UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR ROBOTIK
UND KOGNITIVE SYSTEME

Künstliche Intelligenz
Sommersemester 2022
16. Mai 2022

Niclas Bockelmann
bockelmann@rob.uni-luebeck.de

Übungsblatt 5

Abgabe bis zum 24. Mai 2022, 10 Uhr, im Moodle.

Aufgabe 1: To Pass or not to Pass: That is the question! (10 Punkte)

Wie hoch ist die Wahrscheinlichkeit, dass Sie Ihr nächstes KI-Übungsblatt bestehen? Finden Sie es mithilfe des *Naive Bayes*-Klassifikators heraus! Folgende Attribute beeinflussen Ihr **Bestehen**:

- Ihr **Wissen** zu dem Stoff auf dem Übungsblatt: gut, mittel, schlecht
- Ihre **Lust**, das Übungsblatt zu bearbeiten: keine, geht so
- Die **Deadline** zur Abgabe: < 2 Tage, ≥ 2 Tage
- **Rick and Morty**: "Mist, leider nur eine Wiederholung", "Yeah, endlich Staffel 6!"

Nehmen wir an, dass die Auswertung der ersten 11 Übungsblätter folgendes ergab:

Übungsblatt	**Wissen**	**Lust**	**Deadline**	**Rick and Morty**	Bestehen
1	gut	geht	≥ 2	Wiederholung	ja
2	gut	geht	≥ 2	Wiederholung	ja
3	mittel	keine	< 2	neue Folge	ja
4	schlecht	geht	< 2	neue Folge	nein
5	gut	geht	< 2	Wiederholung	ja
6	mittel	keine	≥ 2	neue Folge	nein
7	schlecht	geht	≥ 2	neue Folge	ja
8	schlecht	keine	≥ 2	neue Folge	ja
9	gut	geht	< 2	Wiederholung	ja
10	schlecht	geht	≥ 2	Wiederholung	ja
11	gut	keine	< 2	neue Folge	nein

a) Bestehen Sie das 12. Übungsblatt, wenn Ihr Wissen gut ist, Sie keine Lust haben, nur noch weniger als 2 Tage zur Verfügung stehen und eine neue Folge der Staffel 6 von Rick and Morty läuft ("neue Folge")? (7 Punkte)

b) Wie sieht es aus, wenn Ihre Einstellung stattdessen *geht so* ist? (3 Punkte)

Aufgabe 2: Predicting Diabetes (10 Punkte)

Sie wollen voraussagen, ob eine Patientin Diabetes hat oder nicht. Hierfür möchten Sie auf den kNN-Algorithmus zurückgreifen. Der Datensatz besteht aus 768 Patientinnen und beinhaltet 8 Features und die Zielklasse mit den Werten: *tested_positiv* und *tested_negative*. Die 8 Feature sind allsamt numerisch und setzen sich zusammen aus:

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/mli)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

Laden Sie sich das juypter notebook aus dem Moodle herunter. Für diese Aufgabe benötigen Sie die Packages *scikit-learn*, *pandas*, *matplotlib* und *seaborn*.

a) Wandeln Sie zunächst die Zielklassenlabels in 0 und 1 um. Das Label *tested_positive* soll dabei 1 entsprechen. (1 Punkt)

b) Plotten Sie die Verteilung aller Feature als Histogramme innerhalb einer Figure. Nutzen Sie dafür die *seaborn* Funktion `histplot`. Speichern Sie die Figure als png-Datei. (1 Punkt)

c) Plotten Sie die Verteilung der Klassenlabel als Histogramm. Speichern Sie die Figure als png-Datei ab. (1 Punkt)

d) Kümmeren Sie sich um die Skalierung der Daten. Hierfür dürfen Sie auf die Funktionen von *scikit-learn* zurückgreifen. (2 Punkte)

e) Trainieren Sie ihre kNN-Modelle von *scikit-learn* `KNeighborsClassifier`. Nutzen Sie hierfür die durch den Training-Test-Split erzeugte Trainings- und Testdaten, Trainieren Sie nun ihre kNN-Modelle mit verschiedenen k-Werten von 1 bis 200 und tracken Sie jeweils die folgenden Metriken auf den Testdaten: F1, Accuracy, AUPRC und AUROC. Hierfür können die entsprechenden *scikit-learn* Funktionen verwendet werden. (3 Punkte)

f) Was sind die besten Ergebnisse die Sie erreichen können? Plotten Sie die vier Metriken über die k-Werte in einer Figure mit der *seaborn* Funktion `relplot`. Speichern Sie die Figure als png-Datei ab. Diskutieren Sie Ihre Ergebnisse im Code. (2 Punkte)

Geben Sie Ihre Implementierung und alle Figures als zip-Datei im Moodle ab.