Bearbeitungshinweise Data Science-Projekt, WI, Kurs 2019

Dr. Ralf Höchenberger

1. Organisatorisches

Jede der von Euch gebildeten Gruppen hat von mir einen Datensatz zur Verfügung gestellt bekommen. Ab diesem Zeitpunkt habt Ihr bis zum (noch zu bestimmenden) Präsentationstermin Zeit, um damit zu arbeiten und eine Präsentation der Ergebnisse zu erstellen. **Im Wesentlichen erstellt Ihr während der Bearbeitungszeit ein Python-Notebook das Ihr auch in der Präsentation auszugsweise oder vollständig verwenden könnt.** Es ist Eurem Ermessen überlassen, ob Ihr zusätzlich eine (PowerPoint-)Präsentation erstellt. Während der Präsentation ist es ratsam, einen Rechner zur Verfügung zu haben, auf dem die Ergebnisse live vorgeführt werden können, da wir gegebenenfalls als Teil der Diskussion einzelne Schritte „live" programmieren werden. **Die Bewertung setzt sich dann aus der Präsentation sowie aus der Leistung der einzelnen Gruppenmitglieder in der Beantwortung meiner Fragen zusammen. Diese Fragen beziehen sich sowohl auf das behandelte Themengebiet als auch auf die nicht im zugeteilten Projekt, aber in der Vorlesung besprochenen Themen zusammen**!

Solltet Ihr während der Bearbeitungszeit Fragen haben, könnt Ihr mir eine Mail schreiben (ralf.hoechenberger@outlook.de). **Es muss kein definierter Stand bis zum (noch zu definierenden) Zwischentermin präsentiert werden – selbst wenn Ihr nur zu einer Ideensammlung ohne Implementierung im Notebook kommt, ist das in Ordnung. Ich gebe dann beim Zwischentermin Hinweise.** Sollten sich insgesamt sehr viele Fragen ergeben, würde ich etwa in der Mitte der Bearbeitungszeit noch mal ein Teams-Meeting mit allen anbieten, wo Ihr mir Eure Fragen stellen könnt.

2. Inhalt

a) Grundidee

Die Grundidee des Projektes besteht darin, mit dem jeweiligen Datensatz das jeweils zugeteilte Thema durchzuarbeiten, so wie es in der Veranstaltung im Anwendungsteil unterrichtet wurde, und sogar darüber hinauszugehen. Zeigt, was Ihr im Kurs gelernt habt, indem Ihr so viele Methoden wie möglich auf Euren speziellen Datensatz anwendet. **Hinter jedem Datensatz verbergen sich inhaltliche Fragestellungen bzw. ein reales Problem, das sinnvoll untersucht werden kann.**
Ein Ziel des Kurses soll auch sein, dass Ihr Euch **auch mit noch unbekannten Methoden beschäftigt**; dies schult Eure Fähigkeit zu recherchieren und sich mit neuen Problemstellungen auseinanderzusetzen. Bemüht auch Google und gebt dort Schlüsselworte des Problems (z.B. auch Fehlermeldungen im Code) ein, um zur Lösung des Problems zu gelangen. Es gibt im Internet sehr viele Foren und weitere Plattformen, die zahlreiche bekannte Fragestellungen ausführlich behandeln (und dies stellenweise auch von sehr fortgeschrittenen Usern).

b) Vorgehensweise

Wie könnte nun der Ablauf des Projektes konkret aussehen? Eine mögliche Vorgehensweise ist diese:

- **Macht Euch sich zunächst mit den Daten selbst vertraut. Worum geht es überhaupt?** Lest den Datensatz ein (entweder bereits direkt in der Programmierumgebung oder bspw. in Excel) und versucht, den Inhalt der Spalten zu verstehen.

- Ein nächster typischer Schritt ist, die sogenannten **„summary statistics"** zu ermitteln. Welche Werte nehmen die Spalten/Variablen an? Wie ist deren Mittelwert/Varianz, Minimum/Maximum? Eventuell findet Ihr auch noch weitere, dem Mittelwert ähnliche Maße.

- Der wichtigste Schritt ist, eine zentrale Fragestellung zu erkennen: **Welche praxisrelevanten Fragestellungen könnte in dem Datensatz adressiert werden?** Je nach vergebenem Themengebiet (kNN, kMC oder NN) zielen diese natürlich bereits grundsätzlich auf etwas anderes ab. **Versetzt Euch in die Lage von Unternehmensberatern, die vom Kunden beauftragt wurden, Muster in den Daten zu erkennen, und Handlungsempfehlungen für effizientere Unternehmensentscheidungen abzuleiten!** Alles Konkrete besprechen wir dann in den einzelnen Gruppenbesprechungen.

c) Anmerkungen

Ihr werdet feststellen, dass sich die Bearbeitung des Projektes an sehr wenigen festen Regeln orientiert und daher auch in hohem Maße kreativ bearbeitet werden kann. **Zeigt, dass Ihr die Aufgabe selbständig in der Gruppe bearbeitet habt!** Geben mehrere Gruppen in zu starkem Ausmaß sehr ähnliche Antworten ab, so bewerte ich alle mit „nicht bestanden". Bitte beachtet, dass es keine einzig korrekte Lösung gibt. **Jede Vorgehensweise, die Kenntnis der Daten und des Materials der Vorlesung demonstriert, sowie eine nachvollziehbare Argumentation aufweist, bewerte ich positiv.** Im Zweifel ist es weniger aufwändig, selbst etwas über den Datensatz zu lernen und sich ein paar Gedanken zu machen, als die Lösung von Mitstudierenden zu überarbeiten, sodass das Plagiat nicht auffällt.
Zu den meisten Datensätzen und den jeweiligen Fragestellungen findet Ihr im Internet fertige Lösungen. Diese kenne ich auch und ein solches Plagiat wird genauso bewertet, wie wenn Ihr von einem Mitstudierenden kopiert. Dort durchgeführte Schritte können zwar vereinzelt nachprogrammiert werden, da es oft die einzige sinnvolle Möglichkeit ist, jedoch werde ich durch gezielte Fragen meist feststellen, ob die Inhalte dann tatsächlich auch verstanden wurden oder nicht.

Viel Freude und Erfolg bei der Bearbeitung des Projektes!