Grundlagen der Statistik – Projektarbeit – Übungsbeispiel WS 2022/23

Betreuung: K. Horneck

Projekt 1: Versicherungskosten

Eine Krankenversicherungsgesellschaft muss über die Prämien Geld für die medizinische Versorgung ihrer Kunden zur Verfügung stellen. Daher ist die Entwicklung von Modellen, die die medizinischen Ausgaben genau prognostizieren, sehr wichtig.
Die medizinischen Kosten sind schwer abzuschätzen, da die teuersten Behandlungen selten und scheinbar zufällig sind. Dennoch sind einige Krankheiten für bestimmte Bevölkerungsgruppen häufiger. Zum Beispiel ist Lungenkrebs bei Rauchern wahrscheinlicher als bei Nichtrauchern, und Herzerkrankungen können bei Fettleibigen wahrscheinlicher sein.
Anhand von Patientendaten werden die durchschnittlichen Kosten für die medizinische Versorgung dieser Bevölkerungsgruppen geschätzt. Diese Schätzungen könnten verwendet werden, um versicherungsmathematische Tabellen zu erstellen, in denen der Preis für jährliche Prämien festgelegt wird.

Daten: `insurance.csv`

Variablen:

- `age`: Alter des Versicherungsnehmers
- `sex`:
- `bmi`: Bodymassindex
- `children`: Anzahl der Kinder / von der Versicherung abgedeckte Angehörige.
- `smoker`: (`yes`|`no`)
- `region`: Wohnregion des Versicherten: `northeast`, `southeast`, `southwest`, or `northwest`.
- `charges`: medizinischen Kosten, die für das Kalenderjahr in Rechnung gestellt wurden

Projekt 2: Fischgewicht

Fische, die vor der Kamera schwimmen, können anhand des Videobildes gemessen und das Gewicht der Fische mithilfe des linearen Regressionsmodells geschätzt werden. So ist kein Wiegen des Fisches notwendig, was zB als Anwendung in Fischfarmen sinnvoll ist.

Daten: `Fish.csv`

Variablen:

- `Species: Fischart`
- `Length1: Vertikale Länge in cm`
- `Length2: Diagonale Länge in cm`
- `Length3: Kreuzlänge in cm`
- `Height: Höhe in cm`
- `Width: Breite in cm`
- `Weight: Gewicht des Fisches in Gramm`

Projekt 3: NO2

Stickstoff-Dioxyd (NO_2) ist ein Luftschadstoff, der vom Verkehr, genauer von Verbrennungsmotoren, ausgestoßen wird. Es ist ein Reizgas, das die Schleimhäute angreift. Seine Konzentration wird, zusammen mit anderen Schadstoffen und Wettergrößen kontinuierlich von automatischen Messstationen erfasst. Daten für die Schweizer Stationen erhält man von www.bafu.admin.ch/bafu/de/home/themen/luft/zustand/daten/datenabfrage-nabel.html

Um gezielte Maßnahmen zur Verringerung der Belastung zu planen, ist es nützlich, die Einflüsse auf die Schadstoff-Konzentration zu kennen.

Daten: `NO2_ZUE.csv`

Variablen:

- `Datum_Zeit`
- `NO2` (Stickstoffdioxid)
- `PM2.5` (Feinstaub)
- `TEMP` (Temperatur)
- `PREC` (Niederschlag)
- `RAIN` (Regen JA/NEIN) [diese Variable wurde aus der Variable „Niederschlag“ erzeugt]

Projekt 4: Prüfungsleistung von Studierenden

Dies ist eine Untersuchung der Faktoren, die die Leistung der Studierenden bei Prüfungen auf Universitätsniveau beeinflussen. Dies ist ein fiktiver Datensatz und sollte nur für datenwissenschaftliche Schulungszwecke verwendet werden. Dieser Datensatz enthält Ergebnisse aus drei Prüfungen und persönliche, soziale und wirtschaftliche Faktoren, die sich auf die Interaktion auswirken.

Data: `StudentsPerformance.csv`

Variablen:

a. `Gender`
b. `race.ethnicity:` Gruppenzugehörigkeit
c. `parental.level.of.educatio`
d. `lunch:` Die Universität stellt kostenfreies bzw kostenreduziertes Essen zur Verfügung oder das Essen hat einen Standartpreis
e. `test.preparation.course` es wurde ein Vorbereitungskurs besucht
f. `math.score:` erreichte Punkte in Mathematik
g. `reading.score`
h. `writing.score`