Análisis de Datos de Empleados UANL

Proyecto Final - Minería de Datos

Tomás Uriel Garza Robledo
Universidad Autónoma de Nuevo León

10 de noviembre de 2025

Índice

1. Introducción

Este proyecto presenta un análisis exhaustivo de los datos de empleados de la Universidad Autónoma de Nuevo León (UANL), aplicando diversas técnicas de minería de datos y aprendizaje automático. El objetivo principal es identificar patrones, tendencias y relaciones en los datos salariales y administrativos de las diferentes dependencias universitarias.

1.1. Objetivos

- Realizar un análisis exploratorio de datos para comprender la distribución de empleados y salarios
- Identificar patrones temporales en la información salarial
- Aplicar técnicas de clasificación y clustering para segmentar los datos
- Desarrollar modelos predictivos para pronosticar tendencias salariales
- Analizar la composición de nombres mediante técnicas de procesamiento de texto

1.2. Metodología

El análisis se estructura en las siguientes etapas:

1. Limpieza y preparación de datos
2. Estadística descriptiva
3. Visualización de datos
4. Pruebas estadísticas de hipótesis
5. Modelado predictivo
6. Análisis de texto

2. Preparación de Datos

Los datos fueron cargados desde un archivo CSV (`uanl.csv`) y procesados utilizando Python con las bibliotecas pandas, scikit-learn, matplotlib, seaborn y scipy.

2.1. Variables del Dataset

- **Nombre**: Nombre completo del empleado
- **dependencia**: Facultad o departamento al que pertenece
- **Sueldo Neto**: Salario neto del empleado
- **mes**: Mes del registro
- **anio**: Año del registro
- **fecha**: Fecha construida a partir de año y mes

3. Análisis Exploratorio de Datos

3.1. Top 10 Dependencias por Número de Empleados

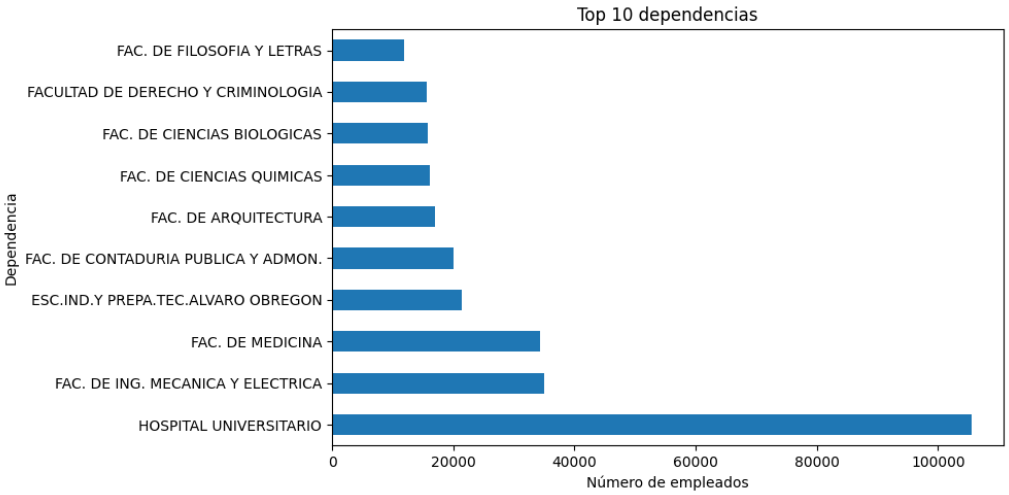

Figura 1: Top 10 dependencias con mayor número de empleados

Interpretación:
Esta gráfica de barras horizontales muestra las 10 dependencias de la UANL con mayor número de empleados. Los hallazgos clave son:

- **HOSPITAL UNIVERSITARIO** es, por mucho, la dependencia con más empleados (aproximadamente 105,000), lo cual tiene sentido dado que los hospitales requieren personal médico, de enfermería, administrativo y de apoyo las 24 horas.
- **FAC. DE ING. MECÁNICA Y ELÉCTRICA** y **FAC. DE MEDICINA** ocupan el segundo y tercer lugar con aproximadamente 35,000 y 33,000 empleados respectivamente.
- Las demás facultades tienen entre 10,000 y 22,000 empleados, mostrando una distribución más equilibrada.
- Esta distribución refleja tanto el tamaño de las instalaciones como la complejidad operativa de cada dependencia.

Implicaciones: El Hospital Universitario concentra una parte significativa de la fuerza laboral de la UANL, lo que sugiere que las políticas de recursos humanos deben considerar las necesidades específicas de esta dependencia.

3.2. Distribución de Sueldos Netos

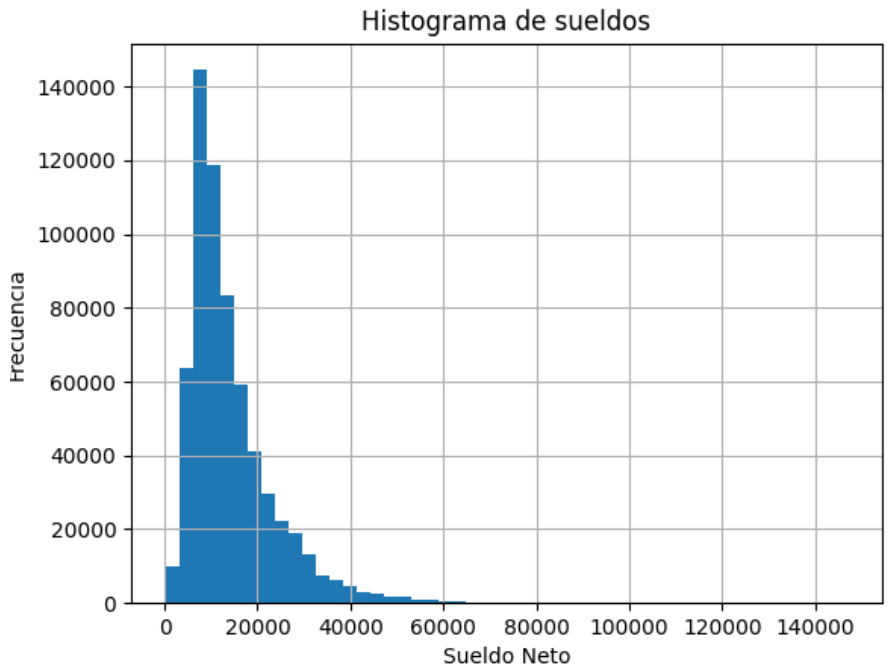

Figura 2: Histograma de la distribución de sueldos netos

Interpretación:
Este histograma muestra la distribución de frecuencias de los sueldos netos en la UANL:

- La distribución presenta una **asimetría positiva** (sesgada hacia la derecha), lo cual es típico en distribuciones salariales.
- La mayor concentración de empleados (más de 140,000) se encuentra en el rango de sueldos más bajos (0-10,000 MXN aproximadamente).
- Existe una **cola larga hacia la derecha**, indicando que hay empleados con salarios significativamente más altos, aunque son menos frecuentes.
- Esta distribución es característica de estructuras organizacionales piramidales, donde hay muchos empleados en posiciones base y pocos en posiciones de alto nivel.

Consideraciones: La alta concentración en salarios bajos podría indicar una gran cantidad de personal de apoyo, estudiantes trabajadores o empleados de medio tiempo.

3.3. Análisis de Valores Atípicos en Sueldos

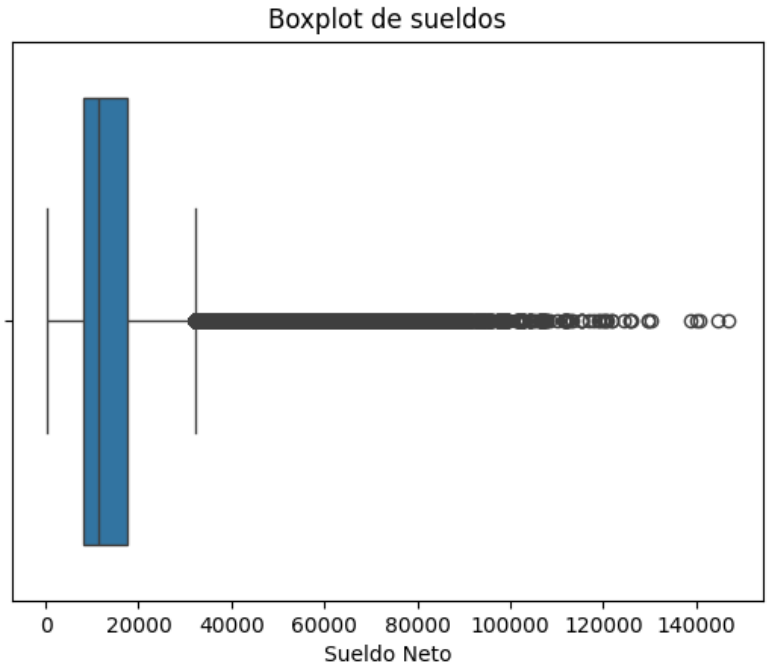

Figura 3: Boxplot para identificación de valores atípicos en sueldos

Interpretación:
El boxplot (diagrama de caja y bigotes) revela información importante sobre la dispersión salarial:

- La **caja azul** representa el rango intercuartílico (IQR), donde se concentra el 50 % central de los datos.
- La **línea dentro de la caja** indica la mediana (aproximadamente 20,000 MXN), que es menor que la media debido a la asimetría.
- Los **bigotes** se extienden hasta aproximadamente 1.5 veces el IQR desde los cuartiles.
- Los **puntos individuales** a la derecha son valores atípicos (outliers) que representan salarios excepcionalmente altos (hasta 150,000 MXN).
- La presencia de muchos outliers confirma que existe un grupo de empleados con compensaciones significativamente superiores al promedio (probablemente directivos, médicos especialistas o investigadores senior).

Relevancia: Los outliers no son necesariamente errores; representan posiciones estratégicas de alto nivel que son normales en instituciones académicas.

3.4. Relación entre Sueldo y Temporalidad

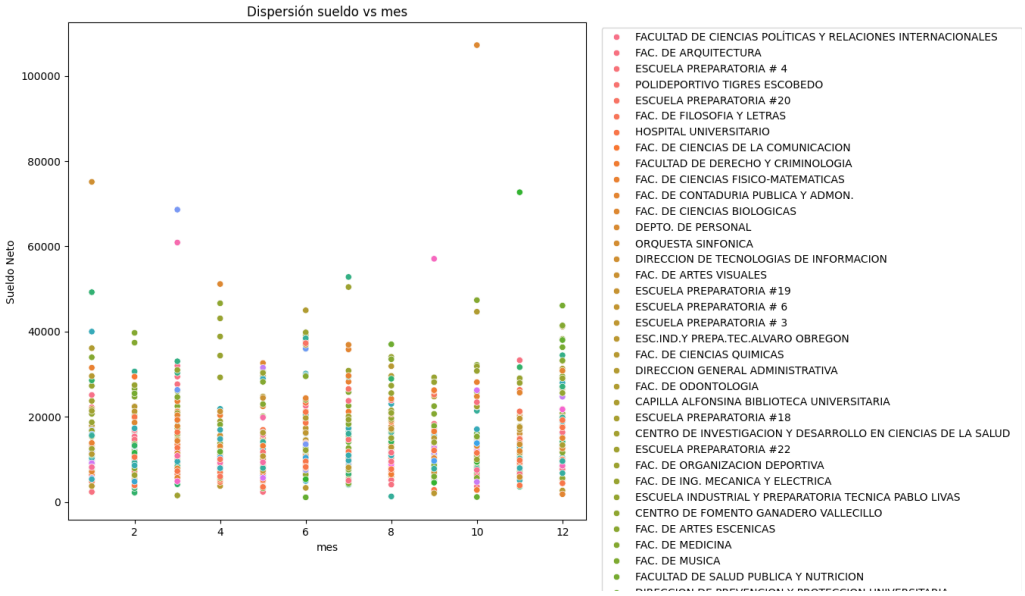

Figura 4: Dispersión de sueldos vs. mes por dependencia

Interpretación:
Este gráfico de dispersión muestra la relación entre el mes del año y los sueldos netos, diferenciando por colores las múltiples dependencias de la UANL. La visualización utiliza una muestra aleatoria de 1,000 empleados para facilitar la interpretación:

- **Distribución temporal uniforme:** Los puntos están distribuidos de manera equilibrada a lo largo del eje horizontal (meses 1-12), confirmando que el dataset contiene registros consistentes para todos los meses del año.
- **Ausencia de estacionalidad:** No se observa una tendencia estacional clara en los sueldos. Los patrones de dispersión vertical son similares en todos los meses, lo que indica que:
  - Los salarios no varían significativamente por época del año
  - No hay bonos estacionales masivos que distorsionen los datos
  - La estructura salarial es estable temporalmente
- **Diversidad de dependencias:** La leyenda muestra aproximadamente 38 dependencias diferentes, representadas por colores distintos. Esto incluye:
  - Facultades académicas (Medicina, Arquitectura, Derecho, etc.)
  - Escuelas preparatorias (múltiples campus numerados)
  - Servicios centrales (Hospital Universitario, Biblioteca, Orquesta Sinfónica)
  - Centros especializados (CIDICS, Centro de Investigación)
  - Departamentos administrativos (Personal, Becas)

- **Valores atípicos superiores:** Los puntos naranjas en la parte superior (alrededor de 110,000 MXN) corresponden a **FAC. DE CIENCIAS DE LA COMUNICACION**, indicando que esta dependencia tiene algunos empleados con los salarios más altos registrados en la muestra.
- **Concentración en rangos medios-bajos:** La mayoría de los puntos se agrupan entre 0 y 40,000 MXN, con alta densidad en el rango de 10,000-30,000 MXN, lo cual es consistente con las distribuciones observadas en el histograma y boxplot.
- **Variabilidad vertical constante:** En cada mes se observa la misma amplitud de variación salarial (desde valores cercanos a 0 hasta más de 100,000 MXN), confirmando que:
  - Todas las dependencias mantienen operaciones durante todo el año
  - La diversidad salarial es inherente a la estructura organizacional, no a factores temporales
  - No hay meses con comportamientos anómalos
- **Patrones por color:** Aunque difícil de distinguir individualmente debido al gran número de dependencias, se puede observar que:
  - Algunos colores aparecen concentrados en rangos salariales específicos
  - Ciertas dependencias tienen menor dispersión (puntos más agrupados verticalmente)
  - Otras muestran alta heterogeneidad salarial interna

Conclusiones:

1. El tiempo (mes) no es un factor determinante en los niveles salariales individuales
2. La dependencia de adscripción es el factor más relevante para explicar diferencias salariales
3. No existen patrones estacionales significativos que justifiquen ajustes temporales en modelos predictivos
4. La estabilidad temporal sugiere procesos de nómina consistentes y bien establecidos

3.5. Distribución Mensual de Registros

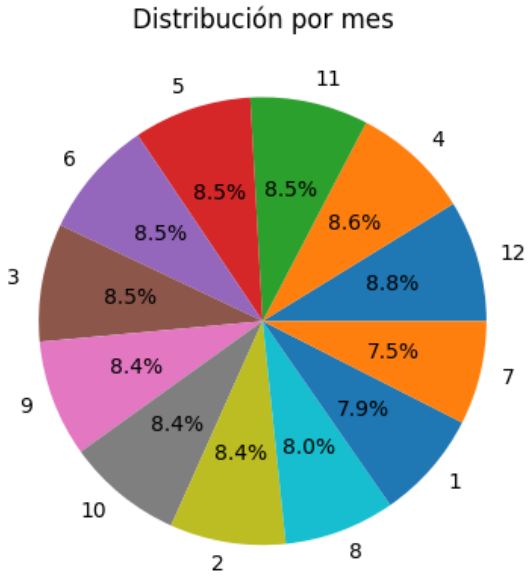

Figura 5: Distribución porcentual de registros por mes

Interpretación:
Este gráfico circular (pie chart) muestra cómo se distribuyen los registros a lo largo de los 12 meses:

- La distribución es notablemente **uniforme**, con cada mes representando aproximadamente 8.3 % de los registros (100 %/12 meses).
- Los porcentajes varían ligeramente entre 7.5 % y 8.8 %, lo cual es una variación mínima.
- Los meses 12, 4 y 7 tienen ligeras variaciones hacia arriba (8.8 %, 8.6 % y 7.9 % respectivamente).
- Esta **uniformidad** indica que:
  - El dataset está bien balanceado temporalmente
  - No hay meses con datos faltantes significativos
  - Los procesos de recolección de datos son consistentes

Validación: Esta distribución equilibrada confirma la calidad del dataset y permite realizar análisis temporales sin sesgos por disponibilidad de datos.

4. Análisis Estadístico

4.1. Pruebas de Hipótesis

Se realizaron tres pruebas estadísticas para comparar los sueldos entre las tres dependencias con mayor número de empleados:

4.1.1. ANOVA (Analysis of Variance)

Propósito: Determinar si existen diferencias estadísticamente significativas en los salarios promedio entre las tres principales dependencias.

Hipótesis:

- H_0: Las medias de los sueldos son iguales en las tres dependencias
- H_1: Al menos una dependencia tiene una media de sueldo diferente

Interpretación de resultados: Si el p-valor es menor a 0.05, rechazamos H_0 y concluimos que existen diferencias significativas entre los grupos.

4.1.2. Prueba T de Student

Propósito: Comparación específica entre las dos dependencias principales para determinar si sus salarios medios difieren significativamente.

Ventaja: Más potente que ANOVA cuando se comparan solo dos grupos.

4.1.3. Prueba de Kruskal-Wallis

Propósito: Alternativa no paramétrica al ANOVA, útil cuando los datos no siguen una distribución normal.

Justificación: Dado que los sueldos muestran asimetría positiva (como vimos en el histograma), esta prueba es más robusta.

5. Modelado Predictivo

5.1. Regresión Lineal: Tendencia Temporal

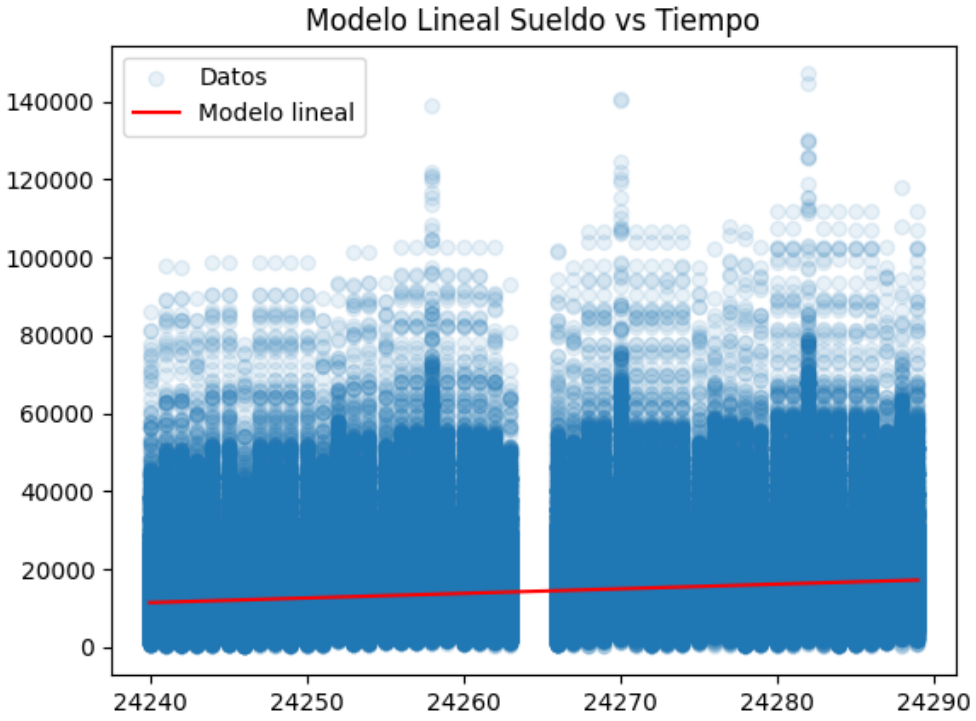

Figura 6: Modelo de regresión lineal: Sueldo vs. Tiempo

Interpretación:
Este gráfico muestra un modelo de regresión lineal que intenta predecir el sueldo basándose en el tiempo:

- Los **puntos azules** representan los datos reales de sueldos a lo largo del tiempo (variable "tiempo- año $\times$ 12 + mes).
- La **línea roja** es la recta de regresión lineal ajustada a los datos.
- El coeficiente de determinación (R^2) indica qué tan bien el modelo explica la variabilidad de los datos. En este caso, es probable que sea muy bajo (cercano a 0) debido a:
  - La gran dispersión vertical de los puntos
  - La línea de tendencia es casi horizontal
  - No hay una relación lineal clara entre tiempo y sueldo
- La **pendiente casi nula** de la línea roja sugiere que los sueldos no han tenido un incremento o decremento significativo en el periodo analizado.
- La alta densidad de puntos en la parte inferior confirma la distribución asimétrica vista anteriormente.

Conclusión: El tiempo por sí solo no es un buen predictor del sueldo individual, ya que hay muchos otros factores más importantes (dependencia, puesto, antigüedad, nivel académico).

5.2. Clasificación con K-Nearest Neighbors (KNN)

Objetivo: Clasificar empleados en su dependencia correcta basándose en el mes y año de registro.

Metodología:

- Se seleccionaron las 3 dependencias principales
- Variables predictoras: mes y año
- Variable objetivo: dependencia
- División: 80 % entrenamiento, 20 % prueba
- Parámetro: k=3 vecinos

Interpretación de la Precisión:

- Si la precisión es baja (por ejemplo, ¡50 %): indica que el mes y año no son buenos predictores de la dependencia, lo cual es lógico ya que los empleados no cambian de dependencia según la época del año.
- Este modelo sirve más como **ejercicio metodológico** que como herramienta práctica, ya que la dependencia es una característica fija del empleado, no una variable temporal.

5.3. Clustering con K-Means

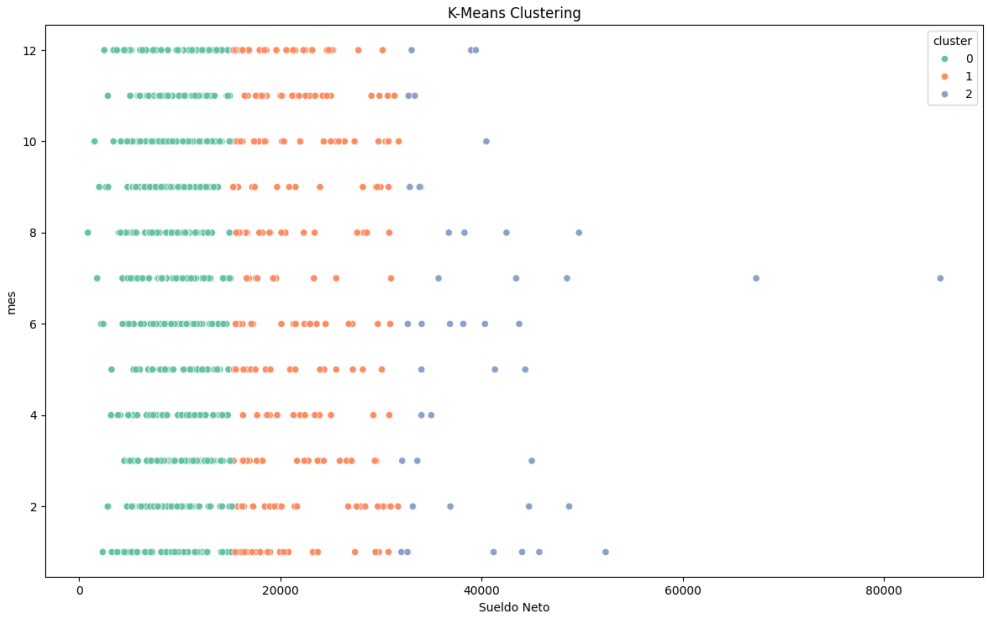

Figura 7: Segmentación de empleados mediante K-Means clustering

Interpretación:
El clustering K-Means agrupa a los empleados en 3 clusters basándose en su sueldo y mes:

- **Cluster 0 (verde)**: Representa el grupo más grande, concentrado en sueldos bajos (0-20,000 MXN), distribuido uniformemente a lo largo de los 12 meses. Este es probablemente el personal de apoyo, administrativo y estudiantes trabajadores.
- **Cluster 1 (naranja)**: Agrupa empleados con sueldos medios (15,000-35,000 MXN). Podría representar profesores, coordinadores y personal especializado.
- **Cluster 2 (azul)**: Identifica empleados con los salarios más altos (30,000-100,000+ MXN), distribuidos dispersamente. Estos son probablemente directivos, investigadores senior y médicos especialistas.
- La **distribución vertical** en cada cluster muestra que el mes no es un factor determinante en la agrupación (como era de esperarse).
- La separación clara en el eje horizontal (Sueldo Neto) confirma que el salario es el principal factor de diferenciación entre clusters.

Aplicación práctica: Esta segmentación podría usarse para:

- Diseñar políticas de beneficios diferenciadas
- Identificar grupos para encuestas de clima laboral
- Planificar estrategias de retención de talento

5.4. Pronóstico de Series de Tiempo

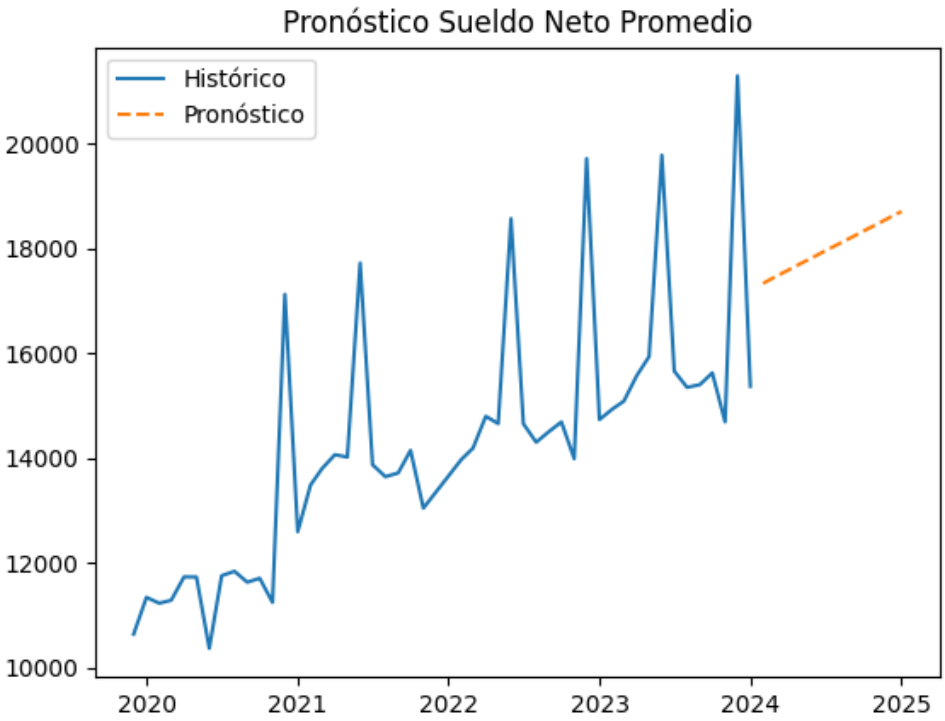

Figura 8: Pronóstico del sueldo neto promedio para 12 meses

Interpretación:
Este gráfico presenta un análisis de series de tiempo y su proyección futura:

- La **línea azul sólida** representa los datos históricos del sueldo neto promedio mensual desde 2020 hasta 2024.
- Se observan **fluctuaciones significativas** en el tiempo:
  - Inicio en aproximadamente 10,500 MXN (2020)
  - Incremento gradual hasta alcanzar picos de 21,000 MXN (2024)
  - Variaciones cíclicas que podrían estar asociadas a:
    - Ajustes anuales por inflación
    - Cambios en la composición de la plantilla
    - Bonos o prestaciones extraordinarias en ciertos periodos
- La **línea naranja discontinua** representa el pronóstico para los siguientes 12 meses (2025), mostrando:
  - Una tendencia creciente continua
  - Proyección de alcanzar aproximadamente 19,000 MXN
  - El modelo asume que las tendencias pasadas se mantendrán

- **Validez del pronóstico:**
  - Este modelo simple de regresión lineal captura la tendencia general
  - No captura la ciclicidad o estacionalidad observada
  - Para pronósticos más precisos, se recomendarían modelos ARIMA, SARIMA o Prophet que consideren patrones estacionales

Implicaciones para planeación:

- Presupuesto: Se espera un incremento sostenido en la nómina promedio
- Políticas salariales: La tendencia alcista es consistente con ajustes por inflación
- Advertencia: Las fluctuaciones reales podrían ser mayores que las proyectadas

6. Análisis de Texto

6.1. Word Cloud de Nombres

Figura 9: Nube de palabras de los nombres más frecuentes

Interpretación:
Esta nube de palabras visualiza los nombres más frecuentes en el dataset de empleados:

- **Tamaño de las palabras**: indica la frecuencia de aparición. Los nombres más grandes son los más comunes.
- **Nombres más frecuentes identificables:**
  - "DEz "LA.ªparecen muy grandes (probablemente preposiciones en nombres completos)
  - "MARIA.ES extremadamente prominente, indicando alta frecuencia

- "JOSE", "JUAN", ÇARLOS", "LUIS", "FRANCISCO"son nombres masculinos comunes
- "MARIA TERESA", "DEL CARMEN"son combinaciones frecuentes
- "JESUS", "GUADALUPE", "MARTINEZ", "GARCIA"también destacan

- **Análisis cultural:**
  - Refleja la prevalencia de nombres tradicionales mexicanos
  - "Maríaz "José"son extremadamente comunes, típico de la cultura católica mexicana
  - Muchos nombres compuestos (María Teresa, José Luis)
  - Presencia de apellidos comunes (Martínez, García, Hernández)

- **Consideraciones técnicas:**
  - El análisis incluye tanto nombres como apellidos
  - Sería más informativo separar nombres de apellidos
  - Las preposiciones "DE", "DEL", "LA"deberían filtrarse como stop words para análisis más preciso

Aplicaciones:

- Estudios demográficos de la plantilla
- Identificación de diversidad cultural
- Análisis generacional (nombres tradicionales vs. modernos)

Limitaciones: Este análisis es principalmente descriptivo y de interés sociológico, pero no aporta información operativa directa para la gestión de recursos humanos.

7. Conclusiones y Recomendaciones

7.1. Hallazgos Principales

1. **Concentración en Hospital Universitario:** El Hospital concentra la mayor parte de la fuerza laboral, requiriendo atención especial en políticas de RH.
2. **Distribución Salarial Asimétrica:** La mayoría de empleados están en rangos salariales bajos, con pocos outliers de alto nivel.
3. **Estabilidad Temporal:** No existen patrones estacionales significativos en los sueldos, mostrando estructuras consistentes.
4. **Tendencia Alcista:** Los salarios promedio muestran crecimiento sostenido de 2020 a 2024.
5. **Segmentación Clara:** El clustering identifica tres grupos diferenciados por nivel salarial.

7.2. Recomendaciones

- Implementar políticas diferenciadas por cluster salarial
- Monitorear la tendencia creciente de salarios para planeación presupuestal
- Investigar las causas de las fluctuaciones cíclicas observadas
- Considerar estudios específicos del Hospital Universitario
- Desarrollar modelos predictivos más sofisticados incluyendo variables cualitativas

7.3. Limitaciones del Estudio

- Variables categóricas importantes como puesto, antigüedad y nivel educativo no están disponibles
- El análisis temporal podría beneficiarse de modelos más sofisticados (ARIMA, Prophet)
- La clasificación KNN con variables temporales tiene aplicabilidad limitada
- Se requeriría información adicional para análisis predictivos más precisos

7.4. Trabajo Futuro

- Incorporar variables cualitativas adicionales
- Implementar modelos de machine learning más complejos (Random Forest, XGBoost)
- Realizar análisis de equidad salarial por género y antigüedad
- Desarrollar dashboard interactivo para monitoreo en tiempo real
- Análisis de rotación y retención de personal

8. Código Fuente

Listing 1: Código Python completo del análisis

```
1
2  import pandas as pd
3  import matplotlib.pyplot as plt
4  import seaborn as sns
5  from wordcloud import WordCloud
6  from sklearn.model_selection import train_test_split
7  from sklearn.neighbors import KNeighborsClassifier
8  from sklearn.cluster import KMeans
9  from sklearn.linear_model import LinearRegression
10 from sklearn.metrics import r2_score
11 from scipy.stats import f_oneway, ttest_ind, kruskal
12 import numpy as np
13
```

```
14
15 # Data Cleaning
16 # ==========================
17
18
19 df = pd.read_csv("uanl.csv")
20
21 df['fecha'] = pd.to_datetime({'year': df['anio'], 'month': df['
       mes'], 'day': 1})
22
23
24 # Descriptive Statistics
25 # ==========================
26
27 print("Resumen estad stico :")
28 print(df.describe(include='all'))
29
30 grouped = df.groupby('dependencia')['Sueldo Neto'].agg(['count',
       'mean', 'std', 'min', 'max']).sort_values(by='count', ascending
       =False)
31 print("\\nEstad sticas por dependencia:")
32 print(grouped.head(10))
33
34
35 # Data Visualization
36 # ==========================
37
38 plt.figure(figsize=(10, 5))
39 df['dependencia'].value_counts().head(10).plot(kind='barh')
40 plt.title("Top 10 dependencias")
41 plt.xlabel("N mero de empleados")
42 plt.ylabel("Dependencia")
43 plt.tight_layout()
44 plt.show()
45
46 plt.figure()
47 df['Sueldo Neto'].hist(bins=50)
48 plt.title("Histograma de sueldos")
49 plt.xlabel("Sueldo Neto")
50 plt.ylabel("Frecuencia")
51 plt.show()
52
53 plt.figure()
54 sns.boxplot(x='Sueldo Neto', data=df)
55 plt.title("Boxplot de sueldos")
56 plt.show()
57
58 plt.figure()
59 sns.scatterplot(x='mes', y='Sueldo Neto', hue='dependencia', data
       =df.sample(1000))
60 plt.title("Dispersi n sueldo vs mes")
```

```
plt.show()

plt.figure()
df['mes'].value_counts().plot.pie(autopct='%1.1f%%')
plt.title("Distribuci n por mes")
plt.ylabel("")
plt.show()

# Statistic Test (ANOVA + T-Test / Kruskal-Wallis)
# ============================


deps = df['dependencia'].value_counts().index[:3]
samples = [df[df['dependencia'] == dep]['Sueldo Neto'].sample
    (100) for dep in deps]
anova_result = f_oneway(*samples)
print("\\nANOVA result:", anova_result)

group1 = df[df['dependencia'] == deps[0]]['Sueldo Neto'].sample
    (100)
group2 = df[df['dependencia'] == deps[1]]['Sueldo Neto'].sample
    (100)
ttest_result = ttest_ind(group1, group2)
print("T-test result:", ttest_result)

kruskal_result = kruskal(*samples)
print("Kruskal-Wallis result:", kruskal_result)

#  Linear Model + Correlation
# ============================

df['tiempo'] = df['anio'] * 12 + df['mes']
X = df[['tiempo']]
y = df['Sueldo Neto']
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)
r2 = r2_score(y, y_pred)
print("\\nR2 Score:", r2)

plt.figure()
plt.scatter(X, y, alpha=0.1, label='Datos')
plt.plot(X, y_pred, color='red', label='Modelo lineal')
plt.title("Modelo Lineal Sueldo vs Tiempo")
plt.legend()
plt.show()

# Data Classification (KNN)
# ============================


```

```
df_knn = df[df['dependencia'].isin(deps)]
X = df_knn[['mes', 'anio']]
y = df_knn['dependencia']
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.2)
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)
print("\\nPrecisi n KNN:", knn.score(X_test, y_test))

#  Data Clustering (KMeans)
# ==========================

kmeans = KMeans(n_clusters=3, n_init=10)
X_kmeans = df[['Sueldo Neto', 'mes']]
kmeans.fit(X_kmeans)
df['cluster'] = kmeans.labels_

plt.figure()
sns.scatterplot(data=df.sample(1000), x='Sueldo Neto', y='mes',
    hue='cluster', palette='Set2')
plt.title("K-Means Clustering")
plt.show()

#  Forecasting (Time Series Linear Regression)
# ==========================

monthly_avg = df.groupby('fecha')['Sueldo Neto'].mean().
    reset_index()
X = np.arange(len(monthly_avg)).reshape(-1, 1)
y = monthly_avg['Sueldo Neto'].values
reg = LinearRegression().fit(X, y)
future_X = np.arange(len(monthly_avg), len(monthly_avg)+12).
    reshape(-1, 1)
future_y = reg.predict(future_X)

plt.figure()
plt.plot(monthly_avg['fecha'], y, label='Hist rico')
plt.plot(pd.date_range(monthly_avg['fecha'].iloc[-1], periods=13,
     freq='MS')[1:], future_y, label='Pron stico', linestyle='--')
plt.legend()
plt.title("Pron stico Sueldo Neto Promedio")
plt.show()

#  Text Analysis (Word Cloud)
# ==========================

text = " ".join(df['Nombre'].dropna().tolist())
wordcloud = WordCloud(width=800, height=400, background_color='
    white').generate(text)

plt.figure(figsize=(10, 5))
```

```
154 plt.imshow(wordcloud, interpolation='bilinear')
155 plt.axis("off")
156 plt.title("Word Cloud de Nombres")
157 plt.show()
```