DúoDinámico
Ariel Santiago Tovar – 202013754
Juan Felipe Sinisterra - 201921647

ANALÍTICA COMPUTACIONAL PARA LA TOMA DE DECISIONES

PROYECTO 1 – PRODUCTIVIDAD EN MANUFACTURA

1. Descripción general

Se ha seleccionado como usuario final el área de producción de la empresa, interesada especialmente en la productividad y tasas de producción.

2. Roles

Ariel Santiago Tovar: Ingeniería de datos, análisis de datos y tablero de datos.

Juan Felipe Sinisterra: Ciencia de datos, análisis de negocio y despliegue.

3. Preguntas de negocio y plan de acción

Tarea 1

Para el desarrollo de la tarea 1 se consideró el área de producción de la empresa, considerando la perspectiva del área de producción. A partir de esto, se considera importante explorar la productividad de la empresa considerando diferentes factores o variables propuestas. Se consideran 2 instancias de productividad, la productividad objetivo y la productividad real. Considerando esto, se buscó realizar modelos que comprendieran y pudieran llegar a predecir el comportamiento de estas productividades.

Se propone realizar los modelos y a partir de estos ofrecer una tablero interactivo y agradable la cual permita cambiar valores de parámetros pertinentes y visualizar resultados de interés.

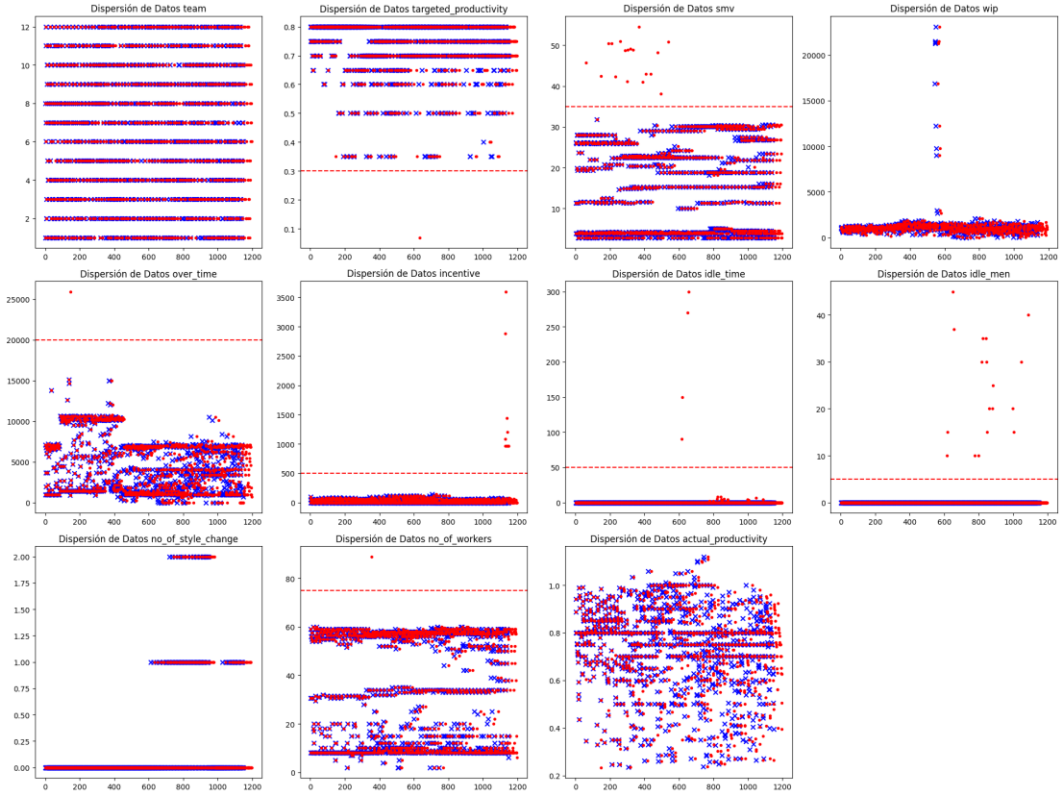
4. Datos

Tarea 2 – Limpieza y alistamiento de datos

Para la limpieza y alistamiento de datos se importaron estos y se empezó conociendo la forma y el tipo de datos que se encuentran. Para iniciar la exploración, se creó un DataFrame, encontrando que los datos estaban compuestos por 15 columnas de datos y 1197 entradas. Entre los tipos de las entradas, había variables tanto numéricas, como fechas y strings. Después de esto, se notó que había un valor nulo en los primeros 5 datos para la columna WIP. Al revisar en la totalidad del DataFrame de datos, se encontró que en total había 506 valores nulos únicamente para la columna WIP.

Antes de corregir los valores nulos encontrados, se decidió iniciar la limpieza de los datos disponibles que fuesen numéricos. Para esta limpieza, se realizaron gráficas de dispersión para cada una de las variables con el fin de poder identificar y seguidamente eliminar valores atípicos en cada una de las columnas (se eliminaba la fila del dato atípico). Al realizar este procedimiento, fue posible realizar un gráfico de dispersión comparativo en el que se ve fácilmente los datos que fueron eliminados.

También, en el siguiente gráfico es posible encontrar los datos originales en rojo y los datos limpios en azul, los cuales ayudan a entender desde qué valor se decidió que los datos eran atípicos y por ende el valor referencia para eliminar estos valores.



Ahora, con respecto a los valores nulos del WIP, se pensaron dos estrategias: eliminar la columna de datos o estimar el WIP a partir de las demás variables explicativas (todas las variables menos "targeted_productivity" y "actual_productivity") haciendo uso de un modelo de regresión lineal.

Al implementar la primera estrategia, se obtuvo un nuevo set de datos que se llamó *dflimpio*. Luego, se desarrolló la segunda estrategia, encontrando los coeficientes del modelo de regresión lineal y seguidamente llenando los valores nulos de la columna WIP con el resultado estimado haciendo uso de los coeficientes ya mencionados. A este set de datos se le llamó *dfarreglado* y se verificó que los valores nulos para la columna WIP pasara de 496 a 0. Por otro lado, al realizar la estimación del WIP y filtrar por valores atípicos, se encontró que en este set de datos se tienen 1140 entradas en 15 columnas.

Mientras se realizaba este proceso, se encontró que, para la columna del departamento había dos posibles valores: "finishing" y "sweing". Sin embargo, se encontró un error en algunos de los valores que tomaba "finishing" pues en varias entradas, en vez de encontrar "finishing", se encontraba "finishing " (con espacio al final). Esto se corrigió para ambos sets de datos con el fin de que esto no tuviese repercusiones en un futuro.

Tarea 3 – Exploración de datos

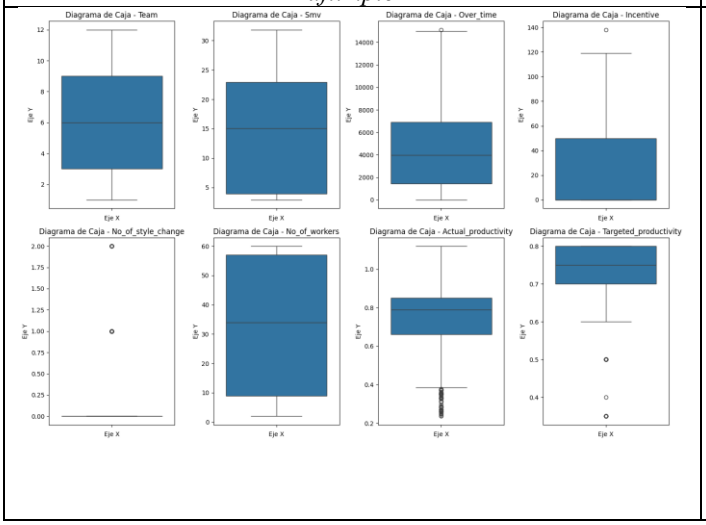
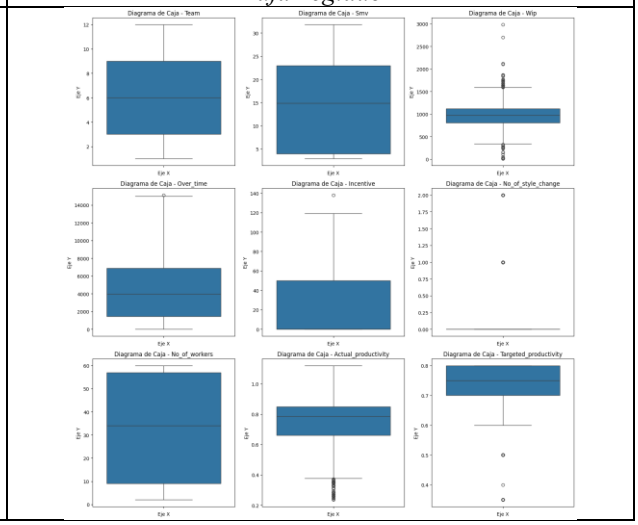
Teniendo en cuenta que a partir de la tarea anterior se obtuvieron 2 grupo de datos (*dflimpio y dfarreglado*), se realizó un análisis de exploración de datos para cada uno de estos. Esto debido a que, en este punto, no se sabe cuál de los dos grupos de datos es con el que es más conveniente trabajar.

Primeramente, se encontró que *dflimpio* cuenta con 1148 entradas y *dfarreglado* con 1140. Por otro lado, como se sabe el *dfarreglado* contiene una columna más (15 columnas en total), correspondiente al WIP. Sabiendo la forma de los sets de datos, se pudieron calcular unas primeras estadísticas descriptivas para las variables numéricas para cada uno de los grupos de datos obteniendo que las columnas idle_time e idle_men no tenían valores máximos, mínimos, promedio, etc… Al eliminar estas columnas se volvieron a obtener las siguientes estadísticas descriptivas:

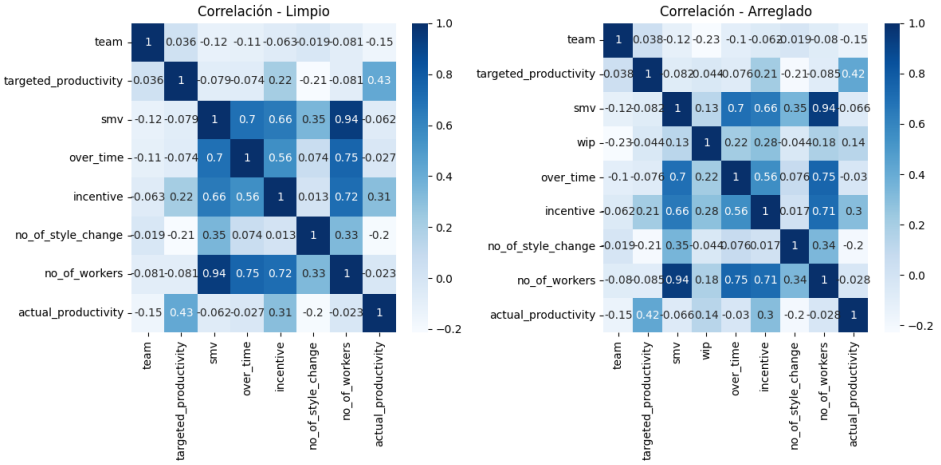
dflimpio								
Estadística	team	targeted_productivity	smv	over_time	incentive	no_of_style_change	no_of_workers	actual_productivity
# Datos	1148,000	1148,000	1148,000	1148,000	1148,000	1148,000	1148,000	1148,000
Promedio	6,414	0,731	14,488	4528,162	25,980	0,146	34,051	0,742
Desviación Estándar	3,482	0,095	10,245	3261,981	30,631	0,421	22,116	0,170
Mínimo	1,000	0,350	2,900	0,000	0,000	0,000	2,000	0,236
25%	3,000	0,700	3,940	1440,000	0,000	0,000	9,000	0,662
50%	6,000	0,750	15,090	3960,000	0,000	0,000	34,000	0,789
75%	9,000	0,800	22,940	6900,000	50,000	0,000	57,000	0,850
Máximo	12,000	0,800	31,830	15120,000	138,000	2,000	60,000	1,120

dfarreglado									
Estadística	team	targeted_productivity	smv	wip	over_time	incentive	no_of_style_change	no_of_workers	actual_productivity
# Datos	1140,000	1140,000	1140,000	1140,000	1140,000	1140,000	1140,000	1140,000	1140,000
Promedio	6,414	0,731	14,434	984,721	4515,570	25,647	0,147	33,910	0,741
Desviación Estándar	3,476	0,095	10,256	295,568	3268,675	30,433	0,422	22,120	0,170
Mínimo	1,000	0,350	2,900	10,000	0,000	0,000	0,000	2,000	0,236
25%	3,000	0,700	3,940	808,211	1440,000	0,000	0,000	9,000	0,660
50%	6,000	0,750	14,890	974,003	3960,000	0,000	0,000	34,000	0,787
75%	9,000	0,800	22,940	1122,919	6900,000	50,000	0,000	57,000	0,850
Máximo	12,000	0,800	31,830	2984,000	15120,000	138,000	2,000	60,000	1,120

Con el fin de visualizar las estadísticas descriptivas presentadas en la tabla anterior, se realiza un boxplot para las variables tanto para el set limpio como el arreglado respectivamente:

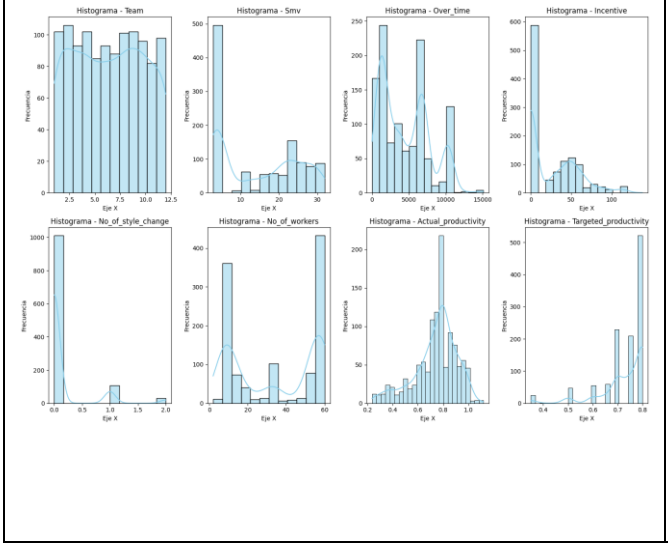
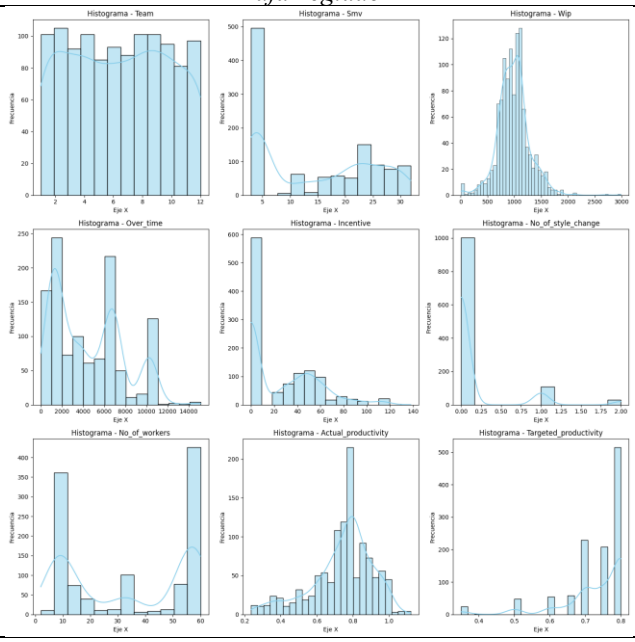
dflimpio	*dfarreglado*
	

Una vez obtenidas las estadísticas presentadas, fue de importancia encontrar la correlación entre variables pues esto puede ser de interés para la sección de modelos. Por esto, se realizaron los siguientes gráficos que muestras la correlación entre las variables para cada grupo de datos:



A partir de estos gráficos es posible encontrar que existe una gran correlación entre la variable no_of_workers con smv, over_time e incentive. Esto es realmente importante para tener en cuenta pues al realizar los modelos, puede que esto anterior genere algún tipo de sesgo o problema.

Por último, se realizó un histograma con el fin de poder representar la distribución de frecuencia de los datos por variable explicativa. Tanto para el grupo limpio y arreglado se obtuvieron los siguientes gráficos:

dflimpio	*dfarreglado*
	

A partir de la exploración de datos, se pudo determinar la naturaleza de las variables, así como sus valores máximos, mínimos, promedio y desviación estándar para cada conjunto de datos. Además, se encontró la frecuencia de estos datos, lo cual se relaciona con las mismas estadísticas descriptivas encontradas. Se considera que esta exploración de datos es realmente valiosa, ya que permite tener una comprensión más clara de la información proporcionada por cada una de las variables explicativas. Este entendimiento es clave para los pasos siguientes, especialmente para el desarrollo de modelos.

5. Modelos

Tarea 4 – Modelamiento

Para el modelamiento se exploraron los dos DataFrames, tanto *dflimpio y dfarreglado.* Se utilizo statsmodels.formula.api como librería para los modelos. También, se utilizó una significancia del 5% para evaluar todas las variables y modelos. Se realizo primero una etapa exploratoria para decidir qué DataFrame utilizar. Primero los modelos de la productividad actual.

Primero se utilizó el *dfarreglado*. Se estimo un modelo con todas las variables, poniendo la fecha, el quarter, el departamento y el equipo como variables categóricas. Por más de que el equipo tome valores numéricos, para todos los modelos se consideró una variable categórica debido a que cada equipo puede ser su categoría ya que cada equipo puede ser diferente y comportarse diferentemente. Adicionalmente se consideraron el resto de las variables (smv, incentivo, wip, over_time, idle_time, idle_men, no_of_style_change y no_of_workers). En el primer modelo se obtuvo que el intercepto, el departamento, el equipo, el smv, el incentivo, el over_time y el no_of_workers eran variables significativas. Se obtuvo un R^2 de 0.356 pero más importante un R^2_{adj} de 0.310. Con esto, se decidió eliminar todas las variables no significativas.

En el segundo modelo se eliminaron todas estas variables, se obtuvo un modelo donde la productividad actual seguía el departamento, el equipo, el smv, el over_time y el no_of_workers. Este modelo tiene un R^2 de 0.307 y un R^2_{adj} de 0.297. Para confirmar que este modelo reducido era mejor se realizó una prueba F-Parcial, donde la hipótesis nula afirma que el modelo reducido es mejor, mientras que la hipótesis alterna afirma que el modelo completo es mejor. Se obtuvo un p-valor de 0.9569, de tal manera que se puede aceptar la hipótesis nula, confirmando que es mejor el modelo reducido.

Además, se realizó un modelo para la productividad objetivo, considerando las mismas variables que se usaron al modelar la productividad real. En este se obtuvo un R^2 de 0.201 y un R^2_{adj} de 0.190.

Después de esto se tomaron en cuenta los datos del *dflimpio*. Se siguió un procedimiento similar, en el primer modelo, siendo este el completo se obtuvo un R^2 de 0.357 y un R^2_{adj} de 0.312. Se quitaron las mismas variables que con los datos que incluían el WIP, debido a que estas mismas no fueron significativas. El siguiente modelo tuvo un R^2 de 0.310 y un R^2_{adj} de 0.300. Se realizo la respectiva F-parcial para comprobar estadísticamente que modelo era mejor y se obtuvo un p-valor de 0.9536, significando que el modelo reducido era mejor para describir la productividad actual. De igual manera se hizo un modelo para el targeted productivity, con un R^2 de 0.202 y un R^2_{adj} de 0.191.

Se decidió utilizar *dflimpio*, los modelos realizados con esta arrojaron mayor R^2 y R^2_{adj}. Esto se debe a que tiene más observaciones, porque unos outliers del WIP se eliminaron y tenían aún menos datos.

Se continuo con la modelación, se hizo un split 80-20 para datos de entrenamiento y datos de prueba. Se continuo con el modelo previamente propuesto, donde la productividad actual siguiendo el departamento, el equipo, el smv, el over_time y el no_of_workers. Utilizando los datos de entrenamiento se obtuvo un modelo con un R^2 de 0.315 y un R^2_{adj} de 0.302. En la salida del modelo se menciona la existencia de posibles problemas de multicolinealidad. Entonces se calcularon los VIFS del modelo reducido. Se obtuvo que no_of_workers tenía el mayor VIF, teniendo un valor de 32.17 y smv también presentaba problemas, con un VIF de 24.34. Para corregir esto se decidió eliminar la variable no_of_workers. Obteniendo un modelo con un R^2 de 0.308 y un R^2_{adj} de 0.296. Al realizar esto se obtuvo que el smv tenía un p-valor de 0.365, mayor que la significancia considerada y por ende se terminó eliminando.

En el modelo en el que se eliminó el smv, se encontró un R^2 de 0.307 y un R^2_{adj} de 0.296, de igual manera para comprobar cual modelo era mejor, se realizó una F-parcial y con un p-valor de 0.6355 se afirma que el modelo reducido, que no incluye smv, es mejor. De este modelo igualmente se revisó multicolinealidad, en ambos casos los VIFS fueron menores a 2.5.

Finalmente, esto resulto en un modelo en el cual el actual productivity se modela a través del departamento, los equipos, que son variables categóricas y el over_time e incentivo. Con las mismas variables del actual productivity se modelo la targeted productivity.

En una revisión bibliográfica, se encontraron diversas propuestas de aspectos que afectan la productividad. Algunos de los que se mencionan son el ambiente de trabajo, la cultura del lugar, la carga de trabajo, el entrenamiento y desarrollo, el liderazgo, la comunicación y los incentivos [1]. Se encuentra que los incentivos pueden ser positivos y estimular la productividad mientras que la alta carga de trabajo o el overtime puede ser un aspecto negativo en la productividad de un trabajador. Una variable que se elimino fue la de número de trabajadores, y más allá de que tuviera multicolinealidad, fue pertinente removerla, debido a que el número de trabajadores puede llegar a un punto en el que es improductivo seguir agregándolos, debido a que pueden empezar a atravesarse entre ellos y a limitar su espacio de trabajo y puede afectar la productividad. Esto se justifica con la ley de rendimientos marginales decrecientes, donde se explica que un aumento en la fuerza laboral continuo puede llevar a una disminución en producción [2]. Debido a esto, también se considera adecuado y ventajoso eliminar dicha variable por el patrón real que puede tener su comportamiento.

Para el targeted productivity, se hizo un modelo con las mismas variables que el actual productivity, este tiene un R^2 de 0.209 y un R^2_{adj} de 0.196. Finalmente, se obtienen los siguientes betas para los dos modelos:

	Actual	Target
Intercepto	0.794920	0.731224
Departamento (Sewing)	-0.185258	-0.089341
Team 2	-0.018902	0.012373
Team 3	-0.003449	0.014434
Team 4	0.004133	-0.002691
Team 5	-0.032679	-0.02166
Team 6	-0.062889	0.025077
Team 7	-0.054918	0.016419
Team 8	-0.051835	0.006372
Team 9	-0.051450	0.038684
Team 10	-0.045606	0.024695
Team 11	-0.061901	0.002289

Team 12	-0.021129	0.045522
over_time	-0.000005	-0.000003
incentive	0.004122	0.001895

Cabe aclarar que estos modelos son toman como base el equipo 1 en finishing, por consecuente el intercepto incluye esto. Por ende, el beta del departamento (sewing) es el efecto adicional por que el equipo este en sewing y no en finishing. Los betas referentes a los equipos, es el efecto adicional que se incurre por utilizar dicho equipo. Otros aspectos generales son que el incentivo tiene un beta positivo, teniendo un efecto positivo en la productividad, mientras que el over_time tiene un efecto negativo. Se observa que en la productividad actual la mayoría de los equipos son menos productivos que el equipo 1, siendo el equipo 4 la excepción. En el modelo para el target productivity, se observa que el beta de la mayoría de los equipos es positivo, lo que puede significar que a estos equipos les proponen un objetivo de productividad mayor, pero mantienen una productividad menor que el equipo 1. Se sacaron medidas de error (MAE, MSE, RMSE y MAPE) con los datos de prueba. Se obtuvo lo siguiente:

	MAE	MSE	RMSE	MAPE
Actual_productivity	0.0991	0.0198	0.1409	17.42%
Targeted_productivity	0.0557	0.0063	0.0793	8.68%

Se observa que la mayoría de las medidas de error tienen un valor pequeño, lo cual es muy favorable dado que da a entender que los modelos desarrollados si son pertinentes. La medida que se considera como la más importante es el MAPE, dado que nos da la mejor percepción de nuestros errores, en el caso del targeted productivity el MAPE es adecuadamente bajo, mientras que en el actual productivity es un poco más alto, pero sin ser una situación problemática o preocupante.

En conclusión, sobre la modelización, se desarrollaron modelos que permitieron conocer el comportamiento de la productividad actual y la productividad objetivo, fue posible explorar diferentes variables y realizar un modelo el cual incluyera variables pertinentes y permitiera evaluar dicha productividad de la manera más objetiva posible.

6. Producto

Tarea 5 – Diseño y desarrollo del tablero

Para el diseño del tablero en Dash, primeramente, se pensó en que la disposición de este será de manera vertical en el navegador, de acuerdo con los contenidos de este. En orden de inicio a fin, los contenidos son los siguientes: título del tablero, objetivo del producto, instrucciones de uso, inicio de la aplicación (ingreso de los parámetros del modelo), resultados preliminares y finales de la aplicación.

Conociendo la disposición del tablero, se pensó en que el objetivo del proyecto en el Dash fuese realmente claro y aporte al usuario a entender para qué sirve el producto. En cuanto a las instrucciones del producto, estas se muestran seguidamente del objetivo y son realmente claras en cuanto a lo que el usuario debe realizar para poner en funcionamiento el producto. También, las instrucciones dan al usuario la información sobre los posibles valores que puede tomar los inputs del producto.

Al tener claras todo lo preliminar al funcionamiento del producto, es importante determinar qué valores son los que se le permitirá al usuario ingresar y qué resultados se generarán incluyendo visualizaciones importantes. Para todo esto anterior, es vital tener un buen conocimiento de todo lo realizado en la sección de modelos.

Teniendo en cuenta el modelo final obtenido para la productividad real y la objetivo, se pensó que los valores que el usuario va a poder ingresar son los relacionados a las variables explicativas de los modelos. Es interesante encontrar que ambos modelos finales encontrados comparten las mismas variables explicativas y por consiguiente coeficientes de modelo. Teniendo en cuenta lo anterior, los parámetros de entrada del tablero son el valor del incentivo, tiempo extra, equipo y departamento.

En la sección de resultados preliminares, hemos optado por presentar de manera gráfica los efectos que cada uno de los valores tiene en la productividad. Esto se realiza para cada modelo, lo que nos brinda la posibilidad de crear una comparación gráfica del efecto de las variables de entrada tanto en la productividad real como en la objetiva.

Por último, se tienen los resultados del producto, los cuales son la estimación que se tienen para la productividad real como la objetivo a partir de los valores ingresados como inputs del producto.

Es interesante encontrar que los valores de respuesta se actualizan inmediatamente ante cualquier cambio en los parámetros. Esto anterior genera una mayor facilidad de uso y entendimiento de la herramienta pues su comprensión permite el desarrollo de análisis robustos.

En imágenes, el tablero realizado se puede visualizar de la siguiente manera:

Tarea 6 – Despliegue

Para el despliegue, se lanzó una instancia EC2 en AWS. En esta se subió el tablero de Dash y se corrió el tablero para permanecer disponible el enlace mediante se mantenga la instancia. El enlace que da acceso al tablero de Dash creado es el siguiente: http://3.214.39.41:8050/.

Por otro lado, el enlace del repositorio en el que se puede evidenciar el desarrollo detallado de cada una de las tareas según los requerimientos encontrados en el enunciado del proyecto es el siguiente: https://github.com/JuanSinisterra/Proyecto1_ACTD_ATJS

Referencias

[1] Stephanie, Peoplelogic, [En línea]. Available: https://peoplelogic.ai/blog/factors-that-affect-employee-productivity.

[2] Economipedia, «Ley de rendimientos decrecientes: Qué es y ejemplos,» economipedia, [En línea]. Available: https://economipedia.com/definiciones/ley-de-rendimientos-decrecientes.html.

DúoDinámico
Ariel Santiago Tovar – 202013754
Juan Felipe Sinisterra - 201921647