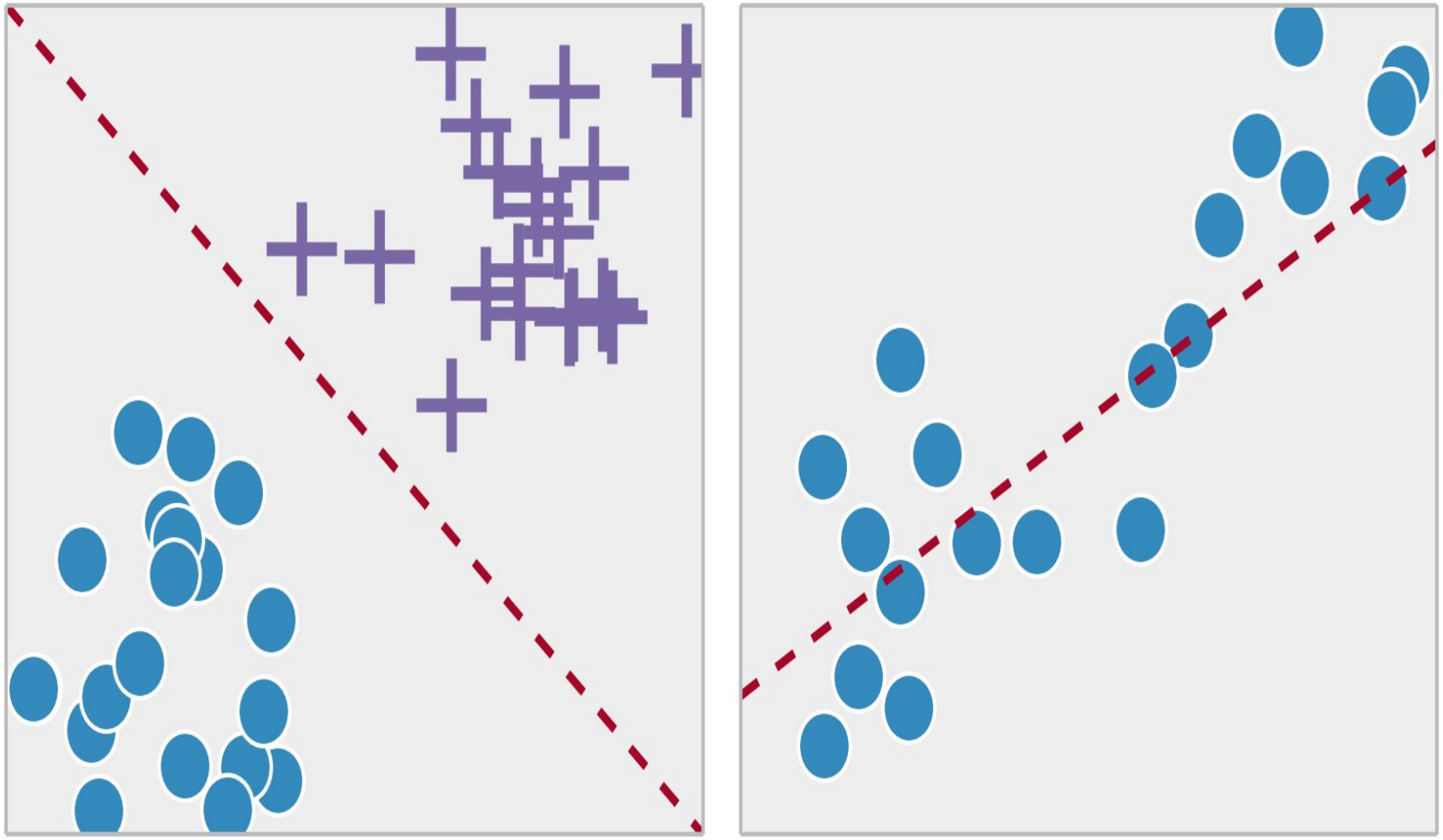
# Aprendizaje Automático

## Árboles de Decisión
## Regresión

# Classification vs Regression



# Árbol de regresión

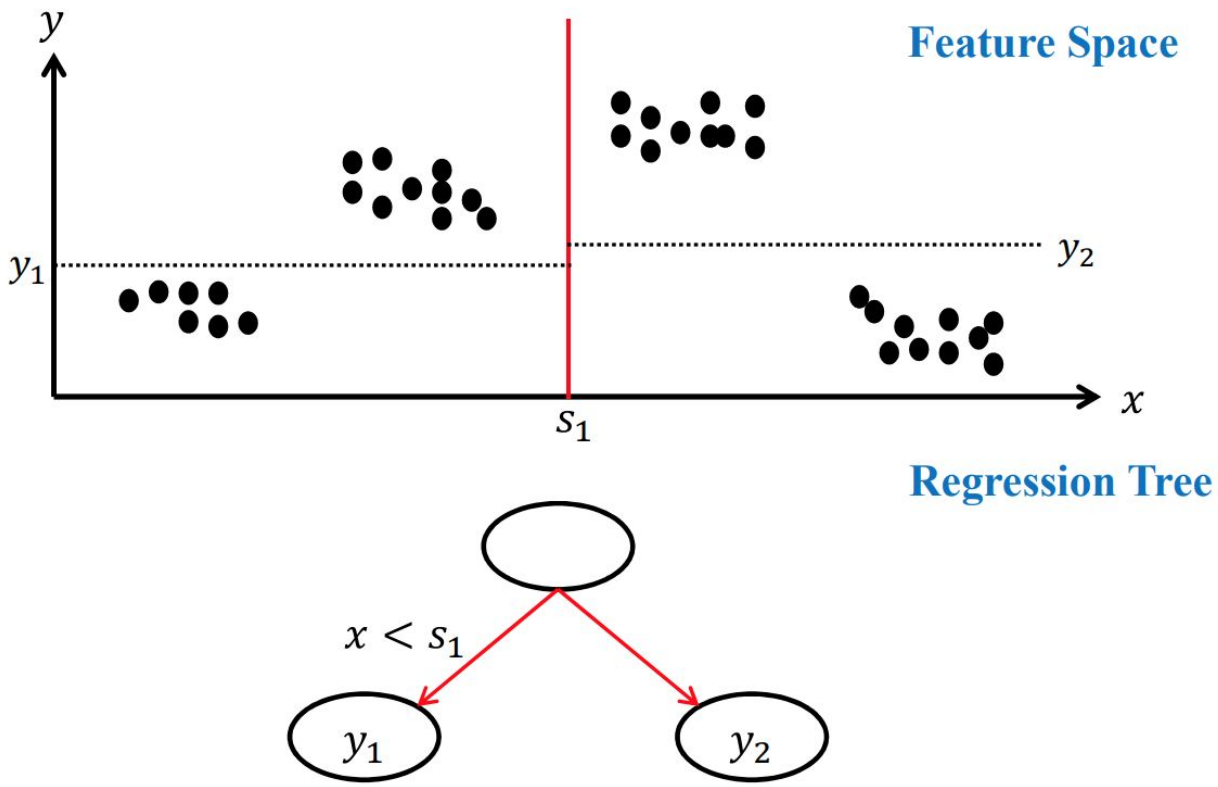
- Dado un conjunto de datos D = (x1, y1), …., (xn, yn) donde xi, yi € R
- Donde el target es un valor continuo
- El objetivo es predecir este valor continuo

```
data = pd.read_csv("PlayTennisR.csv")
```

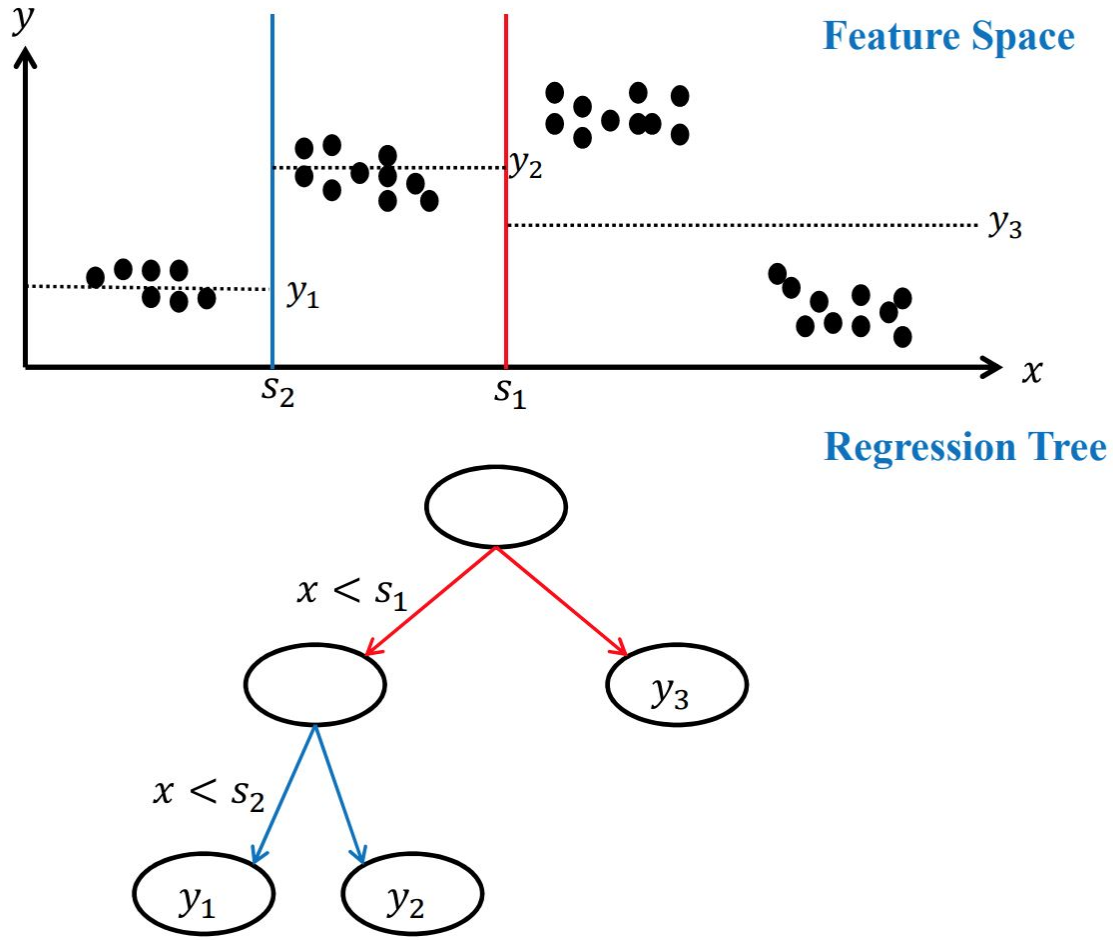
```
data.head()
```

| | Day | Outlook | Temperature | Humidity | Wind | Hours Played |
|---|---|---|---|---|---|---|
| **0** | D1 | Sunny | Hot | High | Weak | 26 |
| **1** | D2 | Sunny | Hot | High | Strong | 30 |
| **2** | D3 | Overcast | Hot | High | Weak | 48 |
| **3** | D4 | Rain | Mild | High | Weak | 46 |
| **4** | D5 | Rain | Cool | Normal | Weak | 62 |

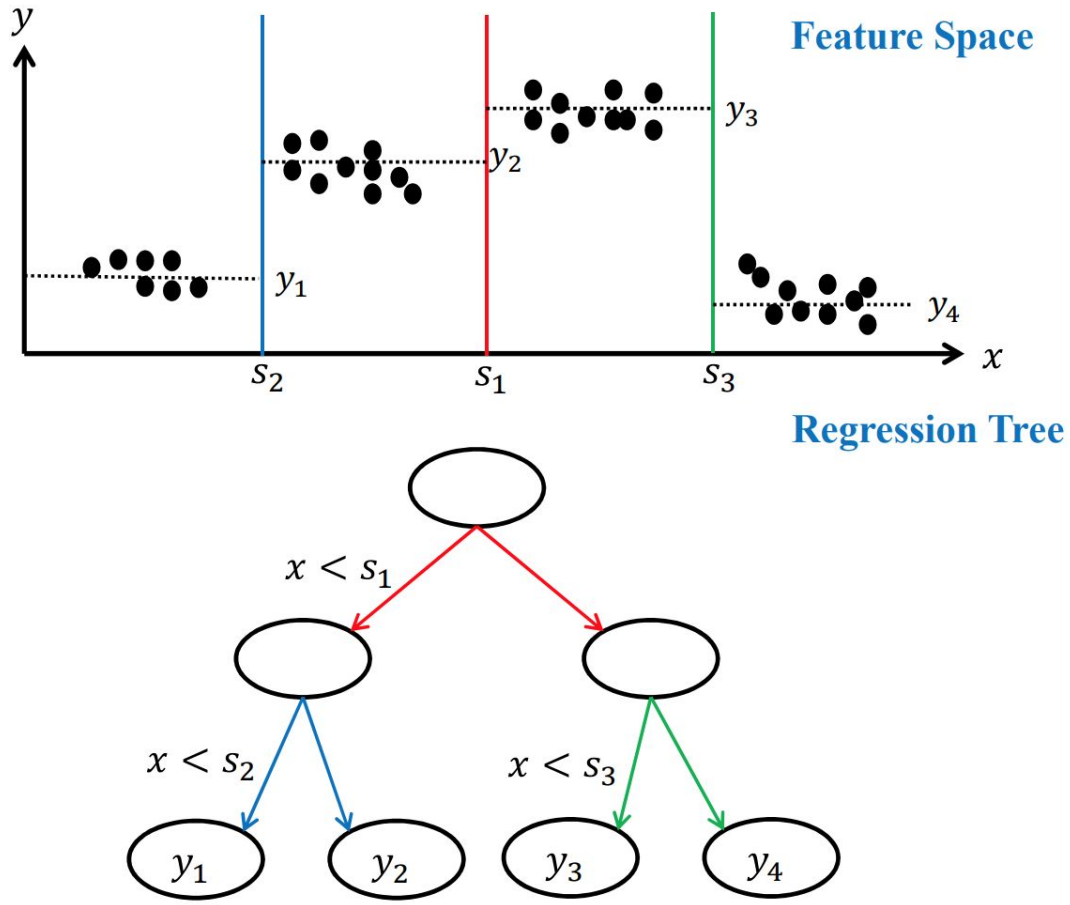
# Ejemplo de árbol de regresión



# Ejemplo de árbol de regresión



# Ejemplo de árbol de regresión



## Criterio de Separación

Mean Square Error - MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Escogemos el mínimo de los MSE

$$\hat{y} = \arg\min_{y\in\mathbb{R}} \sum_{i\in\mathcal{R}_k} (y - y_i)^2 = \frac{1}{\sum \mathbb{I}(x_i \in \mathcal{R}_k)} \cdot \sum \mathbb{I}(x_i \in \mathcal{R}_k) \cdot y_i$$

# Ejemplo de Play Tennis

## Conversión a valores numéricos

```
data = pd.read_csv("PlayTennisR.csv")
```

```
data.head()
```

| | Day | Outlook | Temperature | Humidity | Wind | Hours Played |
|---|---|---|---|---|---|---|
| **0** | D1 | Sunny | Hot | High | Weak | 26 |
| **1** | D2 | Sunny | Hot | High | Strong | 30 |
| **2** | D3 | Overcast | Hot | High | Weak | 48 |
| **3** | D4 | Rain | Mild | High | Weak | 46 |
| **4** | D5 | Rain | Cool | Normal | Weak | 62 |

| | Outlook | Temperature | Humidity | Wind | Hours Played |
|---|---|---|---|---|---|
| **0** | 2 | 1 | 0 | 1 | 26 |
| **1** | 2 | 1 | 0 | 0 | 30 |
| **2** | 0 | 1 | 0 | 1 | 48 |
| **3** | 1 | 2 | 0 | 1 | 46 |
| **4** | 1 | 0 | 1 | 1 | 62 |

# Cálculo de MSE

Variable: **X0 = Outlook**  Valores [0, 1, 2]

Posibles cortes (Split): [0.5, 1.5]

Con split = 0.5, entonces

| | Outlook | Temperatu | Humidity | Wind | Hours Played |
|---|---|---|---|---|---|
| 0 | 2 | 1 | 0 | 1 | 26 |
| 1 | 2 | 1 | 0 | 0 | 30 |
| 2 | 0 | 1 | 0 | 1 | 48 |
| 3 | 1 | 2 | 0 | 1 | 46 |
| 4 | 1 | 0 | 1 | 1 | 62 |
| 5 | 1 | 0 | 1 | 0 | 23 |
| 6 | 0 | 0 | 1 | 0 | 43 |
| 7 | 2 | 2 | 0 | 1 | 36 |
| 8 | 2 | 0 | 1 | 1 | 38 |
| 9 | 1 | 2 | 1 | 1 | 48 |
| 10 | 2 | 2 | 1 | 0 | 48 |
| 11 | 0 | 2 | 0 | 0 | 62 |
| 12 | 0 | 1 | 1 | 1 | 44 |
| 13 | 1 | 2 | 0 | 0 | 30 |

Si Outlook < 0.5
media = 49.25
MSE = 57.68

Si Outlook > 0.5
media = 38.7
MSE = 133.61

MSE total = 57.68 + 133.61
= 191.29

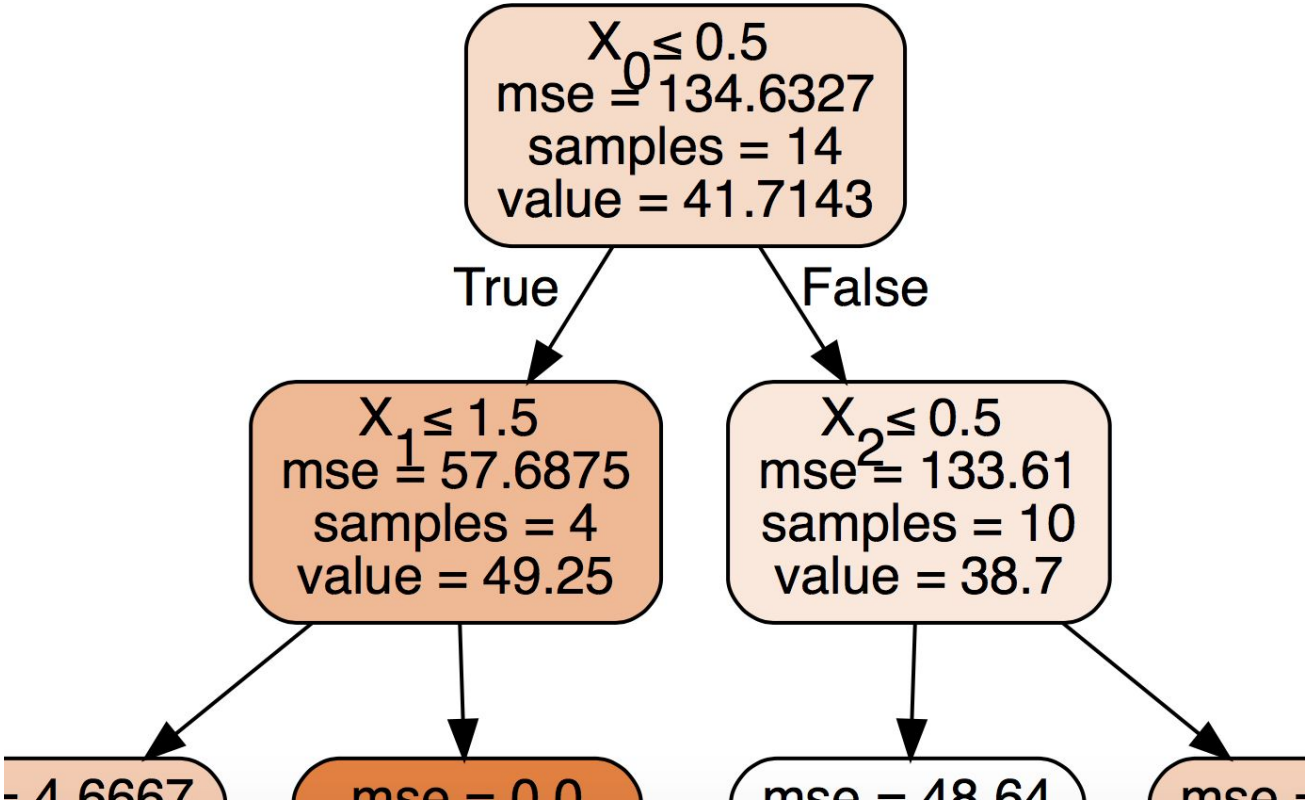
# Cuál es el menor MSE?

Realizando los cálculos:

| Variable | Split | Media1 | MSE1 | Media2 | MSE2 | MSE total |
|---|---|---|---|---|---|---|
| Outlook | 0.5 | 49.25 | 57.69 | 38.70 | 133.61 | **191.30** |
| | 1.5 | 45.11 | 145.65 | 35.60 | 288.08 | 433.73 |
| | | | | | | |
| Temperatura | 0.5 | 41.5 | 194.25 | 41.80 | 110.76 | 305.01 |
| | 1.5 | 45.00 | 102.33 | 39.25 | 144.68 | 247.01 |
| | | | | | | |
| Humidity | 0.5 | 43.71 | 119.06 | 39.71 | 142.20 | 261.26 |
| | | | | | | |
| Wind | 0.5 | 43.50 | 97.75 | 39.33 | 173.89 | 271.64 |

Outlook con split 0.5 sería el más óptimo para ser el nodo raíz

# Arbol generado



# Algoritmo

- Start with $\mathcal{R}_1 = \mathbb{R}^d$
- For each feature $j = 1, ..., d$, for each value $v \in \mathbb{R}$ that we can split on:
  - Split the data set:

$$I_< = \{i : x_{ij} < v\} \text{ and } I_> = \{i : x_{ij} \geq v\}$$

  - Estimate parameters:

$$\beta_< = \frac{\sum_{i \in I_<} y_i}{|I_<|} \text{ and } \beta_> = \frac{\sum_{i \in I_>} y_i}{|I_>|}$$

  - Quality of split is measured by the squared loss:

$$\sum_{i \in I_<} (y_i - \beta_<)^2 + \sum_{i \in I_>} (y_i - \beta_>)^2$$

- Choose split with minimal loss.
- Recurse on both children, with $(x_i, y_i)_{i \in I_<}$ and $(x_i, y_i)_{i \in I_>}$.

# Notebook

2 regression example

# Cómo evaluar el score en regresión?

$R^2$ o coeficiente de determinación: El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

# $R^2$: Coeficiente de determinación

$$\textit{Coefficient of Determination}, R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

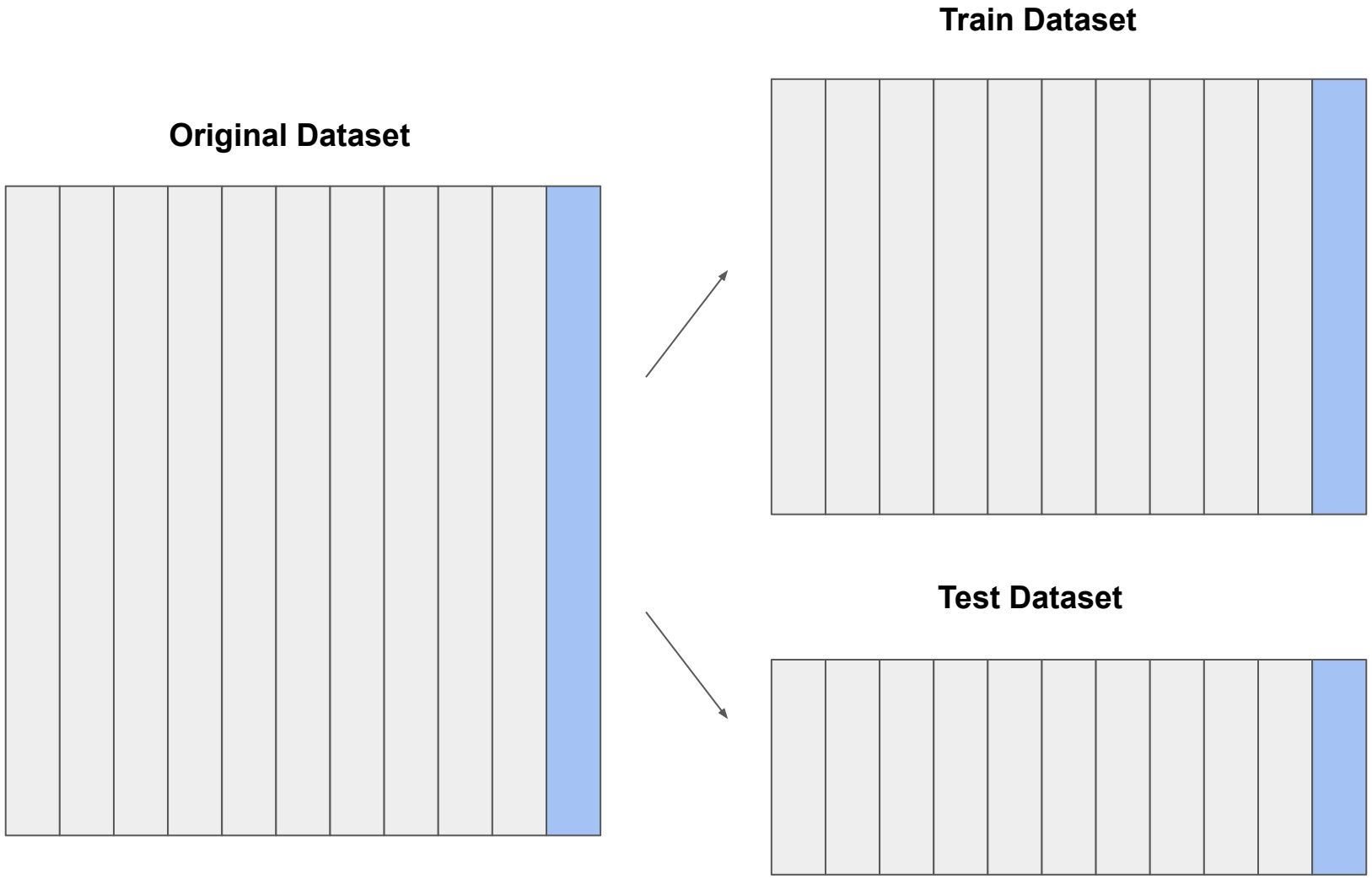
$$\textit{Sum of Squares Regression}, SSR = \sum_i (f_i - \bar{y})^2$$

$$\textit{Sum of Squares Total}, SST = \sum_i (y_i - \bar{y})^2$$

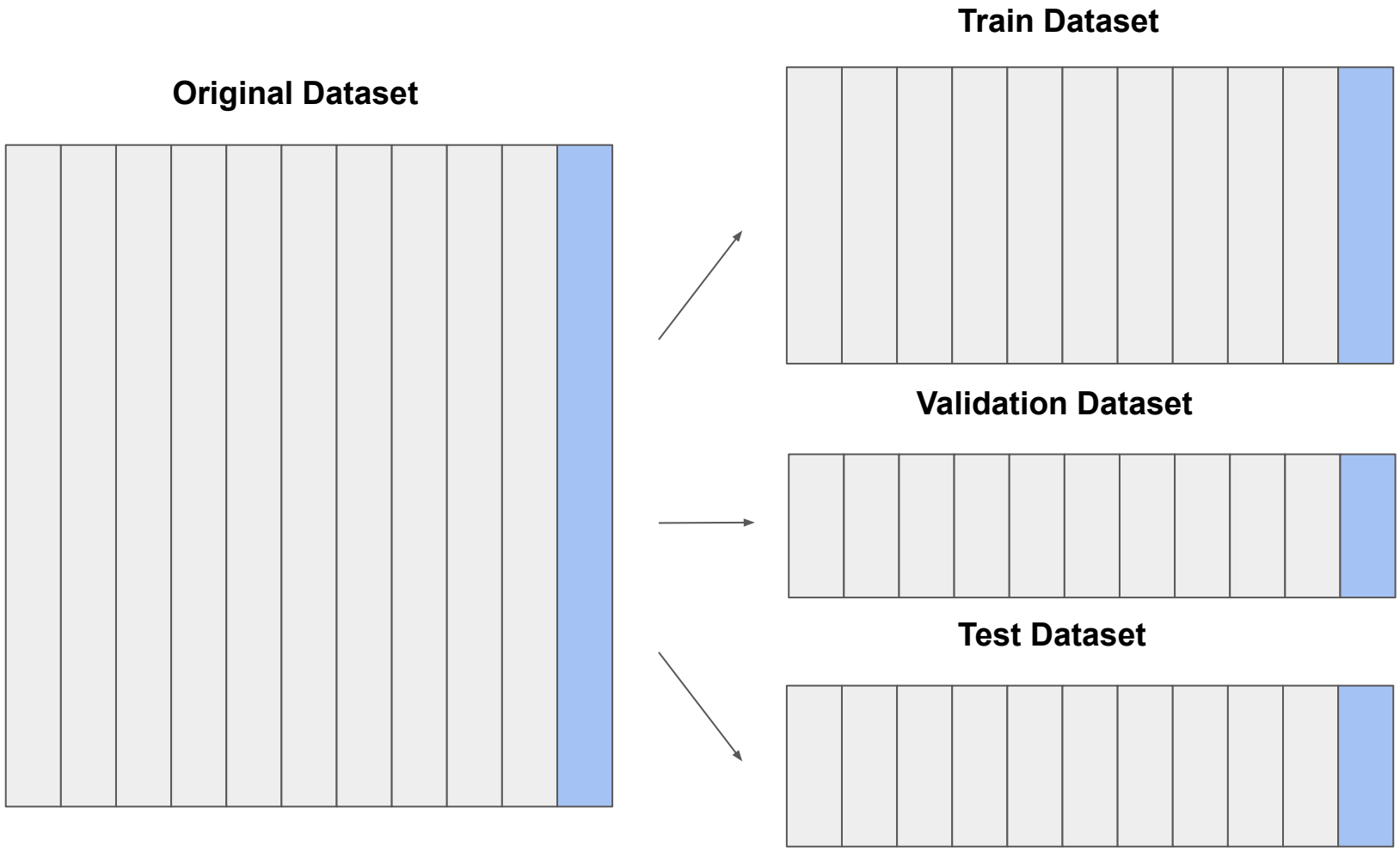
$$\textit{Sum of Squares Error}, SSE = \sum (y_i - f_i)^2$$

**Cuál es el rango de $R^2$?**

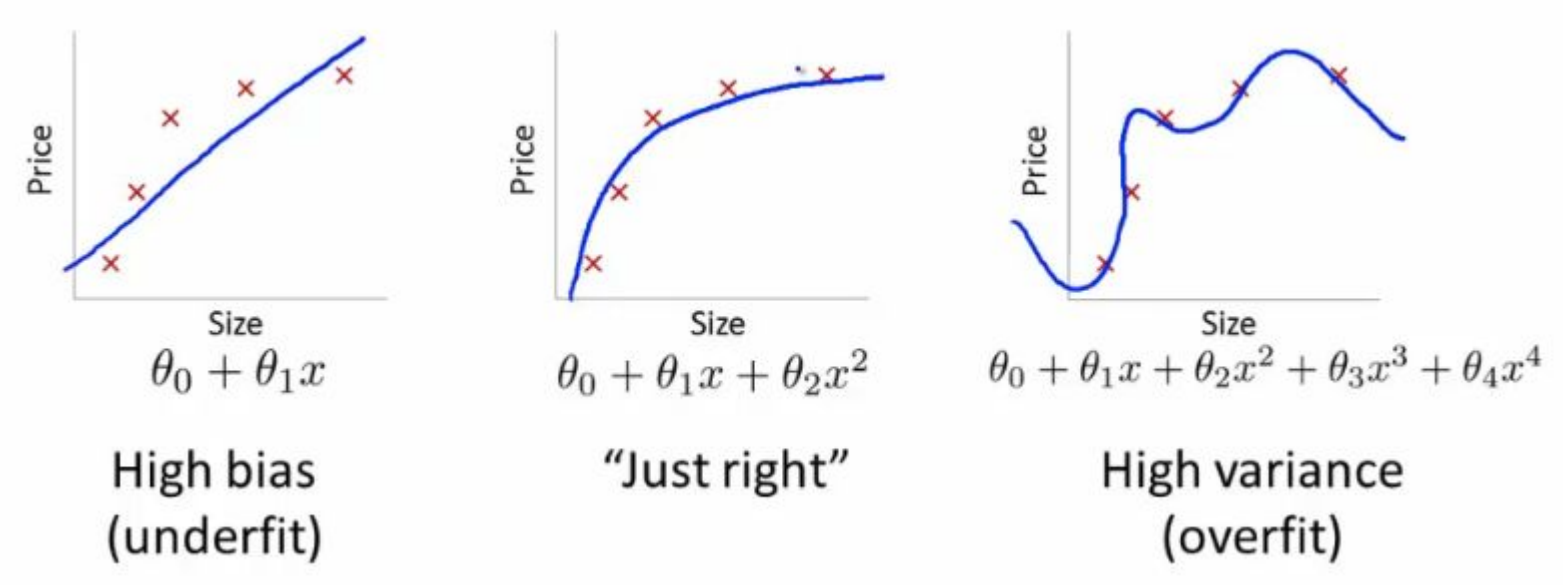
# Train and test split



# Even better: Train, validation and test split



# Overfitting



# Sobreentrenamiento

- Se debe evitar el sobreentrenamiento
  - Parar de crecer el árbol temprano.
  - Postprocesamiento del árbol (poda)

  Cómo?
  - Usar un conjunto de ejemplos de validación
  - Usar estadísticas