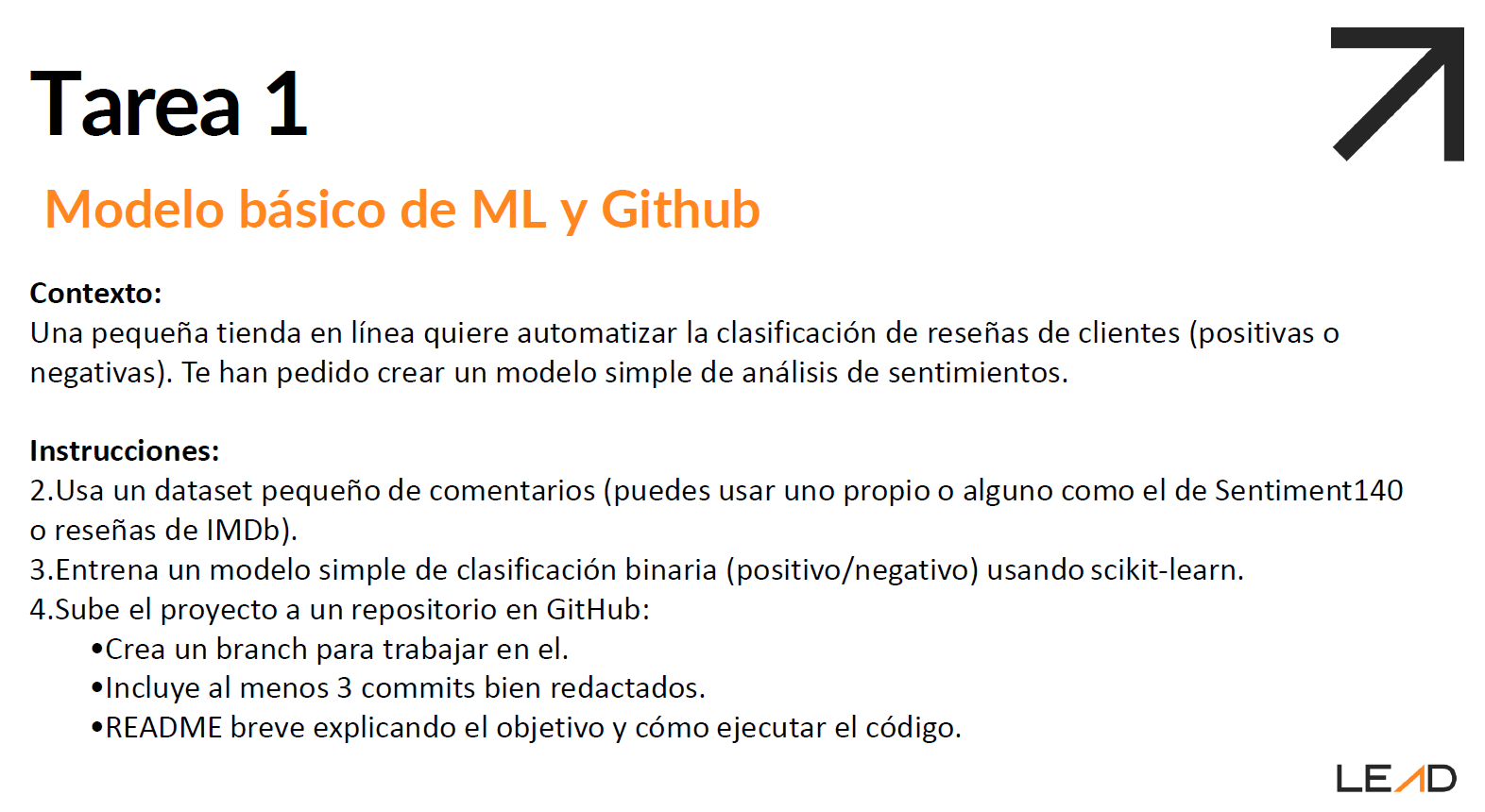
Tarea 1: Juan Carlos Cantillo Soto



**Solución:**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.naive\_bayes import MultinomialNB

from sklearn.metrics import classification\_report

# Cargar el dataset (solo una muestra por simplicidad)

df = pd.read\_csv('training.1600000.processed.noemoticon.csv', encoding='latin-1', header=None)

df = df[[0, 5]] # 0 = Sentimiento, 5 = Texto

df.columns = ['sentiment', 'text']

# Filtrar solo 5000 positivos y 5000 negativos

df = df[df['sentiment'].isin([0, 4])]

df = df.groupby('sentiment').head(5000)

# Convertir 0 → 0 (negativo), 4 → 1 (positivo)

df['sentiment'] = df['sentiment'].replace({0: 0, 4: 1})

# Separar datos

X = df['text']

y = df['sentiment']

# Dividir en entrenamiento y prueba

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Vectorizar texto

vectorizer = CountVectorizer(stop\_words='english')

X\_train\_vec = vectorizer.fit\_transform(X\_train)

X\_test\_vec = vectorizer.transform(X\_test)

# Entrenar modelo

model = MultinomialNB()

model.fit(X\_train\_vec, y\_train)

# Evaluar

y\_pred = model.predict(X\_test\_vec)

print(classification\_report(y\_test, y\_pred))