Der k-nächste-Nachbarn-Algorithmus

Anleitung zur Verwendung der TK-Mappen

Material zur Lehrerfortbildung KI@Informatik11

Konzeption der Anleitung:

Die Anleitung dient zur Erklärung der in den Lehrerfortbildungen „KI@Informatik 11 – was, wozu, wie, womit unterrichten“ der Didaktik der Informatik der Universität Passau zum Thema „Künstliche Intelligenz“ in der 11. Jahrgangsstufe verwendeten Tabellenkalkulationsmappen zum Thema „k-nächste-Nachbarn-Algortihmus“.

Dr. Wolfgang Pfeffer

Dominicus-von-Linprun-Gymnasium Viechtach
E-Mail: schule@pfeffer-wolfgang.de

Tobias Fuchs

Universität Passau
E-Mail: fuchs_unipa@outlook.de

Weiteres Material findet sich im Mebis-Kurs „Material zu den KI-Fortbildungen der Universität Passau“ mit der `ID`-Nummer `1324361`. Den Einschreibeschlüssel zum Kurs erhalten Sie auf Anfrage.

Die auf dem Deckblatt enthaltene Grafik ist der Plattform `www.pixabay.com` entnommen und wurde weiter bearbeitet. Diese steht unter der dort aufgeführten Lizenz.
Die im Handbuch abgebildeten Screenshots zeigen sämtlich die in Microsoft Excel geöffneten Tabellenkalkulationsmappen. Gemäß des Leitfadens zur Verwendung von Screenshots von Microsoft-Produkten wird folgende Erklärung aufgeführt: *Used with permission from Microsoft.*

Inhaltsverzeichnis

Überblick

Für die Lehrerfortbildung *KI@Informatik11 - Was, wozu, wie, womit unterrichten?* wurden zur Visualisierung und Simulation des maschinellen Lernprozesses am Beispiel des **k-nächste-Nachbarn-Algorihmus** (kNN-Algorithmus) Tabellenkalkulationsmappen (TK-Mappen) (speziell für Microsoft Excel und LibreOffice Calc) erstellt, welche es ermöglichen, einzelne Schritte im maschinellen Lernprozess zu simulieren und selbst auszuprobieren. Diese Mappen stehen in zwei Versionen mit unterschiedlichen Anwendungskontexten zur Verfügung.

Verfügbare Kontexte:

- **Klassifikation von T-Shirtgrößen** (S, M, L) anhand der Merkmale *Körpergröße* und *Brustumfang*
- **Klassifikation von Irispflanzenarten** (Setosa, Versicolor, Virginica) anhand der Merkmale *(Kelch-) Blattlänge* und *(Kelch-) Blattbreite*

Eine ausführliche Erläuterung der Anwendungskontexte finden Sie im Mebis-Kurs zur Fortbildung. Nicht für Eingaben vorgesehene Bereiche der Mappen sind gesperrt. Das Passwort zur Entsperrung der Mappen lautet **KI@11**

Hinweis:
Für eine Entsperrung der Versionen mit dynamischer Skalierung sollte das Setzen des Passworts auch im VBA-Code auskommentiert werden).

Aufbau der TK-Mappen zur Anwendung Klassifikation

Die TK-Mappen sind jeweils in Tabellen unterteilt, welche jeweils einen Aspekt des maschinellen Lernprozesses des kNN-Algorithmus veranschaulichen.

1. Klassifikation eines neuen, einzugebenden Datenpunkts mit Hilfe des kNN-Algorithmus
2. Training des kNN-Modells
3. Bestimmung des Hyperparameters k für den kNN-Algorithmus allgemein
4. Bestimmung des Hyperparameters k für den kNN-Algorithmus mit Hilfe des *Leave-One-Out-Verfahrens*
5. Testen des trainierten Modells und Darstellung der Ergebnisse in Form einer Konfusionsmatrix

Des Weiteren wird am Ende der Anleitung eine weitere TK-Mappe vorgestellt, welche die Simulation einer weiteren Anwendung des k-nächste-Nachbarn-Algorithmus, die **Regression** ermöglicht. Hier wurde der Anwendungskontext *Vorhersage eines Hauspreises* in Abhängigkeit des Merkmals *Fläche in* m^2 gewählt.

1. Klassifikation eines neuen Datenpunkts

Der kNN-Algorithmus eignet sich zur Klassifikation von Datenpunkten, d.h. die Einordnung eines Datenpunkts anhand seiner Merkmale in eine vorhandene Klasse. Im verwendeten Kontext wird ein neuer Datenpunkt mit den Merkmalen *Körpergröße* und *Brustumfang* in eine der Klassen *S, M, L* eingeordnet. Mit Hilfe der TK-Mappe kann dies in der Tabelle **Klassifikation** (Reiterauswahl im unteren Bereich der Mappe) simuliert werden.

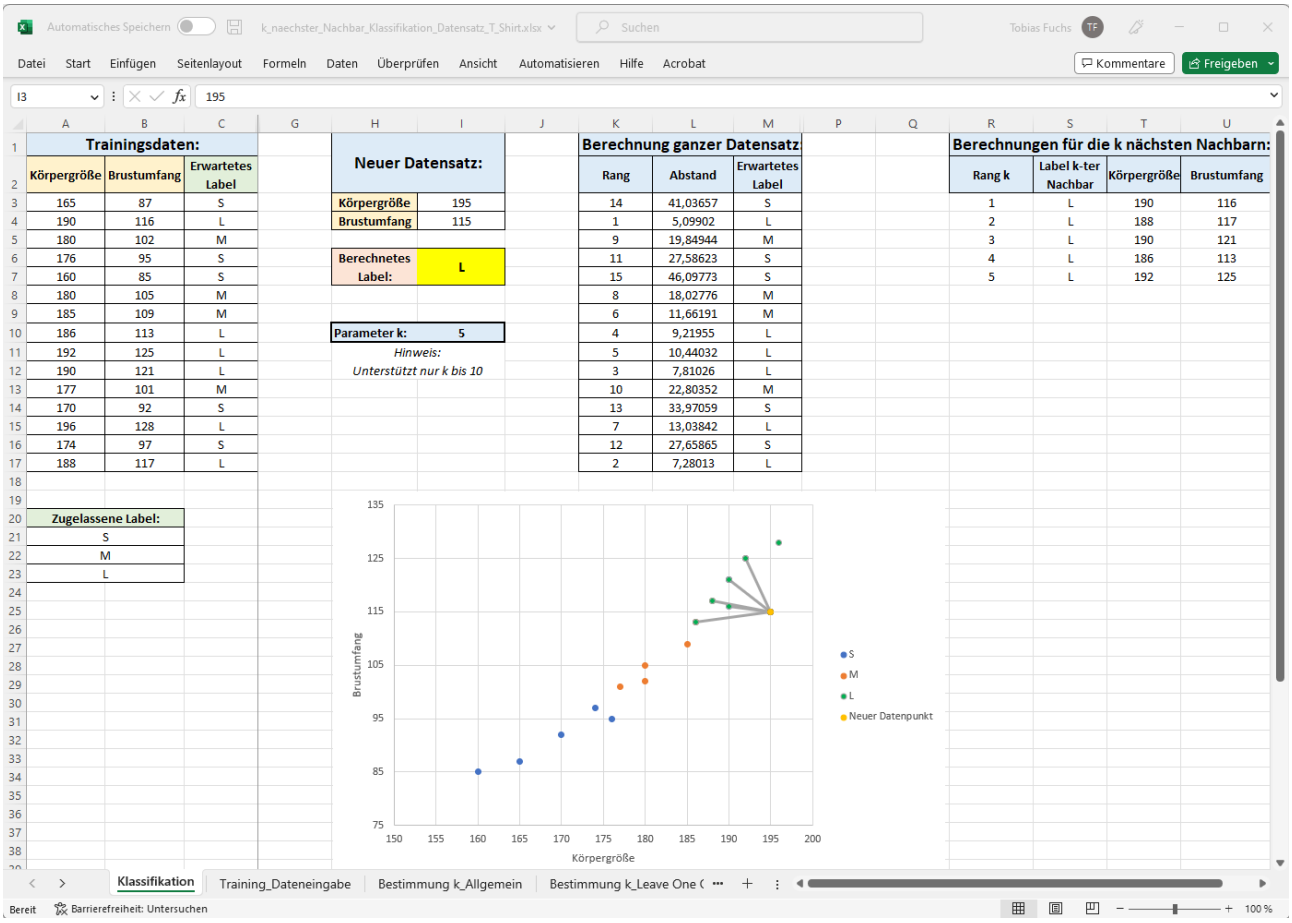


Abbildung 1: Screenshot der Tabelle zur Klassifikation eines neuen Datenpunkts (geöffnet in Microsoft Excel).

Vorgehen:

1. **Eingabe** des zu **klassifizierenden Datenpunkts**:
   Geben Sie unter *Neuer Datensatz* die von Ihnen gewünschten Werte für die Merkmale *Körpergröße* und *Brustumfang* ein.

2. **Eingabe** des **Parameters** k (d.h. Festlegen der Anzahl an betrachteten Nachbarn):
   Geben Sie neben der Zelle mit Text *Parameter k* den von Ihnen gewünschten Wert für k ein. Es werden nur ganzzahlige Werte von 1 bis 10 unterstützt.

3. **Ablesen** der **gefundenen Klasse**:
   Neben der Zelle mit Text *Berechnetes Label* wird die berechnete Klasse für den eingegebenen Datenpunkt angezeigt(gelbes Feld).

Im unteren Bereich werden Ihnen die vorhanden Trainingsdaten und der zu klassifizierende Datenpunkt (gelb) in einem Koordinatensystem angezeigt. Des Weiteren werden die k Trainingsdatenpunkte, die dem zu klassifizie-

renden Datenpunkt am nächsten liegen, mit diesem durch Linien verbunden. Die k nächsten Nachbarn des zu klassifizierenden Datenpunkts sind also direkt sichtbar.

Im Bereich **Trainingsdaten** können Sie die verwendeten Trainingsdaten verändern. Verwenden Sie in der Spalte *Erwartetes Label* nur Werte, welche im Bereich *Zugelassene Label* spezifiziert sind.

Den einzelnen Trainingsdatenpunkten wird dann ein Rang entsprechend des jeweiligen Abstands zu dem zu klassifizierenden Datenpunkt zugewiesen. Der diesem Punkt am nächsten liegende Trainingsdatenpunkt erhält Rang 1, der Trainingsdatenpunkt mit dem nächstgrößeren Abstand erhält Rang 2, usw.
Im Bereich **Berechnung für die k nächsten Nachbarn** werden die k nächsten Nachbarn aufgrund des festgelegten Rangs angezeigt.

Im Bereich **Berechnung ganzer Datensatz** werden die Abstände zwischen dem zu klassifizierenden Datenpunkt und jedem Trainingsdatenpunkt mit Hilfe des euklidischen Abstands bestimmt. Sie können die Formeln für die Bestimmung der Abstände auch anpassen, um eine andere Metrik, wie etwa die Manhattan-Metrik für die Abstandsberechnung zu verwenden.
Die folgende Formel realisiert die Abstandsberechnung unter Verwendung des euklidischen Abstands. Diese Formel muss in Zelle `L3` eingetragen werden und kann aufgrund der Verwendung von absoluten und relativen Zellbezügen einfach mit der *Automatisch Ausfüllen* - Funktionalität („kleines schwarzes Kästchen“) des Tabellenkalkulationsprogramms nach unten ausgefüllt werden.

```
=WURZEL(($I$3-A3)^2+($I$4-B3)^2)+ZEILE()/1000000
```

Anmerkung zur Addition von `ZEILE()/1000000`
Da es sein kann, dass zwei Abstände exakt gleich sind, dann aber die Auswahl eines eindeutigen n-ten Nachbarn nicht mehr möglich ist, wird in obiger Formel zu jedem Abstand ein minimaler, von der aktuellen Zeile abhängiger Wert mit ZEILE()/1000000 addiert. Dies reduziert die Wahrscheinlichkeit des Auftretens von exakt gleichen Abständen wesentlich und hat auf die Vergleichbarkeit der Abstände insgesamt kaum einen Einfluss.

2. Training des k-nächste-Nachbarn-Modells

Das Trainieren des Modells beim k-nächste-Nachbarn-Algorithmus ist, im Vergleich zu anderen Algorithmen maschinellen Lernens, denkbar einfach. Hier müssen nur die Trainingsdaten geladen werden und das Training des Modells ist abgeschlossen.

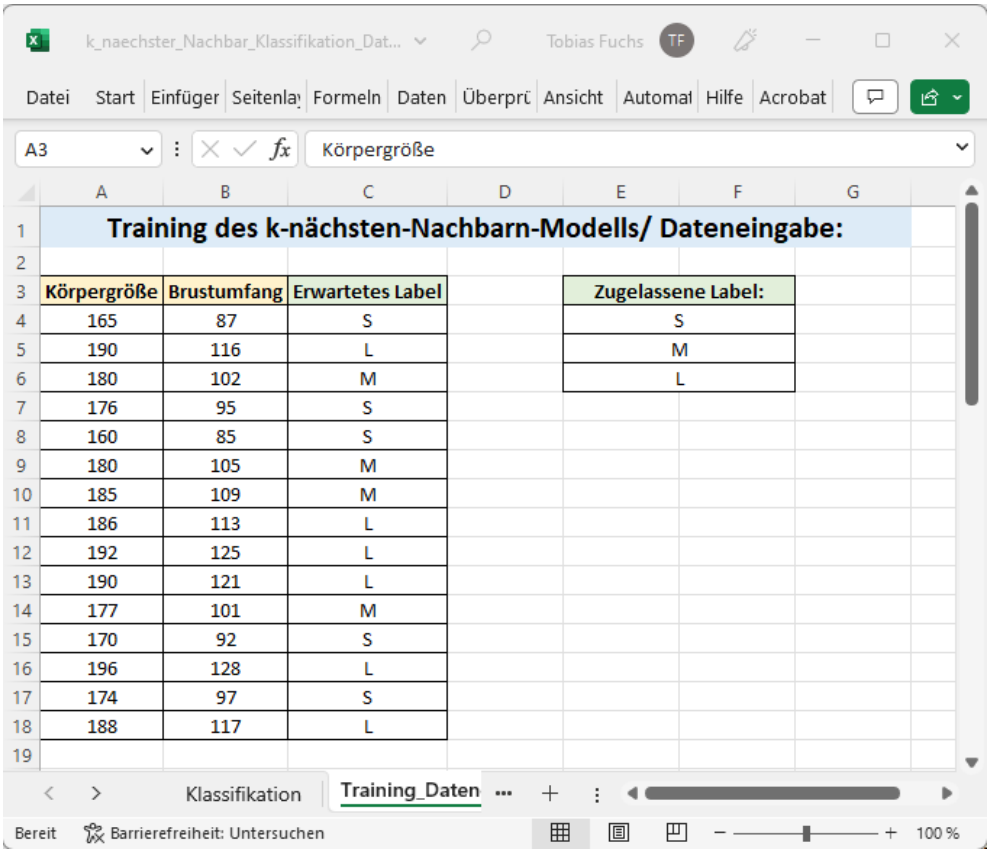


Abbildung 2: Screenshot der Tabelle zum Training des kNN-Modells (geöffnet in Microsoft Excel).

Vorgehen:

1. **Füllen** der Tabelle **mit** (vorbereiteten) **Trainingsdaten**:
   - Tragen Sie in der ersten Zeile der Tabelle (*gelbe Zellen*) die betrachteten Merkmale ein.
   - Tragen Sie darunter die Werte der Trainingsdaten dieser Merkmale ein. In der letzten Spalte der Tabelle (unterhalb *Erwartetes Label*) soll das für die Merkmalswerte erwartete Label eingetragen werden. Dabei sind nur Werte aus dem Bereich *Zugelassene Label* zulässig.
2. *(Optional)* **Anpassung** der **zugelassenen Label**:
   Sollten Sie andere Label in Ihren Trainingsdaten verwenden wollen, so ersetzen Sie dafür die Einträge im Bereich *Zugelassene Label*. Bisher werden nur drei Label gleichzeitig unterstützt.

3. Bestimmung des Parameters k – Allgemein

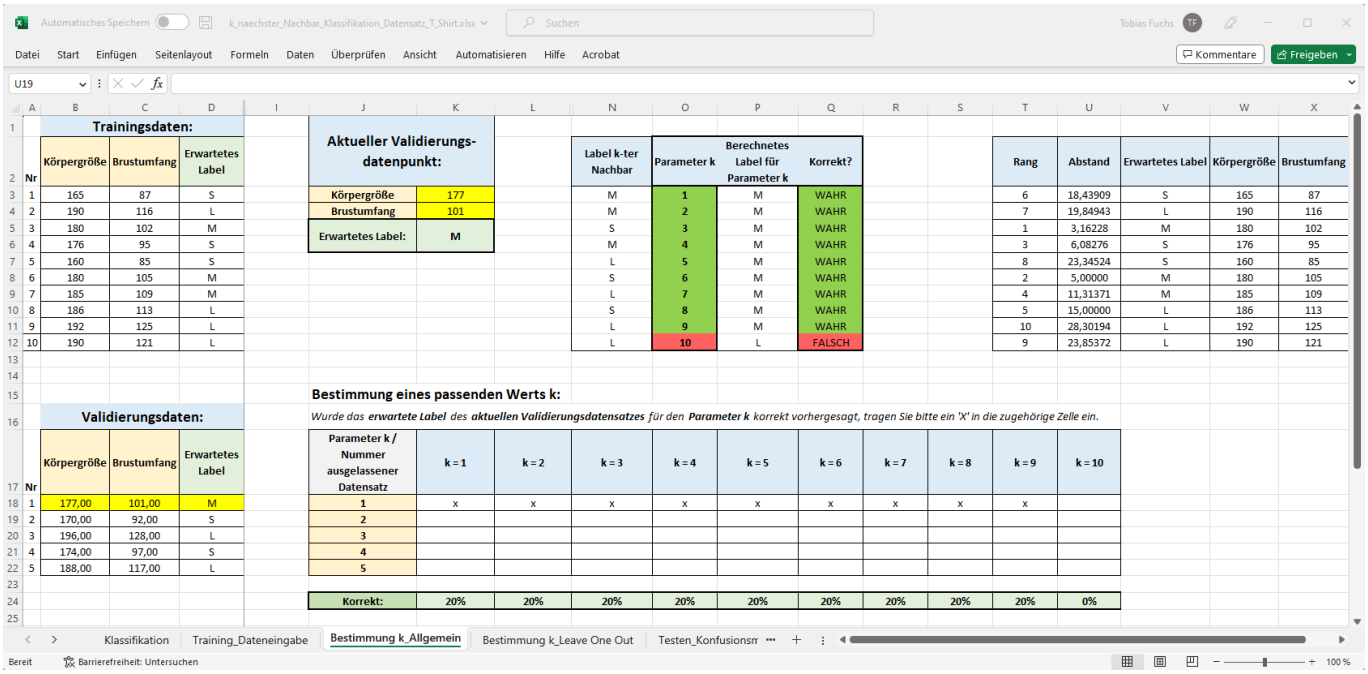
Der Hyperparameter k legt fest, wie viele Nachbarn eines zu klassifizierenden Datenpunkts für die Klassifikation, d.h. die Einordnung in eine Klasse, herangezogen werden. Die Wahl des Parameters k ist also von zentraler Bedeutung für den k-nächste-Nachbarn-Algorithmus. Die Auswahl des Parameters erfolgt in der Regel nicht zufällig, sondern es werden passende Werte unter Verwendung der **Validierungsdaten** bestimmt.

Das Verfahren kurz gefasst...

Die einzelnen Validierungsdatenpunkte werden der Reihe nach für verschiedene Werte von k mit Hilfe der Trainingsdaten / des trainierten Modells klassifiziert. Im Anschluss daran wird geprüft, für welche Werte von k der aktuell betrachtete Validierungsdatenpunkt korrekt klassifiziert wurde. Dieses Vorgehen wird für alle Validierungsdaten sukzessive durchgeführt. Die Werte für k, welche am häufigsten zu einer korrekten Klassifikation geführt haben, sind unsere Kandidaten für den Wert von k.
Für eine detaillierte Betrachtung des Verfahrens wird auf die Fortbildung verwiesen. In diesem Dokument liegt der Fokus auf der Verwendung der TK-Mappen um das Vorgehen zu simulieren.

Situation:

In der Tabelle werden die im vorherigen Schritt eingegebenen Trainingsdaten bereits automatisch in *Trainingsdaten* und *Validierungsdaten* aufgeteilt. Diese aufgeteilten Daten werden in der Tabelle links in den jeweiligen Bereichen angezeigt.



Trainingsdaten:

Nr	Körpergröße	Brustumfang	Erwartetes Label
1	165	87	S
2	190	116	L
3	180	102	M
4	176	95	S
5	160	85	S
6	180	105	M
7	185	109	M
8	186	113	L
9	192	125	L
10	190	121	L

Aktueller Validierungsdatenpunkt:

Körpergröße	177
Brustumfang	101
Erwartetes Label:	M

Label k-ter Nachbar	Parameter k	Berechnetes Label für Parameter k	Korrekt?
M	1	M	WAHR
M	2	M	WAHR
S	3	M	WAHR
M	4	M	WAHR
L	5	M	WAHR
S	6	M	WAHR
L	7	M	WAHR
S	8	M	WAHR
L	9	M	WAHR
L	10	L	FALSCH

Rang	Abstand	Erwartetes Label	Körpergröße	Brustumfang
6	18,43909	S	165	87
7	19,84943	L	190	116
1	3,16228	M	180	102
3	6,08276	S	176	95
8	23,34524	S	160	85
2	5,00000	M	180	105
4	11,31371	M	185	109
5	15,00000	L	186	113
10	28,30194	L	192	125
9	23,85372	L	190	121

Bestimmung eines passenden Werts k:

*Wurde das **erwartete Label** des **aktuellen Validierungsdatensatzes** für den **Parameter k** korrekt vorhergesagt, tragen Sie bitte ein 'X' in die zugehörige Zelle ein.*

Validierungsdaten:

Nr	Körpergröße	Brustumfang	Erwartetes Label
1	177,00	101,00	M
2	170,00	92,00	S
3	196,00	128,00	L
4	174,00	97,00	S
5	188,00	117,00	L

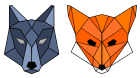
Parameter k / Nummer ausgelassener Datensatz	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
1	x	x	x	x	x	x	x	x	x	
2										
3										
4										
5										
Korrekt:	20%	20%	20%	20%	20%	20%	20%	20%	20%	0%

Abbildung 3: Screenshot der Tabelle zur Bestimmung des Hyperparameters k (geöffnet in Microsoft Excel)

Vorgehen:

1. **Auswahl** des aktuellen **Validierungsdatenpunkts**:

   Wählen Sie einen Validierungsdatenpunkt aus den Validierungsdaten aus und tragen Sie diesen als



Aktueller Validierungsdatenpunkt ein. Es müssen dabei lediglich die Werte für die Merkmale *Körpergröße* und *Brustumfang* eingetragen werden. Wird der eingetragene Datenpunkt in den Validierungsdaten gefunden, wird das *erwartete Label* automatisch übernommen und der entsprechende Datenpunkt in den Validierungsdaten gelb hervorgehoben.
Im Screenshot wird dies für den Validierungsdatenpunkt mit Körpergröße 177 und Brustumfang 101 dargestellt.

2. **Ablesen** der **Klassifikation für verschiedene Werte von k**:

   Die Tabelle klassifiziert den eingetragenen Validierungsdatenpunkt automatisch für $k = 1, ..., 10$ und vergleicht das *berechnete Label* (Spalte F) mit dem *erwarteten Label* des Datenpunkts. Des Weiteren wird direkt angezeigt, ob für einen Wert von k die Klassifikation korrekt war, sprich ob das berechnete Label mit dem erwarteten Label des Validierungsdatenpunkts übereinstimmt(Spalte G).

   Hinweis:
   Die Einträge im Bereich *Label k-ter Nachbar* stellen das Label des einzelnen k-ten Nachbarn des aktuell betrachteten Validierungsdatenpunkts dar. So ist es nachvollziehbar, wie das Modell auf die jeweilige Klassifikation für den betrachteten Wert von k kommt.

3. **Eintragen** des **Klassifikationsergebnisses für verschiedene Werte von** k:
   Tragen Sie im Bereich *Bestimmung eines passenden Werts k* für die Werte von k ein **X** ein, für die die Klassifikation korrekt war.
   Im Screenshot sehen Sie die Eintragung für den ersten Validierungsdatenpunkt. Für $k = 1, ..., 9$ ist die Klassifikation korrekt (Eintrag X). Für $k = 10$ ist die Klassifikation fehlerhaft (Kein Eintrag in der entsprechenden Zelle).

Führen Sie dieses Vorgehen für alle Validierungsdatenpunkte durch.

4. **Ablesen** der **besten Werte** für k:

   In Zeile 24 wird prozentual dargestellt, wie viele Validierungsdatenpunkte für die jeweiligen Werte von k korrekt vorhergesagt wurden (grün). Die Werte von k, welche hier den höchsten Prozentsatz aufweisen, sind unsere Kandidaten für k.

Im rechten, oberen Bereich der Tabelle werden Informationen zu den 10 nächsten Nachbarn des aktuell betrachteten Validierungsdatenpunkts angezeigt. Diese dienen lediglich der zusätzlichen Information und haben keine direkte Bedeutung für die Bestimmung der Kandidaten für den Parameter k.

4. Bestimmung des Hyperparameters k – Leave-One-Out-Verfahren

Wie im vorherigen Kapitel beschrieben, werden die Validierungsdaten verwendet um geeignete Werte für den Parameter k zu finden. Nachdem dieser Parameter k gefunden wurde, sind die Validierungsdaten nutzlos.
Es gibt aber Verfahren der Kreuzvalidierung mit deren Hilfe keine Aufteilung in Trainings- und Validierungsdaten notwendig ist und somit nach der Bestimmung des Parameters k keine Daten nutzlos werden. Ein Beispiel für ein Verfahren der Kreuzvalidierung ist das **Leave-One-Out - Verfahren**.

Das Verfahren kurz gefasst...

Bei diesem Verfahren wird der Reihe nach jeweils ein Punkt P aus den Trainingsdaten entnommen (*Leave One Out*) und dieser Punkt P dient für diesen einen Durchlauf als Validierungsdatenpunkt. Für P läuft der Validierungsdurchlauf identisch zum vorherigen Kapitel ab, d.h. P wird für verschiedene Werte von k mit Hilfe der verbleibenden Trainingsdaten klassifiziert. Dann wird geprüft, für welche Werte von k der aktuell betrachtete Validierungsdatenpunkt P korrekt klassifiziert wurde.
Im Anschluss daran wird P wieder zu den Trainingsdaten hinzugefügt und der nächsten Trainingsdatenpunkt wird als Validierungsdatenpunkt für einen weiteren Durchlauf entnommen. Dieses Vorgehen wird für alle Trainingsdaten sukzessive durchgeführt. Die Werte für k, welche am häufigsten zu einer korrekten Klassifikation geführt haben, sind unsere Kandidaten für den Wert von k.

Für eine detaillierte Betrachtung des Verfahrens wird auf die Fortbildung verwiesen. In diesem Dokument liegt der Fokus auf der Verwendung der TK-Mappen um das Vorgehen zu simulieren.

Situation:
In der Tabelle werden links die eingegebenen Trainingsdaten angezeigt.

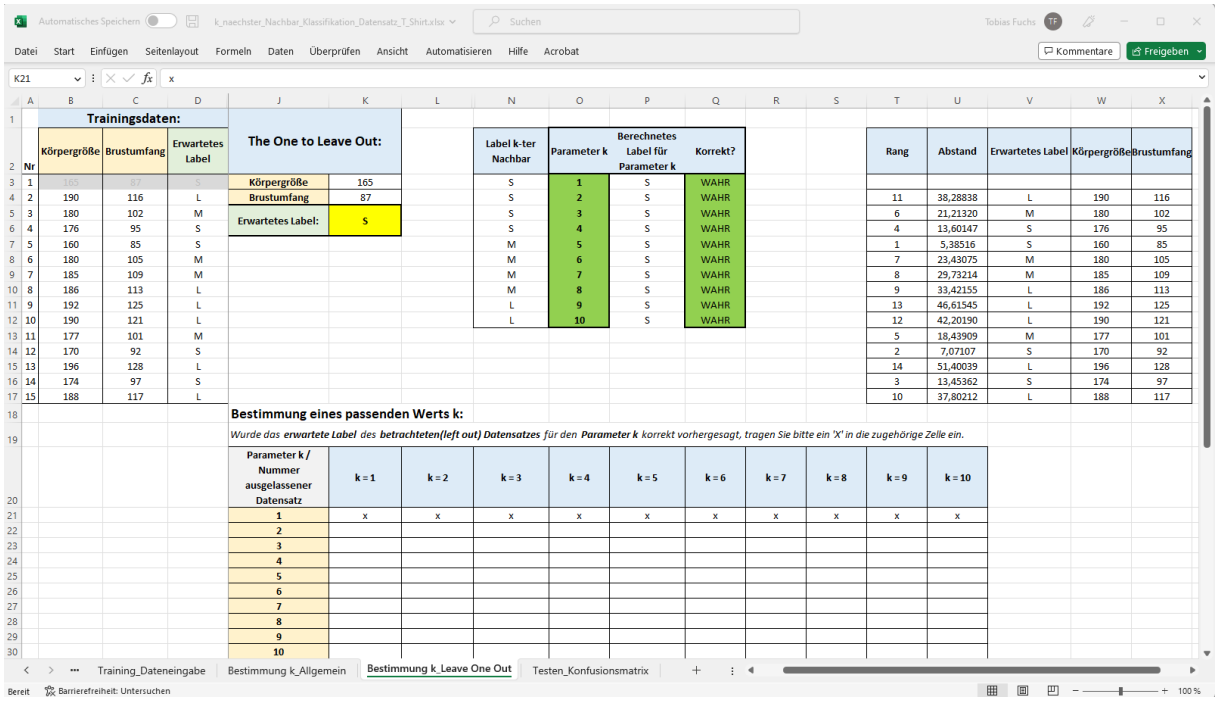
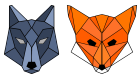


Abbildung 4: Screenshot der Tabelle zur Bestimmung des Hyperparameters k mit Hilfe des Leave-One-Out - Verfahrens (geöffnet in Microsoft Excel)



Vorgehen:

1. **Auswahl** des aktuellen **Validierungsdatenpunkts** aus den Trainingsdaten:

   Wählen Sie einen Validierungsdatenpunkt aus den Trainingsdaten aus und tragen Sie diesen als *The One to Leave Out* ein. Es müssen dabei lediglich die Werte für die Merkmale *Körpergröße* und *Brustumfang* eingetragen werden. Wird der Validierungsdatenpunkt in den Trainingsdaten gefunden, wird das *erwartete Label* automatisch übernommen und der entsprechende Datenpunkt in den Trainingsdaten ausgegraut.
   Im Screenshot wird dies für den Validierungsdatenpunkt mit Körpergröße 165 und Brustumfang 87 dargestellt.

2. **Ablesen** der **Klassifikation für verschiedene Werte von k**:

   Die Tabelle klassifiziert den eingetragenen Validierungsdatenpunkt automatisch für $k = 1, ..., 10$ und vergleicht das *berechnete Label* mit dem *erwarteten Label* des Datenpunkts. Des Weiteren wird direkt angezeigt, ob für einen Wert von k die Klassifikation korrekt war, sprich ob das berechnete Label mit dem erwarteten Label des Validierungsdatenpunkts übereinstimmt.

3. **Eintragen** des **Klassifikationsergebnisses für verschiedene Werte von k**:

   Tragen Sie im Bereich *Bestimmung eines passenden Werts k* für die Werte von k ein **X** ein, für die die Klassifikation korrekt war.
   Im Screenshot sehen Sie die Eintragung für den ersten Validierungsdatenpunkt. Für $k = 1, ..., 10$ ist die Klassifikation korrekt (Einträge X).

Führen Sie dieses Vorgehen für alle Validierungsdatenpunkte durch.

4. **Ablesen** der **besten Werte** für **k**:

   In Zeile 37 wird prozentual dargestellt, wie viele Validierungsdatenpunkte für die jeweiligen Werte von k korrekt vorhergesagt wurden (grün). Die Werte von k, welche hier den höchsten Prozentsatz aufweisen, sind unsere Kandidaten für k.

Im rechten, oberen Bereich der Tabelle werden Informationen zu den aktuell verbleibenden Trainingsdaten angezeigt. Diese dienen lediglich der zusätzlichen Information und haben keine direkte Bedeutung für die Bestimmung der Kandidaten für den Parameter k.

5. Testen des trainierten Modells (Konfusionsmatrix)

Nachdem mit Hilfe maschinellen Lernens ein Modell trainiert wurde, muss die Güte des Modells geprüft werden, bevor es in den produktiven Betrieb übergehen kann. Das Testen stellt neben dem Training den zweiten großen Bestandteil des maschinellen Lernprozesses dar.

Das Verfahren kurz gefasst...

Eine Menge an **Testdaten** wird mit Hilfe des trainierten Modells klassifiziert und die *erwarteten Labeln* der Testdaten werden mit den vom Modell für die Testdaten *berechneten Label* verglichen. Die Ergebnisse dieser Vergleiche können in Form einer **Konfusionsmatrix** übersichtlich dargestellt werden.
Des Weiteren können für diese Ergebnisse Maßzahlen berechnet werden, welche es ermöglichen die Güte des trainierten Modells einzuschätzen. Eine gängige Maßzahl ist die **Genauigkeit** des Modells. Hierzu wird die Anzahl der vom Modell korrekt klassifizierten Testdaten durch die Anzahl aller Testdaten dividiert. Das Ergebnis stellt die Genauigkeit des trainierten Modells dar.

Für eine detaillierte Betrachtung des Verfahrens wird auf die Fortbildung verwiesen. In diesem Dokument liegt der Fokus auf der Verwendung der TK-Mappen um das Vorgehen zu simulieren.

Situation:

In der Tabelle werden links eine Menge an Testdaten angezeigt. Diese können von Ihnen beliebig verändert werden. Achten Sie dabei aber darauf, dass Sie nur *Erwartete Label* aus den bei der Eingabe der Trainingsdaten zugelassenen Label verwenden.

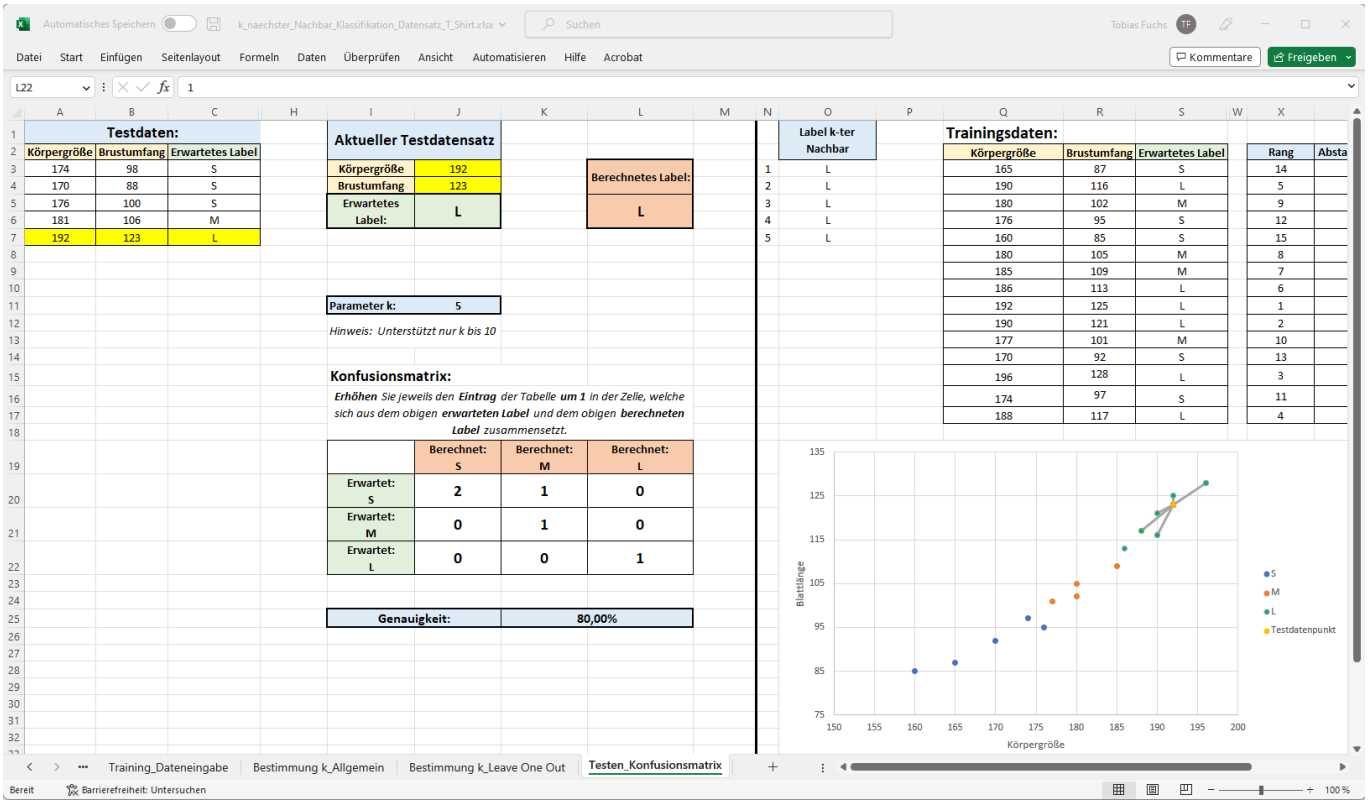
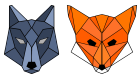


Abbildung 5: Screenshot der Tabelle zum Testen des Modells und der Ergebnisdarstellung in Form einer Konfusionsmatrix (geöffnet in Microsoft Excel)



Vorgehen:

1. **Eingabe** des **Parameters** k (d.h. Festlegen der Anzahl an betrachteten Nachbarn):

   Geben Sie neben der Zelle mit Text *Parameter k* den von Ihnen gewünschten Wert für k ein. Es werden nur ganzzahlige Werte von 1 bis 10 unterstützt.

2. **Auswahl** des aktuellen **Testdatenpunkts**:

   Wählen Sie einen Testdatenpunkt aus den Testdaten aus und tragen Sie diesen als *Aktueller Testdatensatz* ein. Es müssen dabei lediglich die Werte für die Merkmale *Körpergröße* und *Brustumfang* eingetragen werden. Wird der Testdatenpunkt in den Testdaten gefunden, wird das *erwartete Label* automatisch übernommen und der entsprechende Datenpunkt wird in den Testdaten gelb hervorgehoben.
   Im Screenshot wird dies für den Testdatenpunkt mit Körpergröße 192 und Brustumfang 123 dargestellt.

3. **Vergleich berechnetes Label - erwartetes Label** und **Eintragung** in die **Konfusionsmatrix**:

   Vergleichen Sie den Eintrag neben *Berechnetes Label*(rötlich) mit dem Eintrag neben *Erwartetes Label*(grün). Erhöhen Sie den zugehörigen Eintrag in der Konfusionsmatrix um 1.

Führen Sie dieses Vorgehen für alle Validierungsdatenpunkte durch.

4. **Ablesen** der **Genauigkeit**:

   Die Genauigkeit des trainieren Modells für die Testdaten wird automatisch berechnet und in der entsprechenden Zelle angezeigt (blau).

Im rechten Bereich der Tabelle werden noch zusätzliche Informationen zu den Trainingsdaten, den zum aktuellen Testdatenpunkt P k nächsten Nachbarn sowie eine graphische Darstellung der k nächsten Nachbarn von P angezeigt. Diese Informationen sind für das Testen nicht direkt notwendig und dienen nur der zusätzlichen Information.