Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2022

Visualisierung des Datensatzes Student Behaviour

Paul Tebbe, Matrikelnummer: 216237588

21.12.2022

1 Einleitung

..... Viele verschiedene Faktoren haben einen Einfluss auf die Note der Studenten. Dies geht bereits damit los in welche Familie das Kind hineingeboren wird, welchen akademischen Standart die Eltern haben, welche finanziellen Mittel usw. Weiter spielt die Erziehung und der spätere soziale Umgang eine Rolle. [**Earthman**] Diese Einflüsse sind schwer zu messen, vor allem schwer genau zu messen. Diese Faktoren sind kaum von alleine beeinflussbar, es lässt sich nicht bestimmen, in welche Familie ich geboren werde und welche Voraussetzungen ich also für später mit auf den Weg gegeben bekomme. Doch auf welche Faktoren kann der Student selber später Einfluss nehmen. Die Zeit, die täglich in oder eben nicht in die schulische Weiterbildung investiert wird, sollte in der Theorie so einen selbst beeinflussbaren Faktor darstellen. In dieser Arbeit werden sich also folgende Fragen und folgende gestellt Zielstellungen gesetzt:

- Inwieweit spielt die eigene Zeitgestaltung eine Rolle bei der Note der Befragten
- Wie können die verschiedenen Korrelationen übersichtlich veranschaulicht werden um dem Leser einen schnellen Überblick zu geben
-

Tipps zu Latex und Koma-Script für Hausarbeiten sind im LaTeX Reference Sheet for a thesis with KOMA-Script von Marion Lammarsch und Elke Schubert zusammengefasst. Der Bericht fällt in die Kategorie von InfoVis-Paper, die Tamara Munzner Design Study nennt [**Munzner2008**]: In der Einleitung sollen sie zuerst das Zielproblem beschrieben. Daraus sollen sie Fragestellungen motivieren, die mittels Techniken der Informationsvisualisierung beantwortet

werden können. In dem Abschnitt direkt unter der Überschrift Einleitung sollen Sie nach einer kurzen Einleitung Fragestellungen und das Zielproblem motivieren und beschreiben. ...

1.1 Anwendungshintergrund

Wer in der Schule und im College gute Noten hat und sich somit einen besseren Abschluss erarbeitet hat im späteren Leben deutlich bessere Chancen auf „individuelle Lebenschancen, Selbstverwirklichung, beruflichen Erfolg sowie soziale, politische und kulturelle Teilhabe. “

1.2 Zielgruppen

Die Zielgruppe für die Daten umfasst ein weites Spektrum an Personen. Die Visualisierung sind sowohl für Schüler und Studenten, als auch für Lehrer und Eltern interessant, also jegliche Personengruppe, die Berührungspunkte mit der Notengebung und deren Einflussfaktoren aufweist. Die Informationen, die aus den Visualisierungen gewonnen werden können, können sehr gut dafür genutzt werden zu sehen, wie grade der Einflussfaktor Lernzeit und auch die vorangegangen Noten aus der Schulzeit einen Einfluss auf die Noten im College nehmen. Differenziert werden kann hierbei auch nochmal unter den Geschlechtern. Auch interessant könnte dieses Projekt aber auch für Forschende im Bereich soziale Ungleichheitne in der Bildung sein. Wenn keine eindeutigen Anzeichen zu sehen sind für eine Korrelation zwischen investierter Lernzeit und besseren Noten, so könnte dies ein Anzeichen dafür sein, dass doch andere Einflussgrößen mehr Auswirkung auf die Noten haben als der reine Fleiß. Dies wäre ein starker Indikator für soziale Ungleichheiten bei der Bildung, wie es bereits in der Einleitung angedeutet wurde.

Beschreiben sie die Personengruppe oder Personengruppen, die das von ihnen benannte Anwendungsproblem lösen möchte. Auf welches Vorwissen können sie in dieser Gruppen von Anwenderinnen aufbauen? Welche Informationsbedürfnisse werden durch die Visualisierungen adressiert?

1.3 Überblick und Beiträge

Der in dieser Arbeit verwendete Datensatz differenzieren die Studenten nach verschiedenen Merkmalen. Diese gehen von allgemeinen Unterscheidungen, wie Geschlecht, Alter und Gewicht, über die Aufteilung ihrer Freizeit und die aufgewendete Lernzeit bis hinzu dem körperlichen Wohlbefinden ausgedrückt mittels der Variable ‚stress level ‘.
Die für diese Visualisierung mit am interessantesten Daten sind die der verschiedenen Noten in der Schule und im college. Die Noten, gerade mit der von den Studenten angegebenen täglich Lernzeit sollte in einem positiven Zusammenhang stehen. Also steigt die Lernzeit am Tag, steigt auch die Note. Das gleiche sollte auch für die Beziehung der Noten untereinander gelten, also welcher Student schon in der 10.ten und 12.Klasse gute Noten hatte, sollte nun auch im College gut abschneiden. Die körperlichen Voraussetzungen sowie das Department sollten eigentlichen keine größeren Einflüssen auf die Noten haben. Die Zeit verbracht auf sozialen Medien könnte

eine negative Korrelation mit den Noten aufweisen, genauso wie der finanzielle Status und das Stress Level bei den Studenten.

In diesem Abschnitt geben sie einen kurzen Überblick über die Daten und verwendeten Visualisierungen. Dann benennen sie die Beiträge ihres Projekts. Diese Beiträge müssen sie in den hinteren Teilen des Berichts genauer ausführen und belegen.

2 Daten

Wie bereits in den Abschnitten vorher erwähnt, handelt es sich bei den vorliegenden Daten um den Datensatz *Student Behaviour* von *kaggle* . Die darin aufgeführten und gesammelten Daten geben nähere Auskunft über das Verhalten der Studenten und deren Noten. Die Verwendung des Datensatzes zu wissenschaftlichen Zwecken ist abzuraten, da der Verfasser angibt die Daten selber von Universitäten gesammelt zu haben und bis auf diesen Satz keine näheren Informationen zur Datenbeschaffung liefert. Seiner Meinung nach kann der Datensatz dazu genutzt werden das Verhalten der Studenten in Zukunft besser zu deuten. Er enthält Informationen von 235 Studenten in Form einer CSV-Datei. Diese CSV-Datei hat 19 Spalten mit den Informationen der Studenten über *Certification* als Boolean, der ausgibt ob der Student einen certification Kurs besucht hat oder nicht, *Gender* mit Boolean über das Geschlecht, *Department*, welcher Student welchem Derpartment angehört, *Height(CM)* und *Weight(KG)* als Float, die Größe und Gewicht aufzeigen, *10th Mark*, *12th Mark* und *College Mark*, die als Float die Note der Studenten zeigen, *Hobbies*, sowie *stress level* und *financial status* als String, *part-time job* und *Do you like your degree* als Boolean, *salary expectation* als integer, *willingness to pursue a career based on their degree* als String in Form einer Prozentangabe (möglich sind hier nur die Angaben: 0,25,50,75 und 100 Prozent) und *travelling time*, *social media time* und *daily studying time* als String in Form einer Klassenangabe (also z.B. 30-60 Minuten ist eine Angabe).
Beschreiben Sie vorhandenen Daten. Gehen sie kritisch darauf ein, in wie weit sich die Daten für die Bearbeitung der Fragestellungen und dem Erreichen von Lösungen für die oben beschriebene Zielgruppen eignen. Haben sie die Daten sinnvoll mit weiteren Datenquellen ergänzt? Wenn ja, wie? Erklären sie die technische Bereitstellung der Daten. Wie sind die Daten zugänglich? Welche Formate werden genutzt. Gibt es Besonderheiten beim Lesen der Formate? Beschreiben sie die Datenvorverarbeitung. Welche Datenvorverarbeitungsschritte sind notwendig? Beschreiben Sie die einzelnen Schritte und begründen sie sie, z.B. warum werden manche Daten weggelassen, über welche Mengen werden Durchschnitte berechnet, warum sind die so berechneten Werte aussagekräftiger als andere Werte. Wenn möglich sollen sie die Datenvorverarbeitung in Elm programmieren, so dass ihre Anwendung auf eine Änderung der Rohdaten reagieren kan.

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Analysieren sie die konkreten Anwendungsaufgaben, die die Lösung des Zielproblems durch die Anwender:innen bearbeitet werden müssen. Welche sinnvollen mentale Modelle helfen den Personen bei der Bearbeitung. Sind diese mentalen Modelle für sie notwendig, um die Aufgaben lösen zu können? Gehen sie bei ihrer Argumentation von den Anwendungsaufgaben aus und kommen sie dann zu den mentalen Modellen, deren Aufbau durch Visualisierungen unterstützt wird.

3.2 Anforderungen an die Visualisierungen

Leiten sie Anforderungen an das Design der Visualisierungen ab, die sich durch ihre Analyse des Zielproblems ergeben.

3.3 Präsentation der Visualisierungen

Präsentieren sie die visuelle Abbildungen und Kodierungen der Daten und Interaktionsmöglichkeiten. Sie müssen begründen, warum und wie gut ihre Designentscheidungen die erstellten Anforderungen erfüllen. Weiterhin müssen sie begründen, warum die gewählte visuelle Kodierung der Daten für das zulösenden Problem passend ist. Typische Argumente würden hier auf Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen. Die besten Begründungen diskutieren explizit die konkrete Auswahl der Visualisierungen im Kontext von mehreren verschiedenen Alternativen. Machen sie hier nicht den Fehler, einfach nur Visualisierung aus den vorgegebenen Bereichen zu diskutieren, weil das in der Regel nicht sinnvoll ist. Wenn sie sich für einen Scatterplot entschieden haben, ist ein Zeitreihendiagramm in der Regel keine Alternative. Diskutieren sie also nicht einfach Zeitreihendiagramme, weil sie in den Anforderungenen an das Projekt neben Scatterplots stehen, sondern suchen sie nach echten alternativen Visualisierungen, die zum Aufbau eines vergleichbaren mentalen Modells führen. Diskutieren sie die Expressivität und die Effektivität der einzelnen Visualisierungen.

Die eben beschriebenen Präsentationen und Begründungen sollen für jede der drei folgenden Visualisierungen durchgeführt werden.

3.3.1 Visualisierung Eins

3.3.2 Visualisierung Zwei

3.3.3 Visualisierung Drei

3.4 Interaktion

Die präsentierten Visualisierungstechniken müssen interaktiv zu einer Anwendung verknüpft werden. Die Interaktion mit einer Visualisierung soll in den anderen Visualisierungen zu einer

Änderung führen. Erklären sie die möglichen Interaktionen mit den einzelnen Visualisierungen und die möglichen Verknüpfungen zwischen ihnen. Begründen Sie warum die konkreten Interaktionen umgesetzt wurden und welche Zwecke für die Anwenderinnen mit ihnen unterstützt werden. Begründen sie ebenfalls warum sie andere Interaktionsmöglichkeiten nicht umgesetzt haben. Wenn sie keine der geforderten Interaktionen umsetzen, erhalten Sie im gesamten Projekt deutlichen Punktabzug.

4 Implementierung

Beschreiben Sie die Implementierung ihrer Visualisierungsanwendung in Elm. Stellen die Gliederung ihres Quellcodes vor. Haben Sie verschiedene Elm-Module erstellt. Was war aufwändig umzusetzen, was ließ sich mit dem vorhanden Code aus den Übungen relativ einfach umsetzen?

Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können.

5 Anwendungsfälle

Präsentieren sie für jede der drei Visualisierungen einen sinnvollen Anwendungsfall in dem ein bestimmter Fakt, ein Muster oder die Abwesenheit eines Musters visuell festgestellt wird. Begründen sie warum dieser Anwendungsfall wichtig für die Zielgruppe der Anwenderinnen ist. Diskutieren sie weiterhin, ob die oben beschriebene Information auch mit anderen Visualisierungstechniken hätte gefunden werden können. Falls dies möglich wäre, vergleichen sie die den Aufwand und die Schwierigkeiten ihres Ansatzes und der Alternativen.

5.1 Anwendung Visualisierung Eins

5.2 Anwendung Visualisierung Zwei

5.3 Anwendung Visualisierung Drei

6 Verwandte Arbeiten

Führen sie eine kurze Literatursuche in der wissenschaftlichen Literatur zu Informationsvisualisierung und Visual Analytics nach ähnlichen Anwendungen durch. Diskutieren sie mindestens zwei Artikel. Stellen sie Gemeinsamkeiten und Unterschiede dar.

7 Zusammenfassung und Ausblick

Fassen sie die Beiträge ihre Visualisierungsanwendung zusammen. Wo bietet sie für die Personen der Zielgruppe einen echten Mehrwert.

Was wären mögliche sinnvolle Erweiterungen, entweder auf der Ebene der Visualisierungen und/oder auf der Datenebene?

Anhang: Git-Historie