INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Projeto Capacitação em Ciência de Dados para o SEN e AP Investimento apoiado pelo Plano de Recuperação e Resiliência (PRR) e pelos Fundos Europeus Next Generation EU[i]

Programação Metodológica

1. ORGANIZADOR

Instituto Nacional de Estatística

2. IDENTIFICAÇÃO DO/A FORMADOR/A

Nome	N.º Emp (aplicável a trabalhadores do INE)
Bartholomeus Johannes Schoenmakers	12432

N.º CCP	Data de emissão	Data de validade (se aplicável)[1]
F730743/2022	27/07/2022	Clique ou toque para introduzir uma data.

Nome	N.º Emp (aplicável a trabalhadores do INE)
Luis Osório do Carmo Ferreira	16667

N.º CCP	Data de emissão	Data de validade (se aplicável)[1]
F730358/2022	20/07/2022	Clique ou toque para introduzir uma data.

Nome	N.º Emp (aplicável a trabalhadores do INE)
Sónia Patrícia Folgado C. Borges Quaresma Gonçalves	14508

N.º CCP	Data de emissão	Data de validade (se aplicável)[1]
EDF 460645/2007 DC	14/11/2007	Clique ou toque para introduzir uma data.

3. IDENTIFICAÇÃO DA AÇÃO

Designação

Python para estatísticas oficiais - Advanced Data Science

Formato	Local
Presencial	Delegação do Porto do INE

Data de Início	Data de Fim	Horário	Duração
06/05/2024	14/05/2024	ver observações	30 horas

[1] Conforme indicado na Portaria n.º 994/2010, de 29 de setembro, os CAPs, emitidos ao abrigo do Decreto regulamentar n.º 66/94, de 18 de novembro, com as alterações introduzidas pelo Decreto Regulamentar n.º 26/97, de 18 de junho, incluindo aqueles que tenham sido renovados nos termos do disposto na Portaria n.º 1119/97, de 5 de novembro, consideram-se emitidos sem dependência de qualquer período de validade, não carecendo de ser objeto de renovação.

Observações sobre a calendarização, horário ou outra

Módulo 1
Formador Luís Ferreira
06/05/2024 - 9:30 -12:30/14:00 – 17:00 (6 horas)

Módulo 2
Formadora Sónia Quaresma
07/05/2024 - 9:30 -12:30/14:00 – 17:00 (6 horas)
08/05/2024 - 9:30 -12:30/14:00 – 17:00 (6 horas)

Módulo 3
Formador Bartholomeus Schoenmakers
13/05/2024 - 9:30 -12:30/14:00 – 17:00 (6 horas)
14/05/2024 - 9:30 -12:30/14:00 – 17:00 (6 horas)

Objetivos

No final do curso os formandos deverão ser capazes de:

- Programar em Python usando Classes (Herança e Polimorfismo) e Módulos
- Definir Túplos, Sequências e Sets em Python
- Criar funções em Python sobre Colecções
- Utilizar Programação recursiva em Python
- Explorar Expressões regulares em Python Comprehensions
- Utilizar o package Scikit
- Lidar em Python con Missing Values: Discarding instances, acquiring missing values or performing imputation
- Aplicar Métodos de Imputação: predictive value imputation, unique-value imputation or using reduced-feature Models
- Efectuar Normalização e Standardização dos Dados usando Pipelines de pré-processamento
- Distinguir entre Modelos de Regressão e Classificação
- Utilizar Modelos de Regressão com Regularização: Ridge e Lasso
- Utilizar Modelos de Classificação (K Nearest Neighboors, Random forests, Support Vector Machines)
- Ser capazes de avaliar os Modelos gerados: Training e Testing Errors, K-Fold cross validation e Receiver Operating Characteristics – ROC (and AUC)
- Fazer manipulações avançadas de dados geográficos
- Utilizar JSON e GEOJSON
- Aceder dados por API
- Obter dados geográficos utilizando o Address Matching com Open Street Map, Bing ou Google
- Utilizar Técnicas de Analise geográfica espacial

Programa

Configuração da Formação - O tema da formação apresenta um grau de complexidade elevado, requerendo foco e capacidade de abstração.
A formação terá uma configuração presencial, de forma a potenciar o envolvimento e motivação dos formandos.

Estrutura Modular e Respectiva Carga Horária - Os conteúdos programáticos do curso de Advanced Data serão organizados em 3 módulos, segundo um desenvolvimento sequencial.

Módulo I - 6 horas - O módulo consiste no aprofundamento do conceito de Classe e na introdução a ferramentas mais complexas da programação.
Os tópicos abordados no módulo são:

- Classes (Herança e Polimorfismo)
- Túplos, Sequências e Sets
- Colecções
- Programação recursiva
- Expressões regulares
- Tratamento de Exceções
- Zen Of Python
- ChatGPT

Módulo 2 – 12 horas - O modulo é dedicado à aprendizagem supervisionada usando o scikit. A aprendizagem supervisionada é, uma subcategoria de machine learning e de inteligência artificial. Caracteriza-se pelo uso de labeled datasets que são usados para treinar algoritmos que classificam ou preveem resultados com precisão. Os dados de input são inseridos no modelo, e a precisão do algoritmo é medida através de uma função de perda, que minimizando o erro permite que o modelo seja ajustado adequadamente. Além dos algoritmos essenciais para a aprendizagem supervisionada serão discutidos os modos de avaliação dos diferentes tipos de modelos: regressão e classificação, bem como o pré-processamento dos dados essencial para a qualidade do modelo.

- Scikit package
- Missing Values: Discarding instances, acquiring missing values or performing imputation
- Métodos de Imputação: predictive value imputation, unique-value imputation or using reduced-feature Models
- Multi-collinearity: highly correlated variables
- Class Imbalance: Undersampling, oversampling and synthetic data generation (SMOTE)
- Normalização e Standardização dos Dados
- Pipelines de pré-processamento
- Aprendizagem Supervisionada: Modelos de Regressão e Classificação
- Regularização: Ridge e Lasso
- Entropy e Information gain: CART trees
- Other models – K Nearest Neighboors, Random forests, Support Vector Machines, Neural Networks
- Avaliação de modelos: Training e Testing Errors, K-Fold cross validation e Receiver Operating Characteristics – ROC (and AUC)

Módulo 3 – 12 horas - Neste módulo, aprofundamos os diferentes aspectos da manipulação de dados geográficos. Faz-se a introdução de analise espacial de dados geográficos e apresenta-se algumas técnicas de análises espaciais. Outros dos objetivos é aprender como obter a geografia de endereços, mostrar os formatos de dados JSON e GEOJSON e como obter dados através de um API. Os tópicos abordados no módulo são:

- Manipulação avançada de dados geográficos
- Análise De Dados Espaciais (Geoestatística)
- Autocorrelação Espacial
- JSON
- Obtenção de dados via API no formato JSON
- Georreferenciação (Geocoding) de Moradas

Metodologia

Será usada uma mistura de métodos, recorrendo maioritariamente ao método demonstrativo como fio condutor da narrativa possibilitando o desenvolvimento sequencial da apresentação dos conteúdos.
Numa primeira fase, sincrética, o formador proporcionará uma visão global da operação que o formando deverá aprender, explicando globalmente o conteúdo, o processo e o tempo necessário para o realizar. Esta demonstração / explicação é feita
num ritmo normal de execução da tarefa em aprendizagem.
Este método será intercalado com o método activo, para levar os formandos a experimentar e manipular os materiais e permitir aferir a aquisição de alguns conceitos no decurso da formação.
Serão utilizadas ainda técnicas pedagógicas de tipo expositivo (por exemplo, em momentos de introdução de temas, sínteses parciais e finais).

Recursos Pedagógicos

A acção de formação terá em sala um computador portátil para o formador e um para cada formando, um videoprojector e um quadro branco com as respectivas canetas.

Recursos Materiais e Logísticos

Critérios de seleção dos espaços de formação:

- Computadores funcionais equipado com placa de som, microfone, colunas de som e com ligação à Internet e à intranet e pré-instalação de oracle, Ambiente Anaconda com os pacotes de Python instalados NumPy, Pandas, Scikit-learn, Seaborn , Matplotlib, SciPy, PyPlot, Getpass, cx_Oracle, instalação de Visual Studio Code com Plugin de Pyhton da Microsoft
- Sala com boa luminosidade, ventilação, temperatura e isolada de ruídos perturbadores ao bom funcionamento;
- Espaço equipado com todos os recursos didáticos necessários;
- Mobiliário que respeite as regras de ergonomia dos formadores e dos formandos;
- Espaço amplo o suficiente para permitir a concretização de dinâmicas de grupo e da disposição das mesas em IUI, no sentido da facilitação da comunicação;
- Local de fácil acesso;
- Espaço cuja limpeza é assegurada diariamente.

Avaliação

Não haverá avaliação. Será emitido um certificado de participação no curso se o formando tiver registado uma assiduidade mínima de 95% sobre a duração global do curso.

4. DESTINATÁRIOS

Destinatários	N.º Participantes
Técnicos do INE e das EDCs	12 (n.º máximo)

Requisitos Prévios

[i] www.recuperar Portugal.gov.pt