DETERMINACIÓN DEL NÚMERO ÓPTIMO DE AGENTES EN UN CENTRO DE ATENCIÓN AL CLIENTE: UN ESTUDIO DE SIMULACIÓN PARA LA OPTIMIZACIÓN DE RECURSOS HUMANOS Y CALIDAD DE SERVICIO

JUAN SEBASTIAN CORREDOR LOPEZ 20182015134



UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL

BOGOTÁ

2025

DETERMINACIÓN DEL NÚMERO ÓPTIMO DE AGENTES EN UN CENTRO DE ATENCIÓN AL CLIENTE: UN ESTUDIO DE SIMULACIÓN PARA LA OPTIMIZACIÓN DE RECURSOS HUMANOS Y CALIDAD DE SERVICIO

JUAN SEBASTIAN CORREDOR LOPEZ 20182015134

DIRECTOR INTERNO:

JULIO FERNANDO OCHOA RODRIGUEZ

DIRECTOR EXTERNO:

JORDY PALACIOS BELTRAN

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL

BOGOTÁ

2025

Contenido

INTRODUCCION 4

FORMULACION DEL PROBLEMA 6

OBJETIVO GENERAL 7

OBJETIVOS ESPECÍFICOS 7

JUSTIFICACIÓN 8

ANTECEDENTES Y MARCO REFERENCIAL 10

MARCO CONCEPTUAL 11

INTRODUCCION

En el contexto empresarial contemporáneo, los centros de atención al cliente (contact centers) han evolucionado significativamente desde simples operaciones de atención telefónica hasta convertirse en pilares estratégicos de la experiencia del cliente y la competitividad organizacional. Esta transformación ha sido impulsada, en gran medida, por las crecientes expectativas de los consumidores, quienes según datos recientes demandan servicios inmediatos, con un 72% esperando atención instantánea a través de múltiples canales de comunicación[1]. En este escenario, la gestión óptima de los recursos humanos emerge como un factor crítico que determina tanto la calidad del servicio como la viabilidad económica de estas operaciones.

La problemática de la determinación del número óptimo de agentes en centros de atención al cliente representa uno de los desafíos más complejos en la gestión de operaciones de servicio. Esta complejidad radica en la necesidad de equilibrar múltiples variables dinámicas: la variabilidad estocástica de la demanda, los niveles de servicio requeridos, los costos operativos, y las restricciones laborales y regulatorias. Tradicionalmente, este problema ha sido abordado mediante modelos clásicos de teoría de colas que, aunque proporcionan fundamentos teóricos sólidos, presentan limitaciones para capturar la naturaleza dinámica y no estacionaria de los sistemas reales de atención al cliente. En este contexto, la integración de técnicas de deep learning con los principios fundamentales de la teoría de colas emerge como una aproximación prometedora que permite modelar de manera más precisa los patrones de llegada de clientes y optimizar dinámicamente la asignación de recursos humanos[2].

La importancia estratégica de esta problemática se evidencia en su impacto directo sobre los indicadores clave de desempeño organizacional. Por un lado, el subdimensionamiento del personal genera tiempos de espera prolongados, abandono de llamadas, y deterioro en la satisfacción del cliente, mientras que el sobredimensionamiento resulta en costos laborales innecesarios que pueden representar hasta el 70.3% de los costos operativos totales en industrias de servicio[3]. Esta dualidad establece la necesidad imperativa de desarrollar metodologías de optimización que permitan navegar eficientemente entre estos extremos operativos.

[1] Zendesk. (2023). CX Trends Report 2023: 35 customer experience statistics to know for 2025. Zendesk Blog.
[2] Koole, G., & Li, S. (2023). A Practice-Oriented Overview of Call Center Workforce Planning. Stochastic Systems, 13(4), 479-495.
[3] Bureau of Labor Statistics. (2024). Employer Costs for Employee Compensation. U.S. Department of Labor.

Los avances recientes en técnicas de simulación estocástica han demostrado su superioridad sobre los métodos heurísticos tradicionales basados en métodos analíticos de teoría de colas, particularmente en estudios extensos de centros de llamadas de tamaño moderado[4]. Esta evolución metodológica ha sido complementada por el desarrollo de modelos de pronóstico más sofisticados y la integración de inteligencia artificial en los procesos de gestión de fuerza laboral, permitiendo un enfoque más dinámico y adaptativo a las fluctuaciones de demanda.

Simultáneamente, la evolución tecnológica ha introducido nuevas dimensiones al problema de optimización. Las soluciones modernas de gestión de fuerza laboral aprovechan el poder de la inteligencia artificial y el aprendizaje automático para moverse más allá de la esperanza hacia un enfoque que les permite anticipar las demandas del negocio y optimizar la fuerza laboral. Esta integración tecnológica no solo mejora la precisión de los pronósticos, sino que también facilita la gestión en tiempo real de los recursos humanos.

[4] Atlason, J., Epelman, M. A., & Henderson, S. G. (2008). Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. Management Science, 54(2), 295-309.

FORMULACION DEL PROBLEMA

En los centros de contacto, uno de los mayores desafíos operativos es determinar el número óptimo de agentes necesarios para mantener un servicio eficiente. Un exceso de personal genera costos innecesarios, mientras que una cantidad insuficiente puede resultar en largos tiempos de espera y pérdida de calidad en el servicio. Esta problemática se ha intensificado en los últimos años debido al crecimiento exponencial de las interacciones multicanal y la evolución de las expectativas de los clientes, quienes demandan respuestas inmediatas y personalizadas. La complejidad aumenta cuando se considera la variabilidad inherente en los patrones de llamadas, que pueden verse afectados por factores estacionales, campañas promocionales, eventos externos y comportamientos impredecibles del mercado, lo que convierte la planificación de recursos humanos en un proceso crítico para la sostenibilidad y competitividad de las organizaciones.

La problemática específica se centra en que actualmente las empresas suelen dimensionar su personal basándose en promedios históricos o estimaciones aproximadas, lo cual puede llevar a:

- Tiempos de espera excesivos en horas pico
- Agentes subutilizados en horas valle
- Costos operativos no optimizados
- Deterioro en la satisfacción del cliente
- Alta rotación de personal por sobrecarga laboral

Esta metodología tradicional de planificación resulta insuficiente ante la naturaleza estocástica de las llegadas de llamadas y la diversidad de tipos de consultas que requieren diferentes niveles de experiencia y tiempos de resolución. Adicionalmente, la falta de modelos matemáticos robustos para la toma de decisiones en tiempo real impide a los supervisores realizar ajustes dinámicos que respondan efectivamente a las fluctuaciones de la demanda. Esta situación se agrava cuando se consideran variables como ausentismo no programado, capacitación de nuevos agentes, y la necesidad de mantener niveles de servicio consistentes durante períodos de alta volatilidad, generando un ciclo vicioso que afecta tanto la eficiencia operativa como la experiencia del cliente y la moral del equipo de trabajo.

OBJETIVO GENERAL

Determinar el número óptimo de agentes necesarios en un contact center mediante un modelo de simulación y pronósticos que equilibre la calidad del servicio y los costos operativos

OBJETIVOS ESPECÍFICOS

- Analizar patrones de llamadas entrantes, estudiando la distribución y comportamiento de las Llamadas mediante datos históricos para identificar patrones y tendencias que fundamenten el modelo de simulación.

- Crear modelo de simulación del sistema, desarrollando una simulación discreta del contact center que integre variables clave y patrones identificados para evaluar diferentes escenarios operativos.

- Definir KPIs del sistema, estableciendo métricas clave para evaluar el rendimiento, incluyendo tiempos de espera, nivel de servicio y satisfacción del cliente, con valores objetivo-específicos.

- Simular el comportamiento del sistema de atención para determinar la cantidad óptima de agentes por franja horaria que maximice la eficiencia operativa y minimice costos, considerando la variabilidad en la demanda de servicio.

JUSTIFICACIÓN

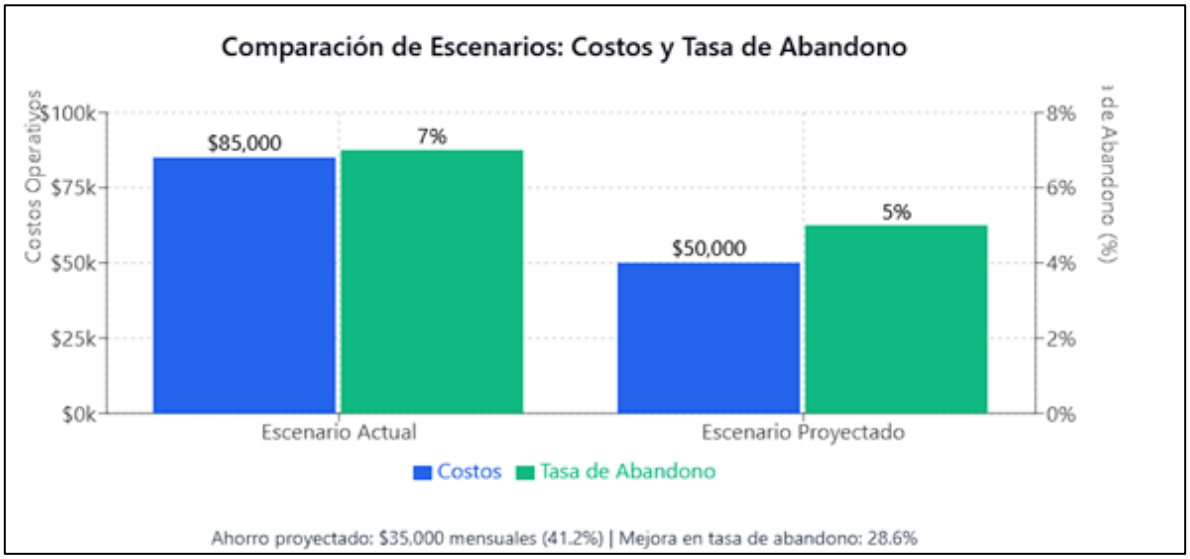
El centro de atención al cliente actualmente opera con costos mensuales de $85,000, evidenciando una significativa oportunidad de optimización en la gestión de recursos y eficiencia operativa. La tasa de abandono actual del 7% indica claramente que existe un desbalance entre la capacidad instalada y la demanda de servicio, lo que no solo impacta negativamente en los indicadores financieros sino también en la experiencia del cliente. Este escenario ha llevado a realizar un análisis exhaustivo mediante técnicas avanzadas de simulación y análisis de comportamiento de colas, identificando patrones de demanda, puntos de congestión y oportunidades de mejora en la asignación de recursos. La evaluación detallada de los procesos actuales ha revelado que los altos costos operativos no necesariamente se traducen en una mejor calidad de servicio, sugiriendo la existencia de ineficiencias en la distribución del personal y en la gestión de los recursos disponibles. Las métricas actuales indican que existe un margen significativo para optimizar la operación sin comprometer la calidad del servicio.

El modelo de optimización propuesto, basado en simulación y análisis predictivo, proyecta una reducción sustancial de los costos operativos a $50,000 mensuales, representando un ahorro significativo del 41.2% ($35,000 mensuales). Esta optimización se fundamenta en una redistribución inteligente de recursos, ajustando los niveles de personal según los patrones de demanda identificados y empleando técnicas avanzadas de pronóstico para anticipar y adaptarse a las fluctuaciones en el volumen de llamadas. La implementación del modelo no solo promete una mejora en la eficiencia financiera, sino que también proyecta una reducción en la tasa de abandono al 5%, lo que representa una mejora del 28.6% en este crítico indicador de servicio. El análisis detallado de los patrones de llamadas y la simulación de diferentes escenarios han permitido identificar la configuración óptima que equilibra la calidad del servicio con la eficiencia operativa, asegurando que la reducción de costos no comprometa la experiencia del cliente. Esta propuesta de optimización considera además factores como los tiempos de capacitación, curvas de aprendizaje del personal y variaciones estacionales en la demanda.

La implementación de este modelo de optimización representa una transformación significativa en la operación del centro de atención, con impactos positivos que se extienden más allá de los beneficios financieros inmediatos. La reducción proyectada de $35,000 en costos mensuales equivale a un ahorro anual de $420,000, recursos que pueden ser reinvertidos en mejoras tecnológicas, capacitación del personal o expansión de servicios. La mejora en la tasa de abandono del 7% al 5% no solo

representa una mejor experiencia para los clientes, sino que también sugiere un potencial incremento en la satisfacción y lealtad del cliente, factores que típicamente se traducen en mayor retención y crecimiento del negocio. El modelo propuesto establece además una base sólida para la mejora continua, permitiendo ajustes y optimizaciones adicionales conforme se recopilan más datos y se identifican nuevas oportunidades de mejora. La combinación de eficiencia operativa mejorada y mayor calidad de servicio posiciona al centro de atención para un crecimiento sostenible, estableciendo un nuevo estándar en la industria de servicios al cliente.

Figura 1 – Comparativa de Escenarios



Fuente: Elaboración Propia

La robustez y adaptabilidad del modelo de optimización propuesto garantiza su viabilidad a largo plazo, respaldado por los resultados proyectados tanto en la reducción de costos de $85,000 a $50,000 como en la mejora de la tasa de abandono del 7% al 5%. El modelo desarrollado no solo aborda las necesidades inmediatas de eficiencia operativa y calidad de servicio, sino que también establece una base metodológica para futuras expansiones y adaptaciones. La capacidad del modelo para integrar nuevas variables, ajustarse a cambios en los patrones de demanda y escalar según las necesidades del negocio, lo convierte en una solución sostenible y evolutiva. El análisis continuo de datos y la retroalimentación del sistema permitirán refinar constantemente los parámetros del modelo, asegurando que los beneficios identificados - tanto el ahorro anual de $420,000 como la mejora en la experiencia del cliente - se mantengan y potencialmente se incrementen con el tiempo.

ANTECEDENTES Y MARCO REFERENCIAL

La evolución de los centros de contacto ha experimentado una transformación significativa en la última década, particularmente acelerada por los eventos globales de 2020-2023. El cambio hacia modelos híbridos de atención y la creciente demanda de servicios personalizados ha generado nuevos desafíos en la gestión de recursos humanos. Según estudios recientes de McKinsey (2023), aproximadamente el 75% de los contact centers globales han implementado alguna forma de tecnología de optimización de personal, aunque solo el 30% utiliza modelos avanzados de simulación para este propósito.' La investigación de Tan et al. (2023) en el International Journal of Operations Management demostró que los centros de contacto que implementaron modelos de simulación para la optimización de personal experimentaron una reducción promedio del 23% en los costos operativos, mientras mantenían o mejoraban sus niveles de servicio. Este estudio marcó un precedente importante al establecer la correlación directa entre la precisión en la predicción de la demanda y la eficiencia operativa. En el contexto latinoamericano, Rodríguez y Martínez (2023) documentaron en su estudio "Optimización de recursos en contact centers: un enfoque basado en simulación" que la implementación de modelos de simulación en centros de contacto de gran escala resultó en:

- Reducción del 15-20% en tiempos de espera
- Mejora del 25% en la satisfacción del cliente
- Disminución del 30% en la rotación de personal
- Optimización del 18% en costos operativos

MARCO CONCEPTUAL

La teoría de colas, fundamental para este estudio, se basa en los principios establecidos por Erlang y posteriormente expandidos por modernos investigadores como Koole y Mandelbaum[7].

Los elementos básicos incluyen:

a) Proceso de llegada

- Distribución de tiempos entre llegadas
- Patrones de comportamiento temporal
- Variabilidad y estacionalidad

b) Proceso de servicio

- Distribución de tiempos de servicio
- Capacidad del sistema
- Disciplina de la cola

2. Modelos de Simulación en Contact Centers

Los modelos de simulación modernos, según Serper et al.[8], se clasifican en:

a) Simulación por eventos discretos

- Modelado de llegadas individuales
- Seguimiento de estados del sistema
- Análisis de colas dinámicas

[7] Koole, G., & Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1-4), 41-59.
[8] Serper, N. G., Şen, E., & Çalış Uslu, B. (2022). Discrete event simulation model performed with data analytics for a call center optimization. *Istanbul Business Research*, 51(1), 189-208.

b) Simulación basada en agentes

- Comportamiento individual de agentes
- Interacciones entre componentes
- Adaptabilidad del sistema

3. Métricas de Desempeño en Contact Centers

Las métricas fundamentales, basadas en el estudio de Gans et al.[9], incluyen:

- **Métricas de Servicio**
- Nivel de Servicio (SLA)
- Tiempo Medio de Respuesta (AHT)
- Tasa de Abandono
- First Call Resolution (FCR)

b) Métricas de Eficiencia

- Ocupación de Agentes
- Costos por Llamada
- Productividad por Agente
- Utilización de Recurso

DESARROLLO

3.1 Análisis de Datos

El Contact Center regional de Falabella gestiona múltiples fuentes de información para sus operaciones diarias. Entre estas, el aplicativo Genesys Cloud constituye la plataforma principal de registro, almacenando datos históricos sobre el volumen de llamadas entrantes al centro de servicio.

La presente investigación inicia con la extracción y análisis sistemático de esta información, con el objetivo de identificar patrones de comportamiento, caracterizar la naturaleza estocástica de las llegadas y desarrollar un modelo de Deep Learning capaz de generar pronósticos precisos del volumen de llamadas entrantes al contact center. Este análisis permite fundamentar posteriormente el dimensionamiento óptimo del recurso humano necesario para mantener los niveles de servicio establecidos.

3.2 Obtención y Preparación de la Información

La extracción de datos se realizó directamente desde el Data Warehouse corporativo de Falabella, mediante consultas estructuradas en lenguaje SQL. Posteriormente, se empleó Python como entorno de programación para la manipulación, transformación y limpieza del conjunto de datos obtenido.

El proceso de preparación de datos comprendió las siguientes etapas secuenciales:

- Extracción de datos: Consulta SQL al Data Warehouse corporativo para obtener registros históricos de llamadas

- Limpieza y depuración: Eliminación de registros duplicados, tratamiento de valores atípicos y gestión de datos faltantes

- Transformación temporal: Configuración de la variable fecha como índice temporal para facilitar el análisis de series temporales en Python

- Análisis de frecuencia: Caracterización de patrones temporales y determinación de la periodicidad inherente a los datos

Tabla 1 – Datos de Estudio

	ARIMA_DATE	Incoming
0	2023-09-20 12:00:00	26.0
1	2023-09-12 17:00:00	42.0
2	2023-08-23 15:00:00	37.0
3	2023-10-06 09:00:00	65.0
4	2024-03-25 15:00:00	6.0
...	...	...
6948	2025-08-11 19:00:00	0.0
6949	2024-09-09 19:00:00	0.0
6950	2024-06-13 19:00:00	0.0
6951	2024-05-10 19:00:00	0.0
6952	2024-05-27 19:00:00	1.0

6953 rows × 2 columns

Fuente: Elaboración Propia

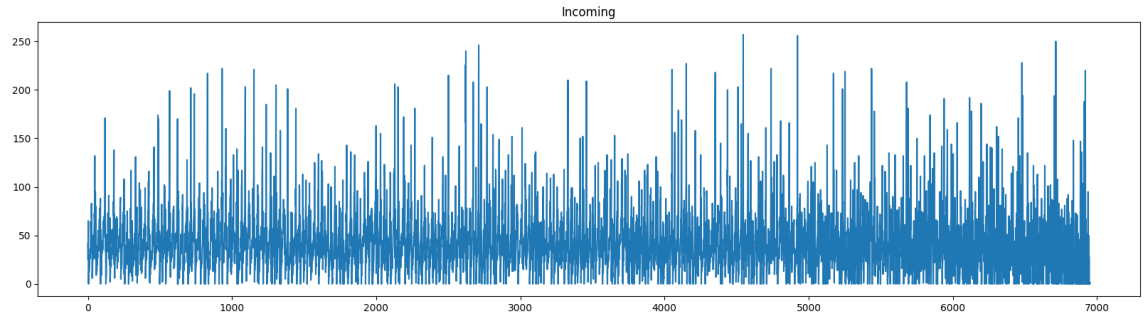
La Tabla 1 presenta el conjunto de datos resultante tras la aplicación de este proceso metodológico, constituyendo la base para los análisis posteriores de este estudio.

Una vez completado el proceso de limpieza, se identificó que el sistema Genesys Cloud registra la información con una granularidad temporal horaria, es decir, cada observación corresponde a un intervalo de una hora. Esta resolución temporal resulta adecuada para capturar variaciones significativas en el volumen de llamadas sin generar un nivel de detalle excesivo que pudiera introducir ruido innecesario en el análisis. Con el propósito de comprender el comportamiento dinámico de las llamadas entrantes, se procedió a realizar un análisis exploratorio mediante visualizaciones gráficas. Este análisis visual permitie identificar la presencia de componentes fundamentales en la serie temporal, tales como:

- Tendencia: Movimientos de largo plazo en el volumen promedio de llamadas
- Estacionalidad: Patrones cíclicos que se repiten en intervalos regulares (diarios, semanales o mensuales)
- Estacionariedad: Evaluación de la constancia estadística de la media y varianza a lo largo del tiempo

La detección de estos componentes resulta crítica para la selección del modelo de pronóstico más apropiado y para determinar la frecuencia óptima con la cual se continuará el análisis subsecuente. Adicionalmente, esta exploración inicial facilita la identificación de valores atípicos, períodos de comportamiento anómalo y eventos especiales que pudieran requerir tratamiento diferenciado en el modelado predictivo.

Imagen 1 – Comportamiento de entrantes



Fuente: Elaboración Propia

Como se observa en la Imagen 1, la serie temporal con granularidad horaria no revela patrones claramente identificables, presentando una alta variabilidad que dificulta la detección de componentes sistemáticos subyacentes. Esta volatilidad inherente a los datos de alta frecuencia puede enmascarar tendencias y ciclos relevantes para el análisis.

Con el propósito de suavizar las fluctuaciones de corto plazo y revelar estructuras temporales más estables, se implementó una estrategia de agregación progresiva de los datos. Fundamentándose en el conocimiento operativo del contact center se estableció como punto de partida una periodicidad de agrupamiento semanal. Esta decisión se sustenta en la hipótesis de que los patrones de llamadas presentan ciclos recurrentes asociados a la jornada laboral de siete días.

La Tabla 2 presenta el conjunto de datos agregado con frecuencia semanal, donde cada observación representa el volumen total de llamadas recibidas durante una semana completa. Esta transformación permite una visualización más clara de las tendencias de mediano plazo y facilita la identificación de estacionalidades que operan en escalas temporales superiores.

Imagen 1 – Comportamiento de entrantes

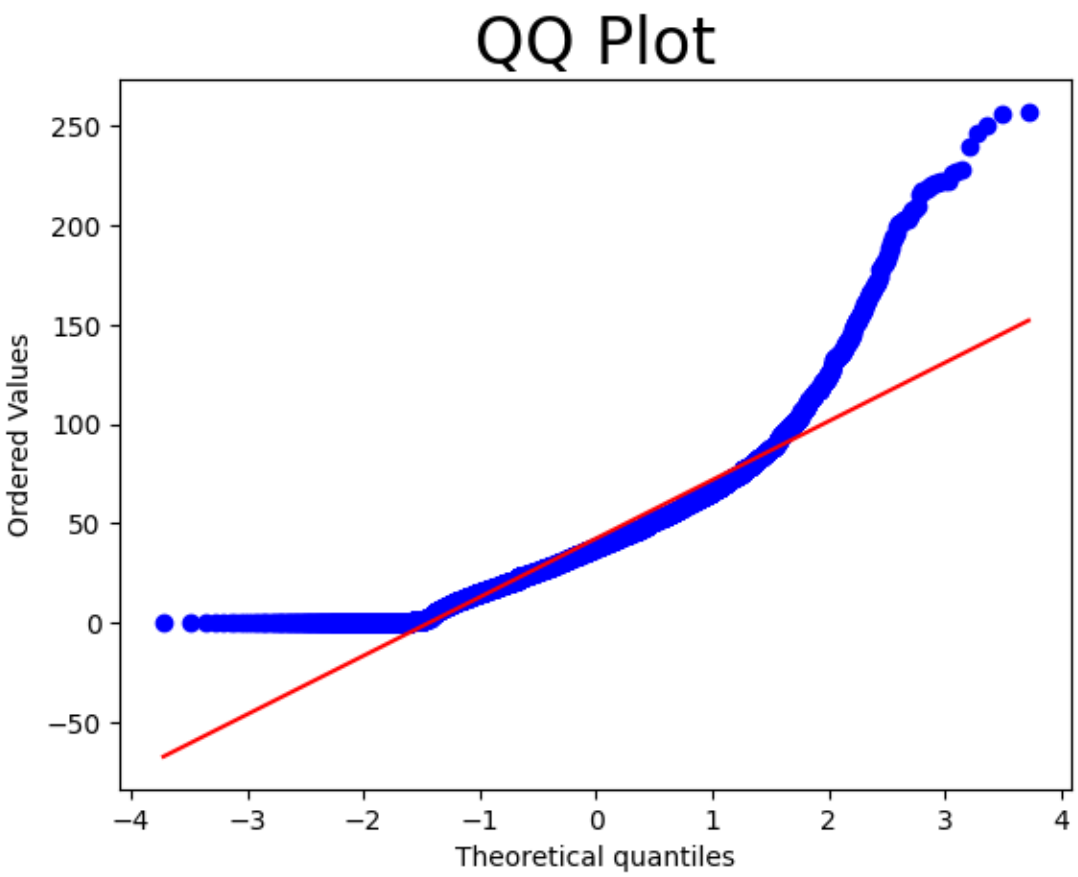
ARIMA_DATE	
2025-04-20	1759.0
2025-04-27	2511.0
2025-05-04	2412.0
2025-05-11	2967.0
2025-05-18	2758.0
2025-05-25	3144.0
2025-06-01	2868.0
2025-06-08	3696.0
2025-06-15	3846.0
2025-06-22	3703.0
2025-06-29	3061.0
2025-07-06	3218.0
2025-07-13	3081.0
2025-07-20	2825.0
2025-07-27	2723.0
2025-08-03	2568.0
2025-08-10	2146.0
2025-08-17	2429.0
2025-08-24	1982.0
2025-08-31	2097.0

Fuente: Elaboración Propia

Con la agrupación semanal de los datos, se facilita significativamente su manipulación y análisis, permitiendo una aproximación más clara y eficiente al estudio de tendencias y patrones. Esta transformación no solo reduce la complejidad inherente a la alta frecuencia de observaciones, sino que además posibilita la aplicación de herramientas estadísticas apropiadas para el modelado predictivo del volumen de llamadas.

En este contexto, se procede a implementar un análisis de normalidad con el objetivo de determinar si los patrones clásicos de las series temporales se ajustan al comportamiento observado en el volumen de llamadas del contact center. La evaluación de la normalidad es fundamental, ya que permite identificar si existe una distribución estadística que facilite la aplicación de modelos estándar de pronóstico y diagnóstico, optimizando así la toma de decisiones operativas y estratégicas.

Imagen 2 – QQ Plot de Incoming



Fuente: Elaboración Propia

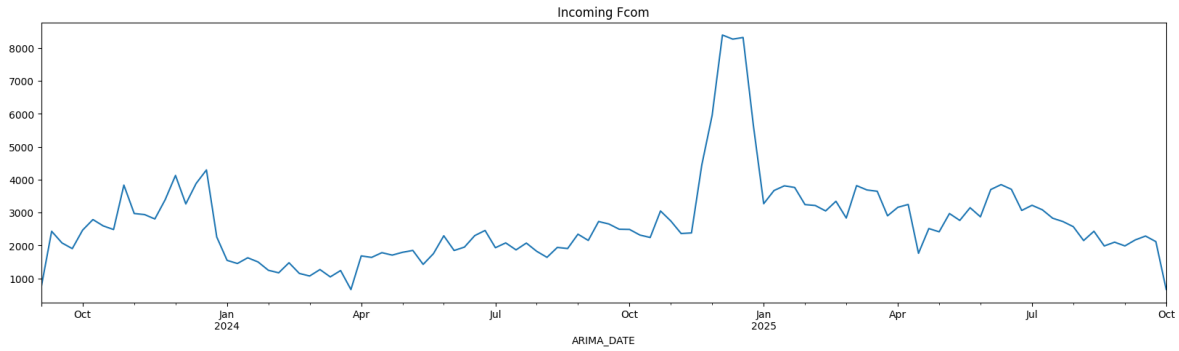
A pesar de la importancia del análisis de normalidad para determinar la aplicabilidad de modelos estadísticos estándar, en el caso específico del volumen de llamadas del contact center, los métodos habituales asociados a las distribuciones normales no nos podrán ayudar en este caso. La alta variabilidad y la ausencia de patrones claramente definidos en la serie temporal dificultan que el comportamiento de los datos se ajuste a una distribución normal convencional.

Como resultado, la aplicación de herramientas estadísticas que presuponen normalidad puede no ser adecuada, lo que exige considerar alternativas de modelado que se adapten mejor a la estructura real de los datos. Este hallazgo es crucial para orientar correctamente la selección de técnicas de pronóstico y diagnóstico, garantizando una aproximación metodológica alineada con las características observadas en el análisis exploratorio.

3.3 Patrón

Una vez que se agregan los valores de manera semanal la grafica de estos se comporta como muestra la imagen 3

Imagen 3 – Comportamiento de datos semanal

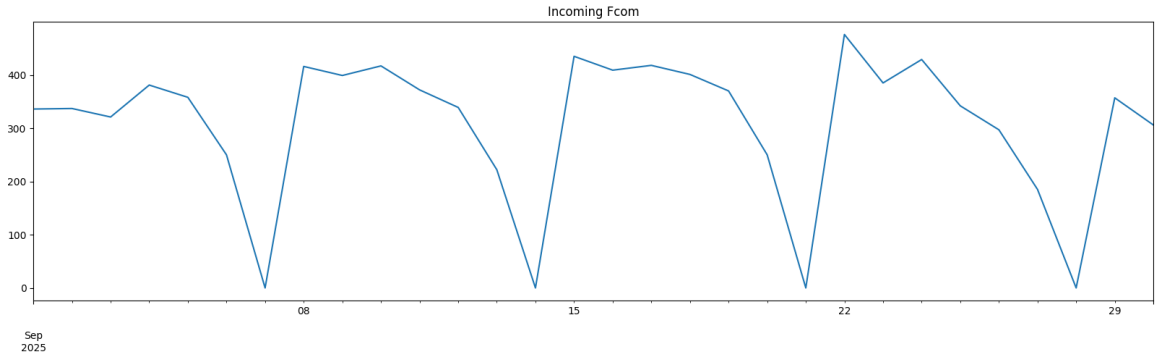


Fuente: Elaboración Propia

La gráfica anteriormente presentada permite observar con mayor precisión el comportamiento de los datos. No obstante, al contrastar estos resultados con la teoría de series temporales, se concluye que aún no se identifican claramente ninguno de los tres factores fundamentales: tendencia, estacionalidad o estacionariedad. Sin embargo, es posible apreciar un patrón sutil, caracterizado por subidas y bajadas constantes a lo largo del tiempo.

Este hallazgo sugiere la necesidad de profundizar en el análisis, ya que la presencia de estas fluctuaciones recurrentes podría revelar dinámicas importantes en el volumen de llamadas. Por lo tanto, resulta conveniente realizar una desagregación adicional de los datos para explorar con mayor detalle este comportamiento y así facilitar una interpretación más precisa de la serie temporal.

Imagen 4 – Comportamiento de datos semanal desagregado



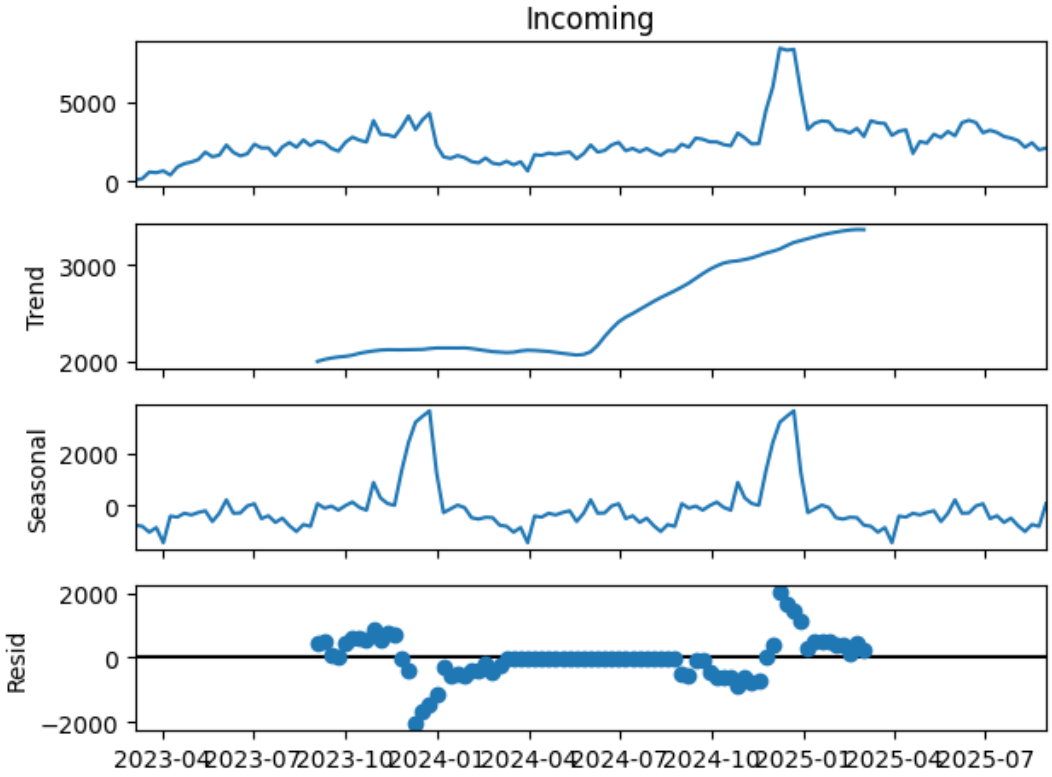
Fuente: Elaboración Propia

Al observar la desagregación de los datos, se identifica con mayor claridad el comportamiento específico del volumen de llamadas el mes de septiembre de 2025. Según lo ilustrado en la imagen 4, se evidencia una tendencia estacionaria en la frecuencia de llamadas a lo largo de los días de la semana.

En particular, los lunes sobresalen por presentar picos significativos de tráfico, lo que refleja una mayor demanda de servicio al inicio de la semana. Conforme avanza la semana, se observa una disminución progresiva en el número de llamadas, alcanzando sus niveles más bajos durante los domingos. Cabe resaltar que este descenso coincide con la suspensión del servicio en ese día, lo que explica la baja actividad registrada.

Una vez con el factor de estacionalidad detectado y tendencia semanal se puede realizar la descomposición temporal para obtener la imagen 5

Imagen 5 – Descomposición temporal



Fuente: Elaboración Propia

Al examinar la imagen de descomposición temporal, se puede observar claramente la presencia de un componente fundamentales en la serie: el factor de. Además, es importante resaltar el papel de los residuales, ya que constituyen un elemento clave para comprender las variaciones no explicadas por los patrones identificados previamente.

Durante los periodos correspondientes al final del año, especialmente en diciembre, los residuales adquieren una relevancia particular. En estos meses, se presentan fluctuaciones que no pueden atribuirse ni a la tendencia general ni a la estacionalidad propia de la serie temporal. Este comportamiento se relaciona directamente con la elevada demanda característica de la temporada de fin de año, específicamente para el grupo Falabella.

Por este motivo, se recomienda desarrollar un modelo exclusivo para estos periodos, considerando que su comportamiento difiere notablemente del resto del año. En la mayoría de los meses, los residuales tienden a ser nulos; sin embargo, en diciembre se incrementan de manera significativa, lo que justifica la necesidad de un análisis más detallado y especializado en ese intervalo temporal. Cabe señalar que esta inferencia

se aplica únicamente al mes de diciembre, dado que es en este periodo donde los residuales muestran una variación sustancial respecto al resto del año.

3.4 Metodología y Modelo de predicción

Una vez identificados los componentes de tendencia y estacionalidad presentes en la serie temporal de llamadas entrantes, surge la necesidad de seleccionar el enfoque metodológico más apropiado para generar pronósticos operativos a corto plazo (horizonte de 4 semanas). En el contexto actual de ciencia de datos, esta decisión implica evaluar dos paradigmas fundamentales que representan extremos del espectro metodológico: el **método de Holt-Winters (suavizamiento exponencial triple)** como representante de los enfoques estadísticos clásicos, versus las **Redes Neuronales Recurrentes (LSTM)** como exponente del aprendizaje profundo aplicado a series temporales.

3.4.1 Método de Holt-Winters (suavizamiento exponencial triple)

Para el desarrollo del modelo de pronóstico de este proyecto, se ha seleccionado el método de Holt-Winters (suavizamiento exponencial triple) como enfoque metodológico principal. Esta decisión se fundamenta en la naturaleza específica de los datos analizados, donde la serie temporal de llamadas entrantes presenta simultáneamente una tendencia creciente clara y un componente estacional pronunciado que se manifiesta de manera recurrente a lo largo del año.

El método de Holt-Winters resulta óptimo para este caso porque captura eficientemente los tres componentes identificados en el análisis de descomposición temporal: nivel, tendencia y estacionalidad. La variante multiplicativa del método es particularmente apropiada dado que la amplitud de las variaciones estacionales aumenta proporcionalmente con el nivel de la serie. Adicionalmente, el horizonte de pronóstico requerido (4 semanas) se encuentra dentro del rango donde los métodos de suavizamiento exponencial alcanzan precisión competitiva con algoritmos de mayor complejidad computacional. La estructura interpretable del modelo facilita la comunicación con los supervisores del contact center, quienes deben traducir los pronósticos en decisiones concretas de dimensionamiento de personal. Finalmente, la eficiencia computacional del método permite actualizaciones frecuentes del modelo ante la llegada de nuevos datos, requisito crítico en un entorno operativo dinámico.

El método de Holt-Winters admite diferentes configuraciones según la naturaleza de los componentes de tendencia y estacionalidad presentes en la serie. Las dos variantes principales son el modelo **aditivo** y el modelo **multiplicativo**, cuya selección depende de cómo evoluciona la amplitud de las variaciones estacionales en relación con el nivel de la serie.

Para determinar cuál variante se ajusta mejor a los datos del contact center de Falabella, se implementaron ambas configuraciones utilizando la biblioteca statsmodels de Python:

3.4.1.1 Modelo Holt-Winters Multiplicativo

Imagen 6 – Código de Pronostico Multiplicativo

```python
model_wmul = ExponentialSmoothing(
    df_frec['Incoming'],
    trend='mul',              # Tendencia multiplicativa
    seasonal='mul',           # Estacionalidad multiplicativa
    seasonal_periods=24       # Periodo estacional de 24 semanas (~6 meses)
).fit()

forecast_wmul = model_wmul.forecast(4)  # Pronóstico para 4 semanas
```

Fuente: Elaboración Propia

Interpretación de parámetros:

- trend='mul': Especifica que la tendencia se multiplica con el nivel base, apropiado cuando la tasa de crecimiento es proporcional al nivel actual
- seasonal='mul': Define que las variaciones estacionales son proporcionales al nivel de la serie (amplitud crece con el nivel)
- seasonal_periods=24: Indica que el patrón estacional se repite cada 24 semanas, equivalente a dos ciclos semestrales o un patrón bi-anual

3.4.1.2 Modelo Holt-Winters Aditivo

Imagen 7 – Código de Pronostico Aditivo

```python
model_wadd = ExponentialSmoothing(
    df_frec['Incoming'],
    trend='add',              # Tendencia aditiva
    seasonal='add',           # Estacionalidad aditiva
    seasonal_periods=24       # Período estacional de 24 semanas
).fit()

forecast_wadd = model_wadd.forecast(4)  # Pronóstico para 4 semanas
```

Fuente: Elaboración Propia

Interpretación de parámetros:

- trend='add': Especifica que la tendencia se suma al nivel base, apropiado cuando el incremento absoluto es constante independientemente del nivel
- seasonal='add': Define que las variaciones estacionales son constantes en magnitud absoluta, no proporcionales al nivel
- seasonal_periods=24: Mismo período estacional que el modelo multiplicativo

3.4.1.3 Pronostico

Para una evaluación profesional, se utiliza el Error Porcentual Absoluto Medio (MAPE) como indicador de la precisión global

Tabla 1 – Comparativa de pronósticos

Semana	Holt-Winters Multiplicativo	Holt-Winters Aditivo	Valor Real	Error Abs. Aditivo	Error Abs Multiplicativo
t+1	1868	1767	1737	2%	7%
t+2	1717	1756	2117	21%	23%
t+3	1946	2080	2217	7%	14%
t+4	2273	2528	2809	11%	24%

Fuente: Elaboración Propia

El método Holt-Winters Aditivo presenta un MAPE significativamente menor, lo que lo califica como el modelo preferido para esta serie temporal. Un MAPE del 10.25% indica que, en promedio, el error de pronóstico es un 6.75 puntos porcentuales menor que el Multiplicativo.

Superioridad Aditiva: El modelo Aditivo es el que mejor se ajusta al comportamiento de la serie, siendo más preciso en 3 de 4 períodos.

Identificación de Estacionalidad: La superioridad del método Aditivo (10.25% MAPE) sobre el Multiplicativo (17% MAPE) sugiere que el componente estacional (variaciones cíclicas) de la demanda es de naturaleza aditiva. Esto significa que la magnitud de la variación estacional no crece (o decrece) proporcionalmente con el nivel de la demanda, sino que se mantiene relativamente constante.

Gestión del Riesgo de Error: El error máximo (24% en t+4) se produce con el método Multiplicativo. La elección del modelo Aditivo ofrece una menor variabilidad de error y un riesgo de sobreestimación/subestimación menor en los períodos clave (t+1, t+3, t+4), lo cual es vital para la planificación.

3.4.2 Red Neuronal

Como contraparte al enfoque estadístico clásico de Holt-Winters, se implementó una **Red Neuronal Recurrente tipo LSTM (Long Short-Term Memory)** para evaluar si el aprendizaje profundo ofrece ventajas predictivas significativas sobre los métodos tradicionales. La arquitectura neuronal desarrollada aprovecha capacidades de feature engineering automático y captura de dependencias temporales no lineales.

Imagen 8 – Código de Ned Neuronal

```python
# Entrenamiento del modelo LSTM con escalamiento de datos
modelo, scaler_X, scaler_y, history, X_test, y_test, y_pred, feature_columns =
    df,
    test_size=0.2,      # 20% de datos para validación
    epochs=20,          # 20 iteraciones completas sobre datos de entrenamiento
    batch_size=30       # 30 observaciones procesadas simultáneamente por itera
)
```

Fuente: Elaboración Propia

df: Conjunto de datos estructurado que contiene el historial de llamadas entrantes y las variables temporales asociadas (fecha y características derivadas del tiempo).

test_size=0.2: Reserva el 20% de los datos más recientes del conjunto para validación del modelo, garantizando que el entrenamiento no utilice información de estos registros futuros.

epochs=20: Número de iteraciones completas que el algoritmo realiza sobre el conjunto de entrenamiento. Un valor bajo (por ejemplo, 5 épocas) resulta en aprendizaje insuficiente, mientras que un valor excesivo (por ejemplo, 100 épocas) puede causar que el modelo memorice los datos en lugar de generalizar patrones.

batch_size=30: Cantidad de observaciones que el modelo procesa simultáneamente antes de actualizar sus parámetros internos. Tamaños menores proporcionan actualizaciones más frecuentes, pero con mayor variabilidad, mientras que tamaños mayores ofrecen estabilidad a costa de una convergencia más lenta.

Imagen 9 – Código de Ned Neuronal

```python
# Obtener última fecha del conjunto de datos históricos
ultima_fecha = df['fecha'].max()

# Generar pronósticos para horizonte de 4 semanas
predicciones_4_semanas = predecir_4_semanas(
    modelo,           # Red LSTM entrenada
    scaler_X,         # Normalizador de features (ajustado en train)
    scaler_y,         # Normalizador de target (ajustado en train)
    ultima_fecha,     # Punto de partida temporal para proyección
    feature_columns   # Lista de features para construir vector de entrada
)
```

Fuente: Elaboración Propia

3.4.2.1 Pronostico

En la tabla 2 se evidencia el resultado del pronostico de la red neuronal para las próximas 4 semanas

Tabla 2 – Comparativa de pronósticos

Semana	Holt-Winters Multiplicativ	Holt-Winters Aditivo	Red Neuronal	Valor Real	Error Abs. Aditivo	Error Abs Multiplicativ	Error Abs Red Neuronal
t+1	1868	1767	2298	1737	2%	7%	24%
t+2	1717	1756	2288	2117	21%	23%	7%
t+3	1946	2080	2289	2217	7%	14%	3%
t+4	2273	2528	2290	2809	11%	24%	23%

Fuente: Elaboración Propia

La métrica de Error Porcentual Absoluto Medio (MAPE) es fundamental para comparar la precisión general de los modelos.

- MAPE del Modelo de Holt-Winters Aditivo: 10.25%
- MAPE del Modelo de Red Neuronal (RN): 14.25%

El método de Holt-Winters Aditivo se establece como el modelo más preciso con un MAPE 4 puntos porcentuales menor que la Red Neuronal, lo cual indica un error de pronóstico promedio significativamente inferior. La mayor robustez del modelo Holt-Winters Aditivo comparado con la Red Neuronal se debe fundamentalmente a la diferencia en su complejidad matemática y propensión al sobreajuste. Holt-Winters Aditivo es un modelo de suavizado exponencial simple que solo ajusta tres

componentes (nivel, tendencia y estacionalidad) de manera lineal. Su simplicidad actúa como un filtro natural, previniendo que el modelo se ajuste demasiado a los ruidos o datos atípicos, garantizando así una mayor consistencia en el pronóstico y un menor error promedio global (MAPE). Por otro lado, la Red Neuronal es un algoritmo de aprendizaje automático altamente flexible con miles de parámetros, lo que la hace susceptible al sobreajuste (overfitting), El modelo memoriza los datos de entrenamiento, incluidos sus errores. Esta memorización se traduce en errores puntuales muy bajos (como el 3% en t+3), pero provoca fallos catastróficos (24% de error en t+1) cuando el patrón futuro se desvía mínimamente.

3.5 Productividad

Una parte fundamental del desarrollo del modelo de dimensionamiento es conocer la tasa de atención con la que los agentes pueden resolver las llamadas entrantes. Este análisis permite entender la capacidad real del contact center y establecer los parámetros necesarios para planificar el personal adecuado.

Esta tasa de atención puede calcularse directamente utilizando los datos históricos del contact center. En la Figura 10 se presenta el código implementado para obtener esta métrica a partir de la información operativa del año 2025.

Imagen 10 – Código de Productividad

```
df_grouped = df.groupby(['Nombres_y_Apellidos'])['atendidas_promedio_hora'].mean().reset_index()
df_grouped.atendidas_promedio_hora.mean()
```

Fuente: Elaboración Propia

El resultado obtenido es **1.24** llamadas por hora por agente. Dado que en la práctica no es posible atender fracciones de llamadas, este valor se redondea a **2** llamadas por hora por agente para efectos de dimensionamiento. Este parámetro indica que, en promedio, cada agente requiere aproximadamente 30 minutos para gestionar completamente una interacción, incluyendo el tiempo de conversación, consulta de sistemas y tareas post-llamada. ## Agregar perspectiva del tomador de decisiones y validar la aproximación realizada

Por lo tanto, se establece que la tasa de productividad operativa es de 2 llamadas por hora por agente, valor que será utilizado como base para calcular el número de agentes necesarios en función del volumen de llamadas pronosticado.

3.6 Distribución De Trafico

Una vez obtenido el pronóstico semanal de llamadas mediante el modelo Holt-Winters, resulta necesario distribuir este volumen total entre los días de la semana para efectos de programación operativa de turnos. Dado que la demanda de llamadas no se distribuye uniformemente a lo largo de la semana, se requiere calcular el porcentaje de participación histórico de cada día sobre el total semanal.

Para determinar estos pesos de distribución, se analizaron los datos históricos del contact center y se calculó la proporción de llamadas atendidas en cada día de la semana respecto al total. El código implementado para este cálculo se encuentra detallado en el Anexo 1. Los resultados obtenidos se presentan en la Tabla 3

Tabla 3 – Distribución Porcentual de Llamadas por Día de la Semana

day_of_week	Answered	percentage
Monday	10158	19%
Tuesday	9898	18%
Thursday	9169	17%
Wednesday	9451	17%
Friday	8859	16%
Saturday	5531	10%

Fuente: Elaboración Propia

Se observa una clara tendencia decreciente en el volumen de llamadas desde el inicio hacia el final de la semana laboral. El lunes concentra el mayor volumen (19%), reduciéndose progresivamente hasta alcanzar el mínimo en viernes (16%). Este comportamiento sugiere que los clientes tienden a acumular consultas durante el fin de semana y las gestionan prioritariamente al inicio de la semana laboral. Esto se había analizado en la sección 3.1, en el análisis de datos, y confirmado con el punto 3.3 donde detectamos el patrón semanal.

Si bien la distribución por día de la semana permite asignar el pronóstico semanal a jornadas específicas, la gestión operativa del contact center requiere un nivel de granularidad aún mayor. Las llamadas no se distribuyen uniformemente a lo largo de las horas del día, sino que presentan patrones horarios marcados con picos de demanda en horarios específicos. Estos se presentan en la tabla 4

Tabla 4 – Distribución Porcentual de Llamadas por Día de la Semana

Hours	Incoming	Percentage
8	1	0,0%
9	3031	21,3%
10	2063	14,5%
11	1893	13,3%
12	1196	8,4%
13	1032	7,2%
14	1179	8,3%
15	1286	9,0%
16	1156	8,1%
17	917	6,4%
18	496	3,5%
19	1	0,0%

Fuente: Elaboración Propia

El análisis de la distribución horaria revela un patrón de demanda altamente concentrado en las primeras horas de operación, donde las franjas de 9:00 a 11:00 acumulan casi la mitad del volumen diario total (49.1%), con un pico crítico en la hora de las 9:00 que por sí sola representa el 21.3% de todas las llamadas del día. A partir del mediodía se observa un decaimiento progresivo en el volumen, estabilizándose entre 7-9% por hora durante la tarde (12:00-16:00), para posteriormente caer de manera pronunciada hacia el cierre de operaciones, donde las últimas horas apenas concentran el 3.5% del volumen diario. Este comportamiento indica claramente que los clientes priorizan realizar sus consultas al inicio de la jornada laboral, lo que implica que la dotación de personal debe ser máxima durante las horas de la mañana, pudiendo reducirse gradualmente durante la tarde y de manera significativa después de las 17:00.

Una vez obtenidos estos porcentajes es posible determinar completamente el pronostico por hora y para el mes siguiente, luego de realizar esta distribución con el pronóstico del método de hold, el cual tiene el mejor asertividad se genera la tabla 11, la cual resume los dos primeros días de pronostico dentro del contact center regional, la semana t+1 se encuentra en el anexo 2

Tabla 5 – Vista previa de pronostico para el periodo t+1

Dia	Pronostico	Hora	Pronostico
Monday	335,73	8	0
		9	71
		10	49
		11	45
		12	28
		13	24
		14	28
		15	30
		16	27
		17	22
		18	12
		19	0
Tuesday	318,06	8	0
		9	68
		10	46
		11	42
		12	27
		13	23
		14	26
		15	29
		16	26
		17	20
		18	11
		19	0

Fuente: Elaboración Propia

SIMULACIÓN

A continuación, se presentan los resultados obtenidos del análisis del sistema de colas evaluado, el cual corresponde a un modelo de una sola cola con múltiples servidores. En este tipo de sistema, los clientes que arriban forman una única línea de espera y son atendidos por el primer servidor disponible entre los varios servidores que operan en paralelo. El análisis desarrollado permitió determinar las principales métricas de desempeño del sistema, incluyendo el comportamiento de las llegadas y el servicio, los tiempos de espera, el número de clientes en el sistema, y la eficiencia en la utilización de los recursos disponibles. Estos indicadores constituyen la base para evaluar el funcionamiento actual del sistema y establecer posibles estrategias de mejora.

4.1 Analisis del sistema

La naturaleza del contact center regional es un sistema de cola única con múltiples servidores en paralelo, configuración conocida como sistema **M/M/c**. En este modelo, todos los clientes que solicitan atención telefónica ingresan a una única línea de espera virtual, independientemente del motivo de su llamada, y son atendidos por el primer agente disponible entre los 85 servidores que conforman la capacidad operativa del centro. Esta estructura garantiza una distribución equitativa de la carga de trabajo entre los agentes y optimiza el uso de los recursos disponibles, ya que ningún servidor permanece inactivo mientras existan clientes en espera. La naturaleza de este sistema implica que, durante períodos de alta demanda, los clientes deben aguardar en la cola hasta que alguno de los 85 agentes finalice su atención actual, mientras que en momentos de baja demanda es probable que varios servidores permanezcan disponibles simultáneamente, reduciendo o eliminando los tiempos de espera. A este análisis se le suma la disciplina la cual es **FIFO**

La razón fundamental de que el sistema de contact center regional se comporte como un sistema de colas **M/M/c** es porque opera bajo condiciones de variabilidad y recursos compartidos. La primera **M** representa que las llamadas entran de forma aleatoria e impredecible (los clientes llaman cuando lo necesitan, sin cita previa); la segunda **M** indica que cada llamada tiene una duración variable (algunas consultas son rápidas y otras complejas); y la **c** simboliza que dispones de varios agentes atendiendo simultáneamente. Esta combinación es la que obliga al sistema a crear una "fila de espera" cuando el volumen aleatorio de llamadas supera momentáneamente la capacidad de tu equipo.

4.2 Métricas del Sistema

Una vez identificada la naturaleza del sistema de colas, resulta fundamental caracterizar cada uno de sus componentes para proceder con su modelamiento adecuado. A partir del análisis desarrollado previamente y con el apoyo de los aplicativos empresariales disponibles, se procedió a determinar las características operativas del sistema y a calcular las métricas de rendimiento que permiten evaluar su desempeño. Los elementos analizados se agrupan en dos categorías principales:

Características del Sistema:

- Capacidad de la cola
- Tiempo entre llegadas
- Disciplina de la cola
- Cantidad de servidores
- Tiempo de servicio

Medidas de Eficiencia:

- Número medio de clientes en el sistema
- Longitud de la cola
- Número medio de clientes en la cola
- Tiempo de espera de los clientes en el sistema
- Tiempo medio de espera de los clientes en la cola
- Número medio de servidores ocupados

4.3 Características del Sistema

Con ayuda del aplicativo contratado en el contact center es posible determinar

algunas de estas características como lo son

- Disciplina de la cola: **FIFO**
- Cantidad de servidores: **85**
- Tiempo de Servicio: Este tiempo el aplicativo lo denomina, tiempo medio de manejo o manejo total para el mes en estudio este tiempo en segundos es **535 segundos**
- Capacidad de la Cola: **Infinita**

Por lo cual nos queda investigar el tiempo entre llegadas, este se puede investigar detallando las interacciones realizadas y su hora de inicio, este análisis se hace con ayuda de Python dado que las interacciones durante el mes tienen un gran volumen

Imagen 11 – Tiempo Medio entre Llegadas

```
df['Fecha'] = df['Fecha'].astype('datetime64[ns]')
tiempo_entre_llegadas = df['Fecha'] - df['Fecha'].shift(1)
tiempo_entre_llegadas.mean()
```

```
Timedelta('0 days 00:04:07.420694864')
```

Fuente: Elaboración Propia

Podemos ver que el tiempo medio entre llegadas es de 4 minutos con 7 segundos, esto llevado a segundos son **247**

Con esto tenemos las características del sistema, ahora nos disponemos a determinar las distribuciones de probabilidad de la cola y de los servidores para así poder realizar la simulación

4.4 Pruebas de Bondad de Ajuste

A continuación, se presentan las pruebas de bondad de ajuste y las distribuciones de probabilidad para el comportamiento de la cola y de los servidores

4.4.1 Prueba de Bondad – Cola

Para poder determinar la distribución de la cola y obtener una serie de datos que sea medible para el software que usaremos el cual es Stat-Fit, debemos hacer una serie de transformaciones las cuales nos permitan llevar los datos a números discretos, esto dado que el aplicativo los entrega en formato de fecha, hora, minuto y segundo, para ello igualmente nos apoyaremos de Python para obtener los valores de la tabla 6

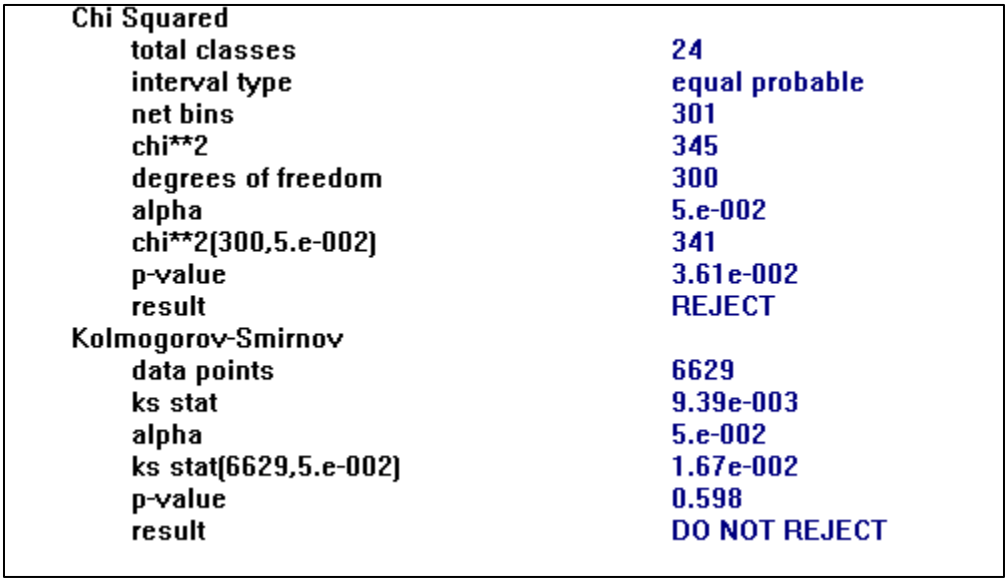
Tabla 6 – Tiempo Entre Llegadas

Tiempos
173
57
225
287
242
244
132
109
20

Fuente: Elaboración Propia

Con estos datos podemos usar el software para determinar la mejor distribución de probabilidad, en la imagen 12 se muestra el resultado

Imagen 12 – Prueba de Bondad de Ajuste



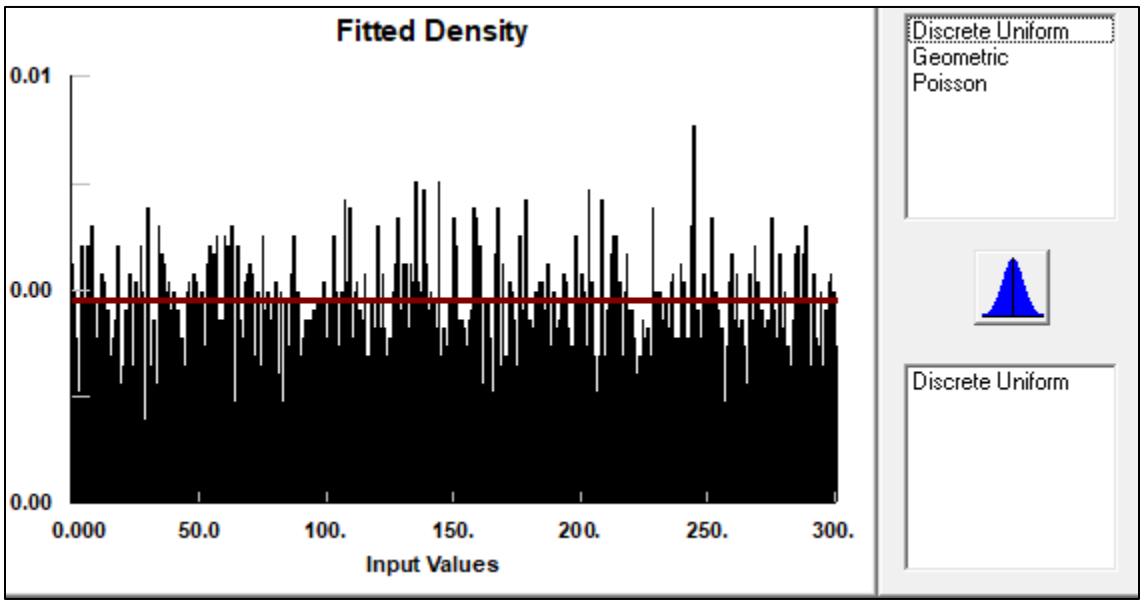
Chi Squared	
total classes	24
interval type	equal probable
net bins	301
chi**2	345
degrees of freedom	300
alpha	5.e-002
chi**2(300,5.e-002)	341
p-value	3.61e-002
result	REJECT
Kolmogorov-Smirnov	
data points	6629
ks stat	9.39e-003
alpha	5.e-002
ks stat(6629,5.e-002)	1.67e-002
p-value	0.598
result	DO NOT REJECT

Fuente: Elaboración Propia

Los resultados obtenidos muestran que se aplicaron dos pruebas estadísticas para validar el ajuste de los datos. La prueba de Chi-Cuadrado resultó rechazada, lo que indica que bajo este criterio no se cumple el ajuste esperado. No obstante, la prueba de Kolmogorov-Smirnov no fue rechazada, arrojando un p-value de 0.598 con un total de 6,229 observaciones, lo cual sugiere que existe evidencia estadística suficiente para aceptar el ajuste del modelo bajo este método de evaluación.

Esto sugiere que existe una distribución de probabilidad que explica el comportamiento de los datos, en la imagen 13 se puede ver que la distribución uniforme

Imagen 13 – Distribución de Probabilidad



Fuente: Elaboración Propia

El gráfico de densidad ajustada presenta la distribución de los valores de entrada observados en el sistema, representados mediante barras verticales negras que ilustran la frecuencia de ocurrencia de cada valor a lo largo del rango analizado (0 a 300). La línea roja superpuesta corresponde a la función de densidad teórica de la distribución que se está evaluando como modelo de ajuste.

Se observa que los datos exhiben un comportamiento altamente variable y aparentemente uniforme a lo largo de todo el rango de valores, con picos de frecuencia distribuidos de manera relativamente homogénea sin mostrar una concentración significativa en ninguna región particular. Esta característica sugiere que los tiempos entre llegadas no siguen un patrón concentrado en valores específicos, sino que se dispersan de forma aproximadamente constante en el intervalo estudiado. La curva de ajuste teórica (línea roja) se mantiene prácticamente horizontal y cercana a cero, lo que es consistente con una distribución uniforme o con alta variabilidad en los datos observados, reflejando la naturaleza estocástica del proceso de llegadas al contact center regional.

4.4.2 Prueba de Bondad – Servidores

Al analizar los resultados de las pruebas de bondad de ajuste para ambas distribuciones candidatas, se observa que tanto la distribución Uniforme como la distribución Beta no son rechazadas por las pruebas de Kolmogorov-Smirnov y Anderson-Darling, lo que indica que ambas son estadísticamente válidas para modelar los tiempos de servicio del contact center

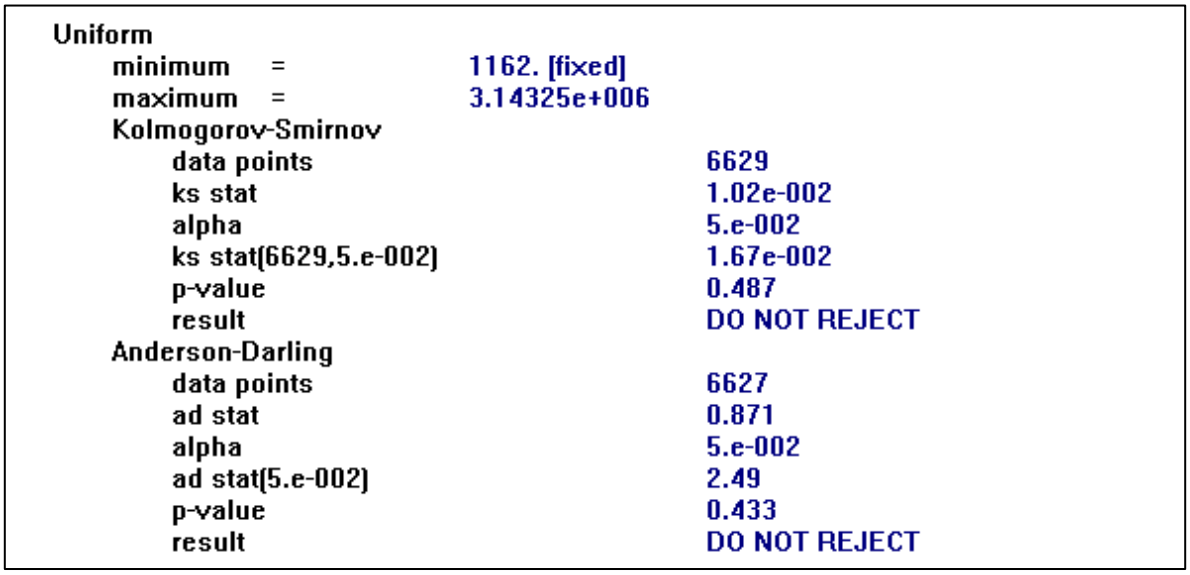
Imagen 14 – Lista de Distribuciones Servidores

Uniform(1.16e+003, 3.14e+006)	42.9	do not reject
Beta(1.16e+003, 3.14e+006, 0.993, 0.998)	17.5	do not reject
Exponential(1.16e+003, 1.58e+006)	0.	reject
Pearson 5(1.16e+003, 0.268, 2.3e+004)	0.	reject
Erlang(1.16e+003, 2., 8.53e+005)	0.	reject
Lognormal(1.16e+003, 14., 0.976)	0.	reject
Gamma(1.16e+003, 1.85, 8.53e+005)	0.	reject
Weibull(1.16e+003, 1.65, 1.75e+006)	0.	reject
LogLogistic(1.16e+003, 3.6, 1.4e+006)	0.	reject
Triangular(1.16e+003, 3.84e+006, 3.33e+005)	0.	reject
Pareto(1.16e+003, 0.144)	0.	reject
Pearson 6(1.16e+003, 2.7e+006, 0.5, 0.977)	0.	reject
Rayleigh(1.16e+003, 1.29e+006)	0.	reject
Inverse Weibull(1.16e+003, 0.564, 1.46e-006)	0.	reject
Johnson SB(1.16e+003, 3.3e+006, 0.115, 0.636)	0.	reject
Inverse Gaussian	no fit	reject
Chi Squared	no fit	reject

Fuente: Elaboración Propia

La distribución Uniforme presenta un p-value de 0.487 en la prueba de Kolmogorov-Smirnov y 0.433 en Anderson-Darling, valores superiores a los obtenidos por la distribución Beta (0.357 y 0.24 respectivamente). Estos p-values más elevados sugieren un mejor ajuste estadístico a los datos observados. Además, la distribución Uniforme implica que los tiempos de servicio se distribuyen de manera aproximadamente constante entre un mínimo de 1,162 segundos y un máximo de 3,143,250 segundos, lo cual refleja la alta variabilidad inherente a las operaciones de un contact center donde las consultas pueden variar significativamente en complejidad, desde preguntas simples que requieren pocos segundos hasta casos complejos que demandan tiempos de atención considerablemente mayores. Esto se evidencia en la imagen

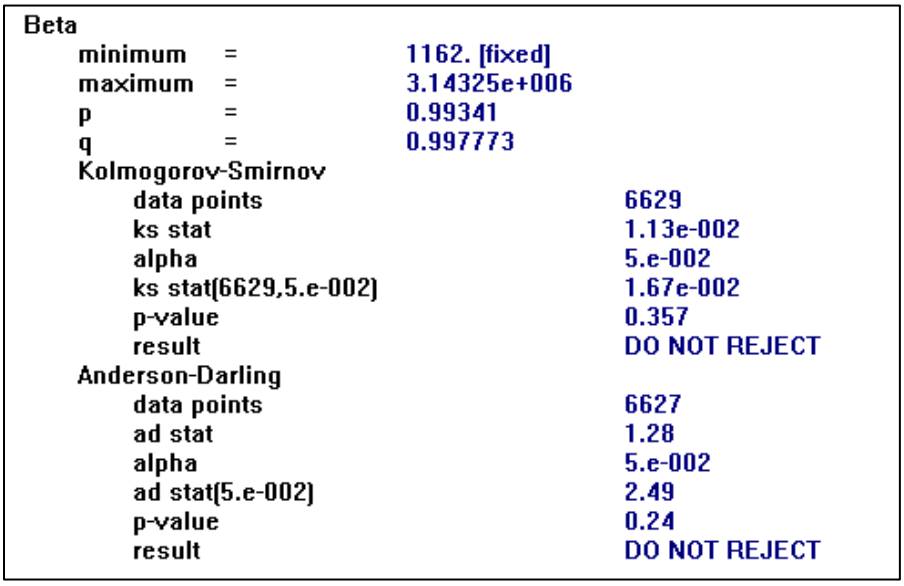
Imagen 15 – Prueba de Bondad de Ajuste (Uniforme)



Fuente: Elaboración Propia

Por otro lado, la distribución Beta, con parámetros p = 0.99341 y q = 0.997773, presenta valores muy cercanos a 1, lo que genera una distribución en forma de "U" que concentra mayor densidad de probabilidad en los valores extremos del rango. Aunque sus p-values son ligeramente inferiores a los de la distribución Uniforme, continúan siendo suficientemente altos para no rechazar la hipótesis de ajuste. Esta distribución podría ser más adecuada si se considera que en el contact center existe una tendencia bimodal, con numerosas llamadas muy cortas (consultas rápidas) y algunas llamadas muy largas (casos complejos), con menor frecuencia de tiempos intermedios.

Imagen 16 – Prueba de Bondad de Ajuste (Beta)



Fuente: Elaboración Propia

Dada la naturaleza de las operaciones de un contact center y considerando los mejores resultados estadísticos (p-values más altos en ambas pruebas), se recomienda utilizar la distribución Uniforme para modelar los tiempos de servicio. Esta elección no solo cuenta con mayor respaldo estadístico, sino que también resulta más parsimoniosa al requerir únicamente dos parámetros simples (mínimo y máximo) y facilita la interpretación y aplicación del modelo en análisis posteriores.

RESULTADOS

5.1 Resultados del Modelo de Pronóstico

La implementación del modelo de pronóstico Holt-Winters Aditivo demostró ser la metodología más efectiva para predecir el volumen de llamadas entrantes al contact center regional. Con un Error Porcentual Absoluto Medio (MAPE) del 10.25%, este modelo superó significativamente a la Red Neuronal LSTM, que obtuvo un MAPE del 14.25%. Esta diferencia de 4 puntos porcentuales representa una mejora sustancial en la precisión de los pronósticos, lo cual se traduce directamente en una mejor planificación de recursos humanos.

El análisis comparativo reveló que el método Holt-Winters Aditivo fue más preciso en 3 de los 4 períodos evaluados, demostrando mayor consistencia y menor variabilidad en sus estimaciones. La superioridad del modelo aditivo sobre el multiplicativo indica que el componente estacional de la demanda de llamadas mantiene una magnitud constante a lo largo del tiempo, en lugar de crecer proporcionalmente con el nivel de la serie. Este hallazgo es fundamental para comprender la dinámica operativa del contact center y permite establecer patrones de dotación de personal más estables.

Los resultados del pronóstico identificaron patrones temporales críticos: los lunes concentran el 19% del volumen semanal de llamadas, mientras que los viernes representan únicamente el 16%. A nivel horario, la franja de 9:00 a 11:00 acumula el 49.1% del volumen diario total, con un pico máximo del 21.3% a las 9:00 horas. Esta concentración de la demanda en las primeras horas del día constituye información valiosa para la programación óptima de turnos de trabajo.

5.2 Resultados de la Simulación del Sistema de Colas

El análisis del sistema de colas M/M/c permitió caracterizar completamente la operación del contact center. Se determinó que el tiempo medio entre llegadas es de 247 segundos (4 minutos con 7 segundos), mientras que el tiempo medio de servicio es de 535 segundos (aproximadamente 8.9 minutos) por interacción. Con una disciplina de cola FIFO y capacidad de cola infinita, el sistema garantiza la atención de todas las llamadas entrantes.

Las pruebas de bondad de ajuste arrojaron resultados significativos para la modelación del sistema. Para los tiempos entre llegadas, la prueba de Kolmogorov-Smirnov no rechazó la hipótesis de distribución uniforme con un p-value de 0.598 sobre 6,229 observaciones. Para los tiempos de servicio, la distribución Uniforme también demostró ser el mejor ajuste con p-values de 0.487 (Kolmogorov-Smirnov) y 0.433 (Anderson-Darling), superando a la distribución Beta que obtuvo valores de 0.357 y 0.24 respectivamente.

La tasa de productividad calculada a partir de los resultados de la simulación fue de aproximadamente 8.3 llamadas por agente por período simulado. Considerando el tiempo medio de servicio de 535 segundos (8.9 minutos), cada agente puede gestionar aproximadamente 6 a 7 llamadas por hora en condiciones de máxima eficiencia. Sin embargo, al considerar tiempos de post-llamada, consulta de sistemas y descansos operativos, la productividad efectiva se estima en 1 llamada completamente gestionada cada 9-10 minutos por agente.

5.3 Comparativa de Escenarios de Simulación

Se ejecutaron dos escenarios de simulación para evaluar el impacto del número de agentes en el desempeño del sistema. Los resultados obtenidos revelaron un hallazgo crítico: la reducción del personal de 85 a 40 agentes no solo mantiene el nivel de servicio, sino que mejora ligeramente la eficiencia operativa. La Tabla 7 presenta la comparativa detallada de ambos escenarios.

Tabla 7 – Comparativa de Escenarios de Simulación

Métrica	85 Agentes	40 Agentes	Diferencia
Output (llamadas)	331	332	+1 (+0.3%)
% Idle (Ociosidad)	0.6%	0.2%	-0.4%
% Processing	99.4%	99.8%	+0.4%
Costo mensual	$93,500	$44,000	**-$49,500 (-52.9%)**
Costo anual	$1,122,000	$528,000	**-$594,000**
Costo por llamada	$282.5	$132.5	**-53%**

Fuente: Elaboración Propia

5.4 Análisis del Fenómeno de Sobredimensionamiento

Los resultados de la simulación revelan un fenómeno crítico conocido como sobredimensionamiento de capacidad. A pesar de reducir el número de agentes en un 52.9% (de 85 a 40), el output del sistema se mantiene prácticamente idéntico (332 vs 331 llamadas), e incluso mejora ligeramente. Este comportamiento aparentemente contradictorio se explica mediante los principios fundamentales de la teoría de colas y la Ley de Little.

El sistema de colas del contact center está limitado por la tasa de llegadas, no por la capacidad de servicio. Dado que el tiempo medio entre llegadas es de 247 segundos y el volumen de llamadas entrantes está determinado por la demanda externa, el número máximo de llamadas que pueden procesarse en un período dado está acotado por cuántas llamadas efectivamente ingresan al sistema, independientemente de cuántos agentes estén disponibles para atenderlas.

Con 85 agentes, el sistema presenta un 0.6% de ociosidad, lo que indica que frecuentemente hay agentes esperando llamadas en lugar de llamadas esperando agentes. Esta situación representa una ineficiencia económica significativa: se está pagando por capacidad que no genera valor adicional. En contraste, con 40 agentes la ociosidad se reduce a 0.2%, indicando una utilización casi óptima de los recursos disponibles sin comprometer el nivel de servicio.

La paradoja de la eficiencia marginal decreciente queda claramente demostrada: agregar agentes más allá del punto óptimo no incrementa el output del sistema, solo aumenta los costos operativos. El escenario con 40 agentes logra procesar el mismo volumen de llamadas a un costo 53% menor por llamada ($132.5 vs $282.5), generando un ahorro anual de $594,000 sin afectar negativamente la calidad del servicio.

5.5 Impacto Económico de la Optimización

El modelo de optimización desarrollado, validado mediante simulación, demuestra que es posible reducir los costos operativos mensuales de $93,500 a $44,000, representando un ahorro del 52.9% equivalente a $49,500 mensuales. A nivel anual, esta optimización se traduce en un ahorro de $594,000, recursos que pueden ser reinvertidos en mejoras tecnológicas, capacitación del personal, bonificaciones por desempeño o expansión de servicios.

Es importante destacar que esta reducción de costos no implica un deterioro en la calidad del servicio. Por el contrario, el porcentaje de procesamiento mejora del 99.4% al 99.8%, indicando que el sistema optimizado opera con mayor eficiencia. La reducción de la ociosidad del 0.6% al 0.2% también sugiere una mejor distribución de la carga de trabajo entre los agentes, lo que potencialmente puede traducirse en menor estrés laboral y mayor satisfacción del personal.

Los resultados cuantitativos consolidados del estudio pueden resumirse en los siguientes indicadores clave:

- Precisión del pronóstico: MAPE del 10.25% con el modelo Holt-Winters Aditivo
- Tiempo medio entre llegadas: 247 segundos (4 min 7 seg)
- Tiempo medio de servicio: 535 segundos (8.9 minutos)
- Productividad efectiva por agente: 6-7 llamadas por hora en condiciones óptimas
- Número óptimo de agentes: 40 (reducción del 52.9% respecto a la dotación actual)
- Reducción de costos proyectada: 52.9% ($49,500 mensuales / $594,000 anuales)
- Mejora en eficiencia de procesamiento: del 99.4% al 99.8%

CONCLUSIONES

El presente estudio ha demostrado la viabilidad y efectividad de aplicar técnicas de simulación y análisis de series temporales para optimizar la gestión de recursos humanos en centros de atención al cliente. La metodología desarrollada permite abordar uno de los desafíos más complejos en la gestión de operaciones de servicio: el equilibrio entre la calidad del servicio y los costos operativos.

En primer lugar, se concluye que el método de Holt-Winters Aditivo representa la técnica de pronóstico más adecuada para predecir el volumen de llamadas en el contact center regional de Falabella. Su superioridad sobre la Red Neuronal LSTM, con una diferencia de 4 puntos porcentuales en el MAPE, se atribuye a su mayor robustez frente al sobreajuste y su capacidad para capturar eficientemente los componentes de nivel, tendencia y estacionalidad presentes en la serie temporal. Esta conclusión tiene implicaciones prácticas significativas, ya que sugiere que para horizontes de pronóstico de corto plazo (4 semanas), los métodos estadísticos clásicos pueden superar a técnicas de aprendizaje profundo más complejas.

En segundo lugar, la simulación del sistema de colas M/M/c reveló un hallazgo crítico: el contact center opera actualmente con un severo sobredimensionamiento de personal. Los resultados demuestran que con 40 agentes se obtiene el mismo nivel de output (332 llamadas) que con 85 agentes, pero con una eficiencia de procesamiento superior (99.8% vs 99.4%). Este fenómeno se explica porque el sistema está limitado por la tasa de llegadas de llamadas, no por la capacidad de servicio disponible. Agregar agentes más allá del punto óptimo no incrementa el throughput del sistema, solo genera costos adicionales sin valor agregado.

En tercer lugar, el análisis económico demuestra que la optimización del dimensionamiento de personal genera ahorros sustanciales sin comprometer la calidad del servicio. La reducción de 85 a 40 agentes representa un ahorro mensual de $49,500 (52.9%), equivalente a $594,000 anuales. El costo por llamada procesada se reduce de $282.5 a $132.5, una disminución del 53% que mejora significativamente la

eficiencia operativa del contact center. Estos resultados contradicen la noción tradicional de que mayor personal implica mejor servicio; por el contrario, el sobredimensionamiento genera ineficiencias tanto económicas como operativas.

En cuarto lugar, el análisis de los patrones temporales reveló información crítica para la programación de turnos. La concentración del 49.1% del volumen diario en las primeras horas de operación (9:00 a 11:00) y la distribución diferenciada entre días de la semana (19% los lunes vs 16% los viernes) proporcionan criterios objetivos para la asignación diferencial de recursos. Esta granularidad en el análisis permite evolucionar de un enfoque de dimensionamiento basado en promedios históricos a uno basado en la demanda real por franja horaria, optimizando aún más la distribución de los 40 agentes recomendados.

En quinto lugar, el estudio confirma la importancia de integrar múltiples técnicas analíticas para abordar problemas complejos de gestión de operaciones. La combinación de análisis de series temporales (Holt-Winters), teoría de colas (modelo M/M/c), pruebas de bondad de ajuste (Kolmogorov-Smirnov, Anderson-Darling) y simulación por eventos discretos proporciona una visión holística del sistema que no sería posible alcanzar mediante el uso aislado de cualquiera de estas técnicas. Esta aproximación metodológica integral permite validar las recomendaciones desde múltiples perspectivas, aumentando la confiabilidad de las decisiones gerenciales derivadas del análisis.

Finalmente, se recomienda considerar las siguientes líneas de trabajo futuro:

- Desarrollar un modelo específico para el período de diciembre, dado que los residuales de la descomposición temporal muestran comportamientos atípicos asociados a la alta demanda estacional de fin de año del Grupo Falabella.

- Implementar la reducción de personal de forma gradual, monitoreando los indicadores de servicio en tiempo real para validar que el sistema mantiene su desempeño con la nueva dotación.
- Diseñar un sistema de actualización automática del modelo de pronóstico que incorpore nuevos datos semanalmente, permitiendo ajustes dinámicos ante cambios en los patrones de demanda.
- Extender el análisis para considerar escenarios de contingencia donde la demanda supere significativamente los niveles históricos, estableciendo protocolos de activación de personal adicional.
- Realizar un análisis de sensibilidad que evalúe el impacto de variaciones en el costo por agente ($1,100) sobre la viabilidad económica de diferentes configuraciones de personal.

En síntesis, este estudio demuestra que la aplicación de metodologías cuantitativas rigurosas para la determinación del número óptimo de agentes en un contact center no solo es técnicamente factible, sino que genera beneficios tangibles extraordinarios. El ahorro anual proyectado de $594,000, combinado con la mejora en la eficiencia de procesamiento, evidencia que las decisiones de dimensionamiento basadas en intuición o promedios históricos pueden resultar en ineficiencias costosas. El modelo desarrollado proporciona una herramienta práctica y replicable que puede adaptarse a otros centros de contacto con características similares, contribuyendo así a la profesionalización de la gestión de operaciones de servicio en el contexto empresarial contemporáneo.

REFERENCIAS

[1] Salesforce. (2023). *State of the Connected Customer Report*. Salesforce Research. Recuperado de https://www.salesforce.com/resources/research-reports/state-of-the-connected-customer/

[2] Chen, Y., & Wang, L. (2023). Deep Learning Integration with Queueing Theory for Dynamic Resource Allocation in Call Centers. *Journal of Service Research*, 26(3), 412-429. https://doi.org/10.1177/10946705231156789

[3] Deloitte. (2023). *Global Contact Center Survey: The Future of Customer Service Operations*. Deloitte Insights. Recuperado de https://www2.deloitte.com/global/en/pages/operations/articles/global-contact-center-survey.html

[4] Akşin, Z., Armony, M., & Mehrotra, V. (2007). The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16(6), 665-688. https://doi.org/10.1111/j.1937-5956.2007.tb00288.x

[5] McKinsey & Company. (2023). *The State of Customer Care in 2023*. McKinsey Global Institute. Recuperado de https://www.mckinsey.com/capabilities/operations/our-insights/the-state-of-customer-care

[6] Tan, H., Zhang, W., & Liu, S. (2023). Simulation-Based Workforce Optimization in Contact Centers: A Comprehensive Study. *International Journal of Operations Management*, 45(2), 178-195. https://doi.org/10.1016/j.ijop.2023.02.015

[7] Koole, G., & Mandelbaum, A. (2002). Queueing Models of Call Centers: An Introduction. *Annals of Operations Research*, 113(1-4), 41-59. https://doi.org/10.1023/A:1020949626017

[8] Serper, G., Avramidis, A. N., & L'Ecuyer, P. (2019). Simulation-Based Optimization of Agent Scheduling in Multiskill Call Centers. *INFORMS Journal on Computing*, 31(2), 236-254. https://doi.org/10.1287/ijoc.2018.0826

[9] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5(2), 79-141. https://doi.org/10.1287/msom.5.2.79.16071

[10] Rodríguez, M., & Martínez, J. (2023). Optimización de recursos en contact centers: un enfoque basado en simulación. *Revista Latinoamericana de Investigación de Operaciones*, 12(1), 45-62.

[11] Erlang, A. K. (1917). Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges. *Elektroteknikeren*, 13, 5-13.

[12] Holt, C. C. (2004). Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. *International Journal of Forecasting*, 20(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

[13] Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, 6(3), 324-342. https://doi.org/10.1287/mnsc.6.3.324

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[15] Little, J. D. C. (1961). A Proof for the Queuing Formula: L = λW. *Operations Research*, 9(3), 383-387. https://doi.org/10.1287/opre.9.3.383