# Ahle Wurst trifft KI

## Aufdecken neuer Zusammenhänge und Auswahl benötigter Sensoren mittels Data Mining und künstlicher Intelligenz

Seminararbeit
am FG Kommunikationstechnik - ComTec
der Universität Kassel

Alexander Elia Rode
Matr.Nr. 35542571
alexander.rode@student.uni-kassel.de

Philip Laskowicz
Matr.Nr. 35569183
philip.laskowicz@student.uni-kassel.de

Sommersemester 2024

**Zusammenfassung**

Künstliche Intelligenz bietet zur Untersuchung von Datensätzen Möglichkeiten, um in Daten Vorhersagen zu treffen und neue Zusammenhänge zu entdecken. In der folgenden Ausarbeitung werden einige Möglichkeiten der Zusammenhangsfindung und der Vorhersage von Daten wiedergegeben und anhand eines Beispieldatensatzes aus der Wurstproduktion angewandt. In diesem Zusammenhang sollen die Fragen beantworten werden, ob mit Hilfe von KI der Reifeprozess einer Rohwurst vorhergesagt und der Reifegrad anhand der vorliegenden Daten bestimmt werden kann. Zudem soll geprüft werden, welche Features der Daten in Zusammenhang stehen und inwiefern sich die Auswahl an Sensoren beschränken lässt.

# 1 Einführung

Daten sind in der Informatik von grundlegender Bedeutung. Daten beinhalten in einer formalisierten Darstellung Informationen, aus denen unter Verwendung verschiedener Methoden in den meisten Fällen neues Wissen generiert werden kann. Durch den Zugewinn an Wissen stellt sich ein positiver Aspekt heraus, sodass durch dieses Wissen Prozesse und Abläufe erklärt, vereinfacht oder optimiert werden können. Die folgende Ausarbeitung beschäftigt sich zum einen mit der Korrelation zwischen den Merkmalen eines Datensatzes, um Beziehungen in den Daten zu erkennen und zum anderen mit der Frage, wie sich unter Hinzunahme von gelabelten Daten die Auswahl und Anzahl an Sensoren beschränken lässt. In den folgenden Abschnitten werden verschiedene Methoden des Data Minings und der künstlichen Intelligenz aus der Fachliteratur vorgestellt. Das Ziel dieser Ausarbeitung ist neben der Auswahl an Methoden zur Untersuchung von Daten, auch eine Anwendung der Methoden an einen Datensatz aus dem Bereich der Wurstproduktion aufzuzeigen. Der vorliegende Datensatz enthält verschiedene Fakten zum Zustand und den Umgebungsbedingungen einer „Ahlen Wurst“. Abschließend soll die Frage untersucht werden, ob gemäß den vorliegenden Daten der Reifestatus einer Wurst bestimmt werden kann.

Im nächsten Abschnitt werden die notwendigen Grundlagen des Data Minings erläutert. Anschließend werden im dritten und vierten Abschnitt verschiedene Methoden zur weiteren Untersuchung vorgestellt. Der Untersuchungsdatensatz wird im fünften Abschnitt erläutert. Die Ergebnisse der Untersuchung und Auswertungen des Datensatzes werden im abschließenden sechsten Abschnitt dargestellt.

# 2 Konzepte und Grundlagen

Daten sind formalisierte Fakten, die beobachtbare oder messbare Werte über Ereignisse in einem Zustandsraum darstellen. Durch die Zuweisung einer Bedeutung lassen sich aus Daten Informationen ableiten, woraus Wissen generiert werden kann. Daten können auf verschiedene Arten und Weisen erfasst werden, beispielsweise über die Betrachtung eines Ereignisses durch den Menschen oder automatisiert durch Maschinen und entsprechende Sensoren.

Ein Datensatz enthält mehrere Datenpunkte. Die Features stellen dabei die Spalten bzw. die Merkmale des Datensatzes dar.

Künstliche Intelligenz ist ein Teilgebiet der Informatik und zugleich die Fähigkeit einer Maschine „intelligent“ zu interagieren. Unter dieser Interaktion wird meist die Fähigkeit menschlich zu agieren verstanden und so unter anderem logisches Denken, Lernen, planen und Kreativität zu imitieren.

Data Mining wird meist als Teilprozess einer Datenanalyse verstanden. In diesem Teilprozess steht der Einsatz von systematischen Methoden auf große Datenbestände im Fokus, sodass durch diese Methoden Verbindungen erkannt oder neue Vorhersagen für ähnliche Probleme getroffen werden können. Die folgende Abbildung erläutert den Prozess einer Datenanalyse nach Fayyad. In der vorliegenden Darstellung wird deutlich, dass

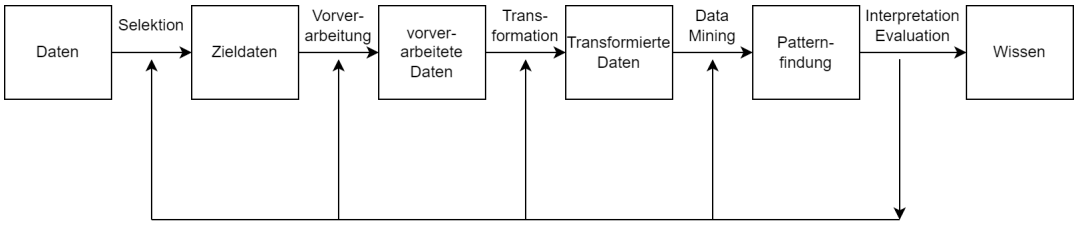


Abbildung 1: Darstellung der Datenanalyse nach Fayyad

zu einer vollständigen Untersuchung eines Datensatzes verschiedene Schritte abgearbeitet werden müssen. Folgt man der Abbildung werden im ersten Schritt Daten selektiert, sodass Daten vollständig, genügend vorhanden und verwendbar sein müssen. Im zweiten Schritt werden die Daten vor-verarbeitet. In diesem Schritt wird sichergestellt, dass keine fehlerhaften Daten die weiteren Ergebnisse verfälschen. In der Transformation werden die Daten auf die Methoden des Data Minings abgestimmt. In diesem Schritt erfolgt auch eine Prüfung, welche Feature eines Datensatzes für die Fragestellung eine Rolle spielen. Im anschließenden Pattern-Matching wird nach wiederkehrenden Mustern oder Zusammenhängen in den Daten gesucht. Aus den Ergebnissen lassen sich Schlussfolgerungen für neues Wissen ziehen. Unter Umständen müssen die vorherigen Schritte erneut iteriert werden, falls sich in den bisherigen Ergebnissen keine plausiblen Schlussfolgerungen ableiten lassen konnten.

Im Data Mining werden im allgemeinen sogenannte Lernstrategien verwendet, um aus den vorhandenen Daten zu lernen. In diesem Zusammenhang tauchen die Begriffe überwachtes und das unüberwachte Lernen wiederkehrend auf. Unüberwachte Lernmethoden verwenden keine gelabelten Daten, sondern versuchen Zusammenhänge und Vorhersagen auf Basis der Rohdaten zu treffen. Das Clustern ist ein großer Vertreter dieser Verfahren. Beim überwachten Lernen haben wir einen Datensatz, von dem wir bereits die Ausgabe kennen. Mit diesen Daten können wir ein Modell trainieren und anschließend dieses Modell auf neue Daten anwenden, um zu klassifizieren und Vorhersagen zu treffen. Neuronale Netz sind hier oft verwendete Modelle.

Bei einem neuronalen Netz handelt es sich um ein Netzwerk, welches aus Schichten besteht und dem biologischen Gehirn nachempfunden ist. Jede Schicht besteht dabei aus Perzeptronen, die biologischen Neuronen im Gehirn nachahmen.

# 3 Unüberwachte Methoden und Ansätze: Aufdecken unbekannter Zusammenhänge

Das Gebiet des Unüberwachten Lernens enthält verschiedene Methoden zur Untersuchung von Daten. Diese Methoden lassen sich in Verfahren unterteilen, die Daten veranschaulichen und erklären und Verfahren, die Vorhersagen ermöglichen. Diese Verfahren können in verschiedene Bereiche, wie die Dimensionsreduktion, die Clusteranalyse oder auch die Regression eingeteilt werden.

## 3.1 Grundlegende Statistiken und Korrelation

Bereits grundlegende Funktionen aus der Statistik bieten die Möglichkeiten der Datenanalyse und schaffen eine Zugänglichkeit zu den Daten. Zu diesen Methoden zählen Histogramme, Boxplots und Diagramme über den zeitlichen Verlauf. Um den Zusammenhang zwischen den verschiedenen Merkmalen zu untersuchen, existieren Korrelationsarten, um die Beziehung zwischen zwei Merkmalen zu messen. Gängige Korrelationskoeffizienten sind die Pearson-, Kendall und Spearman Korrelation. Die Pearson Korrelation wird zur Messung linear verwandter Variablen eingesetzt. Die Anwendung erfolgt unter der Voraussetzung, dass die Daten metrisch und linear zusammenhängend sind. Der Koeffizient wird mit der folgenden Formel berechnet: $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}*\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$ mit $r_{xy}$ Koeffizient zwischen x und y, n: Anzahl der Werte, $x_i$ bzw. $y_i$: i-ter Wert von x bzw. y und $\overline{x}$ bzw. $\overline{y}$ arithmetisches Mittel von x bzw. y

Die Verfahren nach Kendall und Spearman setzen eine ordinale Skalierung voraus. Der Koeffizient nach Kendall errechnet sich aus: $\tau = \frac{K-D}{K+D}$, wobei K und D die Anzahl der konkordanten bzw. diskordanten Paare sind. Ein Paar von Beobachtungen $(x_i, y_i), (x_j, y_j)$ ist genau dann konkordant, wenn $x_i < x_j$ und $y_i < y_j$ oder $x_i > x_j$ und $y_i > y_j$, ansonsten diskordant Der Koeffizient nach Spearman errechnet sich aus: $r_s = 1 - \frac{6\sum(R_{x_i}-R_{y_i})^2}{n(n^2-1)}$ mit $R_{x_i}$ als Rang von $x_i$; $R_{y_i}$ als Rang von $y_i$.

## 3.2 Dimensionsreduktion

Eine Herausforderung, die im Data Mining und in der Visualisierung von Daten eine entscheidende Rolle spielen, ist die Dimension des Datensatzes, wodurch Datensätze enorm anwachsen können. Die Dimension eines Datensatzes entspricht der Anzahl der Features der Daten. Das Ziel in der Dimensionsreduktion ist die Reduzierung von Daten auf wesentliche Inhalte, sodass die Dimensionen reduziert und Algorithmen effizienter angewendet werden können Zwei wesentliche Verfahren sind die Hauptkomponentenanalyse (PCA) und der Ansatz über einen Autoencoder.

Bei der Hauptkomponentenanalyse werden hochdimensionale Daten auf eine geringere Dimension reduziert. Die abgeleiteten Komponenten bilden die Hauptkomponenten, die neuen Features, des Datensatzes. PCA erfordert eine Skalierung, damit sich die Daten in einem gleichen relativen Bereich befinden. Dieser Ansatz findet besonders bei hochdimensionalen Daten, wie Bildern und Videos Anwendung.

Ein weiterer Ansatz handelt über Autoencoder, die Daten durch ein Neuronales Netz auf wesentliche Inhalte komprimieren, um die Daten in anderen Verfahren weiter analysieren zu können. Autoencoder bestehen im Wesentlichen aus einem Encoder, der Daten komprimiert und einen Decoder, der die Ausgangsdaten versucht nachzubilden. Das Neuronale Netz ist derart ausgeprägt, dass die Neuronen-Anzahl der Eingabeschicht der Feature-Anzahl $k$ der Eingabedaten oder einer größeren Zahl entspricht. Die verborgenen Schichten besitzen eine kleinere Neuronen-Anzahl als $k$. Der sogenannte Engpass ist eine verborgene Schicht, die die Daten am stärksten komprimiert. Diese Schicht stellt die Ausgabe des Encoders und gleichzeitig die Eingabe des Decoders da. Der Decoder be-

steht im Gegensatz zum Encoder aus Schichten mit wachsender Neuronen-Anzahl. Das Neuronale Netz wird nun so trainiert, dass die Eingabedaten, den Ausgabedaten entsprechen. Nach dem das Netz trainiert ist, wird der Decoder verworfen und die Ausgabe des Engpasses als Ausgabe des neuronalen Netzes gewählt. Die Idee ist, dass durch den Engpass das neuronale Netz seine Daten zwangsläufig komprimieren muss, aber dennoch die wesentlichen Informationen beibehalten muss, um die Ausgangsdaten im Decoder gut nachzubilden.

## 3.3 Clusteranalyse

Beim unüberwachten Lernen stehen keine Labels, also keine Eingruppierung der Daten zur Verfügung. Für den späteren Datensatz aus der Wurstproduktion ist eine interessante Fragestellung, ob die Wurst gemäß den Daten in Reifestadien eingeteilt werden kann. Diese Aufgabenstellung lässt sich auf das Clustern zurückführen. Ziel des Clusterns ist, dass die Daten aufgrund ihrer (Un-)Ähnlichkeit in Gruppen eingeteilt werden und Muster erkannt werden. In der Clusteranalyse existieren mehrere Verfahren, die verschiedene Ansätze verfolgen.

Eine weitverbreitete Methode zur Clusteranalyse ist der Algorithmus K-Means. Bei diesem Algorithmus wird jeder Datenpunkt in eine von k Gruppen eingeteilt. Die Idee des Algorithmus ist k Centroide, die jeweils ein Mittelpunkt eines Clusters darstellen zu initialisieren und die Centroide schrittweise anzupassen, bis die Daten in ihre Gruppen optimal eingeteilt sind. K-Means versucht dafür den euklidischen Abstand zwischen jeden Punkt und dem Mittelpunkt des Clusters zu minimieren. Daraus resultieren auch Nachteile gegenüber Datensätze mit unterschiedlichen Dichten und Gruppen, die nicht konvex sind.

Ein Ansatz, um dichtebasierte Cluster in Daten zu finden, ist der Ansatz DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN ist ein dichtebasierter Algorithmus und gruppiert Daten, danach wie dicht sie beieinander liegen. Dafür hat der Algorithmus zwei wesentliche Parameter. Der Parameter *eps* gibt die maximale Distanz zwischen zwei Punkten an, damit dieser Punkt als in der gleichen Nachbarschaft liegend gilt. $Min_{samples}$ gibt die Anzahl der Datenpunkte an, damit eine Gruppe von in der Nachbarschaft liegenden Punkten als Cluster gilt. Dieses Verfahren ist robust gegenüber Ausreißern und kann nicht-konvexe Cluster ermitteln. Des Weiteren ist für dieses Verfahren keine feste Anzahl an Clustern notwendig, jedoch besteht die Schwierigkeit bei der richtigen Wahl der Parameter *eps* und $Min_{samples}$.

Eine Erweiterung von DBSCAN ist HDBSCAN, die den Algorithmus um eine hierarchische Struktur durch verschiedene Dichte-Schwellwerte ergänzt, wodurch Clustern mit variierenden Dichten erkannt werden.

## 3.4 Regression

In der Regression wird versucht mit verschiedenen Regressionsmodellen funktionale Abhängigkeiten zwischen Merkmalen zu erkennen, um die Daten hinreichend gut approximieren zu können. Modelle, wie die Lineare und Polynomielle Regression suchen dafür

mathematische Funktionen, die die Abhängigkeiten beschreiben und Vorhersagen von Werten ermöglichen

# 4 Überwachte Methoden und Ansätze: Auswahl benötigter Daten und Sensoren

In diesem Abschnitt werden verschiedene Methoden und Ansätze beschrieben, die zur Reduktion der Sensoranzahl verwendet werden können. Die Reduktion der Sensoranzahl ist ein wichtiger Schritt, um die Messungen in größeren Maßstäben durchführen zu können. Durch die Verwendung weniger Sensoren können Kosten für die Anschaffung und den Betrieb der Sensorik verringert werden. Zudem wird die Datenmenge verkleinert, was die Verarbeitung und Aufbewahrung der Daten vereinfachen. Für diesen Anwendungsfall wurden drei verschiedene Methoden des überwachten Lernens verwendet und die Ergebnisse im sechsten Abschnitt miteinander verglichen. Bei den hier verwendeten Methoden handelt es sich um ein neuronales Netzwerk, den Random Forest (RF) sowie die Support Vector Machine (SVM).

## 4.1 Neuronales Netzwerk

Ein neuronales Netzwerk, auch Multilayer Perceptron (MLP) (deutsch Mehrlagiges Perzeptron) genannt, ist ein Netzwerk, welches aus mehreren Schichten besteht und im Grundsatz einem biologischen Gehirn nachempfunden ist.

Jede Schicht besteht in sich aus mehreren Perzeptronen, welche die biologischen Neuronen im Gehirn nachahmen und einen Eingabevektor auf einen Wert zwischen 0 und 1 abbilden. Diese Perzeptronen bestehen aus Gewichtungen für die Eingänge und einer Aktivierungsfunktion. Die Gewichtungen geben dabei an, zu welchen Graden die Eingaben des Perzeptrons zueinander gewichtet werden. Die Aktivierungsfunktion bestimmt die Ausgabe des Perzeptrons basierend auf den gewichteten Eingaben und gibt einen Wert zwischen 0 und 1 aus.

In einem neuronalen Netzwerk werden Schichten aus, in der Regel mehreren parallelgeschalten, Perzeptronen in Reihe geschaltet. Dabei wird die Eingabe einer Schicht auf alle Perzeptronen dieser Schicht gegeben, welche verschiedene Gewichtungen besitzen. Die Aktivierungsfunktion ist in der Regel für alle Perzeptronen im neuronalen Netzwerk gleich. Die Ausgabe der Schicht ist dann wieder ein Vektor mit einer Größe, die der Anzahl der Perzeptronen in dieser Schicht entspricht.

Ein neuronales Netzwerk besteht aus einer Eingabeschicht, einer oder mehreren versteckten Schichten und einer Ausgabeschicht. Die Ausgabeschicht besitzt dabei so viele Perzeptronen, wie es verschiedene Gruppen, auch Klassen genannt, gibt, in die die Daten eingeteilt werden sollen. Dabei wird in der Regel die Klasse mit dem höchsten Ausgabewert des zugehörigen Perzeptrons gewählt.

Während des Trainings wird dem neuronalen Netzwerk eine Eingabe gegeben und die daraus folgende Ausgabe mit den tatsächlichen Werten verglichen. Basierend auf dem Fehler zwischen den Werten werden die Gewichtungen der einzelnen Perzeptronen

angepasst. Dieser Vorgang wird so lange wiederholt, bis der Fehler minimiert, oder eine maximale Anzahl an Trainingszyklen erreicht wurde.

## 4.2 Random Forest

Der Random Forest (deutsch Zufallswald) ist ein sogenanntes Ensemble-Verfahren, bei dem mehrere Entscheidungsbäume parallel verwendet werden.

Bei einem Entscheidungsbaum handelt es sich um eine Baumstruktur, bei der jeder Knoten ein Feature des Datensatzes und die Kanten die mögliche Entscheidung auf diesem bzw. Ausprägungen des Features darstellen. Die Blätter des Baumes entsprechen dann den möglichen Klassen, in die die Daten eingeteilt werden können.

Der Vorteil eines Random Forests gegenüber einem einzelnen Entscheidungsbaum besteht darin, dass die Bäume unabhängig voneinander trainiert werden und verschiedene Features und Teile des Datensatzes verwenden. Dadurch wird die Gefahr von Overfitting, dem zu starken Anpassen des Modells an den Trainingsdatensatz, wodurch das Modell auf neuen Daten ungenau wird, verringert. Die Ausgabeklasse des Random Forests wird dann durch die Berechnung es Mittelwertes der Ausgabe der einzelnen Bäume bestimmt.

Das Training des Random Forests erfolgt, indem für jeden Baum zufällig eine Teilmenge der Features und des Trainingsdatensatzes ausgewählt wird. Die Entscheidungsbäume werden dann trainiert bzw. gebaut, indem für jeden Knoten des Baums die beste Aufteilung des Datensatzes in Bezug auf die Klassen festgelegt wird. Dieser Vorgang wird so lange wiederholt, bis ein Abbruchkriterium erreicht wird.

Diese Abbruchkriterien können beispielsweise die maximale Tiefe des Baums oder die minimale Anzahl an Datenpunkten für einen weiteren Knoten sein und werden für den gesamten Random Forest festgelegt. Ebenfalls wird die Anzahl der Bäume des Random Forests festgelegt.

## 4.3 Support Vector Machine

Die Support Vector Machine (deutsch Stützvektormaschine) ist ein Verfahren zur Klassifikation von Daten, bei dem die Daten mit Hilfe von Hyperebenen in die verschiedenen Klassen aufgeteilt werden.

Die Datenpunkte des Trainingsdatensatzes werden als Vektoren im Raum der Features dargestellt. Die Hyperebenen werden dann so angepasst, dass die Abstände dieser zu den nächstgelegenen Datenpunkten maximal sind.

Ein Vorteil ist dabei, dass nicht alle Datenpunkte für die Anpassung der Hyperebenen benötigt werden, vielmehr sind nur die nächstgelegenen Datenpunkte, auch Stützvektoren (engl. support vector), von Interesse. Von diesen Stützvektoren leitet sich auch der Name 'Stützvektormaschine' her.

Da die Hyperebenen nicht gebogen werden können, wird häufig der sogenannte Kernel-Trick verwendet. Dabei wird der Vektorraum der Features in einen Raum mit mehr Dimensionen abgebildet. Bei einer genügend hohen, unter Umständen unendlichen, Anzahl an Dimensionen lassen sich die Daten dann durch Hyperebenen trennen.

Nach der Anpassung der Hyperebenen wird die Dimensionalität des Raumes dann wieder reduziert, wodurch Hyperflächen entstehen, die die Daten klar trennen.

### 4.4 Ansätze zur Reduktion der Sensoranzahl

Für die Reduktion der Sensoranzahl wurden drei verschiedene Ansätze getestet. Zwei dieser Ansätze verwendeten dabei den gesamten Datensatz, geteilt in Trainings- und Testdaten, während der dritte Ansatz den Datensatz bis zu bestimmten Zeitpunkten verwendete. Bei allen Ansätzen wurden jeweils die drei Methoden neuronales Netzwerk, Random Forest und Support Vector Machine verwendet. Im ersten Ansatz wurde jede Kombination der Features getestet, um die beste Kombination, unabhängig von den tatsächlich verwendeten Sensoren, zu finden. Im zweiten und dritten Ansatz wurden dann die Kombinationen der Sensoren, mit den zugehörigen Features, getestet.

### 4.5 Bewertung der Ergebnisse

Für die Bewertung der Ergebnisse wurden drei Metriken verwendet, die im Bereich des überwachten Lernens häufig verwendet werden. Dabei handelt es sich um die Sensitivität, die Korrektklassifikationsrate und den positiven Voraussagewert.

Die Sensitivität gibt dabei an, wie groß die Wahrscheinlichkeit ist, dass ein tatsächlich positives Ergebnis auch als dieses erkannt wird. Sie wird wie folgt berechnet:
$Sensitivität = \frac{TP}{TP+FN} = \frac{TP}{P}$, wobei $TP$ die Anzahl der richtigen positiven, $FN$ die Anzahl der falschen negativen und $P = TP + FN$ die Anzahl der tatsächlich positiven Ergebnisse ist.

Die Korrektklassifikationsrate gibt den Anteil der korrekt Klassifizierten Daten an und wird durch $Korrektklassifikationsrate = \frac{TP+TN}{TP+FN+FP+TN} = \frac{TP+TN}{N}$ berechnet, wobei $TN$ die Anzahl der richtigen negativen, $FP$ die Anzahl der falschen positiven und $N = TP + FN + FP + TN$ die gesamte Anzahl der Datenpunkte ist.

Der positive Voraussagewert, auch Genauigkeit genannt, gibt an, wie groß der Anteil der korrekt positiv klassifizierten Daten an der Gesamtanzahl der positiv klassifizierten Daten ist. Er wird durch $PositiverVoraussagewert = \frac{TP}{TP+FP}$ berechnet.

Da diese Metriken jeweils bestimmte Aspekte der Klassifikation bewerten, wurde ein sogenannter Score gebildet. Dieser Score wurde als arithmetisches Mittel der drei Metriken als $Score = \frac{Sensitivität+Korrektklassifikationsrate+PositiverVoraussagewert}{3}$ berechnet. Ein Vorteil dieses Scores ist, dass so die verschiedenen Aspekte berücksichtigt werden, da je nach Anwendungsfall unterschiedliche Metriken von Interesse sein könnten.

## 5 Untersuchs-Datensatz: Reifeprozess einer Wurst

Mittels Künstlicher Intelligenz können Zusammenhänge in Daten erkannt und Merkmale bestimmt werden, die einen besonderen Einfluss auf weiterführende Fragestellungen besitzen. Aus einer Messreihe stehen Daten bereit, die verschiedene Umgebungsbedingungen und Zustände von einzelnen Würsten während ihres Reifeprozesses darstellen. In der folgenden Darstellung werden die verwendeten Sensoren aufgezeigt und der Aufbau

der Messung in einem Schema verdeutlicht. Wie in Abbildung 2 erkennbar ist, werden die Sensoren in Umgebungssensoren und Wurst-Sensoren unterschieden. Die Umgebungssensoren dienen zur Bestimmung der Umgebungsverhältnisse, wie Helligkeit, Bewegungsstatus, Luftfeuchtigkeit und der Temperatur im Raum, in dem die einzelnen Wurstprodukte reifen. Für die Würste gibt es Sensoren, die jeweils für jede einzelne Wurst Temperatur, Feuchtigkeit, elektrische Leitfähigkeit und den PH-Wert messen. Ein Teil der Sensoren war dabei nicht für den Einsatz in Lebensmitteln konzipiert, sondern vielmehr für die Erhebung von Daten im Erdreich. Die Daten wurden über ein LoRaWAN-Netzwerk eingesammelt, welches aus den einzelnen Endgeräten (Sensoren), einem Gateway und einem Netzwerkserver besteht. LoRaWAN überträgt die Daten über eine Funkverbindung sehr energieeffizient im 868 MHz-Bereich. Die ausgewählten Sensoren und ihre Funktionen werden in der Abbildung ersichtlich.

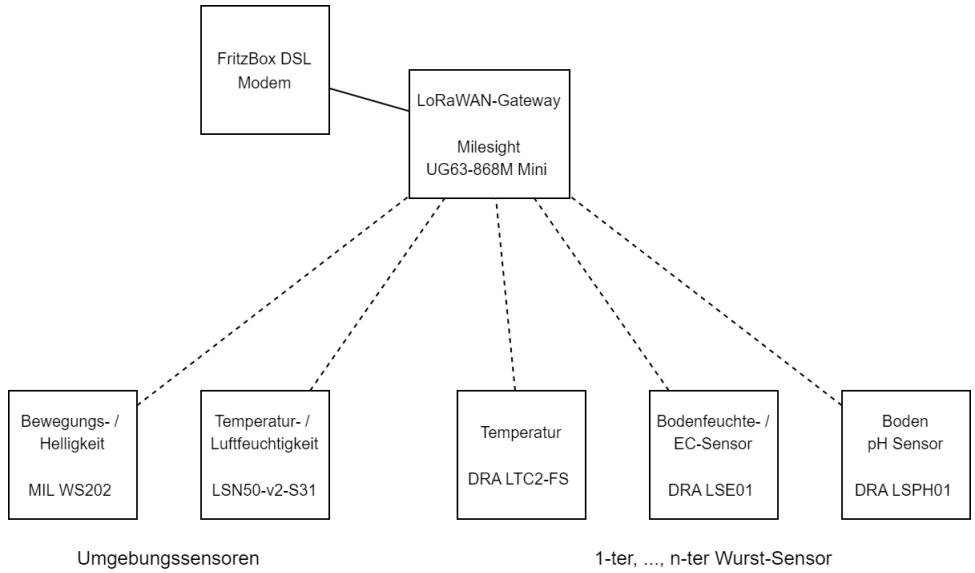


Abbildung 2: Darstellung des LoRaWAN-Netzwerks zur Übertragung der Messwerte

Der Reifeprozess der Rohwurstherstellung ist gekennzeichnet durch Säuerung in Form von PH-Wert Senkung und durch die Trocknung, sodass der aW-Wert der Wurstprodukte sinkt. Der aW-Wert ist ein Maß für nicht festgebundenes Wasser in Lebensmitteln. Bei Wasser beträgt dieser Wert 1. Richtwerte für diesen Prozess sind PH-Werte unter 5,4 und ein aW-Wert kleiner als 0,85. Die PH-Wert Reduzierung ist abhängig von Starterkulturen, die in den einzelnen Würsten verwendet werden. Einen besonderen Einfluss auf den Reifeprozess haben die äußeren Umgebungsbedingungen, wie Temperatur und Luftfeuchtigkeit und innere Faktoren, wie die Zusammensetzung der Wurst. Das Ziel bei dem Reifeprozess besteht aus Haltbarmachung der Würste, der Umrötung, Aromabildung und das Herstellen einer Schnittfestigkeit.

# 6 Zusammenfassung und Ergebnisse - Ist die Wurst gemäß den Daten reif?

Die vorgestellten Methoden und Ansätze können bei der Untersuchung eines Datensatzes, wie dem vorliegenden Datensatz über die Reifung einer Wurst, hilfreich sein. Durch die Methoden wurde untersucht, welche Zusammenhänge in dem vorliegendem Datensatz existieren und ob es Modelle gibt, die den Reifegrad einer Wurst anhand der vorliegenden Daten-Features vorhersagen können.

## 6.1 Auswertung und Aufdecken unbekannter Zusammenhänge

Der vorliegende Datensatz besteht aus zwei Messreihen mit je 5 unterschiedlichen Würsten. Der Datensatz besteht aus ca. 875000 Datenpunkten mit 9 unterschiedlichen Features und wurde über zwei Zeiträume aufgenommen. Der erste Zeitraum dauert vom 30.03.2023 und reicht bis zum 02.08.2023. Der zweite Zeitraum reicht vom 31.07.2023 bis zum 03.04.2024. Die statistische Auswertung ergibt, dass die mittlere Luftfeuchtigkeit bei der ersten Messreihe ca. 72,5 % und bei der zweiten Messreihe ca. 78,7 % beträgt. Der zweite Datensatz erhält keine Annotationen und somit keine eindeutigen Informationen zum Reifegrad der Wurst.

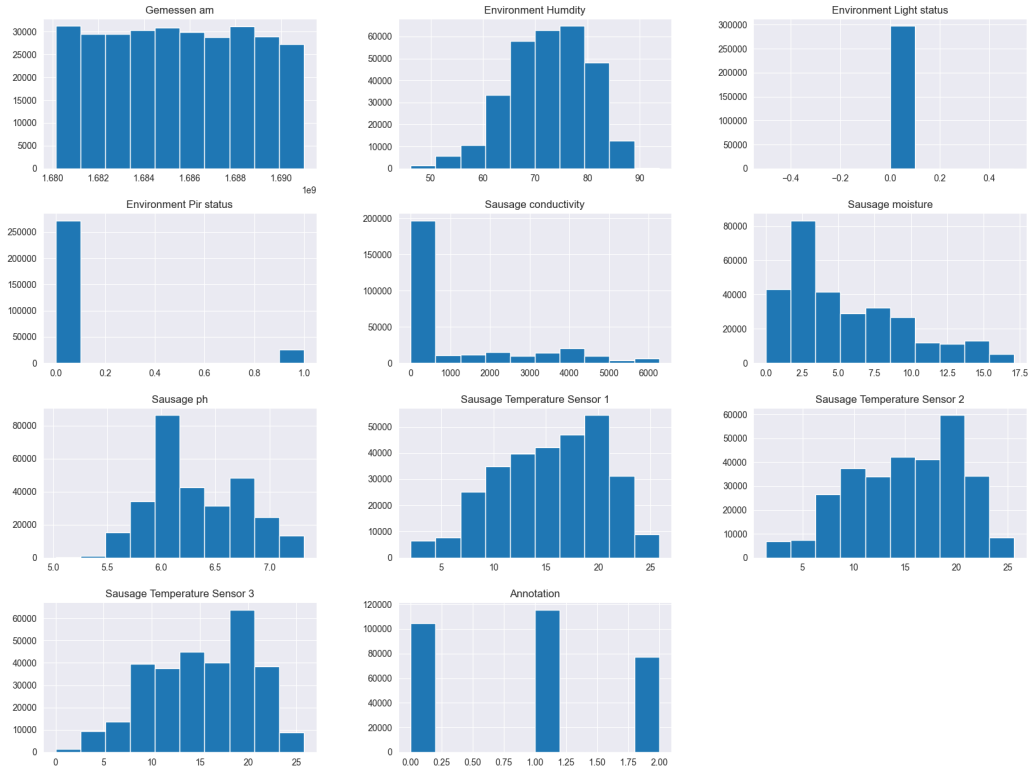


Abbildung 3: Histogramm Messreihe 1

In Abbildung 3 werden Histogramme basierend auf der ersten Messreihe mit fünf Würsten dargestellt. Aus den Histogrammen geht unter anderem hervor, dass die Würste durchgehend in einer dunklen Umgebung gereift sind. Die Anzahl der Messdaten sind auf den Zeitraum gleichmäßig verteilt. Die Mehrzahl der Messdaten besitzen eine niedrige Leitfähigkeit. Es gibt einige starke Ausreißer nach oben. Die Mehrzahl der Datenpunkte besitzen einen PH-Wert von 6. Die Umgebungstemperatur beträgt im Mittel 15,39 Grad Celsius und 15,09 Grad Celsius im Mittel bei der Temperatur der Wurst.

In der nachfolgenden Abbildung 4 sind zeilenweise ein Teil der Messdaten der ersten drei Würste aus der ersten Messreihe abgebildet. In der ersten Spalte ist die Leitfähigkeit, in der zweiten Spalte die Feuchtigkeit und in der dritten Spalte der PH-Wert einer Wurst abgebildet. Die roten Striche parallel zu Y-Achse kennzeichnen die Übergänge der Reifegrad-Phasen Umrötung, Phase 1 und 2. Auf der X-Achse ist die Erfassungszeit als Unix Timestamp und auf der Y-Achse der Wert zum zugehörigen Feature dargestellt.

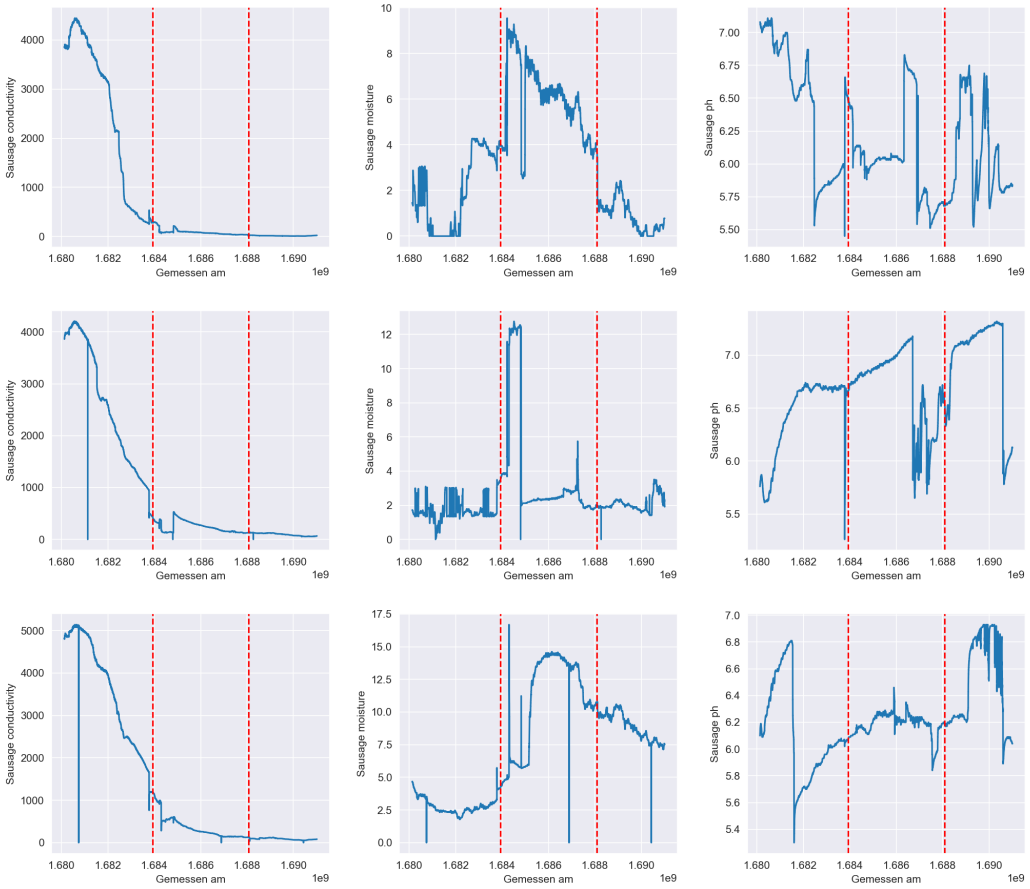


Abbildung 4: Zeitreihe der Messdaten von Wurst 1 bis 3

Aus dieser Abbildung lässt sich eine erste Vermutung über den Zusammenhang zwi-

schen Reifezustand und den Daten herstellen. In den Messreihen ist erkennbar, dass der erste Reifegrad, die Umrötung der Rohwurst, sich dadurch kennzeichnet, dass die Leitfähigkeit von einem hohen Niveau aus leicht ansteigt und dann mit der Zeit gegen Null abfällt. In der zweiten Phase flacht dieser Abfall leicht ab bis in der abschließenden 3. Phase die Leitfähigkeit gegen Null tendiert. Die Feuchtigkeit nimmt in der Abbildung leicht zu. Zu dem PH-Wert lässt sich schwierig eine klare Aussage treffen, da hier die Werte sehr stark schwanken und ausreißen. In diesem Fall besteht die Möglichkeit, dass hier die Messdaten verfälscht sind. In der Abbildung 5 wird die zweite Messreihe dargestellt. Hier sind viele Ausreißer zu erkennen, die die Daten verfälschen und Aussagen über den Verlauf erschweren. Jedoch ist auch hier erkennbar, dass die Leitfähigkeit im Laufe der Zeit gegen Null abfällt.

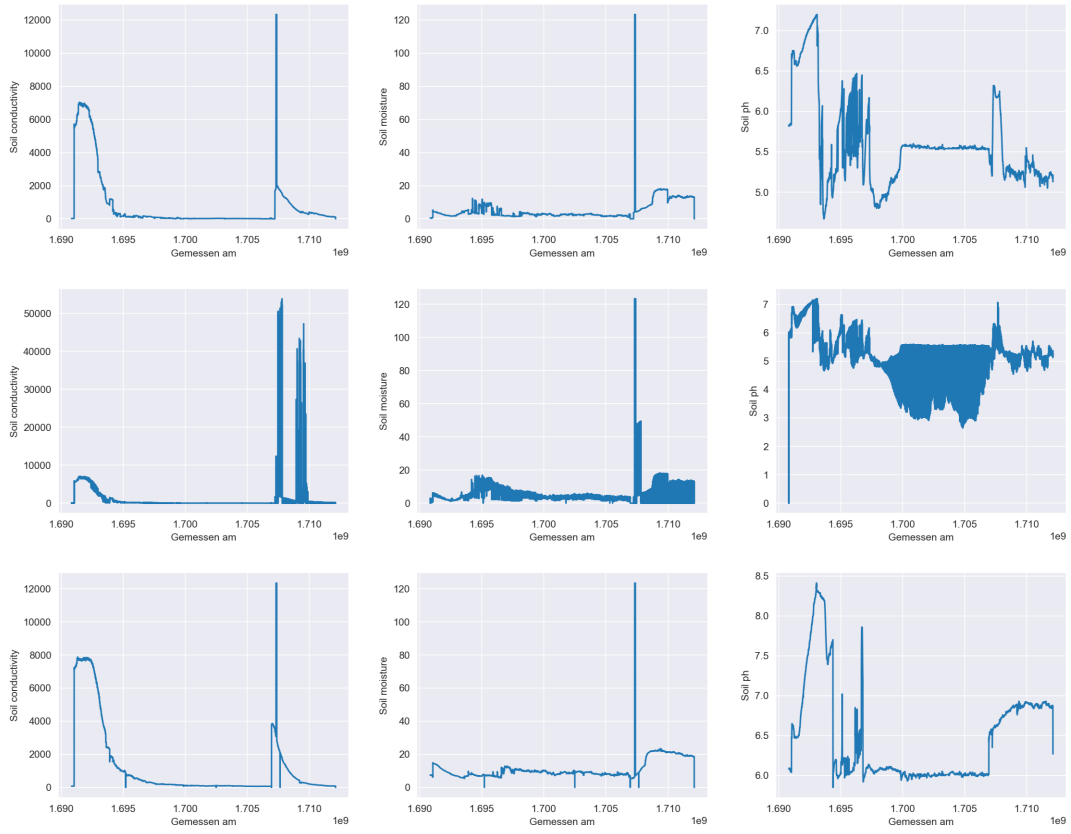


Abbildung 5: Zeitreihe der Messdaten von Wurst 6 bis 8

Um die ersten Zusammenhänge, die sich in den Zeitreihen erkennen lassen, weiter zu untersuchen, lassen sich Korrelationsmethoden heranziehen. In der nachfolgenden Abbildung 6 ist die Korrelation nach der Methode Spearman durchgeführt. Auf der horizontalen Achse ist die Reihenfolge der Features von links nach rechts identisch zu der vertikalen Achse von oben nach unten. Ein Wert nahe -1 und 1 deutet auf einen linea-

ren Zusammenhang zwischen jeweils zwei Parametern hin. Hier lässt sich entnehmen, dass besonders die Werte Zeit, Leitwert, PH-Wert und die Temperaturwerte im Zusammenhang mit dem Reifegrad (Annotation) stehen. Die Temperatur und die Annotation könnten große Zusammenhangswerte besitzen, da die Temperatur durch den gewählten Zeitraum ansteigen. Daher kann man hier keinen direkten Zusammenhang zwischen dem Feature Temperatur und dem Reifegrad ableiten.

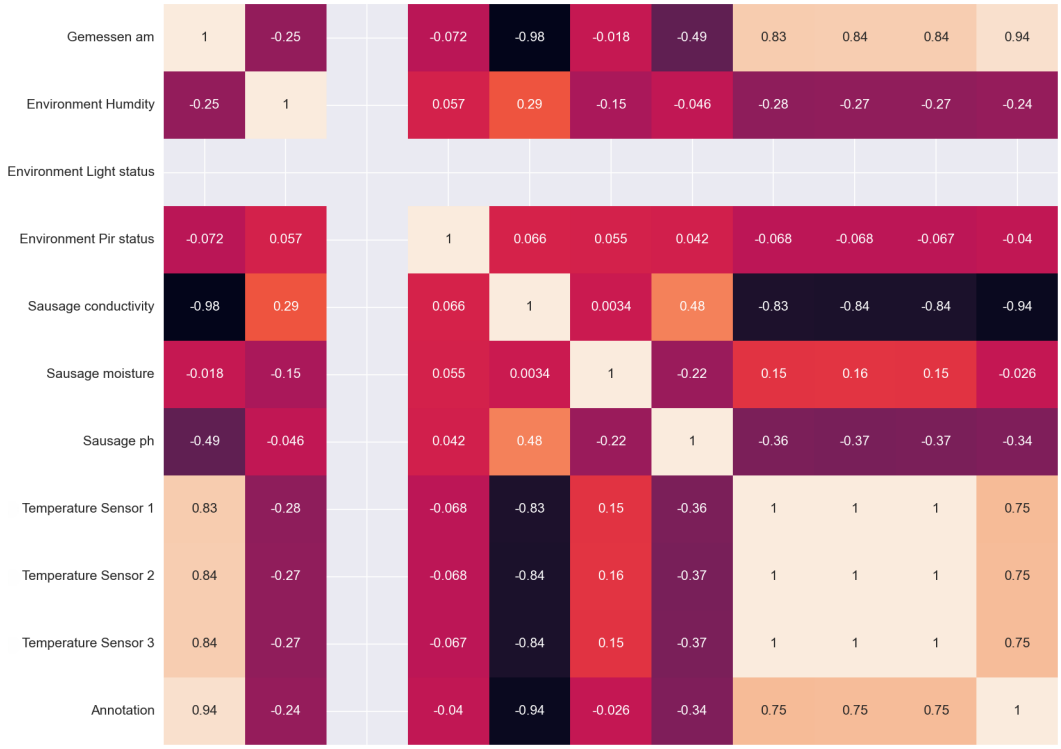


Abbildung 6: Korrelationsanalyse nach Spearman

In der folgenden Abbildung wurde eine Clusteranalyse mit K-Means für die Datenfeatures Leitfähigkeit und PH-Wert durchgeführt. Die zweite Abbildung zeigt die wahre Verteilung der Daten nach Reifegrad. Weitere Clusteranalysen wurden mit DBSCAN, HDBSCAN und Self-Organized-Maps durchgeführt. Aus diesen Daten sind jedoch keine neuen Zusammenhänge erkannt worden. In den erstellten Clustern konnte keine konkrete Übereinstimmung zwischen Annotation und Rohdaten ermittelt werden.

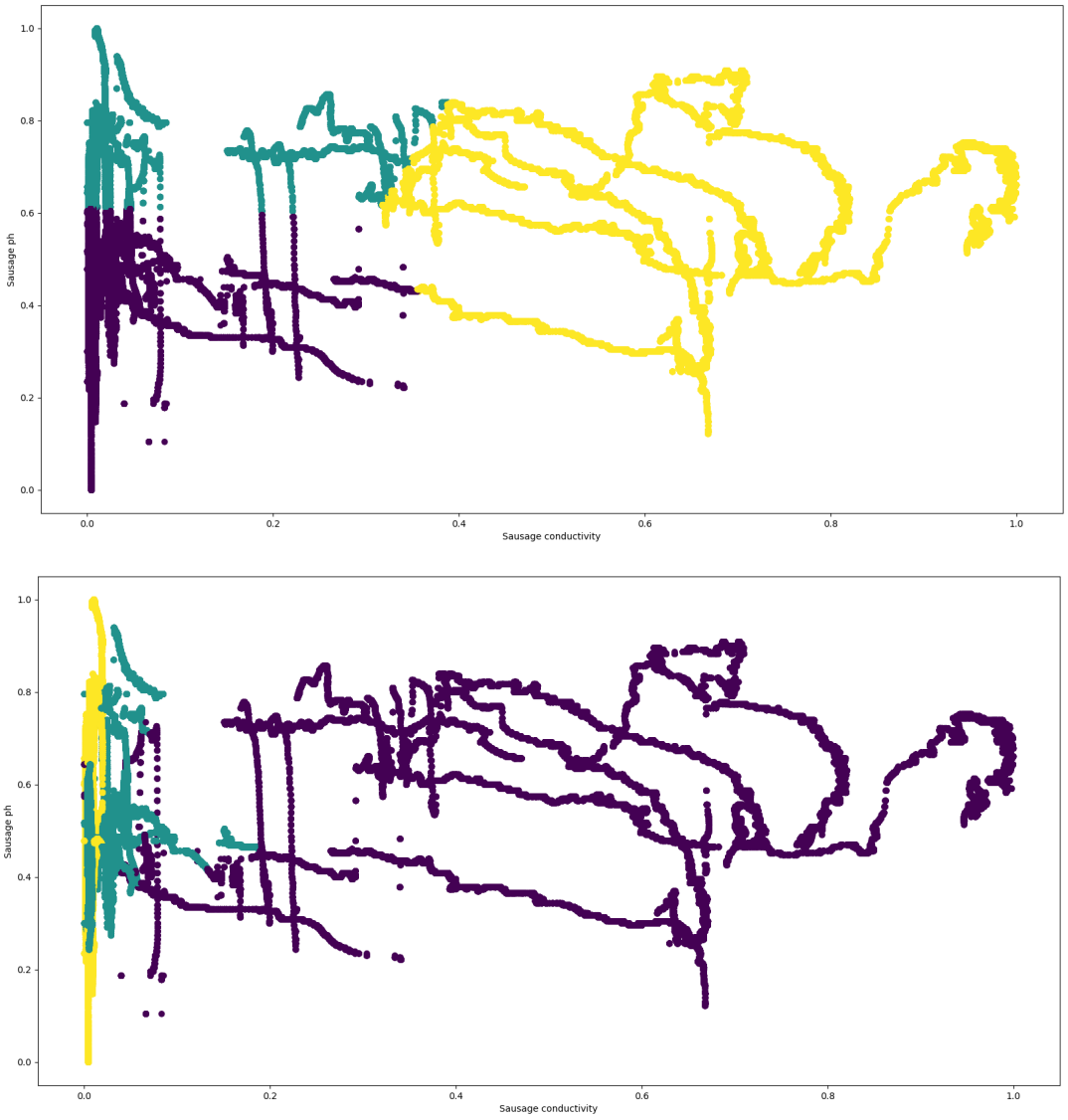


Abbildung 7: Clusteranalyse K-Means

Bis zu diesem Punkt stellt sich das Feature Leitfähigkeit als am geeignetsten heraus, um eine Aussage über den Reifegrad der Wurst zu geben. Daher wurde mit Regression ein Polynom bestimmt, welches aus den Zeitlinien der fünf Würste der ersten Zeitreihe folgt. Dieser Verlauf ist in der nachfolgenden Abbildung zu sehen.

Das Polynom besitzt den folgenden Aufbau: $y = c_1 * x^3 + c_2 * x^2 + c_3 * x^1 + c_4 * x^0$ mit den Koeffizienten: $[-7.80063911 * 10^{-18}; 3.95377876 * 10^{-08}; -6.67994362 * 10^{+1}; 3.76193671 * 10^{10}]$

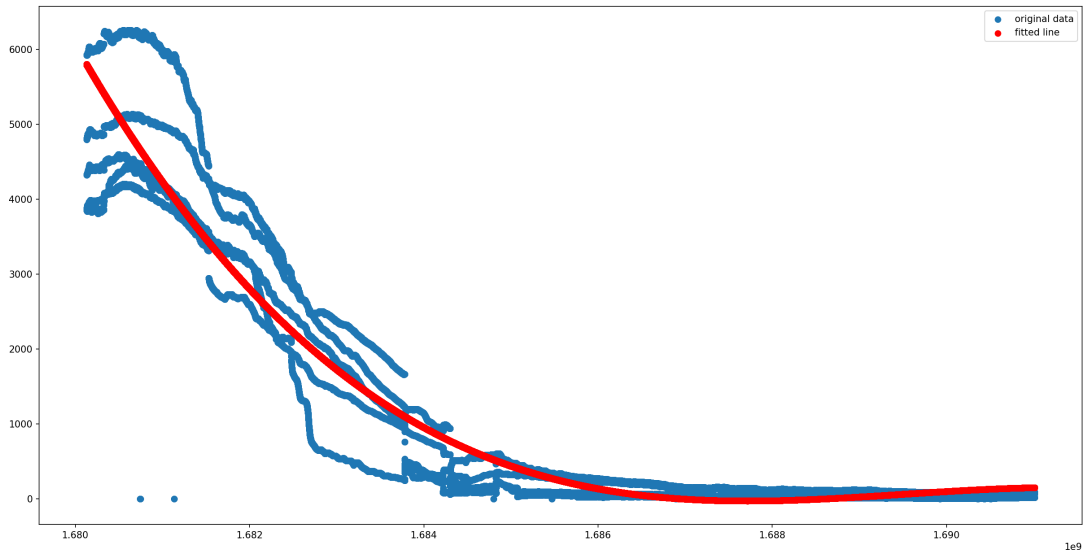


Abbildung 8: Polynomialregression Leitfähigkeit

Des Weiteren wurden in der Untersuchung des Beispieldatensatzes die Methoden der Dimensionsreduktion angewandt, um die mehrdimensionalen Daten in einem dreidimensionalen Datensatz in der Clusteranalyse zu untersuchen. Weitere Cluster-Darstellungen können den zugrundeliegenden Jupyter-Notebooks entnommen werden.

## 6.2 Auswahl benötigter Daten und Sensoren

Für die Untersuchung der Reduktion der Sensoranzahl wurde damit begonnen, die Daten aufzubereiten. Dabei wurden Label-Encoder verwendet, um nicht numerische Daten, wie die Helligkeit oder den Reifegrad, in numerische Daten umzuwandeln. Zudem wurden Robust-Scaler verwendet, welche die Daten auf Werte zwischen 0 und 1 skalieren und dabei weniger empfindlich gegenüber Ausreißern sind als andere Scaler. Von den beiden erhobenen Messreihen ließ sich leider nur die erste verwenden, da für die zweite Messreihe keine Annotationen vorhanden waren. Dadurch lassen sich die Modelle aus dem Bereich des überwachten Lernens nicht anwenden, da diese eine vorgegebene Klassifikation für das Training und die Bewertung benötigen.

### 6.2.1 Hyperparameter-Optimierung

Als Vorbereitung für die verschiedenen Ansätze wurde eine sogenannte Hyperparameter-Optimierung durchgeführt. Dabei werden die optimalen Hyperparameter (Parameter, die vor dem Training für ein Modell festgelegt werden) auf Basis eines Teils des Datensatzes ermittelt. Diese Hyperparameter wurden wie folgt festgelegt:

- **Neuronales Netzwerk**: Zwei versteckte Schichten mit je 50 Neuronen, Logistische Funktion als Aktivierungsfunktion, Stochastisches Gradientenverfahren zur

Gewichtsoptimierung nach Kingma und Lei Ba, Alpha = 0.0001, Konstante Lernrate von 0.01, maximal 100 Iterationen und Frühes Stoppen, wenn der Fehler auf einem automatisch erzeugten Validierungsdatensatz von 30% nicht mehr sinkt.

- **Random Forest**: 100 Entscheidungsbäume, Entropie als Kriterium für die Qualität der Aufteilung, Keine maximale Tiefe der Bäume, Mindestens 2 Datenpunkte für einen weiteren Knoten, Mindestens 1 Datenpunkt für ein Blatt, Keine maximale Anzahl an Features für die Aufteilung und Nutzung des gesamten Datensatzes für jeden Baum.
- **Support Vector Machine**: Regularisierungsparameter C = 1, Lineare Kernel-Funktion, Verwendung einer Heuristik zur Reduzierung der Anzahl an Stützvektoren, Toleranz für die Klassifizierungsgrenze = 1 und maximal 300 Iterationen für die Konvergenz.

Die Hyperparameter wurden dann für die verschiedenen Ansätze zur Reduktion der Sensoranzahl verwendet. Dabei wurde der Datensatz in Trainings- und Testdaten aufgeteilt, wobei die Trainingsdaten die Daten von drei der fünf Würste enthielten und die verbleibenden zwei Würste für die Testdaten verwendet wurden. So wurde sichergestellt, dass sowohl die Trainings- als auch die Testdaten den vollständigen Reifeprozess umfassen. Durch die Aufteilung der Daten basierend auf den Würsten wurde eine Teilung von 60/40 für Trainings- und Testdaten verwendet, dies kommt der häufig verwendeten Teilung von 70/30 sehr nahe. Mit diesen Daten wurden dann die drei Ansätze aus Abschnitt 4.4 durchgeführt.

### 6.2.2 Reduktion der Features

Begonnen wurde mit der Bestimmung der optimalen Kombination der Features. Dabei wurden alle möglichen Kombinationen von Features getestet und die jeweiligen Scores der verschiedenen Modelle bestimmt. In der nachfolgenden Abbildung 9 sind die Scores dieser Modelle für die unterschiedlichen Features dargestellt, wobei jeweils die Scores derjenigen Kombinationen einbezogen wurden, die das jeweilige Feature enthalten. Daraus lässt sich erkennen, welche Features in welchen Modellen einen besonders hohen Einfluss auf die Genauigkeit der Vorhersage des Reifegrads haben.

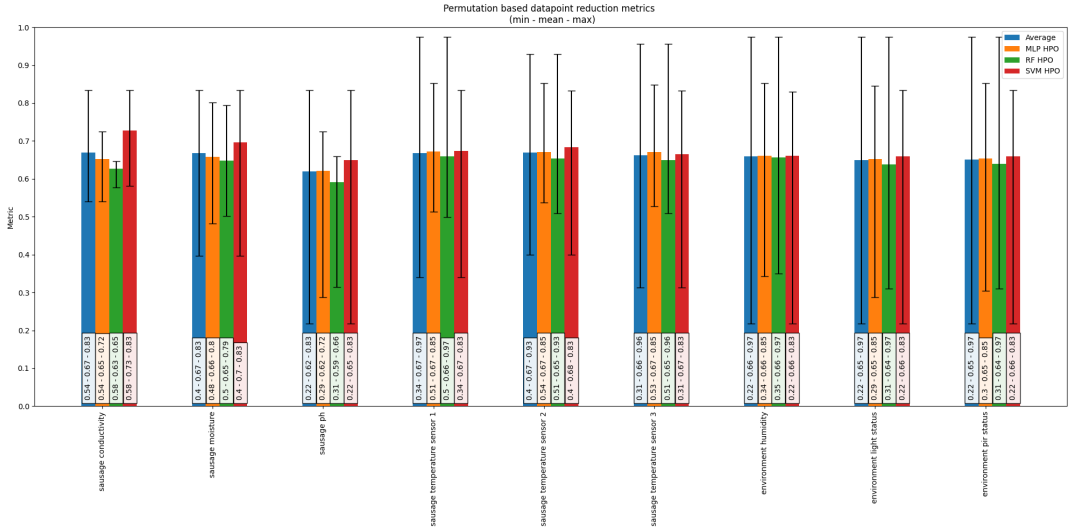


Abbildung 9: Scores der verschiedenen Modelle für die verschiedenen Features

In dieser Abbildung ist zu erkennen, dass die Features *Temperatur 1 (Umgebungstemperatur)*, *Luftfeuchtigkeit*, *Helligkeit* und *Bewegungsstatus* den höchsten score haben. Dieser beträgt 0.97 und wurde von allen Features mit dem Random Forest Modell erreicht. Da es sich bei diesen vier Messwerten um Umgebungssensoren handelt, ist hier der Einfluss des Zeitraums von Frühling bis Sommer zu bedenken. Dadurch existiert zwar eine hohe Korrelation zwischen der *Umgebungstemperatur* und der *Luftfeuchtigkeit* und dem Reifegrad, welche jedoch nicht durch den Reifegrad bedingt ist. Der hohe Einfluss des *Bewegungsstatus* und der *Helligkeit* lässt sich jedoch schwer erklären, da sich der Wert für die *Helligkeit* im zeitlichen Verlauf nicht ändert und der *Bewegungsstatus* dauerhaft wechselt.

Interessant ist auch, dass die Features *Leitfähigkeit*, *Wurst-Feuchtigkeit*, sowie *Temperatur 1 (Umgebungstemperatur)* und *Temperatur 2 (Wursttemperatur)* die höchsten durchschnittlichen Scores besitzen. Diese Features sind interessant, da sie mit Ausnahme der Umgebungstemperatur alle in den Würsten gemessen wurden.

In der nachfolgenden Tabelle 1 sind die jeweils besten Kombinationen von Features für die unterschiedlichen Modelle dargestellt.

Dabei lässt sich erkennen, dass der Random Forest mit den Features *Luftfeuchtigkeit*, *Helligkeit*, *Bewegungsstatus* und *Temperatur 1 (Umgebungstemperatur)* mit einem score von 0.97 am besten abschneidet. Die Support Vector Machine erreicht dagegen mit deutlich mehr Features als die anderen Modelle, den schlechtesten score von 0.83.

Besonders interessant an dieser Tabelle ist, dass die beiden Features *Bewegungsstatus* und *Umgebungstemperatur* von allen Modellen verwendet werden. Dies verdeutlicht den großen Einfluss der Umgebungstemperatur auf die Klassifikation des Reifegrads, auf Grund des bereits erwähnten Zusammenhangs zwischen der Umgebungstemperatur und

der Zeit. Der Einfluss des Bewegungsstatus wirft dagegen erneut Fragen auf.

| Score | Features | Model |
|---|---|---|
| 0.97 | humidity; light; pir status; temperature 1 | RF HPO |
| 0.85 | humidity; pir status; temperature 1; temperature 2 | MLP HPO |
| 0.83 | conductivity; light; pir status; moisture; ph; temperature 1 | SVM HPO |

Tabelle 1: Beste Kombinationen der Features für die verschiedenen Modelle

### 6.2.3 Reduktion der Sensoren

Für den zweiten Ansatz, der Reduktion der Sensoranzahl, wurden zunächst die Features den verschiedenen Sensoren zugewiesen. Die Zuweisung wurde verwendet, um im weiteren Verlauf die Features bestimmter Sensoren auswählen zu können. In der nachfolgenden Tabelle 2 sind diese Zuordnung dargestellt.

Wichtig das dabei zu erwähnen, dass auf Grund des Aufbaus des Datensatzes die Sensoren für die Leitfähigkeit (*DRA LSE01*) und den pH-Wert (*DRA LSPH01*) nicht differenziert werden können. Dies liegt daran, dass beide Sensoren Temperaturwerte messen und diese im Datensatz in einem Feature zusammengefasst wurden. Dies ist besonders störend, da diese beiden Sensoren die meisten Messwerte aus den Würsten liefern und somit besonders interessant sind.

| Name | Beschreibung |
|---|---|
| LSN50-v2-S31 | Environment temperature & humidity |
| DRA LSE01 + DRA LSPH01 | Sausage moisture; temperature & conductivity + ph |
| DRA LTC2-FS | Sausage temperature |
| MIL WS202 | Milesight environment PIR & lightintensity |

Tabelle 2: Zuordnung der Features zu Sensoren

Für die Untersuchung der Reduktion der Sensoranzahl wurden dann die Kombinationen aller Features der jeweils ausgewählten Sensoren getestet. Dadurch wurden nur Featurekombinationen echter Sensoren getestet, wobei immer alle Features eines Sensors enthalten waren.

In der nachfolgenden Abbildung 10 sind die Scores der verschiedenen Modelle für die Sensoren dargestellt. Dabei wurden wieder, wie bereits bei der featurebasierten Untersuchung, alle Kombination einbezogen, in denen der jeweilige Sensor enthalten war.

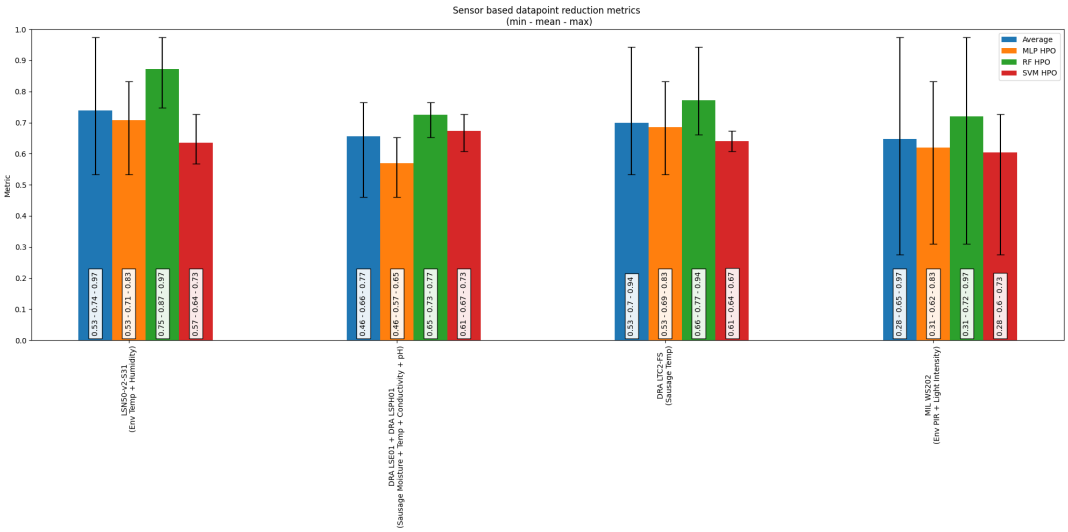


Abbildung 10: Scores der verschiedenen Modelle für die unterschiedlichen Sensoren

In der Abbildung ist zu erkennen, dass der sensorbasierte Ansatz höhere durchschnittliche Scores erreicht, als der featurebasierten Ansatz ($0.69 > 0.66$). Dies ist unter anderem darauf zurückzuführen, dass die Sensoren jeweils mehrere Features liefern und somit auch die durchschnittliche Anzahl an Features in den Kombinationen höher ist.

Weiter ist zu erkennen, dass die Umgebungssensoren *LSN50-v2-S31* und *MIL WS202* die höchsten maximalen Scores von 0.97 erreichen, auch wenn der Wurst-Temperatursensor *DRA LTC2-FS* mit einem maximalen Score von 0.94 nur knapp dahinter liegt.

Der hohe Score des Umgebungssensors für Temperatur und Luftfeuchtigkeit *LSN50-v2-S31* ist dabei wieder auf die bereits in der featurebasierten Untersuchung erläuterten hohen Korrelation zwischen diesen Daten und der Zeit zurückzuführen.

Auch der hohe Score des Wurst-Temperatursensors *DRA LTC2-FS* ist unter Umständen auf die steigende Umgebungstemperatur zurückzuführen, da sich hierdurch auch die Temperatur der Wurst erhöht.

Der hohe Score des Bewegungs- und Lichtsensors *MIL WS202* ist auch hier wieder schwer zu erklären, da diese Daten keinerlei erkennbaren Einfluss auf den Reifegrad haben. Dieser fehlende Einfluss ist auch in der Abbildung anhand der geringen durchschnittlichen Scores, sowie der großen Differenz zwischen maximalen und minimalen Scores zu erkennen. Der Sensor erreicht dabei den niedrigsten minimalen wie auch durchschnittlichen Score von 0.28 und 0.65. Der minimale Score wurde dabei bei der alleinigen Verwendung dieses Sensors erzielt.

In der nachfolgenden Tabelle 3 sind wieder die besten Kombinationen an Sensoren für die verschiedenen Model dargestellt. Dabei ist zu erkennen, dass erneut der Random Forest den höchsten Score von 0.97 erreicht und dabei die Sensoren *LSN50-v2-S31* und *MIL WS202* verwendet.

Zudem ist zu erkennen, dass die beiden Umgebungssensoren *LSN50-v2-S31* und *MIL WS202* von allen Modellen für den höchsten Score einbezogen werden. Diese stellt auch eine Bestätigung der Ergebnisse der featurebasierten Untersuchung auch Tabelle 1 dar. Dort waren die Umgebungstemperatur und der Bewegungsstatus aller Modelle verwendet worden, welche von den beiden Umgebungssensoren gemessen werden.

Interessant ist auch, dass das neuronale Netzwerk und die Support Vector Machine unterschiedliche weitere Sensoren verwenden und somit nicht ein bestimmter Sensor das Ergebnis verbessert.

| Score | Sensors | Model |
|---|---|---|
| 0.97 | LSN50-v2-S31; MIL WS202 | RF HPO |
| 0.83 | LSN50-v2-S31; MIL WS202; DRA LTC2-FS | MLP HPO |
| 0.72 | LSN50-v2-S31; MIL WS202; DRA LSE01 + DRA LSPH01 | SVM HPO |

Tabelle 3: Beste Kombinationen der Sensoren für die verschiedenen Modelle

### 6.2.4 Zeitlich begrenzte Reduktion der Sensoren

Als Grundlagen für den dritten Ansatz, der zeitlich begrenzten Reduktion der Sensoranzahl, wurde zunächst festgelegt, dass diese zeitliche Begrenzung im vierwöchigen Takt erfolgen soll.

Dazu wurden dann die Trainingsdaten jeweils nur bis zu einem bestimmten Zeitpunkt für das Training der Modelle verwendet. Der Testdatensatz wurde jedoch vollständig genutzt, da die Modelle auch für zukünftige Daten anwendbar sein sollen.

Leider konnten nur die Zeiträume ab acht Wochen genutzt werden, da für die ersten ca. 6 Woche, bis zum 13.05.2023, nur ein Reifegrad vorhanden war und die Modelle als Klassifikatoren mindestens zwei Klassen benötigen. Zusätzlich wurde auch der vollständige Datensatz mit einer Dauer von ca. 18 Wochen einbezogen.

In der nachfolgenden Abbildung 11 sind die Scores der verschiedenen Modelle jeweils für die unterschiedlichen Zeiträume dargestellt. Dabei gibt die Zeit in Wochen an, bis zu welchem Zeitpunkt die Daten für das Training verwendet wurden.

In dieser Abbildung ist ein nahezu linearer Anstieg der durchschnittlichen Scores zu erkennen, welcher durch $y = 0.0176x + 0.3553$ approximiert werden kann. Dies zeigt, dass die Modelle mit einer zunehmenden Anzahl an Datenpunkten besser trainiert werden und besonders auch den weiteren Verlauf besser klassifizieren können. Ein weiterer linearer Anstieg nach dem Ende des Datensatzes lässt sich daraus jedoch nicht direkt herleiten, da alle weiteren Datenpunkte dem letzten Reifegrad zugeordnet würden. Dies würde vorrausichtlich zu einem Overfitting der Modelle auf den letzten Reifegrad führen.

Vielmehr müssten weitere Messreihen zu unterschiedlichen Jahreszeiten aufgenommen werden, um die starke Korrelation zu der Umgebungssensoren zu reduzieren. Leider konnte die zweite aufgenommene Messreihe, wie bereits zu Anfang erwähnt, nicht verwendet werden, da keine Annotationen vorhanden waren.

Interessant ist, dass nach 16 Wochen ein starker Anstieg der maximalen Scores zu verzeichnen ist. Dies lässt sich darauf zurückführen, dass nach ca. 13 Wochen, am 30.06.2023,

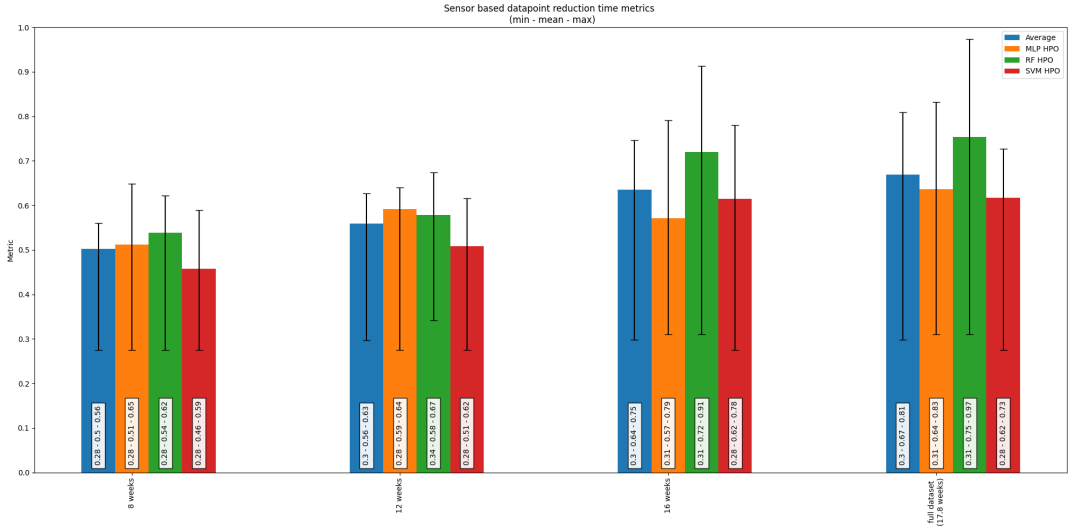


Abbildung 11: Scores der verschiedenen Modelle nach unterschiedlichen Zeiträumen

zum ersten Mal der dritte Reifegrad vorkommt und so die Klassifikation der späteren Daten erst möglich wird.

Auch zu den verschiedenen Klassifikatoren sind interessante Ergebnisse zu erkennen. So erreicht das neuronale Netzwerk für den ersten Zeitraum den höchsten Score von 0.65, was darauf hinweisen könnte, dass dieses Modell für einen stark eingeschränkten Datensatz besser geeignet sein könnte als die anderen.

Ab der 12. Woche erzielt der Random Forest den höchsten Score mit 0.67 und bleibt danach auch der beste Klassifikator. Dies deckt sich mit den Ergebnissen der ersten beiden Untersuchungen, bei denen der Random Forest immer am besten abgeschnitten hat. Der Random Forest benötigt demnach mehr Daten, um seine volle Leistung zu erreichen und verzeichnet auch zwischen den letzten beiden Zeiträumen den größten Anstieg des Scores um 0.06, von 0.91 auf 0.97.

Die Support Vector Machine schneidet, wie bereits in den ersten beiden Untersuchungen, durchgehend am schlechtesten ab und erreicht nur einen maximalen Score von 0.78. Zudem stellt die Support Vector Machine das einzige Modell dar, welches einen fallenden Score zwischen zwei Zeiträumen verzeichnet. So fällt der maximale Score zwischen der 16 Wochen und dem Ende des Datensatzes in der 18. Woche um 0.05 von 0.78 auf 0.73, wobei der durchschnittliche und minimale Score in diesem Zeitraum gleichbleiben. Dies könnte ein Indiz dafür sein, dass es hier zu Overfitting kommt, was sich jedoch nicht klar bestätigen lässt.

## 6.3 Ergebnisse und Ausblick

Die vorgenommene Untersuchung hat gezeigt, dass sich mit den vorgestellten Methoden Zusammenhänge erkennen und geeignete Merkmale analysieren lassen. In der Un-

tersuchung wurde festgestellt, dass verschiedene Sensoren im Zusammenhang mit dem Reifeprozess stehen. Zusammengefasst kann festgestellt werden, dass mit zunehmender Reife die Leitfähigkeit der Wurst nach einem geringen Anstieg gegen $0uS/cm$ abfällt, der PH-Wert unterlag nur einem geringen Abstieg. Die Korrelationsanalyse ergab einen Zusammenhang mit der Temperatur. Hier ist jedoch zu sagen, dass die Temperatur maßgeblich durch den gewählten Zeitraum durch den Sommerbeginn beeinflusst wird und so ein Trugschluss ermöglicht. Insgesamt ergab das Clustern mit den vorgestellten Methoden keinen umfassenden Erfolg. Jedoch konnten Teilsegmente des Reifegrades mit den Cluster-Methoden eingruppiert werden, sodass besonders hohe und niedrige Leitfähigkeitswerte, wie auch in der Annotation, unterschiedlichen Reifegraden zugeordnet wurde.

Zusätzlichen haben die drei Untersuchungen zu den verschiedenen Ansätzen des überwachten Lernens gezeigt, dass eine Reduktion der Features bzw. eine Auswahl der benötigten Sensoren mit Hilfe von künstlicher Intelligenz durchaus möglich ist. Mit Hilfe des ersten, featurebasierten, Ansatzes können die aussagekräftigsten Messwerte bestimmt werden. In der tatsächlichen Anwendung ließen sich so beispielsweise Messwert-Kombination bestimmen, für die dann passende Sensorik gesucht werden kann.

Da dies jedoch nicht immer möglich ist, stellt der zweite, sensorbasierte, Ansatz eine gute Möglichkeit dar, die Sensorauswahl auf die tatsächlich vorhandenen Sensoren zu beschränken. Auch eine Kombination der beiden Ansätze könnte in der konkreten Anwendung einen Mehrwert bieten, da so einerseits die benötigten Sensoren aus den vorhandenen ausgewählt werden können und andererseits basierend auf den gefundenen Messwert-Kombinationen neue Sensoren entwickelt oder angeschafft werden können, welche diese Messwerte liefern.

Der dritte Ansatz, die zeitliche Begrenzung des Datensatzes, hat gezeigt, dass es durchaus möglich ist, auch früher schon die benötigten Sensoren zu bestimmen und so die Einsatzdauer der anderen Sensoren zu reduzieren. Diese Ansätze müssen jedoch noch weiter untersucht werden, um den Einfluss unterschiedlicher Faktoren zu bestimmen und mögliche weitere Auswahlkriterien, sowie optimale Modelle zu finden.

Die Analyse des Datensatzes wurde durch verschiedene Kriterien beeinflusst. Um Trugschlüsse vorzubeugen und präzisere Ergebnisse zu erstellen, sollten für eine weitere Untersuchung von Messreihen die kritischen Einflüsse minimiert werden, wie beispielsweise der Einfluss der Temperatur durch einen Jahreszeitenwechsel. Daher hat sich gezeigt, dass ein diverser Datensatz als Grundlage notwendig ist, der auch verschiedene Zyklen, beispielsweise von Umwelteinflüssen und Temperaturschwankungen abdeckt. Des Weiteren muss der Datensatz sämtliche Klassen der Klassifikation in einem ausgewogenen und ausreichenden Verhältnis abdecken. Eine weitere Möglichkeit ist die Optimierung von Sensoren. Einige Sensoren waren in diesem Fall nicht für Lebensmittel konzipiert, sodass hierdurch fehlerbehaftete Daten und Ausreißer die Ergebnisse beeinflussen können. Letztlich existieren auch weitere Verfahren und Techniken, um weitere Untersuchungen zu ermöglichen. In diesem Fall wäre beispielsweise eine KI-Bildauswertung von Rohwürsten denkbar, um das Reifestadium einer Wurst mit KI zu bestimmen.

# Literatur

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth; *From Data Mining to Knowledge Discovery in Databases*; 1996

[2] Jürgen Cleve, Uwe Lämmel; *Data Mining*; 2020

[3] Jason Brownlee; *Autoencoder Feature Extraction for Classification*; 2020; `https://machinelearningmastery.com/autoencoder-for-classification/`; Letzter Zugriff 09.04.2024

[4] MeatMedia — metzgerfleisch.de; *Rohwurstherstellung*; `https://metzgerfleisch.de/fleisch-online-kaufen/fleisch-und-wurstgefluester/tipps-tricks-helferlein/rohwurst-herstellung-tipps-reifetabelle.html`; Letzter Zugriff 09.06.2024

[5] Jörg Frochte; *Maschinelles Lernen - Grundlagen und Algorithmen in Python*; 2021

[6] Ankur A. Patel; *Praxisbuch Unsupervised Learning*; 2020

[7] Thomas A. Runkler; *Data Mining - Modelle und Algorithmen intelligenter Datenanalyse*; 2. aktualisiere Auflage; 2015