Universidad del Norte
Departamento de Matemáticas y Estadística
Profesor: Dr. Lihki Rubio
Parcial 1 de Machine Learning
Alumno:_____

Duración: 48 horas Fecha: / /

Indicaciones: La presente evaluación debe ser entregada usando solamente un link de pagina web de GitHub creado mediante JBook. Cada figura, tabla, resultado, debe ser interpretado. Resultados y visualizaciones que no cuenten con su respectivo análisis serán evaluados con la nota mas baja.

Ejercicio 1

Ejercicio Notas de Clase (1.0 puntos). Considere los ejemplos vistos en clase, en los que se analizó la implementación de los modelos: *k-nn, regresión lineal, regresión ridge, regresión lasso, regresión logística*. Realice hiperparametrización y validación cruzada usando *(i) GridSearchCV y Pipeline* vs *(ii) Manualmente, sin usar GridSearchCV y Pipeline (Use ciclos FOR, WHILE, etc,...)*. Verifique que los scores obtenidos en los ítems *(i)-(ii)* son los mismos. Los ejercicios son los siguientes:

- Breast Cancer: (KNN, LogisticRegression) Decida cual es la métrica de mayor importancia en la aplicación de detección de cancer (*métrica de negocio*). Utilice esta métrica para la evaluación y selección del modelo y justifique su respuesta. Los resultados deben ser presentados usando el Cuadro 1. *El estudiante que obtenga el mejor score con esta métrica, será premiado con una décima para el corte.* Los datos deben ser cargados mediante el siguiente par de líneas:
  - from sklearn.datasets import load_breast_cancer
  - cancer = load_breast_cancer()
- Boston Housing: (LinearRegression, Ridge, Lasso). Utilice la métrica R^2 y $RMSE$ durante la evaluación y selección del modelo. Los resultados deben ser presentados usando el Cuadro 2. Realice un gráfico en el que muestre en los conjuntos de entrenamiento y test, el precio original y el predicho. *El estudiante que obtenga el mejor score con esta métrica, será premiado con una décima para el corte.* Los datos deben ser cargados mediante el siguiente par de líneas:
  - import mglearn
  - X, y = mglearn.datasets.load_extended_boston())

Ejercicio 2

Análisis Exploratorio de Datos (1.0 puntos). Dado los siguientes conjuntos de datos: Wind Speed y Fraud Detection, realizar un análisis exploratorio de datos el cual incluya lo siguiente:

- Descripción de tipos de variables, *reducción de nombres extensos en columnas*, calcular *número de observaciones, media, desviación estándar, mínimo, máximo, cuartiles*, realizar conteo de *datos faltantes y su porcentaje, histograma o diagrama de barras para la variable respuesta e independientes según corresponda*, seleccionar un mínimo de 4 variables independientes. Análisis de *simetría, datos atípicos y dispersión, etc,...* por medio de boxplot(). *Análisis bivariado*. Trazado de scatterplot() y regplot() para un mínimo de 4 pares de variables explicativas. En cada figura *agregar un análisis y descripción*. Para el conjunto de datos de *detección de fraude* hacer un merge entre tablas basado en *TransactionID*. Para esto, debe usar la función merge(). Esto es: pd.merge(train_transaction, train_identity, on='TransactionID', how='left'). Haga lo mismo para el *conjunto de prueba, el cual debe usar para evaluar el modelo final.*
- Según corresponda, realizar imputación de datos faltantes con la mediana (ver impute()).Realizar reducción de dimensionalidad por medio de eliminiación de columnas altamente correlacionadas usando *Variance Inflation Factor (VIF)*. Para esto se recomienda usar la siguiente librería variance_inflation_factor(). Un $VIF \geq 5$ indica alta multicolinealidad entre la correspondiente variable

independiente y las demás variables. *Recomendación: Eliminar una columna a la vez. Aquella con el máximo* $VIF \geq 5$. *Luego, para el nuevo pandas, calcular nuevamente VIF e identificar nuevas columnas con* $VIF \geq 5$ *máximo, y así sucesivamente hasta obtener solo valores de* $VIF < 5$. Según corresponda, variables categóricas deben previamente codificarse usando por ejemplo OneHotEncoder(). Pueden mantener las variables categóricas antes de la codificación previa al entrenamiento del modelo y *reducir multicolinealidad usando la prueba* chi2_contingency().

Ejercicio 3

Modelos de Clasificación (1.5 puntos). Considere el conjunto de datos Fraud Detection. Implemente la versión de clasificación para cada uno de los modelos estudiados en clases, y prediga la variable respuesta *isFraud*. Construir una tabla de error que contenga las métricas usuales de clasificación: *precission, recall,* f_1*-score, AUC*. Además, agregue *matrices de confusión* (ver confusion_matrix) y *curvas ROC* (ver plot_roc). Puede utilizar la librería GridSearchCV y Pipeline para evaluar cada modelo. Verifique que la validación cruzada seleccionada es la adecuada, y justifíquelo. *Utilice la métrica AUC, para seleccionar el mejor modelo de clasificación (maximizar AUC)*. Los resultados deben estar registrados en una tabla de error (ver Tabla 1) que resuma cada score obtenido por modelo implementado.

Modelo	*precission*	*recall*	f_1*-score*	*AUC*
K-NN	···	···	···	···
Linear Regression	···	···	···	···
Ridge	···	···	···	···
Lasso	···	···	···	···
Logistic Regression	···	···	···	···

Cuadro 1: Modelo de clasificación para detección de fraude.

Ejercicio 4

Modelos de Regresión (1.5 puntos). Considere el conjunto de datos Wind Speed. Implemente la versión de regresión de cada uno de los modelos estudiados en clases, para *predecir velocidad del viento horaria (VENTO, VELOCIDADE HORARIA (m/s))* en el conjunto de datos suministrado. Construir una *tabla de error* con las métricas usuales de regresión, *MAPE, RMSE, R2* (ver Table 2). Realice *particiones de entrenamiento y validación*, con base en lo descrito en la Figura 1. Estas particiones siguen la tendencia de la *velocidad del viento*. *Utilice la métrica RMSE* en la evaluación y validación, *para seleccionar el mejor modelo de regresión*. El *pliegue de validación en cada partición, debe estar siempre ubicado en el porcentaje final de cada partición*, debido a que el tiempo es fundamental en dichas predicciones. *Entre los periodos* $T = 7, 14, 21$, *indique cual corresponde a la mejor ventana de predicción para el entrenamiento*

Modelo	*MAPE*	*RMSE*	R^2
K-NN	···	···	···
Linear Regression	···	···	···
Ridge	···	···	···
Lasso	···	···	···
Logistic Regression	···	···	···

Cuadro 2: Modelo de regresión para velocidad del viento.

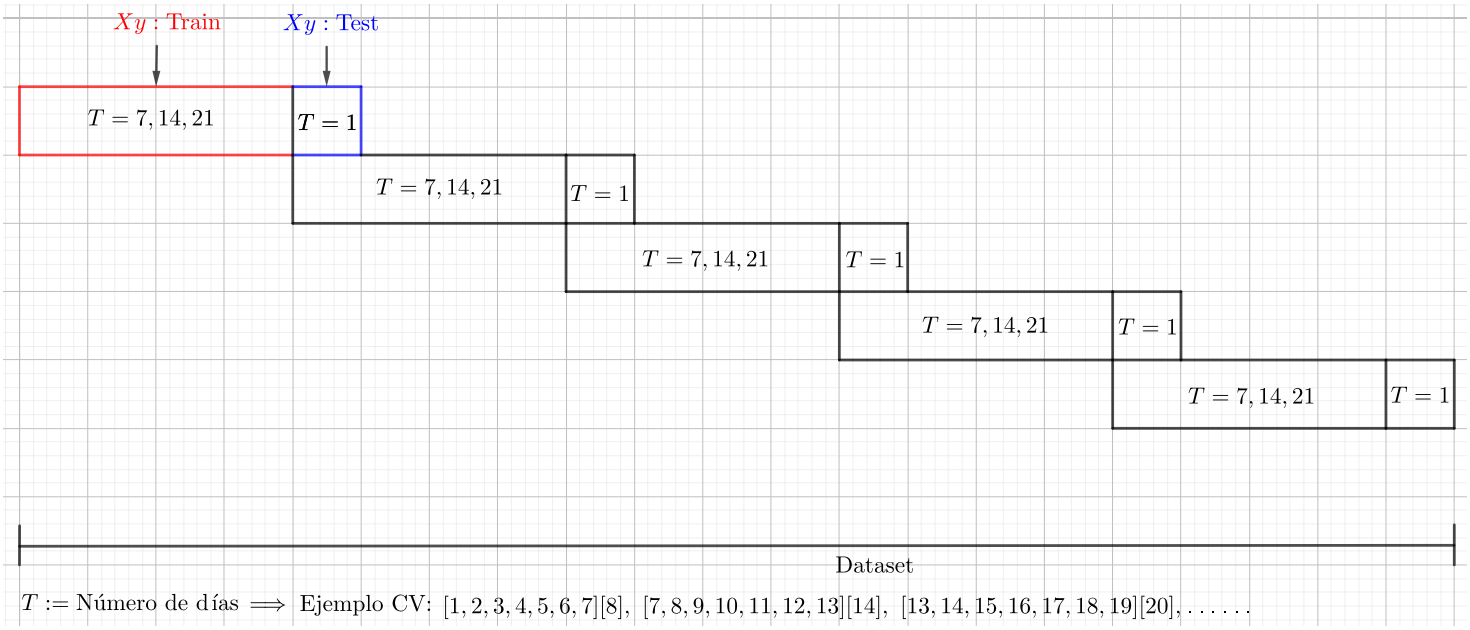

Figura 1: Particiones de entrenamiento y prueba. Modelo de regresión.

Diccionario de variables

Detección de fraude

Los datos proceden de transacciones reales de *comercio electrónico de Vesta* y contienen una amplia gama de características, desde el tipo de dispositivo hasta las características del producto. El objetivo principal es mejorar la eficacia de las alertas de transacciones fraudulentas para millones de personas en todo el mundo, ayudando a cientos de miles de empresas a reducir sus pérdidas por fraude y aumentar sus ingresos. Y, por supuesto, ahorrará a muchas personas la molestia de los falsos positivos.

- *TransactionDT*: Intervalo de tiempo a partir de una fecha y hora de referencia
- *TransactionAMT*: Importe del pago de la transacción en USD
- *ProductCD*: Código de producto, el producto de cada transacción
- *card1 - card6*: Información de la tarjeta de pago, como tipo de tarjeta, categoría de tarjeta, banco emisor, país, etc.
- *addr*: Dirección
- *dist*: Distancia
- *P_ and (R_) emaildomain*: Dominio de correo electrónico del comprador y del destinatario
- *C1-C14*: Recuento, cuántas direcciones se encuentran asociadas a la tarjeta de pago, etc. El significado real está codificado.
- *D1-D15*: Intervalo de tiempo, como los días transcurridos entre la transacción anterior, etc.
- *M1-M9*: Coinciden, como los nombres en la tarjeta y la dirección, etc.
- *Vxxx*: Vesta ofrece una gran variedad de funciones, como la clasificación, el recuento y otras relaciones entre entidades.
- *DeviceType*: Codificada. Información de identidad o conexión de red (IP, ISP, Proxy, etc) o firma digital
- *DeviceInfo*: Codificada. Información de identidad o conexión de red (IP, ISP, Proxy, etc) o firma digital
- *id_12 - id_38*: Codificada. Información de identidad o conexión de red (IP, ISP, Proxy, etc) o firma digital

Velocidad del viento

El pronostico de la velocidad del viento es fundamental, sobre todo por sus implicaciones en: *seguridad en la aviación y la navegación, generación de energía eólica, agricultura, construcción, meteorología, recreación y deporte*. Los datos suministrados, reportan diferentes mediciones que pueden explicar y permitir realizar la predicción de la velocidad del viento. Suponga que las *mediciones presentadas, son obtenidas cada 24 hrs* (ver Wind Speed). Además, suponga que desea *pronosticar, cual será la la velocidad del viento, durante las próximas 24 hrs, fuera de la muestra.* El objetivo principal es, *identificar que cantidad de energía eólica se puede generar durante este tiempo (24 hrs), para posteriormente, poder comercializarla* a empresas que producen por ejemplo hidrógeno verde.

- *HORA (UTC)*: Hora
- *VENTO, DIREï¿½ï¿½O HORARIA (gr) (ï¿½ (gr))*: Dirección del viento horaria
- *VENTO, VELOCIDADE HORARIA (m/s)*: Velocidad horario del viento (m/s)
- *UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)*: Humedad rel. máx. hora anterior (AUT) (%)
- *UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)*: Humedad rel. mín. hora anterior (AUT) (%)
- *TEMPERATURA Mï¿½XIMA NA HORA ANT. (AUT) (ï¿½C)*: Temperatura máx. hora anterior (AUT) (°C)
- *TEMPERATURA Mï¿½NIMA NA HORA ANT. (AUT) (ï¿½C)*: Temperatura mín. hora anterior (AUT) (°C)
- *UMIDADE RELATIVA DO AR, HORARIA (%)*: Humedad relativa horaria (%)
- *PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)*: Presión atmosférica a nivel de estación, horaria (mB)
- *PRECIPITAï¿½ï¿½O TOTAL, HORï¿½RIO (mm)*: Precipitación total por hora (mm)
- *VENTO, RAJADA MAXIMA (m/s)*: Máxima ráfaga de viento (m/s)
- *PRESSï¿½O ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)*: Presión atmosférica máx. hora anterior (AUT) (mB)
- *PRESSï¿½O ATMOSFERICA MIN.NA HORA ANT. (AUT) (mB)*: Presión atmosférica mín. hora anterior (AUT) (mB)