**UNIVERSIDAD TÉCNICA NACIONAL**

**CARRERA: INGENIERÍA DEL SOFTWARE**

CURSO: MINERÍA DE DATOS

**I PARCIAL- CÓDIGO: ISW-911**

**Instrucciones generales:**

1. Ningún grupo podrá contener respuestas parecidas o iguales por que será calificado con un 0 y una carta al expediente
2. El tiempo máximo para realizar la prueba es hasta el próximo miércoles 12 a las 11:59:59 pm
3. **Trabaje en orden y en silencio.**
4. **Total de puntos del examen 100pts.**

**Desarrollo: Resuelva lo que le solicita el siguiente enunciado** (100 pts.).

Recuerde que debe entregar un **notebook en R** utilizando la metodología **CRISP-DM**

**Tabla de calificación**

|  |  |  |
| --- | --- | --- |
| ***Rubro*** | ***Valor*** | ***Obtenido*** |
| **Análisis de problema** | 10pts |  |
| **Entendimiento de los datos** | 10pts |  |
| **Exploración de los datos**  **(Al menos 3 gráficos)** | 25pts |  |
| **Modelado del algoritmo** | 30pts |  |
| **Evaluación** | 15pts |  |
| **Conclusiones** | 10pts |  |
| ***Total*** | 100pts |  |

**Enunciado**

Se desea **agrupar** las películas en 3 grupos partir de las características presentadas más adelante en este enunciado, para esto se obtuvieron 6820 películas entre los años 1986-2016 del sitio web IMDb. **20pts**

**budget**: - El presupuesto de una película. Algunas películas no tienen esto, por lo que aparece como 0

**company**: - La productora

**country**: - País de origen

**director**: - El director

**genre**: - Género principal de la película.

**gross**: - Los ingresos de la película

**name**: - Nombre de la película

**rating**: - Clasificación de la película (R, PG, etc.)

**released**: - Fecha de lanzamiento (YYYY-MM-DD)

**runtime**: - Duración de la película

**score**: - Calificación de usuario de IMDb

**votes**: - Número de votos de los usuarios

**star**: - Actor principal / actriz

**writer**: - Escritor de la película

**year**: - Año de lanzamiento

**Enunciado 2**

Se desea **predecir** el precio de las casas en el área de Boston Mass a partir de las características presentadas más adelante en este enunciado. Recuerde que se quiere predecir utilizando la mayor cantidad de variables. Por favor entregue el mejor modelo de predicción con las características presentadas. Este conjunto de datos contiene información recopilada por el Servicio de Censos de los EE. **40pts**

**Características:**

**CRIM:** - tasa de delincuencia per cápita por ciudad

**ZN**: - proporción de tierra residencial zonificada para lotes de más de 25,000 pies cuadrados.

**INDUS:** - proporción de acres de negocios no minoristas por ciudad.

**CHAS:** - Variable ficticia del río Charles (1 si el trecho delimita al río;

**NOX:** - concentración de óxidos nítricos (partes por 10 millones)

**RM:** - número promedio de habitaciones por vivienda

**EDAD**: - proporción de unidades ocupadas por el propietario construidas antes de 1940

**DIS**: - distancias ponderadas a cinco centros de empleo de Boston

**RAD**: - índice de accesibilidad a autopistas radiales

**IMPUESTO**: - tasa de impuesto a la propiedad de valor total por $ 10,000

**PTRATIO**: - Proporción alumnos por profesor por ciudad.

**BLACK:** - porcentaje de negros por ciudad

**LSTAT**: porcentaje de población pobre

**MEDV**: valor medio de las viviendas ocupadas por sus propietarios en $ 1000

**Enunciado 3**

Se desea **clasificar** en 2 posibles clases los hongos (venenosos o seguros) de los hongos presentes en Norte América. Este conjunto de datos incluye descripciones de muestras hipotéticas correspondientes a 23 especies de champiñones en el hongo familiar Agaricus y Lepiota, extraído de la Guía de campo de la Sociedad Audubon sobre los hongos norteamericanos. Cada especie se identifica como definitivamente comestible, definitivamente venenosa, o de comestibilidad desconocida. Por favor entregue el mejor modelo de clasificación con las características presentadas **40pts.**

**Características:**

**Variable:** - posibles valores que tomara cada variable

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**clase**: - comestible=e, venenoso=p

**cap-shape**: - bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s

**cap-surface**: - fibrous=f, grooves=g, scaly=y, smooth=s

**cap-color**: - brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y

**bruises**: - bruises=t, no=f

**odor**: - almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s

**gill-attachment:** - attached=a, descending=d, free=f, notched=n

**gill-spacing: -** close=c, crowded=w, distant=d

**gill-size: -** broad=b, narrow=n

**gill-color:** - black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y

**stalk-shape**: - enlarging=e, tapering=t

**stalk-root:** - bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?

**stalk-surface-above-ring:** - fibrous=f, scaly=y, silky=k, smooth=s

**stalk-surface-below-ring:** - fibrous=f, scaly=y, silky=k, smooth=s

**stalk-color-above-ring:** brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y

**stalk-color-below-ring:** -brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y

**veil-type**: - partial=p, universal=u

**veil-color:** - brown=n, orange=o, white=w, yellow=y

**ring-number**: - none=n, one=o, two=t

**ring-type:** - cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z

**spore-print-color: -** black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y

**population**: - abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y

**habitat:** - grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d