Karlsruher Institut für Technologie

**Bachelorarbeit**

# Didaktisch-methodische Entwicklung von computergestütztem Lehr- und Lernmaterial zur Mathematik des Satzes von Bayes am Beispiel der Spam-Klassifizierung


| | |
|---|---|
| Vorgelegt von | Jonathan Kantz |
| Erstprüfer | Prof. Dr. Martin Frank<br>Steinbuch Centre for Computing (SCC)<br>Karlsruher Institut für Technologie (KIT) |
| Zweitprüferin | Dr. Ingrid Lenhardt<br>Fakultät für Mathematik<br>Karlsruher Institut für Technologie (KIT) |
| Koreferentin | Stephanie Hofmann<br>Steinbuch Centre for Computing (SCC)<br>Karlsruher Institut für Technologie (KIT) |


Karlsruhe, den 27. März 2023

.

# Inhaltsverzeichnis

# 1. Einleitung

Im Rahmen dieser Arbeit wurde ein Workshop zu Bayes Spamfiltern entwickelt, durchgeführt und evaluiert. Ziel des Workshops ist es Schüler*innen[1] einen Alltagsbezug, zu bedingten Wahrscheinlichkeiten zu bieten. Außerdem soll den Schülern Mathematische Modellierung und der Satz von Bayes näher gebracht werden.

Ziel dieser Arbeit ist es die Hintergründe der Entwicklung des Workshops darzulegen, die Grundlegenden Konzepte zu erläutern und wichtige Schritte in der Entwicklung des Workshopmaterials zu dokumentieren.

In Kapitel 2 werden Grundkonzepte von CAMMP und der Mathematischen Modellierung erklärt sowie der Bezug zum Bildungsplan Mathematik Baden-Württemberg (2016) hergestellt.

Daraufhin wird in Kapitel 3 der mathematische und technische Hintergrund eines Naiven Bayes Spamfilter dargelegt, um die mathematischen Konzepte, die in dem Workshop benutzt werden, zu erklären. Weiterhin werden wichtige technische Details der Umsetzung und des Datensatzes beschrieben, die im Material des Workshops, das im Anhang zu finden ist, nicht offensichtlich sind.

Im Kapitel 4 geht es um einzelne, konkrete didaktische und methodische Entscheidungen, die während der Entwicklung getroffen wurden und deren Rechtfertigung.

Im Kapitel 5 wird die Struktur einer Durchführung des Workshops beschrieben.

In Kapitel 6 wird eine erste Durchführung beschrieben und evaluiert, um dann in Kapitel 7 Möglichkeiten aufzuzeigen, wie der Workshop noch erweitert werden könnte.

## 1.1. Motivation

Dieser Workshop wurde erstellt, um Schülern der Oberstufe die Mathematik des Bayes-Spamfilters näher zu bringen. Dadurch können sie ein besseres Verständnis für datengetriebene Algorithmen entwickeln und lernen eine Anwendung der Schul-Stoachastik im Alltag kennen.

[1] Nachfolgend werden Schüler*innen unter der Bezeichnung „Schüler“ zusammengefasst. Analog wird mit den Personengruppen Lehrer*innen, Dozent*innen u. a. verfahren.

# 2. Didaktischer Hintergrund

## 2.1. CAMMP

CAMMP (Computational and Mathematical Modeling Program, Computergestütztes Mathematisches Modellierungsprogramm) ist ein Projekt für außerschulische Lernangebote, das sich hauptsächlich an Schüler der Oberstufe richtet. Zur Zeit gibt es einen Standort an der RWTH Aachen. Dort ist CAMMP ein eigenes Schülerlabor. Am KIT Karlsruhe arbeitet CAMMP mit dem Schülerlabor Mathematik zusammen. Gegründet wurde das Projekt 2011 von Prof. Martin Frank, Prof. Ahmed Ismail und Dr. Nicole Faber(vgl. CAMMP, 2019).

Ziel von CAMMP ist es mithilfe Mathematischer Modellierung Schülern zu helfen einen Zusammenhang zwischen der Schulmathematik und dem Alltag herzustellen.

Ein Angebot des CAMMP-Projekts sind die CAMMP-Days. Das ist ein Workshop, der 3-6 Stunden umfasst oder auch im Rahmen von mehreren Doppelstunden durchgeführt werden kann. Im Verlauf eines CAMMP-Days werden die Schüler, mithilfe eines thematischen Beispiels, an die Mathematische Modellierung heran geführt. Dabei werden Themen aus dem Schüleralltag, wie Filmvorschläge bei Netflix, Positionsbestimmung per GSP oder der Google Pagerank genutzt. Diese dienen der Veranschaulichung und bieten den Schülern einen Rahmen, um Mathematische Modellierung verstehen und erleben zu können. Ein Workshop besteht aus mehreren kurzen Vorträgen, die die nötigen Hintergrundinformationen zur Bearbeitung des Workshops vermitteln. Zusätzlich gibt es in jedem Workshop mehrere interaktive Arbeitsblätter. Diese werden von je zwei bis drei Schüler gemeinsam an einem Computer bearbeitet. Das Format, als digitale Arbeitsblätter in der Arbeitsumgebung Jupyter Lab, erlaubt es den Schülern, innerhalb des Arbeitsblattes, mit realen Daten zu arbeiten, zu Programmieren und eine automatische Überprüfung ihrer Antworten vorzunehmen. Die Arbeitsblätter sollen möglichst selbständig bearbeitet werden, wobei gegenseitige Unterstützung, auch über Gruppengrenzen hinweg, gewünscht ist. Zusätzlich ist während der Durchführung eine Lehrkraft oder ein Mitarbeiter von CAMMP anwesend, der bei Fragen helfen kann.

## 2.2. Mathematische Modellierung

Der Modellierungskreislauf bietet sich besonders für Probleme an, bei denen komplexe mathematische Zusammenhänge bei der Lösung helfen können.

Sowohl in Industrie, als auch in der Forschung wird mathematische Modellierung angewendet, um Experimente oder Prototypen im Vorhinein theoretisch zu betrachten. Durch Mathematische Modellierungen werden Material und Arbeitsaufwand gespart. Viele vollständig virtuelle Vorgänge lassen sich gut durch Mathematische Modellierung beschreiben.

Der Modellierungskreislauf ist ein Metamodell der Mathematischen Modellierung. Und kann benutzt werden um den Prozess zu visualisieren und strukturieren.

Eine wesentliche Grundlage eines CAMMP-Workshops ist der Modellierungskreislauf. Die von CAMMP verwendete Version ist eine vereinfachte Form des Modellie-

rungskreislaufs von Blum und Leiß. Sie umfasst nur die vier Schritte vereinfachen, mathematisch beschreiben, berechnen und interpretieren. Diese Form hat sich über die letzten Jahre bei CAMMP bewährt und findet deswegen auch im Workshop Anwendung.

Vom Modellierungskreislauf mit vier Schritten lassen sich Parallelen zu der prozessbezogenen Kompetenz der Modellierung aus Bildungsplan Mathematik Baden-Württemberg (2016) ziehen.

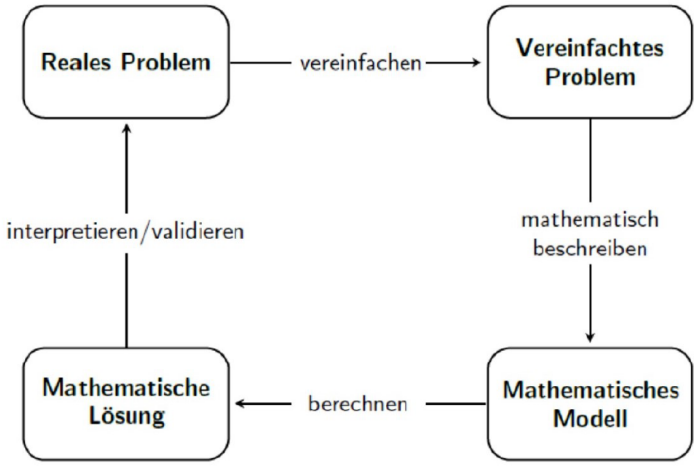



Abbildung 1: Vereinfachter Modellierungskreislauf, angelehnt an Blum und Leiß(vgl. Greefrath et al., 2013, S. 17)

Der Einstieg in den Modellierungskreislauf erfolgt über das reale Problem.

Im ersten Schritt, der Vereinfachung, wird das reale Problem analysiert und auf wichtige Faktoren reduziert. Um die Realität mathematisch abbilden zu können, müssen bestimmte Annahmen getroffen und bestimmte Faktoren ignoriert werden. Dieser Vorgang erfolgt nach der selben Struktur, wie das Äufbereiten und Analysieren der Realsituationïm Bildungsplan.

Der zweite Schritt, die mathematische Beschreibung, verläuft wie das Mathematisieren, wie es im Bildungsplan beschrieben ist. Dabei liegt der Fokus darauf die Daten, mit denen gearbeitet wird, aufzubereiten und die Zusammenhänge in eine Form zu bringen, die der Computer verarbeiten kann.

Der dritte Schritt, die Berechnung, ist im Bildungsplan sehr offen beschrieben. Dabei sollen Hilfsmittel, Algorithmen und Konstruktionen verwendet werden. Bei CAMMP sind die Hilfsmittel der Wahl der Computer und die Programmierung. Es wird ein stärkerer Fokus auf Algorithmen und Automatisierung gelegt, um zu einer mathematischen Lösung zu kommen.

Im vierten Schritt, der Interpretation, geht es um das übertragen der mathematischen Lösung auf das reale Problem. So wird zum Beispiel aus einer Gleichung wieder

ein Mischverhältnis und aus einer Wahrscheinlichkeit eine Schätzung oder Entscheidung. Dieser Schritt ist besonders wichtig, da er offenbart, ob einer der vorangegangenen Schritte fehlerhaft war.

Mögliche Fehler können sein: Eine zu stark vereinfachte Situation, eine ungenaue oder inkorrekte mathematische Beschreibung oder auch ein Fehler in der Implementierung und Durchführung des Lösungsalgorithmus.

Falls ein Fehler gefunden wurde, muss an der entsprechenden Stelle nachgebessert werden. Falls das Ergebnis noch nicht den Ansprüchen genügt, wird der Kreislauf ein weiteres Mal durchlaufen. Das wiederholte Durchlaufen des Kreislaufs findet im Bildungsplan Erwähnung, ist dort aber nur durch den Satz „gegebenenfalls Überlegungen zur Verbesserung der Modellierung anstellen“angedeutet.

# 3. Mathematischer und technischer Hintergrund

## 3.1. Der Satz von Bayes

Der Satz von Bayes wurde im Werk Bayes (1763) posthum zum ersten Mal erwähnt.

Er beschreibt die folgende Formel, mit deren Hilfe sich für zwei Ereignisse A und B mit $P(B) > 0$ aus den Wahrscheinlichkeiten $P(A)$, $P(B)$ und $P(B|A)$ die Wahrscheinlichkeit $P(A|B)$ berechnen lässt.

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{3.1}$$

Der Satz von Bayes findet in verschieden Gebieten Anwendung, in denen es um bedingte Wahrscheinlichkeiten geht. Eines dieser Gebiete ist das medizinische Testen. In diesem Fall wäre $P(A)$ die Wahrscheinlichkeit einer Erkrankung und $P(B)$ die Wahrscheinlichkeit eines Positiven Testergebnisses. $P(B|A)$ wäre somit die Wahrscheinlichkeit, dass der Test anschlägt, wenn der Patient erkrankt ist. Mit Hilfe dieser Daten und dem Satz von Bayes lässt sich dann die Wahrscheinlichkeit $P(A|B)$, mit der ein Patient der ein positives Testergebnis hat erkrankt ist, berechnen.

Der Satz von Bayes ist ebenso hilfreich, um ein Verständnis für Wahrscheinlichkeiten und den Zusammenhang von Wahrscheinlichkeiten zu entwickeln. Wir setzen in das Beispiel von oben konkrete Zahlen ein. Mit der Wahrscheinlichkeit von $P(A) = 0.001$ ist ein Patient erkrankt. Mit der Wahrscheinlichkeit von $P(B|A) = 0.95$ schlägt der Test an. $P(B)$ die Wahrscheinlichkeit für einen positiven Test berechnen wir aus

$$P(A) \cdot P(B|A) \cdot P(\overline{A} \cdot P(B|\overline{A})) \tag{3.2}$$

$P(B|\overline{A}) = 0.1$ und kommen somit auf $P(B) = 0.10085$.

Wenn wir nun den Satz von Bayes anwenden erhalten wir folgenden Formel für $P(A|B)$, die Wahrscheinlichkeit, dass ein Patient wirklich erkrankt ist bei einem positiven Test.

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} = \frac{0.001 \cdot 0.95}{0.10085} = 0,0094 \tag{3.3}$$

Die Wahrscheinlichkeit $P(B|A)$ ist kleiner als 1%, was bei einer Falschnegativwahrscheinlichkeit von 0.05 und einer Falschpositivwahrscheinlichkeit von 0.1 erst einmal unrealistisch erscheint. Hierbei ist zu beachten, dass Tests häufig nicht flächendeckend durchgeführt werden. Meist werden Tests nur durchgeführt wenn es, durch Symptome, konkrete Anhaltspunkte, für eine Erkrankung, gibt. Dadurch verschieben sich die Wahrscheinlichkeiten deutlich.

## 3.2. Der Naive Bayes Klassifikator

Der Naive Bayes Klassifikator, hat die Aufgabe Mails in eine von zwei Klassen, Spam oder Ham, einzuordnen. Dazu wird der Satz von Bayes angewendet. Das besondere des

Naiven Klassifikators ist, dass die Auftretenswahrscheinlichkeit der Wörter in Spam und Ham Mails als bedingt unabhängig angenommen wird. Der genaue Klassifikationsprozess wird in diesem Kapitel erklärt.

## 3.3. Problemstellung und Rahmenbedingungen der Modellierung

In diesem Workshop wird sich vor allem mit der Spam-Erkennung mithilfe von Schlüsselwörtern befasst. Das Konzept ist auch ohne Vorkenntnisse in der Informatik verständlich.

Die Spamerkennung kann man als ein Klassifizierungsprozess in die beiden Klassen Spam und Ham(nicht Spam) sehen.

Um den Klassifizierungsprozess zu vereinfachen, wird die Annahme getroffen, dass die Wahrscheinlichkeit, dass ein Wort in einer E-Mail vorkommt, bedingt unabhängig von den anderen Wörtern der Mail und dem Kontext ist. Obwohl diese Annahme nicht den tatsächlichen Gegebenheiten entspricht, vereinfacht sie die Berechnungen erheblich, ohne das Ergebnis stark zu verfälschen.

### 3.3.1. Der Datensatz

Der Datensatz besteht aus 600 Spam und 600 Ham Mails. Die Mails wurden privat gesammelt und anonymisiert. Die Daten wurden als zwei CSV-Dateien strukturiert, eine für Ham und eine für Spam Mails. Beide Dateien enthalten jeweils eine Liste mit aller Wörter, wobei nur die Wörter aufgenommen wurden, in mindestens 6 Spam bzw. Ham Mails vorkommen. In diesen Listen ist angegeben, in wie vielen Mails ein bestimmtes Wort vorkommt. Falls ein Wort mehrfach in einer Mail vorkommt wird diese Mail trotzdem nur einmal gezählt.

Ausschnitt aus der Tabelle der Ham Mails und Spam Mails:

| Ham | | Spam | |
|---|---|---|---|
| Word | Count | Word | Count |
| den | 307 | den | 332 |
| Jetzt | 16 | Jetzt | 222 |
| € | 11 | € | 171 |
| Hallo | 224 | Hallo | 73 |
| Logo | 8 | Logo | 58 |
| News | 0 | News | 37 |

### 3.3.2. Aufbau der Datenstruktur

Der vorliegende Datensatz besteht zur Hälfte aus Spam Mails. Es gilt somit $P(Spam) = 0,5$. Diesen Wert kann später, je nach Datensatz oder neuen Erfahrungswerten, verändern werden. Für jedes Wort, das im Datensatz erfasst wurde, wird nun die Auftretenswahrscheinlichkeit in den Mails $P(Wort)$ bestimmt. Sie beschreibt die Wahr-

scheinlichkeit, dass eine Mail dieses Wort enthält und kann über die relative Häufigkeit geschätzt werden:

$$P(Wort) = \frac{Anzahl\,der\,Mails\,mit\,Wort}{Gesamtanzahl\,der\,Mails} \tag{3.4}$$

Zusätzlich wird für jedes Wort die Auftretenswahrscheinlichkeit in einer Spam Mail $P(Wort|Spam)$ berechnen.

$$P(Wort|Spam) = \frac{Anzahl\,der\,Spam\,Mails\,mit\,Wort}{Gesamtanzahl\,der\,Spam\,Mails} \tag{3.5}$$

Dieser Schritt wird in Teil 2 auf B.1 Arbeitsblatt 1 für drei Wörter manuell durchgeführt. In Teil 3 des Arbeitsblattss wird der Prozess automatisiert.

### 3.3.3. Umkehrung bedingter Wahrscheinlichkeiten

Die Klassfikation kann als zweistufiges Zufallsexperiment betrachtet werden. Deswegen lässt sich die Klassfikation der Mails in zwei Baumdiagrammen (s.Abb. 2) darstellen.

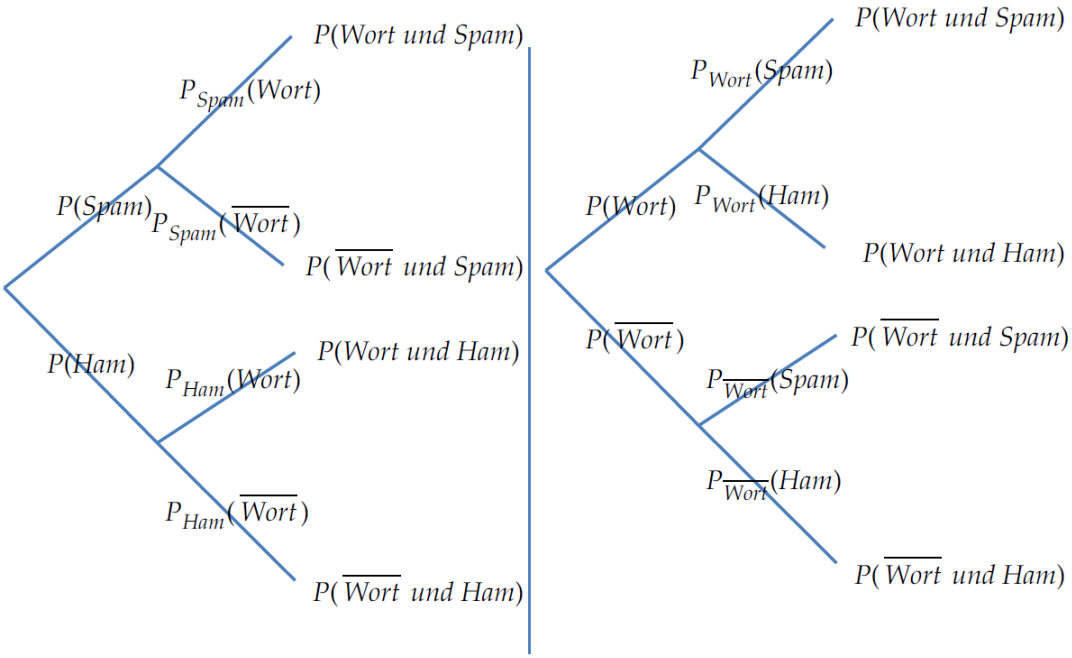



Abbildung 2: zwei Baumdiagramme

Da beide Baumdiagramme zu den selben Wahrscheinlichkeiten führen, kann man die Baumdiagramme zu einem doppelten Baumdiagramm (s.Abb. 3)verknüpfen.

Gesucht ist $P(Spam|Wort)$, die Wahrscheinlichkeit, dass eine Mail eine Spam Mail ist, wenn sie ein bestimmtes Wort enthält. Aus 3.3.2 sind die Wahrscheinlichkeiten

$P(Spam)$, $P(Wort)$ und $P(Wort|Spam)$ bekannt. Mit dem Satz von Bayes 3.1 lässt sich folgenden Formel aufstellen

$$P(Spam|Wort) = \frac{P(Spam) \cdot P(Wort|Spam)}{P(Wort)} \tag{3.6}$$

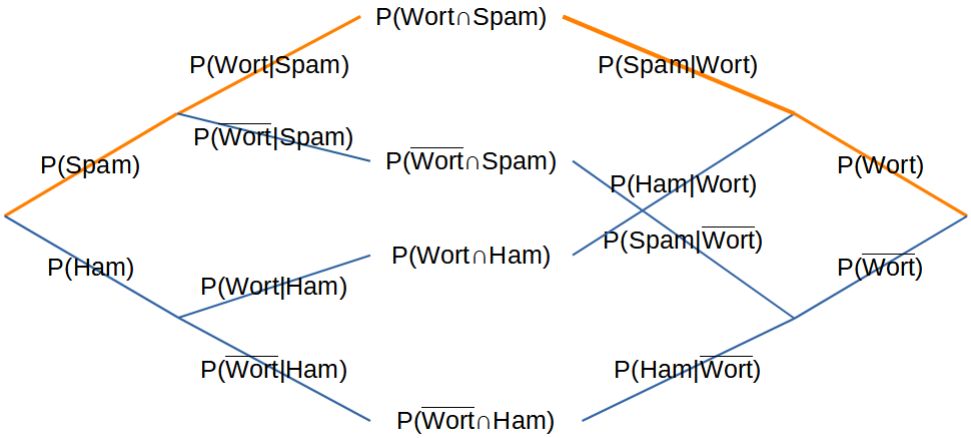



Abbildung 3: Doppeltes Baumdiagramm

Dieser Zusammenhang lässt sich anschaulich anhand des Doppelten Baumdiagramms herleiten (s. Abb. 3). Dazu wird der orangene Pfad betrachtet. Über die Pfade von links und von rechts ergeben sich, mit der ersten Pfadregel, zwei Möglichkeiten die Wahrscheinlichkeit $P(Spam \cap Wort)$ zu berechnen. Die erste Pfadregel besagt, dass die Wahrscheinlichkeit am Ende eines Pfades das Produkt der Wahrscheinlichkeit entlang des Pfades ist.

von links:

$$P(Spam \cap Wort) = P(Spam) \cdot P(Wort|Spam) \tag{3.7}$$

von rechts:

$$P(Spam \cap Wort) = P(Wort) \cdot P(Spam|Wort) \tag{3.8}$$

Durch Gleichsetzen der Terme auf der rechten Seite ergibt sich:

$$P(Spam) \cdot P(Wort|Spam) = P(Wort) \cdot P(Spam|Wort) \tag{3.9}$$

Teilen mit $P(Wort)$ ergibt die Formel, in der Form des Satzes von Bayes, wie er auf das Spamfilter Problem angewendet wird.

$$P(Spam|Wort) = \frac{P(Spam) \cdot P(Wort|Spam)}{P(Wort)} \tag{3.10}$$

Mit der Wahrscheinlichkeit $P(Spam|Wort)$ lässt sich eine Abschätzung treffen, ob eine Mail Spam oder Ham ist. Darauf basierend kann die Mail als Spam oder Ham klassifiziert werden.

Diesen Zusammenhang erarbeiten die Schüler auf B.2 Arbeitsblatt 2 und erstellen mit diesem Wissen einen ersten Klassifikator.

### 3.3.4. Anwendung auf mehrere Wörter

Die meisten Mails bestehen nicht nur aus einem Wort. Eine mögliche Modellverbesserungendie in diesem Workshop angewendet wird, ist die Erweiterung auf mehrere Wörter.

Wie in 3.3.3 beschrieben, wird die Annahme getroffen, dass die Wahrscheinlichkeit für das Auftreten eines Wortes bedingt unabhängig von den anderen Worten in einer Mail ist. Dies erlaubt nicht, die einzelnen $P(Spam|WortX)$ mit $X = 1, 2, \dots$ zu multiplizieren um $P(Spam|Wort1 \cap Wort2 \cap \dots)$ zu berechnen. Nach Anwendung des Satzes von Bayes, um die Wahrscheinlichkeit $P(Spam|Wort1 \cap Wort2 \cap \dots)$ zu berechnen, ergibt sich:

$$P(Spam|Wort1 \cap Wort2 \cap \dots) = \frac{P(Spam) \cdot P(Wort1 \cap Wort2 \cap \dots|Spam)}{P(Wort1 \cap Wort2 \cap \dots)} \tag{3.11}$$

Es lässt sich nun ein Wahrscheinlichkeitsverhältnis aus $P(Spam|Wort1 \cap Wort2 \cap \dots)$ und $P(Ham|Wort1 \cap Wort2 \cap \dots)$ bilden, wodurch die Wahrscheinlichkeit $P(Wort1 \cap Wort2 \cap \dots)$ eliminiert wird.

Da die Unterscheidung in Spam und Ham diese Wahrscheinlichkeit nicht beeinflusst, lässt sich schließen, dass sie auch keinen Einfluss auf die Entscheidung für Spam oder Ham hat. Daraus folgt die Formel:

$$\frac{P(Spam|Wort1 \cap Wort2 \cap \dots)}{P(Ham|Wort1 \cap Wort2 \cap \dots)} = \frac{P(Spam) \cdot P(Wort1 \cap Wort2 \cap \dots|Spam)}{P(Ham) \cdot P(Wort1 \cap Wort2 \cap \dots|Ham)} \tag{3.12}$$

Die Wahrscheinlichkeiten $P(Wort1 \cap Wort2 \cap \dots|Spam)$ und $P(Wort1 \cap Wort2 \cap \dots|Ham)$ lassen sich mithilfe der bedingten Unabhängigkeit zu

$$\frac{P(Spam|Wort1 \cap Wort2 \cap \dots)}{P(Ham|Wort1 \cap Wort2 \cap \dots)} = \frac{P(Spam) \cdot P(Wort1|Spam) \cdot P(Wort2|Spam) \dots}{P(Ham) \cdot P(Wort1|Ham) \cdot P(Wort2|Ham) \dots} \tag{3.13}$$

auflösen. Diese Formel besteht nur aus Wahrscheinlichkeiten die sich aus den Daten direkt ablesen lassen.

Diese Formel leiten die Schüler auf B.3 Arbeitsblatt 3 her und erstellen somit einen Klassifikator für mehrere Wörter.

### 3.3.5. Bedingte und allgemeine Unabhängigkeit

Während der Erstellung des Workshop kam die Frage auf warum $P(Wort1 \cap Wort2 \cap \dots)$ nicht durch $P(Wort1) \cdot P(Wort2) \cdot \dots$ berechnet werden kann. Im Folgenden wird der Zusammenhang erläutert. Auf den Arbetisblättern wurde dieser Inhalt auf das Zusatsblatt B.4 ausgelagert.

In 3.3.4 lässt sich die Unabhängigkeit der Wörter nicht auf die Wahrscheinlichkeit $P(Wort1 \cap Wort2 \cap \dots)$ anwenden. Dies hängt damit zusammen, wie die Daten gewonnen wurden und wie die Wahrscheinlichkeiten berechnet wurden.

Zur Veranschaulichung kann folgendes fiktives Beispiel aus 10 Spam Mails und 10 Ham Mails betrachtet werden:

| Wort | Stadt | Haus |
|---|---|---|
| Anz. Spam Mails | 10 | 9 |
| Anz. Ham Mails | 5 | 1 |
| $P(Wort\|Spam)$ | 1 | 0.9 |
| $P(Wort\|Ham)$ | 0.5 | 0.1 |
| $P(Wort)$ | 0.75 | 0.5 |

Wenn wir die Auftretenswahrscheinlichkeiten für Haus und Stadt als unabhängig betrachten, gilt:

$$P(Haus\ und\ Stadt) = 0.75 \cdot 0.5 = 0.375 \tag{3.14}$$

Wenn wir bedingte Unabhängigkeit annehmen, gilt mit dem Satz der Totalen Wahrscheinlichkeit:

$$\begin{aligned} P(Haus\ und\ Stadt) = P(Spam) \cdot P(Haus\ und\ Stadt|Spam) \\ +P(Ham) \cdot P(Haus\ und\ Stadt|Ham) \end{aligned} \tag{3.15}$$

$$\begin{aligned} P(Haus\ und\ Stadt) = P(Spam) \cdot P(Haus|Spam) \cdot P(Stadt|Spam) \\ +P(Ham) \cdot P(Haus|Ham) \cdot P(Stadt|Ham) \end{aligned} \tag{3.16}$$

$$P(Haus\ und\ Stadt) = 0.5 \cdot 0.9 \cdot 1 + 0.5 \cdot 0.5 \cdot 0.1 = 0.475 \tag{3.17}$$

Wir sehen, dass sich die beiden Werte für $P(Haus\ und\ Stadt)$ unterscheiden. Entweder muss 0.375 oder 0.475 korrekt sein. Aus der Beschreibung der Situation wissen wir, dass es entweder 9 oder 10 Mails geben kann, die die Worte „Haus“ und „Stadt“

enthalten. Sicher sind es 9 Spam Mails, da jede Spam Mail das Wort „Stadt“ enthält und 9 dieser Mails auch das Wort „Haus“. Für die Ham Mails können wir keine sichere Aussage treffen. Die Wahrscheinlichkeit, dass die Mail, die das Wort „Haus“ enthält, auch „Stadt“ enthält ist 0.5. Für eine genaue Schätzung von $P(Haus\ und\ Stadt)$ brauchen wir also den Mittelwert zwischen 10/20 und 9/20. Dieser Mittelwert ist 0.475. Unser Modell wird folglich durch bedingte und nicht allgemeine Unabhängigkeit beschrieben. Das können wir auch intuitiv erkennen, da die Werte in unserer Tabelle in Abhängigkeit von Spam und Ham angegeben sind.

# 4. Didaktisch-methodisches Konzept

## 4.1. Ziel des Workshops

Ziel des Workshops ist es, dass die Schüler Mathematische Modellierung kennen lernen und am Beispiel des Bayes-Spamfilter anwenden. Die Schüler sollen selbst die einzelnen Schritte der Mathematischen Modellierung durchführen und sich dabei das Konzept des Satzes von Bayes herleiten und auf ein alltagsnahes Beispiel anwenden.

## 4.2. Jupyter Lab als digitale Lernumgebung

Bei CAMMP wird Jupyter Lab als digitale Lernumgebung genutzt. Über die Webseite CAMMP-Workshops können die Schüler auf den Workshopserver von CAMMP zugreifen. Dazu erstellen sie sich bei der ersten Anmeldung einen Accout, den sie danach weiter nutzen. Jeder Account hat einen eigenen Speicher auf dem Server. In diesem Speicher können Workshops geladen werden und der Fortschritt auf den Arbeitsblätter gespeichert werden. Durch eine browser- und serverbasierte Arbeitsumgebung, sind keine lokalen Installationen nötig. So ist es möglich mit fast jedem Gerät einen CAMMP-Workshop zu bearbeiten, solange eine Internetverbindung sichergestellt ist. Jupyter Lab kann Kernals für verschiedene Programmiersprachen anbieten. In diesem Workshop wird Julia genutzen. Ein Arbeitsblatt ist aus mehreren Zellen aufgebaut, jede Zelle hat dabei einen von drei Typen.

Der erste Typ ist ein unformatiertes freies Textfeld, Raw genannt.

Der zweite Typ ist eine Textumgebung, genannt Markdown, in der unter anderem Überschriften, Unterstreichungen und LaTeXformeln genutzt werden könne.

Der dritte Typ ist einen Codezelle. In Codezellen kann Code der ausgewählten Programmiersprache geschrieben werden. Diese Codefelder werden direkt für die Programmierung oder als Antwortfelder genutzt. Zu jedem Arbeitsblatt gehört eine check-Datei. In diesen check-Dateien sind Funktionen implementiert, die Antworten überprüfen und Rückmeldung geben. Außerdem sind Funktionen implementiert, die Programme aufrufen und enthalten Defintionen von Variabeln, welche im Workshops genutzt werden. Die Programmiung, in den Workshops von CAMMP, hat meist eine Lückentext-Struktur, sodass keine besonderen Programmierkentnisse erforderlich sind und Syntax-Fehler möglichst vermieden werden.

## 4.3. Hilfestellungen

Ein Ziel bei der Gestaltung des Workshop ist es, dass die Schüler möglichst eigenständig die Arbeitsblätter bearbeiten können. Wie in Kapitel 6 zu lesen gab es bei der ersten Durchführung Probleme mit den Notationen und dem Schreiben von Code. Um diesen Problemen entgegen zu wirken, wurde eine B.6 Variablen und Befehlsübersicht erstellt, welche die Schüler über Verlinkungen aufrufen können. Sie dient als Vorlage aus der Befehle und Variablen kopiert werden können. Durch den Kopiervorgang werden Tippfehler minimiert und Probleme beim Finden der korrekten Notation eliminiert. Damit

die Schüler sich auf die Mathematik fokussieren können, wird die Programmierung als Baukastensystem zur Verfügung gestellt.

## 4.4. Wahl des Datensatzes

Es gab zum Zeitpunkt der Erstellung des Workshops keinen öffentlichen deutschen Datensatz aus Spam und Ham Mails. Im Englischen lagen dagegen mehrere vor. Trotzdem wurde sich dagegen entschieden, einen englischen Workshop zu erstellen oder mit englischen Daten zu arbeiten. Dem liegen mehrere Faktoren zu Grunde.

Der erste Faktor ist, dass die Schüler selbst viele deutsche Spam Mails kennen und auch typische Begriffe, die in solchen vorkommen. Im Englischen wäre dies in diesem Umfang nicht der Fall.

Der zweite Faktor ist, die Sprachbarriere beim Erkennen von Spam und Ham Mails.

Der letzte Faktor hängt mit der Idee hinter der Zusatzaufgabe auf dem B.3 Arbeitsblatt 3 zusammen. Das Formulieren einer Spammail ergibt nur in der Muttersprache Sinn, alles andere wäre durch beschränktes Vokabular und Kenntnis der Sprache so sehr verzerrt, dass es nur schwer möglich wäre damit einen Spamfilter ernsthaft zu testen.

Auf Grund dieser Faktoren wurde für diesen Workshop ein eigener Datensatz angelegt. Dieser Datensatz wurde aus einer Sammlung von E-Mails zusammen gestellt und anonymisiert. Dieser Datensatz wurde beschränkt auf 1200 E-Mails, um die Ladezeiten der Arbeitsblätter und bei der Ausführung mehrerer Befehle, in einem angemessenen Rahmen zu halten. Eine Erweiterung oder Modifizierung des Datensatzes wird in Kapitel 7 diskutiert.

## 4.5. Doppeltes Baumdiagramm

Der Satz von Bayes ist für die meisten Schüler nicht intuitiv erfassbar. Die Veröffentlichung Wassner et al. (2002) möchte den Satz von Bayes in eine Form bringen, sodass er leichter verstanden werden kann. Wichtig dabei ist, eine Darstellung zu finden, mit der Schüler die Zusammenhänge anschaulich erfassen können. Dazu wurde ein Doppeltes Baumdiagramm genutzt. Im Gegensatz zur Veröffentlichung wurde in der Umsetzung in diesem Workshop das Doppelte Baumdiagramm (s.Abb. 3) mit Wahrscheinlichkeiten und nicht mit absoluten Häufigkeiten verwendet. Baumdiagramme, die erste Pfadregel und das Umstellen von Gleichungen ist den Schülern ab der 10. Klasse bekannt. Davon ausgehend können sie auf den Satz von Bayes schließen. Auf die gleiche Art und Weise werden Schüler, auf B.2 Arbeitsblatt 2, anhand des Doppelten Baumdiagramms, Schritt für Schritt an den Satz von Bayes herangeführt.

## 4.6. Von manuellem zu automatisiertem Arbeiten

Durch die Digitalisierung eröffnen sich viele Möglichkeiten der Automatisierung. Heute ist es zentrale Praxis alles zu automatisieren, von Planungen, Produktionen, Arbeiten

u.v.m.. Auch der hier entwickelte Workshop zielt darauf ab, die manuell ausgeführte Tätigkeit des Aussortieren von Spam Mails zu automatisieren. Das gleiche Prinzip wird in 3.3.2 verwendet, um die Daten aufzubereiten. Im Verlauf des B.1 Arbeitsblatt 1 bestimmen die Schüler zunächst manuell die Auftretenswahrscheinlichkeiten für drei Wörter, um die Struktur der Daten zu verstehen und danach den Prozess zu automatisieren. Die Schüler können so, erst intuitiv an einem konkreten Beispiel, Prinzipien erkennen und erlernen. Das Problem wird abstrahiert und in Programmcode implementieren, um dann die Arbeit durch einen Computer ausführen zu lassen.

Das gleiche Prinzip wird im B.2 Arbeitsblatt 2 angewendet. In Teil 2 des Arbeitsblattes 2 wird der Satz von Bayes zuerst auf ein einzelnes konkretes Wort anwendet. Das aus dieser Anwendung erlernte Muster wird genutzt, um eine Automatisierung zu erstellen, die für alle Wörter die nötigen Berechnungen durchführt.

## 4.7. Anknüpfung der Workshopinhalte an den Bildungsplan

Auch der Inhalt des Workshops orientiert sich an der „Leitidee Daten und Zufall“für die Klassenstufe 9/10 aus dem Bildungsplan Mathematik Baden-Württemberg (2016). Das Thema bedingten Wahrscheinlichkeiten und die stochastische Unabhängigkeit sind sowohl im Workshop, als auch im Bildungsplan relevant. Der Satz von Bayes kommt nicht explizit im Bildungsplan vor. Er bietet eine Möglichkeit, zur Vertiefung von bedingten Wahrscheinlichkeiten. Die in der Schule erworbenen Kompetenzen, aus der Einheit, zum Thema Daten und Zufall, für die Klassenstufe 9/10, können im Workshop vertieft und auf ein konkretes alltagsnahes Beispiel angewendet werden. Da die Zielgruppe des Workshops Schüler der 10. Klasse und der Oberstufe sind, wurde im Workshop die in der Schule übliche Notation verwendet. Auf die Universitätsnotation wurde verzichtet.

## 4.8. Binnendifferenzierung durch Zusatzaufgaben

CAMMP-Workshops stellen für die meisten Schüler eine neue Herausforderungen dar, da viele wenig bis keine Erfahung mit Programmierung und dem Aufbau der interaktiven Arbeitsblätter eines CAMMP-Days haben. Je nach Vorkenntnissen kommen die Schüler unterschiedlich gut mit dem Material klar. Deswegen ist es besonders wichtig, dass Binnendifferenzierung stattfindet. Dies gibt den Schülern, die gut mit dem Material zurecht kommen, eine Möglichkeit ihre Fähigkeiten anzuwenden. Schüler, die größere Schwierigkeiten haben, werden trotzdem nicht abgehängt.

Um das zu gewährleisten hat jedes Arbeitsblatt eine Zusatzaufgabe. Diese Zusatzaufgaben sind so aufgebaut, dass das zusätzliche Wissen nicht besonders wichtig ist für den weiteren Verlauf des Workshops, deswegen können sie ohne Probleme übersprungen werden. Die Zusatzaufgaben fordern zusätzlich tieferes Verständnis der Zusammenhänge heraus. Zum Beispiel, brauchen die Schüler für die Zusatzaufgabe auf B.1 Arbeitsblatt 1 ein gutes Verständnis davon, was die Wahrscheinlichkeiten, die sich berechnet haben, bedeuten und wie man sie in Beziehung stellen kann. Auf B.2 Arbeitsblatt 2 müssen sie eigenständig recherchieren und Zusammenhänge herstellen. Diese Aufgaben

sind somit besondere Herausforderungen für Schüler, die gut mit dem Material klar kommen, trotzdem verpassen die anderen Schülern nichts.

# 5. Struktur des Workshops

Der Workshop richtet sich an Schüler der 10. Klasse und der Oberstufe.

Notwendige Vorkenntnisse sind bedingte Wahrscheinlichkeiten und der Begriff der stochastischen Unabhängigkeit.

Geplant ist der Workshop für drei Zeitstunden.

Falls der Workshop in zwei Doppelstunden durchgeführt wird, sollte der Teil 1 von B.3 Arbeitsblatt 3 zu Wahrscheinlichkeitsverhältnissen übersprungen werden und kurz von dem Mitarbeiter oder dem Lehrer zusammen gefasst werden. Um zu verhindern, dass der Workshop aus Zeitmangel nicht beendet werden kann.

Falls der Workshop in zwei Doppelstunde durchgeführt wird, endet die erst Doppelstunde mit der Sicherung 1 und die Zweite beginnt mit einer Wiederholung Anhand der Sicherung 1. Die Zeit für die Erarbeitung des B.3 Arbeitsblatt 3 wird auf 20 Minuten verkürzt.

Die Tabelle zeigt den Stundenverlaufsplan des Workshops.

Im Anhang liegen die Musterlösungen vor. In den Schüler Versionen sind nur die Codefelder verändert. An jeder Stellen, an der ein Kommentar, in der Form

„#= NaN =#“

steht, wird die Lösung durch NaN ersetzt. Im Fall von B.3 Arbeitsblatt 3 Teil 2 b) werden zusätzlich auch die Wahrscheinlichkeiten angegeben.

| **Phase** | **Inhalt** | **Schulbezug** | **weitere math. Inhalte** | **Medien/ Materialien** | **Zeit (Min.)** |
|---|---|---|---|---|---|
| Begrüßung | Begrüßung, Einführung in die Problemstellung, Einführung in die mathematische Modellierung | | Modellierungskreislauf | Folien Einführung, Notizen Einführung | 20 |
| Erarbeitung AB 1 | Aufbau einer Datenstruktur, Umwandeln von Häufigkeiten in Wahrscheinlichkeiten | Wahrscheinlichkeiten basierend auf relativen Häufigkeiten, Bedingte Wahrscheinlichkeiten | | AB1-SuS | 45 |
| Sicherung 1 | Besprechung Blatt 1, Einführung Doppeltes Baumdiagramm | Baumdiagramme | Verknüpfung zweier Baumdiagramme | Folien Zwischenvortrag 1 Notizen Zwischenvortrag 1 | 15 |
| Erarbeitung AB 2 | Herleiten des Satzes von Bayes, Erstellung eines ersten Klassifikators | Gleichungen umstellen | Satz von Bayes, durch 0 Teilen | AB2-SuS | 40 |
| Sicherung 2 | Besprechung Blatt 2, Modellverbesserungen | | | Folien Zwischenvortrag 2 Notizen Zwischenvortrag 2 | 10 |

| | | | | | |
|---|---|---|---|---|---|
| Erarbeitung AB 3 | Einführung des Wahrscheinlichkeitsverhältnis, Erstellen und Testen eines Klassifikators für mehrere Wörter | Gleichungen umstellen, unabhängige Wahrscheinlichkeiten | Wahrscheinlichkeitsverhältnisse, Bedingte Unabhängigkeit (auf einem Ergänzungsblatt) | AB3-SuS | 40 |
| Abschlusspräsentation | Zusammenfassung der Ergebnisse Evaluation Verabschiedung | | | Folien Abschluss, Notizen Abschluss | 10 |

# 6. Erprobung und Evaluation

Der Workshop wurde zur Erprobung in zwei Doppelstunden am Markgrafen-Gymnasium in Durlach durchgeführt. Die erste Doppelstunde war am 31.01.2023 die zweite eine Woche später am 07.02.2023. Die Schülergruppe war ein gemischte 10. Klasse, die momentan Stochasik nicht als Thema hatten. Die Klasse bestand aus 19 Schülern.

## 6.1. Ablauf

### 6.1.1. Erste Doppelstunde

- Zu Beginn der ersten Doppelstunde wurden die Begriffe bedingte Wahrscheinlichkeit und Schnittwahrscheinlichkeit und die dazugehörigen Notationen eingeführt. In dieser Version des Workshops wurden noch die Notationen $P(A|B)$, $P(A \cap B)$ im Text und P_A_B, P_A_schnitt_B im Code verwendet.

- In CAMMP-Days wird das Thema Modellierung meist in einen eigenen Vortrag eingeführt. Für die Durchführung in zwei Doppelstunde wurde aus Zeitgründen auf diesen verzichtet, da alle für den Workshop relevanten Modellierungsaspekte im Workshop aufgegriffen werden.

- In der A.1 Einstiegspräsentation wurde das Thema Spamfilter, und der Modellierungskreislauf eingeführt. Zusätzlich wurde der Monty-Python-Spam-Sketch genutzt, um die Herkunft des Begriffes Spam zu erklären.

- In der Erarbeitungsphase des B.1 Arbeitsblattes 1 wurden fünf Beispiele für Spam Mails betrachtet. Aus diesen suchten die Schüler fünf Schlüsselwörter heraus, für die sie mithilfe des Datensatzes die Auftretenswahrscheinlichkeiten bestimmten. Diesen Prozess automatisierten sie man Ende des Arbeitsblattes.

  Die Zweiergruppen benötigten 5 Minuten länger als die geplanten 45 Minuten.

- Zum Abschluss der ersten Doppelstunde wurden die Ergebnisse aus dem B.1 Arbeitsblatt 1 gemeinsam besprochen. Am Ende dieser Doppelstunde wurde in der Präsentation A.2 das Doppelte Baumdiagramm eingeführt, das eine Möglichkeit darstellt, sich dem Prinzip des Satzes von Bayes zu nähern und ihn zu erfassen.

### 6.1.2. Zweite Doppelstunde

- Die zweite Doppelstunde begann mit einer Zusammenfassung der wesentlichen Inhalte der ersten Doppelstunde. Insbesondere das Doppelte Baumdiagramm wurde noch einmal erklärt.

- Im B.2 Arbeitsblatt 2 leiteten sich die Schüler den Satz von Bayes anhand des Doppelten Baumdiagramms her und wendeten ihn auf das Beispiel aus der ersten Einheit an. Dabei erstellten sie einen ersten Klassifikator.

Mehrere Schüler wurden nicht in den vorgegeben 40 Minuten mit dem Arbeitsblatt 2 fertig. Um den Rest des Workshops, zumindest in Teilen, zu erproben wurde trotzdem nach dem Zeitplan fortgefahren.

- Im Anschluss wurde die Herleitung des Satz von Bayes in der dritten Präsentation A.3 besprochen. Zum Abschluss der Präsentation wurde gemeinsam mit den Schülern Probleme des Modells besprochen und Verbesserungsmöglichkeiten gesammelt.

- Die Schüler bearbeiteten das B.3 Arbeitsblatt 3, auf dem sie das Wahrscheinlichkeitsverhältnis kennenlernten und der erstellte Klassifikator ausführlich getestet werden soll. In der Erprobung sind die Schüler aus Zeitgründen nur teilweise bis zur Testung des Klassifikators mit Testmails gekommen.

- In der Abschlusspräsentation A.4 wurden die Inhalte des Workshops noch einmal kurz zusammengefasst.

## 6.2. Erkenntnisse aus der Durchführung

Das entwickelte Material ist auf drei Zeitstunden ausgelegt. Die für die erste Erprobung zu Verfügung stehenden zwei Doppelsunden im Abstand von einer Woche mit festgelegten Pausenzeiten bilden einen recht starren Rahmen.

Die Notation in den Codefeldern hat die Schüler sehr gefordert und viel Zeit gekostet. Das Verständnis der Notation und das Beheben syntaktischer Fehler war sehr aufwändig. In der durchgeführten Version existierte die Variablen- und Befehls-Übersicht B.6 noch nicht. Diese Erklärungen gab es nur in exemplarischer Form eingebettet in die Aufgabenstellungen.

Der Teil 2: Den Wörtern Wahrscheinlichkeiten zuordnen auf B.1 Arbeitsblatt 1 hat sich als etwas langwierig und repetitiv dargestellt. Deswegen wurde für die jetzt vorliegende Version die Anzahl der zu untersuchenden Wörter von fünf auf drei reduziert.

Vor allem Schüler, die davon überzeugt waren, dass sie wissen was sie tun, haben oft nicht die ganze Aufgabenstellung gelesen. Dadurch brauchten sie letztendlich länger als Schüler, die sich selbst als schlechter einschätzten haben und deswegen die Aufgabenstellung genauer gelesen haben.

Es gab teilweise kleine Probleme in der technischen Umsetzung einzelner Aufgaben auf den Arbeitsblättern. Unter anderem war die Überprüfung des Klassifikators auf dem Arbeitsblatt 1 noch fehlerhaft.

Die Schüler haben insgesamt sehr motiviert gearbeitet, was die Durchführung sehr angenehm gestaltet hat.

## 6.3. Rückmeldung der Lehrperson und der Koreferentin

Die Lehrerin der Klasse, Dr. Maren Hattebuhr, ist eine ehemalige Mitarbeiterin bei CAMMP. Sie hat vergleichbare Workshops selbst schon entwickelt und durchgeführt

und konnte sehr fundierte Rückmeldung geben. Folgenden Rückmeldungen der Lehrerin und Koreferentin wurden umgesetz.

Die Mails B.5, die während der Bearbeitung des ersten Arbeitsblattes betrachtet werden, sind nicht Teil des Trainingsdatensatzes. Die Abfrage in Teil 2 des ersten Arbeitsblatts gab somit für Worte, die in den Beispiel-Spam-Mail vorkamen, aus das sie 0 mal im Datensatz vorkommen. Das hat zu Verwirrung geführt. Im Kapitel 7 wird eine mögliche Lösung für das Problem skizziert.

Die Relevanz des Videosketchs und dessen Einbindung in die Einführung wurde in Frage gestellt. Mit einem etwas klareren Konzept, wie genau der Videosketchs zu benutzen ist sollte das aber kein Problem mehr sein.

Im Teil 2 auf dem B.1 Arbeitsblatt 1 gab es mehrere Stellen die grammatikalisch inkorrekt waren und verwirrende Nebensätze enthielten

Eine bessere Einführung in die Arbeitsumgebung von Jupyter Lab hätte den Schüler geholfen.

Es gab gemeinsame Überlegungen die Einführung des Doppelten Baumdiagramms interaktiver zu gestalten und die Schüler mehr in den Entstehungsprozess miteinzubeziehen.

Die Notationen für Befehle sollten vereinfach werden.

## 6.4. Rückmeldung der Schüler

Nach der Durchführung wurde der Workshop mithilfe des Standard-Evaluationsbogens C für einen CAMMP-Day evaluiert. Im Folgenden sind interessante Ergebnisse der Evaluation aufgeführt.

Die Rückmeldungen der Schüler waren insgesamt durchschnittlich bis positiv.

Besonders schwierig war der Umgang mit Jupyter Lab, der digitalen Arbeitsumgebung, die nicht ausreichend eingeführt wurde.

Besonders positiv wurde die Betreuung und aktive Hilfe währende des Workshops empfunden.

Da es keinen Modellierungsvortrag gab und Modellierung an sich nicht besonders Thema der Durchführung war, war auch das Interesse daran und das Verständnis dafür wie erwartet niedrig.

Die Schüler empfanden den Workshop mathematisch als nicht besonders anspruchsvoll, die Herausforderungen lagen vor allem in der technischen Umsetzung.

Der Workshop wurde mit einer Durschnittsnote von 2,55 von den Schülern, für eine Erstdurchführung, gut bewertet. Die Betreuung mit einer Durschnittsnote von 1,44 als sehr gut.

Das weitere Interesse am Thema und der Mathematischen Modellierung konnte leider nicht im besonderen geweckt werden. Fragen die darauf abzielten wurden unterdurchschnittlich bewertet.

Ein weiterer Wunsch war es, dass die Tabelle in Teil 3 Aufgabe b) auf B.1 Arbeitsblatt 1, in der die Schüler die Bestimmung der Wahrscheinlichkeiten automatisieren, auch die Worte enthält, die die Schüler selbst in ihrer Tabelle verwendet haben.

# 7. Ausblick

Der Workshop, der Hauptbestandteil dieser Arbeit ist, wurde auf Grundelage der aus der ersten Erprobung gewonnenen Erkenntnisse 6, angepasst und verbessert. Das Material ist so konzipiert, dass sowohl Mitarbeiter von CAMMP als auch Lehrer den Workshop durchführen können. Dazu gibt es neben den angehängten Arbeitsblättern und Präsentationen ein Basic Paper und einen Stundenverlaufsplan, welche Teile aus Kapitel 3 und Kapitel 5 enthalten.

Im Rahmen der Arbeit konnten nicht alle Verbessungsvorschläge und zusätzliche Ideen zum Material umgesetz werden

- Um die Verwirrung bzgl. der vorgegeben Mails, die nicht Teil des Trainingsdatensatzessind sind, aus Teil 2 von B.1 Arbeitsblatt 1 zu lösen, besteht die Möglichkeit die Mails B.5 durch Mails zu ersetzen, die nicht auf B.3 Arbeitsblatt 3, zum Test des Klassifikators, verwendet werden. Diese Mails müssten in den Datensatz mit aufgenommen werden.

- Ein größerer Datensatz würde die Qualität des Spamfilters erhöhen. Mit diesem Datensatz ist es möglich ein aussagekräftiges Ergebnis zu erreichen, dies sieht man auf B.3 Arbeitsblatt 3 an den 10 Testmails, die alle korrekt klassifiziert werden, ohne mit diesem Ziel ausgewählt worden zu sein. Es liegt ein Skript zum Auswerten von E-Mails im Ordner data in der CAMMP cloud unter dem Namen word_counter.py vor. Mit diesem Skript ist es möglich einen neuen Datensatz zu erstellen, der größer oder diverser ist.

- Um den Wunsch der Schüler für die automatisch generiert Liste 6.4 nachzukommen, wäre es nötig die Worte mithilfe einer Liste zu speichern und dann in die Ausgabe miteinzupflegen.

- Um die Notation in Teil 2 auf B.3 Arbeitsblatt 2 zu verbessern würde sich eine Funktion eignen, die den Befehl ausführt, und durch einen einfachen Namen aufzurufen ist.

- Die Idee aus 6.3 könnte in der Präsentation A.2 umgesetz werden.

- In 5 genannt kann der Workshop auch in einer Doppelstunde durchführt werden. Dazu wäre es möglich eine Kurzversion von B.3 Arbeitsblatt 3 erstellen.

- Bisher wurde daruf verzichtet das Thema im KI-Kontext ein zu betten. Das hing vor allem mit der Entscheidung zusammen, dass der Workshop auch in zwei Doppelstunde durchgeführt werden kann. Bei einer Auswertung des Lernmaterials ist zu überlegen ob der Naive-Bayes-Spam-Filter, als klassische KI-Anwednung, auch in diesen Kontext eingebettet werden sollte.

- Um einen reibungsloseren und einfacheren Verlauf des Workshops zu gewährleisten, ist es möglich weitere Hilfekarten zu ergänzen und die automatisierten Rückmeldungen genauer zu gestalten.

# Anhang

# A. Präsentationen

## A.1. Präsentation Spamfilter Einfürung


**Wie funktionieren eigentlich Spamfilter und was hat das mit Mathe zu tun?**

Einführung in die Problemstellung



CAMMP

### Woher kommt der Begriff Spam?




### Welche Spam Mails kennt ihr?



### Welche Spam Mails kennt ihr?

Gewinne

Hallo !

Du wurdest heute von **Netflix** als **einer der drei glücklichen Gewinnern** eines **Premium-Jahresabo** ausgewählt! Sie können anlagonzlt Filme und Serien auf Ihrem Smartphone, Tablet, Laptop und Fernseher ansehen – **ohne Extragebühren**.

Das Angebot gilt nur noch diese Woche und nur in Deutschland. Klick hier, um sich einen der verbleibenden gratis Abos zu sichern und die Weltsaag zu verfolgen!

» JETZT TEILNEHMEN «

Viel Spaß und Erfolg wünscht, Ihr Gewinnerteam Team!

Lieferungen

TRACKING-ID: #34632900-371

**-Wir konnten Ihr Paket nicht zustellen, da niemand anwesend war, um für die Lieferung zu unterschreiben.**

**-Wir möchten Sie darauf hinweisen, dass wir eine Adressbestätigung benötigen, um den Paketversand erneut zu bestätigen.**

HIER PRÜFEN

Finanzen

Grüße,

Ich bin Stacey Reynolds, die Filialleiterin einer Investmentbank.

Ich persönlich habe während der Jahresabschlussprüfung unserer Bank ein ruhendes Konto im Wert von 48.000.000,00 GBP (nur achtundvierzig Millionen Britische Pfund) von unseren verstorbenen Kunden entdeckt.

Seit dem Tod des Verstorbenen hat bis heute niemand auf diesem Konto gearbeitet. Unser Ethikkodex für Banken wird diesen Fonds auf jeden Fall beschlagnahmen, wenn er für einen Zeitraum von 10 Jahren ohne Ansprüche ruht.

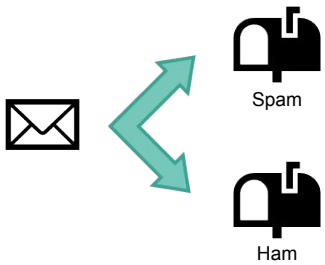
In diesem Zusammenhang benötige ich dringend Ihre volle Unterstützung bei der Überweisung dieses Geldes von unserer Bank, um zu vermeiden, dass unsere Bank dieses Geld beschlagnahmt. Unser Anteilsverhältnis beträgt 50 % für Sie und 50 % für mich.

Sollten Sie Interesse haben, mit mir in diesem Projekt zusammenzuarbeiten? Bitte kontaktieren Sie mich für weitere Informationen und Richtlinien.

Mit freundlichen Grüßen,
Stacey Reynolds
Abteilungsleiter
Investmentbank



### Problem und Ziel des Workshops






### Eigenschaften für einen Spamfilter

Betreff: Anlässlich des Lidl Jubiläums erhalten Sie Monsieur Cuisine Connect
Von: "LIDL" <hello@oelogid.LIDL.de>
Datum: 01.11.2022, 16:36

Monsieur Cuisine Connect
SIE SIND EIN GLÜCKLICHER BENUTZER
BRANDNEU Monsieur Cuisine Connect
KOSTENLOS FORTSETZEN KOSTENLOS In Zusammenarbeit mit logo
Sehr geehrter Lucky-Kunde,
Herzliche Glückwünsche!
Wir geben die Gelegenheit, eine unserer begrenzten Belohnungen zu erhalten KOSTENLOS (Versand nicht inbegriffen).
Beantworten Sie die Fragen zu Ihrer Erfahrung mit uns auf der nächsten Seite, um die Chance zu erhalten Brandneu Monsieur Cuisine Connect!
Ihre Meinung ist sehr wertvoll. Klicken Sie auf OK, um zu beginnen.
OK
Mit freundlichen Grüße,

## Eigenschaften für einen Spamfilter

Betreff: Anlässlich des Lidl Jubiläums erhalten Sie Monsieur Cuisine Connect
Von: "LIDL" <hello@oelogid.LIDL.de>
Datum: 01.11.2022, 16:36

Black und White-List

Monsieur Cuisine Connect
SIE SIND EIN GLÜCKLICHER BENUTZER
BRANDNEU Monsieur Cuisine Connect
KOSTENLOS FORTSETZEN KOSTENLOS In Zusammenarbeit mit logo
Sehr geehrter Lucky-Kunde,
Herzliche Glückwünsche!
Wir geben die Gelegenheit, eine unserer begrenzten Belohnungen zu erhalten KOSTENLOS (Versand nicht inbegriffen).
Beantworten Sie die Fragen zu Ihrer Erfahrung mit uns auf der nächsten Seite, um die Chance zu erhalten Brandneu Monsieur Cuisine Connect!
Ihre Meinung ist sehr wertvoll. Klicken Sie auf OK, um zu beginnen.
OK
Mit freundlichen Grüße,

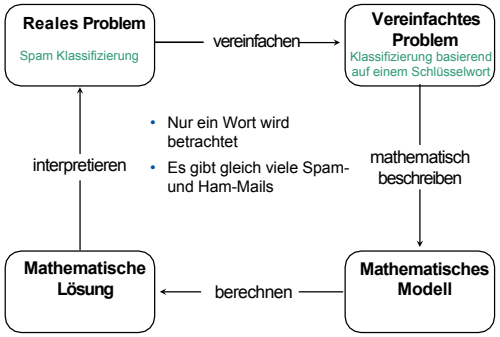
Satzbau und Rechtschreibung

Schlüsselwörter

Distributed Checksum Clearinghouse



## Modellierungskreislauf





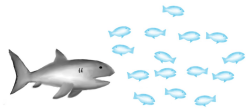

## Ablauf des Workshops

**Ablauf:**

1. Die aus den Daten gewonnenen Informationen in eine geeignete Struktur bringen.
2. Ein Klassifikator abhängig von einem Schlüsselwort erstellen
3. Den Klassifikator auf mehrere Schlüsselworte erweitern
4. Den entwickelte Klassifikator bewerten.



## Jetzt seid ihr dran …



Keine Panik!

ORGANISIEREN!

- Bearbeitet die Arbeitsblätter!
- Aufgabenstellung sorgfältig lesen!
- Teamwork!
- Nutzt die Tipps!
- Nutzt das Internet!
- Fragt die Betreuer:innen!



## Schritte zum Arbeitsmaterial

1. Öffne ***workshops.cammp.online***
2. Auf „Zugriff auf Lernmaterial“ und dann auf „Registrieren!“ klicken
3. Account erstellen: der Username muss Präfix **cammp_** enthalten (z. B. **cammp_laura1234**)
4. Auf „Anmelden!“ klicken, Accountdaten eingeben und einloggen
5. Öffne die Datei *Willkommen_CAMMP*
6. Im Dropdown Menü *Spamfilter* auswählen und herunterladen
7. Ordner 🗀 *spamfilter* und dann Ordner 🗀 *worksheets* öffnen
8. Los geht’s mit Arbeitsblatt 1!

# A.2. Präsentation Spamfilter Diskussion 1


**Wie funktionieren eigentlich Spamfilter und was hat das mit Mathe zu tun?**

Diskussion nach AB 1



## Klassifikator-Code

Mail Wort Tabelle

| Wort | Anzahl |
|---|---|
| Interesse | 21 |
| wirklich | 39 |
| eure | 34 |
| Leben | 36 |

Spam Mail Wort Tabelle

| Wort | Anzahl |
|---|---|
| Interesse | 10 |
| wirklich | 16 |
| eure | 0 |
| Leben | 24 |

$$P(\text{Wort}) = \frac{\text{Anzahl des Wort}}{\text{Anzahl der Mails}} \qquad P\_Spam(\text{Wort}) = \frac{\text{Anzahl des Wort in Spam Mails}}{\text{Anzahl der Spam Mails}}$$

P(Wort) Tabelle

| Wort | Anzahl |
|---|---|
| Interesse | 0,01666667 |
| wirklich | 0,02666667 |
| eure | 0 |
| Leben | 0,04 |

P(Wort|Spam) Tabelle

| Wort | Anzahl |
|---|---|
| Interesse | 0,0175 |
| wirklich | 0,0325 |
| eure | 0,02833333 |
| Leben | 0,03 |



## Nötige Umformung

**gegeben:**

P_Spam(Wort)

P(Wort)

P(Spam) = 0.5

?

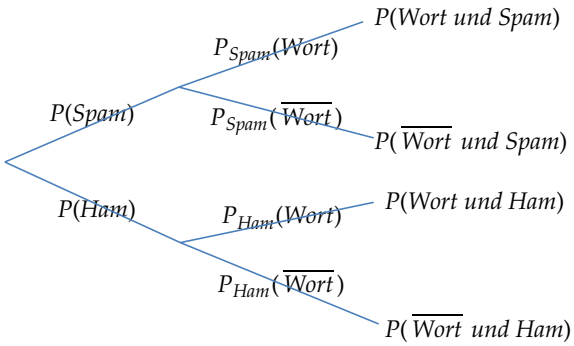
**gesucht:**

P_Wort(Spam)

Modell: Zweistufiges Zufallsexperiment

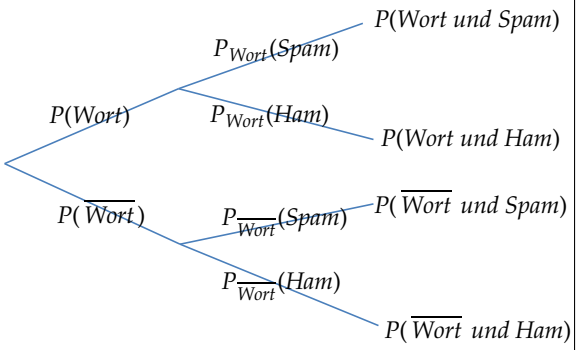


## Wahrscheinlichkeiten im Baumdiagramm



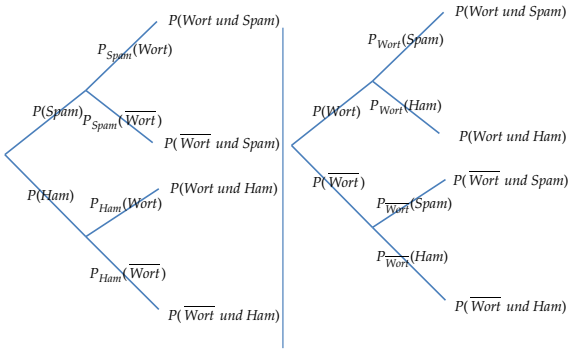



## Wahrscheinlichkeiten im Baumdiagramm

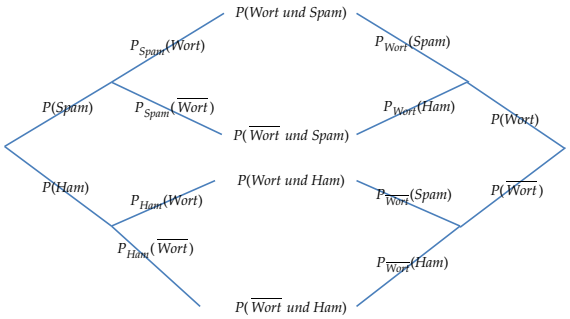





## Zwei „gleich“ Baumdiagramme

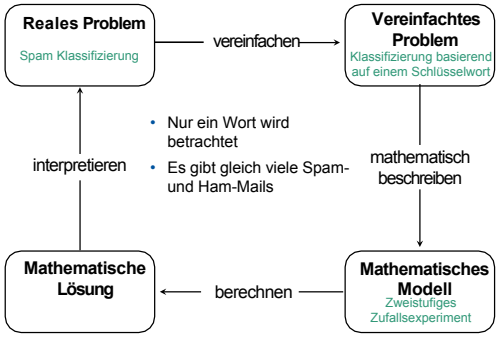


## Doppeltes Baumdiagramm



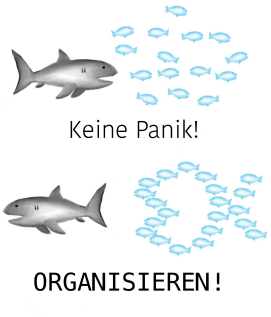



## Modellierungskreislauf






## Jetzt seid ihr dran …




- Bearbeitet die Arbeitsblätter!
- Aufgabenstellung sorgfältig lesen!
- Teamwork!
- Nutzt die Tipps!
- Nutzt das Internet!
- Fragt die Betreuer:innen!



## Schritte zum Arbeitsmaterial

1. Öffne ***workshops.cammp.online***
2. Auf „Zugriff auf Lernmaterial“ und dann auf „Registrieren!“ klicken
3. Account erstellen: der Username muss Präfix **cammp_** enthalten (z. B. **cammp_laura1234**)
4. Auf „Anmelden!“ klicken, Accountdaten eingeben und einloggen
5. Öffne die Datei *Willkommen_CAMMP*
6. Im Dropdown Menü *Spamfilter* auswählen und herunterladen
7. Ordner *spamfilter* und dann Ordner *worksheets* öffnen
8. Los geht’s mit Arbeitsblatt 2!

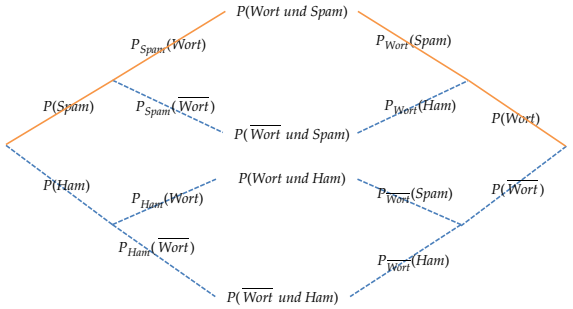
# A.3. Präsentation Spamfilter Diskussion 2


**Wie funktionieren eigentlich Spamfilter und was hat das mit Mathe zu tun?**

Diskussion nach AB 2



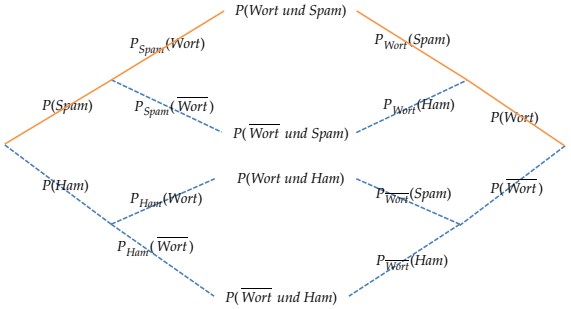
## Doppeltes Baumdiagramm




$P(Spam\ und\ Wort) = P(Spam) * P_{Spam}(Wort)$ $P(Spam\ und\ Wort) = P(Wort) * P_{Wort}(Spam)$



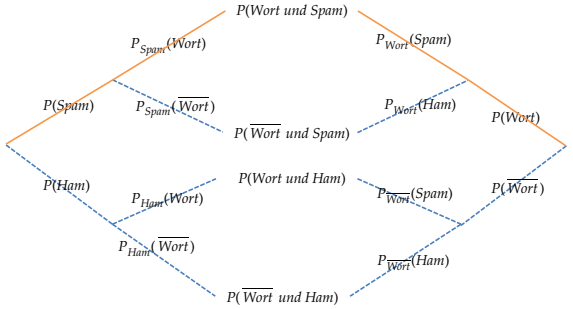
## Doppeltes Baumdiagramm




$P(Spam) * P_{Spam}(Wort) = P(Wort) * P_{Wort}(Spam)$



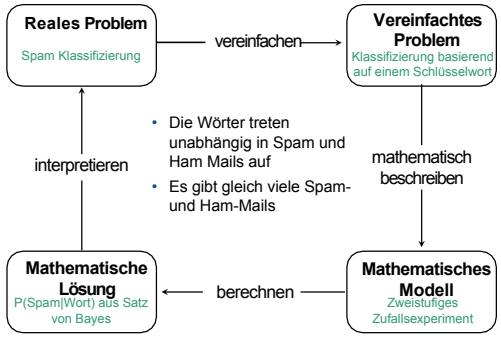
## Doppeltes Baumdiagramm




$$\frac{P(Spam) * P_{Spam}(Wort)}{P(Wort)} = P_{Wort}(Spam)$$



## Modellierungskreislauf






## Probleme des Modells

**Problem**: Worte die nicht im Datensatz vorkommen führen zu Ausgaben mit denen wir nicht rechen können.

| Wort | Pferd |
|---|---|
| P(Wort) | 0 |
| P(Wort\|Spam) | 0 |
| P(Wort\|Ham) | 0 |
| P(Spam\|Wort) | (0.5 * 0) / 0 = NaN |

**Problem**: Wir betrachten nur ein Wort bei der Klassifizierung einer Mail aus mehreren Wörtern.

## Probleme des Modells

**Problem**: Worte die nicht im Datensatz vorkommen führen zu Ausgaben mit denen wir nicht rechen können.

| Wort | Pferd |
|---|---|
| P(Wort) | 0 |
| P(Wort\|Spam) | 0 |
| P(Wort\|Ham) | 0 |
| P(Spam\|Wort) | (0.5 * 0) / 0 = NaN |

| Wort | Pferd |
|---|---|
| P(Wort) | 2 / 1000 = 0.02 |
| P(Wort\|Spam) | 1 / 500 = 0.02 |
| P(Wort\|Ham) | 1 / 500 = 0.02 |
| P(Spam\|Wort) | (0.5 * 0.02) / 0.02 = 0.5 |

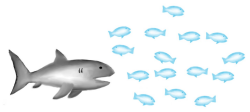
**Lösung:** Wir nehmen an, dass jedes Wort mindestens einmal in jeder Spam und jeder Ham Mail vorkommt.

**Problem**: Wir betrachten nur ein Wort bei der Klassifizierung einer Mail aus mehreren Wörtern.

**Lösung:** Auf Arbeitsblatt 3



## Jetzt seid ihr dran …



Keine Panik!

ORGANISIEREN!

- Bearbeitet die Arbeitsblätter!
- Aufgabenstellung sorgfältig lesen!
- Teamwork!
- Nutzt die Tipps!
- Nutzt das Internet!
- Fragt die Betreuer:innen!



## Schritte zum Arbeitsmaterial

1. Öffne ***workshops.cammp.online***
2. Auf „Zugriff auf Lernmaterial“ und dann auf „Registrieren!“ klicken
3. Account erstellen: der Username muss Präfix **cammp_** enthalten (z. B. **cammp_laura1234**)
4. Auf „Anmelden!“ klicken, Accountdaten eingeben und einloggen
5. Öffne die Datei *Willkommen_CAMMP*
6. Im Dropdown Menü *Spamfilter* auswählen und herunterladen
7. Ordner *spamfilter* und dann Ordner *worksheets* öffnen
8. Los geht’s mit Arbeitsblatt 3!

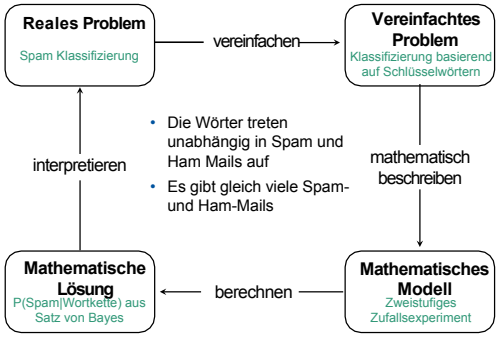
# A.4. Präsentation Spamfilter Abschluss


**Wie funktionieren eigentlich Spamfilter und was hat das mit Mathe zu tun?**

Abschluss



## Modellierungskreislauf






## Ablauf des Workshops

**Ablauf:**

1. Die aus den Daten gewonnenen Informationen in eine geeignete Struktur bringen.

| Wort | P(Wort) | P(Wort\|Spam) |
|---|---|---|
| Millionen | 0.0143218 | 0.0285714 |
| Haus | 0.020219 | 0.0252101 |
| Lieferung | 0.031171 | 0.0403361 |
| JETZT | 0.0235889 | 0.0470588 |

2. Ein Klassifikator abhängig von einem Wort erstellen

```
(P_Spam_Wort, entscheidung) = bayes_klassifikator("Millionen") #Hier dein Wort einsetzen
println("Die Mail wird als $entscheidung klassifiziert mit einer Spam-Wahrscheinlichkeit von $P_Spam_Wort")
```

```
Die Mail wird als Spam klassifiziert mit einer Spam-Wahrscheinlichkeit von 0.9974789915966387
```

3. Den Klassifikator auf mehrere Worte erweitern

```
(wsv, entscheidung) = bayes_klassifikator_2(0.5, ["Ich","bin","eine","Email"], P_Wort_Ham_liste, P_Wort_Spam_liste)
print("Die Mail wird als $entscheidung klassifiziert mit einem Wahrscheinlichkeitsverhältnis von $wsv")
```

```
Die Mail wird als Ham klassifiziert mit einem Wahrscheinlichkeitsverhältnis von 0.11355826230958697
```

4. Den entwickelte Klassifikator bewerten.



## Evaluation

**Euer Feedback ist gefragt!**

# B. Arbeitsblätter

## B.1. Arbeitsblatt 1



# Spamfilter | Arbeitsblatt1

In [2]:
```
#Hier nichts ändern!
include("../code/CheckAB1.jl");
```

## Teil 1: Welche Wörter eignen sich als Schlüsselwörter?

Wir wollen Schlüsselwörter nutzen, um Spam Mails und Ham Mails zu unterscheiden. Dafür ist es wichtig, dass wir herausfinden, welche Wörter sich dafür eignen.

Betrachte dazu die Mails, die du hier findest.

### a) Schlüsselwörter finden

*Benenne 3 Wörter, die deiner Meinung nach besonders für eine Spam Mail sprechen. Trage sie in der Tabelle in der ersten Spalte an Stelle der NaN ein. Die Spalten $P(wort)$ und $P*{Spam}(wort)$ werden später gefüllt._*

Durch einen Doppelklick kannst du in den Bearbeitungsmodus der Tabelle wechseln und durch einen Klick auf den -Button wieder in die normale Ansicht.

| Wort | P(wort) | P_Spam(wort) |
|---|---|---|
| NaN | NaN | NaN |
| NaN | NaN | NaN |
| NaN | NaN | NaN |

---

### b)

*Woran hast du die Schlüsselwörter erkannt?*

## Teil 2: Den Wörtern Wahrscheinlichkeiten zuordnen

Einzelne Schlüsselwörter können verschieden stark dafür sprechen, dass eine Mail eine Spam oder Ham Mail ist. Damit der Computer das versteht, müssen wir diese Aussage mathematisch formulieren.

Unsere Daten bestehen aus drei Listen. In der ersten Liste stehen, für jedes Wort das mindestens 6 mal in unseren über 1000 Mails vorgekommen ist, in wie vielen Mails es vorgekommen ist. In der zweiten Liste steht in wie vielen Spam Mails das Wort vorgekommen ist. In der dritten Liste steht in wie vielen Ham Mails das Wort vorgekommen ist.

Indem du den Befehl "kennzahlen(wort)" Befehl nutzt, kannst du dir für ein beliebiges Wort folgende Informationen aus unseren Daten ausgeben lassen:

- Anzahl der Mails in denen das Wort vorkommt
- Anzahl der Mails in unserem Datensatz
- Anzahl der Spam Mails in unserem Datensatz
- Anzahl der Spam Mails in denen das Wort vorkommt

### a) Daten auslesen

*Führe das Codefeld für deine 3 Wörter aus Teil 1 aus.*

*Ersetzte hierfür das NaN im folgenden Codefeld durch dein Wort. Bearbeite dabei für jedes deiner Wort die Aufgabenteile b) und c).*

*Die Anführungszeichen sich wichtig, damit der Computer das Wort erkennt.*

In [ ]:
```
wort = "NaN" # ersetze das NaN durch dein Wort

#Hier nichts ändern
kennzahlen(wort)
```

### b) Die Auftretenswahrscheinlichkeit der Wörter berechnen

Die Wahrscheinlichkeit, dass ein konkretes Wort in einer zufälligen Mail auftritt, ist $P(Wort)$. In den Codefeldern wird $P(Wort)$ als P_Wort geschrieben, da der Computer das leichter versteht.

*Bestimme im nächsten Codefeld die Wahrscheinlichkeit $P(Wort)$ für deine fünf Wörter aus Teil 1. Notiere die Ergebnisse in deiner Tabelle in der zweiten Spalte. Du kannst auf 4 Nachkommastellen runden.*

Nutze dazu das nächste Codefeld, wie du einen Taschenrechner nutzen würdes.

In [ ]:
```
P_wort = #=NaN=# gib_Anzahl_Mails_mit(wort) / gib_Anzahl_Mails() # ersetze NaN durc

#Hier nichts ändern
check_t2_b(wort, P_wort);
```

### c) Die Auftretenswahrscheinlichkeit der Wörter in Spam Mails berechnen

Die bedingte Wahrscheinlichkeit, dass ein Wort in einer Spam Mail vorkommt, ist $P_{Spam}(Wort)$; im Codefeld als P_Spam_Wort geschrieben.

*Bestimme $P*{Spam}(Wort)$ im nächsten Codefeld für deine fünf Wörter. Die Wahrscheinlichkeit kann ebenso wie oben über die relative Häufigkeit geschätzt werden. Die Ergebnisse kannst du gerundet in der dritten Spalte notieren._

In [ ]:
```
P_Spam_Wort = #=NaN=# gib_Anzahl_Spam_Mails_mit(wort) / gib_Anzahl_Spam_Mails() # e

#Hier nichts ändern
check_t2_c(wort, P_Spam_Wort);
```

## Teil 3: Automatisierung der Wahrscheinlichkeitsberechnung

Im Folgenden interessieren uns nur noch die Wahrscheinlichkeiten $P(Wort)$ und $P_{Spam}(Wort)$. Deswegen wollen wir eine Liste erstellen, die direkt unsere Wahrscheinlichkeiten speichert. Die händische Berechung der Wahrscheinlichkeiten, so wie sie in Teil 2 durchgeführt wird, würde für mehr als 3 Wörter sehr lange dauern. Deshalb wollen wir im Folgenden das Berechnen dieser Wahrscheinlichkeiten und das Erstellen einer Tabelle, wie in Teil 1, automatisieren. Dafür nutzen wir eine for-Schleife.

## Erklärung der for-Schleife

Eine Schleife wiederholt einen Codeabschnitt mehrfach mit sich verändernden Variablen. Ein einfaches Beispiel ist, alle Worte aus einer Liste an Worten, auszugeben. Das siehst du im nächsten Codeblock. Wenn du noch mehr über die for-Schleife erfahren willst, gibt es hier eine Hilfekarte dazu, die das Prinzip nochmal genauer an einem Beispiel erklärt.

## Beispiel einer for-Schleife

*Führe das nächste Codefeld aus, um dich mit der Funktionsweise einer for-Schleife vertraut zu machen.*

In [ ]:
```
#Hier nichts ändern
wort_liste = ["Ich", "bin", "eine", "Wortliste"] # Liste mit Elementen mit denen di

for wort in wort_liste # Schleifenkopf mit sich ändernden Variable

    println(wort) # Anweisung im Schleifenrumpf die wiederholt wird

end # hier endet die Schleife
```

## a) Programmierung des Algorithmus

Wir wollen einen Algorithmus programmieren der $P(Wort)$ und $P_{Spam}(Wort)$ für alle Wörter aus unserem Datensatz bestimmt. Die Struktur für die Schleife ist in der nächsten Codefeld schon aufgebaut. Eine genauere Beschreibung, was im Code passiert, findest du unter dem Codefeld.

*Ergänze im untenstehenden Codefeld die Formel für die Berechnung von $P(Wort)$ und $P*{Spam}(Wort)$* Hierbei kannst du die folgenden Befehle nutzen:_

```
gib_Anzahl_Mails_mit(wort)
gib_Anzahl_Mails()
gib_Anzahl_Spam_Mails()
gib_Anzahl_Spam_Mails_mit(wort)
```

Wenn du einen der Befehle benutzt, berechnet der Computer den angefragten Wert. Du kannst diese Werte auch miteinander verrechnen.

Alle Befehle findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
#Hier nichts ändern
wort_liste = gib_wort_liste()
P_Wort_liste = Dict()
P_Spam_Wort_liste = Dict()
for wort in wort_liste


    #Ersetze die NaN´s durch die Formeln mit denen du P_Wort und P_Wort_Spam berech
    P_Wort = #=NaN=# gib_Anzahl_Mails_mit(wort) / gib_Anzahl_Mails()
    P_Spam_Wort = #=NaN=# gib_Anzahl_Spam_Mails_mit(wort) / gib_Anzahl_Spam_Mails()

    #Hier nichts ändern
    P_Wort_liste[wort] = P_Wort
    P_Spam_Wort_liste[wort] = P_Spam_Wort
end

check_t3_a(wort_liste, P_Wort_liste, P_Spam_Wort_liste);
```

Kurze Beschreibung des Codes in Worten:

Zeile 1-3: Wir holen uns unsere Wortliste aus dem Datensatz und legen zwei leere Listen an, in die wir unsere Ergebnisse speichern wollen.
Zeile 4-6: Dann durchlaufen wir für jedes Wort aus unserer Wortliste den Anweisungsabschnitt der for-Schleife und berechnen hierbei $P(Wort)$ und $P_{Spam}(Wort)$ berechnen.
Zeile 7-8: Zum Schluss hängen wir diese Werte unsereren Listen an.

## Zusatz

*Überlege dir, welche Werte der Wahrscheinlichkeiten $P(Wort)$ und $P*{Spam}(Wort)$ dafür sprechen, dass es sich bei unserem Wort um ein gutes Schlüsselwort für eine Spam Mail handelt._

## B.2. Arbeitsblatt 2



# Spamfilter | Arbeitsblatt 2

In [3]:
```
include("../code/CheckAB2.jl");
```

## Teil 1: Herleitung des Satz von Bayes mithilfe des doppelten Baumdiagramms

In der Besprechung des letzten Arbeitsblatt haben wir das unten abgebildete doppelte Baumdiagramm kennen gelernt. Mit Hilfe dieses Baumdiagramms können wir nun eine Formel entwickeln, mit der wir unsere gesuchte Wahrscheinlichkeit $P_{Wort}(Spam)$ berechnen können.

An dieser Stelle ist ein Bild des Doppelten Baumdiagramms

### a) Gleichungen finden

*Stelle zwei Gleichungen auf, die einmal von rechts und einmal von links (siehe Baumdiagramm) mit den Pfadregeln die Wahrscheinlichkeit $P(Wort\ und\ Spam)$ berechnen.*

Alle Variabel findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
P_W_u_Spam_1 = #=NaN=# P_Wort * P_Wort_Spam # Ersetze das NaN durch die Gleichung v
P_W_u_Spam_2 = #=NaN=# P_Spam * P_Spam_Wort # Ersetze das NaN durch die Gleichung v

# Hier nichts ändern
check_t1_a(P_W_u_Spam_1, P_W_u_Spam_2)
```

### b) Gleichung umstellen: der Satz von Bayes

*Nutze die beiden oberen Gleichungen um einer Formel für die Wahrscheinlichkeit $P*{Wort}(Spam)$ aufzustellen. Die Wahrscheinlichkeit $P(Wort\ und\ Spam)$ kann nicht aus den Informationen, die wir über die Mails haben, berechnet werden. Sie soll in der Formel daher nicht auftreten._

Alle Variablen findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
P_Wort_Spam = #=NaN=# P_Spam * P_Spam_Wort / P_Wort # Ersetze das NaN durch die For

# Hier nichts ändern
check_t1_b(P_Wort_Spam)
```

Die Verallgemeinerung der hier bestimmten Formel wird als **Satz von Bayes** bezeichnet. Sie gilt nicht nur für unseren konkreten Fall, sonder für alle Zufallsexperimente, die mit einem zweistufigen Baumdiagramme, welches auf jeder Stufe nur zwei Zweige hat, dargestellt werden können. Sobald du Aufgabe 1 b) gelöst hast, kannst du durch einen Klick auf die drei Punkte die allgemeine Form anschauen.

> Satz von Bayes: Wir betrachten die beiden Ereignisse A und B. Der Satz von Bayes ermöglicht es die bedingte Wahrscheinlichkeit von $P_B(A)$ auszurechnen, wenn nur die umgekehrte bedingte Wahrscheinlichkeit $P_A(B)$ und die Wahrscheinlichkeiten der einzelnen Ereignisse $P(A)$ und $P(B)$ bekannt sind: $$P_B(A) = \frac{P(A) * P_A(B)}{P(B)}$$ $P_B(A)$ bedeutet Wahrscheinlichkeit von Ereignis A unter der Bedingung von B

## Teil 2: Satz von Bayes auf unser Beispiel anwenden

Unten ist ein Ausschnitt aus der Tabelle, die wir auf Arbeitsblatt 1 bestimmt haben, angegeben.

Zur Erinnerung: Die Wahrscheinlichkeit, dass ein Wort in einer zufälligen Mail auftritt ist $P(Wort)$. Die Wahrscheinlichkeit, dass ein Wort in einer Spam Mail auftritt, ist $P_{Spam}(Wort)$.

In [4]:
```
print_tabel_section()
```

| Wort | P(Wort) | P_Spam(Wort) |
|---|---|---|
| Millionen | 0.0143218 | 0.0285714 |
| Haus | 0.020219 | 0.0252101 |
| Lieferung | 0.031171 | 0.0403361 |
| JETZT | 0.0235889 | 0.0470588 |

- Der Zugriff auf die Wahrscheinlichkeit $P(Wort)$ funktioniert mit get(P_Wort_liste, wort, 0) wobei für den Platzhalter "wort" nun ein konkretes Wort eingesetzt wird. Dieses muss in Anführungszeichen stehen.
- Der Zugriff auf die bedingten Wahrscheinlichkeiten $P_{Spam}(Wort)$ funktioniert mit get(P_Spam_Wort_liste, wort, 0) auch hier ist "wort" wieder ein Platzhalter.
- Der Befehl gibt 0 aus, falls es für das konkrete Wort noch keinen Eintrag in der Tabelle gibt

## a) Den Satz von Bayes auf ein Wort anwenden

_Berechne mit den oben gegebenen Befehlen und dem Satz von Bayes die Wahrscheinlichkeit, dass es sich bei einer Mail um eine Spam Mail handelt, wenn das Wort `Millionen` in dieser vorkommt._

Beachte, dass du Kommazahlen mit einem Punkt trennen musst. z.B. 1,2 wird als 1.2 geschrieben.

Wir nehmen an, dass die Hälfte aller Mails Spam Mails sind. Folglich gilt $P(Spam)$ = 0.5.

Alle Befehle findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
P_Millionan_Spam = #=NaN=# 0.5 * get(P_Spam_Wort_liste,"Millionen",0) / get(P_Wort_

# Hier nichts ändern
check_t2_a(P_Millionan_Spam)
```

## b) Einen ersten Klassifikator entwickeln

Bisher bestimmen wir nur Wahrscheinlichkeiten. Unser Ziel ist letztendlich die Mail in Spam und Ham zu klassifizieren. Wir müssen somit eine Funktion entwickeln, welche Mails der Klasse Spam oder Ham zuordnet. Diese bezeichnen wir als Klassifikator. Dafür definieren wir uns die Funktion

```
bayes_klassifikator(wort)
```

Diese Funktion gibt uns die berechnete Wahrscheinlichkeit aus und darauf basierend eine Entscheidung, ob die Mail als Spam oder Ham klassifiziert wird.

Für die Entscheidung nutzen wir eine Fallunterscheidung. Genau genommen verwenden wir eine if-Bedinung. Falls du mehr über if-Bedingung wissen willst, gibt es hier noch eine kleine Erklärung.

*Im nächsten Codefeld fehlt nur noch die Bedingung für die Fallunterscheidung. Setze sie ein sodas für alle Mails, die wir als Spam klassifizieren wollen, die Bedingung* `wahr` *und für Ham Mails* `falsch` *ist.*

Um Werte zu vergleichen, kannst du folgende Zeichen verwenden:

```
A <  B  steht für "A kleiner B"
A <= B  steht für "A kleiner gleich B"
A == B  steht für "A gleich B"
A >= B  steht für "A größer gleich B"
A >  B  steht für "A größer B"
```

Wenn du die Bedingung eingesetzt hast kannst du den Klassifikator für ein paar Worte ausprobieren, indem du in der letzten Zeile das NaN durch ein Wort deiner Wahl ersetzt.

Alle Befehle und Variablen findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
function bayes_klassifikator(wort)
    P_Wort_Spam = 0.5 * get(P_Spam_Wort_liste,wort,0) / get(P_Wort_liste,wort,0)

    #Bis hier nicht verändern

    if #=NaN=# P_Wort_Spam >= 0.5 #Hier das NaN durch die Bedinung ersetzen

    #Ab hier nicht verändern

        entscheidung = "Spam"
    else
        entscheidung = "Ham"
    end

    return P_Wort_Spam, entscheidung
end

(P_Wort_Spam, entscheidung) = bayes_klassifikator("NaN") #Hier dein Wort einsetzen

println("Die Mail wird als $entscheidung klassifiziert mit einer Spam-Wahrscheinlic
check_t2_b(bayes_klassifikator)
```

Unterschied relative Häufigkeit und geschätzte Wahrscheinlichkeiten

Die Wahrscheinlichkeiten, die wir benutzen stammen aus unserem Datensatz. Auf Arbeitsblatt 1 haben wir, basierend auf der relativen Häufigkeit eines Worte, die zugehörigen Wahrscheinlichkeiten berechnet. Wenn wir nur Mails aus unserem Datensatz betrachten passen die Zahlen genau. Wenn wir neue Mails betrachten, die nichts mit unserem Datensatz zu tun haben, sind unsere Wahrscheinlichkeiten nur noch Schätzungen. Wir gehen davon aus das unser Datensatz die Realtiät gut repäsentiert und die Wahrscheinlichkeiten, die wir berechnet haben, sich auf alle Mails verallgemeinern lassen können.

## Zusatz

Wenn du das Wort "Pferd" in unseren Klassifikator einsetzt, bekommtst du eine besondere Ausgabe.

&

*Überlege und recherchiere im Internet, was die Ausgabe bedeutet und wie es dazu kommen kann.*
*Dieses Ergebnis ist unerwünscht. Wie kann es verhindert werden?*

## B.3. Arbeitsblatt 3



# Spamfilter | Arbeitsblatt 3

In [1]:
```
include("../code/CheckAB3.jl");
```

## Teil 1 | Herleitung der Formel für mehrere Wörter

Für die Entscheidung, ob eine Mail als Spam oder Ham zu klassifizieren ist, genügt es meist nicht, nur ein Wort zu betrachten. Wir suchen eine Formel, mit der wir die Spam-Wahrscheinlichkeit abhängig von mehreren Wörtern berechnen können.

Wir wollen also die folgende Wahrscheinlichkeit bestimmen: $P_{Wort1\ und\ Wort2\ und\ \dots} (Spam)$

Aus unserem Datensatz können wir ohne großen Aufwand $P(Spam), P(Ham), P(Wort), P_{Spam}(Wort)$ und $P_{Ham}(Wort)$ bestimmen.

Zuerst betrachten wir den Fall mit zwei Wörtern.

Dafür können wir jedoch nicht einfach die beiden Wahrscheinlichkeiten $P_{Wort1}(Spam)$ und $P_{Wort2}(Spam)$ miteinandern multiplizieren, da die beiden Ereignisse nicht wie in einem zweistufigen Baumdiagramm getrennt betrachtet werden können.

In Arbeitsblatt 2 haben wir den Satz von Bayes kennengelernt.

### a) Satz von Bayes anwenden

*Wende den Satz von Bayes an um eine erste Formel für $P*{Wort1\ und\ Wort2}(Spam)$ zu bestimmen._

Alle Variablen findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
P_Wort1_schnitt_Wort2_Spam = #=NaN=# P_Spam * P_Spam_Wort1_schnitt_Wort2 / P_Wort1_

# Hier nichts ändern
check_t1_a(P_Wort1_schnitt_Wort2_Spam)
```

*Überlege dir dabei welche Wahrscheinlichkeiten bekannt und welche unbekannt sind.*

Klappe nun den nächsten Teil aus.

---

Unsere neue Formel für $P_{Wort1\ und\ Wort2}(Spam)$ lautet:

$$P_{Wort1\ und\ Wort2}(Spam) = \frac{P(Spam) \cdot P_{Spam}(Wort1\ und\ Wort2)}{P(Wort1\ und\ Wort2)}$$
Leider haben wir jetzt zwei unbekannte Wahrscheinlichkeiten $P_{Spam}(Wort1\ und\ Wort2)$ und $P(Wort1\ und\ Wort2)$. Wir könnten durch Abzählen herausfinden, wie viele Mails es gibt, die genau diese Wortkombination enthalten. Das würde aber dazu führen, dass wir bei der Berücksichtigung von mehreren Wörtern irgendwann nur noch testen ob es diese Mail in unserem Datensatz schon gibt. Zudem wäre dieses Verfahren viel zu aufwändig.

Eine der beiden Wahrscheinlichkeiten können wir aber mit einem Trick kürzen. Dafür betrachten wir die Wahrscheinlichkeit $P_{Wort1\ und\ Wort2}(Ham)$, die wir auch mit dem Satz von Bayes bestimmen können.

$$P_{Wort1\ und\ Wort2}(Ham) = \frac{P(Ham) * P_{Ham}(Wort1\ und\ Wort2)}{P(Wort1\ und\ Wort2)}$$

## b) Eliminieren einer unbekannten Wahrscheinlichkeit

*Gib im nächsten Codefeld an, wie man $P{Wort1\ und\ Wort2}(Spam)$ und $P{Wort1\ und\ Wort2}(Ham)$ verrechnen muss, um eine der beiden unbekannten Wahrscheinlichkeiten kürzen zu können. Ersetze dafür das NaN durch das passende Rechenzeichen.*

Alle Variablen findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [ ]:
```
wahrscheinlichkeitsverhaeltnis = #= P_Wort1_schnitt_Wort2_Spam NaN P_Wort1_schnitt_

# Hier nichts ändern
check_t1_b(wahrscheinlichkeitsverhaeltnis)
```

---

Wir nennen das Ergebnis `Wahrscheinlichkeitsverhältnis` , kurz `wsv` , und erhalten nach dem Kürzen folgenden Formel:

$$\frac{P_{Wort1\ und\ Wort2}(Spam)} {P_{Wort1\ und\ Wort2}(Ham)} = \frac{P(Spam) * P_{Spam}(Wort1\ und\ Wort2)} {P(Ham) * P_{Ham}(Wort1\ und\ Wort2)}$$

Da $P(Wort1\ und\ Wort2)$ nicht von Spam oder Ham abhängt, können wir uns sicher sein, dass sie auch für die Entscheidung, ob eine Mail Spam oder Ham ist, unwichtig ist.

Um unsere Situation zu vereinfachen haben wir angenommen, dass die Auftretenswahrscheinlichkeit unserer Wörter unabhänig ist. In der Realität ist das natürlich anders. Wenn das Wort "Baden" auftaucht, ist das Wort "Württemberg" auch sehr wahrscheinlich. Wir wenden diese Vereinfachung an, da der Effekt auf das Ergebnis minimal ist.

Mithilfe der Unabhängigkeit können wir $P_{Spam}(Wort1\ und\ Wort2)$ in $P_{Spam}(Wort1) * P_{Spam}(Wort2)$ aufteilen. Genau so können wir auch mit $P_{Ham}(Wort1\ und\ Wort2)$ verfahren. Das führt uns zu folgender Formel:

$$\frac{P_{Wort1\ und\ Wort2}(Spam)} {P_{Wort1\ und\ Wort2}(Ham)} = \frac{P(Spam) * P_{Spam}(Wort1) * P_{Spam}(Wort2)} {P(Ham) * P_{Ham}(Wort1) * P_{Ham}(Wort2)}$$

Wir haben eine Formel gefunden, die nur noch Wahrscheinlichkeiten enthält, die wir leicht aus unseren Daten bestimmen können.

Da wir angenommen haben, dass die Auftretenswahrscheinlichkeit für alle Worte unabhängig ist, können wir die Formel auf mehr als zwei Wörter verallgemeinern:

$$\frac{P_{Wort1\ und\ Wort2\ und\ Wort3\ und\ \dots}(Spam)} {P_{Wort1\ und\ Wort2\ und\ Wort3\ und\ \dots}(Ham)} = \frac{P(Spam) * P_{Spam}(Wort1) * P_{Spam}(Wort2) * P_{Spam}(Wort3) * \dots} {P(Ham) * P_{Ham}(Wort1) * P_{Ham}(Wort2)* P_{Ham}(Wort3) * \dots}$$

Wenn du wissen willst, warum wir die Unabhängigkeit des Auftretens der Wörter nicht nutzen können um $P(Wort1\ und\ Wort2\ und\ \dots)$ umzuformen, kannst du dir hier eine kurze Erklärung anschauen.

# Teil 2 | Erstellen des Klassifikators

Die Berechungen die im Teil 1 durchgeführt wurden sind im nächsten Codefeld schon programmiert.

Wie schon bei unserem ersten Klassifikator fehlt nur noch die Bedingung um Spam und Ham Mails zu unterscheiden.

## a) Bedingung des Klassifikators festlegen

*Formuliere eine Bedingung für das Klassifizieren einer Mail als Spam oder Ham abhängig vom Wahrscheinlichkeitsverhältnis wsv setze die Bedingung im nächsten Codefeld ein.*

Alle Befehle und Variablen findest du auch auf der Variablen und Befehls Übersicht von dort kannst du sie in die Codeblöcke kopieren.

In [18]:
```
spam = 0.5
ham = 1 - spam
e_mail = ["Das", "ist", "eine", "Testmail"]

function bayes_klassifikator_2(spam, e_mail, P_Ham_Wort_liste, P_Spam_Wort_liste)
    P_Ham_Wort_Kette = 1
    P_Spam_Wort_Kette = 1
    for wort in unique(e_mail)
        P_Ham_Wort_Kette = P_Ham_Wort_Kette * get(P_Ham_Wort_liste, wort, 1)
        P_Spam_Wort_Kette = P_Spam_Wort_Kette * get(P_Spam_Wort_liste, wort, 1)
    end
    wsv = (spam * P_Spam_Wort_Kette) / (ham * P_Ham_Wort_Kette)

    #Bis hier nicht verändern

    if #=NaN=# wsv >= 1 #Hier das NaN durch die Bedingung ersetzen

    #Ab hier nicht verändern

        entscheidung = "Spam"
    else
        entscheidung = "Ham"
    end
    return wsv, entscheidung
end

check_t2(bayes_klassifikator_2, spam, e_mail, P_Ham_Wort_liste, P_Spam_Wort_liste)
```

```
✓ Deine Bedinung ist korrekt
Die Mail wird als Spam klassifiziert mit einem Wahrscheinlichkeitsverhältnis von 1.
9497403444753998
```

# Teil 3 | Test des Klassifikators

## a) Beispielmails testen

Nun haben wir einen Klassifikator erstellt, der Mails als Spam oder Ham klassifiziert. Wie gut der Klassifikator funktioniert, können wir nun mit ein paar Beispielmails testen. Dafür gibt es eine Liste aus 10 Beispiel-Mails.

Der Befehl println(mail_texts[e_mail_index]) gibt dir den Text der Mail aus. Danach wird mithilfe unseres Algorithmus das Wahrscheinlichkeitsverhältnis berechnet, um die Mail als Spam oder Ham Mail zu klassifizieren.

*Teste den Klassifikator für mindestens drei Mails. Ersetze "NaN" durch eine Zahl zwischen 1 und 10, um eine der Beispiel-Mails auszuwählen.*

In [ ]:
```
# ersetze das NaN durch den Index eine E-Mail aus der Liste
e_mail_index = NaN

# Hier nichts ändern
println(mail_texts[e_mail_index])
(wsv, entscheidung) = bayes_klassifikator_2(spam, split(mail_texts[e_mail_index]),
print("Die Mail wird als $entscheidung klassifiziert mit einem Wahrscheinlichkeitsv
```

## b) Ergebnis interpretieren

*Interpretiere die Ergebnisse. Wie gut funktioniert unser Spamfilter? Was für Probleme könnten auftreten?*

# Zusatz

Nachdem wir in a) Beispielmails klassifiziert haben, wollen wir nun unseren Algorithmus auf eigene Mails anwenden. Dafür hast du im nächsten Codefeld die Möglichkeit, eine Mail zu schreiben, ob Spam oder Ham ist dir überlassen. Ersetze das NaN durch den Text deiner Mail, achte dabei darauf, die Anführungszeichen nicht zu entfernen.

Mit dem Ausführen des Codefeldes wird der von uns erstellte Klassifikator die Mail klassifizieren.

In [ ]:
```
# ersetze das NaN durch deine E-Mail
e_mail_text = "NaN"

# Hier nichts ändern
bayes_klassifikator_2(0.5, split(e_mail_text), P_Ham_Wort_liste, P_Spam_Wort_liste)
```

## B.4. Bedingte Unabhängigkeit



# Die bedingte Unabhänigkeit

## Allgemeine Unabhängigkeit

Zwei Ereignisse A und B sind unabhängig, wenn gilt $P(A\ und\ B) = P(A) * P(B)$. Das bedeutet, dass die Wahrscheinlichkeit, dass A und B eintritt das Produkt der Einzelwahrscheinlichkeiten ist. Zusätzlich gilt auch $P_B(A) = P(A)$, das heißt die Wahrscheinlichkeit von A verändert sich nicht, egal ob B eintritt oder nicht. Das gleich können wir auch für $P_A(B) = P(B)$ sagen.

## Bedingte Unabhänigkeit

Zwei Ereignisse A und B sind dann bedingt unabhängig unter C, wenn gilt $P_C(A\ und\ B) = P_C(A) * P_C(B)$. Das bedeutet solange C gilt beeinflussen sich A und B nicht gegenseitig.

## Ein Rechenbeispiel:

Wir betrachten zur Veranschaulichung der Unterschiede der beiden Definitionen folgendes ausgedachtes Beispiel. Wir haben 10 Spam und 10 Ham Mails, es gilt folglich $P(Spam) = P(Ham) = 0.5$. Die Verteilung der Wörter kannst du aus der Tabelle unten entnehmen.

| Wort | Anz. Spam Mails mit | Anz. Ham Mails mit | P_Spam(Wort) | P_Ham(Wort) | P(Wort) |
|---|---|---|---|---|---|
| Stadt | 10 | 5 | 1 | 0.5 | 0.75 |
| Haus | 9 | 1 | 0.9 | 0.1 | 0.5 |

Wenn wir die Auftretenswahrscheinlichkeiten für Haus und Stadt als unabhänig betrachten, gilt $P(Haus\ und\ Stadt) = 0.75 * 0.5 = 0.375$.

Wenn wir bedingte Unabhängigkeit annehmen, gilt mit dem Satz der Totalen Wahrscheinlichkeit (siehe unten) $$P(Haus\ und\ Stadt) = P(Spam) * P_{Spam}(Haus\ und\ Stadt) + P(Ham) * P_{Ham}(Haus\ und\ Stadt)$$ $$P(Haus\ und\ Stadt) = P(Spam) * P_{Spam}(Haus) * P_{Spam}(Stadt) + P(Ham) * P_{Ham}(Haus) * P_{Ham}(Stadt)$$ $$P(Haus\ und\ Stadt) = 0.5 * 0.9 * 1 + 0.5 * 0.5 * 0.1 = 0.475$$

Wir sehen, dass sich die beiden Werte für $P(Haus\ und\ Stadt)$ unterscheiden. Entweder muss 0.375 oder 0.475 korrekt sein. Aus der Beschreibung der Situation wissen wir, dass es entweder 9 oder 10 Mails geben kann, die die Worte Haus und Stadt enthalten. Sicher sind es 9 Spam Mails, da jede Spam Mail das Wort Stadt enthält und 9 dieser Mails auch das Wort Haus. Für die Ham Mails können wir keine sicher Aussage treffen. Die Wahrscheinlichkeit, dass die Mail, die Haus enthält, auch Stadt enthält ist 0.5. Für eine genaue Schätzung von $P(Haus\ und\ Stadt)$ brauchen wir also den Mittelwert zwischen 10/20 und 9/20. Dieser

Mittelwert ist 0.475. Unser Modell wird folglich durch bedingte und nicht allgemeine Unabhängigkeit beschrieben. Das können wir auch intuitiv erkennen, da die Werte in unserer Tabelle in Abhängigkeit von Spam und Ham angegeben sind.

Mit dieser Erkenntnis könnten wir auch eine Formel für $P_{Haus\ und\ Stadt}(Spam)$ aufstellen. Das Wahrscheinlichkeitsverhältnis ist aber die einfachere und intuitivere Lösung.

---

## Totale Wahrscheinlichkeit

Den Satz der Totalen Wahrscheinlichkeit kann man sich am einfachsten über ein Baumdiragramm herleiten.

An dieser stelle ist das Bild: Baumdiagramm_Totale_Wahrscheinlichkeit

Aus dem Baumdiagramm erhalten wir folgende Gleichung:

$$P(Spam) * P_{Spam}(Wort1\ und\ Wort2) + P(Ham) * P_{Ham}(Wort1\ und\ Wort2) = P(Wort1\ und\ Wort2) * P_{Wort1\ und\ Wort2}(Spam) + P(Wort1\ und\ Wort2) * P_{Wort1\ und\ Wort2}(Ham)$$
Hier können wir $P(Wort1\ und\ Wort2)$ ausklammern und erhalten

$$P(Spam) * P_{Spam}(Wort1\ und\ Wort2) + P(Ham) * P_{Ham}(Wort1\ und\ Wort2) = P(Wort1\ und\ Wort2) * (P_{Wort1\ und\ Wort2}(Spam) + P_{Wort1\ und\ Wort2}(Ham))$$
Die Summe der Wahrscheinlichkeiten an einer Verzweigung im Baumdiagramm muss immer 1 ergeben. Daraus folgt das $P_{Wort1\ und\ Wort2}(Spam) + P_{Wort1\ und\ Wort2}(Ham) = 1$. Mit 1 multiplizieren verändert das Ergebnis nicht und wir erhalten folgende Gleichung:

$$P(Spam) * P_{Spam}(Wort1\ und\ Wort2) + P(Ham) * P_{Ham}(Wort1\ und\ Wort2) = P(Wort1\ und\ Wort2)$$

## B.5. Beispiel Mails



# Spam Mails

## Spam Mail 1.

Hallo

Ich bin Susanne Klatten und ich komme aus Deutschland, ich kann dich kontrollieren

finanzielle Probleme ohne Rückgriff auf Banken im Kreditbereich

Geld . Wir bieten Privatkredite und Geschäftskredite an, ich bin ein

zugelassener und zertifizierter Kreditgeber mit langjähriger Erfahrung im Kreditgeschäft

Kredite und wir vergeben besicherte und unbesicherte Darlehen

Beträge von 2.000,00 € ($) bis maximal

500.000.000,00 € mit einer festen Verzinsung von 3 % auf jährlicher Basis.

Brauchen Sie einen Kredit? Senden Sie uns eine E-Mail an: susannelegitfirm155@gmail.com

Sie können auch meinen Link anzeigen und mehr über mich erfahren.

https://en.wikipedia.org/wiki/Susanne_Klatten https://www.forbes.com/profile/susanne-klatten

E-Mail: susannelegitfirm155@gmail.com Unterschrift, Vorstandsvorsitzender Susanne Klatten

## Spam Mail 2.

Sehr geehrter Kunde,

Bitte beachten Sie, dass wir bei unserem Online-Service der Internetbanking - ING-DiBa aufgrund einiger Angriffe auf unsere Kunden eine neue Authentifizierungsmethode eingeführt haben. Zum Schutz der privaten Daten unserer Kunden haben wir neue Sicherheitsfunktionen entwickelt.

Sie müssen Ihre Angaben bei uns bestätigen.

Klicken Sie hier, um Ihre Angaben zu best_tigen: https://access.ing.de/delogin/de/konto/verifizierung

Die Nichtbeachtung dieser Anweisungen kann dazu führen, dass Ihr Konto vorübergehend gesperrt wird, bis die überprüfung abgeschlossen ist. Dies bedeutet lediglich, dass Sie nicht auf Ihr Konto zugreifen können, bis der überprüfungsvorgang abgeschlossen ist.

Hinweis: Bitte beachten Sie, dass wir nicht für Verlust oder Diebstahl in Ihrem Konto verantwortlich sind, wenn Sie heute keine Aktualisierung durchführen.

Danke für Ihre Kooperation.

Mit freundlichen Grüßen,

## Spam Mail 3.

Mein Name ist Vera Wilfred, ich möchte meinen Fonds 8,6 Millionen US-Dollar für eine Wohltätigkeitsorganisation spenden, um den armen Menschen zu helfen

dieses Ziel und erhalten Sie diese Überweisung auf Ihr Bankkonto.

Meine besten Grüße an Sie und Ihre ganze Familie. Ich erwarte Ihre Antwort. Grüße, Frau Vera Wilfred.

# Spam Mail 4.

K2 - DER KOHLEHYDRATBLOCKER

K2 Diät

Die Höhle der Löwen

14 Kilo in einem Monat sind hiermit kein Problem. Diese kleinen Tropfen verändern die Abnehmmarkt.

K2 Kundenservice



Warum diese kleinen Tropfen die Lösung für viele Gewichtsprobleme sind.

Dazu müssen Sie nur diese neue revolutionäre Methode ausprobieren, die das Fett automatisch verbrennt, und schon nach wenigen Wochen bekommen Sie eine schlanke und attraktive Figur.. Nehmen Sie ab und erhalten Sie dieses Jahr diese schlanke Figur

JETZT ZUM ANGEBOT

Anna fleurer - K2 Kundenservice,

## B.6. Variablen- und Befehlsübersicht



# Variablen und Befehls Übersicht

Auf diesem Blatt sind alle Befehle und Variable, die ihr für die Bearbeitung des Workshops Spamfilter benötigt, aufgelistet und beschrieben. Ihr könnte einen Befehl/ eine Variable kopieren, in dem ihr ihn markiert und anschließend die Tastenkombiation `Strg + c` drückt. Mit der Tastenkombination `Strg + v` könnt ihr ihn / sie wieder einfügen.

Du benötigst meist **nicht** alle Variablen oder Befehle.

## Befehls- und Variablenliste

### AB1 Teil 3

Die folgenden Befehle kannst du statt konkreten Zahlenwerten einsetzen.

Diese Befehle setzen für wort automatisch das Wort ein, welches weiter oben im Code als wort definiert wurde. Du brauchst es nicht zu ersetzen.

In [ ]:
```
gib_Anzahl_Mails_mit(wort)      # Anzahl der Mails die wort enthalten
gib_Anzahl_Mails()              # Anzahl aller Mails
gib_Anzahl_Spam_Mails()         # Anzahl aller Spam Mails
gib_Anzahl_Spam_Mails_mit(wort) # Anzahl aller Spam Mails, die wort enthalten
```

### AB2 Teil 1

Die folgenden Variabeln kannst du nutzen um Wahrscheinlichkeiten im Code darzustellen.

In [ ]:
```
P_Wort      # = P(Wort) die Wahrscheinlichkeit, dass Wort in einer Mail vorkommt.
P_Ham       # = P(Ham) die Wahrscheinlichkeit, dass eine Mail Ham ist.
P_Spam      # = P(Spam) die Wahrscheinlichkeit, dass eine Mails Spam ist.

P_Ham_Wort  # = P_Ham(Wort) Die Wahrscheinlichkeit, dass Wort in einer Mail
            # vorkommt, unter der Bedingung, dass die Mail eine Ham Mail ist.
P_Spam_Wort # = P_Spam(Wort) Die Wahrscheinlichkeit, dass Wort in einer Mail
            # vorkommt, unter der Bedingung, dass die Mail eine Spam Mail ist.
P_Wort_Ham  # = P_Wort(Ham) Die Wahrscheinlichkeit, dass eine Mail Ham Mail
            # ist, unter der Bedingung, dass die Mail das Wort enthält.
P_Wort_Spam # = P_Wort(Spam) Die Wahrscheinlichkeit, dass eine Mail Spam Mail
            # ist, unter der Bedingung, dass die Mail das Wort enthält.
```

## AB2 Teil 2 a)

Die folgenden Befehle kannst du statt konkreten Zahlenwerten einsetzen.

Die 0 am Ende der Befehle steht für den Wert der zurück gegeben wird falls es in unseren Daten das wort nicht gibt. Verändere sie nicht!

Du kannst wort durch konkrete Worte in anführungszeichen ersetzen

In [ ]:
```
get(P_Wort_liste, wort, 0)      # = P(wort) mit diesem Befehl kannst du dir aus
                                # unser Liste, die wir in Blatt 1 erstellt haben,
                                # die Wahrscheinlichkeit P(wort) für ein
                                # konkretes Wort ausgeben lassen.
get(P_Spam_Wort_liste, wort, 0) # = P_Spam(wort) mit diesem Befehl kannst du dir
                                # aus unser Liste, die wir in Blatt 1 erstellt
                                # haben,die Wahrscheinlichkeit P_Spam(wort) für
                                # ein konkretes Wort ausgeben lassen.
#Beispielnutzung
get(P_Wort_liste, "JETZT", 0)      # gibt dir die Wahrscheinlichkeit P(JETZT) aus.
get(P_Wort_Spam_liste, "JETZT", 0) # gibt dir die Wahrscheinlichkeit P(JETZT|Spam)
```

## AB2 Teil 2 b)

In [ ]:
```
P_Wort_Spam # = P_Wort(Spam) Die Wahrscheinlichkeit, dass eine Mail eine Spam Mail
            # ist, unter der Bedingung das die Mail das Wort enthält.
            # Dieser Wert wurde weiter oben im Code mit der von uns bestimmten
            # Formel berechnet. Mit P_Wort_Spam kannst du darauf zugreifen
```

## AB3 Teil 1

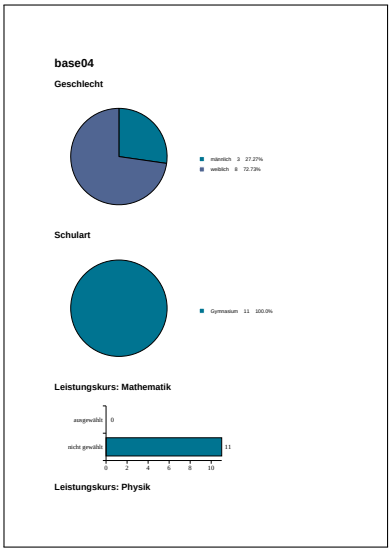
Die folgenden Variabeln kannst du nutzen um Wahrscheinlichkeiten im Code darzustellen

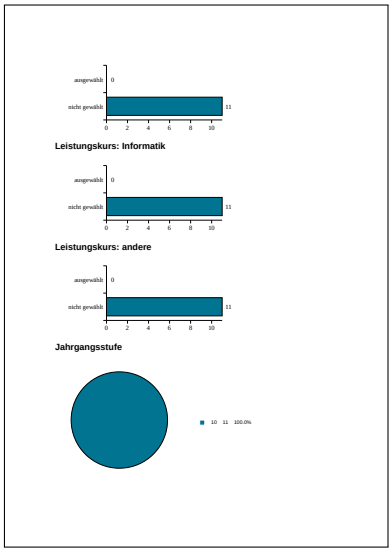
In [ ]:
```
P_W1_u_W2       # = P(Wort1 und Wort2) die Wahrscheinlichkeit, dass sowohl das
                #   Wort1 als auch das Wort2 in einer Mail vorkommen.
P_Spam_W1_u_W2 # = P_Spam(Wort1 und Wort2) die Wahrscheinlichkeit, dass sowohl
                # das Wort1 als auch das Wort2 in einer Mail vorkommen, unter
                # der Bedingung das diese Mail eine Ham Mail ist.
P_Ham_W1_u_W2  # = P_Ham(Wort1 und Wort2) die Wahrscheinlichkeit, dass sowohl
                # das Wort1 als auch das Wort2 in einer Mail vorkommen, unter
                # der Bedingung das diese Mail eine Ham Mail ist.

P_Wort          # = P(Wort) die Wahrscheinlichkeit, dass ein Wort in einer
                # Mail vorkommt.
P_Ham           # = P(Ham) die Wahrscheinlichkeit, dass eine Mail Ham ist.
P_Spam          # = P(Spam) die Wahrscheinlichkeit, dass eine Mails Spam ist.
```

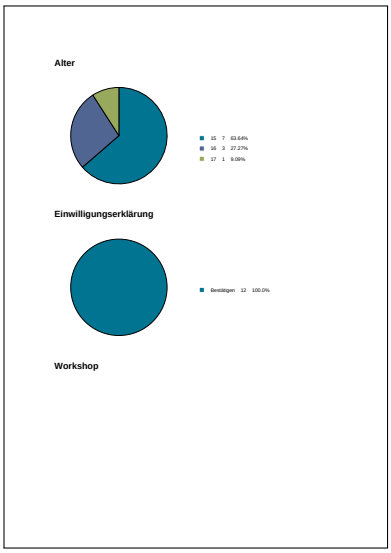
## AB3 Teil 2

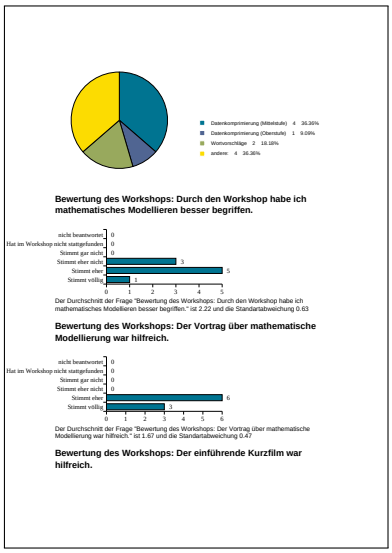
In [ ]:
```
wsv # = Wahrscheinlichkeitsverhältniss. Das Wahrscheinlichkeitsverhältniss
    # wurde weiter oben im Code mit der von uns bestimmten Formel berechnet.
    # Mit wsv kannst du auf den berechneten Wert zugreifen.
```

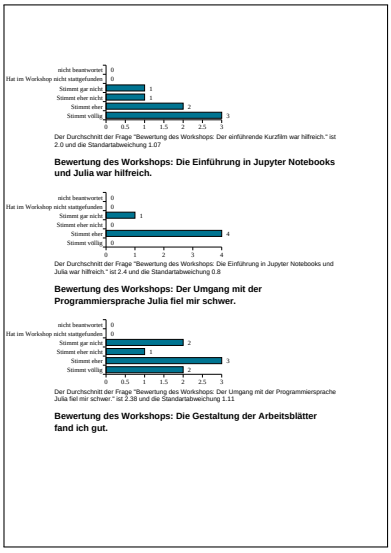
# C. Evaluation

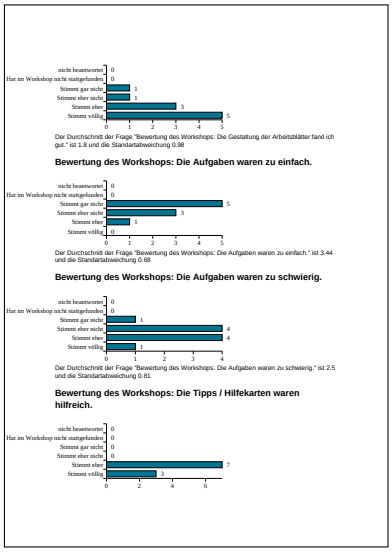


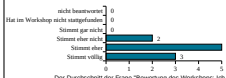



Der Durchschnitt der Frage "Bewertung des Workshops: Die Tipps / Hilfekarten waren hilfreich." ist 1.7 und die Standartabweichung 0.46

**Bewertung des Workshops: Ich konnte die Aufgaben eigenständig bearbeiten.**

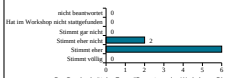

Der Durchschnitt der Frage "Bewertung des Workshops: Ich konnte die Aufgaben eigenständig bearbeiten." ist 2.5 und die Standartabweichung 0.81

**Bewertung des Workshops: Ich fand die automatischen Rückmeldungen, die ich nach der Eingabe meiner Lösungen im Code erhalten habe, hilfreich.**



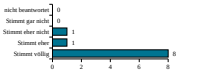

Der Durchschnitt der Frage "Bewertung des Workshops: Ich fand die automatischen Rückmeldungen, die ich nach der Eingabe meiner Lösungen im Code erhalten habe, hilfreich." ist 1.9 und die Standartabweichung 0.7

**Bewertung des Workshops: Die Zusatzaufgaben haben mir sehr gefallen.**



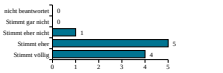

Der Durchschnitt der Frage "Bewertung des Workshops: Die Zusatzaufgaben haben mir sehr gefallen." ist 2.25 und die Standartabweichung 0.43

**Weiterführende Fragen: Die Lern- und Arbeitsatmosphäre war angenehm.**



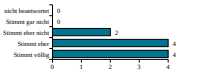

Der Durchschnitt der Frage "Weiterführende Fragen: Die Lern- und Arbeitsatmosphäre war angenehm." ist 1.3 und die Standartabweichung 0.64

**Weiterführende Fragen: Die Inhalte wurden klar vermittelt.**



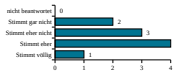

Der Durchschnitt der Frage "Weiterführende Fragen: Die Inhalte wurden klar vermittelt." ist 1.7 und die Standartabweichung 0.64

**Weiterführende Fragen: Die Lern- und Arbeitszeiten waren angemessen (nicht zu lang oder zu kurz, mit ausreichenden Pausen).**



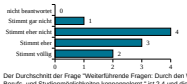

Der Durchschnitt der Frage "Weiterführende Fragen: Die Lern- und Arbeitszeiten waren angemessen (nicht zu lang oder zu kurz, mit ausreichenden Pausen)." ist 1.8 und die Standartabweichung 0.75

**Weiterführende Fragen: Das Projekt hat mein Interesse an Themen der Naturwissenschaften und Technik gesteigert.**



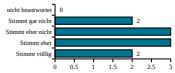

Der Durchschnitt der Frage "Weiterführende Fragen: Das Projekt hat mein Interesse an Themen der Naturwissenschaften und Technik gesteigert." ist 2.6 und die Standartabweichung 0.92

**Weiterführende Fragen: Durch den Workshop habe ich interessante Berufs- und Studienmöglichkeiten kennengelernt.**



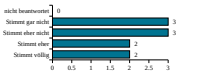

Der Durchschnitt der Frage "Weiterführende Fragen: Durch den Workshop habe ich interessante Berufs- und Studienmöglichkeiten kennengelernt." ist 2.4 und die Standartabweichung 0.92

**Weiterführende Fragen: Ich habe in diesem Workshop viel Neues gelernt, was mir für die Schule, für ein Studium und / oder Beruf weiterhelfen kann.**



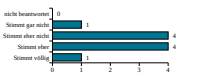

Der Durchschnitt der Frage "Weiterführende Fragen: Ich habe in diesem Workshop viel Neues gelernt, was mir für die Schule, für ein Studium und / oder Beruf weiterhelfen kann." ist 2.5 und die Standartabweichung 1.02

**Weiterführende Fragen: Ich kann mir vorstellen ein Studium oder eine Ausbildung im Bereich der Naturwissenschaften oder der Technik zu beginnen.**



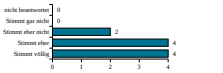

Der Durchschnitt der Frage "Weiterführende Fragen: Ich kann mir vorstellen ein Studium oder eine Ausbildung im Bereich der Naturwissenschaften oder der Technik zu beginnen." ist 2.7 und die Standartabweichung 1.1

**Weiterführende Fragen: Ich würde so einen Workshop gerne noch einmal besuchen.**



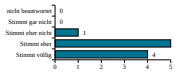

Der Durchschnitt der Frage "Weiterführende Fragen: Ich würde so einen Workshop gerne noch einmal besuchen." ist 2.5 und die Standartabweichung 0.81

**Weiterführende Fragen: Ich würde so einen Workshop auch anderen weiterempfehlen.**



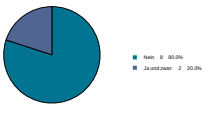

Der Durchschnitt der Frage "Weiterführende Fragen: Ich würde so einen Workshop auch anderen weiterempfehlen." ist 1.8 und die Standartabweichung 0.75

**Weiterführende Fragen: Alles in allem hat mir der Workshop gut gefallen.**




Der Durchschnitt der Frage "Weiterführende Fragen: Alles in allem hat mir der Workshop gut gefallen." ist 1.7 und die Standartabweichung 0.64

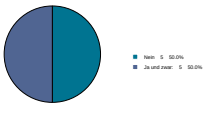
**Nicht gefallen**




**Nicht gefallen: Ja und zwar**

Einträge bei der Frage "Nicht gefallen: Ja und zwar" waren:
Dass das Programmiersystem nicht wirklich erklärt wurde
Programme nicht genug erklärt

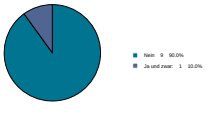
**Besonders gefallen**




**Besonders gefallen: Ja und zwar**

Einträge bei der Frage "Besonders gefallen: Ja und zwar" waren:
Das selbstständige arbeiten
Der Einstieg
Arbeit, Rechnen

**Anderes gesehen**



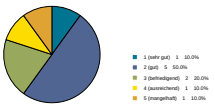

**Lernzuwachs: [01]**

Einträge bei der Frage "Lernzuwachs: [01]" waren:
Wie das aussortieren von Soam Mails funktioniert und wie man ungefähr programmiert
Man kann sehr viel mit programmieren machen
/
Das alles Mathemathik ist
Etwas Programmieren
Wie man Spam von Ham Mails unterscheidet
Selbständiges erarbeiten von neuen Dingen
Wie Spam gefilter wird und der Begriff „Spam" von einem Sketch kommt

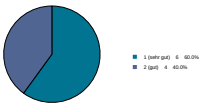
**Modellierung: [01]**

Einträge bei der Frage "Modellierung: [01]" waren:
Keine Ahnung
Das man etwas erstellt
Wanderung von Wallfischen
keine ahnung
Problem in der Realität in der Mathematik beschreiben. Vereinfachen von Sachzusammenhängen
1.2.3.
K

**Note Workshop**




**Note Betreuung**

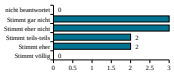



**Abschließender Kommentar: [01]**

Einträge bei der Frage "Abschließender Kommentar: [01]" waren:
Betreuer waren sehr lieb und haben immer versucht zu helfen wenn es nötig war
Leiter waren freundlich
war super

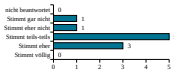
Sehr gut
Lob

**Teilnahme am CAMMP Workshop: Ich werde in Büchern, dem Internet oder anderen Medien recherchieren, um mehr Informationen über die im Workshop behandelten Themen zu bekommen.**



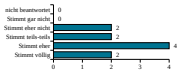

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Ich werde in Büchern, dem Internet oder anderen Medien recherchieren, um mehr Informationen über die im Workshop behandelten Themen zu bekommen." ist 3.7 und die Standartabweichung 1.1

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich in der Lage war, die Aufgaben alleine zu bearbeiten.**



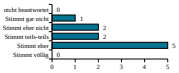

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich in der Lage war, die Aufgaben alleine zu bearbeiten." ist 3.0 und die Standartabweichung 0.89

**Teilnahme am CAMMP Workshop: Die Beschäftigung mit den Themen des Workshops hat Spaß gemacht.**



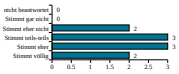

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Die Beschäftigung mit den Themen des Workshops hat Spaß gemacht." ist 2.4 und die Standartabweichung 1.02

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass der Workshop so war, wie ich es mir wünsche.**



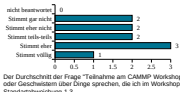

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass der Workshop so war, wie ich es mir wünsche." ist 2.9 und die Standartabweichung 1.04

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Mitschüler*innen meine Leistungen anerkannt haben.**



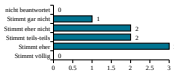

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Mitschüler*innen meine Leistungen anerkannt haben." ist 2.5 und die Standartabweichung 1.02

**Teilnahme am CAMMP Workshop: Ich werde mit Freunden, Eltern oder Geschwistern über Dinge sprechen, die ich im Workshop erlebt habe.**



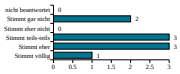

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Ich werde mit Freunden, Eltern oder Geschwistern über Dinge sprechen, die ich im Workshop erlebt habe." ist 3.1 und die Standartabweichung 1.3

**Teilnahme am CAMMP Workshop: Dass wir im Workshop Modellierungen durchgeführt haben, ist mir persönlich wichtig.**



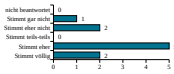

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Dass wir im Workshop Modellierungen durchgeführt haben, ist mir persönlich wichtig." ist 3.12 und die Standartabweichung 1.05

**Teilnahme am CAMMP Workshop: Der Workshop hat meine Neugier geweckt.**




Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Der Workshop hat meine Neugier geweckt." ist 2.89 und die Standartabweichung 1.29

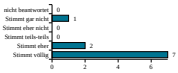
**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass mich meine Mitschüler*innen ernst genommen haben.**




Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass mich meine Mitschüler*innen ernst genommen haben." ist 2.5 und die Standartabweichung 1.28

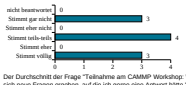
**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Betreuer*innen mich an schwierigen Stellen**

**im Workshop unterstützt haben.**



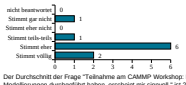

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Betreuer*innen mich an schwierigen Stellen im Workshop unterstützt haben." ist 1.6 und die Standartabweichung 1.2

**Teilnahme am CAMMP Workshop: Während des Workshops haben sich neue Fragen ergeben, auf die ich gerne eine Antwort hätte.**



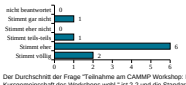

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Während des Workshops haben sich neue Fragen ergeben, auf die ich gerne eine Antwort hätte." ist 3.0 und die Standartabweichung 1.55

**Teilnahme am CAMMP Workshop: Dass wir im Workshop Modellierungen durchgeführt haben, erscheint mir sinnvoll.**



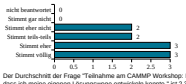

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Dass wir im Workshop Modellierungen durchgeführt haben, erscheint mir sinnvoll." ist 2.2 und die Standartabweichung 1.08

**Teilnahme am CAMMP Workshop: Insgesamt fühlte ich mich in der Kursgemeinschaft des Workshops wohl.**



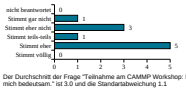

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Insgesamt fühlte ich mich in der Kursgemeinschaft des Workshops wohl." ist 2.2 und die Standartabweichung 1.08

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich meine eigenen Lösungswege entwickeln konnte.**



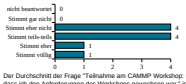

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich meine eigenen Lösungswege entwickeln konnte." ist 2.3 und die Standartabweichung 1.1

**Teilnahme am CAMMP Workshop: Die Inhalte des Workshops sind für mich bedeutsam.**



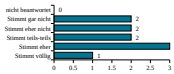

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Die Inhalte des Workshops sind für mich bedeutsam." ist 3.0 und die Standartabweichung 1.1

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich den Anforderungen des Workshops gewachsen war.**



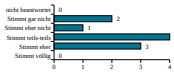

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich den Anforderungen des Workshops gewachsen war." ist 3.1 und die Standartabweichung 0.94

**Teilnahme am CAMMP Workshop: Die Modellierungen im Workshop haben mich fasziniert.**



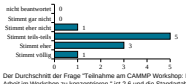

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Die Modellierungen im Workshop haben mich fasziniert." ist 3.1 und die Standartabweichung 1.3

**Teilnahme am CAMMP Workshop: Ich werde außerhalb des Unterrichts über Dinge nachdenken, die wir im Workshop gesehen oder angesprochen haben.**



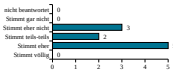

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Ich werde außerhalb des Unterrichts über Dinge nachdenken, die wir im Workshop gesehen oder angesprochen haben." ist 3.2 und die Standartabweichung 1.08

**Teilnahme am CAMMP Workshop: Es fiel mir leicht, mich auf meine Arbeit im Workshop zu konzentrieren.**



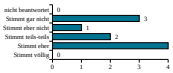

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Es fiel mir leicht, mich auf meine Arbeit im Workshop zu konzentrieren." ist 2.6 und die Standartabweichung 0.8

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass dieser meinen Zielen für den Workshop entsprach.**



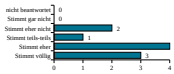

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass dieser meinen Zielen für den Workshop entsprach." ist 2.8 und die Standartabweichung 0.87

**Teilnahme am CAMMP Workshop: Ich würde gerne mehr über die Modellierungen lernen, die wir im Workshop durchgeführt haben.**



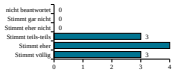

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Ich würde gerne mehr über die Modellierungen lernen, die wir im Workshop durchgeführt haben." ist 3.3 und die Standartabweichung 1.27

**Teilnahme am CAMMP Workshop: Während des Workshops verging die Zeit wie im Flug.**



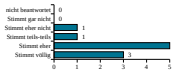

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Während des Workshops verging die Zeit wie im Flug." ist 2.2 und die Standartabweichung 1.08

**Teilnahme am CAMMP Workshop: Während des Modellierens im Workshop hatte ich ein Aha-Erlebnis.**



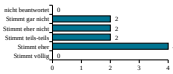

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Während des Modellierens im Workshop hatte ich ein Aha-Erlebnis." ist 2.0 und die Standartabweichung 0.77

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich die Möglichkeit hatte selbstständig zu arbeiten.**



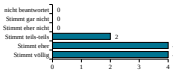

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich die Möglichkeit hatte selbstständig zu arbeiten." ist 2.0 und die Standartabweichung 0.89

**Teilnahme am CAMMP Workshop: Die Themen des Workshops sind mir wichtig.**



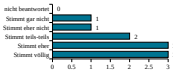

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Die Themen des Workshops sind mir wichtig." ist 3.2 und die Standartabweichung 1.17

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Betreuer*innen mich ernst genommen haben.**



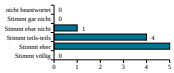

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass meine Betreuer*innen mich ernst genommen haben." ist 1.8 und die Standartabweichung 0.75

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich selbst entscheiden konnte, wie ich eine Aufgabe bearbeite.**



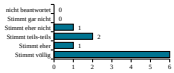

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich selbst entscheiden konnte, wie ich eine Aufgabe bearbeite." ist 2.4 und die Standartabweichung 1.28

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich auch den schwierigen Stoff verstanden habe.**



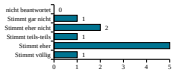

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich auch den schwierigen Stoff verstanden habe." ist 2.6 und die Standartabweichung 0.66

**Teilnahme am CAMMP Workshop: Insgesamt fühlte ich mich bei meinen Betreuer*innen wohl.**



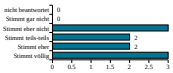

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Insgesamt fühlte ich mich bei meinen Betreuer*innen wohl." ist 1.8 und die Standartabweichung 1.08

**Teilnahme am CAMMP Workshop: Die Beschäftigung mit den Inhalten des Workshops war für mich nützlich.**



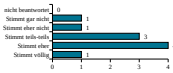

Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Die Beschäftigung mit den Inhalten des Workshops war für mich nützlich." ist 2.7 und die Standartabweichung 1.19

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich neue Inhalte selbstständig erarbeiten konnte.**




Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass ich neue Inhalte selbstständig erarbeiten konnte." ist 2.5 und die Standartabweichung 1.2

**Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass der Workshop so war, wie ich es mir vorstelle.**




Der Durchschnitt der Frage "Teilnahme am CAMMP Workshop: Im Workshop hatte ich das Gefühl, dass der Workshop so war, wie ich es mir vorstelle." ist 2.7 und die Standartabweichung 1.1

# Abbildungsverzeichnis

# Literatur


Bayes, T. (1763). *An essay towards solving a problem in the doctrine of chances* (Bd. 53). The Royal Society. (Communicated by Richard Price) doi: 10.1098/rstl.1763.0053

Bildungsplan Mathematik Baden-Württemberg. (2016). *Bildungsplan Mathematik.* Stuttgart. Zugriff auf `http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GYM_M.pdf` (letzter Aufruf am 04.03.2023)

CAMMP. (2019). *Willkommen bei CAMMP!* Zugriff auf `https://www.cammp.online/index.php` (letzter Aufruf am 14.04.2021)

Greefrath, G., Kaiser, G., Blum, W. & Ferri, R. B. (2013). Mathematisches Modellieren - Eine Einführung in theoretische und didaktische Hintergründe. In W. Blum, R. B. Ferri, G. Greefrath & G. Kaiser (Hrsg.), *Mathematisches Modellieren für Schule und Hochschule - Theoretische und didaktische Hintergrüde* (S. 11-37). Wiesbaden: Springer Spektrum.

Wassner, C., Krauss, S. & Martignon, L. (2002). Muss der satz von bayes schwer verständlich sein? *Praxis der Mathematik*, *44* (1), 12–16.

## Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus der Arbeit anderer unverändert oder mit Abänderungen entnommen wurde, sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 27.03.2023