Vergleich von häufigsten Clustering-Algorithmen auf Kaggle

Florian Merlau Friedrich-Alexander Universität
Erlangen, Deutschland
Email: florian.merlau@fau.de

***Zusammenfassung*—Clustering-Algorithmen spielen eine zentrale Rolle in der Data Science, insbesondere bei der explorativen Datenanalyse und im unüberwachten Lernen. Kaggle, eine führende Plattform für Data-Science-Wettbewerbe, ermöglicht detaillierte Einblicke in die Nutzung und Leistungsfähigkeit dieser Algorithmen. Diese Arbeit untersucht die am häufigsten verwendeten Clustering-Verfahren in Kaggle-Wettbewerben und analysiert deren Erfolgsraten anhand von Wettbewerbsscores. Ziel ist es, Zusammenhänge zwischen Popularität und Performance aufzuzeigen und zu bewerten, inwiefern die auf Kaggle bevorzugten Algorithmen auch für industrielle Anwendungen relevant sind. Die Ergebnisse zeigen, dass klassische Verfahren wie K-Means häufig verwendet werden, während komplexere Modelle unter bestimmten Bedingungen eine bessere Performance erzielen. Zudem wird analysiert, welche Faktoren die Wahl eines Algorithmus beeinflussen und welche Herausforderungen sich für den praktischen Einsatz ergeben. Diese Untersuchung liefert wertvolle Erkenntnisse für Data Scientists, die Clustering-Algorithmen sowohl in Wettbewerben als auch in industriellen Anwendungsfeldern einsetzen.**

I. Einführung

Clustering-Algorithmen stellen eine zentrale Methode in der Data Science dar und ermöglichen es, Datenpunkte anhand ihrer Ähnlichkeiten zu gruppieren. Sie kommen in vielfältigen Anwendungsbereichen zum Einsatz, beispielsweise bei der Segmentierung von Bild- und Textdaten, in der Kundensegmentierung oder in Empfehlungssystemen. über die Jahre hinweg haben sich zahlreiche Clustering-Verfahren etabliert, darunter auf Prototypen basierende Ansätze wie -Means, dichtebasierte Verfahren wie DBSCAN sowie hierarchische Methoden.

Kaggle, eine etablierte Plattform für Data-Science-Wettbewerbe, bietet häufig Wettbewerbe an, in denen Clustering-Aufgaben im Mittelpunkt stehen. Diese Wettbewerbe erlauben nicht nur einen Einblick in die derzeit bevorzugten Clustering-Algorithmen, sondern auch in deren Leistungsfähigkeit, gemessen an spezifischen Wettbewerbsscores. Darüber hinaus ist es von Interesse, ob sich die auf Kaggle beobachteten Präferenzen und Erfolge der Algorithmen auch in industriellen Anwendungen widerspiegeln, wo Aspekte wie Interpretierbarkeit, Skalierbarkeit und Ressourcenaufwand oftmals eine entscheidende Rolle spielen.

Dieser Beitrag entstand im Rahmen des “Big Data Seminar”s, das im Wintersemester 2024/2025 vom Lehrstuhl für Informatik 6 (Datenmanagement) der Friedrich-Alexander Universität Erlangen-Nürnberg durchgeführt wurde.

Das Ziel dieser Arbeit ist es, die am häufigsten verwendeten Clustering-Algorithmen aus jüngsten Clustering-Wettbewerben auf Kaggle systematisch zu untersuchen und zu bewerten. Neben der Häufigkeit des Einsatzes wird insbesondere auch der erzielte Wettbewerbsscore analysiert, um mögliche Zusammenhänge zwischen Beliebtheit und Erfolg einzelner Verfahren aufzuzeigen. Darauf aufbauend soll diskutiert werden, inwieweit sich aus den Ergebnissen Rückschlüsse auf industrielle Anwendungsszenarien ziehen lassen. Die Ergebnisse dieser Untersuchung liefern damit einen Überblick über aktuelle Trends und bieten Anhaltspunkte für die Wahl geeigneter Clustering-Verfahren in Forschung und Praxis.

Der Aufbau dieser Arbeit ist wie folgt gegliedert: Nach der *Einleitung* in Abschnitt I werden in Abschnitt II die relevanten Arbeiten aus dem Bereich *Related Work* vorgestellt. Abschnitt III bietet einen *Hintergrund* zum Thema Clustering, wobei zentrale Konzepte und gängige Verfahren erläutert werden. In Abschnitt IV wird anschließend die *Methodik* dieser Arbeit vorgestellt, einschließlich Datenerhebung und Analysevorgehen. Darauf folgt in Abschnitt V eine *Analyse der Ergebnisse*, bevor in Abschnitt VIII ein *Vergleich der Clustering-Algorithmen* durchgeführt wird. In Abschnitt **??** werden die Resultate ausführlich *evaluiert* und Schlussfolgerungen für Praxis und Forschung gezogen. Abschließend bietet Abschnitt XII eine *Conclusion*, während Abschnitt XIII einen *Ausblick* auf mögliche zukünftige Forschungs- und Anwendungsperspektiven gibt.

II. Verwandte Arbeiten

Die wissenschaftliche Auseinandersetzung mit Clustering-Algorithmen umfasst drei zentrale Bereiche: (1) grundlegende Taxonomien und Methodenvergleiche, (2) plattformbasierte Analysen von Data-Science-Wettbewerben sowie (3) Leistungsbewertungen in spezifischen Anwendungsszenarien.

A. Grundlegende Taxonomien

Jain et al. [1] liefern eine weit verbreitete Klassifikation, in der partitionierende (z. B. *K-Means*), hierarchische (z. B. *Agglomeratives Clustering*), dichtebasierte Verfahren (z. B. *DBSCAN*) und modellbasierte Ansätze unterschieden werden. Ergänzend untersuchen Xu und Wunsch [2] die Verfahren hinsichtlich ihrer Skalierbarkeit, Robustheit gegen Rauschen und Eignung im Umgang mit hochdimensionalen Daten.

B. Kaggle-spezifische Analysen und Einführung in Kaggle

Untersuchungen der Plattform zeigen, dass erfolgreiche Lösungsansätze oft auf der Kombination mehrerer Modelle – also Ensemble-Methoden – basieren. Dieser Ansatz, der in der traditionellen Clustering-Literatur weniger stark im Fokus steht, wird in zahlreichen Fallstudien belegt [3]. Zudem hebt Hillenbrand hervor, dass es signifikante methodische und zielorientierte Unterschiede zwischen den auf Kaggle erarbeiteten Lösungen und den in industriellen Anwendungen üblichen Ansätzen gibt [4].

III. Grundlagen

Clustering ist eine Methode des unüberwachten Lernens, bei der Daten in Gruppen aufgeteilt werden. Dabei werden ähnliche Objekte in dieselbe Gruppe sortiert, während unterschiedliche Objekte getrennt werden. Im Gegensatz zum überwachten Lernen, bei dem ein Modell mit bereits bekannten Kategorien trainiert wird [2], erkennt Clustering Muster in den Daten, ohne dass vorher festgelegt wurde, welche Gruppen es gibt. „Das Ziel des Clusterings ist es, eine Menge nicht gekennzeichneter Daten in verschiedene natürliche Gruppen zu unterteilen“ [2].

Es gibt verschiedene Clustering-Methoden, die je nach Art der Daten und dem Ziel der Analyse unterschiedlich gut geeignet sind.

A. Partitionierende Verfahren

Ein bekanntes Beispiel ist der K-Means-Algorithmus. Dieser teilt einen Datensatz in eine vorgegebene Anzahl von Clustern, indem er versucht, Punkte so zu gruppieren, dass sie nahe an einem mittleren Wert (dem Clusterzentrum) liegen. Praktisch wird K-Means etwa in der E-Commerce-Kundensegmentierung eingesetzt, um Kunden anhand ihres Kaufverhaltens zu gruppieren [1].

B. Hierarchische Verfahren

Hierbei werden Daten schrittweise zusammengeführt (agglomeratives Clustering) oder aufgeteilt (divisives Clustering), sodass eine baumartige Struktur (Dendrogramm) entsteht. Dies hilft zum Beispiel bei der Analyse von Genexpressionsdaten, wo die hierarchische Ähnlichkeit von großer Bedeutung ist [5].

C. Dichtebasierte Methoden

Das DBSCAN-Verfahren gruppiert Datenpunkte, die in ihrer Umgebung dicht beieinander liegen. Dadurch können auch Cluster mit unregelmäßiger Form erkannt und Ausreißer identifiziert werden. Anwendungen finden sich etwa in der Geodatenanalyse, wo räumliche Daten in sinnvolle Gruppen unterteilt werden [6].

D. Weitere Clustering-Ansätze

Neben den genannten Methoden existieren modellbasierte Ansätze wie Gaussian Mixture Models, die davon ausgehen, dass die Daten aus einer Mischung verschiedener Wahrscheinlichkeitsverteilungen bestehen. Auch unscharfe Clustering-Verfahren, bei denen ein Datenpunkt mehreren Clustern zugeordnet werden kann, finden Anwendung – beispielsweise in der Bildsegmentierung [1].

E. Bewertung der Clustering-Ergebnisse

Die Bewertung der Clustering-Ergebnisse erfolgt anhand verschiedener Metriken, die die Qualität der Clusterstruktur analysieren. Eine häufig verwendete Methode ist der Silhouettenkoeffizient, der misst, wie gut ein Punkt innerhalb seines Clusters liegt, indem er die durchschnittliche Distanz zu Punkten desselben Clusters mit der Distanz zu Punkten anderer Cluster vergleicht. Ein hoher Wert deutet auf eine gute Clustertrennung hin. [6]

Zusätzlich wird der Adjusted Rand Index (ARI) verwendet, um das gefundene Clustering mit einer bekannten Referenzklassifikation zu vergleichen. Der ARI bewertet, inwieweit die Zuordnungen der Datenpunkte zu Clustern mit einer vorgegebenen Klasseneinteilung übereinstimmen. Ein Wert nahe eins zeigt eine hohe Übereinstimmung an, während ein Wert nahe null auf eine zufällige Clusterbildung hindeutet. [6]

F. Kaggle

Kaggle ist eine Online-Community-Plattform für Datenwissenschaftler und Machine-Learning-Enthusiasten. Sie ermöglicht es Nutzern, miteinander zu kollaborieren, Datensätze zu finden und zu veröffentlichen, GPU-integrierte Notebooks zu verwenden und an Wettbewerben teilzunehmen, um datenwissenschaftliche Herausforderungen zu lösen. [7]

Ein zentrales Element von Kaggle sind die Wettbewerbe, bei denen Unternehmen und Organisationen große Mengen an Daten und herausfordernde Aufgabenstellungen bereitstellen. Teilnehmer konkurrieren dabei, um die besten Modelle zur Lösung dieser Aufgaben zu entwickeln. Diese Wettbewerbe bieten nicht nur die Möglichkeit, praktische Erfahrungen zu sammeln, sondern auch, von anderen zu lernen und innovative Ansätze zu entdecken [8].

Zusätzlich zu den Wettbewerben bietet Kaggle eine Vielzahl von Ressourcen, darunter öffentlich zugängliche Datensätze, die Möglichkeit, eigene Datensätze hochzuladen, und die Nutzung von Kaggle Notebooks, die es ermöglichen, Code direkt im Browser zu schreiben und auszuführen. [7]

IV. Methodik

Dieser Abschnitt beschreibt das methodische Vorgehen zur Identifizierung und Analyse von Kaggle-Wettbewerben mit Fokus auf Clustering-Verfahren. Ziel ist es, die Häufigkeit des Einsatzes verschiedener Clustering-Algorithmen zu ermitteln und deren Leistung im Vergleich zu bewerten.

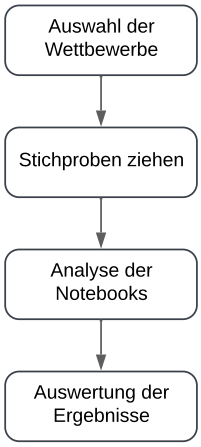

Abbildung 1. Ablaufdiagramm der Mehodik, nach eigener Darstellung

A. Fokus auf Clustering-Wettbewerbe

Der Schwerpunkt dieser Untersuchung liegt auf Wettbewerben, die speziell für Clustering-Aufgaben konzipiert sind. Clustering ist eine Methode des unüberwachten Lernens, bei der ähnliche Datenpunkte ohne vorherige Labels in Gruppen eingeteilt werden (siehe Grundlagenkapitel III). Durch die Konzentration auf solche Wettbewerbe wird sichergestellt, dass die analysierten Notebooks relevante Clustering-Algorithmen implementieren. Dies minimiert das Risiko, Datensätze einzubeziehen, die für andere Analyseformen vorgesehen sind, und gewährleistet aussagekräftige Ergebnisse für den Bereich des unüberwachten Lernens.

B. Auswahl der Wettbewerbe

Es sollen die letzten 20 vergangenen Kaggle-Wettbewerbe ausgewählt werden, die für die Anwendung von Clustering-Methoden geeignet sind. Die Auswahlkriterien umfassen:

- **Datencharakteristik:** Unbeschriftete und nicht zeitreihenbasierte Datensätze, die unüberwachtes Lernen ermöglichen.
- **Bewertungsmetriken:** Verwendung von Metriken, die die Qualität der Clusterbildung bewerten, wie z.,B. der Silhouettenkoeffizient oder der Adjusted Rand Index.
- **Teilnehmerzahl:** Mindestens 100 teilnehmende Teams pro Wettbewerb, um sicherzustellen, dass genügend Notebooks für die Analyse verfügbar sind. Es ist jedoch zu beachten, dass eine hohe Teilnehmerzahl keine Garantie für eine ausreichende Anzahl von Notebooks bietet.

Bei der Auswahl der Wettbewerbe sollen die Beschreibungen und Dateninformationen sorgfältig gelesen werden, um sicherzustellen, dass die genannten Metriken, das Ziel des Wettbewerbs und die bereitgestellten Daten für Clustering-Aufgaben geeignet sind.

C. Stichprobenziehung der Notebooks

Für jeden identifizierten Wettbewerb wird eine Stichprobe von 50 öffentlich zugänglichen Notebooks gezogen. Diese Stichprobengröße wurde gewählt, um eine repräsentative Analyse der angewandten Methoden sicherzustellen, ohne den praktischen Rahmen der Untersuchung – insbesondere hinsichtlich Zeit, Personal und Rechenressourcen – zu überschreiten. Die Festlegung dieser Größe orientiert sich an etablierten Empfehlungen zur statistischen Power und Generalisierbarkeit von Forschungsergebnissen [9]–[11].

Die Stichprobe wird systematisch in einer Exceltabelle dokumentiert (siehe Abbildung 2). Diese Tabelle enthält folgende Informationen zu jedem Notebook:

- Name der Competition
- Anzahl der insgesamt verfügbaren Notebooks für die Competition
- Direktlink zum jeweiligen Notebook

Diese strukturierte Erfassung ermöglicht eine übersichtliche Verwaltung der Notebooks und bildet die Grundlage für die anschließende Analyse.

Competition Slug	Total Notebooks	Notebook Link	Score	Algo
dmassign1	64	https://www.kaggle	0.40427	KMeans
dmassign1	64	https://www.kaggle	-	AgglomerativeClustering
dmassign1	64	https://www.kaggle	-	AgglomerativeClustering

Abbildung 2. Exceltabelle mit den gesammelten Stichprobennotebooks, inklusive Wettbewerb, Notebook-Link und Analyseergebnissen (eigene Darstellung).

D. Analyse der Notebooks

Die ausgewählten Notebooks werden einer manuellen Analyse unterzogen, wobei sowohl Code-Segmente als auch begleitende Texte berücksichtigt werden. Der Fokus liegt dabei auf den implementierten Clustering-Algorithmen. Zusätzlich wird der in den Notebooks angegebene Score der jeweiligen Algorithmen erfasst, um deren relative Leistungsfähigkeit zu bewerten.

Die Ergebnisse dieser Analyse – d.h. die verwendeten Algorithmen sowie die entsprechenden Scores – werden ebenfalls in die Exceltabelle (Abbildung 2) eingetragen. Dies ermöglicht eine systematische Auswertung der eingesetzten Methoden und deren Performance über verschiedene Wettbewerbe hinweg.

E. Vergleich der Clustering-Algorithmen

Nach der Analyse der Ergebnisse erfolgt ein systematischer Vergleich der fünf bis sechs am häufigsten eingesetzten Clustering-Algorithmen. Der Vergleich basiert sowohl auf den erzielten Scores als auch auf spezifischen Charakteristika der Algorithmen und der verwendeten Daten.

Zunächst werden die Ergebnisse der Algorithmen gegenübergestellt, um deren relative Leistung zu bewerten. Dabei wird analysiert, welcher Algorithmus in welchen Szenarien besser abschneidet und welche Faktoren diese Unterschiede beeinflussen. Insbesondere wird untersucht, wie sich unterschiedliche Datenverteilungen, Cluster-Formen und Hyperparametereinstellungen auf die Leistung der Algorithmen auswirken.

Darüber hinaus erfolgt eine detaillierte Betrachtung der Vor- und Nachteile der einzelnen Algorithmen. Hierbei werden Aspekte wie Rechenkomplexität, Skalierbarkeit und Sensitivität gegenüber Ausreißern berücksichtigt.

F. Einschränkungen in den Wettbewerben

Es ist zu beachten, dass der in einem Notebook angegebene Score nicht zwingend dem tatsächlich eingereichten Modell entspricht, da die Einreichung als Team erfolgt und das Notebook möglicherweise nicht das finale Modell widerspiegelt. Daher wird in dieser Analyse ausschließlich der in den Notebooks dokumentierte Score berücksichtigt.

V. Durchführung

Im Folgenden wird die praktische Umsetzung der in Abschnitt IV definierten Methodik beschrieben. Ziel war es, geeignete Kaggle-Wettbewerbe zu identifizieren, die als Plattform für die Analyse von Clustering-Methoden dienen, und anschließend relevante Notebooks systematisch auszuwerten, um Erkenntnisse über die Verbreitung und Performance verschiedener Clustering-Algorithmen zu gewinnen.

VI. Auswahl der Wettbewerbe

Zunächst wurden **80 der zuletzt veröffentlichten Kaggle-Wettbewerbe** dahingehend untersucht, ob sie grundsätzlich als Clustering-Herausforderung geeignet sind. Die Wahl dieser Anzahl ergab sich aus zeitlichen und ressourcenbedingten Einschränkungen, wobei angenommen wurde, dass in diesem Umfang ein repräsentativer Querschnitt aktueller Fragestellungen abgebildet werden kann. Für jeden Wettbewerb wurden das zugehörige Datenset sowie die offizielle Wettbewerbsbeschreibung überprüft und anhand der in Abschnitt IV definierten Kriterien (z. B. Vorhandensein unlabeled Daten, Ausschluss von Zeitreihen, eindeutige Bewertungsmetrik) bewertet.

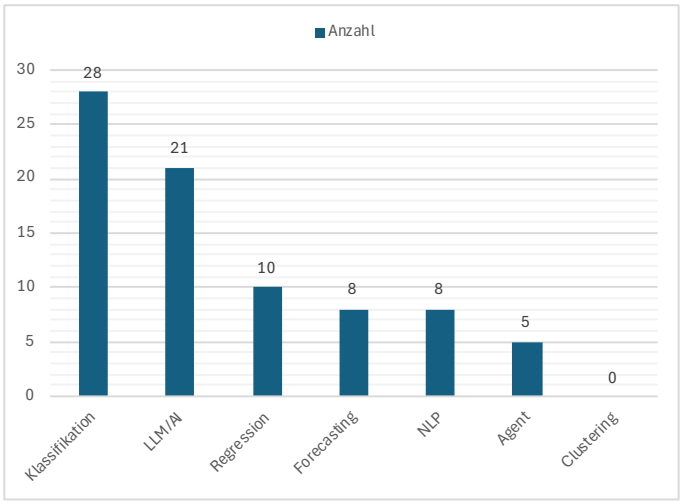

Abbildung 3. Verteilung der analysierten Kaggle-Wettbewerbe nach Typ (eigene Darstellung)

Wie Abbildung 3 zeigt, lag der Schwerpunkt der ausgewerteten Wettbewerbe nahezu ausschließlich auf LLM-bezogenen oder klassifikationsorientierten Aufgaben. Es konnte kein Wettbewerb identifiziert werden, der unmittelbar die Anforderungen einer Clustering-Herausforderung erfüllte. Daher wurde eine **gezielte Recherche** unter Zuhilfenahme der Kaggle-Such- und Filterfunktionen durchgeführt. Bei dieser Suche standen die Schlagwörter `Clustering` und `Unsupervised` im Vordergrund, um Wettbewerbe mit unüberwachtem Lernfokus zu identifizieren. Die Suche lieferte **94 Treffer**, die anschließend anhand zusätzlicher Mindestanforderungen (mindestens 100 teilnehmende Teams und mindestens 50 öffentlich verfügbare Notebooks) weiter gefiltert wurden.

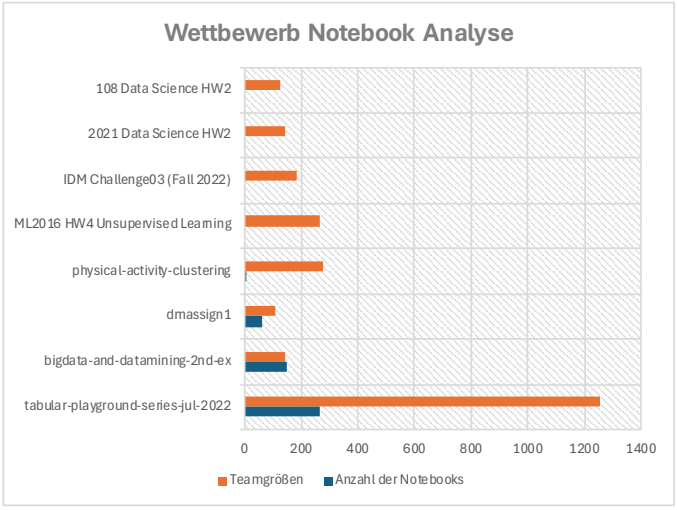

Abbildung 4. Anteil geeigneter Clustering-Wettbewerbe nach Filterkriterien (eigene Darstellung)

Letztlich konnten drei geeignete Clustering-Wettbewerbe identifiziert und für die weitere Analyse ausgewählt werden:

- `tabular-playground-series-jul-2023`
- `bigdata-and-datamining-2nd-ex`
- `dmassign7`

Alle drei Wettbewerbe erfüllen die definierten Kriterien und weisen eine ausreichende Anzahl an teilnehmenden Teams und öffentlich verfügbaren Notebooks auf, sodass eine aussagekräftige Analyse der Clustering-Ansätze möglich ist.

VII. Stichprobenziehung und Analyse der Notebooks

Gemäß der im Methodikteil IV beschriebenen Vorgehensweise wurden die Links zu den relevanten Notebooks sowie die zugehörigen Namen in einer Excel-Tabelle dokumentiert. Die Stichprobenziehung erfolgte vollständig automatisiert mithilfe eines Python-Skripts, das in Kapitel Implementierung IX erläutert wird. Die ermittelten Ergebnisse wurden ebenfalls in der Excel-Tabelle festgehalten.

In einem weiteren Schritt wurden die in der Excel-Tabelle gespeicherten Notebook-Links einzeln geöffnet und systematisch untersucht. Dabei wurde überprüft, welche Clustering-Algorithmen implementiert wurden (beispielsweise anhand von Import-Anweisungen wie `sklearn.KMeans`) und welche Informationen in der Beschreibung der Notebooks zu finden sind. Darüber hinaus wurden, sofern verfügbar, die

erreichten Scores der jeweiligen Modelle erfasst. Die gewonnenen Erkenntnisse wurden abschließend in der Excel-Tabelle zusammengeführt und bilden die Grundlage für die anschließende Analyse der Clustering-Ansätze.

VIII. Vorstellung der ausgewählten Wettbewerbe

In diesem Kapitel werden die zuvor ausgewählten Wettbewerbe hinsichtlich ihrer inhaltlichen Schwerpunkte und Herausforderungen im Bereich des unüberwachten maschinellen Lernens näher untersucht. Dabei werden die Aufgabenstellungen, verwendeten Datensätze und Bewertungsmethoden in den einzelnen Wettbewerben analysiert, um ein tieferes Verständnis der praktischen Anforderungen in diesem Forschungsfeld zu gewinnen.

A. Tabular Playground Series – Juli 2022

Im Wettbewerb *Tabular Playground Series – Juli 2022* wurden tabellarische Datensätze bereitgestellt, bei denen jede Zeile einem bestimmten Cluster zugeordnet war [12]. Die Teilnehmenden hatten die Aufgabe, für jede Zeile den richtigen Cluster zu identifizieren. Da die Anzahl der Cluster nicht vorab festgelegt wurde, musste mit verschiedenen Clustering-Methoden experimentiert werden.

Der Datensatz enthielt sowohl kontinuierliche als auch kategoriale Merkmale, was den Einsatz unterschiedlicher Vorverarbeitungstechniken erforderte. Zur Bewertung der Ergebnisse kam der *Adjusted Rand Index* (ARI) zum Einsatz. Dieser Index misst die Übereinstimmung der ermittelten Cluster mit den tatsächlichen Clustern und berücksichtigt zufällige Übereinstimmungen, wodurch er eine robuste Bewertung der Clustering-Qualität ermöglicht [12].

B. BigData & DataMining – 2nd Ex

Der Wettbewerb *BigData & DataMining – 2nd Ex* fokussiert sich auf die Identifikation von Gruppen in einem Datensatz mittels verschiedener Clustering-Algorithmen [13]. Eine zentrale Herausforderung in diesem Wettbewerb ist die Vorverarbeitung der Daten. Die Normalisierung der Daten ist essenziell, um Verzerrungen durch unterschiedliche Skalierungen zu vermeiden.

Ein weiteres Problem stellt die hohe Dimensionalität der Daten dar, die klassische Clustering-Methoden oft vor Probleme stellt. Daher sind effiziente Algorithmen und Techniken zur Reduktion der Datenkomplexität gefragt. Zusätzlich führt die moderate Unausgewogenheit der Klassen zu Herausforderungen, da einige Clustering-Verfahren empfindlich auf unterschiedliche Clustergrößen reagieren. Die Aufgabenstellung erlaubt es, die Anzahl der Cluster entweder im Voraus festzulegen oder mithilfe externer Validierungsmetriken zu bestimmen [13].

C. DM-Assignment 1

Beim Wettbewerb *DM-Assignment 1* stand die alleinige Anwendung von Clustering-Methoden im Mittelpunkt, um Instanzen fünf vorgegebenen Klassen zuzuordnen [14]. Der bereitgestellte Datensatz umfasste 13 000 Instanzen mit 199 Attributen, die sowohl numerische als auch kategoriale Merkmale enthielten. Eine explorative Datenanalyse war notwendig, um die relevanten Strukturen in diesen hochdimensionalen und teilweise verrauschten Daten zu identifizieren.

Die Aufgabenstellung verlangte, geeignete Clustering-Algorithmen zu finden, die die Instanzen so gruppieren, dass die resultierenden Cluster den vorgegebenen Klassen möglichst entsprechen. Dabei waren insbesondere die Wahl der Vorverarbeitungsschritte (wie Normalisierung und Dimensionsreduktion) und die Feinabstimmung der Hyperparameter von großer Bedeutung. Ein besonderes Merkmal des Wettbewerbs war die Möglichkeit, mehrfach Einsendungen vorzunehmen: Täglich konnten bis zu fünf Lösungen eingereicht werden. Die Bewertung erfolgte in zwei Phasen, wobei 50 % der Daten im *Public Leaderboard* und die restlichen 50 % im *Private Leaderboard* zur Beurteilung herangezogen wurden. Diese Aufteilung verhindert eine Überanpassung an das öffentliche Bewertungssystem und stellt sicher, dass generalisierbare Lösungen bevorzugt werden. [14]

IX. Implementierung

In dieser Arbeit wurde ein Python-Skript entwickelt, das die Kaggle-API nutzt, um automatisch Notebooks aus verschiedenen Wettbewerben abzurufen und in einer Excel-Datei zusammenzufassen. Zunächst werden die benötigten Bibliotheken importiert und die Authentifizierung für die Kaggle-API durchgeführt. Ein Dictionary speichert die URLs der Wettbewerbe und die zugehörigen Slugs, um die API-Abfragen zu erleichtern. Das Skript erstellt einen Ausgabeordner für die Excel-Datei. Für jeden Wettbewerb im Dictionary werden mittels Pagination (Aufteilung der Daten in mehrere Seiten) alle verfügbaren Notebooks abgerufen, wobei pro Seite bis zu 100 Notebooks angefordert werden. Fehler wie ein 404-Fehler bei fehlendem Zugriff oder nicht vorhandenen Wettbewerben werden abgefangen und entsprechend behandelt. Nach der Erfassung der Notebooks für einen Wettbewerb wählt das Skript zufällig bis zu 50 Notebooks aus. Für jedes dieser Notebooks werden Informationen wie die Wettbewerbs-URL, der Wettbewerb-Slug, die Gesamtzahl der gefundenen Notebooks sowie der direkte Link zum jeweiligen Notebook extrahiert und in einer Liste gespeichert. Abschließend werden die gesammelten Daten in ein Pandas DataFrame überführt und als Excel-Datei im definierten Ausgabeordner gespeichert.

Weitere Details und den vollständigen Code ist im GitHub-Repository[1] zu finden.

X. Evaluation der Clustering-Algorithmen

In dieser Arbeit wurde die Performance sowie die Nutzungshäufigkeit verschiedener Clustering-Algorithmen in mehreren Kaggle-Wettbewerben untersucht. Im Folgenden werden zunächst die aggregierten Ergebnisse dargestellt und anschließend die Ergebnisse einzelner Wettbewerbsreihen beschrieben.

[1] https://github.com/XplorodoX/BigData

A. Aggregierte Ergebnisse

Die in Tabelle I aufgeführten aggregierten Werte, sortiert nach der Nutzungshäufigkeit, zeigen, dass **KMeans** mit 64 Einsätzen und einem durchschnittlichen Score von 0.5233 den höchsten Stellenwert besitzt. Es folgen **AgglomerativeClustering** (21 Einsätze, Score 0.3167), **SpectralClustering** (17 Einsätze, Score 0.2967) sowie **GaussianMixture** (16 Einsätze, Score 0.29). **BayesianGaussianMixture** wurde 14-mal eingesetzt und erzielte dabei einen Score von 0.1225. Die weiteren Algorithmen werden seltener verwendet und erzielen teilweise einen Score von 0.

Model	Score	Anzahl
KMeans	0.5233	64
AgglomerativeClustering	0.3167	21
SpectralClustering	0.2967	17
GaussianMixture	0.29	16
BayesianGaussianMixture	0.1225	14
Birch	0.43	6
DBSCAN	0	3
MiniBatchKMeans	0.3467	2
OPTICS	0	1

Tabelle I
AGGREGIERTE ERGEBNISSE ÜBER ALLE WETTBEWERBE, SORTIERT NACH DER NUTZUNGSHÄUFIGKEIT.

Auf Basis dieser Ergebnisse wurden für die detaillierte Analyse die Algorithmen **KMeans**, **AgglomerativeClustering**, **SpectralClustering**, **GaussianMixture** und **BayesianGaussianMixture** ausgewählt, da diese sowohl in der Häufigkeit als auch in der Performance hervorstechen.

B. Einzelwettbewerbsanalysen

Im Folgenden werden die Resultate der einzelnen Wettbewerbsreihen in prägnanter Form erläutert.

1) dmassign-Reihe: In der dmassign-Reihe dominiert **KMeans**: In *dmassign1* wurde es in 23 Fällen mit einem Score von 0.4 eingesetzt. Im Gegensatz dazu erzielte **AgglomerativeClustering** in *dmassign2* trotz 16 Einsätzen einen Score von 0, was darauf hindeutet, dass es hier die spezifische Datenstruktur nicht adäquat abbilden konnte. Außerdem erreichte **Birch** in *dmassign3* mit 3 Einsätzen einen Score von 0.33, während **DBSCAN** in *dmassign4* – wenn auch nur in einem Fall – ebenfalls einen Score von 0 erzielte. Die übrigen Algorithmen wurden in dieser Reihe entweder gar nicht eingesetzt oder führten zu einem Score von 0.

Model	Score	Wettbewerb	Anzahl
KMeans	0.4	dmassign1	23
AgglomerativeClustering	0	dmassign2	16
Birch	0.33	dmassign3	3
DBSCAN	0	dmassign4	1
SpectralClustering	0	dmassign5	0
GaussianMixture	0	dmassign6	0
MiniBatchKMeans	0	dmassign7	0
OPTICS	0	dmassign8	0
BayesianGaussianMixture	0	dmassign9	0

Tabelle II
ERGEBNISSE FÜR DIE DMASSIGN-WETTBEWERBSREIHE.

2) bigdata-and-datamining-2nd-ex: Im Wettbewerb *bigdata-and-datamining-2nd-ex* wurden deutlich höhere Scores erzielt. **KMeans** wurde in 22 Fällen mit einem Score von 0.87 verwendet, während **SpectralClustering** in 17 Einsätzen einen Score von 0.89 erreichte. Besonders hervorzuheben ist, dass **AgglomerativeClustering** mit 4 Einsätzen einen Score von 0.95 erzielte und **Birch** in 3 Einsätzen mit einem Score von 0.96 überzeugen konnte. Im Gegensatz dazu erzielte **GaussianMixture** einen moderaten Score von 0.47 und **MiniBatchKMeans** einen Score von 0.84. Dicht-basierte Methoden wie **DBSCAN**, **OPTICS** und **BayesianGaussianMixture** wurden in diesem Wettbewerb nicht erfolgreich eingesetzt.

Model	Score	Wettbewerb	Anzahl
KMeans	0.87	bigdata-and-datamining-2nd-ex	22
SpectralClustering	0.89	bigdata-and-datamining-2nd-ex	17
AgglomerativeClustering	0.95	bigdata-and-datamining-2nd-ex	4
Birch	0.96	bigdata-and-datamining-2nd-ex	3
GaussianMixture	0.47	bigdata-and-datamining-2nd-ex	1
MiniBatchKMeans	0.84	bigdata-and-datamining-2nd-ex	1
DBSCAN	0	bigdata-and-datamining-2nd-ex	0
OPTICS	0	bigdata-and-datamining-2nd-ex	0
BayesianGaussianMixture	0	bigdata-and-datamining-2nd-ex	0

Tabelle III
ERGEBNISSE FÜR DEN BIGDATA-AND-DATAMINING-2ND-EX WETTBEWERB.

3) tabular-playground-series: Die Auswertung der tabular-playground-Serie ergab, dass **BayesianGaussianMixture** in 14 Einsätzen mit einem Score von 0.49 den höchsten Score erreichte, gefolgt von **GaussianMixture** mit 15 Einsätzen und einem Score von 0.4. **KMeans** wurde in 19 Fällen eingesetzt und erzielte dabei einen Score von 0.3. **MiniBatchKMeans** wurde nur einmal verwendet und erreichte dabei einen Score von 0.2. Die anderen Algorithmen, insbesondere dicht-basierte Verfahren wie **DBSCAN**, **AgglomerativeClustering** und **OPTICS**, zeigten in diesem Wettbewerb keine zufriedenstellende Performance.

Model	Score	Wettbewerb	Anzahl
KMeans	0.3	tabular-playground-series-jul-2022	19
GaussianMixture	0.4	tabular-playground-series-jul-2027	15
BayesianGaussianMixture	0.49	tabular-playground-series-jul-2030	14
DBSCAN	0	tabular-playground-series-jul-2025	2
AgglomerativeClustering	0	tabular-playground-series-jul-2023	1
MiniBatchKMeans	0.2	tabular-playground-series-jul-2028	1
OPTICS	0	tabular-playground-series-jul-2029	1
Birch	0	tabular-playground-series-jul-2024	0
SpectralClustering	0	tabular-playground-series-jul-2026	0

Tabelle IV
ERGEBNISSE FÜR DIE TABULAR-PLAYGROUND-SERIES.

C. Gesamteinschätzung

Die Ergebnisse zeigen, dass die Wahl des optimalen Clustering-Algorithmus stark von den Eigenschaften des jeweiligen Wettbewerbs und der zugrunde liegenden Datensätze abhängt. Zwar dominiert **KMeans** in der aggregierten Betrachtung mit 64 Einsätzen und einem durchschnittlichen Score von etwa 0.52, in einzelnen Wettbewerbsreihen – wie beispielsweise in der dmassign-Reihe (Score 0.4) oder der tabular-playground-Serie (Score

0.3) – erzielt dieser Ansatz jedoch lediglich moderate Ergebnisse. Im Wettbewerb *bigdata-and-datamining-2nd-ex* konnten hingegen Algorithmen wie **AgglomerativeClustering** und **Birch** sehr hohe Scores (0.95 bzw. 0.96) erreichen, was auf eine besonders gute Anpassung an die dortigen Datenstrukturen hinweist. In der tabular-playground-Serie zeigten sich hingegen **BayesianGaussianMixture** und **GaussianMixture** als überlegen, während dicht-basierte Methoden wie **DBSCAN** und **OPTICS** in allen Wettbewerben überwiegend einen Score von 0 erzielten.

Diese differenzierten Ergebnisse verdeutlichen, dass keine universell überlegene Clustering-Methode existiert. Vielmehr muss die Wahl des Algorithmus stets in Abhängigkeit von den spezifischen Anforderungen und Eigenschaften des Datensatzes erfolgen. Die vorliegenden Resultate bieten somit eine solide Grundlage, um in zukünftigen Arbeiten die Eignung der einzelnen Verfahren weiter zu untersuchen.

XI. Clustering Algos imi Vergleich

XII. Fazit

Insgesamt lässt sich festhalten, dass Clustering-Algorithmen auf Kaggle nur in wenigen dedizierten Wettbewerben eine zentrale Rolle spielen. Die ermittelten Beispiele zeigen jedoch, dass verschiedene Verfahren – insbesondere probabilistische Modellierungen (Bayesian- und Gaussian Mixture) sowie klassische Methoden wie KMeans – bei passender Datenaufbereitung und Parameterauswahl durchaus wettbewerbsfähige Ergebnisse erzielen können. Dabei ist kein einzelner Algorithmus universell überlegen; vielmehr bestimmen die Beschaffenheit der Daten, die Bewertungsmetrik und das Ziel des Wettbewerbs, welche Methode besonders erfolgreich ist.

Gleichzeitig verdeutlicht die geringe Zahl an reinen Clustering-Challenges auf Kaggle, dass das unüberwachte Lernen im Kontext von Data-Science-Wettbewerben derzeit eine eher untergeordnete Rolle einnimmt. Zwar ermöglicht die Plattform einen leichten Zugang zu vielfältigen Datensätzen und Community-Beiträgen, doch stehen in der Praxis meist klassifikations- und regressionsorientierte Probleme im Vordergrund.

Für die Industrie bieten die gewonnenen Erkenntnisse dennoch wertvolle Anhaltspunkte: Algorithmen wie KMeans oder Birch sind bei klar abgegrenzten Klassen oft hinreichend performant und zeichnen sich durch relative Einfachheit und Effizienz aus. Modelle mit probabilistischer Natur (z. B. Gaussian Mixture) können hingegen in komplexeren Situationen Vorteile bringen, erfordern aber oftmals mehr Rechenaufwand und Parameterabstimmung.

Zusammenfassend zeigt die vorliegende Analyse, dass die Wahl des „richtigen“ Clustering-Algorithmus stets vom Datenumfeld, der Bewertungsmetrik und den verfügbaren Ressourcen abhängt. Zukünftige Arbeiten könnten die Untersuchung ausweiten, indem sie weitere Wettbewerbsplattformen oder umfangreichere Notebooks analysieren, um ein noch differenzierteres Bild der Anwendungs- und Erfolgsfaktoren von Clustering-Algorithmen zu gewinnen.

XIII. Ausblick

Für zukünftige Arbeiten ergeben sich mehrere Erweiterungsmöglichkeiten. Eine naheliegende Fortsetzung dieser Analyse wäre die Ausweitung der Untersuchung auf weitere Plattformen, wie beispielsweise GitHub oder andere wissenschaftliche Repositorien, um eine umfassendere Übersicht über die Anwendung von Clustering-Algorithmen zu erhalten.

Darüber hinaus könnte anstelle einer Fokussierung auf Wettbewerbsbeiträge eine systematische Analyse aller öffentlich verfügbaren Notebooks erfolgen, die sich mit Clustering befassen. Hierbei wäre insbesondere von Interesse, welche Datensätze verwendet werden und welche Faktoren zur Performanzsteigerung bestimmter Clustering-Algorithmen beitragen.

Eine weitere Erweiterung bestünde darin, alle derzeit aktiven Wettbewerbe auf Plattformen wie Kaggle zu analysieren, um systematisch zu erfassen, welche Clustering-Methoden in den aktuellen Top-Notebooks zum Einsatz kommen. Dies könnte Aufschluss über die Präferenzen der Community sowie über Trends und Entwicklungen im Bereich des unüberwachten Lernens geben.

Diese Erweiterungen könnten dazu beitragen, ein umfassenderes Bild über den Einsatz und die Effektivität von Clustering-Algorithmen zu zeichnen und zukünftige Forschungen in diesem Bereich gezielt zu unterstützen.

Literatur

[1] A. K. Jain, M. N. Murty und P. J. Flynn, “Data clustering: A review”, *ACM Computing Surveys*, 1999.

[2] R. Xu und D. Wunsch, “Survey of Clustering Algorithms”, *IEEE Transactions on Neural Networks*, Jg. 16, Nr. 3, S. 645–678, 2005.

[3] J. Bois und A. Rokem, “Kaggle as a Platform for Machine Learning Education”, *Journal of Data Science Education*, Jg. 1, Nr. 1, 2021.

[4] T. Hillenbrand, “Clustering Methods in Industry: A Survey”, *Journal of Machine Learning Applications*, Jg. 18, S. 45–59, 2021.

[5] L. Kaufman und P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990, ISBN: 978-0471878766.

[6] P. J. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”, *Journal of Computational and Applied Mathematics*, Jg. 20, S. 53–65, 1987.

[7] Çağlar Uslu. (2022). What is Kaggle? Zugriff am 5. Februar 2025, Adresse: https://www.datacamp.com/blog/what-is-kaggle.

[8] ——, (2022). Kaggle Competitions: The Complete Guide. Zugriff am 5. Februar 2025, Adresse: https://www.datacamp.com/blog/kaggle-competitions-the-complete-guide.

[9] R. V. Krejcie und D. W. Morgan, “Determining sample size for research activities”, *Educational and Psychological Measurement*, Jg. 30, Nr. 3, S. 607–610, 1970.

[10] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[11] D. J. Biau, S. Kernéïs und R. Porcher, “Statistics in brief: the importance of sample size in the planning and interpretation of medical research”, *Clinical Orthopaedics and Related Research*, Jg. 466, Nr. 9, S. 2282–2288, 2008.

[12] W. Reade und A. Chow, *Tabular Playground Series – Juli 2022*, Kaggle Competition, https://kaggle.com/competitions/tabular-playground-series-jul-2022, 2022.

[13] JLU_LiuYun, *BigData & DataMining – 2nd Ex*, Kaggle Competition, https://kaggle.com/competitions/bigdata-and-datamining-2nd-ex, 2022.

[14] N. Shah, *DM-Assignment 1*, Kaggle Competition, https://kaggle.com/competitions/dmassign1, 2020.