



UNIVERSIDAD DEL AZUAY

DEPARTAMENTO DE POSGRADOS

MAESTRÍA EN ESTADÍSTICA APLICADA

Análisis económico de los sectores empresariales del Ecuador en el
periodo 2012--2023

Trabajo, previo a la obtención del título de: MAGÍSTER EN
ESTADÍSTICA APLICADA

Autor: Pedro Fernando Vizhco Sigua

Director: Prof. Luis Santiago Sarmiento Moscoso

Cuenca, Ecuador

2025

DEDICATORIA

Quiero dedicar este trabajo a mi familia, quienes me acompañaron a lo largo de este camino lleno de grandes desafíos. Siempre estuvieron ahí, motivándome y preocupándose por mi bienestar mientras trabajaba por alcanzar mis objetivos. A ellos les debo todo y les agradezco profundamente por su amor y apoyo incondicional.

También deseo dedicar este trabajo a la Lcda. Sarita Arpi y al Eco. Manuelito Guamán, personas muy importantes para mí, quienes me han hecho sentir parte de su familia y siempre han estado dispuestos a brindarme su ayuda cuando más la he necesitado.

A todos ellos, mi inmensa gratitud y sincero agradecimiento.

AGRADECIMIENTO

Expreso mi más sincero agradecimiento a Dios, quien me dio la fortaleza, fuerza, sabiduría y perseverancia para culminar esta etapa académica que inicié con dudas, pero lo estoy culminando ya.

Manifiesto mi gratitud al Mgst. Santiago Sarmiento, tutor de este proyecto integrador, por su orientación, compromiso y sugerencias brindadas que me guiaron en el desarrollo de este proyecto; así también a todo el personal docente, académico y directivo de la Universidad del Azuay, quienes me brindaron apoyo y contribución para fortalecer mi perfil profesional y humano. Finalmente, agradezco a mis compañeros de la maestría, con quienes compartí grandes experiencias en este camino de formación profesional y académico, sin dejar de lado lo personal.

RESUMEN

El comportamiento económico de los sectores empresariales es variable y en ocasiones sensible a sucesos imprevistos que afectan su dinámica. El presente estudio se centra en analizar la evolución de las empresas de carácter jurídico de los diferentes sectores económicos del Ecuador en el periodo 2006–2023, con el objetivo de identificar tendencias y patrones clave que contribuyan a la comprensión del desarrollo económico y laboral en estos sectores. Se inicia con un análisis descriptivo de las ventas, exportaciones y empleo; posteriormente, se aplican modelos de series de tiempo y técnicas k-means para identificar tendencias y agrupaciones sectoriales. Luego, se examinan las brechas de género en remuneraciones y plazas de empleo. Finalmente, se validan los modelos generados. Los resultados evidencian una afectación generalizada en todos los sectores económicos durante determinados periodos, así como desigualdades persistentes y comportamientos diferenciados entre sectores. Los datos utilizados para este estudio provienen del Instituto Nacional de Estadística y Censos (INEC).

PALABRAS CLAVE

Comportamiento económico, sectores empresariales, análisis descriptivo, series de tiempo, k-means, brecha de género, remuneración.

ABSTRACT

The economic behavior of business sectors is variable and sometimes sensitive to unforeseen events that affect their dynamics. This study focuses on analyzing the evolution of legally constituted companies across different economic sectors in Ecuador from 2006 to 2023, aiming to identify key trends and patterns that contribute to understanding economic and labor development in these areas. It begins with a descriptive analysis of sales, exports, and employment, followed by the application of time series models and *k-means* techniques to identify sectoral trends and clusters. Subsequently, gender gaps in wages and employment are examined. Finally, the generated models are validated. The results reveal that all economic sectors were affected during specific periods, as well as persistent inequalities and differentiated behaviors among them. The data used in this study were obtained from the National Institute of Statistics and Census (INEC).

KEYWORDS

Economic behavior, business sectors, descriptive analysis, time series, *k-means*, gender gap, remuneration.

ÍNDICE DE CONTENIDOS

<i>ÍNDICE DE FIGURAS</i>	v
<i>ÍNDICE DE TABLAS</i>	vi
<i>INTRODUCCIÓN</i>	7
<i>OBJETIVO GENERAL</i>	9
<i>LABORATORIO 1: ENFOQUE DESCRIPTIVO E INFERENCIAL</i>	10
Objetivos	10
Métodos	10
Resultados y Discusión	11
Conclusiones Parciales	21
Limitaciones del Estudio	21
<i>LABORATORIO 2: ENFOQUE MACHINE LEARNING</i>	23
Objetivos	23
Métodos	23
Resultados y Discusión	25
Series de Tiempo	25
Aprendizaje no Supervisado (Clustering)	44
Conclusiones Parciales	47
Limitaciones del Estudio	47
<i>LABORATORIO 3: ENFOQUE TOMA DE DESICIONES</i>	48
Objetivos	48
Métodos	48
Resultados y Discusión	50
Conclusiones parciales y Limitaciones del Estudio	55
<i>LABORATORIO 4: ENFOQUE DE ROBUSTEZ Y VALIDACIÓN</i>	56
Objetivo	56
Métodos	56
Discusión y Resultados	56
Conclusiones	61
Limitaciones del Estudio	61
<i>CONCLUSIÓN GENERAL</i>	62
<i>BIBLIOGRAFÍA</i>	63

ÍNDICE DE FIGURAS

Figura 1	Boxplot de las ventas totales con valores extremos	13
Figura 2	Boxplot de las ventas por sector económico.....	15
Figura 3	Comportamiento del logaritmo de ventas en el periodo 2012-2020.....	17
Figura 4	Boxplot de las remuneraciones totales por sectores económicos	18
Figura 5	Evolución de las remuneraciones en el periodo 2012-2020	20
Figura 6	Serie de Tiempo del sector de servicios 2012-2023	26
Figura 7	Descomposición de la Serie de tiempo del sector de servicios	26
Figura 8	Autocorrelación parcial de la serie diferenciada del sector de servicios	27
Figura 9	Autocorrelación de la serie diferenciada del sector de servicios.....	27
Figura 10	Pronósticos de la serie de tiempo del sector de servicios.....	28
Figura 11	Serie de Tiempo del sector de Agricultura 2012-2023.....	29
Figura 12	Descomposición de la serie de tiempo del sector de Agricultura	30
Figura 13	Autocorrelación parcial de la serie diferenciada del sector de Agricultura	30
Figura 14	Autocorrelación de la serie diferenciada del sector de Agricultura	30
Figura 15	Pronósticos de la Serie de tiempo del sector de Agricultura	31
Figura 16	Serie de Tiempo del sector de Construcción 2012-2023	32
Figura 17	Descomposición de la serie de Tiempo del sector de Construcción	33
Figura 19	Autocorrelación de la serie diferenciada del sector de Construcción.....	33
Figura 18	Autocorrelación parcial de la serie diferenciada del sector de Construcción	33
Figura 20	Pronósticos de la Serie de tiempo del sector de Construcción	34
Figura 21	Serie de Tiempo del sector de Comercio 2012-2023	35
Figura 22	Descomposición de la serie de Tiempo del sector de Comercio.....	36
Figura 23	Autocorrelación parcial de la serie diferenciada del sector de Comercio.....	36
Figura 24	Autocorrelación de la serie diferenciada del sector de Comercio	36
Figura 25	Pronósticos de la serie de tiempo del sector de Comercio	37
Figura 26	Serie de Tiempo del sector de Explotación de Minas 2012-2023	38
Figura 27	Descomposición de la serie de Tiempo de sector de Explotación de Minas.....	39
Figura 28	Autocorrelación de la serie diferenciada del sector de Explotación de Minas.....	39
Figura 29	Autocorrelación de la serie diferenciada del sector de Explotación de Minas.....	39
Figura 30	Pronósticos de la serie de tiempo del sector de Explotación de minas.	40
Figura 31	Serie de Tiempo del sector de Industrias Manufactureras 2012-2023.....	41
Figura 32	Descomposición de la serie de Tiempo de sector de Explotación de Minas.....	42
Figura 33	Autocorrelación de la serie diferenciada del sector de Industrias Manufactureras	42
Figura 34	Autocorrelación de la serie diferenciada del sector de Industrias Manufactureras. ...	42
Figura 35	Pronósticos de la serie de tiempo del sector de Industrias Manufactureras	43
Figura 36	Clusters de Remuneraciones, Ventas totales y Plazas Totales.....	45
Figura 37	Clusters de Plazas Hombres vs Remuneración.....	46
Figura 38	Clusters de plazas Mujeres vs Remuneración	46
Figura 39	Mapa de Calor anual sobre el índice de la brecha de plazas de empleo por sector económico	50
Figura 40	Diagrama de Pareto de la brecha de empleo por género en los sectores económicos	51
Figura 41	Mapa de Calor anual sobre el índice de brecha de remuneración por género y por sector económico	52
Figura 42	Diagrama de Pareto de la brecha de remuneración de género por cada subsector del sector económico.	53

ÍNDICE DE TABLAS

Tabla 1	Variables de la Base de Datos.....	10
Tabla 2	Tabla de datos faltantes y sus porcentajes.....	12
Tabla 3	Prueba de Kolmogorov Smirnov para la variable Ventas Totales.....	13
Tabla 4	Medianas de las ventas por sectores económicos.....	14
Tabla 5	Prueba Dunn de las ventas por sector económico.....	16
Tabla 6	Medianas del logaritmo de ventas por sectores económicos en el periodo 2012-2020.....	16
Tabla 7	Medianas del logaritmo de remuneraciones por sector económico.....	18
Tabla 8	Prueba Dunn de las remuneraciones por sector económico.....	19
Tabla 9	Medianas de la remuneración total por género.....	20
Tabla 10	Descripción de las variables utilizadas para el estudio de Series de tiempo y Aprendizaje no supervisado.....	23
Tabla 11	Métricas de error para los diferentes modelos del sector de servicio.....	28
Tabla 12	Coeficientes del modelo SARIMA del sector de Servicios.....	29
Tabla 13	Métricas de error para los diferentes modelos del sector de Agricultura.....	31
Tabla 14	Coeficientes del modelo AR del sector de Agricultura.....	32
Tabla 15	Métricas de error para los diferentes modelos del sector de Construcción.....	34
Tabla 16	Coeficientes del modelo SAR2 del sector de Construcción.....	35
Tabla 17	Métricas de error para los diferentes modelos del sector de Comercio.....	37
Tabla 18	Coeficientes del modelo ARMA del sector de Comercio.....	38
Tabla 19	Métricas de error para los diferentes modelos del sector de Explotación de minas....	40
Tabla 20	Coeficientes del modelo SARIMA del sector de Explotación de Minas.....	41
Tabla 21	Métricas de error para los diferentes modelos del sector de Industrias Manufactureras.....	43
Tabla 22	Coeficientes del modelo SARIMA del sector de Industrias Manufactureras.....	44
Tabla 23	Centroides de los clústers generados.....	45
Tabla 24	Variables utilizadas en la toma de decisiones.....	48
Tabla 25	51
Tabla 26	Brecha total de remuneración acumulativa por género en cada sector económico....	53
Tabla 27	Matriz de Confusión de la clasificación de las empresas.....	54
Tabla 28	Métricas estadísticas por tamaño de empresa.....	54
Tabla 29	Métricas estadísticas del modelo global de clasificación.....	55
Tabla 30	Resultados de los supuestos de validez para la serie de tiempo del sector de Servicios.....	56
Tabla 31	Resultados de los supuestos de validez para la serie de tiempo del sector de Agricultura.....	57
Tabla 32	Resultados de los supuestos de validez para la serie de tiempo del sector de Construcción.....	57
Tabla 33	Resultados de los supuestos de validez para la serie de tiempo del sector de Comercio.....	58
Tabla 34	Resultados de los supuestos de validez para la serie de tiempo del sector de Industrias Manufactureras.....	59
Tabla 35	Resultados de los supuestos de validez para la serie de tiempo del sector de Explotación de Minas.....	59
Tabla 36	Índices de Cohesión interna y Separación Interna.....	60
Tabla 37	Índice Davies- Bouldin.....	60
Tabla 38	Coeficientes de correlación de variables.....	61

INTRODUCCIÓN

El desarrollo de un país siempre está vinculado con el objetivo de convertirse en potencia mundial y para ello hay varios factores que son importantes en este camino. El analizar el comportamiento económico de los diferentes sectores empresariales ayuda a comprender cómo se da el desarrollo de las naciones y países; pues las diferentes empresas generan rentabilidad, recursos y contribuyen a la economía nacional. Algunas investigaciones en Ecuador han permitido ver cómo este comportamiento tiene influencia directa en la distribución de la riqueza y el bienestar de la población, y lo más importante, identificar qué sectores empresariales presentan mayor fuerza y apoyo en las diferentes economías de la zona. (Moreira y otros, 2020)

La economía ecuatoriana, a pesar de que ha tenido avances, todavía sigue dependiendo de los ingresos que provienen de la venta y exportación de productos primarios que no aportan con un gran valor agregado, trayendo como consecuencia que el desempeño productivo del país se mantenga estancado o presente avances muy lentos, dificultando su evolución. (Proaño y otros, 2019)

Según Cevallos (2021) en Ecuador, los diferentes sectores económicos han aportado, en diferentes porcentajes, al crecimiento del PIB, dinamizando el mercado laboral presentándose en el siguiente orden: industrias manufactureras (11,84%), comercio (10,50%), petróleo y minas (10,38%), construcción (8,57%), enseñanza servicios sociales y de salud (8,22%), agricultura (7,86%), otros servicios (7,37%) y transporte (6,73%); toda esta información analizada desde el año 2000 hasta 2018. Con los datos se puede ver que todos estos sectores aportan en un 71,47% el crecimiento del PIB. Por lo tanto, es importante analizar el comportamiento individual de los diferentes sectores, pues permitirá comprender e identificar patrones de crecimiento y tomar decisiones sobre políticas a aplicar para mejorar su desempeño.

En la última década, el país ha cruzado por varias transformaciones en el ámbito económico, político y social que ha llegado a tener un impacto en el rendimiento económico. Las protestas en el año 2019, así como la crisis sanitaria del COVID-19 en el año 2020 son factores que han podido afectar la evolución de los diferentes sectores económicos, permitiendo también, reconocer que el país no estuvo preparado para enfrentar crisis de índoles similares. (Becerra y otros, 2021)

El objetivo que tiene este proyecto se centra en analizar el comportamiento de los sectores empresariales ecuatorianos y su evolución a través del tiempo desde 2006 al 2023, enfatizándose en los sectores de: servicios, agricultura, comercio, construcción, minería e industrias manufactureras. Este análisis identificará patrones de crecimiento, productividad e índices de empleo que han afectado el desempeño de cada sector.

Estudios similares, contrastan la importancia de realizar análisis de la misma índole. En el sector de servicios, Cueva (2022) muestra que en 1990 existía una amplia brecha en los niveles de productividad entre el sector de servicios y el de industrias manufactureras, pero conforme

pasó el tiempo esa diferencia se ha ido incrementando, a favor del sector de industrias, mientras que el sector de servicios presenta un comportamiento decreciente en el transcurso del tiempo. Por otra parte, el sector de agricultura es inferior al resto de sectores económicos, pero eso no implica que muestre un crecimiento constante, aunque lento.

El sector industrial manufacturero se ha convertido en uno de los pilares fundamentales del desarrollo económico, pues en años anteriores el Ecuador ha sido dependiente del sector de agricultura y extracción de petróleo. De acuerdo con Sánchez & Soriano (2021), la dolarización, fue uno de los momentos, en donde el sector manufacturero presentó un incremento en las exportaciones y un alza en la balanza comercial, trayendo así un crecimiento económico y creación de empleo contribuyendo al aumento de los estándares de la sociedad.

La agricultura ha sido una actividad esencial en toda economía, teniendo un papel fundamental en la seguridad alimentaria y el crecimiento económico. Según Hernández & Bravo (2025), en Ecuador, este sector ha sido imprescindible para el abastecimiento del mercado interno y ha tenido su participación en el mercado internacional; por otra parte, este sector ha representado el 30% de la mano de obra a nivel nacional. Sin embargo, este sector ha enfrentado varios desafíos, con en 2017 la falta de incentivos afectó a productores pequeños y medianos, y en 2020, la pandemia del COVID-19 introdujo restricciones que complicaron mucho más sus actividades, generando una dinámica muy variable.

Estos estudios de diferentes autores reflejan que el sector empresarial ecuatoriano labora bajo condiciones socioeconómicas complejas y muy variables, que hacen pertinente su estudio, permitiendo así tomar diferentes decisiones para mejorar su desempeño, afrontar crisis espontáneas y lograr alcanzar un nivel competitivo en el mercado internacional.

De esta manera, el estudio proporciona información relevante sobre el comportamiento de los diferentes sectores económicos y los factores que son más notorios en su dinámica a través del tiempo. El problema central se basa en realizar un estudio que analice de manera conjunta la evolución de las plazas de empleo por sector económico, así como las brechas salariales y de empleo según el género. Con esto se logrará comprender, con una mayor precisión, como ha sido el cambio y dinamismo que han mostrado los diferentes sectores económicos.

El presente documento se compone de 4 laboratorios: el Laboratorio 1 el análisis exploratorio de la base de datos utilizada para el estudio; el Laboratorio 2 aplica técnicas de series de tiempo y machine learning para analizar las plazas de empleo disponibles en los diferentes sectores económicos y cómo se pueden realizar diferentes agrupaciones considerando factores como plazas de empleo, remuneraciones, ventas totales, exportaciones, etc; el Laboratorio 3 utiliza un enfoque para la toma de decisiones en base a las plazas de empleo por género y el Laboratorio 4 expone la validación y robustez de los modelos generados en el estudio.

OBJETIVO GENERAL

Analizar la evolución de las empresas jurídicas que llevan contabilidad en cinco sectores económicos: Agricultura, ganadería, silvicultura y pesca, Comercio, Construcción, Explotación de Minas y Canteras, Industrias Manufactureras y Servicios; evaluando su desempeño en ventas, exportaciones y generación de empleo total y por género desde 2006 hasta 2023, con el fin de identificar tendencias y patrones clave que contribuyan a la comprensión del desarrollo económico y laboral en estos sectores.

LABORATORIO 1: ENFOQUE DESCRIPTIVO E INFERENCIAL

La base de datos recopila información estadística sobre la estructura empresarial ecuatoriana a partir de registros administrativos, registro de movimientos económicos, personal ocupado y afiliado. Se inició con el estudio desde el periodo 2006 hasta el 2023, pero al momento de cargar las bases de datos, se identificó que la base de datos correspondiente al periodo 2006-2011 contenía errores en sus registros, por lo cuál para todos los laboratorios siguientes, incluido el laboratorio 1, se excluyó este periodo.

Objetivos

- Comparar el desempeño de los diferentes sectores económicos de Ecuador en términos de ventas totales durante el periodo 2012-2023.
- Analizar la evolución de las ventas totales por sector económico en Ecuador durante el periodo 2012-2023.
- Cuantificar la diferencia salarial entre hombres y mujeres en las empresas de carácter jurídico del Ecuador.

Métodos

Descripción de las variables de la base de Datos.

La base de datos constaba de 96 variables y 1048575 observaciones, de allí se generó una nueva base de datos para trabajar con un total de 19 variables que se describen en la Tabla 1:

Tabla 1

Variables de la Base de Datos

Variable	Escala de medición
Año	Año en el que se realizó la observación. (Numérico)
Tipo_uni_legal	Tipo de contribuyente: persona natural o jurídica. (Categórico)
Contabilidad	La institución lleva o no contabilidad. (Dicotómica)
Sector Económico	Sector económico en el que se desarrolla la empresa. (Categórico)
Ventas_Totales	Valores registrados de las ventas del año. Medida en dólares. (Numérico)
Ventas_Nacionales	Valores registrados de las ventas del año a nivel nacional. Medida en dólares. (Numérico)
Exportaciones	Valores registrados de las exportaciones del año. Medida en dólares. (Numérico)
Plazas_reg_total	Número de plazas totales disponibles registradas. (Numérico)
Plazas_reg_total_H	Número de plazas para hombres registradas. (Numérico)

Plazas_reg_total_M	Número de plazas para mujeres registradas. (Numérico)
Empleados_reg_total	Número de empleados totales registradas. (Numérico)

Variable	Escala de medición
Empleados_reg_H	Número de empleados hombres registradas. (Numérico)
Empleados_reg_M	Número de empleados mujeres registradas. (Numérico)
Estrato_ventas_total	Clasificación de las empresas por estratos de acuerdo con sus ventas. (Categórico)
Estrato_plazas_total	Clasificación de las empresas por el número de plazas registradas. (Categórico)
Estrato_empleo_total	Clasificación de las empresas por el número de empleados registrados. (Categórico)
Remuneración	Valores registrados sobre las remuneraciones totales dadas. (Numérico)
Remuneración_H	Valores registrados sobre las remuneraciones totales dadas al sector masculino. (Numérico)
Remuneración_M	Valores registrados sobre las remuneraciones totales dadas al sector femenino. (Numérico)

Procesamiento y Análisis de los Datos

- Filtrado de Datos: Se realizó un filtrado de datos con el fin de enfocarse en el sector de contribuyente jurídico, reduciendo así la base de 1048575 observaciones a 93031.
- Manejo de datos Faltantes: Mediante el uso de histogramas y tablas de porcentaje, se identificó la cantidad de datos faltantes y se procedió a eliminarlos de acuerdo a diferentes criterios del investigador, los cuales se detallan en los resultados del estudio.
- Manejo de datos atípicos: Se creó diagramas de cajas y bigotes para la identificación de datos atípicos, así también se aplicó pruebas de normalidad para analizar el comportamiento de los datos y escoger de forma correcta las diferentes pruebas paramétricas o no paramétricas a aplicar.
- Coeficientes de Correlación: De acuerdo con la distribución de los datos, se aplicó las diferentes pruebas de correlación para establecer la relación entre variables.

Resultados y Discusión

Manejo de Datos Faltantes

Se inició con la identificación de datos faltantes y el porcentaje que representa, teniendo así los siguientes resultados:

Tabla 2

Tabla de datos faltantes y sus porcentajes.

Variable	N° datos Faltantes	Porcentaje de Datos Faltantes(%)
Ventas_Totales	15155	12,5
Ventas_Nacionales	15143	12,5
Exportaciones	15361	12,7
Plazas_reg_total	18537	15,3
Plazas_reg_total_H	31004	25,6
Plazas_reg_total_M	42015	34,8
Empleados_reg_total	21437	17,7
Empleados_reg_H	31004	25,6
Empleados_reg_M	44035	36,4
Estrato_ventas_total	30576	25,3
Estrato_plazas_total	18537	15,3
Estrato_empleo_total	21437	17,7
Remuneración	18537	15,3
Remuneración_H	31004	25,6
Remuneración_M	42015	34,8

Como se puede observar en la Tabla 2, al tener un alto porcentaje de datos faltantes, se procedió a realizar un conteo de datos faltantes por sector económico y por cada año, generando tablas similares a la “Tabla 2”.

De acuerdo con las tres formas que se ha utilizado para visualizar los datos faltantes, en todos ellos el porcentaje de datos faltantes es elevado. Cuando se realiza un análisis estadístico, cada observación no debe superar el 40% de variables con datos faltantes (> 7 columnas), caso contrario dicha observación deberá ser eliminada; acción que se realizó. Finalmente, aquellas observaciones que no están dentro del 40%, se les mantendrá sus datos faltantes, pues estos no son resultado de errores de registro, sino más bien situaciones que por diferentes razones, las empresas no quisieron declarar, lo que podría estar involucrando situaciones legales irregulares que deberán ser investigadas.

Identificación de Valores Extremos y Pruebas de Normalidad

En el análisis de los valores extremos, los datos faltantes no se considerarán debido a que pueden generar inconsistencias en los resultados; por el contrario, las variables que involucran valores numéricos como: ventas totales y remuneraciones, se les aplicará una conversión logarítmica con el fin de controlar la dimensión de los datos y trabajar con una escala manejable. Esta medida se

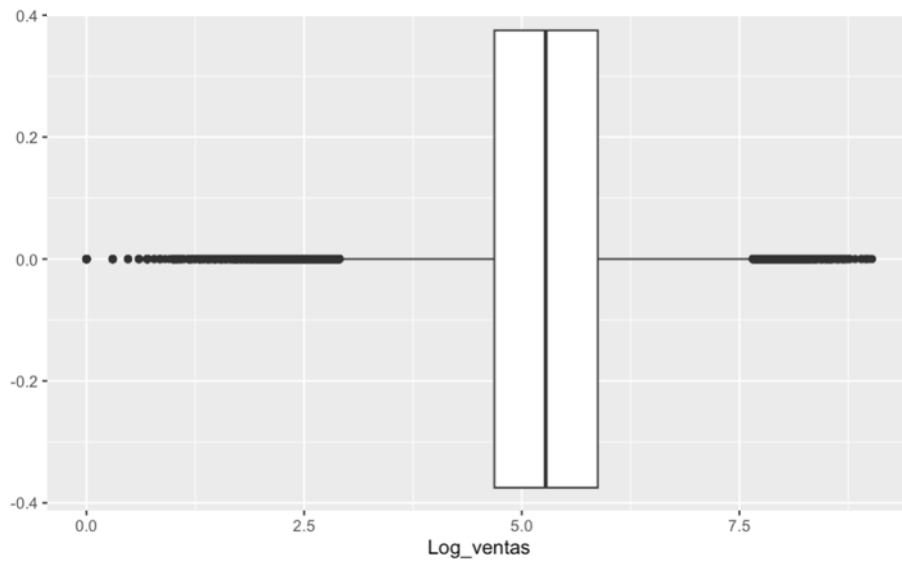
toma debido a que las observaciones registradas en estas variables son valores numéricos muy grandes y su visualización de forma gráfica no permite identificar su distribución. Para aquellos datos que tenían un valor de 0, se les procedió a sumarle a esa observación un valor de 0.001, con el fin de obtener un valor definido para el logaritmo.

Se procedió a crear boxplots para las variables numéricas y aplicar la prueba Kolmogorov-Smirnov para determinar el comportamiento que siguen los datos.

Ventas Totales

Figura 1

Boxplot de las ventas totales con valores extremos



La Figura 1 muestra la distribución que siguen los datos, identificando su mediana por un valor cercano a 5. Existen datos atípicos por ambos extremos de la distribución, presentando una mayor acumulación en el extremo izquierdo, es decir que presenta un sesgo en este sentido. Para establecer si estos datos siguen una distribución normal se procede a desarrollar pruebas de normalidad.

Tabla 3

Prueba de Kolmogorov Smirnov para la variable Ventas Totales

Hipótesis de Normalidad de los Datos	
h ₀ =Los datos siguen una distribución normal.	
h _a =Los datos siguen una distribución no normal.	
Variable	Prueba Kolmogorov-Smirnov
Ventas Totales	$D = 0.81694, p - value < 2.2e - 16$

En la Tabla 3 se pueden observar los resultados de la prueba Kolmogorov Smirnov, en donde se concluye que los datos no siguen una distribución normal, pues se obtuvo un p -value menor a 0,05 que es el nivel de significatividad.

De manera similar, se ha trabajado con los boxplots y pruebas de Kolmogorov Smirnov de las variables: remuneración total, remuneración del sector masculino, remuneración del sector femenino. Se ha identificado una distribución no normal en los datos al obtener un p -value $< 2.2e - 16$; se han identificado datos extremos en ambos lados de la distribución, con excepción de las variables que involucran las plazas de empleo, quienes presenta una distribución sesgada hacia la derecha con una concentración de los datos en valores bajos. Los datos extremos de estas variables no han sido eliminados, pues la ausencia de registros puede significar situaciones de evasión de impuestos, incoherencias en declaraciones de impuestos, aseguraciones sociales de los empleados entre otras; que son motivo de investigaciones severas en dichas empresas.

Comparación del desempeño económico de los diferentes sectores económicos en términos de Ventas.

Se inicia con la comparación de las medianas de los valores logarítmicos de las ventas con el fin de determinar una similitud entre estas medidas, obteniendo así los siguientes resultados:

Tabla 4

Medianas de las ventas por sectores económicos.

Sector Económico	Mediana
Agricultura, ganadería, silvicultura y pesca	5,585
Comercio	5,508
Construcción	4,886
Explotación de Minas y Canteras	4,800
Industrias Manufactureras	5,577
Servicios	4,793

En la Tabla 4, al obtener las medianas de los diferentes sectores económicos, se puede inferir que los sectores de agricultura, comercio e industrias manufactureras son las que presentan una cercana similitud en su nivel de ventas, por el contrario, los sectores de construcción, explotación de minas y servicios muestran una cercana similitud. Con el boxplot que se genere, se podrá identificar de forma más clara estas inferencias.

Figura 2

Boxplot de las ventas por sector económico

Podemos notar, que de acuerdo con la Figura 2, las medianas del logaritmo de las ventas totales tienen a ser posiblemente similares entre algunos sectores, tal como lo inferíamos en la Tabla 4. Esto se podrá concluir de forma más concisa al aplicar una prueba de Kruskal Wallis, en donde se establecen las siguientes hipótesis:

H_0 =Los sectores económicos presentan igualdad en sus ventas del periodo 2012-2020.

H_a =Los sectores económicos presentan diferencias significativas en sus ventas totales del periodo 2012-2020.

Obteniendo en la prueba de Kruskal Wallis un $p - value < 2.2e - 16$, se puede concluir que las medianas de las ventas totales de los diferentes sectores económicos presentan diferencias significativas. Para ser más específico, se aplicará una prueba Dunn con el fin de identificar que sectores específicos poseen diferencias.

Tabla 5

Prueba Dunn de las ventas por sector económico

Sector Económico	Agricultura	Comercio	Construcción	Explotación	Industrias
Comercio	-0,247712 1,00000				
Construcción	31,21230 0,0000*	43,94378 0,0000*			
Explotación	14,74321 0,0000*	15,86534 0,0000*	-0,987168 1,0000		
Industrias	-3,405756 0,0049*	-4,316535 0,001*	-38,57460 0,0000*	-16,93163 0,0000*	
Servicios	40,87381 0,0000*	75,40814 0,0000*	4,395550 0,0001*	2,592027 0,0716	52,84019 0,0000*

De acuerdo a los resultados obtenidos en la prueba Dunn de la Tabla 5, se puede concluir que todos los sectores presentan diferencias significativas en sus niveles de ventas totales con excepción de comercio con agricultura y servicio con explotación minera, los cuales al presentar un valor mayor a 0,05, se concluyen que no hay una diferencia significativa en sus niveles de ventas.

Evolución de las ventas totales por sector económico en Ecuador durante el período 2012-2020.

Se realizó el cálculo de las medianas del logaritmo de ventas por cada sector de acuerdo a cada año, especificados en la Tabla 6 que se presenta a continuación:

Tabla 6

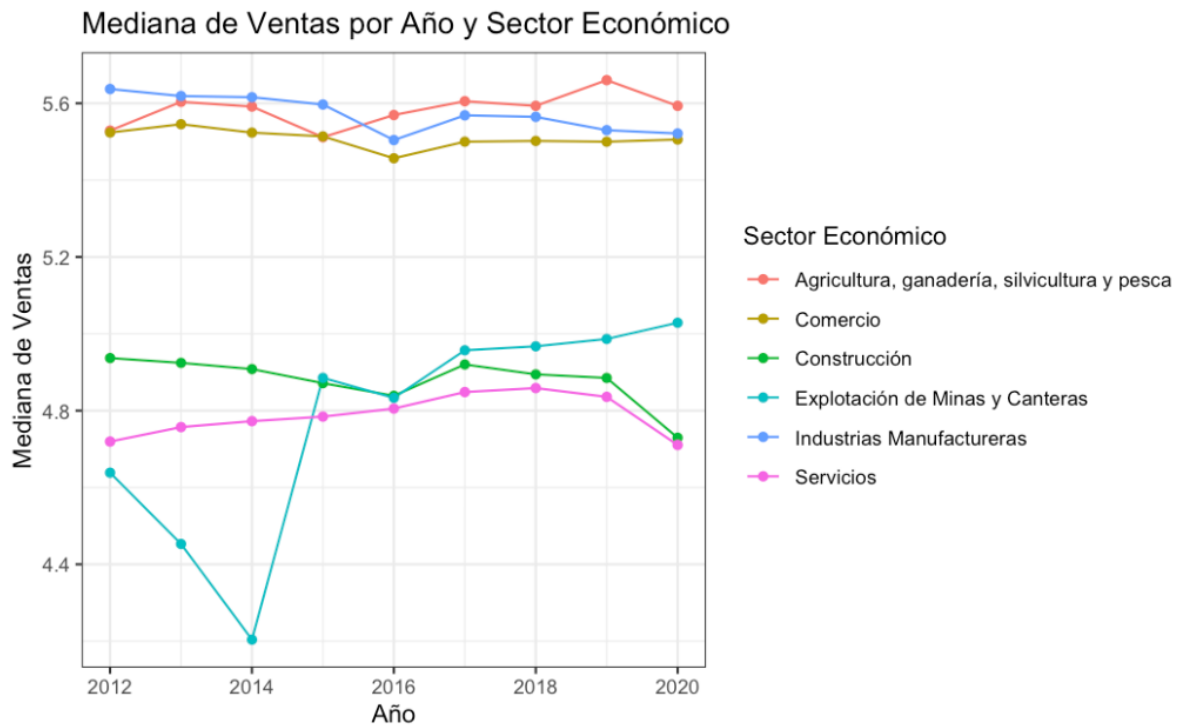
Medianas del logaritmo de ventas por sectores económicos en el periodo 2012-2020

Año	Agricultura	Comercio	Construcción	Explotación	Industrias	Servicios
2012	5,528	5,524	4,936	4,638	5,637	4,719
2013	5,603	5,545	4,924	4,453	5,618	4,757
2014	5,591	5,523	4,908	4,203	5,615	4,772
2015	5,511	5,513	4,871	4,885	5,596	4,784
2016	5,57	5,466	4,838	4,833	5,504	4,805
2017	5,61	5,500	4,920	4,957	5,568	4,848
2018	5,59	5,502	4,894	4,967	5,564	4,858
2019	5,66	5,500	4,885	4,986	5,530	4,835
2020	5,59	5,505	4,729	5,028	5,521	4,710

Para poder visualizar mejor su comportamiento, la Figura 3 nos permitirá comprender mejor dicho comportamiento:

Figura 3

Comportamiento del logaritmo de ventas en el periodo 2012-2020



Se pudo observar en la Figura 3 que el nivel de ventas en 2012 está liderado por la industria, seguido del comercio y agricultura que son casi similares. La evolución de las ventas en estos sectores se ha mantenido en un nivel estable, con algunos aumentos y descensos poco pronunciados. Entre estos tres sectores, el que ha terminado con un nivel más alto en 2020 es la agricultura.

Por otra parte, la construcción, servicios y explotación empezaron con un nivel de ventas más bajo que los otros sectores, siendo el sector de explotación el que más variaciones ha sufrido, pero termina siendo el más alto de entre estos tres sectores. Por el contrario, en 2020 la construcción y los servicios terminan con el nivel más bajo en ventas.

Comparación entre las remuneraciones totales por los diferentes sectores económicos.

Se inicia con la comparación de las medianas de los valores logarítmicos de las remuneraciones con el fin de determinar una similitud entre estas medidas, obteniendo así los siguientes resultados:

Tabla 7

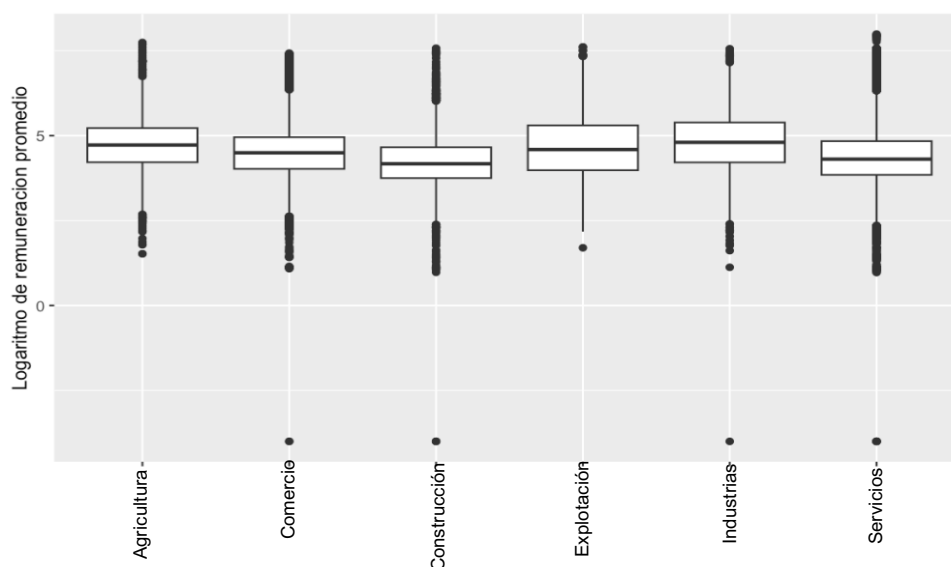
Medianas del logaritmo de remuneraciones por sector económico

Sector Económico	Mediana
Agricultura, ganadería, silvicultura y pesca	4,723
Comercio	4,495
Construcción	4,171
Explotación de Minas y Canteras	4,590
Industrias Manufactureras	4,802
Servicios	4,308

En la Tabla 7 se pueden visualizar los valores correspondientes a las medianas del logaritmo de las remuneraciones, identificando que estos están en un intervalo de 4 y 5. Las medianas del logaritmo de las remuneraciones, tienen a ser posiblemente similares en algunos sectores. Con la Figura 4, se puede identificar que, si bien están en un mismo intervalo, se podrían considerar diferentes, además de que no presentan una dispersión muy alta entre los datos.

Figura 4

Boxplot de las remuneraciones totales por sectores económicos



La similitud de las medianas se podrá concluir con una prueba de Kruskal Wallis, en donde se establecen las siguientes hipótesis:

H_0 = Los sectores económicos presentan igualdad en los niveles de remuneración a su personal del periodo 2012-2020

H_a = Los sectores económicos presentan diferencias significativas en los niveles de remuneración a su personal del periodo 2012-2020

Obteniendo en la prueba de Kruskal Wallis un $p - value < 2.2e - 16$, se puede concluir que las medianas de la remuneración de los diferentes sectores económicos presentan diferencias significativas. Para ser más específico, se aplicará una prueba Dunn con el fin de identificar que sectores específicos poseen diferencias.

Tabla 8

Prueba Dunn de las remuneraciones por sector económico

Sector Económico	Agricultura	Comercio	Construcción	Explotación	Industrias
Comercio	16,69588 0,0000*				
Construcción	35,47922 0,0000*	29,03865 0,0000*			
Explotación	3,892126 0,0007*	-3,771390 0,0012*	-14,43486 0,0000*		
Industrias	-3,257416 0,0084*	-23,56369 0,0000*	-43,15908 0,0000*	-5,735545 0,0000*	
Servicios	30,45279 0,0000*	23,54984 0,0000*	-15,61641 0,0000*	9,579657 0,0000*	40,46119 0,0000*

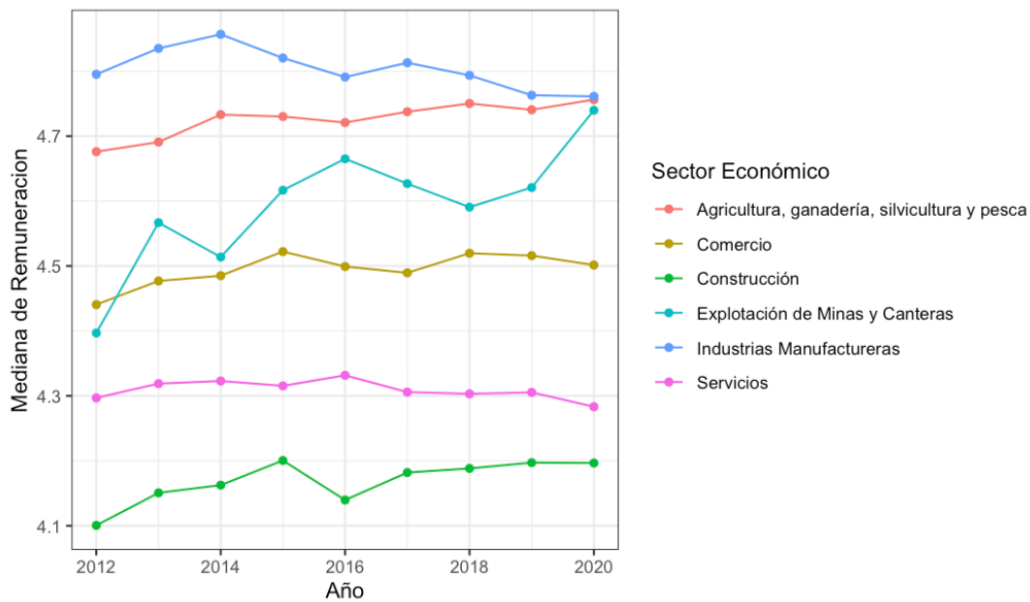
La prueba Dunn de la Tabla 8 ayuda a concluir que la remuneración en los sectores económicos es diferente.

Evolución de las remuneraciones por sector económico en Ecuador durante el período 2012-2020.

Se procede a realizar un análisis del comportamiento de las remuneraciones a partir de las medianas en el transcurso del tiempo.

Figura 5

Evolución de las remuneraciones en el periodo 2012-2020



En la Figura 5, el sector de explotación de minas ha sido el que mayor variación ha presentado en este periodo de tiempo con respecto a la remuneración, mientras que el sector de servicios muestra una mayor estabilidad en el avance del tiempo.

Los sectores que han ofrecido mayor remuneración al terminar el año 2020 son Agricultura, Industrias y Explotación de minas. Se puede mencionar que dichos sectores seguirán presentando un comportamiento muy similar en los años posteriores.

Remuneración total destinada al género masculino y femenino en el periodo 2012-2023.

Se realizará una comparación entre las medianas de la remuneración total de los empleados de género masculino y femenino para identificar el género que gana más. Se toma este proceso pues la distribución de datos no sigue una distribución normal.

Tabla 9

Medianas de la remuneración total por género

Mediana de Log_Remuneración total Masculino	Mediana de Log_Remuneración total Femenino
4,255	4,081

Con los resultados obtenidos en la Tabla 9, se puede identificar una diferencia entre las medianas de las remuneraciones. Para corroborar esto, se realizará una prueba Wilcoxon para muestras independientes estableciendo las siguientes hipótesis:

H_0 = La remuneración total para el género masculino y femenino es igual.

H_a = La remuneración total para el género masculino y femenino es diferente.

El resultado de la prueba Wilcoxon, obteniendo $p - value < 2.2e - 16$, se concluye que las remuneraciones de acuerdo con el sector son diferentes, siendo mayor el para el género masculino.

Mediante el exponencial se calculará la remuneración para cada género y se determinará en qué porcentaje la remuneración masculina supera a la femenina.

$$med_H = 4,255 \rightarrow Med_H = 10^{4,255} = \$17988,71$$

$$med_M = 4,081 \rightarrow Med_M = 10^{4,081} = \$12050,36$$

Cálculo del porcentaje de diferencia:

$$\% inferior_M = 1 - \frac{12050,36}{17988,71} = 0.33$$

Se concluye que la remuneración total destinada para el sector femenino es un 33% menor a la remuneración total destinada para el sector masculino.

Conclusiones Parciales

- En todos los diferentes sectores económicos, existen un alto porcentaje de empresas que omiten declarar sus niveles de ventas al REEM, lo que puede implicar en realizar una investigación más profunda para conocer los motivos de dichas omisiones.
- El nivel de ventas en el transcurso del tiempo presenta un comportamiento variable, pero en la mayoría de los sectores terminan con el mismo nivel con el que empezaron en el año 2012.
- Todos los sectores económicos que presentan remuneraciones totales diferentes, sin embargo, los sectores que presenta mayor remuneración son: Agricultura, Industria y Explotación de minas, terminando con valores muy cercanos en el año 2020.
- El género masculino es aquel que ha recibido mayor remuneración total en todo el periodo analizado, con un valor de \$17988,71; presentando así una posible desigualdad con el género femenino que tiene una remuneración de \$12050,36.

Limitaciones del Estudio

- El alto porcentaje de datos faltantes, si bien se analizaron por los diferentes sectores económicos, talvez se podría esclarecer más la ausencia de esos valores si se analizaran que tipos de actividad específica realizan. Con ello se podría justificar la ausencia de dichos valores.

- El analizar la igualdad de las plazas de trabajo por género, podría ayudar a justificar la diferencia en las remuneraciones de los diferentes géneros.
- La remuneración total invertida por los diferentes sectores podría ser más representativa si en los registros de empleados, se mencionara el número de empleados discapacitados con los cuales cuenta la empresa y la remuneración que se designa para este grupo.
- La base de datos correspondiente al periodo del 2006 al 2011 presenta errores en sus registros, los cuales dificultan poder realizar el análisis respectivo.

LABORATORIO 2: ENFOQUE MACHINE LEARNING

Análisis de Series de Tiempo de Plazas Disponibles y Clustering para la Agrupación de Sectores Empresariales Ecuatorianos (2012-2023)

Objetivos

- Desarrollar y evaluar modelos de series de tiempo para predecir trimestralmente el número de plazas disponibles en los diferentes sectores económicos del Ecuador durante el periodo 2012-2023, con el fin de comparar la oferta laboral entre dichos sectores.
- Aplicar técnicas de aprendizaje no supervisado (clustering) para agrupar empresas de diferentes sectores económicos del Ecuador, durante el periodo 2012-2023, en función de sus remuneraciones, ventas totales, plazas de empleo totales, plazas de empleo para hombres y plazas de empleo para mujeres.

Métodos

Descripción de las variables de la base de Datos para la Serie de Tiempo.

La base de datos consta de 10 variables y 1157075 observaciones que se presentan a continuación en la Tabla 10:

Tabla 10

Descripción de las variables utilizadas para el estudio de Series de tiempo y Aprendizaje no supervisado

Series de Tiempo	
Variable	Escala de medición
Año	Año en el que se realizó la observación. (Numérico)
Sector Económico	Sector económico en el que se desarrolla la empresa. (Categórico)
Situación	Funcionamiento de la empresa (Dicotómica)
Plazas_iess_trim1	Valores registrados del promedio de las plazas vacantes de empleo que registra la empresa en el IESS en el 1° trimestre. (Numérica)
Plazas_iess_trim2	Valores registrados del promedio de las plazas vacantes de empleo que registra la empresa en el IESS en el 2° trimestre. (Numérica)
Plazas_iess_trim3	Valores registrados del promedio de las plazas vacantes de empleo que registra la empresa en el IESS en el 3° trimestre. (Numérica)
Plazas_iess_trim4	Valores registrados del promedio de las plazas vacantes de empleo que registra la empresa en el IESS en el 4° trimestre. (Numérica)
Empleo_iess_trim1	Valores registrados del promedio de los empleados registrados de acuerdo a las plazas de empleo que registra la empresa en el IESS en el 1° trimestre. (Numérica)
Empleo_iess_trim2	Valores registrados del promedio de los empleados registrados de acuerdo a las plazas de empleo que registra la empresa en el IESS en el 2° trimestre. (Numérica)
Empleo_iess_trim3	Valores registrados del promedio de los empleados registrados de acuerdo a las plazas de empleo que registra la empresa en el IESS en el 3° trimestre. (Numérica)

Empleo_iesstrim4	Valores registrados del promedio de los empleados registrados de acuerdo a las plazas de empleo que registra la empresa en el IESS en el 4° trimestre. (Numérica)
Aprendizaje no Supervisado	
Variable	Escala de medición
Año	Año en el que se realizó la observación. (Numérico)
Sector Económico	Sector económico en el que se desarrolla la empresa. (Categórico)
Situación	Funcionamiento de la empresa (Dicotómica)
Ventas totales	Valores registrados de las ventas del año. Medida en dólares. (Numérico)
Ventas nacionales	Valores registrados de las ventas del año a nivel nacional. Medida en dólares. (Numérico)
Exportaciones	Valores registrados de las exportaciones del año. Medida en dólares. (Numérico)
Plazas Totales	Plazas de empleo totales registradas
Plazas_Hombres	Plazas de empleo registradas para hombres
Plazas_Mujeres	Plazas de empleo registradas para mujeres
Remuneración	Valores registrados sobre las remuneraciones totales dadas. (Numérico)

Procesamiento y Análisis de los Datos para las series de tiempo y clustering

- **Filtrado de datos:** De la base original de empresas que se encontraba separada en diferentes archivos, se realizó el filtrado de los datos manteniendo las variables a utilizar en el modelado de las series de tiempo y aprendizaje no supervisado. Posteriormente se realizó la unión de la base de datos.
- **Análisis exploratorio de datos:** Se realizó un análisis exploratorio para comprender las características de las series de tiempo de cada sector. Esto incluirá la visualización de las series, el análisis de la autocorrelación, para identificar posibles patrones de estacionalidad y tendencia, y pruebas de estacionariedad para determinar si las series requieren diferenciación.
- **Selección y Modelado de series de tiempo:** Se consideró el uso de diversos modelos de series de tiempo como ARIMA y SARIMA los cuales toman en cuenta la estacionariedad de las series de tiempo y la captura de los diferentes comportamientos de las series.
- **Estimación y Evaluación de los modelos de series de tiempo:** Para cada sector, se estimó los parámetros de los modelos seleccionados. La evaluación del rendimiento de los modelos se realizó mediante la división de los datos en un conjunto de entrenamiento (70% de los datos) y un conjunto de prueba (30% de los datos). Se utilizarán métricas como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el Error Porcentual Absoluto Medio (MAPE) para cuantificar la precisión de las predicciones en el conjunto de prueba.

- **Cuantificación o codificación de variables:** Para las técnicas de aprendizaje no supervisado, se procedió a la transformación de variables polinomiales en valores numéricos, como lo son: sectores económicos y la situación.
- **Aplicación de algoritmos de Clustering:** Mediante el algoritmo K-Means, se procedió a dividir los datos en k grupos considerando diferentes variables a comparar.

Resultados y Discusión

Series de Tiempo

- *Tratamiento de datos faltantes*

Una vez que se inició con el filtrado de las variables a utilizar en las series de tiempo, se presentaron datos faltantes. Aquellas observaciones que poseían valores faltantes en todas las variables que involucraban las plazas de empleo y empleos registrados, se procedieron a eliminar, pues no aportaban información para el estudio. Las observaciones que poseían 1 o 2 valores perdidos, fueron sustituidos por 0.

- *Cálculo de plazas disponibles por Trimestre*

Para determinar el número de plazas que realmente quedaban disponibles en cada registro, se procedió a restar los empleos registrados en el IESS por trimestre de las plazas registradas en el IESS en el mismo trimestre.

- *Manejo de valores negativos en las Plazas Disponibles*

Al generar la variable que contenía información sobre las plazas reales disponibles, se obtuvieron resultados negativos, los cuales se interpretaron como que la empresa contrató más empleados de los que registró en las plazas del IESS, por tal motivo, dichos valores negativos fueron considerados como 0.

- *Generación de bases trimestrales por Sector Económico*

Al tener una base de datos muy extensa, se procedió a realizar el análisis por cada sector económico. Si bien para el análisis de las series de tiempo se necesitan una gran cantidad de datos históricos, este no era el caso de la base, pues los registros contenían información desde el año 2012 al 2023, por tal motivo se procedió a sumar todas las plazas disponibles por cada trimestre de cada año, obteniendo así un total de 48 observaciones para cada sector.

- *Análisis de Series de Tiempo*

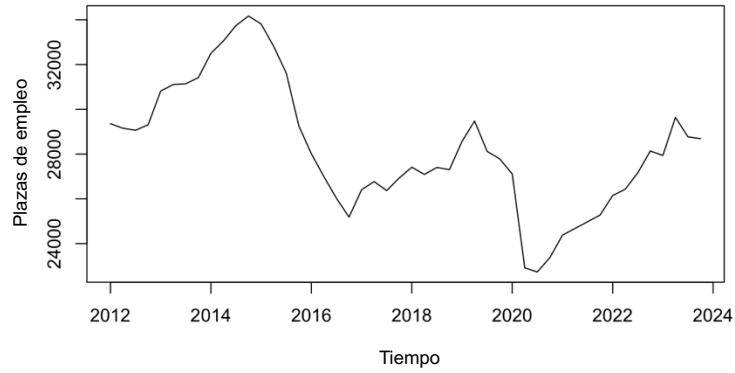
Para cada sector económico, se creó su serie de tiempo con todos los registros con el fin de observar el comportamiento de las plazas disponibles reales que tenía cada sector económico en el transcurso del tiempo.

Sector de Servicios

Se procede a generar la serie de tiempo del sector de servicios, haciendo uso de los registros trimestrales por cada año.

Figura 6

Serie de Tiempo del sector de servicios 2012-2023

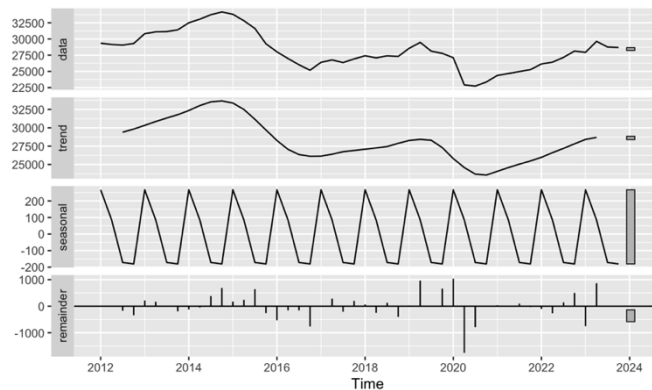


En la Figura 6, se puede observar que el sector de servicios a inicios del periodo 2012 hasta el 2015 presenta un comportamiento creciente, posteriormente se identifica un decrecimiento hasta el segundo trimestre de 2016 para posteriormente presentar un crecimiento más suavizado. En el primer trimestre del año 2020 se identifica un decrecimiento muy notorio a contrario del resto de la serie, provocado por la pandemia del COVID19. Finalmente, a partir del tercer trimestre del año 2020 se identifica un crecimiento, manteniendo dicho comportamiento hasta 2023. Con los límites de la serie de tiempo, se puede identificar que es el sector que ha presentado mayor número de plazas de empleo disponibles.

Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 7

Descomposición de la Serie de tiempo del sector de servicios



En la Figura 7 se puede identificar una tendencia decreciente de la serie de tiempo y un patrón estacional anual (cada 4 trimestres). Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,1269, aceptando la hipótesis nula de no estacionariedad. Debido a este resultado, se procedió a aplicar 2 diferenciaciones a la serie para lograr su estacionariedad. Una vez obtenida su estacionariedad, se procedió a la creación de los modelos de las series de tiempo mediante modelos autorregresivos (AR), medias móviles (MA) y SARIMA, realizando su entrenamiento con el 70% de los datos.

Figura 9

Autocorrelación de la serie diferenciada del sector de servicios

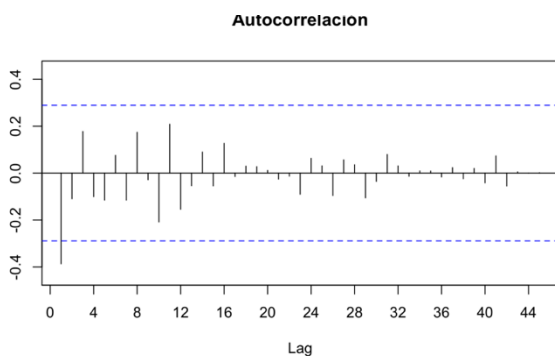
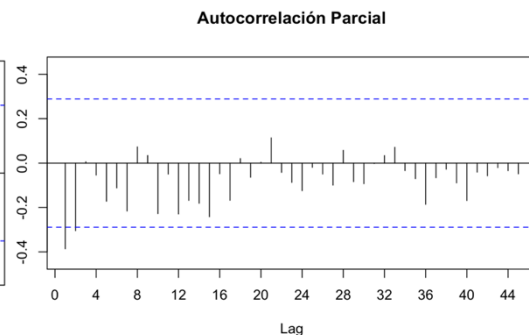


Figura 8

Autocorrelación parcial de la serie diferenciada del sector de servicios

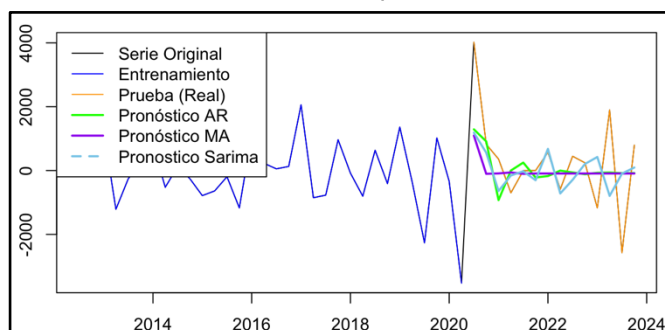


Con la Figura 8 y la Figura 9, al analizar la autocorrelación y autocorrelación parcial de la serie, permite identificar el orden que se puede aplicar para establecer los modelos de series de tiempo. En la autocorrelación se puede identificar un decaimiento gradual, con un pico significativo en el primer rezago, y estos parecen presentar una repetición cada 12 rezagos. Por otra parte, la autocorrelación parcial presenta también un pico en el primer rezago seguido de un decaimiento más acelerado y picos notables cada 12 rezagos. Con estas características, el modelo de la serie de tiempo sugiere un SARIMA, para capturar la estacionalidad que está presente en dicha serie. Los modelos aplicados para este sector fueron: $AR(2,0,0)(1,0,0)$, $MA(0,0,1)$ y $SARIMA(2,0,1)(1,0,1)$.

Posteriormente, se procedió al pronóstico de estos modelos seleccionados, utilizando el conjunto de prueba que era el 30% de los datos.

Figura 10

Pronósticos de la serie de tiempo del sector de servicios



De acuerdo con la Figura 10, se puede notar que el periodo 2021 fue significativo al presentar una variación muy notoria. Con el 30% de los datos utilizados para la prueba se puede observar que el modelo SARIMA es el que se acopla más al comportamiento de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 11

Métricas de error para los diferentes modelos del sector de servicio

Modelo	MASE	RMSE	MAE	MAPE
AR	0,9422	1281,538	992,2859	2097,1403
MA	0,9224	1287,016	971,4493	931,2875
SARIMA	0,9111	1374,821	959,5797	2274,3821

Al comparar las distintas métricas de la Tabla 11, el MASE de los tres modelos presentan valores menores a 1, lo que significa que son modelos mejores que uno de pronóstico simple random. Por otra parte, el MAE del modelo SARIMA es el más pequeño de todos (959,5797), lo que implica que es el que se equivocó menos al momento de las predecir, pero en contraparte, existen puntos donde se producen errores grandes ya que posee un RMSE mayor que los otros modelos. El tener un valor del MAPE demasiado elevado para los diferentes modelos da a entender que no será útil para determinar un pronóstico debido a que este es muy sensible a valores cercanos a 0, los cuales se obtienen al realizar una diferenciación de la serie.

Siendo así, el modelo de la serie de tiempo para el sector de servicios, queda determinado por la siguiente información de la Tabla 12:

Tabla 12

Coefficientes del modelo SARIMA del sector de Servicios

Coefficientes	ar1	ar2	ma1	sar1	sma1	Intercepto
	-0,77	-0,53	0,62	0,99	-0,97	-86,79
s.e.	0,2219	0,1915	0,1905	0,0100	0,1850	234,5809

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Servicios:

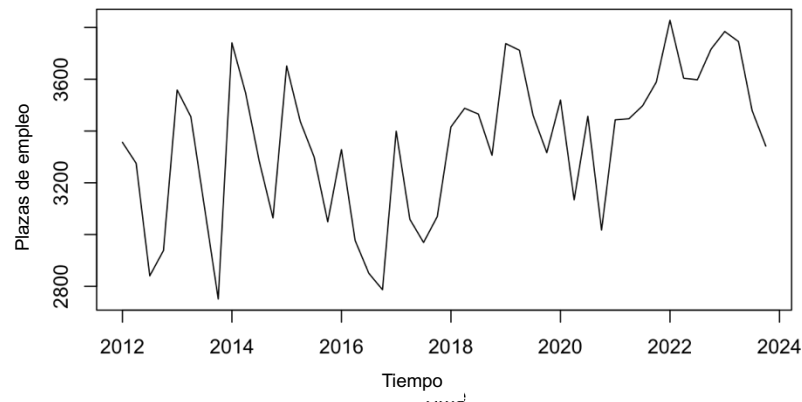
$$z_t = -0,77z_{t-1} - 0,53z_{t-2} + 0,99z_{t-4} + 0,62\varepsilon_{t-1} - 0,97\varepsilon_{t-4} + \varepsilon_t - 86,79$$

Sector de Agricultura, ganadería, silvicultura y pesca

Se procede a generar la serie de tiempo del sector de agricultura, haciendo uso de los registros trimestrales por cada año.

Figura 11

Serie de Tiempo del sector de Agricultura 2012-2023

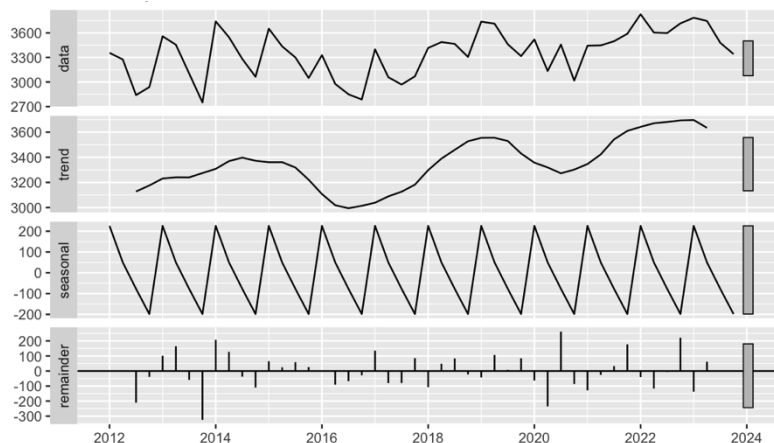


En la Figura 11, se puede observar que el sector de agricultura a inicios del periodo 2012 inicia con un total de plazas arriba de 3200, presentando una variación decreciente y creciente, presentando picos bajos en el último trimestre del año 2013 y un pico también bajo en el último trimestre del año 2017. Posterior a este periodo se inicia con una tendencia creciente con diferentes fluctuaciones, terminando el último trimestre de 2023 con un valor similar al inicio de 2012.

Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 12

Descomposición de la serie de tiempo del sector de Agricultura



En la Figura 12 se puede identificar una tendencia creciente de la serie de tiempo y un patrón estacional anual (cada 4 trimestres). Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,7174, aceptando la hipótesis nula de no estacionariedad. Debido a este resultado, se procedió a aplicar 2 diferenciaciones a la serie para lograr su estacionariedad. Una vez obtenida su estacionariedad, se procedió a la creación de los modelos de las series de tiempo mediante modelos autorregresivos (AR), medias móviles (MA) y SARIMA, realizando su entrenamiento con el 70% de los datos.

Figura 14

Autocorrelación de la serie diferenciada del sector de Agricultura

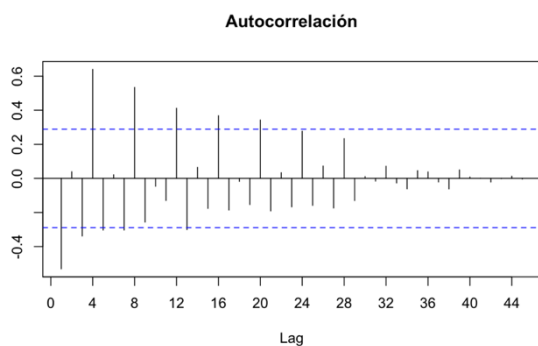
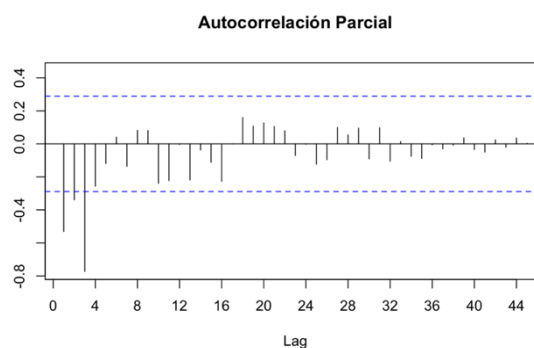


Figura 13

Autocorrelación parcial de la serie diferenciada del sector de Agricultura



Con la Figura 13 y la Figura 14, al analizar la autocorrelación y autocorrelación parcial de la serie, permite identificar el orden que se puede aplicar para establecer los modelos de series de tiempo. En la autocorrelación se puede identificar un decaimiento gradual, con picos significativos cada 4 rezagos, por otra parte, la autocorrelación parcial presenta valores significativos hasta el tercer resago y continúa con una tendencia decreciente.

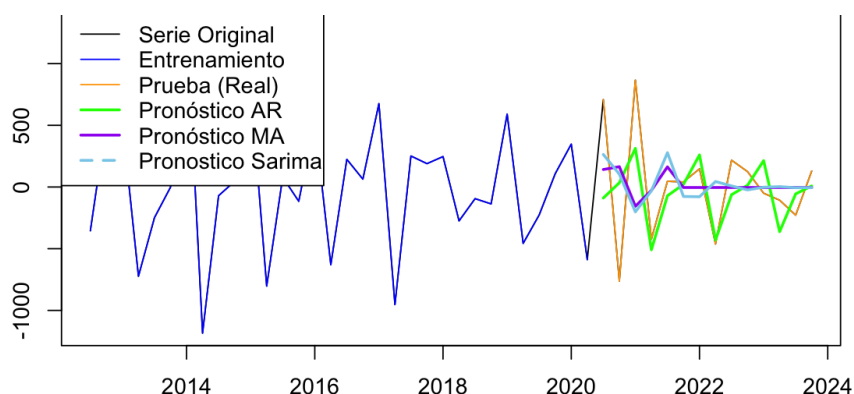
Con estas características, el modelo de la serie de tiempo sugiere un AR, MA o SARIMA, para capturar la estacionalidad que está presente en dicha serie.

Los modelos aplicados para este sector fueron: AR(1,0,0)(1,0,0), MA(0,0,1) y SARIMA(2,0,1)(0,0,1).

Posteriormente, se procedió al pronóstico de estos modelos seleccionados, utilizando el conjunto de prueba que era el 30% de los datos.

Figura 15

Pronósticos de la Serie de tiempo del sector de Agricultura



De acuerdo con la Figura 15, se puede notar que el periodo 2014-2015 fueron significativos al presentar una variación muy notoria, por el contrario, desde el 2020 no ha presentado variaciones notorias comparándolas con el resto de la serie. Con el 30% de los datos utilizados para la prueba se puede observar que el modelo AR es el que se acopla más al comportamiento de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 13

Métricas de error para los diferentes modelos del sector de Agricultura

Modelo	MASE	RMSE	MAE	MAPE
AR	0,9498531	366,4480	265,5107	131,1553
MA	1,1593301	445,1136	324,0655	112,2483
SARIMA	1,2066298	444,1242	337,2871	147,0138

Al comparar las distintas métricas de la Tabla 13, el MASE del modelo AR presenta un valor menor a 1, lo que significa que este modelo es mejor que uno de pronóstico simple random, los modelos MA y SARIMA no cumplen dicho supuesto. Por otra parte, el MAE del modelo AR es el más pequeño de todos (265,51), lo que implica que es el que se equivocó menos al momento de las

predecir, y respaldando esta información, es el modelo que menor RMSE presenta (366,4480). El tener un valor del MAPE demasiado elevado para los diferentes modelos da a entender que no sería útil para determinar un pronóstico debido a que este es muy sensible a valores cercanos a 0, los cuales se obtienen al realizar una diferenciación de la serie.

Siendo así, el modelo de la serie de tiempo para el sector de agricultura queda determinado por la siguiente información de la Tabla 14:

Tabla 14

Coefficientes del modelo AR del sector de Agricultura

Coefficientes	ar1	sar1	sma1	Intercepto
	-0,5067	0,8350	-0,97	-21,1203
s.e.	0,1448	0,0839	0,1850	126,4914

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Agricultura:

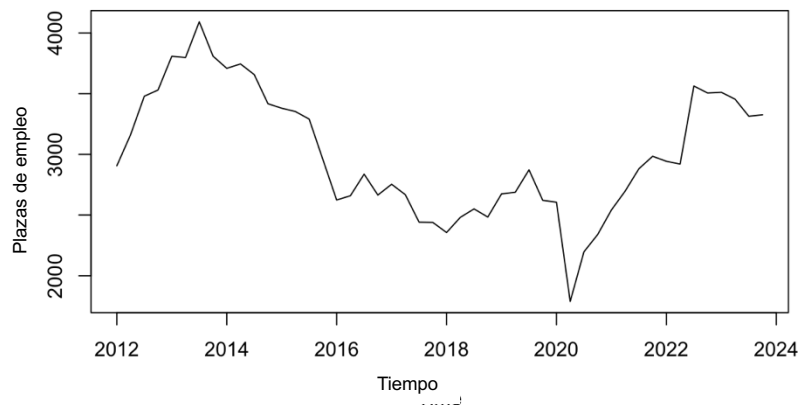
$$y_t = -0,5067y_{t-1} + 0,835y_{t-4} + 0,423y_{t-5} - 21,1203 + \varepsilon_t$$

Sector de Construcción

Se procede a generar la serie de tiempo del sector de construcción, haciendo uso de los registros trimestrales por cada año.

Figura 16

Serie de Tiempo del sector de Construcción 2012-2023



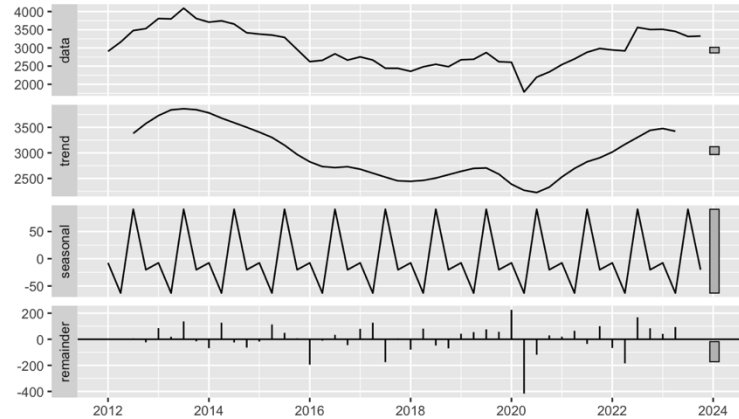
En la Figura 16, se puede observar que el sector de construcción a inicios del periodo 2012 hasta el último trimestre de 2013 presenta un comportamiento creciente, posteriormente se identifica un decrecimiento lento, con pequeños intervalos de crecimiento y decrecimiento. En el primer trimestre del año 2020 se identifica un decrecimiento muy notorio a contrario del resto de la serie, provocado por la pandemia del COVID19. Finalmente, a partir del tercer trimestre del año 2020 se identifica un crecimiento, manteniendo dicho comportamiento hasta 2023. Con los límites de

la serie de tiempo, se puede identificar que este sector ha presentado una cantidad de plazas disponibles algo similares al sector de agricultura.

Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 17

Descomposición de la serie de Tiempo del sector de Construcción



En la Figura 17 se puede identificar una tendencia decreciente de la serie de tiempo y un patrón estacional anual (cada 4 trimestres). Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,8583, aceptando la hipótesis nula de no estacionariedad. Debido a este resultado, se procedió a aplicar 1 diferenciación para lograr su estacionariedad. Una vez obtenida su estacionariedad, se procedió a la creación de los modelos de las series de tiempo mediante modelos SARIMA de diferente orden realizando su entrenamiento con el 70% de los datos.

Figura 18

Autocorrelación de la serie diferenciada del sector de Construcción

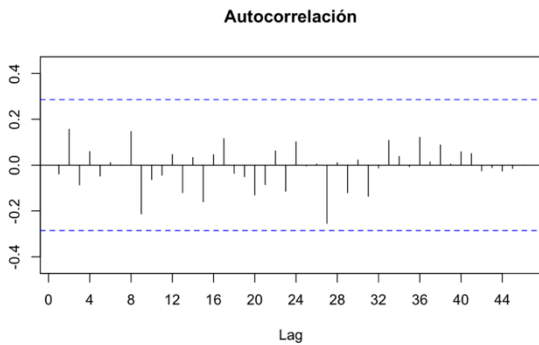
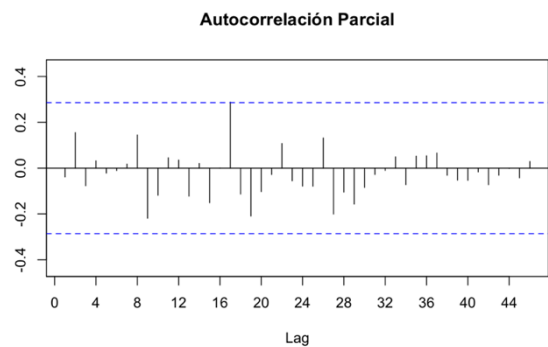


Figura 19

Autocorrelación parcial de la serie diferenciada del sector de Construcción



Con la Figura 18 y la Figura 19, al analizar la autocorrelación y autocorrelación parcial de la serie, estas permiten identificar el orden que se puede aplicar para establecer los modelos de series de tiempo. En la autocorrelación se puede identificar un decaimiento gradual, sin picos significativos, pero manteniendo un patrón cada 4 rezagos.

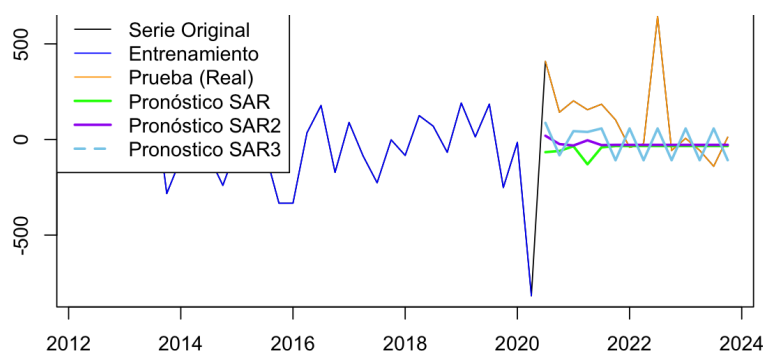
Con estas características, el modelo de la serie de tiempo sugiere un SARIMA de diferentes órdenes para analizar aquel que se adapta mejor al comportamiento. Los modelos evaluados fueron: SAR1(0,0,1)(0,0,1), SAR2(1,0,0)(0,0,1) y SAR3(1,0,1)(0,0,1).

Posteriormente, se procedió al pronóstico de estos modelos seleccionados, utilizando el conjunto de prueba que era el 30% de los datos.

Figura

20

Pronósticos de la Serie de tiempo del sector de Construcción



De acuerdo con la Figura 20, se puede notar que el periodo 2021 fue significativo al presentar una variación muy notoria. Con el 30% de los datos utilizados para la prueba se puede observar que el modelo SAR es el que se acopla más al comportamiento promedio de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 15

Métricas de error para los diferentes modelos del sector de Construcción

Modelo	MASE	RMSE	MAE	MAPE
SAR	0,8529204	260,3586	178,3530	158,2943
SAR2	0,8797492	268,4604	183,9631	162,5710
SAR3	0,9974783	295,0561	208,5813	158,2943

Al comparar las distintas métricas de la Tabla 15, el modelo SAR se considera como el mejor predictor, debido a que es el mejor si se compara con un modelo de predicción simple (MASE), presenta una menor penalización a los errores grandes (RMSE) y presenta los errores absolutos más pequeños (MAE). El tener un valor del MAPE demasiado elevado para los diferentes

modelos da a entender que no sería útil para determinar un pronóstico debido a que este es muy sensible a valores cercanos a 0, los cuales se obtienen al realizar una diferenciación de la serie. Siendo así, el modelo de la serie de tiempo para el sector de construcción queda determinado por la siguiente información de la Tabla 16:

Tabla 16

Coefficientes del modelo SAR2 del sector de Construcción

Coefficientes	ma1	sma1	Intercepto
	0,072	0,1213	-34,2984
s.e.	0,1695	0,2430	45,7932

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Construcción:

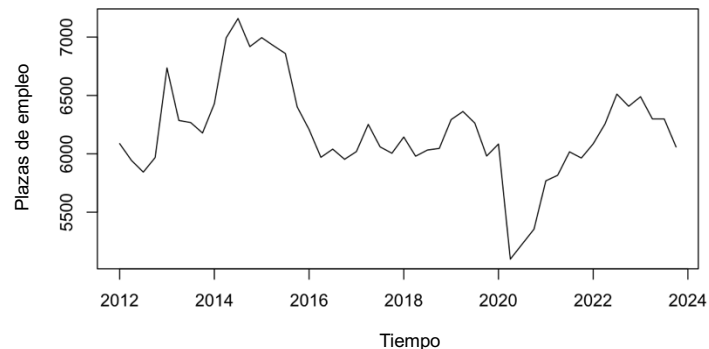
$$C_t = -34,2984 + \epsilon_t + 0,072\epsilon_{t-1} + 0,1213\epsilon_{t-4}$$

Sector de Comercio

Se procede a generar la serie de tiempo del sector de comercio, haciendo uso de los registros trimestrales por cada año.

Figura 21

Serie de Tiempo del sector de Comercio 2012-2023

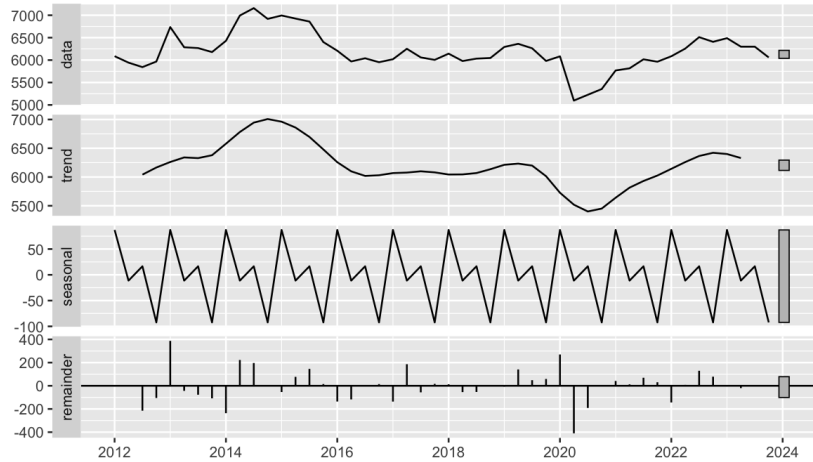


En la Figura 21, se puede observar que el sector de comercio a inicios del periodo 2012 hasta el último trimestre de 2014 presenta un comportamiento creciente con pequeñas variaciones, posteriormente se identifica un decrecimiento hasta inicios de 2016; con pequeños intervalos de crecimiento y decrecimiento a partir de 2016 hasta inicios de 2020 se mantiene un comportamiento constante. En el primer trimestre del año 2020 se identifica un decrecimiento muy notorio a contrario del resto de la serie, que podría ser provocado por la pandemia del COVID19. Finalmente, a partir del tercer trimestre del año 2020 se identifica un crecimiento, manteniendo dicho comportamiento hasta 2023. Con los límites de la serie de tiempo, se puede identificar que este sector ha presentado una cantidad de plazas disponibles mayor que los

sectores de agricultura y construcción. Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 22

Descomposición de la serie de Tiempo del sector de Comercio



En la Figura 22 se puede identificar una tendencia decreciente de la serie de tiempo y un patrón estacional anual (cada 4 trimestres). Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,3012, aceptando la hipótesis nula de no estacionariedad. Debido a este resultado, se procedió a aplicar una diferenciación para lograr su estacionariedad. Una vez obtenida su estacionariedad, se procedió a la creación de los modelos de las series de tiempo mediante modelos SARIMA de diferente orden realizando su entrenamiento con el 70% de los datos.

Figura 24

Autocorrelación de la serie diferenciada del sector de Comercio

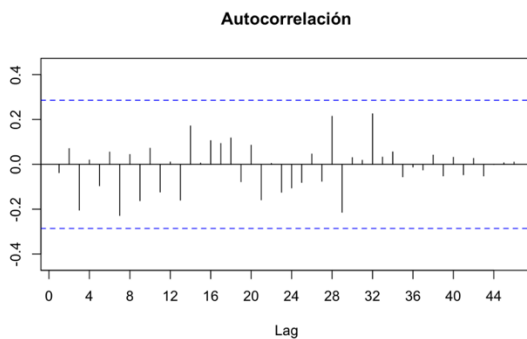
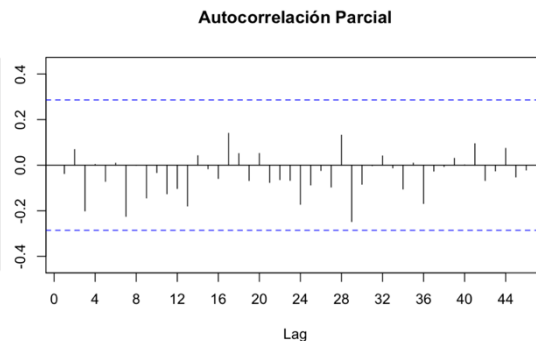


Figura 23

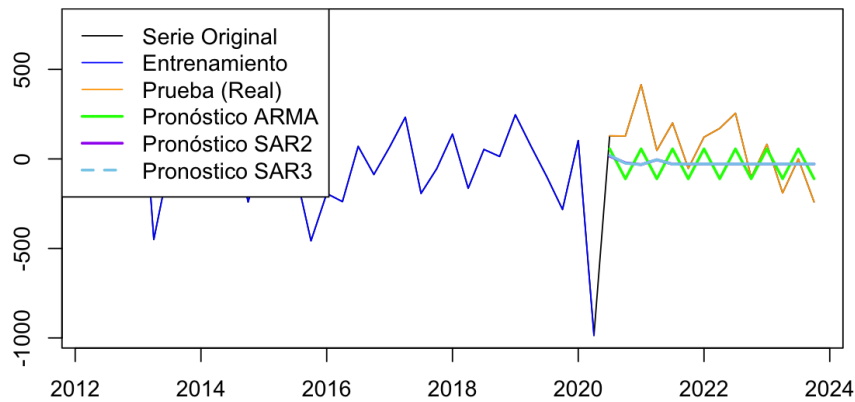
Autocorrelación parcial de la serie diferenciada del sector de Comercio



Con la Figura 23 y la Figura 24, al analizar la autocorrelación y autocorrelación parcial de la serie, permite identificar el orden que se puede aplicar para establecer los modelos de series de tiempo. En ambos casos no se presentan cortes claros y se identifica una oscilación de valores alrededor de cero, por tal motivo se descarta la posibilidad que el modelo se trate de un MA o AR puro. Con estas características, el modelo de la serie de tiempo sugiere un modelo ARMA o SARIMA de diferentes órdenes para analizar aquel que se adapta mejor al comportamiento. Los modelos evaluados fueron: $ARMA(1,0,1)(0,0,0)$, $SAR2(0,0,1)(0,0,1)$ y $SAR3(1,0,0)(0,0,1)$.

Figura 25

Pronósticos de la serie de tiempo del sector de Comercio



De acuerdo con la Figura 25, se puede notar que el periodo 2021 fue significativo al presentar una variación muy notoria. Con el 30% de los datos utilizados para la prueba se puede observar que el modelo ARMA es el que se acopla más al comportamiento de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 17

Métricas de error para los diferentes modelos del sector de Comercio

Modelo	MASE	RMSE	MAE	MAPE
ARMA	0,4583724	166,9790	133,9021	2474,682
SAR2	0,5469532	193,2524	159,7787	1291,923
SAR3	0,5463499	193,1690	159,6025	2474,682

Al comparar las distintas métricas de la Tabla 17, el modelo ARMA se considera como el mejor predictor, debido a que presenta los valores más bajos en las métricas MASE, RMSE y MAE, permitiendo así poder considerarlo como el que da una mayor precisión en la estimación y pronóstico de la serie. El MAPE no se considera como una métrica útil en este caso, debido a que la serie al estar diferenciada contiene algunos valores cercanos a 0, lo que genera una inflación en este último parámetro.

Siendo así, el modelo de la serie de tiempo para el sector de comercio queda determinado por la siguiente información de la Tabla 18:

Tabla 18

Coefficientes del modelo ARMA del sector de Comercio

Coefficientes	ar1	ma1	Intercepto
	-0,9995	0,9897	-27,4558
s.e.	0,0040	0,0425	48,9811

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Comercio:

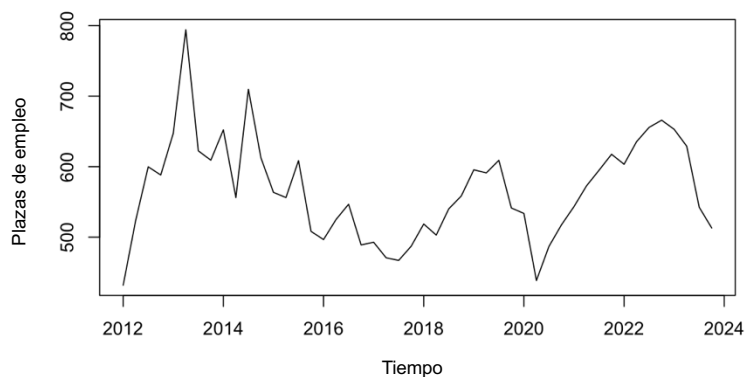
$$Y_t = -27,4558 - 0,9995Y_{t-1} + A_t - 0,9897A_{t-1}$$

Sector de Explotación de Minas y Calderas

Se procede a generar la serie de tiempo del sector de explotación, haciendo uso de los registros trimestrales por cada año.

Figura 26

Serie de Tiempo del sector de Explotación de Minas 2012-2023



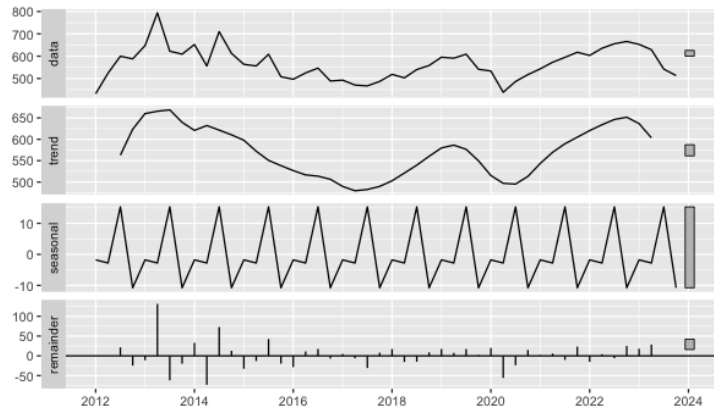
En la Figura 26, se puede observar que el sector de explotación de minas a inicios del periodo 2012 hasta el inicio de 2013 presenta un comportamiento creciente con pequeñas variaciones, posteriormente se identifica un decrecimiento hasta el segundo trimestre de 2017 con pequeños intervalos de variabilidad; a partir del tercer trimestre de 2017 hasta el segundo trimestre de 2019 se presenta una recuperación lenta con un comportamiento creciente. En el primer trimestre del año 2020 se identifica un decrecimiento muy notorio llegando a niveles muy similares que, en 2012, dicho comportamiento pudo ser provocado por la pandemia del COVID19. Finalmente, a partir del segundo trimestre del año 2020 se identifica un crecimiento, manteniendo dicho comportamiento hasta inicios de 2023 y nuevamente presenta un decrecimiento. Con los límites

de la serie de tiempo, se puede identificar que este sector ha presentado la menor cantidad de plazas disponibles de los sectores analizados hasta el momento.

Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 27

Descomposición de la serie de Tiempo de sector de Explotación de Minas



En la Figura 27 se puede identificar una tendencia decreciente de la serie de tiempo y un patrón estacional anual (cada 4 trimestres) dando a entender que estos influyen de manera constante cada año. Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,312, aceptando la hipótesis nula de no estacionariedad. Debido a este resultado, se procedió a aplicar dos diferenciaciones para lograr su estacionariedad. Una vez obtenida su estacionariedad, se procedió a la creación de los modelos de las series de tiempo mediante modelos AR, MA, SARIMA de diferente orden realizando su entrenamiento con el 70% de los datos.

Figura 29

Autocorrelación de la serie diferenciada del sector de Explotación de Minas

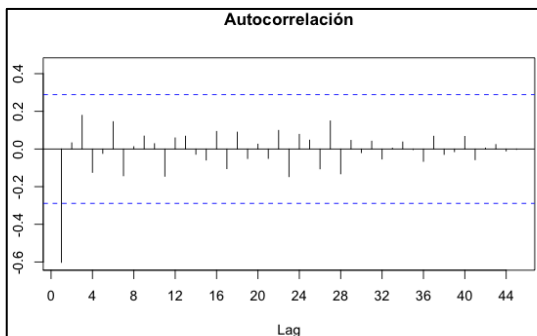
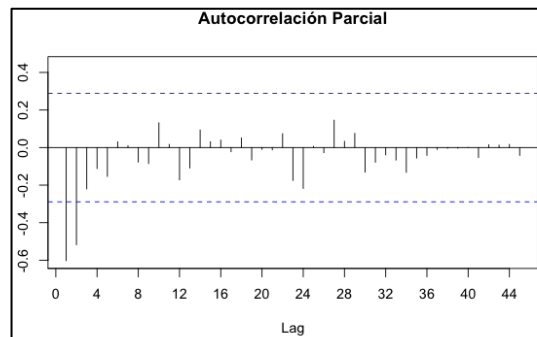


Figura 28

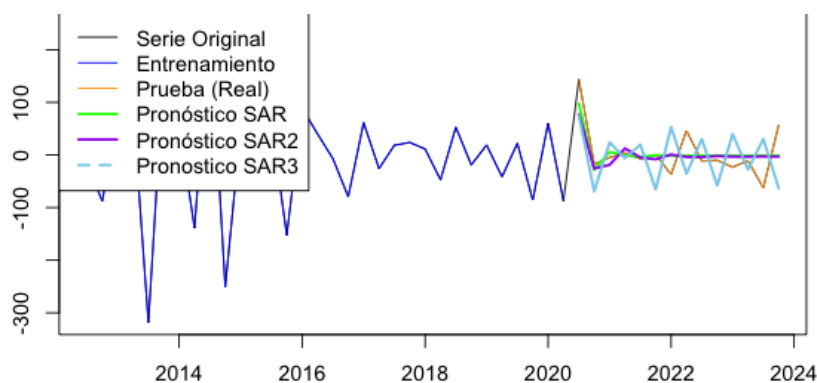
Autocorrelación de la serie diferenciada del sector de Explotación de Minas



Con la Figura 28 y la Figura 29, al analizar la autocorrelación y autocorrelación parcial de la serie, estas permiten identificar el orden que se puede aplicar para establecer los modelos de series de tiempo. En la autocorrelación se puede identificar un corte muy notorio después del primer rezago, lo que puede sugerir un componente de medias móviles de orden 1, por otra parte, la autocorrelación parcial presenta un corte a partir del segundo rezago, sugiriendo un componente autorregresivo de orden 2. Se puede identificar también un patrón estacional anual, pero como los registros eran trimestrales, se incorporó un componente estacional de periodicidad 4. Con estas características, el modelo de la serie de tiempo sugiere un modelo SARIMA de diferentes órdenes para analizar aquel que se adapta mejor al comportamiento. Los modelos evaluados fueron: SAR(1,0,1)(0,0,1), SAR2(2,0,1)(1,0,1) y SAR3(2,0,0)(0,1,1).

Figura 30

Pronósticos de la serie de tiempo del sector de Explotación de minas.



De acuerdo la Figura 30, con el 30% de los datos utilizados para la prueba se puede observar que el modelo SAR es el que se acopla más al comportamiento de la serie, pues, aunque pareciera que el modelo SAR3 captura la dinámica, es importante notar que sus picos son opuestos a los valores originales de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 19

Métricas de error para los diferentes modelos del sector de Explotación de minas

Modelo	MASE	RMSE	MAE	MAPE
SAR	0,2210591	31,45601	23,63774	102,6385
SAR2	0,2405179	34,45557	25,71847	211,2220
SAR3	0,5396621	65,34673	57,70581	102,6385

Al comparar las distintas métricas de la Tabla 19, el modelo SAR se considera como el mejor predictor, debido a que presenta los valores más bajos en las métricas MASE, RMSE y MAE,

permitiendo así poder considerarlo como el que da una mayor precisión en la estimación y pronóstico de la serie. El MAPE no se considera como una métrica útil en este caso, debido a que la serie al estar diferenciada contiene algunos valores cercanos a 0, lo que genera una inflación en este último parámetro.

Siendo así, el modelo de la serie de tiempo para el sector de explotación de minas queda determinado por la siguiente información de la Tabla 20:

Tabla 20

Coefficientes del modelo SARIMA del sector de Explotación de Minas

Coefficientes	ar1	ma1	sma1	Intercepto
	-0,3012	-1	-0,0322	-1,1618
s.e.	0,1740	0,1233	0,2046	0,9515

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Explotación de Minas:

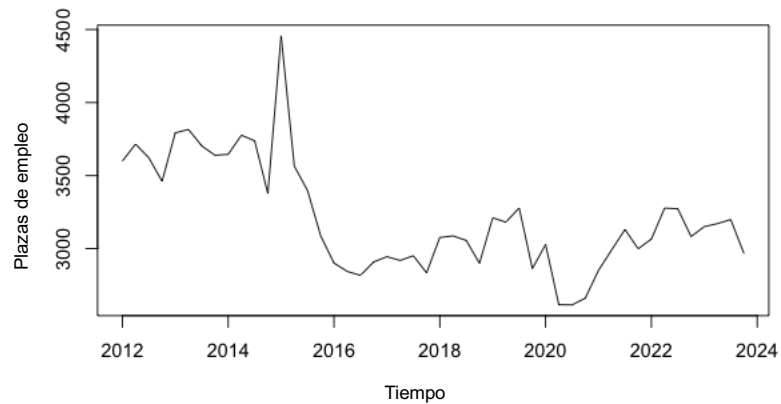
$$Y_t = -1,1618 - 0,3012Y_{t-1} + A_t - 1A_{t-1} - 0,0322A_{t-4}$$

Sector de Industrias Manufactureras

Se procede a generar la serie de tiempo del sector de industrias manufactureras, haciendo uso de los registros trimestrales por cada año.

Figura 31

Serie de Tiempo del sector de Industrias Manufactureras 2012-



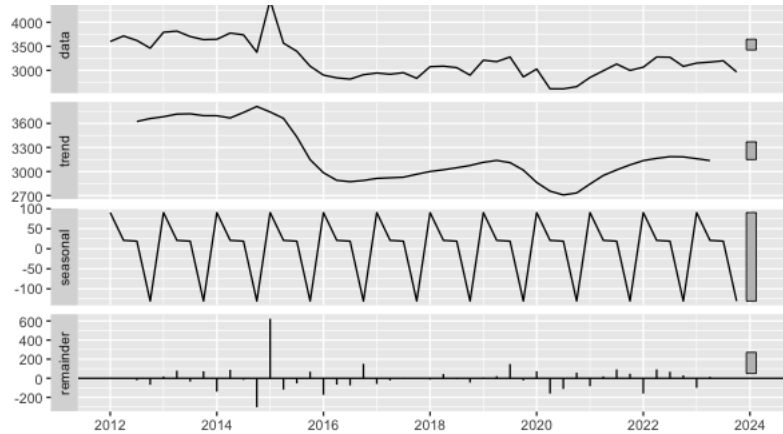
De acuerdo con la Figura 31, se puede identificar que desde el año 2012 hasta el segundo trimestre de 2014 existe un comportamiento constante de las plazas de empleo con algunos intervalos crecientes y decrecientes. Para el año 2015 se identifica el pico más alto y en inicios del 2016 cae a un nivel más bajo que en periodos anteriores, manteniendo dicho comportamiento hasta 2020 donde se puede identificar su punto más bajo y posteriormente un crecimiento lento

hasta 2023. Este sector económico presenta un número de plazas algo similar con el de comercio y agricultura.

Se procede a realizar el análisis de la tendencia, estacionalidad y residuos de las series, así como su estacionariedad, con el fin de poder generar modelos de series de tiempo que permitan estudiar y pronosticar comportamientos futuros.

Figura 32

Descomposición de la serie de Tiempo de sector de Explotación de Minas



En la Figura 32, se identifica una tendencia decreciente en la mayor parte de la serie, además se puede notar fluctuaciones regulares en la estacionalidad con un patrón anual. Al aplicar la prueba de Dickey Fuller para la estacionariedad, se obtuvo un p-value de 0,5879, concluyendo que la serie es no estacionaria. Debido a este resultado, se procedió a aplicar una diferenciación para lograr su estacionariedad. Posteriormente, se procedió a la creación de los modelos de las series de tiempo mediante modelos AR, MA, SARIMA de diferente orden realizando su entrenamiento con el 70% de los datos.

Figura 34

Autocorrelación de la serie diferenciada del sector de Industrias Manufactureras.

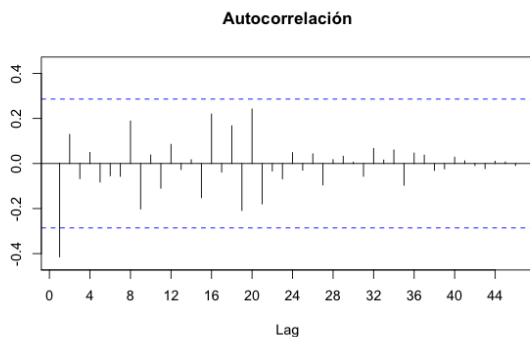
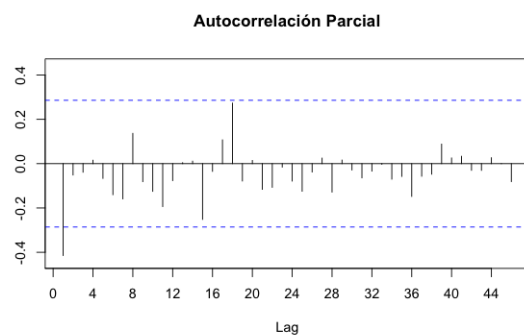


Figura 33

Autocorrelación de la serie diferenciada del sector de Industrias Manufactureras

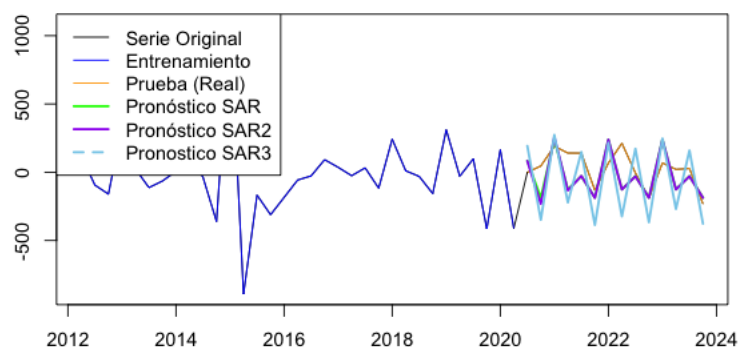


Con la Figura 33 y la Figura 34 al analizar los gráficos de autocorrelación y autocorrelación parcial se identifica un patrón regular que muestra una evidencia de la estacionalidad anual. A su vez, de acuerdo, a sus comportamientos se puede evidenciar un componente autorregresivo y de medias móviles de orden 1, lo cual permite dar un indicio de los modelos de series de tiempo a aplicar. Con estas características, el modelo de la serie de tiempo sugiere un modelo SARIMA de diferentes órdenes para analizar aquel que se adapta mejor al comportamiento. Los modelos evaluados fueron: SAR(0,0,1)(0,1,1), SAR2(1,0,0)(0,1,1) y SAR3(1,0,1)(1,1,0).

Figura 35

Pronósticos de la serie de tiempo del sector de Industrias

Manufactureras



De acuerdo con la Figura 35, con el 30% de los datos utilizados para la prueba se puede observar que el modelo SAR y SAR2 son los que se acoplan más al comportamiento de la serie. Para determinar el mejor modelo de pronóstico se procede a validarlos de acuerdo con las métricas respectivas.

Tabla 21

Métricas de error para los diferentes modelos del sector de Industrias Manufactureras

Modelo	MASE	RMSE	MAE	MAPE
SAR	0,4798715	161,7206	129,3113	666,2587
SAR2	0,4985139	167,9515	134,3349	669,3762
SAR3	0,8171723	256,7625	220,2040	666,2587

Al comparar las distintas métricas de la Tabla 21, el modelo SAR se considera como el mejor predictor, debido a que presenta los valores más bajos en las métricas MASE, RMSE y MAE, permitiendo así poder considerarlo como el que da una mayor precisión en la estimación y pronóstico de la serie. El MAPE no se considera como una métrica útil en este caso, debido a

que la serie al estar diferenciada contiene algunos valores cercanos a 0, lo que genera una inflación en este último parámetro.

Siendo así, el modelo de la serie de tiempo para el sector de industrias manufactureras queda determinado por la siguiente información de la Tabla 22:

Tabla 22

Coefficientes del modelo SARIMA del sector de Industrias Manufactureras

Coefficientes	ma1	sma1
	-0,34	-1
s.e.	0,1659	0,2211

Con la información obtenida, se puede escribir la ecuación matemática que modela a la serie establecida para el sector de Industrias Manufactureras:

$$y_t = y_{t-4} + \epsilon_t - 0,34\epsilon_{t-1} - \epsilon_{t-4} + 0,34\epsilon_{t-4}$$

Aprendizaje no Supervisado (Clustering)

- *Base de Datos para Técnicas de aprendizaje no supervisado*

La extensión inicial de la base de datos fue de alrededor de 34000 observaciones y la limitante del software utilizado solo permite como máximo 10000 observaciones, por tal motivo se procedió a realizar un muestreo estratificado proporcional, con el cual se creó una muestra representativa de la población, asegurando que cada subgrupo importante esté incluido en la muestra contando con la misma proporción que existe en la población.

- *Tratamiento de datos faltantes*

Una vez que se inició con el filtrado de las variables a utilizar para las técnicas de aprendizaje no supervisado, se presentaron datos faltantes. Aquellas observaciones que poseían valores faltantes en todas las variables que involucraban ventas, remuneraciones, plazas de empleo, plazas de hombres y mujeres, se procedieron a eliminar, pues no aportaban información para el estudio. Las observaciones que poseían 1 o 2 valores perdidos en algunas de estas variables, fueron sustituidos por 0.

- *Codificación de variables Categóricas*

En nuestra base de datos, tenemos una variable de sectores económicos, lo cual se procedió a codificarla para poder proceder con el algoritmo de K-means para poder generar diferentes grupos.

Agrupación por sectores económicos, plazas totales, plazas de hombres, plazas de mujeres, remuneraciones y ventas totales

Se procedió a realizar una clusterización comparando las variables mencionadas y realizando una agrupación en 3 grupos. Siendo así, se obtuvo:

Figura 36

Clusters de Remuneraciones, Ventas totales y Plazas Totales



Podemos visualizar que, de acuerdo con la Figura 36, que sin importar el sector económico al que pertenezcan las empresas, se pueden clasificar en 3 grupos específicos que se detallan a continuación:

Tabla 23

Centroides de los clústers generados

Atributo	Clúster 0	Clúster 1	Clúster 2
Sector Económico	1,920	2,786	2,762
Ventas Totales	24873641,800	153221399,143	1098015026,952
Plazas	164,500	1517,929	2722
Plazas Hombres	114,373	1183,571	2308,667
Plazas Mujeres	50,127	334,357	413,333
Remuneraciones	823309304145,589	763705848061165,400	246303784519418,440

La Tabla 23 muestra los centroides de los clústers y nos indica las coordenadas que se han tomado para formar los siguientes grupos; esta información, apoyada con la Figura 36 la cual permite determinar los siguientes resultados:

- **Clúster 0:** Se identifican empresas que gastan en remuneraciones desde 1100 dólares hasta 122 mil millones de dólares anuales aproximadamente, así mismo poseen ventas totales anuales que van desde los 540 dólares hasta alrededor de los 12 mil millones de dólares aproximadamente. En este grupo los límites de plazas van desde 2 hasta alrededor de 12000 plazas de empleo.
- **Clúster 1:** Se identifican empresas que gastan en remuneraciones desde 150 mil millones de dólares hasta 415 billones de dólares anuales aproximadamente, así mismo poseen ventas totales anuales que van desde los 7 millones de dólares hasta alrededor de los 11 mil millones de dólares aproximadamente. En este grupo los límites que se presentan van desde 202 hasta alrededor de 8370 plazas de empleo.

- Clúster 2:** Se identifican empresas que gastan en remuneraciones desde 560 billones de dólares hasta 980 billones de dólares anuales aproximadamente, así mismo poseen ventas totales anuales que van desde los 2 millones de dólares hasta alrededor de los 940 mil millones de dólares aproximadamente. En este grupo se tienen límites que van desde 105 hasta alrededor de 7095 plazas de empleo.

Con estas agrupaciones, se puede inferir que mientras mayor sea el valor de la remuneración, menores plazas de empleo registradas van a tener y a la vez sus ventas no necesariamente deben ser muy altas, es decir, que se puede tener una relación inversamente proporcional entre el total de remuneración y el número de plazas de empleo disponibles registradas.

Figura 37

Clusters de Plazas Hombres vs Remuneración

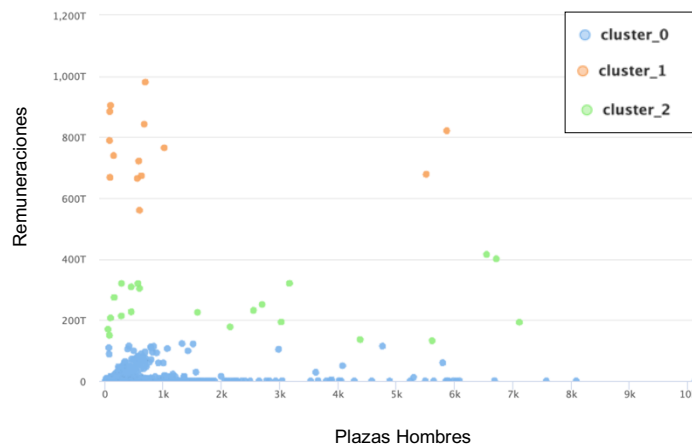
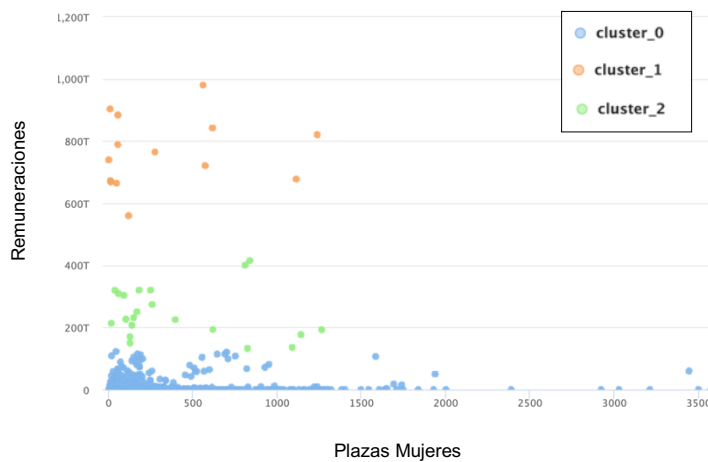


Figura 38

Clusters de plazas Mujeres vs Remuneración



Formando otras agrupaciones que comparan las remuneraciones con las plazas de acuerdo con el género, se puede identificar nuevamente tres grupos, tal y como indican la Figura 37 y Figura 38. Para los hombres, en primera instancia, se puede notar que existe un mayor registro para este grupo a comparación de las mujeres. Pues estos van desde 1 plaza hasta alrededor de 10000 plazas (1 observación), mientras tanto que el grupo de las mujeres como máximo llegan a tener 3500 plazas. Se puede identificar una desigualdad entre el personal femenino y el masculino, dando más aceptación a este último grupo en los diferentes sectores económicos. Por otra parte, las remuneraciones entre hombres y mujeres están equilibrados en cantidad, pues ambos géneros alcanzan los mismos límites.

Conclusiones Parciales

- Comparando los diferentes sectores económicos en el transcurso del tiempo desde 2012-2023, se puede notar que el periodo del año 2020 resultó muy significativo en las plazas de empleo disponibles, presentando un descenso en las plazas disponibles, con lo cual se puede inferir que uno de los factores causantes de este suceso es la pandemia del COVID19, lo cual puede afectar a las series de tiempo.
- Comparando los diferentes sectores económicos, el sector de servicios presentó una mayor disponibilidad de plazas de empleo, alcanzando un pico de 34000 plazas, en segundo lugar, está el sector de comercio con 7000 plazas, en tercer lugar, se encuentra el sector de industrias manufactureras, en cuarto lugar, al sector construcción con 4000 plazas, en quinto lugar, encontramos al sector agricultura y en última posición al sector de explotación de minas con 800 plazas.
- El sector de agricultura, si bien se encuentra en quinto lugar, ha sido el único sector que no tuvo una afectación tan grande en el periodo de la pandemia del COVID19, pues a pesar de presentar diferentes oscilaciones, siempre mantuvo una tendencia creciente en las plazas de empleo a comparación de los otros sectores económicos.
- El aplicar técnicas de aprendizaje supervisado ha permitido identificar tres grupos distintos entre la remuneración, ventas totales, plazas disponibles totales y plazas por género.
- Se puede visualizar que el género masculino en los diferentes sectores económicos presenta mayor disponibilidad de empleo a comparación del género femenino.

Limitaciones del Estudio

- Una limitante que se presenta en este estudio es la información recopilada en el periodo 2020, al tener en cuenta la pandemia del covid19, si bien se dispusiera de variables exógenas, estas podrían ayudar a mejorar el modelo de series de tiempo de los diferentes sectores económicos, con el fin de entender que factores influyeron para ese comportamiento característico de este periodo.

LABORATORIO 3: ENFOQUE TOMA DE DECISIONES**Análisis de la brecha de desigualdad en las remuneraciones y plazas de empleo de acuerdo con el género de los trabajadores.****Objetivos**

- Calcular el índice de brecha de género y de remuneraciones en los diferentes sectores económicos para aplicar estrategias que mejoren las posibles desigualdades presentes en los sectores que mayor índice presenten.
- Desarrollar un modelo k-NN que clasifique el tamaño empresarial (micro, pequeña, mediana o grande) a partir de variables como plazas de empleo, brecha salarial por género y presencia exclusiva de remuneración femenina, con el fin de identificar patrones organizativos vinculados a condiciones laborales.

Métodos**Descripción de la Base de Datos**

La base de datos consta de 8 variables y 1157075 observaciones, las cuales se describen en la Tabla 24 presentada a continuación.

Tabla 24

Variables utilizadas en la toma de decisiones.

Variable	Escala de medición
Año	Año en el que se realizó la observación. (Numérico)
Sector Económico (g_sectores)	Sector económico en el que se desarrolla la empresa. (Categorico)
Código Sección	Código de clasificación de la actividad económica. (Categorica)
	(A) Agricultura, ganadería, silvicultura y pesca.
	(B) Explotación de minas y canteras.
	(C) Industrias manufactureras.
	(D) Suministro de electricidad, gas, vapor y aire acondicionado.
	(E) Distribución de agua; alcantarillado, gestión de desechos y actividades de saneamiento.
	(F) Construcción.
	(G) Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas.
	(H) Transporte y almacenamiento.
	(I) Actividades de alojamiento y de servicio de comidas.
	(J) Información y comunicación.
	(K) Actividades financieras y de seguros.
	(L) Actividades inmobiliarias.
	(M) Actividades profesionales, científicas y técnicas.
	(N) Actividades de servicios administrativos y de apoyo.
	(O) Administración pública y defensa; planes de seguridad social de afiliación obligatoria.
	(P) Enseñanza.

(Q) Actividades de atención de la salud humana y de asistencia social.
 (R) Artes, entretenimiento y recreación.
 (S) Otras actividades de servicios.

Variable	Escala de medición
Plazas de empleo equivalente de hombres (Plazas_equi_hombres)	Plazas de empleo registrado equivalente hombres (Numérico)
Plazas de empleo equivalente de mujeres (Plazas_equi_mujeres)	Plazas de empleo registrado equivalente mujeres (Numérico)
Tamaño de empresas con plazas de empleo registrado (tamanoe_plazas)	Tamaño de empresas con plazas de empleo registrado (Categórico)
Remuneraciones de Hombres (remuneraciones_hombres)	Remuneraciones mujeres en miles de dólares corrientes (Numérico)
Remuneraciones de Mujeres (remuneraciones_mujeres)	Remuneraciones hombres en miles de dólares corrientes (Numérico)

Procesamiento y Análisis de los Datos

- **Filtrado de Datos:** A partir de la base original que posee más de 7 millones de datos, se realizó un filtrado de datos para utilizar información solo de empresas activas desde el periodo 2012 al 2023, debido a que el periodo 2006-2011 presenta errores en sus registros.
- **Manejo de datos faltantes:** De acuerdo con las variables seleccionadas, se identificó la ausencia de información simultánea en las variables de plazas de empleo y remuneraciones tanto de hombres como de mujeres, por lo que dichas observaciones fueron eliminadas debido a que no aportarían información en el estudio. Para aquellos datos que presentaban ausencia de registros en un solo género, se procedió a rellenarlos con un valor de 0.

Índice de brecha de plazas de empleo por Género

El índice de brecha de plazas de empleo es una métrica que permite identificar la desigualdad existente entre hombres y mujeres con respecto al acceso a oportunidades de empleo. Mientras el índice sea más cercano a un valor de 1, indica una mayor desigualdad, y los valores cercanos a 0 indican una equidad entre género. Para determinar este índice se hizo uso de la siguiente fórmula:

$$\text{índice}_{\text{empleo}} = \frac{|plazas_{\text{hombres}} - plazas_{\text{mujeres}}|}{\text{Total de plazas}}$$

Índice de brecha de remuneración económica por género

El índice de brecha de remuneración es una métrica que permite identificar la desigualdad existente entre el salario que reciben hombres y mujeres. Mientras el índice sea más cercano a un valor de 1, indica una mayor desigualdad, y los valores cercanos a 0 indican una equidad entre género. Para determinar este índice se hizo uso de la siguiente fórmula:

$$\text{índice}_{\text{remuneración}} = \frac{|\text{rem. promedio}_{\text{hombres}} - \text{rem. promedio}_{\text{mujeres}}|}{\text{rem. promedio}_{\text{hombres}}}$$

Resultados y Discusión

La desigualdad de género ha sido un fenómeno global que ha venido desde años atrás en todos los campos del desarrollo profesional humano, comúnmente siempre inclinado a una desigualdad para las mujeres. En el ámbito laboral, también es evidente este tipo de desigualdad de género, y más aún cuando se trata de las plazas de empleo y la remuneración económica. Según Barcena (2021), la situación de desigualdad de género tiene sus raíces en estereotipos y patrones culturales que se presentaron en la antigüedad y que aún en la actualidad sigue limitando el desarrollo profesional de las personas.

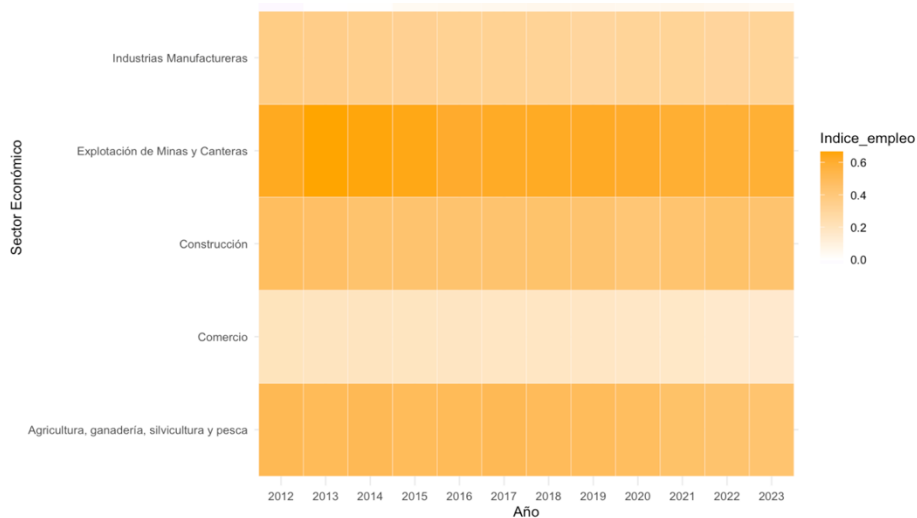
Análisis de la Brecha de Género

Índice de brecha de Plazas de empleo por género

El analizar los diferentes sectores económicos en el transcurso del tiempo permitirá identificar aquellos que mantienen un comportamiento inequitativo con respecto a las plazas de empleo.

Figura 39

Mapa de Calor anual sobre el índice de la brecha de plazas de empleo por sector económico



Con la información visual que proporciona la Figura 39, se puede identificar claramente que el sector de explotación de minas ocupa el primer lugar en presentar un índice alrededor de 0,6, lo cual implica que, en este sector hay una brecha del 60% aproximadamente a favor de los hombres. En segundo lugar, se puede identificar a los sectores de agricultura y construcción que presenta un índice alrededor de 0,4 al 0,5; presentando así nuevamente una inclinación a favor

de los hombres. En tercer lugar y con índices muy bajos, cercanos a 0, están los sectores de servicios y comercio, concluyendo así que estos sectores han sido los más equitativos.

El análisis realizado ha recopilado información anual para cada sector, pero es interesante analizar de forma conjunta cada sector, sumando su aporte total sin tener en cuenta el año de registro. Para ello se tomará en cuenta el total de la brecha existente por género en cada sector.

Tabla 25

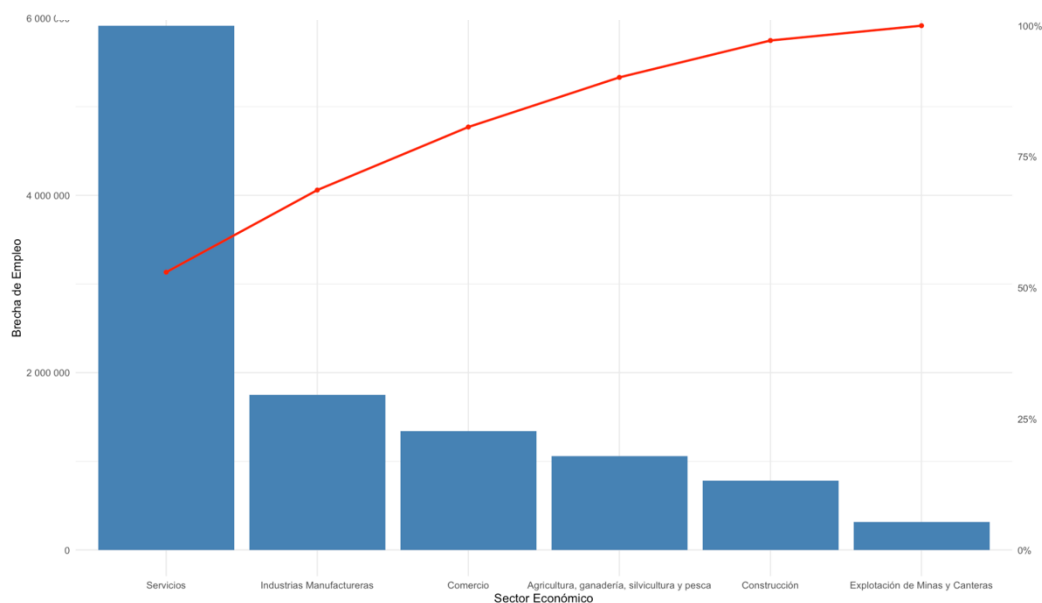
Brecha total acumulativa de empleo por género en cada sector económico.

Sector Económico	Brecha Total	Porcentaje Individual
Servicios	5913923,7	52,98%
Industrias Manufactureras	1747919,4	15,66%
Comercio	1342184,4	12,02%
Agricultura, ganadería, etc.	1057213,8	9,47%
Construcción	785055	7,03%
Explotación de minas	315920,3	2,84%

Con la información proporcionada en la Tabla 25, se pudo identificar que el sector que presenta un mayor aporte a la brecha de empleo es el sector de servicios, seguido del sector de industrias, comercio y agricultura. A continuación, se analizó la proporción que ocupan estos sectores destacados.

Figura 40

Diagrama de Pareto de la brecha de empleo por género en los sectores económicos



Con la información visual proporcionada en la Figura 40, se pudo identificar que aproximadamente el 90% de la brecha de empleo se centra en los 4 sectores mencionados de la Tabla 25.

Si se juntan los resultados de las Figuras 39, 40 y la Tabla 25, se puede presenciar una contradicción para el sector de servicios. Si bien el sector de servicios es el que presenta una mayor equidad de forma anual; también es el sector que maneja un mayor volumen con respecto a las plazas de empleo en total, lo cual ayuda a contrastar esta contradicción. Las industrias manufactureras, el comercio y la agricultura son sectores que, en volumen, manejan menores plazas de empleo, aunque sean los que presenten un mayor índice de brecha de plazas de empleo por género de forma anual.

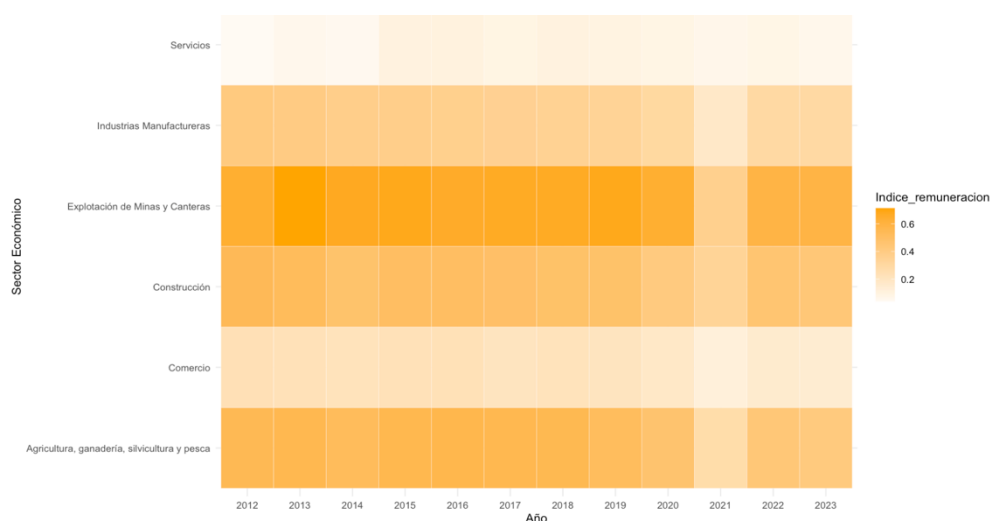
Por lo tanto, como el objetivo es aplicar estrategias para disminuir la desigualdad de género, el sector de servicios es el que debe tener una primera intervención, seguida de los 3 sectores que acumulan el 90%.

Índice de brecha de Remuneración económica por género.

Ahora viendo este estudio desde otra perspectiva, el análisis de la remuneración económica es otra variable que puede proporcionar información valiosa sobre una brecha de desigualdad económica por género.

Figura 41

Mapa de Calor anual sobre el índice de brecha de remuneración por género y por sector económico



Con la información visual que proporciona la Figura 41, se pudo identificar claramente resultados muy similares a los obtenidos en la Figura 39; tenemos al sector de explotación de minas ocupando el primer lugar con un índice alrededor de 0,6. En segundo lugar podemos identificar a los sectores de agricultura y construcción que presenta un índice alrededor de 0,4 al 0,5. En tercer lugar y con índices muy bajos, cercanos a 0, están los sectores de servicios y comercio, concluyendo así que estos sectores son los más equitativos. Todos estos índices presentan una inclinación hacia el género masculino. Sin embargo, el año 2021 presenta para todos los sectores

un índice bajo de aproximadamente 0,2, esto como posible consecuencia de la pandemia del COVID19.

Al analizar estos mismos sectores, sumando el total de la brecha de remuneración por cada uno sin tomar en cuenta el año, se obtiene una distribución muy característica.

Tabla 26

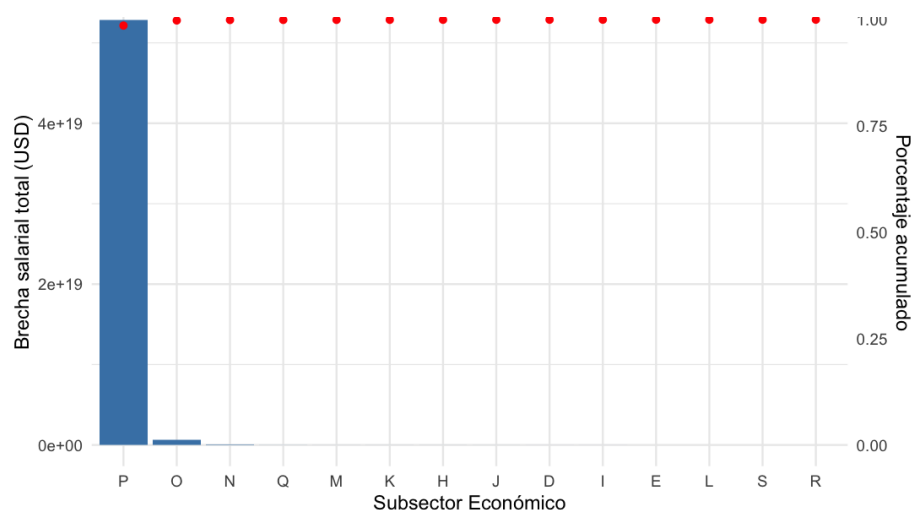
Brecha total de remuneración acumulativa por género en cada sector económico

Sector Económico	Brecha Total de Remuneración (Dólares)	Porcentaje Individual
Servicios	$5,36 \times 10^{19}$	99,78%
Industrias Manufactureras	$3,95 \times 10^{16}$	0,07%
Comercio	$3,14 \times 10^{16}$	0,05%
Agricultura, ganadería, etc.	$2,20 \times 10^{16}$	0,04%
Contrucción	$1,90 \times 10^{16}$	0,03%
Explotación de minas	$6,39 \times 10^{15}$	0,01%

Con la información proporcionada en la Tabla 26, el sector económico de servicios es el que abarca un 99,7% de la brecha salarial entre los diferentes sectores. Por lo cual resulta óptimo realizar un análisis más profundo en este sector, con el fin de identificar cómo se distribuye esa brecha de remuneración entre los diferentes subsectores del sector de servicio.

Figura 42

Diagrama de Pareto de la brecha de remuneración de género por cada subsector del sector económico.



En el sector de servicios, de acuerdo a la Figura 42, se pudo identificar 14 subsectores que fueron descritos en la Tabla 24. El subsector de enseñanza es el que acumula la mayor brecha de

remuneración dentro de este sector económico y en segundo lugar el subsector que abarca la administración pública.

Juntando los resultados de las Figuras 41, 42 y la Tabla 26; así como los resultados del índice de plazas de empleo, el sector de servicios es el que debe ser prioridad al momento de aplicar estrategias para reducir la brecha de género.

Clasificación del Tamaño Empresarial mediante K-NN a partir de Variables Operativas y de Género

A partir de la base de datos, para comprender mejor los factores operativos y estructurales que pueden llegar a afectar el tamaño organizacional de la empresa, se procedió a construir un modelo de clasificación utilizando el algoritmo k-NN. A partir de la brecha de remuneración económica (numérica), el total de plazas ofertadas (numérica) y la existencia única de personal femenino (dicotómica) se espera poder clasificar a las empresas en 4 categorías: Microempresa, Pequeña Empresa, Mediana Empresa y Grande Empresa. Además, con este análisis, se espera poder identificar posibles vínculos entre el tamaño empresarial y desigualdades estructurales en las condiciones laborales.

Utilizando un 80% de la base de datos como conjunto de entrenamiento y un 20% como conjunto de prueba, se obtuvo los siguientes resultados:

Tabla 27

Matriz de Confusión de la clasificación de las empresas

Predicción /Real	Grande Empresa	Mediana Empresa	Microempresa	Pequeña Empresa
Grande Empresa	5074	2589	48	747
Mediana Empresa	2664	8859	500	7000
Microempresa	290	2521	115394	26576
Pequeña Empresa	1322	8797	13272	35762

Con los resultados obtenidos en la Tabla 27, se puede identificar que el modelo k-NN realiza una clasificación mucho más exacta para las clases más definidas y frecuentes como lo son las microempresas, aunque también nos muestra resultados factibles para la clase de grandes empresas. Sin embargo, presenta un nivel medio en las categorías de mediana y pequeña empresa; esto puede darse debido a la cercanía de las características laborales que ocurren en estas 2 últimas.

Tabla 28

Métricas estadísticas por tamaño de empresa.

Tamaño de la Empresa	Sensibilidad	Especificidad	Precisión
Grande Empresa	0,54	0,98	0,76
Mediana Empresa	0,39	0,95	0,67
Microempresa	0,89	0,71	0,80
Pequeña Empresa	0,51	0,86	0,68

En la Tabla 28 se pueden identificar las diferentes métricas estadísticas del modelo para cada tamaño de empresa, teniendo así que el modelo funciona bien para clasificar a: microempresas y grandes empresas. Las pequeñas y medianas empresas son las que presentan el desafío más fuerte para el modelo al presentar métricas bajas de 0,39 y moderadas de 0,51 (sensibilidad). Ahora se procede a evaluar el modelo de forma global tomando en cuenta las métricas respectivas.

Tabla 29

Métricas estadísticas del modelo global de clasificación

Métrica Global	Resultado
Precisión	71,34%
Kappa	0,4915
P-value	< 2.2e-16
Balanced Accuracy Promedio	~73%

De acuerdo con los resultados de la Tabla 29, el modelo realizó una clasificación eficiente en un 71,34%, el coeficiente Kappa de 0,49 permite interpretar que el modelo es mejor que una selección al azar, apoyando a esta conclusión el p-value inferior a 0,05 el cual permite reiterar que el modelo es significativamente mejor a uno que clasifique todo como la clase más frecuente de la base de datos. Finalmente, al presentar un balanced accuracy promedio de 73% permite inferir que el modelo no se está orientando a la clase mayoritaria, sino que presenta una capacidad de clasificación consistente.

Conclusiones parciales y Limitaciones del Estudio

- El análisis realizado permite identificar una brecha sostenida a favor de los hombres tanto en plazas de empleo como en niveles salariales en múltiples sectores económicos a lo largo de los años; manifestándose de forma más notoria en sectores como la explotación de minas, agricultura y construcción.
- A pesar de que el sector de servicios muestra los índices anuales más equitativos en la remuneración y plazas de empleo, de forma acumulada representa el sector que más brecha presenta, específicamente en el subsector de enseñanza; por lo que es pertinente posicionarlo como sector de prioridad para realizar intervenciones de política laboral y aplicación de estrategias que reduzcan dicha brecha de género.
- El modelo de clasificación al incorporar variables como brechas salariales, presencia única de personal femenino y plazas de empleo ha permitido determinar que están vinculados al tamaño de la empresa, implicando que estas desigualdades no son solo de cultura organizacional, sino de la dinámica operativa de la empresa.
- Es pertinente aplicar un programa que tenga como fin reducir la brecha de género en el sector económico y específicamente en las empresas de tamaño mediano debido a que es el grupo que presentó mayor dificultad para realizar su clasificación.

LABORATORIO 4: ENFOQUE DE ROBUSTEZ Y VALIDACIÓN

Objetivo

- Determinar la validez de los diferentes modelos de series de tiempo para los diferentes sectores económicos y clustering utilizados en el laboratorio 2 para el periodo 2012-2023.

Métodos

En el laboratorio 2 se construyeron series de tiempo para analizar las plazas disponibles por sector económico.

La validación de los modelos se realizó mediante:

- Autocorrelación de residuos
- Normalidad
- Homocedasticidad
- Métricas de error

Para el aprendizaje no supervisado se aplicaron:

- Cohesión interna y separación externa
- Independencia relativa de variables

Discusión y Resultados

Series de Tiempo

Se analizaron las plazas disponibles por sector con registros trimestrales, ajustando y validando un modelo de serie de tiempo para cada sector económico.

- *Sector de Servicios*

Para el sector de servicios se utilizó un modelo SARIMA (2,0,1)(1,0,1). Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 30

Resultados de los supuestos de validez para la serie de tiempo del sector de Servicios

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,8291	Se considera que los residuos siguen una distribución normal.
	<i>Lilliefors</i> = 0,2954	
	<i>Anderson – Darling</i> = 0,5573	
Homocedasticidad	<i>ArchTest</i> = 0,5299	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,603	Los residuos no están autocorrelacionados.

De acuerdo con los resultados de la Tabla 30, los residuos del modelo del sector de servicios si cumplen el supuesto de normalidad. Se cumplen los supuestos de homocedasticidad y autocorrelación de los residuos, por lo tanto, se puede considerar el modelo como válido.

Las métricas de errores obtenidas en la Tabla 11, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados y capturando de mejor manera la dinámica de la serie.

- *Sector de Agricultura*

Para el sector de agricultura se utilizó un modelo de autorregresivo de orden 1. Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 31

Resultados de los supuestos de validez para la serie de tiempo del sector de Agricultura

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,4539	Se considera que los residuos siguen una distribución normal.
	<i>Lilliefors</i> = 0,2375	
	<i>Anderson – Darling</i> = 0,4885	
Homocedasticidad	<i>ArchTest</i> = 0,7128	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,09965	Los residuos no están autocorrelacionados.

Para el modelo AR del sector de agricultura se cumplen todos los supuestos de acuerdo con la información de la Tabla 31, por lo que se puede considerar adecuado y válido.

Las métricas de errores obtenidas en la Tabla 13, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados y capturando de mejor manera la dinámica de la serie.

- *Sector de Construcción*

Para el sector de construcción se utilizó un modelo SARIMA (0,0,1)(0,0,1). Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 32

Resultados de los supuestos de validez para la serie de tiempo del sector de Construcción

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,01621	Se considera que los residuos siguen una distribución normal.
	<i>Lilliefors</i> = 0,5129	
	<i>Anderson – Darling</i> = 0,2531	
Homocedasticidad	<i>ArchTest</i> = 0,8355	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,9528	Los residuos no están autocorrelacionados.

De acuerdo con los resultados de la Tabla 32 los residuos del modelo del sector de construcción si cumplen el supuesto de normalidad. Se cumplen los supuestos de homocedasticidad y autocorrelación, por lo tanto, se puede considerar el modelo como válido.

Las métricas de errores obtenidas en la Tabla 15, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados, capturando de mejor manera la dinámica de la serie y permitiendo considerarlo adecuado y confiable.

- Sector de Comercio

Para el sector de comercio se utilizó un modelo SARIMA (1,0,1)(0,0,0). Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 33

Resultados de los supuestos de validez para la serie de tiempo del sector de Comercio

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,007841 <i>Lilliefors</i> = 0,1166 <i>Anderson – Darling</i> = 0,00624	Los residuos no siguen un comportamiento normal
Homocedasticidad	<i>ArchTest</i> = 0,9712	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,9758	Los residuos no están autocorrelacionados.

De acuerdo con los resultados de la Tabla 33 los residuos del modelo del sector de comercio no cumplen el supuesto de normalidad; como implicación las predicciones puntuales son útiles pero los intervalos de confianza pueden generar incertidumbre. Se cumplen los supuestos de homocedasticidad y autocorrelación de los residuos, por lo tanto, se puede considerar el modelo como válido teniendo en cuenta el intervalo de confianza.

Las métricas de errores obtenidas en la Tabla 17, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados, capturando de mejor manera la dinámica de la serie y permitiendo considerarlo adecuado y confiable.

- *Sector de Industrias Manufactureras*

Para el sector de industrias manufactureras se utilizó un modelo SARIMA (0,0,1)(0,1,1,). Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 34

Resultados de los supuestos de validez para la serie de tiempo del sector de Industrias

Manufactureras

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,03128 <i>Lilliefors</i> = 0,01035 <i>Anderson – Darling</i> = 0,0111	Los residuos no siguen un comportamiento normal
Homocedasticidad	<i>ArchTest</i> = 0,8575	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,7926	Los residuos no están autocorrelacionados.

De acuerdo con los resultados de la Tabla 34 los residuos del modelo del sector de industrias manufactureras no cumplen el supuesto de normalidad; como implicación las predicciones puntuales son útiles pero los intervalos de confianza pueden generar incertidumbre. Se cumplen los supuestos de homocedasticidad y autocorrelación de los residuos, por lo tanto, se puede considerar el modelo como válido teniendo en cuenta el intervalo de confianza.

Las métricas de errores obtenidas en la Tabla 21, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados y capturando de mejor manera la dinámica de la serie.

- *Sector de Explotación de Minas y Calderas*

Para el sector de explotación de minas y calderas se utilizó un modelo SARIMA (1,0,1)(0,0,1). Para determinar su validez se procede a analizar los supuestos mencionados en la metodología con sus residuos:

Tabla 35

Resultados de los supuestos de validez para la serie de tiempo del sector de Explotación de

Minas

Supuesto	p-value	Conclusión
Normalidad	<i>Shapiro – Test</i> = 0,2968 <i>Lilliefors</i> = 0,2492 <i>Anderson – Darling</i> = 0,3183	Se considera que los residuos siguen una distribución normal.
Homocedasticidad	<i>ArchTest</i> = 0,2999	Los residuos son homocedásticos.
Autocorrelación	<i>Box – Ljung test</i> = 0,8617	Los residuos no están autocorrelacionados.

Para el modelo del sector de explotación se cumplen todos los supuestos de acuerdo con la información de la Tabla 15, por lo que se puede considerar adecuado y válido.

Las métricas de errores obtenidas en la Tabla 19, muestran la precisión y error que presenta el modelo seleccionado, considerado estos valores los más pequeños con respecto a otros modelos evaluados y capturando de mejor manera la dinámica de la serie.

Clustering Mediante el algoritmo K-means

Se procede a evaluar la validez de los clusters aplicados en función de sus remuneraciones, ventas_totales, plazas, plazas_hombres y plazas_mujeres.

- Cohesión Interna y Separación Externa

Tabla 36

Índices de Cohesión interna y Separación Interna

Grupo	Índice Average Silhouette
Datos Generales	0,991
Clúster 0	0,993
Clúster 1	0,728
Clúster 2	0,580

Con los resultados obtenidos en la Tabla 36, se puede determinar que la base, en general, logra una separación clara entre los grupos. El clúster 0 presenta una cohesión interna muy buena al acercar su índice a 1, concluyendo que sus miembros son homogéneos. El clúster 1 presenta un índice aceptable, con posibilidad de la existencia de elementos en fronteras con otros clústers. Finalmente, el clúster 2 es el que presenta menor homogeneidad.

Tabla 37

Índice Davies- Bouldin

Métrica	Valor Obtenido	Rango Esperado
Davies–Bouldin	-0,296	$0 \leq x \leq 1$

En la Tabla 37, el índice Davies–Bouldin presentó un valor negativo, lo que indica posibles problemas de normalización o diferencias de escala entre variables. No obstante, este resultado no invalida la clasificación, ya que se encuentra respaldada por los hallazgos positivos obtenidos en la Tabla 36.

- Interdependencia de variables

Tabla 38

Coefficientes de correlación de variables

Atributo	Ventas	Plazas	Plazas	Plazas	Remuneraciones
	Totales		Hombres	Mujeres	
Ventas Totales	1	0,479	0,500	0,320	0,105
Plazas	0,479	1	0,978	0,842	0,205
Plazas Hombres	0,500	0,978	1	0,712	0,215
Plazas Mujeres	0,320	0,842	0,712	1	0,136
Remuneraciones	0,105	0,205	0,215	0,136	1

La Tabla 38 evidencia redundancia entre las variables de plazas de empleo (totales, hombres y mujeres), lo que podría distorsionar el clustering. Se recomienda eliminar alguna de ellas, ya que esta autocorrelación pudo influir en el bajo índice Silhouette del clúster 2.

Conclusiones

- Los modelos de las series de tiempo fueron considerados válidos y confiables al cumplir, en su mayoría, los diferentes supuestos establecidos, pero se debe tener en cuenta que el periodo del año 2020-2021 generó efectos significativos en el comportamiento, lo que daba como consecuencia en algunos casos no normalidad.
- El clúster generado realizó una clasificación adecuada teniendo en cuenta las variables asignadas, pero recalando la existencia de variables que presentaban multicolinealidad y otras que debían manejarse con una normalización o mejora de escalas.

Limitaciones del Estudio

La base de datos brindada en un inicio para el estudio ha presentado en algunas variables utilizadas, errores de registros lo que genera dificultades y comportamientos anómalos que pueden haber afectado a algunos los modelos construidos.

CONCLUSIÓN GENERAL

El manejo adecuado de bases de datos es de suma importancia para garantizar la eficiencia, seguridad y confiabilidad de la información de cualquier tipo de institución. Pero en los registros de dichas bases de datos pueden presentarse fallas humanas, técnicas o de diseño que pueden generar interpretaciones erróneas; lo cual sucedió en el periodo 2006-2011 de la base de datos utilizada, por lo que para garantizar un análisis más confiable se omitió el análisis de este periodo. El presente estudio permitió analizar el comportamiento de los diferentes sectores económicos del Ecuador desde el año 2012-2023, permitiendo identificar la dinámica que posee cada sector económico.

Por medio de un análisis descriptivo, temporal y la aplicación de herramientas para la generación y validación de modelos matemáticos, se pudo obtener resultados globales sobre desempeños laborales, desigualdades de género en empleo y remuneración, así como otras tendencias que son únicas en cada sector económico.

Cada sector económico presenta dinámicas únicas debido a factores internos o externos, generando así diferentes variaciones en el empleo, las ventas, remuneración, entre otros. Un evento que marcó un efecto notorio en todos los sectores económicos fue la pandemia del COVID19, permitiendo identificar puntos débiles en algunos sectores, pues no estuvieron preparados para un evento de tal magnitud; siendo así, a partir de esta experiencia se sugiere que cada sector económico debe siempre tener un plan de contingencia que visualice el peor de los escenarios y así estar preparados a responder de la manera más efectiva. A pesar de ello, varios sectores mostraron una capacidad de recuperación y adaptación, confirmando lo mencionado acerca del plan de contingencia.

Finalmente, la identificación de brechas laborales y económicas entre el género masculino y femenino fue otro aspecto bastante notorio en este estudio, permitiendo inferir que esta situación sigue siendo un reto en todos los sectores económicos y se debería implementar políticas públicas que estén orientadas a la equidad y competencia en el campo laboral sin una preferencia de género.

BIBLIOGRAFÍA

- Proaño, S., Alvarado, E., Molina, C., & Mejía, O. (2019). Desarrollo económico local en Ecuador: Relación entre producto interno bruto y sectores económicos. *Revista de Ciencias Sociales (Ve)*, 82-98.
- Cevallos, J. (2021). Relación entre el PIB agregado y los sectores económicos: Un análisis de series de tiempo para Ecuador periodo 2000-2018. Riobamba, Chimborazo, Ecuador.
- Becerra, M., Valencia, E., & Revelo, R. (2021). Análisis del desempleo durante la pandemia COVID-19 y el impacto en diferentes sectores económicos del Ecuador. *Digital Publisher*, 442-451.
- Cueva, L. (2022). Productividad del sector servicios y el crecimiento económico ecuatoriano, 1990-2018 . *Productividad del sector servicios y el crecimiento económico ecuatoriano, 1990-2018* . Quito, Pichincha, Ecuador.
- Sanchez, J., & Soriano, J. (2021). Análisis del desarrollo de las empresas del sector manufacturero y su incidencia en el PIB del Ecuador. *Análisis del desarrollo de las empresas del sector manufacturero y su incidencia en el PIB del Ecuador*. Guayaquil, Guayas, Ecuador.
- Hernández, O., & Bravo, D. (2025). COVIDysuimpacto en la agricultura ecuatoriana. *Revista Económica*, 35-43.
- Barcena, A. (2021). La autonomía económica de las mujeres en la recuperación sostenible y con igualdad. CEPAL.
- Instituto Nacional de Estadística y Censos (INEC). (2023). *Registro Estadístico de Empresas (REEM)*. Instituto Nacional de Estadística y Censos: <https://www.ecuadorencifras.gob.ec/directoriodeempresas/>
- Moreira, M., Carvajal, A., & Barreno, M. (2020). ANÁLISIS DEL COMPORTAMIENTO ECONÓMICO-FINANCIERO DE LOS SECTORES EMPRESARIALES DE MILAGRO, ECUADOR. *ECA Sinergia*, 81-90.