Time	Group	Submission in Moodle; Mails with subject: [SMD2023]
Th. 12:00–13:00	A	tristan.gradetzke@udo.edu and samuel.haefs@udo.edu
Fr. 09:00–10:00	B	lucas.witthaus@udo.edu and david.venker@udo.edu

Exercise 25 *Datenaufbereitung* **0 p.**

(a) Wie sollten nicht-numerische Datentypen wie beispielsweise Strings vor der Analyse behandelt werden müssen?

(b) Kann es hilfreich sein Attribute zu normieren? Wenn ja, wieso?

(c) Wie kann mit Lücken in den Daten oder NaNs und Infs verfahren werden?

(d) Was ist beim Zusammenführen von Datensätzen zu beachten?

(e) Welche Attribute sollten vor dem Trainieren des Klassifizierers aus dem Datensatz entfernt werden. Wie kann dabei eine Reduktion redundanter Informationen erreicht werden? Was muss speziell bei simulationsbasierten Methoden berücksichtigt werden?

Exercise 26 *Multivariate Regression* **0 p.**

In dieser Aufgabe sollen Sie eine 2 dimensionale multivariate Regression mit *sklearn* durchführen und die Ergebnisse anschaulich darstellen.

(a) Erstellen sie ein Dataframe mit 10^5 uniform zwischen 0 und 1 verteilten Zufallszahlen x_1, x_2.

(b) Berechnen sie aus diesen Attributen ein drittes Attribut x_3 mit der Funktionsvorschrift:

$$x_3 = 15\sin(4\pi x_1) + 60(x_2 - 0.5)^2$$

Addieren Sie auf diese Zahl eine standardnormalverteilte Zufallszahl, um Rauschen zu simulieren. Das x_3 Attribut ist von nun an Ihr Zielattribut.

(c) Teilen Sie das Dataframe in einen Trainings- und Test-Datensatz auf.

(d) Wählen Sie einen Random-Forest-Regressor mit 200 Bäumen und trainieren Sie diesen auf dem Trainingsdatensatz um x_3 zu schätzen.

(e) Stellen Sie die erstellten Daten und die Vorhersagen des Regressors in einem dreidimensionalen Plot und mehreren 2 dimensionalen Projektionen dar um die Vorhersage mit der Wahrheit zu vergleichen. Geben sie außerdem den *mean-squared-error* der Vorhersage zu den wahren Werten an.

(f) Erstellen Sie einen weiteren Datensatz, bei dem x_1 und x_2 uniform verteilte Zufallszahlen zwischen 1 und 2 sind. Testen sie nun das in **(d)** trainierte Modell auf dem neu erstellten Datensatz. Trainieren sie das Modell NICHT auf dem neuen Datensatz. Was für Probleme können hier auftreten und was ist die Vorhersage des Regressors? Stellen Sie das Ergebnis wie in Aufgabe **(e)** dar.