Avance propuesta proyecto final (semana 4)

Grupo 4

Iván Camilo Barriga Gómez Código: 202121106
Andrés Felipe Arteta Isaacs. Código: 201124652
David Santiago Muñoz Carrillo. Código: 202121443
Edison Arcángel Giraldo Martínez. Código: 202124624

1. IDENTIFICACION DE CLUSTERS EN RED CELULAR LTE MEDIANTE APRENDIZAJE NO SUPERVISADO(ML)

2. Resumen:

Un proveedor de servicio de telefonía celular LTE requiere orientar las estrategias de planificación y optimización de la red nacional con el fin de maximizar la calidad del servicio entregado a los usuarios.
El proveedor ha recolectado información física de las estaciones radiantes, parámetros de configuración y desempeño de indicadores de calidad tales como: velocidad de navegación, disponibilidad de la celda y sesiones exitosas.
Para el desarrollo de este trabajo, la información recolectada se procesa mediante técnicas de aprendizaje no supervisado (ML), que se encargan de explorar y agrupar la información con cierto grado de similitud, identificando patrones denominados cluster. Estos patrones encontrados brindan al área de negocio información valiosa sobre las características de celdas de la red, las cuales sirven como punto de partida para la optimización de la red.
Este trabajo presenta el tratamiento de la información disponible, los modelos de aprendizaje aplicados y los resultados obtenidos.

3. Introducción:

La red de telefonía celular está compuesta por múltiples estaciones radiantes, que brindan cobertura en áreas delimitadas conocida como celdas. La celda permite la comunicación inalámbrica por radiofrecuencia entre los equipos móviles y la estación radiante.
La red de telefonía móvil celular se encuentra en constante cambio, principalmente por los trabajos de expansión de infraestructura, optimización de la red y cambios en la demanda de servicios de voz y datos, por lo que realizar labores de monitoreo y control son fundamentales para garantizar el correcto funcionamiento y aseguramiento de la calidad del servicio por parte del operador.

Un operador de telecomunicaciones privado de Colombia tiene más de 70 mil celdas que soportan la operación en la red LTE. Cada celda brinda el servicio de voz y datos a los usuarios generando métricas de calidad del servicio. Además, existe un registro con información de Identificación y características físicas de la celda y otro registro de los valores de parametrización, los cuales serán utilizados para la exploración e identificación de patrones que permitan encontrar insights para la gerencia de negocio.

4. Revisión preliminar de antecedentes en la literatura:

Debido a la necesidad de procesar información en las múltiples áreas de una red celular, existen diversas aplicaciones con técnicas de machine learning, según el caso de negocio a resolver.

Abdelfatteh Haidine, Fatima Zahra Salmam, Abdelhak Aqqal y Aziz Dahbi (2021) en su publicacion *“Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives”*[i], [1] presentan de manera detallada como la inteligencia artificial y las técnicas de machine learning son fundamentales para avanzar hacia la próxima generación 5G, con la tendencia marcada hacia las denominadas redes inteligentes. Se presentan diferentes casos de aplicación en planificación, optimización y gestión de redes. Se mencionan algunos casos de aprendizaje no supervisado como detección de anomalías, reducción de latencia, retransmisión.

Nikita Butakov, Loren Jan Wilson, Wenting Sun, Angel Barranco(2021) en el blog del proveedor de hardware Ericsson "*Machine learning use cases: how to design ML architectures for today's telecom systems*" [2] se reconoce la importancia del machine learning en las telecomunicaciones móviles, destacando aplicaciones tales como la detección de anomalías, análisis de causa raíz, administración y optimización de la red, mantenimiento predictivo, gestión de incidencias, abandono churn, con un enfoque en la infraestructura requerida para llevar a cabo estas aplicaciones.

H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng y X. Zhu,(2021) en la publicación "Unsupervised Learning-Based Fast Beamforming Design for Downlink MIMO" [3] presenta una aplicación del aprendizaje no supervisado y el aprendizaje profundo, que se encarga de encontrar una configuración óptima para múltiples lóbulos de radiación y la energía asignada en la celda.

Ordoñez, Marco A. *"Optimización de redes UMTS soportada en Machine Learning"(* 2021) en la tesis de Maestría [4] presenta diferentes metodologías de aprendizaje supervisado como arboles de decisión y redes neuronales aplicado a una red de tercera generación en Colombia. En trabajos futuros se menciona la posibilidad de explorar aplicaciones en la red de cuarta generación LTE.

5. **Descripción detallada de los datos**:

Para capturar la información de la data cruda, se consultaron 2 fuentes de información:

- Base de datos A: Base de datos columnar con información estructurada, que permite consulta directa mediante lenguaje SQL. Se realiza una consulta SQL para cada indicador de desempeño, que tiene granularidad horaria para todas las celdas de la red nacional. En la consulta SQL se agrega la cláusula GROUPBY para obtener la agregación diaria y se filtra con WHERE para obtener los datos del día 23 de agosto de 2022 y se exporta en formato .csv. (Información obtenida:3 Variables)
- Repositorio B: Repositorio con frecuencia de generación diaria, que contiene información de identificación y de parametrización de todas las celdas de la red nacional. Para garantizar la consistencia de la data se elige el repositorio del día 23 de agosto de 2022 y se exporta en formato de excel .xlsb. (Información obtenida:10 Variables).

Luego de la captura se agrupa la data preprocesada en un formato de excel que permite la fácil lectura e identificación de las 3 dimensiones elegidas por el área de negocio, que fueron propuestas para este caso de aplicación.

1.Información de Identificación y características físicas de la celda:

- **Nombre:** Identificador único en la red que tiene como prefijo tres caracteres que identifican la ciudad o departamento en donde está ubicada. Variable cualitativa de naturaleza categórica nominal de tipo str. Para el caso de aplicación no se debe tener en cuenta para el procesamiento de los modelos de ML, solamente para efectos de presentación de resultados asociando el index al Nombre.
- **Frecuencia de radiación:** Característica que indica la banda de la frecuencia de radiación utilizada en el espectro electromagnético. Variable cualitativa de naturaleza categórica nominal de tipo str .
- **Modelo del hardware:** Característica que indica la referencia de la Antena, en donde existen varios proveedores de hw con múltiples modelos. Variable cualitativa de naturaleza categórica nominal de tipo str .
- **Altura Antena [metros]:** Característica que indica la altura a la que se encuentra la Antena, medida desde el piso de la torre de telecomunicaciones hasta el lugar de anclaje. Variable cuantitativa de naturaleza continua de tipo dec
- **Inclinación Antena[grados]:** Característica que indica el grado de inclinación de la antena con respecto al anclaje vertical. Esta inclinación permite orientar el patrón de radiación hacia la zona de cobertura deseada. Variable cuantitativa de naturaleza discreta de tipo int.

2. Información de parametrización de la celda:

- **Potencia de radiación [dbm]:** Característica que indica el nivel de potencia de radiación configurado para la celda. La ratio de potencia al ser tan pequeña se expresa en dbm que es una medida logarítmica. Variable cuantitativa de naturaleza continua de tipo dec.
- **Ancho de banda[Mhz]:** Característica que indica el rango de frecuencias autorizadas para la operación. Se puede entender como una medida de capacidad de los canales utilizados para prestar el servicio. Variable cualitativa de naturaleza ordinal, en donde el valor más pequeño 5Mhz se encuentra en un extremo y el valor más grande 20Mhz se encuentra en el otro extremo.
- **Máximo número de Usuarios [#]:** Característica que indica el límite máximo de número de usuarios admitidos en conexión simultanea por la celda. Variable cuantitativa de naturaleza discreta de tipo int.
- **Código de Zona:** Característica que asocia celdas por cercanía geográfica. Un código de zona contiene múltiples celdas. Variable cualitativa de naturaleza nominal de tipo STR.
- **Multiantena:** Característica que indica el tipo de configuración admitido de la celda, de acuerdo a la cantidad de antenas y puertos de radiofrecuencia disponibles. Variable cualitativa de naturaleza nominal de tipo STR.

3.Información de indicadores de desempeño del servicio de la red LTE:

- **Velocidad de navegación promedio [Megabits/s]:** Indicador de calidad del servicio que determina la velocidad de navegación en datos, ofrecida por la celda a los teléfonos móviles de los usuarios. En este caso el indicador es un promedio de todas las muestras durante las 24 horas del día. Variable continua de naturaleza continua de tipo decimal.
- **Disponibilidad de celda [%]:** KPI que indica la tasa de disponibilidad de la celda en el tiempo, es decir, que porcentaje del tiempo la celda se encuentra operativa. Para efectos prácticos la celda se podría encontrar en estado operativo o en estado fuera de servicio en diferentes intervalos del tiempo durante el día. El 100% indica que la celda durante las 24horas no estuvo fuera de servicio en ningún intervalo de tiempo. Variable cuantitativa de naturaleza continua de tipo dec. Para garantizar la sensibilidad del kpi se recomienda trabajar por lo menos con 4 decimales.
- **Navegación exitosa [%]:** KPI que indica la tasa de éxito de prestación del servicio ofrecido por la celda, en cuanto a navegación en datos durante las sesiones de los usuarios en sus teléfonos móviles (bajo ciertas condiciones).

  De manera practica este KPI se define matemáticamente como:

  (Navegaciones totales - Navegaciones no exitosas) / Navegaciones totales

  Este KPI es conocido como retenibilidad y se expresa en porcentaje donde 100% es el éxito total en el servicio. Esta es una variable cuantitativa de naturaleza continua de tipo dec. Para garantizar la sensibilidad del kpi se recomienda trabajar por lo menos con 4 decimales.

5.1. Preprocesamiento de datos:

Debido a que la base de datos original contiene unos datos vacíos, y estos tienen diferentes posibles causas, entre las que cuales se encuentran:

- El KPI que tiene mayor jerarquía es el de la disponibilidad de la celda, ya que nos informa si la celda está operativa o está fuera de servicio. Existen 1.188 Registros en donde el KPI de disponibilidad está vacío, lo que nos dice que el gestor no está recolectando estadísticas de esas celdas. Existen 2 escenarios:
  - El primero es que existe un grupo de celdas nuevas que ya tiene creado la identificación y la parametrización, pero están "apagadas" hasta que se reciba la autorización de encendido, por lo tanto, ya existen, pero todavía no brindan servicio a los usuarios.

- La segunda, (que debería ser el caso de la mayor parte de celdas vacías), son celdas que están desconectadas, probablemente por alguna falla recurrente de transmisión, o fallas del core mucho más complicadas de resolver. De cualquier forma, estas celdas no brindan servicio al usuario.

Debido a que los registros de interés están ligados a los KPIs de la prestación efectiva del servicio al usuario, los registros vacíos que no brindan servicio no son de interés, y, por lo tanto, se realiza un proceso de limpieza eliminando los registros con datos nulos (NaN).

5.2. Estadísticas básicas de los datos:

Luego de haber realziado el proceso de limpieza se presentan algunas tablas que describen el contenido de la base de datos:

En la tabla 1 se muestran los dos primeros registros de la base de datos, para dar un ejemplo del contenido de la misma. Nota: Se transpone para una mejor visualización.

Tabla 1. Visualización de los 2 primeros registros

	0	1
Nombre	BOG.11 de Noviembre_L1	BOG.11 de Noviembre_L2
Frecuencia de radiación	2600.0	2600.0
Modelo del hardware	HWXX6516DS	HWXX6516DS
Altura Antena	16.0	16.0
Inclinación Antena	4	4
Potencia de radiacion	43	43
Ancho de banda	15 MHz	15 MHz
Máximo número de Usuarios	640	650
Código de Zona	20026	20026
Multiantena	Closed Loop Mimo	Closed Loop Mimo
Velocidad de navegación promedio [Megabits/s]	20,25	16,37
Disponibilidad de celda [%]	100,00%	100,00%
Navegación Exitosa [%]	99,94%	99,93%

En la tabla 2 se presentan estadísticas básicas de las variables numéricas de la base de datos. Evidenciamos que no parecen haber datos muy fuera del rango de cada variable, aunque es de notar las siguientes particularidades:

- Los KPI de disponibilidad de celda y Navegación Exitosa parecen tener mucha concentración en valores altos (ambos percentiles 25% están en 99%). Esto podría denotar una alta sensibilidad hacia grandes valores.
- La desviación de la Altura de la antena y la inclinación de la Antena parecen ser más amplias con respecto al promedio que las demás variables.

Tabla 2. Estadísticas básicas de las variables numéricas

	Frecuencia de radiación	Altura Antena	Inclinación Antena	Velocidad de navegación promedio [Megabits/s]	Disponibilidad de celda [%]	Navegación Exitosa [%]
Cuenta	70.462	70.462	70.462	70.462	70.462	70.462
Promedio	2.048	31	6	17	97,85%	99,31%
Desviación Estándar	703	14	3	6	9,71%	6,31%
Mínimo	700	0	0	0	0,00%	0,00%
Percentil 25%	1.900	21	4	12	99,91%	99,68%
Percentil 50%	2.600	30	6	16	100,00%	99,86%
Percentil 75%	2.600	40	7	21	100,00%	99,92%
Máximo	2.600	100	19	56	100,00%	100,00%

En cuanto a las variables categóricas, en la tabla 3 se encuentran estadísticas de la cuenta de valores únicos (cantidad de categorías o textos diferentes), la categoría que más se repite y la frecuencia de la categoría que más se repite. De estas estadísticas podemos concluir que:

- La variable nombre en general tiene una gran cantidad de valores únicos, pero hay (70462-70371) = 91 categorías que se repiten.
- La segunda variable que más categorías tiene es “código de zona”, seguida de modelo de Hardware.

Tabla 3. Estadísticas básicas de las variables categóricas

	Cuenta	Cuenta de valores únicos	Valor de mayor frecuencia	Frecuencia del mayor
Nombre	70462	70371	MED.San Blas_M1	3
Modelo del hardware	70462	167	ASI4517R3v18	10357
Potencia de radiación	70462	52	43	47877
Ancho de banda	70462	5	15 MHz	40557
Máximo número de Usuarios	70462	41	1200	22424
Código de Zona	70462	252	27012	1671
Multiantena	70462	7	Closed Loop MIMO (4x4)	34219

6. **Propuesta metodológica:**

Debido a que se tienen 13 variables es interesante aplicar metodologías de reducción de dimensionalidad como Análisis de componentes principales. En este punto es importante revisar como influyen los 3 agrupadores de las variables.

Para la identificación de patrones ocultos y segmentación de la información, la aplicación de algoritmos de clustering basado en los indicadores de desempeño, pueden ayudar a revelar grupos de celdas con ciertas características de interés para el negocio. Por ejemplo, el conjunto de celdas con un excelente performance en indicadores de calidad, son de interés ya que podrían convertirse en un “golden parameter” que pudiera ser deseado en otros sitios. Asimismo, las celdas con mal desempeño en indicadores de calidad son de gran interés para aplicar acciones correctivas en estas características comunes. Se propone la exploración de algoritmos como K-medias, K-medoides y clustering Jerárquico.DBSCAN podría servir para outliers.

7. **Bibliografía:**

[1]
Haidine, Abdelfatteh et al. "Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives". Moving Broadband Mobile Communications Forward - Intelligent Technologies for 5G and Beyond, edited by Abdelfatteh Haidine, IntechOpen, 2021. 10.5772/intechopen.98517.
Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives | IntechOpen

[2]
H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng and X. Zhu, "Unsupervised Learning-Based Fast Beamforming Design for Downlink MIMO," in IEEE Access, vol. 7, pp. 7599-7605, 2019, doi: 10.1109/ACCESS.2018.2887308.
https://ieeexplore.ieee.org/abstract/document/8586870/

[3]
Nikita Butakov, Loren Jan Wilson, Wenting Sun, Angel Barranco(2021) "*Machine learning use cases: how to design ML architectures for today's telecom systems*" . Recuperado de https://www.ericsson.com/en/blog/2021/5/machine-learning-use-cases-in-telecom

[4]
Ordoñez, Marco A. *"Optimización de redes UMTS soportada en Machine Learning"*. 2021. Universidad Distrital Francisco José De Caldas, tesis de Maestria. Recuperado de Optimización de redes UMTS soportada en Machine Learning (udistrital.edu.co)

---

i