

Proyecto Final

Fecha límite de entrega: Domingo 2 de Junio

Intr. Sistemas Inteligentes - 2024-2

Objetivo:

De manera individual desarrollar un modelo de clasificación que permita predecir el rendimiento de los estudiantes en las pruebas Saber Pro a partir de un conjunto de variables.

Las categorías de rendimientos se dividen en:

- Baja
- Media-baja
- Media-alta
- Alta

Descripción

El desarrollo del modelo se basará en la plataforma Kaggle (<https://www.kaggle.com>). A través de Campuswire se enviará el link de la competencia. Deben registrarse siguiendo las instrucciones al final de este documento. En esta plataforma encontrarán 3 archivos que necesitarán para el desarrollo del modelo.

- **train.csv:** Son los datos con los cuales deben **entrenar** su modelo, cada fila es un estudiante con su respectivo **ID**, e información académica y socioeconómica que les permitirá hacer la predicción, estos datos están etiquetados con la columna target que es **RENDIMIENTO_GLOBAL**.
- **test.csv:** Son los datos que su modelo debe **predecir**, al igual que los estudiantes de **train.csv**, se tienen el ID y la información de cada uno, sin embargo estos estudiantes no tienen la columna target(**RENDIMIENTO_GLOBAL**). Puesto que son ustedes quienes deben predecirlo.

- **submission.csv:** Es un ejemplo de el envío que debe hacer a la plataforma Kaggle, contiene tan solo 2 columnas, el **ID** y **RENDIMIENTO_GLOBAL**, este les sirve de ejemplo para que sepan cómo organizar su csv para subirlo a Kaggle. Nótese que el juez de Kaggle evalúa **exactamente** los IDs que aparecen en el test.csv. Su predicción debe tener cada uno de los IDs del test. junto con la columna **RENDIMIENTO_GLOBAL** que su modelo haya determinado.

Etapas del modelo

El notebook que usen para resolver el problema debe tener las siguientes secciones

A. Importación de datos

En esta sección, se importan las bibliotecas necesarias y se cargan los datos desde la fuente correspondiente, ya sea un archivo CSV, una base de datos, u otro formato. Es el primer paso para comenzar a trabajar con los datos.

B. Limpieza de datos

Aquí se lleva a cabo el proceso de limpieza de datos para abordar posibles problemas, como valores faltantes, duplicados, errores de formato, etc. También puede incluir la codificación de variables categóricas, normalización de datos, y cualquier otro paso necesario para preparar los datos para el modelado.

C. Exploración de datos

En esta sección, se realizan análisis exploratorios de los datos para comprender mejor su estructura y distribución. Puede incluir visualizaciones, estadísticas descriptivas y la identificación de patrones interesantes en los datos.

D. Modelos

Aquí se selecciona y entrena el modelo de machine learning. Pueden probarse varios modelos, dependiendo de la naturaleza del problema. Se divide el conjunto de datos en conjuntos de entrenamiento y prueba para entrenar y evaluar el modelo.

E. Exploración de hiperparámetros.

Se lleva a cabo la búsqueda y ajuste de los hiper parámetros del modelo para mejorar su rendimiento.

F. Evaluación del desempeño sistemática del modelo final seleccionado

Una vez que se ha seleccionado el modelo final, se evalúa su rendimiento de manera sistemática. Esto implica utilizar métricas de evaluación relevantes (como precisión, recall, AUC-ROC) y, posiblemente, realizar validación cruzada para obtener estimaciones más robustas del rendimiento.

G. Conclusiones y resultados

A partir de los resultados obtenidos analizar los resultados de las etapas del modelo.

¿Qué se puede entender de los datos y las predicciones?

¿Qué variables fueron determinantes en el modelo?

¿Cómo se comportó el modelo?

¿Qué aspectos se tuvieron en cuenta para mejorar el rendimiento del modelo?

H. Generación archivo de envío

Proceso de estructuración del dataframe de envío, de acuerdo el ejemplo de **submission.csv**, posteriormente exportación a csv.

NOTA: Habrá un leaderboard en tiempo real con todos los equipos de la competencia, sin embargo este leaderboard está calculado con el 50% de todos los datos de test. El leaderboard definitivo se calcula con otro 50% de datos los cuales son privados. Kaggle lo hace de esta manera para evitar que se haga overfitting de los datos públicos

Entregable

Además de realizar los envíos en kaggle, se debe entregar:

- **Jupyter Notebook:** con todo el código del proyecto. El notebook debe estar debidamente explicado usando celdas de texto. Todos los pasos de carga, limpieza de datos, análisis exploratorio, creación del modelo. Asegúrese antes de hacer el envío de que el notebook esté libre de errores, se haya guardado correctamente y se visualiza apropiadamente al volver a cargarse.

Se debe enviar el notebook antes de la media noche de la fecha límite a través del siguiente link: <https://www.dropbox.com/request/2GFWgLm2dZXldjlx79ka>. El archivo debe nombrarse isi-proj--userunal.ipynb, donde <userunal> es el nombre de usuario asignado por la universidad. Envíos que no sigan este formato o que se envíen después de la hora límite no se tendrán en cuenta.

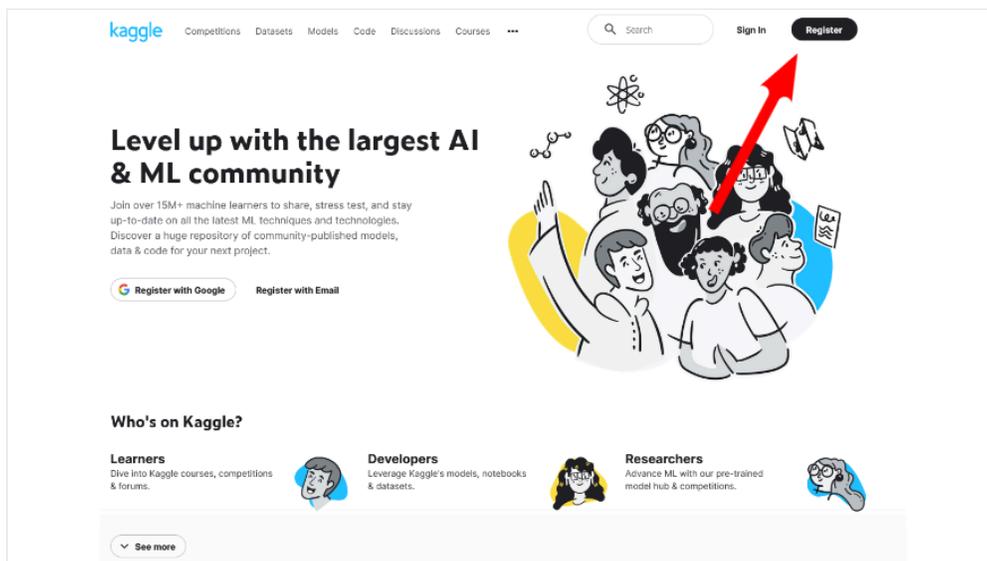
Evaluación

Todos los estudiantes deben hacer envíos a Kaggle y deben obtener al menos un 40% de exactitud (accuracy). Si esto no se cumple no se evaluará el notebook.

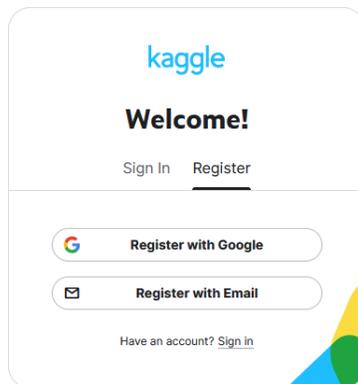
- **80% Notebook**
- **20% Desempeño en la tarea.** Se asignará una calificación de acuerdo con el desempeño en la competencia de Kaggle.

Instructivo plataforma Kaggle

1. Ingresar a <https://www.kaggle.com/>.
2. Dar click en el botón Register.



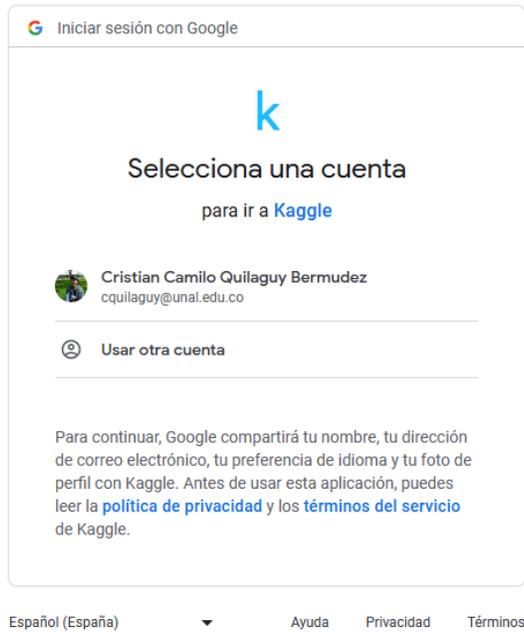
3. Elegir la opción de registrarse con google.



When you link your Facebook, Google, or Yahoo account, Kaggle collects certain information stored in that account that you have configured to make available. By linking your accounts, you authorize Kaggle to access and use your account on the third party service in connection with your use of kaggle.com.

[Contact Us / Support](#)

4. Elegir la cuenta con su correo UNAL. **Su cuenta debe estar con el correo UNAL, puesto que solo sus correos institucionales estarán habilitados para la competencia.**



5. Elegir nickname, aceptar términos y condiciones.

