Programa de formación Machine Learning and Data Science MLDS

Facultad de
INGENIERÍA
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Módulo 1
Análisis y visualización de datos con *Python*

Proyecto aplicado y Rúbrica de evaluación

Facultad de INGENIERÍA
Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA

Proyecto aplicado

Objetivo:

Ejecutar un proyecto de análisis de datos de forma efectiva usando la metodología y las herramientas presentadas en el curso con el fin de hallar características relevantes y relaciones entre los datos.

Descripción

Se espera que utilice la metodología de trabajo propuesta en el curso y las herramientas de análisis de datos y visualización de información para llevar a cabo la planeación y ejecución de un proyecto aplicado. El conjunto de datos sobre el que trabajará puede ser seleccionado por usted de acuerdo con sus intereses en el ámbito personal o profesional. También, podrá elegir alguno de los *dataset* propuestos al final de este documento para realizar el trabajo. En cualquier caso, se busca que poder llegar a obtener conclusiones con información valiosa que aporte en procesos de toma de decisiones en un dominio de aplicación particular.

El proyecto se desarrollará utilizando el lenguaje de programación *Python* y su entorno de herramientas para la computación científica, en forma de *Notebook* en el formato *.ipynb*. Se debe presentar el proyecto tomando como referencia las etapas previas al modelado de la metodología CRISP-DM para análisis de datos. Para las visualizaciones de información se espera que describa y sustente el proceso de abstracción de datos, tarea y definición de codificación visual e interacción y presente el resultado final obtenido con alguna de las librerías presentadas en el curso.

Para la entrega debe preparar un **video** en formato *.mp4* de máximo 5 minutos donde deberá describir y sustentar brevemente la aplicación de las primeras fases de la metodología CRISP-DM, describir el proceso realizado y presentar los hallazgos del análisis apoyándose de las visualizaciones de información y el análisis de los resultados realizado en el *notebook* del proyecto, que también deberá ser entregado en formato *.ipynb*.

Contenido

1. Entendimiento del negocio.
    a. Objetivos de negocio y situación actual.
    b. Objetivo del proyecto.
    c. Planeación del proyecto.

2. Entendimiento de los datos.
    a. Adquisición e integración de datos.
    b. Análisis exploratorio de datos.
    c. Descripciones generales.
    d. Análisis con estadística descriptiva.
    e. Visualización de datos estadísticos.

Proyecto aplicado

3. Preparación de los datos.
    a. Limpieza de datos.
    b. Selección de características.
    c. Preprocesamiento y transformación.

4. Análisis de datos.
    a. Técnicas estadísticas tradicionales.
    b. Análisis de correlación y/o análisis de regresión.
    c. Análisis de resultados y cumplimiento de objetivos.

5. Visualización de datos.
    a. Abstracción y definición de las visualizaciones.
    b. Visualización de información estática e interactiva.
    c. (Opcional) Visualización de mapas coropléticos.

Conjuntos de datos

Como se mencionó anteriormente, el planteamiento y desarrollo del proyecto no es restrictivo en la elección del conjunto de datos a analizar. Si lo desea puede realizar un proyecto personal/laboral que aplique los conceptos y herramientas discutidas en el transcurso del curso. Si tiene dificultad en encontrar un tema para su proyecto, dejamos a su disposición algunos temas propuestos de los que se puede guiar para llevar a cabo la actividad.

A continuación, se presentan algunas alternativas con distintos *dataset* disponibles con datos a nivel local o regional:

COVID-19 - Reporte diario

La coyuntura provocada por la enfermedad por coronavirus COVID-19 es un evento global que es seguido de cerca por múltiples instituciones y entidades, que disponen al público la información recolectada del reporte diario a nivel mundial, regional y local. Algunas de estas instituciones son:

- World Health Organization - Coronavirus Official Data
- John Hopkins University - CSSE COVID-19 Data Repository
- Datos abiertos MinTICs - Casos positivos de COVID-19 en Colombia
- Datos abiertos Alcaldía Mayor de Bogotá - Número de casos confirmados por el laboratorio de COVID- 19 - Bogotá D.C.

Proyecto aplicado

Educación básica en Colombia

El Ministerio de Educación Nacional dispone de un conjunto de datos con estadísticas e indicadores por municipio o departamento de los niveles educativos de Colombia, con cifras de aprobación, repitencia, deserción y cobertura de las instituciones educativas del país.

- Datos Abiertos MinTICs - Estadísticas en educación básica por municipio
- Datos Abiertos MinTICs - Estadísticas en educación básica por departamento

Accidentalidad y seguridad vial

La Secretaría Distrital de Movilidad y el Ministerio de Transporte disponen de información de seguridad vial, con un registro de siniestros viales registrados en los distintos municipios y ciudades de Colombia, categorizando el tipo de siniestro y demás información del suceso.

- Datos abiertos MinTICs - Registro nacional de accidentes de tránsito
- Datos abiertos Bogotá - Siniestros Viales Consolidados Bogotá D.C.

Ubicación geográfica

Con la ayuda de librerías dedicadas a conjuntos de datos geográficos, es posible construir visualizaciones de información representada en mapas coropléticos, u otro tipo de visualizaciones compuestas. Si se quiere trabajar con datos regionales, como municipios, barrios o localidades, puede utilizar alguno de los siguientes conjuntos de datos:

- John Guerra - Mapa de Colombia con ciudades y departamentos (GEO Json)
- Datos abiertos Bogotá - Mapa de referencia para Bogotá D.C.

Rúbrica de evaluación

Criterio	[0.0 - 1.0)	[1.0 - 3.0)	[3.0 - 4.0)	[4.0 - 5.0]	%
Entendimiento del negocio - Objetivos de negocio. - Valoración de la situación actual. - Metas del proyecto de análisis de datos. - Planeación del proyecto de análisis de datos.	- No se describen los objetivos del negocio, ni la situación actual. - No se describen las metas del proyecto de análisis de datos. - No se realiza ni describe la planeación del proyecto.	- Se describen los objetivos del negocio y la situación actual de manera superficial y/o con poca claridad. - Se describen las metas del proyecto sin una justificación adecuada y con poca claridad. - Se realiza una planeación aceptable del proyecto, identificando las fases del proyecto, pero no se describen con claridad.	- Se describen los objetivos del negocio y la situación actual de manera clara. - Se describen las metas del proyecto de manera clara y con una justificación aceptable. - Se realiza una planeación aceptable del proyecto, identificando las fases del proyecto, las cuales se describen de forma preliminar.	- Se describen los objetivos del negocio y la situación actual de manera precisa. - Se describen, conceptualizan y justifican las metas del proyecto de una manera clara y ordenada y coherente con el negocio. - Se realiza una planeación organizada y detallada del proyecto y se describe de manera clara y ordenada.	**10%**
Entendimiento de los datos - Adquisición e integración de datos. - Análisis exploratorio de datos: caracterización de los datos. - Análisis exploratorio de datos: descripción de los datos por medio de estadística descriptiva. - Análisis exploratorio de datos: visualizaciones de datos estadísticos.	- Se carga un único archivo para realizar el análisis, y no queda claro su fuente ni su pertinencia con relación a los objetivos del proyecto. - No se caracteriza el conjunto de datos. - No se describe el conjunto de datos por medio de estadística descriptiva. - No se exploran los datos por medio de visualizaciones de medidas estadísticas.	-Se carga un único archivo para realizar el análisis, sin integrar ni consolidar distintas fuentes de datos previamente. - Se caracteriza el conjunto de datos de forma insuficiente, ignorando detalles que pudieran ser importantes. - Se describe el conjunto de datos por medio de estadística descriptiva, sin intentar interpretar los resultados o sacar conclusiones a partir de ellos. - Se exploran los datos por medio de visualizaciones de medidas estadísticas, sin intentar interpretar los resultados o sacar conclusiones a partir de ellas.	- Se llevan a cabo tareas de adquisición e integración del conjunto de datos, respetando su integridad mediante herramientas distintas al lenguaje de programación Python. - Se caracteriza el conjunto de datos de forma adecuada. - Se describe el conjunto de datos por medio de estadística descriptiva, aportando alguna posible interpretación que ayude a entender los datos. - Se exploran los datos por medio de visualizaciones de medidas estadísticas, aportando alguna interpretación que ayude a entender los datos.	- Se llevan a cabo tareas de adquisición e integración del conjunto de datos, respetando su integridad mediante Python y sus librerías especializadas. - Se caracteriza el conjunto de datos de forma adecuada, aportando información que es relevante para las etapas posteriores del proyecto. - Se describe el conjunto de datos por medio de estadística descriptiva, aportando interpretaciones precisas que ayuden a entender los datos. - Se exploran los datos por medio de visualizaciones de medidas estadísticas, aportando interpretaciones que ayuden a entender la naturaleza de los datos.	**15%**

Rúbrica de evaluación

Preparación de los datos - Limpieza del conjunto de datos. - Preprocesamiento y transformación de los datos. - Selección de datos finales para el análisis posterior.	- No se realiza el proceso de limpieza del conjunto de datos. - No se realiza el proceso preprocesamiento y transformación de los datos. - No se lleva a cabo una selección final del conjunto de datos para el análisis posterior.	- Se realiza el proceso de limpieza del conjunto de datos con algunos errores que podrían afectar los resultados de los análisis posteriores. - Se realiza el proceso preprocesamiento y transformación de los datos con algunos errores que podrían afectar los resultados de los análisis posteriores. - Se describe la selección final del conjunto de datos para el análisis posterior con algunos errores que podrían afectar los resultados de los análisis posteriores.	- Se realiza el proceso de limpieza del conjunto de datos, pero no se justifica adecuadamente. - Se realiza el proceso preprocesamiento y transformación de los datos, pero no se justifica adecuadamente. - Se describe la selección final del conjunto de datos para el análisis posterior, pero no se justifica adecuadamente.	- Se realiza y justifica adecuadamente el proceso de limpieza del conjunto de datos. - Se realiza y justifica adecuadamente el proceso preprocesamiento y transformación de los datos. - Se describe y justifica de manera clara la selección final del conjunto de datos para el análisis posterior.	**15%**
Análisis de datos - Aplicación de estadística descriptiva y/o inferencial para sacar conclusiones a partir de los datos. - Aplicación de análisis de correlaciones entre los datos y/o análisis de regresiones para modelar relaciones entre los datos. - Cumplimiento de los objetivos del proyecto de análisis de datos.	- No se aplican conceptos de estadística descriptiva y/o inferencial en el análisis de datos. - No se realiza un análisis de correlaciones y/o regresiones entre los datos. - No se cumplen los objetivos del proyecto de análisis de datos.	- Se aplican conceptos de estadística descriptiva y/o inferencial en el análisis de datos, pero se cometen algunos errores que pueden afectar los resultados. - Se realiza un análisis de correlaciones y/o regresiones, pero se cometen errores que pueden afectar los resultados. - Se cumplen los objetivos del proyecto de forma preliminar sin añadir información ni conclusiones interesantes para el negocio.	- Se aplican conceptos de estadística descriptiva y/o inferencial en el análisis de datos de manera correcta, pero no se interpretan los resultados. - Se realiza un análisis de correlaciones y/o regresiones de manera correcta, pero no se interpretan los resultados. - Se cumplen los objetivos del proyecto, pero no se evidencian conclusiones que puedan ser interesantes para el negocio.	- Se aplican conceptos de estadística descriptiva y/o inferencial en el análisis de datos de manera correcta, y se interpretan los resultados de forma precisa. - Se realiza un análisis de correlaciones y/o regresiones de manera correcta, y se interpretan los resultados de forma precisa. - Se cumplen los objetivos del proyecto, y se llega a conclusiones y recomendaciones que puedan ser interesantes para el negocio.	**25%**

Rúbrica de evaluación

Visualización de información - Visualizaciones estáticas - Visualizaciones interactivas - Visualización de datos estadísticos - (Opcional) Visualización de mapas coropléticos	- No se realiza visualización de información en ninguna etapa del proyecto.	- Se construyen algunas visualizaciones estáticas de información y su contribución es limitada. - Las visualizaciones utilizadas no tienen un objetivo claro.	- Se realizan algunas visualizaciones de información estáticas para conocer detalles de los datos. - Opcionalmente, se construyen algunas visualizaciones de información interactivas para apoyar las actividades del proyecto. - Se construyen algunas visualizaciones para representar datos estadísticos. - Las visualizaciones utilizadas sirven para contribuir en el análisis de los datos.	- Se construyen algunas visualizaciones de información estáticas e interactivas para apoyar las actividades del proyecto. - Se construyen visualizaciones para representar datos estadísticos y detalles de los datos que evidencian los resultados de un análisis basado en los principios de visualización de información. - Las visualizaciones utilizadas sirven como apoyo importante para el análisis de los datos y las conclusiones del proyecto. - (Opcional) Se realizan visualizaciones estáticas y/o interactivas de mapas coropléticos.	**25%**
Forma - Calidad de la presentación y recursos visuales de apoyo del *notebook*. - Calidad del video de presentación. - Posicionamiento de los elementos gráficos y textuales. - Ortografía y gramática.	- Mala calidad visual en general. - No se realiza la entrega del video de presentación. - Posicionamiento y proporciones nada claras. - Numerosos errores ortográficos y gramaticales.	- Calidad visual mediocre en general. -El video de presentación tiene errores de edición y una calidad visual y de contenido mediocre. - Posicionamiento y proporciones poco claras en los elementos gráficos y textuales. - Algunos errores ortográficos y gramaticales.	- Buena calidad visual en general. - El video de presentación presenta los resultados adecuadamente. - Buen posicionamiento y proporciones de los elementos gráficos y textuales. - Pocos errores ortográficos y gramaticales.	- Excelente calidad visual. - El video de presentación presenta los resultados adecuadamente tiene una gran calidad visual y de presentación. - Buen posicionamiento y proporciones de los elementos gráficos y textuales. - Muy pocos o ningún error ortográfico y/o gramatical.	**10%**
Total					**100%**

Programa de formación
Machine Learning and Data Science MLDS

Facultad de
INGENIERÍA
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA