Liebe Studierende,

im 4. Semester steht das Data Exploration Project an. Da Sie es offensichtlich kaum erwarten können, loszulegen, habe ich mir etwas Zeit genommen, die Rahmenbedingungen auszuarbeiten:

Modulbeschreibung
- Präsenzzeit: 27 h
- Selbststudium: 47 h
- Anwendung von Methoden und Verfahren des Maschinellen Lernens auf eine vorgegebene Datenbasis unter Laborbedingungen. Verwendung von üblichen Repositorien wie Hadoop/Spark/Flink/Mahout, Python-RASBT, R, etc. Ein besonderer Fokus soll auf einer ganzheitlichen wirtschaftsinformatischen Betrachtung liegen. Es soll dabei neben der informatischen Betrachtung auch der betriebswirtschaftliche Nutzen, z.B. anhand eines Use Cases, betrachtet werden.

Grundsätzliches
- Bearbeitung in Gruppen von 2-3 Studierenden
- Organisation der Gruppen durch Studierende. Bei Problemen bitte bei mir melden!
- Jede Gruppe gibt sich einen Namen.
- Jede Gruppe bearbeitet ein anderes Thema.
- Gruppen dürfen Themen selbst vorschlagen - bitte mit mir abstimmen.
- Jede Gruppe organisiert sich selbst (jede Gruppe ist also in „Mini-Scrum-Team“)

Ablauf in Vorlesungsphase
- Regelmäßige Abstimmungen zu den Vorlesungszeiten mind. einmal/Woche: Hier haben wir die Möglichkeit Probleme oder Zwischenergebnisse zu diskutieren.
- Nach der Hälfte der Vorlesungszeit hält jede Gruppe eine Zwischenpräsentation über den derzeitigen Stand (ca. 15 min - hier sind wir zeitlich aber flexibel). Die Zwischenpräsentation geht **nicht** in die Abschlussbewertung mit ein.

Abgaben am Ende des Projekts
- Erstellter Quellcode (bevorzugt R - sollten Sie eine andere Sprache wählen, bitte mit mir abstimmen. Ziel des Projekts ist es **nicht** eine neue Sprache zu lernen)
- Projektreport (5-10 Seiten)

Abschlusspräsentation
- Jede Gruppe präsentiert am Ende die Ergebnisse (20 min je Gruppe + 10 min für Fragen. Die Zeit sollte einigermaßen gleichmäßig auf die Studierende verteilt werden, d.h. in einer 2er-Gruppe jeder ca. 10 min.).

Bewertung
- Abgaben & Präsentation werden bewertet und ergeben die Note für diese Lehrveranstaltung (genaue Gewichtung legen wir noch fest, tendenziell: 50% Projektreport, 30% Präsentation, 20% Code)

Bemerkungen
- Quellcode:
  - Die Qualität des Codes wird bewertet
  - Kommentieren Sie Ihren Code gut
  - Verwenden Sie sprechende (und sinnvolle) Namen für Variablen und Funktionen.
  - Vermeiden Sie Spaghetti-Code.
  - Clean Code
  - Grundregeln des Machine Learnings, die wir in der Vorlesung besprochen haben, einhalten.
  - Eine readme-Datei sollte beschreiben, wie der Code auszuführen ist, wenn Sie mehr als ein R-Skript abgeben

Projektreport
- wissenschaftliche Arbeit
- Inhalte

- Thema und Motivation
- Referenzen? Welche Publikationen gibt es zum Thema?
- Grundlagen
- Ergebnisse
- Diskussion der Ergebnisse

Beispiele für Projekte & Projektreports (Stanford University):
http://cs229.stanford.edu/projects.html

Quelle für Datensätze
- Kaggle
- UCI Machine Learning Library
- RKI (z.B. Covid-19)
- …

Sonstiges (hat nichts mit dem Projekt zu tun) - eine aktuelle Studie

Vielleicht haben Sie von der Studie
https://www.kit.edu/kit/pi_2020_114_signifikanter-effekt-von-schulschliessungen.php
am KIT gehört. Dort wurden Covid-19-Daten aus einigen Ländern mit „KI-Methoden“ analysiert.

Die Studie erlangte ziemliche Aufmerksamkeit, weil sie von Drosten erwähnt wurde:
https://twitter.com/c_drosten/status/1337081420490141700
Allerdings mit dem (heutzutage bemerkenswerten) Zusatz „Ich kann das methodisch nicht bewerten“.

Die Studie wurde dann von einigen Statistikern kritisiert:
https://twitter.com/domliebl/status/1337315379413200897
https://twitter.com/domliebl/status/1337322521457274888
https://twitter.com/domliebl/status/1337430954374258691

https://twitter.com/JohannesTextor/status/1337708442224635908

https://twitter.com/christoph_rothe/status/1337303605456613378

Was lernen wir hieraus?
- auch Promovierte machen Fehler
- Veröffentlichungen müssen kritisch gelesen werden - insbesondere natürlich, wenn es keinen Begutachtungsprozess gibt. Aber selbst dann!
- Korrelation impliziert nicht Kausalität (wohl das Hauptproblem der Studie); sie erinnern sich vielleicht noch an unseren Zufallszahlendatensatz - dort erhielten wir auch leichte Korrelationen. Je mehr Variablen man hat, desto wahrscheinlicher wird es.
- Ein Plausibilitätscheck kann nicht schaden (vgl. Tweet von C. Rothe)

Ich wünsche Ihnen eine schöne Weihnachtszeit und einen guten Start ins neue Jahr!

Viele Grüße
Andreas Weber

--
Prof. Dr. Andreas Weber
Zentrum für Wirtschaftsinformatik

Duale Hochschule Baden-Württemberg Karlsruhe
Baden-Wuerttemberg Cooperative State University Karlsruhe
Erzbergerstr. 121

76133 Karlsruhe

Tel.: +49 (0)721 9735 - 931
E-Mail: andreas.weber@dhbw-karlsruhe.de
https://www.karlsruhe.dhbw.de