Ciencias Sociales
Universidad de la República
URUGUAY

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Programa de Curso

Limpieza y transformación de datos

Especialización en Economía, opción Ciencia de Datos

Docente/s y correo electrónico: Gladys Cardozo - ing.gladyscardozo@gmail.com

Créditos: 4 créditos (20 horas)

Régimen de cursado: Presencial

Carga y distribución de horas estimada: 4 horas semanales

Conocimientos previos recomendados: Python, Jupyter notebook, Estadística.

Objetivos: Esta materia tiene como objetivo que los/las estudiantes aprendan técnicas de limpieza de datos y transformación que les permitan definir el modelo de datos que mejor se adapte a sus fines de análisis.

Contenidos y organización del curso: Esta materia tiene como objetivo que los/las estudiantes aprendan técnicas de limpieza de datos y transformación que les permitan definir el modelo de datos que mejor se adapte a sus fines de análisis. A lo largo del curso se presentarán una cantidad de ejemplos y ejercicios prácticos para la evaluación continua del estudiante. La idea es que los estudiantes puedan trabajar sobre los ejercicios prácticos durante la clase y al final del curso ser capaces de construir y depurar un conjunto de datos, dejándolo listo para el próximo paso (un análisis o el input de un modelo de aprendizaje automático).

Programa:

- Introducción a la limpieza de datos. Objetivo e importancia de la misma. Importancia de la calidad de los datos. Introducción a herramientas y librerías. Primeros pasos de manipulación de datos. Introducción a técnicas básicas de limpieza (variables únicas, variables con baja varianza, valores duplicados).

- Tratamiento de outliers. Diferentes técnicas de identificación y tratamiento. Identificación de valores faltantes y métodos de imputación simples. Métodos de imputación avanzados.

- Selección de features ordinales y categóricas. Tratamiento de diferentes tipos de variables y transformación de las mismas. Uso de “Feature Importance”.

- Como transformar y normalizar variables numéricas. Como transformar y codificar variables categóricas. Normalización y estandarización de otros casos.

- Manejo de variables de tipo fecha y texto. Uso de expresiones regulares. Extracción, manipulación y creación de nuevas variables.

- Manejo de sub dataset. Filtrado de datos condicional. Almacenamiento y recuperación.

- Reducción de dimensionalidad. Introducción al Análisis de componentes principales (PCA) y a la descomposición en valores singulares (SVD).

- Introducción al armado de bases de datos automáticas. Automatización del proceso de limpieza y transformación.

- Estudio de un caso real aplicando todos los conocimientos abordados y presentación del trabajo final.

Método de enseñanza: Teórico - Práctica

Sistema de evaluación: La evaluación consiste en la entrega de los ejercicios prácticos abordados en clase (20%) y un trabajo final individual (80%).

Bibliografía:

[1] Brownlee, Jason (2020). “Data Preparation for Machine Learning – Data Cleaning, Feature Selection, and Data Transforms in Python”

[2] Google Colab: https://colab.research.google.com/

[3] Pandas: https://pandas.pydata.org/docs/reference/frame.html

[4] Numpy: https://numpy.org/doc/stable/

[5] Scikit-learn: https://scikit-learn.org/stable/

[6] Kaggle: https://www.kaggle.com/