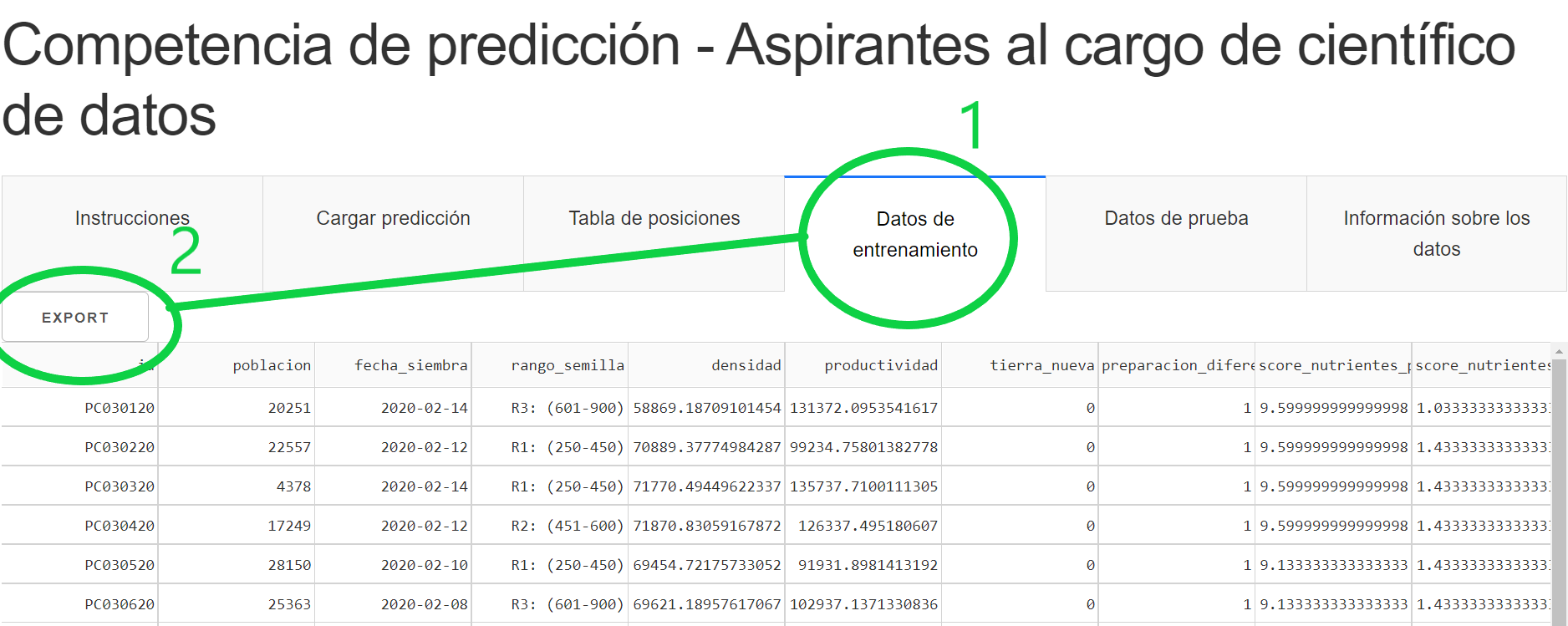
#### Objetivo

El objetivo de esta competencia es estimar un modelo de regresión con el dataset de entrenamiento y predecir la variable dependiente en el dataset de prueba. Los candidatos que logren las predicciones más precisas serán considerados para siguiente etapa de entrevista.

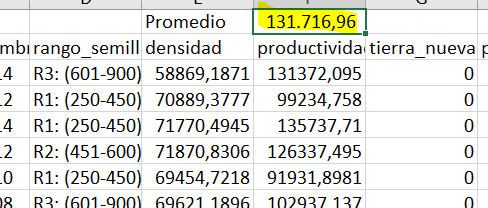
#### Procedimiento

Para empezar la competencia se recomienda descargar los datos de entrenamiento en la pestaña de este aplicativo destinada para eso:



**Asegúrese de que el dataset de entrenamiento tenga 984 registros**

En los datos de entrenamiento se encuentra la variable dependiente llamada **productividad** Las demás variables son los regresores que deben ser procesados para entender la relación entre ellos y la variable objetivo. La aproximación base para resolver un problema de regresión es predecir utilizando la media de la variable de interés en los datos de entrenamiento:



Partiendo de esta aproximación inicial, se va a predecir la productividad de todos los bloques del dataset de prueba con el valor constante de 131716,96 Los datos de prueba se pueden descargar igual que los de entrenamiento, pero en la pestaña de Datos de prueba



**Asegúrese de que el dataset de prueba tenga 80 registros**

Una vez desarrollado el modelo en los datos de entrenamiento, se debe aplicar en los datos de prueba para predecir la productividad cuyo valor real desconocen. Para el ejemplo planteado se muestra la siguiente tabla con dos columnas: id y productividad. El valor de productividad se llena con su valor promedio como se mencionó anteriormente

EXPORT

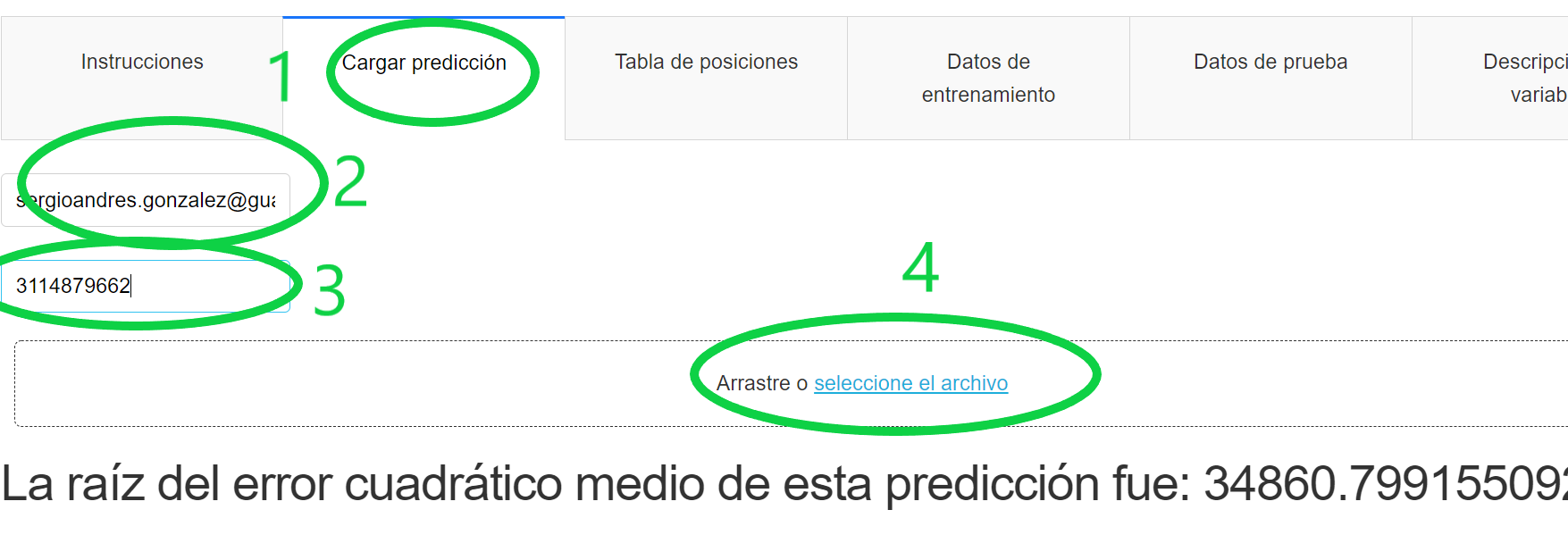
| id | productividad |
| --- | --- |
| PC010120 | 131716.96 |
| PC010320 | 131716.96 |
| PC010720 | 131716.96 |
| PC010820 | 131716.96 |
| PC010920 | 131716.96 |
| PC011020 | 131716.96 |
| PC011120 | 131716.96 |
| PC011220 | 131716.96 |
| PC011320 | 131716.96 |
| PC011420 | 131716.96 |
| PC011520 | 131716.96 |
| PC011620 | 131716.96 |
| PC011720 | 131716.96 |
| PC011820 | 131716.96 |
| PC011920 | 131716.96 |

1

/ 6

Al exportar esta tabla va a disponer de la estructura requrida para el cargue de predicciones: un archivo en formato xlsx de una sola hoja que contiene dos columnas: id y predicción. id es de tipo string y no puede eliminar o agregar registros, mientras que todas las predicciones deben estar en formato numérico.

Para realizar el cargue de una predicción, diríjase a la pestaña de cargar predicción ingrese su correo y contraseña informada vía email y cargue el archivo de predicciones en formato .xlsx



En caso de haber algún error en el cargue, el texto en la parte inferior le indicará que pudo haber sucedido. Si no hay ningún error, se realiza el cargue y se actualiza la información en la tabla de posiciones. Cualquier duda adicional será respondida por correo

#### Descripción de variables

- La variable productividad es la variable objetivo del problema y representa los kilos que se cosechan de una unidad de terreno identificada por su id

- La población corresponde a la cantidad de semillas sembradas por unidad de terreno

- Los rangos de semilla hacen referencia a los rangos de peso de las semillas sembradas

- Tierra nueva y preparación diferencial son marcas que caracterizan tratamientos especiales de la tierra

- Los scores de nutrientes, crecimiento, etc hacen referencia al número de aplicaciones realizadas de cada categoría ponderadas por un factor de oportunidad

- Las variables terminadas en max o min se refieren a porcentajes de enfermedades halladas en muestreos el valor min corresponde al primer valor registrado de la variable en la unidad de terreno, y el valor max corresponde al dato de muestreo más reciente

- Las variables de botones y meristemo corresponden al porcentaje de plantas que llegaron a cada uno de esos estados que se relacionan con una inducción para dar fruto correcta.

- Las variables mean\_pp y var\_pp representan la media y la varianza del muestreo de peso previo al proceso de inducción

- La edad de forza corresponde al número de meses transcurridos desde la siembra hasta la inducción de la planta

muchas gracias, ahora voy a realizar el trabajo de la etapa 3 "Evaluación del modelo" tal cual como me dices, entonces dame acompañamiento detallado paso a paso para lograr el mejor trabajo de mi vida, que algoritmo me recomiendas usar que pueda llegar a generar el mejor resultado, o explicame como hago para evaluar cada resultado , tambien dame un paso a paso de como se debe hacer la prueba, por ahora enfocate en darme el paso a paso detallado para latercera etapa "Evaluación del modelo",En este caso, sugiero evaluar los cinco algoritmos mencionados anteriormente: regresión lineal, regresión polinomial, regresión Ridge, regresión Lasso y regresión Elastic Net. una vez evaluados como hago para escoger el mejor resultado?

sin olvidar el objetivo:

El objetivo es estimar un modelo de regresión con el dataset de entrenamiento y predecir la variable dependiente en el dataset de prueba.

la estructura requrida para el cargue de predicciones: un archivo en formato xlsx de una sola hoja que contiene dos columnas: id y predicción. id es de tipo string y no puede eliminar o agregar registros, mientras que todas las predicciones deben estar en formato numérico.

Descripción de variables

- La variable productividad es la variable objetivo del problema y representa los kilos que se cosechan de una unidad de terreno identificada por su id

- La población corresponde a la cantidad de semillas sembradas por unidad de terreno

- Los rangos de semilla hacen referencia a los rangos de peso de las semillas sembradas

- Tierra nueva y preparación diferencial son marcas que caracterizan tratamientos especiales de la tierra

- Los scores de nutrientes, crecimiento, etc hacen referencia al número de aplicaciones realizadas de cada categoría ponderadas por un factor de oportunidad

- Las variables terminadas en max o min se refieren a porcentajes de enfermedades halladas en muestreos el valor min corresponde al primer valor registrado de la variable en la unidad de terreno, y el valor max corresponde al dato de muestreo más reciente

- Las variables de botones y meristemo corresponden al porcentaje de plantas que llegaron a cada uno de esos estados que se relacionan con una inducción para dar fruto correcta.

- Las variables mean\_pp y var\_pp representan la media y la varianza del muestreo de peso previo al proceso de inducción

- La edad de forza corresponde al número de meses transcurridos desde la siembra hasta la inducción de la planta

Predicción en conjunto de prueba

dame el paso a paso detallado y codigo de esto:

finalmente voy a realizar el trabajo de la etapa 4 Predicción en conjunto de prueba teniendo en cuenta que ya tenemos los pasos de Preprocesamiento de los datos, Selección del modelo y Evaluación del modelo este paso es especificamente para Predicción en conjunto de prueba, entonces dame acompañamiento detallado paso a paso,

* que algoritmo me recomiendas usar que pueda llegar a generar el mejor resultado?
* explicame como hago para obtener la mejor predicción del conjunto de prueba posible
* dame un paso a paso de como se debe hacer la prueba, por ahora enfocate en darme el paso a paso detallado para Predicción en conjunto de prueba,En este caso, como hago para escoger el mejor resultado?

sin olvidar el objetivo:

El objetivo es estimar un modelo de regresión con el dataset de entrenamiento y predecir la variable dependiente en el dataset de prueba.

la estructura requrida para el cargue de predicciones: un archivo en formato xlsx de una sola hoja que contiene dos columnas: id y predicción. id es de tipo string y no puede eliminar o agregar registros, mientras que todas las predicciones deben estar en formato numérico.

Descripción de variables

- La variable productividad es la variable objetivo del problema y representa los kilos que se cosechan de una unidad de terreno identificada por su id

- La población corresponde a la cantidad de semillas sembradas por unidad de terreno

- Los rangos de semilla hacen referencia a los rangos de peso de las semillas sembradas

- Tierra nueva y preparación diferencial son marcas que caracterizan tratamientos especiales de la tierra

- Los scores de nutrientes, crecimiento, etc hacen referencia al número de aplicaciones realizadas de cada categoría ponderadas por un factor de oportunidad

- Las variables terminadas en max o min se refieren a porcentajes de enfermedades halladas en muestreos el valor min corresponde al primer valor registrado de la variable en la unidad de terreno, y el valor max corresponde al dato de muestreo más reciente

- Las variables de botones y meristemo corresponden al porcentaje de plantas que llegaron a cada uno de esos estados que se relacionan con una inducción para dar fruto correcta.

- Las variables mean\_pp y var\_pp representan la media y la varianza del muestreo de peso previo al proceso de inducción

- La edad de forza corresponde al número de meses transcurridos desde la siembra hasta la inducción de la planta