UNIVERSIDAD
SERGIO ARBOLEDA

DESARROLLO DE SOFTWARE INTELIGENTE

PROYECTO

Armando Rafael Acuña Martínez

E-mail: armando.acuna01@usa.edu.co

DOCENTE

Marco Tulio Terán De La Hoz

ESCUELA DE CIENCIAS EXACTAS E INGENIERÍA

MAESTRÍA EN INTELIGENCIA ARTIFICIAL

2023

INTRODUCCIÓN

Este proyecto busca aplicar los temas vistos en la asignatura de Desarrollo de Software Inteligente, utilizando la metodología CRISP-DM y aplicando técnicas y algoritmos de Machine Learning y Deep Learning. Para ello, se debe seleccionar y comprender un conjunto de datos que aparece referenciado en el documento "Proyecto de aula: Deep Learning y Series de tiempo ".

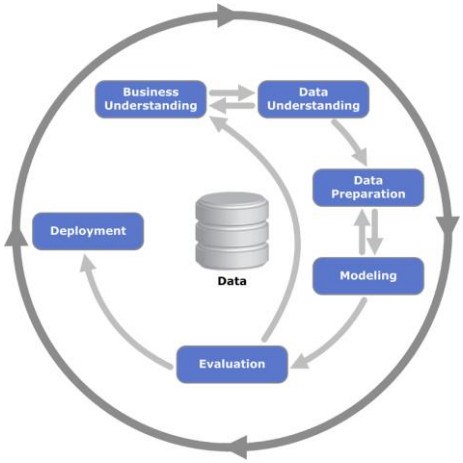
Después de revisar la documentación y el contenido de los diferentes DataSets que allí se referencian, decidí trabajar con:
"**Students Performance in Exams:** Notas obtenidas por estudiantes en varias asignaturas. Estos datos se basan en la demografía de la población. Los datos contienen varias características como el tipo de comida que se le da al estudiante, el nivel de preparación para el examen, el nivel de educación de los padres y el rendimiento de los estudiantes en Matemáticas, Lectura y Escritura. Con los datos se pueden resolver varios tipos de problemas de regresión y clasificación. También se puede utilizar para encontrar qué factores pueden conducir a mejores resultados en los exámenes." (Terán, 2023, pág. 2)

Analizando las diferentes características del DataSet se puede plantear un problema de Regresión, en que se busque estimar los resultados promedio que obtendrá un estudiante considerando las demás características como el género, el tipo de comida que se le da al estudiante, el nivel de preparación para el examen, el nivel de educación de los padres o su raza o tipo étnico.

Para el desarrollo del proyecto se seguirá la Metodología **CRISP-DM**, que incluye los siguientes pasos:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue

Ilustración 1. Metodología CRISP-DM



Fuente: Imagen tomada de (Vallalta Rueda)

Cada paso de la metodología CRISP_DM y la forma como fue desarrollado en el proyecto, será descrito a lo largo del documento.

DESCRIPCIÓN DEL MÉTODO

Comprensión del negocio

Este paso de la metodología CRISP-DM se enfoca en la comprensión de los objetivos y exigencias del proyecto desde una perspectiva de negocio.

En ese sentido, se desea conocer los factores más significativos que intervienen en los resultados obtenidos en las evaluaciones de varias materias de estudiantes y las relaciones que pueden existir entre estos factores.

Algunos de los factores o la relación entre ellos permitiría establecer, por ejemplo, si un género en particular obtiene mayores resultados que otro, o si la preparación previa permite alcanzar mejores resultados, entre otros.

Una pregunta interesante sería: ¿Qué patrones o interacciones en las características de los datos se pueden encontrar para mejorar los resultados de los estudiantes en cada una de las pruebas?

Comprensión de los datos

Este paso de la metodología CRISP-DM se encarga de la recolección de datos inicial y continúa con las actividades que permiten familiarizarse primero con los datos, identificar sus problemas de calidad, descubrir conocimiento preliminar en los mismos, y/o descubrir subconjuntos interesantes para formular hipótesis.

Dos puntos clave en esta fase: conocer los datos, estructura y distribución, y la calidad de los mismos.

En esta fase debemos ser capaces de:

- Ejecutar procesos de captura de datos.
- Proporcionar una descripción del juego de datos.
- Realizar tareas de exploración de datos.
- Gestionar la calidad de los datos, identificando problemas y proporcionando soluciones.

Para el desarrollo del proyecto se cuenta con el DataSet StudentsPerformance.csv, que tiene las siguientes características generales:

Fuente: Students Performance in Exams | Kaggle (SESHAPANPU, 2019)
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams
Peso del archivo: 72.04 kB
Número de registros: 1.000
Descripción: El archivo contiene las calificaciones obtenidas por los estudiantes de secundaria de los Estados Unidos.

Algunos artículos que han publicado la utilización de este DataSet son:

- DESIGN AND DEVELOPMENT OF HYBRID PRINCIPAL COMPONENT ANALYSIS (HPCA) ALGORITHM FOR ACADEMIC PERFORMANCE PREDICTION (Chitra & Rashmi, 2021). El DataSet se utilizó para crear la Fig. 5 Performance-based on the accuracy of StudentsPerformance.csv, mostrando el rendimiento basado en la precisión.
- DEVELOPMENT OF ARTIFICIAL INTELLIGENCE APPLICATIONS (Studi Kasus & Implementasi AI Menggunakan Berbagai Bahasa Pemrograman) (Prihantoro, y otros, 2023, pág. 86) El DataSet se utilizó como fuente de datos para desarrollar los modelos.

- Analysis of Academic Performance Based on Hierarchical Clusters: First Notes. (Lozada, Maldonado, Pullas, & Soria, 2021) El DataSet se utilizó como fuente de datos para desarrollar los modelos.

Este DataSet descargado de la ruta original (descrita anteriormente) y ubicado, en formato .zip en la ruta https://raw.githubusercontent.com/armaacum/data/main/archive.zip
El programa dl_project_ArmandoAcuna03_.ipynb toma el archivo de esa ruta, lo descomprime y lo deja para listo para ser procesado desde el NoteBook.

Composición del archivo:

Campo	Descripción de los datos
Gender	Género
race/ethnicity	Grupo étnico o raza
parental level of education	Nivel de educación de los padres
Lunch	Almuerzo
test preparation course	Curso de preparación para exámenes
math score	Puntaje de matemáticas
reading score	Puntaje de lectura
writing score	Puntaje de escritura

Fuente: Elaboración propia

Estos datos se basan en la demografía de la población. Los datos contienen varias características como el género, tipo de almuerzo que se le da al estudiante, el nivel de preparación para el examen, el nivel de educación de los padres y el rendimiento de los estudiantes en Matemáticas, Lectura y Escritura.

Una muestra de los datos es la siguiente:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Fuente: Elaboración propia

Con los datos se pueden resolver varios tipos de problemas de regresión y clasificación.
Para este caso, se puede utilizar para encontrar qué factores pueden conducir a mejores resultados en los exámenes de los estudiantes.

La exploración de los datos mostró:

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

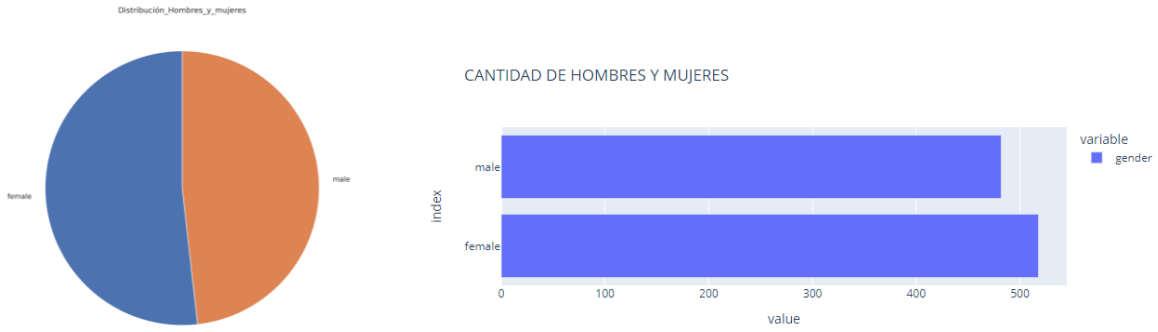
Fuente: Elaboración propia

Esto permitió identificar que el tamaño del DataSet es de 1000 instancias * 8 características.
Así mismo, permitió identificar que las características no tenían valores nulos (ausentes) y el tipo de datos de cada característica.

Las características numéricas encontradas fueron: 'math score', 'reading score' y 'writing score'.

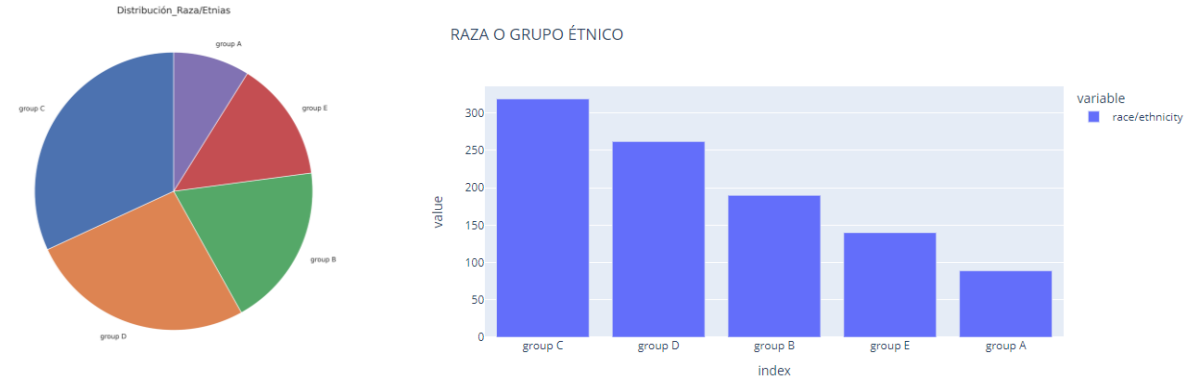
Las características categóricas encontradas fueron: 'gender', 'race/ethnicity', 'parental level of education', 'lunch' y 'test preparation course'.

La revisión de la característica “gender” – género permitió establecer que sólo presentaba dos valores, con la siguiente distribución:



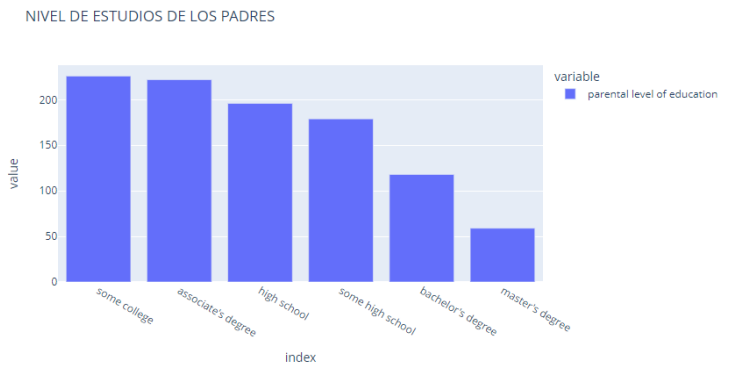
Fuente: Elaboración propia

La revisión de la característica “race/ethnici” – “raza o Grupo Étnico” permitió establecer que contenía cinco valores, cada uno correspondiente a un grupo étnico, con la siguiente distribución:



Fuente: Elaboración propia

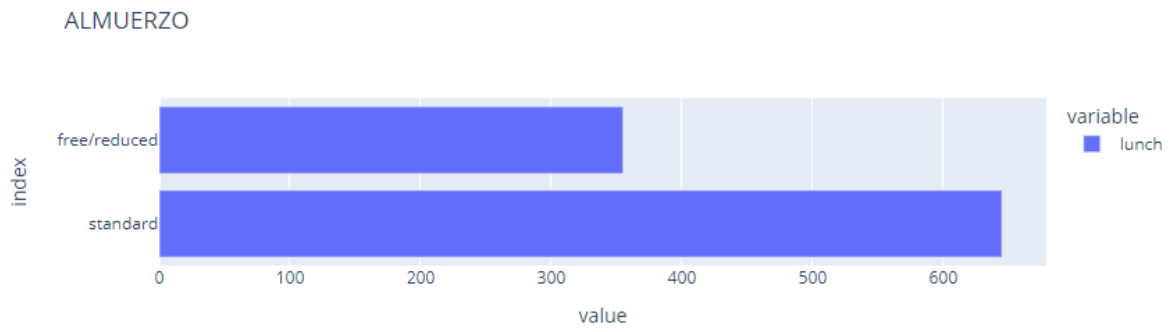
La revisión de la característica “parental level of education” – “Nivel de Estudios de los Padres” permitió establecer que contenía seis valores, cada uno correspondiente a un nivel, con la siguiente distribución:



Fuente: Elaboración propia

Se observa que la característica parental level of education representa 6 valores, donde el nivel de eduación de los padres "some college" se presenta para 226 estudiantes, "associate's degree" 222, "high school" 196, "some high school" 179, "bachelor's degree" 118 y "master's degree" 50. Existe un predominio de estudiantes cuyos padres están en la categoría some college.

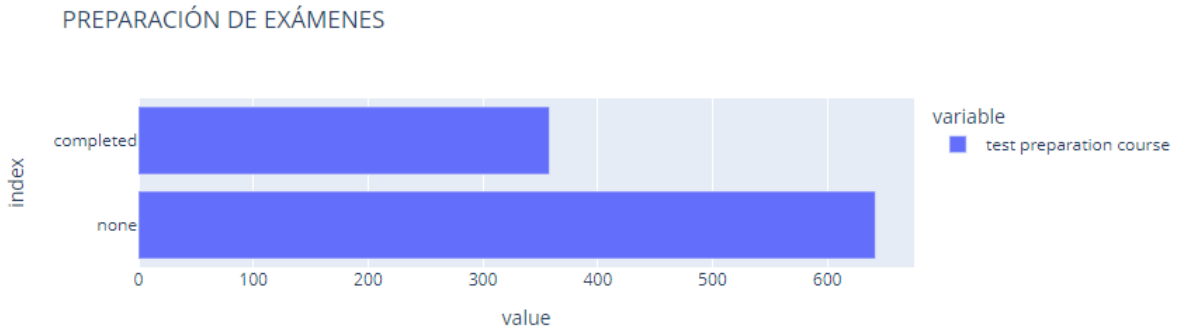
La revisión de la característica “lunch” – “Almuerzo” permitió establecer que contenía seis valores, cada uno correspondiente a un nivel, con la siguiente distribución:



Fuente: Elaboración propia

Se observa que la característica “lunch” representa 2 valores, donde standard contiene 645 estudiantes y free/reduced 355. Existe un predominio de estudiantes cuyo almuerzo es standard.

La revisión de la característica “test preparation course” – “Preparación de los exámenes” permitió establecer que contenía dos valores con la siguiente distribución:



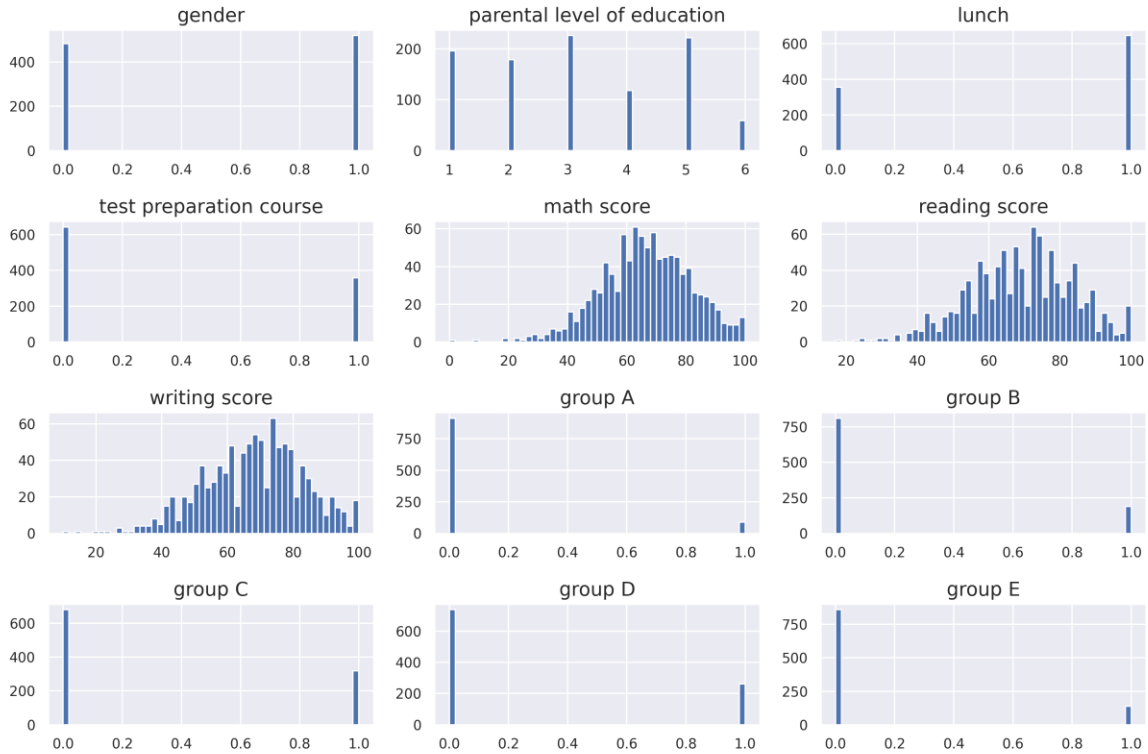
Fuente: Elaboración propia

Se observa que la característica "test preparation course" representa 2 valores, donde none contiene 642 estudiantes y completed 358. Existe un predominio de estudiantes que no realizaron el curso de preparación para exámenes.

Para entender el tipo de datos con los que se está trabajando se grafica un histograma para cada atributo numérico.

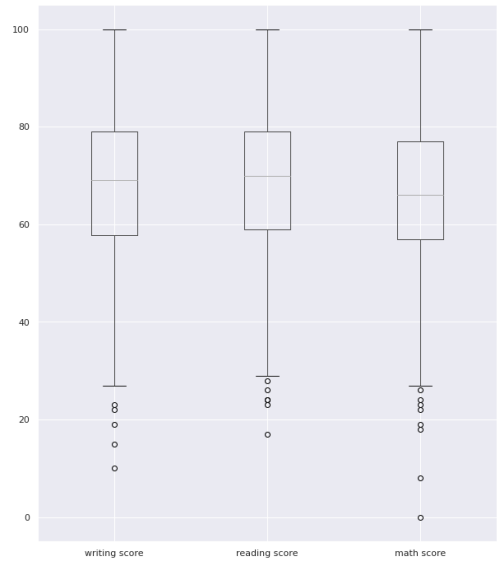
El histograma muestra el número de instancias (en el eje vertical) que tienen un rango de valores dado (en el eje horizontal).

Veamos ahora la distribución y los valores atípicos usando las gráficas de distribución y los diagramas de cajas



Fuente: Elaboración propia

No se identifican problemas asociados a ausencia de valores en las características, sin embargo, se aprecian algunos valores atípicos en los resultados de las evaluaciones:



Fuente: Elaboración propia

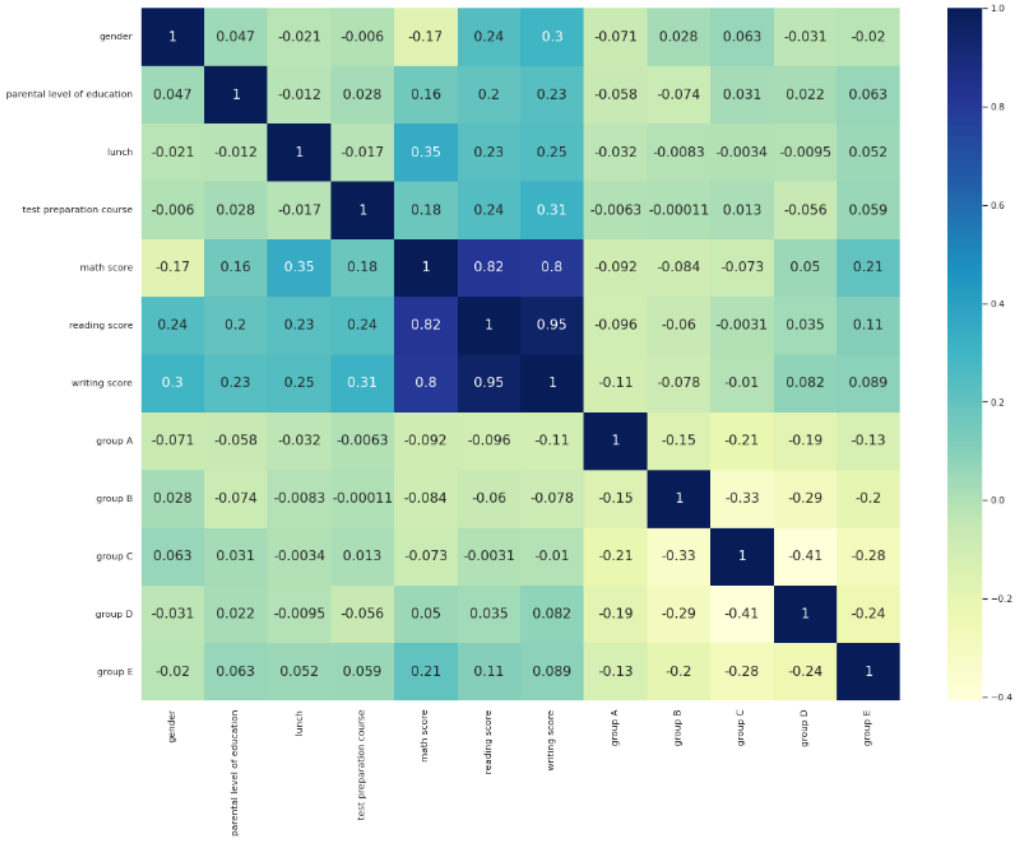
Se puede observar que los resultados de las evaluaciones están concentrados entre las puntuaciones de 55 y 78. Las puntuaciones por fuera de ese rango no son muchas y no deberían eliminarse para asegurar un resultado más real.

La visualización del conjunto de datos permite ver la uniformidad y el equilibrio sin la presencia de valores nulos. Hay mínimos valores atípicos y la distribución también es bastante buena.

Las características categóricas se convertirán a numéricas, de acuerdo con su análisis individual, para que sean manejados por los modelos.

- Dado que la característica "gender" sólo tiene dos valores, se reemplazará por 0 para Hombre(male) y 1 para Mujer(female)
- Dado que la característica "lunch" sólo tiene dos valores, se reemplazará por 0 para free/reduced y 1 para standard
- Dado que la característica "test preparation course" sólo tiene dos valores, se reemplazará por 0 para none y 1 para completed
- La característica "race/ethnicity" se tratará con OneHotEncoder() dado que son pocos valores diferentes y no representan un orden
- Para la característica "parental level of education" se establecerá un orden numérico

Luego de este tratamiento se puede apreciar la correlación de manera visual.



Fuente: Elaboración propia

Se observa que hay una relación alta en los resultados de las evaluaciones tanto de Escritura, Lectura y Matemáticas. También se observa una relación entre el grupo Etnico E y sus resultados en las evaluaciones, aunque no es una relación alta.

Preparación de los datos

Este paso de la metodología CRISP-DM cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos para las herramientas de modelado), incluye
Extracción y Preprocesamiento.

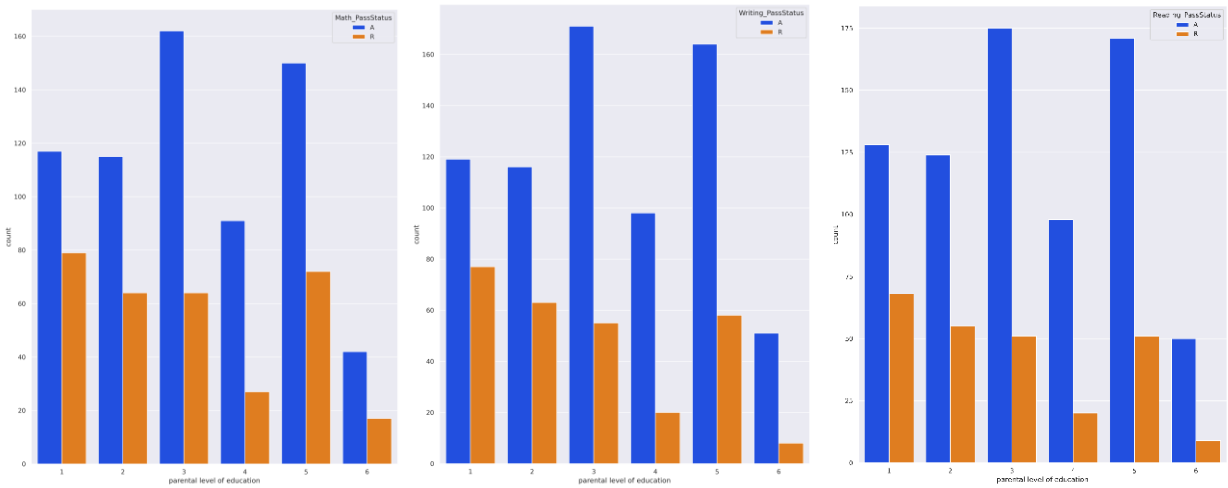
El objetivo de esta fase es obtener los datos finales sobre los que se aplicarán los modelos.

En esta fase debemos ser capaces de:

- Establecer el universo de datos con los que trabajar.
- Realizar tareas de limpieza de datos.
- Construir un juego de datos apto para ser usado en modelos.
- Integrar datos de fuentes heterogéneas si es necesario.

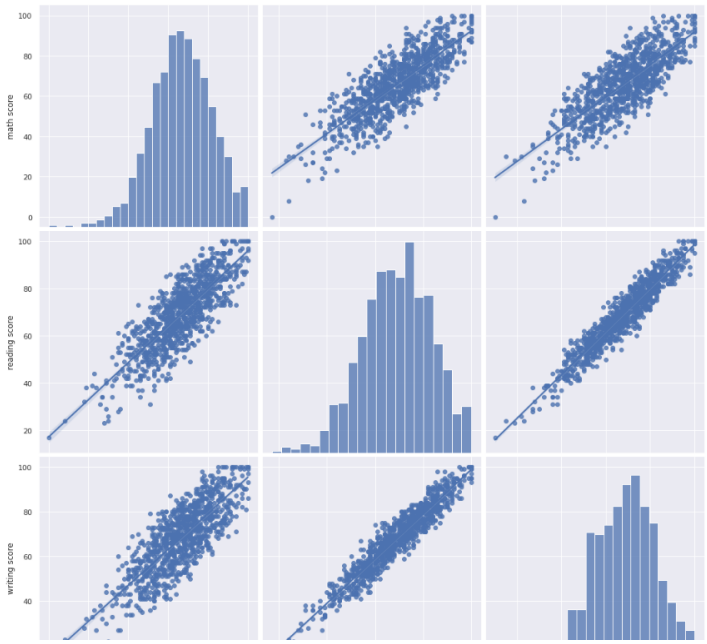
Analizando el negocio, se aprecia que todas las características son importantes, conceptualmente, para poder establecer su aporte en el resultado de los exámenes de los estudiantes.

Por ejemplo, si consideramos que para aprobar una evaluación se requiere una calificación a 60 puntos, en las siguientes gráficas se aprecia cómo el nivel de estudio de los padres se relaciona con la cantidad de estudiantes que aprobaron las evaluaciones de matemáticas, escritura y lectura.



Fuente: Elaboración propia

En la siguiente gráfica se puede observar que hay una gran relación lineal entre los resultados de las diferentes evaluaciones.



Fuente: Elaboración propia

Teniendo en cuenta que las características: 'math score', 'reading score' y 'writing score' son las que tienen una mayor relación, creé una característica ‘promedio_evaluacion’ con la media aritmética de las tres características anteriores. Esta característica ‘promedio_evaluacion’ se convertirá en la variable objetivo de los modelos.

Las características 'math score', 'reading score' y 'writing score' serán excluidas del DataSet, como parte de la limpieza de los datos.

Finalmente, se separarán los datos para entrenamiento (70%) y para Pruebas (30%) y luego de los datos de entrenamiento se deja un 10% para validación, incluyendo procesos de transformación y escalamiento.

Modelado

El objetivo último de esta fase es construir un modelo que nos permita alcanzar los objetivos del proyecto.

En esta fase debemos ser capaces de:

- Seleccionar las técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos.
- Fijar una estrategia de verificación de la calidad del modelo.
- Construir un modelo a partir de la aplicación de las técnicas seleccionadas sobre el juego de datos.
- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos.

Dentro de los diferentes modelos que se pueden utilizar para problemas de regresión, utilicé el modelo "RandomForestRegressor", con los siguientes parámetros iniciales:

- n_estimators = 10, # número de árboles
- criterion   = 'squared_error',
- max_depth   = None,
- max_features = 'auto',
- oob_score   = False,
- n_jobs   = -1,
- random_state = 20

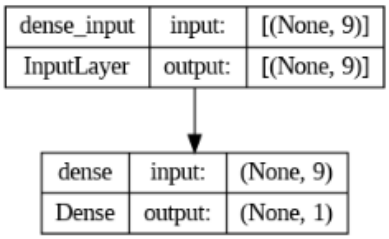
También se realizó la definición de la GRILLA con Random Forest.

Un segundo modelo clásico que se utilizó fue la Regresión Lineal Ridge con param_distributions = {'modelo__alpha': np.logspace(-5, 5, 500)} y el grid se definió como:

```
grid = RandomizedSearchCV(
       estimator  = pipe,
       param_distributions = param_distributions,
       n_iter     = 20,
       scoring    ='neg_root_mean_squared_error',
       n_jobs     = -1, #multiprocessing.cpu_count() - 1,
       cv         = RepeatedKFold(n_splits = 5, n_repeats = 3),
       refit      = True,
       verbose    = 0,
       random_state = 123,
       return_train_score = True
      )
```

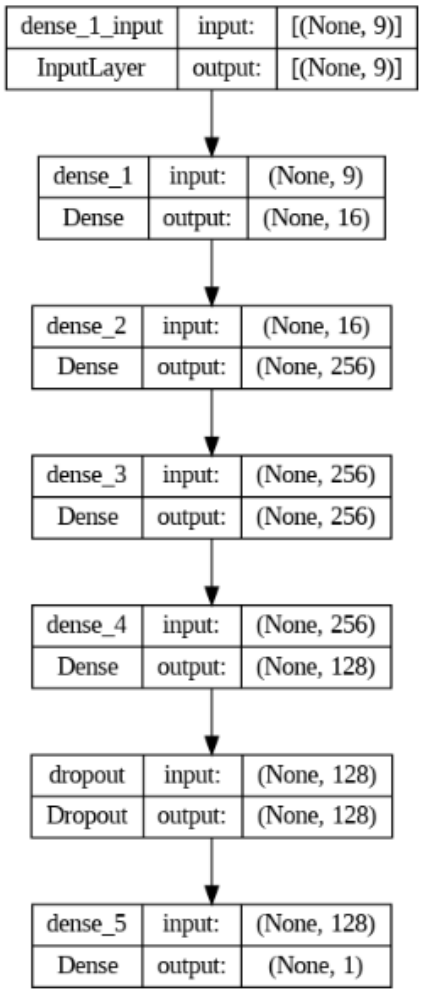
También se desarrolló un código para comparar los modelos clásicos de 'Linear Regression', 'Decision Tree Regressor' y 'Random Forest Regressor'.

A continuación, se utilizaron dos modelos de Redes Neuronales, el primero una red sencilla con 10 épocas



Fuente: Elaboración propia

y luego una red densa con 50 épocas:



Fuente: Elaboración propia

Para la compilación de las redes se estableció: optimizer='adam' y loss='mean_squared_error'

Evaluación del modelo

En esta fase se evalúa el grado de acercamiento del modelo o modelos a los objetivos de negocio.

En esta fase debemos ser capaces de:

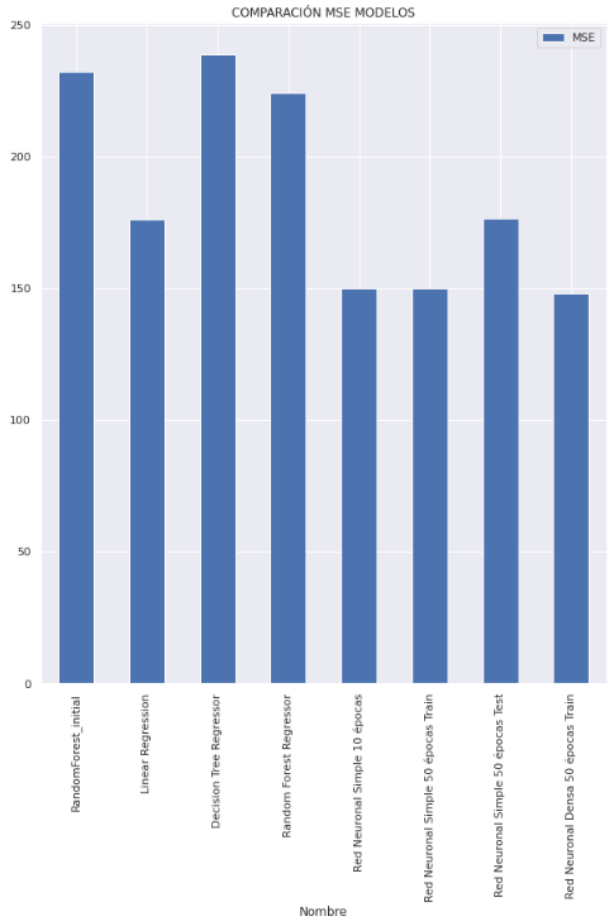
- Evaluar el modelo o modelos generados hasta el momento.
- Revisar todo el proceso que nos ha llevado hasta este punto.
- Establecer los siguientes pasos a tomar, tanto si se trata de repetir fases anteriores como si se trata de abrir nuevas líneas de desarrollo del proyecto.

La evaluación de los modelos seleccionados se realizó considerando el error cuadrático medio MSE.

Inicialmente se compararon los resultados de los modelos clásicos de Machine Learnig, obteniendo los siguientes resultados:

Modelo RandomForest_initial y su valor de mse: 232.18047101160445
Modelo Linear Regression y su valor de mse: 176.01489602687045
Modelo Decision Tree Regressor y su valor de mse: 235.10569597164857
Modelo Random Forest Regressor y su valor de mse: 224.7496509345097

La evaluación, incluyendo los modelos de redes, se puede apreciar a continuación.



Fuente: Elaboración propia

Modelo RandomForest_initial y su valor de mse: 232.18047101160445
Modelo Linear Regression y su valor de mse: 176.01489602687045
Modelo Decision Tree Regressor y su valor de mse: 238.7286487608575
Modelo Random Forest Regressor y su valor de mse: 223.92956023914243
Modelo Red Neuronal Simple 10 épocas y su valor de mse: 150.02391052246094
Modelo Red Neuronal Simple 50 épocas Train y su valor de mse: 149.98777770996094
Modelo Red Neuronal Simple 50 épocas Test y su valor de mse: 176.29908752441406
Modelo Red Neuronal Densa 50 épocas Train y su valor de mse: 148

Despliegue

El objetivo de esta fase es realizar el despliegue de los resultados obtenidos de forma que sea propagado a los usuarios finales, así como el mantenimiento del mismo una vez el despliegue haya finalizado.

En esta fase debemos ser capaces de:

- Diseñar un plan de despliegue de modelos y conocimiento sobre nuestra organización.
- Realizar seguimiento y mantenimiento de la parte más operativa del despliegue.
- Revisar el proyecto en su globalidad con el objetivo de identificar lecciones aprendidas.

En este caso, sólo se llegó a guardar los modelos de Redes neuronales.

CONCLUSIONES

Se concluye que la utilización de redes neuronales ofrece mejores resultados; ya se trata de hacerle tunning a los diferentes modelos para optimizar los resultados.

El trabajo desarrollado durante toda la asignatura fue interesante, desde el seguimiento de la metodología CRISP-DM, hasta la aplicación de los conceptos de Desarrollo de Software Inteligente.

Considero que faltó más tiempo para poder probar más cambios en los hiperparámetros de los modelos e ir generando más métricas para evaluar mejor el funcionamiento de los modelos en el caso particular de este DataSet.

Recomendaría separar un espacio del tiempo de clase para poder ir revisando los avances del proyecto, despejar dudas y hacer recomendaciones particulares a los maestrantes.

Agradecimientos al Ing. Marco Terán por si dedicación, guía y explicaciones para que desarrolláramos este proyecto.

REPOSITORIO

El cuaderno y la información asociada se deja en el repositorio https://github.com/armaacum/data.git para que pueda ser consultada.
La estructura del repositorio contiene las carpetas:

- PROYECTO_DSI
    - CÓDIGO

- DOCUMENTOS
- IMÁGENES

BIBLIOGRAFÍA

Chitra, M., & Rashmi, A. (2021). DESIGN AND DEVELOPMENT OF HYBRID PRINCIPAL COMPONENT ANALYSIS (HPCA) ALGORITHM FOR ACADEMIC PERFORMANCE PREDICTION. *Journal of Tianjin University Science and Technology*, 402. doi:https://doicatalog.org/19.4102/jtus.v54i6.3477

Lozada, E., Maldonado, R., Pullas, P., & Soria, L. (2021). Analysis of Academic Performance Based on Hierarchical Clusters: First Notes. *Information and Communication Technology for Competitive Strategies (ICTCS 2020). Lecture Notes in Networks and Systems, vol 191.*(ISBN: 978-981-16-0738-7). doi:10.1007/978-981-16-0739-4_5

Prihantoro, H., Widyaningsuh, T., Englishistina, I., Eko, N., S.Kom, R., S.Kom, E., . . . Dema, H. (2023). *DEVELOPMENT OF ARTIFICIAL INTELLIGENCE APPLICATIONS (Studi Kasus & Implementasi AI Menggunakan Berbagai Bahasa Pemrograman).* Indonesia: Sonpedia Publishing Indonesia.

SESHAPANPU, J. (2019). *Kaggle.* Recuperado el 2023, de https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Terán, M. (2023). *Modulo Aprendizaje profundo y Series de tiempo.* Recuperado el 2023, de https://github.com/marcoteran/deeplearning

Vallalta Rueda, J. (s.f.). *Escuela de formación en inteligencia artificial en salud.* Recuperado el 2023, de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/