Quality Guideline

Ausbildung

Data-Mining Aufbaukurs

European
Six Sigma Club Deutschland e.V.
www.sixsigmaclub.de

Version: 1.0

Stand: 23.09.2021

Inhalt

Vorwort

Hintergründe

Die Digitalisierung ist in vollem Gange und beschert uns eine ungeahnte Verfügbarkeit von Daten. Es stehen mehr Daten, sowohl strukturierte als auch unstrukturierte, aus einer steigenden Anzahl von Quellen in unterschiedlicher Qualität zur Verfügung.

Dies verändert unter anderem die Vorgehensweisen der datengestützten Problemlösung und Prozessverbesserung und somit auch Six Sigma. Bisher stand die Frage nach der projektbezogenen Erhebung der richtigen Daten zur Lösung eines Problems im Vordergrund. Bereits heute geht es jedoch zunehmend darum, aus großen Mengen vorhandener Daten effizient die relevanten Daten auszuwählen und das bestmögliche Ergebnis zu erzielen.

Will man dieser Entwicklung Rechnung tragen, ist es erforderlich, den Six-Sigma-Werkzeugkasten in Richtung der Datenwissenschaften, insbesondere Big Data, Data-Mining und Machine Learning, zu erweitern und die bislang bewährte Six-Sigma-Ausbildung zu ergänzen.

Die im Folgenden beschriebenen Ausbildungsinhalte beschreiben die Qualifikationen, die in den Bereichen der Qualitäts- und Prozessverbesserung zunehmend Anwendung finden.

Entstehung

Die vorliegende Richtlinie wurde im Arbeitskreis „Six Sigma Weitergedacht“ entwickelt. Ausgehend von den Erfahrungen aus zahlreichen Projekten und Schulungen hat der Arbeitskreis die Auswirkungen von „Digitalisierung“ und „Big Data“ auf die bewährte Prozessverbesserungsmethode Six Sigma diskutiert.

Basis hierfür waren verschiedene Szenarien der Datenverfügbarkeit (von „keine Daten verfügbar“ bis „alle Daten liegen kontextbezogen vor“). Schon zu einem frühen Zeitpunkt hat der Arbeitskreis den DMAIC-Zyklus mit anerkannten Vorgehensmodellen zur Datenanalyse verknüpft. Ziel war es, in Abhängigkeit von Phase und Szenario die Ist- und Zielsituation zu definieren und damit die notwendigen Änderungen und Ergänzungen des Six Sigma Werkzeugkastens zu identifizieren.

Die Ergebnisse wurden auf den Fachkonferenzen des Six Sigma Clubs vorgestellt und diskutiert. Darüber hinaus sind zahlreiche themenrelevante Veröffentlichungen und Vorträge entstanden. Final wurden Arbeitsergebnisse bzw. die als neu identifizierten Schulungsinhalte den Markennehmer*innen und den Mitgliedern in getrennten Umfragen zur Bewertung vorgelegt. Diese Ergebnisse und die Ergebnisse aus aktuellen Forschungsprojekten der Hochschule Koblenz und der FAU Erlangen wurden gemeinsam als Grundlage für diese Ausbildungsrichtlinie herangezogen.

Die nachstehende Grafik veranschaulicht den Schulungsbedarf in den identifizierten Kompetenzfeldern. Theoretische Grundlagen und praktische Anwendungen sollen diese Lücken schließen. Entstanden ist das Netzdiagramm auf Basis der bereits erwähnten Quellen. Kleine Zahlen bedeuten dabei keine/wenig Kompetenz bzw. geringe Notwendigkeit für Zusatzausbildung. Hohe Zahlen eine bereits hohe Ausbildungstiefe oder hohe Anforderungen. In anderen Worten: 0 bedeutet keine/wenig Kompetenz und 3 bedeutet hohe Kompetenz.

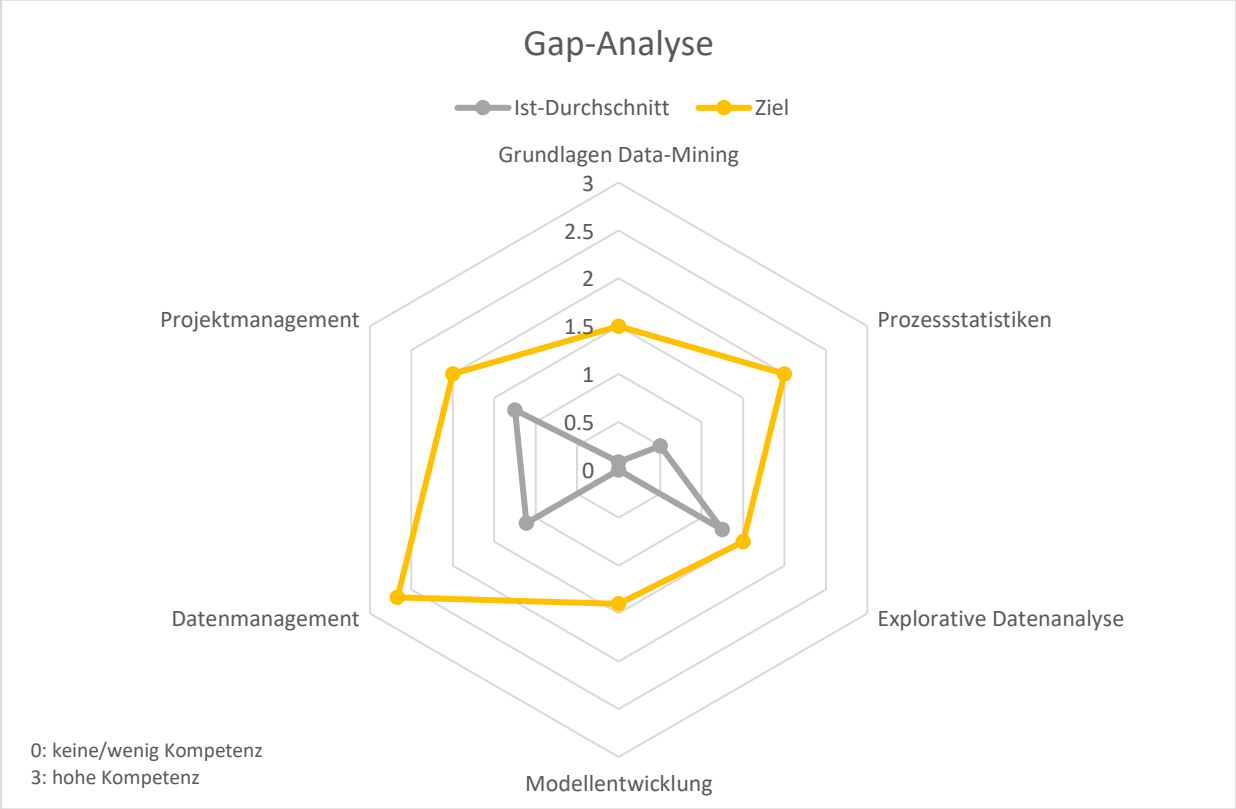

Abbildung 1: Gap-Analyse Methodenkompetenzen

Ziel und Zielgruppe

Ziel des Aufbaukurses Data-Mining ist es, Methodenexpert*innen Handwerkzeuge aufzuzeigen, mit denen auch komplexere Datenstrukturen auswertbar werden. Teilnehmer*innen kennen am Ende dieses Kurses die Anforderungen an analysierbare Datenstrukturen und können diese, auch über Datenquellen hinweg, herstellen. Sie kennen wesentliche Begriffe aus dem Bereich des Data-Mining (DM) und deren Bedeutung. Sie können gängige Data-Mining-Methoden und -Algorithmen anwenden, um Zusammenhänge in Daten zu identifizieren und entwickelte Modelle zu verifizieren. Dabei werden sowohl Regressionsmodelle als auch Klassifikationsmodelle gemeinsam entwickelt, einem Benchmark unterzogen und beurteilt. Auch mit Hilfe von Ansätzen des Machine Learning (ML) können sie Zusammenhangs- und Prognosemodelle entwickeln, in Applikationen nutzen und überwachen.

Dieser Kurs ist speziell für Methodenexpert*innen (nicht nur Six Sigma Belts) konzeptioniert, die während ihrer Arbeit oder in Projekten immer wieder mit Fragestellungen zu Zusammenhängen in Daten konfrontiert werden. Teilnehmer*innen dieses Kurses kennen sich in ihrem Fachgebiet aus und führen bereits regelmäßig statistische Datenanalysen durch.

CRISP-DM und Ähnlichkeiten zu Six Sigma DMAIC

Mit der veränderten Verfügbarkeit von Daten rücken oft auch neue Vorgehensmodelle in den Fokus. Eines der prominentesten Vorgehensmodelle im Umfeld von Big Data und I4.0 stellt CRISP-DM dar. Tatsächlich existieren unzählige vergleichbare Vorgehensmodelle mit unterschiedlichen Namen und nur minimalen Anpassungen. Der Schwerpunkt von CRISP-DM liegt auf den verschiedenen Tätigkeiten bzw. den Zielen der jeweiligen Phasen. Bezüglich der verwendbaren Werkzeuge gibt CRISP-DM hingegen keine oder wenig Hilfestellung.

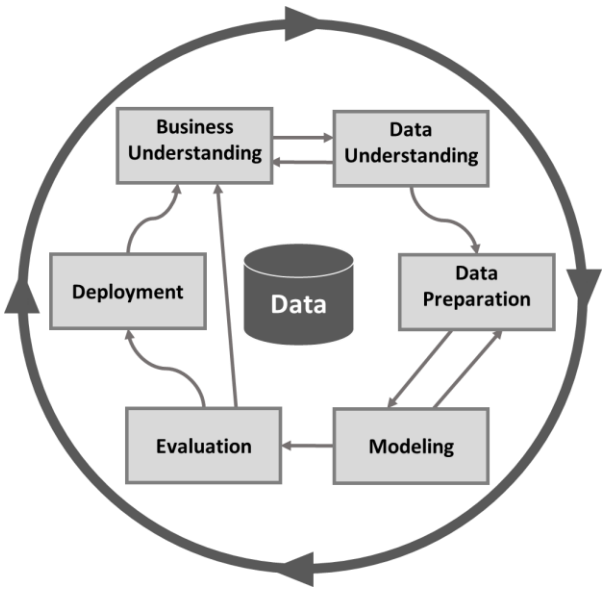

Abbildung 2: Prozessmodell CRISP-DM

Bei näherer Betrachtung wird deutlich, dass sich die Phasenmodelle DMAIC und CRISP-DM wunderbar ergänzen. Der DMAIC liefert dabei vielfach Werkzeuge und Methoden für die jeweiligen Phasen und Phasenziele von CRISP-DM. Durch das Übereinanderlegen bzw. das Integrieren von CRISP-DM in den DMAIC wird also der Kreis der Experten erweitert, die in Abhängigkeit von der jeweiligen Phase eines entsprechenden Projektes mit hinzugezogen werden müssen. Die interdisziplinäre Zusammenarbeit erweist sich für Projektteams im Umfeld von Big Data als wichtiger denn je.

Abbildung 3: Integration von CRISP-DM und DMAIC

1 Einleitung

Die vorliegende Richtlinie gliedert sich in folgende Teile auf.

Zunächst erfolgt eine kurze Definition des Anwendungsbereichs des Aufbaukurses Data-Mining, gefolgt von einer Erläuterung zur Dauer des Kurses.

Danach wird eine Einordnung der Trainingsinhalte in die verschiedenen Projektphasen gemäß dem DMAIC vorgenommen.

Im nächsten Abschnitt werden die Trainingsinhalte hinsichtlich ihres Umfangs (Vermittlung) und ihrer Ziele im Rahmen des Aufbaukurses klassifiziert, wobei das gebräuchliche ESSC-D-eigene Bewertungsschema Anwendung findet.

An dieser Stelle ist es wichtig anzumerken, dass das Schulen der Werkzeuge nicht in der jeweilig angegebenen Phase vorgegeben ist. Die Aufzählung versteht sich vielmehr als Vorschlag bzw. Anregung. Da viele Tools in mehreren Phasen genutzt werden können, liegt es in der Verantwortung der Trainer*innen, die Inhalte entsprechend der angewendeten Didaktik zum passenden Zeitpunkt zu vermitteln.

Im Anschluss werden die einzelnen Schulungsthemen genauer beschrieben, um den engen Zusammenhang zwischen Digitalisierung bzw. Big Data auf der einen und Six Sigma auf der anderen Seite aufzuzeigen. Die aktuellen Anforderungen an die veränderten Qualifikationen des Six Sigma Belts werden verdeutlicht.

Abschließend wird das Muster eines möglichen Trainingsablaufes zur Verfügung gestellt, an dem Trainer*innen sich ggf. orientieren können.

2 Anwendungsbereich

Die Richtlinie beschreibt die empfohlenen Ergänzungen der Ausbildung für Six Sigma Green Belts, Six Sigma Black Belts und allgemein für Methodenexpert*innen. Umfang und Vermittlungstiefe werden klassifiziert beschrieben und dienen dem Abgleich mit bestehenden oder neu zu entwickelnden Trainings. Da es sich um einen Grundlagenkurs handelt, gibt es keine Unterschiede in Umfang und Vermittlungstiefe. Teilnehmer*innen lernen grundlegende Datenanalysetechniken selbständig anzuwenden. Aufbauend auf den vermittelten Grundlagen von Data-Mining und Machine Learning ist es ein leichtes das Erlernte gezielt auszubauen. Aber auch um in Projekten die Sprache der Spezialisten ausreichend verstehen zu können, wird eine Unterscheidung der Lerntiefe nicht empfohlen.

3 Dauer der Ausbildung

Für die Zusatzausbildung Data-Mining für Six Sigma Green Belts, Six Sigma Black Belts und Methodenexpert*innen sind für das Vermitteln der in der Folge beschriebenen Inhalte und das Erreichen der erforderlichen Vermittlungstiefe mindestens 3 Unterrichtstage mit mindestens 30 Unterrichtseinheiten (UE) à 45 Minuten plus Pausen zu absolvieren.

In der Praxis kann es sinnvoll sein, die Ausbildungsdauer auf 4 Unterrichtstage mit in Summe 40 Unterrichtseinheiten à 45 Minuten plus Pausen zu verlängern.

An Hochschulen ist es zulässig, die erforderliche Vermittlungstiefe durch Aufteilen der Unterrichtseinheiten auf Präsenzzeit (Vorlesung) und anteiliges Selbststudium zu erreichen. Der empfohlene Anteil des Selbststudiums beträgt nicht mehr als 25 % der Gesamtunterrichtseinheiten. Dabei stellt der Minimalumfang des oben beschriebenen Standardtrainings (30 UE) die Basis dar. Die im Selbststudium zu leistenden Unterrichtseinheiten werden mit dem Faktor drei berechnet. Für die Zusatzausbildung Data-Mining ergeben sich damit bei maximaler Ausnutzung des Selbststudienanteils von 25 % und minimaler Anzahl von Unterrichtseinheiten gerundet 24 Unterrichtseinheiten Präsenzzeit und zusätzlich 24 Unterrichtseinheiten (8x3) Selbststudium.

Für die Ausgabe einer Teilnahmebescheinigung an die Teilnehmer*innen sollten diese mindestens 85 % der für dieses Training geplanten Gesamtstundenzahl anwesend gewesen sein.

4 Trainingsinhalte, projektphasenorientiert

Im folgenden Kapitel werden die einzelnen Trainingsinhalte und ihr jeweiliger Nutzen den DMAIC-Phasen zugeordnet.

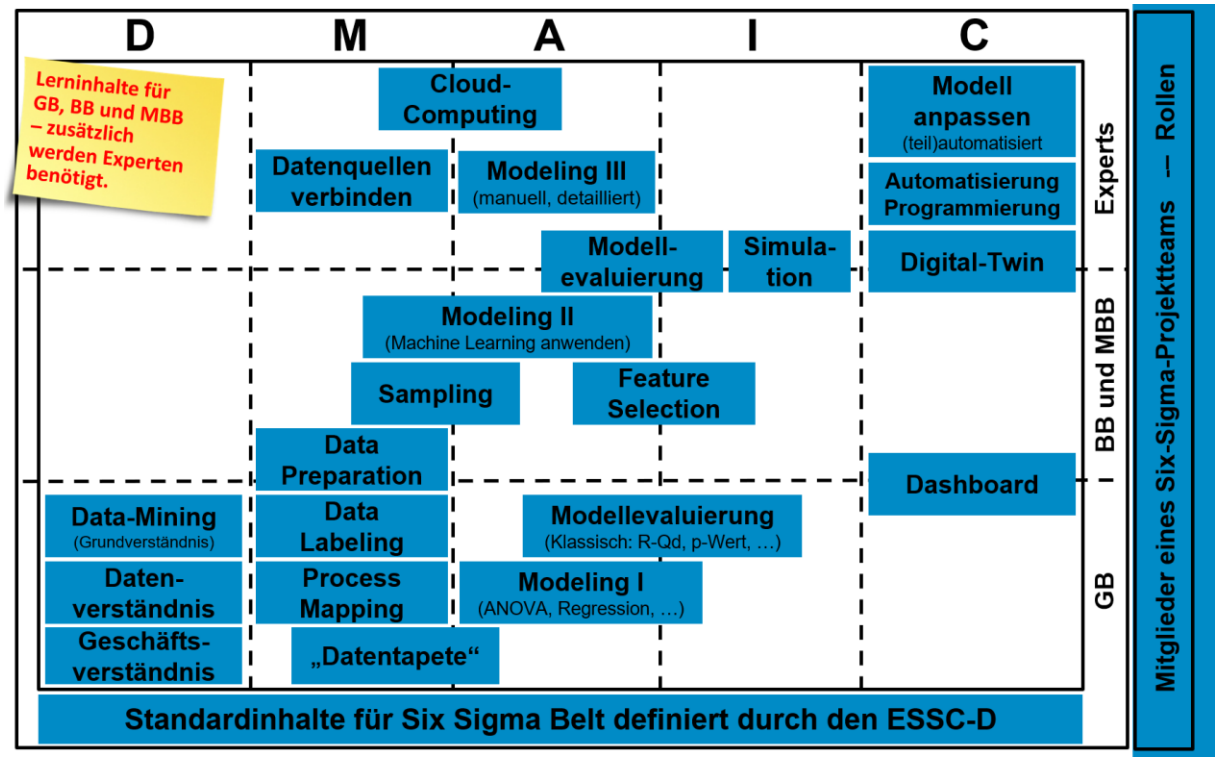

Abbildung 4: Übersicht – Neue Trainingsinhalte und Phasenzuordnung

4.1 Kick-off (K)

- Notwendigkeit und Ziel des Aufbaukurses Data-Mining
  - Im Fokus: Deskriptive Analysen und Zusammenhangsmodelle in komplexeren Datenstrukturen, Prädiktive Analytik und Präskriptive Analytik
- Grundbegriffe der Digitalisierung (KI, AI, ML, AR, I4.0, IOT)
  - Zielgerichtetes Anwenden von Schlagworten und ihrer Bedeutung und den damit verbundenen Ansätzen
- Moderne Projektmanagementkonzepte in der digitalen Transformation
  - DMAIC und moderne Projektmanagementansätze, wie z. B. Lean, Scrum, Agile, Design Thinking
  - Einbinden moderner Managementmethoden in den klassischen DMAIC

4.2 DEFINE (D) – Business Understanding

- Stakeholderanalyse (ggf. erweitert und/oder kombiniert mit RACI)
  - Erweiterter Prozessfokus durch moderne computergestützte Analyseverfahren
  - Zunehmende Bedeutung von umfassenden Stakeholdermanagement
- Data SIPOC / Process Data Map
  - Abgrenzung, aus welchen Bereichen der Wertschöpfungskette Daten betrachtet werden
  - Identifikation und Beschreibung der auf die Prozesse und Produkte wirkenden Parameter und Datenquellen

4.3 MEASURE (M) – Data Understanding and Preparation

- Erweiterter Datenerhebungsplan
  - Übersicht zu projektrelevanten Kennwerten mit Datenherkunft und Beispielen mit Anforderungen an die Datenqualität
  - Identifikation und Aufbereitung der stark anwachsenden Anzahl strukturierter und unstrukturierter Daten
- Grundlagen von ETL und gängigen Werkzeugen
  - Extrahieren der relevanten Daten aus verschiedenen Quellen
  - Transformieren der Daten in das Schema und das Format der Zieldatenbank
  - Laden der Daten in die Zieldatenbank
- Feature Selection
  - Schnelle und einfache Identifikation potenzieller Faktoren aus einer großen Anzahl von Parametern
- Erweiterte C&E Matrix
  - Abgleich der Experten- und Datensicht als Basis für eine Diskussion zur Unterscheidung von Korrelation und Kausalität
- Werkzeuge für die Datenaufbereitung (Data Preparation)
  - Aggregieren, Pivotisieren, Transponieren, Binning
  - Umgang mit fehlenden Werten
  - Umgang mit Multikollinearität und Ausreißern

4.4 ANALYSE (A) – Modeling

- Unterschiede zwischen überwachtem und unüberwachtem Lernen (am Beispiel Anomalie-Erkennung und Vorhersagemodell)
    - Supervised Learning: Prädiktives Modell auf Basis von Eingabe- und Ausgabedaten entwickeln
    - Unsupervised Learning: Daten auf Basis von Eingabedaten gruppieren und interpretieren
- Häufig verwendete Data-Mining-Methoden
    - Bootstrapping (bessere und robustere Schätzungen für die Streuung der Parameter)
    - Labeling (Datenkategorisierung)
    - Ensemble (mehrere Modelle parallel verwenden und entscheiden lassen)
    - Feature Selection (automatisches Erkennen von Einflussgrößen)
- Explorative Datenanalyse (EDA): Grafiken und deskriptive Statistiken
    - Aufdecken von Regelmäßigkeiten, Abhängigkeiten oder besonderen Zusammenhängen zwischen scheinbar vollkommen ungleichen Daten
    - Multidimensionale grafische Darstellungen
- Umgang mit unbalancierten Daten
    - Aufbereiten von Daten, wenn nur wenige Fehler und viele „Gut“-Daten vorliegen, mit Hilfe von Verfahren wie Oversampling bzw. Undersampling
- Häufig auftretende Aufgaben im Data-Mining und dazugehörige Verfahren (Algorithmen)
    - Klassifikation und Regression (lineare und logistische Regression, Entscheidungsbäume, künstliche neuronale Netze)
    - Clustering und Segmentierung (K-Means, künstliche neuronale Netze, Clusterverfahren) zur Gruppierung anhand vorgegebener Kriterien
    - Abhängigkeitsanalysen (Assoziationsanalyse)
    - Diskriminanzanalyse (Unterscheidung von zwei oder mehreren Gruppen, die mit mehreren Merkmalen (Variablen) beschrieben werden)
- Typische Data-Mining-Werkzeuge
    - Python, R-Studio, KNIME, Rapid Miner, SPM, …
    - Konkretes Beispiel mit einem der vorgestellten Werkzeuge
- Strategien für die Auswahl von Trainings- und Testdaten
    - Vorstellen unterschiedlicher Ansätze je nach Anwendungsfall
- Kenngrößen zur Bewertung der Modellqualität für Trainings- und Testdaten
    - R-Qd, ROC, BIC, AIC, Confusion Matrix: Bedeutung und Vor- und Nachteile der verschiedenen Kenngrößen
- Cloud-Computing
    - Vor- und Nachteile des Cloud-Computings, die externe Infrastruktur für eine rechenintensive Modellerstellung, i. d. R. inklusive Speicherplatz, Rechenleistung, Anwendungssoftware als Dienstleistung etc.
- Übersicht der Standardlösungen auf Basis von Big Data und KI
    - Verfügbare Bibliotheken, Plattformen, Bild- und Sprachverarbeitung sowie Textverarbeitung (überwiegend als Open Source verfügbar)
- Nutzungsmöglichkeiten komplexerer Data-Mining-Modelle
    - Einstellungsempfehlungen für Modelle
    - Praxisbeispiele zu Projekten mit Prädiktiver Analytik und Präskriptiver Analytik

4.5 IMPROVE (I) – Verification

- Grundprinzipien des Model Deployment
    - CICD - Continuous Integration Continuous Deployment als ein mögliches Grundprinzip

4.6 CONTROL (C) – Deployment

- Performance Monitoring und Re-Training der Modelle
    - Überwachen der Leistungs- bzw. Prognosefähigkeit des verwendeten Modells bei sich technisch relevant verändernden Prozessen
    - Kontrollieren der Trainings- und Testdaten und gegebenenfalls Anstoßen eines neuen Anlernens des Modells
- SPC 4.0
    - Erweiterte Prozesskontrolle unter Verwendung moderner Verfahren bis hin zur Vollautomatisierung
    - Verwenden eines statistischen Modells zur Prozessüberwachung

5 Umfang und Ziele der einzelnen Themen

Die oben definierten Themen, Methoden und Tools beschreiben welche Inhalte für das Training mindestens gefordert sind. In diesem Abschnitt werden der Umfang und die Ziele dieser Themen anhand von Klassifizierungen spezifiziert. Das Ergebnis des jeweiligen Trainings muss die benannte Klasse oder höhere erreichen, um der Guideline zu entsprechen.

5.1 Legende zur Klassifizierung

5.1.1 Klassifizierung für den Umfang (Vermittlung)

Klasse	Bedeutung
A	Methode wurde erklärt
B	Methode wurde gemeinsam genutzt
C	Methode wurde allein oder in Gruppe geübt
D	Methode wurde geübt inkl. Feedback zur Übung

5.1.2 Klassifizierung der Ziele

Klasse	Bedeutung
1	Der Teilnehmer hat das Prinzip der Anwendung verstanden
2	"1" und Teilnehmer kann Tool auswählen & anwenden
3	„2“ und Teilnehmer kann wichtige Ergebnisse interpretieren
4	"3" und Teilnehmer kennt detailliert die Berechnungshintergründe
5	"4" und Teilnehmer kann Ergebnis auch von Hand errechnen

5.2 Klassifizierung für den Data-Mining Aufbaukurs

	Thema	Phase	Umfang (Vermittlung)	Ziel
1	Smart Knowledge Picking	K	A	2
2	DMAIC und moderne Projektmanagementansätze (Lean, Scrum, Agile, Design Thinking, ...)	K	B	1
3	Grundbegriffe der Digitalisierung (KI, AI, ML, AR, I4.0, IOT)	K	B	1
4	Möglichkeiten und Grenzen der induktiven Statistik und des modernen Data-Mining	K	B	1
5	Stakeholderanalyse (ggf. erweitert und/oder kombiniert mit RACI)	D	B	3
6	Data SIPOC / Process Data Map	D	C	3
7	Erweiterter Datenerhebungsplan (für projektrelevante Kennwerte mit Datenherkunft, Datenqualität und Beispielen)	M	B	3
8	Anforderungen an die Datenqualität (Messwerte, Kurven, Bilder, Text) für das Data-Mining	M	C	4
9	Feature Selection zur Komplexitätsreduktion	M	B	3
10	Erweiterte C&E Matrix (Feature Selection: ausgewählte Features mit Experten plausibilisieren)	M	B	3
11	Werkzeuge für die Datenaufbereitung (Data Preparation) (Aggregieren, Pivotisieren, Transponieren, Binning; Umgang mit fehlenden Werten, Multikollinearität und Ausreißern)	M	C	3
12	Grundlagen ETL/ELT und gängige Werkzeuge	M	B	1
13	Unterschiede überwachtes/unüberwachtes Lernen	M	B	1
14	Häufig verwendete Data-Mining-Methoden (Bootstrapping, Labeling, Komitee, Feature Selection, ...)	M	A	1
15	Explorative Datenanalyse (Grafiken und deskriptive Statistiken)	M	C	2
16	Umgang mit unbalancierten Daten (nur wenige Fehler, aber viele "Gut"-Daten)	M	A	1

17	Häufig verwendete Data-Mining-Algorithmen (Entscheidungsbaum, Neuronale Netze, Naive Bayes, SVM, K-Nearest, K-Means, Hauptkomponentenanalyse, Lasso, ...)	A	A	1
18	Erweiterte multivariate Analysen (Diskriminanzanalyse, Clusteranalyse, …)	A	A	1
19	Typische DM-Werkzeuge (kurze Übersicht über Tools: Python, R-Studio, KNIME, Rapid Miner, SPM…)	A	B	3
20	Multidimensionale grafische Darstellungen	A	A	1
21	Strategien für die Auswahl von Trainings- und Testdaten	A	C	3
22	Kenngrößen zur Bewertung der Modellqualität (R-Qd, ROC, BIC, AIC, Confusion Matrix) für Trainings- und Testdaten	A	B	3
23	Vor- und Nachteile von Cloud-Computing	A	A	1
24	Übersicht Standardlösungen auf der Basis von Big Data und KI (verfügbare Bibliotheken, Plattformen, Bildverarbeitung, Sprachverarbeitung, ...)	I	A	1
25	Nutzungsmöglichkeiten komplexerer DM-Modelle (Einstellungsempfehlungen, Prädiktive Analytik, Präskriptive Analytik) mit Beispielen aus der Praxis	I	A	1
26	Grundprinzipien von Model Deployment und Re-Training (Performance Monitoring, KI-Modelle, CICD – Continuous Integration Continuous Deployment)	C	C	3

6 Nutzen und typische Anwendungsbereiche

1. **Smart Knowledge Picking (SKP)**
   Das Konzept des Lernens unterlag in den letzten Jahren einem grundlegenden Wandel. Mit diesem Thema hat sich der Arbeitskreis „Six Sigma Weitergedacht“ im Rahmen des ESSC-D beschäftigt und dazu zahlreiche Projekte, Umfragen und Facharbeiten in seine Untersuchung einbezogen. Die Erfahrung aller Beteiligten und die Vorstellung bzw. Veröffentlichung der Ergebnisse zeichnen ein klares Bild. Für akute Aufgaben ist es notwendig, sich das Wissen/Spezialwissen häufig erst anzueignen. Unter SKP versteht man das gezielte und aufgabenbezogene Zusammentragen und Nutzen von Wissen.

2. **DMAIC und moderne Projektmanagementansätze**
   Die Six-Sigma-Methodik hat seit ihrer Einführung nichts von ihrer Aktualität und Flexibilität verloren. Je nach Unternehmensphilosophie kommen vielfältige Projektmanagementansätze und Methoden zum Einsatz. Viele Unternehmen entwickeln aus den verfügbaren Verfahren ihre eigene favorisierte Vorgehensweise. Auch Six Sigma Belts mit ihrem Methodenwissen arbeiten oft in Teams, die andere Philosophien/ Methoden verfolgen. Damit der Belt der Zukunft sich in diesem Umfeld sicher und unterstützend einbringen kann, sollten die Grundprinzipien gängiger Projektmanagementmethoden bekannt sein.

3. **Grundbegriffe der Digitalisierung**
   Projektteams, Sponsoren und das Management eines Unternehmens werden zunehmend mit den Herausforderungen der Digitalisierung konfrontiert. Im Idealfall gestalten sie diese auch aktiv mit. In beiden Fällen ist es notwendig, dass der Belt der Zukunft Begriffe und Ansätze aus der Digitalisierung richtig einordnen, bewerten und anwenden kann.

4. **Möglichkeiten und Grenzen von induktiver Statistik und modernem Data-Mining**
   In Qualitäts- und Digitalisierungsprojekten kann man, Stand heute (2021), sehr viele Datenanalyseaufgaben mit den klassischen statistischen Verfahren adressieren. Bestimmte Aufgaben, nicht nur im Kontext von Big Data, lassen sich mit modernen Verfahren des Data-Mining jedoch deutlich schneller und zuverlässiger bearbeiten. Bei bestimmten Datenlagen und Aufgaben sind Data-Mining-Algorithmen und -Methoden häufig die einzige Möglichkeit, Daten auszuwerten.

5. **Stakeholder (erweitert)**
   Optimierungen und Zusammenhangsanalysen werden infolge der Digitalisierung häufiger über die komplette Wertschöpfungskette hinweg angegangen. Dadurch steigt automatisch die Anzahl der zu berücksichtigenden Parteien und Personen.

6. **Data SIPOC / Process Data Map**
   Ein wichtiger Teil der Datenanalyse ist die Kenntnis darüber, welche konkreten Daten an welcher Stelle im Prozess aufgenommen werden. Damit kann der Prozessablauf mit den verschiedenen Datenquellen synchronisiert werden („Data Understanding“).

7. **Erweiterter Datenerhebungsplan/Datenkatalog**

   In der Process Data Map sind alle Merkmale visualisiert und den Prozessschritten zugeordnet. Auf dieser Grundlage wird im Datenerhebungsplan das Domänenwissen zusammengetragen und kategorisiert. Relevant sind dabei Informationen wie Erfassungsfrequenz, Skalenniveau, Spezifikationen, Stör- und Steuergrößen und alle weiteren eventuell für das Projekt relevante Themen. Die hieraus resultierende Tabelle wiederum kann als Basis für die Auswahl der richtigen Merkmale zur Modellierung dienen und als Basis für FMEAs oder andere klassische Werkzeuge herangezogen werden.

8. **Anforderungen an die Datenqualität für das Data-Mining**

   Zur Diskussion der Datenqualität gehören auch Überlegungen zur Eignung des Messsystems. Auch im Kontext von Data-Mining spielen die Datenqualität und die Repräsentativität eine entscheidende Rolle.

   Im Zusammenhang mit der Digitalisierung und der damit verbundenen Datenmenge sind auch digitale Übertragungsfehler und Kompatibilitätsthemen immer häufiger anzutreffen.

   Um eine analysierbare Datenstruktur herzustellen, benötigt man unterschiedliche Werkzeuge und Methoden zur Datenorchestrierung. Die zu erreichende Zielstruktur ist in der Regel lesbar.

9. **Feature Selection zur Komplexitätsreduktion**

   Durch die Betrachtung gesamter Wertschöpfungsketten und die Möglichkeiten, große Mengen an Informationen zu speichern, erhält man rasch sehr viele Features (Spalten mit Variablen). Häufig entstehen auf diese Weise Tabellen mit mehreren hundert Spalten, die für Zusammenhangsmodelle infrage kommen. Mit Hilfe von Machine-Learning-Verfahren kann man die potenziell wichtigsten und unwichtigsten Features zum Beantworten der jeweiligen Fragestellung identifizieren.

10. **Erweiterte C&E Matrix**

    Nach der Feature Selection sollte man die Ergebnisse mit Experten validieren. Ziel ist es, einerseits zu prüfen, ob die wichtigsten Features technisch plausibel sind. Andererseits muss geprüft werden, ob technisch erwartete Features ungewollt entfernt wurden. Eine erweiterte C&E Matrix kann verwendet werden, um einzelne Features von den Experten und parallel von einem statistischen Modell gewichten zu lassen. So können Widersprüche identifiziert, diskutiert und aufgelöst werden.

11. **Werkzeuge für die Datenaufbereitung (Data Preparation)**

    Das Herstellen einer analysierbaren Datenstruktur beansprucht den Großteil der für die Datenanalyse verwendeten Zeit. Analysedaten mit vielen bis sehr vielen Merkmalen in eine analysierbare Struktur mit „Bordmitteln“ zu überführen, ist nicht zu empfehlen. Der Umgang mit fehlenden Werten, unterschiedlichen Informationsdichten, dem Berechnen von Ersatzgrößen, dem Aggregieren von Daten und stark korrelierenden Features sollte geübt werden. Zudem ist die Art und Weise der Datenaufbereitung unbedingt nachvollziehbar zu dokumentieren und häufig auch Bestandteil der Ergebnisdiskussion.

    Typische Werkzeuge sind dabei Aggregieren, Pivotisieren, Transponieren, Binning, aber auch Methoden zum Umgang mit fehlenden Werten, Multikollinearität und Ausreißern.

12. **Grundlagen von ETL/ELT und gängigen Werkzeugen**
Beim ETL-Prozess (Extract-Transform-Load) handelt es sich um mehrere Einzelschritte, durch die sich Daten aus verschiedenen Datenquellen (strukturiert oder unstrukturiert) in ein Data Warehouse (DW) integrieren lassen. Der Prozess kommt häufig zur Verarbeitung großer Datenmengen im Umfeld von Big Data und Business Intelligence zum Einsatz.

13. **Unterschiede zwischen überwachtem und unüberwachtem Lernen**
Im Machine Learning unterscheidet man hauptsächlich (aber nicht ausschließlich) zwischen zwei Arten von Lerntypen: Supervised (überwachtes Lernen) und Unsupervised Learning (unüberwachtes Lernen). Während es beim Supervised Learning darum geht, „Zusammenhänge zwischen X und Y zu erkennen“ oder die „Zukunft vorherzusagen“, lautet das Ziel beim Unsupervised Learning, Muster in den vorhandenen Daten zu entdecken und zu verstehen.

14. **Häufig verwendete Data-Mining-Methoden**
In der Analyse komplexerer Datenstrukturen gehören auch für Six Sigma Belts einige Methoden und Verfahren des Data-Mining zum grundlegenden Handwerkszeug. Das schließt zum Beispiel Bootstrapping, Labeling, Ensemble und Feature Selection ein.

15. **Explorative Datenanalyse (Grafiken und deskriptive Statistiken)**
Die Visualisierung von Eingangs- und Zielgrößen sowie Ergebnissen, inklusive statistischer Kennzahlen, stellt einen wesentlichen Bestandteil der Datenanalyse dar. Bei komplexeren Datenstrukturen und Ergebnissen benötigt man geeignete Software und Grafiktypen zur Visualisierung der Sachverhalte (Stichworte: Heatmap, Korrelationsmatrix, Parallelkoordinaten und andere multidimensionale Darstellungen).

16. **Umgang mit unbalancierten Daten (nur wenige Fehler, aber viele "Gut"-Daten)**
Für die Entwicklung von Zusammenhangsmodellen ist es in der Regel vorteilhaft, wenn balancierte Daten zum Trainieren eines Modells vorliegen. Entwickelt man Modelle auf Basis historischer Daten, sind diese typischerweise nicht balanciert.
Ausgehend vom Beispiel „Nur wenige Fehler, aber viel „Gut“-Daten“ können folgende Methoden unterstützen:
    - Entferne „Gut“-Daten, um balancierte Daten zu bekommen.
      (Gleich viele Gute wie Fehler)
    - Erzeuge künstliche Fehler durch Kopieren oder leichte Variation. (Gleich viele Fehler wie Gute)
    - Wähle einen Algorithmus, der auf unbalancierte Daten optimiert wurde.

    Wie man damit umgehen und einen soliden Trainingsdatensatz entwickeln kann, sollte man als Belt wissen.

17. **Häufig verwendete Data-Mining-Algorithmen**
Ein Belt sollte die gängigsten Algorithmen und typische Anwendungsbeispiele in der Praxis kennen, um diese bei konkreten Fragestellungen benennen und eventuell auch anwenden zu können. Dazu gehören zum Beispiel Entscheidungsbäume, Neuronale Netze, Naive Bayes, SVM, K-Nearest, K-Means, Hauptkomponentenanalyse und Lasso.

18. **Erweiterte multivariate Analysen**
Das klassische Training zum Six Sigma Green Belt oder Six Sigma Black Belt endet im Bereich der Modellentwicklung bei einfacher und multipler Regression. Um Nutzen aus komplexeren Datenstrukturen ziehen zu können oder auch um diese zu vereinfachen, sollten multivariate Verfahren bekannt sein und angewendet werden können. Dazu gehören Clusteranalysen, Dendrogramme, Hauptkomponentenanalyse, Diskriminanzanalyse und andere.

19. **Typische Data-Mining-Werkzeuge**
Typische Softwarepakete oder kostenfreie Skriptsprachen werden im Kurs vorgestellt und angewendet. Dazu zählen zum Beispiel Python, R-Studio, KNIME, Rapid Miner und SPM.

20. **Multidimensionale grafische Darstellungen**
Bei komplexeren Datenstrukturen und Ergebnissen benötigt man geeignete Software und Grafiktypen zur Analyse der Sachverhalte (Stichworte: Heatmap, 3D-Diagramme, Konturdiagramme, Korrelationsmatrix, Parallelkoordinaten und andere).

21. **Strategien für die Auswahl von Trainings- und Testdaten**
Komplexere Datenstrukturen beinhalten häufig auch viele Beobachtungen (Zeilen). Klassische statistische Verfahren, wie t-Tests, aber auch die Varianzanalysen, zeigen überwiegend statistische Signifikanz an. Die „p-Werte“ liefern hier häufig keine nützliche Information. Daher ist es üblich, eine Teilmenge der verfügbaren Daten zu bilden. Auf der einen Teilmenge (Trainingsdaten) werden statistische Analysen durchgeführt und eine Kreuzvalidierung gegen die andere Teilmenge (Testdaten) vorgenommen. Die Güte der Kreuzvalidierung entdeckt ein mögliches „Overfitting“ und liefert ein weiteres Kriterium zur Nutzbarkeit des statistischen Modells. Für die Aufteilung in Trainings- und Testdaten gibt es unterschiedliche Philosophien, die ein Belt kennen und anwenden können sollte.

22. **Kenngrößen zur Bewertung der Modellqualität für Trainings- und Testdaten**
Ein Belt sollte in die Lage versetzt werden, die Güte und damit Anwendbarkeit statistischer Modelle zu bewerten und kritisch hinterfragen zu können. Wichtige Kenngrößen sind zum Beispiel R-Qd, ROC, BIC, AIC und Confusion Matrix, deren Vor- und Nachteile vermittelt werden.

23. **Vor- und Nachteile von Cloud-Computing**
Je nach Anwendungsbereich und der damit notwendigen Datenverfügbarkeit spielen unterschiedliche Systeme ihre Stärken aus.
Im Hochverfügbarkeitsbereich und im Umgang mit sensiblen Daten fallen Cloud-Systeme bspw. häufig aus technischen Überlegungen heraus. Sobald man lokale oder Cloud-Systeme verwendet, müssen immer auch die rechtlichen Aspekte berücksichtigt werden.
In Bezug auf die Skalierbarkeit der Rechenleistung ist man mit Cloud-Systemen jedoch deutlich flexibler und günstiger, da man nur für die tatsächliche Rechenzeit bezahlt.

24. **Übersicht Standardlösungen auf der Basis von Big Data und KI**
Im Umfeld von Big Data und KI gibt es viele Open-Source-Lösungen, die für einen Belt effektiv und praxisnah eingesetzt werden können. Im Allgemeinen ist hierbei zu empfehlen, auf Standardlösungen zurückzugreifen, die sich im Six-Sigma-Umfeld bereits bewährt haben. Dazu gehören Bibliotheken unter anderem zur Datenanalyse, zur Bild- und zur Spracherkennung.

25. **Nutzungsmöglichkeiten komplexerer Data-Mining-Modelle**
Konkrete Anwendungsbeispiele von Modellen aus dem Bereich Prädiktive Analytik und Präskriptive Analytik aus der Praxis sollten den Teilnehmer*innen vorgestellt werden. Benennung der Erfolgsfaktoren bei der Datenaufbereitung und Parametrierung der Modelle.

26. **Grundprinzipien von Model Deployment und Re-Training**
Wird ein Modell regelmäßig im Betrieb genutzt, müssen Regeln und Prinzipien definiert werden, die die Aktualität und damit die Performance des Modells sicherstellen. Es wird ein kontinuierlicher Verbesserungsprozess für die Modelle eingeführt, die häufig der CICD-Pipeline (Continuous Integration Continuous Deployment) folgen.

7 Möglicher Trainingsverlauf

Bei der Entwicklung eines Trainings zu diesen Themen kann folgender Vorschlag gerne zur Inspiration herangezogen werden.

Tag 1:

Vormittag (Fokus *Auffrischung*):

- Erfahrungsaustausch, Erwartungen und Ziele
- Explorative Datenanalyse und Datenvorbereitung
- Entwickeln von Modellen als Teil der Induktiven Statistik (Korrelation, Regression und ANOVA)
- Bewerten der Modell- und Prognosequalität mit Vertrauensbereichen

Nachmittag (Fokus *Begriffe, Möglichkeiten und Grenzen*):

- Industrie 4.0 in Produktionsprozessen – Transformation von Daten in Nutzen
- Typische Datenstrukturen und die sich daraus ergebenen Herausforderungen
- Typische Verfahren, Ansätze und Begriffe des Data-Mining
- Ein Vorgehensmodell zur strukturierten Datenanalyse (bspw. CRISP-DM)
- Model Lifecycle Management – „SPC der Zukunft“

Tag 2:

Vormittag (Fokus *Datenaufbereitung und Datenerkundung*):

- Vorstellen von Fallbeispiel(en)
- Explorative Datenanalyse und Datenvorbereitung
- Sicherstellen, dass Aufgabe und Daten verstanden sind (Business und Data Understanding)
- Strategien zum Entwickeln von Trainings- und Testdatensätzen

Nachmittag (Fokus *Supervised Learning → Regressionsmodelle*):

- Entwickeln von Regressionsmodellen und Entscheidungsbäumen
- Absichern von Modellen mit Hilfe der Kreuzvalidierung
- Modellauswahlverfahren
- Feature Selection, Bootstrapping und andere
- Überführen der Ergebnisse/Erkenntnisse in eine stringente Aussagelogik (klare Handlungsempfehlung/klare und einfache Ergebnispräsentation)
- Model Lifecycle Management für das gemeinsam entwickelte Modell

Tag 3:

Vormittag (Fokus *Supervised Learning → Klassifikationsmodelle*):

- Vorstellen von Fallbeispielen
- Explorative Datenanalyse und Datenvorbereitung
- Sicherstellen, dass Aufgabe und Daten verstanden sind (Business und Data Understanding)
- Strategien zum Entwickeln von Trainings- und Testdatensätzen, speziell bei unbalancierten Zielgrößen

Nachmittag (Fokus *Unsupervised Learning*):

- Einblick in die multivariate Datenanalyse zur Identifikation von Clustern
- Klassische Data-Mining-Ansätze zur Identifikation von Clustern
- Absichern von Erkenntnissen mit Hilfe der Kreuzvalidierung
- Überführen der Ergebnisse/Erkenntnisse in eine stringente Aussagelogik

„Lernen ist wie Rudern gegen den Strom. Hört man damit auf, treibt man zurück."
(Laozi, chinesischer Philosoph, 6. Jh. v. Chr.)