Trabajo Práctico 1: Business Case

Integrantes: Fridman, Axel Hsueh, Noé Salas, Héctor.

[**LINK**](https://www.overleaf.com/9917288324cjxmmyckvqrc) del overleaf para la entrega final

# CONSIGNA TP 1: Business Case

En este trabajo práctico se espera la confección de un caso de negocio. El mismo incluirá motivaciones, explicación del dominio, formulación inicial de hipótesis, costos, riesgos, y deberá incluir algunas exploraciones iniciales que motiven y sustenten la viabilidad del trabajo futuro sobre el dominio elegido.

El trabajo es grupal, en **grupos de 3 personas**.

Se deberá entregar el mismo **impreso en papel y subir al campus**, en el documento se espera encontrar:

1. Descripción del caso de negocio (Máximo: 1/2 carilla):
   1. Producto / Problemática
   2. KPIs
   3. Métricas offline
2. Presentación del dataset elegido, incluir link de descarga al mismo (Máximo 1/2 carilla)
3. Análisis de viabilidad (costo, riesgos, tamaño de oportunidad)
   1. Presentar una tabla con más de 5 hipótesis priorizadas con las siguientes columnas:
      1. Hipótesis (máximo 140 caracteres)
      2. Experimento para validar/refutar (máximo 140 caracteres)
      3. Costo de hacer el experimento
      4. Impacto esperado de saber la respuesta
   2. Tomar 2 de las hipótesis anteriores, extender su descripción y realizar el experimento propuesto para luego incluir para cada una:
      1. Argumentos sobre importancia/relevancia de la misma (máximo 3 párrafos)
      2. Ilustrar graficamente aspectos a explorar/explorados (máximo 2 gráficos)
      3. Conclusiones preliminares y acciones a seguir
4. Link al Jupyter notebook con el código *limpio* de los 2 experimentos

Fecha límite de entrega: 14/09/2022 - 10:00 AM

**Motivación y explicación del dominio**

Un accidente cerebrovascular isquémico ocurre cuando se interrumpe o se reduce el suministro de sangre a una parte del cerebro, lo que impide que el tejido cerebral reciba oxígeno y nutrientes. \cite{MayoClinic}[[1]](#footnote-0) Cada año, casi 800.000 personas tienen un accidente cerebrovascular, más de 140.000 mueren y muchos sobrevivientes quedan con discapacidades. \cite{cdc}[[2]](#footnote-1) Un ACV es una emergencia médica cuyo tratamiento inmediato es crucial para reducir el daño cerebral y futuras complicaciones. Si bien los síntomas son altamente reconocibles, su identificación en el momento es difícil. En el caso de Argentina, la Cruz Roja cuenta con 50.000 personas capacitadas para la identificación del ACV, \cite{CruzRoja}[[3]](#footnote-2) el 0.11% de la población.

**Descripción del caso de negocio**

Una clínica importante quiere reducir las muertes de sus pacientes por ACV por medio de una campaña de concientización de sus pacientes de alto riesgo. Para lograr una temprana detección, debemos capacitar a las personas con mayor riesgo y a su entorno a identificar los síntomas más comunes. Eso nos dejaría con una incógnita, ¿quiénes son las personas con más riesgo de tener un ACV? En este caso, nuestro objetivo será identificar los usuarios con riesgo de ACV y proveerles una notificación temprana.

Nuestro KPI será la reducción de las muertes por accidentes cardiovasculares anuales en un 20% en un lapso de 5 años en nuestro hospital. La métrica offline que trataremos de mejorar y refinar será la F1 de nuestro modelo predictor de probabilidad de accidente. Claramente poder predecir de manera correcta nuestros pacientes de alto riesgo, nos permitirá tomar mayores recaudos y reducir las muertes; consecuentemente, esto impactará en nuestro KPI. Cabe destacar que para nuestro caso, consideramos que es más importante la detección de aquellos pacientes con riesgo de ACV frente a los que no tendrán ACV pese a que el modelo los considere como tal.

**Presentación del dataset**

El dataset elegido se encuentra en el siguiente [link](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) y cuenta con 5000 observaciones y diversas variables, algunas muy relacionadas a la salud y fisiología, como por ejemplo: si es hipertenso, edad, sexo, si tuvo alguna enfermedad del corazón, nivel de glucosa, e índice de masa corporal. Mientras que otras variables son más relacionadas a cuestiones sociales o de hábitos: estado civil, tipo de trabajo, tipo de residencia y si es fumadora o no lo es/ lo dejó.

No escapa nuestra atención, que si bien es posible encontrar que la residencia o el tipo de trabajo pueden correlacionarse bien con si tuvo o no un ACV, podría deberse a otro factor en común como pobreza o falta de acceso al sistema de salud. Es por eso que si bien puede que exploremos su relación con las otras variables, tenemos cautela en su valor explicativo.

Este dataset fue descargado 87.535 veces con lo cual ya fue sumamente estudiado y se lo considera de alta calidad para el desarrollo de un modelo simple.

**Análisis de viabilidad**

|  | Hipótesis | Experimento para validar/refutar | Costo de hacer el experimento | Impacto esperado de saber la respuesta |
| --- | --- | --- | --- | --- |
| **Experimento 1** | La edad es el mayor predictor de riesgo. | Implementar árboles de decisión y validar si es de las preguntas que ‘más arriba’ esta / reduce más la entropía. | Realizar este tipo de experimento tiene solo costo temporal. Estimamos 3 hs totales entre capacitarnos en las herramientas y tomar conclusiones. | Poder interpretar mejor nuestro modelo y darnos capacidad explicativa sobre por qué pondera el riesgo de una persona de tener un ACV de una manera u otra. |
| Experimento 2 | El ‘work\_type’ no es relevante. | Agregar la variable a un modelo ya implementado y observar que la métrica usada no mejora sustancialmente | Ídem experimento (1). | Ídem |
| Experimento 3 | El impacto del modelo asumiendo un modelo perfecto | Elaboración de un modelo perfecto asumiendo que se puede obtener todos los datos necesarios de los pacientes para que el modelo funcione | Acceso a los datos por parte del hospital, realización del modelo. | Capacidad de evitar en el país los 126 mil casos de ACV por año, de los cuales 18 mil terminan en muerte. [[4]](#footnote-3) |
| Experimento 4 | Fumar causa ACV | Establecer un modelo y estudiar su causalidad. | Capacitación en estadística para determinar causalidad entre variables. | ídem |
| **Experimento 5** | El género del paciente varía en importancia según la edad. | Dividir el dataset en 4 grupos etarios. 0-20, 20-40, 40-60, 60-80. Luego entrenar un modelo para cada dataset y medir la precisión con el data test de cada uno de los modelos, pero sin considerar la feature género. Luego explorar cómo varía la precisión del modelo al agregar esa feature en cada uno de mis datasets. | Realizar este tipo de experimento tiene solo costo temporal. Estimamos 3 hs totales entre capacitarnos en las herramientas y tomar conclusiones. | Determinar si el género es un buen predictor de riesgo. |
| Experimento 6 | Los pacientes están dispuestos recibir una notificación de posible ACV siendo este un posible falso positivo | Llamar a una cantidad entre 50 y 100 pacientes de un hospital preguntándoles si estarían dispuestos a recibir un llamado de advertencia de ACV sabiendo que nuestro modelo puede equivocarse. | Aproximadamente 5-6 hs de trabajo entre llamadas y documentación. | Nos permitirá saber cuán preciso deberíamos hacer nuestro modelo para no “sobreestimar” la cantidad de potenciales ACV y tener una viabilidad del proyecto. |

**Comprobación de Hipótesis:**

**Experimento 1:**

El árbol balanceado ofrece un accuracy del 92 %, lo que permite establecer la importancia de variables de una forma confiable (no sé exactamente cuál sería un umbral para decir esto, tal vez 70 %). Según la importancia de variables, obtenida tanto con el árbol como con el algoritmo Random Forest, la edad es el predictor más relevante para decidir si una persona tiene un factor de riesgo elevado de sufrir un ACV o no. Los gráficos también corroboran esta información, puesto que el riesgo varía según el rango etario: entre 0 y 20 años, no existe prácticamente ningún riesgo de sufrir un ACV, mientras que entre 20 y 40, se observa un incremento en este sentido y entre 60 y 80, se produce el mayor número de casos. Por lo tanto, hay suficiente evidencia para concluir que la edad es un factor de alta relevancia.

experimentar la viabilidad

Referencias links y etc:

Apellido, Nombre. *Título.* [Lugar de publicación] url [Tipo de texto]

* Ordenar pacientes según severidad (métrica sobre ranking)
* Tiempo de notificación
* Pensar en cómo encajaría el modelo en la realidad, sobre el proyecto
* Verificar los hipótesis con visualizaciones
* Complementar con datos de interne

1. Fuente:

   Mayo Clinic. \textit{Accidente Cardiovascular}.[Artículo] <https://www.mayoclinic.org/es-es/diseases-conditions/stroke/symptoms-causes/syc-20350113?utm_source=Google&utm_medium=abstract&utm_content=Stroke&utm_campaign=Knowledge-panel> [↑](#footnote-ref-0)
2. Fuente:

   Centers for Disease Control and Prevention [CDC]. \textit{Prevención de muertes por accidentes cerebrovasculares.}

   <https://www.cdc.gov/spanish/signosvitales/accidentes-cerebrovasculares/index.html#:~:text=Cada%20a%C3%B1o%2C%20casi%20800%20000,los%20accidentes%20cerebrovasculares%20son%20prevenibles>. [↑](#footnote-ref-1)
3. Fuente: Cruz Roja Argentina

   <https://www.cruzroja.org.ar/?utm_source=Google&utm_medium=Search&utm_campaign=General&gclid=CjwKCAjwsMGYBhAEEiwAGUXJaf-R79LwD9BZNeEDYXhPjg_shEReaOlDj-V6mz-4_DYByRZ-kdkXwRoC4kkQAvD_BwE>

   ¡Hola red! Les dejo un breve artículo sobre unos proyectos que como estudiantes estuvimos llevando a cabo en las clases de Modelado Continuo de la [Facultad de Ciencias Exactas y Naturales, UBA](https://www.linkedin.com/post/edit/6976255719079591936/#). Cuéntenme que les parece. [↑](#footnote-ref-2)
4. Chavez, Valeria. En la Argentina se produce un ACV cada 9 minutos: tres señales de alerta para detectarlo a tiempo. En Infobae <https://www.infobae.com/salud/2019/10/29/en-la-argentina-se-produce-un-acv-cada-9-minutos-tres-senales-de-alerta-para-detectarlo-a-tiempo/> [↑](#footnote-ref-3)