La Ciencia de Datos como Elemento que aporta al Sector Gubernamental de la Sociedad

Sebastián Meza Pastrana

# Introducción

Este informe se adentra en el emocionante mundo de la Ciencia de Datos aplicada a problemas del mundo real, explorando tres casos de estudio que demuestran el poder y la versatilidad de las técnicas de análisis y predicción en el ámbito gubernamental. Cada proyecto representa un esfuerzo valioso para mejorar la toma de decisiones y la eficiencia operativa a través de la aplicación ingeniosa de algoritmos de Machine Learning y análisis de datos. El caso 1 nos mostrará como minimizar los delitos fiscales, seguido del caso 2, un modelo de predicción de deuda pública y por último, el caso 3, modelo para optimizar el recaudo de impuestos. A medida que avancemos en este informe, descubriremos cómo estas iniciativas han abordado desafíos específicos y han logrado resultados impactantes.

# Modelo de Caso 1: Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of Sao Paulo. [1]

El proyecto se llevó a cabo en el municipio de São Paulo, Brasil, con el objetivo de aplicar técnicas de Ciencia de Datos para predecir crímenes fiscales relacionados con el sistema de impuestos sobre servicios. Se realizó en un período que abarcó los años 2016, 2017 y 2018. La meta principal fue mejorar la toma de decisiones en los planes de auditoría fiscal del gobierno, aumentando la eficiencia en la detección de delitos fiscales y, en última instancia, incrementando los ingresos fiscales.

El proyecto se dividió en varias etapas clave. En primer lugar, se realizó una selección de características relevantes, eligiendo atributos específicos que podrían influir en la predicción de los crímenes fiscales. Estos atributos incluyeron el valor anual de impuestos declarados en facturas, categorizados según si el contribuyente se identificó como una empresa especializada en un servicio único o como una pequeña empresa que utiliza procedimientos simplificados.

Luego, se extrajo y procesó la información de las auditorías fiscales realizadas durante los años mencionados. Esta información se organizó en un conjunto de datos que se utilizó para entrenar y probar diversos algoritmos de Aprendizaje Automático, como Redes Neuronales, Naive Bayes, Árboles de Decisión, Regresión Logística, Random Forests y Ensemble Learning.

Se realizó un análisis detallado de cada algoritmo según métricas de rendimiento como precisión, exactitud, sensibilidad y especificidad. Los hallazgos indicaron que Random Forests se destacó por su desempeño en la predicción de crímenes fiscales, demostrando una precisión notable del 64%, exactitud del 67% y una especificidad del 40% en comparación con los demás algoritmos.

Finalmente, se validó el modelo ajustado por Random Forests utilizando datos fiscales no vistos de 2018. Los resultados en esta etapa de validación confirmaron la capacidad del modelo para generalizar y predecir crímenes fiscales en un contexto real empíricos para cada uno de los modelos para calcular los ahorros porcentuales en capital regulatorio. Los impactos positivos y negativos del modelo para cada stakeholder identificado se presentan en la Tabla I (Ver Tabla I).

TABLA I

Impactos del Proyecto caso 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupos de interés:** | **Tipos de impacto** | **Impacto positivo** | **Impacto negativo** |
| **Contribuyentes** | Cumplimiento fiscal mejorado | Aumento de confianza en el Sistema Tributario | Mayor carga administrativa |
| Reducción de multas y sanciones | Ahorro de recursos económicos | Mayor escrutinio por parte de las autoridades competentes |
| **Gobierno** | Eficiencia operativa | Aumento de recaudo fiscal traducido en mayor presupuesto | Posible resistencia o rechazo de los contribuyentes |
| Decisiones basadas en datos | Mayor capacidad para planificar y asignar los recursos | Desafíos en la implementación técnica |
| **Empresarios - Inversionistas** | Reducción de incertidumbre y confianza | Facilita la planificación financiera | Mayor carga administrativa |
| Identificación de negocios y contratación con estado | Descubrimiento de nuevas áreas de negocio. | Mayor exposición a regulaciones y auditorías |

*Fuente: Elaboración propia*

# Modelo de Caso 2: Machine Learning Algorithm for Mid-Term Projection of EU Member States’ Indebtedness. [2]

El proyecto fue realizado en Svishtov, Bulgaria, por 5 investigadores del Departamento Crédito y Finanzas de la Escuela de Economía de Tsenov. Se enfocó en el análisis y pronóstico a mediano plazo de las deudas públicas de los estados miembros de la Unión Europea. El objetivo principal del proyecto fue determinar la viabilidad de predecir las deudas públicas utilizando un modelo de aprendizaje automático. El proceso se desarrolló en varias etapas clave, destacando la recopilación de una amplia gama de indicadores macroeconómicos, fiscales, monetarios, globales y de convergencia que influencian los niveles de deuda pública. Estos indicadores fueron utilizados para desarrollar un modelo de predicción utilizando la técnica de regresión de bosques aleatorios.

El proyecto se llevó a cabo en dos iteraciones. En la primera iteración, se utilizó un conjunto completo de 33 indicadores para entrenar el modelo y analizar la relación entre estos indicadores y la relación deuda/PIB. En la segunda iteración, se redujo el conjunto de indicadores a los 8 más significativos, con el objetivo de evaluar su impacto en la predicción de la deuda pública. La técnica de Random Forest se utilizó para modelar y predecir el comportamiento futuro de la relación deuda/PIB en los estados miembros de la UE.

Los resultados del proyecto mostraron que los indicadores más influyentes en la relación deuda/PIB eran los pagos de intereses sobre la deuda pública, los gastos fiscales en relación con el PIB y la balanza de cuenta corriente en relación con el PIB. Además, se observó un cambio en la composición de los cuatro estados miembros más endeudados en el mediano plazo, lo que tuvo implicaciones tanto para esos países como para la UE en general.

No se presentaron resultados cuantitativos del modelo usado, pero, el modelo demostró una alta precisión y una baja variabilidad en las predicciones. Los valores de error medio absoluto, error cuadrático medio y error cuadrático medio se mantuvieron dentro de rangos aceptables, lo que indicó la robustez y eficacia del modelo de regresión de bosques aleatorios en este contexto. Adicionalmente, se identificaron que países aumentarían o disminuirían la relación deuda/PIB. A continuación, se presentan los aspectos positivos y negativos identificados para cada stakeholder (Ver Tabla II).

TABLA II

Impactos del Proyecto caso 2

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupos de interés:** | **Tipos de impacto** | **Impacto positivo** | **Impacto negativo** |
| **Estados miembros de UE** | Indicadores de deuda | Identificación de riesgos de deuda | Aumento de la carga de deuda para algunos países y aumento de riesgo sistémico en UE. |
| Estabilidad Financiera | Contribución a la estabilidad financiera en países con menor deuda | Posible contagio de crisis de deuda a países con relaciones comerciales |
| **Mercado Global** | Rendimiento de los bonos | Valoración más precisa de los riesgos asociados a tasas de interés | Aumento de la volatilidad en los mercados financieros |
| Mercado de valores | Identificación de oportunidades de inversión | Posible disminución en la confianza de los inversores en países con alta deuda |

*Fuente: Elaboración propia*

# Modelo de Caso 3: Optimización en el recaudo de impuestos en el Sector Público.

El proyecto fue realizado por Andrés Moreno, ingeniero de sistemas de profesión. Dicha investigación en alianza con CAOBA de la cual es miembro. El objetivo del proyecto es acompañar a la Secretaría de Impuestos Distritales de la ciudad de Bogotá, para detener los diferentes problemas relacionados con el cumplimiento de impuestos por parte de los contribuyentes mediante el uso de la analítica, para mitigar los problemas y proporcionar herramientas a los fiscalizadores y enfocarse en los casos de los contribuyentes con un estado más crítico.

El proyecto se desarrolló en varias etapas en las cuales, inicia desde la comprensión profunda de los requisitos y la recopilación de datos pertinentes hasta la presentación de resultados validados. La fase de entendimiento del negocio y datos permitió un análisis exhaustivo de los requisitos, junto con la aplicación de técnicas de análisis exploratorio de datos (EDA) para descubrir patrones, tendencias y desafíos potenciales dentro de los conjuntos de datos. La preparación de datos fue crucial para asegurar la calidad y coherencia de la información, involucrando tareas de limpieza, imputación de valores y normalización de formatos.

Posteriormente, se crearon modelos descriptivos que visualizaron de manera efectiva elementos clave como el número de contribuyentes, recaudaciones en el tiempo y tipos de obras, utilizando gráficos y estadísticas descriptivas. Además, modelos analíticos creativos, basados en análisis de series temporales y minería de datos, se emplearon para abordar cuestionamientos específicos y detectar patrones excepcionales. Estos enfoques avanzados, como el análisis de valores atípicos y análisis de tendencias en el tiempo para la alineación urbana, permitieron una comprensión más profunda y refinada de los datos. La validación final se llevó a cabo a través de la evaluación con expertos del negocio, donde los resultados fueron presentados y verificados por profesionales con experiencia en el área fiscal, asegurando así la coherencia y relevancia de los hallazgos y ofreciendo oportunidades para mejorar y perfeccionar los modelos implementados.

Finalmente, los resultados significativos que se obtuvieron fueron de aumento de eficiencia operacional y en el recaudo, en especial el adecuado análisis en la liquidación de impuestos de delineación urbana donde se identificaron 5.200 obras y el impuesto de la sobretasa a la gasolina. Se lograron identificar los impactos positivos y negativos para cada uno de los involucrados (Ver Tabla III).

TABLA III

Impactos del Proyecto caso 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupos de interés:** | **Tipos de impacto** | **Impacto positivo** | **Impacto negativo** |
| **Autoridad Fiscal** | Eficiencia operacional | Mejora en la priorización y enfoque en contribuyentes críticos, lo que agiliza y optimiza la fiscalización. | Posible resistencia o descontento por cambios en procesos de fiscalización. |
| Incremento potencial de recaudos | Mayor recaudación al identificar y cobrar obras sin liquidar impuestos. | Desacuerdo o rechazo de los contribuyentes afectados por las multas y cobros pendientes. |
| **Contribuyentes** | Cumplimiento de obligaciones | Información clara sobre las obligaciones | Posible carga adicional o ajustes |
| Mayor confianza en el proceso de declaración | Mayor seguridad y transparencia en el cumplimiento de las obligaciones fiscales. | Posible incertidumbre o dudas sobre la interpretación de las normativas fiscales. |

*Fuente: Elaboración propia*

# CONCLUSIÓN.

En conclusión, los modelos expuestos en los tres casos prácticos resaltan el impacto transformador de la Ciencia de Datos en diferentes sectores. Desde la predicción de crímenes fiscales en São Paulo hasta la proyección de la deuda pública en la Unión Europea y la optimización del recaudo de impuestos en Bogotá, estos proyectos han demostrado la capacidad de las técnicas analíticas avanzadas para mejorar la toma de decisiones y la eficiencia operativa. A través de la comprensión profunda de los datos, la aplicación de modelos descriptivos y el empleo de técnicas analíticas creativas como análisis de series temporales y minería de datos, se ha logrado una comprensión más rica de patrones y tendencias, proporcionando conocimientos valiosos para expertos en áreas fiscales y gubernamentales. Esto se traduce en eficiencia y optimización en la obtención de los recursos por parte del estado que finalmente, financia el gasto público.

# IV. BIBLIOGRAFÍA.

|  |  |
| --- | --- |
| [1] | A. Ippolito y A. C. Garcia Lozano, «Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of Sao Paulo,» *SciTePress,* Vols. %1 de %2 In Proceedings of the 22nd International Conference on Enterprise Information Systems - Volume 1: ICEIS, pp. 452-459, 2020. |
| [2] | S. Zarkova, D. Kostov, P. Angelov, T. Pavlov y A. Zahariev, «Machine Learning Algorithm for Mid-Term Projection of the EU Member States’ Indebtedness,» *Risks,* vol. 11, nº 11040071, 2023. |