DuocUC
Vicerrectoría Académica
ET3
FORMA A

Examen Transversal
Alumno(a)

MDY7001	MINERIA DE DATOS

PUNTAJES Y NOTA

ENTREGA DE ENCARGO CON PRESENTACIÓN		
PUNTAJE TOTAL:	35 pts	NOTA: 7.0
PUNTAJE:	21 pts	NOTA: 4.0

INSTRUCCIONES GENERALES:

Entrega de encargo con presentación

Lea atentamente lo solicitado para cada una de las secciones del examen transversal.

- Los equipos de trabajo tendrán tres semanas para desarrollar este encargo a partir de la fecha de entrega de las instrucciones por parte del docente.
- Los estudiantes en grupos de 3 a 5 integrantes, deben realizar un proyecto de minería de datos, considerando:
    - Necesidades de la organización respecto de sus datos disponibles.
    - Identificar patrones de comportamiento dentro de los datos.
    - Aplicar modelos de minería de datos y realizar la evaluación de los algoritmos aplicados.
    - Comparar los resultados obtenidos por modelo.
    - Proyección del comportamiento de las diferentes áreas de la organización, de acuerdo con los resultados.
    - Realizar conclusiones para apoyar la toma de decisiones, de acuerdo con las necesidades de la organización.
- El docente le hará entrega de un dataset al cual le realizará aplicación de Minería de Datos en base a un enunciado para realizar proyección y predicción de información.
- En el equipo que ha elegido para rendir el examen transversal, revise que tienen a disposición la plataforma Anaconda Navigator y el IDE Jupyter Notebook para programación en Python.
- Comprima en un archivo .rar los siguientes archivos:
    - Archivo Word con un análisis del caso y capturas de pantalla sobre lo realizado durante la experiencia de aplicación.
    - Archivo Word con la evaluación de los algoritmos aplicados, y comparación de resultados de acuerdo a los diferentes parámetros.
    - Notebook con el código Python ejecutado para el proyecto debidamente documentado.
- La presentación debe durar máximo 15 minutos, y se debe realizar la justificación de las acciones realizadas en el producto.

Requisitos específicos:

Para el desarrollo del sistema solicitado, los equipos de trabajo deben considerar los siguientes aspectos:

- Desarrolla un proyecto de minería de datos utilizando la herramienta Jupyter Notebook.
- Realice un análisis de la organización y de los datos que disponen antes de comenzar a trabajar los registros.
- Realizar el análisis exploratorio de los datos (*EDA*), obtener representación gráfica de estos para tomar decisiones acertadas sobre la utilidad de los datos conforme al comportamiento observado.
- Generar información y conocimiento útil desde las fuentes de datos y la experiencia, para apoyar la toma de decisiones de acuerdo a las necesidades de la organización.
- Haz experimentos usando como mínimo 4 algoritmos revisados en clases para realizar las predicciones.
- Evalúa el algoritmo en base a las métricas que corresponda para determinar cuál es el más adecuado al contexto.

CONTENIDO DEL INFORME

- Información del proyecto
- Propósito y justificación del proyecto
- Realizar análisis de los datos disponibles y encontrar patrones de comportamiento.
- Explicar paso a paso lo realizado y adjuntar el notebook generado por la herramienta Jupyter Notebook con el código Python usado para el manejo de datos.
- Realizar razonamiento y justificación respecto del análisis de los datos
- Definición de fases, tareas y entregables a partir del dataset y el proyecto en general.
- Sugerir decisiones estratégicas para apoyar la toma de decisiones que ayudarán a la organización.

Formatos de entrega:

- Hoja tamaño carta o A4
- Tipo de letra: Títulos Arial 14 Negrita, Contenido Arial 12
- Interlineado: 1,5.
- Párrafo: Justificado.

NOTA: Dentro de la evaluación del examen serán considerados los siguientes puntos:

- Formato Presentación Informe.
- Ortografía.
- Redacción.

PRESENTACIÓN ORAL

- Presentación 15 minutos (10 de presentación y 5 de preguntas).
- Quien no cuente con el material entregable no podrá realizar la presentación.
- La presentación debe basarse en la estructura solicitada por el docente.

CASO: PROYECTO COVID-19

Objetivo General: Diagnosticar si individuos que presentan síntomas y características específicas, padecen la enfermedad ***COVID-19*** y están con riesgo de muerte. Este diagnóstico se debe lograr gracias a un modelo predictivo basado en minería de datos y aprendizaje automático con una exactitud de clasificación **mayor o igual al 95%**, lo que eventualmente, permitirá a los médicos tomar medidas proactivas y priorizar de mejor manera la hospitalización de los casos con mayor riesgo vital.

Contexto: El 31 de diciembre de 2019, la Organización Mundial de la Salud (OMS) recibió una alerta sobre varios casos de neumonía en la ciudad de Wuhan, provincia de Hubei en China. El virus no coincidía con ningún otro virus conocido. Esto generó preocupación porque cuando un virus es nuevo, no se sabe cómo afecta a las personas.

El nuevo coronavirus 2019 (2019-nCoV) es un virus (más específicamente, un coronavirus) identificado como la causa de un brote de enfermedad respiratoria detectado por primera vez en Wuhan, China. Al principio, muchos de los pacientes en el brote en Wuhan, China, según los informes, tenían algún vínculo con un gran mercado de mariscos y animales, lo que sugiere la propagación de animal a persona. Sin embargo, un número creciente de pacientes, según los informes, no han estado expuestos a los mercados de animales, lo que indica que se está produciendo una propagación de persona a persona. En este momento, no está claro qué tan fácil o sostenible se está propagando este virus entre las personas.

Solución global esperada: Con la información desagregada a nivel diario sobre los individuos afectados en el mundo, se requiere que los alumnos de la asignatura de Minería de Datos de DUOC UC puedan obtener un modelo predictivo que permita anticipar y priorizar la atención de los pacientes con alta probabilidad de muerte. Específicamente se espera que con la utilización de un algoritmo de clasificación seleccionado y evaluado, entre al menos cuatro vistos en la asignatura, se obtengan las predicciones con el riesgo de muerte de los pacientes que consultan, ya sea individualmente o masivamente.

Fuente de Datos: Use para el entrenamiento de los modelos el archivo ***COVID19_2020_open_line_list.xlsx*** entregado con el caso. Estos datos varían día a día por lo que puede usar los datos actualizados para hacer *scoring* del modelo. La URL del dataset actualizado está en:

https://docs.google.com/spreadsheets/d/1itaohdPiAeniCXNlntNztZ_oRvjh0HsGuJXUJWET008/edit#gid=0.

Descripción de Datos:

Columna	Descripción
ID	Identificador Correlativo
age	Edad del individuo
sex	Género del individuo
city	Ciudad donde fue detectado
province	Región donde fue detectado
country	País donde fue detectado
wuhan(0)_not_wuhan(1)	Indica si estuvo o no en la ciudad de Wuhan
latitude	Coordenada Latitud donde se detectó
longitude	Coordenada Longitud donde se detectó
date_onset_symptoms	Síntomas iniciales presentados por el individuo
date_admission_hospital	Fecha de admisión en el hospital
date_confirmation	Fecha de confirmación de la enfermedad
symptoms	Síntomas predominantes
lives_in_Wuhan	Indica si vivió en Wuhan o no
travel_history_dates	Historial de viajes con sus fechas
travel_history_location	Historial de viajes con los lugares
reported_market_exposure	Exposición en el mercado que se haya reportado
additional_information	Información extra
chronic_disease_binary	Indica si tiene o no tiene enfermedad crónica
chronic_disease	Indica cuál es la enfermedad crónica que tiene
source	Fuente informativa del caso
outcome	Resultado del diagnóstico
date_death_or_discharge	Fecha de fallecimiento o Alta
notes_for_discussion	Notas de discusión del caso
location	Ubicación del individuo

Instrucciones Generales:

- Importe y analice el dataset ***COVID19_2020_open_line_list.xlsx*** entregado en el caso usando la librería *Pandas* para Python.

- Documente y fundamente las decisiones que tome respecto de la limpieza y transformación de los datos, tomando en consideración los casos en que sea necesario limpiar filas o columnas, manipular el tipo de datos, eliminar datos redundantes, excluir variables según criterios de correlación, tratamiento de datos faltantes, tratamiento de *outliers* (valores atípicos), uso de variables *dummies* y agrupación de datos (*data binning*).

- Genere una nueva planilla de datos llamada ***COVID19_Datos_Preprocesados.xlsx*** que contendrá los datos depurados listos para ser procesados posteriormente por un algoritmo de Machine Learning.

- Realice el análisis exploratorio de los datos para verificar y validar que las conjeturas realizadas en la fase de limpieza de datos fueron implementadas correctamente, y en congruencia con las decisiones tomadas en el preprocesamiento de los datos. Si producto de este análisis exploratorio debes hacer cambios, éstos deberán quedar documentados en el informe.

- Implemente y ejecute el entrenamiento/test de al menos 4 algoritmos de clasificación vistos durante el semestre para las predicciones que permitan satisfacer los objetivos del proyecto. En esta etapa deberá contemplar al menos 4 experimentos que se diferencien por el ajuste de los parámetros de los modelos aplicados.

- Confeccione una tabla por cada experimento con las métricas de rendimiento obtenidas de los 4 algoritmos. De la misma tabla obtener las conclusiones a priori y explicar los resultados en lenguaje natural.

Algoritmo	AUC	CA	F1	Precisión	Recall
<<indicar algoritmo utilizado>>	<<Área bajo la curva ROC>>	<<Exactitud de Clasificación>>	<<F1>>	<<Precisión>>	<<Sensibilidad>>

- Documente las conclusiones para la selección del mejor modelo con apoyo de gráficas (de las métricas) desarrolladas en Python y explique los resultados en lenguaje natural (interpretación de las métricas).

- Una vez que seleccione el mejor modelo según las métricas de evaluación, realice la predicción real usando una planilla con datos actualizados (no etiquetada) ubicada en:

  https://docs.google.com/spreadsheets/d/1itaohdPiAeniCXNlntNztZ_oRvjh0HsGuJXUJWET008/edit#gid=0

- Documente y presente ante el curso los resultados del proyecto y extraiga conclusiones respecto del aporte real del modelo respecto de satisfacer los objetivos de negocio e implemente en Python la presentación gráfica de dicha información de manera adecuada para la audiencia con su respectiva explicación en lenguaje natural.

Entregables:

1. Un notebook desarrollado en Jupyter Notebook con el código Python ejecutado y debidamente documentado con todos los pasos desarrollados (***proyecto_covid19.ipynb***)
2. Planilla con los datos preprocesados que sirvieron de insumo para los modelos de minería de datos y aprendizaje automático, con el nombre ***COVID19_Datos_Preprocesados.xlsx*** y su respectiva metadata.
3. Informe en formato Word que cumpla con las especificaciones entregadas en el formulario ET3.