1.360.2-79

ACTA DE REUNIÓN INNOVACIÓN PÚBLICA DIGITAL CIENCIA DE DATOS SECRETARÌA DE LAS TIC

ACTA No. 126 A

FECHA: Santiago de Cali, 18 de septiembre de 2020

HORA: 9:00 am a 10:00 am

ASUNTO: Reunión ORACLE Análisis de Datos Socializar a Rentas lo que se hizo.

LUGAR: Sesión virtual

ASISTENTES: Sonia Yamileth Castro Yama- Asesora – Gobernadora, Martha Carrasquilla, Augusto Mendonca, Antonio Cantillo, Guilherme Diniz, Adalberto Martínez, Cristian Petro, Mariela Ivonne Sinisterra Muñoz.

ORDEN DEL DÍA:

1. Saludo.
2. Entrega y demostración de puesta en marcha del servidor Oracle, y el análisis de los archivos de rentas en el servidor de Oracle.

Para empezar la Dra. Sonia Yamileth Castro Yama manifiesta que se han logrado muchos objetivos, y hay mucho por hacer, más en gestión pública y sin recursos.

1) A continuación, Evelin comenta que es un Workshop muy interactivo, tiene un nivel de detalle mayor, más profundo y más extenso. La propuesta era pedirle a Augusto que le diera para la próxima semana 2 o 3 opciones de agenda y hacerles la presentación final de cierre al Secretario Carlos Hernán Ocampo Ramírez y a Hernán para que también se conozcan, en donde sería invitada Sandra, que fue pieza clave.

2) Augusto Mendonca será encargado de coordinar con Evelin y Guilherme Diniz para la súper presentación que tenemos el día de hoy sobre Workshop. De la misma manera, Augusto nos comparte hoy, en el transcurso del día, la

quién fue el contacto que le dieron; en esa medida, tratarán de ajustarse a esas agendas.

3) Por un lado, la Dra. Sonia dice que tenían pendiente el tema del código, por lo tanto, no vieron "R" en el servidor. Es necesario saber que cuando ingresan en el servidor, ahí se encuentra una tabla, pero como les digo estamos más enfocados en los resultados y en generar un modelo de ejemplo y no tanto en las visualizaciones. Después del Workshop se podría revisar un poco, aunque no es prioritario por el momento.

4) Dra. Sonia continúa desarrollando el tema sobre R estudio, debemos realizar el script y él nos da unas visualizaciones, con las cuales nos damos cuenta si el modelo está funcionando o no, o si por lo contrario tenemos un tipo de correlación. Se debe recalcar que tener las visualizaciones es importante.

5) Antonio Cantillo, responde que ahora la visualización es buena, y la instalación de la librería que mencionó Evelin fue exitosa. Sin embargo, "ggplot" no se pudo instalar, es necesario señalar la importancia de esta herramienta porque en un futuro va ser de mucha ayuda, pues facilita el trabajo, y con lo básico que tenemos funciona. Desde el primer día que la probamos se resolvió el problema tanto con Oracle analiti como con ggplot, y esa es la parte que en realidad usamos para ejecutar el código. Como menciona la Dra. Sonia las visualizaciones son buenas para eso, no para ver los resultados. Entonces para suplir esa necesidad que teníamos, hasta ayer, lo que estamos realizando es exportar las tablas generadas y hacemos las visualizaciones, sin problema alguno.

6) Augusto, le comenta a la Dra. Sonia que la idea, por un lado, es tener un Workshop práctico, para mostrar que trabajar es más fácil. Por otro lado, el objetivo es evidenciar que si se trabajan con las dos herramientas se pueden obtener muchos logros.

7) La Dra. Sonia le manifiesta a Martha Carrasquilla, sus dudas y preocupaciones sobre la facilidad de adquirir la combinación de ORACLE y R en nuestro medio. Sonia, a su vez le dice a Antonio que la sugerencia es instalar esas dos recomendaciones que dio Evelyn con la imperiosa necesidad de llevarlas a cabo hoy. También indica que se debe mostrar esta presentación funcionando con el servidor al secretario con las dos alternativas, aunque nos faltan detalles, pero se debe subrayar la conveniencia de esta herramienta para todo el Departamento.

8) Luego Augusto expone que algunas cosas no van ser nuevas, pero nosotros vamos a intentar poner una dinámica distinta porque nosotros queremos cambiar la dinámica de Workshop no solamente mostrar los resultados, sino intentar evidenciar un trabajo real, lo más posible, debido a que este Workshop es muy práctico.

9) Enseguida Evelin va a compartir pantalla, lo primero que vamos a observar es el primer proyecto, o primer conjunto de datos con el que habíamos trabajado, anteriormente de su propiedad. Es interesante analizarlo por la cantidad de datos que tiene, es importante resaltar que por ejemplo hay más de 95.000 nit, para nosotros significa que son más de 95.000 mil empresas y entonces al analizar esa cantidad de datos estadísticamente, podemos ver con ustedes que esos nit pertenecían a 4 municipios , también observamos que existen diferentes actividades económicas, las cuales reporta cada empresa, y la mayor parte de los ingresos pertenecían al formato de la DIAN 110 , en ese sentido vimos en los gráficos que ese formato 110 era el responsable por el comportamiento global de todo el conjunto, incluso cuando contábamos cuatro DIAN, que serían cuatro formatos de la DIAN 110, 210, 220, 240. Se puede evidenciar que el comportamiento global estaba pasado en el DIAN 110, según la cantidad de datos, entonces, por ejemplo, aquí podemos observar que la actividad económica con mayor ingreso es la 6412, si hacemos una actividad global esta es la actividad que más ingresos en media genera y si analizamos solamente el 110 es esa misma, mientras que, si estudiamos los otros DIAN como el DIAN 210, actividad económica 90.

Entonces pudimos observar aparentemente que ellos tienen comportamientos diferentes, luego es interesante intentar siempre tener en cuenta que, a la hora de crear modelos, no es solamente cuestión de juntar cuatro DIAN y formar un solo modelo, eso puede servir para algunos tipos de datos, en especial si son datos bastante controlados, es decir que todos tienen los comportamientos similares u homogéneos. En este gráfico intentamos representar eso, ahora pregunto lo siguiente: ¿será que realmente en estos cuatro años, sin estudio tienen comportamientos similares? Aquí se evidencia que no. Por ejemplo, en el DIAN 110 la media ha ido aumentando, la media de los ingresos frutos y netos no ha cambiado, está igual o ha disminuido se ha mantenido estable, ya si vemos los totales brutos podemos decir que tienen comportamientos parecidos, pero el DIAN 110 siempre proporciona mayor cantidad de ingresos comparado con los otros. En esa medida si colocamos junto todo eso en un modelo para predecir la situación actual, probablemente él se va a pasar y va tomar como referencia el DIAN 110, y vimos también que éste es el porcentaje de aporte de ingresos en media de ingresos están concentrados en el DIAN 110 y los otros poco aportan.

10) Evelin continúa su presentación afirmando que este fue el primer ejercicio que hicimos, el objetivo de este estudio era tener una idea de cómo estaban distribuidos los DIAN, para después hacer MARCH con los datos de licores y rentas, lamentablemente no se pudo hacer eso porque fue un problema de datos, por esta razón no encontramos una manera de juntarlos, así que se estudiaron de manera individual .Cabe mencionar que tenemos un proyecto DIAN y un proyecto de licores que vamos a hacer más adelante, y uno de rentas.

Ahora, vamos a intentar crear un proyecto basado en los datos de licores: primero creamos el proyecto nuevo, tenemos la lista de las tablas que se han creado y que se han importado, pero en nuestro caso, vamos a utilizar esta tabla de licores e impuestos. Observaremos que nuestros datos están clasificados en dos tipos, pueden ser atributos, variables categóricas y los numéricos, por ejemplo, los años son categóricos. Por lo tanto si quiero visualizar por año, hemos observado que esta es una columna que viene con datos vacíos, una tabla simple que tiene datos es fondo de cuentas., Si quiero unas barras conocer cómo es la tendencia pueden ver, aunque yo no sé exactamente qué datos hay en fondo de cuenta, puedo tener una idea sobre lo que pasó con esos datos; Se puede constatar que ellos están disminuyendo, excepto por este 2017, algo debe haber pasado durante ese año que pudo haber cambiado, y eso podemos observarlo si intentamos buscarlo, es imprescindible que cuando se vaya a realizar un análisis, debemos tener mucho cuidado con estos datos porque puede que haya habido un error de datos o puede haber habido una empresa o conjunto de empresas que declararon mucho, o fueron penalizadas en ese tiempo, lo cual hace que el comportamiento global cambie y tendríamos que aplicar métodos específicos para resolver ese problema, uno de ellos seria coger la media o desconsiderar esa empresa o ese conjunto de empresas que pagaron más o menos, dependiendo de lo que represente el fondo de cuenta.

Con el objetivo de clarificar la Dra. Sonia pregunta si hay una gráfica que determine cuando haya muchos datos atípicos y si se llama grafica de bigotes, Es decir, esa gráfica es muestra si hay datos muy atípicos, o sea, que se salgan de la media, y pues, en un solo año hayan vendido más. Por ejemplo, en el 2016 se tuvo que vender en todas las empresas porque a partir del 2017 empezaba una ley, por ello compraron una cantidad de licor.

11) Dicho lo anterior, Evelin responde: este primer contacto que tenemos con los datos es para tomar acciones futuras, éstos gráficos nos ayudan a identificar, pero hay algunos gráficos más específicos que en R los puedes explotar; esos gráficos de los cuales está hablando Sonia los puedes hacer específicamente y allí puedes tomar acciones, para ilustrar , lo que se hizo en el otro ejemplo, sería que una empresa aparentemente pagó toda su deuda acumulada de varios años, debido a esto el ingreso se aumentó

mucho en un año, pero ese no es comportamiento global, no es un comportamiento patrón de varias empresa, es un comportamiento especifico, raro y no deberíamos considerarlo nuestro modelo. Por tanto, podríamos fácilmente eliminarlo y trabajar el restante de los datos, eso no significa que vamos a eliminar todos los datos de 2017, solo significa que identifiquemos dentro de ese año cuáles son los declarados, cuáles son los que tienen comportamiento extraño o diferente y vemos si aplica alguna.

Se prosigue el análisis. Hablando en fondo de cuenta, aquí hay una columna que se llama Ciudad Destino es la que nos puede brindar bastante información, nuevamente no sabemos qué ciudades llegan los licores, dónde importan o se declaran, pero ya tengo una idea, hay algunas ciudades a las que principalmente llegan, podemos ver que aquí en la columna hay varios errores, lo que debe haber pasado es que hubo un error en script.

En consonancia con lo dicho por Evelin, Augusto comenta que eso es común, pues los datos nunca están perfectos. Yo creo que nueve de cada 10 van a responder que tienen datos buenos. Entonces, es importante utilizar las herramientas para calificar los datos.

12) Evelin retoma, lo que hemos hecho es básicamente eso. Hemos creado un filtro con la Ciudad, en donde podemos observar que hemos eliminado una fecha en el nombre de la Ciudad, es decir, los desconsideramos, luego se escoge lo que son Ciudades, por ejemplo, las cinco Ciudades destino con mayor ingreso en rentas por impuestos de licores. Obteniendo los siguientes resultados: Yumbo es la ciudad que tiene más, seguida por Palmira y Buenaventura.

De acuerdo con lo expuesto anteriormente, Sonia aporta que estas son las ciudades que más impuestos pagan, y pregunta: ¿cómo se llama este programa?, porque éste no es R. A lo cual Evelin responde Oracle OAF.

13) Augusto se suma al diálogo, Sonia mira que muchas cosas podemos hacer acá con mucha simplicidad.

14) La Dra. Sonia responde: sí, eso estoy viendo, pero necesitamos tener código. Por favor, me pueden enviar por el chat el nombre de los programas.

Referente al tema que se discute, Evelin afirma lo que queríamos hacer aquí es mostrarles a ustedes que este proceso se puede hacer con código, no necesitamos de un programador para generar tipos de gráficos. Siguiendo las etapas del proyecto, en la primera etapa van a tener datos, pero como les sucede a ustedes, aún no tienen contratados los programadores, no obstante, ya tienen la capacidad de hacer un análisis

inicial de los datos para poder decir aquí hay 50 contribuyentes. Pero con 50 contribuyentes no puedo hacer nada, se debe hacer con una cantidad más grande.

Ahora veamos que en el 2019 Yumbo tuvo un mayor movimiento, La Unión también tuvo una mayor participación.

15) Enseguida Augusto relata: conseguimos hacer una medición muy similar a la relación de estas variables, a pesar que juntamos todo, cuando ponemos las ciudades no están similares, pues La Unión tenía una relación distinta de Yumbo.

16) Conforme a esto Evelin manifiesta: vamos a intentar generar otro tipo de gráfico, como los impuestos y el fondo de cuenta, porque por lo que vimos allí, hay una gran cantidad de impuestos, supongo que cada código representa un uno diferente, entonces, en este grafico lo que queremos ver es cuales de estos impuestos son los que tienen mayor fondo de cuenta.
Aquí podemos observar que estos cuatro tipos de impuestos iniciales equivalen a estos puntos, muy probablemente, el comportamiento global está centralizado en estos que figuran al comienzo. Si, por ejemplo, queremos hacer un análisis solo de un impuesto que está aquí en el nivel más profundo, hay que hacer probablemente análisis más específicos, solo considerando esos datos, porque si hacemos un estudio global va a estar predominado por esto.

Podemos ver, en este caso que un análisis interesante seria involucrar unos de estos cuatro primeros, que parecen tener una relevancia mayor y aportan más o menos proporcionalmente. Entonces, se podría hacer una clasificación con estos cinco, los cuales cada uno van a aportar un porcentaje similar y no vamos a crear una clasificación desbalanceada, los amarillos, los rojos, los verdes y azules, sería como intentar balancear nuestros algoritmos de acuerdo a esto. Vamos a guardarlo para que ustedes lo tengan, licores Workshop, así se va llamar y ustedes cuando accedan ya van a poder continuar haciendo sus análisis con esos datos.

17) Después la Dra. Sonia interroga a Evelin por el clúster que hizo: ¿uno, dos y tres son representaciones de nuestros equipos de fondo cuenta? ¿qué variables cogiste allí? No me quedó claro.

18) Posteriormente Evelin responde, que cogió los impuestos, por ejemplo, este aquí 76P02002 es un Impuesto; este de aquí 76P00134 es otro impuesto; estos son todos los impuestos que ustedes tiene en esa tabla, el tamaño del cuadrado es de impuesto y el tipo de cuenta, puedes ver que estos 5 de aquí generan un fondo de cuenta parecido entre ellos, estos naranjas de aquí tiene un fondo de cuenta parecido; si te pones a analizar es que no sabemos qué impuestos son esos, pero es, por ejemplo, P79, 79P02843 es

parecido lo que genera al 76P688, lo cual significa que estos de aquí tienen comportamientos parecidos, todos los que son amarillos son comportamientos similares. La idea es que cuando hagas tus estudios utilices este tipo de visualizaciones para que te guíes en la decisión futura que vas a tomar, para este caso no voy a mezclar estos de aquí con estos pequeños porque generan mil y generan billones.

19) En ese sentido, la Dra. Sonia pide dos tips sobre la normalización o la estandarización que se tiene que hacer para los análisis, pues hay ocasiones que se tienen como porcentajes y hay otra variable con miles de millones de pesos, ¿cómo es tu criterio? Tips de experiencia en ORACLE.

20) Por consiguiente, Evelin responde, cuando tienes cantidades monetarias y quieres trabajar con las mismas se debe predecir cuál es el futuro, la declaración que va a generar el próximo año, si tu estándar de las opciones es mover el rango entre cero y uno, lo que va pasar es que, si tienes una empresa que declara un billón y una empresa que declara cien en valores normales van hacer cien y un billón, si las normalizas van hacer 0.01.01001 y la otra va ser 0.99999 muy cerca de uno. Entonces si tu normalizando estos datos esa proporción se va mantener, solo que tú quieres saber el resultado, si tú quieres que tu resultado sea un monto y si normalizas tu target, tu resultado va ser entre un valor entre cero y uno, en esa medida nosotros queremos un valor real. Cada una de las columnas se tiene que trabajar independientemente, es interesante si se hace normalización dependiendo del modelo que vas a utilizar, por ejemplo, si vas a utilizar un SVM el cual trabaja con valores numéricos y la varianza es muy importante para ese modelo, por esa razón es muy importante normalizar o estandarizar porque si no van a ver variables que pesan muchísimo, pues son muy relevantes y tu clasificación te va a dar un clúster con un único contribuyente que aporta dos millones y un segundo clúster que va a hacer todos los otros puntos. Es necesario normalizar, eso va de la mano con la elección del algoritmo que vas a usar, si no es un SVM y es un árbol de decisión, el árbol de decisión trabaja independientemente con sus variables y él hace un análisis con la ganancia y la homogeneidad de sus datos, entonces no es tan importante hacer una normalización, pero siempre me parece que el análisis de los datos siempre debe pasar por una estandarización.

A veces no necesita ser tan estricto para colocarlos entre cero y uno, específicamente tu media puede variar y tu variante también puede variar un poco más, eso tiene significado al interpretar tus datos, así que no necesitas forzarlo entre cero y uno, no es necesario.

21) Por otro lado, Evelin manifiesta que se va a empezar a visualizar otro aspecto de su presentación, vamos a empezar con lo bonito, comenzaremos desde los datos, nosotros tenemos las tablas que fueron

creadas de DIAN, de rentas, licores, el objetivo de este ejercicio es que podamos utilizar los datos históricos para entrenar un modelo, y a su vez ese modelo nos ayude a predecir un año que no está en la base datos. Tenemos aquí datos desde 2014, 2015, 2016, 2017, 2018, 2019, sin embargo, para 2020 no tenemos. En nuestro mundo ideal, si no hubiera habido pandemia la declaración teóricamente debería seguir el patrón de los años anteriores claramente no va ser así, porque este año debe haber habido facilidades de pago, aumentado el límite de tiempo de pago y todo lo demás. Bueno, como es un ejercicio vamos a intentar predecir para el 2020 cuáles serían las declaraciones, para eso lo que hemos hecho es preparar resultados, he creado una lista allí que se llama tendencia que va a contener tres años de historia, voy a entrenar con el año 2019 y los tres años anteriores al 2019 me van a servir como referencia para la predicción, estas van a ser mis variables 2016, 2017, 2018 y mi target va a ser 2019, y con eso creo una vista que se llama rentas tendencia, voy a crear una vista con mi conjunto de test porque estoy entrenando el modelo, pero quiero ver si puedo predecir si es la etapa del test, una etapa de la aplicación o la ejecución del modelo.. Entonces, para hacer ello voy a coger otros datos, cuando yo ejecuto el modelo la idea es predecir no tengo un target, no sé cuál va a ser la contribución del año 2020, no tengo esa última variable; como pueden ver en esos tres no los tengo, pero como referencia para predecir ese target voy a utilizar los tres años anteriores el 2019, 2018 2017, de manera análoga a lo que hicimos o sea, para predecir lo que pasó 2019 pueden observar qué pasó hace un año, qué pasó hace dos años, qué pasó hace tres años, esto al momento de ejecutar y al momento de entrenar por datos que si existen y son el 2019. Estoy observando igual qué pasó hace un año, en el2019, era el 2018 y pasó esto, eso fue lo que declaraste; hace dos años 2019 o sea el 2017 me declaraste esto; hace tres años 2019 o sea, 2016 me declaraste esto. Eso es lo que hacemos, participar a la historia en el proceso de entrenamiento por qué se hace esto, porque si agrupas tus datos por año puedes observar la tendencia, por ejemplo, si en el 2016 esta persona pagaba poco en 2017 aumentaba un poco más, en el 2018 aumentó más, eso muestra una tendencia clara que está en aumento su predicción, y esa persona probablemente este año, en el 2020 va continuar aumentando o de lo contrario puede haber tendencias ascendentes, puede haber tendencias descendentes de empresas que están yendo a falencia, o que cambiaron alguna política, cada empresa individualmente tiene su comportamiento, pero recordemos que cada tipo de estudio son estadísticos y por tanto toman en cuenta el comportamiento global de todos los datos que estamos observando.

22) A continuación vamos a ver los datos de esta tabla de rentas, se hizo un ejercicio con Adalberto y Sonia, fuimos a ver los datos y observamos que, dentro de ellos, en el año 2016 hay un pico aquí extrañísimo, parece que las declaraciones están siguiendo un patrón, un comportamiento, pues todas están creciendo un poquito aquí, disminuyó ésta, pero más o menos

el comportamiento es parecido, aunque aquí hay un pico enorme, entonces, fuimos atrás de donde salió ese pico, esto es solamente los meses en los cuales más se declaran como en diciembre. Les comparto para observen el 2016 y vean qué pasó con este pico, hubo un NIT, que declaró 26 billones él solito hizo ese pico, entonces, podemos asumir que tuvo un comportamiento excepcional, pero es solo un NIT, como ven, no está siguiendo el patrón de todo el resto, lo sacamos en nuestro estudio porque no aporta al análisis global.

Por eso estoy haciendo este filtro de este NIT, lo estoy sacando de donde lo encontré porque él tiene un comportamiento extraño, éste segundo tiene nulos y hay otro NIT que está vacío sin nombre, en total son tres casos. Luego creamos nuestro "DATA" y con esto ya tenemos nuestro entrenamiento y nuestro test.

Ahora, si ya venimos a "R" y aquí en "R" lo que vamos hacer es descargar la librería que nos ayuda a conectarnos con "Hi" y con "Spark", que es nuestro recurso de procesamiento, aquí podemos ver que ya tenemos acceso a la lista de todas las tablas, las cuales están en HI tenemos todas las que están aquí y están son las dos que vamos a utilizar en este ejercicio, es decir, las dudas que han creado y las van a utilizar, las traemos a nuestro contexto. Una vez más recuerden que al efectuar el hecho de aquí no significa que me estoy trayendo todos los datos a memoria, sino que es una referencia, un comando, cargo todos mis datos ahora que tengo esas referencias, mi intención ponerlo a disposición de mi procesamiento, colocarlo en contexto de "SPARK", utilizando "HDP" porque "SPARK" solo conoce "HDP" y necesita que los datos que él va a procesar estén distribuidos, ponemos "TRIM" como "DATA" y el test como "DATA" test. Todo lo que "SPARK" conoce está en "HDP", no está en memoria.

En "OAS" también vamos hacer lo mismo que podemos hacer en "R", en realidad, el procedimiento normal sería ver los datos en "OAS", ver que existen datos que son exorbitantemente grandes y tienen comportamientos únicos, especiales, pero hay un grupo de datos que tienen un comportamiento parecido y los podemos aprovechar, probamos algunos algoritmos MACHIN y OAS tienen para después implementarlos, ya luego en nuestro clúster en R, ya hemos identificado nuestro status, los limites inferior y superior, hasta dónde vamos a considerar nuestros datos y ahora vamos a implementar un algoritmo de una regresión lineal.

Bueno, vamos a crear un modelo de regresión lineal, vamos a utilizar como Datos nuestro conjunto de entrenamiento, lo que acabamos de hacer es darles clip a nuestros datos, no podemos colocarlos todos en el orden que están, es decir, con todos los datos fuera del patrón general, que hace un año tiene que tener valores como mínimo uno y como máximo 84 millones. Ya tenemos nuestros datos delimitados y vamos a crear un algoritmo de

entrenamiento, en nuestro caso queremos que nos prediga un valor numérico.

23) Después de la explicación la Dra. Sonia dice que sería súper importante, si muestran esta presentación al Secretario de Hacienda y a la Directora de Rentas, seria genial ver todas las visualizaciones. La invitación sería hacer estas diapositivas con esta información. Igualmente afirma que deben hacer un plan de acción para que las empresas a quienes nosotros pagamos esa base de datos tengan las variables que nosotros necesitamos para comenzar a hacer los estudios. Tenemos que hablarles a ellos, pues supongo que ignoran cuál es el Municipio que más paga impuestos de licores de las rentas. La tabla de DIAN es a nivel del Departamento.

24) Posteriormente, Evelin termina su exposición.

Para terminar Augusto dice que espera que le haya gustado y que siempre van a estar a su disposición.

25) Tomando la palabra Martha le pide a Augusto que para la próxima semana les comparta un par de opciones de agenda, igualmente le manifiesta a Sonia que va a contactarla para confirmar a las personas que deberían asistir a la ya anteriormente mencionada reunión. En la medida de lo posible seguiremos apoyándolos si aparecen dudas. Hay una disposición total y un compromiso para seguirlos colaborándolos en este proceso.

26) Agradecemos a ustedes por toda la disposición, el compromiso, la pasión, la entrega como dice Evelin, están empezando y seguramente van a lograr grandes cosas y esperemos que sea de la mano de "ORACLE". El grupo quedará abierto para consultas y nos seguiremos hablando.

27) Por último, la Dr. Sonia pide que por favor les regalen la presentación de Análisis de Datos para socializar a Rentas lo que vieron hoy y los hallazgos. Igualmente informar lo que se va hacer y en la reunión ustedes toman la palabra. Yo quiero que ellos vean que poseemos una herramienta muy poderosa, que podemos levantar en laboratorios de inteligencia artificial con proyecciones infinitas. La Dr. Sonia se despide agradeciendo, por acá siempre estamos a la orden, un abrazo enorme.

Finalmente, Cristian Petro, se despide y agradece enormemente el apoyo que han dado, han recibido más de lo que esperaban.

Atentamente,

estas diapositivas con esta información. Igualmente afirma que deben hacer un plan de acción para que las empresas a quienes nosotros pagamos esa base de datos tengan las variables que nosotros necesitamos para comenzar a hacer los estudios. Tenemos que hablarles a ellos, pues supongo que ignoran cuál es el Municipio que más paga impuestos de licores de las rentas. La tabla de DIAN es a nivel del Departamento.

24) Posteriormente, Evelin termina su exposición.

Para terminar Augusto dice que espera que le haya gustado y que siempre van a estar a su disposición.

25) Tomando la palabra Martha le pide a Augusto que para la próxima semana les comparta un par de opciones de agenda, igualmente le manifiesta a Sonia que va a contactarla para confirmar a las personas que deberían asistir a la ya anteriormente mencionada reunión. En la medida de lo posible seguiremos apoyándolos si aparecen dudas. Hay una disposición total y un compromiso para seguirlos colaborándolos en este proceso.

26) Agradecemos a ustedes por toda la disposición, el compromiso, la pasión, la entrega como dice Evelin, están empezando y seguramente van a lograr grandes cosas y esperemos que sea de la mano de "ORACLE". El grupo quedará abierto para consultas y nos seguiremos hablando.

27) Por último, la Dr. Sonia pide que por favor les regalen la presentación de Análisis de Datos para socializar a Rentas lo que vieron hoy y los hallazgos. Igualmente informar lo que se va hacer y en la reunión ustedes toman la palabra. Yo quiero que ellos vean que poseemos una herramienta muy poderosa, que podemos levantar en laboratorios de inteligencia artificial con proyecciones infinitas. La Dr. Sonia se despide agradeciendo, por acá siempre estamos a la orden, un abrazo enorme.

Finalmente, Cristian Petro, se despide y agradece enormemente el apoyo que han dado, han recibido más de lo que esperaban.

Atentamente,

SONIA YAMILETH CASTRO	CRISTIAN PETRO	LILIANA PLAZAS
Asesora – Gobernadora	Líder Hardware	Líder Economía Digital

(se anexan imágenes)

Transcriptor: Mariela Ivonne Sinisterra Muñoz- Técnico

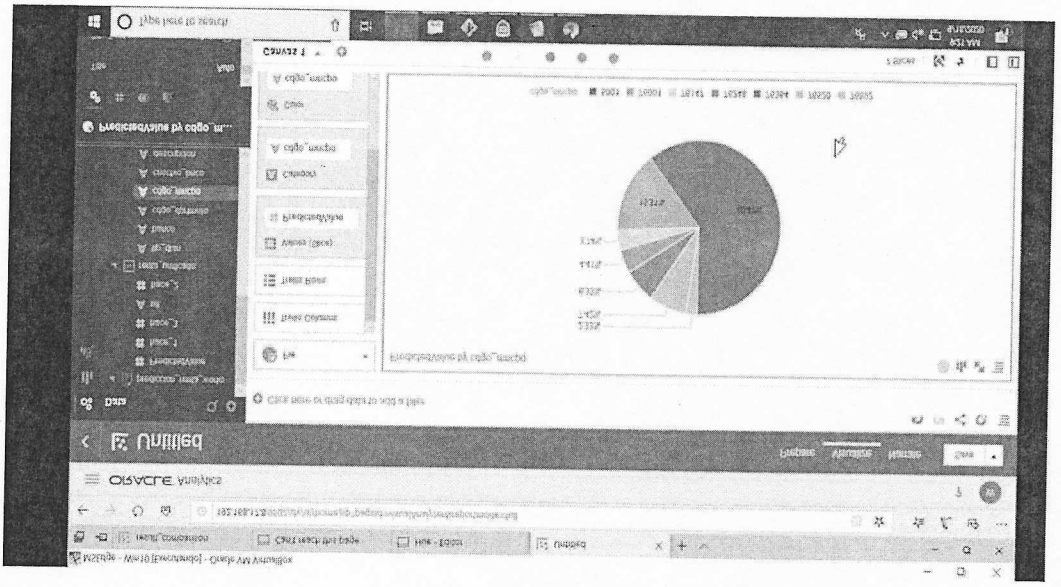
Archivarse en: Carpeta Actas de Reunión de Oficina

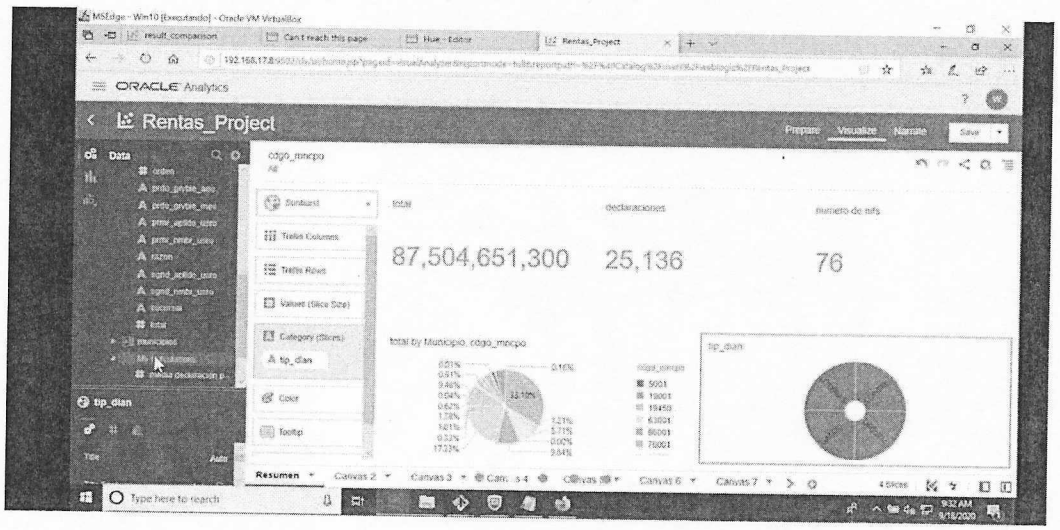
SONIA YAMILETH CASTRO YAMA
Asesora – Gobernadora

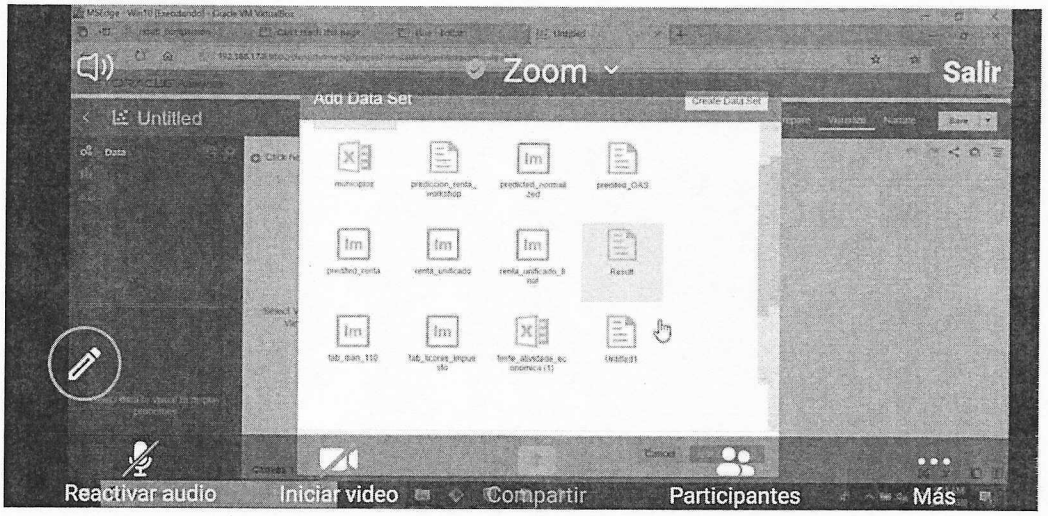
CRISTIAN PETRO
Líder Hardware

(se anexan imágenes)

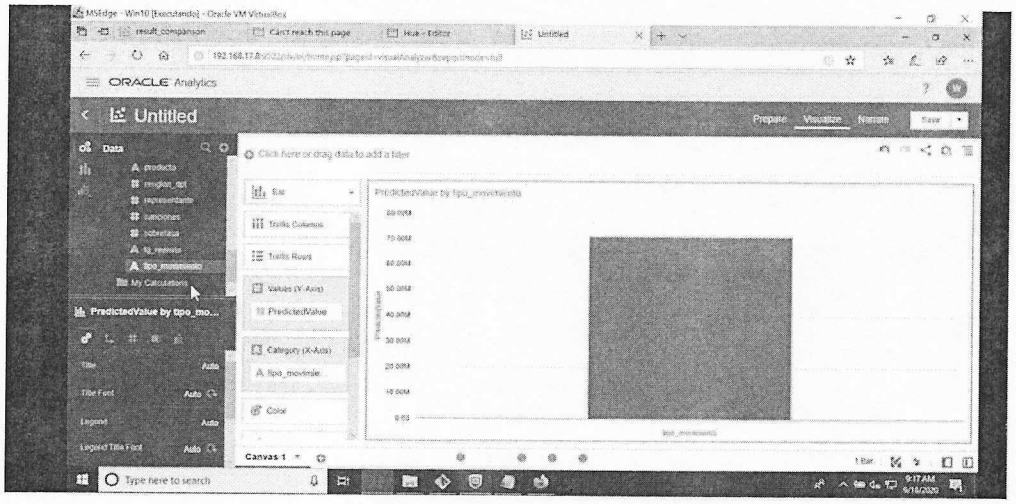
Transcriptor: Mariela Ivonne Sinisterra Muñoz- Técnico

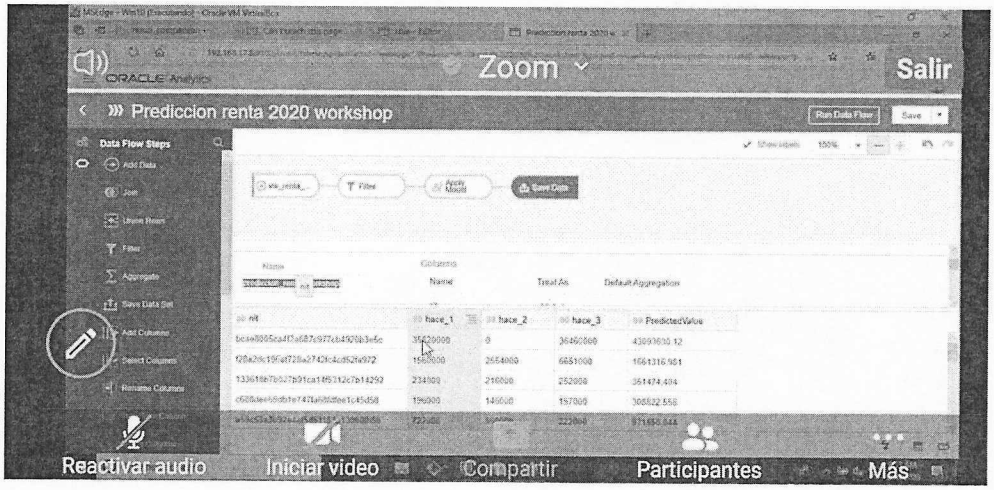


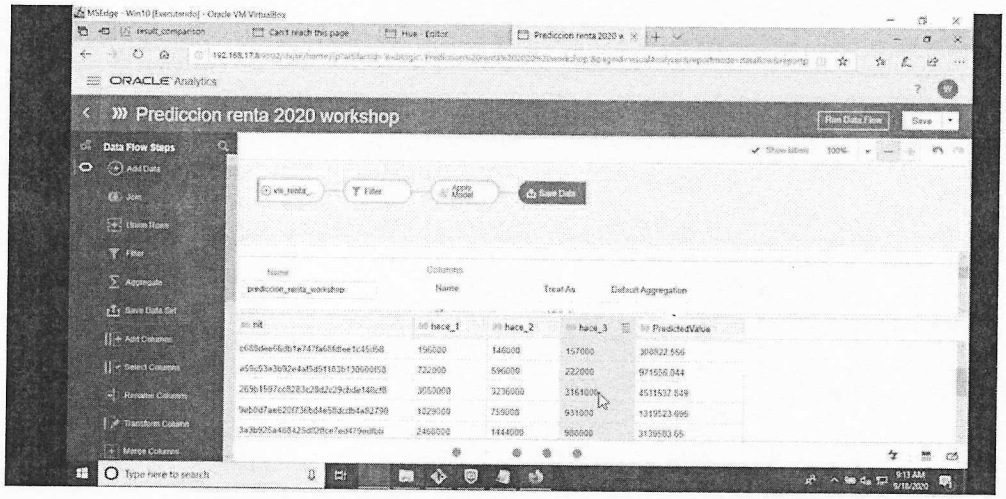

ORACLE Analytics
Rentas_Project
Prepare
Visualize
Narrate
Save
Data
87,504,651,300
25,136
76
Resumen
Canvas 2
Canvas 3
Canvas 6
Canvas 7

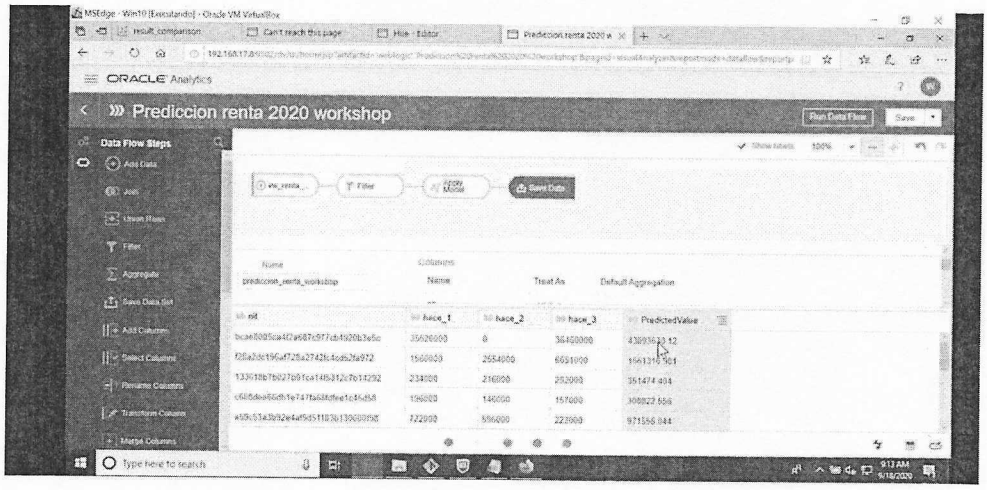

Zoom
Salir
Untitled
Add Data Set
Reactivar audio
Iniciar video
Compartir
Participantes
Más

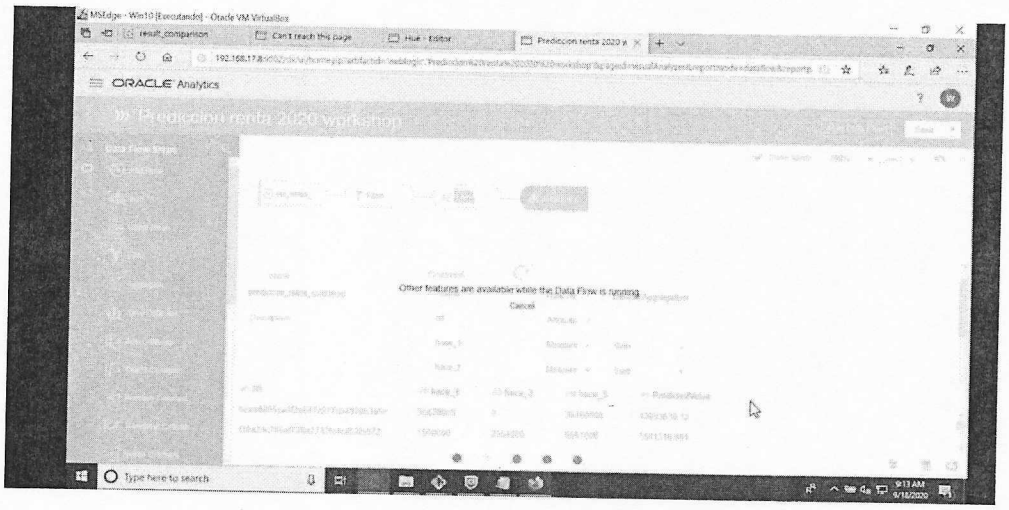

ORACLE Analytics
Untitled
Add Data Set
Data Diagram

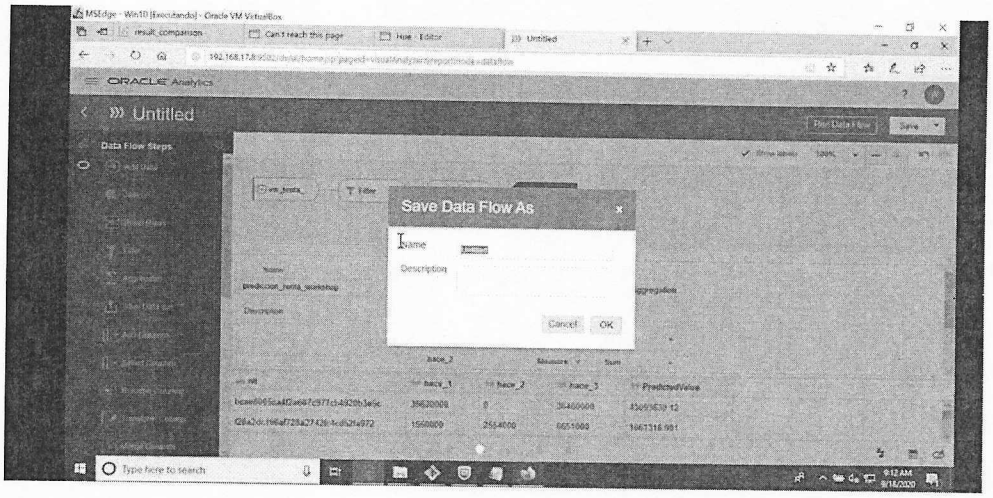

ORACLE Analytics
Untitled
Prepare
Visualize
Narrate
Save
Data
Canvas 1
Type here to search

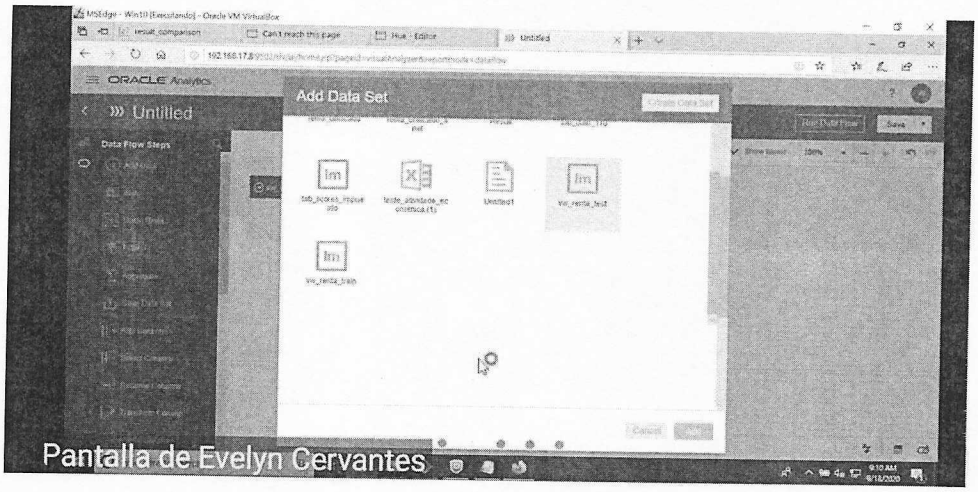

Zoom
Salir
ORACLE Analytics
Prediccion renta 2020 workshop
Run Data Flow
Save
Data Flow Steps
Reactivar audio
Iniciar video
Compartir
Participantes
Más

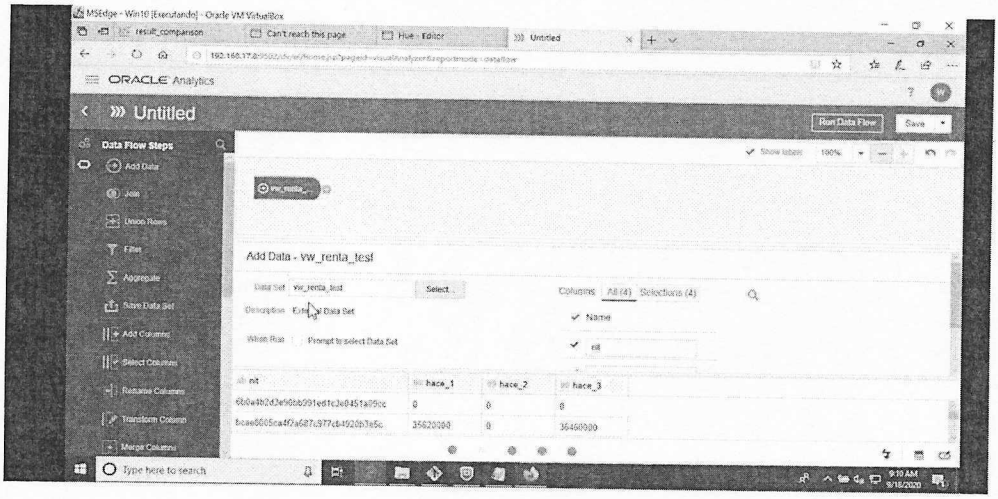

ORACLE Analytics
Prediccion renta 2020 workshop
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
Save Data
Name
Columns
prediccion_renta_workshop
Name
Treat As
Default Aggregation
rid
hace_1
hace_2
hace_3
PredictedValue
196000
146000
157000
722000
596000
222000
3050000
3236000
2161000
1029000
759000
931000
2460000
1444000
980000
Type here to search
9:13 AM
9/18/2020

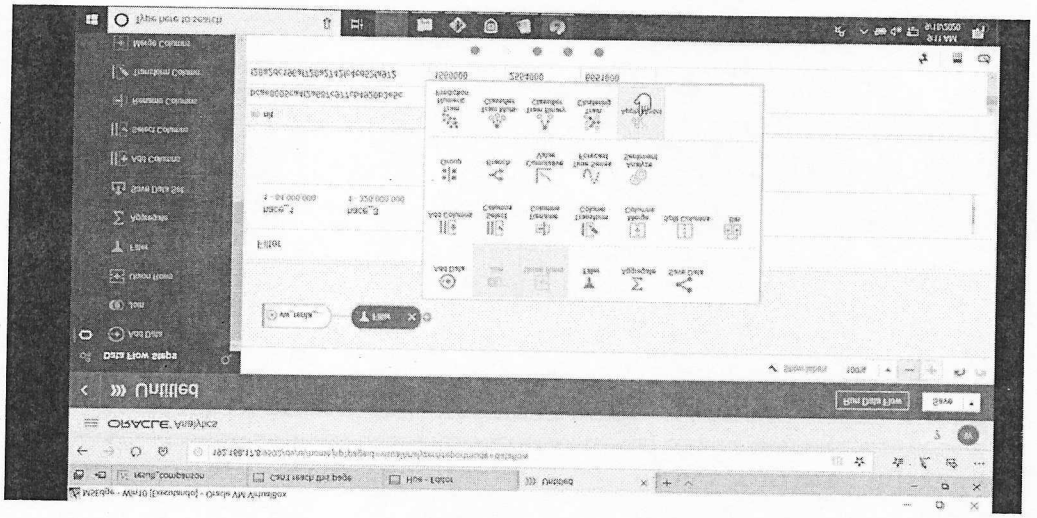

ORACLE Analytics
Prediccion renta 2020 workshop
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
Save Data
Name
Columns
prediccion_renta_workshop
Name
Treat As
Default Aggregation
rid
hace_1
hace_2
hace_3
PredictedValue
196000
146000
157000
722000
596000
222000
Type here to search
9:13 AM
9/18/2020

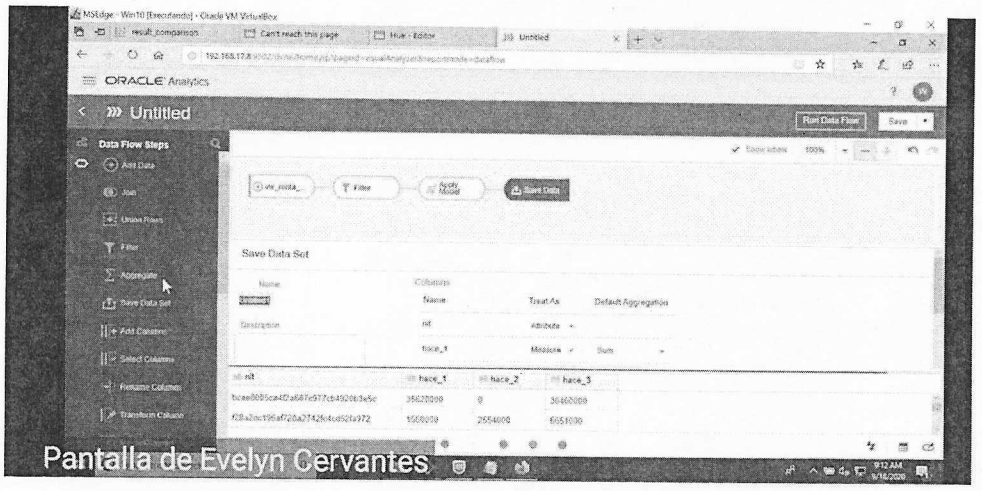

ORACLE Analytics
Other features are available while the Data Flow is running.
Cancel

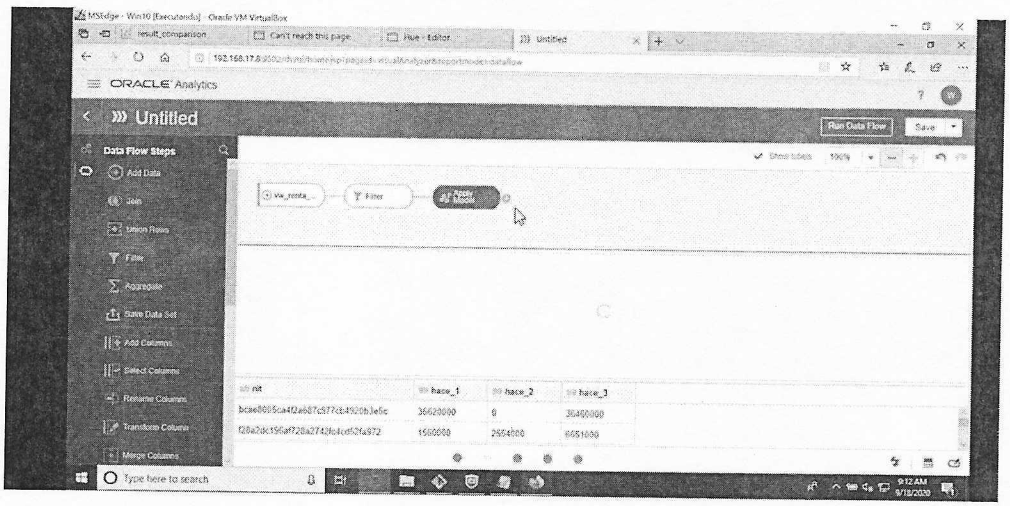

ORACLE Analytics
Untitled
Data Flow Steps
Save Data Flow As
Name
Description
Cancel
OK

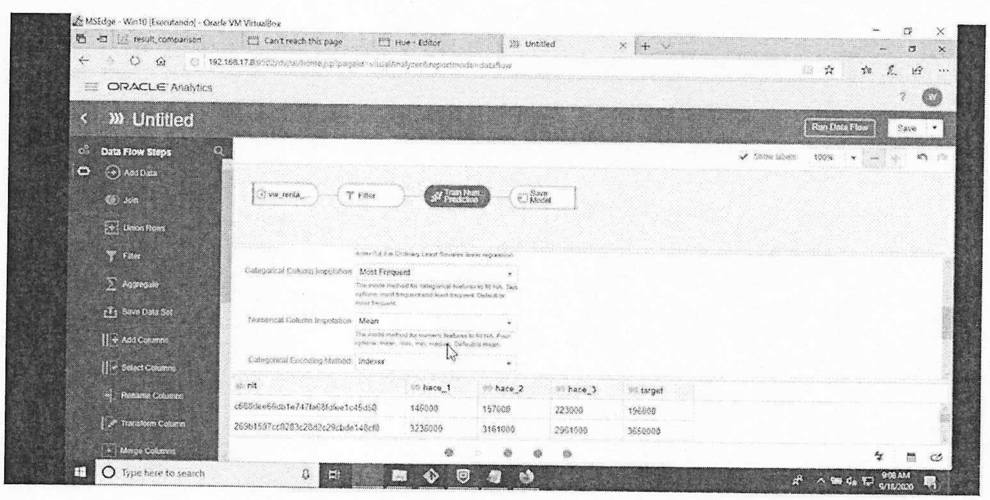

ORACLE Analytics
Untitled
Add Data Set
Pantalla de Evelyn Cervantes

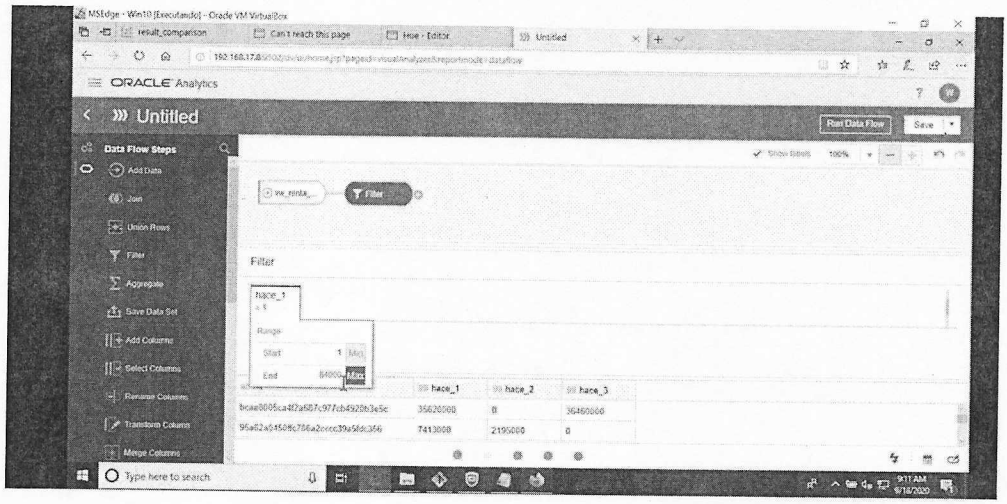

ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data - vw_renta_test

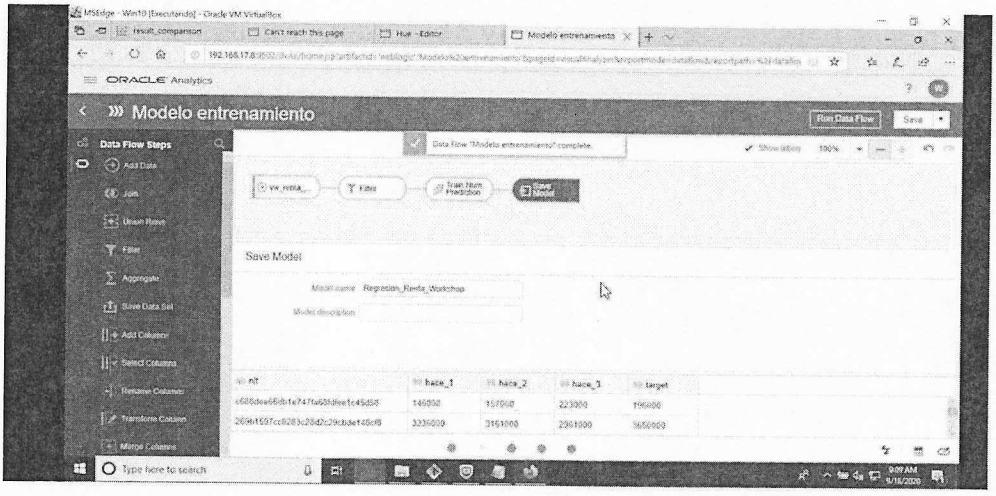


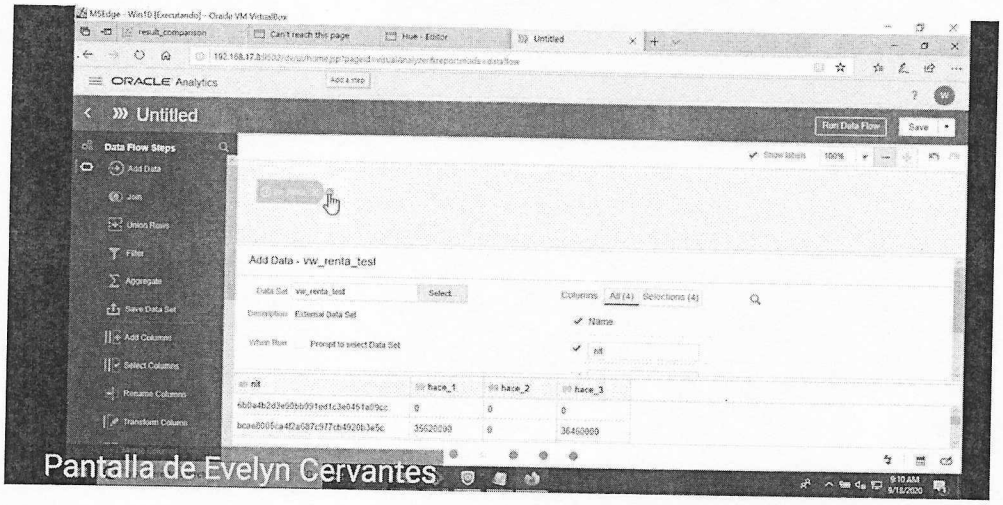

ORACLE Analytics
Untitled
Data Flow Steps
Save Data Set
Pantalla de Evelyn Cervantes

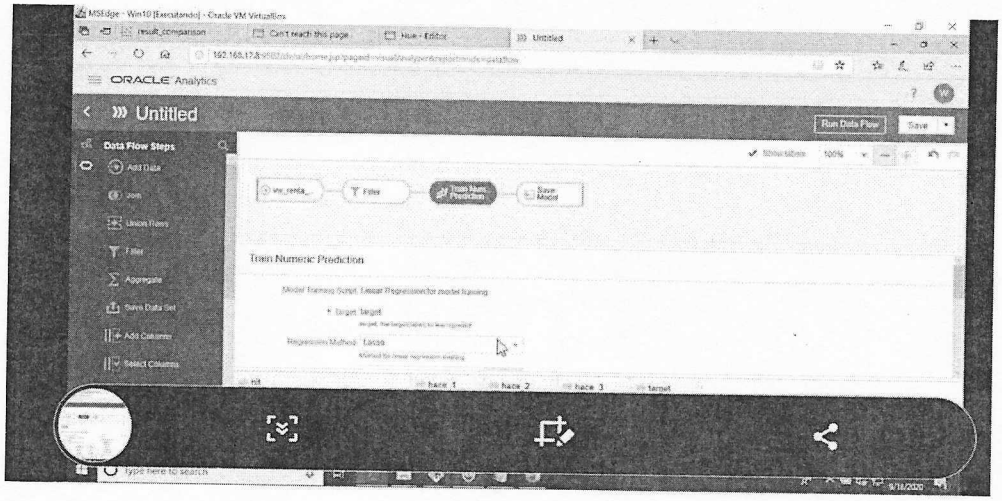

MSEdge - Win10 [Executando] - Oracle VM VirtualBox
result_comparison
Can't reach this page
Hue - Editor
Untitled
ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
Apply Model
nit
hace_1
hace_2
hace_3
Type here to search

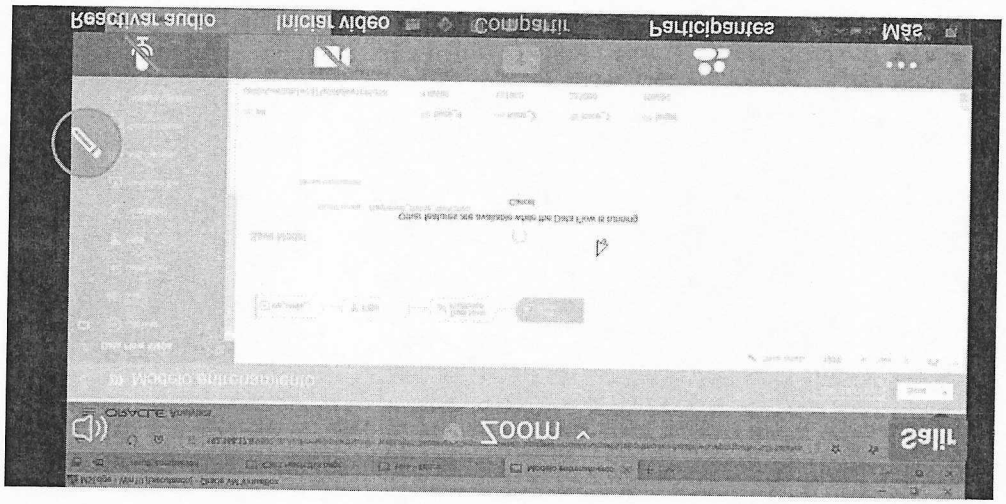

MSEdge - Win10 [Executando] - Oracle VM VirtualBox
result_comparison
Can't reach this page
Hue - Editor
Untitled
ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
Filter
Train Numeric Prediction
Save Model
Categorical Column Imputation
Most Frequent
Numerical Column Imputation
Mean
Categorical Encoding Method
Indexer
nit
hace_1
hace_2
hace_3
target
Type here to search

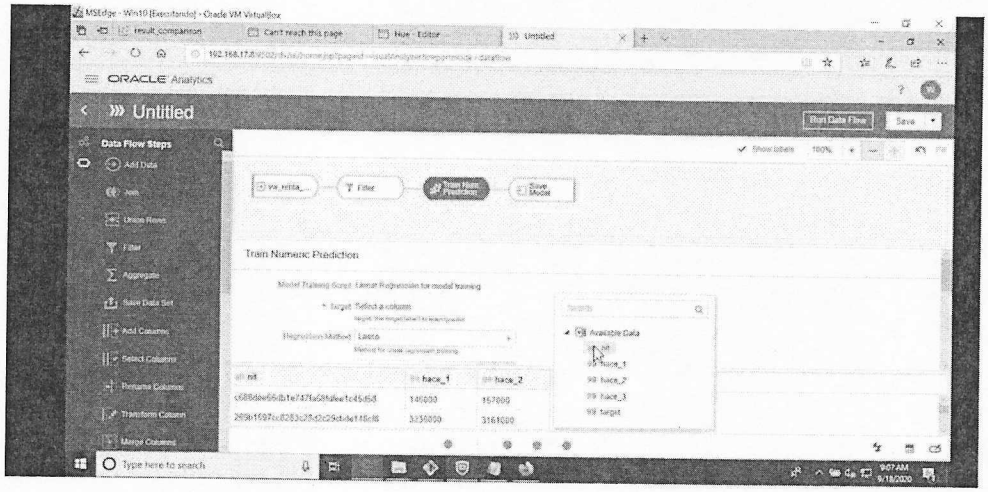

ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
vw_renta_
Filter
Filter
hace_1
hace_2
hace_3
Type here to search

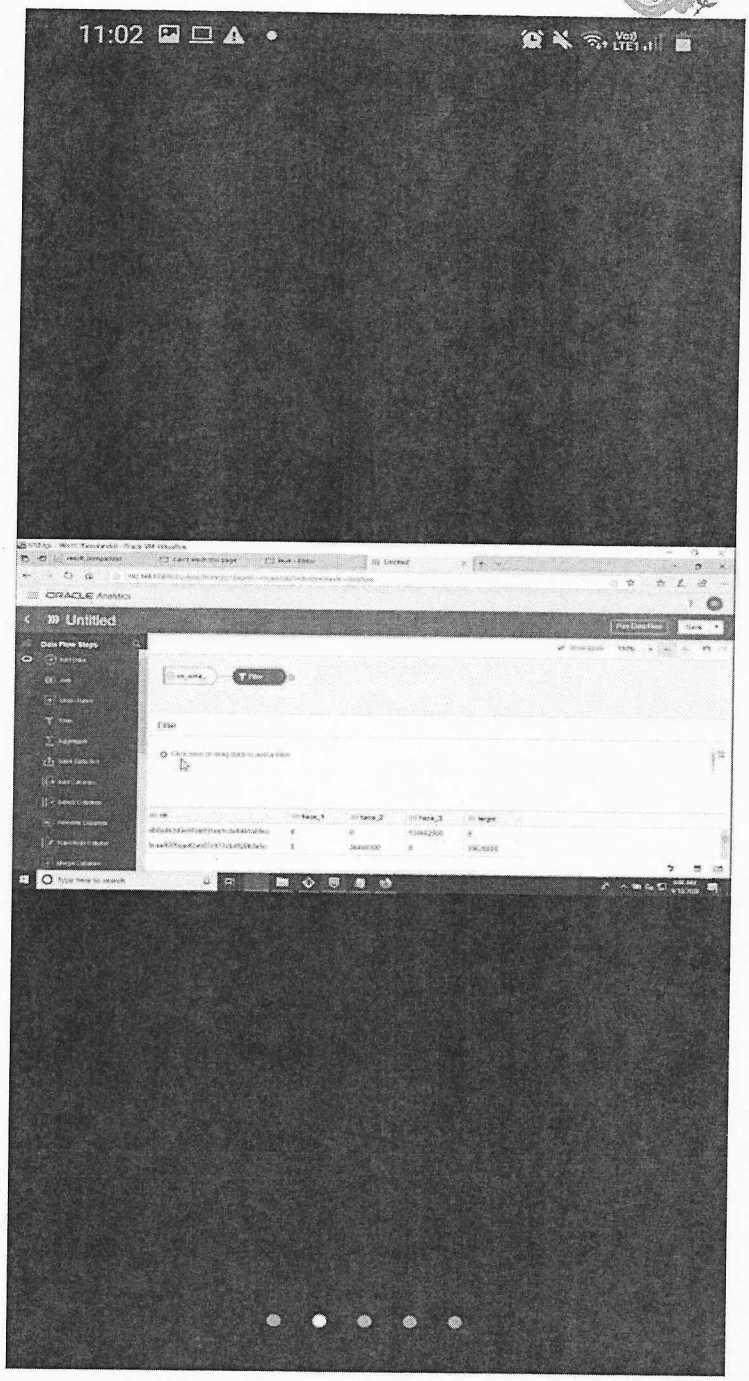

ORACLE Analytics
Modelo entrenamiento
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Column
Merge Columns
vw_renta_
Filter
Train Num Prediction
Save Model
Save Model
Model name
Regresion_Renta_Workshop
Model description
nit
hace_1
hace_2
hace_3
target
Type here to search

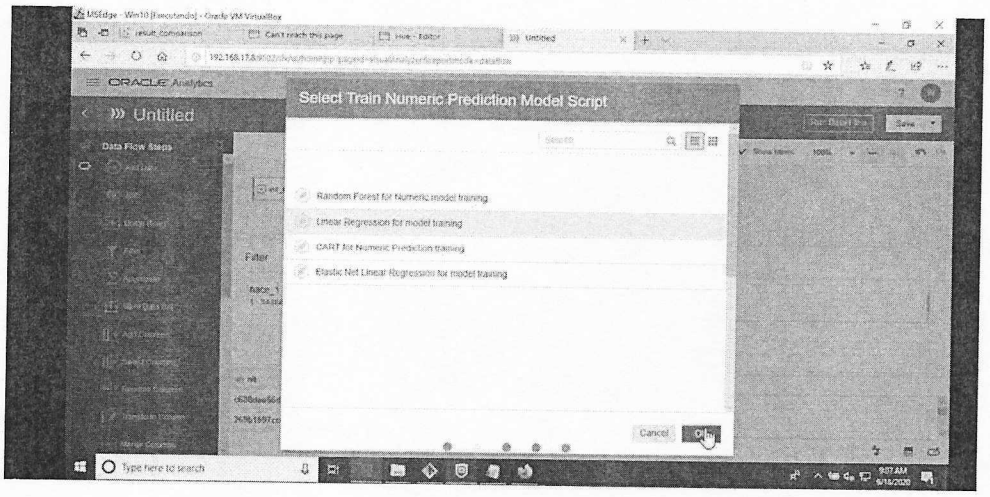

ORACLE Analytics
Untitled
Data Flow Steps
Add Data - vw_renta_test
Pantalla de Evelyn Cervantes

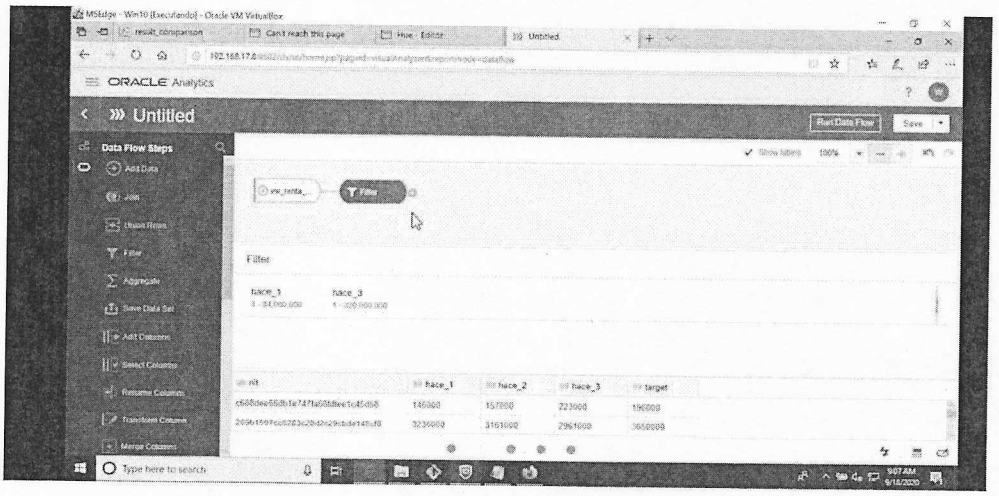

ORACLE Analytics
Untitled
Data Flow Steps
Train Numeric Prediction

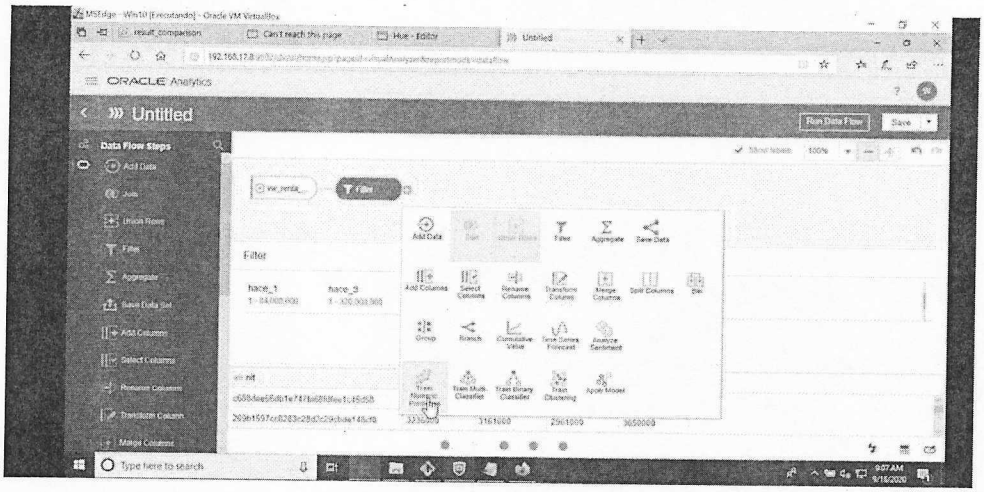

Zoom

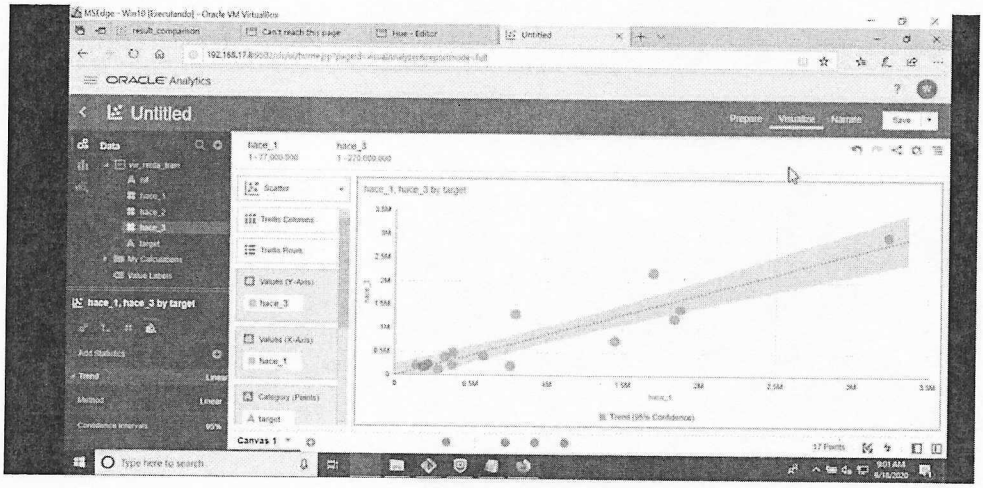

ORACLE Analytics
Untitled
Train Numeric Prediction


11:02
ORACLE Analytics
Untitled


ORACLE Analytics
Untitled
Select Train Numeric Prediction Model Script
Random Forest for Numeric model training
Linear Regression for model training
CART for Numeric Prediction training
Elastic Net Linear Regression for model training
Cancel
OK
Data Flow Steps
Filter


ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Columns
Select Columns
Rename Columns
Transform Columns
Merge Columns
Filter
face_1
face_2
face_3
target


ORACLE Analytics
Untitled
Run Data Flow
Save
Data Flow Steps
Add Data
Join
Union Rows
Filter
Aggregate
Save Data Set
Add Column
Select Columns
Rename Column
Transform Column
Merge Column
Group
Branch
Cumulative Value
Time Series Forecast
Analyze Sentiment
Train Numeric Regression
Train Multi Classifier
Train Binary Classifier
Train Clustering
Apply Model
Filter
Type here to search


ORACLE Analytics
Untitled
Prepare
Visualize
Narrate
Save
Data
Scatter
hace_1, hace_3 by target
Canvas 1
Type here to search

GOBERNACIÓN VALLE DEL CAUCA	REGISTRO DE ASISTENCIA						Código: FO-M8-P1-14 Versión: 01 Fecha de Aprobación: 15/08/2018 Página: 1 de
LUGAR:	REUNIÓN VIRTUAL						TOTAL HORAS: 1 hora
No. DE ACTA: 126A	HORA DE INICIO: 9:00am		HORA DE TERMINACIÓN: 10:00am				CODIGO DEL PROCESO/SUBPROCESO:
FACILITADOR (ES)/RESPONSABLE: SONIA YAMILETH CASTRO YAMA							
NOMBRE DEL EVENTO/TEMA DE REUNIÓN/TEMAS A TRATAR	REUNIÓN ORACLE ANÁLISIS DE DATOS SOCIALIZAR A RENTAS LO QUE SE HIZO						FECHA: 18 09 2020

No.	DEPENDENCIA / ENTIDAD/MPIO	NOMBRES Y APELLIDOS COMPLETOS	CARGO	CÉDULA	No. DE CELULAR - TEL/EXT	CORREO ELECTRÓNICO	FIRMA DE ASISTENCIA
1	Secretaria de las Tenologías de la Información y las comunicaciones	Sonia Yamileth Castro Yama	Asesora - Gobernadora			scastro@valledelcauca.gov.co	Asistío virtual
2	Secretaria de las Tenologías de la Información y las comunicaciones	Martha Carrasquilla					Asistío virtual
3	Secretaria de las Tenologías de la Información y las comunicaciones	Augusto Mendonca					Asistío virtual
4	Secretaria de las Tenologías de la Información y las comunicaciones	Antonio Cantillo					Asistío virtual
5	Secretaria de las Tenologías de la Información y las comunicaciones	Guilherme Diniz					Asistío virtual
7	Secretaria de las Tenologías de la Información y las comunicaciones	Adalberto Martínez					Asistío virtual
8	Secretaria de las Tenologías de la Información y las comunicaciones	Cristian José Petro Petro				cjpetro@valledelcauca.gov.co	Asistío virtual
9	Secretaria de las Tenologías de la Información y las comunicaciones	Mariela Ivonne Sinisterra Muñoz	Técnico	67017814		maisinisterra@valledelcauca.gov.co	Asistío virtual

Valle Invencible | GOBERNACIÓN VALLE DEL CAUCA
Secretaría de las Tecnologías de la Información y las Comunicaciones

1.360.2-79

ACTA DE REUNION INNOVACIÓN PÚBLICA DIGITAL CIENCIA DE DATOS DIVULGACIÓN DE GENERACIÓN DE CONOCIMIENTO APARTIR DE LA APLICACIÓN DEL PROCEDIMIENTO M2-P6-04, SECRETARIA DE LAS TIC

ACTA No. 130 B

FECHA: Santiago de Cali, 24 de septiembre de 2020

HORA: 10:30 am a 1:00 pm

ASUNTO: ACTA SOCIALIZACIÓN RETO 1 Y LA DEMOSTRACIÓN DEL SERVIDOR DEL BIG DATA DE ORACLE

LUGAR: Secretaría de las Tics

ASISTENTES: Dr. Carlos Hernán Ocampo, Liliana Plaza, Zoraida Bravo, Elen Norelia Balanta, Liliana Rodríguez, Marcelo Salgado, Laura Lorena Gaitán, Dra. Sonia Castro, Cristian Petro.

OBJETIVO: SOCIALIZAR LOS PROYECTOS DE BIG DATA, ANALITICA DEL RETO 1 Y LA CLAUSURA DE LA CAPACITACIÓN Y ASESORAMIENTO DE ORACLE EN EL SERVICIOS DE BIG DATA.

DESARROLLO:

En esta reunión se mostró la importancia y la trascendencia de proyectos basados en analítica en la Gobernación del Valle y además del tema en comento se socializaron los hallazgos que se hicieron con las bases de datos de las cámaras de Comercio y con siete rentas Departamentales, la deguello, licores, contribución al deporte, Cervezas, Cigarrillos y Lotería

Atentamente,

CARLOS HERNÁN OCAMPO RAMÍREZ
Secretario

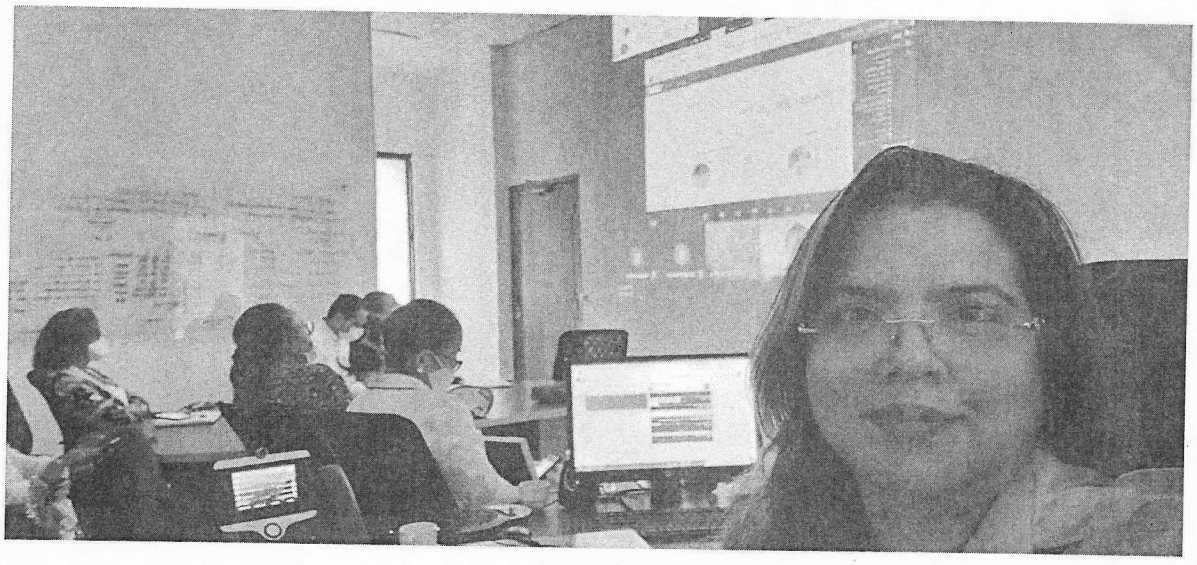
LILIANA MILENA PLAZA ÑUSTE
Líder de Economía Digital

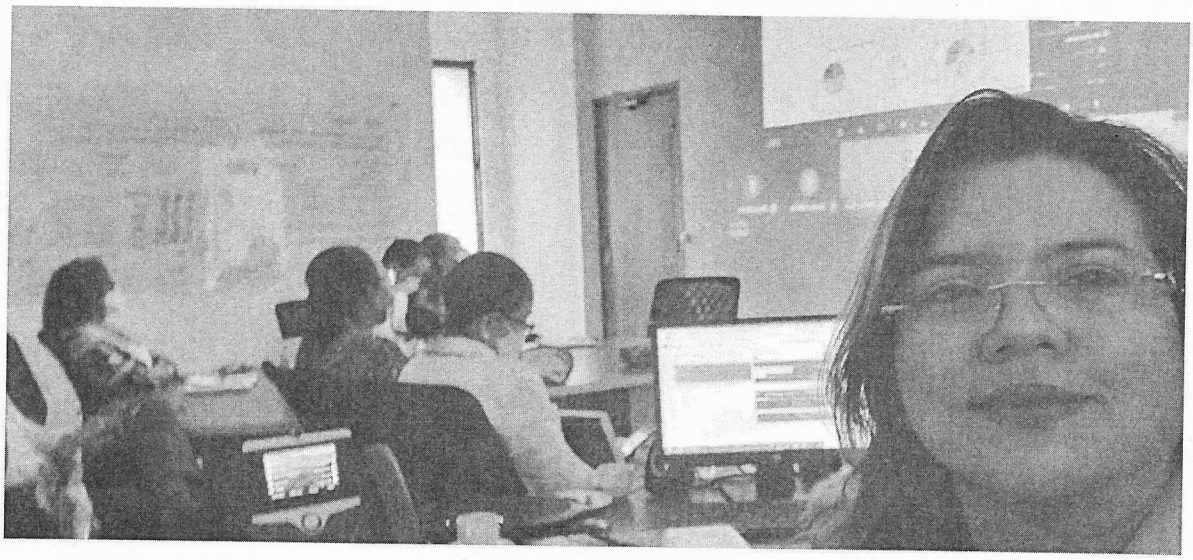
SONIA YAMILETH CASTRO YAMA
Asesora Gobernadora

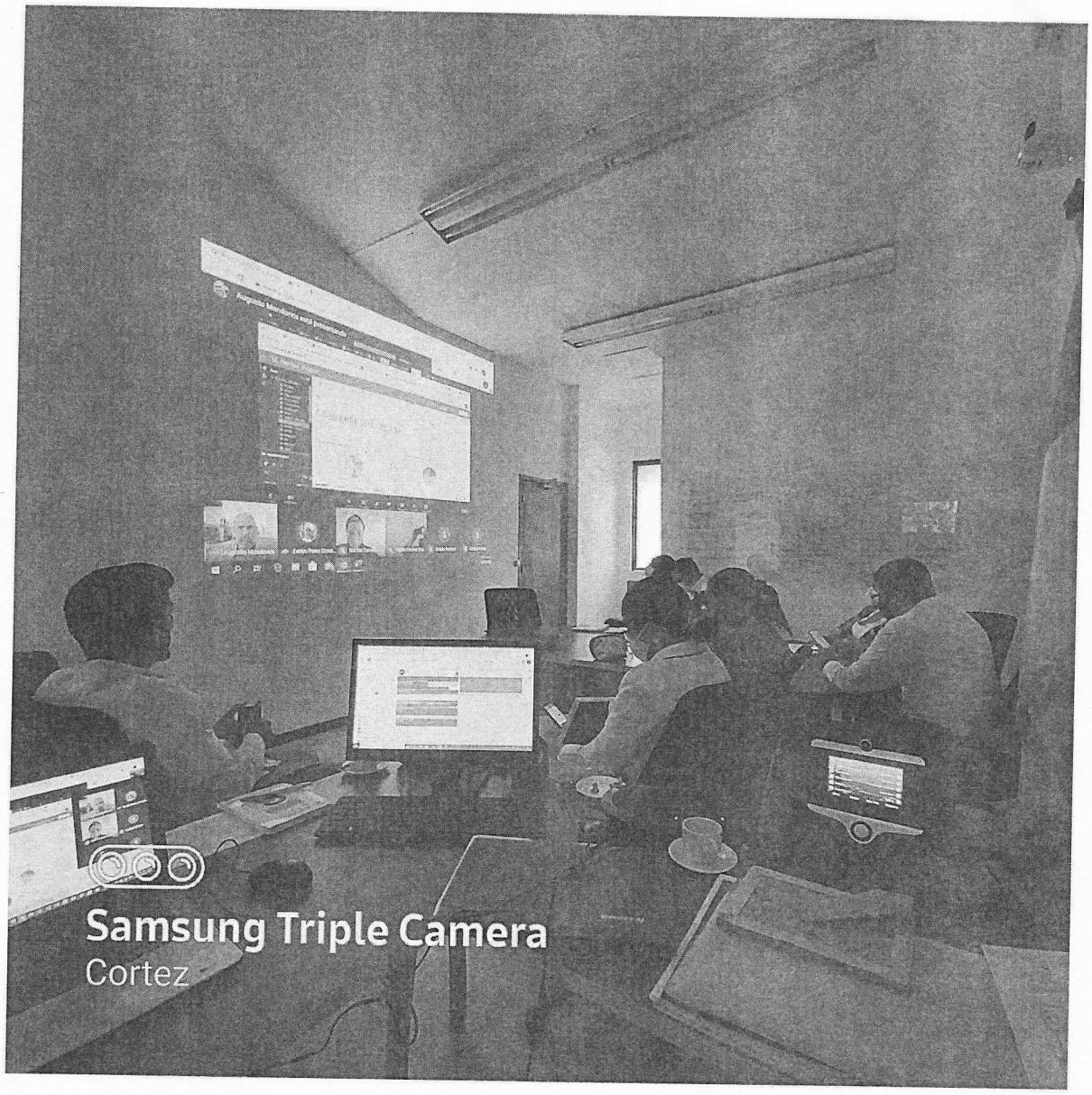
(se anexan imágenes)

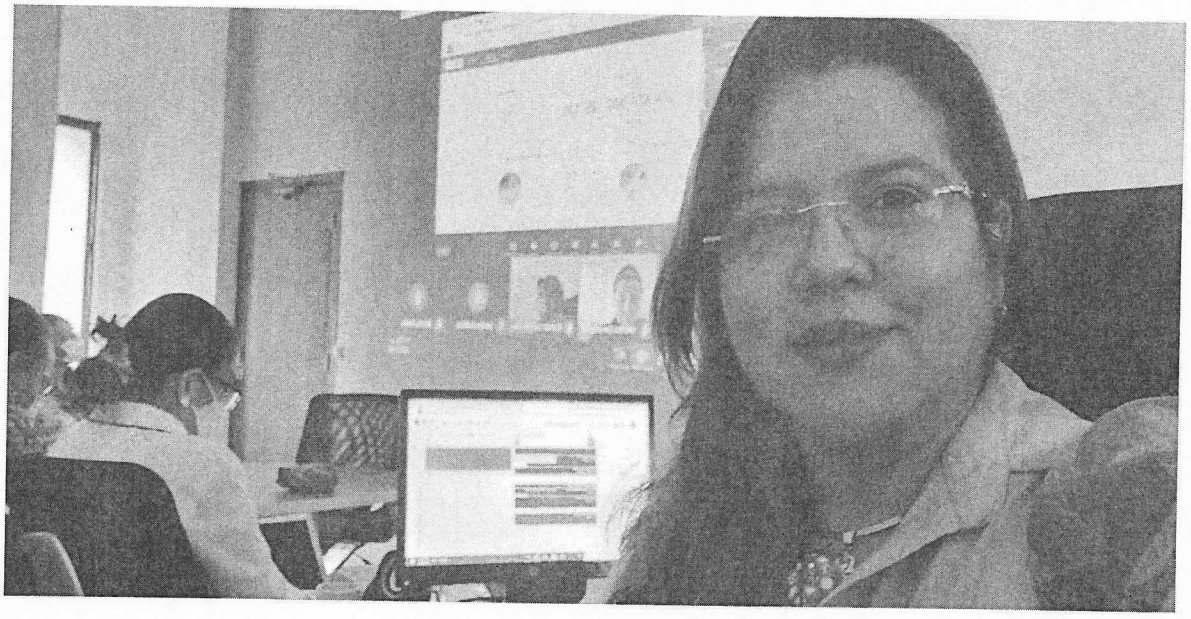
Transcriptor: Mariela Ivonne Sinisterra Muñoz- Técnico

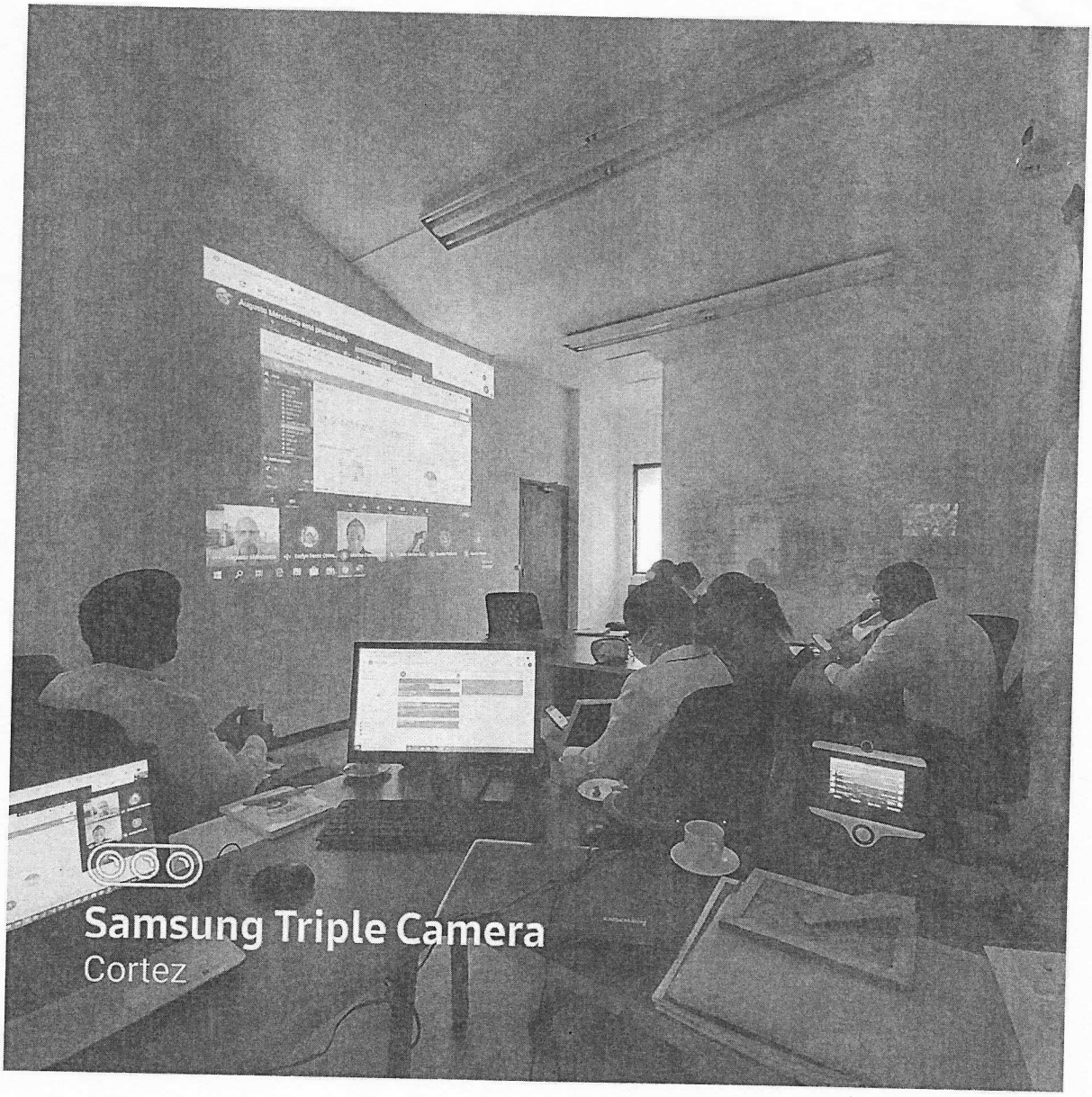
Archivarse en:

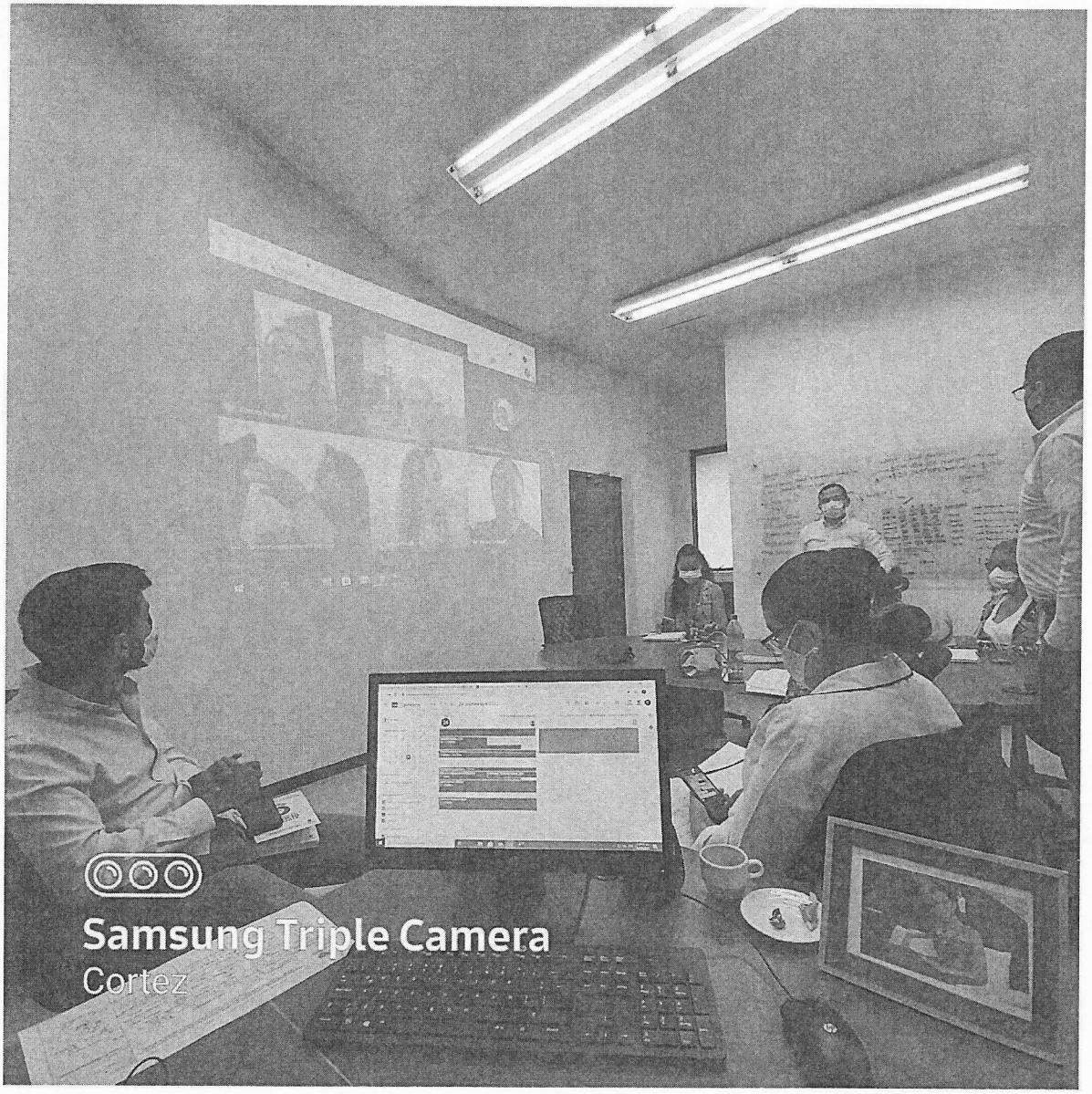


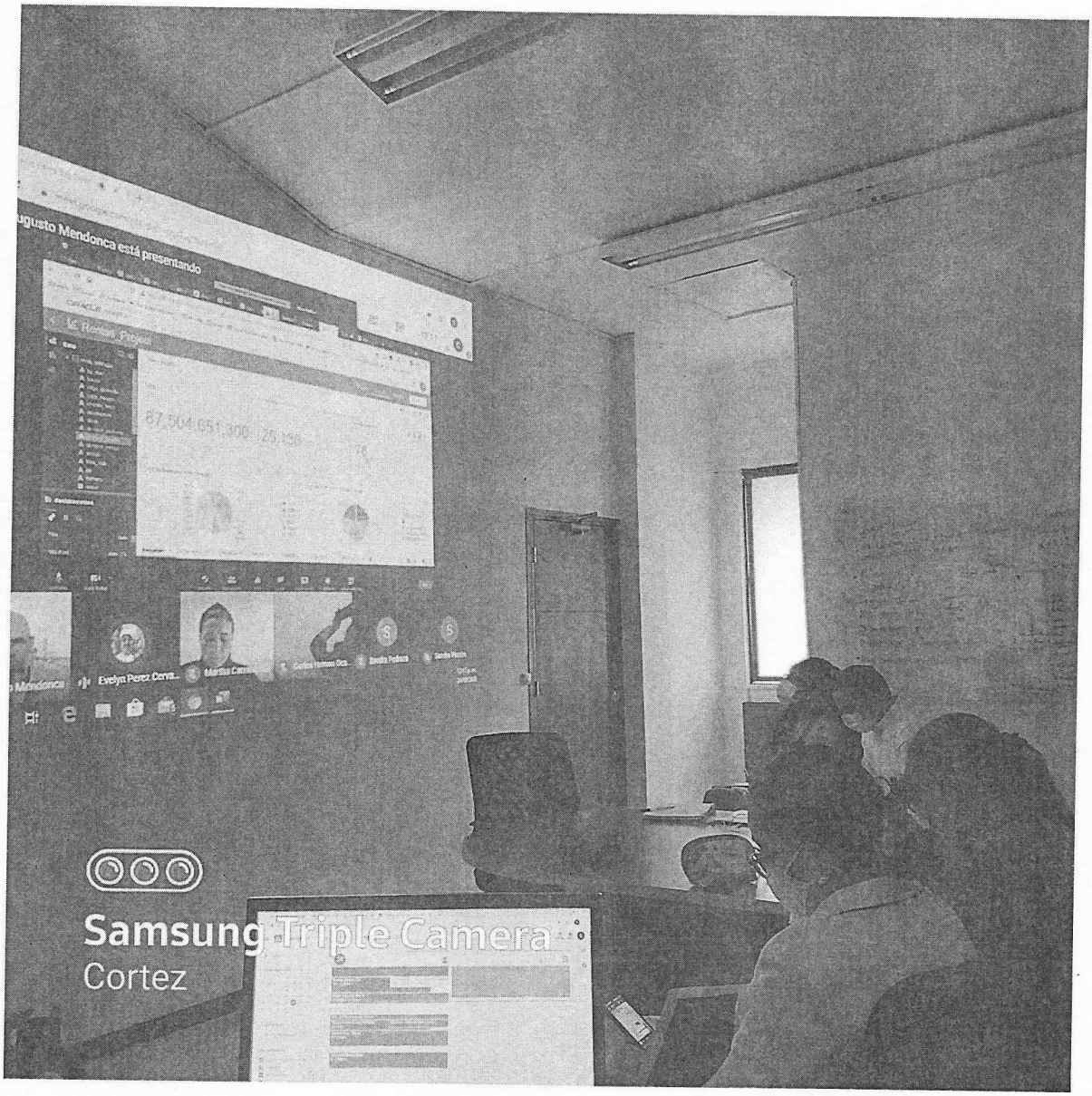


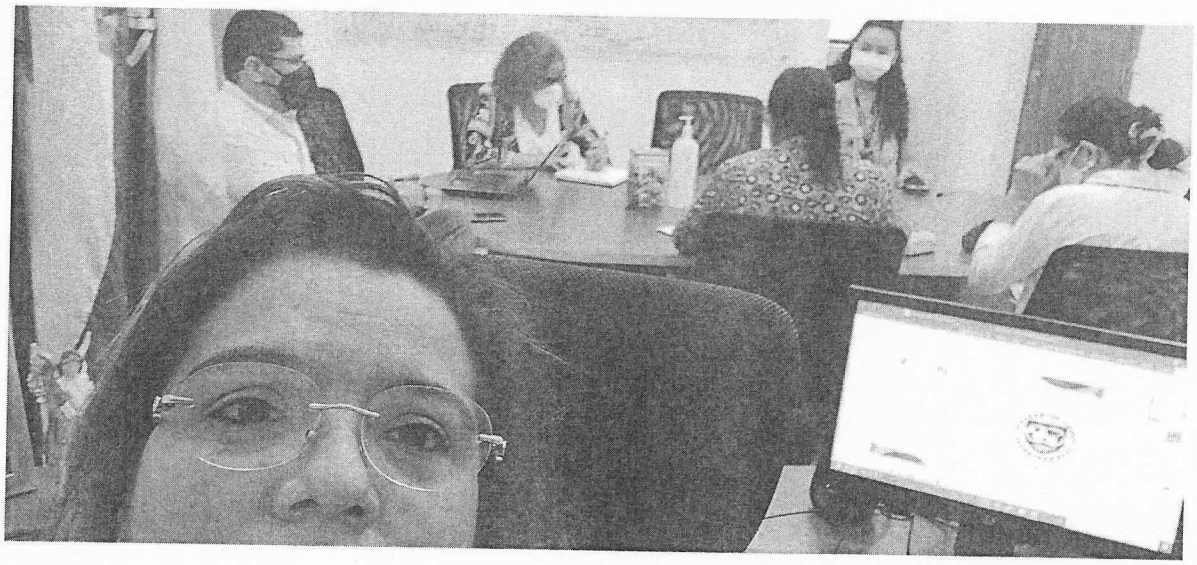

Samsung Triple Camera
Cortez

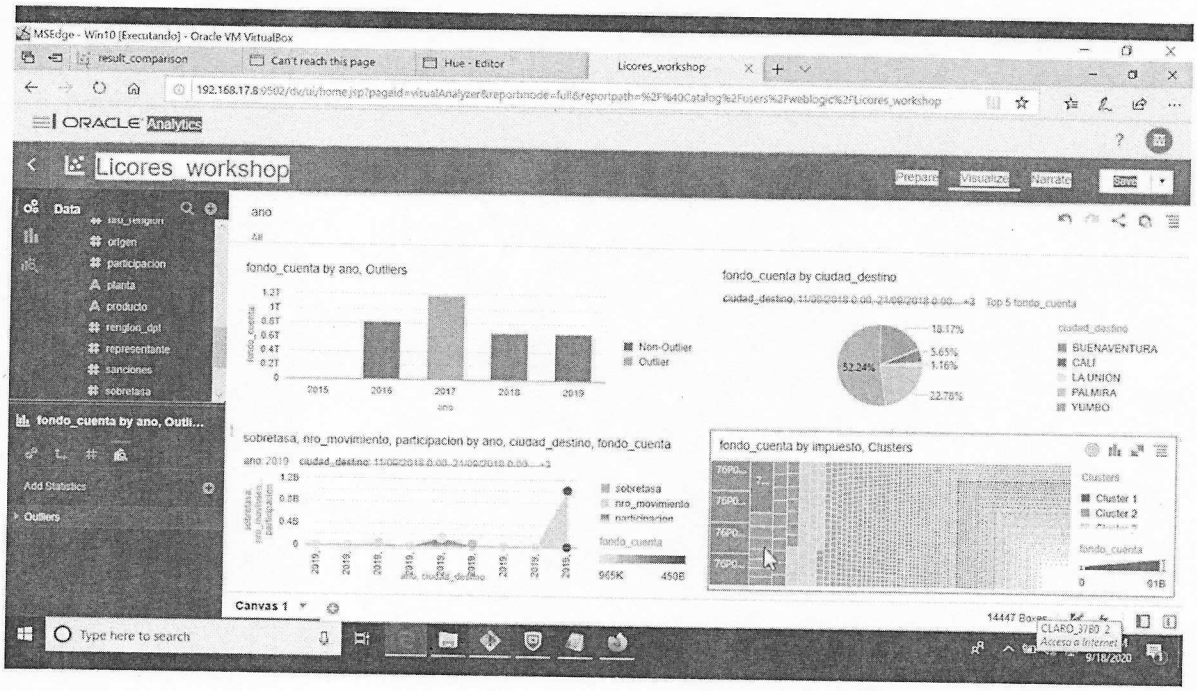



Samsung Triple Camera
Cortez


Samsung Triple Camera
Cortez


Augusto Mendonca está presentando
Samsung Triple Camera
Cortez




MSEdge - Win10 [Executando] - Oracle VM VirtualBox
result_comparison
Can't reach this page
Hue - Editor
Licores_workshop
ORACLE Analytics
Licores_workshop
Prepare
Visualize
Narrate
Save
Data
origen
participacion
planta
producto
region_dpt
representante
sanciones
sobretasa
fondo_cuenta by ano, Outli...
Add Statistics
Outliers
ano
fondo_cuenta by ano, Outliers
Non-Outlier
Outlier
fondo_cuenta by ciudad_destino
Top 5 fondo_cuenta
BUENAVENTURA
CALI
LA UNION
PALMIRA
YUMBO
sobretasa, nro_movimiento, participacion by ano, ciudad_destino, fondo_cuenta
sobretasa
nro_movimiento
fondo_cuenta
fondo_cuenta by impuesto, Clusters
Clusters
Cluster 1
Cluster 2
Canvas 1
Type here to search
9/18/2020

Valle Invencible

GOBERNACIÓN VALLE DEL CAUCA

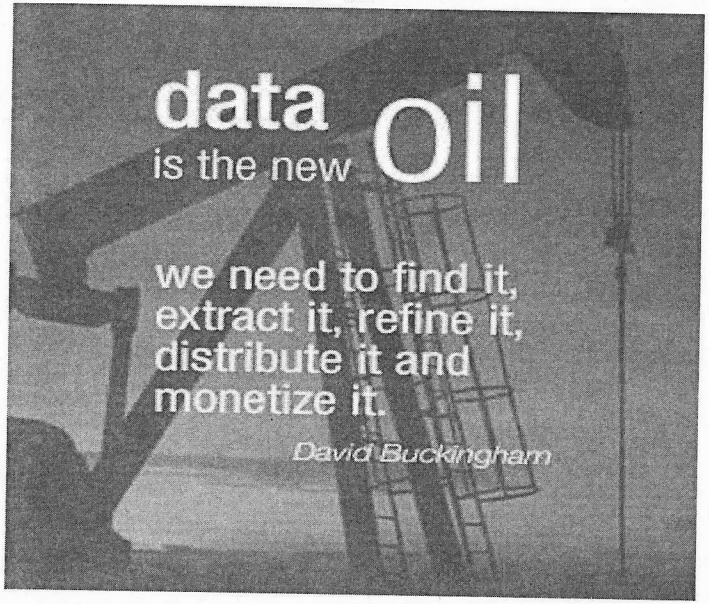
FONDO MONETARIO INTERNACIONAL

Foro Económico Mundial en el documento del The Future of Jobs Report 2018, nos alerta sobre la pérdida de 75 millones de puestos de trabajo para el 2030 por la oleada tecnologías emergentes.

Conferencista
Dependencia

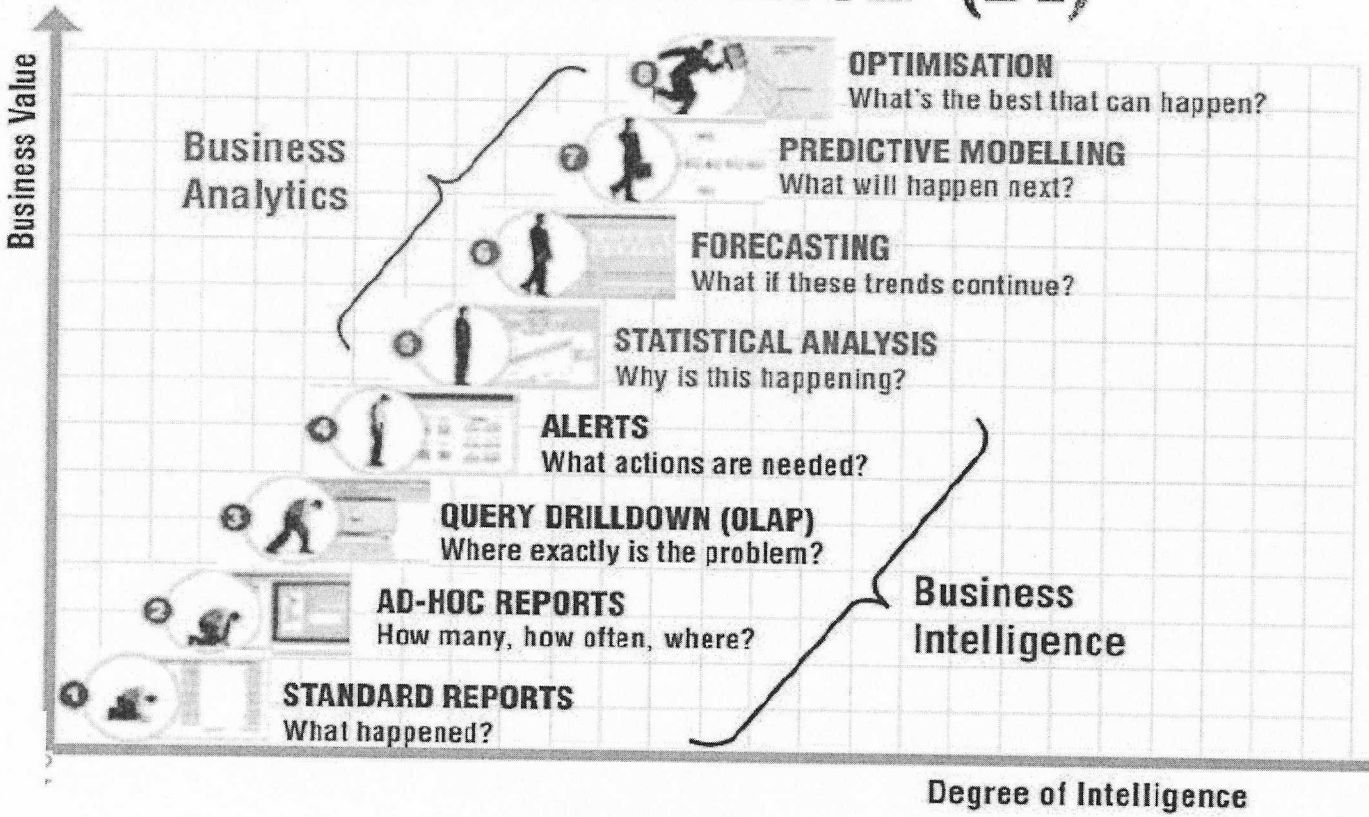
24/11/2020

LOS DATOS



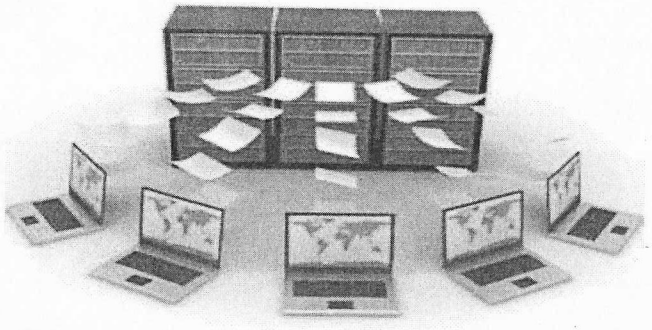
- Materia prima
- "Huella dactilar"
- Necesitan tratamiento
- Activo estratégico
- Monetizable
- Granularidad

BUSINESS INTELLIGENCE (BI)



www.rosebt.com

Datos: activos de la organización



- **Activo (asset):** recurso económico que se puede tener o controlar y que posee o produce valor.
- La monetización de los datos es un aspecto cada vez más común y pronto entrará a la contabilidad de las empresas.
- A medida que las organizaciones dependen más de sus datos, se puede establecer más claramente el valor de éstos.

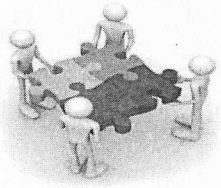
¿Cómo Identificar el valor de los datos?

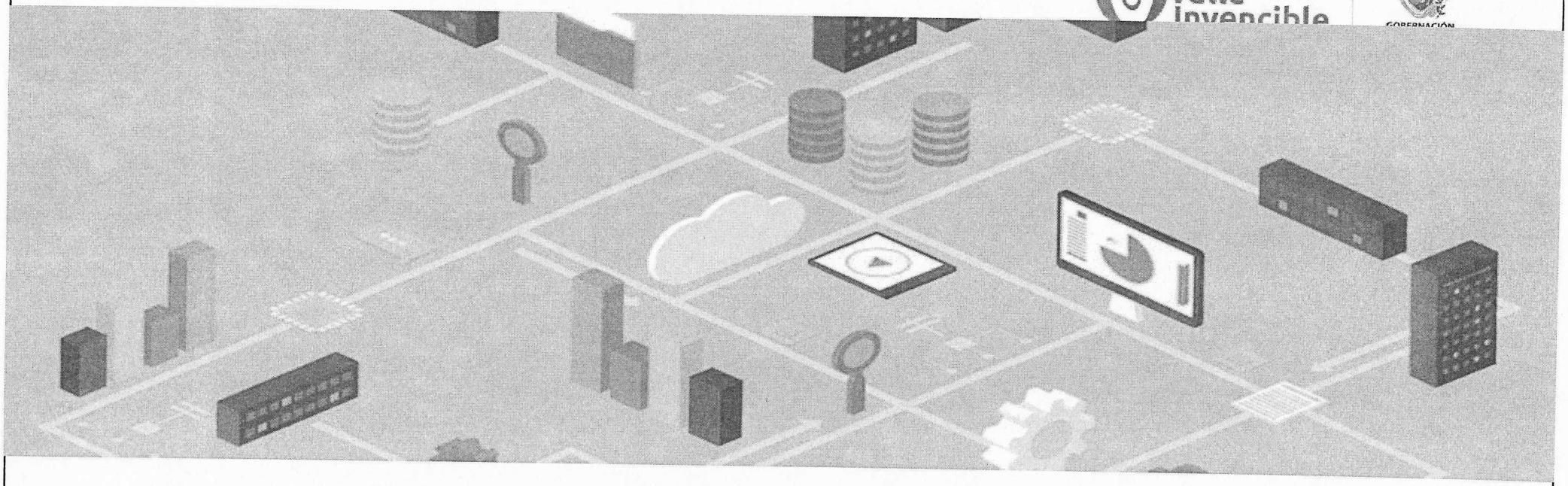
Datos como activos

3.Enumere al menos tres desafíos o problemas a los que se enfrenta su empresa hoy en día en relación con los datos.

4.Para los datos identificados, indique el grado de protección que se les da en la empresa. Defina una escala (1 al 5; alto-medio-bajo; etc).

5.Aún es difícil comprender los datos como activo. Es difícil darles un valor monetario, a menos que su empresa venda los datos. Sin embargo, los datos son un activo. Plantee al menos tres razones por las cuales los datos de su organización **sí** tienen un valor monetario





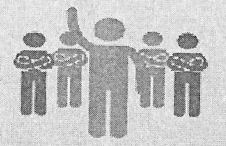
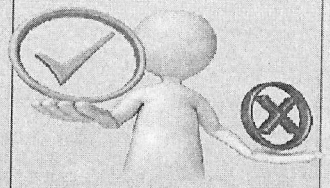
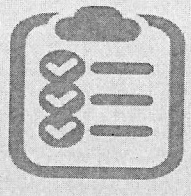
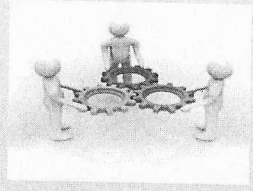
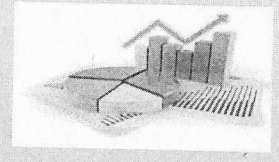
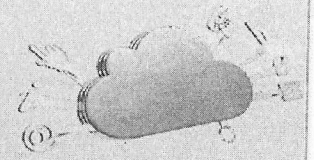
- La gobernanza de datos es para garantizar que la gestión de datos ocurra. Y que se haga correctamente.

Visión general de un programa de DG

Valle Invencible

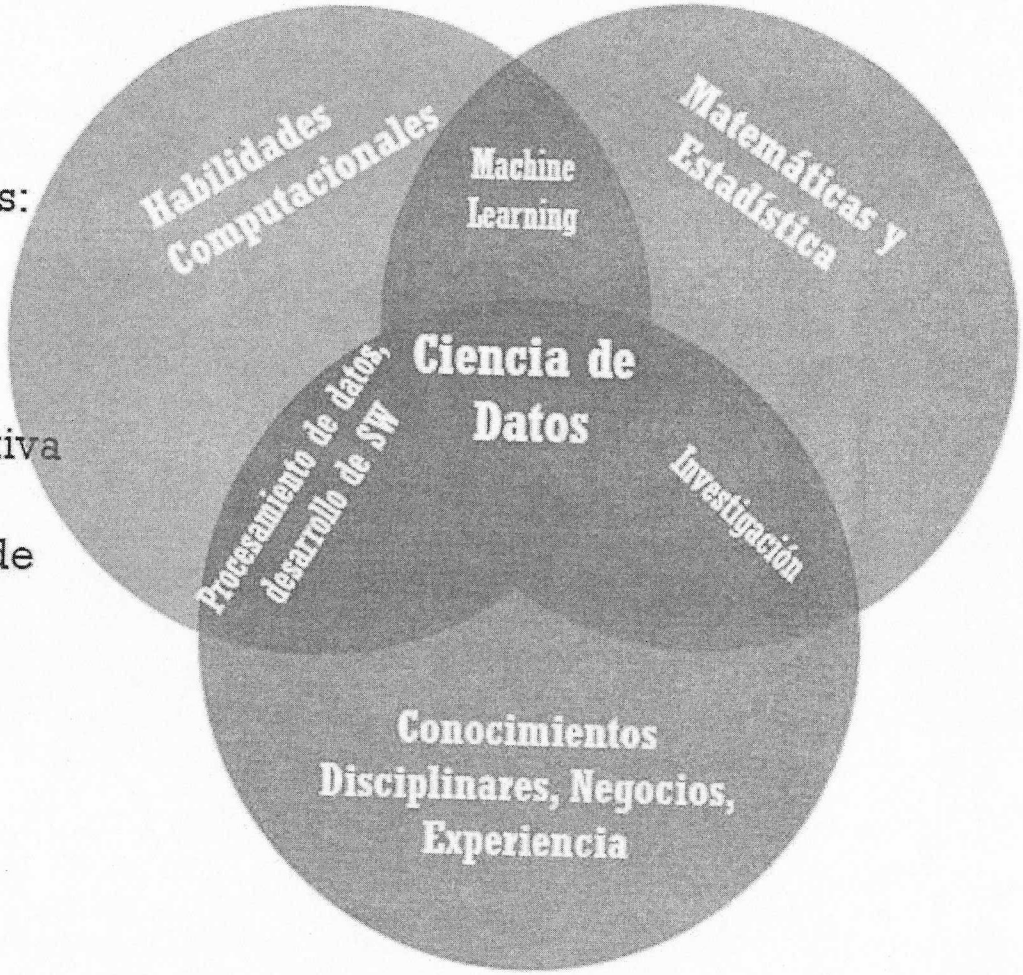
GOBERNACIÓN VALLE DEL CAUCA

Elementos del programa

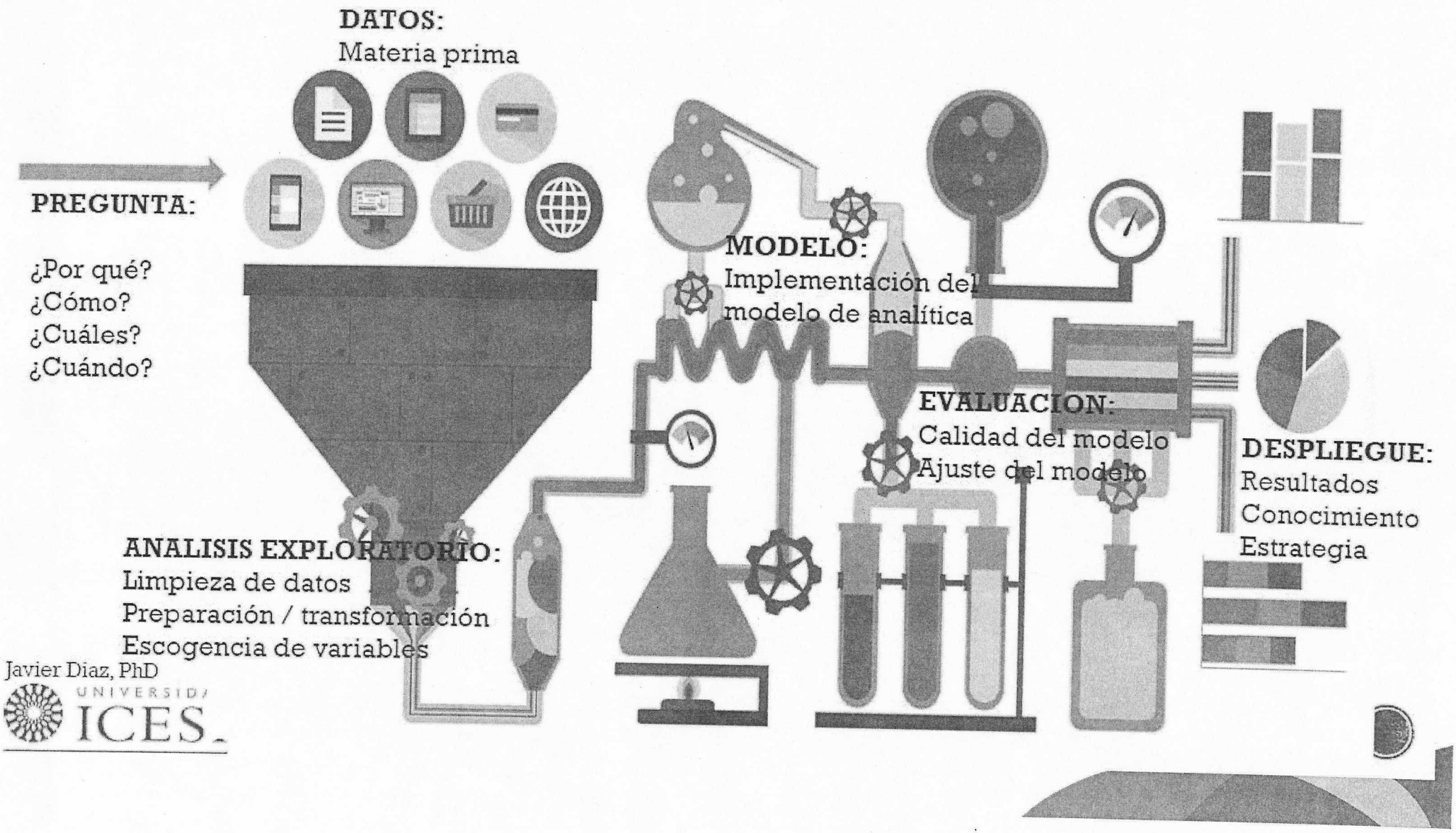
Organización de roles	Principios	Políticas	Funciones	Métricas	Herramientas y tecnología
					

CIENCIA DE DATOS

- La **ciencia de datos** incluye, entre otros:
  - Fundamentación matemática,
  - Pensamiento computacional,
  - Pensamiento estadístico,
  - Preparación, limpieza y gestión efectiva de los datos,
  - Técnicas de descripción y curación de datos,
  - Técnicas de modelos de datos,
  - Comunicación efectiva,
  - Reproducibilidad de experimentos,
  - Mejores prácticas,
  - Dilemas éticos



Javier Diaz, PhD
UNIVERSIDAD ICESI

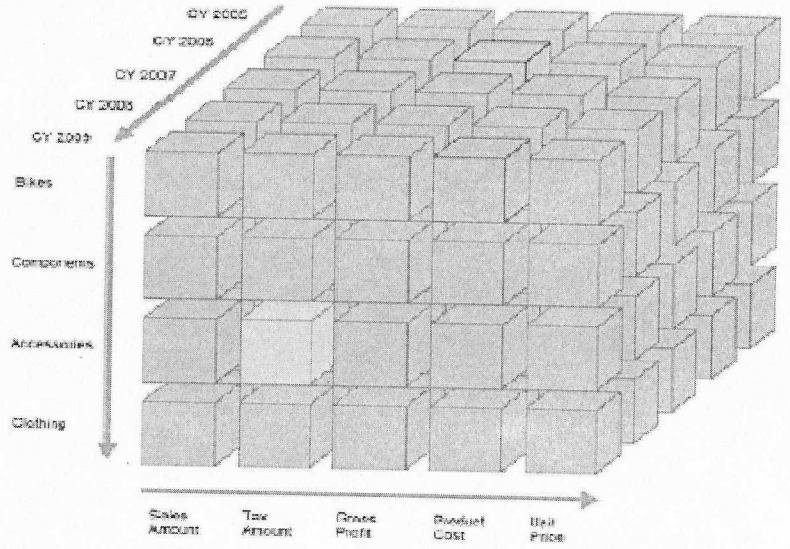


Javier Diaz, PhD
UNIVERSIDAD ICESI

BUSINESS INTELLIGENCE (BI)

OLAP (On-line Analytical Processing): Conjunto de operaciones de alto rendimiento que permiten proporcionar rápidamente información útil para la toma de decisiones a nivel estratégico, utilizando modelos multi-dimensionales

- Diferente del sistema tradicional de operaciones usado por las bases de datos relacionales: **OLTP** (On-Line Transactional Processing)
- Salidas de los procesos en forma de matrices, donde las filas y columnas son compuestas por las **dimensiones** (e.g. productos, clientes, puntos de venta, fechas, etc), y los valores de las celdas corresponden a **hechos** (e.g. ventas, cantidades, gastos). Conceptualmente se ilustra como un cubo
- Usuarios pueden hacer operaciones OLAP, "navegando" sobre el cubo, y especificando cálculos analíticos → self service



Reto 1

Convocatoria Min Tic

Desde la Dirección de Desarrollo de la Industria TI del Ministerio de Tecnologías de la Información y las Comunicaciones, con la convocatoria "Retos de entidades públicas y privadas aplicando Ciencia de Datos", el cual tiene como objetivo gestionar las capacidades de conocimiento tecnológico en inteligencia artificial y procesamiento de datos en beneficio de la productividad y competitividad del sector productivo de la industria y el estado, surge la necesidad de fortalecer y apoyar el desarrollo de competencias y habilidades digitales en el talento humano de la economía nacional.

Conferencista
Dependencia

24/11/2020

objetivo

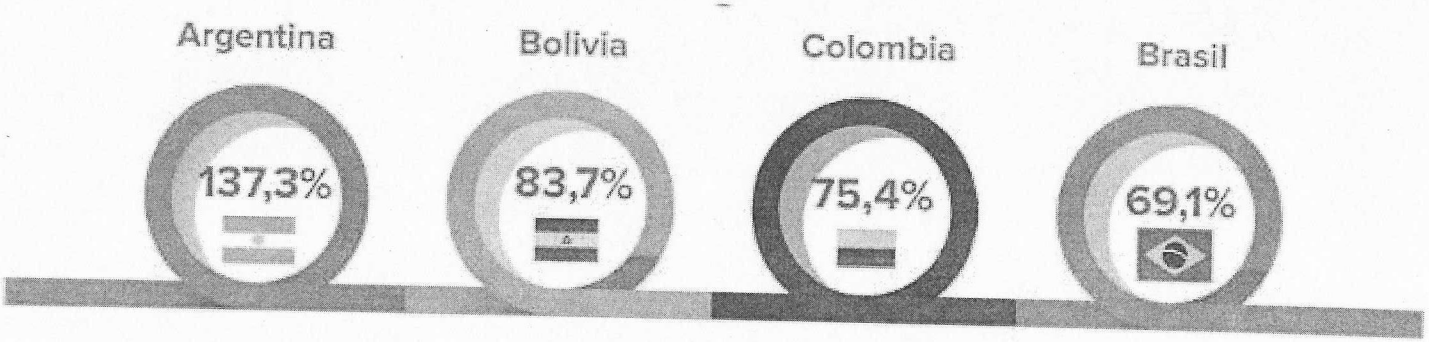
- reto: "Algoritmos para identificación y detección de Contribuyentes omisos, inexactos o evasores de impuestos del Valle del Cauca", un Análisis de información Exógena Nacional en programas de Inteligencia Fiscal Tributaria.
- Prototipo mediante ciencia de datos un modelo predictivo para detectar contribuyentes omisos.


Fiscalización de impuestos en el Valle del Cauca
DS 4A
ENTRENAMIENTO PARA EL FUTURO DIGITAL
GOBIERNO DE COLOMBIA

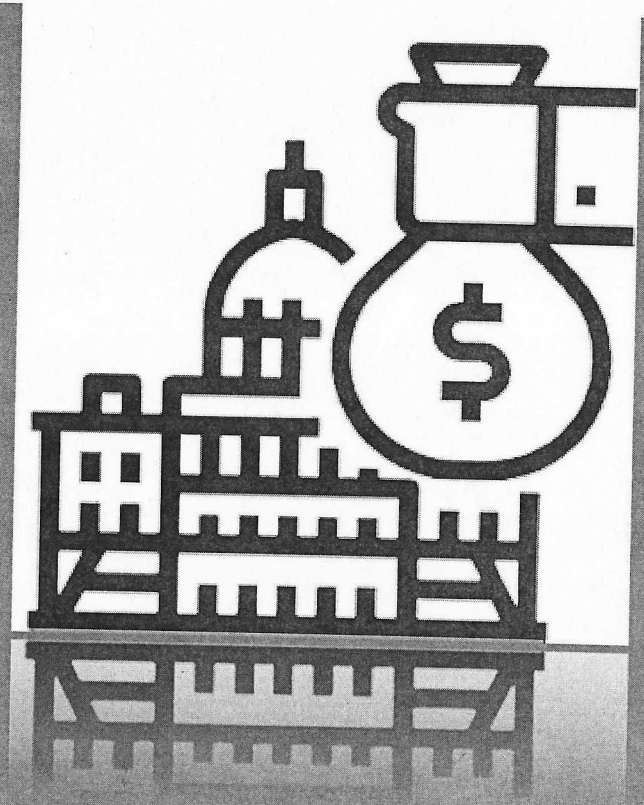
"El nativo está dispuesto a cumplir con sus obligaciones porque reconoce que de ese cumplimiento se desprenden beneficios y privilegios."

Bronislaw Malinowski

“La tasa de impuestos de Colombia es la tercera más alta de la región”



"La evasión equivale a 30% del total de lo que se recauda de impuestos al año"



3% Evasión del total del **PIB**

$15 billones

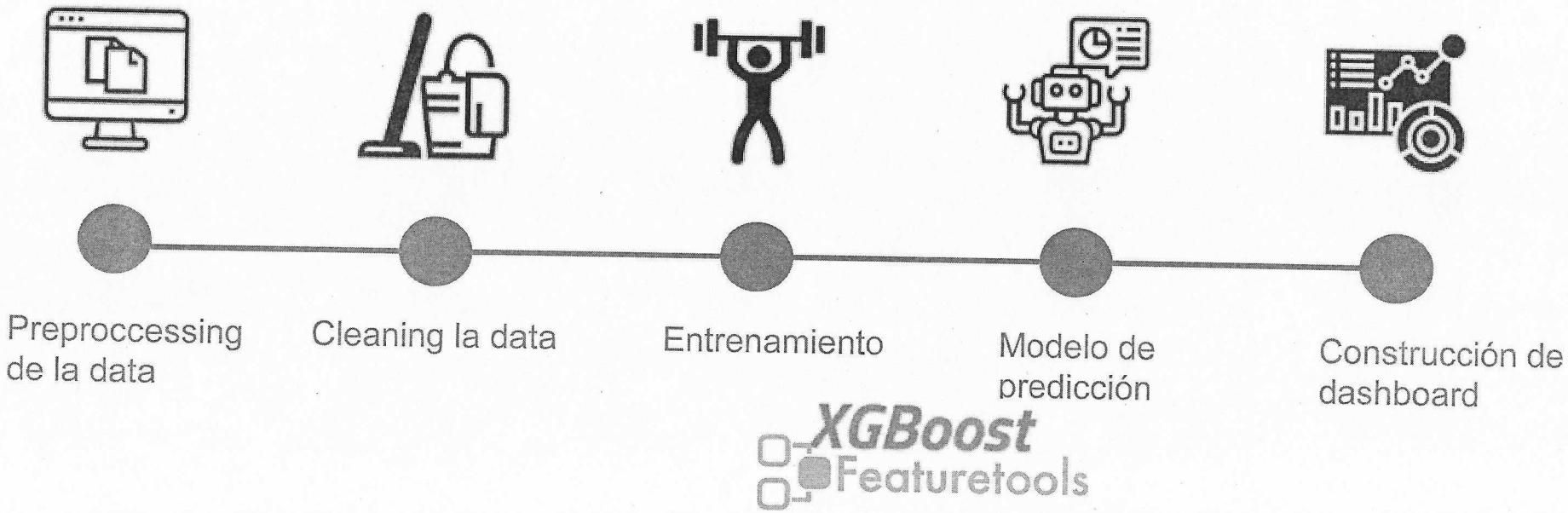
Evasión de impuestos de renta a personas jurídicas llega al año

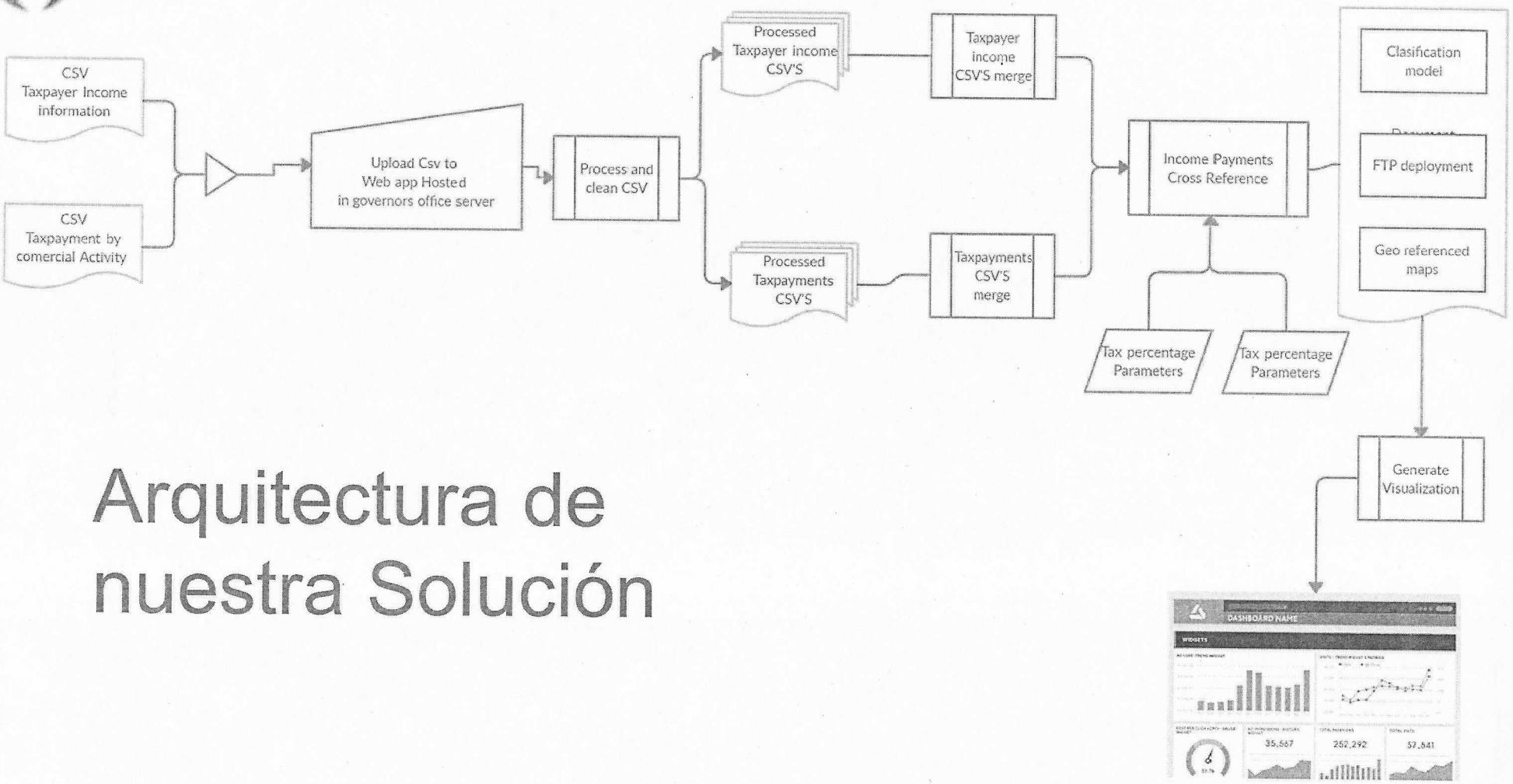

Impuestos fiscalizados en la solución
7 7 7

...Què se hizo?

Procedimiento para la Creación del Modelo.


... la solución

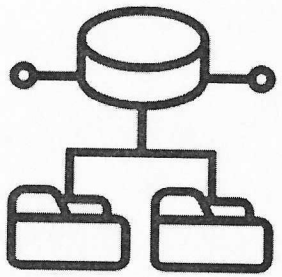

Preproccessing de la data
Cleaning la data
Entrenamiento
Modelo de predicción
Construcción de dashboard
XGBoost
Featuretools



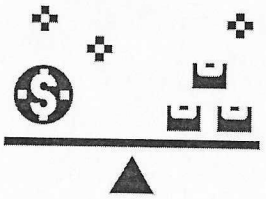
Arquitectura de nuestra Solución

entrenamiento

...dataset para el



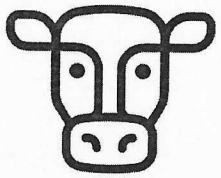
36% de la data entregada se utilizó para generar el modelo



1629 empresas entraron en la revisión de fiscalización de este modelo

85% de los registros corresponden a Cali

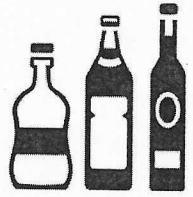
14% Degüello de ganado



34.3% Loterías foraneas

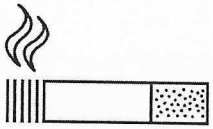
2.4% Imp Cervezas

3.3% Imp Licores

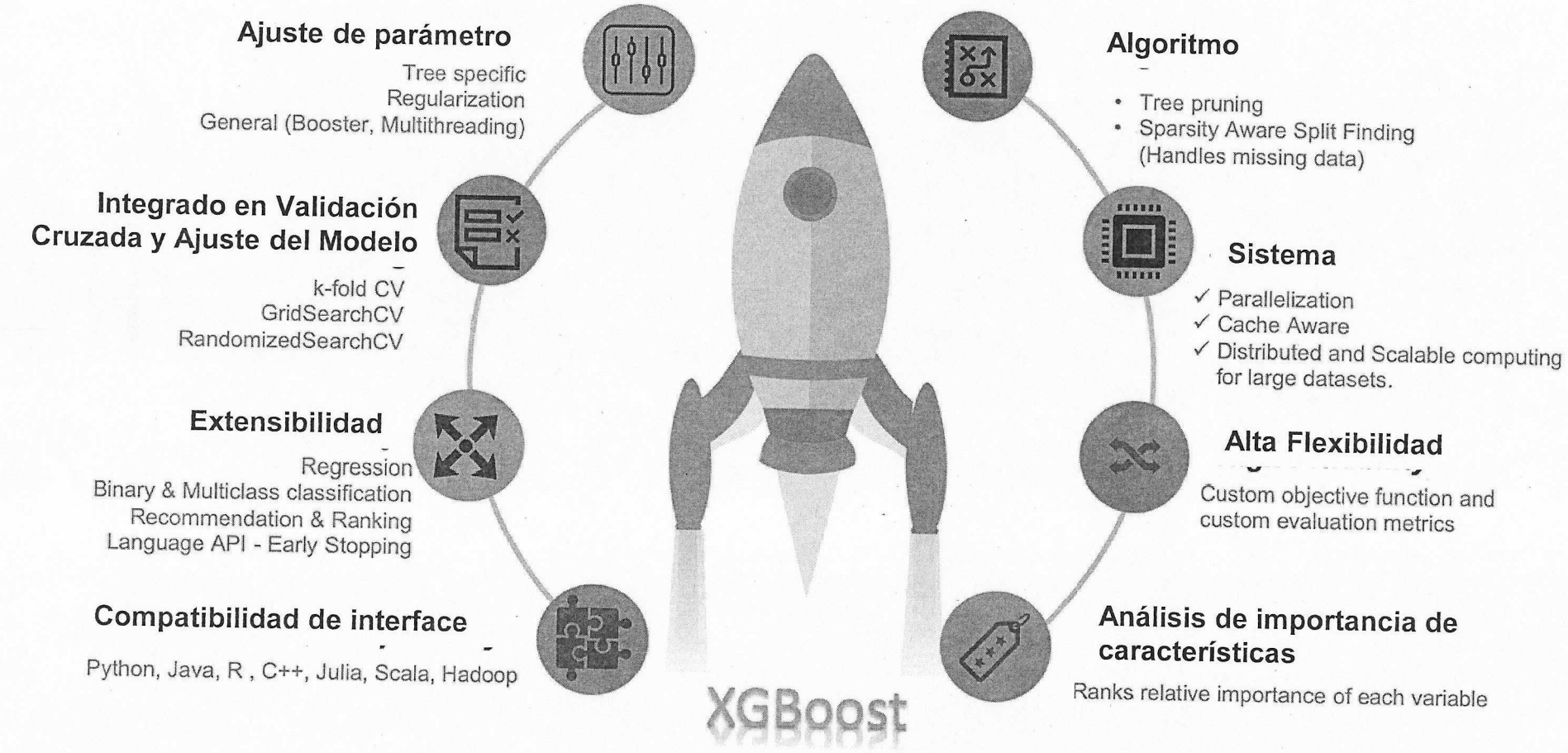


29.9% Imp para recreación y deporte

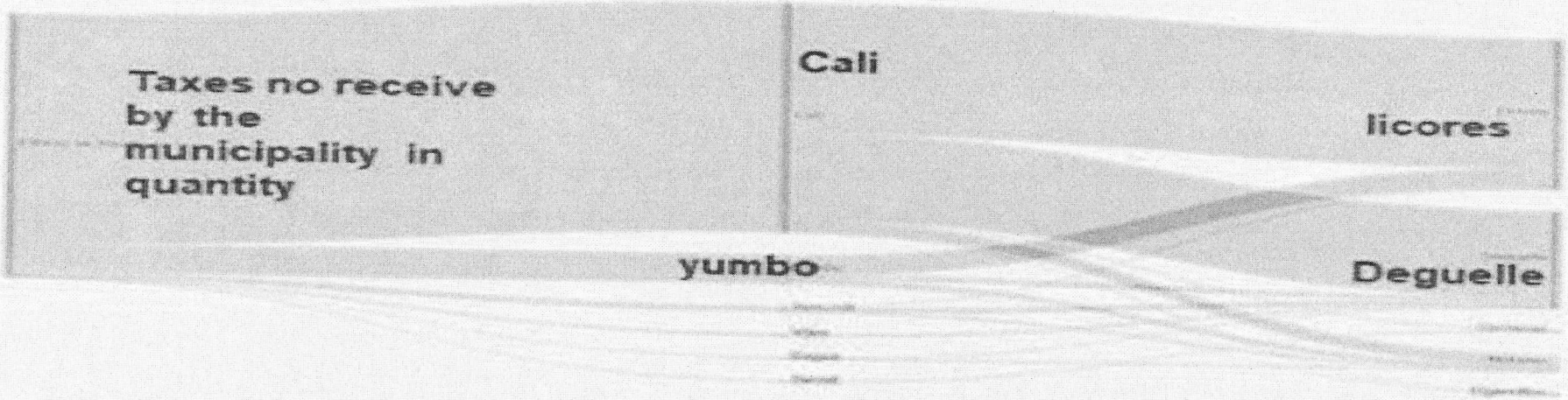
0.12% Imp para recreación y deporte



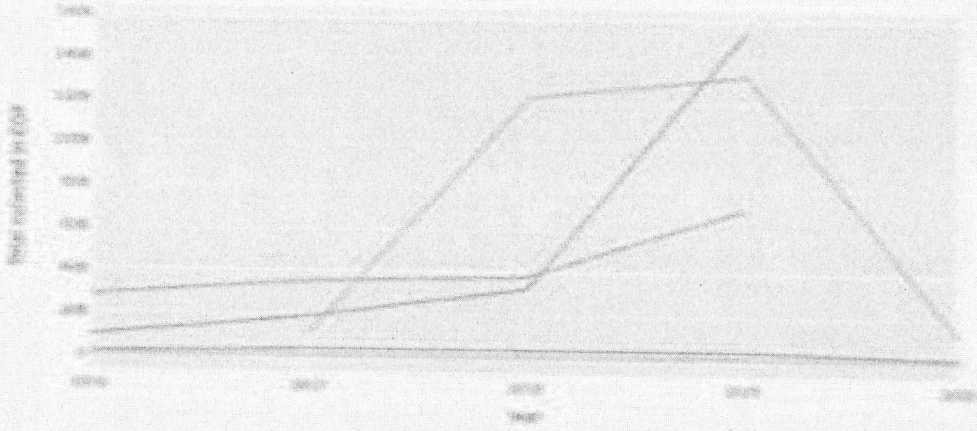
...Modelo de predicción

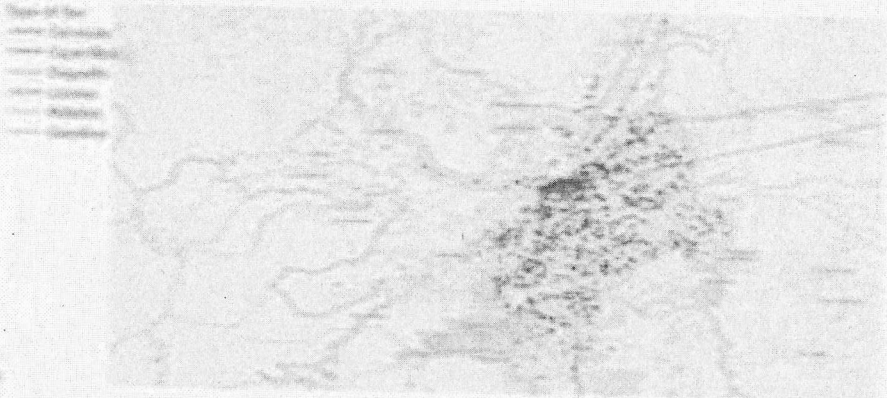

Ajuste de parámetro
Tree specific
Regularization
General (Booster, Multithreading)
Integrado en Validación Cruzada y Ajuste del Modelo
k-fold CV
GridSearchCV
RandomizedSearchCV
Extensibilidad
Regression
Binary & Multiclass classification
Recommendation & Ranking
Language API - Early Stopping
Compatibilidad de interface
Python, Java, R , C++, Julia, Scala, Hadoop
Algoritmo
• Tree pruning
• Sparsity Aware Split Finding (Handles missing data)
Sistema
✓ Parallelization
✓ Cache Aware
✓ Distributed and Scalable computing for large datasets.
Alta Flexibilidad
Custom objective function and custom evaluation metrics
Análisis de importancia de características
Ranks relative importance of each variable
XGBoost

And this is what we got to see:

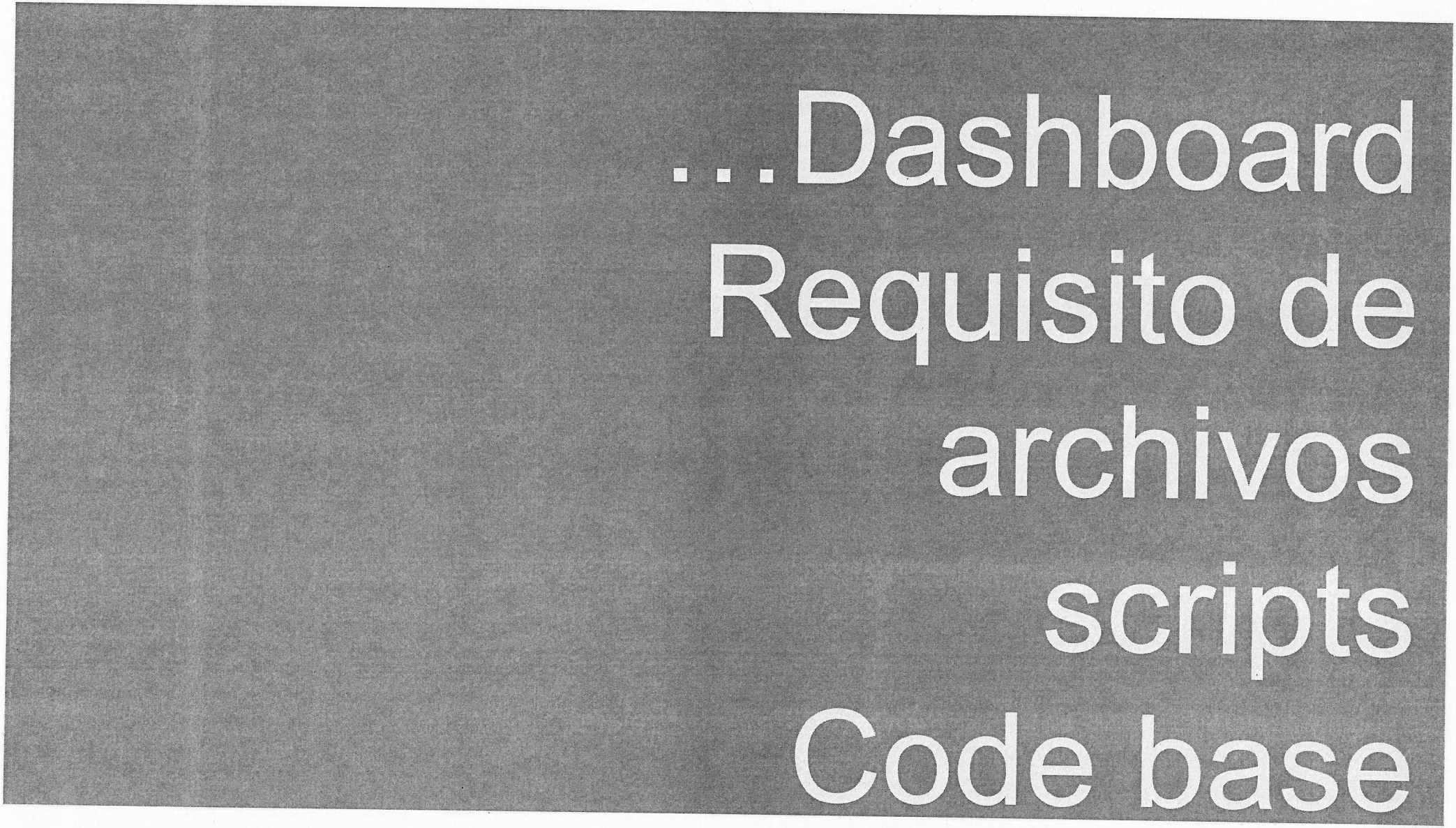


Taxes no receive by year and its location within the municipality





...Dashboard
Requisito de archivos scripts
Code base


...Dashboard
Requisito de
archivos
scripts
Code base

Próximos pasos

1. Mejoramiento de la calidad de la información proporcionada
2. Análisis basado en variables cualitativas para poder tener una perfilamiento de las compañías en la evasión
3. Análisis de priorización en el cobro y esperado de pago

Próximos pasos

1. Mejoramiento de la calidad de la información proporcionada
2. Análisis basado en variables cualitativas para poder tener una perfilamiento de las compañías en la evasión
3. Análisis de priorización en el cobro y esperado de pago

colombia/volando-alto-con-bdo-
to-de-la-evasion-de-impuestos-en
sion-equivale-a-30-del-total-de-lo-que

colombia/volando-alto-con-bdo-colom
s

- https://www.bdo.com.co/es-co/blog-bdo-en-colombia/volando-alto-con-bdo-colombia/septiembre-2017/infografia-el-costo-de-la-evasion-de-impuestos-en
- https://www.larepublica.co/economia/la-evasion-equivale-a-30-del-total-de-lo-que-se-recauda-de-impuestos-al-ano-2945888
- https://www.bdo.com.co/es-co/blog-bdo-en-colombia/volando-alto-con-bdo-colombia/abril-2017/infografia-el-costo-de-evadir-impuestos

GOBERNACIÓN VALLE DEL CAUCA

REGISTRO DE ASISTENCIA (CAPACITACIÓN, INDUCCIÓN, REIDUCCIÓN, ASESORÍA, ASISTENCIA TÉCNICA, REUNIONES, CONSEJO DE GOBIERNO, COMITÉ TECNICO DE LA DEPENDENCIA, EQUIPO DE MEJORA CONTINUA DEL PROCESO)

Código: FO-MI-P1-14
Versión: 01
Fecha de Aprobación: 15/08/2018
Página: 1 de

LUGAR: Secretaría TIC - Reto 1

TOTAL HORAS:

No. DE ACTA:

HORA DE INICIO: 10:30 A.M. HORA DE TERMINACIÓN: 1:00.

CODIGO DEL PROCESO/SUBPROCESO:

FACILITADOR (ES)/RESPONSABLE: Carlos Hernán Ocampo - Sonia Castro - Liliana Plaza.

NOMBRE DEL EVENTO/TEMA DE REUNION/TEMAS A TRATAR: Presentación de resultados primer reto en Ciencia de datos y revisión de participación Segundo reto

FECHA: 24-09-2020

No.	DEPENDENCIA / ENTIDAD/MPIO	NOMBRES Y APELLIDOS COMPLETOS	CARGO	CÉDULA	No. DE CELULAR - TEL/EXT	CORREO ELECTRÓNICO	FIRMA DE ASISTENCIA
1	Setic	Liliana Plaza	Líder Programa	31714613	2023	lmplaza@valledelcauca.gov.co	Liliana Plaza.
2	Rentas	Zoraida Bonilla	Gerente	31901512	—	zbonilla@valledelcauca.gov.co	
3	SETIC	Carlos H. Ocampo	Secretario	94431700	—	chocampo@valledelcauca.gov.co	
4	Rentas	Elsy Marcela Rolando G	Profesional Universitario	66842338		enbalanta@valledelcauca.gov.co	
5	UAEI Rentas y G. Tributaria	Liliana Rodriguez R	Sub. G. Fiscal	31.202.730	1948	lrodriguez@valledelcauca.gov.co	
6	Rentas	Miguel Salazar	Contratista	1130622772	1910	[illegible]	
7	Unidad de Rentas	Cara Cobra Galan	Prof. Universitario	1110512611	1943	Ilgalan@valledelcauca.gov.co	
8	Setic	Sonia Castro	Asesor	0692137	2300	Scastro	
9							
10							
11							
12							
13							

Valle Invencible

GOBERNACIÓN VALLE DEL CAUCA
Secretaría de las Tecnologías de la Información y las Comunicaciones

1.360.2-79

ACTA DE REUNIÓN INNOVACIÓN PÚBLICA DIGITAL CIENCIA DE DATOS
SECRETARÌA DE LAS TIC

ACTA 155

000

FECHA: Santiago de Cali, 19 de octubre de 2020

HORA: 4:23 pm a 5:23 pm

ASUNTO: CONVENIO CON UNIVALLE PARA HABLAR SOBRE EL PROYECTO DEL CÁNCER

ASISTENTES: Dr. Oswaldo Solarte
Liliana Plaza
Ing. Carlos Ocampo
Sonia Yamileth Castro Yama
Mariela Ivonne Sinisterra Muñoz

ORDEN DEL DÍA:

1. SE INICIA LA REUNIÓN CON EL SALUDO
2. RETROALIMENTACIÓN DEL CONVENIO

DESARROLLO:

1. La Dra. Sonia Yamileth Castro Yama, le da la palabra al profesor Solarte, quien informa sobre la necesidad crear un prototipo de inteligencia artificial para extraer información de carácter investigativo para manejo de los oncólogos en el tratamiento del cáncer, en este caso, para acelerar la investigación médica.

   Los datos obtenidos son de carácter investigativo para crear herramientas que le puedan servir a los médicos.

2. El Ing. Carlos Ocampo, interviene y pregunta sobre la información de las historias clínicas ¿esa información está al día?

3. El Dr. Oswaldo Solarte procede a responder e informa que el hospital en España actualmente, les entrega las historias clínicas de 500 pacientes con cáncer y les pide que investiguen qué sucede, les interesaba principalmente los pacientes con cáncer de pulmón. Teniendo en cuenta que los ciudadanos son obsesivos con el uso del cigarrillo desde una edad muy temprana (14 años). Pero aquí en Colombia existen otros tipos de cáncer más comunes.

   También está trabajando con el registro de cáncer de Cali, donde solo se manejan reportes de patología, pero actualmente, está sacando unos resultados donde se demuestra que, en el Valle en los últimos 10 años, se está presentando un incremento en los casos de cáncer de cuello uterino y cáncer de mama; información obtenida gracias a la tecnología y a la inteligencia artificial. También se pudo evidencia que el 80 % de los pacientes con cáncer están en los estratos 1,2 y 3, en una relación de 3 a 1 con los estratos 4,5 y 6.

   Por el momento solo se cruzan esas dos variables, cabe mencionar que uno de los problemas del registro de cáncer es que solo existe el registro del patólogo, pero en la historia clínica es donde está todo lo del paciente, que medicamentos le han aplicado, si tuvo efectos adversos, si el medicamento funcionó.

   Un ejemplo es la investigación que se hizo en España, una vez analizaron todas estas variables, se dieron cuenta que los pacientes de cáncer de pulmón tienen un tiempo de supervivencia de máximo 5 años y el tiempo promedio es de 3 años. Con esa información puede proyectar cuánto costará el tratamiento teniendo en cuenta el tamaño del tumor (1, 2, 3 o 4); ya con esta información se está empezando a hacer los modelos.

   La idea que tiene es apoyar a la Ciudad, que el sistema de salud se pueda utilizar todas estas tecnologías, utilizar la información que existe, pero no se utiliza con nuestro apoyo. En Europa se han creado empresas solo para analizar este tipo de información, ya que no solo beneficia el sistema de salud, también a futuro, se puede empezar a capacitar personas para que empiecen a crear sistema orientados al sector salud.

   Hoy en día la mayoría de información está depositada en los registros médicos y no los están aprovechando, motivo por el cual eligió esto como tesis doctoral, ya que el texto médico que uno lee es diferente a la prosa que uno lee producto de la presión con la que trabajan.

   Tiene una beca con el Ministerio de Educación con un presupuesto asignado para mis experimentos en Madrid y otro presupuesto para apoyar

a personas como en la Universidad del Valle, por eso se necesitan las historias clínicas. En caso de realizar un convenio, la Gobernación o el Hospital Universitario no tendrían que poner dinero, solamente los datos y las publicaciones científicas que saldrían de ahí, estarían la Universidad del Valle, la Gobernación y el Hospital Universitario, cuenta con el apoyo de 2 estudiantes de maestría y solo se necesitaría las historias clínicas de Cali, del Valle o de Colombia, no importa, pero que esas historias clínicas sean de pacientes con cáncer.

4. Interviene el Ing. Carlos Ocampo afirmando, que si ya tiene las historias clínicas que es lo más importante.

5. La Doctora Liliana Plaza solicita información sobre ¿cuál es el papel que desempeña la secretaría de las TIC´s, que debe poner sobre la mesa, quién proporciona los datos?

6. Responde el Dr. Oswaldo Solarte reafirmando que necesitan los datos, que en este caso los proporciona el hospital Universitario, entonces ustedes serán como el área técnica, lo que necesito son datos, no presupuesto, si logramos el convenio, necesitamos alguien de la secretaría de salud con unas horas a la semana, definir requerimientos y si es posible asignar un oncólogo para validar la información al final del proceso. Ahora bien, el Valle del Cauca sería pionero tanto en Colombia como en América Latina

7. El Interviene la representante de las TIC´s e informa que, si la información la proporciona las secretarías de salud, no necesitan un soporte técnico al proyecto.

8. Interviene el Ing. Carlos Ocampo, e informa que los dueños de la información en cierta manera, no son la Secretarías de la salud sino las ESES, quienes actualmente están en un proyecto de implementación de una historia clínica electrónica; me han encomendado desde la gobernadora para que esté al frente del proyecto, se estima que para el 31 de diciembre, estamos negociando para inter operar las 53 ESES, ya se está negociando y cuadrando los temas legales con casi el 90 % para interoperabilidad, es decir, para que se interconectan las 53 ESES, una vez este interoperar todo el tema, ya podríamos decir que hay una única fuente de información a la que podemos acceder, y no sería sino hasta el otro año.

9. Interviene el Dr. Oswaldo Solarte, por mi no hay problema, pero es necesario que los datos estén para el primer trimestre del próximo año.

10. Interviene la doctora Liliana, es necesario que antes de empezar a trabajar con esa información, y que pueda el Dr. Solarte iniciar el proyecto con apoyo de la gobernación, revisar el tema de la autorización de los pacientes, el manejo de sus datos, etc.

11. Interviene el Ing. Carlos Ocampo, e informa que es muy importante trabajar muy fuerte en las políticas de protección de datos personales.

12. Responde el Dr. Oswaldo Solarte, todo los que son datos personales, se eliminan, a nosotros no nos interesa, una vez se tiene el registro médico, el paciente pasa a ser, paciente 1, 2 o 3.

13. Pregunta el Ing. Carlos Ocampo, al trabajarlo de forma anónima ¿cuál sería la finalidad del estudio o el objetivo final con todas esas estadísticas?

14. Responde el Dr. Oswaldo Solarte, el objetivo sería conocer por qué al paciente 1 se le aplicó el mismo medicamento que al paciente 2 y uno duró 5 años y el otro 3, se empiecen a explorar esas variables, que historia personal o familiar, que hábitos tenia, y con esa información poder tratar a los nuevos pacientes.

    Ya este proceso en Estado Unidos llamado medicina de precisión, los oncólogos diseñan un tratamiento preciso para cada paciente.

    A diferencia de aquí, que los oncólogos manejan pacientes con historias clínicas de 1000 hojas, pero si le pasamos al Oncólogo una tabla estructurada, datos con las gráficas, poder mostrar la similitud entre pacientes, saber que médicamente fue más efectivo y con esto reducir costos a nivel administrativo.

15. Interviene Sonia Yamileth Castro Yama y agrega que lo más importante es predecir la enfermedad con diagnóstico cáncer.

16. Continúa el Dr. Oswaldo Solarte, es muy importante teniendo la información, realizar campañas de prevención enfocadas en la población más susceptible a contraer algún cáncer, según su edad, características y estrato, teniendo en cuenta que el cáncer es una enfermedad muy costosa, afectando no solo al paciente, sino a toda la familia a nivel psicológico y económico.

    Se puede identificar hasta el barrio que está presentando más casos. Al final se les podría ofrecer a los Oncólogos un sistema que se puede alimentar con todas las historias clínicas y que puedan tener todas esas variables que les puede ayudar a ellos a mejorar la investigación. Solo

necesitaríamos Oncólogos quienes pueden interpretar esa información y determinan que

variable tiene sentido, es decir, el Oncólogo se presenta en dos fases, al principio para decir qué variables necesita y al final para validar la información. Si sacamos este proyecto, necesitaríamos a un Oncólogo por una hora semanal, reuniones cortas y puntuales.

Debemos aprovechar que lo de la política de datos ya está en el CONPES y podemos presentar propuestas, además de mi investigación científica, quiero empezar a formar personas, crear empresas, fichas técnicas. Yo aquí les dejo la idea y empezaré a trabajar con Sonia sobre una propuesta a la Universidad del Valle.

Podríamos empezar a trabajar con el cáncer que está afectando más, como en España, que iniciaron con el cáncer de pulmón que es el que más afecta a la población, pero aquí en Colombia es otro y estos prototipos se los podemos presentar al MINSALUD.

17. Interviene el Ing. Carlos Ocampo y le solicita a Daniela que inicie averiguando cómo se pueden proteger es cuestión de las políticas de protección de datos frente a esto y paralelamente podríamos ir mirando con una de las ESES amigas, para plantearle la idea y ver de qué manera podríamos obtener información primaria.

18. Interviene Daniela y le informa al secretario que con ese insumo hay que crear un convenio, materializarlo y que quede organizado jurídicamente. La experiencia que hemos tenido es que antes de compartir información básica, tenga todos los requerimientos que pide la SIC para el manejo de información básica.

19. Interviene el Dr. Oswaldo Solarte informando que se puede ir adelantando como dice el secretario sobre en que nos podemos soportar para que los datos estén y sean legales. La clave aquí con los datos es que el único objetivo será investigativo, inclusive, en España hay historias clínicas sobre el COVID, que han puesto a disposición de todo el mundo, porque es la única forma de investigar, pero esto del COVID puede que el próximo ya esté resuelto, pero el cáncer no.

20. Interviene la Dra. Sonia e informa que la universidad proporcionará el Oncólogo y nosotros haríamos el acompañamiento de datos, o sea, que no pasaría por la secretaría de salud, serían la secretaria de las TIC´s, el Hospital Universitario y la Universidad del Valle, para que queden de una vez las ESES incorporadas. Es necesario que pase por jurídica.

21. Interviene el secretario y comunica mediante una llamada de celular en altavoz al Sr. Alfredo Cordoba quien escucha la explicación nuevamente del proyecto por parte del Dr. Oswaldo Solarte.

22. Responde el Sr. Alfredo me informa que tiene acceso a historias clínicas de nivel 1 y 2 del departamento del Valle del Cauca, no especializadas en oncología, tenemos mucha información para aportarles a ustedes en su estudio.

23. Interviene el Dr. Oswaldo Solarte, y aclara que con esa información de nivel 1 y 2, también es posible trabajar, poder determinar los problemas por año y fecha que más consultan los pacientes, perfilar la época de consulta y posteriormente, planear mejor los recursos; si nos pueden facilitar esas historias clínicas de nivel 1 y 2 también nos sirven a nivel de diagnóstico, no a nivel de tratamiento, pero si podríamos realizar un estudio descriptivo.

    Entonces quedo atento a cualquier decisión y me pueden contactar a través de la gobernación por medio de Sonia.

24. Interviene el secretario e informa que el Sr. Alfredo tiene 11 ESES y podría ser de gran ayuda.

25. Interviene el Dr. Oswaldo Solarte e informa que con la información de los hospitales nivel 1 y 2, se puede crear un proyecto, empezar a estructurar esa información y conocer a qué está yendo la gente a las ESES conocer por temporadas climáticas, perfiles etc., es decir, si yo conozco a qué vinieron el año anterior, el director de esa ESE se puede preparar y planear mejor sus recursos para este año o el próximo.

    Con esta información no pueden hacer lo del cáncer porque es más complejo y necesitaría información de hospitales nivel 4 y 5, pero con este podrían hacer otro prototipo.

26. Interviene el secretario e informa que hay 2 cosas pendientes, el tema de Daniela y lo otro interno, ya le pasé el contacto a Sonia de Alfredo con quien ya tenemos luz verde y lo otro a nivel técnico, ir preparando para acondicionar el ambiente.

27. El Dr. Oswaldo Solarte informa que, si van a trabajar más adelante con datos de cáncer, es necesario trabajar con el Hospital Universitario, y empezar a trabajar con los recursos; en Europa, para este proyecto de cáncer el presupuesto era de 4 millones de euros.

Otro punto, es la cantidad de empleos que este proyecto puede llegar a generar, ya que la I.A, es la electricidad de hace 100 años, esto apenas está

empezando y será positivo para el Valle. En la Universidad del Valle ya montamos una maestría en ciencia de datos y ha tenido buena acogida.

Le dejo estos datos de empresas en España llamadas IOMED y Sabana dedicadas al análisis de datos, estos son los tipos de empresas que se pueden crear aquí, generando empleos; los propietarios de esas empresas son médicos, quienes vieron el potencial en la I.A y decidieron invertir.

Da las gracias por la atención recibida, y explica que siempre está en contacto con la doctora Sonia, y pueden seguir avanzando en la parte legal, hacer los contactos con el Hospital Universitario, con la secretaría de salud para obtener las historias clínicas d pacientes de cáncer, si no se pueden con ese tipo de pacientes de cáncer, trabajar con el señor que hablamos ahora. Se podría crear un prototipo en 6 meses con la información de hospitales nivel 1 y 2. En el momento no tengo ninguna información para iniciar.

El 80 % data a nivel mundial no está estructurada, son datos escritos en lenguaje natural y el 20 % restantes está en bases de datos estructurales, solo pocas compañías como Google, Facebook y Twitter están explotando esa información.

En dos años regresará a Colombia, y empiezo a trabajar con la facultad de salud de la Universidad del Valle.

28. Interviene la Dra. Sonia y pregunta si alguien tiene algo más para aportar, pero nadie interviene.

29. Finalmente, se despide la presentadora y participantes.

Atentamente,

Carlos Hernan Ocampo
Secretario TIC
(se anexan imágenes)

Sonia Castro
Asesora

Liliana Plazas
Líder Economía Digital

Transcriptor: Mariela Ivonne Sinisterra Muñoz- Técnico

Archívese en:

GOBERNACIÓN VALLE DEL CAUCA

REGISTRO DE ASISTENCIA (CAPACITACIÓN, INDUCCIÓN, REIDUCCIÓN, ASESORÍA, ASISTENCIA TÉCNICA, REUNIONES, CONSEJO DE GOBIERNO, COMITÉ TECNICO DE LA DEPENDENCIA, EQUIPO DE MEJORA CONTINUA DEL PROCESO)

Código: FO.SM.01.18
Versión: 01
Fecha de Aprobación: 15/04/2015
Página: 1 de

LUGAR:

No. DE ACTA: HORA DE INICIO: 4:00 HORA DE TERMINACIÓN:

TOTAL HORAS:

FACILITADOR (ES)/RESPONSABLE: Sonia Y. Castro - Acta 155

CÓDIGO DEL PROCESO/SUBPROCESO:

NOMBRE DEL EVENTO/TEMA DE REUNIÓN/TEMAS A TRATAR: Convenio con Univalle para Hablar sobre Proy. Cáncer.

FECHA: 19-10-2020

No.	DEPENDENCIA / ENTIDAD/MPIO	NOMBRES Y APELLIDOS COMPLETOS	CARGO	CÉDULA	No. DE CELULAR - TEL/EXT.	CORREO ELECTRÓNICO	FIRMA DE ASISTENCIA
1	Setic	Sonia Castro	Asesor.	66921377	2300	scastro@valle.	
2	Setic	Liliana Plaz	Líder Proyec	31711613	2372	lmplaza@Valle.	
3	Univalle	Oswaldo S.	Profesor	94 V5116	4357	oswaldo.solarte@	
4	SETIC	Carlos Ocampo	Secretario	94431700		[redacted] chocampo@valledelcauca.gov.co	
5							
6							
7							
8							
9							
10							
11							
12							
13							