Universidad de los Andes Colombia — Acreditación institucional de alta calidad 10 años — Mineducación Resolución 582 del 9 de enero de 2015	**Maestría en Ingeniería de Información (MINE)** **Ingeniería de Sistemas y Computación** **Curso: MINE 4101-Ciencia de datos aplicada** **Parcial I** **Periodo: 2021-20**

"Al entregar la solución de este examen, yo, **[nombre y apellido]** con código [#] me comprometo a no conversar durante el desarrollo de este examen con ninguna persona diferente a los profesores del curso sobre aspectos relacionados con el examen; tampoco utilizaré algún medio de comunicación por voz, texto o intercambio de archivos, para consultar o compartir con otros, información sobre el tema del examen antes de entregarlo. Soy consciente y acepto las consecuencias que acarreará para mi desempeño académico cometer fraude en este examen".

Reglas

1. El examen es estrictamente individual y tiene 150 minutos para resolverlo.
2. Utilice procedimientos explícitos y justifique su respuesta.
3. Este examen deber ser entregado por BloqueNeon en el enlace destinado para tal fin, hasta la fecha máxima establecida (30 de septiembre, 9:15pm).
4. Se debe entregar un archivo en formato html en el espacio destinado para este fin en BloqueNeón

Enunciado del Parcial

Descripción del caso:

La cafeína puede provocar un aumento breve pero drástico de la presión arterial, incluso en personas que no tienen presión arterial alta. Sin embargo, no es claro qué causa este aumento en la presión arterial. Además, la respuesta de la presión arterial a la cafeína difiere de una persona a otra. Algunas personas que beben regularmente bebidas con cafeína tienen una presión arterial promedio más alta que las que no beben ninguna. Otros que beben bebidas con cafeína con regularidad desarrollan tolerancia a la cafeína. Como resultado, la cafeína no tiene un efecto a largo plazo sobre la presión arterial.

El efecto del café y la cafeína sobre la presión arterial (PA) y las enfermedades cardiovasculares (ECV) es incierto. Se le ha encomendado el estudio de los efectos a largo plazo de la ingesta de cafeína y café sobre la PA y la asociación entre el consumo habitual de café y el riesgo de ECV.

Para esto se ha desarrollado un estudio sobre 100 000 personas. Se les ha preguntado sobre su consumo promedio de café diario (en tazas de café) y se ha medido su presión arterial sistólica. Se han incluido otros datos como edad, colesterol en sangre, índice de masa corporal (imc), identidad racial/étnica (0 – blanco, 1- negro, 2 – hispano), género (masculino o femenino) y si es fumador o no.

I. [25%] Exploración de los datos

A partir de la exploración inicial de los datos:

1. **[10%]** Describa el resultado de la exploración de datos (distribuciones de una variable) e indique el tipo de cada variable (categórica ordinal, categórica nominal, cuantitativa).
2. **[10%]** Describa los problemas de calidad de los datos (datos faltantes, datos duplicados) y su estrategia para corregirlos. Indique si existe algún patrón en los datos faltantes. Puede usar la librería *missingno* de Python.
3. **[10%]** A partir de visualizaciones bivariadas, responda ¿El alto consumo de café puede estar relacionado con una mayor presión sistólica? Identifique otros factores que pueden estar asociados con una mayor presión sistólica. **Nota:** La presión sistólica se considera normal hasta el valor de 120.

II. [30%] Pruebas de hipótesis y correlación

4. **[15%]** Analice dos factores que pueden estar relacionados con diferencias en el promedio de presión sistólica observado. Exprese la prueba de hipótesis y valídela usando pruebas por permutación o t-test. Escriba sus conclusiones de las pruebas.
5. **[10%]** Calcule y analice la correlación entre las variables y la presión sistólica. Escriba sus conclusiones al respecto. ¿Existen factores protectores? ¿Cuáles factores de riesgo puede descubrir?
6. **[5%]** ¿Existen problemas de multicolinealidad en los datos? ¿Cómo se pueden interpretar? ¿Qué problemas se pueden presentar en el análisis? ¿Cómo puede corregirlos? Sea específico con respecto al caso de estudio.

III. [45%] Análisis con regresión lineal

Para los siguientes puntos recuerde separar el conjunto de datos en datos de entrenamiento, validación y pruebas. También puede usar validación cruzada.

1. **[15%]** ¿Existe una relación entre el número de tazas de café y la presión sistólica? Utilice una regresión lineal simple (con una sola variable dependiente) para responder esta pregunta. Analice el resultado, interpretando el intercepto y el coeficiente, las medidas de error y ajuste y haciendo una prueba de hipótesis sobre el coeficiente de la variable dependiente.
2. **[25%]** Teniendo en cuenta el análisis de los datos, realice una regresión lineal multivariable, intentando obtener una mayor medida de ajuste (R cuadrado). Puede realizar transformaciones a las variables, crear nuevas características y hacer selección de características. Justifique sus decisiones. Cuando seleccione un modelo, reporte el error sobre los datos de prueba.
   Considerando el resultado, ¿existe una relación entre el número de tazas de café y la presión sistólica de la persona?
3. **[5%]** De acuerdo con su análisis, ¿qué recomendaciones podrían darse a las personas para reducir el riesgo de presión alta?

Descripción de variables

Variable	Descripción
id	Un identificador de la persona que participó en el estudio
tazas_cafe	Número de tazas de café al día que toma la persona en promedio (según reporte de la persona)
Edad	Edad de la persona al momento del estudio
Genero	Genero reportado de la persona
Fumador	0 – No fumador 1- Sí fumador
Imc	Índice de masa corporal de la persona. IMC = $\frac{peso}{altura^2}$
Tc	Colesterol en sangre el día del estudio El colesterol en la sangre es una medida compuesta del colesterol LDL, HDL y 20% de los triglicéridos.
Etnicidad	Auto-reporte de la identidad étnica o racial. No se permitieron varias selecciones.
Presión_sis	Presión sistólica el día del estudio.

Nota: estos datos fueron simulados para este parcial