TALLER DE MACHINE LEARNING PARA BOOSTING DE UNA NOTA

A. OBJETIVO GENERAL

- El objetivo principal de este proyecto es proporcionar una oportunidad única de aprendizaje y recuperación a los alumnos. Mediante la resolución de problemas prácticos en el ámbito del Machine Learning, los estudiantes tendrán la oportunidad de mejorar su rendimiento académico en el curso. Específicamente, este proyecto servirá como una vía para mejorar notas deficientes obtenidas en talleres o pruebas de cátedra, conforme a las normativas y reglas establecidas en la primera clase del curso.

B. OBJETIVOS ESPECÍFICOS

1. **Desarrollo de Caso Práctico en ML**: Los estudiantes deberán implementar un caso práctico utilizando Python y algoritmos de aprendizaje supervisado como **Arboles de Decisión, SVM, Regresión Logística, Random Forest (RF), AdaBoost (ADA), Gradient Boosting Machine (GBM), XGBoost (XGB) o LightGBM (LGB)**. Este desarrollo involucra, llevar a cabo un Análisis Exploratorio de Datos (EDA), preprocesamiento de la información, ingeniería y selección de características, pruebas con 3 de los algoritmos mencionados y la interpretación de la importancia de las variables finales utilizadas para generar modelos que respondan al desafía de la realidad planteado.
2. **Aplicación de Enfoque CRISP-DM**: Los alumnos deberán adoptar el enfoque CRISP-DM para estructurar su proceso de modelado, lo que implica, a lo menos: "entender el negocio", preparar los datos, modelar, evaluar y desplegar la solución.
3. **Comparación de Algoritmos y Justificación de Decisiones**: Se espera que los estudiantes modelen el problema utilizando siempre, al menos tres algoritmos de ML diferentes para determinar cuál usarán, aplicando métricas objetivas. Deberán seleccionar el mejor algoritmo basado en métricas de desempeño y en el criterio analítico personal. Cada decisión tomada en el proceso de modelado que conduzca a un mejor rendimiento del algoritmo deberá ser debidamente justificada.

C. LA HISTORIA: La Odisea del Sommelier Digital

En un pequeño y pintoresco pueblo de España, donde el vino es casi tan importante como el aire que se respira, vive María. Ella es una apasionada del vino con un problema único: ha heredado una bodega de su abuelo, llena de vinos exquisitos, pero sin etiquetas (las cajas y botellas contienen solo información parcial sobre los vinos). Para complicar más las cosas, tiene un evento importante: un festival de vinos que podría hacer o deshacer el legado de su familia.

María necesita urgentemente dos cosas:

1. **Identificar el Tipo de Vino**: Tiene que saber qué vinos son tintos, blancos, rosados, etc., para poder presentarlos adecuadamente y ofrecerlos a un precio justo en el festival.
2. **Estimar el Precio**: Necesita saber cuánto cobrar por cada botella sin subvalorar (para no perder plata) o sobrevalorar el vino (para no perder negocios y retomar la tradición de su familia).

Ella recurre a usted, un experto en ML, para que la ayude a resolver estos dilemas utilizando la ciencia de datos y machine learning.

Los datos cuyo abuelo disponía se los estaban en el archivo **naria_vinos.csv** el cual tiene la siguiente descripción:

Variable	Descripción
vignedo	Nombre de la viña o viñedo que produce el vino.
vino	Indica la cantidad el nombre del vino producido por el viñedo
agno	Indica la añada (año) de producción del vino.
rating	Indica el rating dado por los consumidores del vino en la aplicación Vivino.
nro_revis	Indica el número de revisiones que ha tenido el vino desde su inclusión en la aplicación.
país	País de origen del vignedo donde se produce el vino.
region	Indica la región o valle donde se produce este vino.
precio	El precio en euros que se vende internacionalmente este vino.
tipo	Indica que variedad de vino es.
cuerpo	Puntuación del cuerpo, definida como la riqueza y peso del vino en la boca [de 1 a 5].
acidez	Puntuación de acidez, definida como el punto "más alto" del vino o su tartáridez; es lo que hace que un vino sea refrescante y que tu lengua salive y quieras otro sorbo [de 1 a 5].

Objetivos del Proyecto:

Ayudar a María a clasificar los vinos por 'tipo' utilizando sus características. Y asesorarla en la determinación del precio justo para cada botella de vino.

Aspectos Metodológicos (mínimos):

- El alumno deberá demostrar la comprensión de las necesidades de María y plantear cómo el ML puede resolverlas. Debe desarrollar un enfoque que deberá explicar en el documento de modelado.
- Deberá analizar el dataset para identificar patrones de cualquier tipo que realmente ayuden a María a tipificar sus vinos y a asignarles el precio justo.
- Deberá utilizar algoritmos más adecuados para clasificar el 'tipo' y predecir el 'precio'. (Siempre debe comparar al menos 3 algoritmos)
- Y finalmente deberá desarrollar modelos funcionales que María pueda utilizar para sus objetivos.

Entregables para María:

1. Un mecanismo simple que María pueda usar para escanear una botella de los cientos de cajas que no están etiquetadas cuya información base se encuentra en el archivo **maria_sintipo_sinprecio.csv** y obtener su tipo y precio sugerido.
2. Un resumen ejecutivo que describa el enfoque utilizado para resolver el desafío de maría desde la problemática inicial hasta la solución final. Adicionalmente, se le deberá indicar a María cuales son los posibles desafíos que la solución entregada presenta (con alguna métrica como la exactitud de los modelos)

Resultado Ideal Esperado: (por tanto, es lo que el alumno debe generar para ayudar a María)

María utilizará los modelos de ML creados por usted para clasificar y valorar con éxito todas las botellas en su bodega. El festival es un gran éxito, y el legado de su familia estará a salvo. Gracias al poder del machine learning, María no solo salvará su negocio, sino que también se debe convertir en la estrella del festival de vinos.

D. EN QUE CONSISTE EL TALLER – Esto es lo que usted debe responder con el trabajo

Como se mencionó en los objetivos, usted debe construir los mejores modelos predictivos posibles que María utilizará para estimar tipificar sus vinos y establecer su precio. Recuerde que María debería utilizar de forma automatizada su modelo para etiquetar sus vinos. Para lograrlo, se le pide lo siguiente:

a) Utilice Python con Jupyter Notebook para construir sus análisis de los datos y el modelo propuesto.
b) Realice un Análisis Exploratorio Detallado de los Datos (EDA) EDA completo que sirva como fundamento para cualquier modelo predictivo posterior. Su Análisis EDA debe incluir los siguientes elementos como mínimo:
   i) Comprensión del Dataset:
      (1) Describan cada variable, incluyendo su tipo (nominal, ordinal, intervalo, ratio) y el papel que juega en el dataset (por ejemplo, característica, target).
      (2) Identifiquen si existen valores faltantes y desarrollen una estrategia para tratar con ellos (eliminación, imputación, etc.).
   ii) Estadísticas Descriptivas:
      (1) Calculen estadísticas descriptivas básicas (media, mediana, modo, rango intercuartil, etc.) para las variables cuantitativas.
      (2) Presenten la frecuencia de categorías para las variables cualitativas.
   iii) Visualización de Datos:
      (1) Utilicen gráficos de barras y gráficos de torta para variables categóricas.
      (2) Para variables numéricas, generen histogramas, diagramas de caja y gráficos de violín para ver la distribución de los datos.
      (3) Implementen diagramas de dispersión y gráficos de correlación para identificar relaciones entre variables.
   iv) Análisis de Variables:
      (1) Analicen las variables objetivo para entender su distribución y cómo se relacionan con otras características del dataset.
      (2) Para las variables categóricas, consideren usar tablas de contingencia y pruebas de chi-cuadrado para explorar asociaciones.
   v) Mantengan un registro detallado de sus hallazgos y de las decisiones tomadas durante el EDA.
   vi) Asegúrense de comentar su código adecuadamente para que otros puedan entender su razonamiento.
c) Preprocese los datos para asegurarse de que su modelo tenga la mejor entrada para el modelamiento y genere resultados confiables.
d) Realice la ingeniería y selección de características para el modelamiento, justificando cada creación y eliminación de característica mediante algún análisis estadístico, un gráfico o basándose en lo que algún algoritmo indique que no aporta para la clasificación.
e) Busque alguna optimización de hiper parámetros de cada algoritmo para que su modelo alcance un mejor desempeño, ya que su posición está en juego durante su periodo de prueba.
f) Utilice siempre 3 algoritmos de ML que se hayan visto en clases: un subconjunto de 3 de los siguientes algoritmos {Regresión Logística, Support Vector Machine, Árboles de Decisión, Random Forrest, Adaboost, Gradient Boosting Machine, XGBoost o Light GBM } para realizar la clasificaciones y/o regresiones que el desafío le exige.
g) Use su mejor modelo de clasificación para determinar el tipo de los vinos y su mejor modelo de regresión para determinar el precio con lo cual ayudará a María en su desafío. Y complete el archivo maría_sintipo_sinprecio.csv.

E. ENTREGABLES – Incluya N° del equipo/grupo en nombres de archivos tal como se indica.

<table>
<tr><th>Entregable</th><th>Descripción</th></tr>
<tr><td>1. Código Python en jupyter</td><td>Gran parte del proyecto se debe entregar en el archivo. ipynb donde se deberá documentar lo que se hace con “markdown”, con secciones claras de lo que se hace.</td></tr>
<tr><td>2. Archivos con mejores modelo regresión (.joblib)</td><td>Por cada predicción que debe realizar deberá comparar al menos dos algoritmos para elegir algún modelo que mejor le resulte para su enfoque de solución. Sus mejores modelos los deberá grabar en un archivo con el nombre “bestmodelXXXXXX-GNN.joblib” ya que se le recomienda usar la librería joblib de Scikit-learn con la que podrá generar un archivo con la extensión ‘.joblib’ que contiene su modelo. [XXXXXX = breve texto que indique el objetivo del modelo]</td></tr>
<tr><td>3. Archivos .csv con las predicciones para los 169 cajas de vinos que María no tiene identificados</td><td>Deberá entregar el archivo maria_sintipo_sinprecio.csv con los valores predichos para tipo y para precio usando sus mejores modelos. En función de los desempeños se asignará el puntaje según siguiente tabla:
Sea PC = Predicciones correctas en tipo y precio (50% y 50%)
<table><tr><td>Si PC >= 136</td><td>10 pts</td></tr><tr><td>Si PC >= 102 & < 136</td><td>07 pts</td></tr><tr><td>Si PC >= 85 & < 102</td><td>05 pts</td></tr><tr><td>Si PC >= 68 & < 85</td><td>03 pts</td></tr><tr><td>Si PC < 68</td><td>00 pts</td></tr></table></td></tr>
<tr><td>4. Informe del Proceso de Modelamiento</td><td>Documento en pdf, basado en Word o Powerpoint donde se explique los enfoques usados en todo el proceso, se justifiquen las decisiones sobre características, sobre métricas y modelos seleccionados.</td></tr>
</table>

F. QUE SE EVALUA y NOTA FINAL– así se determinará la calificación final de este proyecto!

La nota final de este taller de boosting será calculada como la suma de los puntos obtenidos en cada sección que será evaluada de la siguiente forma:

Aspecto Evaluado	Puntaje Máximo	Observaciones
EDA	13	Haga las interpretaciones
Preprocesamiento	13	Debe estar orientado a los objetivos
Ingeniería de Características	12	Busque características que le ayuden a mejorar las predicciones
Selección de Características	12	Elija por algún método las características que mejor desempeño generan
Optimización de hiper parámetros	8	Use grillas para buscar los hiper parámetros que mejor desempeño generan.
Modelos finales y determinación de métricas de desempeño e interpretación de métricas finales	12	Haga selección de sus mejores modelos e interprete sus resultados en las métricas
Desempeño de modelo en predecir	10	Profesor tiene los valores reales de Tipo y Precios y se comparará contra ellos.
Presentación / Documentación Código	8	Justifique librería, documento bloques de código, no línea por línea.
Documento de Modelado (resumen Ejecutivo a María, Enfoque de Modelamiento, justificación de Ingeniería y de selección de características, interpretación de métricas, selección y justificación de modelos	12	Tome en consideración la formalidad, la ortografía, la redacción, que sea conciso pero completo y todo
TOTAL	**98**	**Nota Final = (0.06 * Puntaje) +1.0**

BUEN DESARROLLO DE PROYECTO