Pontificia Universidad Católica de Chile
Educación Profesional - Escuela de Ingeniería
Diplomado en Big Data y Ciencias de Datos
Minería de Datos
Relator: Sebastián Raveau

Evaluación 1 –Análisis de Datos y Regresión Lineal

Fecha de Entrega: domingo 16 de julio

El objetivo de esta evaluación es entender y modelar cómo el Producto Interno Bruto (PIB) per cápita de un país se puede ver afectado por factores externos. Para esto cuenta con información de 28 potenciales variables explicativas de distinta índole provenientes de 188 países. La descripción de las variables de la base de datos se encuentra en la siguiente página.

La evaluación consta de tres etapas:

1. Revisar la presencia de potenciales *outliers* en la base de datos (elija al menos cuatro variables a analizar, que le parezcan más relevantes y/o interesantes para modelar el PIB per cápita). Para esto, deberá aplicar los tres métodos vistos en clases (percentiles, intervalos de variabilidad y valor-z robusto) y comparar los resultados. En base a los resultados, debe analizar y concluir sobre la presencia de *outliers*. Recuerde que los métodos no identifican *outliers*, sino que simplemente identifican candidatos a ser *outliers*.
2. Plantear hipótesis preliminares, indicando cuáles variables explicativas cree usted que podrían contribuir a explicar el PIB per cápita y cuál debiera ser su signo. Fundamente sus hipótesis en un análisis preliminar de los datos.
3. Proponer y estimar un modelo de regresión lineal que le parezca adecuado. Explique su procedimiento, indicando a grandes rasgos los pasos que siguió para llegar a su modelo final. No es necesario que reporte todos los modelos intermedios, simplemente basta con reportar el modelo final y explicar cómo llegó a él.

Debe generar un breve reporte y entregarlo al correo: mineria.datos.PUC@gmail.com. El asunto del correo y el nombre del archivo deben comenzar con [Evaluación 1] seguido con los apellidos de los estudiantes. Por ejemplo: [Evaluación 1] Gutiérrez y Soto. El objetivo de la evaluación es demostrar que es capaz de aplicar los contenidos del curso. Por favor presente sólo la información relevante, sin llenar múltiples páginas con gráficos, códigos y estadísticas. Extensión recomendada: 10 planas de contenido (i.e. sin contar portada, índices o anexos, los cuales no son obligatorios).

Pontificia Universidad Católica de Chile
Educación Profesional - Escuela de Ingeniería
Diplomado en Big Data y Ciencias de Datos
Minería de Datos
Relator: Sebastián Raveau

Descripción de la Base de Datos

Variable	Descripción
PIB	PIB per cápita - miles de millones de USD / habitante
POB	Población - millones
IDH	Índice de desarrollo humano
GINI	Coeficiente de Gini
IPC	Índice del precio al consumidor
FAO	Índice de precios alimenticios de la FAO
GENERO	Índice de desigualdad de género
ELECTRICIDAD	Tasa de electrificación
ESCOLARIDAD	Años de escolaridad promedio
SUICIDIOFEM	Tasa de suicidios femeninos - cada 100.000 personas
SUICIDIOMAS	Tasa de suicidios masculinos - cada 100.000 personas
BOSQUE	Porcentaje de superficie que corresponde a bosques
FOSIL	Porcentaje de uso de combustibles fósiles
DIOXIDO	Emisiones de dióxido de carbono - Toneladas per cápita
DESASTRE	Población afectada por desastres naturales - miles
AFECTADOS	Población sin hogar por desastres naturales - miles
HOMICIDIO	Tasa de homicidios - cada 100.000 personas
MORTINF	Mortalidad de niños menores de 5 años - miles
MORTMAT	Tasa de mortalidad maternal - cada 100 nacimientos
TURISMO	Turistas internacionales - millones
INTERNET	Porcentaje de uso de internet
VIOLENCIA	Porcentaje de víctimas de violencia entre parejas
VIDA	Expectativa de vida - años
CELULAR	Subscripciones a telefonía celular - cada 100 personas
DESERCION	Tasa de deserción escolar primaria
PRISION	Tasa de encarcelamiento - 100.000 personas
RENOVABLE	Porcentaje de uso de energías renovables
PARLAMENTO	Porcentaje de escaños parlamentarios ocupados por mujeres
INMIGRANTES	Porcentaje de población que es inmigrante

Su entrega será evaluada de acuerdo con los siguientes aspectos:

Claridad [1,5 puntos]

Se espera que el reporte sea claro y bien redactado, con las ideas debidamente desarrolladas. Las tablas y figuras deben estar debidamente presentadas y explicadas. El procedimiento de modelación se explica adecuadamente. El lenguaje utilizado debe ser apropiado para un reporte técnico.

Aplicación de métodos de detección de *outliers* [1,5 puntos]

Se deben aplicar los contenidos vistos en la clase de manera adecuada. Se deben justificar los parámetros utilizados en cada método. El resultado de los métodos (i.e. cuántas y cuáles observaciones son potenciales *outliers*) debe ser explícito. Se comparan los resultados de cada método.

Análisis preliminar [1 punto]

Se analiza y discute el impacto esperado de las potenciales variables explicativas. Se analiza y discuten potenciales transformaciones de variables explicativas (ya sean no-lineales o categóricas). Se plantean hipótesis a ser probadas en la modelación

Modelación [2 puntos]

La describe la selección de variables explicativas. Se describe el proceso de especificación del modelo. Se analizan los resultados del modelo.