Electiva Aprendizaje automático

Maestría en Gestión de Tecnologías de Información y Conocimiento

Facultad de Ingeniería

Universidad de Nariño

Guía 2

Análisis Exploratorio

El conjunto de datos Sick, el cual registra enfermedades de la tiroides, proporcionados por: el Instituto Garavan, Instituto de Nueva Gales del Sur, Syndney, Australia. El archivo contiene estudios de pacientes, donde se pretende determinar si tienen o no una enfermedad. Clase (1) Enfermo Clase (0) No Enfermo.

Se han registrado sobre 955 individuos las variables siguientes:

caso	Número de registro
edad	Edad del paciente
genero	Genero del paciente
depresion	El paciente tiene depresión
embarazada	Paciente en estado de embarazo
cirugía_de_tiroides	Tiene cirugías de tiroides
nivel_TSH	nivel TSH
nivel_T3	nivel T3
nivel_TT4	nivel TT4
nivel_T4U	nivel T4U
nivel_FTI	nivel FTI
clase	Enfermo tiroides 1

Guía

1. Clasificar las variables de la base.
2. Construir la tabla de frecuencias y un diagrama de barras de las variables categóricas.
3. Calcular las medidas de resumen de las variables cuantitativas.
4. ¿Cuáles de las variables le parecen simétricas a partir de estos resúmenes? Confirme estas observaciones mediante un boxplot.
5. Calcular la desviación intercuartil y detectar presencia de valores salvajes moderados y severos en cada una de las variables cuantitativas.
6. Realizar un boxplot comparativo para cada una de estas variables particionando por la clase. ¿Le parece que alguna de estas variables está relacionada con la enfermedad, es decir que toma valores muy distintos en ambos grupos? Analizar en todos los casos la presencia de outliers.
7. Hallar las matrices de varianzas y covarianzas y de correlaciones.
8. Aplicar la estandarización z-score y calcular la matriz de varianzas y covarianzas.
9. Graficar un dispersograma y un mapa de calor con la matriz de correlaciones.

Actividad

El conjunto de datos Ecaes.csv registra los resultados de las pruebas Ecaes en estudiantes de los programas de INGENIERIA CIVIL, INGENIERIA ELECTRONICA, INGENIERIA DE SISTEMAS, INGENIERIA AGRONOMICA, INGENIERIA EN PRODUCCION ACUICOLA, INGENIERIA AMBIENTAL e INGENIERIA AGROFORESTAL de la universidad de Nariño en el año 2018.

El archivo contiene 220 estudiantes sobre los cuales se han medido las siguientes variables:

ESTU_CONSECUTIVO	Identificador único de registro
ESTU_GENERO	Genero del estudiante
ESTU_ESTADOCIVIL	Estado civil estudiante
FAMI_HOGARACTUAL	Tipo de hogar del estudiante
FAMI_CABEZAFAMILIA	Es cabeza de familia
FAMI_NUMPERSONASACARGO	Número de personas a cargo
FAMI_ESTRATOVIVIENDA	Estrato de la vivienda
FAMI_PERSONASHOGAR	Número de personas en el hogar
FAMI_CUARTOSHOGAR	Número de cuartos del hogar
FAMI_TIENECOMPUTADOR	Tiene computador
ESTU_HORASSEMANATRABAJA	Horas semanales de trabajo
INST_NOMBRE_INSTITUCION	Nombre institución
ESTU_PRGM_DEPARTAMENTO	Departamento del programa
ESTU_PRGM_ACADEMICO	Nombre del programa
MOD_RAZONA_CUANTITAT_PUNT	puntaje de razonamiento
MOD_LECTURA_CRITICA_PUNT	puntaje de lectura
MOD_COMPETEN_CIUDADA_PUNT	puntaje competencias ciudadanas
MOD_INGLES_PUNT	puntaje ingles
PUNT_GLOBAL	puntaje promedio

1. ¿Cuáles variables de la base le parecen categóricas?
2. ¿Cuáles variables de la base le parecen cuantitativas discretas?
3. ¿Cuáles variables de la base le parecen cuantitativas continuas?
4. Construir la tabla de frecuencias y un diagrama de barras de las variables categóricas y las cuantitativas discretas.
5. Calcular las medidas de resumen de las variables cuantitativas.
6. Confirme la presencia de outliers graficando los boxplot de cada variable cuantitativa.
7. ¿Cuáles de las variables le parecen simétricas a partir de estos resúmenes? Confirme estas observaciones mediante un boxplot.
8. Encuentre outliers moderados y severos de cada variable cuantitativa continua.
9. Realizar un boxplot comparativo para el puntaje global discriminando por la variable género. ¿Qué genero obtuvo mejores resultados?
10. Realizar un boxplot comparativo para el puntaje global discriminando por la variable FAMI_ESTRATOVIVIENDA. ¿Qué estrato obtuvo mejores resultados? ¿Cuál tiene una mediana mayor?
11. Realizar un boxplot comparativo para el puntaje de razonamiento cuantitativo discriminados por programa. ¿Qué programa obtuvo mejores resultados en la prueba?
12. Hallar las matrices de varianzas y covarianzas y de correlaciones.

13. Graficar un dispersograma y un mapa de calor con la matriz de correlaciones. ¿Qué variables están correlacionadas positivamente?, ¿cuáles negativamente? y ¿cuáles consideran que no tienen correlación?