**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. **Título del proyecto.**

**IDENTIFICACION DE CLUSTERS RED CELULAR LTE mediante aprendizaje no supervisado ML**

**DETECCION DE PATRONES EN RED CELULAR LTE mediante aprendizaje no supervisado ML**

**DETECCION DE PATRONES EN RED DE TELEFONIA CELULAR LTE utilizando aprendizaje no supervisado ML**

**CLUSTERIZACION EN RED CELULAR LTE**

**APLICACIÓN DE APRENDIZAJE NO SUPERVISADO (ML) EN DETECCION DE PATRONES DE UNA RED DE TELEFONIA CELULAR LTE**

1. **Resumen**. Una síntesis de lo hecho hasta el momento. Ésta es una parte critica de un trabajo, de hecho, es a veces lo único que se lee. Es por ello, que es una parte integral del proyecto y en esta primera entrega se espera que ustedes tengan una primera aproximación que irán mejorando con el tiempo

Ideas principales a desarrollar:

* Mencionar el problema
* Mencionar objetivo ppal
* Mencionar alcance
* Metodología
* Resumir resultados y generalizar

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.

Por lo tanto, se desea la identificación de clusters basado en informacion 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.

Mediante el aprendizaje no supervisado que pertenece a una de las áreas de machine learning, (se aplicaron métodos como Kmeans y Kcentroides….) para descubrir los patrones en la red LTE.

1. **Introducción**. Donde se establece el problema o la pregunta bien definida con una motivación clara para resolverla.

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.

La aplicación que proponemos para estos datos es el descubrimiento o segmentación de grupos de celdas basado en tres agrupadores de información disponible: Identificación de la celda, Parametrización de la celda e Indicadores de desempeño.

1. **Revisión preliminar de antecedentes en la literatura**. Revisión de literatura nacional e internacional, para enriquecer el proyecto con lo que ya se haya escrito sobre el tema que se va a tratar.

[Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives | IntechOpen](https://www.intechopen.com/chapters/77411) citada

[Exploring machine learning use cases in telecom - Ericsson](https://www.ericsson.com/en/blog/2021/5/machine-learning-use-cases-in-telecom)

[Firms' knowledge profiles: Mapping patent data with unsupervised learning - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0040162516303651)

[Unsupervised Learning-Based Fast Beamforming Design for Downlink MIMO | IEEE Journals & Magazine | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/8586870) citada

1. **Descripción detallada de los datos**. Descripción que incorpore estadísticas descriptivas y/o visualizaciones.

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).

\*Disclaimer: Esta información es privada y debe ser utilizada únicamente para el ejercicio académico y sin compartir información sensible ni de georeferenciación.

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

1.Información de Identificación y características físicas de la celda (color azul):

* 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. Al ser el identificador único de la celda se considera como la etiqueta Y del conjunto de datos o variable independiente. 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 (color amarillo):

* 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: Caracteristica 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 dia. Variable continua de naturaleza continua de tipo dec.
* 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. 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 sesiónes de los usuarios en sus teléfonos moviles (bajo ciertas condiciones).

De manera practica este KPI se define matematicamente:

(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.

Se realiza proceso de limpieza de data eliminando los registros con datos nulos o con el simbolo de guion.

A continuación, se presentan algunas visualizaciones:

1. **Propuesta metodológica**. En esta sección debe mostrarse cómo el proyecto incorpora aprendizaje no supervisado y qué algoritmo planean utilizar. (Este algoritmo puede ser preliminar y pueden modificarlo para la entrega final, si ven que a lo largo del curso incorporan nuevas herramientas que pueden resultar más apropiadas)
2. **Bibliografía**. Incorpora los artículos mencionados en el texto.

(Ejemplo utilizando MLA)

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](https://www.intechopen.com/chapters/77411)

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/>

------Información entrega 3 Para copiar y editar------

1.**Elegir una de las preguntas utilizando la retroalimentación recibida en el avance anterior y expandir sobre su motivación a responderla.**

**Monitoreo de Red de telefonía celular**

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.

La aplicación que proponemos para estos datos es el descubrimiento o segmentación de grupos de celdas basado en tres agrupadores de información disponible: Identificación de la celda, Parametrización de la celda e Indicadores de desempeño.

**Motivación**: La red LTE tiene más de 70mil celdas para dar cobertura a nivel nacional y por lo tanto se requieren herramientas que puedan procesar y generar información valiosa para los ingenieros de radiofrecuencia, que les permita trabajar de manera más eficiente. Las tareas de auditoría de la red son un requerimiento constante para mantener identificado los cambios en valores atípicos, tanto de parametrización como de indicadores de desempeño.

1. **Describir si el problema pertenece a una tarea de reducción de dimensión, clustering, o una combinación de los dos, explicando el por qué.**

Análisis de Clustering porque con más de 70mil registros y 12 variables de diferente naturaleza, se quieren encontrar patrones que puedan responder preguntas como:

¿Existe algún patrón en las celdas con mayor velocidad de navegación? ¿y con menor velocidad?

¿Qué tanto influye la disponibilidad de las celdas en la velocidad? ¿y en las llamadas exitosas?

¿Existe algún patrón en las celdas con mayor porcentaje de llamadas exitosas? ¿y con menor porcentaje?

¿Existen celdas atípicas en cuanto a parámetros de identificación?

¿Existen celdas atípicas en cuanto a parámetros de configuración?

¿Cómo es el desempeño (velocidad, disponibilidad y llamadas exitosas) en estas celdas atípicas?

1. **Proponer al menos un algoritmo/técnica de aprendizaje no supervisado que ustedes crean que es la adecuada para responder la pregunta.**

k-medoides, para establecer grupos de K clústeres que representen la población con diferentes características según la información disponible. Esto nos podría proporcionar insights sobre la relación entre las características propias de los registros y las variables que describen resultados o características asociadas al desempeño del registro.

1. **Presentar estadísticas descriptivas utilizando tablas y/o visualizaciones de los datos crudos que tengan a su disposición. Describir el plan que van a seguir para tener los datos listos. Por ejemplo, cómo van a limpiarlos, que van a hacer con los datos faltantes, etc.**

Origen de la data cruda:

* Base de datos privada en donde se puede extraer la información de indicadores de desempeño de la red móvil a través de consultas SQL. (Información solicitada en proceso de extracción: 3 variables).
* Repositorio privado con información de identificación y de parametrización. (Información obtenida:10 Variables).

Estas dos fuentes de información no tienen barreras, sin embargo, al ser información privada debe ser utilizada únicamente para el ejercicio académico y sin compartir información sensible.

Información de indicadores de desempeño:

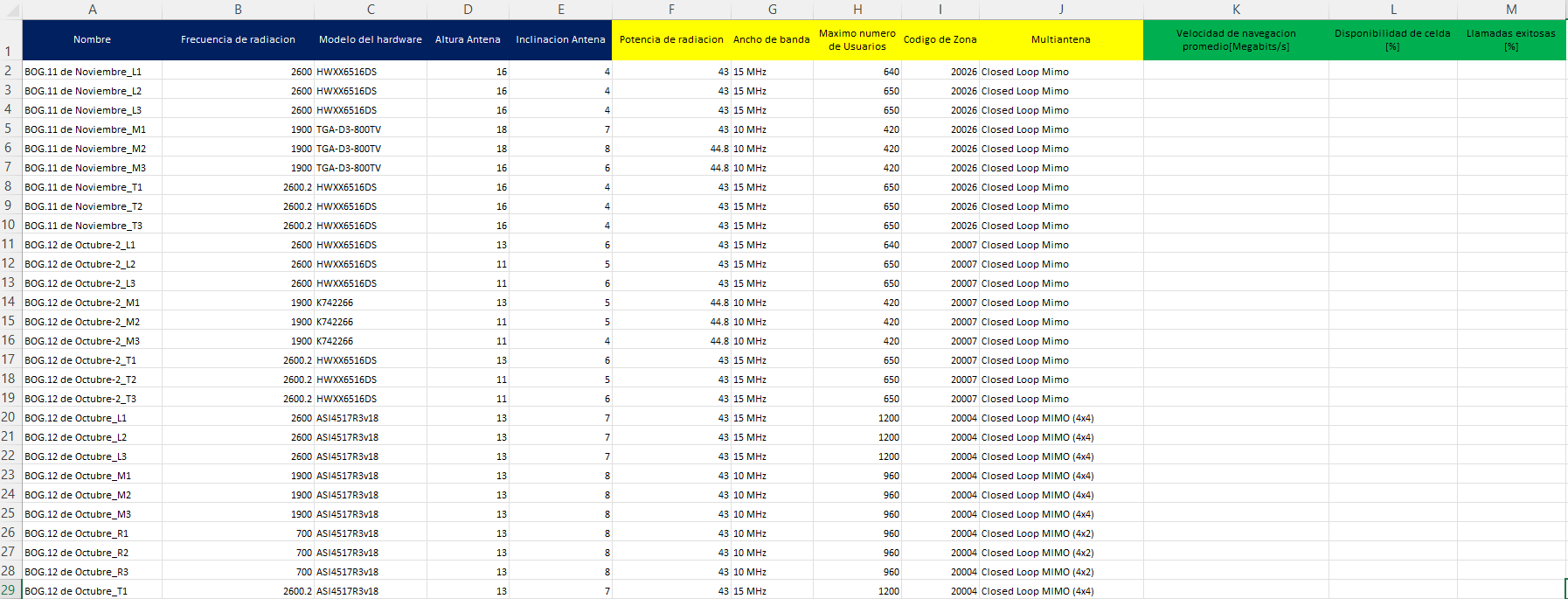
* Velocidad de navegación promedio [Megabits/s]
* Disponibilidad de celda [%]
* Navegacion exitosa [%]

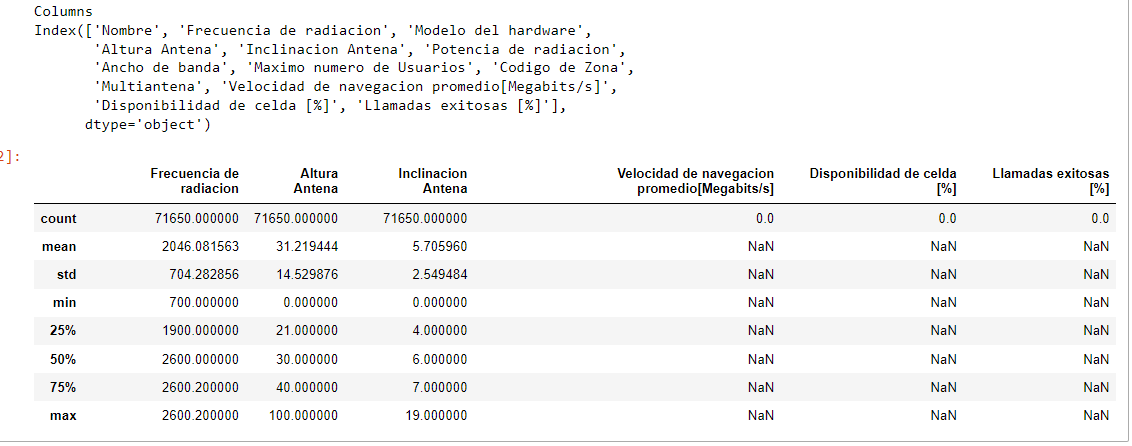
Información de parametrización de la celda:

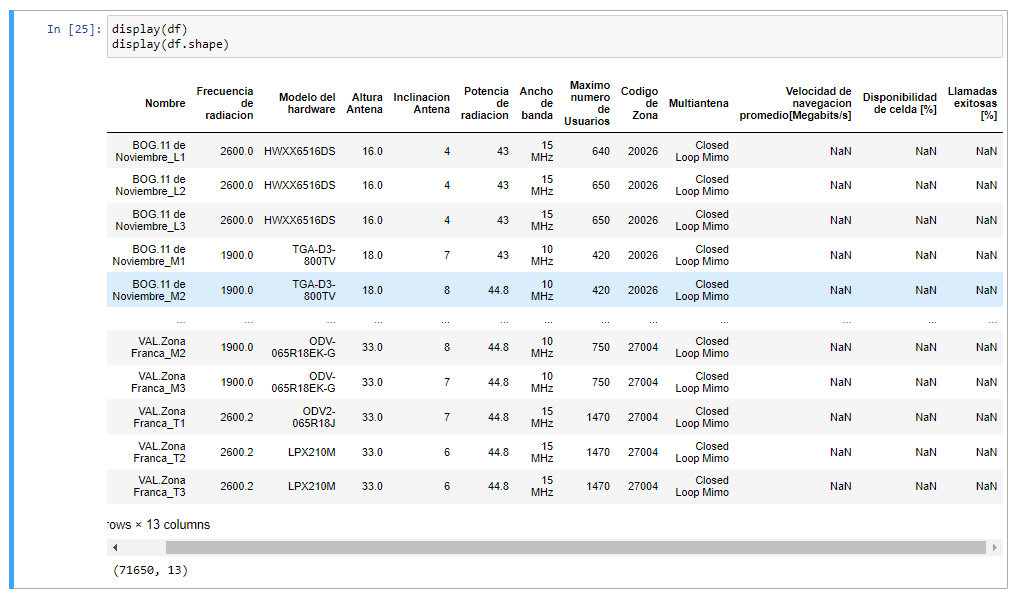
* Potencia de radiación
* Ancho de banda
* Máximo número de Usuarios
* Código de Zona
* Multiantena

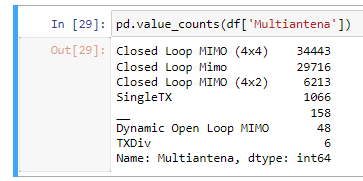
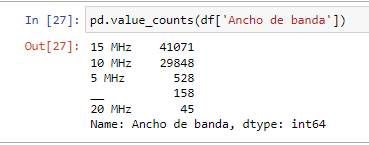
Información de Identificación de la celda:

* Nombre
* Frecuencia de radiación
* Modelo del hardware
* Altura Antena
* Inclinación Antena









Se debe realizar proceso de limpieza de datos faltantes sin imputación, ya que el objetivo es mantener sin alteración los registros existentes.

1. **Con base en los roles definidos en el documento de la semana anterior, delinear las actividades que se llevaron y llevarán a cabo para la primera entrega calificada del proyecto. Ser preciso, y tomar este punto como un contrato entre los miembros del equipo.**

* Exploración de datos, preprocesamiento: Iván Barriga y Edison Giraldo
* Conceptualización del modelo a implementar: Todos
* Modelación e implementación en Python: Andrés Arteta y Santiago Muñoz
* Análisis de resultados: Todos

Referencias

[Colombia, Julio 2022, Experiencia de la Red Móvil Report | Opensignal](https://www.opensignal.com/es/reports/2022/07/colombia/mobile-network-experience)

[¿Qué es LTE? Diferencias entre LTE y 4G // SoftwareLab](https://softwarelab.org/es/lte-4g/#:~:text=En%20concreto%2C%20el%20LTE%20permite,y%20cuatro%20en%20el%20equipo.)

[What is a Radio Access Network (RAN)? (techtarget.com)](https://www.techtarget.com/searchnetworking/definition/radio-access-network-RAN)

<https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Territorial/Guia-Corta-Metodlogia-Medicion-Nuevo-IDF.pdf>