# Data Preparation

- Scalars
- One hot encoding

מתי ולמה משתמשים בשינוי קני מידה (scalars)

מודלים של למידת מכונה (כמו SVM , מסווגים לוגיסטיים ורשתות עצביות)

רגישים מאוד לסקאלת נתונים שונות.

משתנים בסקאלות שונות יכולים להשפיע על תוצאות המודל ולגרום לו להעדיף

משתנים גדולים.

מטרות שינוי קנה המידה:

- **שיפור ביצועי המודל** - מוודא שהמודל מתייחס לכל המשתנים באותה מידה.
- **האצת תהליך הלמידה** - נירמול מייעל את האופטימיזציה.
- **מניעת בעיות חישוביות** - ערכים גדולים עלולים לגרום לבעיות דיוק.

scalars

# Standard Scaler

## מה הוא עושה?

סטנדרטיזציה של הנתונים, כלומר החלפת כל נתון בציון התקן שלו כך שמתקבל ממוצע (μ) של 0 וסטיית תקן ( σ) של 1.

## נוסחא לחישוב ציון תקן:

$$Z = \frac{x_i - \mu}{\sigma}$$

## מתי להשתמש?

- כאשר הנתונים מפולגים בצורה **נורמלית (Normal Distribution)**
- מתאים למודלים מבוססי מרחק (כגון KNN, PCA)

**יתרונות:**

1.מבטל סקאלות שונות בין משתנים.

2.מתאים למודלים שמניחים פיזור נורמלי.

**חסרונות:**

1.רגיש מאוד לערכים קיצוניים (outliers)

2.ייתכן שלא יתאים לנתונים שאינם מפולגים נורמלית.

# Standard Scaler

# Min-Max Scaler

## מה הוא עושה?

נרמול הנתונים לסקאלה בין מינימום ומקסימום, לרוב בין 0 ל-1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

## מתי להשתמש?

- כאשר אינך מניח פיזור נורמלי.
- שימושי במיוחד למודלים שמושפעים מתכונות חיוביות בלבד (כמו רשתות עצביות).

**יתרונות:**

1.מתאים לנתונים שאינם נורמליים.

2.שומר על הערכים בטווח מוגדר (לרוב 0–1).

**חסרונות:**

1.רגיש לערכים קיצוניים, כי הם מגדירים את טווח הנירמול.

2.עלול לעוות את המשמעות של נתונים אם יש פיזור רחב מאוד.

# Min-Max Scaler

| מאפיין | Min-Max Scaler | Standard Scaler |
|---|---|---|
| פורמולה | $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ | $Z = \frac{x_i - \mu}{\sigma}$ |
| מטרה | סקאלה בטווח קבוע (לרוב 0–1) | ממוצע 0 וסטיית תקן 1 |
| רגישות ל-Outliers | גבוהה | גבוהה |
| מתאים לנתונים נורמליים | לא | כן |
| שימושים עיקריים | רשתות עצביות, מודלים לא ליניאריים | PCA, מודלים סטטיסטיים |

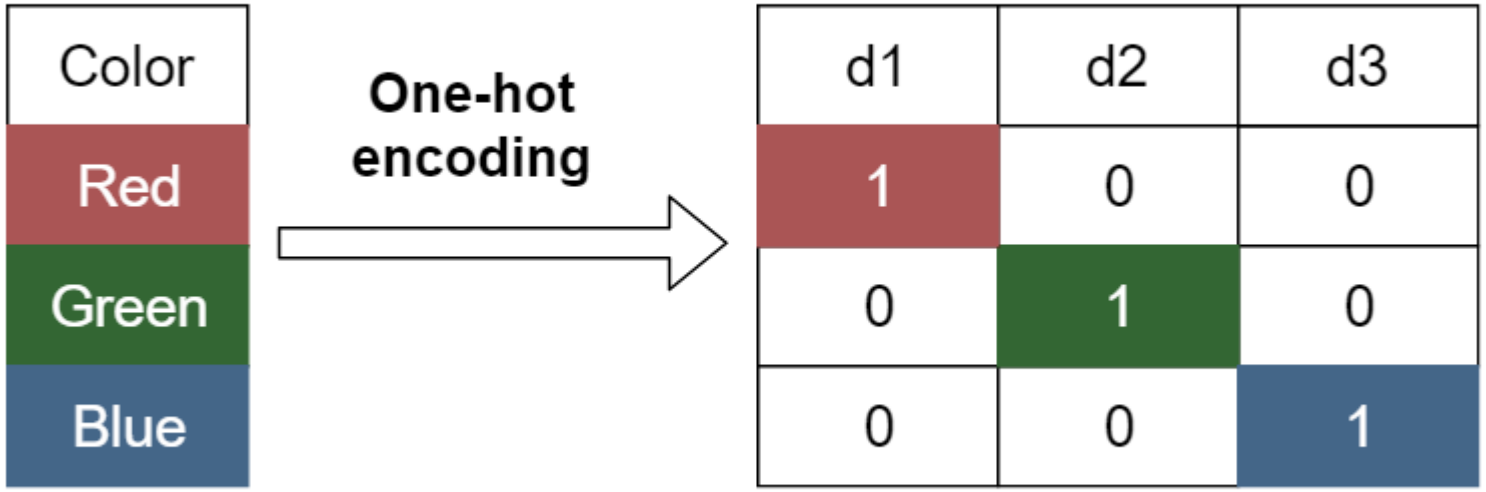
## למה צריך Encoding?

**מודלים סטטיסטיים ומודלים מבוססי למידה חישובית**

(כמו רגרסיה לינארית) לא יודעים להתמודד עם משתנים קטגוריים ישירות. הם זקוקים למספרים משתנים כמו "צבע" (אדום, כחול, ירוק) או "סוג חדר" (דירה, בית פרטי) צריכים לעבור המרה לערכים נומריים כדי שהמודל יוכל לעבוד איתם.

## מה זה OneHotEncoding?

- ממיר כל ערך בקטגוריה לעמודה חדשה.
- אם בעמודה יש nnnקטגוריות, נקבל nnnעמודות (כל אחת עם ערכים 0 או 1).



**pd.get_dummies**

- פונקציה של pandas שמבצעת **One-Hot Encoding** בצורה פשוטה ומהירה.
- יוצרת עמודות חדשות לכל קטגוריה, בדיוק כמו OneHotEncoder.

להלן שתי הדרכים לבצע את הפעולה:

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder
```

```
df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first =True)
```

אופציונאלי
להסרת עמודה אחת מיותרת

```
ohe = OneHotEncoder(sparse_output=False, drop='first')
encoded = ohe.fit_transform(df[['Sex', 'Embarked']])
encoded_df = pd.DataFrame(encoded,
                          columns=ohe.get_feature_names_out(['Sex', 'Embarked']))
df = pd.concat([df,encoded_df], axis = 1).drop(['Sex', 'Embarked'], axis=1)
```

| תכונה | pd.get_dummies (pandas) | OneHotEncoder (sklearn) |
|---|---|---|
| **דרישות נתונים** | מקבלת נתוני pandas DataFrame או Series | דורשת נתונים בפורמט קטגוריה / אובייקט |
| **גמישות** | פחות גמישה | יותר גמישה |
| **חלק מצינור עבודה** | משמשת לעיבוד מהיר ואינטואיטיבי | משתלבת מצוין ב- Pipeline של sklearn |