

Artificial Intelligence Skills Alliance (ARISA)



www.aiskills.eu



ARISA, the project

ARISA supports the development of Artificial Intelligence (AI) skills in Europe.



Focus on **Artificial Intelligence**



4-year Erasmus+ funded project (2022-2026)

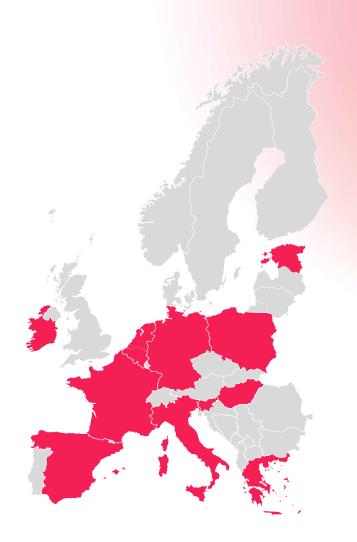


20 partners from 12 EU countries



Consortium

- DIGITALEUROPE
- CCIS
- IVSZ
- DIGITAL SME
- EXELIA
- CIMEA
- ASIIN
- Uni of Utrecht Applied Technology
- Budapest Uni of technology
- Uni of Ljubljana
- Warsaw School of Computer Science
- Uni Internacional De La Rioja Sa
- DTSL
- Adecco Formazione
- Global Knowledge
- BCS Koolitus



List of Associated Partners: https://aiskills.eu/the-alliance/





• Who to upskill and reskill?

ARISA focuses on 4 occupational domains. Under those domains, we look at current and emerging professional roles that are relevant to ARISA.



Business Leader



Technology Leader



Technology Practitioner



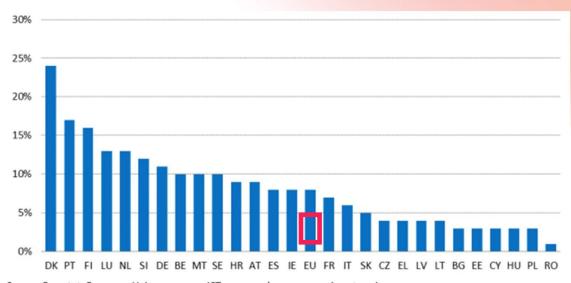
Policymaker



• Why this project?

- Urgent market needs for skilled employees in Al
- Al skills mismatches between the education & training offerings and industry requirements
- Discrepancies in AI technology uptake across Member States

Enterprises using an AI technology (% of enterprises), 2021



Source: Eurostat, European Union survey on ICT usage and e-commerce in enterprises.



You are here because...

ARISA is testing curricula and training programmes related to emerging job profiles and roles in the field of AI.

9 Testing sites in Europe and Ukraine

 You are among the 200+ learners testing the ARISA training programmes!

Data Science for Engineers and Economists using Python

March-April 2025

How you can contribute

- Participate actively
- Assess your learning experience by answering to the evaluation questionnaire at the end of the course
- Provide your testimonial (article, video, podcast)











www.aiskills.eu



ARISA Learning Material

Educational Profile and EQF level: DATA SCIENTIST EQF 6

PLO: 1, 2, 3, 4, 5

Learning Unit (LU): MACHINE LEARNING: SUPERVISED

Topic: 1. INTRODUCTION



www.aiskills.eu

Copyright © 2024 by the Artificial Intelligence Skills Alliance

All learning materials (including Intellectual Property Rights) generated in the framework of the ARISA project are made freely available to the public under an open license Creative Commons Attribution—NonCommercial (CC BY-NC 4.0).

ARISA Learning Material 2024

This material is a draft version and is subject to change after review coordinated by the European Education and Culture Executive Agency (EACEA).

Authors: Universidad Internacional de La Rioja

Disclaimer: This learning material has been developed under the Erasmus+ project ARISA (Artificial Intelligence Skills Alliance) which aims to skill, upskill, and reskill individuals into high-demand software roles across the EU.



This project has been funded with support from the European Commission. The material reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



About ARISA

- The Artificial Intelligence Skills Alliance (ARISA) is a four-year transnational project funded under the EU's Erasmus+ programme. It delivers a strategic approach to sectoral cooperation on the development of Artificial Intelligence (AI) skills in Europe.
- ARISA fast-tracks the upskilling and reskilling of employees, job seekers, business leaders, and policymakers into AI-related professions to open Europe to new business opportunities.
- ARISA regroups leading ICT representative bodies, education and training providers, qualification regulatory bodies, and a broad selection of stakeholders and social partners across the industry.

ARISA Partners & Associated Partners I LinkedIn I Twitter



FECHAS: Data Science for Engineers and Economists using Python

- 11 Marzo, 13 Marzo, 18 Marzo, 20 Marzo, 25 Marzo, 27 Marzo, 1 Abril, 3 Abril
- 10-12 Colombia (16-18 CET)
- Después de cada clase se subirá presentación y código
- Trabajo Final estará disponible la semana del 1 de Abril. Se podra entregar hasta el 20 de Abril. Se entregara un notebook.
- El 3 de Abril haremos una encuesta.
- ¿Alguna fecha es problematica?



TEMARIO (APROX)

1. Capítulo 1: Introducción a Ciencia de Datos y Series Temporales

- Clase 1: Introducción a Ciencia de Datos y ST
- Clase 2: Análisis de Series Temporales (Prophet)

2. Capítulo 2: Aprendizaje Automático Supervisado

- Clase 3: K-NN, Modelos Lineales y Sobreajuste
- Clase 4: Validación cruzada y arboles de decisión
- Clase 5: Métricas y Optimización de Modelos
- Clase 6: Redes Neuronales y DL

3. Capítulo 3: Aprendizaje Automático No Supervisado

Clase 7: Conceptos Básicos + PRESENTACION TRABAJO

4. Capítulo 4: IA Responsable

• Clase 8: IA Responsable y Ética + ENCUESTA FINAL



FECHAS: Data Science for Engineers and Economists using Python

- Duración 2 MESES
- CLASES PRESENCIALES: 1 MES
 - 11 Marzo, 13 Marzo, 18 Marzo, 20 Marzo, 25 Marzo, 27 Marzo, 1 Abril, 3
 Abril
- TRABAJO PRACTICO: 2 Semanas
- TUTORAS: 2 Semanas
- Es un curso de upskilling pensado para EQF 6 y 7



Portal Educativo OpenED

- https://kt.unir.net/opened/course/view.php?id=27
- Grabaciones y contenido de las Sesiones
- Capitulo 0. Introducción a Python.
- Encuesta final.
- Trabajo final.
- Anuncios
- Foro de Debates (usadlo, somos muchos...)



Python

- Introducción a los Notebooks
- Tipos de datos básicos
- Listas y Diccionarios
- Control
- Pandas
- Numpy
- Matplotlib



Conda y Notebooks

- https://www.anaconda.com/products/individual
- Entornos:
- https://docs.anaconda.com/navigator/tutorials/manageenvironments/
- Anaconda cloud + Anaconda Assistant (error correction)
- PyCharm u otros (IDEs)



Evaluación

- Trabajo final
- 3 partes
- 0-5, con 3 suficiente?, A/B/C/D?
- Aplicación a un dataset real que os sera proporcionado
- Se entregarÁ por la Plataforma mediante un notebook
- Dos Semanas despues de la finalizacion de la clase (17 de Abril)
- Necesario para el curso
- EVALUACION FINAL: 40% ASISTENCIA PARTICIPACION + 60% TRABAJO
- SE EMITE EL CERTIFICADO



scikit-learn

scikit-learn Homepage: http://scikit-learn.org/

- scikit-learn User Guide <u>http://scikit-learn.org/stable/user_guide.html</u>
- scikit-learn API reference http://scikit-learn.org/stable/modules/classes.html



Pandas: Data Manipulation

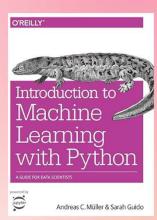
Provides key data structures like DataFrame

Support for reading/writing data in different formats



References

- "Introduction to Machine Learning with Python"
- "Tensorflow for Deep Learning: From Linear Regression to Reinforcement Learning"





 Content transformed and build upon "Applied Machine Learning in Python" from Kevyn Collins-Thomson, University of Michigan under CC BY-NC 3.0



Ciencia de datos y aprendizaje automático

- Ciencia de datos: proceso de extraer información de los datos. Esto incluye la recopilación de datos, la limpieza, la exploración, la visualización, el análisis estadístico y la aplicación de modelos de aprendizaje automático.
- Aprendizaje automático (ML): el ML es un subconjunto de la ciencia de datos centrado específicamente en la creación de algoritmos que permiten a los ordenadores aprender de los datos y hacer predicciones o decisiones sin necesidad de programarlos explícitamente.



¿Que es el Aprendizaje Máquina?

- El estudio de programas informáticos que aprenden mediante ejemplos
- Los algoritmos de aprendizaje automático pueden generalizar a partir de ejemplos:

dado un conjunto de entrenamiento de imágenes etiquetadas, un clasificador de imágenes puede averiguar cómo aplicar etiquetas de manera precisa a nuevas imágenes no vistas previamente



Machine Learning aprende mediante ejemplos

La IA tradicional se basa en reglas...

Los algoritmos aprenden reglas a partir de ejemplos etiquetados

Un conjunto de ejemplos etiquetados usado para el aprendizaje se llama datos de entrenamiento

Las reglas aprendidas también deberían ser capaces de generalizar para reconocer o predecir correctamente nuevos ejemplos que no están en el conjunto de entrenamiento



Machine Learning se basa en experiencia...

- Ejemplos etiquetados (spam de correo electrónico)
- Opiniones de usuario
- Etiquetado de imágenes medicas
- Estos datasets son costosos de obtener



El aprendizaje automático es la suma de

- Métodos estadísticos
- Informática
 - Algoritmos para capturar, manipular, indexar, combinar, recuperar y realizar predicciones sobre datos
 - Arquitecturas informáticas a gran escala
- Economía, biología, psicología, etc.



El aprendizaje automático está en todas partes....

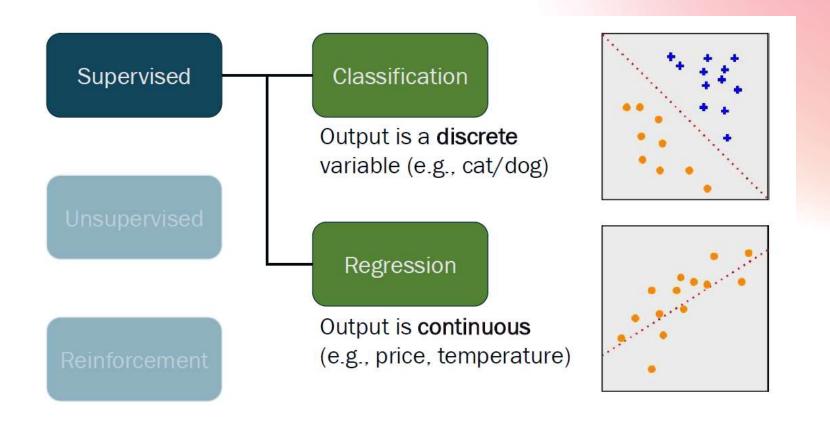
- Finanzas: detección de fraudes, calificación crediticia
- Búsqueda en la web
- Reconocimiento de voz
- eCommerce: Recomendaciones de productos
- Filtrado de correo no deseado
- Aplicaciones sanitarias: diseño y descubrimiento de fármacos
- Educación: Puntuación automatizada de ensayos



Tipos de aprendizaje automático

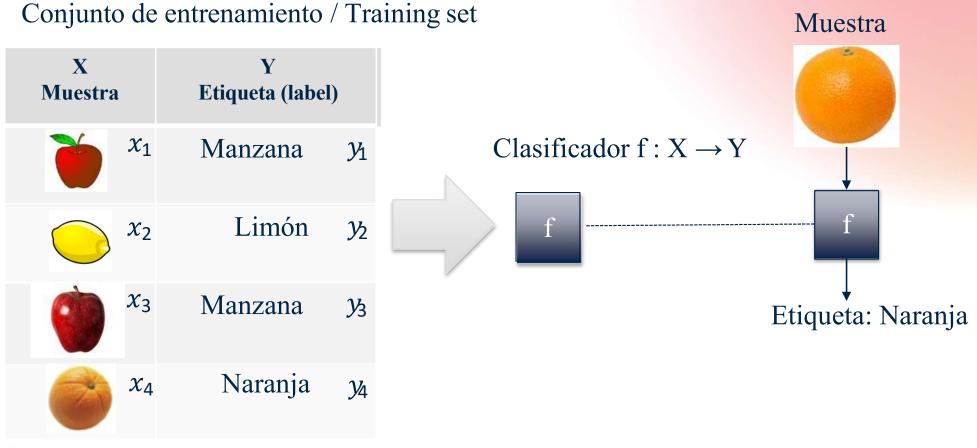
- Aprendizaje automático supervisado: aprenda a predecir los valores objetivo a partir de datos etiquetados.
 - Clasificación (los valores objetivo son clases discretas)
 - Regresión (los valores objetivo son valores continuos))
- Aprendizaje automático no supervisado: encuentre la estructura en datos sin etiquetar.
 - Buscar grupos de instancias similares en los datos (agrupación en clústeres)
 - Búsqueda de patrones inusuales (detección de valores atípicos)







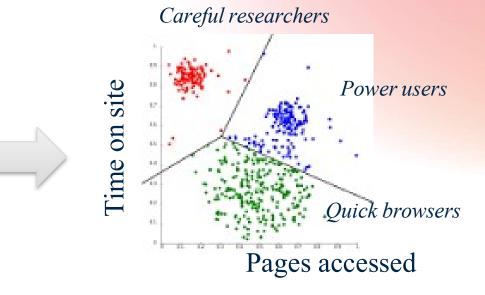
Aprendizaje Supervisado: Clasificación





Aprendizaje automático: no supervisado

Búsqueda de clústeres de usuarios similares (agrupación en clústeres)





Pasos básicos del aprendizaje automático



Elegir:

Una representación de características

Tipo de clasificador para ustedse

Elegir:

¿Qué criterio distingue a los clasificadores buenos de los malos?

Elegir:

Cómo buscar la configuración/parámetros que proporcionan el mejor clasificador

para este criterio de evaluación



Representación de características

Email

To: Chris Brooks From: Daniel Romero

Subject: Next course offering

Hi Daniel,

Could you please send the outline for the
 next course offering? Thanks! -- Chris



<u>Feature</u>	Count
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2

Feature representation

A list of words with their frequency counts

Picture







A matrix	ofcolor
values	(pixels)

Sea Creatures



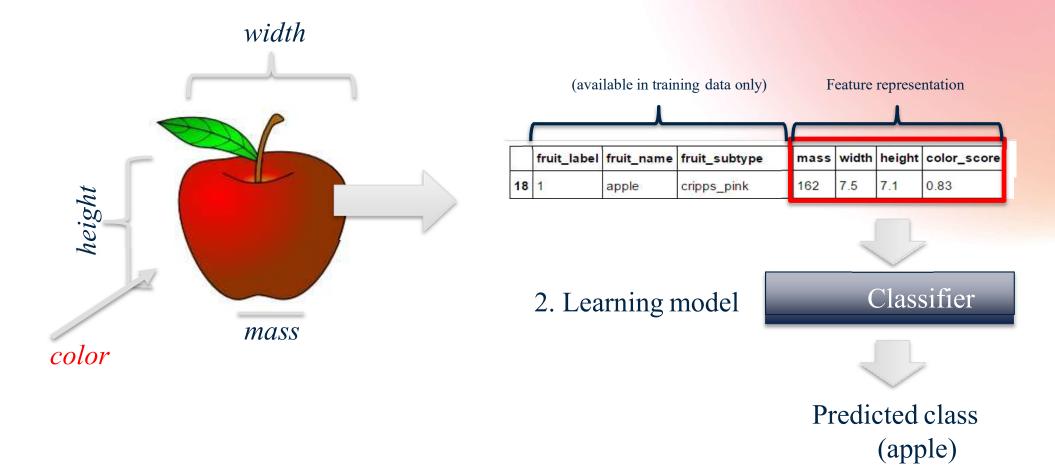


<u>Feature</u>	<u>Value</u>			
DorsalFin	Yes			
MainColor	Orange			
Stripes	Yes			
StripeColor1	White			
StripeColor2	Black			
Length	4.3 cm			

A set of attribute values



Feature Example...





Los pasos de ML son un ciclo que se refina

Representación:

Extraer y
seleccionar
características
de objetos

Entrenar
modelos: Ajustar
el estimador a los
datos
(TRAINING)

Refinamiento

<u>de</u>

<u>características</u>

<u>y modelos</u>

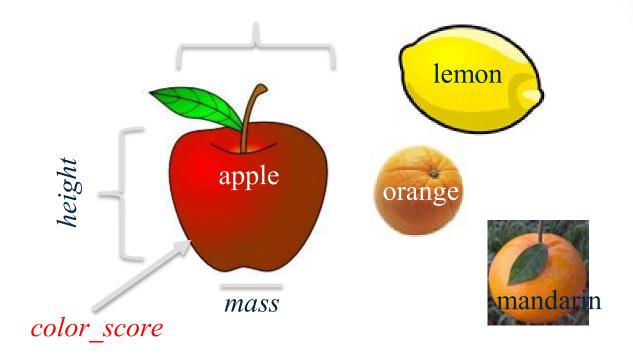
Evaluación (PRUEBAS)



Un ejemplo de problema de aprendizaje automático



The Fruit Dataset



	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	aolden delicious	156	7.6	7.5	0.67

fruit_data_with_colors.txt

Dataset by Dr. Iain Murray, Univ. of Edinburgh https://github.com/rocksaint/fruit-data-with-colours



Colour Scale

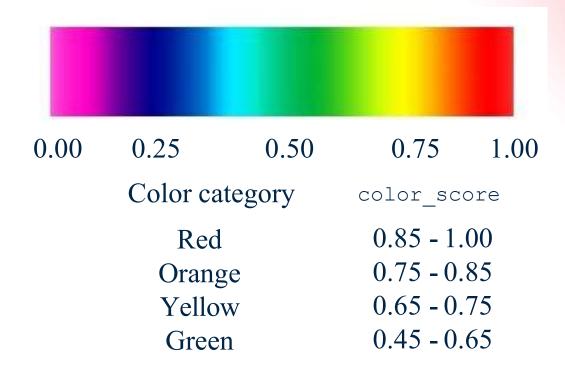
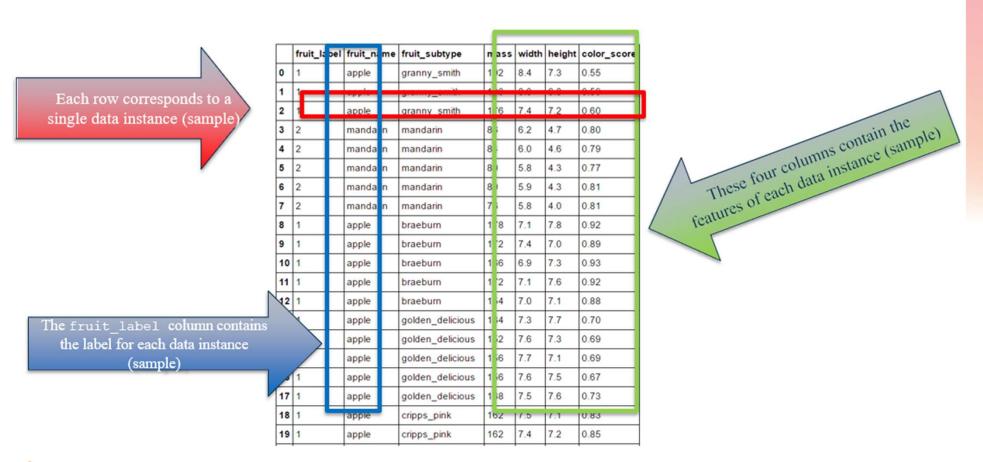




Tabla como datos de entrada para algoritmos de aprendizaje automático...





Pruebas vs. Entrenamiento (Testing vs. Training)

X					У		X_train			y_train		X_test					y_test			
,	heigh	width	mas	s color_score			1 12	T	_	T- 1										
0	7.3	8.4	192	0.55	0	1	1 66	t width		color_score			8	height	width	mass	color_score			
1	6.8	8.0	180	0.59	0 1	1	42 7.2	7.2	154	0.82	42	3	26	9.2	9.6	362	0.74		26	3
2	7.2	7.4	176	0.60	2	1	48 10.1	7.3	174	0.72	48	4	35	7.9	7.1	150	0.75		35	3
3	4.7	6.2	86	0.80	3	2	7 4.0	5.8	76	0.81	7	2	43	10.3	7.2	194	0.70		43	4
4	4.6	6.0	84	0.79	4	2	14 7.3	7.6	152	0.69	14	1	28	7.1	6.7	140	0.72	2	28	3
5	4.3	5.8	80	0.77	5	2 2	32 7.0	7.2	164	0.80	32	3	11	7.6	7.1	172	0.92		11	1
6	4.3	5.9	80	0.81	6	2	49 8.7	5.8	132	0.73	49	4	2	7.2	7.4	176	0.60		2	1
7	4.0	5.8	76	0.81	7	2	29 7.4	7.0	160	0.81	29	3	34	7.8	7.6	142	0.75		34	3
8	7.8	7 1	178	0.92	Q	1	37 7.3	7.3	11	0.79	37	3	46	10.2	7.3	216	0.71		46	4
9	7.0									0.73	56	4	40	7.5	7.1	154	0.78	i.	40	3
10	7.3	+	X t	rain.	X test.	y_train, y_test	- :		\neg /	,	•		22	7.1	7.3	140	0.87	-0	22	1
_	7.6					split(X, y)	39 7.4	6.8	144	0.75	39	3	4	4.6	6.0	84	0.79	-6	4	2
	7.1			CI			39 7.4	0.8	144	0.75	39	3			-	$^{\rm H}/$	0.93		10	1
	7.7	-			1.5		0 47	100	loc	0.80	,	,	130	17.5	11.1	п /	0.79	T.	30	3
	7.3	7.6	152	0.69	14	1	3 4.7	6.2	86		0	1	41	8.2	7.6	180	0.79	8	41	3
5000	7.1	7.7	156	0.69	15	1	0 7.3	8.4	192	0.55	0	1	33	8.1	7.5	190	0.74	is .	33	3
7.50	7.5	7.6	156	0.67	16	1	53 8.4	6.0	120	0.74	53	4		700	1.000.00	Patrick	(Charles of A			
100	7.6	7.5	168	0.73	17	1	47 9.7	7.3	196	0.72	47	4								
_	7.1	7.5	162	0.83	18	1	44 10.5	7.3	200	0.72	44	4								_
	7.2	7.4	162	0.85		1														
19	1.2	1.4	102	0.65	19	1			_	. !										
Original data set				Training set				Test set												



39

Preprocesamiento de datos

- La inspección de los valores de las características puede ayudar a identificar qué limpieza o preprocesamiento aún debe realizarse una vez que pueda ver el rango o la distribución de valores que es típica para cada atributo.
- Datos faltantes o ruidosos
- Inconsistencias, como el uso de un tipo de datos incorrecto para una columna
- Unidades de medida incorrectas para una columna en particular
- No hay suficientes ejemplos de una clase en particular.

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	192
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	apple	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn		7.4	7.0	0.89
10	1	apple	braeburn		6.9	7.3	0.93
11	1	apple	braeburn		7.1	7.6	0.92
12	1	apple	braeburn		7.0	7.1	0.88
13	1	apple	golden_delicious	io.	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69





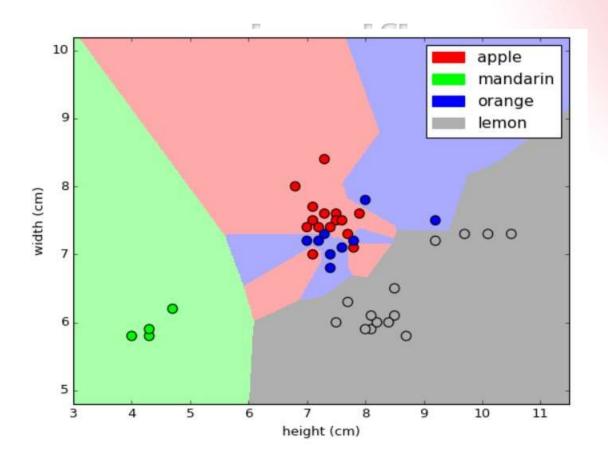
Algoritmo K-NN: K- Vecino más cercano

Dado un conjunto de entrenamiento X_train con etiquetas y_train y dada una nueva instancia x_test se va a clasificar:

- 1. Encuentre las instancias más similares (llamémoslas X_NN) a x_test que están en X_train.
- 2. Obtenga las etiquetas y_NN para las instancias de X_NN
- 3. Prediga la etiqueta para x_test combinando las etiquetas y_NN por ejemplo, mayoría simple de votos



Algoritmo K-NN: K- Vecino más cercano



Límites de decisión del conjunto de datos de frutas con k = 1

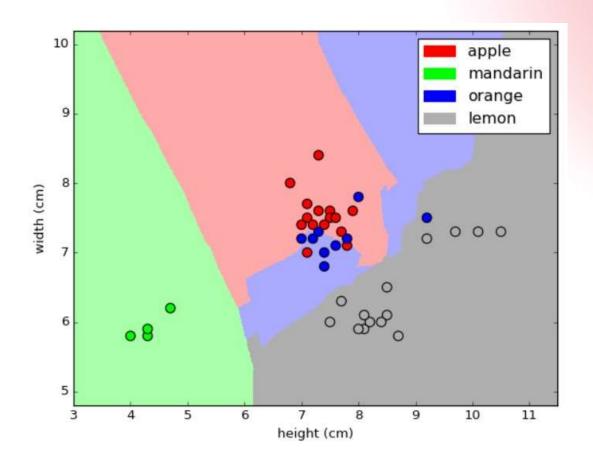


Algoritmo K-NN: K- Vecino más cercano

- 1. Una métrica de distancia (euclidea por ejemplo)
- 2. ¿Cuántos vecinos "más cercanos" hay que mirar? K=1, 2, 3...
- 3. Función de ponderación opcional en los puntos vecinos (opcional)
- 4. Método para agregar las clases de puntos vecinos (voto)

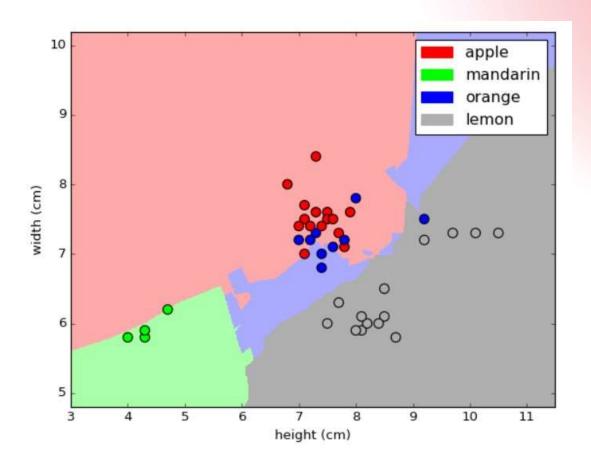


K-NN with K=5





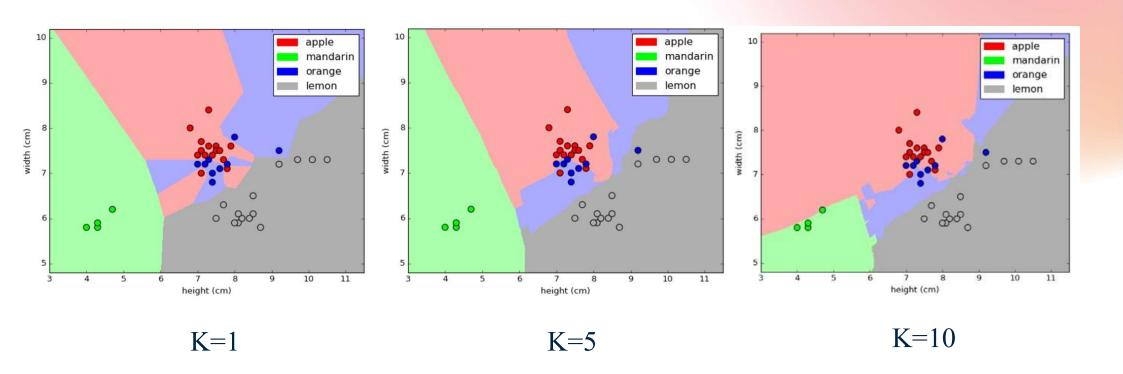
K-NN with K=10





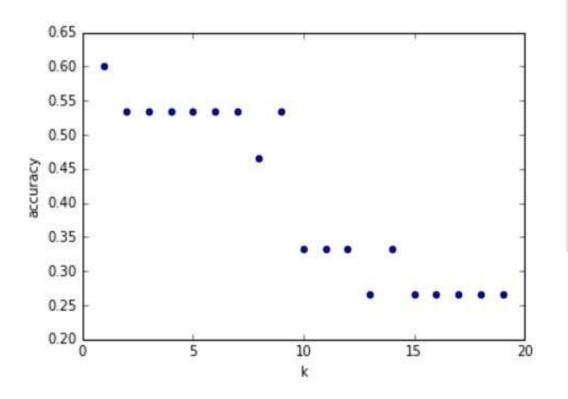
45

Papel de K en la sensibilidad de los límites ...





Precisión frente al valor k para k-NN

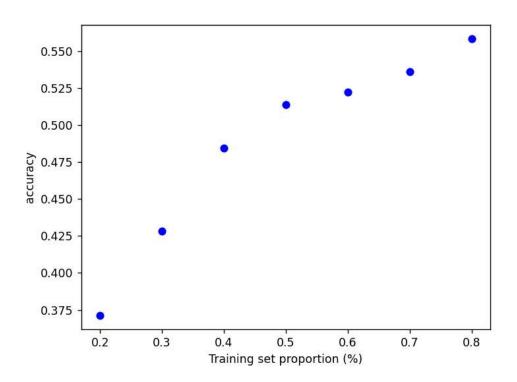


```
k_range = range(1,20)
scores = []
for k in k range:
    knn = KNeighborsClassifier(n_neighbors = k)
    knn.fit(X train, y train)
    scores.append(knn.score(X test, y test))
plt.figure()
plt.xlabel('k')
plt.ylabel('accuracy')
plt.scatter(k_range, scores)
plt.xticks([0,5,10,15,20]);
```

Conjunto de datos de frutas con división de prueba de entrenamiento del 75%/25%



Relación entre la sensibilidad y el entrenamiento en las pruebas



```
t = [0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2]
knn = KNeighborsClassifier(n_neighbors = 5)
plt.figure()
for s in t:
    scores = []
    for i in range(1,1000):
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1-s)
        knn.fit(X_train, y_train)
        scores.append(knn.score(X_test, y_test))
    plt.plot(s, np.mean(scores), 'bo')

plt.xlabel('Training set proportion (%)')
plt.ylabel('accuracy');
```



ANALISIS DE SERIES TEMPORALES

- Muchas señales se expresan en función del tiempo y requieren un tratamiento especial (económica, sociales, consumo...)
- Originalmente es un área que surge como parte de la estadística
- El análisis de series temporales se ha incorporado también al área de aprendizaje automático – y ha sufrido una explosión.
 - Aplicaciones a seguridad (patrones de conexión)
 - Aplicación a procesamiento de voz
 - Librerias Comerciales
 - Aprendizaje profundo supervisado y no supervisado
 - Detección de anomalias



ANALISIS DE SERIES TEMPORALES



■ ВООК

Machine Learning for Time Series Forecasting with Python

By Francesca Lazzeri



■ BOOL

Hands-on Time Series Analysis with Python: From Basics to Bleeding Edge Techniques

By B V Vishwas and ASHISH PATEL



ВООК

Practical Time Series Analysis

By Aileen Nielsen



BOO!

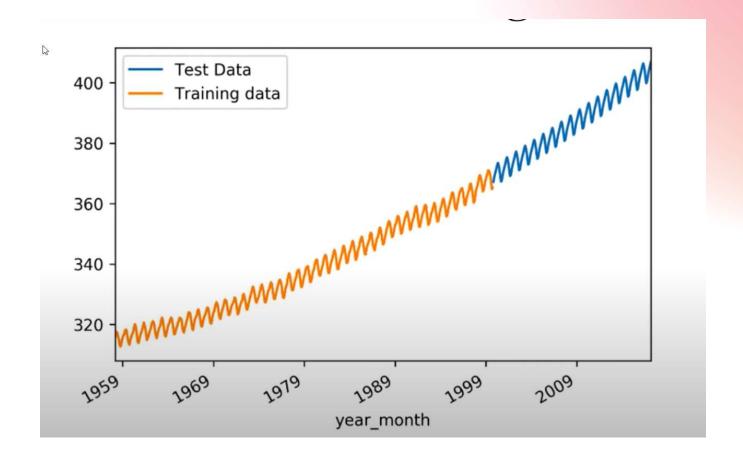
Machine Learning for Time-Series with Python

By Ben Auffarth



...

Definición





ANALISIS DE SERIES TEMPORALES

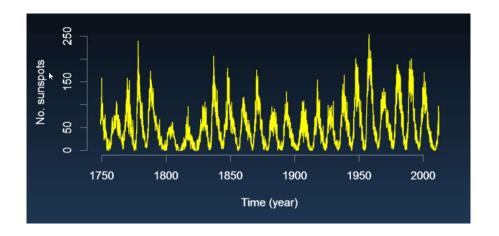
- La periodicidad con la que los datos se registran se conoce como la frecuencia del conjunto de datos
- Frecuencias alta –Periodos cortos y viceversa
- <u>Para poder analizar series temporales</u> de una manera significativa si tiene que cumplir que todos losperiodos de tiempo deben ser iguales; i.e. la frecuencia tiene que se constante.
- Puede haber valores faltantes





Ejemplos

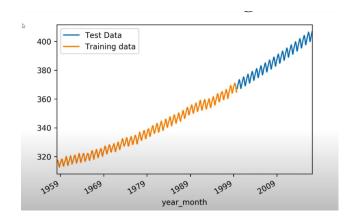
- Predicción de Tiempo: Temperatura, viendo, lluvia...
- Inversión: bolsa, crecimiento económico, PIB
- Medicina: vista como serie temporal, ritmos cardiacos, ECG, EEG, epidemiologia
- Astronomia: Numero de manchas solares, nivel de brillo de una estrella

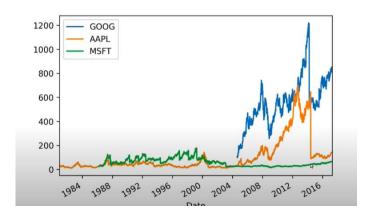


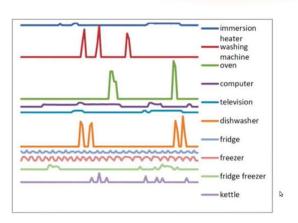


Aplicaciones

- Predicción (univariante)
- Predicción multivariante
- Clasificación de Series Temporales
- Clustering de Series Temporales









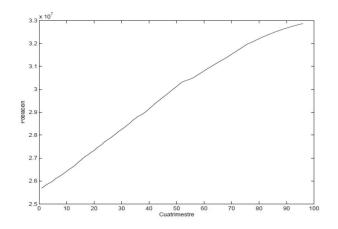
Estacionariedad

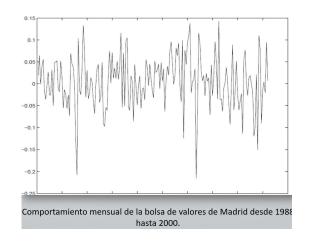
- Una serie es estacionaria cuando es estable a lo largo del tiempo, es decir, cuando la media y la varianza son constantes en el tiempo y, además, no presenta tendencia. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo.
- La series NO estacionarias son series en las cuales la tendencia y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que no oscilaría alrededor de un valor constante.
- EXISTEN TEST PARA PROBARLO COMO VEREMOS.
- La mayoría de métodos requieren estacionariedad > aun así se pueden aplicar.



Estacionariedad

- Serie no estacionaria. Tiene una tendencia positiva (crece)
- En general las series sociales, no son estacionarias muestran una tendencia (mas o menos constante)
- En este caso la tendencia es variable:i.e. crecimiento anual cambia con el tiempo





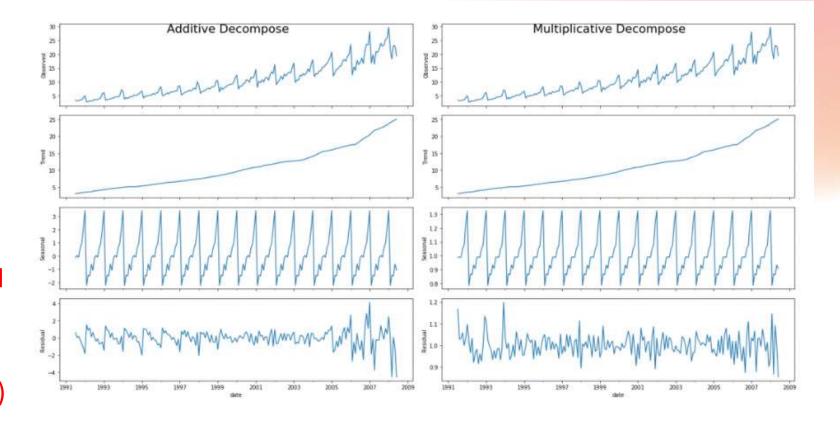


Elementos de una ST

Tendencia

Estacionalidad

• Ruido (Blanco)





Notación

- Variables: Letras Mayúsculas del Alfabeto (X,Y,Z)
- EL periodo de tiempo cubierto se denota con T
- Un único periodo de tiempo se denota con t
- X(t) valor de la serie en un intervalo especifico
- X(t-1) periodo anterior de tiempo al actual
- X(t+1) periodo posterio de tiempo al actual
- X(t+k)
- X(t-k)
- arisa

Peculiaridades

- Los intervalos entre observaciones deben ser idénticos.
 - Pueden faltar valores para un periodo lo que causa que el intervalo cambie
- Se puede ajustar la frecuencia en función de los valores que interesen.
 - Si tenemos valores diarios se puede pasar a frecuencia semanal
- Las series de tiempo requieren un orden
 - desde el punto de viste de ML no se puede separar train-test de cualquier manera (se hace aleatoriamente)
 - Se elije un punto de corte en el tiempo



Peculiaridades

$$(x_{t_1}, y_{t_1}), \ldots (x_{t_m}, y_{t_m})$$
 $t_1 < t_2 < \ldots < t_m$

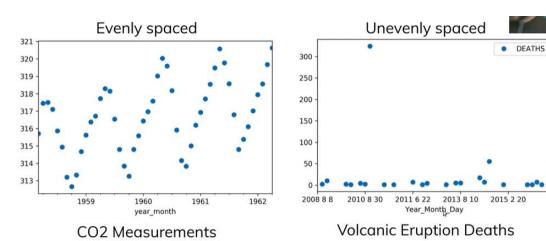
Equally spaced:

$$t_{i+1} - t_i = const$$

Unevenly spaced:

$$t_{i+1} - t_i$$
 varies

aka "Point processes" in statistics





60

Paquetes de Series Temporales

- 1. Nixtla is a set of state-of-the-art time series forecasting methods that are embedded in a set of libraries and models, most famous are: statsforecast., neuralforecast, TimeGPT
- 2. **DARTS**. Darts offers a consistent *fit*() and *predict*() interface across various forecasting models, from classical methods like ARIMA to deep learning models.
- Prophet. Prophet is specifically designed for users with limited statistical knowledge, allowing them to easily handle seasonal effects and missing data.
- 4. **SKTime**. SKtime provides a scikit-learn-compatible framework for time series analysis, enabling users to apply machine learning techniques seamlessly.











www.aiskills.eu