compensar | fundación universitaria

Machine Learning

Especialización en Big Data

Abril 20 de 2022

Temas

1 Formulación del problema

2 Modelamiento gráfico del problema

3 Conceptos formulación problema

4 Conceptos Machine Learning

4.1 Clasificación supervisada

4.2 Clasificación no supervisada

5 Protocolo experimental

compensar | fundación universitaria

compensar | fundación universitaria

GOAL

Conocer los problemas solucionables con Machine Learning

Conocer los algoritmos y conceptos básicos de ML

Ver los tipos de problemas de ML por medio de workflows en pandas

ALFONSO AYALA PALOMA

MAGISTER EN INGENIERÍA – AREA SISTEMAS Y COMPUTACIÓN

PERFIL PROFESIONAL

Magíster en Ingeniería en el área de Sistemas y computación. Especialista en Seguridad de la Información, Especialista en Docencia Universitaria, Profesional en Ingeniería de Sistemas. Catedrático en las Universidades Cooperativa y del Tolima. Amplia experiencia en proyectos de desarrollo de sistemas de información, herramientas de soporte a toma de decisiones, procesos de Transformación Digital y coaching de Innovación.

Historia 1

Como un usuario de data, deseo conocer el problema de ML, para poder iniciar mi formación en ML

Formulación del problema

Dado un conjunto de entradas X, se quiere predecir un conjunto de salidas **y**

AI

"Es una foto de un gato"
"Estoy 95% seguro que esta respuesta es correcta"

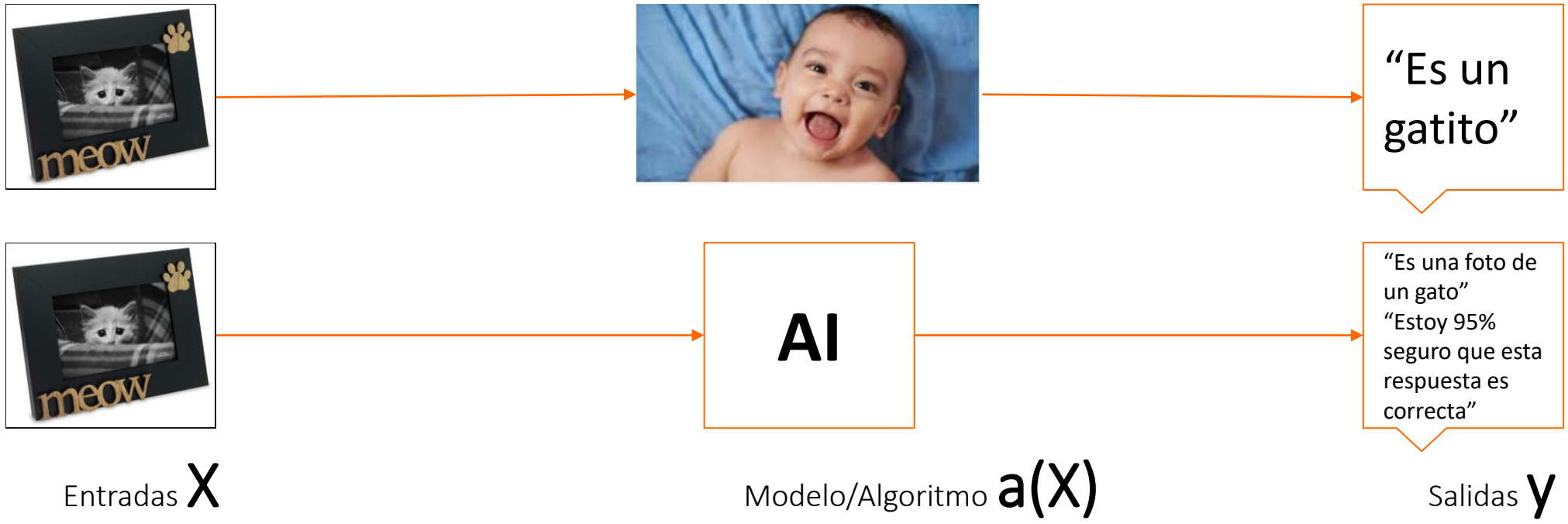
Entradas X

Modelo/Algoritmo a(X)

Salidas y

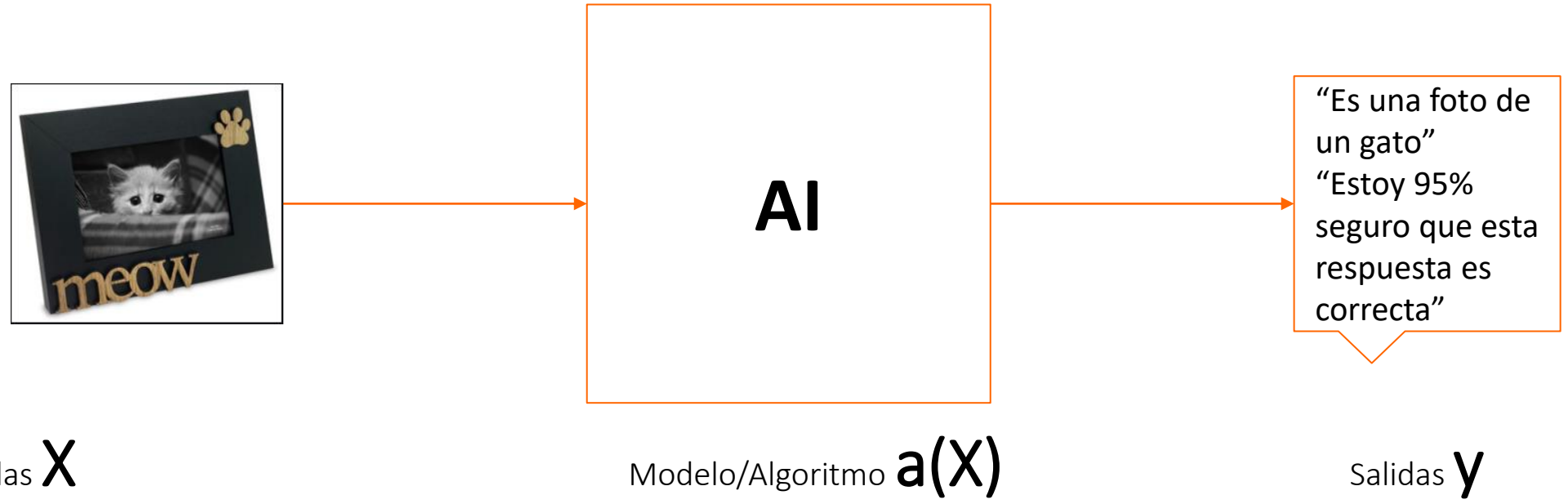
Formulación del problema

Se debe reducir el error entre la salida del algoritmo y la salida esperada (casos reales).



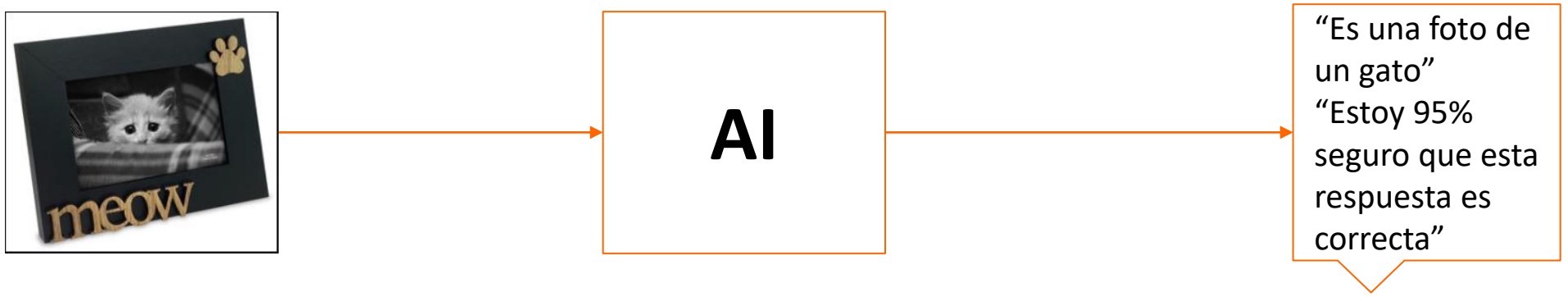
Formulación del problema

Se debe encontrar el algoritmo a(X) tal que sus salidas minimicen su diferencia con las salidas esperadas



“El problema”

Entrenar el algoritmo a(X) tal que sus salidas minimicen su diferencia con las salidas esperadas, para un conjunto de entradas X



Entradas X

Modelo/Algoritmo a(X)

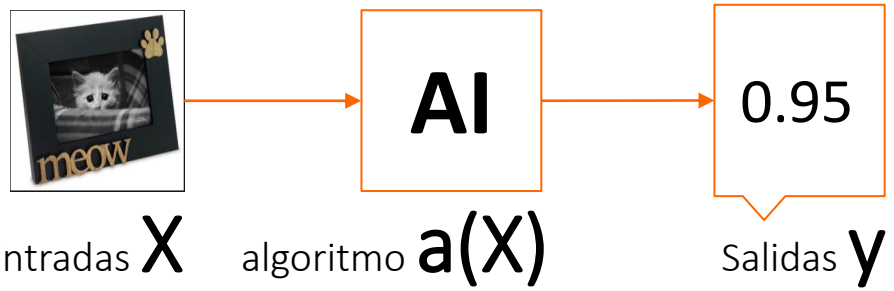
Salidas y

Historia 2

Como un usuario de data, deseo conocer los tipos de problemas solubles con ML, para poder empezar a encontrarlos y darles solución.

Problema de regresión

Predecir una cantidad (Variable numérica)



Ejemplos:

Dada la velocidad del viento, presión atmosférica, predecir la temperatura.

Predecir el precio de una casa basándose en área, estrato, cercanía de escuelas y otras características

Predecir el ingreso dadas las ventas anteriores.

También se usa en “forecasting” financiero, análisis de tendencias, marketing, predicción de series de tiempo.

Problema de regresión

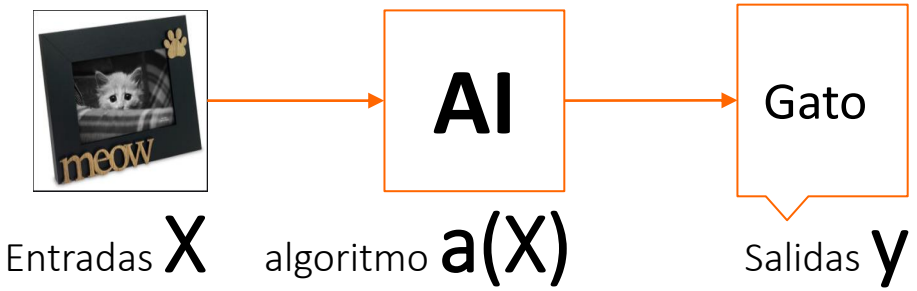
Tenemos más de 50 ejemplos.
Queremos encontrar un valor numérico.
Tenemos data ejemplo (con el valor buscado)

Dado: X -> y (y en los números reales)

Encontrar a(X) tal que a(X) = y

Problema de clasificación con labeled data

Predecir un label de clase (Variable texto)



Ejemplos:

Determinar si un correo es spam.

Determinar si una persona sufre de una enfermedad.

Problema de clasificación

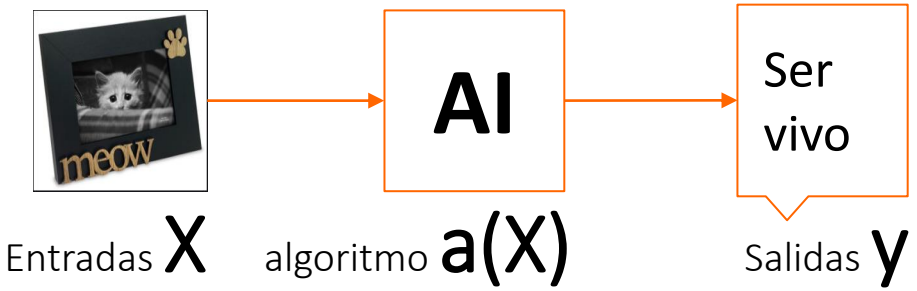
Tenemos más de 50 ejemplos
Queremos encontrar una categoría/label (variable texto).
Tenemos data ejemplo (marcada con la categoría buscada –labeled--)

Dado X -> y (y es una categoría)

Encontrar a(X) tal que a(X) = y

Problema de clasificación sin labeled data (clustering)

Predecir a que grupo pertenece (Variable texto)



Ejemplos:

Encontrar segmentos de clientes.

Encontrar patrones de consumo similar.

Clasificar personas por el uso de correo y determinar que tipo de marketing utilizar

Tomado de:

Problema de clustering

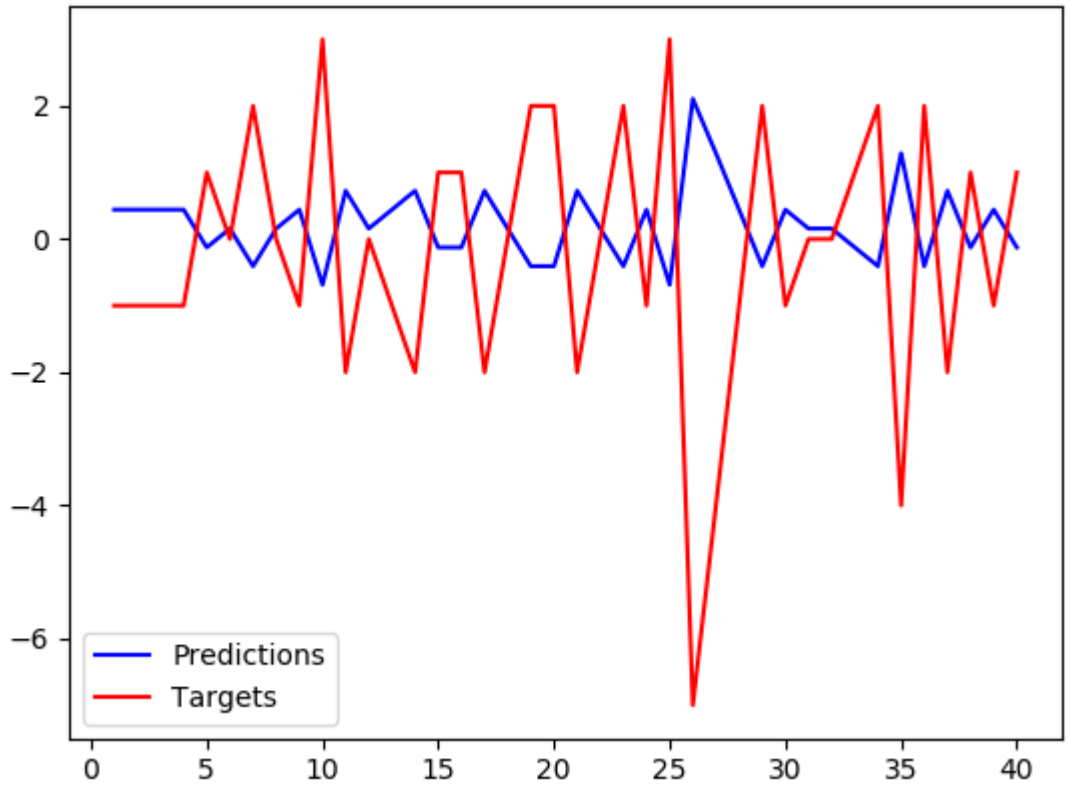
Tenemos más de 50 muestras
Queremos encontrar a que clase/categoría pertenece una muestra.
NO Tenemos data ejemplo (NO marcada con la categoría –non labeled---)

Dado X

Encontrar a(X) tal que a(X)=y (Donde y es una categoría)

Modelamiento gráfico del problema

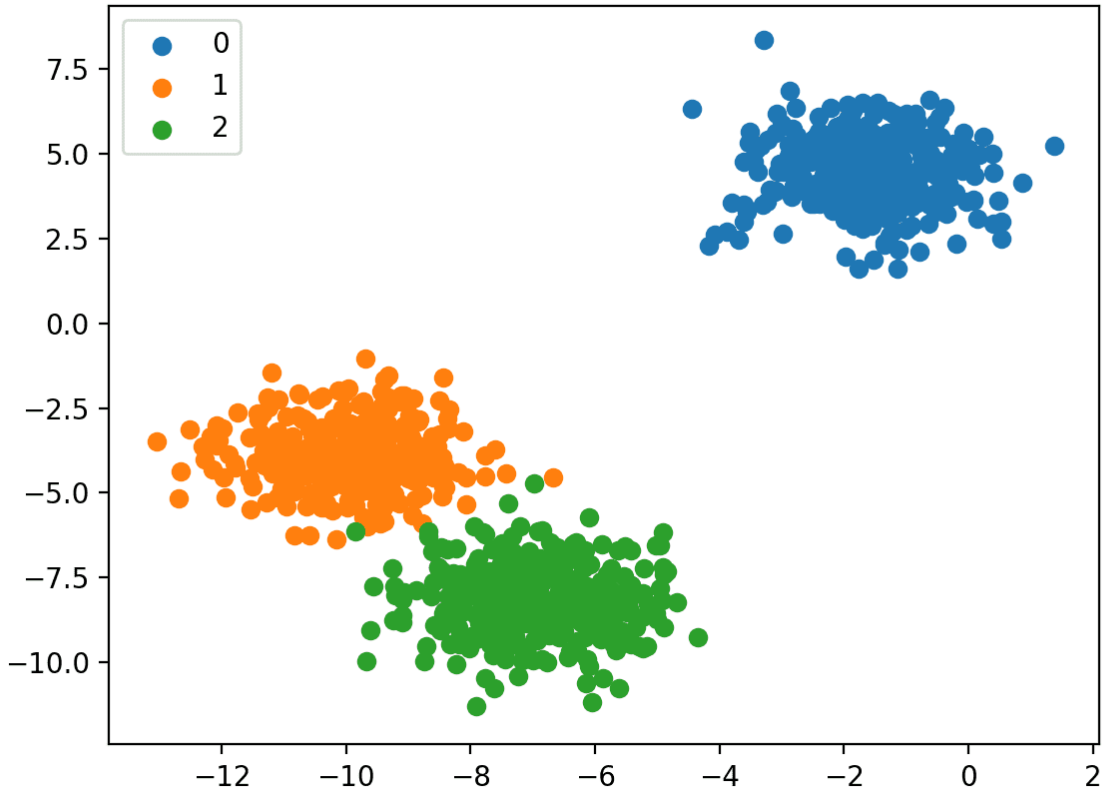
Regresión



Tomado de: https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

Modelamiento gráfico del problema

Clasificación de labeled data

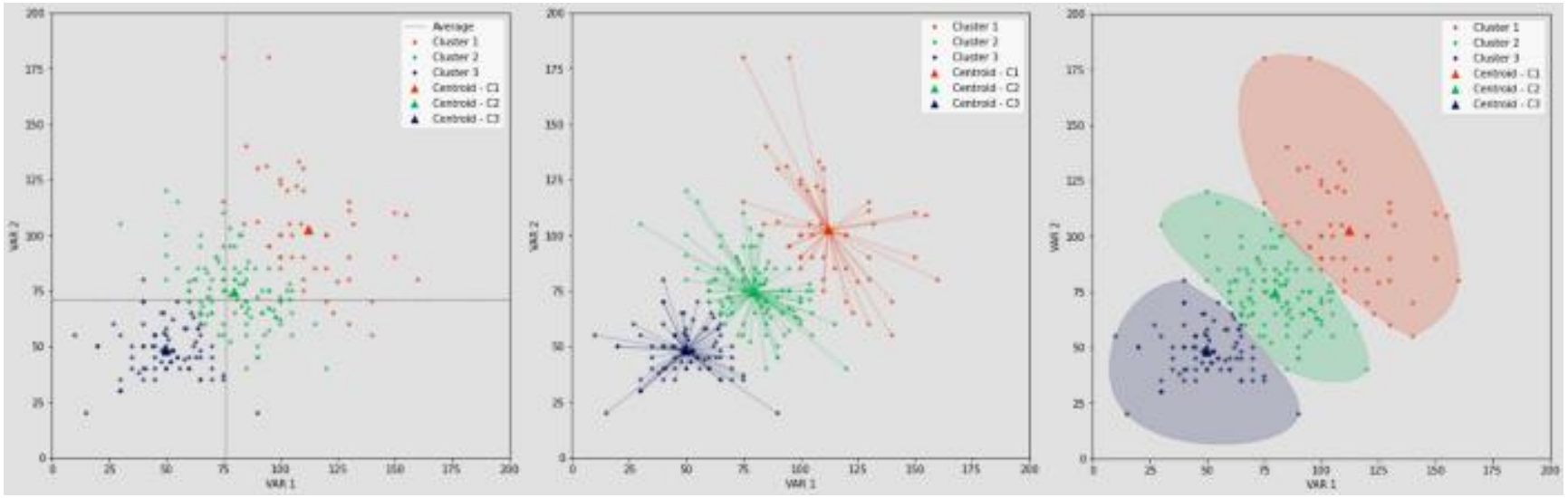


Tomado de: https://machinelearningmastery.com/types-of-classification-in-machine-learning/

Modelamiento gráfico del problema

compensar | fundación universitaria

Clasificación de unlabeled data



Tomado de: https://towardsdatascience.com/visualizing-clusters-with-pythons-matplolib-35ae03d87489

Conceptos de formulación del problema

Problema

Variable independiente: X

Variable dependiente: y

Algoritmo: a(X)

Muestras: Conjunto de ejemplos

Solución Lineal

Una solución LINEAL, asume que la F se ajusta a una forma conocida, entonces el proceso consiste en:

1. Escoger la forma de la función
2. Aprender los coeficientes de la función.

Ejemplos: Regresión Lineal y Regresión Logística.

Los algoritmos que no asumen formas determinadas, son más flexibles y se denominan NO LINEALES.

1. Estudiar el problema

Problema: Se desea predecir el salario en un cargo específico dados los años de experiencia en el mismo. Se cuenta con un Dataset: (experiencia, salario).

Análisis: se tiene una variable independiente y una dependiente. Se quiere predecir una variable numérica, entonces es un problema de regresión.

1. Estudiar el problema

Importar/obtener el dataset.

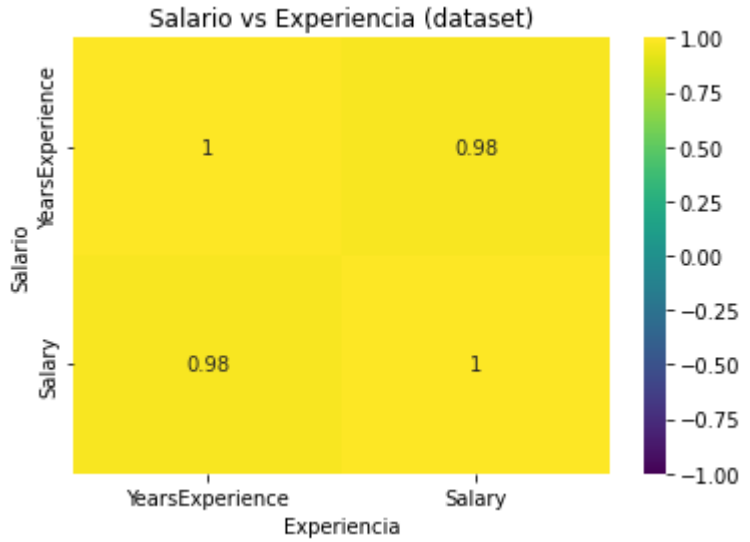
```
# 1. Analisis del problema
# Tenemos dos variables: AñosdeExperiencia = INDependiente, Salario = Dependiente (La que se va a predecir).
# Nuestra variable DEPendiente (Salario) es numérica.
# Entonces Este es un problema de REGRESION
# Usemos un modelo de REGRESION LINEAL
```

```
# 2. Importar el dataset
print("2. Importando Dataset.")
url="https://raw.githubusercontent.com/alfonsoayalapaloma/datasets/a4327dbd4334c0150af732995a470a32fc9e05c6/Salary_Data.csv"
df = pd.read_csv(url)
```

1. Estudiar el problema

Analizar el dataset: Dos variables numéricas.

```
# 3. Data Analysis
print("3. Revision Data.")
# Ver las primeras filas
print("Primeras filas:")
print(df.head())
print("Informacion del DATASET:")
# Ver informacion del dataset
print(df.info())
print(df.describe())
print("Tamaño del DATASET:")
print(df.shape)

print("Grafica del dataset")
plt.scatter( df["YearsExperience"] , df["Salary"], color = 'red')
plt.title('Salario vs Experiencia (dataset)')
plt.xlabel('Experiencia')
plt.ylabel('Salario')
plt.show()
```



1. Estudiar el problema

Feature engineering:

X=[experiencia] años

Y=[salario] USD

```
# 4. Feature Engineering: Definicion de variables independientes(X) y dependiente (y)
print("4. Definiendo variables X y Y.")
# Tomar todas las filas y todas las columnas menos la ultima
X = df.iloc[:, :-1].values
print("Tamaño de X")
print(X.shape)
# Tomar todas las filas de la ultima columna.
y = df.iloc[:, -1].values
print("Tamaño de y")
print(y.shape)
```

2. Fase de entrenamiento

Modelo: lineal

Partir dataset en Train/test

```
# 5. Partir el dataset en TRAINING y TEST
# se tomará 20% para pruebas.
print("5. Partiendo el dataset en entrenamiento y pruebas.")
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
print("Set de entrenamiento")
print(X_train.shape)
print("y entrenamiento")
print(y_train.shape)
print("Set de pruebas")
print(X_test.shape)
print("y pruebas")
print(y_test.shape)
```

2. Fase de entrenamiento

Entrenar.

```
# 6. Entrenamiento
print("6. Entrenando el modelo.")
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
# Print out the statistics
#regressor.summary()
print("Coeficientes: \n", regressor.coef_)
```

3. Inspeccionar la solución

Predicciones de prueba

```
# 7. Hacer predicciones de prueba
print("7. Predicciones de prueba.")
y_pred = regressor.predict(X_test)
dataset = pd.DataFrame(data=X_test, columns=['YearsExperience'])
dataset["PREDICCION_Salary"]=list(y_pred)
print(dataset)
```

3. Inspeccionar la solución

Validar modelo

```
# 8. Calcular el score del modelo R-squared
from sklearn.metrics import mean_squared_error, r2_score
print("8. Calificacion del modelo")
print("R-squared score:")
model_r2_score = regressor.score(X_test,y_test)
print("Coeficiente de determinacion (1=Prediccion Perfecta):")
print( r2_score(y_test, y_pred)*100 ,"%" )
print("Mean Squared Error : %f" % mean_squared_error(y_test, y_pred))
```

```
8. Calificacion del modelo
R-squared score:
Coeficiente de determinacion (1=Prediccion Perfecta):
98.8169515729126 %
Mean Squared Error : 12823412.298127
```

3. Inspeccionar la solución

Visualizar resultado

```
# 9. Visualizar el resultado de la prueba
print("9. Visualizacion del resultado de la prueba")
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.scatter(X_test, y_pred, color = 'green')
plt.title('Salario vs Experiencia (Entrenamiento)')
plt.xlabel('Experiencia')
plt.ylabel('Salario')
plt.show()
```

3. Inspeccionar la solución

Conclusiones

```
#10 Conclusiones
print("10. Interpretacion")
print(" Con un " + str(model_r2_score*100) +" de probabilidad, nuestro modelo puede predecir el salario dado los años de experiencia.")
```

4. Entender mejor el problema

Existe una relación entre las variables.

Se alcanza la solución del problema por lo que no se continúa con el paso 5. Volver a estudiar el problema.

Sesgo vs. Varianza

Sesgo (BIAS): que tan lejos está la respuesta de la esperada. Se debe a las suposiciones para simplificar el problema

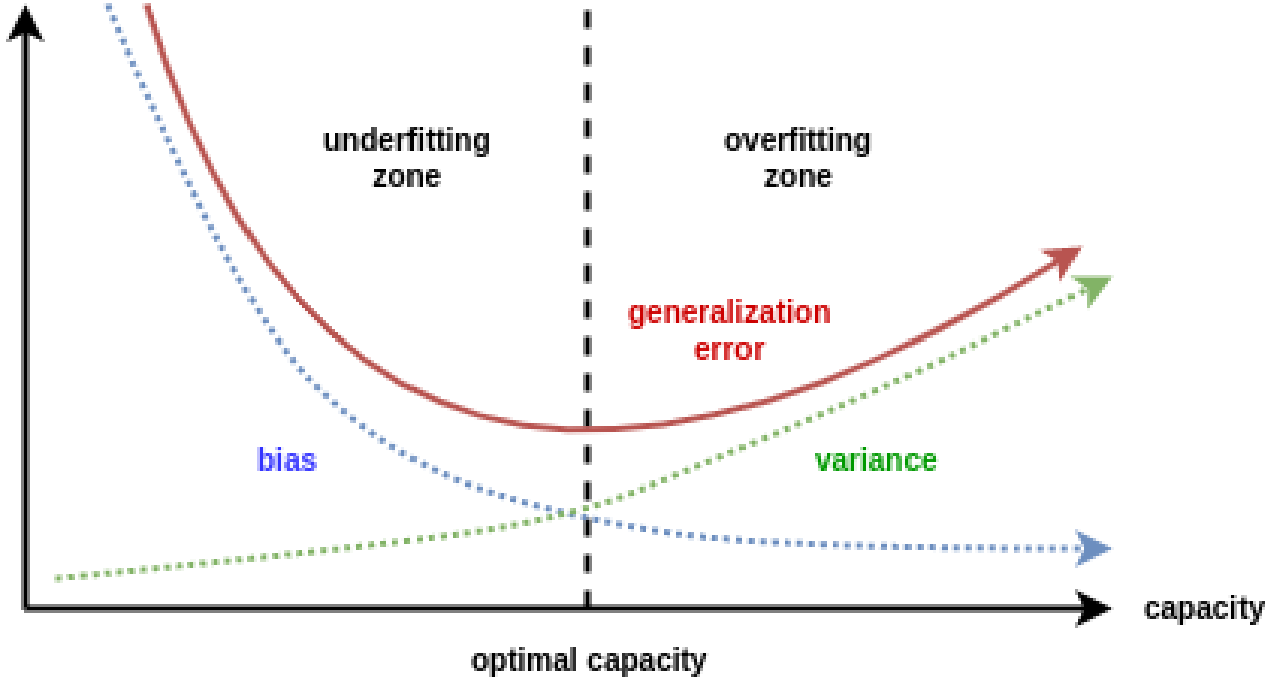
Ejemplo: Decision Trees=Bajo Sesgo, Regresion Lineal=Alto sesgo

Varianza: Es la medida de la dispersión de la data. Es la medida de cambio del estimado de la función F si se usa otra data de entrenamiento.

Ejemplo: K-nearest neighbor=Alta varianza, Linear Discriminant Analysis=Baja varianza

Sesgo vs. Varianza

META: Bajo Sesgo y baja Varianza, pero al subir una, baja la otra. (trade-off)



Sesgo vs. Varianza

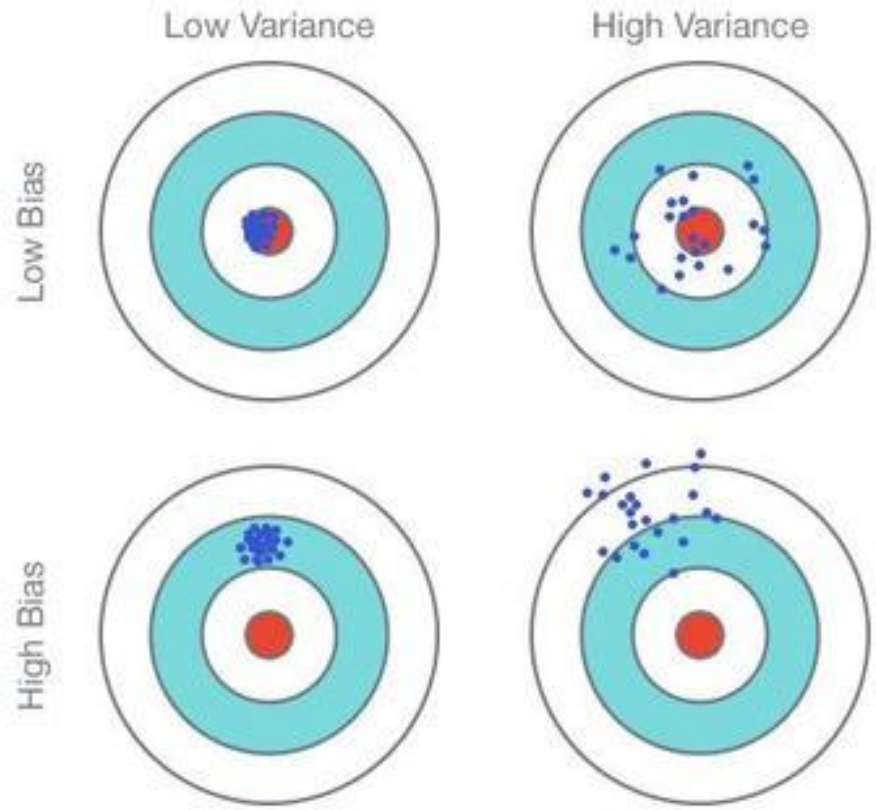


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

Tomado de: https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83

Data

- Data: hechos puros y simples sin estructura u organización.
- Information: data estructurada, que le da significado y contexto.
- Knowledge: Habilidad de usar la información para alcanzar objetivos estratégicos
- Wisdom: la capacidad de escoger los objetivos consistentes con los valores propios dentro de un contexto social mayor.

Tomado de: https://ecolabsblog.com/2010/11/29/data-information-knowledge-and-wisdom/

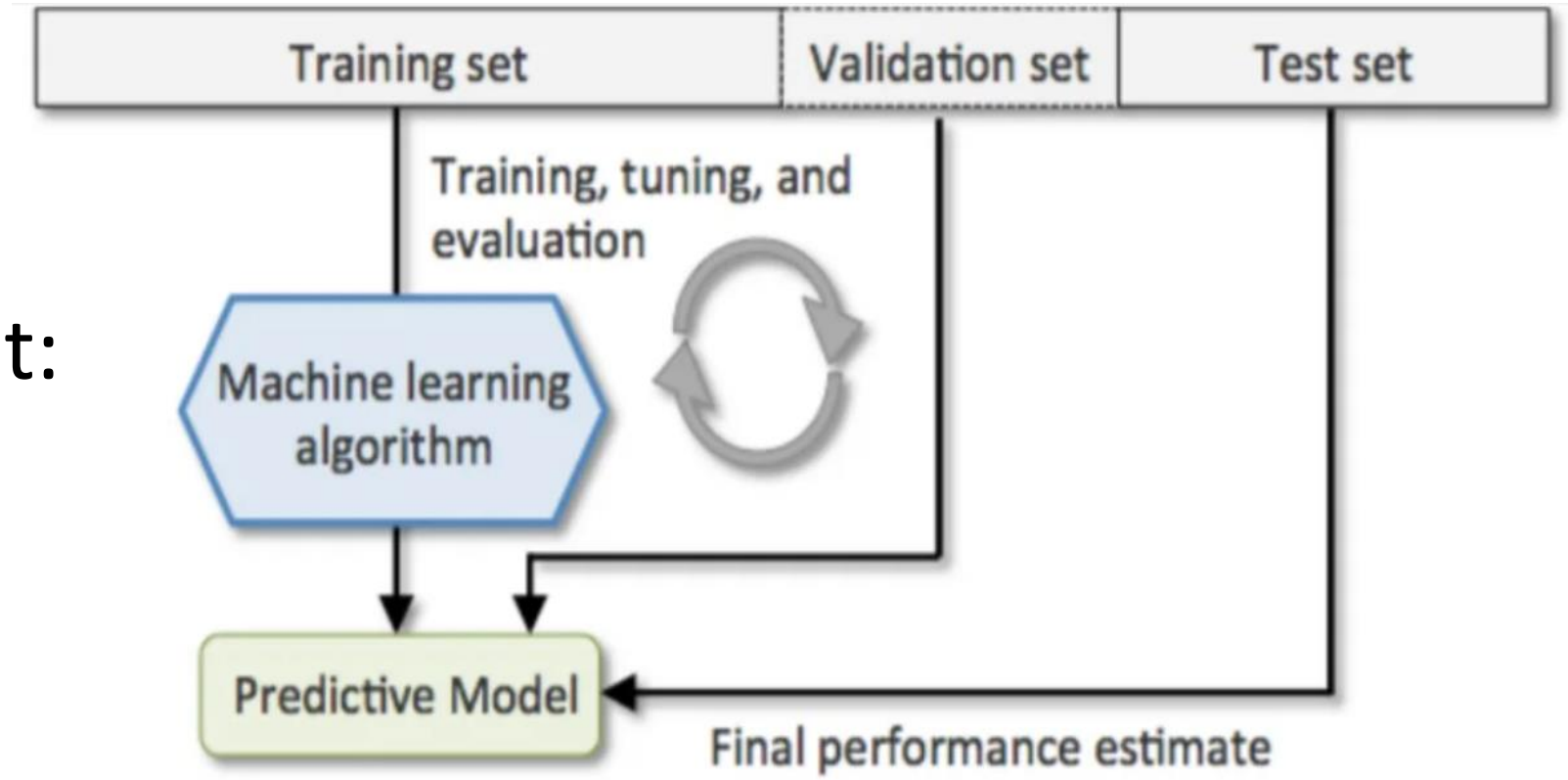
Pruebas

Dataset:

Training Dataset:

Validation Dataset:

Test Dataset:

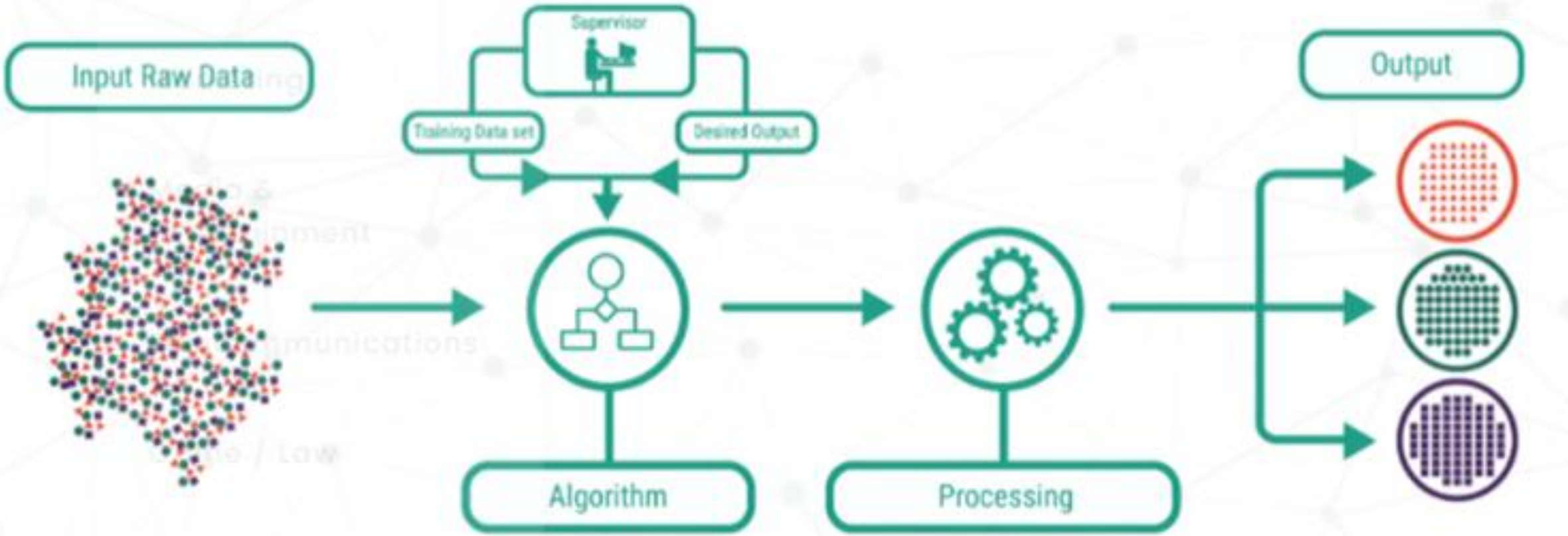


Tomado de: https://medium.com/@danieldesposito/difference-between-data-information-and-knowledge-2aac18419f58

Clasificación supervisada

compensar | fundación universitaria

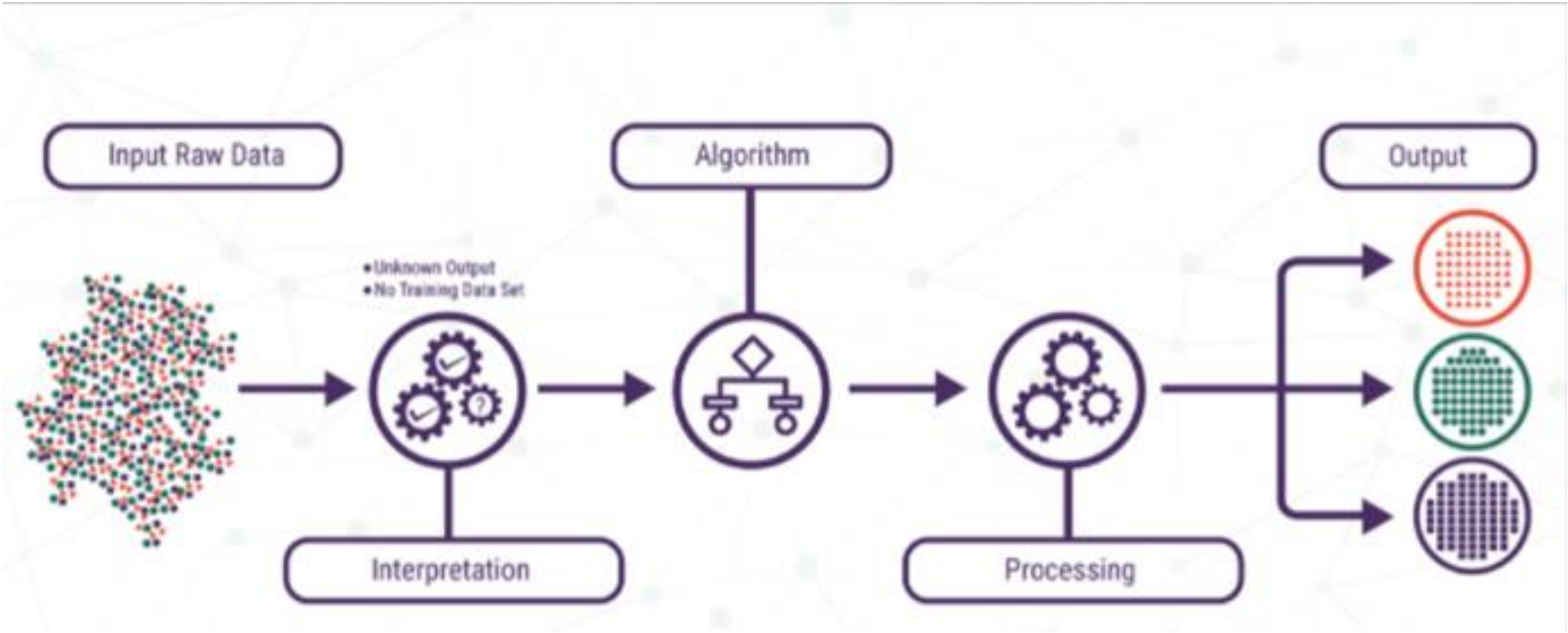
Labeled data creada por un supervisor



Clasificación no supervisada

compensar | fundación universitaria

Non labeled data / No supervisor



Errores

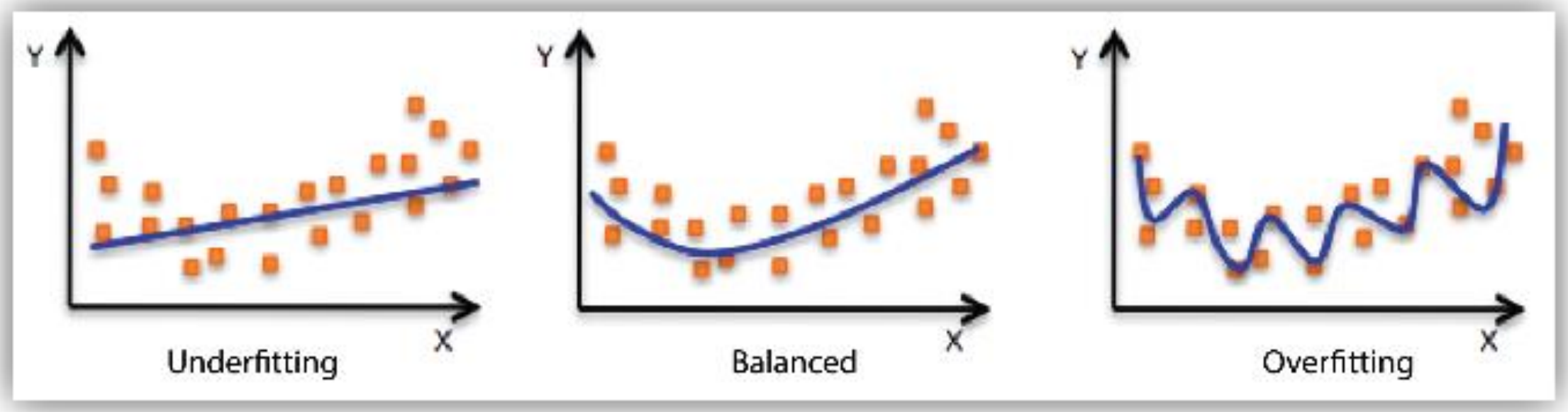
Error: Medida natural del rendimiento de un clasificador

Tasa de error: proporción del número de errores sobre el número de instancias

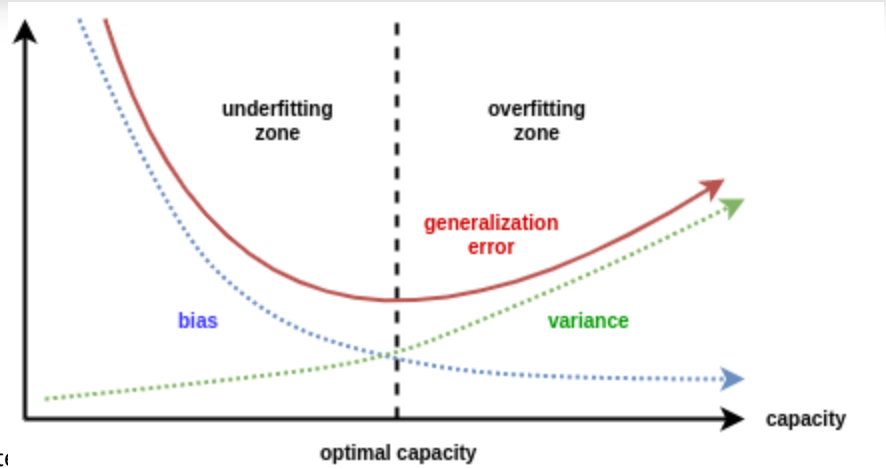
Tomado de: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Protocolo experimental

Underfitting / Balanced / Overfitting



Simple



Complejo

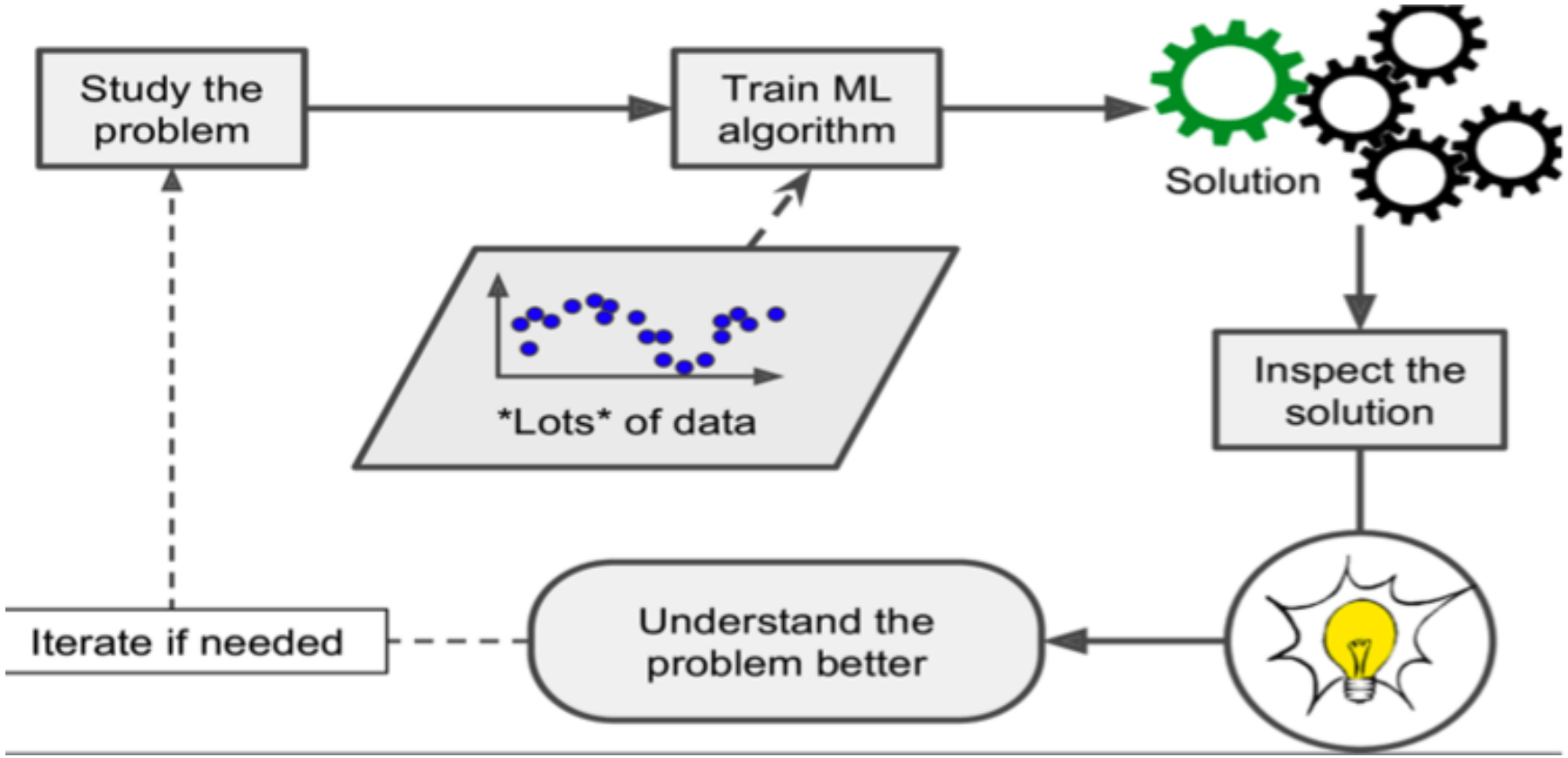
Tomado de: https://docs.aws.amazon.com/machine-learning/lat

Protocolo experimental

1. Estudiar el problema
    1. Análisis
    2. Importar dataset
    3. Data analysis
    4. Feature engineering
2. Fase de Entrenamiento
    1. Seleccionar modelo.
    2. Dividir los datos originales en train y test, de forma aleatoria, generalmente 2/3 para train y 1/3 para test.
    3. Entrenar
3. Inspeccionar la solución:
    1. Hacer predicciones de prueba
    2. Validación del modelo
    3. Visualizar el resultado
    4. Conclusiones
4. Entender mejor el problema
5. Si no se cumplen los valores esperados, regresar a 1.

Tomado de: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Protocolo de experimento



Formato de informe

Abstract

Resumen consiso del documento.

Introducción

Resumen del propósito del reporte y resumen de la data/tema. Incluir contexto. Resumir las preguntas de análisis, conclusiones y la estructura del documento.

Cuerpo – Cuatro secciones

Sección Data – Incluya descripciones escritas de la data con hojas de calculo relevantes.

Sección Métodos – Explica como se recogió y analizó la data.

Sección Análisis - Explica el análisis realizado. Presenta los modelos utilizados. Incluir gráficas.

Resultados - Describir los resultados del análisis..

Conclusiones

Respuestas a las preguntas planteadas en la introducción. Presentación de los resultados más relevantes. Recomendaciones.

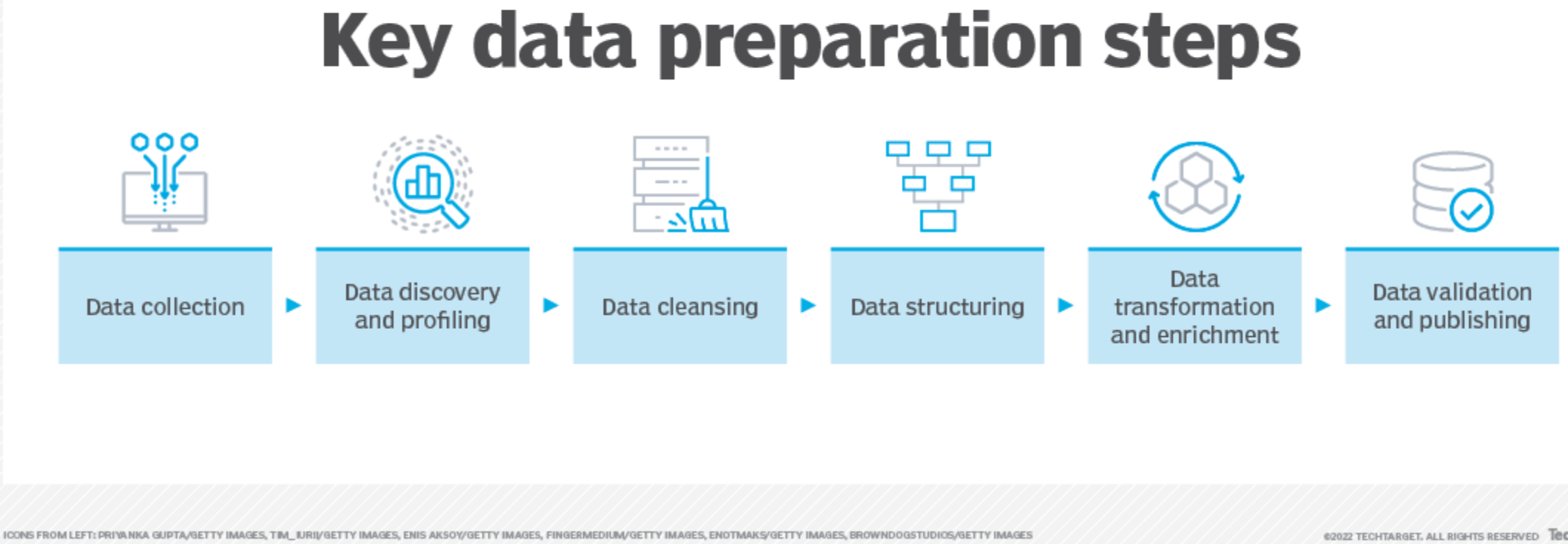
Anexos

Detalles del proceso de los datos, data secundaria, incluyendo referencias

Construcción de conjunto de datos y muestreo

compensar | fundación universitaria

Construcción Dataset

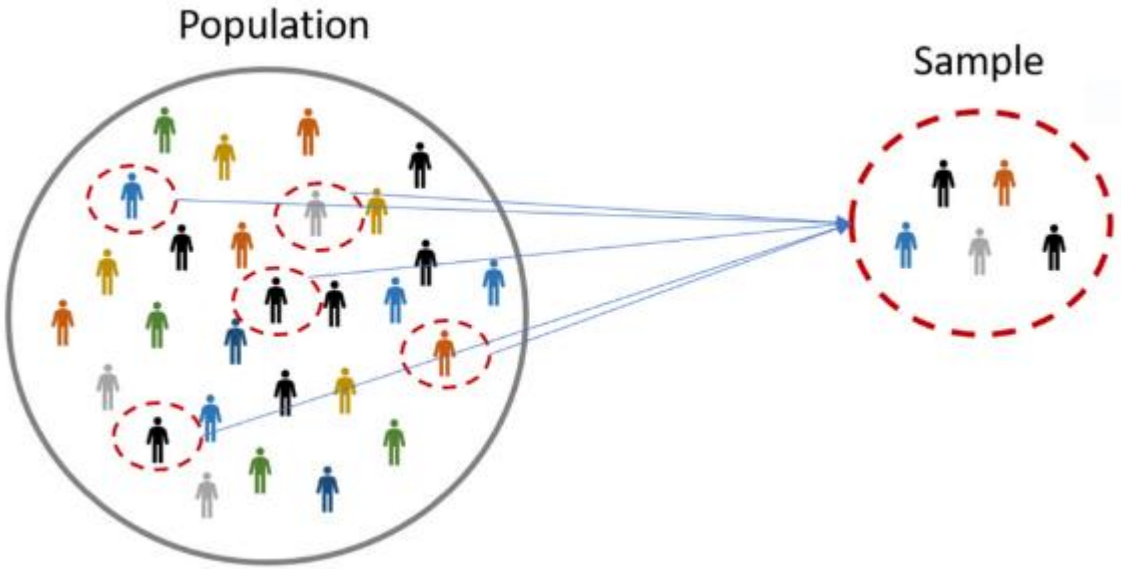


Tomado de: https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation

Data Review

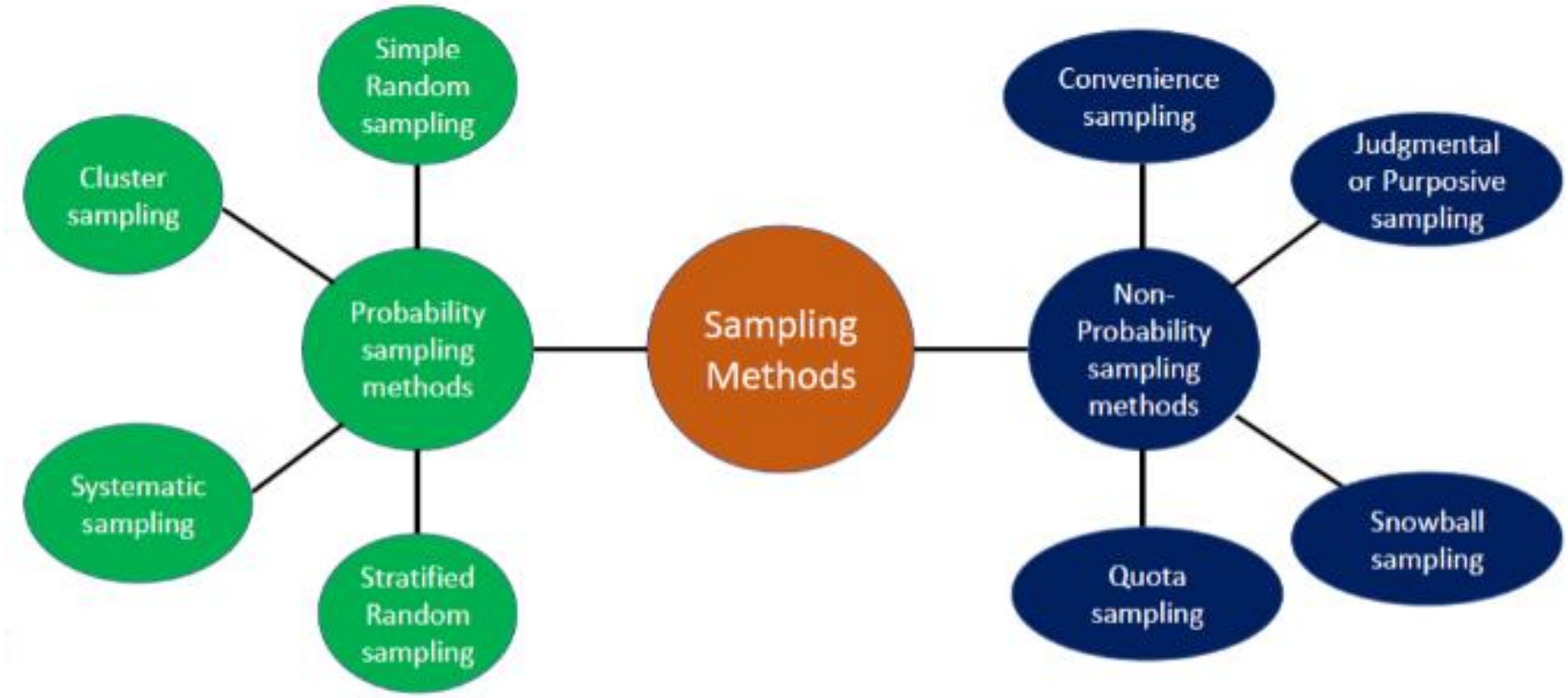
```
# Data Review
# Se va a analizar:

# Variable dependiente (y)
# tipos de variables (categorical y numerical)
# data faltante
# Variables numericas:
#   Discretas
#   Continuas
#   Distribuciones
#   Transformaciones
# Variables categoricas
#   Cardinalidad
#   Labels "raros"
#   Mappings especiales
```

Metodología de muestreo en ML



Tomado de: https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Métodos de muestreo en ML



Tomado de: https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Validación cruzada (Cross-validation)

Validación: decidir si nuestro modelo describe la data. Para ello hacemos pruebas.

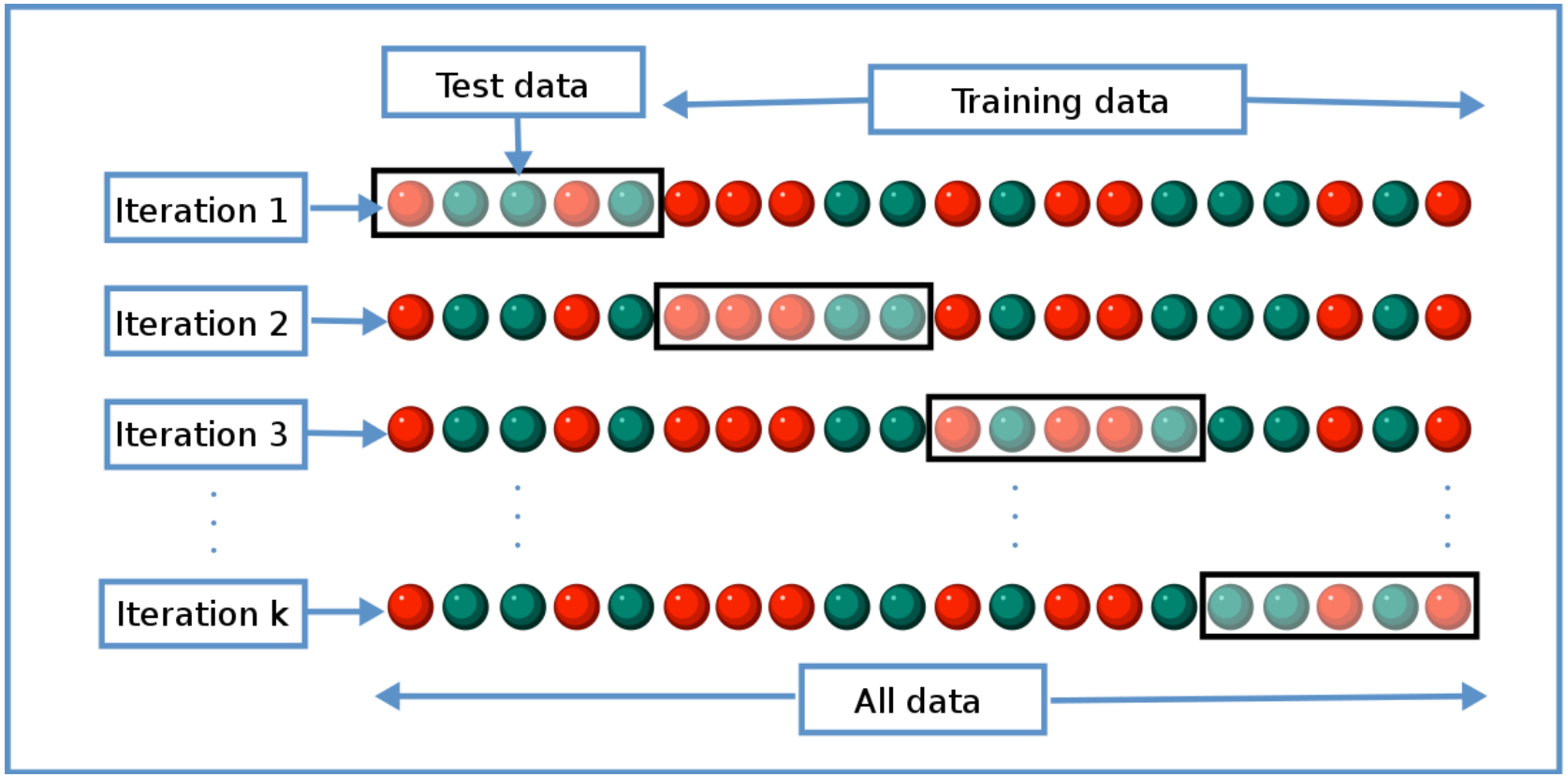
Validación cruzada: Cómo se comporta el modelo con data que aún no ha visto?

Hold out: Reservar una parte para entrenar

Y otra para pruebas.

k-fold cross-validation

Procedimiento de re-sampling usado para validar modelos ML con una muestra de datos limitada.



Tomado de: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

k-fold cross-validation

Procedimiento de re-sampling usado para validad modelos ML con una muestra de datos limitada.

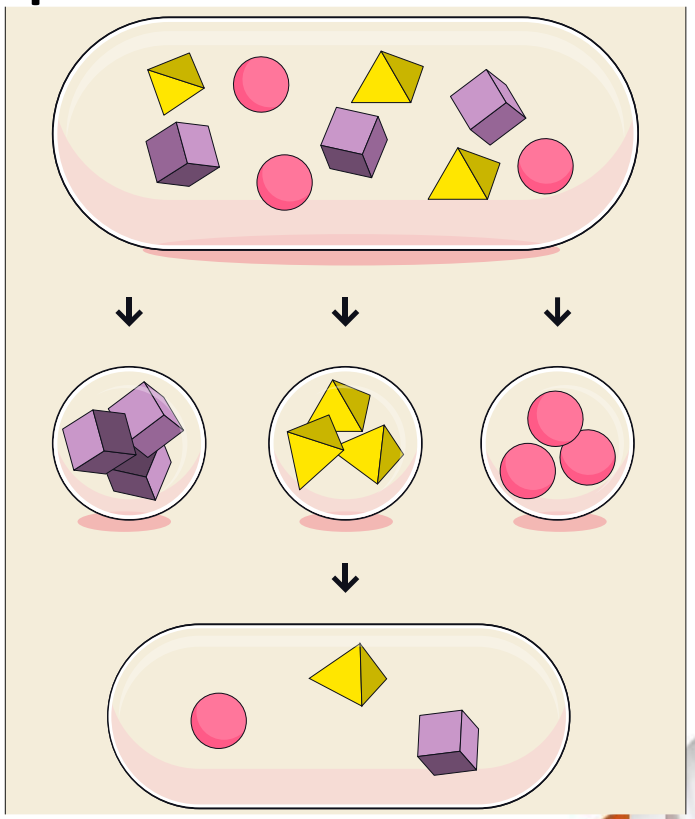
El procedimiento consiste en:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Tomado de: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Stratified K-fold cross-validation

Crea folds a partir de subgrupos (estratos/categorías) que reflejan la conformación proporcional de la población.



Tomado de:https://learn.robinhood.com/articles/2oCoidMzwxIjU3DaySZQuC/what-is-stratified-random-sampling/

Varianza=Dispersión y sesgo=Alejarse del blanco

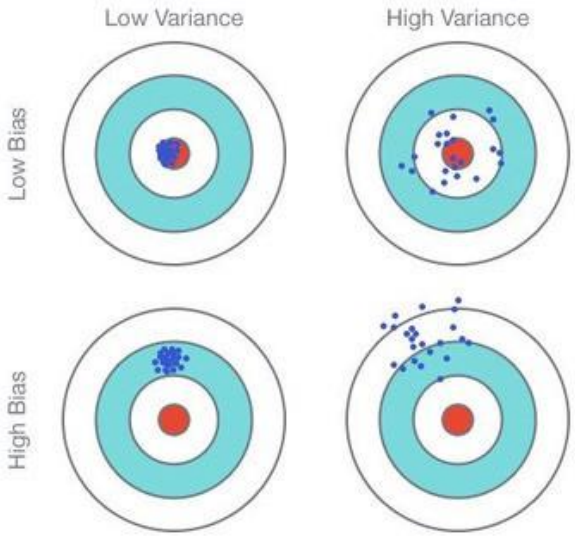
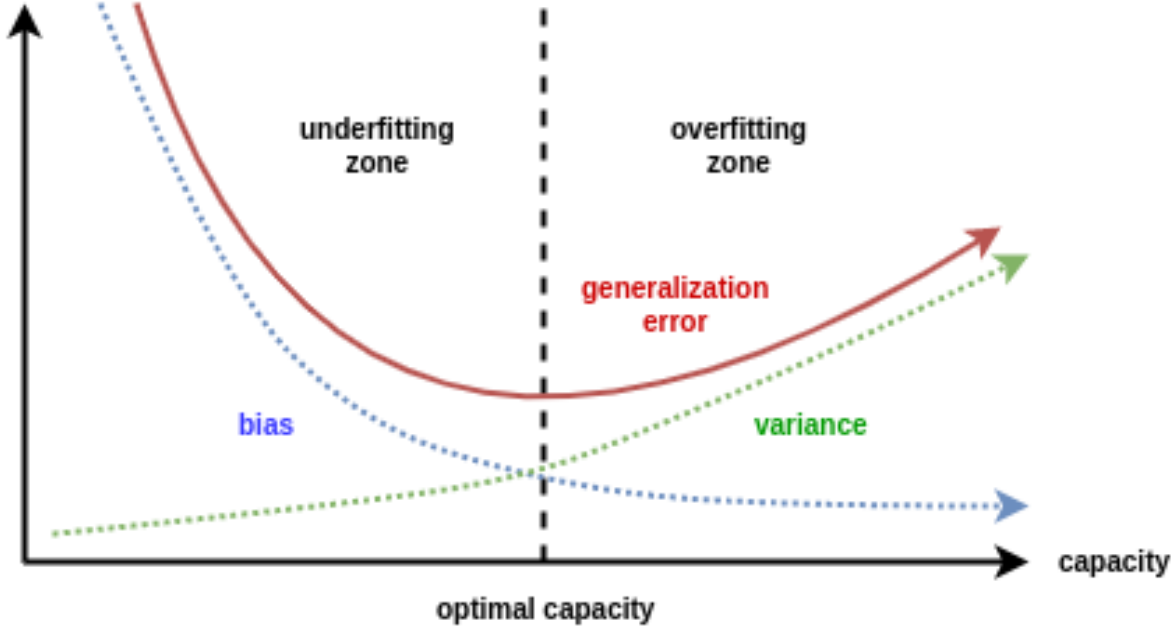


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off



Tomado de: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

SPRINT REVIEW

Conocer los problemas solucionables con Machine Learning

Conocer los algoritmos y conceptos básicos de ML

Ver los tipos de problemas de ML por medio de workflows en pandas

SPRINT RETROSPECTIVE

POR HACER:

compensar | fundación universitaria

VIGILADA MINEDUCACIÓN

Av. Calle 32 No. 17 - 30
Pbx: 555 82 10
ucompensar.edu.co
Bogotá, D.C. - Colombia

ucompensar