MIAD

Maestría
en Inteligencia
Analítica de Datos

PROGRAMA DEL CURSO

Machine Learning y Procesamiento de Lenguaje Natural - 2022-12

Generalidades del curso

Datos generales del curso

- Tipo de programa: Posgrados-Maestría en inteligencia Analítica de Datos
- Nombre del curso: Machine Learning y Procesamiento de Lenguaje Natural
- Código (NCR): 10016
- Facultad o Departamento: Departamento Ingeniería Industrial
- Periodo académico: 2022-12
- Horario Sesiones Sincrónicas: Martes 7:00 am – 8:45 am y Jueves 5:30 pm – 7:15 pm

Equipo docente

Profesor a cargo

- Nombre profesor (a): Alejandro Correa Bahnsen
- Correo electrónico: alej-cor@uniandes.edu.co

Líder de Tutores

- Nombre profesor (a): Luisa Fernanda Roa Ballen
- Correo electrónico: lf.roa10@uniandes.edu.co

Tutores

- Nombre profesor (a): Ana María Quintero
- Correo electrónico: am.quintero12@uniandes.edu.co

- Nombre profesor (a): Jaime Acevedo
- Correo electrónico: jd.acevedo244@uniandes.edu.co

- Nombre profesor (a): Alejandro Mantilla
- Correo electrónico: a.mantillar@uniandes.edu.co

- Nombre profesor (a): Camilo Barjas
- Correo electrónico: ca.barajasb@uniandes.edu.co

- Nombre profesor (a): Nicolás Mejía
- Correo electrónico: n.mejia10@uniandes.edu.co

Horarios de atención a estudiantes

Los horarios de atención serán liderados por cada tutor del curso, en la siguiente tabla se especifica los días de las semana y horarios respectivos.

Tutor	Correo	Día	Hora
Ana Quintero	am.quintero12@unaindes.edu.co	Miercoles	8:00 am
Jaime Acevedo	jd.acevedo244@unaindes.edu.co	Jueves	8:00 am
Alejandro Mantilla	a.mantillar@uniandes.edu.co	Sábado	7:00 pm
Camilo Barajas	ca.barajasb@uniandes.edu.co	Martes	6:00 pm
Nicolas Mejia	n.mejia10@uniandes.edu.co	Viernes	10:00 am
Luisa Roa	lf.roa10@unaindes.edu.co	Lunes	6:00 pm

Introducción al curso

Machine Learning y Procesamiento de Lenguaje Natural

Descripción del Curso

El machine learning es un campo que ha revolucionado el mundo al proveer soluciones efectivas a diferentes problemas a través de algoritmos capaces de aprender patrones en un conjunto particular de datos. Así mismo, el procesamiento de lenguaje natural utiliza diferentes técnicas y algoritmos para comprender el lenguaje humano, por lo que es una de las áreas de aprendizaje automático más aplicadas. Específicamente, en las organizaciones, el uso de machine learning constituye una ventaja competitiva, ya que permite generar soluciones innovadoras y precisas a diferentes problemáticas.

Por lo anterior, en este curso estudiaremos los conceptos y técnicas fundamentales del machine learning y del procesamiento del lenguaje natural (NLP), con un enfoque en problemas reales para su implementación en contextos organizacionales.

Objetivos de Aprendizaje

Al finalizar el curso el estudiante estará en capacidad de:

- Implementar sistemas productivos de Machine Learning para que puedan ser usados en diversos ambientes organizacionales.

- Crear modelos de Machine Learning de acuerdo con las necesidades particulares de una organización.
- Implementar modelos de procesamiento de Lenguaje Natural en contextos organizacionales pertinentes.
- Seleccionar modelos predictivos, con base en sus implicaciones técnicas, para que puedan ser usados en contextos organizacionales.

Competencias a desarrollar

Las habilidades que se desarrollan en este curso son:

- Pensamiento crítico y resolución de problemas de analítica de datos y machine learning.
- Comunicación verbal y escrita efectiva al sustentar proyectos de machine learning.
- Colaboración y trabajo en equipo para proyectos de analítica de datos y machine learning.

Contenido de la asignatura

El curso está compuesto por 8 módulos o semanas. Cada módulo tiene sus objetivos y un conjunto de lecciones con distintas actividades para el estudiante.

Durante las semanas 1 y 2 se presentan modelos basados en árboles de decisión y ensamblajes, particularmente se identifica sus aplicaciones y se implementan distintos modelos de estos tipos. En la semana 3, se estudiará el proceso de disponibilización de modelos mediante microservicios, creando APIs locales y alojadas en la nube para disponibilizar modelos en un contexto organizacional. Entre las semanas 4 y 5, se estudiarán los conceptos base del procesamiento de leguaje natural, así como diferentes técnicas de preprocesamiento de textos para crear e implementar modelos predictivos basados, precisamente, en datos de textos. Luego, en la semana 6 se introducen los modelos de redes neuronales artificiales, y finalmente, en las dos últimas semanas del curso, se profundiza en distintos tipos de redes neuronales, su calibración y aplicación para resolver problemas de procesamiento de lenguaje natural.

En la siguiente tabla encontrará los objetivos de aprendizaje asociados a cada uno de los módulos o semanas:

Módulo	Objetivos
1. Árboles de decisión y ensamblaje	• Reconocer las técnicas de uso de árboles de decisión y ensamblaje. • Identificar la relación de cada uno de los parámetros de los modelos de ensamblaje con su ejecución. • Crear e implementar árboles de decisión y ensamblajes de forma manual.

2. Random Forest y XGBoost	• Reconocer las características de los modelos de ensamblaje Random Forest y XGBoost y sus respectivos parámetros de ejecución. • Ajustar los parámetros de los modelos Random Forest y XGBoost para lograr su mejor desempeño. • Crear e implementar modelos de ensamblaje Random Forest y XGBoost.
3. Machine Learning como servicio en la nube	• Seleccionar la combinación de algoritmos que permita obtener el mejor desempeño productivo de un modelo. • Crear e implementar interfaces de Programación de Aplicaciones (APIs) en servicios de la nube para poder disponibilizar un modelo en un contexto organizacional.
4. Introducción al Procesamiento del Lenguaje Natural (NLP)	• Reconocer las características y la utilidad de los modelos de procesamiento de lenguaje natural. • Reconocer e implementar distintas técnicas para el preprocesamiento de textos en modelos NLP.
5. Análisis de sentimientos	• Reconocer e implementar distintas técnicas para el preprocesamiento de textos en modelos NLP. • Crear e implementar modelos predictivos basados en datos de textos.
6. Introducción a Redes Neuronales	• Reconocer el concepto de red neuronal, sus respectivos tipos y aplicaciones. • Reconocer los parámetros para la implementación de redes neuronales. • Crear e implementar redes neuronales para modelos de clasificación, usando librerías especializadas en Python.
7. NLP usando Redes Neuronales I	• Reconocer el concepto de red neuronal, sus respectivos tipos y aplicaciones. • Crear el mejor modelo de predicción de NLP, de acuerdo con el contexto particular de una organización. • Implementar proyectos de NLP que abarquen el proceso de selección de datos, preprocesamiento, modelación, análisis de resultados y disponibilización.
8. NLP usando Redes Neuronales II	• Ajustar los parámetros de ejecución de redes neuronales para lograr su mejor desempeño. • Crear el mejor modelo de predicción de NLP, de acuerdo con el contexto particular de una organización. • Implementar proyectos de NLP que abarquen el proceso de selección de datos, preprocesamiento, modelación, análisis de resultados y disponibilización.

Metodología

Las actividades del curso están diseñadas para comprender las temáticas asociadas a cada una de las semanas, adquirir habilidades para desarrollar modelos en Python e integrar y aplicar los contenidos tratados en la resolución de distintos tipos de problemas predictivos. Cada semana cuenta con videos, lecturas de profundización, tutoriales y cuestionarios, que le permitirá al estudiante prepararse para los talleres individuales y grupales que se han propuesto durante las semanas.

Adicionalmente, en el curso se desarrollarán dos proyectos grupales en los que los estudiantes deberán resolver problemas predictivos usando datos reales. Lo anterior, con el propósito de poder consolidar sus aprendizajes y de aplicar las distintas temáticas en el desarrollo de proyectos de ciencia de datos en contextos organizacionales.

Secciones en vivo

El curso cuenta con sesiones en vivo (sincrónicas) en las que se resolverán dudas, se profundizará en los temas vistos y se retroalimentarán de manera general las actividades realizadas hasta el momento.

Por semana se harán dos sesiones de 1 hora y 45 minutos de duración, en las que se tratarán los mismos temas, pero en horarios diferentes. Lo anterior con el propósito de que los estudiantes tengan la posibilidad de escoger el horario que más se ajuste a sus necesidades para asistir a una sola de las dos sesiones.

Los horarios son los siguientes:

- Martes de 7:00 a 8:45 a.m.
- Jueves de 5:30 a 7:15 p.m.

Herramientas y requerimientos tecnológicos

En el curso se hará uso de lenguaje de programación Python, particularmente se usarán Jupyter nootebooks para implementar bloques de código. Las principales librerías para usar son:

- Scikit-Learn
- Spacy
- Pytorch
- Keras
- Tensorflow
- Flask

Adicionalmente, se usará Amazon Web Services (AWS), Git y Kaggle para complementar el desarrollo del curso.

Criterios de evaluación y aspectos académicos

Evaluación

El curso cuenta con evaluaciones formativas y sumativas semanales, que aportan a la apropiación de los conceptos dados en el curso y permiten dar cuenta del cumplimiento de los objetivos de aprendizaje propuestos en cada semana.

El aporte de cada una de las actividades calificables en la nota definitiva del curso son los siguientes:

Cuestionarios semanales (20%, cada uno 2.5%)

- Árboles de decisión y ensamblajes.
- Random Forest y XGboost.
- Machine Learning como servicio en la nube.
- Introducción al procesamiento de lenguaje natural.
- Análisis de sentimientos.
- Introducción a redes neuronales.
- NLP usando redes neuronales.
- Calibración de parámetros en redes neuronales.

Talleres individuales y grupales (25%, cada uno 5%)

- Taller individual: Construcción e implementación de árboles de decisión y métodos de ensamblaje.
- Taller grupal: Construcción e implementación de modelos Bagging, Random Forest y XGBoost.
- Taller grupal: Tokenización de textos
- Taller grupal: Análisis de sentimientos y técnicas de NLP.
- Taller grupal: Redes Neuronales.

Proyectos grupales (50%, cada uno 25%)

- Proyecto 1: Predicción de precios de vehículos usados.
- Proyecto 2: Clasificación de género de películas.

Actividades de coevaluación del trabajo en equipo (5%, cada uno 2.5%)

- Actividad de coevaluación semana 5.
- Actividad de coevaluación semana 8.

Dedicación

Este es un curso de tres créditos, por lo que tiene una dedicación aproximada de 18 horas semanales. Sin embargo, es importante tener en cuenta que en la plataforma el tiempo estimado por semana es menor, debido a que no contempla el tiempo que los estudiantes podrían dedicar a la toma de apuntes y al repaso de los contenidos que se encuentran en los recursos.

Parámetros de calificación de actividades académicas

Cada una de las actividades calificables cuentan con su respectiva rubrica de evaluación.

Bibliografía

- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".
- Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series* (Vol. 1142, No. 1, p. 012012). IOP Publishing.