

Proyecto Final de Aprendizaje de Máquina

Estudio sobre registros de consumo de datos, voz y sms



Integrantes:

Claudia Hernández Pérez
Joel Aparicio Tamayo
Kevin Márquez Vega
Javier A. González Díaz
José Miguel Leyva Cruz
Luis E. Amat Cárdenas

Grupo: C-412

Tutor: Lic. Roberto Marti Cedeño

Índice

1. Estudio del estado del arte	2
2. Estudio sobre los datos	2
2.1. Descripción general del conjunto de datos	3
2.2. Distribución de los datos	4
3. Detección de patrones	6
3.1. Datos	7
3.2. Voz	8
3.3. SMS	10
4. Reducción de dimensionalidad y ruido	11
5. Framework de AutoML	12
5.1. El Paradigma AutoML	12
5.2. Arquitectura del Mini-Framework	13
5.3. Resultados y Análisis	14
6. Elección de modelos para predicción	14
6.1. Datos	15
6.2. Voz	15
6.3. SMS	16
6.4. Predicción de consumo	17
7. Conclusiones	19

1. Estudio del estado del arte

La investigación sobre predicción y clasificación de tráfico en redes ha evolucionado de manera notable en los últimos quince años. Entre 2010 y 2015, los modelos estadísticos clásicos como ARMA, ARIMA y SARIMA [5] fueron la base para el análisis de series temporales, mientras que las redes neuronales superficiales como MLP (Multi-Layer Perceptron) [2] demostraron capacidad para capturar patrones complejos antes del auge del DL (Deep Learning).

A partir de 2016 los modelos híbridos que combinan CNN (Convolutional Neural Network) y LSTM (Long Short-Term Memory) [8] [9] permitieron capturar tanto correlaciones espaciales como temporales, consolidando el enfoque de Tráfico espacio-temporal en la predicción de tráfico. Además se mantuvo la relevancia de algoritmos clásicos como SVM (Support Vector Machine), K-NN (K-Nearest Neighbors) y árboles [6], en tareas de clasificación.

En el periodo 2019–2022 se diversificaron las técnicas, incorporando clustering y reducción de dimensionalidad por ejemplo: K-means, DBSCAN y PCA (Principal Component Analysis) [7], además de modelos optimizados como SVR (Support Vector Regression) [11], que superó a los estadísticos en escenarios no lineales. Los avances recientes destacan el uso de K-NN (K-Nearest Neighbors) con selección de características en LTE (Long Term Evolution) [3], la incorporación de grafos mediante GNN (Graph Neural Network) combinados con RNN (Recurrent Neural Network) [4] para modelar dependencias espaciales y temporales. También se resaltan alternativas rápidas y robustas como OS-ELM.

Finalmente, las tendencias emergentes (2024–2025) apuntan hacia arquitecturas modernas como Transformer, capaz de capturar dependencias globales y temporales con gran precisión. En paralelo, la detección de anomalías en tráfico de telecomunicaciones ha cobrado relevancia, con el uso de algoritmos como DBSCAN y SVM (Support Vector Machine) [10]. Las revisiones recientes concluyen que el aprendizaje profundo moderno con LSTM (Long Short-Term Memory), GNN (Graph Neural Network) y Transformer [1] domina en problemas Tráfico espacio-temporal complejos, mientras que los métodos ML clásico siguen siendo útiles en escenarios simples o con restricciones de recursos.

2. Estudio sobre los datos

La Empresa de Telecomunicaciones de Cuba S.A. (ETECSA) proporcionó un conjunto de datos con el propósito de desarrollar estudios y modelos

basados en técnicas de Aprendizaje de Máquina (Machine Learning). Dichos datos reflejan el uso de diversos servicios de telecomunicaciones por parte de los usuarios, tales como llamadas telefónicas, mensajes de texto (SMS), recargas de saldo y consumo de datos móviles.

El objetivo principal de este análisis es comprender la estructura, el contenido y las características de los datos, con vistas a su preparación y posterior aplicación en modelos predictivos o de análisis de comportamiento.

2.1. Descripción general del conjunto de datos

El conjunto de datos se encuentra en formato tabular y contiene aproximadamente 10 mil registros pertenecientes a 1102 usuarios y 40 variables, distribuidas en columnas que describen los diferentes aspectos de cada transacción o evento de uso de servicios. Los registros son sobre el consumo de los servicios datos, sms, voz y recarga, en distinta proporción (ver Figura 1).

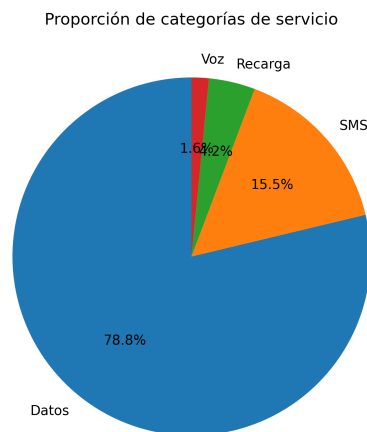


Figura 1: Proporción de los servicios en los datos.

Cada fila representa un registro detallado de uso de servicio (CDR, por sus siglas en inglés: Call Detail Record), que documenta información relacionada con un evento generado por el cliente, como una llamada, el envío de un mensaje o una conexión a internet móvil.

Los registros reportan fechas entre las 12 de la madrugada y 12 del mediodía del día 1 de octubre de 2025 en UTC (lo que serían las 8 de la noche del 30 de septiembre y 8 de la mañana del 1 de octubre hora de Cuba ‘UTC-4’) del día 1 de octubre de 2025.

A continuación, se presenta un resumen de los tipos de variables más relevantes (ver Cuadro 1):

Tipo de variable	Ejemplo de campos	Descripción general
Identificadores	CDR_ID, OBJ_ID, OWNER_CUST_ID	Identifican de manera única cada registro, objeto o cliente asociado.
Temporales	START_DATE, END_DATE	Indican la fecha y hora de inicio y fin del servicio utilizado.
Catégoricas	SERVICE_CATEGORY, FLOW_TYPE, USAGE_SERVICE_TYPE	Especifican el tipo de servicio, su categoría (voz, datos, SMS, recarga) y dirección del tráfico (entrante/saliente).
Núméricas	ACTUAL_USAGE, ACTUAL_CHARGE, TOTAL_TAX_AMOUNT	Miden el volumen de uso (minutos, megabytes, mensajes) y los cargos monetarios asociados.
Listas o estructuras anidadas	CHARGE_LIST, CHARGE_SERVICE_INFO, BALANCE_CHG_LIST	Detallan cargos, impuestos y modificaciones de saldo que se producen en cada evento.

Cuadro 1: Descripciones de los datos.

Estos campos se complementan con información auxiliar relacionada con unidades de medida, identificadores de cuenta, ciclos de facturación y valores reservados para futuras ampliaciones del sistema.

2.2. Distribución de los datos

Con el fin de visualizar el comportamiento de los datos se muestran sus distribuciones (ver Figura 2).

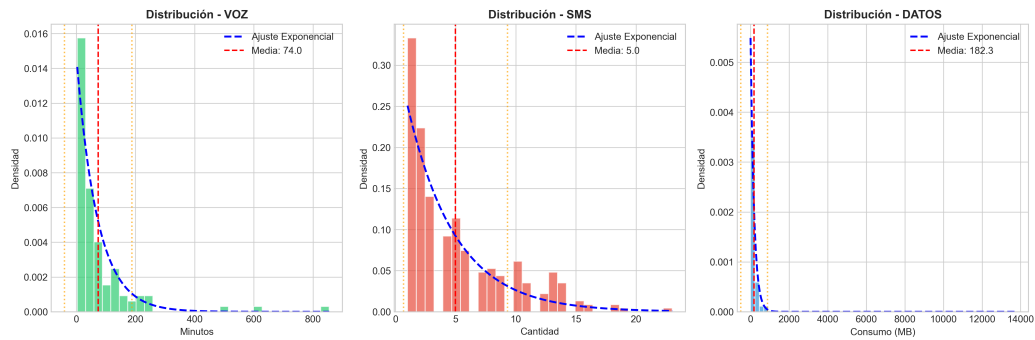


Figura 2: Distribución de los datos

Las gráficas muestran ciertas irregularidades que resultan de interés. La intención es realizar un estudio más profundo de estas anomalías por independiente, dado que no existe ninguna relación entre ellos (ver Figura 3).

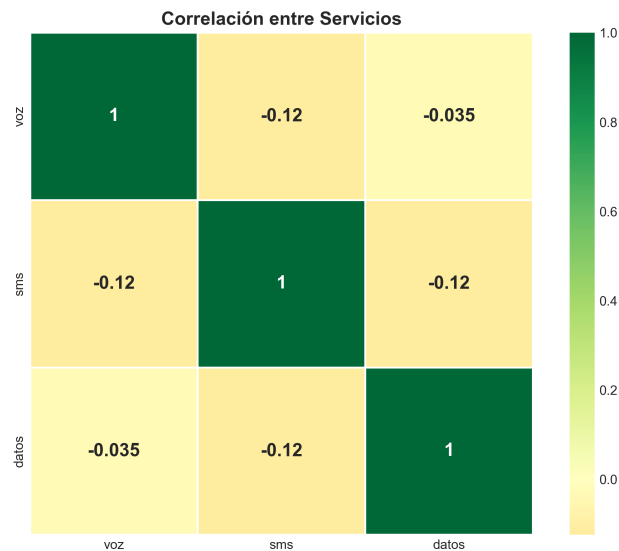


Figura 3: Correlación entre servicios.

Para complementar la visualización de las distribuciones, se presenta un resumen estadístico de los tres servicios en la Cuadro 2. Este cuadro permite observar con mayor claridad las diferencias en el comportamiento de voz, SMS y datos, destacando no solo los valores centrales, sino también la dispersión y los extremos de cada serie.

Métrica	VOZ (min)	SMS	DATOS (MB)
Períodos	114	311	443
Media	74.01	4.99	182.25
Mediana	36.5	3.0	60.86
Desv. Est.	114.46	4.32	694.13
Mínimo	3.0	1.0	0.02
Máximo	855.0	23.0	13 719.91
Total	8437.0	1551.0	80 737.68

Cuadro 2: Resumen estadístico del consumo por servicio.

En el caso de voz, la diferencia marcada entre la media y la mediana revela que existen consumos muy altos que distorsionan el promedio. Aunque la mayoría de los períodos se concentran en valores moderados, algunos clientes alcanzan picos de cientos de minutos, lo que genera una gran dispersión. Esto sugiere que el servicio de voz tiene un grupo reducido de usuarios intensivos que elevan el total acumulado, mientras que la mayoría mantiene un uso más bajo y estable.

En SMS, la media y la mediana son cercanas, lo que indica una distribución más equilibrada y homogénea. La variabilidad es baja y el rango de consumo es estrecho, lo que refleja un patrón de uso bastante uniforme entre clientes.

En datos, la situación es distinta: la media es muy superior a la mediana y la desviación estándar es enorme. Esto evidencia que unos pocos períodos concentran consumos extremadamente altos, mientras que la mayoría se mantiene en valores mucho más bajos. El rango, desde apenas fracciones de MB hasta decenas de GB, confirma la presencia de irregularidades que dominan el total acumulado.

3. Detección de patrones

Las tres gráficas (ver Figura 4) muestran cómo varía el consumo de servicios móviles a lo largo del tiempo, revelando patrones distintos para cada tipo.

En voz, el uso presenta picos aislados pero con largos tramos de baja actividad, lo que sugiere que las llamadas se concentran en momentos específicos. La línea de referencia indica que la mayoría de los períodos están por debajo del promedio, reforzando la idea de un consumo esporádico.

En SMS, el comportamiento es más regular: hay una secuencia continua de envíos con menor variabilidad. La curva se mantiene cerca del promedio,

lo que refleja un patrón de uso más homogéneo entre clientes o dispositivos.

En datos, el consumo es mucho más volátil. Se observan saltos abruptos y valores muy por encima del promedio, lo que indica que algunos períodos concentran un tráfico intensivo, posiblemente por descargas, streaming o actualizaciones. Esta gráfica evidencia que el servicio de datos es el más dinámico y el que más carga representa para la red.

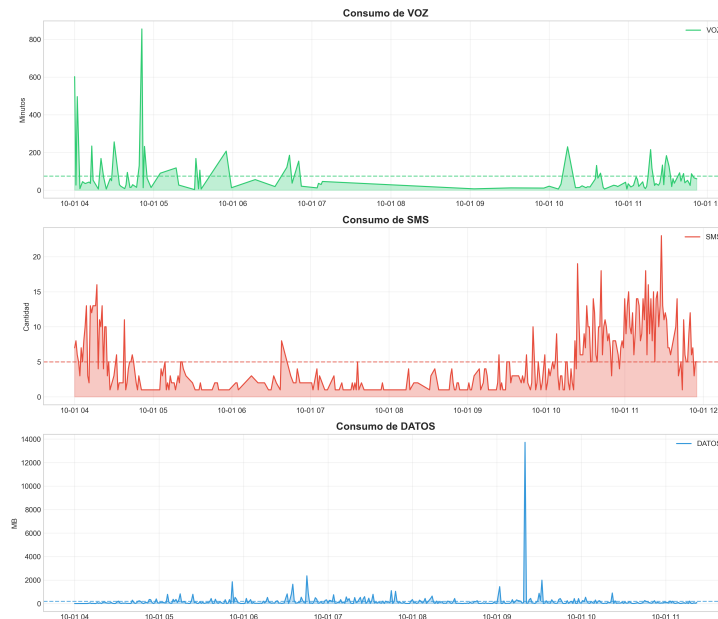


Figura 4: Comparación de consumo de servicios.

Se analizaron los comportamientos más relevantes de cada servicio utilizando DBSCAN y los patrones de horario por servicio usando series de tiempo.

3.1. Datos

En las horas registradas en el horario entre las 4 y 6 de la madrugada (ver Figura 5), se refleja un notable consumo, que pudiera estar relacionado con las promociones de recargas con datos ilimitados en el horario de la madrugada.

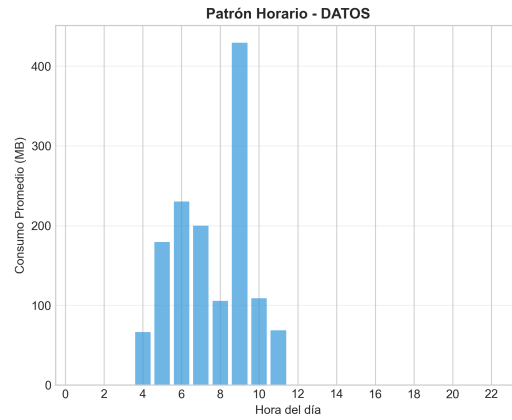


Figura 5: Patrones de horario del servicio de datos.

Tras un análisis realizado sobre el uso del servicio en un intervalo de tiempo se observaron anomalías extremas, habían registros de muy bajo consumo pero también de muy alto. Esto se refleja en una notable diferencia entre la media de consumo de datos en estas anomalías y en los consumidores “normales” (ver Cuadro 3).

Variable	Valor (Byte)
Media Normal	7 853 641.44
Media Outlier	249 439 482.75
Min Outlier	64.0
Max Outlier	14 114 331 638.0

Cuadro 3: Resumen de valores de la variable ACTUAL_USAGE

3.2. Voz

En las horas registradas se muestra un consumo significativo en el horario entre las 12 y 2 de la madrugada (ver Figura 6). Este horario destaca debido a que los otros registros pertenecen a horas de la madrugada donde tiene sentido que disminuya el uso de llamadas.

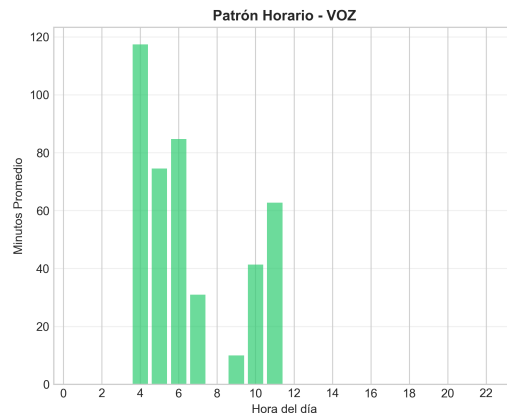


Figura 6: Patrones de horario del servicio de voz.

Se identificaron anomalías significativas en las llamadas. Se consideraron aquellas llamadas cuya duración superaba ampliamente la media, utilizando la columna **ACTUAL_USAGE** (ver Cuadro 4). Además, se examinó la variable **ACTUAL_CHARGE** (ver Cuadro 5), la cual mostró una irregularidad marcada: la mayoría de los registros presentan valores en cero, mientras que algunos alcanzan cifras superiores a 2000. Finalmente, se analizó la cantidad de llamadas recibidas por los usuarios (ver Cuadro 6)

Variable	Valor (min)
Media Normal	36.52
Media Outlier	310.5
Min Outlier	97.0
Max Outlier	831.0

Cuadro 4: Resumen de valores de la variable **ACTUAL_USAGE**.

Variable	Valor (centavos)
Media Normal	0.0
Media Outlier	18 888.89
Min Outlier	2 500.0
Max Outlier	80 500.0

Cuadro 5: Resumen de valores de la variable **ACTUAL_CHARGE**.

Variable	Valor
Media Normal	1.20
Media Outlier	5.0
Min Outlier	4
Max Outlier	6

Cuadro 6: Resumen de valores de la variable llamadas recibidas.

3.3. SMS

En las horas registradas se observa un pequeño aumento del consumo en el horario entre las 6 y 8 de la mañana (ver Figura 7), lo que pudiera estar dado debido a que a esas horas la comunicación por mensajes es más discreta que realizar una llamada. No obstante no se distinguen diferencias significativas.

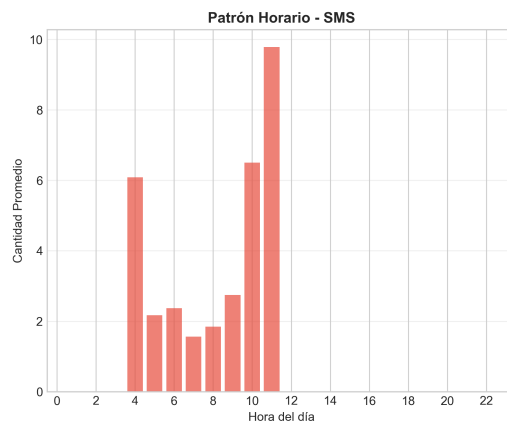


Figura 7: Patrones de horario del servicio de sms.

En el caso de los mensajes de texto se realizó un análisis específico considerando únicamente la cantidad de SMS recibidos, utilizando la columna `OTHER_NUMBER`. Mientras que la mayoría de los registros presentan un número reducido de mensajes recibidos, algunos casos muestran valores significativamente superiores (ver Cuadro 7).

Variable	Valor
Media Normal	2.54
Media Outlier	25.96
Min Outlier	7
Max Outlier	168

Cuadro 7: Resumen de valores de la variable mensajes recibidos.

4. Reducción de dimensionalidad y ruido

Si bien se tienen pocas variables que no deberían suponer un problema al ejecutar los algoritmos clásicos, existe una alta correlación (ver Cuadro 8) entre algunas de ellas, de las que se pueden prescindir.

Variable 1	Variable 2	Correlación
RATING_USAGE	ACTUAL_USAGE	1.000000
DEFAULT_ACCT_ID	OBJ_ID	1.000000
DEFAULT_ACCT_ID	OWNER_CUST_ID	1.000000
OWNER_CUST_ID	OBJ_ID	1.000000
OBJ_ID	CDR_ID	0.913608
DEFAULT_ACCT_ID	CDR_ID	0.913608
OWNER_CUST_ID	CDR_ID	0.913607

Cuadro 8: Correlaciones entre variables seleccionadas

Para abordar este problema se proponen dos enfoques con supuestos conceptuales distintos:

- **Análisis de Componentes Principales (PCA):** considera que la información está contenida en la varianza de los datos.
- **Análisis de Componentes Independientes (ICA):** asume que la información se encuentra en la independencia estadística de las variables.

Los resultados muestran que las dos primeras componentes extraídas por PCA e ICA corresponden principalmente a los identificadores de usuario y al consumo, junto con las columnas altamente correlacionadas con ellos.

Finalmente, se observa que tanto PCA como ICA omitieron las variables temporales. Esto resulta coherente con el conjunto de datos analizado, ya que la información disponible se restringe al intervalo comprendido entre las 12:00 a.m. y las 8:00 a.m. del día 1 de octubre de 2025.

5. Framework de AutoML

Tradicionalmente, el ciclo de vida de un proyecto de Machine Learning (ML) involucra fases iterativas como:

- **Preprocesamiento y Feature Engineering:** transformación y creación de variables para capturar patrones relevantes.
- **Selección del Modelo:** elección del algoritmo o familia de algoritmos más apropiada.
- **Entrenamiento y Validación:** ajuste de los parámetros del modelo y evaluación preliminar.
- **Afinamiento de Hiperparámetros:** optimización de los parámetros que gobiernan el proceso de aprendizaje.
- **Evaluación Final y Despliegue:** validación en conjuntos de prueba e implementación.

La falta de estandarización, derivada de la naturaleza diversa de los problemas abordados por el ML (Machine Learning), junto con el alto costo en *horas de trabajo especializado* y recursos de cómputo, ha generado históricamente ineficiencias significativas. Estas se manifiestan en procesos manuales propensos a errores humanos, una experimentación limitada por las restricciones de tiempo y soluciones con baja reproducibilidad que dificultan su auditoría.

5.1. El Paradigma AutoML

El avance de las técnicas englobadas bajo el término **AutoML** (Automated Machine Learning) ha permitido automatizar gran parte de estas tareas, que antes dependían capitalmente del criterio y la experiencia del especialista (*“ojo del experto”*).

Si bien actualmente no existe un framework universal que, de manera rápida y para el caso general, genere un prototipo de solución óptima para proyectos reales complejos (y teoremas como el *“No Free Lunch”* demuestran que tal herramienta omnipotente no puede existir), la implementación de soluciones AutoML *ad-hoc* persigue objetivos clave:

1. **Reducción del sesgo humano:** minimizar la subjetividad en la selección y configuración de modelos.

2. **Aumento de la productividad:** liberar al experto de tareas repetitivas para que se centre en el diseño del problema y la interpretación de resultados.
3. **Mejora de la exhaustividad:** explorar de forma sistemática y automática espacios de búsqueda (modelos, hiperparámetros) mucho más amplios.

En aras de estos objetivos y para abordar las necesidades específicas de este proyecto (en particular, el trabajo con **series de tiempo**), se optó por diseñar e implementar un *mini-framework* de AutoML especializado (ver Figura 8).

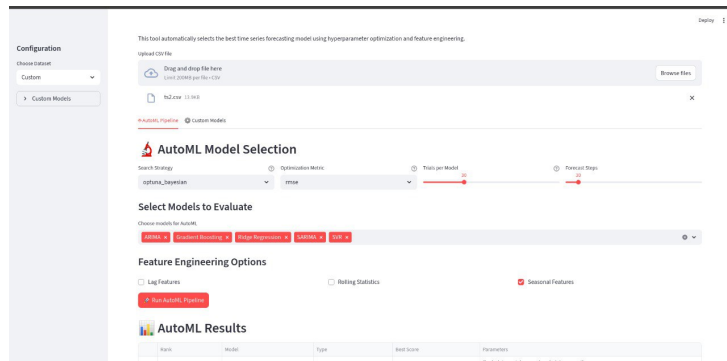


Figura 8: Framework para AutoML.

5.2. Arquitectura del Mini-Framework

El sistema propuesto se estructura en los siguientes componentes modulares:

- **Interfaz Visual de Configuración:** permite al usuario definir los parámetros de alto nivel del *pipeline* AutoML (no los hiperparámetros internos de cada modelo). Desde aquí se seleccionan:
 - Conjuntos de características (*features*) a evaluar.
 - Métricas de evaluación primarias y secundarias (e.g., RMSE, MAE, MAPE).
 - Restricciones de tiempo y recursos computacionales.
- **Adaptadores de Modelo (*Model Adapters*):** capa de abstracción que unifica la interfaz de entrada/salida para diversos tipos de modelos (estadísticos como ARIMA, basados en árboles, redes neuronales, etc.), facilitando su integración en el *pipeline*.

- **Motor de Búsqueda y Optimización:** componente central que implementa el algoritmo de búsqueda. Opera bajo la hipótesis simplificada de que el *pipeline* óptimo, dado un dataset fijo, está compuesto por un **único modelo** (es decir, no explora automáticamente *ensembles* o *stacks* de modelos ya que sería computacionalmente inviable). Realiza una **búsqueda exhaustiva** (o semi-exhaustiva, según espacios muy grandes) sobre el espacio definido de modelos e hiperparámetros, guiada por la métrica objetivo especificada por el usuario.
- **Módulo de Visualización y Comparación:** genera reportes y visualizaciones unificadas (pronóstico vs. real, comparación de métricas entre modelos) que facilitan el análisis comparativo y la toma de decisiones final por parte del experto.

Esta arquitectura permite **eliminar de forma automática variantes de modelo poco efectivas** en las etapas tempranas, filtrando solo las configuraciones más prometedoras para una evaluación detallada. Esto resulta en un **uso más eficiente del tiempo del experto** y de los recursos computacionales.

5.3. Resultados y Análisis

La aplicación del framework al dominio específico de las series de tiempo del proyecto arrojó resultados consistentes con la intuición teórica:

- Los modelos más exitosos fueron, en su mayoría, **modelos estadísticos clásicos** que imponen un sesgo (*inductive bias*) fuerte y apropiado sobre la estructura temporal de los datos (estacionalidad, tendencia, autocorrelación). Estos modelos, al estar bien especificados para el problema, requieren relativamente pocos datos para un ajuste robusto.
- El proceso AutoML permitió descartar rápidamente enfoques más complejos (como ciertas redes neuronales) que, sin un ajuste muy especializado y mayor volumen de datos, no superaban la robustez y simplicidad de los modelos estadísticos.

6. Elección de modelos para predicción

Utilizando los módulos del *framework* se hicieron predicciones de cada uno de los servicios del estudio. Se tuvieron en cuenta los algoritmos: Random Forest, Gradient Boosting, Regresión Lineal y Ridge Regression. La métrica

utilizada fue la raíz del error cuadrático medio (mientras más pequeña mejor). Para todos los servicios se obtuvo que el mejor algoritmo fue Random Forest, aunque hubo pequeñas diferencias en el orden de eficiencia según el servicio.

6.1. Datos

En el conjunto de prueba se obtuvo 762.80 (MB) de RMSE que puede parecer elevado, pero en este caso, el modelo logra capturar adecuadamente la tendencia general del consumo, como se evidencia en la superposición visual entre la serie real y la predicha.



Figura 9: Proyección de los datos de prueba sobre consumo de datos.

Sin embargo, también se observan momentos de subestimación y sobreestimación, lo que sugiere que aún existe margen de mejora en la precisión del modelo, especialmente en los picos de consumo (ver Figura 9).

6.2. Voz

En la Figura 10 se observa que se obtuvo 42.08 (min) de RMSE lo cual indica que el modelo logra una predicción bastante precisa.

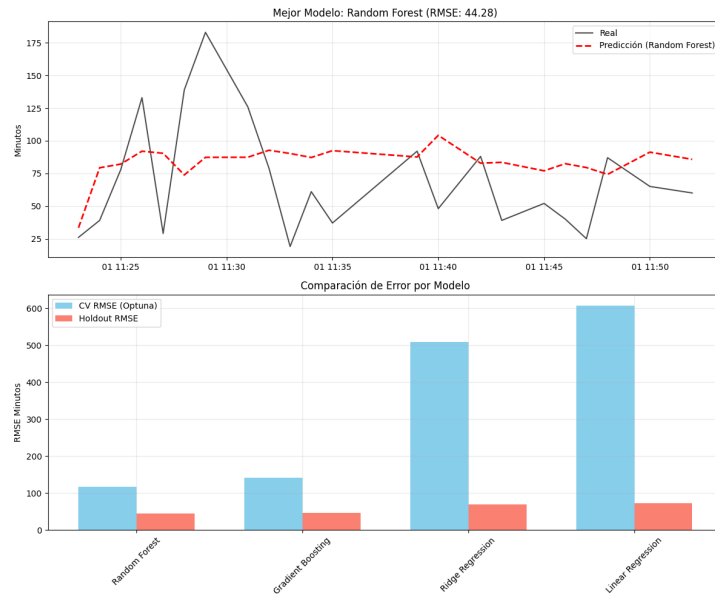


Figura 10: Proyección de los datos de prueba sobre consumo de voz.

La superposición entre la serie real y la predicha muestra una cierta concordancia, lo que no logra capturar las fluctuaciones notables en el transcurso del tiempo.

6.3. SMS

Visualmente, la línea roja discontinua (predicción) sigue de cerca la trayectoria de la línea negra (real), lo que sugiere que el modelo logra capturar con precisión la dinámica del fenómeno observado. No se aprecian desviaciones sistemáticas ni errores persistentes, lo que refuerza la interpretación de que el modelo tiene buen ajuste en este intervalo temporal (ver Figura 11).

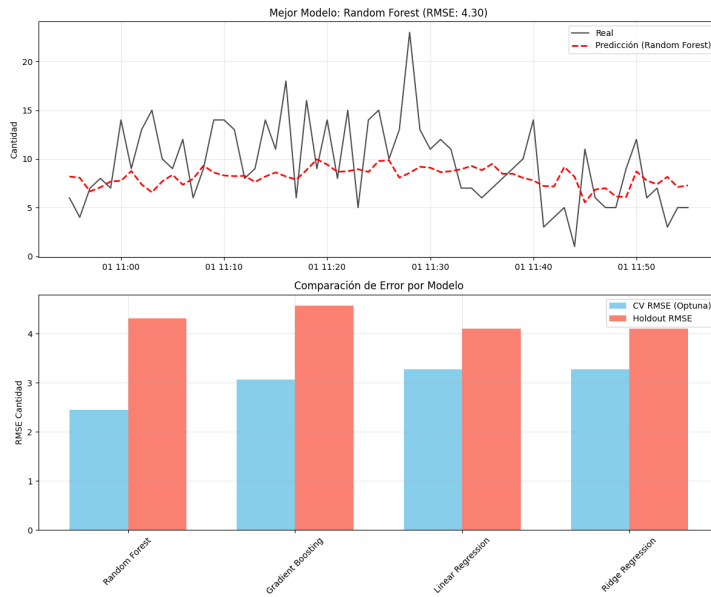


Figura 11: Proyección de los datos de prueba sobre consumo de sms.

El RMSE de 4.36 resulta bajo en términos absolutos, lo que evidencia un margen de error reducido y respalda la confiabilidad del modelo para tareas de predicción en este contexto.

6.4. Predicción de consumo

Como resultado de todo el análisis realizado se pudieron generar predicciones de todos los servicios en un espacio de tiempo de 30 minutos a partir de que finaliza el tiempo registrado en el conjunto de datos inicial.

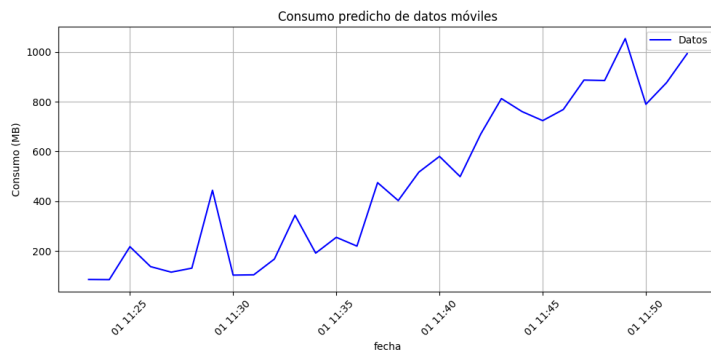


Figura 12: Predicción sobre consumo de datos. Muestra una tendencia creciente en los próximos 30 minutos.

La Figura 12 muestra cómo varía el consumo predicho de datos móviles. Aunque hay fluctuaciones, se observa una tendencia general ascendente, lo que sugiere un incremento progresivo en el uso de datos durante ese período. Este comportamiento podría estar asociado a que inicia el horario laboral y la vida en general (a partir de las 7:30 am).



Figura 13: Predicción sobre consumo de voz. Se espera un brusco descenso pero luego se mantiene constante en los próximos 30 minutos.

La Figura 13 muestra una predicción de consumo de voz con una variación inicial moderada, seguida de una estabilización progresiva en torno a valores medios. El descenso al inicio refleja una leve corrección en el patrón de tráfico, que luego se mantiene relativamente constante, lo que sugiere un comportamiento regular durante el intervalo observado.

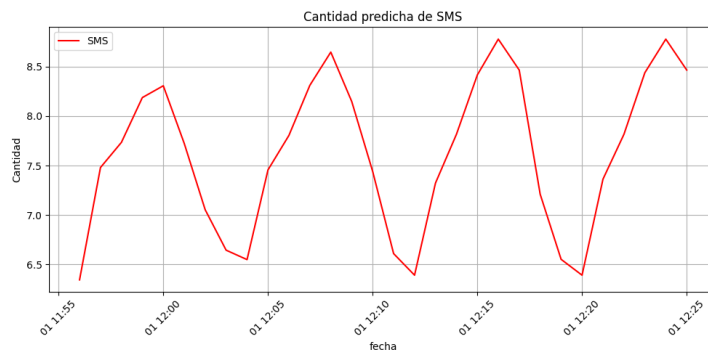


Figura 14: Predicción sobre consumo de sms. Se proyectan variaciones sostenidas en los próximos 30 minutos

En la Figura 14 se observa una variabilidad moderada, sin saltos abruptos ni tendencias marcadamente crecientes o decrecientes. El comportamiento es

relativamente estable, con fluctuaciones suaves que podrían reflejar cambios menores en la actividad de envío de mensajes durante ese periodo. Esto sugiere un patrón de tráfico regular, sin anomalías evidentes.

7. Conclusiones

El análisis conjunto de voz, SMS y datos móviles permite comprender que cada servicio refleja dinámicas distintas de uso, pero en conjunto ofrecen una visión integral del comportamiento de los clientes y de las exigencias que enfrenta la red. La identificación de patrones horarios, la detección de anomalías y la caracterización estadística de los consumos no son hallazgos aislados: constituyen insumos estratégicos para la planificación de capacidad, la segmentación de usuarios y la definición de políticas comerciales.

Las irregularidades encontradas en voz y datos muestran que existen perfiles de consumo extremos que distorsionan las métricas globales. Reconocer estos perfiles es esencial para diseñar esquemas de gestión diferenciados, evitando que unos pocos usuarios intensivos comprometan la calidad de servicio del resto. En contraste, la estabilidad del SMS confirma que ciertos servicios pueden actuar como referencia de comportamiento homogéneo.

La implementación de un mini-framework de AutoML especializado demostró ser una herramienta valiosa dentro del proyecto. No solo agilizó el ciclo de experimentación, sino que también proporcionó una base objetiva y reproducible para la selección del modelo final.

La experiencia reafirma que, si bien el AutoML no reemplaza al experto, lo potencia al permitirle centrarse en la formulación del problema, el diseño de características de alto nivel y la interpretación de resultados, mientras automatiza la búsqueda extensiva y sistemática en el espacio de soluciones técnicas.

En síntesis, las interpretaciones realizadas permiten pasar de una lectura descriptiva de los datos a una comprensión operativa: qué significan los consumos en términos de carga de red, cómo se relacionan con hábitos de los usuarios y qué implicaciones tienen para la gestión futura. El trabajo aporta un marco metodológico que combina análisis estadístico, detección de anomalías y modelado automatizado.

Los datos analizados resultaron poco representativos, pues los registros abarcan únicamente un intervalo de ocho horas de un mismo día, lo que impide observar el comportamiento en otras franjas horarias relevantes. Además, la ausencia de información de distintos días en las mismas horas limita la posibilidad de identificar patrones recurrentes o variaciones temporales. Finalmente, hubiese sido especialmente valioso contar con registros de jornadas

festivas, ya que el consumo de voz, SMS y datos podría diferir significativamente respecto a un día considerado ‘normal’, aportando una perspectiva más completa sobre las dinámicas de uso de los servicios.

Sin embargo, este hecho nunca impuso un problema mayor. Los estudios realizados sobre los datos se hicieron de forma genérica con el propósito de que fueran adaptables a otras muestras de datos.

Glosario

ARIMA Extensión de ARMA que incorpora diferenciación para manejar series no estacionarias. 2

ARMA Modelo estadístico para series temporales que combina componentes autorregresivos y de media móvil. 2

CNN (Convolutional Neural Network) Red neuronal especializada en la extracción de patrones espaciales. 2

DBSCAN Algoritmo de clustering basado en densidad capaz de detectar ruido y anomalías. 2, 7

DL (Deep Learning) Subcampo de Machine Learning basado en redes neuronales profundas con múltiples capas ocultas. 2

DNN (Deep Neural Network) Red neuronal profunda con varias capas ocultas. 2

GNN (Graph Neural Network) Red neuronal diseñada para datos estructurados como grafos. 2

K-means Algoritmo de clustering no supervisado basado en centroides. 2

K-NN (K-Nearest Neighbors) Algoritmo de aprendizaje basado en instancias que utiliza vecinos más cercanos. 2

LSTM (Long Short-Term Memory) Tipo de RNN capaz de capturar dependencias de largo plazo en series temporales. 2

LTE (Long Term Evolution) Estándar de comunicaciones móviles de cuarta generación (4G). 2

ML (Machine Learning) Conjunto de métodos que permiten a un sistema aprender patrones a partir de datos sin programación explícita. 12

ML clásico Algoritmos de Machine Learning no profundos, como k-NN, SVM, Árboles de decisión y modelos estadísticos. 2

MLP (Multi-Layer Perceptron) Red neuronal feed-forward clásica compuesta por múltiples capas totalmente conectadas. 2

OS-ELM Variante de ELM para aprendizaje secuencial u online. 2

PCA (Principal Component Analysis) Técnica de reducción de dimensionalidad basada en varianza. 2

RNN (Recurrent Neural Network) Red neuronal diseñada para procesar datos secuenciales. 2

SARIMA Modelo ARIMA que incluye explícitamente componentes estacionales. 2

SVM (Support Vector Machine) Algoritmo de clasificación y regresión basado en maximización del margen. 2

SVR (Support Vector Regression) Versión de SVM para tareas de regresión. 2

Transformer Arquitectura basada en mecanismos de atención que captura dependencias globales de largo alcance. 2

Tráfico espacio-temporal Modelado del tráfico considerando dimensiones espaciales y temporales. 2

Referencias

- [1] O. Aouedi, V. A. Le, K. Piamrat, and Y. Ji, “Deep Learning on Network Traffic Prediction: Recent Advances, Analysis, and Future Directions”, *ACM Computing Surveys*, vol. 58, no. 1, pp. 1-36, 2025. <https://dl.acm.org/doi/10.1145/3703447>
- [2] N. Chabaa, Y. Zeroual, and J. Antari, “Identification and Prediction of Internet Traffic Using Artificial Neural Networks”, *International Journal of Computer Applications*, vol. 9, no. 10, pp. 19-25, 2010. https://www.academia.edu/21865003/Identification_and_Prediction_of_Internet_Traffic_Using_Artificial_Neural_Networks_Open_Access
- [3] D. Ćurguz, M. Veletić, and M. Zajc, “Machine Learning Models for Prediction of Mobile Network User Throughput in the Area of Trunk Road and Motorway Sections”, *Proceedings of the 27th International Conference on Information, Communication and Automation Technologies (ICAT)*, 2022. <https://atct.ba/proceedings-2022/machine-learning-models-for-prediction-of-mobile-network-user-throughput-in->
- [4] W. Jiang, “Cellular Traffic Prediction with Machine Learning: A Survey”, *Expert Systems with Applications*, vol. 201, p. 117163, 2022. <https://www.sciencedirect.com/science/article/abs/pii/S0957417422011654>
- [5] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, “Traffic Flow Prediction with Big Data: A Deep Learning Approach”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865-873, 2015. <https://bookdown.org/amanas/traficomadrid/docs/Traffic%20flow%20prediction%20with%20big%20data%20-%20A%20deep%20learning%20approach.pdf>
- [6] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, “Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey”, *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1988-2014, 2018. <https://ieeexplore.ieee.org/abstract/document/8543584>
- [7] M. Usama, J. Qadir, A. Raza, H. Arif, K.-L. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, “Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges”, *IEEE Ac-*

- cess, vol. 7, pp. 65579-65615, 2019. <https://ieeexplore.ieee.org/abstract/document/8713992/>
- [8] Y. Wu and H. Tan, “Short-term Traffic Flow Forecasting with Spatial-Temporal Correlation in a Hybrid Deep Learning Framework”, arXiv preprint arXiv:1612.01022, 2016. <https://arxiv.org/abs/1612.01022>
 - [9] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, “Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction”, Proceedings of the AAAI Conference on Artificial Intelligence, 2018. <https://arxiv.org/abs/1802.08714>
 - [10] Y. Zhang, X. Li, and J. Wang, “Outlier Detection in Streaming Data for Telecommunications and Industrial Applications: A Survey”, Electronics, vol. 13, no. 16, p. 3339, 2024. <https://doi.org/10.3390/electronics13163339>
 - [11] K. Zheng, L. Zhao, J. Mei, and W. Wang, “A Novel Method for Improved Network Traffic Prediction Using Enhanced Deep Reinforcement Learning Algorithm”, IEEE Access, vol. 8, pp. 130043-130055, 2020. <https://www.mendeley.com/catalogue/9c49ec6e-4571-3438-b2e3-48c8eac12c0a/>