Tarea 2 INF-398

Para datatests Concrete Data y Fish Market:

Para el dataset Fish Market la variable objetivo es el peso del pescado (Weight) → Dataset Overview: https://www.kaggle.com/datasets/vipullrathod/fish-market/data

Para el dataset Concrete Data la variable objetivo es la resistencia a compresión del hormigón (Concrete compressive strength) → Dataste Overview: https://www.kaggle.com/datasets/maajdl/yeh-concret-data/data

1. Si aplica, pre procesar datos (manejar valores nulos, comprobar tipo de datos, etc).
2. Realizar gráfico de correlación y pairplot. ¿Qué conclusiones relevantes puede sacar de estos datos? HINT: puede usar seaborn para el grafico de pairplot y un heatmap para correlación
3. Separar datos en train test y val. Luego estandarizar
4. Entrenar modelos Regresion Lineal, Ridge, KRidge y Lasso. teniendo en cuenta las siguientes consideraciones, y respondiendo para cada modelo:
    a. ¿Qué pasa con las predicciones cuando el valor del nivel de regularización (alpha) es muy alto en Lasso y Ridge?
    b. ¿Qué pasa con las predicciones cuando el valor de alpha tiende a cero en Lasso y Ridge? ¿A qué modelo de regresión se asemeja y por qué?
    c. Luego de elegir un valor de alpha óptimo para Lasso, Ridge y KRidge, realizar entrenamiento. HINT: pruebe distintos Kernel para KRidge.
    d. Entrenar una regresión lineal mediante Forward Stepwise Selection y Backward Stepwise Selection para elegir los atributos mas relevantes.
    e. Proponer 2 métricas de evaluación final relevantes y evaluar.

f. Revisar que atributos fueron seleccionado por cada modelo. ¿Existe alguna semejanza entre las columnas seleccionas por los distintos modelos?

g. Guardar tiempos de entrenamiento y métricas de cada modelo para compararlos mas adelante. HINT: puede ir guardándolo en una tabla de MARKDOWN o comentario de su notebook, no es necesario que se programe. LINK: https://tablesgenerator.com/markdown_tables

EJEMPLO TABLA:

Tabla Dataset Fish Market

MODELO	TIEMPO EJECUCION [s]	METRICA 1	METRICA 2
Linear Regression (LR)	10	1	1
Lasso	24	2	2
Ridge	35	4	5
KRidge	22	2	2
LR Forward Stepwise Selection	24	1	3
LR Backward Stepwise Selection	12	1	11

5. Diga por qué las escogió dichas métricas. Además para cada dataset responda ¿Cuál modelo se adapta mejor en tiempo y rendimiento?
6. Para cada dataset, según su conocimiento teórico ¿Cuál modelo debió adaptarse mejor en rendimiento? ¿por qué? ¿concuerda con lo obtenido?

SUGERENCIA: Para ordenar la entrega puede hacer los pasos de 1-4 con un dataset y luego con el segundo, finalmente responder 5 y 6 con resultados obtenidos.

Solo para dataset Concrete Data

1. Si aplica, pre procesar datos (manejar valores nulos, comprobar tipo de datos, etc)
2. Realizar gráfico de correlación.
3. Separar datos en train test y val. Luego estandarizar
4. Aplicar PCA al conjunto de train con el numero máximo de dimensiones y realizar grafico de varianza explicada vs nro. dimensiones. ¿Con cuántas dimensiones se puede explicar el dataset sin perder la integridad de este?
5. Aplicar PCA con el numero de dimensiones seleccionado y realizar grafico de correlación a su resultado. ¿Qué particularidad se observa en el gráfico?
6. Realizar predicción con los modelos vistos en clases realizando ajuste de hiperparámetros para optimizar su modelo (ya sea Cross Validation u otra técnica).
7. Proponer 2 métricas de evaluación relevantes diciendo por qué las escogió y elegir el mejor modelo en base a ellas. HINT: si encuentra que el resultado es mejor sin PCA, menciónelo.

SUGERENCIA: al ser lo mismo que la pregunta anterior, no hay problema en copiar y pegar su desarrollo de la pregunta 1-3

Kaggle

Dataset Overview:https://www.kaggle.com/datasets/rehandl23/fifa-24-player-stats-dataset

LINK A DESAFÍO:
https://www.kaggle.com/t/38f9592033f242ab94912051e2dac246

Objetivo del desafío:
El objetivo de este desafío es predecir el valor estimado de los jugadores de fútbol utilizando el conjunto de datos proporcionado. La variable objetivo es "Value", que representa el valor monetario estimado de cada jugador en el mercado.

Descripción del Dataset:
El conjunto de datos incluye una variedad de atributos que representan

diferentes habilidades y características de los jugadores. Cada atributo proporciona información valiosa que puede influir en el valor de un jugador. A continuación se detallan algunos de los atributos más relevantes:

- **Player**: Nombre del jugador.
- **Country**: Nacionalidad o país de origen del jugador.
- **Height**: Altura del jugador en centímetros.
- **Weight**: Peso del jugador en kilogramos.
- **Age**: Edad del jugador.
- **Club**: Club al que está actualmente afiliado el jugador.
- **Ball Control, Dribbling, Marking, Slide Tackle, Stand Tackle**: Habilidades técnicas que pueden afectar el rendimiento y el valor del jugador.
- **Aggression, Reactions, Composure**: Atributos mentales que influyen en el desempeño en situaciones de juego.
- **Value**: La variable objetivo que queremos predecir.

Consideraciones importantes:

1. Se evaluará la métrica R^2
2. Pueden realizar 5 submission por día
3. Es OBLIGATORIO evaluar el uso de PCA con un gráfico de varianza explicada. Si el rendimiento de su modelo sin PCA es mejor, no lo utilice para el submission, pero debe quedar en su notebook que al menos lo implentó como prueba.