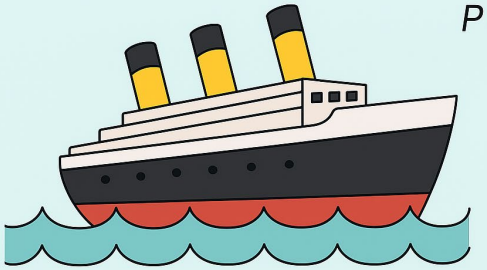
Naive Bayes Classifier



$$P(\text{Survived} \mid \text{Pclass=First}, \text{Sex=male}, \text{Age=27}) \propto P(\text{Survived} \mid \text{Pclass=First}) P(\text{Survived} \mid \text{Sex=male}) P(\text{Survived} \mid \text{Age=27})$$

Big Data & Data Science
Naïve Bayes Classifier

WS 2025/26

Prof. Dr. Klemens Waldhör

Copyright

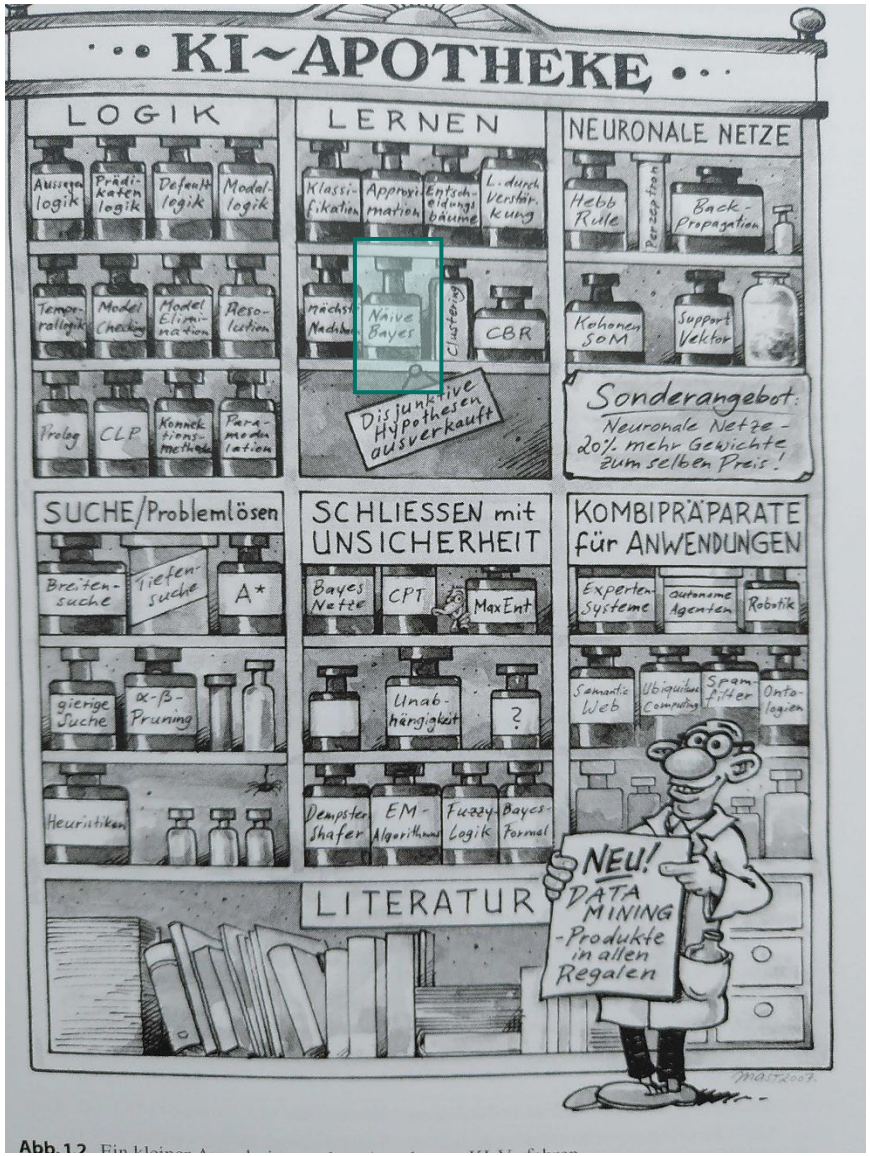


Abb. 1.2

Machine Learning

Naïve Bayes Classifier

Grundidee, Formel und zentrale Eigenschaften

Formel (Satz von Bayes):

$P(C|X) = [\ P(X|C) \cdot P(C)\ ] / P(X)$ oder

$P(\text{Klasse}|\text{Merkmale}) = [P(\text{Merkmale}|\text{Klasse}) \cdot P(\text{Klasse})] / P(\text{Merkmale})$

Die Merkmale werden als unabhängig angenommen (naive Annahme):

$P(X|C) = \prod P(x_i|C)$

Der Naïve-Bayes-Klassifikator ist ein einfaches, aber wirkungsvolles Verfahren der überwachten Klassifikation.

Er schätzt für jede mögliche Klasse die Wahrscheinlichkeit, dass ein Objekt mit gegebenen Merkmalen zu dieser Klasse gehört. Bayes hilft, Wahrscheinlichkeiten zu aktualisieren, wenn neue Informationen (Merkmale) vorliegen.

Bayes hilft, Wahrscheinlichkeiten zu aktualisieren, wenn neue Informationen (Merkmale) vorliegen.

- Klassen: „Spam“ oder „Nicht-Spam“ Merkmale: Wörter, Absender, Betreff usw
- Naiver Bayes-Klassifikator: Annahme unabhängiger Merkmale → $P(\text{Merkmale}|\text{Klasse}) = \prod P(M_i|\text{Klasse})$

Symbol	Bedeutung
C	Klasse (z. B. männlich / weiblich)
$X = (x_1,\ldots,x_n)$	Merkmalsvektor (z. B. Passagiere)
P(C)	Prior – Grundwahrscheinlichkeit der Klasse
P(X|C)	Likelihood – Wahrscheinlichkeit der Merkmale bei Klasse C
P(X)	Evidenz – Wahrscheinlichkeit, die Merkmale insgesamt zu beobachten
P(C|X)	Posterior – gesuchte Wahrscheinlichkeit, dass Objekt zur Klasse C gehört

Vorteile und Nachteile

Vorteile:

- Einfach und sehr schnell zu berechnen
- Funktioniert auch mit kleinen Trainingsmengen
- Gut interpretierbar (jede Komponente hat klare Bedeutung)
- Besonders geeignet für Textklassifikation oder Basisprognosen

Nachteile:

- Unabhängigkeitsannahme oft unrealistisch
- Schätzungen instabil bei seltenen Kombinationen (→ Laplace-Glättung nötig)
- Modell kann Korrelationen zwischen Merkmalen nicht erfassen

Titanic

Angenommen, man möchte die Überlebenswahrscheinlichkeit („*Yes*“) eines Titanic-Passagiers berechnen, der

- in der 1. Klasse gereist ist,
- männlich ist und
- 27 Jahre alt war.

Dies lässt sich als bedingte Wahrscheinlichkeit formulieren:

$$P(Yes|First, Male, Age = 27)$$

Der Naive-Bayes-Ansatz geht davon aus, dass diese Wahrscheinlichkeit aus den bedingten Wahrscheinlichkeiten der einzelnen Merkmale berechnet werden kann:

$$P(Yes|First, Male, Age = 27) = P(Yes|First) \cdot P(Yes|Male) \cdot P(Yes|Age = 27)$$

Titanic – Survived – Male – fiktive Daten

Wahrscheinlichkeit zu überleben: $P(Survived) = 0.7$

Wahrscheinlichkeit, männlich zu sein: $P(Male) = 0.6$

Gemeinsame Wahrscheinlichkeit (Joint Probability / Verbundwahrscheinlichkeit)

- Wahrscheinlichkeit, dass zwei bestimmte Merkmalsausprägungen gleichzeitig auftreten.
- Wahrscheinlichkeit, dass – unter allen Passagieren (männlich und weiblich) – ein männlicher Passagier überlebt hat.

wenn Survived und Male **unabhängig** wären!

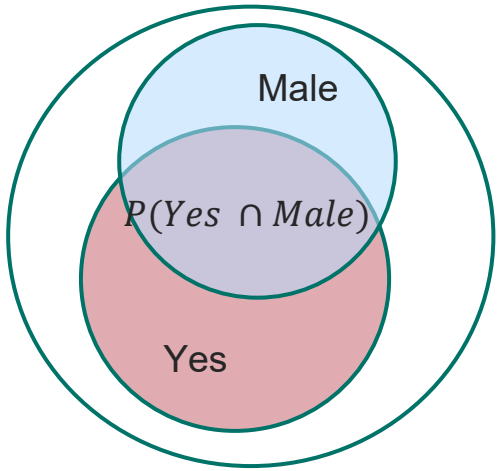
$$P(Survived \cap Male) = P(Survived) \cdot P(Male) = 0.42$$

Bedingte Wahrscheinlichkeit (Conditional Probability)

Wahrscheinlichkeit, dass ein bestimmter Merkmalswert auftritt, unter der Bedingung, dass ein anderer Merkmalswert gleichzeitig gilt.

Wahrscheinlichkeit, dass – unter allen männlichen Passagieren – ein Passagier überlebt hat.

$$P(Survived|Male) = \frac{P(Survived \cap Male)}{P(Male)} = \frac{0.42}{0.6} = 0.7$$



Titanic – Survived – Male – reale Daten

Geschlecht	Überlebt = 1	Gestorben = 0	Zeilensumme
männlich	109/891 = 0.122	468/891 = 0.526	577/891 = **0.647**
weiblich	233/891 = 0.262	81/891 = 0.091	314/891 = **0.353**
Spaltensumme	342/891 = **0.384**	549/891 = **0.616**	891/891 = **1.000**

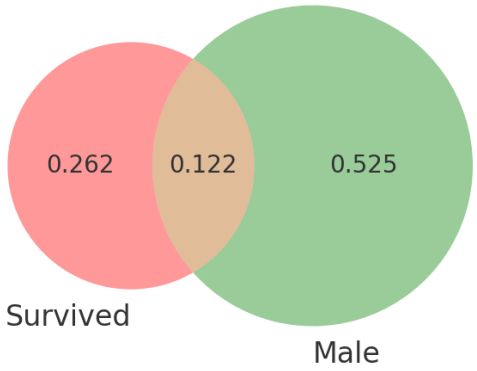
Gemeinsame Wahrscheinlichkeit Überlebt und Männlich

- P(Survived ∩ Male) = 109/891 = 0.122

Bedingte Wahrscheinlichkeit „Überlebt unter den Männern“

- P(Survived|Male) = P(Survived ∩ Male) / P(Male) = 0.122/0.647 = 0.188 (18.8%)

Venn-Diagramm: Survived & Male (reale Titanic-Werte)



Titanic – Survived - First

- Der Naive-Bayes-Klassifikator erfordert, dass wir die Überlebenswahrscheinlichkeit für jedes einzelne Merkmal berechnen.
- Beispiel: Die Wahrscheinlichkeit des Überlebens in Abhängigkeit von der Passagierklasse.
- Wir möchten also berechnen: $P(\text{Yes} \mid \text{First})$, $P(\text{Yes} \mid \text{Second})$ und $P(\text{Yes} \mid \text{Third})$

Passagierklasse	Überlebt
First	Yes
Third	No
Third	No
Second	No
Third	Yes
First	Yes
Second	Yes
Second	Yes
First	Yes
Third	Yes

Häufigkeitstabelle		
Class	**No**	**Yes**
First	80	136
Second	97	87
Third	372	119
Total	**549**	**342**

Titanic – Survived - First

- Berechne und füge Wahrscheinlichkeiten zur Häufigkeitstabelle

Häufigkeitstabelle (mit Wahrscheinlichkeiten)						
Class	**No**	**Yes**	**Total**			
First	80	136	**216**	=216/891	0.2424	$P(First)$
Second	97	87	**184**	=184/891	0.2065	$P(Second)$
Third	372	119	**491**	=491/891	0.5510	$P(Third)$
Total	**549**	**342**	**891**			
	=549/891	**342/891**				
	0.61616	**0.38383**				
	$P(No)$	$P(Yes)$				

$$P(First|Yes) = \frac{n(\text{First} \cap Yes)}{n(Yes)} = \frac{136}{342} = 0.39$$

$$P(Yes) = \frac{342}{891} = 0.3838$$

- Wir berechnen $P(Yes|First)$ aus der Tabelle $P(Yes|First) = \frac{n(\text{First} \cap Yes)}{n(First)} = \frac{136}{216} = \mathbf{0.62}$
- Wir wenden das Bayes Theorem an

likelihood **prior**

$$P(Yes|First) = \frac{P(First|Yes)P(Yes)}{P(First)} = \frac{0.39 \cdot 0.38}{0.24} = \mathbf{0.62}$$

posterior **Evidenz**

Titanic – Survived, Male. First, Age

- Wir wiederholen diesen Vorgang für alle weiteren Merkmale, um die benötigten bedingten Wahrscheinlichkeiten zu berechnen:

$$P(\text{Yes} \mid \text{First}), P(\text{Yes} \mid \text{Male}) \text{ und } P(\text{Yes} \mid \text{Age} = 27)$$

- Unter der „naiven“ Annahme, dass alle Merkmale voneinander unabhängig sind, können wir die einzelnen Ergebnisse multiplizieren, um eine Vorhersage der Überlebenswahrscheinlichkeit zu erhalten:

$$P(\text{Yes} \mid \text{First, Male, Age=27}) \propto P(\text{Yes} \mid \text{First}) \cdot P(\text{Yes} \mid \text{Male}) \cdot P(\text{Yes} \mid \text{Age=27})$$

- Da gilt $P(\text{No}) = 1 - P(\text{Yes})$, können wir die Klassifikation nach folgender Entscheidungsregel vornehmen:

Wenn P(Yes | First, Male, Age=27) > 0.5 → survived

sonst → not survived

Warum Bayes-Formel wirklich notwendig?

$$P(Yes|First) = \frac{n(\text{First} \cap Yes)}{n(First)} = \frac{136}{216} = \mathbf{0.62}$$

Wenn vollständige Kreuztabelle (z. B. „Klasse × Überleben“) mit allen Zellenwerten bekannt ist, kann man direkt jede bedingte Wahrscheinlichkeit berechnen..

Wozu dann?

$$P(Yes|First) = \frac{P(First|Yes)P(Yes)}{P(First)} = \frac{0.39 \cdot 0.38}{0.24} = \mathbf{0.62}$$

Satz von Bayes ist dann relevant, wenn Kreuztabelle oder die Daten nicht vollständig bekannt sind oder nur nur einzelne Wahrscheinlichkeiten oder Schätzungen. Bayes ist damit das Werkzeug, um rückwärtszuschließen.

Bayes: „Ich weiß, 40 % der Überlebenden waren in der 1. Klasse *(P(First/Yes)=0.40)* und 24 % aller Passagiere waren in der 1. Klasse *(P(First)=0.24)*. 38 % aller Passagiere überlebten (*P(Yes)=0.38)*“ → rechne daraus P(Yes|First).

$$P(\text{Yes} \mid \text{First}) = \frac{P(\text{First} \mid \text{Yes}) \cdot P(\text{Yes})}{P(\text{First})}$$

$$P(\text{Yes} \mid \text{First}) = \frac{0.40 \cdot 0.38}{0.24}$$

$$P(\text{Yes} \mid \text{First}) = \frac{0.152}{0.24} = 0.633$$

Obwohl nur 24 % aller Passagiere in der 1. Klasse waren, stammen 40 % der Überlebenden aus dieser Gruppe. Das erhöht die Überlebenswahrscheinlichkeit für 1.-Klasse-Passagiere auf ≈ 63 %. Also immer erster Klasse fahren, wenn man es sich leisten kann – zumindest auf der Titanic.

Vorwärts vs. Rückwärts bedingte Wahrscheinlichkeit

Titanic

P(Yes | female, first)

- Vorwärtsrichtung: Wahrscheinlichkeit zu überleben, wenn Person eine Frau in 1. Klasse ist.
- Klassifikator-Vorhersage (Ziel gegeben Merkmale).

P(female, first| Yes)

- Rückwärtsrichtung: Anteil der Frauen in 1. Klasse unter den Überlebenden.
- Deskriptive Analyse (Merkmale gegeben Ziel).

Beide teilen denselben Zähler P(female, First, Survived), aber unterscheiden sich im Nenner!

P(Klasse|Merkmale) = [P(Merkmale|Klasse) · P(Klasse)] / P(Merkmale)

Zusammenfassung: Schritte vom Satz von Bayes zum Klassifikator

Grundidee

Aus beobachteten Merkmalen (z. B. Testergebnis, Urlaubsland, Symptome) soll entschieden werden, zu welcher Klasse ein Objekt gehört (z. B. krank oder gesund).

Allgemeine Formel

$P(\text{Klasse}|\text{Merkmale}) = [P(\text{Merkmale}|\text{Klasse}) \cdot P(\text{Klasse})] / P(\text{Merkmale})$

Entscheidungsregel

Da P(Merkmale) für alle Klassen gleich ist, wählt man: Klasse = argmax_c [P(Merkmale|c) · P(c)]

Naive Bayes

Nimmt an, dass die Merkmale unabhängig sind:

$P(\text{Merkmale}|\text{Klasse}) = \prod P(M_i|\text{Klasse})$.

Lernphase

Schätze aus Trainingsdaten: P(Klasse) und $P(\text{Merkmal}_i|\text{Klasse})$.

Vorhersagephase

Berechne für neue Daten P(Klasse|Merkmale) und wähle die Klasse mit der größten Wahrscheinlichkeit.

Beispiel

Krankheitstest mit zusätzlichem Merkmal 'Urlaubsland': Kombination aus Basisrate und Merkmalen ergibt die Klassenzuordnung.

Beispiel Irisdatensatz mit Gaussian Naive Bayes (GNB).

$$P(x_i \mid c_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

```
# iris_nb_gaussian_semicolon.py
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# 1) CSV laden: Semikolon + Dezimal-Komma
df = pd.read_csv("iris_cleaned.csv", sep=";", decimal=",")

# 2) Features/Ziel
X = df[["sepal_length","sepal_width","petal_length","petal_width"]].astype(float)
y = df["species"].astype(str)  # z.B. 'iris-setosa', 'iris-versicolor', 'iris-virginica'

# 3) Train/Test
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42, stratify=y
)

# 4) Gaussian Naive Bayes
clf = GaussianNB()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

# 5) Metriken
print("Accuracy:", round(accuracy_score(y_test, y_pred), 4))
print("\nConfusion matrix:\n", confusion_matrix(y_test, y_pred, labels=sorted(y.unique())))
print("\nClassification report:\n", classification_report(y_test, y_pred))
```

Beispiel Irisdatensatz

```
# === 6) Testbeispiele (manuell) ===
# Format: [sepal_length, sepal_width, petal_length, petal_width]
tests = pd.DataFrame([
    [4.9, 3.0, 1.4, 0.2],   # typische Setosa
    [6.0, 2.9, 4.5, 1.5],   # typische Versicolor
    [6.5, 3.0, 5.5, 2.0],   # typische Virginica
    [5.5, 3.8, 1.7, 0.3],   # kurze Petalen → vermutlich Setosa
], columns=["sepal_length","sepal_width","petal_length","petal_width"])

# === 5) Vorhersage ===
pred = clf.predict(tests)
proba = clf.predict_proba(tests)

# === 6) Ausgabe ===
print("Testdaten:")
print(tests)
print("\nVorhergesagte Klassen:")
for i, p in enumerate(pred):
    probs = {str(cls): round(float(prob), 3) for cls, prob in zip(clf.classes_, proba[i])}
```

Beispiel Irisdatensatz

python iris_nb_gaussian.py

Accuracy: 0.8919

Confusion matrix:

```
[[12  0  0]
 [ 0 12  1]
 [ 0  3  9]]
```

Classification report:

	precision	recall	f1-score	support
iris-setosa	1.00	1.00	1.00	12
iris-versicolor	0.80	0.92	0.86	13
iris-virginica	0.90	0.75	0.82	12
accuracy			0.89	37
macro avg	0.90	0.89	0.89	37
weighted avg	0.90	0.89	0.89	37

Testdaten:

	sepal_length	sepal_width	petal_length	petal_width
0	4.9	3.0	1.4	0.2
1	6.0	2.9	4.5	1.5
2	6.5	3.0	5.5	2.0
3	5.5	3.8	1.7	0.3

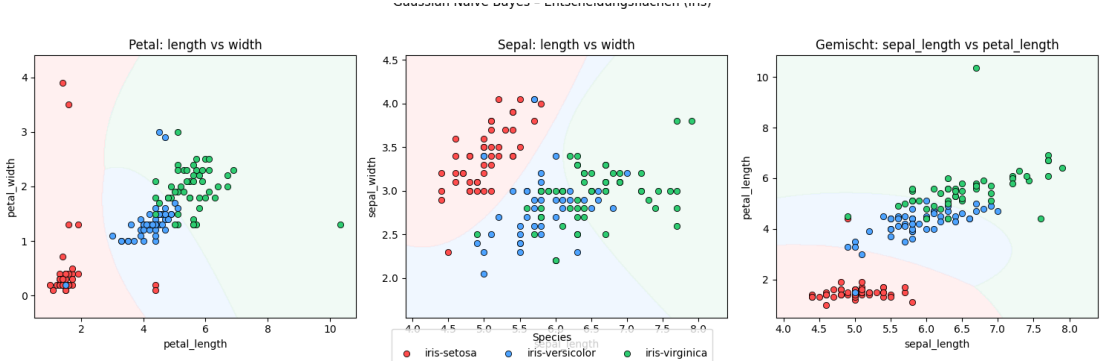
Vorhergesagte Klassen:

Beispiel 1: iris-setosa ⟶ {'iris-setosa': 1.0, 'iris-versicolor': 0.0, 'iris-virginica': 0.0}

Beispiel 2: iris-versicolor ⟶ {'iris-setosa': 0.0, 'iris-versicolor': 0.923, 'iris-virginica': 0.077}

Beispiel 3: iris-virginica ⟶ {'iris-setosa': 0.0, 'iris-versicolor': 0.026, 'iris-virginica': 0.974}

Beispiel 4: iris-setosa ⟶ {'iris-setosa': 1.0, 'iris-versicolor': 0.0, 'iris-virginica': 0.0}



Beispiel 1 (kleine Petalen) → eindeutig *Setosa*.
Beispiel 2 (mittlere Petalen) → *Versicolor*.
Beispiel 3 (lange Petalen) → *Virginica*.
Beispiel 4 → ebenfalls *Setosa*.

Aufgabe: Titanic-Daten

Aufgabe 1:

Die Bayes-Formeln können auch über die Mengenlehre und Venn-Diagramme dargestellt werden.

Erstellen Sie eine Präsentation, die den Zusammenhang Bayes – Mengenlehre / Venn-Diagramme darstellt.

Aufgabe 2:

Ermitteln Sie die Entscheidungs-/Klassifikationsregel für die Merkmalskombination

Yes | First, Female, Age < 30

basierend auf den Titanic Daten.

Aufgabe: Titanic-Daten

Aufgabe 1:

Die Bayes-Formeln können auch über die Mengenlehre und Venn-Diagramme dargestellt werden.

Erstellen Sie eine Präsentation, die den Zusammenhang Bayes – Mengenlehre / Venn-Diagramme darstellt.

1. **Definition der bedingten Wahrscheinlichkeit:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{und} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

2. **Da $A \cap B = B \cap A$, gilt:**

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

3. **Umstellen nach $P(A|B)$:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Aufgabe: Titanic-Daten Lösung

Aufgabe 2:

Ermittle die Entscheidungs-/Klassifikationsregel für die Merkmalskombination

Yes | First, Female, Age < 30

basierend auf den Titanic Daten.

```
== Roh-Kodierungen ==
pclass: pclass
3    491
1    216
2    184
Name: count, dtype: int64
sex   : sex
male      577
female    314
Name: count, dtype: int64
survived: survived
0    549
1    342
Name: count, dtype: int64
age (erste 5): [22.0, 38.0, 26.0, 35.0, 35.0]
```

```
== Nach Normalisierung ==
pclass_n: pclass_n
3    491
1    216
2    184
Name: count, dtype: int64
sex_n   : sex_n
1    577
0    314
Name: count, dtype: int64
survived_n: survived_n
0    549
1    342
Name: count, dtype: int64
age_n nulls: 177

== Ergebnis ==
Gesamtanzahl (First, Female, Age<30): 30
Überlebt (Yes=1): 28
P(Yes | First, Female, Age<30) = 0.933
```

Canine Ovorhoe: Exkurs und Übungen

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Sie machen eine Urlaubsreise in ein exotisches Land. Dabei erfahren Sie, dass in Ihrem Urlaubsland eine tödliche Infektionskrankheit kursiert - "Canine Ovorhoe". Nach Ihrer Rückkehr steuern Sie voller Panik Ihren Hausarzt an.

Er führt mit Ihnen sofort einen Test durch, der positiv ausfällt, also bedauerlicherweise feststellt, dass Sie an der tödlichen Krankheit leiden. Sie beginnen Ihr Testament aufzusetzen.Ihr Hausarzt gibt Ihnen noch folgende Informationen:

- Der Test erkennt in 99% der Fälle die Infektion, nur in einem von hundert Fällen wird die Krankheit fälschlich nicht diagnostiziert.
- Von hundert Nichtinfizierten werden achtundneunzig als gesund erkannt, zwei fälschlicherweise als infiziert.
- Der Test identifiziert also mit 99% die Kranken und mit 98% die Gesunden.
- Außerdem weiß man, dass sich jeder 1000. mit der Krankheit ansteckt.

Mit welcher Wahrscheinlichkeit wird Ihre Frau/Mann in Kürze Witwe/Witwer sein?

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Urlaub…

Sie kehren aus dem Urlaub zurück. Ein Test auf die Krankheit „Canine Ovorhoe“ fällt positiv aus.

Gegebene Wahrscheinlichkeiten aus der Aufgabe:

- Sensitivität (P(Positiv|Krank)) = 99% (99 von 100)
- Spezifität (P(Negativ|Gesund)) = 98% (98 von 100)
- (P(Krank)) = 0,1% (1 von 1000)

Frage: Wie wahrscheinlich ist es, dass Sie wirklich krank sind, wenn der Test positiv ist? Und damit mit welcher Wahrscheinlichkeit wird Ihre Frau/Mann in Kürze Witwe/Witwer sein?

Erläuterung der Formel P(Krank|Positiv)

- P(Krank): Wahrscheinlichkeit, krank zu sein (Prävalenz / a-priori-Wahrscheinlichkeit).
- P(Krank|Positiv): Wahrscheinlichkeit, krank zu sein, wenn der Test positiv ist (gesuchte a-posteriori-Wahrscheinlichkeit).
- P(Positiv|Krank): Wahrscheinlichkeit, dass der Test positiv ist, wenn man krank ist (Sensitivität).
- P(Gesund): Wahrscheinlichkeit, gesund zu sein (1 − P(Krank)).
- P(Positiv|Gesund): Wahrscheinlichkeit, dass der Test positiv ist, obwohl man gesund ist (Falsch-Positiv-Rate = 1 − Spezifität).
- Zähler: [P(Positiv|Krank) · P(Krank)] = Anteil der richtig positiv Getesteten, gewichtet mit der Basisrate.
- Nenner: [P(Positiv|Krank) · P(Krank) + P(Positiv|Gesund) · P(Gesund)] = gesamte Wahrscheinlichkeit für ein positives Testergebnis (richtig + falsch positiv).

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Frage: Wie wahrscheinlich ist es, dass Sie wirklich krank sind, wenn der Canine Ovorhoe Test positiv ist?

$P(Krank|Positiv) = [P(Positiv|Krank) \cdot P(Krank)] / [P(Positiv|Krank) \cdot P(Krank) + P(Positiv|Gesund) \cdot P(Gesund)]$

Mit den Werten:

- $P(Positiv|Krank) = 0,99$
- $P(Positiv|Gesund) = 0,02$
- $P(Krank) = 0,001$
- $P(Gesund) = 0,999$

$\Rightarrow P(Krank|Positiv) = (0,99 \times 0,001) / [(0,99 \times 0,001) + (0,02 \times 0,999)] \approx 0,047$

Nur etwa 4,7 % der positiv Getesteten sind tatsächlich krank!

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Andere Betrachtungsweise

1. Test

	Person	%	Test positiv	Test negativ
Krank	100	99%	99	1
Gesund	100000	2%	2000	98000
Summe	100100		**2099**	**98001**

Verhältnis wirklich Kranker bei einem positiven Test

4,72%

2. Test

	Person	%	Test positiv	Test negativ
Krank	99	99%	98	1
Gesund	2000	2%	40	1960
Summe	2099		**138**	**1961**

Verhältnis wirklich Kranker bei einem positiven Test

71,02%

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Was passiert bei einer besseren Erkennungsrate?

- Der Test erkennt in 99% der Fälle die Infektion, nur in einem von hundert Fällen wird die Krankheit fälschlich nicht diagnostiziert.
- Von hundert Nichtinfizierten werden neunundneunzig als gesund erkannt, eine fälschlicherweise als infiziert.
- Der Test identifiziert also mit 99% die Kranken und mit 99% die Gesunden.
- Außerdem weiß man, dass sich jeder 1000. mit der Krankheit ansteckt.

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Was passiert bei einer besseren Erkennungsrate?

1. Test

	Person	%	Test positiv	Test negativ
Krank	100	99%	99	1
Gesund	100000	1%	1000	99000
Summe	100100		**1099**	**99001**

Verhältnis wirklich Kranker bei einem positiven Test

9,01%

2. Test

	Person	%	Test positiv	Test negativ
Krank	99	99%	98	1
Gesund	1000	2%	20	980
Summe	1099		**118**	**981**

Verhältnis wirklich Kranker bei einem positiven Test

83,05%

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Was passiert bei einer noch besseren Erkennungsrate?

- Der Test erkennt in 99% der Fälle die Infektion, nur in einem von hundert Fällen wird die Krankheit fälschlich nicht diagnostiziert.
- Von 10.000 Nichtinfizierten werden 9.999 als gesund erkannt, eine fälschlicherweise als infiziert.
- Der Test identifiziert also mit 99% die Kranken und mit 99% die Gesunden.
- Außerdem weiß man, dass sich jeder 1000. mit der Krankheit ansteckt.

Reise in ein fremdes Land - Der Fall – Canine Ovorhoe

Was passiert bei einer besseren Erkennungsrate?

1. Test

	Person	%	Test positiv	Test negativ
Krank	100	99,00%	99	1
Gesund	100000	0,01%	10	99990
Summe	100100		**109**	**99991**

Verhältnis wirklich Kranker bei einem positiven Test

90,83%

2. Test

	Person	%	Test positiv	Test negativ
Krank	99	99,00%	98	1
Gesund	10	2,00%	0	10
Summe	109		**98**	**11**

Verhältnis wirklich Kranker bei einem positiven Test

99,80%

Bayes mit zusätzlicher Variable: Urlaubsland

Erweiterung des Krankheitsbeispiels

Verschiedene Urlaubsregionen haben unterschiedliche Infektionsraten.

Länder und angenommene Prävalenzen:

- Land A_Hoch: P(Krank|Land) = 2 %
- Land B_Mittel: P(Krank|Land) = 0,5 %
- Land C_Niedrig: P(Krank|Land) = 0,05 %

Der Test bleibt gleich gut: Sensitivität 99 %, Spezifität 98 %

Bayes-Berechnung je Land

Allgemein für jedes Land L

$$P(\text{Krank}|\text{Pos}, L) = \frac{P(\text{Pos}|\text{Krank}, L) \cdot P(\text{Krank}|L)}{P(\text{Pos}|\text{Krank}, L) \cdot P(\text{Krank}|L) + P(\text{Pos}|\text{Gesund}, L) \cdot P(\text{Gesund}|L)}$$

Mit aus Daten geschätzten Größen (z. B. aus Häufigkeitstabellen je Land):

- P(Krank|Pos,L)
  die Wahrscheinlichkeit, tatsächlich krank zu sein, wenn der Test positiv ist und der Kontext L gilt (z. B. bestimmtes Land oder Labor).
- P(Pos|Krank, L)
  Wahrscheinlichkeit, dass der Test positiv ist, wenn die Person krank ist (→ Sensitivität)
- P(Krank|L)
  Wahrscheinlichkeit, krank zu sein im Land/Umfeld L(→ Prävalenz)
  Das Produkt beschreibt, wie oft kranke Personen in diesem Kontext tatsächlich positiv getestet werden.
- P(Pos|Gesund, L)
  Wahrscheinlichkeit, dass der Test positiv ausfällt, obwohl die getestete Person gesund ist, unter Berücksichtigung des Kontexts LL (z. B. Land, Labor, Altersgruppe).
- P(Gesund | L)
  Wahrscheinlichkeit, dass eine Person gesund ist, gegeben, dass sie sich im Kontext Lbefindet (z. B. in einem bestimmten Land, einer Region, einer Bevölkerungsgruppe oder Altersklasse).

Ergebnisse je Land (Beispielhafte Schätzungen)

Aus Simulation mit 20.000 Personen:

- Land A_Hoch → P(Krank|Pos) ≈ 50 %
  Land B_Mittel → P(Krank|Pos) ≈ 20 %
  Land C_Niedrig → P(Krank|Pos) ≈ 2 %

Je höher die Grundprävalenz (Basisrate), desto höher auch die Wahrscheinlichkeit, dass ein positiver Test wirklich krank bedeutet.

Gesamtbetrachtung und Stratifizierung

Über alle Länder gilt:

$$P(\text{Krank}|\text{Pos}) = \Sigma_L \left[P(\text{Krank}|\text{Pos}, L) \cdot P(L|\text{Pos}) \right]$$

Dabei ist $P(L|\text{Pos})$ der Anteil der Positiven aus Land L.

- Ein Land mit hoher Prävalenz kann überproportional zu positiven Tests beitragen.
- Aggregierte und landweise Betrachtung können stark abweichen (Simpson-Paradoxon).

Bezug zum Bayesschen Klassifikator

Das Merkmal 'Land' wird hier als zusätzlicher Faktor genutzt.

Im Naive Bayes-Klassifikator werden alle Merkmale gemeinsam betrachtet:

$$P(\text{Klasse}|\text{Merkmale}) \propto P(\text{Klasse}) \cdot \prod P(\text{Merkmal}_i|\text{Klasse})$$

Länder-, Alters- oder Verhaltensdaten sind nur verschiedene Merkmale, die probabilistisch verknüpft werden.