Examen Diplomado en Ciencia de Datos — Módulo

6 Diciembre 2025

Conjunto de datos

Considere el siguiente diagrama de datos.

Los datos proporcionados registran información sobre restaurantes y sus diferentes dimensiones: horarios, tipo de cocina, métodos de pago y estacionamientos. Adicionalmente se registra información de usuarios de los mismos restaurantes, así como tipo de cocina predilecta y tipos de pago. Ambas entidades se relacionan a través de las calificaciones que el usuario da al restaurante.

Diccionario de datos

- **users**: La tabla concentra información resumida del usuario.
  - **userID**: Es un identificador del usuario.
  - **latitude**: Es la latitud geográfica del usuario.
  - **longitude**: Es la longitud geográfica del usuario.
  - **smoker**: Indica si el usuario es fumador o no.
  - **drink level**: Indica el nivel o la frecuencia con la que el usuario bebe.
  - **dress preference**: Indica la preferencia de vestimenta que tiene el usuario. – Ambience: Indica el tipo de ambiente preferido por el

usuario. – **transport**: Medio de transporte usado por el usuario.

– **marital status**: Estado civil.
– **hijos**: Dependientes económicos.

– **birth_year**: Año de nacimiento del usuario.

– **interest**: Tópico de interés del usuario.

– **personality**: Personalidad del usuario.

– **religion**: Religión profesada por el usuario.

– **activity**: Actividad profesional practicada por el usuario.

– **color**: Color predilecto por el usuario.

– **weight**: Peso corporal del usuario en kilogramos.

– **budget**: Nivel de presupuesto o socioeconómico del cliente.

– **height**: Altura en metros del usuario.

• **usercuisine**: La tabla registra los tipos de cocina predilectas por el usuario.

– **userID**: Es un identificador del usuario.

– **Rcuisine**: Tipo de cocina.

• **userpayment**: Registra los métodos de pago con los que cuenta el usuario.

– **userID**: Es un identificador del usuario.

– **Upayment**: Tipo de método de pago.

• **ratings**: La tabla registra las calificaciones que el usuario asigna a los restaurantes en diversos rubros.

– **userID**: Es un identificador del usuario.

– **placeID**: Es un identificador del restaurante.

– **rating**: Es la calificación otorgada en general al restaurante.

– **food_rating**: Es la calificación otorgada a la comida.

– **food_rating**: Es la calificación otorgada al servicio.

• **restaurants**: Registra la información resumida de los restaurantes.

– **placeID**: Es un identificador del restaurante.

– **latitude**: Es la latitud geográfica del restaurante.

– **longitude**: Es la longitud geográfica del restaurante.

– **the_geom_meter**: Es un código que representa geográficamente al restaurante.

– name: Es el nombre del restaurante.

– **address**: Dirección postal donde se encuentra el restaurante.

– **city**: Es la ciudad donde se encuentra el restaurante.

– **fax**: Es el fax del restaurante.

– **address**: Es el código postal donde se encuentra el restaurante.

– **alcohol**: Indica el tipo de bebidas alcohólicas servidas por el restaurante.

– **smoking_area**: Indica el tipo de área para fumadores con la que el restaurante

cuenta o no.
  – **dress_code**: Indica el tipo de código de vestimenta requerido por el restaurante.
  – **accessibility**: Indica si el restaurante tiene accesibilidad para personas con sillas de ruedas, etc.
  – **price**: Indica el nivel de precios que maneja el restaurante.
  – **url**: Es la liga al sitio web del restaurante.
  – **Rambience**: Es el tipo de ambiente del restaurante.
  – **franchise**: Muestra si el restaurante pertenece a una franquicia o no.
  – **area**: Indica el tipo de área donde se encuentra el restaurante.
  – **other_services**: Registra si el restaurante ofrece otros servicios además de los alimen ticios.

- **parking**: Registra los tipos de estacionamiento a los que tiene acceso el restaurante.
  – **placeID**: Es un identificador del restaurante.
  – **parking_lot**: Es el tipo de estacionamiento.
- **cuisine**: Indica los tipos de cocina ofertados por el restaurante.
  – **placeID**: Es un identificador del restaurante.
  – **Rcuisine**: Es el tipo de cocina.
- **paymenth methods**: Muestra las formas de pago con las que cuentan los restaurantes.
  – **placeID**: Es un identificador del restaurante.
  – **Rpayment**: Es la forma de pago.
- **hours**: Registra los días y horarios en los que el restaurante ofrece servicios.
  – **placeID**: Es un identificador del restaurante.
  – **hours**: Rango de horas de apertura del restaurante.
  – **days**: Días de la semana de apertura del restaurante.

Solución

Seleccione la unidad muestral con la que se desea trabajar, a partir de este punto se detonan las siguientes actividades. Posteriormente, realice lo siguiente:

- (3 puntos) Ingeniería de datos: Teniendo muy presente la unidad muestral, genere al menos 5 variables adicionales a las proporcionadas individualmente.
- (1 punto) Construcción de variable objetivo: Genera la variable objetivo que quiere estimar, ya sea continua o categórica.
- (2.5 puntos) Limpieza de datos: De acuerdo con lo visto en clase, identifique si es necesario la aplicación de los procesos que se listan a continuación:
  – Detección y remoción de valores extremos.
  – Detección y remoción de variables poco pobladas. 65%
  – Detección y tratamiento de valores ausentes.

– Remoción de variables altamente correlacionadas. Remueva solo aquellas con correlación 1 en valor absoluto.
– Detección y remoción de variables unitarias (unarias).

En caso de que algún proceso no sea necesario, justifique.

- (2.5 puntos) Reducción de dimensiones: Aplique las siguientes técnicas al conjunto de datos resultante del punto anterior. Dichos procesos no deben de ejecutarse de manera secuencial, es decir, no dependen uno de otro.
  – Utilizando PCA, reduzca el conjunto de datos a 2 dimensiones y realice una visualización del conjunto.
  – Utilizando *Clustering de variables*, remueva multicolinealidad.
  – Seleccione las mejores variables utilizando *SelectKBest*.
  – Seleccione las mejores variables utilizando *WoE* y *IV*, solo si aplica.

Los últimos dos puntos solo aplican si se construyó una variable objetivo.

- (1.5 puntos) BI: Adicional al desarrollo anterior, construya un tablero sencillo de BI en la herramienta de su preferencia. Priorice el uso de variables geográficas. Este punto es opcional.

Cuestionario

- (0.5 puntos) ¿Por qué Excel no es una Base de Datos? Elabore.
- (0.5 puntos) ¿Qué diferencia hay entre un Ingeniero de Datos, un Científico de Datos y un Arquitecto de Datos?
- (0.5 puntos) ¿Cómo reduce dimensiones PCA?
- (0.5 puntos) ¿Cuál es la diferencia entre *importancia de variables* y *poder predictivo*?

Feedback

(0 puntos) Por favor, aporte comentarios sobre el avance del curso, el ponente y las clases. El objetivo es poder mejorar los contenidos y el desarrollo del diplomado.

Entregables

- [.ipynb] Código utilizado para la construcción de la solución.
- [.png] Imagenes del conjunto de datos reducido con PCA.
- [.csv] Tablas analíticas de datos identificadas como *restaurantes.csv* y *usuarios.csv*.
- [url/.pdf] Tablero dinámico funcional. Esto solo en caso de haber realizado

tablero.

• [.pdf] Elaboración de un documento entregando todos los puntos.

Consideraciones

• Todos los rubros son obligatorios, el único del cual se puede prescindir es de la sección correspondiente a BI.

• La calificación máxima a obtener en el examen es de 12.5 considerando la sección de BI, sin ella, de 11.

• El feedback no es obligatorio para poder asentar calificación.
• Las únicas arquitecturas válidas a utilizar son aquellas vistas en clase.
• El examen debe resolverse en Python.

• Examine el uso (o no) de todas las fuentes de datos.

• El examen debe resolverse completamente de manera individual. Cualquier sospecha o con formación de trabajo compartido anulará la calificación del examen.

• La entrega debe encontrarse en perfecto orden y de forma entendible.

• El código no debe contener ningún error.

• Todas las dudas sobre el examen deben de realizarse en el grupo de WhatsApp, a reserva de casos extremos.

• La fecha límite para entregar el examen es el 11 de diciembre de 2025 a las 00:00 horas.