# Proyecto Individual 2

¡Bienvenidos al segundo proyecto! Durante estos días estarán poniendo en práctica sus habilidades en el campo de la predicción. Deberán usar cierta métrica para medir la performance del modelo la cual, a su vez, será usada para elegir los mejores modelos.

## Información relevante

Este proyecto es una instancia de evaluación, por lo cual es INDIVIDUAL y OBLIGATORIO para los alumnos de Data Science de Henry. Se disponibilizará un Google Form y pueden cargarse los resultados las veces que quieran. Es obligatorio que todos disponibilicen el código utilizado, para validar los modelos construidos.

## 🏥 Estancia hospitalaria 🏥

La hospitalización, o estancia hospitalaria, cuando es prolongada constituye una preocupación a nivel mundial debido a sus efectos negativos en el sistema de salud, aumentando los costos, generando

deficiencia en la accesibilidad de prestación de servicios de salud, saturación de unidades de hospitalización y urgencias, por consiguiente, mayores efectos adversos como lo son las enfermedades intrahospitalarias.

El estudio de los procesos de atención en salud, así como el conocimiento de las características y perfiles de los usuarios con el objetivo de predecir la ocupación hospitalaria, es uno de los aspectos al que las autoridades de salud han prestado gran interés, pues permite no sólo garantizar los recursos necesarios para la atención del paciente, sino realizar ajustes respecto a la oferta y demanda de los servicios de salud y los implementos asociados.

## Descripción del problema

Un importante Centro de Salud lo ha contratado con el fin de poder predecir si un paciente tendrá una estancia hospitalaria prolongada o no, utilizando la información contenida en el dataset asociado, la cual recaba una muestra histórica de sus pacientes, para poder administrar la demanda de camas en el hospital según la condición de los pacientes recientemente ingresados.

Para esto, se define que un paciente posee estancia hospitalaria prolongada si ha estado hospitalizado más de 8 días. Por lo que debe generar dicha variable categórica y luego categorizar los pacientes según las variables que usted considere necesarias, justificando dicha elección.

## Entrega

Deben tener el código en un script .py o Jupyter Notebook .ipynb, el cual debe incluir un buen EDA, feature engineerging y, de ser posible, un pipeline de Machine Learning para el procesamiento de datos que consideren necesario. Es importante **explicar claramente cada paso realizado** mediante comentarios en el script o textos formato markdown dentro del Notebook, pensar que cualquier persona (en este caso serán los Henry Mentors evaluadores) debe entender de la mejor manera posible cada razonamiento y pasos aplicados.

Recuerden, además, que deben enviar el repositorio que contenga el proyecto, por lo que es importante que le dediquen tiempo también a esta parte, dejando todo ordenado y con un README acorde, que sirva de introducción al contenido dentro de éste.

Por otro lado, es obligatorio que el script genere un archivo .csv sólo con las predicciones, teniendo únicamente **una sola columna** (sin index) que debe llamarse 'pred' y tenga todos los valores de las predicciones, con un valor por fila. De no llamarse así la **única columna**, nuestro script de validación **NO LO VA A TOMAR** y no aparecerán en el dashboard.

El nombre del archivo debe ser su usuario de GitHub, si su usuario de GitHub es 'pjr95', el archivo .csv con las predicciones debe llamarse 'pjr95.csv'. Vamos a validar tanto los datos que suban como el código, por lo que seguir estos pasos es fundamental.

Cuando entreguen les pedimos que verifiquen que su usuario de GitHub aparezca en el dashboard. En caso de que no aparezca, tal como se comentó más arriba, es debido a que el archivo entregado con las predicciones no cumple con los requisitos solicitados.

Recuerden que la columna objetivo **no está presente en el dataset**, deben crearla en base a la consigna. Recuerden verificar que el número de columnas del set de datos que utilizan para entrenar el modelo sea igual que el número de columnas que tiene el set de testeo, con el que harán las predicciones.

## Criterio de evaluación

Deberán crear un repositorio en GitHub con acceso público donde subirán la solución propuesta en un archivo de Jupyter Notebook (.ipynb) o bien un script de Python (.py) y el archivo de texto plano con valores separados por comas (.csv) con las predicciones. Dicho repositorio deben compartirlo en el formulario de entrega, junto al archivo .csv con las predicciones. Deberán escribir su propio readme, describiendo brevemente el problema y la solución que proponen (no debe ser una simple copia de la consigna del PI).

La solución propuesta debe incluir los siguientes ítems, por cada uno cumplido sumará 1 punto, siendo 1 la nota mínima y 5 la nota máxima:

- Entrenamiento y predicción utilizando un Modelo de Machine Learning adecuado al problema (clasificación o regresión).
- Análisis exploratorio de los datos (EDA).
- División de dataset en train y test utilizando train_test_split, CV, KFold o similares.
- Utilización de Pipelines en la producción del modelo.
- Comentarios y redacción con la fundamentación de la solución propuesta, escrita en Markdown en el Jupyter Notebook (.ipynb) o bien en un documento aparte.

## Métrica a utilizar

Como método de evaluación del desempeño del modelo, se utilizará la métrica de Exhaustividad (Recall) para las estadías hospitalarias largas, a partir de la matriz de confusión (Confusion Matrix).

$$ Recall=\frac{TP}{TP+FN}$$

Donde $TP$ son los verdaderos positivos y $FN$ los falsos negativos.

Como métrica adicional para verificar el desempeño de su modelo, también se utilizará la métrica de precisión (Accuracy) para las estadías hospitalarias largas.

$$ Accuracy=\frac{TP+TN}{P+N}$$

siendo $TP$ los verdaderos positivos, $TN$ verdaderos negativos y $P+N$ población total.

## Archivos provistos

Se proveen los siguientes archivos para realizar el proyecto:

- 'hospitalizaciones_train.csv': Contiene 410000 registros y 15 dimensiones, el cual incluye la información **numérica** de la cantidad de días de estancia hospitalaria.
- 'hospitalizaciones_test.csv': Contiene 90000 registros y 14 dimensiones, el cual no incluye la información de la cantidad de días de estancia hospitalaria.

## Descripción de las dimensiones

- Available Extra Rooms in Hospital: Habitaciones adicionales disponibles en el hospital. Una habitación no es igual a un paciente, pueden ser individuales o compartidas.
- Department: Área de atención a la que ingresa el paciente.
- Ward_Facility_Code: Código de la habitación del paciente.
- doctor_name: Nombre de el/la doctor/a a cargo del paciente.
- staff_available: Cantidad de personal disponible al momento del ingreso del paciente.
- patientid: Identificador del paciente.
- Age: Edad del paciente.
- gender: Género del paciente.
- Type of Admission: Tipo de ingreso registrado según la situación de ingreso del paciente.
- Severity of Illness: Gravedad de la enfermedad/condición/estado del paciente al momento del ingreso.
- health_conditions: Condiciones de salud del paciente.
- Visitors with Patient: Cantidad de visitantes registrados para el paciente.
- Insurance: Indica si la persona posee o no seguro de salud.
- Admission_Deposit: Pago realizado a nombre del paciente, con el fin de cubrir los costos iniciales de internación.
- Stay (in days): Días registrados de estancia hospitalaria.

## Sugerencias

- Exploren el dataset. Saquen medidas resumen, vean distribuciones de los datos, analicen bien el tipo de problema, etc.
- Piensen qué tipo de modelo podría ser aplicable según la descripción del problema y el tipo de variable de salida.
- Busquen información sobre la métrica aplicada, cada métrica tiene pros y contras.
- En cuanto a la utilización de GitHub, recuerden que si quieren hacer un cambio experimental pero no quieren romper el modelo, pueden utilizar branching.
- Aprovechen esta instancia de aprendizaje, experimenten y, sobre todo, ¡diviértanse!

## Disclaimer

De parte del equipo de Henry se quiere aclarar y remarcar que los fines de los proyectos propuestos son exclusivamente pedagógicos, con el objetivo de realizar proyectos que simulen un entorno laboral, en el cual se trabajen diversas temáticas ajustadas a la realidad. No reflejan necesariamente la filosofía y valores de la organización. Además, Henry no alienta ni tampoco recomienda a los alumnos y/o cualquier persona leyendo los repositorios (y entregas de proyectos) que tomen acciones en base a los datos que pudieran o no haber recabado. Toda la información expuesta y resultados obtenidos en los proyectos, nunca deben ser tomados en cuenta para la toma real de decisiones (especialmente en la temática de finanzas, salud, política, etc.).