Aufgabe 1: Vergleichbarkeit und Reproduzierbarkeit

a) Warum vergleicht man die Algorithmen, Modelle, Performance-Metriken usw.?

Bei Smart Data Analytics können auf einem gegebenen Datensatz eine Vielzahl von unterschiedlichen Algorithmen angewendet werden, die mit verschiedenen Hyperparametern unterschiedliche Modelle erzeugen können.

Wollen wir nun mit diesen Modellen Erkenntnisse aus den Daten gewinnen und Vorhersagen produzieren, so benötigen wir eine Möglichkeit, die Modelle zu bewerten um die besten Modelle auszusuchen und zu verwenden.

Hierzu benötigt man Performance-Metriken, um verschiedene Modelle miteinander vergleichen zu können.

b) Was versteht man unter Reproduzierbarkeit bezüglich Datenanalytik?

Reproduzierbarkeit bedeutet, dass es für andere Personen möglich sein muss, die gleichen Resultate (unter Verwendung der gleichen Tools und Daten) zu einem beliebigen Zeitpunkt erneut zu erhalten.

c) Was muss man beachten, damit die Ergebnisse vergleichbar und reproduzierbar sind?

Natürlich müssen zunächst alle verwendeten Skripte, Daten und sonstigen Artefakte zusammen mit den Ergebnissen archiviert werden.

Außerdem muss eine Spezifikation der verwendeten Umgebung (insbesondere Tools und Pakete mit Versionen) beigefügt werden.

Die Analyse muss vollständig deterministisch sein, an allen Stellen an denen Zufall verwendet wird muss daher manuell ein random Seed gesetzt werden.