UdeCataluña

DIPLOMADO EN BIG DATA – DATA SCIENCE

MÓDULO 05

Práctica individual Módulo 5

Tema:

Modelamiento

Dinámica de la actividad:

Estimado estudiante en esta oportunidad vamos a desarrollar un caso práctico de modelamiento. Esta fase es el corazón del proceso de Minería de Datos. Tenga en cuenta que el éxito del modelamiento depende del trabajo en fases previas: selección de características, limpieza y transformación de datos.

Objetivo:

Usar las técnicas de Machine Learning vistas en el módulo para entrenar y optimizar al menos 2 modelos usando 2 algoritmos de Machine Learning.

Entregables:

- Jupyter Notebook en formato PDF. Escriba su informe haciendo énfasis en los resultados y su análisis. Organice su informe en secciones empleando las celdas de texto de los notebooks de Jupyter.

Adjunte el link del notebook de Google Colab con permiso de comentarios, la retroalimentación se hará sobre el notebook

Visualizar el Tutorial exportar HTML/PDF que se encuentra en las instrucciones de la actividad

- Si presenta problemas con la conversión a PDF puede enviar el link de su notebook con permisos para hacer comentarios.
- Se recomienda que añada comentarios a su código para documentar el proceso y que sea sencillo en el futuro para el personal técnico entender su código.

Detalles de la entrega:

Se espera que continúe trabajando con el conjunto de datos del PRAI 3 y 4.

Para desarrollar su Proyecto Aplicada tiene 2 opciones:

1. **Proyecto propio:** En esta opción usted escogerá un conjunto de datos y usted deberá hacer el planteamiento de hipótesis, objetivos y preguntas de negocio que permitan obtener valor de sus datos. Puede escoger uno de los 3 conjuntos de datos de su trabajo investigativo del PRAI 1. Recuerde que su conjunto de datos debe tener un problema computacional asociado (regresión, clasificación, agrupamiento). Cualquier duda por favor comuníquese con el experto temático.
2. **Proyecto guiado:** En esta opción el docente le proporcionará un conjunto de datos y un notebook con instrucciones y ejercicios a desarrollar en materia de programación y análisis de datos para cada fase del proyecto. A continuación, se hace el planteamiento del proyecto.

Contexto: Airbnb es una empresa que ofrece una plataforma de software dedicada a la oferta de alojamientos particulares y turísticos mediante la cual los anfitriones pueden publicitar y contratar el arriendo de sus propiedades con sus huéspedes; anfitriones y huéspedes pueden valorarse mutuamente, como referencia para futuros usuarios. Muchos nuevos anfitriones no cuentan con información global de tendencias del mercado por lo que sus precios no son óptimos. Airbnb gana una comisión por cada arrendamiento, por lo tanto, está interesado en que sus anfitriones cobren una

tarifa óptima de acuerdo a las características del hospedaje. Si los anfitriones ganan más… Airbnb también.

Problema de Negocio: La empresa Airbnb lo ha contratado para desarrollar un modelo que permita responder la siguiente pregunta: ¿Cuál es la variable o característica más relevante para determinar el precio de un hospedaje en Airbnb? Y además que permita predecir el precio de un hospedaje dadas ciertas características del mismo.

Sistema de información: El conjunto de datos objetivo posee información acerca de 38.000 hospedajes de la plataforma Airbnb en la ciudad de Nueva York. Los datos a usar son datos públicos creados por Inside Airbnb, para más información puede consultar: http://insideairbnb.com/get-the-data/

Visualizar el Notebook Fase 3: Modelamiento con técnicas de Machine Learning

Contenido:

1. **Descripción del sistema de información real:**

Referenciar a nivel general el sistema de información objeto de estudio, explicar el origen de los datos a analizar y definir claramente cuál es el objetivo de modelamiento. Por ejemplo, a partir de las mediciones de la flor: largo y ancho del pétalo y sépalo predecir la especie de la flor. Es una tarea de clasificación.

2. **Partición de entrenamiento y prueba**

Usar las técnicas vistas para crear un subconjunto de prueba y uno de entrenamiento. Recuerde definir una semilla para que obtenga reproducibilidad durante la experimentación.

3. **Modelamiento**

Escoja un algoritmo de Machine Learning para hacer el entrenamiento de su modelo, puede escoger algoritmos no vistos en clase. Haga un breve resumen de las ventajas y desventajas de los 2 algoritmos escogidos.

4. **Búsqueda de hiperparámetros óptimos**

Escoja los hiperparámetros de experimentación, haga experimentos sistemáticos para encontrar la combinación de hiperparámetros que mejor desempeño obtenga sin excederse en carga computacional. Si su problema es clasificación no olvide mostrar y analizar la matriz de confusión, así mismo discutir cuál de las métricas vistas en clase es la más útil de acuerdo a su aplicación.

5. **Información adicional**

La mayoría de modelos contienen información en forma de atributos que puede ser usada para analizar la toma de decisiones dentro del modelo.

6. **Conclusiones**

- ¿Cuál de los algoritmos le dio el modelo con mejor desempeño? ¿Cuál escogería para usar en producción? Recuerde que las métricas no lo son todo, considere también la interpretabilidad del modelo y el tiempo de entrenamiento. Justifique su respuesta.
- ¿Se cumplió con el objetivo de análisis predictivo? ¿los resultados fueron los esperados? ¿qué falta para que su modelo tenga el desempeño requerido? ¿qué otro algoritmo podría emplear para lograr mejores resultados?

Tenga en cuenta las recomendaciones hechas en la retroalimentación previa a este trabajo.

Enfoque su esfuerzo en analizar, interpretar y discutir los resultados obtenidos.

Recuerde que usar una tabla para resumir resultados es mejor que usar texto. Y a la vez usar una gráfica es mejor que usar una tabla.

UdeCataluña

Bogotá - Colombia.