Programa de Bioingeniera

Probabilidad y estadística

Primer proyecto

Semestre 2025 -2

**Fecha de entrega: 24 de octubre 2025**

ANÁLISIS EXPLORATORIO Y ESTADÍSTICA INFERENCIAL

**Conjunto de datos: Parkinson's Disease Dataset**

Este conjunto de datos comprende información de salud integral de 2.105 pacientes diagnosticados con la enfermedad de Parkinson, cada uno identificado de forma única con IDs que van del 3058 al 5162. El conjunto incluye detalles demográficos, factores de estilo de vida, antecedentes médicos, mediciones clínicas, evaluaciones cognitivas y funcionales, síntomas y un indicador de diagnóstico.

**Información del paciente**

* *PatientID*: Identificador único asignado a cada paciente (3058 a 5162).

**Detalles demográficos**

* *Age:* Edad de los pacientes, entre 59 y 90 años.
* *Gender:* Género de los pacientes, donde 0 representa Hombre y 1 representa Mujer.
* *Ethnicity:* Etnia de los pacientes, codificada como:

0: Caucásico

1: Afroamericano

2: Asiático

3: Otro

* *EducationLevel:* Nivel educativo de los pacientes, codificado como:

0: Ninguno

1: Secundaria

2: Pregrado

3: Posgrado

**Factores de estilo de vida**

* *BMI:* Índice de masa corporal, entre 15 a 40.
* *Smoking:* Estado de tabaquismo, donde 0 indica No y 1 indica Sí.
* *AlcoholConsumption:* Consumo de alcohol semanal en unidades, entre 0 a 20.
* *PhysicalActivity:* Actividad física semanal en horas, entre 0 a 10.
* *DietQuality:* Puntuación de calidad de la dieta, entre 0 a 10.
* ***SleepQuality:*** Puntuación de calidad del sueño, entre 4 a 10.

**Historial médico**

* *FamilyHistoryParkinsons:* Antecedentes familiares de enfermedad de Parkinson, donde 0 indica No y 1 indica Sí.
* *TraumaticBrainInjury:* Antecedente de lesión cerebral traumática, donde 0 indica No y 1 indica Sí.
* *Hypertension:* Presencia de hipertensión, donde 0 indica No y 1 indica Sí.
* *Diabetes:* Presencia de diabetes, donde 0 indica No y 1 indica Sí.
* *Depression:* Presencia de depresión, donde 0 indica No y 1 indica Sí.
* *Stroke:* Antecedente de accidente cerebrovascular, donde 0 indica No y 1 indica Sí.

**Mediciones clínicas**

* *SystolicBP:* Presión arterial sistólica, entre 90 a 180 mmHg.
* *DiastolicBP:* Presión arterial diastólica, entre 60 a 120 mmHg.
* *CholesterolTotal:* Colesterol total, entre 150 a 300 mg/dL.
* *CholesterolLDL:* Colesterol LDL (lipoproteínas de baja densidad), entre 50 a 200 mg/dL.
* *CholesterolHDL:* Colesterol HDL (lipoproteínas de alta densidad), entre 20 a 100 mg/dL.
* *CholesterolTriglycerides:* Triglicéridos, entre 50 a 400 mg/dL.

**Evaluaciones cognitivas y funcionales**

* ***UPDRS:*** Puntuación de la Escala Unificada de Evaluación de la Enfermedad de Parkinson, de 0 a 199. Puntajes más altos indican mayor gravedad de la enfermedad.
* ***MoCA:*** Puntuación de la Evaluación Cognitiva de Montreal, de 0 a 30. Puntajes más bajos indican deterioro cognitivo.
* ***FunctionalAssessment:*** Puntuación de evaluación funcional, de 0 a 10. Puntajes más bajos indican mayor deterioro.

**Síntomas**

* *Tremor:* Presencia de temblor, donde 0 indica No y 1 indica Sí.
* *Rigidity:* Presencia de rigidez muscular, donde 0 indica No y 1 indica Sí.
* *Bradykinesia:* Presencia de bradicinesia (lentitud de movimiento), donde 0 indica No y 1 indica Sí.
* *PosturalInstability:* Presencia de inestabilidad postural, donde 0 indica No y 1 indica Sí.
* *SpeechProblems:* Presencia de problemas del habla, donde 0 indica No y 1 indica Sí.
* *SleepDisorders:* Presencia de trastornos del sueño, donde 0 indica No y 1 indica Sí.
* *Constipation:* Presencia de estreñimiento, donde 0 indica No y 1 indica Sí.

**Información del diagnóstico**

* *Diagnosis:* Estado de diagnóstico de la enfermedad de Parkinson, donde 0 indica No y 1 indica Sí.

**Información confidencial**

* *DoctorInCharge:* Esta columna contiene información confidencial sobre el médico responsable, con "XXXConfid" como valor para todos los pacientes.

El conjunto de datos completo se encuentra disponible para su consulta en el siguiente enlace: <https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis>

**PROCEDIMIENTO**

**Análisis de datos:**

1. Análisis de variables cuantitativas: Seleccione **cinco** variables cuantitativas del conjunto de datos y para **Cada una** realice lo siguiente:

a. Valor mínimo y máximo, medidas de tendencia central, variabilidad, forma y posición.

b. Tabla de frecuencias.

c. Gráfico de caja y bigotes.

d. Histograma.

e. Gráfico de densidad.

g. Identificar si existen valores atípicos en las variables, ¿estos valores podrían generar un falso análisis?

1. Análisis de variables cualitativas: Seleccione **cinco** variables cualitativas del conjunto de datos y para **Cada una** realice lo siguiente:

a. Tabla de frecuencias

b. Gráfico de barras

c. Gráfico de torta

1. Construya la matriz de correlación y el análisis de covarianza usando todas las variables cuantitativas del conjunto de datos. Analicen qué relaciones son fuertes o débiles, positivas o negativas, y comente si los resultados tienen sentido en el contexto de los datos.
2. Utilizar groupby en Python para calcular medidas estadísticas (media, máximo, mínimo, desviación estándar, varianza, conteo de valores nulos) sobre las variables:

['UPDRS', 'FunctionalAssessment', 'MoCA’, SleepQuality’], agrupando por Diagnosis (grupo con y sin Parkinson) y Gender.

Ejemplo:

Imagen que contiene Texto

El contenido generado por IA puede ser incorrecto.

1. Dividir el conjunto de datos según el diagnóstico (Diagnosis) y realizar las siguientes visualizaciones:

**a. Gráfico de barras**: Distribución de pacientes por diagnóstico (con/sin Parkinson) y género.

**b. Histogramas y gráficos de densidad**: Para las siguientes variables: UPDRS, FunctionalAssessment, MoCA y SleepQuality diferenciando por grupo diagnóstico (con/sin Parkinson).

**c. Boxplots**: Comparación de las variables cognitivas por grupo (UPDRS, FunctionalAssessment, MoCA y SleepQuality) diagnóstico (con/sin Parkinson) y género (hombre/mujer).

**Ejemplo**:

**Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.**

**Análisis gráfico**: Cada gráfico deberá estar acompañado de una interpretación clara que permita inferir patrones, diferencias o tendencias relevantes

**Estadística inferencial:**

En el marco de nuestra investigación sobre la enfermedad de Parkinson y el deterioro cognitivo y funcional asociado, se solicita realizar pruebas estadísticas de comparación utilizando el conjunto de datos disponible. El objetivo de estas pruebas es identificar diferencias significativas entre grupos y variables clave, con el fin de profundizar en la comprensión de la enfermedad y sus factores relacionados.

1. **Comparación de UPDRS según diagnóstico**  
   Realizar una prueba estadística para comparar las puntuaciones del **UPDRS** entre los grupos con y sin diagnóstico de Parkinson.
2. **Comparación de MoCA por género (sin Parkinson)**  
   Filtrar los pacientes **sin diagnóstico de Parkinson** y analizar si existen diferencias en la variable **MoCA** entre hombres y mujeres.
3. **Efecto del nivel educativo y género sobre UPDRS**  
   Analizar mediante pruebas estadísticas si las variables independientes **EducationLevel** y **Gender** influyen significativamente sobre la puntuación **UPDRS**.
4. **Hipótesis propia**  
   Plantee una hipótesis basada en el análisis exploratorio del conjunto de datos. A continuación, siga los pasos descritos:

**a. Visualización preliminar:**

Genere primero una visualización pertinente (como un histograma, boxplot, gráfico de barras, etc.) que le permita observar el comportamiento de las variables involucradas en su hipótesis. Esta visualización debe servir como base para formular su hipótesis y justificar la necesidad del análisis estadístico.

**b. Formulación de hipótesis:**

Defina claramente la hipótesis nula (H₀) y la hipótesis alternativa (H₁), de acuerdo con el objetivo planteado y el tipo de variables consideradas.

**c. Selección de la prueba estadística:**

Justifique la elección de la prueba estadística más adecuada para contrastar su hipótesis (por ejemplo: prueba t de Student, ANOVA, etc.), considerando el tipo de datos, su distribución y la presencia o no de homogeneidad de varianza.

**d. Evaluación de supuestos:**

Verifique los supuestos requeridos por la prueba elegida, tales como normalidad de los datos y homocedasticidad.

**e. Análisis de significancia:**

Establezca el nivel de significancia (por ejemplo, α = 0.05) y explique las implicaciones de los posibles errores tipo I y tipo II.

**f. Interpretación de resultados:**

Interprete los resultados obtenidos y determine si puede rechazar o no la hipótesis nula.

**g. Sustentación teórica:**

Fundamente sus conclusiones con referencias bibliográficas actualizadas y relevantes que respalden su interpretación desde el punto de vista clínico, psicológico o epidemiológico según corresponda.

Para cada una de las comparaciones solicitadas tenga en cuenta:

a. Explicar por qué se elige cada prueba estadística en función del tipo de datos, número de grupos, y escalas de medición.

b. Verificación de supuestos: Evaluar los supuestos de normalidad y homogeneidad de varianzas utilizando pruebas como Shapiro-Wilk, Levene o similares.

c. Nivel de significancia: Indicar el nivel de significancia utilizado (por ejemplo, α = 0.05) y justificar su elección en términos del equilibrio entre errores tipo I y tipo II.

d. Formulación de hipótesis: Establecer claramente la hipótesis nula (H₀) y la alternativa (H₁) para cada comparación o modelo estadístico.

**PARÁMETROS DE ENTREGA**

* Formato de entrega: Se deben entregar dos archivos obligatorios:
  + Un notebook (Jupyter, Colab, o equivalente) con el código completo y comentado.
  + Un documento en formato Word (.docx) que contenga los análisis, gráficos, resultados e interpretaciones de manera ordenada y formal.
* El informe debe contener todas conclusiones de cada uno de los resultados presentados, sin el apartado de conclusiones no se revisa. Utilizar el lenguaje estadístico apropiado para establecer las conclusiones.
* Finalmente, la nota estará definida por el trabajo escrito y la sustentación que será presencial, divididos así: Trabajo escrito 70%, sustentación 30%.
* Para la sustentación, se recomienda preparar una **presentación con diapositivas** que incluyan las gráficas generadas y los resultados obtenidos durante el análisis.

**NOTA: La nota del trabajo se define con la sustentación. Trabajo sin sustentación no se califica.**