Machine Learning – Datenanalyse

Aufgabenblatt

Datenqualität

Lern-Story: Bevor man Daten analysiert, muss man wissen, wie gut sie sind. *Datenqualität* hat eine *objektive* Seite — messbare Genauigkeit, Vollständigkeit und Konsistenz — und eine *subjektive* Seite: ob die Daten für den konkreten Zweck eines Nutzers brauchbar und vertrauenswürdig sind. In diesem Teil beurteilen Sie die Qualität eines Datensatzes anhand beider Dimensionen.
Lernziel: Nach Abschluss dieses Teils kann der Absolvent:in die Datenqualität der Daten bestimmen.
Input:

- Datenqualität
- Code Notebooks

Output:

- Die Datenqualität vom Lego Datensatz bestimmen

Lernerfolg: Bestimmen Sie ihren Lernerfolg.

Aufgabe 1: Wiederholung

Fassen Sie in eigenen Worten zusammen, was Sie in diesem Teil gelernt haben. Was unterscheidet objektive von subjektiver Datenqualität, und welche Dimensionen (z. B. Genauigkeit, Vollständigkeit, Konsistenz, Relevanz) gehören dazu?

Antwort

Aufgabe 2: Single Choice Fragen

1. Was beschreibt die “objektive" Datenqualität?

   ☐ Die persönliche Meinung des Datennutzers über die Daten.

   ☐ Die messbare Genauigkeit, Vollständigkeit und Konsistenz der Daten.

   ☐ Die Möglichkeit, die Daten für unterschiedliche Zwecke zu interpretieren.

   ☐ Die Bewertung der Daten durch Experten in einem bestimmten Bereich.

2. Welche der folgenden Aspekte betrifft die “subjektive" Datenqualität?

   ☐ Die Reproduzierbarkeit von Ergebnissen.

   ☐ Die Meinung der Nutzer, ob die Daten für ihre Zwecke geeignet sind.

   ☐ Die Anzahl fehlender Datenpunkte.

   ☐ Die Genauigkeit der Messungen.

3. Wie kann die “Vollständigkeit" von Daten in Bezug auf die objektive Datenqualität am besten beschrieben werden?

   ☐ Die Genauigkeit der erfassten Daten.

   ☐ Der Grad, in dem alle erwarteten Daten verfügbar sind.

   ☐ Die Fähigkeit, die Daten ohne Fehler zu interpretieren.

   ☐ Die Subjektivität der Datenbewertung durch den Nutzer.

4. Welche der folgenden Aussagen beschreibt am besten den subjektiven Aspekt der Datenqualität?

   ☐ Daten sind konsistent, wenn sie in allen Datensätzen die gleichen Werte haben.

   ☐ Die Daten wurden präzise und ohne Fehler gemessen.

   ☐ Der Nutzer empfindet die Daten als nützlich und vertrauenswürdig für seinen spezifischen Zweck.

   ☐ Alle Datenfelder enthalten die erforderlichen Informationen.

5. Ein Nutzer beschwert sich, dass die Daten für seine Analyse unbrauchbar sind, obwohl sie technisch korrekt sind. Welcher Aspekt der Datenqualität wird hier kritisiert?

   ☐ Genauigkeit

   ☐ Subjektive Qualität

   ☐ Vollständigkeit

   ☐ Objektive Qualität

6. Welche der folgenden Aussagen ist ein Beispiel für eine objektive Datenqualitätsprüfung?

   ☐ “Die Daten sind für die Analyse meiner Forschung geeignet."

   ☐ “Ich habe das Gefühl, dass die Daten nützlich sind."

   ☐ “Ich kann die Daten in meinem speziellen Kontext gut nutzen."

   ☐ “Die Daten enthalten keine fehlerhaften Werte oder Widersprüche."

7. Wie kann die Genauigkeit von Daten als Teil der objektiven Datenqualität bewertet werden?

   ☐ Durch das Vertrauen des Nutzers in die Daten.

   ☐ Durch den Vergleich der Daten mit einer bekannten, wahren Referenz.

   ☐ Durch die Einschätzung der Nützlichkeit für den aktuellen Anwendungszweck.

   ☐ Durch den subjektiven Eindruck, den die Daten vermitteln.

8. Welcher der folgenden Begriffe bezieht sich auf die subjektive Einschätzung der Datenqualität?

   - ☐ Konsistenz
   - ☐ Genauigkeit
   - ☐ Verlässlichkeit
   - ☐ Relevanz

9. Ein Datensatz enthält alle notwendigen Datenpunkte, aber einige der Daten sind falsch oder ungenau. Welche Dimension der Datenqualität wird hier beeinträchtigt?

   - ☐ Vollständigkeit
   - ☐ Konsistenz
   - ☐ Genauigkeit
   - ☐ Objektivität

10. Ein Datensatz wird von zwei unterschiedlichen Personen verwendet, aber beide bewerten ihn unterschiedlich hinsichtlich seiner Nützlichkeit. Dies ist ein Beispiel für:

    - ☐ Subjektive Datenqualität
    - ☐ Objektive Datenqualität
    - ☐ Vollständigkeit
    - ☐ Konsistenz

11. Welche der folgenden Aussagen beschreibt die Konsistenz von Daten?

    - ☐ Alle Datenpunkte stimmen mit einem externen Standard überein.
    - ☐ Die Daten sind für alle Nutzer in ihrem spezifischen Anwendungsfall geeignet.
    - ☐ Die Daten sind innerhalb des Datensatzes einheitlich und enthalten keine widersprüchlichen Informationen.
    - ☐ Der Datensatz enthält alle erwarteten Informationen.

12. Welche der folgenden Massnahmen kann zur Verbesserung der subjektiven Datenqualität beitragen?

    - ☐ Sicherstellen, dass die Daten korrekt und konsistent sind.
    - ☐ Erheben der Daten in einem standardisierten Format.
    - ☐ Einholen von Nutzerfeedback, um zu verstehen, wie die Daten für ihre Zwecke besser nutzbar
    - ☐ Vermeidung von fehlenden Datenpunkten.

13. Welcher der folgenden Begriffe gehört zur objektiven Datenqualität?

    - ☐ Nutzerzufriedenheit
    - ☐ Relevanz
    - ☐ Genauigkeit
    - ☐ Zweckmässigkeit

14. Wenn ein Datensatz vollständig und korrekt ist, aber von einem Nutzer als irrelevant empfunden wird, welches Datenqualitätsproblem tritt auf?

    ☐ Ein objektives Problem mit der Vollständigkeit

    ☐ Ein subjektives Problem mit der Relevanz

    ☐ Ein objektives Problem mit der Genauigkeit

    ☐ Ein subjektives Problem mit der Konsistenz

Aufgabe 3: Datenqualität

Bestimmen Sie für den Lego-Datensatz `Data/lego/lego_daten_tidy.csv` die Datenqualität. Beschreiben Sie die Datenqualität

- objektiv (Vollständigkeit, Konsistenz, Genauigkeit) und
- subjektiv (Eignung/Relevanz für einen konkreten Zweck).

Orientieren Sie sich am Code-Notebook `Data_Quality.ipynb` (`Codevorlagen/`).

Aufgabe 4: Tidy Data

1. Beschreiben Sie das Konzept von Tidy Data.
2. Ist der gegebene Lego Datensatz Tidy?