Beschreibung Domänen Projekt 2: Zeitreihen Klassifikation/ Anomalie Erkennung

Vorabversion - kleinere Änderungen und Ergänzungen später sind möglich

1. Thema und Hintergrund

Das Thema des Domänenprojektes 2 wird in diesem Semester von einer Landesforstverwaltung gestellt. Die Forstverwaltung ist verantwortlich für die nachhaltige Bewirtschaftung und Pflege des Staatswaldes. Eine zentrale Aufgabe dabei ist die Kenntnis über die Baumartenverteilung, da diese die Grundlage für Entscheidungen im Waldmanagement, in der Biodiversitätserhaltung und im Klimaschutz bildet. Traditionell wird diese Information über eine groß angelegte Forstinventur gewonnen, die alle zehn Jahre stichprobenweise in einem Raster über ganz Deutschland durchgeführt wird, die sog. Bundeswaldinventur (BWI). Diese liefert zwar wertvolle Daten, erfassen jedoch nicht die gesamte Waldfläche flächendeckend.

Um eine vollständige Baumartenklassifikation für das gesamte Staatsgebiet zu erhalten, soll daher auf Satellitendaten zurückgegriffen werden. Sentinel-2 liefert hierfür multispektrale Zeitreihen mit hoher räumlicher Abdeckung. Die Herausforderung besteht darin, die Datenmenge so gering wie möglich zu halten, da eine Verarbeitung über lange Zeitserien für jede einzelne Fläche rechenintensiv wäre. Ziel ist es, robuste Methoden zu entwickeln, die auch mit „wenigen“ Zeitschritten pro Pixel eine verlässliche Klassifikation erlauben.

2. Daten

Die in diesem Projekt verwendeten Daten stammen von den Sentinel-2-Satelliten und bestehen aus bodennahen Reflexionswerten (BOA) der Erdoberfläche. Sentinel-2 erfasst zwölf Spektralbänder (B1–B12) in unterschiedlichen Wellenlängenbereichen, von denen zehn für die Analyse genutzt werden.

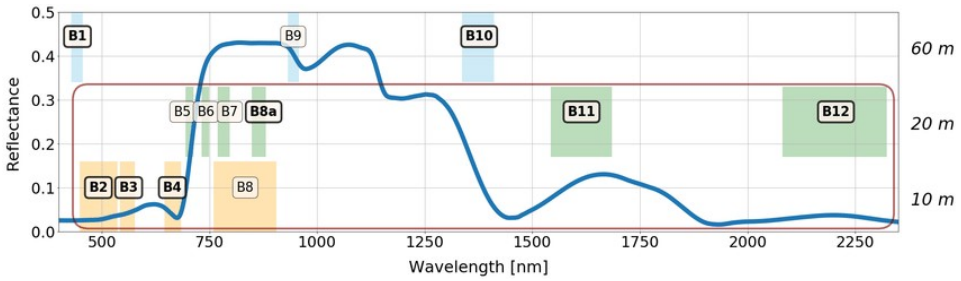


Abbildung 1 nach [1], zeigt die spektrale Abdeckung der Sentinel-2A-Bänder im Vergleich zu einem typischen Vegetationsreflexionsspektrum. Die blaue Linie stellt das Reflexionsmuster gesunder Vegetation dar, während die rote Umrandung jene Bänder markiert, die in Vegetationsanalysen besonders häufig verwendet werden. Rechts ist die jeweilige räumliche Auflösung der Bänder angegeben.

Vegetationstypen unterscheiden sich deutlich in ihren Reflexionswerten. Besonders die Bereiche außerhalb des sichtbaren Spektrums, wie die Red-Edge- und SWIR-Regionen (B5–B7, B8A, B11, B12), sind entscheidend, da sie Rückschlüsse auf Wasser- und Chlorophyllgehalt zulassen.

Die Satelliten liefern großflächige Aufnahmen mit einer räumlichen Auflösung von 10 m sowie einer zeitlichen Wiederholrate von etwa 5 Tagen. Durch die Überlagerung der Aufnahmen entstehen Zeitreihen pro Pixel, welche Satellite Image Time Series (SITS) genannt werden. Siehe Abbildung 2.

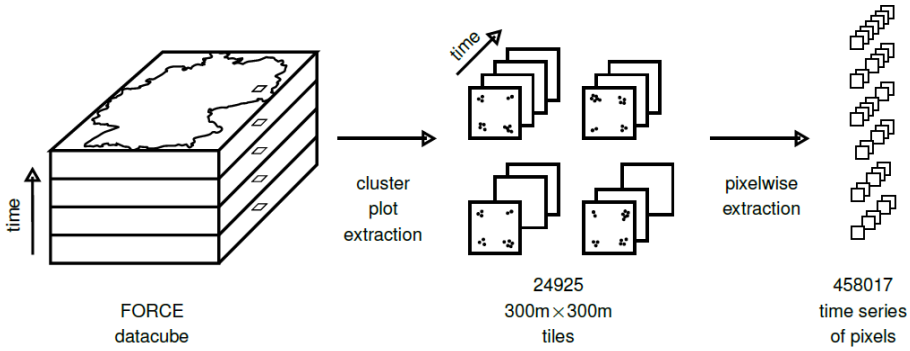
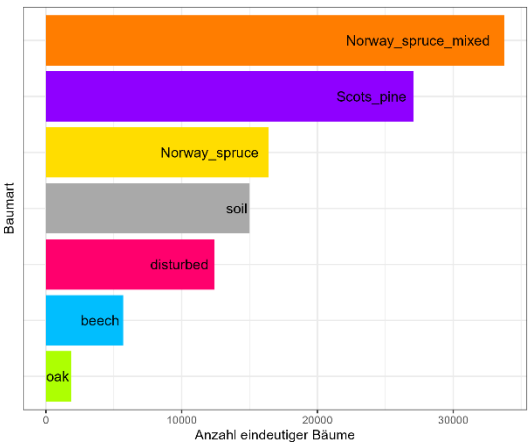


Abbildung 2: SITS-Erstellung auf Basis der BWI daten

Die Label stammen primär aus der Bundeswaldinventur (BWI) 2012: Für jede dort erfasste Baumposition wurde eine zugehörige Pixel-Zeitreihe extrahiert (tree-centric Verknüpfung). Informationen aus der BWI 2022 dienen ergänzend zur Plausibilisierung (z. B. Status „Baum 2022 noch vorhanden“). Da die BWI auf manueller Kartierung basiert, können Fehler auftreten – etwa durch GNSS-Ungenauigkeiten, menschliche Fehleinschätzungen oder Pixel, die mehrere Baumarten abbilden.

Die Verteilung der Baumarten im Datensatz ist deutlich unausgeglichen: Während Nadelbaumarten wie Fichte und Kiefer stark vertreten sind, kommen Laubbaumarten wie Eiche und Buche vergleichsweise selten vor. Dieses Ungleichgewicht spiegelt die realen Waldstrukturen wider, stellt jedoch auch eine Herausforderung für die Modellierung dar, da Modelle dazu neigen können, häufigere Klassen zu bevorzugen



In einer idealen Situation wären die Zeitreihen kontinuierlich, frei von Rauschen und klar zwischen den Baumarten unterscheidbar. Wie beispielsweise in Abbildung 3, welche die Median Werte des ganzen Datensatzes abbildet. In der Praxis führen jedoch Wolken, Schatten oder Schneebedeckung zu teils großen Lücken. Zusätzlich erschweren jahreszeitliche Veränderungen (z. B. Laubfall oder

Schneeperioden) die Analyse. So sehen die einzelnen Zeitreihen sehr lückenhaft, wie in Abbildung 4 aus.

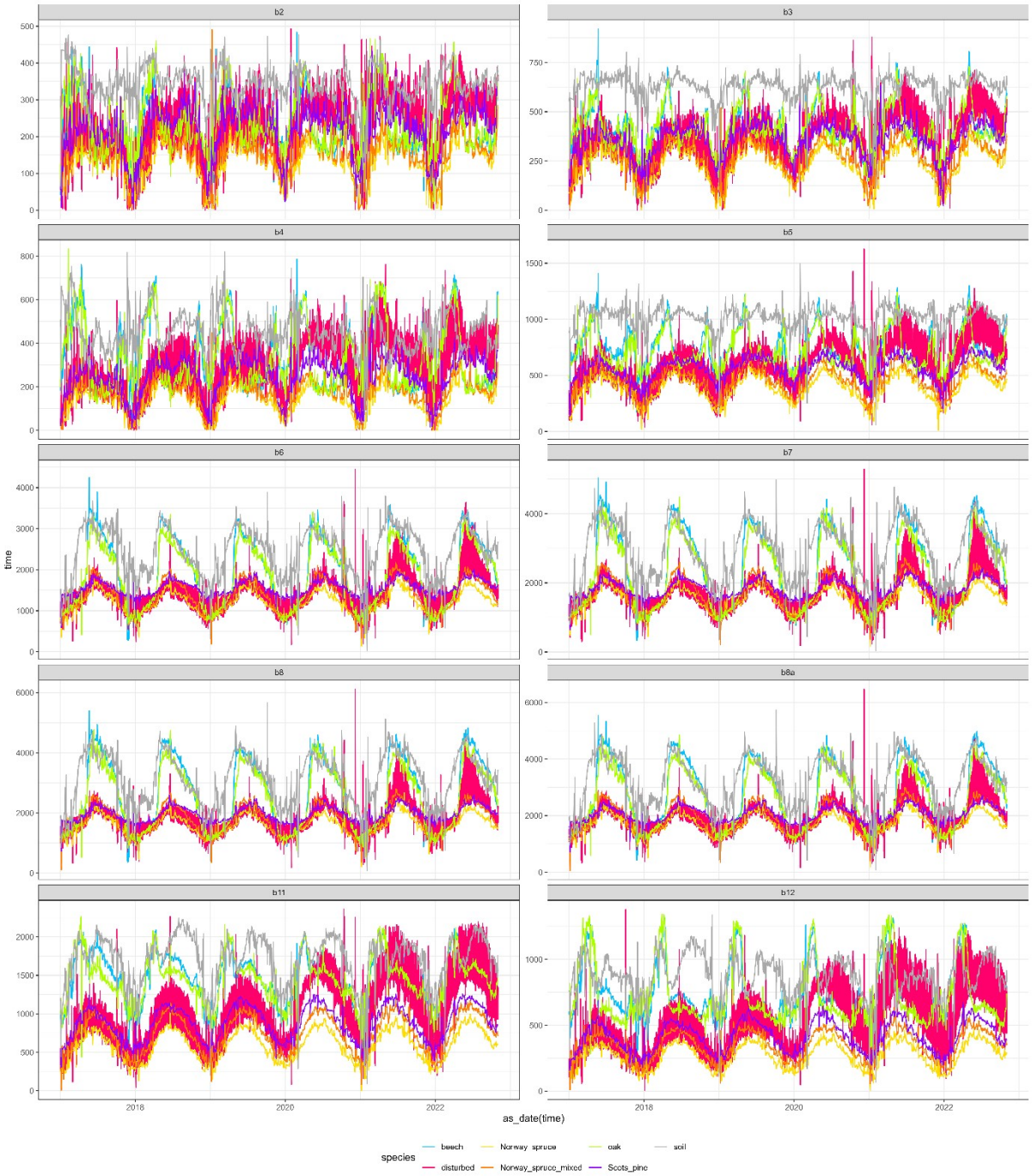


Abbildung 3: Medianwerte des bereinigten Trainingsdatensatzes, dargestellt pro Klasse und Spektralband.

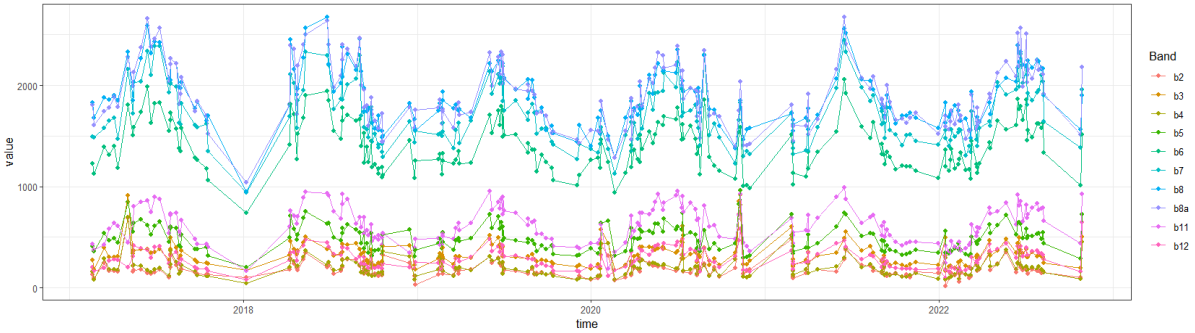
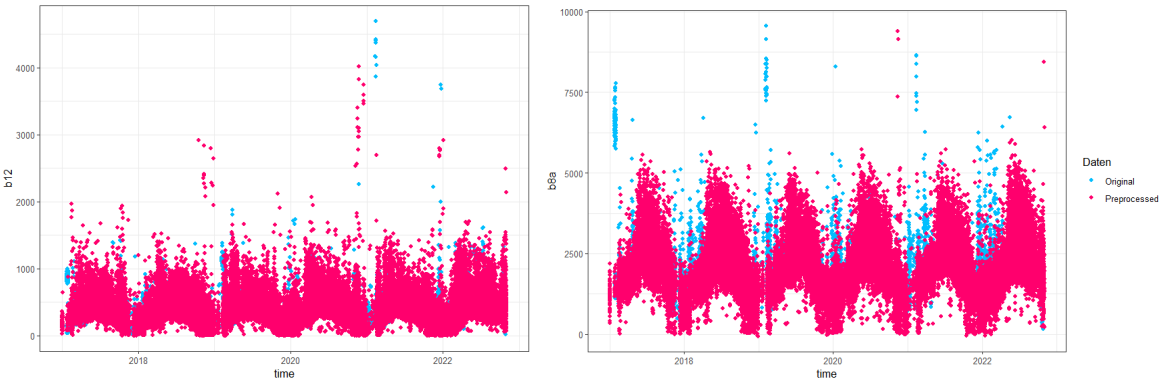


Abbildung 4: Multivariate Zeitreihe eines Pixels

Vorarbeit

In Wäldern mit geschlossenem Kronendach wird die BOA-Reflexion jedoch von der obersten Schicht des Kronendachs dominiert, und Bäume, die von größeren Individuen überschattet werden tragen wenig zur Gesamtreflexion innerhalb eines Pixels bei.

- Um den bereitgestellten Trainingsdatensatz zusammenzustellen, haben wir die Daten auf Bäume gefiltert, die mit hoher Wahrscheinlichkeit aus den Satellitenbildern sichtbar sind (anhand der Baumhöhe).
- Zusätzlich wurden fehlerhafte Beobachtungen (z. B. durch Wolken, Wolkenschatten oder Schnee) mit Hilfe des FORCE-Frameworks [1], das eine automatisierte Qualitätsfilterung und Atmosphärenkorrektur für Sentinel-2-Daten durchführt, entfernt.



- **!! Wichtig:** Ihr Datensatz enthält eine Variable: „disturbance_year“. Nutzen Sie diese nicht für die Klassifikation. Diese Variable kommt nur in dem Trainingsdatensatz vor

3. Aufgabenstellung:

Ziel des Projekts ist die Entwicklung eines Modells zur Klassifikation multivariater Sentinel-2-Zeitreihen, das Baumarten mit hoher Genauigkeit vorhersagt.

- Darüber hinaus untersuchen die Teams, wie sich der für die Modellentscheidung erforderliche Datenumfang (konkret die Anzahl der benötigten Zeitstempel pro Pixel) ohne spürbaren Qualitätsverlust reduzieren lässt, um eine großflächige Anwendung zu ermöglichen. (Tabellarischer vergleich mehrerer Modelle, ROC, …). Für die Klasse „Disturbed“, wählen Sie am besten abschnitte der Zeitreihe um das „disturbance_year“.

- Als optionalen Bonus erstellen die Teams eine wissenschaftlich fundierte Feature-Importance-Analyse (z. B. durch Integrated Gradients oder mittels SHAP) und leiten daraus nachvollziehbare fachliche Einsichten ab.

Zur Bewältigung der Aufgabe beginnen die Teams mit einem klaren Problemverständnis und einem gründlichen Data Understanding: Dazu zählen die Besonderheiten von Satellite Image Time Series (SITS) sowie eine explorative Analyse der Zeitreihen im Hinblick auf Rauschen, Ausreißer und potenziell unsichere Labels. Die Nutzung einschlägiger Literatur und frei zugänglicher Online-Quellen zu SITS ist ausdrücklich erwünscht; alle Annahmen, Entscheidungen und Zwischenergebnisse sind klar und reproduzierbar zu dokumentieren.

Data Preparation:

Die Datenvorverarbeitung ist ein erforderlicher Schritt. Folgende Methoden könnten von Hilfe sein:

- **Zeitliche Auflösung**: Aggregation der Daten, z. B. zu wöchentlichen, zweiwöchentlichen oder monatlichen Zeitintervallen.
- **Bereinigung / Imputation / Glättung:**
  - Trotz des Einsatzes des FORCE-Algorithmus sind weiterhin viele Ausreißer in den Daten vorhanden.
  - Im Gegensatz zu den meisten Zeitreihenanwendungen sind die Zeitreihen in diesem Datensatz stark lückenhaft. Überlegen Sie daher, wie Sie mit diesen Lücken umgehen wollen.
  - In SITS werden häufig folgende Glättungsverfahren eingesetzt: Whittaker-Algorithmus, Kalman-Smoothing, Savitzky–Golay-Glättung.
- **Feature Engineering**: In der Vegetationsanalyse auf Basis von SITS werden häufig sogenannte Vegetationsindizes (VIs) genutzt, um phänologische Eigenschaften von Pflanzen hervorzuheben. Dabei werden aus mehreren spektralen Bändern neue Features berechnet. Einige davon sind in [3] aufgeführt.

Vegetation index	Index acronym	Formula	Reference
Normalized Difference Vegetation Index	NDVI	$\frac{B_8 - B_4}{B_8 + B_4}$	Tucker (1979)
Green Normalized Difference Vegetation Index	GNDVI	$\frac{B_7 - B_3}{B_7 + B_3}$	Gitelson and Merzlyak (1998)
Weighted Difference Vegetation Index	WDVI	$B_8 - 0.5 \times B_4$	Clevers (1989)
Transformed Normalized Difference Vegetation Index	TNDVI	$\sqrt{\frac{B_8 - B_4}{B_8 + B_4} + 0.5}$	Yi (2019)
Soil Adjusted Vegetation Index	SAVI	$\left(\frac{B_8 - B_4}{B_8 + B_4 + 0.5}\right) \times 1.5$	Huete (1988)
Infrared Percentage Vegetation Index	IPVI	$\frac{B_8}{B_8 + B_4}$	Crippen (1990)
Modified Chlorophyll Absorption Ratio Index	MCARI	$((B_5 - B_4) - 0.2(B_5 - B_3)) \times \frac{B_5}{B_4}$	Daughtry et al. (2000)
Red Edge In-flection Point	REIP	$\frac{700 + 40\left(\left(\frac{B_4 + B_7}{2}\right) - B_5\right)}{B_6 - B_5}$	Guyot et al. (1988)
Modified Soil Adjusted Vegetation Index 2	MSAVI2	$\frac{2B_8 - 1 - \sqrt{(2B_8 + 1)^2 - 8}}{2}$	Qi et al. (1994)
Difference Vegetation Index	DVI	$B_8 - B_4$	Jordan (1969)

[3]

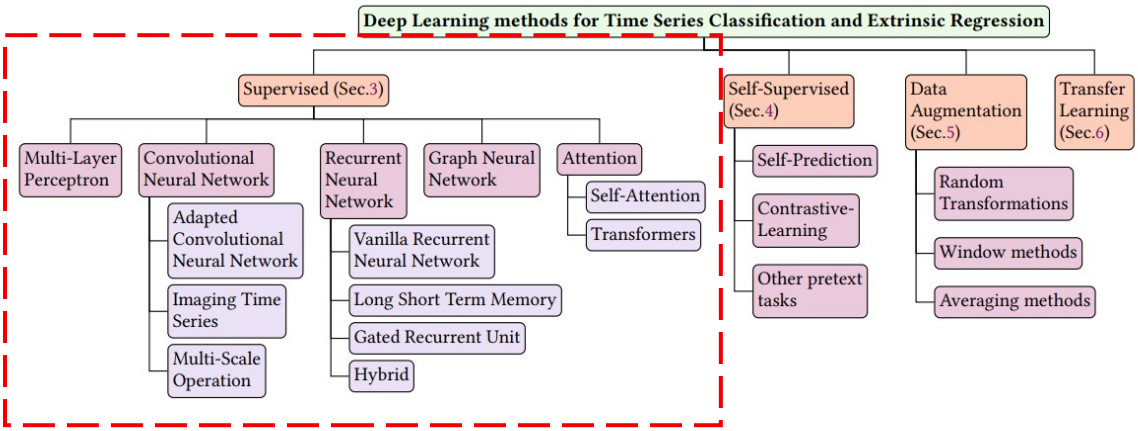
Modelling:

Es handelt sich um eine Multiklassifikationsaufgabe auf Basis von Zeitreihendaten. Zur Lösung können überwachte Klassifikationsverfahren eingesetzt werden. Folgende wichtige Aspekte und Techniken bei der Zeitreihenklassifizierung sind zu berücksichtigt werden:

1. **Feature-Extraktion:** Das Extrahieren relevanter Merkmale aus den Zeitreihendaten ist ein wichtiger Schritt. Dazu gehören statistische Kennzahlen, Merkmale die zeitlichen Aspekte bzw. Veränderungen über die Zeit abbilden oder/und andere (domänenspezifische) Merkmale, die die Charakteristiken der Zeitreihen hervorheben.
2. **Modelle die direkt auf den Zeitreihendaten funktionieren:** Es gibt verschiedene Algorithmen und Modelle, die für die Klassifizierung von Zeitreihen verwendet werden können (für die meisten dieser Ansätze braucht man nicht das klassische Feature Engineering). Einige Beispiele sind:
    a. **Deep Learning basierte Ansätze:** insbesondere Rekurrente neuronale Netzwerke (RNNs) wie LSTM (Long Short-Term Memory) oder GRU (Gated Recurrent Unit) können verwendet werden, um zeitliche Abhängigkeiten in den Daten zu erfassen.
    b. **Distanzbasierte Methoden:** Hierbei werden Distanzen zwischen Zeitreihen als Merkmale verwendet.
    c. **Shapletbasierte Methoden**: Extrahiert Teilsequenzen (Shaplets) aus den Zeitreihen, die wichtige Muster für die Klassifizierung enthalten und berechnet dann die Ähnlichkeit zur Zeitreihe und verwenden diese Ähnlichkeiten als Feature in einem beliebigen Klassifikationsalgorithmus
    d. **Ensemble-Methoden:** Kombinationen verschiedener Modelle können die Leistung verbessern, indem sie unterschiedliche Aspekte der Daten erfassen.

Es ist zu beachten, dass es ungleiche anzahlen an Zeitreihen pro Klasse gibt.

- Links zu Literatur und Algorithmen:
    - Time Series Classification: A Review of Algorithms and Implementations | IntechOpen
    - Package for Time series classification
    - Rocket
    - Mapping temperate forest tree species using dense Sentinel-2 time series: https://doi.org/10.1016/j.rse.2021.112743



Train-Test-Design

Wir stellen Ihnen einen gelabelten Datensatz zur Verfügung. Sie entscheiden selbst den Train-Test Split. Gegen Ende des Projekts werden wir ihnen eine weiteren Datensatz zur Validierung zukommen lassen, ohne Labels, mit den folgenden variablen Parametern:
"time", "id", "doy", "b2", "b3", "b4", "b5", "b6", "b7", "b8", "b8a", "b11", "b12"

Sie solle die Labels vorhersagen. Wir werden dann die Genauigkeit ihrer Klassifikation auf dem Testdaten messen (siehe Evaluation).

Evaluation

- Die Genauigkeit Ihrer Klassifikation wird anhand des folgenden Maß berechnet:
  - **Balanced Accuracy:**
    - $\mathrm{BAcc} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}.$
  - wobei ($k \in \{1, \ldots, K\}$) den Klassenindex bezeichnet und (K) die Gesamtzahl der Klassen ist.

Zusätzlich zu behandelnde Fragestellungen:

1. Bitte prüfen Sie, ob es möglich ist den Data intake in das Model zu reduzieren. Bedenken sie, wir möchten in Zukunft Ihre Methodik verwenden um Bundeslandweit, oder darüber hinaus Baumarten zu klassifizieren, daher wollen wir die Datenmenge pro Pixel möglichst reduzieren. Analysieren Sie dafür den Tradeoff zwischen Daten sparen und Balanced Accuracy.
2. Bonusfrage: Welche Features oder Zeiträume sind besonders wertvoll für das von Ihnen verwendete Modell?

Ein geringer *Data Intake* wird belohnt. Die Anzahl der benötigten Datenpunkte pro Zeitreihe für eine ausreichende Klassifikation, **vor** dem Preprocessing (z.B. Gap-Filling ist nicht intake-relevant), Führt in die Bewertung mit ein.
Maß ist die Anzahl verwendeter *Seasons*, wobei eine Season = 52 Wochen entspricht. Werden Zeiträume aus verschiedenen Jahren kombiniert, werden diese aufsummiert (z.\,B.\ drei Ausschnitte mit 26 Wochen = 1,5 Seasons).

➔ (Die genaue Gewichtung ist noch nicht absolut)

Bonus: Wenn Sie eine aussagekräftige Feature-Importance Ihres Modells bereitstellen.

4. Organisation und Zeitmanagement

- **Direkt**: Teamnamen & Teammitglieder melden; Zugänge klären (Git-Repo, Datenablage, etc.).
- 29.09.2025 09:00 - 13:00 Uhr: Kickoff Workshop in Präsenz (Anwesenheitspflicht): Aufgabenstellung, Bewertungsrahmen, Datenübersicht, Q&A; Start in den Team-Workflows.
- **Oktober**: Wöchentliche Status-Meetings voraussichtlich jeweils Donnerstag Nachmittags (alternierend online und in Präsenz): Kurz-Status je Team (Risiken, nächste Schritte, Blocker);
- **09.10.25 Meilenstein 1** (Domain & Baseline):
    - Kurzreport zum Problemverständnis (Domäne & Business-Ziel), Data Understanding
    - Erste data Pre-processing Pipeline
    - erster Baseline-Klassifikator (mit 70/30 Clusterplot-Split). Code + kurzer Demo-Run.
- **16.10.2025 Meilenstein 2** (Modellreife & Validierung):
    - Optimierte Pre-Processing Pipeline
    - Verbessertes Modell mit sauberer Validierung
    - Vergleich gg. Baseline, klare Fehleranalyse (Confusion Matrix, pro Klasse).
- **23.10.2025 Meilenstein 3** (Dateninput minimieren):
    - Studie zur Datenreduktion (z. B. 1–2-jährige Sequenzen, Sommerfenster, wenige Zeitstempel) inkl. Trade-off-Analyse „Qualität vs. Input“.
    - Bonus: Feature-Importance/Erklärbarkeit (z. B. Permutation, SHAP): wichtigste Bänder/Zeiträume/Lag-Effekte, fachlich begründet.
- **31.10.2025 Ende der Projektarbeit – Abgabe:**
    - **Code** (reproduzierbar; README, env/requirements)
    - **Abschlussbericht** (knapp & klar), Modellkarte (Annahmen, Grenzen), Interpretation der Ergebnisse.
    - **csv-Datei** mit Ihren Vorhersagen auf dem Validierungsdatensatz
    - **Dokumentation der täglichen Arbeitszeit**
    - **Zusammenfassung Ihrer Beiträge zur Teamleistung (1 Seite)**
- [im November] Abschlusspräsentation:
    - 10–15 min Pitch + Q&A; Fokus auf Zielerreichung, Evidenz, Repro-Pfad.
- [im November] Einzelgespräche:
    - Reflexion, Beitrag, Learnings.

Bitte **dokumentieren Sie tagesgenau Ihre tägliche Arbeitszeit** mittels der in ILIAS bereitgestellten Vorlage unter Angabe des Arbeitsgegenstandes - wenn Sie schon im September starten bereits im September

5. Bewertungskriterien

Wir werden folgende Kriterien für die Bewertung heranziehen:

1. Primär bewerten wir die Modellleistung anhand der Balanced Accuracy auf den Validierungsdaten
    1. Sie erhalten einen Datensatz ohne Labels gegen Ende des Projekts, die Ergebnisse der Klassifikation Ihres Models werden bewertet.
    2. Die analyse des Tradeoffs zwischen Daten sparen und Model-Accuracy.
    3. Bonus für eine Feature Importance Analyse.
2. Fachliche Kompetenz und Qualität
    1. Durchführung des Gesamtprozesses

    2. Kenntnis und richtige Anwendung der Algorithmen
    3. Preprocessing und Tuning der Modelle
3. Technische Kompetenz und Qualität
    1. Struktur, Dokumentation und Verständlichkeit des Codes
    2. Robustheit und Korrektheit der Implementierung
4. Dokumentation
    1. Abschlussbericht: Dokumentation ihrer Entscheidungen – Pre-processing, Algorithmen, Bewertung ihrer Modelle
    2. Vollständigkeit und Verständlichkeit
    3. Abschlusspräsentation
5. Kommunikation
    1. In den Statusmeetings
    2. Sonstige Kommunikation mit den Dozenten
    3. Vorstellung des Meilensteins & finale Präsentation
6. Organisation & Zeitmanagement
    1. Fortschritt zu den Zwischenterminen
    2. Aufgabenteilung und Zusammenarbeit im Team

[1] Estimating Crop Primary Productivity with Sentinel-2 and Landsat 8 using Machine Learning Methods Trained with Radiative Transfer Simulations - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Band-settings-of-Sentinel-2A-with-respect-to-a-typical-vegetation-reflectance-spectrum_fig2_347624648 [accessed 2 Sept 2025]

[2] Frantz, D. (2019). FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Remote Sensing*, *11*(9), 1124. https://doi.org/10.3390/rs11091124

[3] Ma, G., Ding, J., Han, L., Zhang, Z., & Ran, S. (2021). Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. *Regional Sustainability*, *2*(2), 177–188. https://doi.org/10.1016/j.regsus.2021.06.001