

Aufgabe 19

a)

Wenn eine ungewichtete Norm wie die euklidische Norm für die Abstandsberechnung verwendet wird und sich die Attribute in der Größenordnung unterscheiden, dann haben die größeren Attribute in der Abstandsberechnung ein deutlich höhere Gewicht. Es ist praktisch nur noch der Abstand des größten Attributs relevant, der Rest spielt keine Rolle mehr. Bspw. Abstand der zwei Punkte $a = (1, 10000, 5)^T$ und $b = (3, 10232, 9)^T$:

$$\sqrt{(a - b)^2} \approx b_2 - a_2 = 232 . \quad (1)$$

b)

Der k -NN-Algorithmus wird als *lazy learner* bezeichnet, weil praktisch der gesamte Rechenaufwand in die Anwendungs-Phase gelegt wird. Anders als bspw. beim Random Forest Verfahren, bei dem in der Lernphase rechenaufwendig Bäume erstellt werden, die sich in der Anwendungsphase mit relativ geringem Rechenaufwand auswerten lassen.

d)

Reinheit:	0,8340
Effizienz:	0,9644
Genauigkeit:	0,9242
Signifikanz:	66,76

e)

Reinheit:	0,8712
Effizienz:	0,9813
Genauigkeit:	0,9454
Signifikanz:	65,03

f)

Reinheit:	0,8590
Effizienz:	0,9859
Genauigkeit:	0,9414
Signifikanz:	66,26

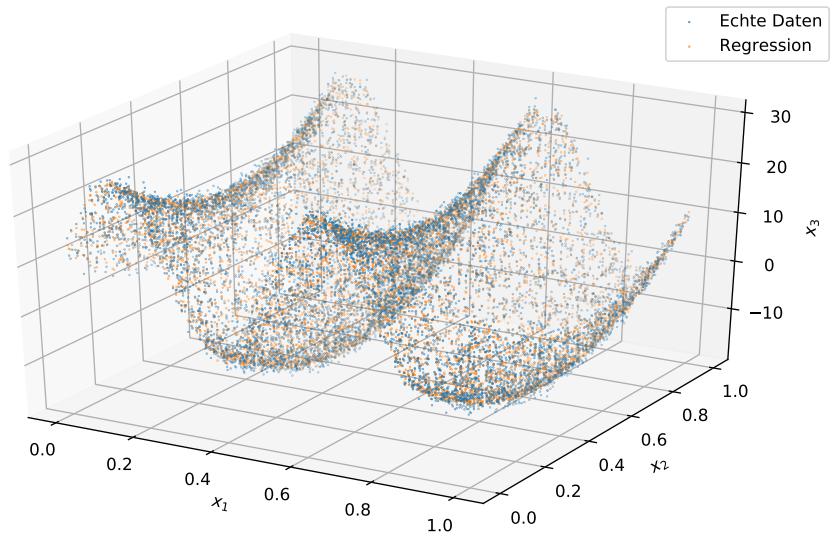


Abbildung 1: 3D-Plot, mit $x_1, x_2 \in \{0, 1\}$.

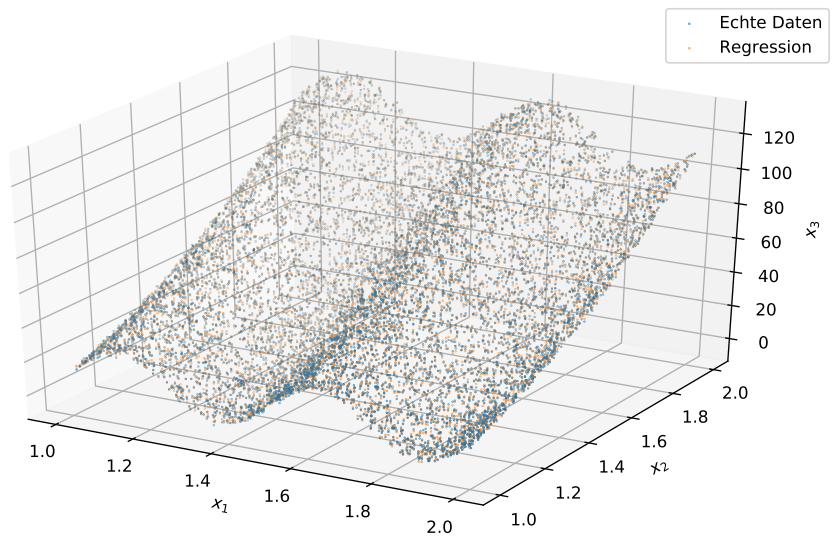


Abbildung 2: 3D-Plot, mit $x_1, x_2 \in \{1, 2\}$.

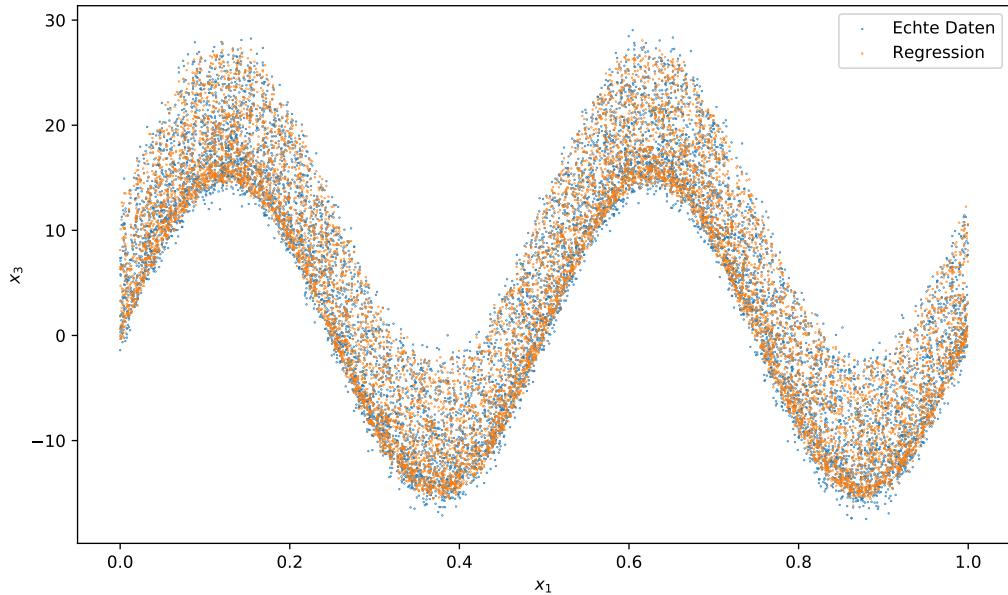


Abbildung 3: Projektion in die x_1, x_3 -Ebene, mit $x_1, x_2 \in \{0, 1\}$.

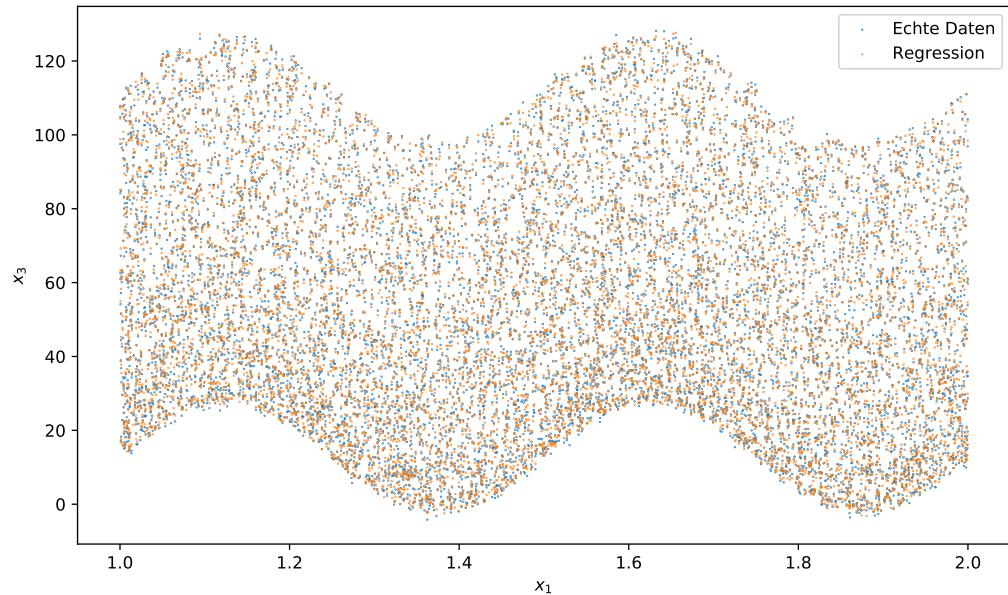


Abbildung 4: Projektion in die x_1, x_3 -Ebene, mit $x_1, x_2 \in \{1, 2\}$.

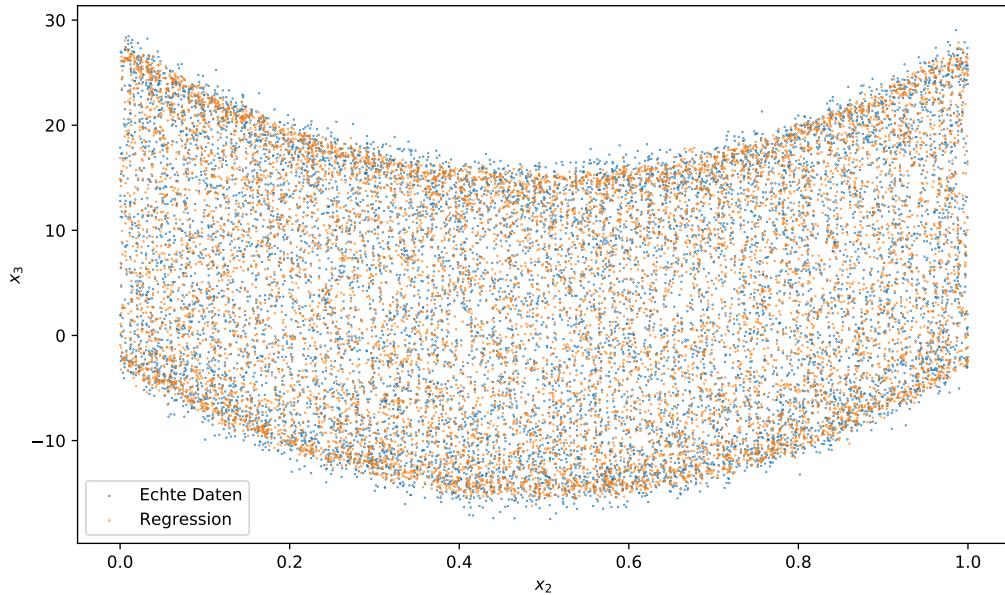


Abbildung 5: Projektion in die x_2, x_3 -Ebene, mit $x_1, x_2 \in \{0, 1\}$.

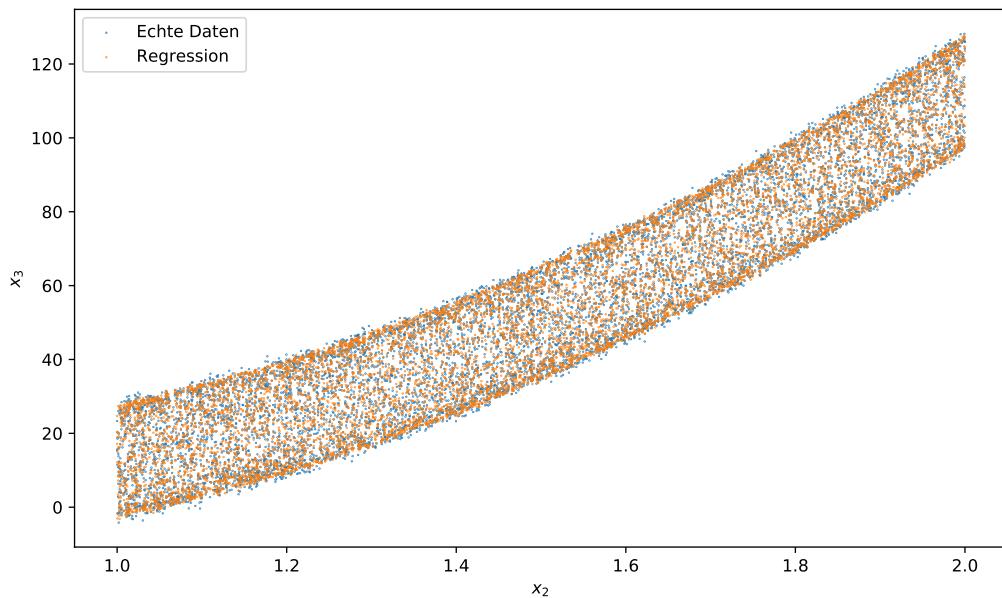


Abbildung 6: Projektion in die x_2, x_3 -Ebene, mit $x_1, x_2 \in \{1, 2\}$.

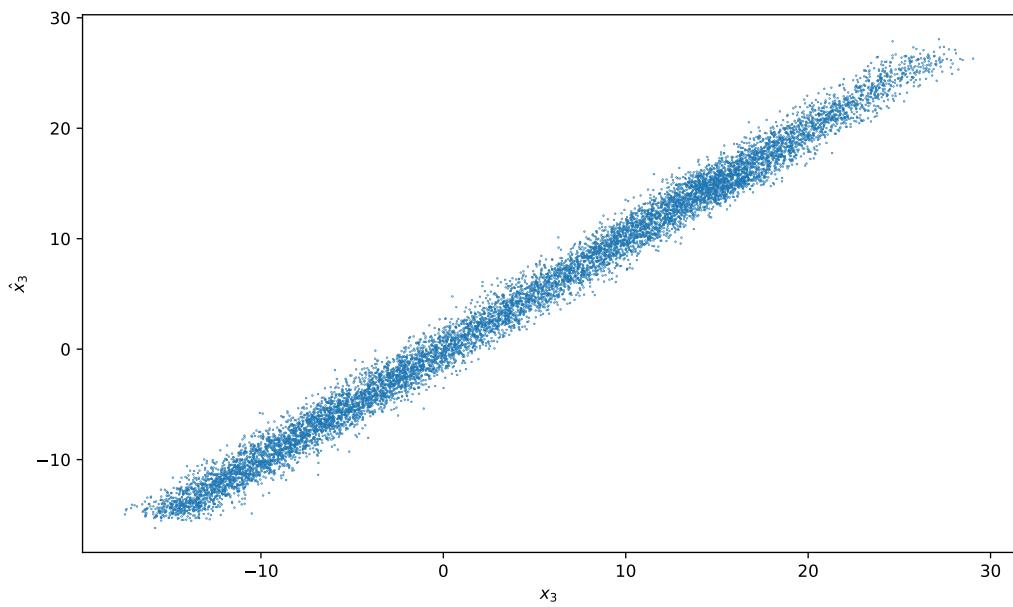


Abbildung 7: Geschätzter x_3 -Wert gegen den echten x_3 -Wert (optimal ist eine Gerade mit Steigung 1), mit $x_1, x_2 \in \{0, 1\}$.

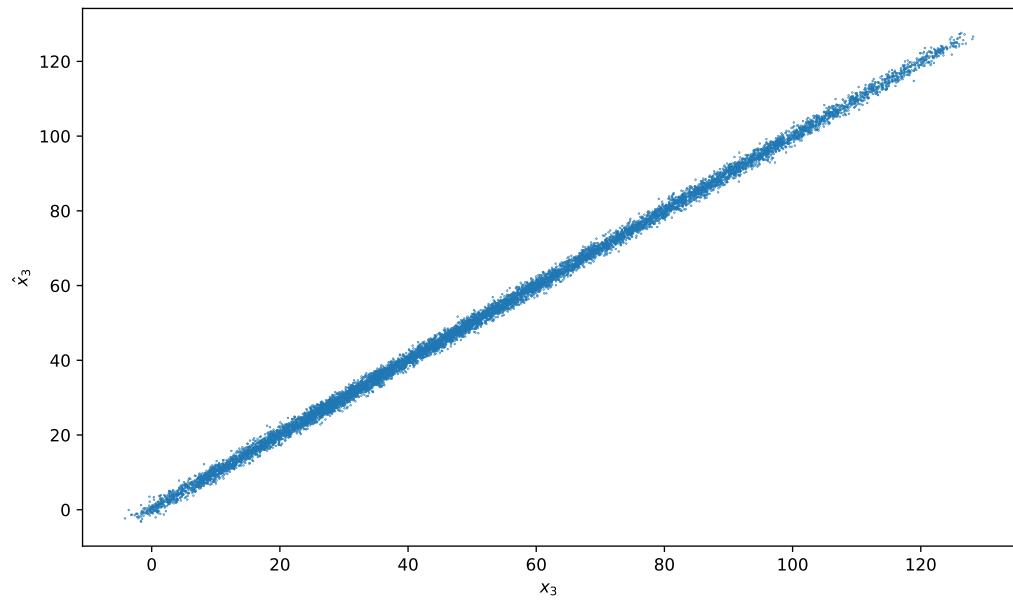


Abbildung 8: Geschätzter x_3 -Wert gegen den echten x_3 -Wert (optimal ist eine Gerade mit Steigung 1), mit $x_1, x_2 \in \{1, 2\}$.

Aufgabe 20

e, f)

$$\begin{aligned}\text{MSE, } x_1, x_2 \in \{0, 1\}: & \quad 1,245 \\ \text{MSE, } x_1, x_2 \in \{1, 2\}: & \quad 1,404\end{aligned}$$

21. a)

$$H(Y) = - \sum_{z \in Z} P(Y=z) \log_2(P(Y=z))$$

$$Z = \{\text{False}, \text{True}\}$$

$$P(Y=\text{False}) = \frac{5}{14}, \quad P(Y=\text{True}) = \frac{9}{14}$$

$$\Rightarrow H(Y) = -\frac{5}{14} \log_2\left(\frac{5}{14}\right) - \frac{9}{14} \log_2\left(\frac{9}{14}\right) \approx 0,9903$$

b) Entropie vorher: $H(Y)$

Entropie nachher: $H(Y|X)$

$$\Rightarrow H(Y|X) = - \sum_{m \in M} P(X=m) \sum_{z \in Z} P(Y=z|X=m) \times \log(P(Y=z|X=m))$$

$$M = \{\text{False}, \text{True}\}$$

$$P(X=\text{False}) = \frac{8}{14}, \quad P(X=\text{True}) = \frac{6}{14}$$

$$P(Y=\text{False}|X=\text{False}) = \frac{2}{14}$$

$$P(Y=\text{False}|X=\text{True}) = \frac{3}{14}$$

$$P(Y=\text{True}|X=\text{False}) = \frac{6}{14}$$

$$P(Y=\text{True}|X=\text{True}) = \frac{3}{14}$$

$$\begin{aligned} \Rightarrow H(Y|X) &= -\frac{8}{14} \left(\frac{2}{14} \log_2\left(\frac{2}{14}\right) + \frac{6}{14} \log_2\left(\frac{6}{14}\right) \right) \\ &\quad - \frac{6}{14} \left(\frac{3}{14} \log_2\left(\frac{3}{14}\right) + \frac{3}{14} \log_2\left(\frac{3}{14}\right) \right) \end{aligned}$$

$$\Rightarrow H(Y|X) \approx 0,9367$$

$$\Rightarrow H(Y) - H(Y|X) \approx 0,00356$$

Aufgabe 21

c)

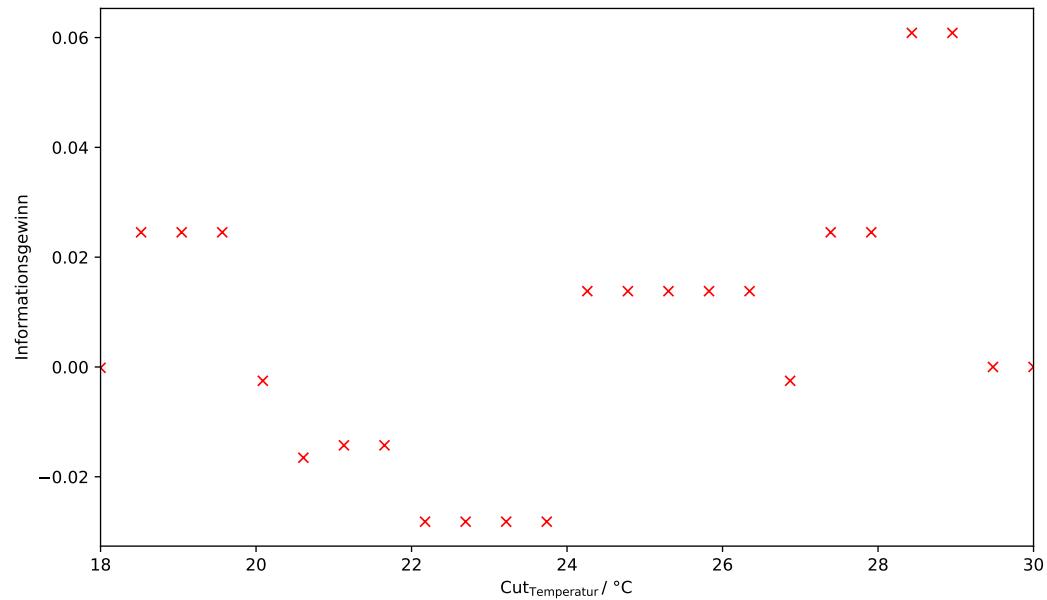


Abbildung 9: Untersuchung des Attributes "Temperatur".

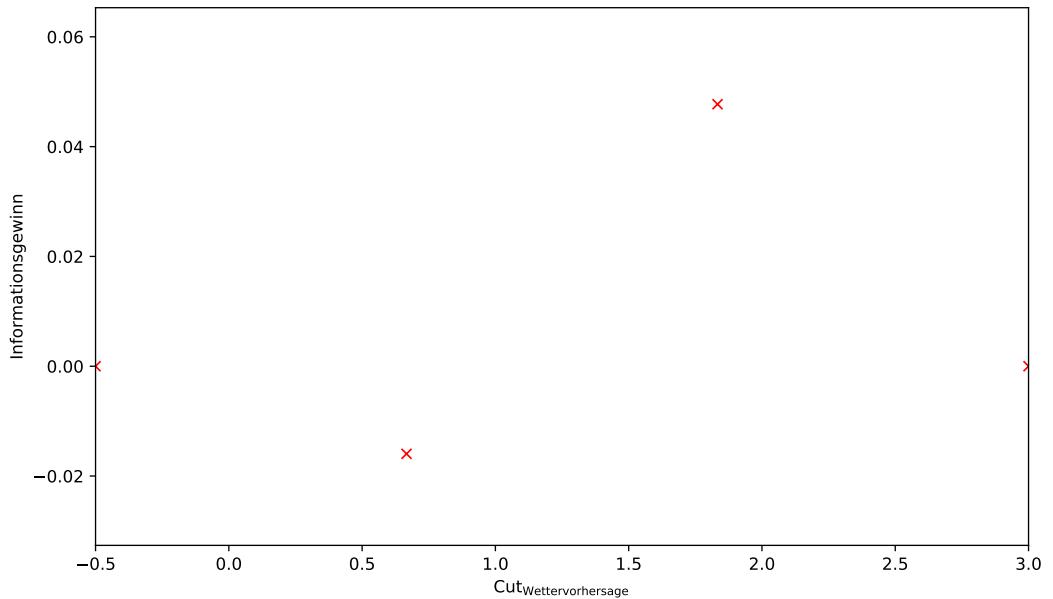


Abbildung 10: Untersuchung des Attributes "Wettervorhersage".

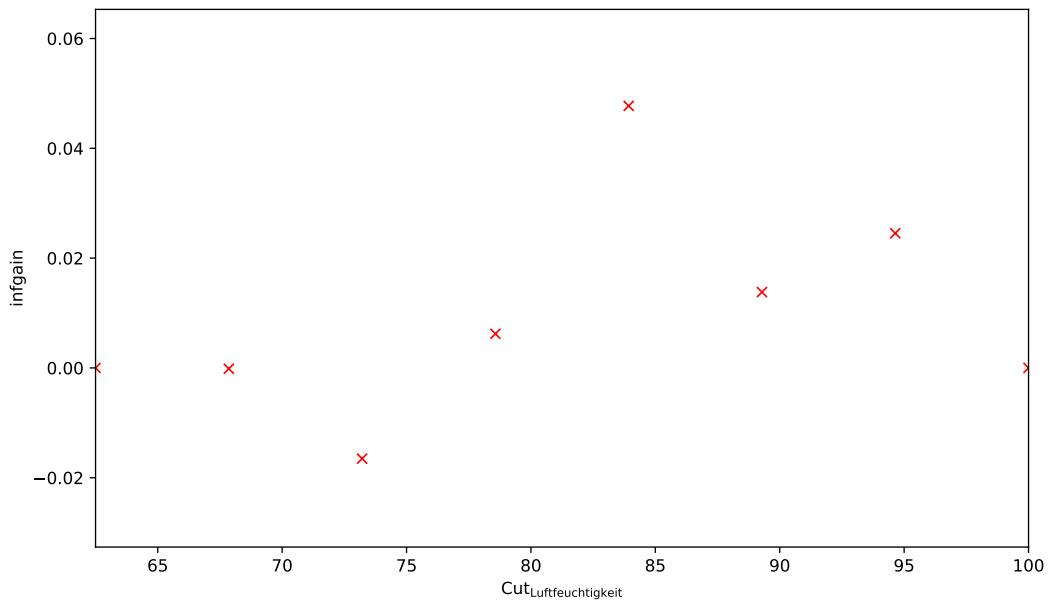


Abbildung 11: Untersuchung des Attributes "Luftfeuchtigkeit".

d)

Da bei der Temperatur der höchste Informationsgewinn vorliegt, ist das das beste Attribut zum Trennen der Daten.