KI programmieren im Informatikunterricht Teil 7: Data Scientist

[B2] [B4] [B3] [B0]

ver: 1.0, 29.05.2023, 09:19:27

Abb.[B0]: „DataScientistSammeltModelliertTrainiert“, A. Schindler Lizenz CC BY-SA 4.0, Lernaufgabe "KI im Unterricht" unter Verwendung weiterer Quellen (s. Bildnachweis)

Inhaltsverzeichnis

A Übersicht

A 1 Zusammenfassung:

Im KI-Bereich entwickeln sich rasant neue Berufsfelder und z.T. unerwartete Berufschanchen. In dieser Lernaufgabe sollen Die SuS einige dieser Berufsfelder besser kennenlernen.

Intention der Lernaufgabe

- Kennnlernen möglicher Berufsbilder im Bereich von KI

Hinweise:

Voraussetzung sind die Grundlagen zur KI, wie sie in der Lernaufgabe „KI programmieren im Informatikunterricht Teil I: Einführung“ vermittelt werden. Für einige Teilaufgaben empfiehlt sich die Kenntnisse aus der Lernaufgabe: „KI programmieren im Informatikunterricht Teil 2: Bilderkennung“.

Desweiteren wird eine Python IDE (z. B. Spyder) mit installiertem TensorFlow und Seaborn empfohlen.

A2 Stundenübersicht

Doppelstunde:

- Hinführung zum Thema
- SuS erarbeiten sich in Gruppenarbeit anhand von Arbeitsblättern und praktischen Übungen ein Blick auf neue Berufsfelder im Bereich KI
- Vorstellung der Gruppenarbeiten
- Fazit

1. Doppelstunde

Zeit	***Phase***	***Beschreibung***	***Methode/Medien/ Anmerkung***
10 Min.	Einstieg	*Zur Einordnung von Berufsbildern im KI – Bereich; i ns Heft:*	Präsentation

Zeit	Phase	Beschreibung	Methode/Medien/ Anmerkung
		• *Was macht eine Data Scientist* • *CRISP-DM*	
50 Min.	Erarbeitung	Grupenarbeit: Jede Gruppe bearbeitet die Aufgaben 1. und 2. anhand eines Arbeitsblattes. Leitfragen dienen zur Orientierung und zur weitergehenden Arbeit am Thema.	Arbeitsblätter, IDE mit vorinstalliertem TensorFlow; Rückgriff auf bestehende Projekte möglich
30 Min.		Besprechung der Gruppenarbeiten	
5 Min.		Abschlussdiskussion und Fazit	

A3 Themeneinstieg und theoretische Grundlagen

Zentrale Aspekte in der Arbeit eines Data Scientist ist auf der einen Seite das **Domänenwissen**, also der fachliche Kontext in dem die Daten Anwendung finden sollen und auf der Softwareentwicklungsseite das beherrschen der passenden **Datenstrukturen**. Grundlegender Prozess dafür ist CRISP-DM. CRISP-DM steht für *Cross-Industry Standard Process for Data Mining* und ist ein weit verbreiteter und anerkannter Prozess zur Durchführung von Data-Mining-Projekten. Er bietet eine strukturierte und systematische Vorgehensweise, um aus Rohdaten wertvolle Informationen und Erkenntnisse zu gewinnen.

Der CRISP-DM-Prozess besteht aus sechs Hauptphasen, die sequenziell durchlaufen werden, aber auch iterativ sein können:

1. **Domänenwissen / Geschäftsverständnis (Business Understanding):**
   In dieser Phase werden die Ziele und Anforderungen des Projekts identifiziert und verstanden. Es werden Fragen gestellt wie: Was sind die Ziele? Welche Fragen sollen beantwortet werden? Wie kann Data Mining dabei helfen?

2. **Datenverständnis (Data Understanding):**

Hier erfolgt eine Erkundung der verfügbaren Daten. Es wird untersucht, welche Daten vorhanden sind, woher sie stammen, wie sie strukturiert sind und welche potenziellen Probleme oder Lücken es gibt. Ziel ist es, ein umfassendes Verständnis der Datenbasis zu erlangen.

3. **Datenvorbereitung (Data Preparation):**
   In dieser Phase werden die Daten für die eigentliche Analyse vorbereitet. Dazu gehört die Auswahl der relevanten Daten, die Bereinigung von fehlerhaften oder inkonsistenten Einträgen, die Integration von verschiedenen Datenquellen und die Transformation der Daten in ein geeignetes Format für die Analyse.

4. **Modellierung (Modeling):**
   In dieser Phase werden die eigentlichen Data-Mining-Techniken angewendet, um Muster und Beziehungen in den Daten zu entdecken. Es werden verschiedene Modelle erstellt, getestet und bewertet, um dasjenige zu finden, das die besten Ergebnisse liefert.

5. **Modell testen / Auswertung (Evaluation):**
   Hier werden die erstellten Modelle und Ergebnisse bewertet. Es wird überprüft, ob die gesteckten Ziele erreicht wurden und ob die Modelle ausreichend genau und nützlich sind. Die Ergebnisse werden mit den Zielen abgeglichen und entsprechende Empfehlungen werden entwickelt.

6. **Bereitstellung (Deployment):**
   In dieser Phase werden die Ergebnisse in den operativen Betrieb überführt. Das bedeutet, dass die entwickelten Modelle, Erkenntnisse oder Visualisierungen in den Geschäftsprozess integriert werden, um einen Mehrwert zu generieren. Es werden Maßnahmen ergriffen, um sicherzustellen, dass die entwickelten Lösungen erfolgreich umgesetzt werden können.

Während des gesamten Prozesses gibt es Rückkopplungen und Iterationen zwischen den einzelnen Phasen, um die Ergebnisse zu verbessern oder Anpassungen vorzunehmen, wenn neue Erkenntnisse gewonnen werden. CRISP-DM ist flexibel und anpassungsfähig, um den unterschiedlichen Anforderungen und Gegebenheiten von Data-Mining-Projekten gerecht zu werden.

Die Lernaufgabe soll es den SuS ermöglichen im Kontext der vorgestellten Berufsbilder CRISP-DM in Teilen kennenzulernen. Im Anschluss an die Doppelstunde bietet sich an, dass die SuS eigenständige Projekte entlang von CRISP-DM durchführen. Hierbei ist bei der Auswahl der Daten darauf zu achten, dass mit relativ kleinen und gut strukturierten

Datensätzen begonnen wird. Mögliche Quellen für Daten finden sich am Ende unter **D3 Quellen / Lektüreliste zum Weiterlesen**.

Der in den Aufgaben angesprochene **Datensatz** enthält alle Überholvorgänge, die im Rahmen des Projekts “Radmesser” von 100 freiwilligen Fahrradfahrerinnen und Fahrradfahrern im Zeitraum (von 23.8.2018 bis 12.11.2018) in Berlin erfasst wurden. (Quelle: https://github.com/tagesspiegel/radmesser/tree/master/opendata)

- Aufbereitete Daten: RadmesserCut_distleft_hour_bezirk.csv
- Mit Fehlern: Radmesser_distleft_hour_bezirk.csv

Seaborn ist eine Pythonbibliothek zur Datenanalyse. Sie versucht etwas einfacher zu sein als Matplotlib. Seaborn benötigt folgende Bibliotheken:

- Matplotlib
- NumPy
- Pandas
- SciPy

B Lernaufgabe

B1 Unterrichtsbegleitende Präsentation

Was macht ein Data Scientist?

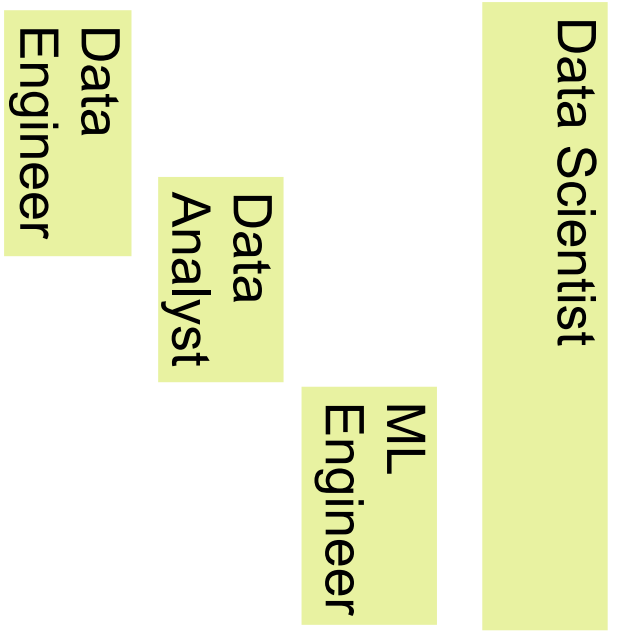
Domänenwissen

+

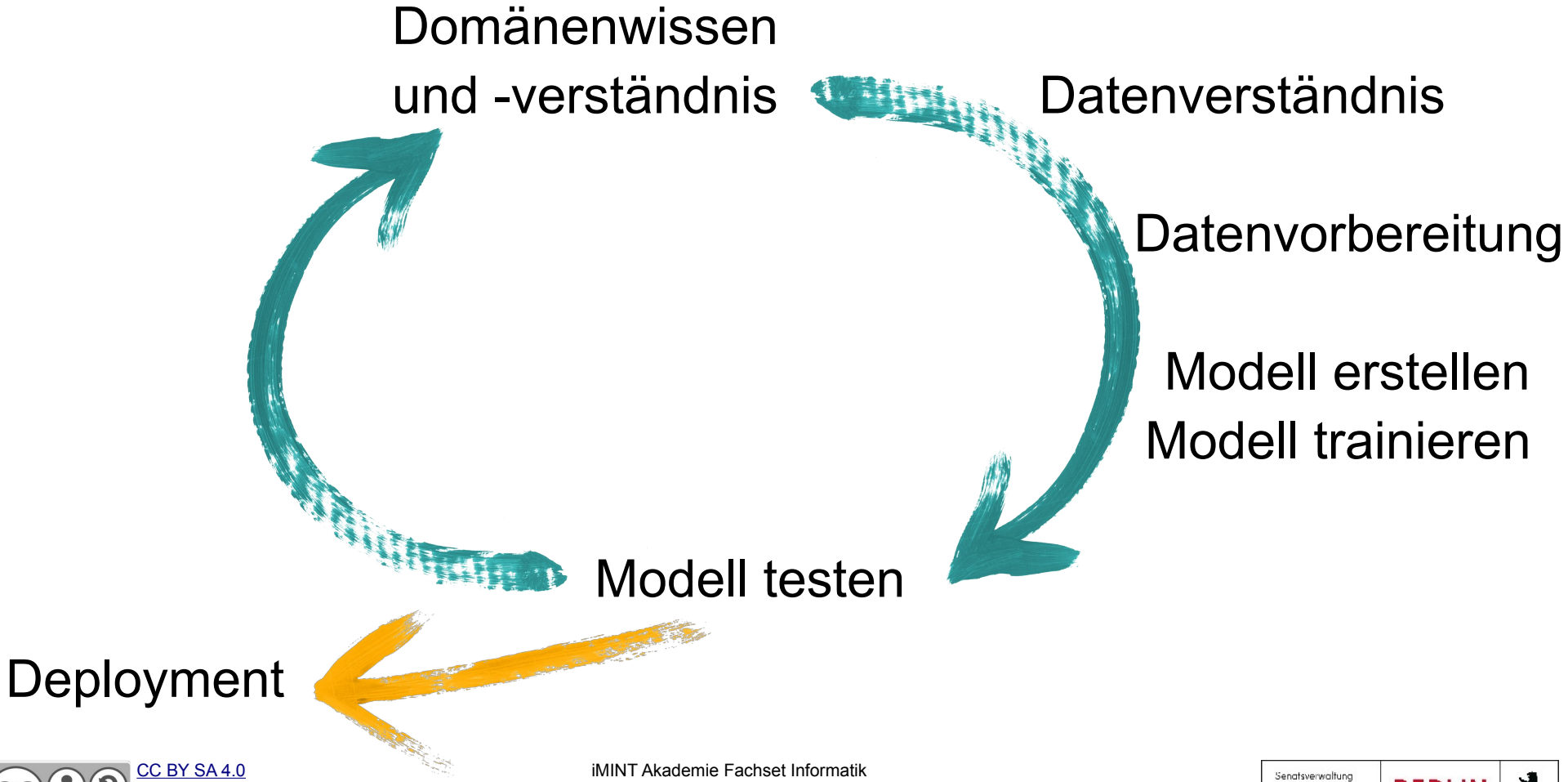
Kenne deine Datenstrukturen

3. Was macht ein Data Scientist?

- Daten sammeln / importieren
- bereinigen / normalisieren / hot encoding
- Datenanalyse / features auswählen
- `model` erstellen
- `model` deployment



3. Data Science Prozess CRISP-DM



4. Data Engineer

Ein Data Engineer ist für die Entwicklung, Wartung und Optimierung von Dateninfrastrukturen verantwortlich. Er ist an der Verarbeitung, Speicherung und Verteilung von großen Datenmengen beteiligt. Dazu gehört die Entwicklung von Datenpipelines, dabei weren die Daten von verschiedenen Quellen gesammelt und in eine zentrale Datenbank oder ein Data Warehouse geladen. Data Engineere arbeiten auch an der Optimierung der Leistung und Skalierbarkeit dieser Dateninfrastrukturen.

Data Engineere arbeiten eng mit Data Scientists und Data Analysten zusammen, um sicherzustellen, dass die Daten in einer Weise bereitgestellt werden, die es ihnen ermöglicht, ihre Analysen durchzuführen.

4. Data Analyst

Ein Data Analyst ist verantwortlich für die Untersuchung und Analyse von Daten, um Trends, Muster und Prognosen zu identifizieren. Sie arbeiten mit großen Datenmengen und verwenden Tools und Technologien wie SQL, Excel und statistische Software, um die Daten zu sammeln, zu bereinigen, zu analysieren und zu visualisieren. Sie können auch statistische Modelle und Machine-Learning-Algorithmen verwenden, um Prognosen und Vorhersagen zu erstellen. Data Analysten arbeiten häufig in Bereichen wie Finanzen, Marketing, Vertrieb und Einzelhandel. Sie arbeiten eng mit Data Scientists und anderen Abteilungen zusammen, um die Geschäftsentscheidungen zu unterstützen und die Leistung der Organisation zu verbessern.

4. ML Engineer

Ein ML Engineer (Machine Learning Engineer) ist verantwortlich für die Entwicklung, Implementierung und Wartung von ML-Modellen und ML-Systemen in einer Organisation. Sie arbeiten eng mit Data Scientists und anderen Experten zusammen, um die Anforderungen an das Modell zu definieren und das Modell zu entwickeln, zu trainieren und zu testen. Sie verwenden Tools und Technologien wie Python, TensorFlow, PyTorch und andere, um die Modellierung und die Implementierung der Modelle durchzuführen.

Sie sind auch verantwortlich dafür, das Modell in die Produktion zu bringen und sicherzustellen, dass es in Echtzeit zuverlässig und skalierbar ausgeführt wird. Sie überwachen die Leistung des Models und optimieren es. ML Engineer arbeiten in vielen Branchen, darunter Finanzen, E-Commerce, Gesundheitswesen, Autonome Systeme und viele mehr.

4. Data Scientist

Data Scientists sind Experten für Datenanalyse, sie sind verantwortlich für die Extraktion von Erkenntnissen aus großen Datenmengen. Sie arbeiten mit Methoden der Statistik, Mathematik und Informatik, um komplexe Probleme zu lösen und Prognosen zu erstellen. Sie sammeln, bereinigen und verarbeiten Daten aus verschiedenen Quellen, um sie zu analysieren und zu visualisieren. Dazu verwenden sie Machine-Learning-Algorithmen und statistische Modelle.
Data Scientists arbeiten in vielen Bereichen z.B. Finanzen, E-Commerce, Gesundheitswesen, Marketing, Verkehr, Maschinenbau und viele mehr. Sie arbeiten eng mit verschiedenen Bereichen zusammen, um Geschäftsentscheidungen zu unterstützen und die Leistung von Organisationen zu verbessern.

4. Fazit

- Vorstellung der einzelnen Berufe im Data Bereich
- Fazit

B2 Arbeitsblätter

1. Data Engineer

Ein Data Engineer ist für die Entwicklung, Wartung und Optimierung von Dateninfrastrukturen verantwortlich. Er ist an der Verarbeitung, Speicherung und Verteilung von großen Datenmengen beteiligt. Dazu gehört die Entwicklung von Datenpipelines, dabei weren die Daten von verschiedenen Quellen gesammelt und in eine zentrale Datenbank oder ein Data Warehouse geladen. Data Engineere arbeiten auch an der Optimierung der Leistung und Skalierbarkeit dieser Dateninfrastrukturen. Data Engineere arbeiten eng mit Data Scientists und Data Analysten zusammen, um sicherzustellen, dass die Daten in einer Weise bereitgestellt werden, die es ihnen ermöglicht, ihre Analysen durchzuführen.

Aufgaben:

1. Fasse obigen Text kurz zusammen und stelle Ihn Deinen Mitschülern vor.
2. Bereinige den Datensatz *Radmesser_distleft_hour_bezirk.csv* und beschreibe mögliche Fehlerquellen, erarbeite Lösungsmöglichkeiten.
   Im Datensatz sind Abstände bei Überholvorgänge (KFZ überholen Fahrräder) in Berliner Bezirken erfasst.

Weitere Leitfragen

3. Lese in der Keras-API die Bedeutung folgender Befehle nach:

```
tf.keras.layers.RandomFlip
tf.keras.layers.RandomCrop
tf.keras.layers.RandomRotation
```

2. Data Analyst

Data Analysten sind verantwortlich für die Untersuchung und Analyse von Daten, um Trends, Muster und Prognosen zu identifizieren. Sie arbeiten mit großen Datenmengen und verwenden Tools und Technologien wie SQL, Excel und statistische Software, um die Daten zu sammeln, zu bereinigen, zu analysieren und zu visualisieren. Sie können auch statistische Modelle und Machine-Learning-Algorithmen verwenden, um Prognosen und Vorhersagen zu erstellen. Data Analysten arbeiten häufig in Bereichen wie Finanzen, Marketing, Vertrieb und Einzelhandel. Sie arbeiten eng mit Data Scientists und anderen Abteilungen zusammen, um die Geschäftsentscheidungen zu unterstützen und die Leistung der Organisation zu verbessern.

Aufgaben:

1. Fasse obigen Text kurz zusammen und stelle Ihn Deinen Mitschülern vor.
2. Ermittle welche Auffälligkeiten in folgendem Datensatz *RadmesserCut_distleft_hour_bezirk.csv* stecken. Stelle die Daten mit Hilfe von Seaborn dar.
   Im Datensatz sind Abstände bei Überholvorgänge (KFZ überholen Fahrräder) in Berliner Bezirken erfasst.

```
import seaborn
import pandas
import matplotlib.pyplot as plt

df = pandas.read_csv("RadmesserCut_distleft_hour_bezirk.csv")
resHour = seaborn.scatterplot(x="hour", y="distleft", data=df)
plt.show()
```

Weitere Leitfragen:

3. Visualisiere weiter Daten aus dem Datensatz um noch mehr Aussagen treffen zu können. Z.B. zu einzelnen Bezirken usw.

3. Machine Learning Engineer

Ein ML Engineer (Machine Learning Engineer) ist verantwortlich für die Entwicklung, Implementierung und Wartung von ML-Modellen und ML-Systemen in einer Organisation. Sie arbeiten eng mit Data Scientists und anderen Experten zusammen, um die Anforderungen an das Modell zu definieren und das Modell zu entwickeln, zu trainieren und zu testen. Sie verwenden Tools und Technologien wie Python, TensorFlow, PyTorch und andere, um die Modellierung und die Implementierung der Modelle durchzuführen.

Sie sind auch verantwortlich dafür, das Modell in die Produktion zu bringen und sicherzustellen, dass es in Echtzeit zuverlässig und skalierbar ausgeführt wird. Sie überwachen die Leistung des Models und optimieren es. ML Engineer arbeiten in vielen Branchen, darunter Finanzen, E-Commerce, Gesundheitswesen, Autonome Systeme und viele mehr.

Aufgaben:

1. Fasse obigen Text kurz zusammen und stelle Ihn Deinen Mitschülern vor.
2. Erstelle mit Hilfe von TensorFlow ein eigenes NN mit mehreren Schichten.

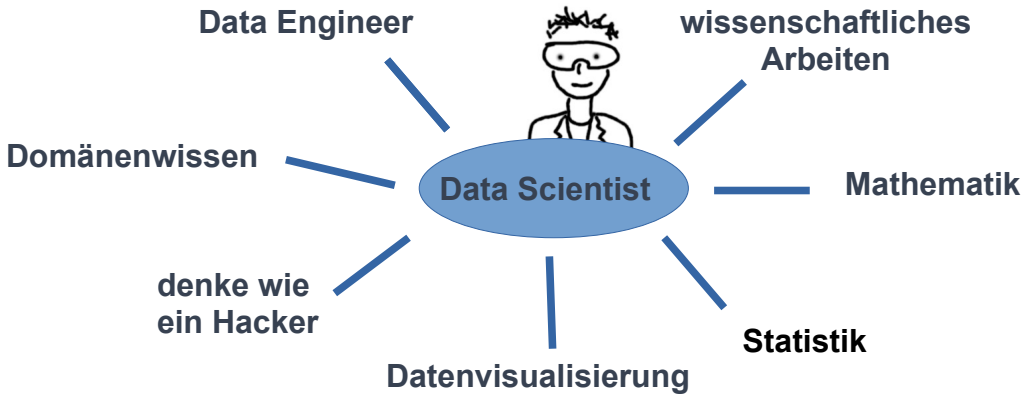
Weitere Leitfragen:

3. Lass Dir anzeigen wie das NN aufgebaut ist, und wie sich die Gewichte verändern. Setze dazu folgende Befehle ein:

```
model.summary()
print (model.get_weights())
```

4. Data Scientist

Data Scientists sind Experten für Datenanalyse, sie sind verantwortlich für die Extraktion von Erkenntnissen aus großen Datenmengen. Sie arbeiten mit Methoden der Statistik, Mathematik und Informatik, um komplexe Probleme zu lösen und Prognosen zu erstellen. Sie sammeln, bereinigen und verarbeiten Daten aus verschiedenen Quellen, um sie zu analysieren und zu visualisieren. Dazu verwenden sie Machine-Learning-Algorithmen und statistische Modelle.
Data Scientists arbeiten in vielen Bereichen z.B. Finanzen, E-Commerce, Gesundheitswesen, Marketing, Verkehr, Maschinenbau und viele mehr. Sie arbeiten eng mit verschiedenen Bereichen zusammen, um Geschäftsentscheidungen zu unterstützen und die Leistung von Organisationen zu verbessern.



[B1]

Abb. 1: Data Science ist eine Mischung aus verschiedenen Fähigkeiten.

Aufgaben:

1. Fasse obigen Text kurz zusammen und stelle Ihn Deinen Mitschülern vor.
2. Stelle zwei der Fähigkeiten aus Abb. 1 vor und erkläre warum sie für einen Data Scientist besonders wichtig sind.

Weitere Leitfragen

3. Ordne den Schritten im CRISP-DM die unterschiedlichen Fähigkeiten aus Abb. 1 zu.

Abb.1 [B1]: „DataScientist“, Alexander Schindler, Lizenz CC BY-SA4.0, „KI programmieren im Informatikunterricht Teil 7: Data Scientist“

C Bezug zum Rahmenlehrplan

C1 Didaktischer Kommentar

Ziel der Aufgabe ist es, den Schülern einen Einstieg in die Berufswelt im KI-Bereich zu geben.

Einige Vorerfahrungen mit NN, python, Softwareentwicklung sollten vorhanden sein, insbesondere die Inhalte der Lernaufgabe: „KI programmieren im Informatikunterricht Teil I: Einführung“.
Die Lernaufgabe ist skalierbar, d.h. sie kann um eigene Ideen (z.B. Datensätze, Problemstellungen) erweitert werden und sie kann und soll somit Basis für eigene Projekte sein. Nicht zuletzt ist mit der Thematisierung von KI ein Ausblick auf die Auswirkungen von KI auf die eigene Lebenswelt und zukünftige Berufswelt der SuS möglich und somit ist die Aufgabe auch im Bereich Informatik und Gesellschaft zu verorten.

Lernervoraus-setzungen	• Die SuS können einfache Algorithmen in Python programmieren. • Die SuS können mit Variablen umgehen. • Die SuS haben eine grundlegende Vorstellung von Algorithmen und Datenstrukturen. • Die SuS können Bibliotheken in Python einbinden • Die SuS können eine IDE benutzen. • Die SuS haben ein grundsätzliches Verständnis für ML wie es in der Lernaufgabe: „KI programmieren im Informatikunterricht Teil I: Einführung“ entwickelt wurde.

C2 Bezüge zum Rahmenlehrplan Informatik

Kompetenzen	Standards (Die Schülerinnen und Schüler können….)
Mit Fachwissen umgehen	Bezug zum RLP Sek I: **Kompetenzbereich**: Informatisches Modellieren **Kompetenz**: Informatische Modelle analysieren und bilden **Standard F**: informatische Modelle als reduzierte Abbildung der realen Welt beschreiben und beurteilen **Standard G**: ein Modell selbst erstellen

C Bezug zum Rahmenlehrplan

Erkenntnisse gewinnen	Die SuS können Modelle selbst erstellen und mit NN umgehen.
Kommunizieren	Bezug zum RLP Sek I: **Kompetenzbereich**: Kommunizieren und Kooperieren – Teamarbeit organisieren und koordinieren **Kompetenz**: Arbeitsergebnisse dokumentieren und präsentieren **Standard G**: adressatengerecht mit Softwareunterstützung präsentieren
Bewerten	Bezug zum RLP Sek I: **Kompetenzbereich**: Informatisches Modellieren **Kompetenz**: Informatische Modelle analysieren und bilden **Standard F**: informatische Modelle als reduzierte Abbildung der realen Welt beschreiben und beurteilen **Standard H**: beurteilen, ob das selbst erstelle Modell problemadäquat ist

Unterrichtsfach	Informatik
Jahrgangsstufe/n	Sek I: 10 Sek II: IN-3
Niveaustufe/n	**Bezug zum RLP Sek I:** **Kompetenzbereich**: Informatisches Modellieren **Kompetenz**: Informatische Modelle analysieren und bilden **Standard F**: informatische Modelle als reduzierte Abbildung der realen Welt beschreiben und beurteilen **Standard G**: ein Modell selbst erstellen **Standard H**: beurteilen, ob das selbst erstelle Modell problemadäquat ist **Bezug zum RLP Sek II:** 3. Kurshalbjahr (**IN-3**) Grundlagen der Informatik und Vertiefungsgebiet: V5 Künstliche Intelligenz
Zeitrahmen	**Eine Doppelstunde:** Erarbeitung und Vorstellung in Gruppen

	anschließend sind Fortsetzungen möglich.
Thema	Berufsbilder im KI-Bereich

Kontext	• Beruf und Arbeitswelt • Gesellschaft
Schlagwörter	Programmierung, Informatik und Gesellschaft, Informatiksystem verstehen, Problemlösen, mit Informationen umgehen, Modellbildung, künstliche Intelligenz, KI, Deep Learning, DL, Python, TensorFlow, Keras, Seaborn, Daten, Datenvisualisierung

C3 Bezüge zum Basiscurriculum Medienbildung[1]

Standards des BC Medienbildung	Die Schülerinnen und Schüler können …
Präsentieren	• Arbeitsergebnisse vorstellen • Reflektierter Technologieeinsatz

C4 Bezüge zu übergreifenden Themen[2]

Berufs-und Studienorientierung	Data Engineer, Data Analyst, Data Scientist, Machine Learning Engineer, Softwareentwickler

1 vgl. Rahmenlehrplan Jahrgangsstufen 1-10, Teil B, S. 15-22, Berlin, Potsdam 2015

2 vgl. Rahmenlehrplan Jahrgangsstufen 1-10, Teil B, S. 24ff, Berlin, Potsdam 2015

D Anhang

D1 „Material“ für den Einsatz dieser Lernaufgabe

- Für SchülerInnen eine entsprechend leistungsfähige Pyhton IDE es empfiehlt sich Spyder
- Python (derzeit Januar 2022 ab Version 3.8, 64 Bit) mit installierter TensorFlow (ab 2.6) Bibliothek und seaborn Bibliothek
- Da die jeweils aktuelle TensorFlowbibliothek in einer nicht aktuellen Pythonversion entwickelt wird, hilft es bei (unklaren) Problemen beim compilieren eine eher ältere Pythonversion zu verwenden.
- Instatllierte Seaborn Bibliothek. Seaborn benötigt: Matplotlib, Numpy, Pandas, SciPy.

D2 Hinweise / Musterlösung der Lernaufgabe

1. Data Engineer

1. Fasse obigen Text kurz zusammen und stelle Ihn Deine Mitschülern vor.
   Erwartungshorizont (beispielhaft)**:**
   Ein Data Engineer entwickelt und optimiert Dateninfrastrukturen, um große Datenmengen zu verarbeiten, speichern und verteilen. Sie erstellen Datenpipelines, um Daten von verschiedenen Quellen zu sammeln um sie z.B. in eine Datenbank zu laden. Data Engineers arbeiten eng mit Data Scientists und Analysten zusammen.

2. Analysiere folgenden Datensatz *Radmesser_distleft_hour_bezirk.csv* und beschreibe mögliche Fehlerquellen, erarbeite Lösungsmöglichkeiten.
   Erwartungshorizont (beispielhaft):
   Ansicht der .csv-Datei z.B. über LibreOffice Calc.
   - Spalte *distleft* nicht einheitlich:
     - eine Zelle ohne Wert
     - eine Zelle mit Einheit mm
     - zwei fehlerhafte große Werte, Einheit unklar.

3. Lese in der Keras-API die Bedeutung folgender Befehle nach:
   ```
   tf.keras.layers.RandomFlip
   ```

```
tf.keras.layers.RandomCrop
tf.keras.layers.RandomRotation
```

Erwartungshorizont (beispielhaft)**:**
In dieser Aufgabe geht es vor allem darum, dass die SuS erkenn, dass aus wenigen Inputdaten zahlreiche neue Daten generiert werden können. Und zwar durch

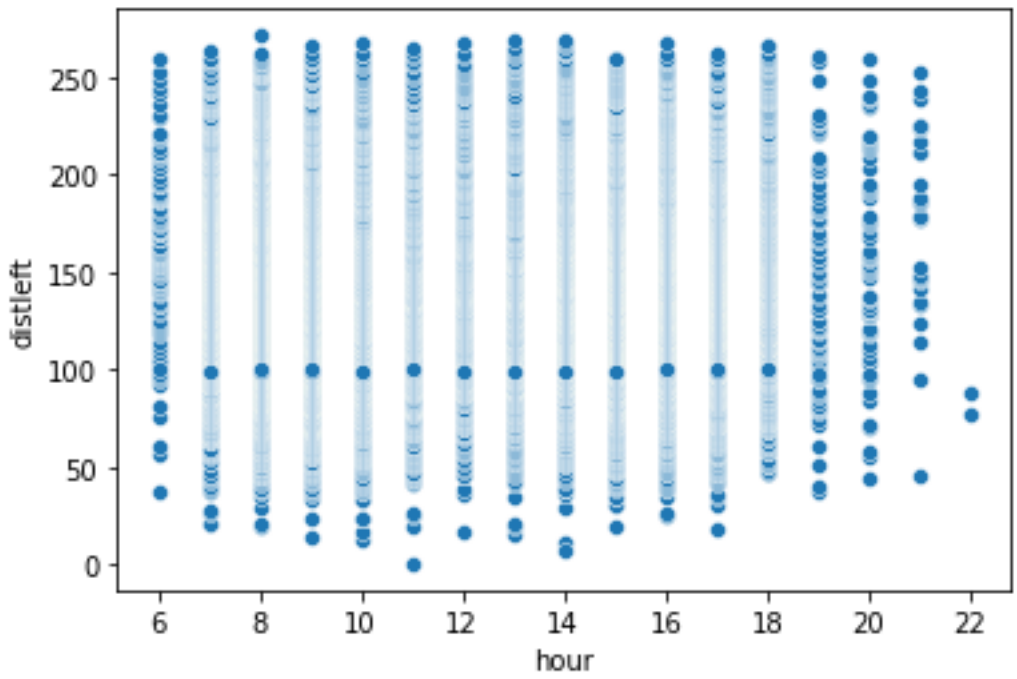
- RandomFlip: zufälliges horizontales oder vertikales spiegeln https://keras.io/api/layers/preprocessing_layers/image_augmentation/random_flip/
- RandomCrop: zufälliges ausschneiden https://keras.io/api/layers/preprocessing_layers/image_augmentation/random_crop/
- RandomRotation: drehen um einen zufälligen Winkel https://keras.io/api/layers/preprocessing_layers/image_augmentation/random_rotation/

Eine mögliche Anwendung ergibt sich im Zusammenfassung mit der Lernaufgabe: „KI programmieren im Informatikunterricht Teil 2: Bilderkennung“
Keras API: https://keras.io/guides/preprocessing_layers/

2. Data Analyst

1. Fasse obigen Text kurz zusammen und stelle Ihn Deine Mitschülern vor.
   Erwartungshorizont (beispielhaft)**:**
   Data Analysten analysieren Daten, identifizieren Trends und Muster und unterstützen Geschäftsentscheidungen. Sie verwenden Tools wie SQL und Excel, um Daten zu sammeln, bereinigen, analysieren und visualisieren. Data Analysten arbeiten in verschiedenen Bereichen und kooperieren mit Data Scientists und anderen Abteilungen.

2. Ermittle welche Auffälligkeiten in folgendem Datensatz *RadmesserCut_distleft_hour_bezirk.csv* stecken. Stelle die Daten mit Hilfe von seaborn dar.
   Erwartungshorizont (beispielhaft)**:**

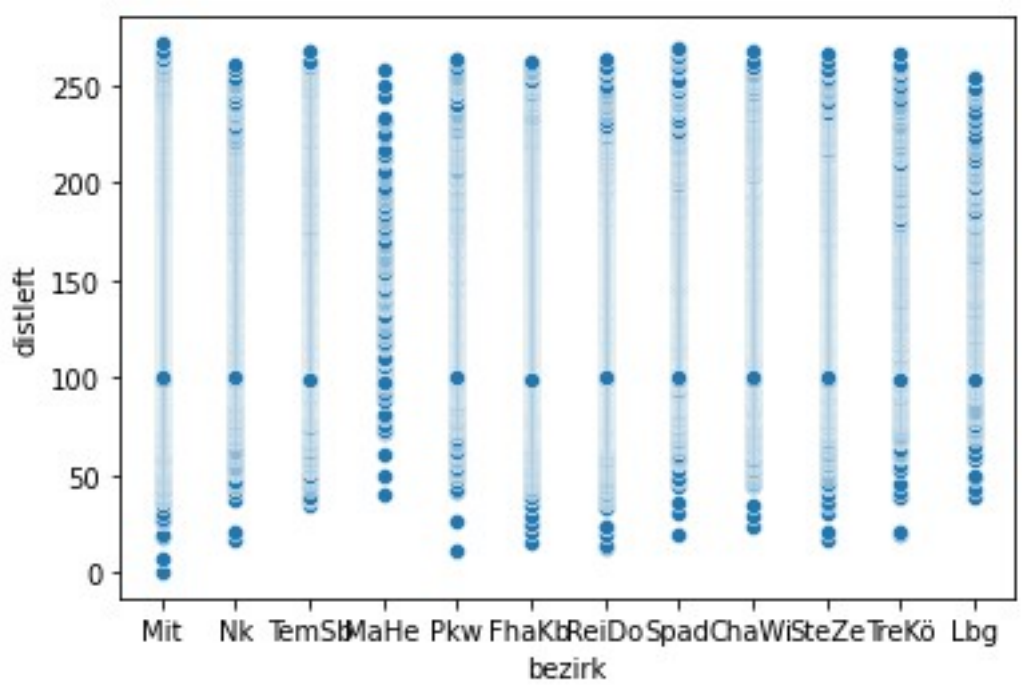


Kürzere Distanzen beim Überholen finden sich eher zu den typischen Zeiten des Berufsverkehrs. Viele Überholvorgänge unterschreiten den gesetzlichen Mindesabstand von innerorts 150cm.

3. Weitere Darstellungsmöglichkeiten: https://seaborn.pydata.org/examples/
   z.B.:

```
import seaborn
import pandas
import matplotlib.pyplot as plt

csvBezirk = pandas.read_csv
("RadmesserCut_distleft_hour_bezirk.csv")
resBezirk = seaborn.scatterplot(x="bezirk", y="distleft",
data=csvBezirk)
plt.show()
```



Mögliche Aussagen:

- In Mitte wird am dichtesten überholt,
- in Marzahn-Hellersdorf mit dem größten Abstand.

3. ML Engineer

1. Fasse obigen Text kurz zusammen und stelle Ihn Deine Mitschülern vor.
   Erwartungshorizont (beispielhaft)**:**
   Ein ML Engineer entwickelt und implementiert ML-Modelle und ML-Systeme. Sie arbeiten mit Data Scientists zusammen, um Anforderungen zu definieren und Modelle zu entwickeln, zu trainieren und zu testen. ML Engineers nutzen Tools wie Python, TensorFlow und PyTorch. Sie bringen die Modelle in die Produktion, überwachen die Leistung und optimieren sie. ML Engineers arbeiten in viele verschiedenen Branchen.

2. Erstelle mit Hilfe von TensorFlow ein eigenes NN mit mehreren Schichten.
   Erwartungshorizont (beispielhaft)**:**
   siehe Lernaufgabe: „KI programmieren im Informatikunterricht Teil 2: Bilderkennung“

3. Lass Dir anzeigen wie das NN aufgebaut ist, und wie sich die Gewichte verändern. Setze dazu folgende Befehle ein...]

Erwartungshorizont (beispielhaft)**:**
`model.summary()`: Gibt eine Zusammenfassung des neuronalen Netzwerkes mit einer Aufzählung der einzelnen Layer. Es ermöglicht eine gute Übersichtl über das *model*. Hier am Beispiel des in derLernaufgabe: „KI programmieren im Informatikunterricht Teil 2: Bilderkennung“ eingeführten NN. Es Lässt sich aber beliebig auf andere mit TensorFlow erstellten NN übertragen

```
Layer (type)                 Output Shape              Param#
=================================================================
rescaling_2 (Rescaling)      (None, 200, 200, 3)       0
conv2d_6 (Conv2D)            (None, 200, 200, 16)      448
max_pooling2d_6 (MaxPooling2D)  (None, 100, 100, 16)   0
conv2d_7 (Conv2D)            (None, 100, 100, 32)      4640
max_pooling2d_7 (MaxPooling2)   (None, 50, 50, 32)     0
conv2d_8 (Conv2D)            (None, 50, 50, 64)        18496
max_pooling2d_8 (MaxPooling2D)  (None, 25, 25, 64)     0
flatten_2 (Flatten)          (None, 40000)             0
dense_4 (Dense)              (None, 128)               5120128
dense_5 (Dense)              (None, 5)                 645
=================================================================
Total params: 5,144,357
Trainable params: 5,144,357
Non-trainable params: 0
```

Die Anzal der Parameter (`Total params`) zeigt wie groß das *model* ist. So lässt sich abschätzen ob das *model* oder einzelne Layer evtl. zu groß/zu klein sind.

`model.get_weights()`: Zeigt die aktuellen Gewichte des Models an. Usw. Hier lassen sich keine weitergehende Aussagen treffen außer: Es sind viele und die Änderung der Werte ist nur schwer nachvollziehbar.

4. Data Scientist

1. Fasse obigen Text kurz zusammen und stelle Ihn Deine Mitschülern vor.
   Erwartungshorizont (beispielhaft)**:**
   Data Scientists sind Experten für Datenanalyse und Prognosen. Sie verwenden Statistik, Mathematik und Informatik, um komplexe Probleme zu lösen. Data Scientists sammeln, bereinigen und analysieren Daten aus verschiedenen Quellen und arbeiten in Bereichen wie Finanzen, E-Commerce und

Gesundheitswesen. Ihr Ziel ist es, Geschäftsentscheidungen zu unterstützen und die Leistung von Organisationen zu verbessern.

2. Stelle zwei der Disziplinen aus der Abbildung vor und erkläre warum sie für einen Data Scientist besonders wichtig sind.
   Erwartungshorizont (beispielhaft):
   - denke wie ein Hacker: im Laufe eines Zyklus des CRISP-DM Prozesses können zahlreiche Probleme auftauchen (z.B. zu lange Trainingszeit des models) diese müssen oft kreativ und unkonventionell gelöst werden. Z.B.: Kann man evtl einen Teil der Daten weglassen? Kann die Auflösung von Bildern reduziert werden usw.
   - Domänenwissen: Ergebnisse müssen immer wieder auf ihre plausibilität überprft werden. Dies geht nicht ohne die entsprechende fachliche Kenntnis.

3. Ordne den unterschiedlichen Fähigkeiten Schritte im CRISP-DM zu.
   Erwartungshorizont (beispielhaft)**:**
   - *Domänenwissen*: Data Engineer, wissenschaftliches Arbeiten, Domänenwissen
   - *Datenverständnis*: Data Engineer, wissenschaftliches Arbeiten, Mathematik, Statistik, Datenvisualisierung, Domänenwissen
   - *Datenvorbereitung*: Data Engineer, Mathematik, Statistik, Datenvisualisierung
   - *Modell erstellen und trainieren*: wissenschaftliches Arbeiten, Datenvisualisierung, denke wie ein Hacker, Domänenwissen
   - *Modell testen*: wissenschaftliches Arbeiten, Mathematik, Statistik, Datenvisualisierung, denke wie ein Hacker, Domänenwissen
   - *Deployment*: denke wie ein Hacker, Domänenwissen

D3 Quellen / Lektüreliste zum Weiterlesen

- MORONEY, Laurence[2020]: AI and Machine Learning for Coders, A Programmer‘s Guide to Artificial Intelligence.
- Seaborn Python Bibliothek zur Datenanalyse https://seaborn.pydata.org/
- Seaborn Beispiele https://seaborn.pydata.org/examples/
- TensorFlow API: www.tensorflow.org/api_docs/python/tf/all_symbols
- Keras API: https://keras.io/guides/preprocessing_layers/

Online verfügbare Daten:

- für die Aufgaben: https://github.com/tagesspiegel/radmesser/tree/master/opendata
- Kaggle Datasets: www.kaggle.com/datasets
- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
- European Data Portal: www.europeandataportal.eu
- Data.gov: www.data.gov
- Open Data Portal Deutschland: www.govdata.de
- **Lizenztext für die Daten der Überholvorgänge:** Radmesser Open Data License: https://github.com/tagesspiegel/radmesser/blob/master/opendata/LICENSE.md
  Die Datensätze werden unter der Open Data Commons Attribution License (ODC-By) v1.0 zur Verfügung gestellt: https://www.opendatacommons.org/licenses/by/1.0/
  Bedingungen sind: Die Nennung des Ursprungsprojekts "Tagesspiegel Radmesser" und ein Link zum Originalprojekt http://radmesser.de.
  Die korrekte Attribution bei allen daraus entstehenden Veröffentlichungen ist also: ODC-By v1.0/Tagesspiegel Radmesser/http://radmesser.de

D4 Bildnachweise

Bildtitel	*Seite*	*Quelle*
[B0]	1	„DataScientistSammeltModelliertTrainiert“, Alexander Schindler, Lizenz CC BY-SA4.0, „KI programmieren im Informatikunterricht Teil 7: Data Scientist“ unter Verwendung weiterer Quellen [B2], [B3], [B4] s. Bildnachweis
[B1]	20	„DataScientist“, Alexander Schindler, Lizenz CC BY-SA4.0, „KI programmieren im Informatikunterricht Teil 7: Data Scientist“
[B2]	1	"Winter Wolf iPhone wallpaper", xploitme, Lizenz CC BY-SA 2.0, Abgerufen: 7.01.2021 https://www.flickr.com/photos/45928872@N08/4272568627
[B3]	1	"Mi perra Heura observando los patos - Dog Heura observing ducks", ferran pestaña, Lizenz CC BY-SA 2.0, verändert (zugeschnitten) Abgerufen: 7.01.2021 www.flickr.com/photos/57956171@N00/249920052
[B4]	1	"fat cat", cuatrok77, Lizenz CC BY-SA 2.0. Abgerufen: 15.01.2020 https://www.flickr.com/photos/69573851@N06/7158001813