

Python Aplicado a la Ciencia de Datos

Franco Calle

Resumen y Objetivos

Python es un lenguaje de programación cuya versatilidad para el procesamiento de datos y capacidad para el análisis funcional lo hace cada vez más popular. Su popularidad ha crecido tanto en los últimos años que en el sector privado y la academia cada vez lo utilizan más para lograr distintos objetivos relacionados a la ciencia de datos como la limpieza y procesamiento de datos, automatización de procesos, hasta predicción y modelamiento del comportamiento humano.

El objetivo del curso es brindar al estudiante los fundamentos y lógica de programación en Python que sirvan como herramienta para abordar y resolver problemas que podrían ser aplicados en el sector privado y en la academia. Específicamente, al finalizar el curso, el estudiante podrá reconocer y aplicar de manera conjunta objetos como listas, dataframes, diccionarios, tuplas y funciones para resolver un problema de predicción y validarlo en la vida real.

Introducción

- ¿Por qué es importante aprender un lenguaje de programación?
- Curva de aprendizaje y alcance de distintos lenguajes de programación (R, Stata, C+, Python)
- ¿Qué es Data Science? ¿Qué es Machine Learning?
- ¿Por qué Python?

Primer Módulo: Elementos Básicos de Python

Primera Clase: Variables, expresiones, y statements

- Values, variable names, and keywords
- Operators, operands, expressions, the order of operations, string operations

- Asking for inputs to the user
- Mnemonic variable names

Segunda Clase: Objetos de Python

- Lists
- Dictionaries
- Tuples
- Multiple assignments with dictionaries

Tercera Clase: Ejecución Condicional

- Boolean Expressions
- Logical operators
- Conditional, alternative, and chained conditional executions
- Nested conditionals
- Guardians: catching expressions using try and except
- Short-circuit evaluation of logical expressions

Cuarta Clase: Iteraciones

- Updating variables
- Definite loops using for
- Double, multiple and nested iteration
- While statement
- List comprehension

Quinta Clase: Funciones

- The flow of execution, arguments, and parameters
- Adding new functions
- Definitions and uses
- Anonymous functions Lambda
- Mapping and filtering
- Numpy module
- Pandas module

Sexta Clase: DataFrames

- Using and creating Dataframes
- Replace and rename columns
- Slicing Dataframes
- Merge, Append
- Import, Export

Segundo Módulo: El problema de predicción y el uso de Machine Learning

Primera Clase: Predicción y Métricas de Performance

- The Prediction Problem
- Confusion Matrix
- Accuracy, Precision, Recall
- The receiving operations curve and Area Under the Curve

Segunda Clase: Métodos de Machine Learning para Clasificación

- Logistic Regression
- K-Nearest Neighbors Classification
- Supervised Machine Learning
- Multi – Class Classification
- Decision Trees
- Random Forests

Tercera Clase: Métodos de Machine Learning para Regresión

- Least Squares
- Ridge
- Polynomial Regression
- R-squared, MSE

Cuarta Clase: Validación Cruzada y Composición del Modelo

- Cross Validation
- Grid Search
- Contribución de la data al modelo
- Contribución de las variables al modelo

Proyecto Final

El proyecto final requiere utilizar todos los conocimientos aprendidos en el Módulo 1 y 2 del curso para resolver un problema de predicción aplicado a las ciencias sociales utilizando data del INEI.