Postulación de proyecto final
Diplomado Big Data para Políticas Públicas
Universidad Adolfo Ibáñez

1. Integrantes y roles dentro del proyecto

Priscila Rodríguez - Jefe o responsable del proyecto
Eduardo Jiménez - Científico de Datos y Analista de Datos
Nicolás Torrealba - Científico de Datos y Analista de Datos
Kiumarz Goharriz - Responsable de visualización de Datos

2. Objetivo del proyecto

El objetivo del proyecto es sistematizar y visualizar información de las causas no contenciosas del Tribunal de Defensa de la Libre Competencia (TDLC) y, posteriormente, **predecir** cuál es la probabilidad de que el organismo emita una sentencia favorable para el denunciante.

3. Objetivos del modelo de ciencia de datos

Para este proyecto se proponen tres (3) objetivos del modelo de ciencia de datos:

Objetivo 1: Realizar web scraping y text mining para descargar y analizar la información disponible desde la página web del TDLC.

Objetivo 2: Desarrollar una visualización en Shiny con estadísticas relevantes de las causas no contenciosas del TDLC.

Objetivo 3: Desarrollar un modelo que permita predecir la probabilidad de que el TDLC emita una sentencia favorable al denunciante.

4. Breve descripción del proyecto

Durante los últimos años, si bien el número de causas no contenciosas que el TDLC ha debido analizar, no ha variado significativamente, si lo ha hecho el impacto mediático de sus sentencias, toda vez que las mismas han afectado de distinta forma la vida de las personas. Casos emblemáticos como la colusión de las tres principales cadenas de farmacias en el precio de 220 medicamentos (2009), la colusión en el precio de los compresores de los refrigeradores entre Whirpool y su competidora Tecumseh (2012), la colusión de seis importantes navieras en el proceso de contratación de transporte marítimo de automóviles (aún en análisis) o bien, los famosos casos de colusión del mercado de los pollos, supermercados, confort y pañales, dan luces de la relevancia del impacto de las sentencias de este organismo en la sociedad.

No obstante, a pesar de la importancia de las decisiones del TDLC, no existen dentro de la página web del organismo módulos de visualización, que permitan de manera rápida tener una primera mirada de los mercados que son analizados con mayor frecuencia, las materias más relevantes, ni el número de veces en que el TDLC acoge o rechaza las solicitudes de los denunciantes en las causas no contenciosas, entre otras muchas estadísticas relevantes. En este sentido, un primer objetivo de este proyecto de Big Data para Políticas Públicas, es que todos aquellos interesados en la información que emite este organismo regulador, puedan visualizar de forma rápida e ilustrativa, las estadísticas más relevantes, para lo cual se propone crear una plataforma con Shiny.

Un segundo objetivo, mucho más ambicioso, por cierto, es poder predecir en base a los expedientes de las causas no contenciosas de los último diez años, la probabilidad de que una nueva causa se acepte o se rechace. La importancia de este ejercicio, radica en que los involucrados podrían ajustar su estrategia ganadora de forma mucho más eficiente, de acuerdo al resultado que se obtenga del modelo.

5. Nivel de madurez de los datos

El siguiente cuadro presenta el nivel de madurez de los datos considerados para el proyecto:

Categoría	Área	Nivel de Madurez	Descripción
Cómo se almacena la información	Acceso	Básico	Se puede acceder a los datos, pero a través de *Web scraping*
	Almacenamiento	Básico	Archivos PDF
	Integración	Avanzado	Toda la información está accesible desde una única fuente, la cual es actualizada de forma permanente
Qué información se recolecta	Relevancia	Avanzado	Existe toda la información relevante y es suficiente para resolver el problema sin hacer transformaciones importantes
	Calidad	Intermedio	Están todos los datos, sin embargo es necesario utilizar técnicas de *text mining* para sistematizar la información
	Frecuencia	Avanzado	Se incorpora nueva información en tiempo real
	Granularidad	Avanzado	Detalle a nivel de resoluciones individuales
	Historia	Avanzado	Todos los datos se mantienen en una única fuente de información
Otros	Privacidad	Avanzado	La información utilizada en el proyecto es de carácter público
	Documentación	Intermedio	Si bien no existe un diccionario de datos, los actos administrativos emitidos por el TDLC, mantienen una estructura que permite realizar una identificación de manera relativamente sencilla