Empleando las librerías *Lazypredict* y *Optuna* para analizar la regresión con el dataset *Alumnos beneficiarios del comedor universitario de la Universidad Nacional Hermilio Valdizán – 2023 (UNHEVAL)*

Lizbeth Estefany Cáceres Tacora

Mayo 2025

1 Introducción

En el contexto actual, donde la toma de decisiones basada en datos se ha convertido en una práctica esencial en diversos campos, el uso de modelos predictivos se vuelve fundamental. La capacidad de prever comportamientos, tendencias o valores futuros a partir de datos históricos permite optimizar recursos, mejorar servicios y anticiparse a necesidades. En este sentido, las técnicas de regresión representan una herramienta poderosa para modelar relaciones entre variables y generar predicciones cuantitativas.

El presente informe tiene como finalidad analizar un conjunto de datos obtenido de un portal de acceso público, aplicando métodos de regresión supervisada. A través de este análisis, se busca evaluar el desempeño de diferentes modelos y encontrar aquel que ofrezca el mejor equilibrio entre precisión y capacidad de generalización.

Para facilitar la exploración inicial de los modelos, se emplea la biblioteca *lazypredict*, que permite ejecutar rápidamente múltiples algoritmos de regresión sin necesidad de configuraciones manuales extensas. Esta herramienta proporciona una visión comparativa del rendimiento de distintos modelos utilizando métricas estándar como el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE).

Sin embargo, la simple comparación entre modelos no siempre es suficiente. La calidad de un modelo predictivo también depende de la adecuada selección de sus hiperparámetros. Por esta razón, se recurre a **Optuna**, una biblioteca de optimización bayesiana que permite ajustar automáticamente los hiperparámetros de los modelos con el fin de mejorar su rendimiento. En particular, se hace uso de Optuna para optimizar el hiperparámetro *alpha* en la regresión Ridge, un tipo de regresión lineal con regularización L2 que es especialmente útil cuando existen problemas de multicolinealidad entre las variables predictoras.

A través de este enfoque combinado —exploración automática de modelos y optimización de hiperparámetros— se busca no solo encontrar el modelo más eficaz, sino también entender cómo influye la regularización en la estabilidad y precisión de las predicciones. Este proceso es esencial cuando se trabaja con conjuntos de datos que presentan relaciones complejas o variables altamente correlacionadas.

En resumen, el presente trabajo ofrece un análisis práctico y detallado del uso de herramientas modernas de aprendizaje automático para la construcción de modelos predictivos robustos y bien ajustados, contribuyendo así a una mejor interpretación y utilización de los datos disponibles.

2 Elección del Dataset

Para este análisis se seleccionó la base de datos de acceso libre titulada *"Alumnos beneficiarios del comedor universitario de la Universidad Nacional Hermilio Valdizán – 2023 (UNHEVAL)"*, debido a que contiene información relevante y estructurada sobre variables relacionadas con el consumo alimentario de los estudiantes universitarios. Esta base de datos ofrece un contexto real que permite evaluar el comportamiento de distintos modelos de regresión, facilitando así la comprensión del funcionamiento y la aplicabilidad de herramientas de aprendizaje automático.

Cabe señalar que no todas las variables incluidas en el dataset original eran pertinentes para el objetivo del presente estudio. Algunas de ellas no guardaban relación directa con las tareas de predicción, por

lo que se procedió a un proceso de depuración y transformación de datos con el fin de preparar un subconjunto más adecuado para el análisis. Las principales acciones realizadas durante esta fase de limpieza y preprocesamiento fueron las siguientes:

- **Conversión de la fecha de nacimiento a formato `datetime`:** Se transformó la columna correspondiente con el fin de facilitar su uso posterior, especialmente para el cálculo de la edad.
- **Cálculo de la edad al 31 de diciembre de 2023:** Dado que se identificaron inconsistencias entre la edad reportada y el año de nacimiento, se optó por calcular una edad uniforme a partir de la fecha de nacimiento, tomando como referencia el cierre del año 2023. Esta transformación permite reducir posibles sesgos y asegurar mayor coherencia en los datos.
- **Eliminación de columnas irrelevantes o redundantes:** Se descartaron aquellas variables que no aportaban información significativa al análisis predictivo, lo que permitió reducir la dimensionalidad del dataset y centrar el estudio en variables verdaderamente útiles para el modelo de regresión.
- **Revisión de valores nulos o atípicos en las variables de raciones servidas:** Se inspeccionaron y depuraron registros con valores faltantes o inconsistentes en las columnas correspondientes a las raciones (desayuno, almuerzo, cena), con el propósito de garantizar la fiabilidad del análisis y la correcta aplicación de los modelos.

Como resultado de este proceso de preparación, se obtuvo un conjunto de datos depurado, coherente y enfocado en variables relevantes, adecuado para ser utilizado en modelos de regresión lineal y para evaluar el impacto de la regularización y la optimización de hiperparámetros en el desempeño predictivo.

3 Uso de la librería *Lazypredict*

Una vez realizado el proceso de limpieza y depuración del dataset, se procedió a su preparación para su análisis mediante la librería *Lazypredict*. Para ello, se seleccionaron únicamente las variables numéricas que aportan valor al modelo de regresión, considerando tanto la variable dependiente como las variables independientes con potencial explicativo.

Las variables utilizadas fueron definidas de la siguiente manera:

- **Variable dependiente:** Cantidad de raciones de almuerzo (*N RA ALMUERZO*)
- **Variables independientes:**
  - Edad
  - Año de la primera matrícula (*ANIO MATRICULA*)
  - Cantidad de raciones de desayuno (*N RAC DESAYUNO*)
  - Cantidad de raciones de cena (*N RAC CENA*)

Una vez establecidas las variables que representarían tanto la variable objetivo como las predictoras, se implementó el código correspondiente para aplicar la librería *Lazypredict*. Esta herramienta permite comparar rápidamente el desempeño de distintos modelos de regresión sobre el mismo conjunto de datos, proporcionando métricas como el R^2 y el error cuadrático medio (RMSE), lo que facilita la identificación de los modelos con mejor ajuste.

El código empleado para esta etapa se presenta a continuación:

```
import pandas as pd; from sklearn.model_selection import train_test_split; from
    lazypredict.Supervised import LazyRegressor
df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv")[lambda x: x['
    N_RAC_ALMUERZO'] > 0]
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']
models, _ = LazyRegressor(verbose=0).fit(*train_test_split(X, y, test_size=0.2,
    random_state=42))
print(f"Modelos:\n{models}")
```

Listing 1: Uso de la libreria Lazypredict

Donde los resultados optenidos fueron los siguientes:

Resultados de Modelos

Modelo	R^2	RMSE	Tiempo
LinearRegression	0.71	3.03	0.01 s

Cada uno de estos valores se interpretan de la sigueinte forma :

- R-Squared ($R^2 = 0.71$): Esto significa que el 71% de la variabilidad en las raciones de almuerzo ($N\_RAC\_ALMUERZO$) se explica por el modelo lineal usando las variables EDAD,$ANIO\_MAT1$, $N\_RAC\_DESAYUNO$ y N_RAC_CENA.. Es un valor aceptable, indica una buena relación lineal entre las variables predictoras y la variable objetivo. Sin embargo, el 29% restante de la variabilidad no es explicada por el modelo, lo que puede deberse a otros factores no incluidos, errores de medición, o relaciones no lineales.
- RMSE (Raíz del Error Cuadrático Medio = 3.03): En promedio, el modelo se equivoca por unas 3 raciones al predecir $N\_RAC\_ALMUERZO$. El RMSE debe evaluarse en el contexto del rango de $N\_RAC\_ALMUERZO$. Por ejemplo: Si los valores típicos están entre 0 y 10, entonces un error de 3 puede ser alto. Si están entre 0 y 50, un RMSE de 3 puede ser bajo.
- Tiempo de ejecución (0.01 s): Extremadamente rápido. Ideal para producción o para ejecutar muchas predicciones al instante.

Por otro lado tenemos que los modelos que mejor se ajsuatan son

Resultados de Modelos

Modelo	R^2	RMSE	Tiempo
KNeighborsRegressor	0.87	2.03	0.01 s
HistGradientBoosting	0.80	2.53	0.22 s
LGBMRegressor	0.80	2.53	0.10 s
ExtraTreesRegressor	0.80	2.55	0.23 s

4 Uso de la libreria Optuna

La regresión lineal simple (LinearRegression de sklearn) no tiene hiperparámetros que valga la pena optimizar con Optuna. Es decir: No tiene un parámetro como alpha, *max_depth*, n_estimators, etc., que pueda ajustarse. Solo ajusta una línea recta a los datos. No puedes "afinarla" más allá de los datos que le das. Por eso, no tiene sentido usar Optuna para buscar los "mejores hiperparámetros" de un modelo que no tiene ninguno relevante que optimizar.

```
import optuna, numpy as np, pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score, KFold

df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv")[lambda x: x['N_RAC_ALMUERZO'] > 0]
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']

def objective(trial): return -cross_val_score(LinearRegression(), X, y, cv=KFold(5, shuffle=True, random_state=42), scoring='neg_root_mean_squared_error').mean()

study = optuna.create_study(direction="minimize").optimize(objective, n_trials=30)
print(f"Mejor RMSE: {study.best_value:.4f}\nMejores hiperparametros: {study.best_params}")
```

Listing 2: Ejemplo de código en Python

significa que Optuna hizo una prueba (Trial) intentando encontrar los mejores hiperparámetros para tu modelo, con un resultado de RMSE = 2.485... (error cuadrático medio raíz). Sin embargo, todos los parameters: están vacíos.

5 Uso de ridge con Optuna

En este caso se utilizó la regresión Ridge, un tipo de regresión lineal mejorada mediante regularización, también conocida como regresión con penalización L2. Este método es particularmente útil cuando se desea mejorar la estabilidad del modelo y su capacidad de generalización.

Una de las principales ventajas de Ridge es su capacidad para manejar la multicolinealidad. Cuando las variables predictoras están altamente correlacionadas entre sí, los coeficientes obtenidos por una regresión lineal tradicional pueden volverse inestables, lo que reduce la fiabilidad de las predicciones. Ridge soluciona este problema al reducir la magnitud de los coeficientes, estabilizando así el modelo.

Otra razón para utilizar Ridge es que ayuda a reducir el sobreajuste (overfitting). En conjuntos de datos pequeños o con ruido, es común que los modelos se ajusten demasiado a los datos de entrenamiento, perdiendo capacidad de generalización. La penalización L2 de Ridge actúa limitando los coeficientes grandes, lo que evita un ajuste excesivo y mejora el rendimiento en datos nuevos.

Además, Ridge cuenta con un hiperparámetro ajustable llamado alpha, que controla la intensidad de la regularización. Esto permite afinar el modelo según las características del conjunto de datos. Herramientas de optimización automática como Optuna pueden utilizarse para encontrar el valor óptimo de alpha y así mejorar el rendimiento del modelo.

5.1 ¿Cuándo es recomendable usar Ridge?

- Cuando se tiene un número elevado de variables predictoras.
- Cuando algunas de esas variables están correlacionadas entre sí.
- Cuando se busca mejorar la capacidad de generalización del modelo, evitando el sobreajuste.

```python
import pandas as pd; from sklearn.model_selection import train_test_split, KFold, cross_val_score
from lazypredict.Supervised import LazyRegressor; from sklearn.linear_model import Ridge; import optuna

df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv").query('N_RAC_ALMUERZO > 0')
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Modelos evaluados:\n", LazyRegressor(verbose=0, ignore_warnings=True).fit(X_train, X_test, y_train, y_test)[0])

def objective(trial): return -cross_val_score(Ridge(alpha=trial.suggest_float('alpha', 1e-4, 100.0, log=True)), X, y, cv=KFold(5, shuffle=True, random_state=42), scoring='neg_root_mean_squared_error').mean()

study = optuna.create_study(direction="minimize").optimize(objective, n_trials=30)
print("\nOptuna - Ridge:\n Mejor alpha:", study.best_params, "\n Mejor RMSE:", study.best_value)
```

Listing 3: Uso del Optuna junto con el modelo Ridge

Datos obtenidos del modelo Ridge

Modelo	R^2	RMSE	Tiempo
Ridge	0.71	3.02	0.03 s

- R-Squared: 72% de la variabilidad en la variable objetivo ($N\_RAC\_ALMUERZO$) puede explicarse por las variables independientes (EDAD, $ANIO\_MAT1$, $N\_RAC\_DESAYUNO$, N_RAC_CENA) Esto indica un buen ajuste del modelo.
- RMSE (Root Mean Squared Error): el error promedio de predicción es de aproximadamente 3 raciones de almuerzo. Esto te indica cuán lejos están las predicciones, en promedio, de los valores reales.
- Time Take: el modelo se entrenó muy rápidamente — solo 0.03 segundos, lo que sugiere eficiencia computacional.

6 Resultados optenidos

Se utilizó Optuna para optimizar el valor del hiperparámetro alpha en un modelo de regresión Ridge, con el objetivo de minimizar el RMSE (Root Mean Squared Error). Este indicador mide el error promedio entre las predicciones del modelo y los valores reales observados, siendo más sensible a errores grandes. El parámetro alpha controla el grado de regularización en la regresión Ridge: valores más altos imponen una penalización mayor a los coeficientes del modelo, ayudando así a prevenir el sobreajuste y mejorar la generalización. En este caso, Optuna encontró el valor óptimo de alpha que proporciona el menor RMSE en el conjunto de validación.

Resultado Obtenido

Mejor valor de α: 71.32
Mejor RMSE: 2.49

Esto significa que el mejor modelo Ridge que Optuna encontró tiene una penalización moderadamente alta (valor de alpha grande) y que predice la cantidad de almuerzos servidos con un error promedio de aproximadamente 2.49 raciones.

- El valor óptimo de alpha es alto (71.32), lo cual indica que el modelo se beneficia de una fuerte regularización. Esto puede deberse a que algunas variables están correlacionadas (como $N\_RAC\_DESAYUNO$ y N_RAC_CENA con $N\_RAC\_ALMUERZO$), y el modelo necesita evitar darles demasiado peso.
- El RMSE de 2.485 significa que, en promedio, el modelo se equivoca en aproximadamente 2.5 almuerzos cuando predice el número de almuerzos servidos. Dependiendo del rango total de $N\_RAC\_ALMUERZO$, este error puede considerarse pequeño o no (por ejemplo, si los valores típicos son entre 20 y 100, sería aceptable).
- Los valores de alpha muy pequeños (por ejemplo, cerca de 0) dieron peores resultados, lo cual confirma que sin regularización el modelo no generaliza tan bien.

7 Conclusión

A lo largo del presente análisis se exploró el uso de técnicas de regresión y optimización de modelos aplicadas a un dataset real, utilizando herramientas como Lazypredict y Optuna. A partir de la limpieza y selección de variables relevantes del dataset de beneficiarios del comedor universitario de la UNHEVAL, se logró construir modelos predictivos capaces de estimar la cantidad de raciones de almuerzo que recibe un estudiante.

Con Lazypredict, se identificaron modelos con buen desempeño, destacando el KNeighborsRegressor con un R^2 de 0.87 y un RMSE de 2.03, lo cual representa una predicción bastante acertada dentro del rango observado. Estos resultados evidencian la utilidad de modelos más complejos frente a una regresión lineal simple.

Por otro lado, al utilizar Optuna, se evidenció que su utilidad depende del tipo de modelo. En el caso de LinearRegression, no fue posible optimizar nada relevante debido a la ausencia de hiperparámetros. Sin embargo, al implementar Ridge Regression, se aprovechó la capacidad de Optuna para ajustar el hiperparámetro alpha, lo que permitió mejorar la estabilidad y capacidad de generalización del modelo frente a problemas como la multicolinealidad y el sobreajuste.

En conclusión, el trabajo demuestra que:

- La calidad de los datos y su adecuada preparación son fundamentales para construir modelos predictivos confiables.
- Herramientas automáticas como Lazypredict permiten evaluar rápidamente múltiples modelos y seleccionar los más prometedores.
- El uso de técnicas de regularización, como Ridge, y su ajuste mediante algoritmos de optimización como Optuna, resulta esencial cuando se trabaja con datos reales, especialmente si existen variables correlacionadas o se desea mejorar la robustez del modelo.
- Este análisis proporciona una base sólida para aplicar técnicas de regresión y optimización en contextos reales, y sugiere que combinar diferentes herramientas puede conducir a resultados más precisos y eficientes.