CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Curso
Minería de Datos con Python

Unidades 1-8

Trabajo Práctico Final

Objetivos

Aplicar los conocimientos adquiridos en el curso

Consignas

Ejercicio 1 (40 puntos)

En este ejercicio, el objetivo es obtener un modelo de regresión lineal múltiple que describa el comportamiento de los datos en el archivo student_performance.csv (data tomada de kaggle.com).

En este archivo de datos se registra la calificación obtenida por estudiantes (Performance Index) en una escala 0-100, en relación con las siguientes variables

- Horas de Estudio (Hours Studied)
- Calificación Previa (Previous Scores)
- Actividades Extracurriculares (Extracurricular Activities). Esta variable puede tomar valores Yes/No, si el estudiante realizó o no, respectivamente, dichas actividades. Para la determinación del modelo,

esta variable categórica debe transformarse a un valor numérico de 0 (No) o 1 (Yes)
- Horas de sueño (Sleep Hours)
- Número de exámenes practicados (Sample Question Papers Practiced)

Desarrolle un Jupyter Notebook que

1) Cree un dataframe de Pandas que contenga los datos del archivo student_performance.csv suministrado
2) Normalice los datos correspondientes a cada variable de entrada, restando la media y dividiendo entre la desviación estándar
3) Utilice el comando train_test_split para separar los datos en conjuntos de entrenamiento y prueba. La fracción de datos de prueba debe estar entre 20% y 30%.
4) Determine un modelo de regresión lineal múltiple utilizando los datos de entrenamiento. Este modelo debe estimar Performance Index a partir de las variables de entrada (valores normalizados)
5) Evalúe el desempeño del modelo de regresión lineal obtenido en el apartado anterior. Para esto
    - Calcule las métricas Error Absoluto Medio, Error Cuadrático Medio, Puntuación R2, Puntuación de Varianza Explicada
    - Realice un gráfico de dispersión de datos pronosticados (eje vertical) vs. datos de prueba (eje horizontal)
    - Realice un gráfico de línea de los errores del modelo sobre los datos de prueba
    - Comente brevemente los resultados

Ejercicio 2 (60 puntos)

En este ejercicio, el objetivo es obtener un modelo para clasificar los datos en el archivo titanic_processed.csv. Este archivo contiene los datos ya procesados del conjunto de datos titanic.

Centro de e-Learning SCEU UTN - BA. Medrano 951 2do piso
(1179) // Tel. +54 11 7078- 8073 / Fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

El registro correspondiente a cada pasajero(a) especifica si sobrevivió (valor de Survived igual a 1) o no (valor de Survived igual a 0), en relación con las siguientes variables

- Clase en la que viajaba (PClass), que puede ser 1, 2 o 3
- Sexo (Sex), Femenino (0) o Masculino (1)
- Edad (Age)
- SibSP, que agrupa el número de familiares como cónyuge, hermanas, hermanos, hermanastras y hermanastros
- Parch, que agrupa el número de familiares como madre, padre, madrastra, padrastro, hijas, hijos, hijastras e hijastros
- Precio pagado por el boleto (Fare)
- La ciudad de embarque (Cherbourg (C), Queenstown (Q) o Southampton (S). Esta información está codificada en la forma one-hot

Desarrolle un Jupyter Notebook que

1) Cree un dataframe de Pandas que contenga los datos del archivo titanic_processed.csv suministrado
2) Utilice el comando train_test_split para separar los datos en conjuntos de entrenamiento y prueba. La fracción de datos de prueba debe estar entre 20% y 30%.
3) Determine un modelo de clasificación de regresión logística utilizando los datos de entrenamiento. Este modelo debe estimar la Supervivencia del pasajero a partir de las variables de entrada
4) Evalúe el desempeño del modelo de regresión logística obtenido en el apartado anterior, a través de la precisión (accuracy) sobre el conjunto de prueba. Trate de obtener un modelo de regresión logística que tenga una precisión (accuracy) de por lo menos 80%
5) Realice un gráfico de dispersión que muestre las clases de los datos de prueba y las clases pronosticadas por el modelo para los mismos datos. El eje y debe mostrar las clases (0 o 1) y el eje x el índice de los datos.

Utilice diferentes marcadores y colores para cada caso. Seleccione un tamaño de figura suficientemente grande para distinguir los marcadores. En base a este gráfico, comente sobre los resultados

6) Determine un modelo de clasificación del tipo árbol de decisión utilizando los datos de entrenamiento. Este modelo debe estimar la Supervivencia del pasajero a partir de las variables de entrada
7) Evalúe el desempeño del modelo de árbol de decisión obtenido en el apartado anterior, a través de la precisión (accuracy) sobre el conjunto de prueba. Trate de obtener un árbol de decisión que tenga una precisión (accuracy) de por lo menos 80%
8) Grafique el árbol de decisión obtenido. En base a este gráfico, comente sobre los resultados

Formato de presentación:

- Los participantes deben entregar un archivo comprimido que contenga los jupyter notebooks desarrollados. El nombre del archivo comprimido debe tener el formato TP_Apellido. Por ejemplo, TP_Canelon. El nombre de cada Jupyter notebook debe tener el formato Apellido_TPF_Ejercicio#.ipynb. Por ejemplo, Canelon_TPF_Ejercicio1.ipynb
- El archivo comprimido debe generarse con la aplicación 7-zip, que puede descargarse gratuitamente desde https://www.7-zip.org/.

Fecha límite de entrega:

Nominal: 04/07/2024 - 23:59 hrs

Recuperatorio: 11/07/2024 - 23:59 hrs

Las fechas de entrega son inapelables, ya que están configuradas automáticamente en el Campus. La plataforma no permitirá que los

Centro de e-Learning SCEU UTN - BA. Medrano 951 2do piso
(1179) // Tel. +54 11 7078- 8073 / Fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

participantes entreguen fuera de la fecha/hora indicada. Si no pueden hacerlo en la primera, podrán hacerlo en el recuperatorio.

Criterios de evaluación

La calificación total del trabajo estará en función del número de consignas realizadas correctamente. Si alguna consigna no funciona de manera correcta o genera un error en el Jupyter notebook, se restarán puntos del total correspondiente a esa consigna