UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Series temporales para el pronóstico de tasas de tributación - Grupo Bancolombia

---

INFORME DE PASANTIA PARA OPTAR POR EL TÍTULO DE MATEMÁTICO
PROGRAMA ACADÉMICO DE MATEMÁTICAS

Adriana Lucia Moreno Ruiz
Jefe inmediato: Alejandro Mejía Restrepo
Dirigido por: Luis Fernando Villarraga

Bogotá DC
Octubre de 2022

Resumen

En el presente documento se expone el desarrollo del proyecto de pasantía realizado en la Dirección de Impuestos del Grupo Bancolombia, el cual es líder financiero con más de 146 años de experiencia promoviendo el desarrollo económico sostenible. En la pasantía se implementa Python como lenguaje de programación para realizar diferentes análisis de datos y a partir de medidas estadísticas determinar e implementar los modelos más adecuados para realizar predicciones de cifras fiscales.

Palabras clave: Series de tiempo, aprendizaje automático, tributo, predicción.

Abstract: This document presents the development of the internship project carried out at the Tax Department of Bancolombia Group, which is a financial leader with over 146 years of experience promoting sustainable economic development.In the internship, Python is implemented as a programming language to perform different data analysis and, based on statistical measures, determine and implement the most suitable models for predicting tax figures.

Agradecimientos: A cada una de las personas de la Dirección de Impuestos Corporativos del Grupo Bancolombia, especialmente al equipo de Transformación Digital y Procesos, por cada una de las enseñanzas y oportunidades que han hecho parte de mi desarrollo personal y profesional, a los docentes del Proyecto Académico de Matemáticas, a mis compañeros y familia, por acompañarme y apoyarme en este enriquecedor proceso.

Índice

1. Objetivo de la pasantía

1.1. Objetivo general

Implementar soluciones matemáticas en el área de impuestos, usando herramientas de programación para aplicar modelos estadísticos y matemáticos en el análisis de datos de series tiemporales.

1.2. Objetivos específicos

- Comprender los conceptos básicos de impuestos en el sector financiero.
- Analizar los datos mediante modelos y medidas estadísticas para el pronóstico de tasas de impuestos.
- Desplegar el desarrollo de los modelos predictivos determinados a partir de las métricas obtenidas.

2. Preliminares

Actualmente, organizaciones, empresas y personas se enfrentan a una era de revolución tecnológica, enfocada en implementar avanzadas técnicas y herramientas con el fin de obtener ventajas competitivas, por lo cuál, es común escuchar cada vez con mayor frecuencia términos como: Machine Learning, ciencia de datos, inteligencia artificial, entre otros. Es allí donde las matemáticas retoman un papel fundamental, ya que permiten implementar diversos modelos aplicables a diferentes campos, aportando la capacidad de optimizar procesos en la industria basados en información. En esta sección, se definirán algunos aspectos relacionados con impuestos para entidades financieras, así como conceptos introductorios al análisis de datos en series de tiempo.

Observaciones: Para el desarrollo de este documento se han consultado diferentes bibliografías, a pesar de haber tomado algunos textos como referencia, las notaciones pueden cambiar con el fin de mantener continuidad y claridad en el informe. Adicionalmente, se notifica que todos los datos utilizados durante el desarrollo de la pasantía, corresponden a información confidencial interna del Grupo Bancolombia, por tal motivo, no es posible mostrar cifras, nombres y/o algún tipo de información real vinculada con la compañía, todas las cifras que se muestran en el informe se eligieron de manera aleatoria para dar una muestra de el proceso realizado.

2.1. Impuesto de renta

Los impuestos son pagos obligatorios que las personas y las empresas hacen al Estado, con el fin de que éste pueda proveer bienes y servicios a todos los colombianos. El impuesto de renta para entidades financieras, corresponde a un tributo en el cual se cobra un porcentaje de las ganancias de las empresas. [4]

2.1.1. Tarifa Nominal

La tarifa nominal corresponde al porcentaje de impuesto establecido en la normatividad colombiana, este sirve como un indicador de la carga tributaria directa que enfrentan las empresas en el país. Sin embargo, esta tarifa no representa el valor real de tributo final para cada empresa, ya que sobre esta tarifa se pueden aplicar deducciones y exenciones para reducir el monto a pagar.

2.1.2. Tasa Efectiva

La tasa efectiva de tributación es el porcentaje obtenido después de aplicar las correspodientes deducciones sobre el impuesto inicial. Este porcentaje puede ser mayor, menor o igual a la tarifa nominal.

2.2. Series de tiempo

Una serie temporal es un conjunto de observaciones generadas secuencialmente a lo largo del tiempo. Si el conjunto de momentos T_0 en los cuales se registran dichas observaciones es continuo, se dice que la serie temporal es continua, si este conjunto es discerto, se dice que la serie temporal es discreta. Por tanto, las observaciones de una serie de tiempo discreta realizadas en los tiempos $t_1, t_2, ..., t_n$, pueden ser denotadas por $x(t_1),\ x(t_2),\ ..., x(t_n)$.[1]
En este documento se consideran únicamente series de tiempo discretas donde las observaciones se realizan en un intervalo continuo h, cuando se tienen N valores sucesivos de una serie de este tipo, con observaciones realizadas en intervalos de tiempo equidistantes $t_0,\ t_0 + h,\ t_0 + 2h,\ ...,\ t_0 + Nh$, dichas observaciones se denotan como $x_1,\ x_2,\ ..., x_n$; Así, en función del intervalo de iempo establecido establecido h, se tienen series temporales con periodicidad diaria, mensual, trimestral, anual, etc.
Ya que el objetivo está centrado en realizar pronósticos a partir de estos conjuntos de observaciones, es posible estudiar características estadísticas que permitan determinar el modelo adecuado; en el caso de las series de tiempo se habla de tres componentes principales que se describen a continuación. [7]

- **Tendencia**: Comportamiento o movimiento suave de la serie a largo plazo, suele clasificarse como creciente o decreciente.
- **Componente estacional**: Representa movimientos de oscilación dentro en un periodo de tiempo discreto h.
- **Componente aleatoria**: Refleja variaciones aleatorias al rededor de los componentes de tendencia y estacionalidad.

Adicionalmente se definen a continuación algunas medidas estadísticas que se implementaron para realizar el análisis inicial de los datos y la construcción del modelo.

2.2.1. Media

Sea $X_t = \{x_1, x_2, x_3, ...x_n\}$ un conjunto de observaciones dado, se define la media como el valor obtenido al dividir la suma de todos los x_i entre la cantidad de observaciones del conjunto, esto es:

$$E[X_t] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2.2.2. Varianza

En términos estadísticos, la varianza es una medida de dispersión que representa las desviaciones o distancias entre un conjunto de observaciones con relación a su media. La varianza se calcula mediante la sumatoria de las desviaciones al cuadrado con respecto a la media, que en este caso se representa como $\bar{x}$, todo esto dividido entre el número total de observaciones. Esto es:

$$Var[X_t] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad [8]$$

2.2.3. Covarianza

La covarianza es una medida estadística que permite determinar el grado de variación conjunta de dos variables aleatorias. En el caso de las series temporales, la covarianza mide el grado de asociación lineal entre los valores de la misma serie que distan en k periodos de tiempo, es decir x_τ y $x_{\tau+k}$, de modo que, en las series temporales, la covarianza esta dada por:

$$Cov(X_t, X_{t+k}) = E(X_t)(X_{t+k}) \quad [3]$$

2.2.4. Serie temporal estacionaria

Una serie temporal se dice estacionaria si su media y varianza se mantienen constantes en el tiempo, esto es, para todo τ en T_0, se tiene:

$$\begin{aligned} E[X_t] &= \mu \\ Var(X_t) &= \sigma^2 \\ Cov(X_t,\ X_{t+k}) &= \gamma_k, \text{ para todo } k \text{ en } T_0 \quad [3] \end{aligned}$$

2.2.5. Ruido Blanco (White noise)

Se denomina ruido blanco a aquel conjunto de observaciones en una serie temporal, cuyos valores son independientes e idénticamente distribuidos a lo largos del tiempo, con media y varianza igual a cero. Este es un ejemplo clásico de serie de tiempo estacionaria.

2.2.6. Autocorrelación

En diversas ocasiones, los valores que toma la serie temporal no son independientes entre sí, es decir, los valores actuales y/o futuros, están explicados por el comportamiento de observaciones anteriores, de modo que, cuando se busca realizar pronósticos resulta fundamental determinar o medir esta dependencia, para ello se recurre a las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF).

2.2.7. Función de autocorrelación (ACF)

La auocorrelación mide la dependencia de las variables, separadas en k periodos, dicha función está determinada por:

$$\rho_j = \text{corr}(X_j,\ X_{j-k}) = \frac{\text{Cov}(X_j,\ X_{j-k})}{\sqrt{V(X_j)}\sqrt{V(X_{j-k})}} \quad [2]$$

2.2.8. Función de autocorrelación parcial (PACF)

La función de autocorrelación parcial permite determinar la correlación entre dos variables separadas por k periodos, a diferencia de la autocorrelación simple, en este caso no se considera la dependencia creada por los retardos intermedios existentes entre ambas.

$$\pi_j = corr(X_j,\ X_{j-k}/X_{j-1}X_{j-2}...X_{j-k+1}) = \frac{\text{Cov}(X_j - \bar{X}_j,\ X_{j-k} - \bar{X}_{j-k})}{\sqrt{V(X_j - \bar{X}_j)}\sqrt{V(X_{j-k} - \bar{X}_{j-k})}} \quad [2]$$

2.2.9. Procesos lineales estacionarios

Una vez se ha verificado que la serie de tiempo cuenta con las características de una serie estacionaria, es posible implementar diferentes procesos cuyo objetivo es determinar el comportamiento general de la serie, con el fin de predecir el comportamiendo de la misma en el futuro.

2.2.10. Procesos autoregresivos

Los modelos autoregresivos se adaptan particularmente a series en las cuales el valor actual X_t, puede explicarse en función de k valores pasados $x_{t-1},\ x_{t-2},...,\ x_{t-k}$, donde k determina el número de rezagos necesarios para pronosticar el valor actual.

2.2.11. Procesos autoregresivos de Orden 1: AR(1)

El proceso autoregresivo de orden uno, expresa que cada valor X_t, depende únicamente del valor pasado x_{t-1}, dicho proceso está expresado de la siguiente manera:

$$x_t = \phi x_{t-1} + \varepsilon_t,$$

donde ε_t es un proceso de ruido blanco con media cero, varianza constante σ^2 y ϕ es el parámetro. Posteriormente, se requiere verificar que el modelo AR(1) es estacionario, por lo cuál se deben satisfacer las condiciones:

- El modelo es estacionario en media:

$$E(X_t) = E(X_t = \phi X_{t-1} + \varepsilon_t) = \phi E(X_{t-1})$$

- El modelo es estacionario en covarianza:

$$\gamma_0 = E(X_t - E(X_t))^2 = E(\phi X_{t-1} + \varepsilon_t - 0)^2 = \phi^2 V(X_{t-1}) + \sigma^2 [6]$$

Este desarrollo autoregresivo de orden uno se empleó como prueba inicial para validar que se podría ajustar la metodología deseada en cada conjunto de datos. Adicionalmente se presenta a continuación la forma general para el desarrollo en orden superior, para estos casos se recurre a herramientas computacionales que facilitan la construcción del modelo.

2.2.12. Procesos autoregresivos de orden p AR(p)

De acuerdo al análisis anterior, un modelo autoregresivo de orden p estará determinado por:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \ldots ,$$

donde ε_t es un proceso de ruido blanco y $\phi_0, \phi_1, ..., \phi_p$ son los parámetros del modelo. [6]

2.2.13. Procesos de medias móviles

Los modelos de medias móviles son permiten determinar o explicar el comportamiento de la serie a partir del promedio ponderado de los errores pasados.

2.2.14. Proceso de medias móviles de orden 1 MA(1)

El proceso de medias móviles de orden 1 está determinado por:

$$X_t = \varepsilon_t - \theta\varepsilon_{t-1},$$

donde ε_t es un proceso de ruido blanco y θ es el parámetro.

Nuevmente se hace necesario verificar que el modelo sea estacionario, es decir:

- El modelo es estacionario en media:

$$E(X_t)E(\varepsilon_t - \theta\varepsilon_{t-1}) = E(\varepsilon_t) - \theta E(\varepsilon_{t-1}) = 0$$

- El modelo es estacionario en covarianza

$$\gamma_0 = E(X_t - E(X_t))^2 = E(X_t)^2 = E(\varepsilon_t - \theta\varepsilon_{t-1})^2 \quad [6]$$

De manera similar a la contrucción realizada en el proceso autoregresivo de orden 1, se realiza el desarrollo de un modelo de medias móviles de orden uno determinando bajo qué condiciones el modelo es estacionario y construir posteriormente el modelo ARIMA deseaso. A continuación se muestra la forma general para un proceso de media móviles de orden superior.

2.2.15. Proceso de medias móviles de orden q MA(q)

El modelo de medias móviles de orden q está dado por:

$$X_t = \theta_0 - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} - \varepsilon_t \cdots,$$

donde ε_t es un proceso de ruido blanco y $\theta_0, \theta_1, ..., \theta_q$ son los parámetros para el modelo. [6]

2.2.16. Proceso autoregresivo de medias móviles

Inicialmente se contruyó por separado un modelo autoregresivo y uno de medias móviles, pero usualmente las series no se comportan bajo una única carácterística, éstas suelen presentar variaciones lo cuál implica que podrían presentar características de un proceso autoregresivo (AR) y de medias móviles (MA) simultáneamente, en este caso, se dice que la serie puede ser modelada por un proceso ARMA. Dicho modelo contará con dos parámetros fundamentales que representan la composición autoregresiva y de medias móviles en la serie, de modo que, si se determina un proceso ARMA(p,q),

esto indica que la representación de la serie cuenta con p términos autoregresivos y q términos de media móvil. Algebráicamente esto es:

$$X_t = c + \underbrace{\phi_1 X_{t-1} + \cdots + \phi_p X_{t-p}}_{AR(p)} + \underbrace{\theta_1 \varepsilon_{t-1} + \cdots + \theta_q X_{t-q} + \varepsilon_t}_{MA(q)},$$

donde ε_t es un proceso de ruido blanco y $c, \phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q$, son los parámetros del modelo. En este caso, las condiciones de estacionariedad, están dadas de manera similar que en los procesos por separado de AR(p) y MA(q). De acuerdo a esta construcción, un modelo ARMA(1,1) viene dado por:

$$X_t = \phi X_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1},$$

donde ε_t es un proceso de ruido blanco y ϕ, θ son los parámetros del modelo.

2.2.17. Procesos lineales no estacionarios

Los modelos presentados hasta ahora funcionan bajo el supuesto de estacionariedad de la serie temporal, pero la mayoría de datos empleados en este estudio, y en general, aquellas series que reflejan fenómenos económicos, no son estacionarias; para modelar series en estas condiciones se hace necesario establecer otro tipo de procesos que permitan alcanzar las características inicales necesarias para contruir el modelo.

2.2.18. Proceso autoregresivo integrado y de media móvil ARIMA(p,d,q)

Un proceso integrado consiste en realizar transformaciones a los valores de la serie temporal, con el fin de conseguir una representación estacionaria de la misma, de modo que sea posile implementar los modelos descritos para dichas series. en el caso de ARIMA, el proceso a realizar es diferenciar d veces la serie y posteriormente a esta serie diferenciada será posible aplicar el modelo ARMA(p,q), luego la serie original estará dada por ARIMA(p,d,q), la cuál se denomina serie de tiempo autoregresiva integrada de media móvil. Allí p denotará el número de términos autoregresivos, d el número de veces que se ha diferenciado la serie para hacerla estacionaria y q el número de términos de media móvil. Este modelo está denotado algebráicamente por:

$$X_t^d = c + \underbrace{\phi_1 X_{t-1}^d + \cdots + \phi_p X_{t-p}^d}_{AR(p)} + \underbrace{\theta_1 \varepsilon_{t-1}^d + \cdots + \theta_q \varepsilon_{t-q}^d + \varepsilon_t^d}_{MA(q)} \quad [6]$$

2.2.19. Proceso estacional autoregresivo integrado y de media movil ARIMA(p,d,q)(P,D,Q)$_s$

Una serie de tiempo se estudian observaciones separadas por un periodo constante h, cuando dicho periodo es menor a un año, es común encontrar patrones sistemáticos cada cierto periodo s, estas variaciones son identificados en la serie temporal como *Factores estacionales* y deben ser incluidas en el modelo asignando parámetros que representen esa componente estacional de la serie. Este tipo de procesos cuenta con dos cmponentes.

- $ARIMA(p,d,q)$: esta componente modela la dependencia regular que está asociada a n observaciones consecutivas.
- $ARIMA(P,D,Q)$ esta componente modela la dependencia estacional, que está asociada a observaciones separadas por s periodos.[2]

el modelo está representado algebráicamente por:

$$X_t = c + \overbrace{\phi X_{t-1} + \cdots + \phi_p X_{t-p}}^{AR(p)} + \overbrace{\theta_1 X_{t-s} + \cdots + \theta_P X_{t-Ps}}^{SAR(P)}$$

$$+ \underbrace{\varepsilon_t - \varphi_1 \varepsilon_{t-1} - \cdots - \varphi_q \varepsilon_{t-q}}_{MA(q)} - \underbrace{\Theta_1 - \varepsilon_{t-s} - \cdots - \theta_Q \varepsilon_{t-Qs}}_{SMA(Q)}$$

2.2.20. Valores atípicos (Outliers)

En un conjunto de datos dado, se denomina valores atípicos a aquellas observaciones cuyos valores se alejan de la línea de mínimos cuadrados o mantienen un comportamiento alejado de la media y la mediana, en otras palabras, son valores muy distantes a las demás observaciones del mismo conjunto de datos. Usualmente los valores atípicos son ocasionados por:

- Errores de procedimiento.
- Acontecimientos extraordinarios.
- Valores extremos ocasionados por eventos particulares.
- Causas no conocidas.

Los valores atípicos distorsionan los resultados de los análisis, y en general alteran las medidas estadísticas de los datos, por lo cuál resulta indispensable identificar su causa y evaluar la manera correcta de trabajar con ellos dentro del conjunto de datos; en algunos casos no presentan un valor de importancia, de modo que pueden ser extraídos de la muestra de datos sin problema, en otros casos será necesario emplear alguna técnica para suavizar el impacto que estos presentan dentro del modelo predictivo.

2.3. Médidas de precisión

En el pronóstico de series temporales, y en cualquier proceso predictivo, es necesario contar con métricas de evaluación de resultados que permitan determinar la precisión del modelo empleado, durante el desarrollo de la pasantía, las métricas empleadas principalmente se describen a continuación.

2.3.1. MAPE

El MAPE es una medida empleada para determinar la precisión en modelos predictivos, esta métrica expresa el error como una distancia porcentual que se calcula de la siguiente manera:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\,\% \quad [5]$$

donde n corresponde al número de observaciones que se tendrán en cuenta para la prueba del modelo, y_i son los valores reales y $\hat{y}_i$ su correspondiente predicción.

2.3.2. MAE

Esta medida corresponde a la media del error absoluto, de manera similar al MAPE, representa el error como una distancia promedio, pero en este caso no será un valor porcentual. Su cálculo se realiza de la siguiente manera:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad [5]$$

donde n corresponde al número de observaciones que se tendrán en cuenta para la prueba del modelo, y_i son los valores reales y $\hat{y}_i$ su correspondiente predicción.

2.3.3. Coeficiente de determinación R cuadrado (R2)

Esta medida permite determinar la proporción de la variación entre los valores observados y la regresión (ajustada), dicho coeficiente está dado por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2}$$

donde el numerador refleja la suma de los cuadrados del error y el denominador toma en cuenta la suma total de los cuadrados corregida, la cual representa la variación en los valores de respuesta que idealmente serían explicados con el modelo. [8]

2.3.4. Error cuadrático medio (RMSE)

Este estadístico determina la desviación estándar en los errores de predicción y se calcula de la siguiente forma : [5]

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

donde n corresponde al número de observaciones que se tendrán en cuenta para la prueba del modelo, y_i son los valores reales y $\hat{y}_i$ su correspondiente predicción.

3. Modelo para la predicción de Tasa Efectivas de Tributación

El proyecto está enfocado en el análisis de comportamiento en series temporales de la tasa efectiva de tributación para cada una de las compañías del Grupo Bancolombia, para ello se desarrollan dos modelos, el primero toma en cuenta la información histórica mensual de tasa efectiva; el segundo modelo se genera a partir de la predicción indivdual de cada uno de los conceptos que explican el cambio de tasa nominal a tasa efectiva, es decir, aquellos montos que corresponden a deducciones y exenciones dentro del impuesto de renta.

3.1. Generalidades y aspectos clave

El impuesto de renta para empresas corresponde a un tributo que se debe pagar anualmente al Estado de acuerdo a las ganancias y movimientos financieros que se hayan realizado en ese periodo de tiempo, es por ello, que cada organización debe realizar una previsión mensual del posible cobro que se va a generar, ya que esto puede resultar determinante para definir de manera estratégica los movimientos financieros a realizar los siguientes meses del año, este cálculo suele realizarse de manera manual, realizando la depuración del impuesto a partir del capital actual, la tarifa nominal y las deducciones y exenciones que se presenten hasta ese momento. Ya que este proceso se realiza de manera repetitiva y no cuenta con una medida de precisión, estaba generando sobrecotos en cuanto al tiempo de ejecución de la tarea, además, se podía correr el riesgo de proporcionar una cifra poco acertada, en ese caso la organización estaría expuesta a un déficit económico con respecto al valor real a pagar. Por esta razón, se hace necesario recurrir a modelos estádisticos y matemáticos avanzados para generar las provisiones de renta a partir del pronóstico en series de tiempo, lo cuál reducirá operabilidad, tiempo de ejecución y, además, cuenta con medidas que permiten evaluar la presición del pronóstico generado.

3.2. Datos

El pago de impuesto de renta se realiza de manera anual, sin embargo, esta periodicidad no resulta muy favorable para ningún modelo predictivo, ya que sería necesario contar con mucha información historica para poder disponer de una cantidad suficiente de datos, por esta razón, los datos utilizados para la predicción corresponden al registro mensual de la tasa efectiva para 32 compañías del Grupo Bancolombia. Este estudio se realizó con un total de 967 registros, correspondientes a los meses desde enero de 2019 hasta mayo de 2022.

3.3. Limitaciones

Durante el análisis de los datos y el desarrollo del modelo, se presentaron algunas limitaciones relacionadas directamente con la naturaleza de los datos. Inicialmente se realizó el análisis descriptivo para 32 compañías, posteriormente, se determinó que solo sería posible desarrollar el modelo predictivo para 6 de ellas, algunas compañías fueron descartadas debido a que son libres de impuestos o tributan dentro de otra compañía; en otros casos, fue necesario descartar compañías debido a la deficiente cantidad de registros, principalmente correspondían a compañías que fueron creadas recientemente. Adicionalmente, durante el año 2020 se presenta una irregularidad en los datos de todas las compañías, esto viene dado como causa directa de una pandemia. Debido a que estos valores atípicos alteraban el comportamiento del modelo, fue necesario normalizarlos mediante técnicas estadísticas para disminuir el impacto que causaban en los resultados.

3.4. Herramientas

Todas las etapas del análisis, desde la limpieza y clasificación de datos, hasta la entrega de resultados y entrega de la herramienta ejecutable para generar los pronósticos mensualmente, fueron desarrolladas empleando el lenguaje de programación Python, el cual, a pesar de estar enfocado principalmente en la programación orientada a objetos, cuenta con gran diversidad de librerías para el desarrollo de modelos predictivos y manejo de grandes cantidades de datos. Algunas de las librerías empleadas en el desarrollo de este proyecto fueron:

- Numpy
- Pandas
- Matplotlib
- Statsmodel
- Scikit-learn

4. Metodología

El proyecto fue desarrollado mediante cuatro etapas que permiten la construcción de un modelo estructurado de acuerdo a la necesidad que se presentaba en la compañía; estas estapas son:

- Entendimiento del problema: Este proceso incluye la comprensión de nuevos conceptos y la naturaleza de los datos.
- Limpieza, transformación y análisis descriptivo de los datos: En este caso se realiza análisis descriptivo para cada compañía y adicionalmente se realiza el mismo análisis para el consolidado de todo el grupo.
- Modelamiento: Esta etapa incluye la evaluación de diferentes características estadísticas para determinar el mejor modelo.
- Entrega de resultados, conclusiones y recomendaciones: Adicional a la entrega de resultados se realiza el desarrollo de una herramienta ejecutable que calculará el pronóstico mensualmente.

4.1. Entendimiento del problema

El entendimiento del problema se ha desarrollado a lo largo del documento, como se explicó anteriormente, uno de los objetivos principales está en disminuir la operatividad y brindar pronósticos más acertados que pueden ser evaluados mediante métricas de precisión. La naturaleza de los datos está explicada principalmente por los movimientos financieros que genera la entidad de manera mensual, teniendo en cuenta que a partir de dicha información se planea generar una provisión anual.

4.2. Limpieza, transformación y análisis descriptivo de los datos

Se toman los datos de una tabla que suministra información al tablero de control de la DIC administrado por Power BI, de allí se extrajo información correspondiente a 32 compañías y un total de 37 columnas que además de brindrar indormación sobre, país, fecha y compañía, hacen referencia a diferentes valores tributarios que se evalúan en la dirección; Para este estudio se extrajeron solamente 7 columnas que contenían información reelevante como año, mes, país, compañía, tasa efectiva acumulada, gasto por impuesto coon eliminacines consolidada, utilidad con eliminaciones consolidada; las demás columnas no se consideraron representativas.

El formato en el cuál se almacena dicha información, y con el cuál se trabajó incialmente es el siguiente:

PAIS	COMPAÑÍA	MES	AÑO	TASA EFECTIVA ACUMULADA	GASTO POR IMPUESTO CON ELIMINACIONES CONSOLIDADA	UTILIDAD CON ELIMINACIONES CONSOLIDADA
PAIS 1	COMPAÑÍA 1	1	2019			
PAIS 2	COMPAÑÍA 2	1	2019			
PAIS 3	COMPAÑÍA 3	1	2019			
PAIS 1	COMPAÑÍA 4	1	2019			
PAIS 2	COMPAÑÍA 5	1	2019			
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
PAIS 1	COMPAÑÍA 29	7	2022			
PAIS 1	COMPAÑÍA 30	7	2022			
PAIS 3	COMPAÑÍA 31	7	2022			
PAIS 4	COMPAÑÍA 32	7	2022			

Figura 1: Estructura inicial del conjunto de datos

En la sección preeliminar de este documento se define la serie temporal como un conjunto de observaciones registradas en tiempos determinados, de modo que para trabajar este conjunto como una serie temporal, se debe contar con una columna o variable que denote la fecha, en programación este tipo de variable se le asigna el formato *DateTime* y en este caso se definiió la estructura dd/mm/yy, este valor cuál se obtuvo a partir de la información almacenada en las columnas MES y AÑO, y luego se adiciona a los datos en el formato indicado de la siguiente manera:

PAIS	COMPAÑÍA	MES	AÑO	FECHA	TASA EFECTIVA ACUMULADA	GASTO POR IMPUESTO CON ELIMINACIONES CONSOLIDADA	UTILIDAD CON ELIMINACIONES CONSOLIDADA
PAIS 1	COMPAÑÍA 1	1	2019	1/01/2019			
PAIS 2	COMPAÑÍA 2	1	2019	1/01/2019			
PAIS 3	COMPAÑÍA 3	1	2019	1/01/2019			
PAIS 1	COMPAÑÍA 4	1	2019	1/01/2019			
PAIS 2	COMPAÑÍA 5	1	2019	1/01/2019			
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
PAIS 1	COMPAÑÍA 29	7	2022	1/07/2022			
PAIS 1	COMPAÑÍA 30	7	2022	1/07/2022			
PAIS 3	COMPAÑÍA 31	7	2022	1/07/2022			
PAIS 4	COMPAÑÍA 32	7	2022	1/07/2022			

Figura 2: Estructura de los datos adicionando la columna fecha

A partir de esta información se realiza el análisis descriptivo implementando python como herramienta para clasficicar y graficar la información de cada una de las compañías; inicando con un conteo de datos se procede a descartar casos de información incompleta, estos correspondían a compañías que dejaron de existir, por tal motivo no era necesario incluirlas en algun estudio, adicionalmente se contaba con compañías que fueron creadas recientemente, estos casos también fueron descartados temporalmente ya que no se contaba con registros suficientes para implementar un modelo predictivo. Al finalizar este proceso se presenta el análisis descriptivo determinando algunas medidas estadísticas que anticipan como proceder en el desarrollo de los modelos.

Dentro del estudio descriptivo se identifica la presencia de outliers, una vez realizada la validación con el equipo experto para verificar que no se debe a errores de digitación, se procede a realizar la evaluación del comportamiento de cada compañía en particular determinando que la mayoría de estos valores atípicos surgieron como efecto de una pandemia, que además de alterar el transcurso natural de los procesos económicos, dio paso a cambios en normativas nacionales, lo cuál se vio reflejado en valores alejados de la media. En otros casos se verificó que se presentaron algunas fusiones entre compañías, lo cuál ocasionó variaciones permanentes en las tasas de tributación, esto causó una división del comportamiento en la serie temporal, por esto se determinó que para ciertas compañías no sería suficiente la calidad y cantidad de datos.

En los casos restantes se considera posible implementar el modelo predictivo realizando un manejo especial con los valores atípicos, para ello se realiza una la evaluación de métodos de imputación o normalización de dichos valores, con el fin de disminuir el impacto que estos presentarían en el modelo; en este estudio se toma en cuenta el coeficiente de determinación $R2$ para definir el método de imputación más efectivo en cada caso. De acuerdo al proceso realizado con cada una de las compañías se tuvieron en cuenta los siguientes métodos para la normalización de outliers:

- Interpolación polinomial: Se trataron los outliers como valores restantes y porteriormente se realizó el ajuste de un polinomio para determinar el valor que mejor se ajustaba al mes correspondiente de acuerdo a las características de la serie. Este método se emplea de manera efectiva en aquellos casos donde los outliers corresponden a pocos casos aislados.
- Normalización por media: El estudio se realizó con datos mensuales, registrando un factor de estacionalidad anual, lo cuál permite deducir que el comportamiento de cada mes presentará variaciones de manera similar en cada año con respecto al anterior. De acuerdo a esto, el procedimiento de normalización fue el siguiente: Si se presenta un outlier en el mes x del año i, denotado como x_i, se calculará la media de los valores registrados en los meses x_{2019}, x_{2020}, x_{2021}, x_{2022} y por último se reemplaza el outlier con la media obtenida.

Este método se implementó para disminuir el impacto de valores atípicos que correspondían a meses consecutivos, se consideró efectivo ya que gráficamente se mantiene el comportamiento de la serie, pero disminuye la escala en los meses en que se registraron dichos valores. Esto se puede observar mediante un ejemplo de la siguiente manera: Considere una serie de observaciones registrada con periodicidad trimestral, se asume que se presentaron outliers durante tres periodos consecutivos, al aplicar el método de normalización por media, se observa gráficamente como los valores siguen representando el comportamiento de la serie pero en menor escala, se esta manera podemos entregar información que se adapta mejor al modelo predictivo y que se asemeja al comportamiento real de los datos.

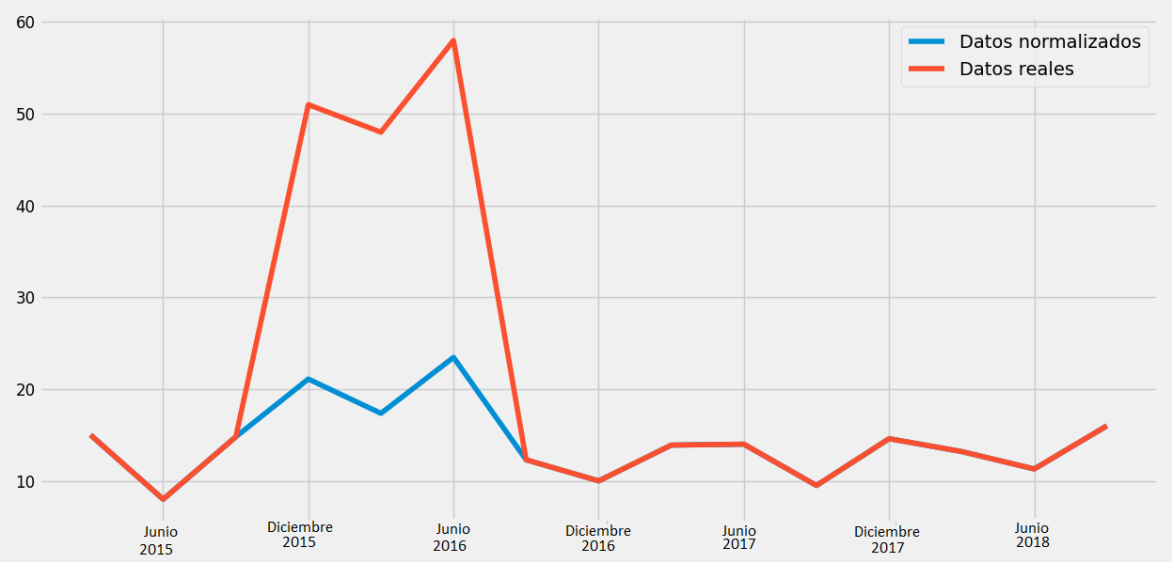

Figura 3: Ejemplo de normalización de datos con media

∗ Estos valores fueron elegidos de manera aleatoria para el ejemplo y no representan ningún tipo de dato perteneciente a la organización.

Finalizado el proceso de limpieza y transformación, se presenta el análisis descriptivo al equipo de expertos identificando las características mencionadas anteriormente, con esto se concluyó que solo 6 de las compañías que cumplían con determinada cantidad y calidad de datos serían sometidas a pruebas en los modelos predictivos.

4.3. Modelamiento

De acuerdo a las características estadísticas determinadas en el paso anterior, se probaron de manera simultánea dos modelos analíticos diferentes, estos fueron, Holt-Winters que consiste en representar la serie mediante la técnica de suavizado exponencial, y SARIMA, que se definió anteriormente como un proceso autoregresivo integrado de medias móviles. Considerando la versatilidad de los modelos, las características de la serie temporal que representa cada compañía y los objetivos de la organización con el desarrollo de estas iniciativas analíticas, se eligió el modelo SARIMA para realizar el pronóstico de las 6 compañías seleccionadas. Teniendo en cuenta las políticas de privacicad y confidencialidad de la organización, se realizará una descripción del proceso efectuado empleando diferentes datos elegidos aleatoriamente con el fin de explicar a detalle cada paso realizado en la construcción del modelo.

Una de las primeras condiciones mencionadas para trabajar estos modelos en series temporales fue la estacionariedad, para ello se determina que la serie sea constante en media y varianza, en la práctica esta es una condición muy poco común, por ello los métodos ARIMA representan una ventaja ya que permiten realizar una transformación a la serie dotándola de esas carácterísticas necesarias para el desarrollo del modelo, dicha trasnformación se ejecuta por el método de diferenciación como un retroceso en el cuál las observaciones son redefinidas como: $y'_t = y_t - y_{t-1}$, este sería el caso de una primera diferenciación, de esta manera se consigue obtener una representación estacionaria de la serie como se observa:

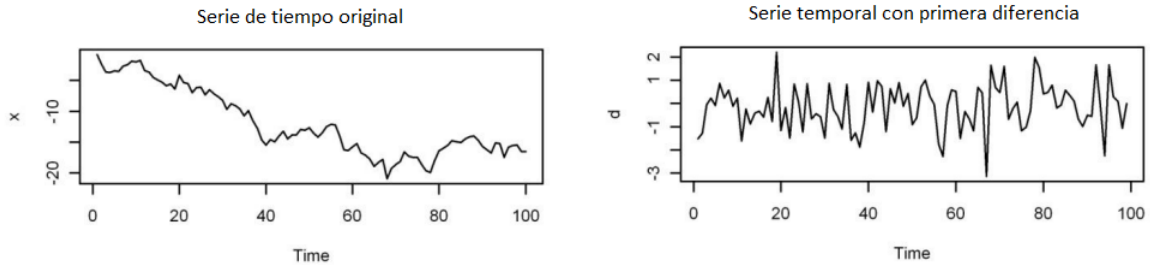

Figura 4: Ejemplo de transformación de serie para hacerla estacionaria

∗ Imágen tomada de: http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D

Dependiendo de cada caso particular fue necesario realizar el proceso de diferenciación más de una vez, es importante tener en cuenta que el número de veces que se realiza el proceso de diferenciación, representa el parámetro d en el proceso $ARIMA(p,d,q)$, por ejemplo, para las series que fueron diferenciadas dos veces, el modelo es de la forma $ARIMA(p,2,q)$.

Después de verificar que la serie temporal es estacionaria, se utilizaron las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) para determinar entender el comportamiento de la serie y de acuerdo a ello, determinar los parámetros p y q del modelo. Recordando que el término p corresponde al comportamiento autoregresivo de la serie, este se ouede determinar mediante la función de autocorrelación parcial ya que esta representa la correlación entre la serie y sus retrasos; en el caso del parámetro q se analizó la de autocorrelación parcial.

Para cada una de las compañías se realizaron las respectivas gráficas (ACF) y (PACF) mediante la librería *statsmodels*, allí se identifican bandas azules que se pueden interpretar como bandas de error, esto quiere decir que los retrasos representados dentro de esa franja no son estadísticamente significativos. Los valores que se encuentran fuera de las bandas azules se consideran retrasos significativos, por tanto, estos serán los candidatos para los parámetros de cada modelo. Un ejemplo de ellos se muestra a continuación:

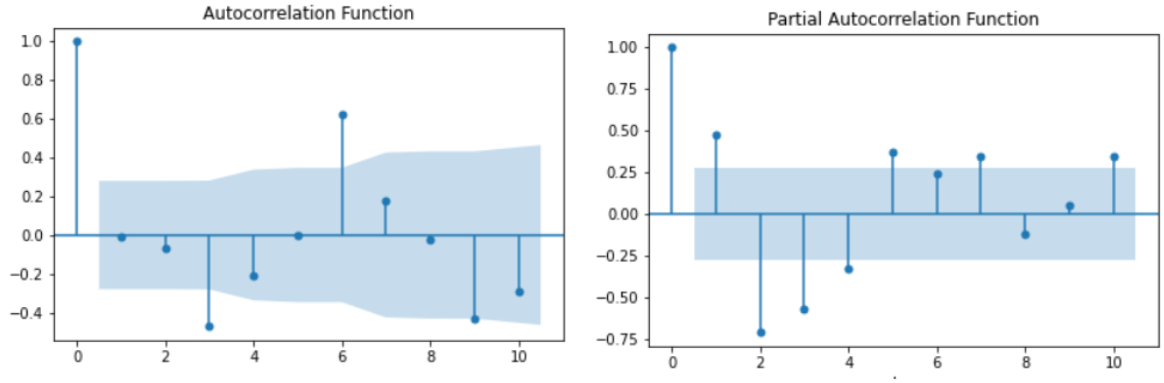

Figura 5: Gráficas ACF y PACF

∗ Estos valores fueron elegidos de manera aleatoria para el ejemplo y no representan ningún tipo de dato perteneciente a la organización.

En la gráfica de autocorrelación se identifican retrasos significativos en 3 y 6, luego el componente de medias móviles para el modelo podría ser de $MA(3)$ o $MA(6)$; ahora, en la gráfica de autocorrelación parcial, los retrasos más significativos están en 2 y 3, entonces el componente autoregresivo estaría determinado por $AR(2)$ o $AR(3)$. De esta manera se establecieron cuatro posibles modelos para cada serie temporal (cada compañía), teniendo en cuenta que, para describir un ejemplo se definió el parámetro de diferenciación como $d = 1$, los cuatro posibles modelos para el ejemplo serían:

- $ARIMA(2,1,3)$
- $ARIMA(2,1,6)$
- $ARIMA(3,1,3)$
- $ARIMA(3,1,6)$

Existen diferentes métodos para seleccionar el modelo adecuado, de acuerdo las prácticas analíticas de la organización y la naturaleza de los datos con los cuáles fue desarrollado el modelo, se determinó la métrica $MAPE$ como factor determinante para elegir el mejor modelo. Para ello se implementa el método de validación cruzada, seleccionando dos

subconjuntos de la siguiente manera, inicalmente se toma el 80% de los datos como conjunto de entrenamiento, el 20% restante se determina como conjunto de pruebas, es decir, se ejecuta el modelo tomando en cuenta solo el 80% de los datos, posteriormente se realiza el pronóstico para las fechas correspondientes al 20% restante, luego se comparan los pronósticos comparan los pronósticos con los valores reales del conjunto de pruebas y se evalúa mediante la métrica $MAPE$, los parámetros que entreguen un mejor resultado en este proceso, serán seleccionados para generar los pronósticos, de acuerdo a las capacidades del modelo y los requerimientos de la organización, se determinó un horizonte de 6 meses, es decir, cada vez que se realiza el proceso, se obtendrán pronósticos para los siguientes 6 meses.La principal métrica que se tiene en cuenta para medir la precisión de los pronósticos es el $MAPE$, en el desarrollo realizado las métricas para las 6 compañías obtuvieron un $MAPE$ en un rango de 5%a 43%; se hace necesario aclarar que el modelo se ejecutará mensualmente entrenándose con más valores, por tanto se espera que dichas métricas mejoren, especialmente para las compañías con $MAPE$ superior a 20%.

Uno de los objetivos de la DIC es generar modelos que no depedan de las personas y que se mantengan en el tiempo, por este motivo se propuso el desarrollo de una herramienta ejecutable que calculará mensualmente los pronósticos de los 6 meses siguientes para cada una de las compañías incluídas en el modelo, los cuáles se generan en un 95% de manera automática ya que el funcionario solamente deberá dar click en la herramienta para ejecutarla, incialmente la aplicación creará una carpeta donde se almacena la información mensual de dichos pronósticos, esto con el fin de que se pueda realizar un control del modelo y se cuente con información suficiente en caso de que sea necesario reajustar el modelo; tras realizar cada ejecucuión el programa creará para cada compañía las siguientes salidas:

- Gráfica de la serie de tiempo original
- Gráfica con la serie temporal más los pronósticos generados para 6 meses
- Archivo.csv con las métricas obtenidas y los parámetros seleccionados para el modelo.

Figura 6: Formato de salida métricas y parámetros

- Archivo .csv con la información de: Tasa real, tasa normalizada, predicción, valor mínimo y valor máximo. Donde las cifras de valor mínimo y valor máximo determinan un intervalo de confianza para el pronóstico realizado. Dicha información está ordenada en el siguiente formato:

Fecha	Tasa real	Tasa normalizada	Pronóstico	Valor mínimo	Valor máximo
1/01/2015	.	.			
1/02/2015	.	.			
1/03/2015	.	.			
1/04/2015	.	.			
1/05/2015	.	.			
1/06/2015	.	.			
1/07/2015	.	.			
1/08/2015			.	.	.
1/09/2015			.	.	.
1/10/2015			.	.	.
1/11/2015			.	.	.
1/12/2015			.	.	.
1/01/2016			.	.	.

Figura 7: Formato de salida de resultados

5. Recomendaciones

- Se recomienda a los equipos encargados, mantener actualizada la información del tablero de control ya que de ello depende el funcionamiento de la herramienta ejecutable que realizará los pronósticos de manera automática.
- Dado que el MAPE es una medida de precisión expresada en términos porcentuales, se solicita que esta no sea la única me dida para determinar si el pronóstico es bueno o malo, siempre será recomendable contar con el juicio experto para evaluar los resultados antes de descartar un modelo.
- Será una ventaja mantener el mismo formato en que se almacena la información ya que esto permitirá construir bases de datos sólidas que faciliten continuar con la implementación de modelos predictivos en la DIC.

6. Conclusiones

- Uno de los prinicipales factores identificados en el desarrollo de la pasantía es como las matemáticas están inmersas en diversas áreas, desde el aporte teórico, pero pincipalmente desde el aporte lógico deductivo, entregando la capacidad de implementar diferentes herramientas con el fin de crear soluciones que favorecen a organizaciones y personas en general.
- Aunque no fue posible incluir todas las compañías del Grupo Bancolombia en el modelo predictivo, se entrega un desarrollo adaptable, en el cuál se podrán seguir incluyendo compañías y procesos a medida que los datos cumplan con las condiciones suficientes para implementar el modelo desarrollado.
- Uno de los retos de la industria es automatizar en mayor medida los modelos, en este proceso se observó como trasciende el rol de las personas; durante el desarrollo de la pasantía, esto se vio reflejado especialmente en la interpretación de pronósticos, ya que el $MAPE$ no es siempre determinante, se hizo necesario entender lo que representa cada métrica y de acuerdo a su experiencia y conocimiento contrastar que tan representativas eran las cifras entregadas por el modelo.
- Este proceso ha representado una experiencia muy enriquecedora personal, profesional y académicamente, he logrado fortalecer muchas habilidades, conocer nuevos conceptos y especialmente enfrentar uno de os retos más comunes entre quiénes nos dedicamos inicalmente a estudiar la teoría, esto es poder transmitir nuestros saberes y conclusiones a personas de otras áreas, lo cuál representa el verdadero valor del conocimiento.

Referencias

[1] Jenkins G. Box, G. *Time Series Analysis: Forecasting and Control.* Probability and Statistics. Wiley, 2016.

[2] Davis R Brockwell, P. *Introduction to Time Series and Forecasting.* Texts in Statistics. Springer, 2016.

[3] Vásquez F. Series de tiempo. https://franciscoariel.github.io/site/estadistica/series_tiempo/. 2022.

[4] Observatorio fiscal. El sistema tributario en colombia. https://www.ofiscal.org/tributacion. 2021.

[5] Sanahuja P. Métricas de evaluación de rendimiento para predicciones de series temporales. https://polmartisanahuja.com/metricas-de-evaluacion-de-rendimiento-para-predicciones-de-series-temporales/. 2022.

[6] R.H. Stoffer S.S., Shumway. *Time Series Analysis and Its Applications, Fourth Edition.* Springer Texts in Statistics. Springer, 2017.

[7] R. Tsay. *Analysis of Financial Time Series (Third Edition).* John Wiley Son, 2010.

[8] R.E. Walpole. *Probabilidad y Estadística Para Ingenierías Y Ciencias (9.a ed.).* Pearson educación, 2012.