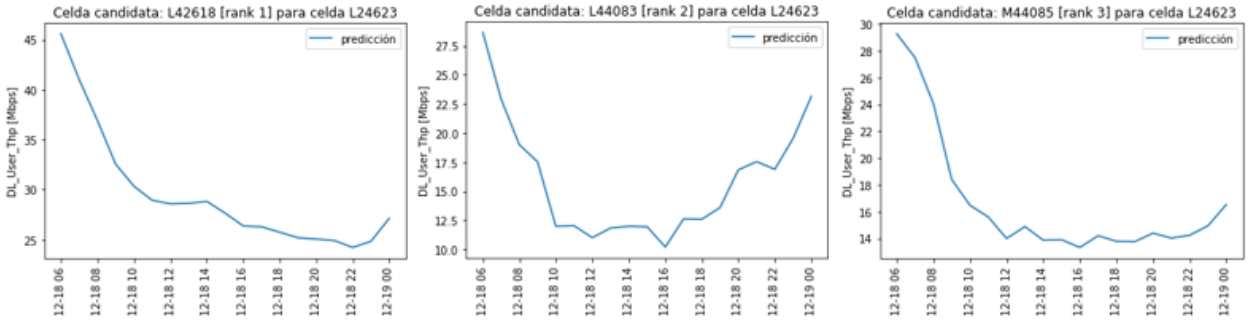


Figura 4.35: Ejemplo de salida del regresor para las 3 mejores celdas vecinas.

Capítulo 5

Conclusiones

Debido al actual crecimiento y aumento en la complejidad de las redes de telefonía móvil, cada vez se hace más difícil el monitoreo y optimización de estas para mantener una calidad de servicio acorde al mercado. Es por esto que empresas como Entel están en busca de soluciones para automatizar procesos y dejar de lado el análisis descriptivo, reemplazándolo por análisis predictivo. Esto es posible utilizando nuevas tecnologías desarrolladas actualmente basadas en inteligencia artificiales, específicamente aprendizaje de máquinas.

Es por esto que el objetivo general de este trabajo consiste en estudiar la factibilidad de una herramienta práctica para el control preventivo, que determine el estado de celdas LTE según el tráfico de red. Efectivamente es factible contar con un modelo que facilite las tareas de análisis predictivo utilizando los datos de la red: El modelo implementado cumple el objetivo propuesto, el cual se desarrolló en base a tres objetivos específicos.

- El modelo logra predecir el comportamiento del throughput en base al historial de indicadores de rendimiento.
- El modelo logra identificar con alta precisión las celdas con buen y mal rendimiento de throughtput utilizando el umbral mínimo aceptado por Entel de 3.3 Mbps.
- El modelo actualmente permite aplicar cambios de parámetro para llevar a cabo el proceso de optimización de la red en base al traspaso de usuarios entre celdas de bajo a buen rendimiento.

De forma especifica en el peor caso el bloque clasificador del modelo logra detectar el 85 % de las celdas con rendimiento critico, mientras que en el mejor caso detecta el 90 % de estas. En cuanto a la totalidad de las celdas, es decir con buen y mal rendimiento, se obtiene una tasa de acierto sobre el 91 % para las 3 bandas de frecuencia LTE.

El bloque de filtro de celdas aptas para el traspaso de usuarios logra identificar correctamente las 3 mejores celdas candidatas para cada celda de bajo rendimiento según la predicción obtenida.

El bloque regresor logra entregar una predicción para celdas de bajo rendimiento con un error absoluto promedio de 0.5 Mbps para las 3 bandas LTE, mientras que para celdas de buen rendimiento este se mide como error relativo y es de 15 % en promedio. En ambos casos es una medida aceptable, pero queda sujeta a mejoras en el futuro.

Los tres bloque son probados integralmente en una prueba real sobre el cluster 4 para un día específico. Se logró detectar el 88 % de las celdas con mal rendimiento, mientras que la tasa de aciertos para cualquier tipo de celda fue de 98 %. En cuanto al bloque de filtro logró encontrar para cada una de las 42 celdas detectadas las 3 mejores celdas candidatas. Finalmente el bloque regresor entrega con bajo error la curva horaria de cada celda candidata para dejar a criterio del usuario la celda mas apta para el traspaso de carga.

Si bien el modelo ayuda al proceso de optimización y automatización de la red, el aporte diferencial de este trabajo de titulo yace en la integración de conocimientos de inteligencia artificial en la industria de las telecomunicaciones, pues es una de las primeras aplicaciones desarrolladas dentro del área de optimización de Entel.

Dentro de las posibles mejoras y trabajos futuro está la optimización del algoritmo regresor para minimizar el error absoluto de las curvas; y la puesta a prueba del modelo con el objetivo de crear una base de datos con los cambios de parámetro realizados, para así poder implementar un modelo que recomiende el mejor parámetro, automatizando aun más el proceso de optimización.

Finalmente este trabajo de titulo logra cumplir a cabalidad todos los objetivos propuestos en un comienzo, mas aun actualmente este modelo esta en modo de prueba dentro de la empresa, dando un valor agregado al análisis de datos y optimización de la red, por lo que cumple con las expectativas de la empresa.

5.1. Trabajo futuro

Para el trabajo llevado a cabo en esta memoria se proponen mejoras para aumentar el rendimiento del algoritmo clasificador y regresor, algunas de estas son:

- Mejorar la selección de atributos con nuevos algoritmos, como por ejemplo RRelief.
- Probar configuraciones y arquitecturas más complejas con redes neuronales.
- Trabajar con muestreo diario, específicamente con datos en hora peak.

Es importante mencionar que esta memoria consistió en una primera etapa del proyecto, a continuación se muestra la planificación general:

- Etapa I: Predicción de throughtput.
- Etapa II: Estimador de cambio de parámetros.
- Etapa III: Automatizacion.

El siguiente paso (Etapa II) es el diseño e implementación de un algoritmo que identifique la parametrización óptima para cada una de las celdas clasificadas con mal rendimiento, para ello se plantea aplicar pruebas manuales sobre la red y registrar el efecto de cada cambio, esto con la finalidad de que un algoritmo pueda aprender la mejor estrategia de cambios considerando el contexto de cada celda LTE, finalmente se propone automatizar el proceso y aplicarlo en toda la red (Etapa III).

Bibliografía

[1] C. Smith and D. Collins, *3G wireless networks*. McGraw-Hill Education, 2002.

[2] R. A. Comes, F. B. Álvarez, F. C. Palacio, R. F. Ferre, J. P. Romero, and O. S. Roig, "Lte: Nuevas tendencias en comunicaciones móviles," *Fundación Vodafone España*, vol. 1, 2010.

[3] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband*. Academic press, 2010.

[4] Apis, *Apis Technical Training AB*. 2012.

[5] 3GPP, *TS 36.410: Evolved Universal Terrestrial Radio Access Network (E-UTRAN) S1 general aspects and principles*. 2014.

[6] 3GPP, *TS 23.401 General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access*. 2016.

[7] 3GPP, *TS 23.228 IP Multimedia Subsystem (IMS) Stage 2*. 2015.

[8] M. Nohrborg, "LTE overview," 2020.

[9] Huawei, "5G knowledge express," 2020.

[10] Ericcson, "Key performance indicators user description 37/1553," 2018.

[11] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[12] M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II*, vol. 10535. Springer, 2017.

[13] S. Skansi, *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer, 2018.

[14] M. A. Nielsen, *Neural networks and deep learning*, vol. 2018. Determination press San Francisco, CA, 2015.

[15] A. Samba, Y. Busnel, A. Blanc, P. Dooze, and G. Simon, "Instantaneous throughput prediction in cellular networks: Which information is needed?," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, IEEE, 2017.

[16] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *European Wireless 2014; 20th European Wireless Conference*, VDE, 2014.

[17] X. Dong, W. Fan, and J. Gu, "Predicting lte throughput using traffic time series," *ZTE Communications*, vol. 13, no. 4, 2015.