



UNIVERSIDAD DE CHILE

Minería de Datos

Welcome to the Machine Learning class

Valentin Barriere

Universidad de Chile – DCC

CC5205, Otoño 2024

Introducción

Definición

La minería de datos tiene como objetivo la **extracción** de conocimiento **a partir de grandes cantidades de datos**, mediante métodos automáticos o semiautomáticos.

Propone usar un conjunto de **algoritmos** [...] para construir modelos a partir de los datos, es decir, encontrar estructuras interesantes o patrones según criterios predefinidos, y **extraer un máximo de conocimientos** de ellos.¹

¹https://fr.wikipedia.org/wiki/Exploration_de_données

Datos

Diferentes tipos de datos:

- Datos estructurados:
 - Datos sociales: Edad, Salario, Color de piel, Lugar de residencia
 - Datos métricos: Likes de una publicación, Tiempo pasado en una página, Número de amigos en común
- Datos no estructurados:
 - Texto: Frase, Párrafo, Documento
 - Sonido: Canción, Discurso
 - Imagen: Foto, Vídeo

Diferentes tipos de Minería:

- Exploración de datos: Detectar valores simples, sesgos
- Tarea de clasificación/regresión: Alimentarse de datos para caracterizar nuevos datos **por clase o con un valor**, de manera supervisada
- Tarea de agrupamiento: Caracterizar datos por clase de manera no supervisada
- Reducción de dimensiones: Desarrollar estructuras comunes para representaciones comprimidas de datos

Outline : Aplicaciones

Aplicaciones

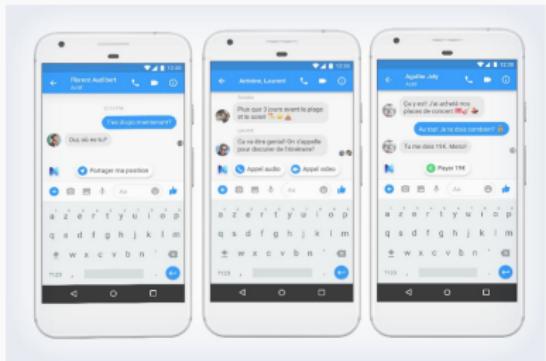
Significación de las termas

Prerrequisitos

TP Exploración de Datos:
MovieLens
Overview

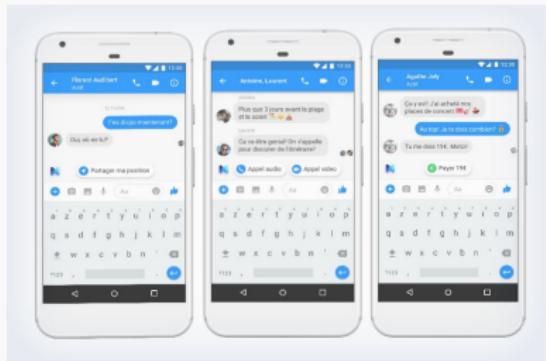
Aplicaciones (I/II)

- Detección de eventos en un texto



Aplicaciones (I/II)

- Detección de eventos en un texto

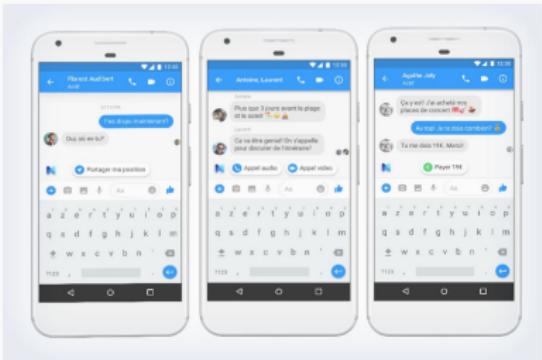


- Procesamiento automático de opiniones de usuarios



Aplicaciones (I/II)

- Detección de eventos en un texto



- Procesamiento automático de opiniones de usuarios



Les connaissez-vous ?



Marina Dunion
Digital Marketing @Air France
& Co-Founder @FlexiFly
● Teddy Viraye-Chevalier et 3 autres relations



Salvatore Anzalone
Post-Doc at ISIR, University Pierre et Marie Curie, Paris
● Thomas Janssoone et 2 autres relations

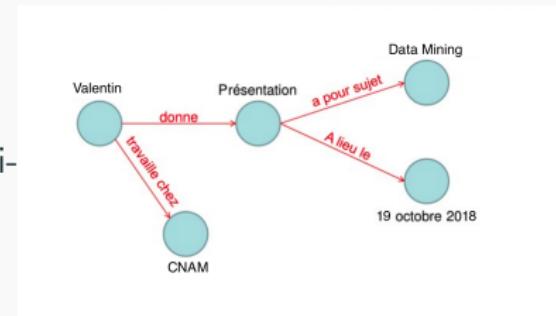


Halla Olafsdottir
Medical Solutions Project Manager | Chef de Projet
● Télécom ParisTech

- Propuesta de recomendaciones a un usuario

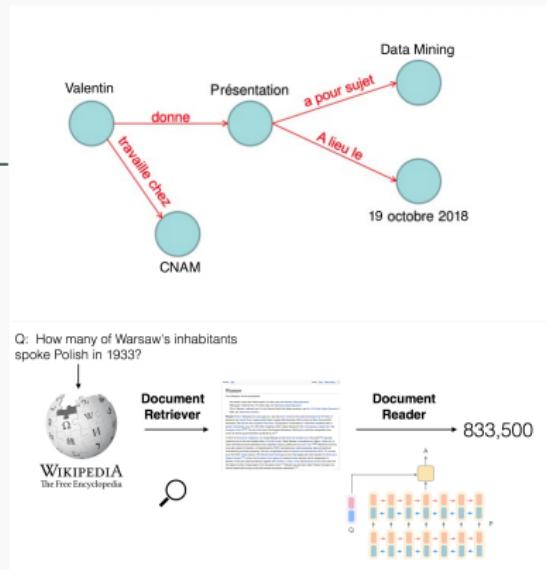
Aplicaciones (II/II)

- Detección de relaciones entre entidades en un texto



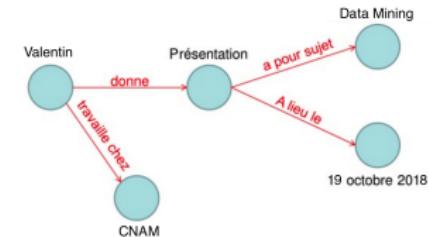
Aplicaciones (II/II)

- Detección de relaciones entre entidades en un texto
- Respuesta a una pregunta

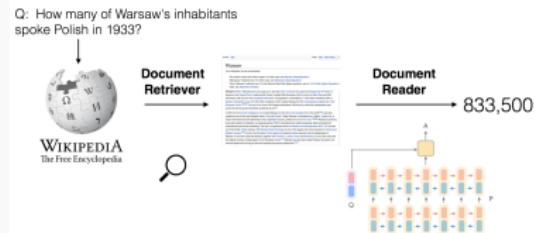


Aplicaciones (II/II)

- Detección de relaciones entre entidades en un texto



- Respuesta a una pregunta



- Módulo de IE para un agente conversacional



Outline : Significación de las termas

Aplicaciones

Significación de las termas

Prerrequisitos

TP Exploración de Datos:
MovieLens
Overview

AI vs. Data Science vs. Machine Learning

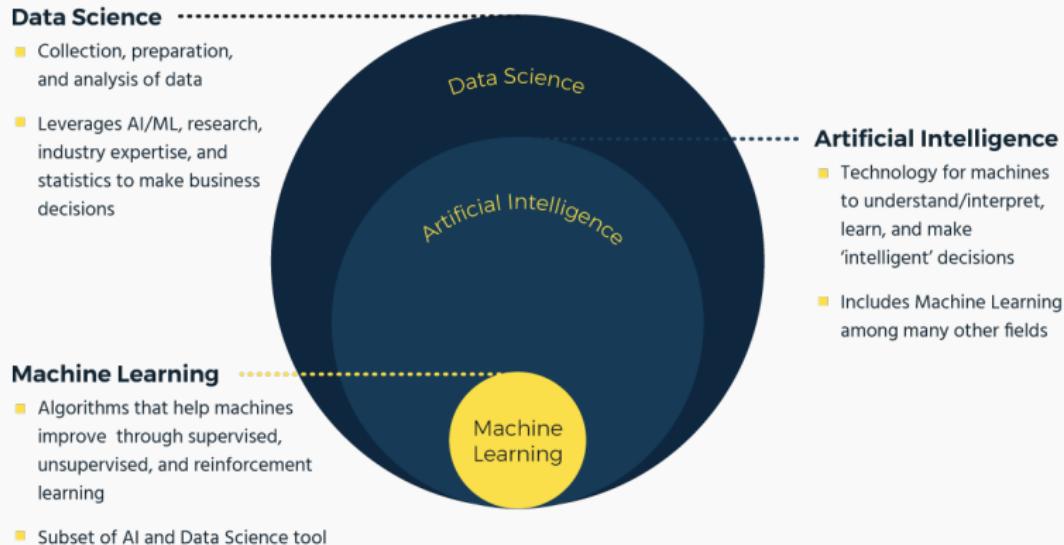


Figure 1: Diferencias entre campos

En resumen

Data Science se centra en el análisis de datos para extraer conocimiento, Machine Learning utiliza algoritmos para hacer predicciones y tomar decisiones basadas en datos, y Artificial Intelligence se refiere al desarrollo de sistemas que pueden realizar tareas inteligentes de manera autónoma.

Definición (sobre) simplista:

- Data mining genera entendimiento.
- Machine learning genera predicciones.
- Artificial intelligence genera acciones.

Ejemplo en plataforma de música

Data Scientist

Recopila y analiza datos de usuarios de plataformas de música para identificar patrones y preferencias musicales.

Machine Learner

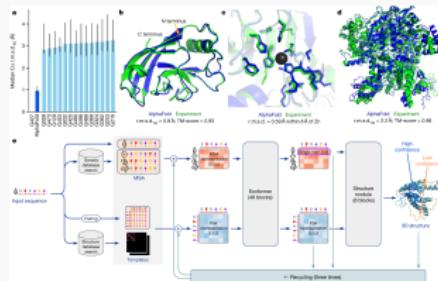
Desarrolla y optimiza un modelo de recomendación de música utilizando algoritmos de aprendizaje automático para predecir las preferencias de los usuarios.

Artificial Intelligence

Implementa un agente social que puede interagir con el usuario, para mejorar la personalización de las recomendaciones musicales y proporcionar una experiencia más precisa y contextualizada.

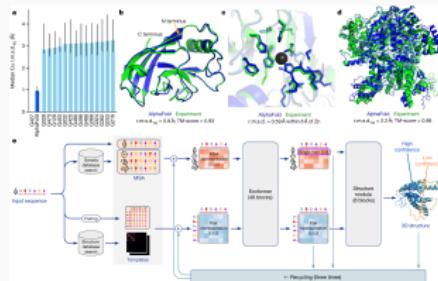
Significación del trabajo: Porque hacer eso?

- Avance científico



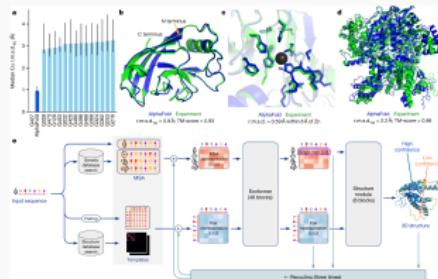
Significación del trabajo: Porque hacer eso?

- Avance científico
- Prevención y gestión de desastres naturales



Significación del trabajo: Porque hacer eso?

- Avance científico
- Prevención y gestión de desastres naturales
- Impacto en la salud pública



Open Chronic

Améliorer la prise en charge des malades chroniques

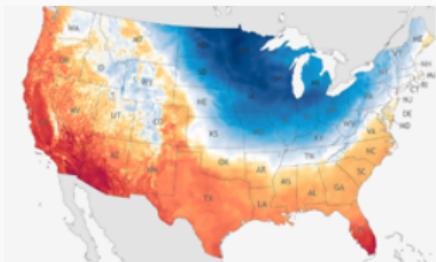
Santé Promotion 3

Ministère de la santé, Direction de la recherche, des études, de l'innovation et des statistiques

Paris Data science

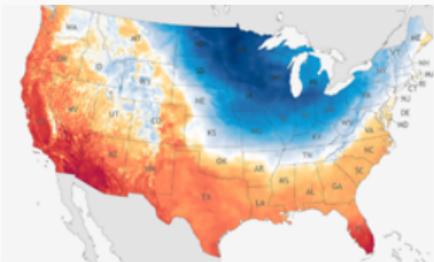
Significación del trabajo: Porque hacer eso?

- Sostenibilidad ambiental

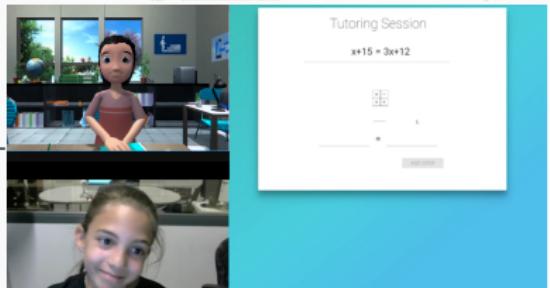


Significación del trabajo: Porque hacer eso?

- Sostenibilidad ambiental

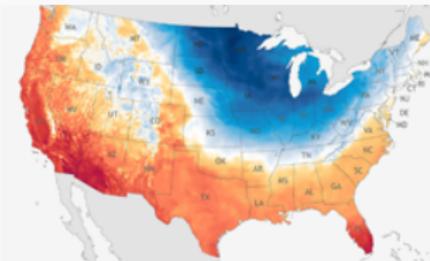


- Impulso a la educación y la investigación

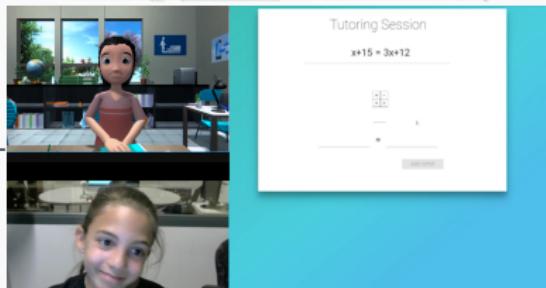


Significación del trabajo: Porque hacer eso?

- Sostenibilidad ambiental



- Impulso a la educación y la investigación



- Democracia participativa



Outline : Prerrequisitos

Aplicaciones

Significación de las termas

Prerrequisitos

TP Exploración de Datos:
MovieLens
Overview

El programa

Partes teóricas

Bases de estadística, álgebra lineal: Presentación general del aprendizaje estadístico, Bases matemáticas de los diferentes modelos, enfoque intuitivo

Partes prácticas

Bases de Python: Uso de herramientas de manipulación de datos, Uso de una biblioteca de DL, Uso de una biblioteca de ML, Análisis de sentimientos, Ranking sobre preferencias de vino, ...²

²Non contractual por este curso

Material

- Computadora
- Jupyter Notebook y Anaconda:
<https://www.anaconda.com/download/>
- Los notebooks y las cheatsheets disponibles online:

Python For Data Science Cheat Sheet
NumPy Basics

Learn Python for Data Science interactively at www.DataCamp.com

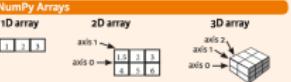
NumPy

The NumPy library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Use the following import convention:
`>>> import numpy as np`

NumPy Arrays

1D array 2D array 3D array



Creating Arrays

Python For Data Science Cheat Sheet
Matplotlib

Learn Python interactively at www.DataCamp.com

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

1. Prepare the Data Also see Lists & NumPy

1D Data

```
>>> import numpy as np  
>>> x = np.linspace(0, 10, 100)  
>>> y = np.cos(x)  
>>> z = np.sin(x)
```

2D Data or Images

```
>>> data = 2 * np.random.random((100, 100))  
>>> data -= data.mean() / 100  
>>> Y, X = np.mgrid[-3:3:100j, -3:3:100j]  
>>> Z = np.exp(-X**2 - Y**2)
```

Python For Data Science Cheat Sheet
Scikit-Learn

Learn Python for data science interactively at www.DataCamp.com

Scikit-learn

Scikit-learn is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.

A Basic Example

```
>>> from sklearn import neighbors, datasets, preprocessing  
>>> digits = datasets.load_digits()  
>>> X_train = digits.data[0:1797].ravel().reshape(-1, 64)  
>>> y_train = digits.target[0:1797]  
>>> X_test = digits.data[1797:1800].ravel().reshape(-1, 64)  
>>> y_test = digits.target[1797]  
>>> knn = neighbors.KNeighborsClassifier(n_neighbors=3)  
>>> knn.fit(X_train, y_train)  
>>> y_pred = knn.predict(X_test)  
>>> accuracy_score(y_test, y_pred)
```

Crea Sup

Lines
>>> t
>>> l
>>> r
>>> N
>>> E
>>> S
>>> K
>>> KNN
>>> k
>>> C
>>> U
>>> P
>>> R
>>> M
>>> Mod

Python For Data Science Cheat Sheet
Pandas Basics

Learn Python for Data Science interactively at www.DataCamp.com

Pandas

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.

Use the following import convention:
`>>> import pandas as pd`

Pandas Data Structures

Series

A one-dimensional labeled array capable of holding any data type



Writing Code And Text

Python For Data Science Cheat Sheet
Jupyter Notebook

Learn More Python for Data Science interactively at www.DataCamp.com

Saving/Loading Notebooks

Create new notebook



Open an existing notebook



Make a copy of the current notebook



Save current notebook and record checkpoint



Preview of the printed notebook



Close notebook & stop running any scripts



Python For Data Science Cheat Sheet
Keras

Learn Python for data science interactively at www.DataCamp.com

Keras

Keras is a powerful and easy-to-use deep learning library for Theano and TensorFlow that provides a high-level neural networks API to develop and evaluate deep learning models.

A Basic Example

```
>>> import numpy as np  
>>> from keras.models import Sequential  
>>> from keras.layers import Dense  
>>> data = np.random.random((1000,100))  
>>> labels = np.random.randint(0,2, size=(1000,1))  
>>> model = Sequential()  
>>> model.add(Dense(12,  
>>> activation='relu',  
>>> input_dim=100))  
>>> model.add(Dense(8, activation='sigmoid'))  
>>> model.compile(optimizer='adam',  
>>> loss='binary_crossentropy',  
>>> metrics=['accuracy'])  
>>> model.fit(data, labels, epochs=10, batch_size=32)
```

Mod Seq

>>> f
>>> m
>>> n
>>> Mult
>>> E
>>> F
>>> m
>>> Multi-C
>>> fci
>>> msi
>>> msi
>>> msi
>>> msi
>>> Regress
>>> msc
>>> msc
>>> Conv

Outline : TP Exploración de Datos: MovieLens

Aplicaciones

Significación de las termas

Prerrequisitos

TP Exploración de Datos:
MovieLens

Overview

Exploración de Datos: MovieLens

Estudio simple de un conjunto de datos de críticas de películas

- 3 millones de notas
- Descriptores sociales: edad, sexo, ...
- Primer enfoque básico de minería de datos

movielens

Introducción a pandas



- Biblioteca de Python para manipular bases de datos:
<https://pandas.pydata.org/>
- Permite realizar operaciones y visualizaciones
- Fácil de usar

Outline : Overview

Aplicaciones

Significación de las termas

Prerrequisitos

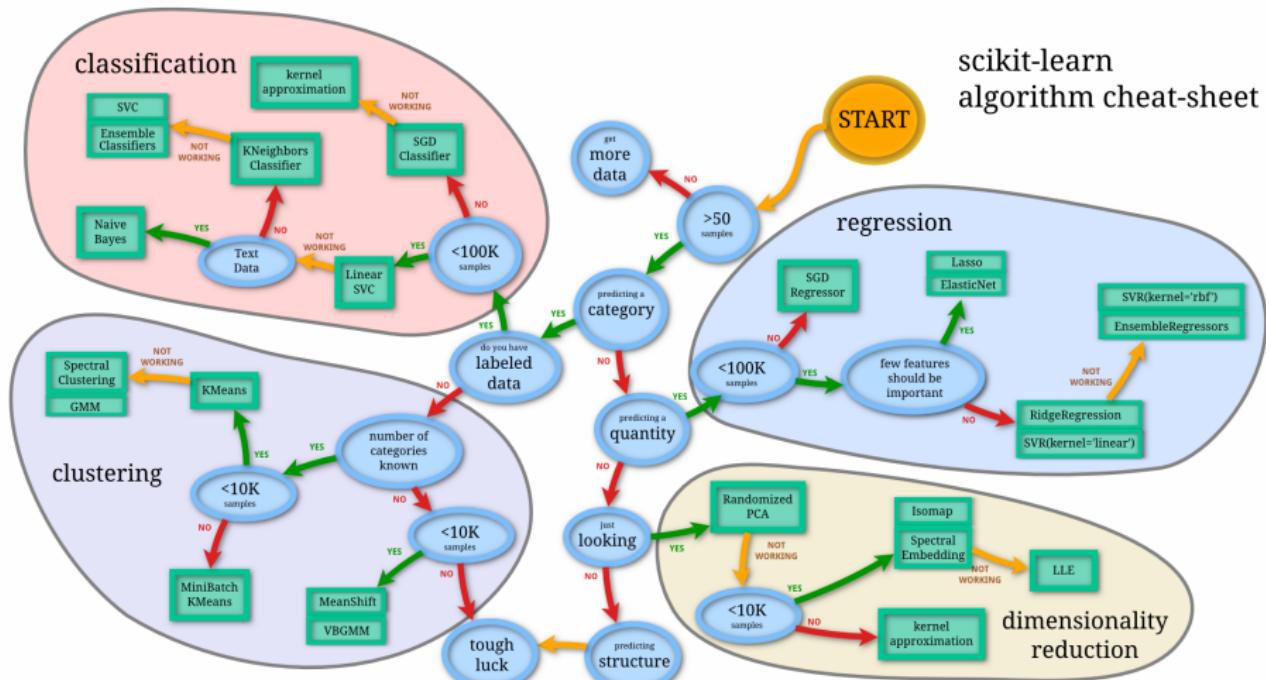
TP Exploración de Datos:
MovieLens
Overview

Los diferentes métodos

- Clasificacion: predecir una clase determinada
- Regresion: predecir un valor
- Clustering: agrupar elementos en cluster (clases no determinadas)
- Reducción de dimension: diminuir el espacio de representación de los datos
- Detección de anomalía

Los diferentes métodos

scikit-learn
algorithm cheat-sheet



Back

scikit
learn

Ejemplos: Clasificación

- Reconocimiento de emociones en el habla:
- Clasificación de especies de animales a partir de imágenes:
- Detección de objetos en imágenes médicas:

Ejemplos: Clasificación

- Reconocimiento de emociones en el habla: **la persona está enojada o feliz**
- Clasificación de especies de animales a partir de imágenes:
- Detección de objetos en imágenes médicas:

Ejemplos: Clasificación

- Reconocimiento de emociones en el habla: **la persona está enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas:

Ejemplos: Clasificación

- Reconocimiento de emociones en el habla: **la persona está enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas: **es un tumor?**

Ejemplos: Clasificación

- Reconocimiento de emociones en el habla: **la persona esta enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas: **es un tumor?**

Binary Classification



Multiclass Classification



Multilabel Classification



Ejemplos: Regresion

- Reconocimiento de emociones en el habla:
- Evaluación de daños a partir de imágenes despues un terremoto:
- Detección de severidad de Alzheimer en la voz:

Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intensidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto:
- Detección de severidad de Alzheimer en la voz:

Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intesidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz:

Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intesidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**

Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intensidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**

Age Prediction via Regression



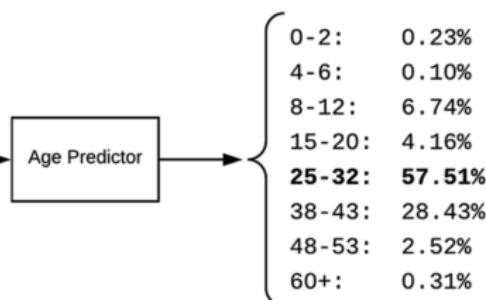
Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intensidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**

Age Prediction via Regression



Age Prediction via Classification



Ejemplos: Clustering

- Topic Mining en forums políticos:
- Clasificación de desinformación en redes sociales:
- Segmentación de clientes:

Ejemplos: Clustering

- Topic Mining en forums políticos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales:
- Segmentación de clientes:

Ejemplos: Clustering

- Topic Mining en forums políticos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **grupamos estos noticias que parecen raras**
- Segmentación de clientes:

Ejemplos: Clustering

- Topic Mining en forums políticos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **grupamos estos noticias que parecen raras**
- Segmentación de clientes: **la gente que le gustan las chelas**

Ejemplos: Clustering

- Topic Mining en forums políticos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **grupamos estos noticias que parecen raras**
- Segmentación de clientes: **la gente que le gustan las chelas**



Questions?