Prueba - Fundamentos Data Science

Objetivo

- Implementar los contenidos aprendidos a lo largo de las 8 unidades para resolver dos problemas de carácter obligatorio.
- Se deben desarrollar dos desafíos aplicando lo aprendido en el módulo Fundamentos de Data Science.
- Ambos desafíos presentarán un enunciado a solucionar, así como una descripción de los datos disponibles a utilizar.
- Cada una de las respuestas deben considerar los requerimientos mínimos y buenas prácticas detalladas a continuación.

Consideraciones Generales

La prueba debe desarrollarse en consideración a los siguientes puntos:

- Una sección llamada **Preliminares** donde se realiza la descripción del problema y objetivos, así como explicar cómo implementarán su solución (debe considerar qué criterios de optimización y métricas de desempeño).
- Una sección llamada **Aspectos computacionales** donde se describirán las librerías y módulos a implementar, así como las funciones generadas y su objetivo.
- Una sección llamada **Descripción** donde se generará un análisis descriptivo considerando el tipo de variables (desde el punto de vista estadístico así como computacional). Esta sección debe considerar medidas univariadas/ frecuencias, datos perdidos y gráficos distributivos sobre las variables a analizar. A partir de ésta se debe clarificar la estrategia de preprocesamiento (datos perdidos, recodificaciones).
- Una sección llamada **Modelación descriptiva**, que buscará definir cuáles son los principales determinantes del objeto de estudio. En base a esta sección se podrá construír o depurar el modelo predictivo.
- Una sección llamada **Modelación predictiva**, donde se implementará una solución analítica que aumente las métricas de desempeño. Se solicitan por lo menos 3 modelos predictivos, donde deberán reportar las principales métricas. Cada modelo predictivo debe tener una reseña sobre el por qué se diseño de esa forma.

Puntuación y corrección

- Puntaje total: **20 puntos** (Cada hito equivale a 5 puntos)
- Para aprobar, se requiere un puntaje igual o superior a 16 puntos.

Requerimientos de buenas prácticas

- Todo el código debe estar escrito siguiendo las buenas prácticas de Python. Esto implica nomenclatura constante en la designación de variables, funciones y comentarios.
- Su análisis debe priorizar el uso de funciones en la medida de lo posible. Las funciones deben estar alojadas en un archivo de funciones auxiliares que debe ser importado al notebook. Las funciones deben contener un `docstring` que detalle claramente el propósito de la función, qué parámetros se ingresan y qué retorna.
- Todo output generado debe estar acompañado por una breve explicación.
- Por razones de sanitización del notebook, se recomienda incluír el módulo `warnings` para evitar avisos de deprecación.
- Puede hablar y discutir su avance con sus compañeros, pero bajo ningún motivo deben compartirse respuestas textuales y/o código. En caso de discutir su avance con su compañeros, agregue al final del notebook con quiénes colaboró en cada respuesta.

Hitos

Hito 1: Sesión Presencial 1, Unidad 7

Completar el punto de **Preliminares**, así como **Aspectos computacionales**.

- Elementos a considerar en éste hito:
    - Los dos enunciados deben estar clarificados, considerando el tipo de problema a resolver (regresión o clasificación). Para cada uno de los enunciados y su problema identificado, se debe justificar el uso de métricas para medir el desempeño del problema. (**3 puntos**)
    - Se debe considerar el uso de las librerías asociadas para la ingesta, preprocesamiento, visualización y modelación, así como métricas de evaluación. (**1 punto**)
    - Se debe detallar y considerar el proceso de preprocesamiento y recodificación de datos. (**1 punto**)
- **Entregable:** Dos notebooks (uno por enunciado) con todos los puntos detallados.

Hito 2: Sesión Presencial 2, Unidad 7

Completar el punto de **Descripción**.

- Elementos a considerar en éste hito:
    - La inspección visual del vector objetivo. (**2 puntos**)
    - La inspección visual de las variables. (**2 puntos**)
    - La inspección de datos perdidos en las variables. (**1 punto**)
    - De ser necesario, se puede iterar en el proceso de preprocesamiento y recodificación de las variables
- **Entregable:** Dos notebooks (uno por enunciado) con todos los puntos detallados e interpretados. De ser necesario, un archivo con extensión `.py` con todas las funciones implementadas.

Hito 3: Sesión Presencial 1, Unidad 8

Completar el punto de **Modelación descriptiva**.

- Elementos a considerar en éste hito:
    - La modelación mediante regresión de ambos problemas. (**2 puntos**)
    - La interpretación de los principales regresores en cada problema. (**2 puntos**)
    - La definición de las estrategias de **Modelación predictiva**. (**1 punto**)
- **Entregable:** Dos notebooks (uno por enunciado) con todos los puntos detallados e interpretados. De ser necesario, un archivo con extensión `.py` con todas las funciones implementadas.

Hito 4: Sesión Presencial 2, Unidad 8

Completar el punto de **Modelación predictiva**.

- Elementos a considerar en éste hito:
    - La preparación del ambiente de trabajo (imports, separación de muestras) para implementar modelos de predicción. (**1 punto**)
    - La implementación de por lo menos tres modelos predictivos. (**2 puntos**)
    - El reporte del mejor modelo predictivo en base a los resultados. (**2 puntos**)
- **Entregable:** Dos notebooks (uno por enunciado) con todos los puntos detallados e interpretados. De ser necesario, un archivo con extensión `.py` con todas las funciones implementadas

Desafío 1: Determinantes del ingreso

Enunciado

Usted trabaja para un organismo no gubernamental que está interesado en las dinámicas socioeconómicas que determinan la desigualdad de ingreso y la erradicación de la pobreza extrema, enmarcado dentro de los objetivos del desarrollo del nuevo milenio del Programa de las Naciones Unidas para el Desarrollo. Le encomiendan el desarrollo de un modelo predictivo sobre la probabilidad que un individuo presente salarios por sobre o bajo los 50.000 dólares anuales, en base a una serie de atributos sociodemográficos.

Descripción de la base de datos

Para desarrollar este desafío se debe utilizar la base de datos `income-db.csv`.

Las variables que componen esta base se detallan a continuación:

- `age`: Edad del individuo.
- `workclass`: Naturaleza de la organización que emplea al individuo.
- `education`: Nivel educacional del individuo:Bachelors (Licenciado), Some-college (Superior incompleta), 11th (3ro medio), HS-grad (Secundaria completa), Prof-school (Escuela profesional), Assoc-acdm (Técnico superior administrativo) , Assoc-voc (Técnico superior vocacional), 9th (1ro medio), 7th-8th (7mo-8vo), 12th (4to medio), Masters (Maestría de postgrado), 1st-4th (1ro-4to básico), 10th(2do medio), Doctorate (Doctorado), 5th-6th (5to-6to), Preschool (Preescolar).
- `capital-gains`: Ingresos generados por inversiones fuera del trabajo asalariado = Ingresos generados por inversiones fuera del trabajo asalariado.
- `capital-losses`: Pérdidas generadas por inversiones fuera del trabajo asalariado.
- `fnlwgt`: Ponderador muestral.
- `marital-status`: Estado civil del individuo: Married-civ-spouse (Casado/a régimen civil), Divorced (Divorciado/a), Never-married (Soltero/a), Separated (Separado/a), Widowed (Viudo/a), Married-spouse-absent (Casado con esposo/a ausente), Married-AF-spouse (Casado/a régimen castrense).
- `occupation`: Ocupación del individuo: Tech-support (Soporte técnico), Craft-repair (Reparaciones), Other-service (Otros servicios), Sales (Ventas), Exec-managerial (Ejecutivo administrativos), Prof-specialty (Profesores), Handlers-cleaners (Aseo y ornato), Machine-op-inspct (Inspectores de maquinarias), Adm-clerical (Administrativos servicio al cliente), Farming-fishing (Pesca-ganadería), Transport-moving (Transporte), Priv-house-serv (Asesor del hogar), Protective-serv (servicios de seguridad), Armed-Forces (Fuerzas armadas).
- `relationship`: Relación respecto a su familia Wife(Esposa), Own-child (hijo único), Husband (Esposo), Not-in-family (No pertenece a la familia), Other-relative (Familiar de otro tipo), Unmarried (Soltero).
- `race`: Raza del encuestado White(Blanco caucásico), Asian-Pac-Islander (Isleño del Asia Pacífico), Amer-Indian-Eskimo (Pertenenciente a pueblos originarios), Other (Otro grupo), Black (Afroamericano).
- `sex`: Sexo del encuestado.

- `hours-per-week` : Cantidad de horas trabajadas por semana.
- `native-country` : País de origen. United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
- `income` : `<=50K` Si el individuo percibe ingresos inferiores a 50.000 dólares anuales, `>50K` si el individuo percibe ingresos superiores a 50.000 dólares anuales. **Este es su vector objetivo**.

Aspectos adicionales a considerar

- La base de datos contiene los valores perdidos como `?`. Deberá transformarlos para poder trabajar de forma adecuada.
- Desde la organización le sugieren que debe recodificar las siguientes variables acorde a las siguientes nomenclaturas:
  - `occupation` debe recodificarse como `collars` siguiendo una nomenclatura similar a:
    - `white-collar` $\leftarrow$ Prof-specialty, Exec-managerial, Adm-clerical, Sales, Tech-support.
    - `blue-collar` $\leftarrow$ Craft-repair, Machine-op-inspct, Transport-moving, Handlers-cleaners, Farming-fishing, Protective-serv, Priv-house-serv.
    - `others` $\leftarrow$ Other-service, Armed-Forces
  - `workclass` debe recodificarse como `workclass_recod` siguiendo una nomenclatura similar a :
    - `federal-gov` $\leftarrow$ Federal-gov.
    - `state-level-gov` $\leftarrow$ State-gov, Local-gov.
    - `self-employed` $\leftarrow$ Self-emp-inc, Self-emp-not-inc
    - `unemployed` $\leftarrow$ Never-worked, Without-pay.
  - `education` debe recodificarse como `educ_recod` siguiendo una nomenclatura similar a :
    - `preschool` $\leftarrow$ Preschool
    - `elementary-school` $\leftarrow$ 1st-4th, 5th-6th
    - `high-school` $\leftarrow$ 7th-8th, 9th, 10th,11th, 12th, HS-grad
    - `college` $\leftarrow$ Assoc-voc, Assoc-acdm, Some-college
    - `university` $\leftarrow$ Bachelors, Masters, Prof-school, Doctorate
  - `marital-status` debe recodificarse como `civstatus` siguiendo una nomenclatura similar a :
    - `married` $\leftarrow$ Married-civ-spouse, Married-spouse-absent, Married-AF-spouse
    - `divorced` $\leftarrow$ Divorced
    - `separated` $\leftarrow$ Separated
    - `widowed` $\leftarrow$ Widowed.
  - `native-country` debe recodificarse como `region` donde cada país debe asignarse a uno de los 5 continentes.
  - `income` debe recodificarse de forma binaria.

Desafío 2: Rendimiento escolar

Enunciado

Lo contactan de una escuela Portuguesa para generar un modelo que identifique aquellos alumnos que presentan un bajo desempeño académico, medido en el promedio final del año escolar. Para ello le envían un archivo con registros sociodemográficos y conductuales de los alumnos dos escuelas para perfilar a los estudiantes.

De manera adicional la psicopedagoga sugiere inspeccionar una batería de preguntas asociadas a aspectos ambientales del alumno (de `famrel` a `health`) y ver si éstas se pueden abstraer en categorías latentes.

Descripción de la base de datos

Para responder esta pregunta deben utilizar el archivo `students.csv`.

Las variables que componen la base se detallan a continuación.

- `school` : Escuela del estudiante. (binaria: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira)
- `sex` : Sexo del estudiante. (binaria: 'F' - Mujer o 'M' - Hombre)
- `age` : Edad del estudiante. (numérica: de 15 a 22)
- `address` : Ubicación de la casa del estudiante. (binaria: 'U' - urbana o 'R' - rural)
- `famsize` : Tamaño de la familia. (binaria: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- `Pstatus` : Estado cohabitacional de los padres. (binaria: 'T' - cohabitando juntos o 'A' - viviendo separados)
- `Medu` : Nivel educacional de la madre. (numérica: 0 - ninguno, 1 - educación básica (4to), 2 - de 5to a 9, 3 - educación media, o 4 - educación superior).
- `Fedu` : Nivel educacional del padre. (numérica: 0 - ninguno, 1 - educación básica (4to), 2 - de 5to a 9, 3 - educación media, o 4 - educación superior).
- `Mjob` : Ocupación de la madre. (nominal: 'teacher' profesora, 'health' relacionada a salud, 'services' (e.g. administración pública o policía), 'at_home' en casa u 'other' otra).
- `Fjob` : Ocupación del padre (nominal: 'teacher' profesor, 'health' relacionado a salud, 'services' (e.g. administración pública o policía), 'at_home' en casa u 'other' otra).
- `reason` : Razón para escoger la escuela (nominal: 'home' cercano a casa, 'reputation' reputación de la escuela, 'course' preferencia de cursos u 'other' otra)
- `guardian` : Apoderado del estudiante (nominal: 'mother' madre, 'father' padre u 'other' otro)
- `traveltime` : Tiempo de viaje entre hogar y colegio. (numeric: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, or 4 - >1 hora).
- `studytime` : Horas semanales dedicadas al estudio. (numérica: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, o 4 - >10 horas)
- `failures` : Número de clases reprobadas. (numérica: n si 1<=n<3, de lo contrario 4)
- `schoolsup` : Apoyo educacional del colegio. (binaria: si o no)
- `famsup` : Apoyo educacional familiar. (binaria: si o no)
- `paid` : Clases particulares pagadas (matemáticas o portugués) (binaria: si o no)
- `activities` : Actividades extracurriculares. (binaria: si o no)

- `nursery` : Asistió a guardería infantil. (binaria: si o no)
- `higher` : Desea proseguir estudios superiores (binaria: si o no)
- `internet` : Acceso a internet desde el hogar (binaria: si o no)
- `romantic` : Relación romántica (binaria: si o no)
- `famrel` : Calidad de las relaciones familiares. (numérica: de 1 - muy malas a 5 - excelentes)
- `freetime` : Tiempo libre fuera del colegio (numérica: de 1 - muy poco a 5 - mucho)
- `goout` : Salidas con amigos (numérica: de 1 - muy pocas a 5 - muchas)
- `Dalc` : Consumo de alcohol en día de semana (numérica: de 1 - muy bajo a 5 - muy alto)
- `Walc` : Consumo de alcohol en fines de semana (numérica: de 1 - muy bajo a 5 - muy alto)
- `health` : Estado de salud actual (numérica: from 1 - muy malo to 5 - muy bueno)
- `absences` : Cantidad de ausencias escolares (numérica: de 0 a 93)
- `G1` : Notas durante el primer semestre (numérica: de 0 a 20). **Este es uno de sus vectores objetivos para el modelo descriptivo.**
- `G2` : Notas durante el segundo semestre (numérica: de 0 a 20). **Este es uno de sus vectores objetivos para el modelo descriptivo.**
- `G3` : Promedio final (numérica: de 0 a 20). **Este es uno de sus vectores objetivos para el modelo descriptivo y el vector a predecir en el modelo predictivo**.

Aspectos adicionales a considerar

- La base de datos presenta una serie de anomalías. En la escuela no tienen buenas prácticas sobre cómo ingresar datos, por lo que existen datos perdidos que están registrados bajo tres categorías: *nulidade*, *sem validade*, *zero*. De manera adicional, hay 3 variables numéricas que se registraron como strings, cuya interpretación en `pandas` devuelve una estructura de datos genérica. Finalmente, la base está con un encoding distinto al normal y los delimitadores son distintos.
- Para simplficar el análisis y su posterior inclusión en un modelo predictivo, se sugiere recodificar las variables binarias como 0 y 1. Se recomienda seguir en criterio de asignarle 1 a aquellas categorías minoritarias.
- El procedimiento también debe aplicarse para aquelas variables nominales con más de 2 categorías siguiendo la misma lógica.
- En la parte de modelación descriptiva, se deben generar modelos saturados por cada una de las notas registradas en `G1` , `G2` y `G3` .
- Para la parte de modelación predictiva, se debe generar un modelo para predecir las notas en `G3` .