RESUMEN MESA TEMÁTICA: “CIENCIA DE DATOS PARA POLÍTICAS PÚBLICAS”

El uso de datos para la toma de decisiones en el sector público se ha visto revolucionado con las herramientas de big data y analytics. La posibilidad de procesar grandes volúmenes de información en un menor tiempo plantea nuevos desafíos para las instituciones públicas, las que hoy tienen la oportunidad de enfrentar antiguos problemas con nuevas herramientas.

Los casos que se revisarán permitirán conocer más de cerca cómo se han llevado a cabo soluciones de ciencia de datos para problemas de política pública, considerando la mirada desde la industria y desde agentes públicos, pues cada uno de estos actores es fundamental para la introducción de innovaciones en la administración pública. Asimismo, las presentaciones muestran la manera en la que herramientas de código abierto como R, se plantean como una alternativa importante para el desarrollo de innovaciones en el sector público, tendientes a mejorar la gestión del Estado.

MARCO GENERAL DE LA SITUACIÓN DE CIENCIA DE DATOS EN RELACIÓN DESARROLLO DE LAS POLÍTICAS PÚBLICAS.

En la actualidad se habla de tópicos como Modernización, Transformación digital, Innovación, Ciencia de datos, Inteligencia artificial, asumiendo que tienen el mismo significado para todas las personas. Sin embargo, se hace necesario aclarar qué se entenderá por cada uno de ellos y cómo se relacionan entre sí. La presentación busca entregar elementos teóricos para entender los conceptos señalados. Esto, con el objeto de abordar las aplicaciones de la ciencia de datos; particularmente, la emergente utilización de R como *software open source* en la gestión pública. Lo anterior, desde un marco conceptual claro y compartido por los distintos actores involucrados en estas tareas.

En primer lugar, modernización considera los elementos en el plano físico: conectividad, servidores, máquinas, inter-operatividad. En estos casos, por lo general, nos referiremos al *hardware* involucrado. Cuando se comenzó con el impulso modernizador en el Estado justamente se abordó este tema como el principal.

En segundo lugar, transformación digital se relaciona con el proceso de digitalización de trámites. También se considera la incorporación de tecnología a los procesos que soportan la operación del Estado.

La innovación, por su parte, tiene relación con cambios en la manera en la que se hacen las cosas y con el modo en el que nos relacionamos. Esto puede venir acompañado de elementos tecnológicos o no. La innovación, como es entendida en instituciones como el Laboratorio de Gobierno, debiese estar centrada en las personas. En este marco conceptual, entonces, la ciencia de datos es comprendida como una disciplina que está al servicio de las personas.

Una segunda dimensión que se abordará en esta presentación son las implicancias políticas y éticas vinculadas a la ciencia de datos, por cuanto los datos y particularmente la ciencia de datos, puede ser entendida como una disciplina que logra representar la realidad únicamente de manera parcial. Al desarrollar modelos en base a datos históricos, se construyen modelos predictivos que reproducen su entrenamiento, es decir, se predice el futuro en base al pasado. El desafío propuesto es trabajar con los datos para modificar trayectorias y evitar reproducciones, es decir, prevenir o generar intervenciones tempranas. En este sentido, se propone dar una reflexión respecto a la utilización de los datos para uso empresarial, campañas políticas, etc. y cuestionarse sobre los límites del uso que estos tienen.

Luego, se mencionarán algunos esfuerzos impulsados desde el Estado, como Transparencia y Datos abiertos, con la mirada puesta en las implicancias que tiene para el Estado una mayor apertura en el acceso a los datos.

Finalmente, en los desafíos pendientes, se menciona la relación entre empresas y Estado en lo relativo al manejo de los datos y sus análisis. Se revisarán temas como datos abiertos, capacidades instaladas, *software open source (R)*, entre otros.

EXPERIENCIAS DE TRABAJO CON GRANDES VOLUMENES DE DATOS EN PROCESAMIENTO DE CENSO 2017

En los censos de población se recoge un volumen de datos importante. No todos estos datos cumplen con los requisitos de calidad establecidos, los errores pueden tener su origen en el informante, en el registro de los datos por parte del censista, en el reconocimiento óptico de los formularios, etc.

Dada la importancia que tienen esos datos para el desarrollo de políticas públicas, tanto a nivel nacional como a nivel local, y el alto costo de obtener la información, se hace un gran esfuerzo para aprovechar los datos obtenidos. Estos son revisados buscando errores o incoherencias. Una vez identificados conflictos en los datos, se busca ejecutar el mínimo de intervenciones que produzca un conjunto coherente, evitando descartar los registros.

Este proceso de validación y edición se aplica también en encuestas, muchas veces en forma manual, pero en los Censos, por el volumen de datos, solo es practicable por medios automáticos.

En Censo 2017 por primera vez se introdujo el uso de R para la validación y edición de los datos.

En esta presentación queremos transmitir algunas experiencias de adoptar R para este proceso, relatar lo que creemos fueron los mayores problemas y las soluciones que el proyecto elaboró para atenderlos. Principalmente en lo que se refiere al manejo de un volumen de datos importante, con sus efectos en los tiempos de procesamiento y la gestión del espacio en disco.

Los temas tratados se relacionan con:

- técnicas generales para aprovechar operaciones en que R es eficiente y evitar operaciones en que el performance de R no es el deseado,
- estrategia de particionamiento para permitir procesamiento paralelo,
- organización de la ejecución de procesos que permita maximizar el paralelismo, dentro de los límites de recursos de los equipos computacionales disponibles,
- selección de formato de datos en disco, balanceando la facilidad de manejo, el rendimiento y el uso de espacio.

También se discutirá el efecto de los requerimientos de reproducibilidad en la organización del trabajo y el uso de BBDD relacionales como repositorios de datos.

Creemos que la presentación será de interés no solo para quienes procesen encuestas y censos, sino para todos quienes requieran aplicar procesos de limpieza y depuración sobre volúmenes importantes de datos.

USO DE APIS Y MODELOS DE SERIES DE TIEMPO PARA ESTIMAR DOTACIÓN DE PERSONAL

Esta presentación conjuga 2 elementos relevantes para el desarrollo de la ciencia de datos en el sector público. Por un lado se encuentra la relevancia del modelo construido, en términos sustantivos; y por otro, el tipo de desafíos de infraestructura informática que es necesario enfrentar en muchos casos.

Parte del proceso de Reforma Procesal Civil (RPC), implementado por el Ministerio de Justicia y Derechos Humanos, involucra determinar la dotación de personal que será destinada a cada territorio jurisdiccional. Para llevar esto a cabo se contó con estudios de cargas de trabajo para los distintos perfiles (e.g. juez, administrador, administrativo de causas, entre otros) y procedimientos (e.g. Ejecutivo, Voluntario, Sumario y otros) además de los ingresos históricos por materia y juzgado proporcionados por la Corporación Administrativa del Poder Judicial.

Con estos inputs se procedió a ajustar un modelo de redes neuronales de dos capas usando la librería `forecast`. Pero surge un problema al momento de calcular el output del modelo, y radica en la capacidad del hardware existente en la administración pública.

Pese a que el problema a abordar difícilmente califica como *big data*, la falta de un servidor o equipos de escritorio adecuados llevó a tener que generar alguna solución alternativa. Esta solución fue crear una base de datos PostgreSQL y usando el paquete `plumber` se realizaron todos los cálculos del lado del servidor, y del lado del cliente bastó con un computador de escritorio con capacidad no mayor a la de un tablet promedio del año 2018.

El resultado del modelo es una salida JSON que indica con cuántos funcionarios se debe contar como mínimo para abordar las cargas de trabajo trimestrales durante todo el año para cada tupla juzgado-perfil-procedimiento, respetando la legislación laboral vigente. Por ejemplo, la solución que entrega el algoritmo, entre otros elementos, indica que el 3er juzgado civil de Santiago requiere N administrativos de causas el año 2020 para resolver todas las tareas relacionadas a los M ingresos en distintas materias que reciben durante el mismo periodo.

Entonces, el desafío enfrentado durante este proceso fue doble: llevar todo a *DigitalOcean*, que fue la plataforma que permitió sortear las dificultades de infraestructura informática del Ministerio; y luego proceder a experimentar modelos que respondieran a la necesidad de optimizar la dotación de personal en los territorios jurisdiccionales. La presentación abordará ambos desafíos en conjunto.

TRES EXPERIENCIAS DE ÉXITO EN EL USO DE LA CIENCIA DE DATOS PARA EL SECTOR PÚBLICO

Un cambio de paradigma muy notorio de los últimos años, es cómo la innovación dejó de provenir exclusivamente de la investigación desarrollada en universidades. En este sentido, la web y cloud han sido dos oleadas tecnológicas trasformadoras, lideradas desde la industria, que han dejado huellas permanentes en términos de metodologías, técnicas, estándares y tecnología.

Esta presentación aborda la manera en la que los cambios tecnológicos recién mencionados pueden contribuir a una mejor gestión del sector público. Uno de los cambios particularmente notables en términos de elaboración de políticas públicas ha sido el surgimiento de grandes volúmenes de datos, fenómeno que ha estado acompañado de la aparición de tecnologías que permiten procesar estas nuevas fuentes de información y sacar provecho de ellas. A partir del denominado Big Data y de las fuentes de información no estructurada es posible enfrentar nuevos problemas que no necesariamente pueden abordarse a partir de fuentes tradicionales, como encuestas y/o datos administrativos.

Para ejemplificar lo anterior se revisan 3 experiencias concretas: 1) las políticas de "cero papel", 2) la reforma al SERNAC y 3) el sistema anti evasión en el pago del transporte implementado por el Metro de Valparaíso.

Respecto a la política de “cero papel” se describe una solución inteligente para detectar de manera automática las instituciones públicas que solicitan dentro de sus trámites alguna documentación en papel. Con esta información es posible generar un dataset que reúne la información producida por todas las instituciones públicas, lo cual disminuye la cantidad de trámites y evita la duplicación de ciertos procedimientos.

La segunda experiencia que se aborda se relaciona con la necesidad de fortalecer el rol fiscalizador del SERNAC. Se ha detectado que ciertos segmentos de la población no utilizan necesariamente los canales formales establecidos por el SERNAC para dirigir sus reclamos. En lugar de ello, utilizan las redes sociales. Considerando esta situación, se encuentra en fase de implementación un sistema automatizado que monitorea constantemente algunas redes sociales, con el objeto de detectar reclamos sistemáticos hacia ciertas empresas. En base a dicha información el SERNAC puede girar hacia una postura más activa en lo que respecta a su rol fiscalizador, sin tener que esperar a que la ciudadanía utilice los canales formales para llevar a cabo reclamos.

La última experiencia abordada corresponde a un sistema de reconocimiento facial implementado en el Metro de Valparaíso, que tiene el objetivo de detectar el mal uso que puedan tener los beneficios relacionados con el costo del pasaje. Así, lo que se intenta es reconocer si el beneficiario corresponde a la persona que efectivamente está utilizando el beneficio. Para ello se recurre a tecnología de visión computacional.

Las tres experiencias abordadas buscan motivar la discusión respecto a las potencialidades que tiene la ciencia de datos en lo relativo al mejoramiento de la gestión pública. Asimismo, se busca poner sobre relieve la capacidad que tiene la industria de generar innovaciones útiles para problemas de política pública.