



## “CURSO DE PYTHON INTERMEDIO Y AVANZADO PARA CIENCIA DE DATOS”

Este curso está basado en conocer los fundamentos de la programación mediante el uso de la herramienta PYTHON, que cuenta con diferentes bibliotecas con funcionalidades estadísticas y gráficas, que permite la programación orientada a Objetos, importación y conexión a base de datos, manipulación de los Datos, técnicas para el preprocesamiento, entendimiento, limpieza, transformación de los datos, técnicas para la visualización e identificación de datos inconsistentes y Outliers, balanceo y Discretización de datos, así como técnicas para la reducción de la dimensionalidad de los datos todo ello con un enfoque hacia la Ciencia de los Datos.

### CONTENIDO:

1. ENTENDIMIENTO Y CODIFICACIÓN DE DATOS (3H)
  - Variables e indicadores
  - Tipos de datos
  - Análisis exploratorio datos
  - Elaboración de Diagrama de Barras.
  - Elaboración de Histogramas.
  - Elaboración de Gráfico de Sectores Circulares.
  - Elaboración de Diagramas de dispersión
  
2. IMPUTACIÓN Y TRATAMIENTO DE OUTLIERS (3H)
  - Identificación de datos faltantes.
  - Visualización de datos faltantes.
  - Imputación de datos faltantes.
  - Identificación de Outliers.
  - Tratamiento de Outliers.
  - Ejercicio de aplicación
  
3. TRATAMIENTO DE INCONSISTENCIAS Y DATA NO BALANCEADA (3H)
  - Tratamiento de data no balanceada.
  - Balanceo de datos.
  - Estrategias de balanceo.
  - Ejercicio de aplicación
  
4. DISCRETIZACIÓN DE DATOS (3H)
  - Discretización supervisada (por Entropía y ChiMerge)
  - Discretización no supervisada (intervalos de igual amplitud e intervalos de igual frecuencia)
  - Discretización con intervalos de igual amplitud.
  - Discretización con intervalos de igual frecuencia.
  - Discretización k medias.
  - Método de Entropía.
  - Método CHIMerge
  - Ejercicio de aplicación
  
5. ESCALAMIENTO Y ANÁLISIS DE COMPONENTES PRINCIPALES (3H)



- Introducción al análisis de componentes principales.
- ¿Qué es el Análisis de Componentes Principales (PCA)?
- ¿Por qué usar PCA?
- Prueba de esfericidad de Barlett.
- Estimación de los componentes principales.
- Varianza total y varianza explicada.
- Interpretación de los componentes principales.
- Ejercicio de aplicación

#### 6. MÉTODOS DE PARTICIONAMIENTO (3H)

- Introducción a los métodos de particionamiento jerárquico.
- ¿Qué es el análisis clúster?
- ¿Qué no es el análisis clúster?
- Medidas de Similaridad, distancia y proximidad.
- Distancia Euclidiana.
- Segmentación con Método K – medias.
- Estableciendo el número óptimo de clúster.
- Suma de cuadrados, silueta y Calinski y Harabasz.

#### 7. TÉCNICAS NO SUPERVISADAS - ANÁLISIS CLÚSTER (3H)

- Elaboración de la matriz de distancias (disimilaridad).
- Método k-medias.
- Procesamiento de datos y aplicación de Método k-medias.
- Caso de uso aplicación técnica Clustering,
- Identificación del número óptimo de clúster.
- Suma de cuadrados, silueta y Calinski y Harabasz.
- Perfilamiento del clúster.

#### 8. TÉCNICAS PAM - CLARA - FANNY (3H)

- Métodos de particionamiento alrededor de Medoides (PAM).
- Clustering in Large Applications (CLARA)
- Fuzzy Analysis (FANNY)
- Identificación del número óptimo de clúster.
- Suma de cuadrados, silueta y Calinski y Harabasz.

#### 9. MÉTODOS JERÁRQUICOS AGLOMERATIVO Y DIVISIVO (3H)

- Método jerárquico aglomerativo: AGNES.
- Métodos jerárquico divisivos: DIANA.
- Estableciendo el número óptimo de clúster.
- Suma de cuadrados, silueta y Calinski y Harabasz.
- Ejercicio de aplicación.

#### 10. REGLAS DE ASOCIACIÓN (3H)

- Conceptos: soporte, confianza, lift
- Algoritmo para la búsqueda de asociaciones: el algoritmo a priori
- Medidas para la elección de reglas de asociación
- Interpretación de resultados

#### 11. REGRESIÓN LINEAL SIMPLE (3H)



- Supuestos para el modelo de regresión lineal.
- Modelo y estimación de coeficientes.
- Indicadores MSE, RMSE, SSE, SST.
- Ejercicio de aplicación.

#### 12. REGRESIÓN LOGÍSTICA (3H)

- Introducción a la regresión logística.
- Modelos y Técnicas de selección de variables.
- Evaluación Modelo haciendo uso de las métricas Matriz de Confusión, accuracy, sensibilidad, especificidad, AUC y curva ROC.
- Ejercicio de aplicación

#### 13. ANÁLISIS DISCRIMINANTE LINEAL (3H)

- Introducción al análisis discriminante lineal (LDA).
- Modelo y estimación de coeficientes.
- Técnica de selección de variables (Lambda de Wilks y tasa de acierto).
- Ejercicio de aplicación.

#### 14. ANÁLISIS DISCRIMINANTE CUADRÁTICO (3H)

- Introducción al análisis discriminante Cuadrático (QDA).
- Modelo y estimación de coeficientes.
- Evaluación Modelo haciendo uso de las métricas de accuracy, sensibilidad, especificidad, AUC y curva ROC.
- Ejercicio de aplicación.

#### 15. APRENDIZAJE BASADO EN ÁRBOLES DE DECISIÓN I (3H)

- Introducción a los árboles de decisión
- Creación y lógica de los árboles de decisión.
- Árboles de decisión basado en el algoritmo CART.
- Medidas de impureza.

#### 16. APRENDIZAJE BASADO EN ÁRBOLES DE DECISIÓN II (3H)

- Combinar arboles de Decisión mediante bosques aleatorios.
- Criterios de particiones, parada
- Poda de los árboles de decisión.
- Maximizar la ganancia en la información.
- Ejercicio de aplicación.

#### 17. RANDOM FOREST (3H)

- Introducción del Algoritmo Random Forest.
- Características del Algoritmo Random Forest.
- Selección de variables en el Algoritmo Random Forest.
- Mediciones del Algoritmo Random Forest
- Ventajas y Desventajas del Algoritmo Random Forest

#### 18. MÁQUINAS DE SOPORTE VECTORIAL (3H)

- Definición de Máquinas de Soporte vectorial (SVM).
- Casos no separables linealmente.
- Definición de hiperplanos, Kernel.
- Tipos de Kernel (Lineal, Radial, Polinómico)



Colegio de Estadísticos del Perú  
Consejo Regional de Lima

- Ejercicio de aplicación

#### 19. REDES NEURONALES (3H)

- Introducción a redes neuronales artificiales
- Funcionamiento de una red neuronal
- Paralelización de entrenamiento de redes neuronales

#### 20. TEXT MINING (3H)

- Introducción a las técnicas de Text Mining
- Procesos de Text Mining
- Terminología de Text Mining
- Uso de TF-IDF
- Nube de palabras

#### 21. EVALUACIÓN DEL CURSO

### **REQUISITOS:**

TENER CONOCIMIENTOS BÁSICOS DE PYTHON. TENER EN CUENTA QUE HAY 24 HORAS DE CLASE DE PYTHON BÁSICO COLGADO EN LA PAGINA WEB DEL COESPE LIMA.

### **INICIO:**

15 DE OCTUBRE

### **HORARIO:**

MIÉRCOLES Y VIERNES DE 7:30 A 10:30PM.

### **DURACION:**

60 HORAS

### **MODALIDAD:**

VIRTUAL

### **COSTO:**

GRATUITO PARA LOS MIEMBROS DEL COESPE DE TODAS LAS REGIONES.

### **CERTIFICADO:**

SE OTORGARÁ 01 CERTIFICADO APROBATORIO CON QR A LOS QUE PASEN LA EVALUACION FINAL; 01 CERTIFICADO DE PARTICIPACION A LOS QUE NO PASEN LA EVALUACION PERO QUE COMPLETEN EL 50 % DE ASISTENCIA. PARA ELLO TENDRAN QUE CANCELAR EL MONTO DE S/ 50.00 SOLES POR ADELANTADO Y NO HAY DEVOLUCION EN CASO ABANDONE EL CURSO

### **INSCRIPCION:**

A TRAVES DEL LINK QUE SE ENCUENTRA EN LA PAGINA WEB DEL COESPE LIMA.