# Instrucciones.

**Examen Parcial Electiva IV**

**Ing. MsC. Marco Javier Suarez Barón PhD.**

* 1. Para este proyecto, tendrá acceso a un dataset de entrenamiento con 36000 cuentas de clientes que aplicaron a un crédito y se sabe si pagaron o no dicho crédito.
  2. El dataset está compuesto por 22 variables predictoras (variables independientes) y 1 variable dependiente a predecir con el nombre de *defaulted*\_*loan*.
  3. El objetivo del proyecto es predecir si un cliente no terminará de pagar su préstamo (*defaulted*\_*loan* = *True*) o si terminará de pagar dicho préstamo (*defaulted*\_*loan* = *False*) en el dataset de evaluación el cual consiste de 24000 observaciones.
  4. Para el entrenamiento de su algoritmo de clasificación binaria, tendrá el siguiente **archivo de entrenamiento**:

train\_dataset.csv

* 1. Para la aplicación del modelo entrenado y la entrega del proyecto requerirá el siguiente **archivo de evaluación**:

test\_dataset.csv

* 1. Note que la columna *defaulted*\_*loan* solo existe en el archivo de entrenamiento, su tarea es generar esta columna para el dataset de evaluación con sus predicciones generadas a partir de entrenar con el dataset de entrenamiento.
  2. Note que la predicción será 1 para cuando *defaulted*\_*loan* = *True* y 0 para cuando *defaulted*\_*loan* = *False*. Es decir, las predicciones deberán quedar de la siguiente manera: *defaulted*\_*loan* = 1 para cuando sea *True* y *defaulted*\_*loan* = 0 para cuando sea *False*. Usted deberá hacer el mapeamiento de esta variable de un espacio [*False, True*] a un espacio [0*,* 1]. Recuerde hacer uso de las operaciones *if* y *else*.
  3. Las variables independientes poseen información acerca de la edad del cliente, si tiene registros criminales, número de hijos, dirección actual, entre otros.
  4. Para este proyecto será fundamental para su realización que emplee las técnicas de análisis exploratorio de datos. Tome en cuenta que algunas variables podrían proporcionar más información que otras. Se le recomienda la construcción de meta-variables que le puedan ayudar a predecir la variable de interés **evitando hacer uso de información externa**. Sólo deberá utilizar variables y derivados de las variables las cuales están dadas en el archivo de entrenamiento.
  5. Para el proyecto es libre de emplear los algoritmos de aprendizaje para clasificación binaria visto al momento; acompañado con validación cruzada.
  6. La evaluación y métrica de desempeño a utilizar para este proyecto consistirá en la métrica *F* 1\_*score*. La cual es mostrada a continuación:

*F* 1\_*score* = 2 *precision × recall*

*×*

*precision* + *recall*

(1)

* 1. *Precision* está dado por la siguiente ecuación en donde *tp* es el número de verdaderos positivos y *fp*

es el número de falsos positivos:

*precision* =

*tp*

*tp* + *fp*

(2)

* 1. *Recall* está dado por la siguiente ecuación en donde *tp* es el número de verdaderos positivos y *fn* es el número de falsos negativos:

*recall* =

*tp*

*tp* + *fn*

(3)

* 1. Mayor información sobre la métrica *F* 1\_*score*, visitar https://scikit-learn.org
  2. Investigar con más detalle sobre las ecuaciones de *precision* y *recall*.
  3. Aplique las métricas de selección de modelos de acuerdo a los resultados experimentales obtenidos por las curvas ROC y las curva lift y matriz de confusión.

### Se le recomienda trabajar en un algoritmo que sea capaz de hacer generalizaciones, para tener una calificación más alta en el proyecto. Por tanto se penalizarán soluciones que tengan sobre-ajuste o sub-ajuste.

* 1. **La calificación del proyecto únicamente tomará en cuenta el desempeño de sus predicciones en el archivo de evaluación.**
  2. Algunos recursos de referencia como [Kaggle](https://www.kaggle.com/benhamner/random-forest-benchmark-r/code) y [Github](https://amunategui.github.io/binary-outcome-modeling/)  le pudieran servir.
  3. Se le recomienda que analice y comprenda los talleres de *R* y *Python* que hicieron para poder completar este ejercicio.

# Septiembre 4 de 2020 antes de las 12 m

* 1. Cuando termine de predecir la columna de *defaulted*\_*loan* en el dataset de evaluación *test*\_*dataset.csv*, exporte el archivo en formato CSV donde solo vengan las siguientes dos columnas *application*\_*id* y *defaulted*\_*loan*].
  2. Para la exportación a CSV Python tiene una función *write.csv*.
  3. El archivo CSV será guardado con el siguiente formato de nombrado. Es necesario que ponga su nombre y primer apellido seguido por el día, el mes y el año (separado por guiones intermedios y usando el formato dd y mm cuando se trate de días o meses que solo tienen un dígito) seguido por:

Proyecto- 1-machine learning.csv

* 1. Por ejemplo, si su nombre es Alfonso Reyes Martínez y realizó el proyecto final el día 4 de septiembre de 2020. El archivo de entrega CSV tendrá el siguiente nombre:

Alfonso Reyes 04-09-2020 PF Proy final machine learning.csv

* 1. Se incluye un archivo ejemplo para que pueda ver como se debe entregar el archivo CSV. El archivo ejemplo tiene el nombre de:

Ejemplo PF desarrollo modelos pred.csv

## Ejemplo archivo de entrega del proyecto visto en MS Excel.

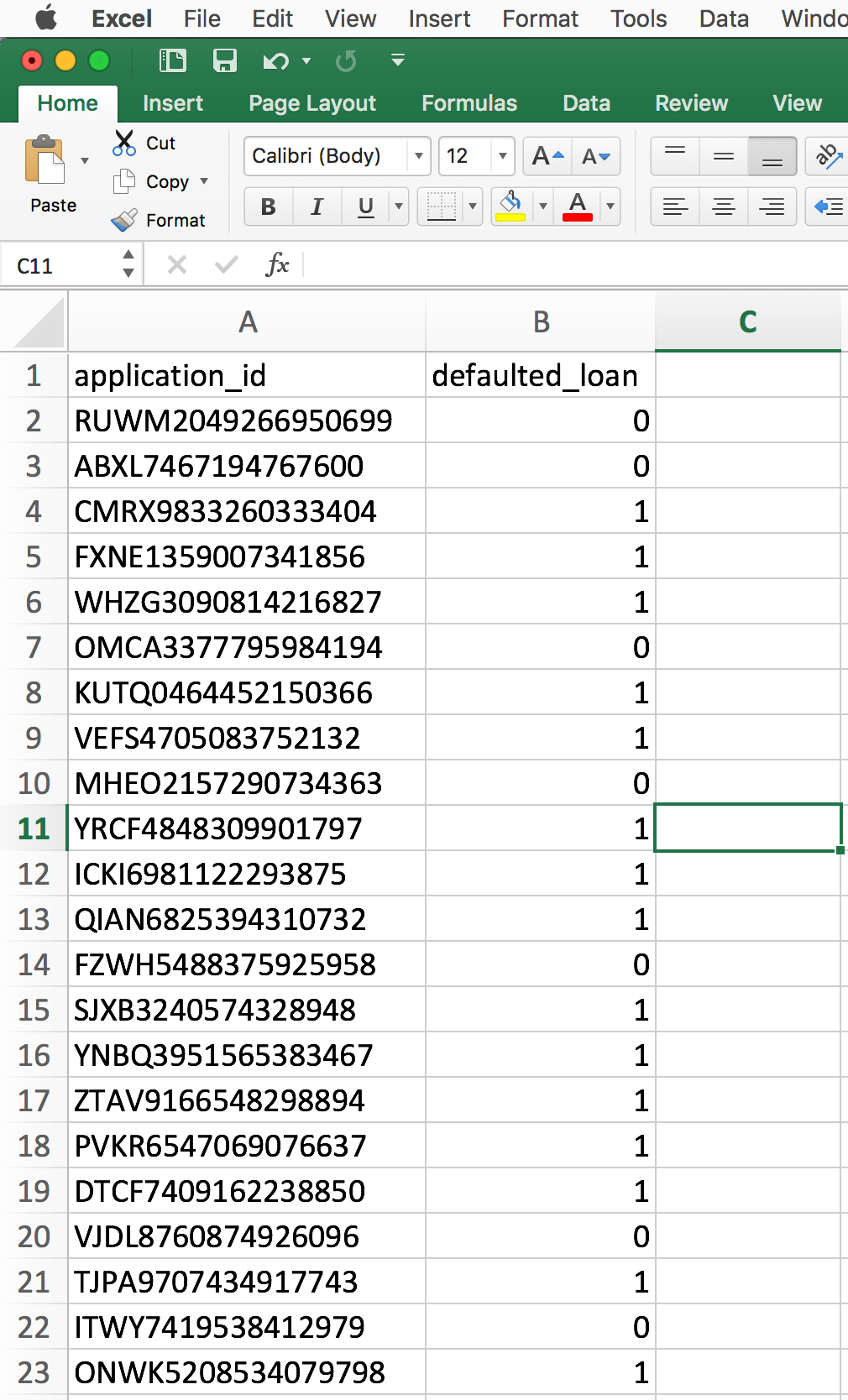


Figura 1: Como se muestra en el archivo ejemplo, deberá entregar un archivo CSV que contenga únicamente las columnas [*application*\_*id, defaulted*\_*loan*]. Note que las predicciones deberán ser *defaulted*\_*loan* = 1 para cuando su predicción sea *True* y *defaulted*\_*loan* = 0 para cuando su predicción sea *False*. 1 indica que su predicción es hacia que dicho cliente no pagará el crédito y 0 implica que su predicción es a que dicho cliente si pagará el crédito. Este archivo tendrá las dimensiones de 24001 filas (1 fila pertenece al nombre de las columnas y 24000 observaciones del dataset de evaluación) y 2 columnas.

## Ejemplo archivo de entrega del proyecto visto en un editor de textos.

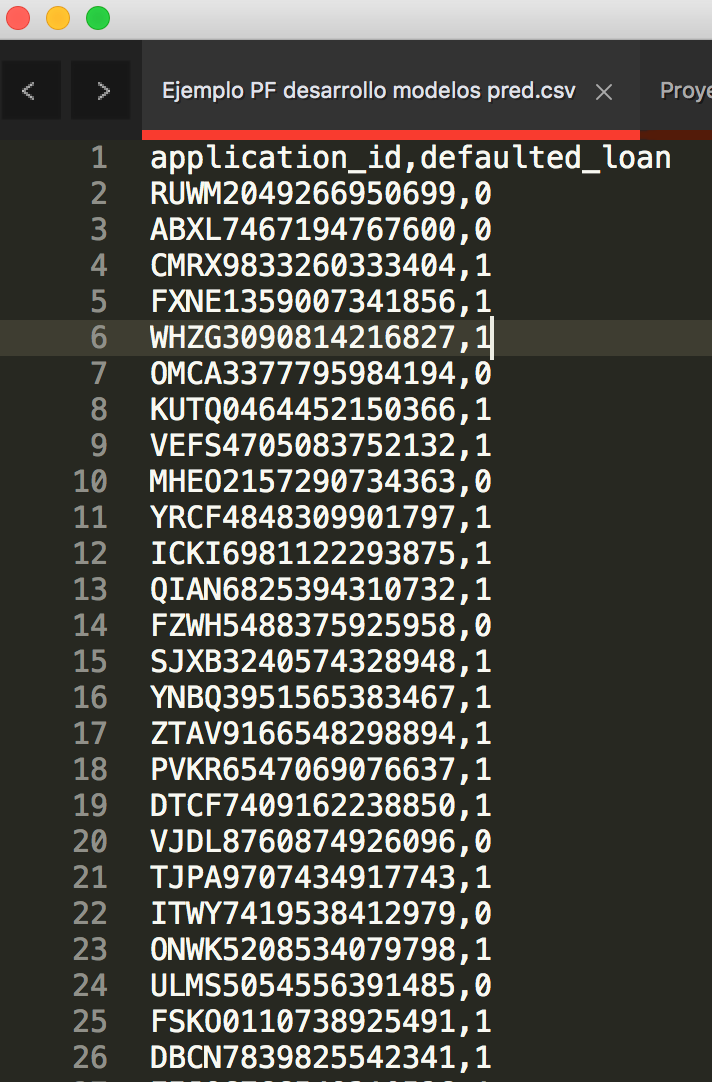


Figura 2: Como se muestra en el archivo ejemplo, deberá entregar un archivo CSV que contenga únicamente las columnas [*application*\_*id, defaulted*\_*loan*]. Note que las predicciones deberán ser *defaulted*\_*loan* = 1 para cuando su predicción sea *True* y *defaulted*\_*loan* = 0 para cuando su predicción sea *False*. 1 indica que su predicción es hacia que dicho cliente no pagará el crédito y 0 implica que su predicción es a que dicho cliente si pagará el crédito. Este archivo tendrá las dimensiones de 24001 filas (1 fila pertenece al nombre de las columnas y 24000 observaciones del dataset de evaluación) y 2 columnas.

# Evaluación.

* 1. Al hacer la entrega de su archivo de evaluación CSV que consiste en las predicciones de morosidad de las 24000 observaciones del dataset de evaluación, un algoritmo automático de evaluación hará un mapeo de sus predicciones con las etiquetas reales que solo el profesor tiene acceso.

### Asegúrese de respetar el formato indicado en la sección de entrega, de lo contrario tendrá una puntuación de 0 en el proyecto por no respetar el formato de entrega.

* 1. La métrica que se utilizará para evaluar sus predicciones con las etiquetas reales será la misma métrica que se le pide para que usted evalué el entrenamiento del modelo. La métrica *F* 1\_*score*.

**3.1. Ponderación de la evaluación.**

El rango de puntuación va de 0 a 100. Donde 0 es restringido cuando el archivo de entrega no respeta las instrucciones de entrega y no puede leerse correctamente por el sistema de evaluación del profesor. Para el caso donde el archivo de entrega respete las instrucciones, la puntuación del proyecto estará determinada por lo siguiente.

Si sus predicciones generan una métrica:

*F* 1\_*score ≥* 0*,*90, su calificación será de 100 puntos.

*F* 1\_*score ≥* 0*,*80 y *F* 1\_*score <* 0*,*90, su calificación del proyecto será de 95 puntos. *F* 1\_*score ≥* 0*,*70 y *F* 1\_*score <* 0*,*80, su calificación del proyecto será de 85 puntos. *F* 1\_*score ≥* 0*,*60 y *F* 1\_*score <* 0*,*70, su calificación del proyecto será de 75 puntos. *F* 1\_*score >* 0*,*50 y *F* 1\_*score <* 0*,*60, su calificación del proyecto será de 65 puntos.

*F* 1\_*score* 0*,*50, su calificación del proyecto será de 50 puntos. Dado que la predicción parece aleatoria. Daría lo mismo tirar una moneda para predecir morosidad en el dataset de evaluación.

*≤*

**3.2 Informe escrito**

Se debe entregar un archivo pdf. que contenga todo el análisis, observaciones, exploración, patrones y descubrimiento de conocimiento encontrado. La metodología, métodos y materiales son indispensables adicionalmente. El informe prioritariamente debe contener las acciones de soporte a toma de decisión que las áreas funcionales y operacionales requieran.