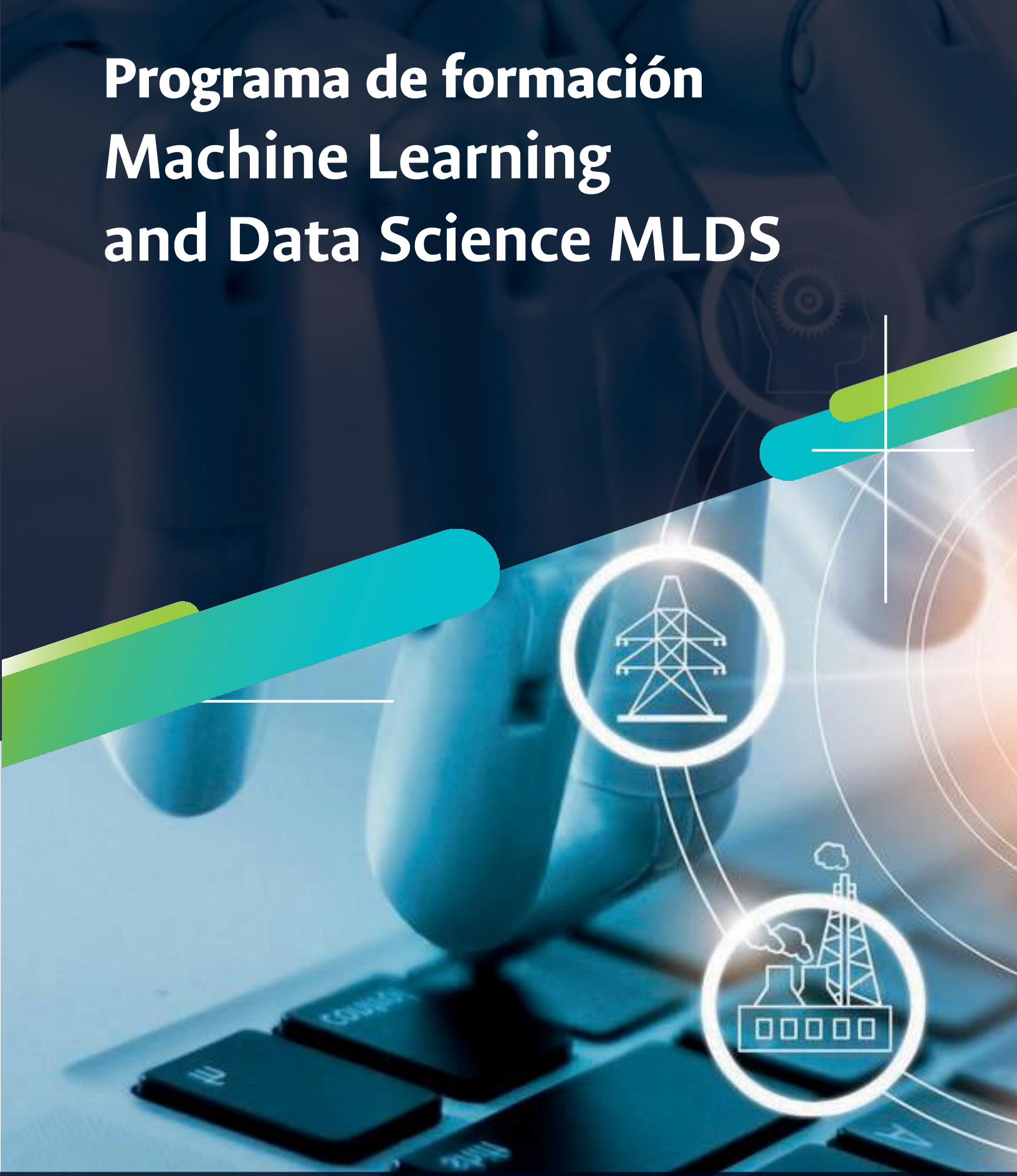
# Programa de formación Machine Learning and Data Science MLDS

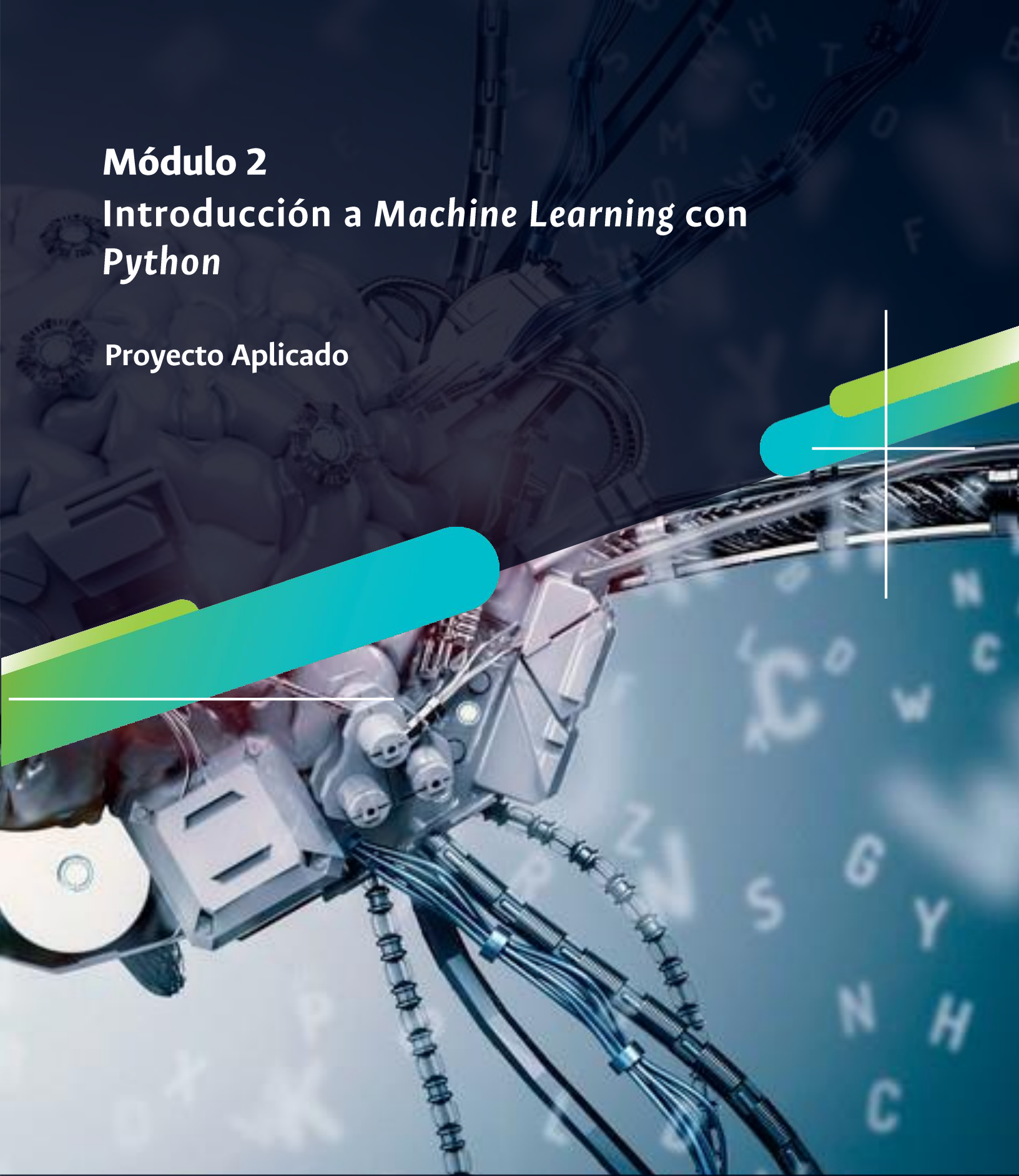


Facultad de
INGENIERÍA
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA

# Módulo 2
# Introducción a *Machine Learning* con *Python*

## Proyecto Aplicado



Facultad de
INGENIERÍA
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA

# Proyecto Aplicado

## Objetivo:

Ejecutar un proyecto de machine learning de forma efectiva usando la metodología y las herramientas presentadas en el curso con el fin de:

1. Hallar características relevantes y relaciones entre los datos.
2. Elaborar y evaluar modelos de machine learning.

## Descripción

Se espera que utilice la metodología de trabajo propuesta en el curso y las herramientas de modelamiento 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 a sus intereses en el ámbito personal o profesional. También, podrá elegir alguno de los *datasets* propuestos al final de este documento para realizar el trabajo. En cualquier caso, se busca, a través de un proceso de extensiva experimentación con modelos de machine learning 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 despliegue de la metodología CRISP-DM para análisis de datos (abordada en el módulo 1).

Para la entrega debe preparar un **video** en formato .mp4 de máximo 5 minutos donde deberá describir y sustentar brevemente el preprocesamiento de los datos, selección de modelos y evaluación de modelos usando *Scikit-learn*; tomando como apoyo el Notebook realizado.

## Contenido

1. Entendimiento del negocio.
    a. Objetivos de negocio
    b. Metas del Proyecto de machine learning

2. Entendimiento de los datos.
    a. Recolección inicial de datos.
    b. Descripción de los datos.
    c. Exploración de los datos.
    d. Verificación de la calidad de los datos.

3. Preparación de los datos.
    a. Selección de Datos finales.
    b. Limpieza de los datos

4. Modelamiento
    a. Selección de un modelo apropiado para el problema (supervisado o no supervisado).
    b. Diseño de la Experimentación (validación cruzada).
    c. Construcción y Entrenamiento de los Modelos.

5. Evaluación
    a. Definición de los criterios o métricas de evaluación.
    b. Análisis de los resultados.

## 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:

## Conjuntos de datos de machine learning

Los siguientes son conjuntos de datos generales que han sido diseñados específicamente para realizar varias tareas de machine learning, podrá encontrar muchos recursos disponibles sobre aplicaciones en estos conjuntos de datos:

### Google Play Store Apps:

Datos de 10 mil aplicaciones de la App Store obtenidas a través de *web scraping* con el objetivo de analizar el mercado de Android.

### Trip Advisor Hotel Reviews:

20 mil reseñas de hoteles extraídas de Tripadvisor. Se puede usar este conjunto de datos para descubrir cómo son los mejores hoteles o usarla en sus propios viajes.

### Avocado Prices:

Datos históricos de los precios del aguacate y volumen de ventas en múltiples mercados de estados unidos. Se puede modelar como una serie de tiempo.

# Proyecto Aplicado

### Fashion MNIST:

Un conjunto de datos similar a MNIST con 70 mil imágenes con tamaño 28x28 de prendas de ropa. Presenta una tarea de clasificación.

### Students Performance in Exams:

Notas obtenidas por estudiantes en varias asignaturas.

### IBM HR Analytics Employee Attrition & Performance:

Prediga el desgaste de sus empleados más valiosos. Descubra los factores que conducen al desgaste de los empleados y explora cuestiones importantes como "La relación entre la distancia de la casa al trabajo por puesto de trabajo y el desgaste" o "La relación entre el ingreso mensual promedio por educación y desgaste". Este es un conjunto de datos ficticio creado por científicos de datos de IBM.

## Otros conjuntos de datos:

- Mall Customers - Agrupamiento
- Fake News Detection - Clasificación
- SOCR data — Heights and Weights - Regresión
- Credit Card Fraud Detection - Clasificación.

## Buscadores de conjuntos de datos

Los siguientes son plataformas y repositorios que facilitan la búsqueda de conjuntos para proyectos de machine learning. Puede utilizar las siguientes plataformas para buscar y seleccionar el conjunto de datos para su proyecto:

- Google Dataset Search
- Kaggle
- UCI Machine Learning Repository
- VisualData
- CMU Libraries

# Rúbrica de evaluación

| Criterio | [0.0 - 1.0) | [1.0 - 3.0) | [3.0 - 4.0) | [4.0 - 5.0] | % |
|---|---|---|---|---|---|
| **Forma**<br><br>- Calidad visual de la presentación.<br>- Posicionamiento y proporcionamiento de los elementos gráficos y textuales.<br>- Ortografía y gramática.<br>- Calidad del video. | - Mala calidad visual en general.<br><br>- Posicionamiento y proporciones nada claras.<br><br>- Numerosos errores ortográficos y gramaticales.<br><br>- El video dura menos de 1 minuto o dura más de 6 minutos.<br><br>- No se puede entender el audio del video.<br><br>- No se alcanza a distinguir nada en el video por mala calidad de la imagen. | - Calidad visual mediocre en general.<br><br>- Posicionamiento y proporciones poco claras en los elementos gráficos y textuales.<br><br>- Algunos errores ortográficos y gramaticales.<br><br>- El video dura menos de 2 minutos.<br><br>- Una gran parte del video tiene audio difícil de entender y con mala calidad de imagen. | - Buena calidad visual en general.<br><br>- Buen posicionamiento y proporciones de los elementos gráficos y textuales.<br><br>- Pocos errores ortográficos y gramaticales.<br><br>- El video dura menos de 3 minutos.<br><br>- Pocos errores en el video, tanto en el audio como la imagen. | - Excelente calidad visual.<br><br>- Buen posicionamiento y proporciones de los elementos gráficos y textuales.<br><br>- Muy pocos o ningún error ortográfico y/o gramatical.<br><br>- El video tiene una duración entre 4 y 5 minutos.<br><br>- El video es totalmente entendible y tiene buena calidad. | **10%** |
| **Entendimiento del Negocio**<br><br>- Objetivos de Negocio. (e.g Aumentar ganancias)<br>- Metas del proyecto de *Machine Learning.* (e.g Identificar, modelar) | No se describen los objetivos de negocio ni metas del proyecto de *machine learning.* | - Se menciona de manera muy general pero no se elabora a nivel detallado.<br><br>- No es clara la relación entre los objetivos de negocios y las metas del proyecto. | - Se menciona y se elabora a nivel detallado, pero no es claro cómo se integra con el resto del proyecto. | - Se explica con detalles claros y se entiende cómo se integra con el resto del proyecto. | **5%** |

# Rúbrica de evaluación

| | | | | | |
|---|---|---|---|---|---|
| **Entendimiento de los Datos**<br><br>- Recolección inicial de datos.<br>- Descripción de los datos.<br>- Exploración de los datos.<br>- Verificación de la calidad de los datos. | - Los datos utilizados no son de una fuente fiable y justificable.<br><br>-No se describe la naturaleza de los datos ni se hace un análisis exploratorio. | - Algunos de los datos utilizados no son de una fuente fiable y justificable.<br><br>-Se describen las variables utilizadas de manera superficial.<br><br>- Exploración de datos muy superficial. | -Las fuentes de los datos utilizados son debidamente justificadas y autenticadas.<br><br>- Se describen las variables utilizadas de manera clara.<br><br>- Exploración de datos sin conclusiones válidas o significativas. | -Las fuentes de los datos utilizados son debidamente justificadas y autenticadas.<br><br>- Se describen las variables utilizadas de manera clara y se justifica su elección.<br><br>- Exploración de datos con conclusiones válidas. | **5%** |
| **Modelamiento**<br><br>- Selección del modelo (supervisado o no supervisado, clasificación o regresión).<br>- Diseño de los experimentos.<br>- Construcción y entrenamiento de los Modelos. | - Se aplica un modelo de forma errónea.<br>- No hay una metodología de experimentación apropiada para el problema.<br>- No hay una descripción del experimento realizado.<br>- Mal uso de las métricas de evaluación. | - Se utiliza un modelo apropiado para el problema, pero no se seleccionan las métricas apropiadas.<br><br>- No es clara la selección de métricas de desempeño de los modelos.<br><br>- Faltan elementos de la metodología de experimentación. | - Se selecciona un modelo de forma apropiada pero no queda claro si las métricas son apropiadas.<br><br>- Se realiza una experimentación con pocos elementos faltantes. | - Se explica la estrategia a nivel detallado y es claro cómo se implementó.<br><br>- Se realiza una cantidad razonable de experimentos.<br><br>- Se identifican claramente las variables de entrada y salida.<br><br>- Se justifica la selección del modelo de aprendizaje. | **50%** |
| **Evaluación**<br><br>-Definición de criterios de evaluación.<br>- Análisis de los resultados. | No se discuten los resultados obtenidos. | - Se menciona de manera muy general pero no se elabora a nivel detallado.<br><br>- No es clara la relación entre la evaluación de los modelos y las metas del proyecto.<br><br>- No se compara con resultados del estado del arte. | - Se menciona la estrategia y se elabora a nivel detallado, pero no es claro cómo se implementó dentro del proyecto.<br><br>- Se evalúa el impacto de los resultados sobre los objetivos de negocio. | - Se explica la estrategia a nivel detallado y es claro cómo se integra con el resto del proyecto.<br><br>- Buenas prácticas en la presentación de resultados (e.g buena cantidad de posiciones decimales).<br><br>- Se evalúa el impacto de los resultados sobre los objetivos de negocio. | **30%** |
| **Total** | | | | | **100%** |

# Recursos adicionales

IBM, (2012) *Manual CRISP-DM de IBM SPSS Modeler*
ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Kevin Luk, (2019)
Downloading Datasets into Google Drive via Google Colab

# Créditos

Facultad de
INGENIERÍA

**Autores**
Fabio Augusto González Osorio, PhD

**Asistente docente**
Miguel Angel Ortiz Marín

**Diseño instruccional**
Claudia Patricia Rodríguez Sánchez

**Diseño gráfico**
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

**Diagramación PDF**
Daniela Duque García

**Fecha**
**2020-II**

# Programa de formación
# Machine Learning and Data Science MLDS

Facultad de
INGENIERÍA
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA