Projektbericht zum Modul
Information Retrieval und Visualisierung
Sommersemester 2022/23

Thema:

Untersuchung des Einflusses von Alkohol auf Schüler mithilfe verschiedener Visualisierungstechniken

Vorgelegt von:

Arsela Leskaj

Matrikelnummer: 216104691

Abgabetermin: 20.12.2022

GitHub Repository: https://github.com/Ella199/Elm_Project-Student-Alcohol-Consumption

Projektwebseite: file:///Users/arselaleskaj/Documents/GitHub/Elm_Project-Student-Alcohol-Consumption/Webseite/index.html

Gliederung

1. Einleitung

Eine Studie aus dem Jahr 2007 vom Eurostat statistischem Amt der Europäischen Union (EU), beobachtet frühzeitige Schul- und Ausbildungsabgänger innerhalb der EU. Dabei wird festgestellt, dass die durchschnittliche Schul- und Ausbildungsabgangrate der prozentualen 18- bis 24-Jährigen in der EU im Jahr 2006 bei 15% lag. Portugal stach mit einer sehr hohen Abgangrate von 40% auf.
Es ergibt sich die Frage, womit der vorzeitige Schulabbruch zu tun hat. Eine Möglichkeit ist ein zu hoher Alkoholkonsum.

Junge Menschen kommen sehr schnell mit Alkohol in Kontakt. Das hat verschiedene Gründe. Soziale Aspekte wie Anerkennung bei anderen Schülern, sich bestimmten Freizeit- oder Lerngruppen zugehörig und verbunden zu fühlen, hängen mit dem Trinkverhalten vieler Jugendlicher zusammen. Viele Schüler lernen in der Gruppe und wollen in wechselseitiger Verbindung mit den anderen Mitschülern stehen, da sie sich gegenseitig austauschen möchten und der Wissenstransfer dabei helfen kann, Fortschritte beim Lernen zu machen und den Unterrichtsstoff besser zu verinnerlichen. Oft trinken manche Jugendliche nach einer Gruppenarbeit, um den erzielten Erfolg von einem langen Lerntag zu feiern. Um sich eine Belohnung zu gönnen, greifen sie zu alkoholischen Getränken. Die familiären Hintergründe spielen auch eine sehr wichtige Rolle, zum Beispiel die oftmals zu hohen Erwartungen der Eltern, die zu Leistungsdruck führen. Diejenigen, die nicht mit dem Lernstress umgehen können sehen das Trinken als einen möglichen Weg, den Stress zu bewältigen. [1]

Eine Studie der Universität Essex belegt, dass diejenigen, die nach einem langen Lerntag Alkohol trinken, Informationen besser im Langzeitgedächtnis speichern können als andere, die auf den Alkoholkonsum verzichten.[2]

Allerdings führt eine regelmäßige übersteigende Trinkmenge an Alkohol zu chronischem Alkoholkonsum, welcher viele Verhaltensstörungen induzieren kann. Solche Verhaltensstörungen können die sozialen Interaktion in der Gruppen beeinträchtigen. Somit kann es zur Reduzierung sozialer Kontakte kommen und das kann wiederum dazu führen, dass manche Schüler von anderen Gruppenmitgliedern ausgeschlossen werden. Viele Studien berichten aber auch, dass der Alkoholkonsum geringere Wirkung auf die Schulnoten hat. Der Alkoholkonsum kann als intrinsische Motivation dienen und sorgt für mehr Interaktion oder weckt das Interesse der

---

[1] Vgl. Koler P., (2014), Stumpp G., Stauber B., & Reinl H. (2009)

[2] Vgl. FOCUS online, (2017)

Gruppenmitglieder, mit anderen Mitgliedern Zeit zu verbringen. Darüber hinaus können die Schüler in Lerngruppen zusammenfinden und nach dem Lernen zusammen Zeit verbringen. Alkohol kann auch die soziale Interaktion in der Gruppe stärken. Das Konsumieren des Alkohols kann bestimmte Bereiche im Belohnungssystems des Gehirns simulieren und dementsprechend das Verhalten der Individuen beeinflussen.[3]

Im Fokus dieser Projektarbeit steht die Untersuchung der Zusammenhänge zwischen dem Alkoholkonsum und den erreichten Schulnoten von zehnter bis zwölfter Klasse anhand von drei Visualisierungstechniken, die im Kapitel 1.1 vorgestellt werden. Darüber hinaus werden in dieser Projektarbeit interessante Erkenntnisse zu der Thematik geliefert. Weiterhin werden auch andere demographische (zum Beispiel Bildungsniveau der Eltern) und soziale (beispielsweise Freizeit nach der Schule) Merkmale in Erwägung gezogen um zu verstehen, ob noch andere Merkmale die Schülerleistungen beeinträchtigen.

1.1 Anwendungshintergrund

Ziel dieser Arbeit ist es, eine Übersicht anhand von drei ausgewählten Visualisierungstechniken zum Thema "Untersuchung des Einflusses von Alkohol auf Schüler mithilfe verschiedener Visualisierungstechniken" zu bieten. Die verwendeten Techniken sind Scatterplot, Parallele Koordinaten und Stickfigureplot.

Ein Scatterplot bildet die Datenpunkte in einem zweidimensionalen Raum ab. Mithilfe dieser Visualisierungstechnik lässt sich die Beziehung zwischen Attributwerten auf den Koordinatenachsen untersuchen.

So können beispielsweise Informationen über zwei numerische Beträge - zum Beispiel Alkoholkonsum unter der Woche und Abschlussnoten aus dem Fachbereich Mathematik - geliefert werden. Diese Technik beinhaltet weitere Vorteile, wie zum Beispiel die Identifizierung bestimmter Muster in den Datensätzen. Somit lassen sich Zusammenhänge zwischen Datensätzen untersuchen und es können Aussagen gemacht werden, ob zwischen den Variablen unter anderem eine positive oder negative Korrelation besteht. Weiterhin können die Anwender zu neuen Erkenntnissen gelangen und interessante Rückschlüsse zu diesem Thema ziehen, indem sie anhand dieser Visualisierung die Möglichkeit erhalten, Ausreißpunkte in den Datensätzen zu identifizieren.[4]

---

[3] Vgl. Soyka (2001), Doerfel A. Helmholtz-Gemeinschaft (2018)

[4] Vgl. Rumsey (2010)

Die zweite Visualisierung innerhalb dieser Forschungsarbeit sind die Parallele Koordinaten. Bei dieser Technik besteht die Möglichkeit, die Datensätze in einem mehrdimensionalen Raum darzustellen. Die Koordinatenachsen dieser Visualisierung bauen sich parallel zueinander auf und stellen die Werte für vier verschiedene numerische Variablen dar. Attributwerte werden durch Linien dargestellt. Vier numerische Beiträge der jeweiligen Variablen lassen sich gleichzeitig anzeigen und durch Anklicken bestimmter Variablen kann man die Linien eindeutig identifizieren. Somit lassen sich einzelne Zusammenhänge zwischen den Attributen feststellen. Weiterhin lässt sich beobachten, wie eng sich Liniengruppen zusammenschließen und in welche Richtung sie sich bewegen.[5]

Die dritte und somit letzte Technik, die zur Visualisierung der Datensätze verwendet wurde, ist Stickfigureplot. Durch diese Technik lassen sich die Daten in einem mehrdimensionalen Raum darstellen. Diese Technik zieht im Vergleich zu den anderen vorgestellten Techniken mehr Variablen in Betracht. Eine von den Eigenschaften dieser Technik sind unter anderem wahrnehmungsbasierte Darstellungen und gleichzeitige Beobachtung der Zusammenhänge und Muster zwischen sieben Variablen. Ähnlich wie bei Scatterplot können hier auch Korrelationsbeziehungen und Ausreißer in den Datensätzen identifiziert werden.[6]

Die für diese Projektarbeit verwendeten Datensätzen stammen aus zwei öffentlichen Sekundarschulen: Gabriel Pereira und Mousinho da Silveira in Portugal. Die Datensätze wurden in den Jahren 2005 und 2006 erhoben. Das Schulsystem in Portugal besteht aus neun Jahren Grundschulbildung, gefolgt von drei Jahren Sekundarschule. Wobei die neun Jahre Grundschulbildung die Voraussetzung für die Sekundarschule sind. Die Bewertungsskala in der Sekundarschule reicht von 0 bis 20 Punkten. 0-9 Punkte gelten als nicht bestanden und 20 ist die beste Note. In den Datensätzen werden die Schulnoten von der zehnten bis zur zwölften Klasse für die Schulfächer Mathematik und Portugiesisch erfasst. Die Abschlussnoten der zwölften Klasse ergeben sich aus dem Durchschnitt der zehnten und elften Klasse. Die Datensätze enthalten wesentliche Informationen unter anderem über schulbezogene Leistungen sowie soziale und demographische Merkmale. Somit können anhand der Analyse bestimmter Variablen in den Datensätzen Erkenntnisse hinsichtlich der Beeinträchtigung der Schülerleistungen in den beiden Schulfächern gewonnen werden.[7]

---

[5] Vgl. Gemignani (2021)

[6] Vgl. Pickett & Grinstein (1988)

[7] Vgl. Cortez & Silva (2008), Taborda (2022)

1.2 Zielgruppen

Für die vorliegenden Daten sind im Rahmen dieser Arbeit drei potenzielle Zielgruppen denkbar: Data Science-Unternehmen, Schuldirektoren und die Forschung.

Data Science-Unternehmen sindals Zielgruppe möglich, um konkrete Vorhersagen zu treffen hinsichtlich der Abschlussnote und um herauszufinden, ob die Absolventen nach einem universitären Hochschulabschluss oder einer betrieblichen Ausbildung streben. Zudem können sie auch langfristige Untersuchungen zur Bevölkerungsentwicklung oder einen Generationsvergleich in Portugal machen.

Auch für die Schuldirektoren können die Ergebnisse dieser Forschungsarbeit von Interesse sein. Beispielsweise können sie sich dafür interessieren, wann welche Präventionsmaßnahmen eingeführt werden sollten, um die Leistung der Schüler zu verbessern. Ein Schuldirektor kann sich auch für eine bessere Lernqualität interessieren und die entsprechenden Maßnahmen treffen, um dies zu ermöglichen.

Anhand der Daten kann erkannt werden, ob die Schüler in naturwissenschaftlichen (Mathematik) oder in geisteswissenschaftlichen (Portugiesisch) Fächern am besten abschneiden. Zudem ist es für den Schuldirektor interessant, wann die Schüler am meisten Alkohol trinken - am Wochenende oder unter der Woche - und wie der Alkoholkonsum mit den Schulnoten zusammenhängt. Denkbar wäre, Unterrichtsstunden für das Thema Alkohol einzuführen, damit sich die Schüler besser informieren können. Wobei zu beachten ist, dass das nur eine Empfehlung ist.

In Kapitel 3 wird anhand der Visualisierungen in Erfahrung gebracht, ob die Schüler ein Problem mit Alkohol haben. Alternativ wäre auch möglich für die Schüler, die mehr als durchschnittlich Alkohol konsumieren, Coaching in der Schule einzuführen.

Eine weitere potenzielle Zielgruppe ist die Forschung. Sie kann untersuchen, ob sozialstrukturelle Aspekte den Alkoholkonsum beeinflussen, zum Beispiel inwieweit der Sozialstatus der Eltern die Schulnoten beeinflusst. So kann die Forschung in Erfahrung bringen, ob diejenigen Kinder, die aus einer Familie stammen, in der die Eltern eine akademisches Bildungsniveau verfügen, bessere Noten haben.

1.3 Überblick und Beiträge

Das Thema Alkoholkonsum bei Jugendlichen wird weltweit diskutiert und es werden Studien beispielsweise an Schulen durchgeführt, welche den Zusammenhang zwischen Schulleistung und Alkoholkonsum untersuchen.[8]

[8] Vgl. Balsa, Giuliano & French (2011), DeSimone & Wolaver (2005), Dee & Evans (2003)

Mithilfe dieser Projektarbeit können mögliche Fragen, die das Interesse der angesprochenen Zielgruppen zu diesem Thema entsprechen, beantwortet werden. Die Projektarbeit beschäftigt sich mit folgenden Fragen: Werden die Abschlussnoten von den früheren erreichten Noten aus der zehnten und elften Klasse beeinflusst? Inwieweit beeinträchtigt der Alkoholkonsum die erreichten Schulnoten der zehnten, elften und zwölften Klasse in den Fächern Mathematik und Portugiesisch? Welche Noten erreichen die Schüler bezogen auf das Ausmaß ihres Alkoholkonsums? Wie kann eine Kombination des Alkoholkonsums mit demographischen und soziablen Variablen die Schülerleistung beeinflussen? Wie sehen die geschlechtsspezifischen Unterschiede hinsichtlich des Zusammenhangs von Alkoholkonsum und der Abschlussnoten aus?

Um viele Interessenten darüber zu informieren und einen guten Überblick zu bieten, wurden in dieser Projektarbeit drei Visulalisierungstechniken verwendet: Scatterplot, Parallele Koordinaten und Stickfigureplot. Scatterplot ermöglicht den Nutzern eine zweidimensionale Darstellung der Datensätze. Um diese Technik anzuwenden, wird von den Anwendern kein Vorwissen im Fachbereich verlangt. Scatterplot stellt eine benutzerfreundliche Anwendung für die Nutzer dar.

Parallele Koordinaten ist eine weitere Visualisierungstechnik, die in dieser Arbeit angewendet wird. Bei dieser Technik können beliebige Variablen ausgewählt und gegenüber gestellt werden. Somit können wesentliche Informationen aus den Datensätzen gezogen werden. Diese Art von Visualisierung fordert keine Vorkenntnisse der Anwender und ist ebenfalls benutzerfreundlich. Im Gegensatz zu Scatterplot weist diese Technik mehr Interaktionsmöglichkeiten zwischen den Variablen auf, da vier Attributwerte ausgewählt werden können.

Bei der dritten Visualisierungstechnik handelt es sich um einen Stickfigureplot und diese Technik weist eine mehrdimensionale Datendarstellung auf. Besonders bei dieser Darstellung ist, dass gleichzeitig sieben Variablen beobachtet werden können. Im Vergleich zu den anderen Visualisierungstechniken wurden bei dieser Technik Variablen - unter anderem soziale Attributwerte - dargestellt, die nicht in den vorherigen zwei Visualisierungstechniken vorkommen. Um diese Technik anzuwenden, ist es von Vorteil, über Vorkenntnisse zu der Thematik zu verfügen.

2. Daten

Für die bereits erwähnten Visualisierungstechniken werden Daten aus der Online-Plattform Kaggle entnommen. Die Daten dienten ursprünglich einem wissenschaftlichen Artikel, der sich mit der Methoden des Data Mining beschäftigt und Vorhersagen über die Schulnoten der Sekundarschüler in Portugal macht.[9] Ähnliche Arbeiten können bei der Suchmaschine Google gefunden werden.

[9] Vgl. Cortez & Silva (2008)

Einige Studien werden im Kapitel 6 als Vergleich zu dieser Projektarbeit genutzt um tiefergreifende Einblicke über die verwendeten Visualisierungstechniken zu liefern.

Die zur Verfügung gestellten Datensätze bei der Plattform Kaggle bestehen aus zwei CSV-Dateien und einer weiteren Datei, die dazu dient, die zwei CSV-Dateien in einer Datei zusammenzuführen.

Die eine CSV-Datei beinhaltet die erhobenen Daten für das Schulfach Portugiesisch und besteht insgesamt aus 649 Datensätzen. Die andere CSV-Datei für das Schulfach Mathematik enthält insgesamt 395 Datensätze. Aus diesem Grund wurde entschieden, den zur Verfügung gestellten R-Code (student-merge.R) bei der Webseite Kaggle auszuführen, um einen einheitlichen Datensatz der beiden Studienfächer zu bekommen. Nach dem Ausführen der mitgelieferten R-Anweisungen von den Autoren, wurde die ursprüngliche CSV-Datei mit 382 Datensätzen verkleinert. Die neue CSV-Datei beinhaltet die gemeinsamen Angabedaten derjenigen Studenten, die die beiden Umfragen sowohl für Mathematik als auch für Portugiesisch ausgefüllt haben. Eine wichtige Eigenschaft der Datensätze besteht darin, dass sich die Schüler nicht eindeutig identifizieren lassen, da die Daten anonym erhoben sind. Aus diesem Grund wurde für die Zusammenführung der beiden Datensätze in die neue CSV-Datei nach identischen Merkmalen der Schüler gesucht, die jeden Schüler charakterisieren.[10]

Die Datensätze lassen sich mit den ausgewählten Visualisierungstechniken gut anwenden und die Zielgruppen können daraus wesentliche Informationen in Erfahrung bringen. Auf das Thema wird im Kapitel 5 ausführlich eingegangen und Praxisbeispiele in Bezug auf die Zielgruppen geben.

2.1 Technische Bereitstellung der Daten

Für die Aufbereitung der Daten wurde die Programmiersprache Python verwendet. Python wurde benutzt, um mögliche Fehler bei den Datensätzen zu vermeiden, damit verlässliche Ergebnisse erzeugt werden können. Zudem wurden mit Hilfe von Python einige Datensätze ausgelassen, die für die Visualisierungstechniken nicht geeignet waren. Insgesamt gab es, wie in dem Kapitel zuvor erwähnt, 649 Befragte, die die Daten vollständig hinsichtlich der Sprache Portugiesisch und 382 für Mathematik eingegeben haben. Nach der Datenzusammenführung wurden die Daten mit Hilfe der Programmiersprache Python bearbeitet. Die Daten befinden sich in einem Ordner Data, der aus zwei unterschiedlichen Unterordnern besteht und zwar aus CSV und Pandas. Die Originaldaten sind im Ordner CSV zur Verfügung gestellt, damit die Anwender die Datensätze jederzeit anschauen können, um ein besseres Verständnis der Thematik zu erlangen. Zudem ist es auch wichtig, dass die

[10] Vgl. Student Alcohol Consumption, update (2016)

Nutzer der Github “Student Alcohol Consumption” den Zugang zu den Daten dieser Forschungsarbeit erhalten, um sowohl die Datenstruktur der originalen Datensätze als auch die aufbereiteten Datensätze zu haben.[11]

2.2 Datenvorverarbeitung

Für die Verwaltung und Analyse von Daten wurde die Software-Bibliothek Pandas von Python verwendet. Mithilfe dieser Programmbibliothek wurden die Datensätze in mehreren Schritten verarbeitet. Bevor die Daten auf Github hochgeladen wurden, wurden einige Schritte vorgenommen, die notwendig für eine Datenbereinigung sind.

Diese Schritte umfassen die Umbenennung, das Entfernen von doppelten Einträgen und NaN-Werten (Not a Number/keine Zahl). Weiterhin wurde mithilfe dieser Programmiersprache eine Analyse hinsichtlich des Datentyps durchgeführt.

Nach dem Zusammenführen der beiden Datensätze in eine gemeinsame CSV-Datei wurden bei den identischen Attributen jeweiligen Schülern automatisch neue Spalten .x für das Schulfach Mathematik und .y für das Schulfach Portugiesisch in den Datensätzen hinzugefügt. In einem weiteren Schritt wurden doppelte Einträge entfernt und nur die identischen Attribute beibehalten. Weiterhin wurden die Schulnoten für die zehnte, elfte und zwölfte Klasse mit der jeweiligen Schulfachbezeichnung umbenannt. Die Analyse hinsichtlich der Datentypen der Attribute ist für die nächsten Schritte bezüglich der Programmierung der Visualisierungstechniken wichtig. Darüberhinaus können Informationen geliefert werden, welche Attribute unter anderem einen numerischen- oder einen Text-Wert zuweisen.

Für das Einlesen der Daten in die Elm-Programmiersprache wurde entschieden, den Dateityp CSV zu verwenden, weil sich die Werte des Dateityps Excel in dieser Programmiersprache nicht einlesen konnten. Der Datensatz, der für die Bearbeitung der vorgestellten Visualisierungstechniken verwendet wurde, wurde als mergedstudent_FINAL_NaN.CSV im Ordner Daten/CSV gespeichert.

3. Visualisierungen

Das folgende Kapitel bietet einen umfassenden Überblick über die drei Visualisierungstechniken, die Anforderungen der Zielgruppen und die jeweiligen Interaktionsmöglichkeiten. Dieses Kapitel besteht aus vier Haupt-Unterkapiteln: Analyse der Anwendungsaufgaben, Anforderungen an die Visualisierungen, Präsentation der Visualisierungen und Interaktion. Wobei das Haupt-Unterkapitel

[11] Vgl. Student Alcohol Consumption, update (2016)

namens Präsentation der Visualisierungen aus drei weiteren Unterkapiteln - Visualisierung Eins, Zwei und Drei - besteht.

3.1 Analyse der Anwendungsaufgaben

Die bereits erwähnten Visualisierungen sollen den Anwendern dabei helfen, interessante Erkenntnisse zu gewinnen und aussagekräftige Rückschlüsse zu ziehen. Die Visualisierung der Datensätze in einem Scatterplot kann dazu verwendet werden um zwei numerische Arttribute miteinander zu vergleichen. Neben den numerischen Attributen wird auch ein Text Attribut namens “sex” angezeigt, um die Schüler geschlechtsspezifisch zu unterscheiden. Die Anwender können auf den ersten Blick erkennen, um welches Geschlecht es sich in der Visualisierung handelt, da die Schülerinnen und Schüler durch zwei Farben, rot für Schülerinnen und blau für Schüler, deutlich unterscheiden lassen. Da die Angabe der Daten anonym durchgeführt wurde, lässt sich nicht erkennen, welche Schüler welche Daten angegeben haben. Bei der Scatterplotvisualisierung können beliebige Attribute miteinander kombiniert werden. Wie bereits im vorherigen Kapitel erwähnt, sind Vorkenntnisse bei dieser Art der Visualisierung nicht erforderlich.

Es wird aber empfohlen, sich mit dem Thema Alkoholkonsum von Schülern auseinander zu setzen, bevor die Visualisierungstechniken angewendet werden können, damit die Kombination der Attribute sinnvoll ist. Diese Herangehensweise ist wichtig, um die entsprechende Frage hinsichtlich der Untersuchung zu definieren. Somit können Erkenntnisse hinsichtlich der Zusammenhänge zwischen zwei beliebig ausgewählten Attributen abgeleitet werden. Mithilfe der Scatterplotvisualisierung können neue Erkenntnisse zu spezifischen Untersuchungszwecken gewonnen werden, indem die Anwender selbst die Auswahl beziehungsweise die Kombination der Variablen festlegen.

Nutzern soll es möglich sein, einen fundierten Einblick in das Thema zu bekommen, indem sie Zusammenhänge zwischen den Variablen vergleichen können. Zum Beispiel können sie die zwei Beobachtungen - beispielsweise den Zusammenhang der Schulnoten der zwölften Klasse in den Schulfächern Mathematik und Portugiesisch - interaktiv vergleichen.

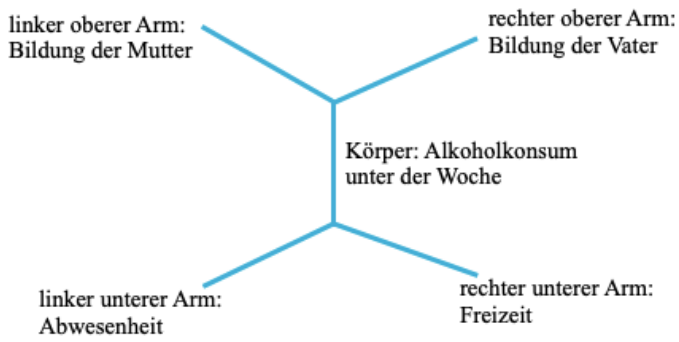
Diese Analyse ist empfehlenswert, weil die Anwender dadurch nicht nur allgemeine Zusammenhänge zwischen Alkoholkonsum und Noten untersuchen können, sondern auch die Besonderheiten des Einflusses des Alkoholkonsums auf die Schulleistung in den genannten Schulfächern. Darüberhinaus lassen sich Muster aufdecken und Trends identifizieren.

Durch die beliebige Auswahl von Kombinationen der Attribute können viele Fragen der Zielgruppen beantwortet werden. Es können beispielsweise Untersuchungen durchgeführt werden, wie der Alkoholkonsum die Noten der Schulfächer Mathematik oder Portugiesisch beinträchtigt oder ob der Alkoholkonsum zur Verbesserung der Schulnoten führt. Die Auswahl der Schulfächer kann auch die Frage beantworten, ob der Alkoholkonsum einen schlechten beziehungsweise keinen oder sogar einen positiven Einfluss auf geisteswissenschaftliche oder naturwissenschaftliche Fächer hat. Interessieren sich die Zielgruppen auch dafür, wie regelmäßig Alkohol eingenommen wird, ob unter der Woche oder am Wochenende, können die entsprechenden Attribute aus der x- und y-Achse ausgewählt werden, um mehr über das Thema zu erfahren.

Mit Hilfe der Visualisierungstechnik Scatterplot lassen sich zwei Attributwerte in einem zweidimensionalen Raum vergleichen. In den Parallele Koordinaten können dagegen mehr als zwei Attribute miteinander in Bezug setzen. Die Anwender können mithilfe dieser Visualisierungstechnik vier Attribute auswählen und Zusammenhänge zwischen den ausgewählten Variablen untersuchen.

Interessieren sich Anwender für die Untersuchung bestimmter Eigenschaften, können auf vier Achsen die Werte ausgesucht und die Ergebnisse durch mit Linien miteinander verbundene Attribute angezeigt werden. Sobald die Attributwerte aus vier Achsen ausgewählt sind, wird die Linie mit der Frage grün markiert und das Geschlecht der Schüler sowie die Namen der jeweiligen Werte aus vier ausgewählten Variablen angezeigt. Durch diese Visualisierungstechnik können mehrere Trends und unterschiedliche Zusammenhänge in den ausgewählten Werten gefunden werden.

Stickfigureplot ermöglicht eine mehrdimensionale Darstellung der Datensätze. Ein Stickfigure ist aus fünf Bestandteilen aufgebaut, und jeder Bestandteil ist einem numerischen Wert zugeordnet. Im Vergleich zu den anderen vorgestellten Visualisierungstechniken werden hier neben den Schulnoten beider Fächern und dem Alkoholkonsum noch weitere Attribute analysiert. Die Entscheidung, noch weitere Attribute der Datensätze in der Analyse miteinzubeziehen, kann beispielsweise damit begründet werden, dass soziale Aspekte einen Einfluss auf die Leistung haben können. Diese Untersuchung soll den Zielgruppen dabei helfen, neue und relevante Zusammenhänge zwischen den Variablen zu erkennen. Wie bereits erwähnt besteht eine Stickfigure aus fünf Eigenschaften: Körper, rechter oberer Arm, linker oberer Arm, rechter unterer Arm und linker unterer Arm. Dem Körper ist der Attributwert der Bildung der Mutter, dem rechten oberen Arm der Attributwert der Bildung des Vaters, dem linken oberen Arm der Attributwert der Freizeit der Schüler, dem rechten unteren Arm der Attributwert der Abwesenheit von der Schule und dem linken unteren Arm der Attributwert des Alkoholkonsums unter der Woche zugeordnet.



Die Anwender können durch Anklicken einzelner Stickfigures die Attributwerte im Detail betrachten. Wobei zu beachten ist, dass durch beliebiges Auswählen der x- und y-Achse eine unregelmäßige Verteilung der Punkte entstehen kann, und somit die Texturen schwer zu erkennen sind. Um dieses Problem zu beheben, wurde eine Linie aufgebaut, die von links nach rechts gezogen werden kann, um die Länge der Liniensegmente jeder Stickfigure zu bestimmen. Dabei empfiehlt es sich, die Länge sieben auszuwählen, um die einzelnen Plots besser zu erkennen. Außerdem lässt sich diese Aussage nicht verallgemeinern, da die Anwender selbst bestimmen können, mit welchen Längen sie gern arbeiten möchten.

3.2 Anforderungen an die Visualisierungen

In Kapitel 3.1 wurden die Anwendungen der jeweiligen Visualisierungstechniken beschrieben und diesbezüglich sollen die zu erfüllenden Anforderungen in diesem Abschnitt definiert werden. Den Anwendern soll es möglich sein von der Hauptseite zu den verschiedenen Visualisierungen zu gelangen. Darüberhinaus soll eine Navigationsleiste aufgebaut werden, die die Navigation zu den verschiedenen Darstellungen ermöglicht.

Die Anforderungen für die Visualisierung der Daten in einem Scatterplot ergeben sich aus dem Vergleich von zwei verschiedenen Attributwerten. Dabei soll dementsprechend eine Auswahl der Attributwerten in der Scatterplot dynamisch aufgebaut werden. Somit sind die Zielgruppen, in der Lage individuell zu entscheiden, welchen Attributwert sie miteinander vergleichen möchten. Um möglichst viele Zusammenhänge in der Daten zu erkundigen, soll den Anwendern eine große Auswahl an Variablen zu Verfügung gestellt werden. Weiterhin soll für Anwender erkennbar sein,, ob die Daten übereinander liegen. Darüberhinaus sollen die übereinander liegenden Kreise durch Betonung der Farbe stärker hervorgehoben werden, damit die Anwender verstehen können, dass auch andere Punkte auf der selben Position in den Achsenkoordinaten stehen. Um geschlechterspezifische Merkmale in der Visualisierung zu analysieren, sollen die Kreise in zwei

Farben Rot und Blau aufgeteilt werden. Die ausgewählten Kreise sollen durch Anklicken farblich markiert und mit darüber schwebenden Werten angezeigt werden. Im Vergleich zu Scatterplot sind die Anforderungen bei Parallele Koordinaten wesentlich anspruchsvoller, da die Nutzer durch das Wechseln der vier Achsen, verschiedene Zusammenhänge in den Parallele Koordinaten erkunden können. Die Visualisierung soll so aufgebaut werden, dass die Anwender die Auswahl der Achsen selbst bedienen können und aus den Daten Zusammenhänge erkennt. Darüberhinaus sollen Buttons und das Dropdown-Menü übersichtlich beschriftet sein und das Geschlecht der Schüler deutlich durch Farben identifiziert sein. Somit können Anwender in den Parallele Koordinaten durch Anklicken der Buttons und Dropdown-Menü selbst festlegen, beispielsweise welche Schulnoten sie aus welchen Klassen in Zusammenhang mit dem Alkoholkonsum - entweder unter der Woche oder am Wochenende - vergleichen möchten. Die vier ausgewählten Attribute sollen durch eine farbig hervorgehobene Linien mit dadrüber liegenden Werten verbunden und dargestellt werde. Diese Anwendung soll benutzerfreundlich gestaltet werden, damit es den Anwendern leicht fällt, die Attribute auszuwählen und somit x-Einheiten auf der x-Achse und y-Einheiten auf der y-Achse zu identifizieren.

Zu den Anforderungen bei den Stickfigureplots gehört unter anderem der Vergleich der Attribute, wobei die Interessenten die Möglichkeit haben im Gegensatz zu anderen Techniken mehrere Merkmale der Daten gleichzeitig analysieren zu können. Neben dieser Visualisierung soll ein Stickfigure als Demonstration visuell dargestellt werden und die einzelnen Körperteile sollen deutlich beschriftet werden. Zudem soll den Anwendern möglich sein die Größe des Stickfigures mit dem Regler so einzustellen, dass eine Textur sichtbar wird. Weiterhin soll eine Dropdown-Menü aufgebaut werden, damit die Klassenstufen, um die Mathematik- und Portugesisch-Noten mit weiteren Daten in der Stickfigure-Darstellung zu erkunden. Darüberhinaus soll für die Anwender möglich sein, durch Auswählen der Attribute in der Dropdown-Liste, die Verschiebung der visualisierten Stickfigures zu beobachten, um Aussagen über die Signifikanz der Werte machen zu können.

Bei der Auswahl mit der Maus sollen Stickfigures farbig angezeigt und die numerischen Variablen sollen über den Stickfigure übersichtlich dargestellt werden. Da in dieser Projektarbeit nicht davon ausgegangen werden kann, dass Schuldirektoren und die Forschung die benötigenden Vorkenntnisse haben, werden die Datenvisualisierungen übersichtlich und verständlich dargestellt. Somit soll die Beschriftung der Buttons, Dropdown, Navigationsliste, Achsen, Stickfigures, Punkte und Linien überschaubar dargestellt und farblich markiert werden.

Zudem verfügen alle Visualisierungen über Dropdown und eine Navigationsleiste, die deutlich beschriftet sind. Mithilfe der Navigationsleiste können die Anwender zu den beliebigen Visualisierung durch das Anklicken der Navigationsleiste ansteuern. Es ist dabei anzumerken, dass aus Gründen des Platzmangels die Beschriftungen abgekürzt benutzt werden. Die Bedeutung jeder verwendeten Abkürzungen wird in der Datei readme.md erklärt. Die Abkürzungen der jeweiligen Variablen wurde sorgfältig ausgewählt, sodass die Anwender sich intuitiv zurecht finden. Somit können die Anwender instinktiv die Dropdown-Liste auswählen, um die x-und y-Achsen mit verschiedenen Attributen zu verwenden. Sobald eine Eigenschaft aus der Dropdown-Liste ausgewählt wird, werden die x-und y-Achsen automatisch beschriftet. Die Beschriftung der Achsen ist optisch gut lesbar. Um die Komplexität der Visualisierungen mit Interaktionsmöglichkeit benutzerfreundlicher darzustellen, wurden die Schüler für alle Visualisierungen in männlich und weiblich durch zwei Farben unterschieden.

3.3 Präsentation der Visualisierungen

In diesem Unterkapitel werden alle Visualisierungstechniken dargestellt und anhand konkreter Beispiele genauer beschrieben.

3.3.1 Visualisierung Eins

Wie in der Abbildung 1 zu sehen ist, können die Anwender beliebig und intuitiv den Dropdon-Menü bedienen und sich schnell und übersichtlich die zugrundeliegende Daten aufzeigen lassen. Die Abbildung zeigt einen ausgewählten Kreis, der farblich markiert ist. Zu dem Kreis gehören die

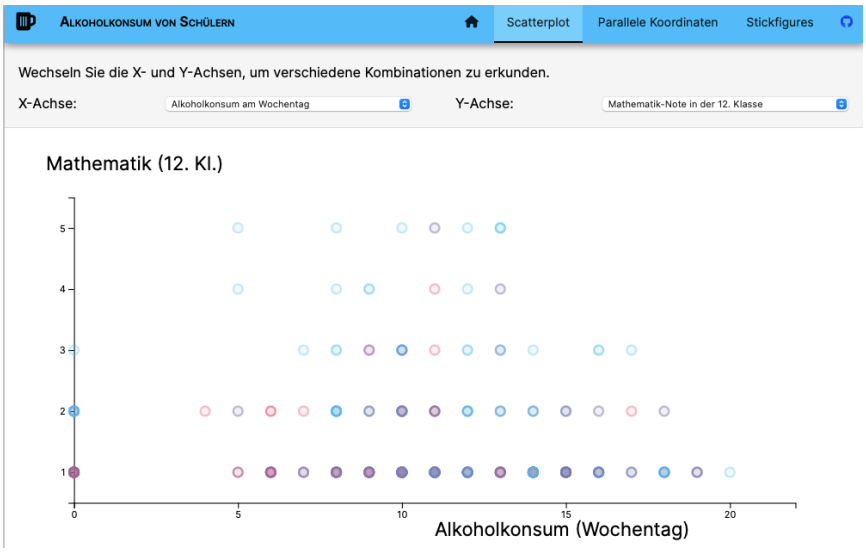


Abbildung 1: Scatterplot (Quelle: eigene Darstellung)

Attributwerte vom Geschlecht des Schülers, Abschlussnote in Mathematik und Alkoholkonsum unter der Woche.

Zudem können Trends beziehungsweise Veränderungen der Entwicklungsrichtung der Daten untersucht werden um daraus Handlungsempfehlungen abzuleiten. Ein wesentlicher Vorteil dieser Visualisierungstechnik besteht darin, dass x- und y-Achsen dynamisch und automatisch angepasst werden können. Diese Visualisierung ist optisch ansprechend für Eine zweidimensionale Darstellung der Daten.

3.3.2 Visualisierung Zwei

Die Anwender können bei der Visualisierung Parallele Koordinaten, neben der Dropdown-Liste Buttons auswählen, um beliebig viele Darstellung der Daten aufzuzeigen. Ein wesentlicher Vorteil dieser Technik ist, dass man vier Attribute auswählen kann. Somit lassen sich mehrere Zusammenhänge von Attributen analysieren und neue Erkenntnisse gewinnen. Wie in der Abbildung 2 zu sehen ist, können beliebige Variablen durch Mausklick ausgewählt werden. Nach dem auswählen der Werte wird langsam in die Visualisierung eine Linie mit rein gezeichnet.

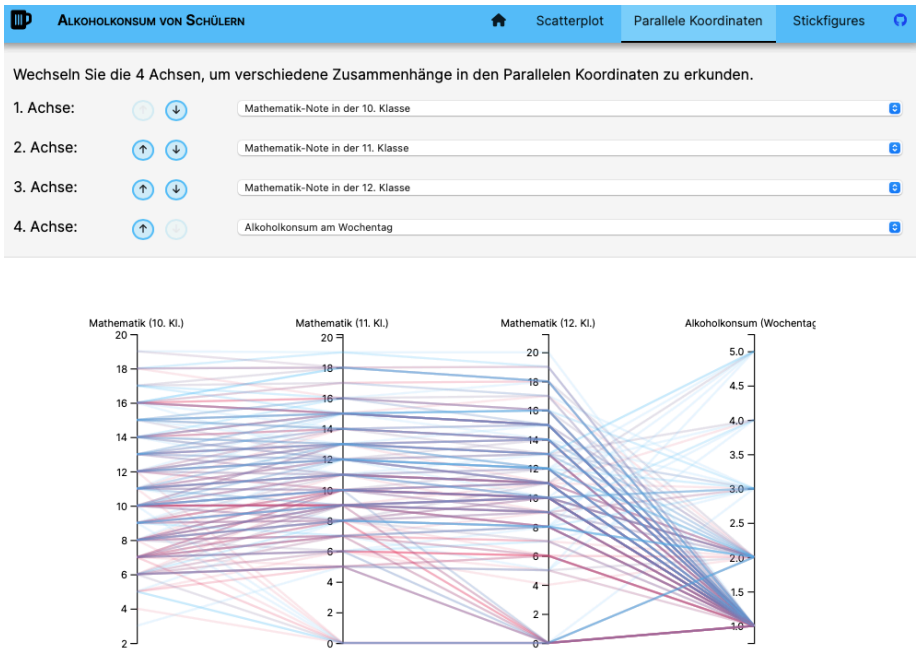


Abbildung 2: Parallele Koordinaten (Quelle: eigene Darstellung)

Diese Linie ist farblich markiert, damit sie von den anderen Farben eindeutig unterschieden werden kann. Die Anwender können die einzelnen Werte der vier gewählten Attribute schnell ablesen. Auf diesem Weg lassen sich Informationen schneller erkennen und einfacher verstehen.

3.3.3 Visualisierung Drei

Bei der Visualisierung Drei geht es um eine Visualisierungstechnik namens Stickfigureplot.
Eine wichtige Eigenschaft dieser Technik besteht darin, dass sich mehrere Attributwerte miteinander vergleichen lassen. Dadurch können mehrere Zusammenhänge zwischen den Variablen untersucht werden, um wichtige Rückschlüsse aus den Datensätze zu ziehen.
Aus der dargestellten Abbildung 3, lässt sich eine ausgewählte Stickfigure und deren Attributwerte deutlich erkennen. Wenn eine bestimmte Stickfigure ausgewählt wurde, wird diese automatisch farblich markiert. Somit kann erkannt werden welche Stickfigure ausgewählt wurde. Weiterhin können Korrelationsbeziehungen identifiziert werden, durch das Vertuschen der x- und y-Achen. Bei dieser Visualisierung werden im Vergleich zu den bisherigen Visualisierungstechniken die x- und y-Achse nicht unabhängig voneinander ausgewählt, sondern die Kombination der Achsen ist hierbei bereits vorgegeben. Diese Entscheidung wird so begründet, dass im Vorfeld viele Alternativen durchgeführt wurden und auf ihre Sinnhaftigkeit mehrfach geprüft wurden, sodass es sich bei diesen Kombinationen bereits um sinnvolle Verknüpfungen handelt.

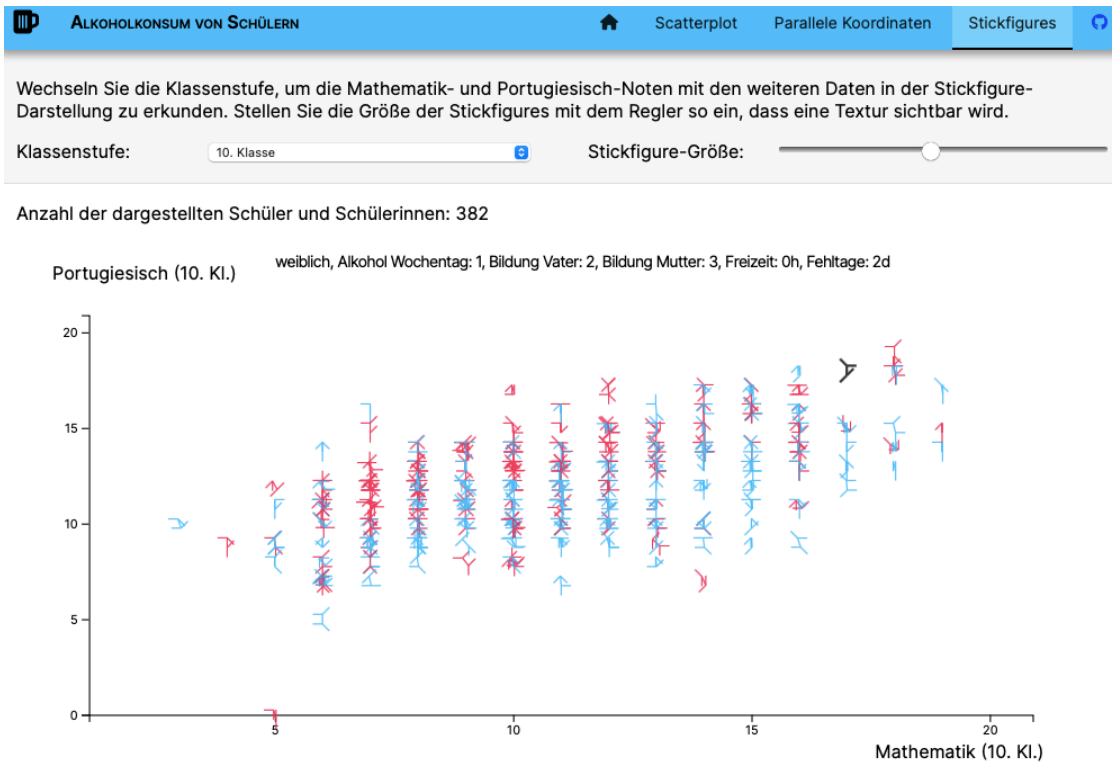


Abbildung 3: Stickfigureplot (Quelle: eigene Darstellung)

3.4 Interaktion

Bei den drei vorgestellten Visualisierungstechniken können viele Interaktionsmöglichkeiten durchgeführt werden. Die Visualisierungstechniken sind so aufgebaut, damit möglichst viele

Interaktion zwischen den Variablen entstehen. Aus der Interaktion der Attributwerte, lassen sich wichtige Erkenntnisse aus den Daten extrahieren. Bei den Visualisierungstechniken können viele Interaktionsmöglichkeiten durch beliebiges auswählen der x- und y-Achse entstehen. Im Gegensatz dazu sind die Interaktionsmöglichkeiten der Stickfigureplot etwas reduzierter, da der Austausch der Achsen vordefiniert ist. Dabei soll betont werden, dass durch das Gegenüberstellen der sieben Attribute die Stickfigureplot möglichst interessante und wichtige Zusammenhänge entstehen können. Daten visuell und interaktiv darzustellen, gibt den Zielgruppen bessere Möglichkeiten, gewisse Muster und Trends zu erkennen. Die Zielgruppen können individuell festlegen, aus welcher Darstellung der Daten sie welche Erkenntnisse ziehen. Viele Attribute aus den Datensätzen weisen wichtige Kennzahlen auf, die besonders für die Forschung interessant sind, um Hypothesen aus der Darstellungen zu generieren. Die Anwender können durch Beobachtung bestimmte Muster beziehungsweise Korrelationsbeziehungen der Datensätze neue Einblicke gewinnen, um beispielsweise Handlungsempfehlungen aus den Visualisierungstechniken zu extrahieren. Aus der Visualisierungen von Scatterplot und Stickfigureplot können zudem erkannt werden, ob ein linearer Zusammenhang zwischen zwei Variablen besteht. Darüberhinaus können Aussagen über bestimmten Merkmale getroffen werden.

Zu den unterschiedlichen Visualiesierungstechniken wird eine Webseite aufgebaut, die alle ausgewählten Visualisierungen der Daten umfasst. Die Anwender können jederzeit die entsprechende Visualisierungen abrufen. Durch Anklicken der Stickfigureplot, Parallele Koordinaten und Scatterplot können die Visulalisierungstechniken ausgewählt werden und die Anwender werden auf die entsprechenden Seiten weitergeleitet, die sie ausgewählt haben. Durch Eingabe des Befehls “Elm make src/Main.elm” zusammen mit dem absoluten Pfad der jeweiligen Visualisierungen in Visual Studio Code, konnte eine Datei mit dem Name index.html erzeugt werden, welche die Darstellung der Stickfigureplot, Parallele Koordinaten und Scatterplot als Webseite ermöglicht.

4. Implementierung

Um die Visualisierungstechniken zu programmieren ist es notwendig Programmierkenntnisse in der Sprache Elm zu besitzen. Zudem werden sowohl mathematische als auch theoretische Fachkenntnisse benötigt. Für die Implementierung waren die erworbenen Kenntnisse in den Übungen und Vorlesungen des Moduls “Information Retrieval und Visualisierung“ sehr hilfreich. Neben des zur Verfügung gestellten Lerninhalts aus diesem Modul wurde eine Literaturrecherche

mithilfe der Online-Bibliothek der Universität Halle-Wittenberg und Google Scholar durchgeführt. Weiterhin wurde die Dokumentation des Elm Codes von der Webseite “elm-lang.org" genutzt. Zudem wurden sowohl YouTube-Videos als auch Github-Verlauf von verschiedenen Authoren als Hilfestellung verwendet. Für die Analyse der Daten aus der Webseite “Kaggle” wurden Fachbücher aus dem Fachbereich der Data Science verwendet.

Hinsichtlich der Programmierung des Scatterplots wurden die Inhalte aus den Übungen eins bis drei verwendet. Zur Erstellung der Parallele Koordinaten wurde als Grundlage für den Code die Lerninhalte aus der Übung sechs und sieben genommen. Zudem wurde auch eine Literaturrecherche durchgeführt, die als Hilfestellung genutzt wurde, um bestimmte Funktionen in der Programmiersprache Elm zu definieren. Der Inhalt des Codes der dritten Visualisierungstechnik Stickfigure basiert auf den Lerninhalten der Übung acht und der im StudIP zur Verfügung gestellten Lerninhalte. Die Programmierinhalte der drei dargestellten Visualisierungstechniken werden so strukturiert, dass die Anwender eine Übersicht zu den geschriebenen Codes erhalten. In vielen Abschnitten in der Programmierung sind Kommentare hinzugefügt, damit der Code für Außenstehende gut lesbar und nachvollziehbar ist.

Für den Aufbau der Visualisierungen wurden verschiedene Module aus der Webseite elm-lang.org lokal auf dem Rechner installiert. Diese Module werden zu dem Projekt im Visual Studio Code (VSC) “Student Alcohol Consumption” hinzugefügt um bestimmte Parameter und Funktionen zwecks der Implementierung zu verwenden. In diesem Abschnitt werden die wichtigsten Funktionen zusammengefasst, die fundamental für die Visualisierung der Daten in Elm sind. Für die drei Techniken wurde eine Datei namens Main gebaut, in welcher Funktionen der Browser.element: init für die Initialisierung, die Sektion view, subscriptions, und update definiert wurden. Diese Funktionen sind innerhalb der Main zu finden. Die Interaktionsmöglichkeiten sind in Main dynamisch aufgebaut und Anwender können anhand der dargestellten Buttons und Navigationsleiste auf die entsprechenden Visualisierungen gelangen.

Die Buttons dienen dazu zwischen den Merkmalen hin und her zu springen um möglichst viele Zusammenhänge in den Daten zu erkunden.

Zusammengefasst ist Main für das Switchen von den Visualisierungen und Bündeln von Nachrichten, die aus der jeweiligen Visualisierungen gesendet werden, verantwortlich. Dieses Bündeln erfolgt durch das Mapping und das Wrapping jedes Models, jeder Message, und jedes Views der drei Visualisierungen.

Für das Laden der Datensätze in Elm wird der Link von Github genommen, in welchem die CSV-Daten hochgeladen sind. Die Decodierung der entnommenen Datensätzen aus Github wird mithilfe

der CSV-Decoders, als “decodingStudentAcoholConsumption” definiert, erfolgen. Weiterhin wird eine weitere Funktion update benutzt, damit die Anwender die x- und y-Achsen beliebig austauschen können. Für die Visualisierung der Darstellungen wurden zwei Techniken verwendet: SVG und CSV.

Die bereits erwähnten Funktionen, entnommen aus den Übungsinhalten, mussten bei der Programmierung der Visualisierungstechniken angepasst werden. Dadurch dass mehrere Attribute zur Analyse herangezogen wurden, mussten Anpassungen der Übungsinhalten für den Projektbericht vorgenommen werden.

Notwendige Parameter wurden definiert, um Interaktion zwischen den Visualisierungen zu ermöglichen. In jeder Visualisierung wurde der Parameter Chosen Data, der innerhalb type alias Data gehört, definiert, das den type Maybe Studentalcoholconsumption hat. Der Parameter Chosen Data wurde für jeden View, Update und Message definiert um bestimmte Ereignisse zu empfangen. Weiterhin wurde der Parameter PointChosen, der innerhalb type MSG zu finden ist, definiert um einen bestimmten Datensatz aus der Visualisierung auszuwählen. Die Parameter sind so definiert, damit Interaktion zwischen den Visualisierungen entstehen kann. Die definierten Parametern und Funktionen ermöglichen das bündeln den Nachrichten zwischen den Visualisierungen und werden durch das Main gesteuert. Die Anwender haben die Möglichkeit bestimmte Datenpunkte durch Mausklick von der Visualisierung Stickfigureplot auszuwählen. Wenn aus dem Stickfiguteplot raus gewechselt wird, werden die gewählten Daten aus dem Stickfigureplot für den Scatterplot updated.

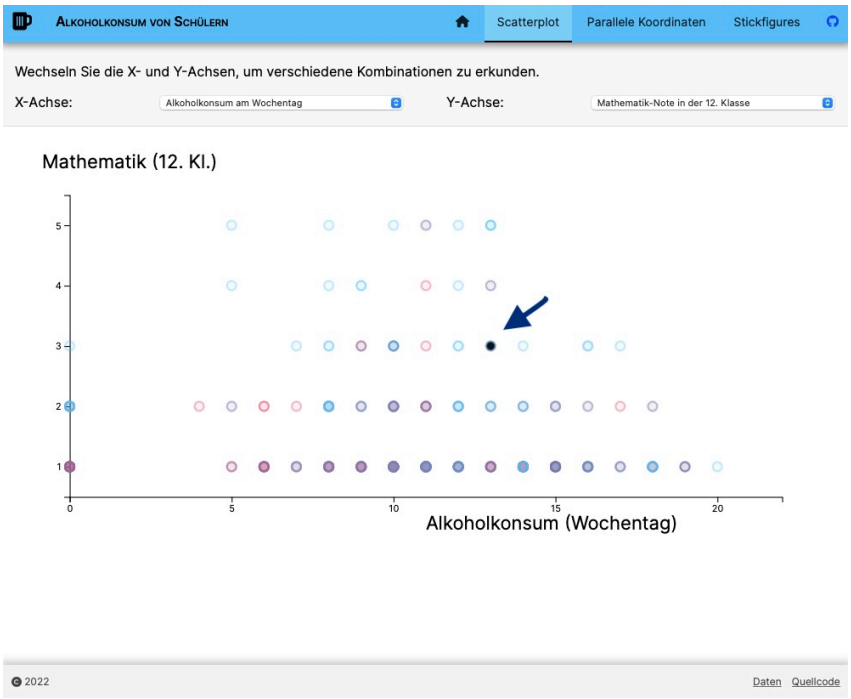


Abbildung 4: Stickfigureplot (Quelle: eigene Darstellung)

Darüber hinaus wurden in der Datei Main mögliche Anpassungen innerhalb Update sowohl für Stickfigureplot als auch Scatterplot vorgenommen.

Abbildungen 4 und 5 stellen die Interaktion zwischen der zwei Visualisierungen dar.

Die Interaktion ist dynamisch aufgebaut und die Anwender können individuell entscheiden, welche Interaktionsmöglichkeit zwischen der Stickfigureplot und Scatterplot sie gerne untersuchen möchten. Es ist dabei anzumerken, dass nur eine Interaktion zwischen der Visualisierungen Stickfigureplot und Scatterplot möglich ist.

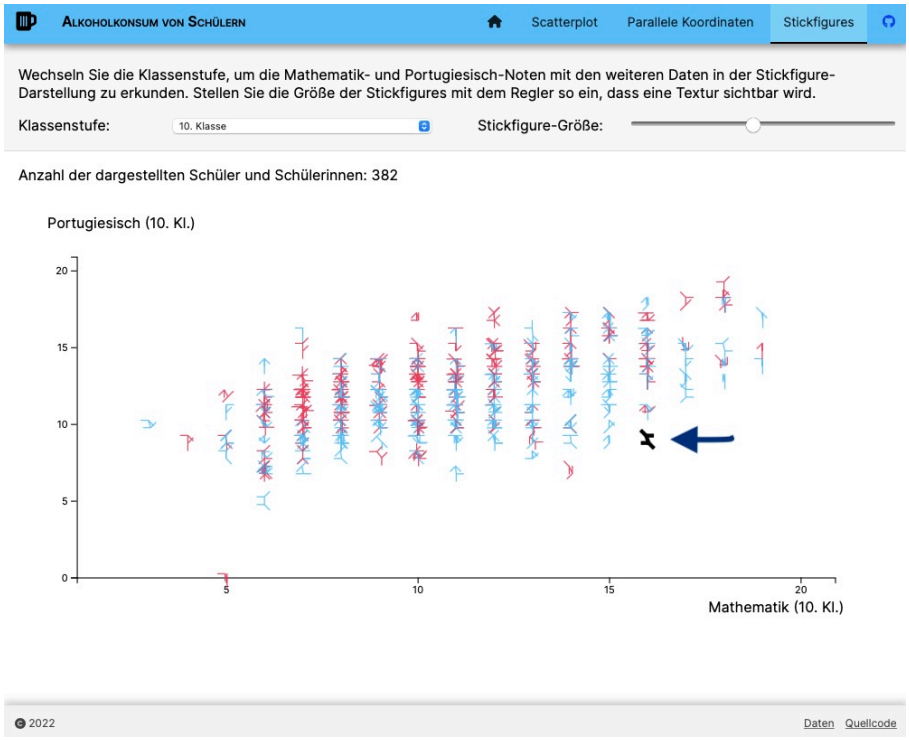


Abbildung 5: Scatterplot (Quelle: eigene Darstellung)

Der Code für alle drei Visualisierungstechniken befindet sich in einem Unterordner mit dem Name src aus dem Gesamtordner GitHub. Als Grundlage für CSS, also das Design der Visualisierungen und Html u.a. dargestellten Button, Dropdown und Navigationsleiste, wurden neben der Inhalte auf der Webseite Elm, eine Literaturrecherche durchgeführt und oft wurden Beispiele in unterschiedlichen GitHub-Projekten die für das Visualisierungsprojekt hilfreich sind, gefunden.

5. Anwendungsfälle

Um den Anwendern mit den vorgestellten Visualisierungen vertraut zu machen, werden in diesem Kapitel Anwendungsbeispiele beschreiben, die für entsprechenden Zielgruppen mögliche Anwendungsfälle darstellen.

5.1 Anwendung Visualisierung Eins

Bei der Visualisierung Eins wird der Output der y-Achse, welche für die Abschlussnote aus dem Schulfach Mathematik steht in Verhältnis zu der x-Achse, welche für den Alkoholkonsum unter der Woche steht dargestellt. Sobald ein Kreis ausgewählt wird, werden die entsprechenden Attributwerte angezeigt. Betrachtet man die Datensätze als Ganzes kann man erkennen im welchen Notenbereich die Schüler liegen, in Abhängigkeit des Ausmaßes des Alkoholkonsums. Zudem kann man aus den Daten erkennen, ob sich mögliche Ausreißpunkte aus den Daten identifizieren lassen.
Die Abbildung 6 dient der Zielgruppen Schuldirektoren, Forschung und Data Science, um beispielsweise die Frage zu beantworten, wie der Alkoholkonsum unter der Woche die Abschlussnoten in dem Studienfach Mathematik beeinflusst. Wie in der Abbildung zu sehen ist, liegen die meisten Datenpunkte zwischen den Alkoholstufen eins und zwei und die erreichten

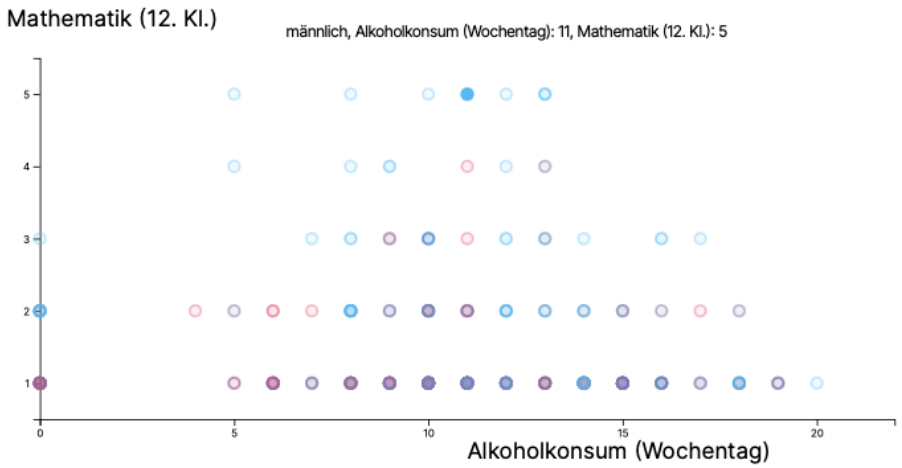


Abbildung 6: Scatterplot (Quelle: eigene Darstellung)

Abschlussnoten zwischen 15 und 17 Punkten. In welchem Bereich die meisten Schüler liegen, ist farblich gekennzeichnet. Die Betonung der Farbe des Außenkreises der dargestellten Datenpunkte, ist ein Hinweis dafür, dass mehrere Datensätze übereinander liegen. In der folgenden Abbildung kann auch erkannt werden, dass bei einigen Kreisen im oberen rechten Bereich die Farbe heller ist was bedeutet, dass nur wenig Schüler die Alkoholstufe fünf erreichen und im Vergleich zu den anderen Schülern, deren Alkoholkonsum in dem Bereich eins bis zwei liegt, schlechtere Noten bekommen.
In der Scatterplot kann durch eine Stichprobe herausgefunden werden, welche Geschlechter die höchsten Alkoholstufen erreichen, indem die Anwender durch das Anklicken der dargestellten Kreise, die im Bereich der Alkoholstufe fünf liegen, das Geschlecht der Schülern in Erfahrung

bringen. Darüberhinaus kann ein Schuldirektor erkennen, welche Alkoholstufe die meisten Schüler erreichen und wie die Abschussnote in Zusammenhang mit dem Alkoholkonsum aussieht. Die Forschung kann aus den dargestellten Datenpunkte den Zusammenhang der Abschlussnoten mit dem höchsten Alkoholkonsum unter der Woche untersuchen. Wie in der Abbildung zu sehen ist erreicht nur ein kleiner Teil der Schüler die Alkoholstufe fünf und die Noten schwanken zwischen 10 und 20. Data Science kann zum Beispiel interessieren, ob Ausreißer zu finden sind oder welche Aussagen allgemein über die Daten gemacht werden können .

Alternativ zu dieser Darstellung kann ein Säulendiagramm dargestellt werden. Allerdings sollen die Datensätzen bei dieser Diagrammart gruppiert werden. Denkbar ist es, die Schülerleistung in Zusammenhang mit dem Alkoholkonsum von Schülern getrennt voneinander zu visualisieren.

5.2 Anwendung Visualisierung Zwei

Durch Analyse der Visualisierungstechnik Parallele Koordinaten, können vier Merkmale in Betracht gezogen werden und interessante Erkenntnisse gewonnen werden.

In der Abbildung sind die Schulnoten von zehnter, elfter und zwölfter Klasse dargestellt in Zusammenhang mit dem Alkoholkonsum unter der Woche.

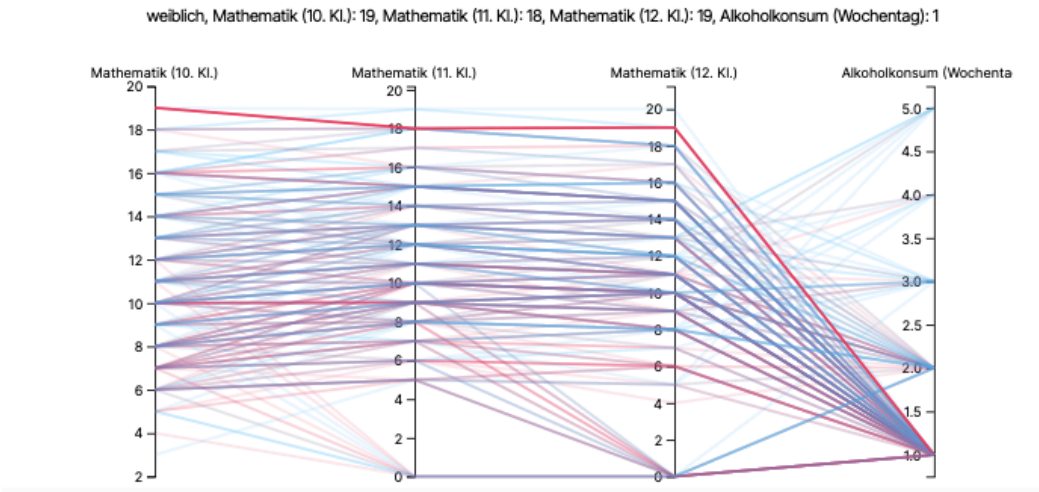


Abbildung 7: Parallele Koordinaten (Quelle: eigene Darstellung)

Die Anwender können aus der dargestellten Technik beobachten, wie die erreichten Schulnoten aus der zehnten und elften Klasse in Zusammenhang mit dem Alkoholkonsum in die Gesamtnote mit einfließen. Die Daten können einzeln und verbunden betrachtet werden.

Der Schuldirektor aus der vordefinierten Zielgruppe kann sich für ein Gesamtbild hinsichtlich der Ergebnisse der Schüler interessieren. Er kann die Entwicklung der Schulnoten aus der Alkoholstufe

1 und 2 beobachten. Auf der Abbildung kann auf den ersten Blick erkannt werden, dass die Schüler bis zur Alkoholstufe 1, die besseren Schulnoten für die zehnte, elfte und zwölfte Klasse erreichen.
So kann auch festgestellt werden, dass diejenigen, die in der zehnten und elften Klasse sehr gute Noten erreichen und in dem Bereich der Alkoholstufe 1 bleiben, somit die besten Noten haben. Aus der dargestellten Visualisierungstechnik kann zudem auch erkannt werden, dass die Schüler, die in der zehnten und elften Klasse in dem Bereich von 0 bis 9 liegen, auch diejenigen sind, die einen Schulabschluss nicht schaffen. Dadurch können die Schuldirektoren zukünftig für diejenigen, die in der zehnten Klasse schlecht abschneiden zusätzlich Schulunterricht in der Schule anbieten, damit die Schüler in der kommenden elften und zwölften Klasse die Möglichkeit erhalten die Schulnoten zu verbessern. Für die Forschung wäre zudem auch interessant zu untersuchen, ob die Schulnoten von der zehnten, elften und zwölften Klasse stark voneinander abhängen. Aus den dargestellten Daten können Stichproben gezogen werden um herauszufinden, ob beispielsweise die Schüler, welche eine schlechte Note in der zehnten Klasse bekommen, sich in den folgenden Klassen verbessert haben oder gleich geblieben sind. Für die Zielgruppe Data Science sind ebenfalls die Untersuchungen der Zusammenhänge der Variablen von Interesse.
So können bestimmte Gruppen von Schülern ausgehend von erreichten Schulnoten in der zehnten, elften und zwölften Klasse in Abhängigkeit mit der Alkoholstufe untersucht werden.
Wie in der Abbildung 7 zu sehen ist, werden die besten Abschlussnoten von den früheren Noten aus der zehnten und elften Klasse beeinflusst und das erreichte Trinkniveau beträgt 1. Die höchste Abschlussnote, die Schüler in der 5. Alkoholstufe erreicht haben beträgt 13. Aus den Daten kann erkannt werden, dass mit der Erhöhung der Alkoholstufe die Abschlussnoten deutlich schlechter ausfallen. So können die gewonnenen Erkenntnisse aus den Daten öffentlich zugänglich gemacht und für spätere Reports genutzt werden. Reports können wiederum für die Wissenschaft sehr hilfreich sein damit mögliche Abweichungen in den untersuchten Variablen aufgedeckt werden. Alternativ zu dieser Darstellung kann eine Liniengraph visualisiert werden. Jedoch ist diese Visualisierungsart nicht gut für die Datensätze geeignet, weil die Linien übereinander liegen können. Denkbar wäre vor der Visualisierung mögliche Änderungen vorzunehmen, zum Beispiel die Datensätze in männliche und weibliche Schüler zu gruppieren.

5.3 Anwendung Visualisierung Drei

Die Visualisierung Drei Stickfigureplot versucht die Frage zu beantworten, welchen Einfluss die Bildung der Eltern in Kombination mit Alkoholkonsum unter der Woche, Freizeit und Abwesenheit auf die Abschlussnote hat. Den Zusammenhang zwischen diesen Variablen zu untersuchen ist sehr

wichtig, da die Abschlussnote ein Entscheidungskriterium für eine Weiterführung in Form eines Studiums an einer Universität ist. Wie in der dargestellten Abbildung 8 zu sehen ist, lassen sich sieben einzelne Variablen der jeweiligen Schüler miteinander vergleichen.

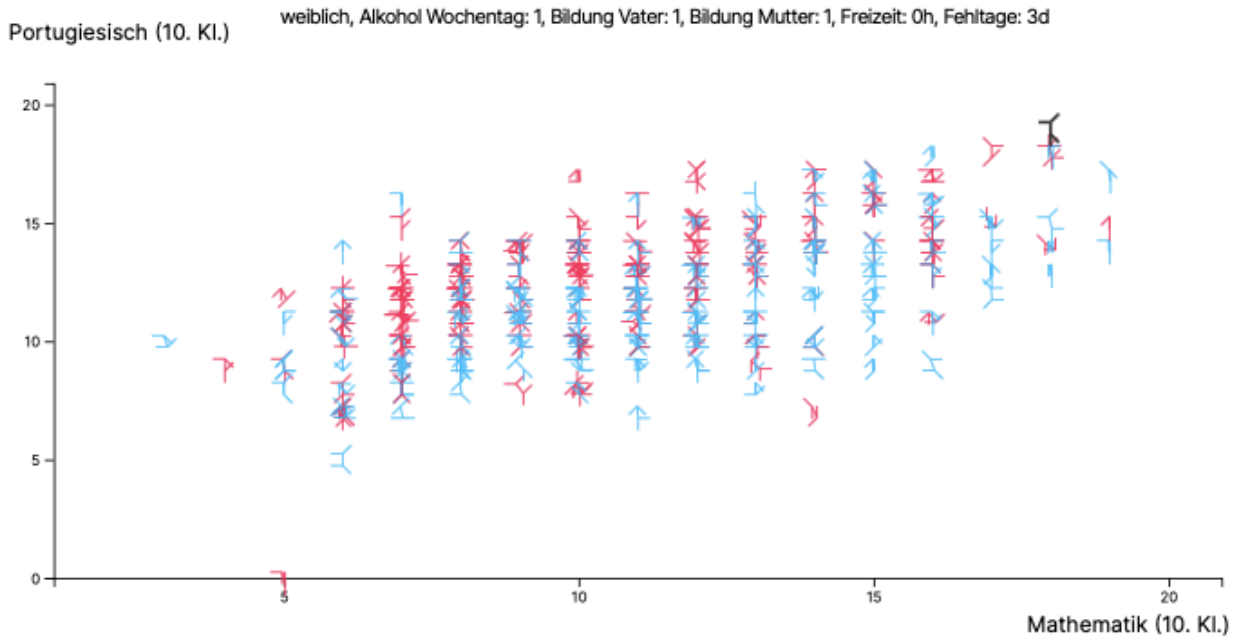


Abbildung 8: Stickfigureplot (Quelle: eigene Darstellung)

Zudem können die Daten als Ganzes betrachten werden um zu sehen welche Korrelationsbeziehungen sich aus der dargestellten Visualisierung ableiten lassen. Darüberhinaus lässt sich aus der Abbildung 8 feststellen, dass die erreichten Schulnoten für Mathematik und Portugiesisch aus der zehnten Klasse mit den anderen dargestellten Variablen (Alkoholkonsum, Freizeit nach der Schule, Abwesenheit und die Bildung der Eltern) positiv korrelieren. Mithilfe dieser Visualisierung lassen sich korrelative Beziehungen identifizieren. In diesem Fall kann aus der dargestellten Abbildung erkannt werden, dass die Datensätze einen positiv linearen Zusammenhang darstellen. Somit kann die Zielgruppe Data Science Erkenntnisse über Korrelationsbeziehungen der Daten gewinnen. Stichprobenartig kann durch das Anklicken der jeweiligen Stickfiguren festgestellt werden, dass diejenigen Schüler, die öfter in der Schule fehlen auch viel Alkohol konsumieren, sich nach der Schule viel Freizeit nehmen und deren Eltern keinen Akademischen Abschluss haben, schlechte Schulnoten in beiden Studienfächer erreichen.. Zudem gibt es auch Schüler, die trotz des niedrigen Bildungsniveau der Eltern gute Note in der Schule erzielen. Jedoch kann aus dieser Visualisierung nicht auf denersten Blick erkannt werden, ob das Bildungsniveau der Eltern ein Problem für die Leistung darstellt. Diese Analyse der Daten kann für die Zielgruppe der Schuldirektoren nützlich sein, damit zukünftig Methoden in der Lehre eingeführt werden, um

Schüler, die nicht aus akademischer Familie stammen, in der Schule zu unterstützen. Zudem könnten diese Erkenntnisse der Forschung dabei helfen, die Fragen zu beantworten, wie und warum die demographischen, sozialen und schulbezogenen Merkmale die Schülerleistungen beeinflussen. Alternativ zu dieser Visualisierung kann eine Graph-Darstellung namens Sugiyama-Methode dargestellt werden, um die Beziehung zwischen den Variablen genauer zu untersuchen. Jedoch wäre es hier von Vorteil die Schülerleistung in Abhängigkeit des Bildungsniveau der Eltern zu untersuchen. Eine weitere Alternative wäre eine Sternplotvisualisierung, die eine Ähnlichkeit mit der dargestellten Visualisierung Stickfigureplot aufweist.

6. Verwandte Arbeiten

Die Datensätzen, die in dieser Forschungsarbeit verwendet wurden, finden auch in anderen wissenschaftlichen Visualisierungen oder vergleichbaren Projektarbeiten Anwendung. In diesem Abschnitt wird ein Vergleich zu einigen veröffentlichen Projektarbeiten, in denen eine ähnlich Analyse durchgeführte wurde, gemacht. Zuerst wird eine wissenschaftliche Arbeit, in welcher die Machine learning Methoden angewendet wurde, um die Abschlussnoten der Sekundarschule zu vorhersagen, beschrieben und mit den vorgestellten Visualisierungstechniken verglichen. Diese Forschungsarbeit verwendet vier Methoden zur Vorhersage der Schülerleistung: Data Mining (DM)-Methoden, Entscheidungsbäume (DT), Random Forests (RF), Neuronale Netze (NN) und Support

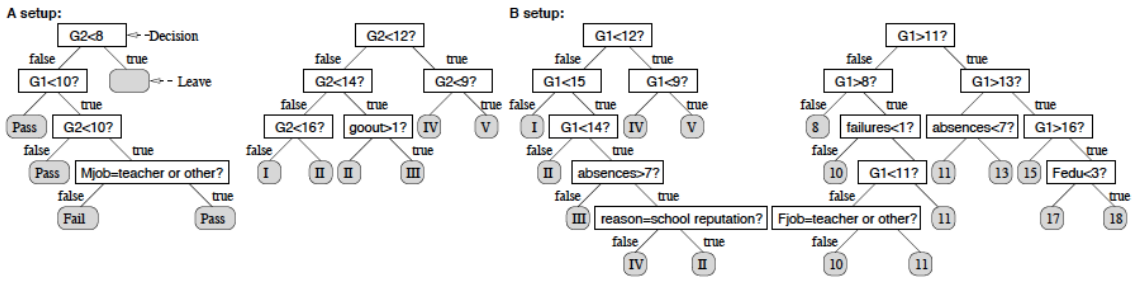


Abbildung 9: Beispiel Decision Trees (A-Por-Bin, A-Por-5L, B-Por-5L and B-Por-Reg) Vgl. Cortez und Silva

Vector Machines (SVM).[12] Eine der verwendeten Methoden Decision tree, ist ein Entscheidungsbaum-Algorithmus, der Kennzahlen der Zielvariablen des Entscheidungsprozess voraussagt.[13] Diese Methode befasst sich in dem wissenschaftlichen Artikel mit der Untersuchung der Zusammenhänge der Attributwerte: Schulnoten von der zehnten bis zwölften Klasse und Bildungsniveau der Eltern, um die Abschlussnote vorauszusagen.

[12] Vgl. Cortez & Silva (2008)

[13] Vgl. Decision Tree Algorithm Examples in Data Mining (2022)

Eine ähnliche Funktionalität weist die Baumhierarchie auf, welche die Forschungsfrage beantwortet, wie die Bildung der Eltern ausgehend von der Schulnoten von erster und zweiter Klasse, die Abschlussnoten beeinflusst. Diese Visualisierungsart wurde aufgrund der hohen Komplexität des Datensatzes in dem Projekt nicht umgesetzt.

Denkbar ist für diese Technik die Schulnoten zu gruppieren, zum Bespiel von 0-9, 10-11,12-13,14-16 und 17-20. Eine Übersetzung der Schulnoten kann wie folgend aussehen: 0-9 soll für 5,0; 10-11 für 3,7; 12-13 für 2,7; 14-16 für 2,3-1,7; ,17-20 für 1,3-1,0 stehen.[14] Vergleichbar zur Baumhierarchie ist die Graph-Darstellung Sugiyama-Methode.

Es ist anzumerken, dass mithilfe der in dieser Arbeit ausgewählten Visualisierungstechniken keine Aussage zur Voraussage der Schülerleistungen getroffen werden kann. Mithilfe dieser Techniken können Attribute verglichen werden oder Zusammenhänge untersucht werden um beispielsweise unter anderem herauszufinden, wie der Alkoholkonsum und die erbrachten früheren Leistungen die Abschlussnote beeinflussen.

Auf der Webseite RPubs von RStudio-Code wurde eine vergleichbare Projektarbeit gefunden, die auf fünf Forschungsfragen aufbaut. Eine dieser Forschungsfragen untersucht unter anderem, ob zugängliches Internet zu Hause und der Alkoholkonsum unter der Woche die Schulnoten beeinflussen. So kann der Fokus der Untersuchung auch auf die Lernumgebung zum Beispiel durch zugängliches Internet zu Hause gelegt werden, um zu neuen Erkenntnisse zu gelangen. Eine interessante Erkenntnisse dieser Projektarbeit ist, dass die weiblichen Schüler mehr dazu neigen Unterstützung bei Lernschwierigkeiten im Vergleich zu den männlichen Schüler zu bekommen.[15]

Neben den vorgestellten Studien bzw. Visualisierungsprojekten wurden drei andere Studien gefunden, in denen einige statistische Methoden verwendet wurden, die den Zusammenhang des Alkoholkonsums mit dem Lerneffekt in der Schule untersuchen. Darüberhinaus wird in den Studien berichtet, dass der Alkoholkonsum nicht immer auf die Beeinträchtigung der Schulnoten hindeutet. Der Lerneffekt in der Schule wird oft von dem Ausmaß an Alkohol beeinflusst und die Schülerleistung kann geschlechtsspezifisch unterschiedlich ausfallen.[16]

7. Zusammenfassung und Ausblick

Mithilfe der dargestellten Visualisierungstechniken erhalten die Anwender einen umfangreichen Überblick zum Thema vom Einfluss des Alkoholkonsums auf die Schulnoten in Portugal. Zudem

---

[14] Vgl. uni-potsdam

[15] Vgl. RPubs. (2020)

[16] Vgl. Balsa, Giuliano & French (2011), DeSimone & Wolaver (2005), Dee & Evans (2003)

ermöglichen die Techniken Interaktive Möglichkeiten ausgewählter Attributwerte und liefern interessante Erkenntnisse bezüglich der Untersuchung vom Einfluss des Alkoholkonsums und dendemographischen und sozialen Merkmalen auf die Schulnoten.

Hinsichtlich der vorgestellten Visualisierung Scatterplot können Informationen über die ausgewählten Attribute angezeigt werden und zudem lässt sich anhand der ausgewählten Farben eindeutig identifizieren, um welches Geschlecht es sich handelt. Weiterhin liefert diese Technik Erkenntnisse über die Korrelationsbeziehungen der visualisierten Datensätze. Die wesentliche Eigenschaft dieser Visualisierungstechnik ist der Vergleich zweier Attributwerte in einem zweidimensionalen Raum. Diese Technik ist so aufgebaut, dass die Anwender die Auswahl der x- und y-Achsen beliebig wählen können. Durch diese Interaktionsmöglichkeit, lassen sich schnell und intuitiv Vergleiche zweier Werte machen. Anhand der Visualisierungstechnik Parallele Koordinaten können gleichzeitig vier Attributwerte miteinander vergleichen werden. Bei dieser Technik können die Anwender entscheiden, welche Interaktionsmöglichkeiten sie untersuchen möchten. Zum Beispiel kann mithilfe dieser Technik, wie bereits im Kapital 5 erwähnt, festgestellt werden, ob ein Einfluss von dem Alkoholkonsum auf die Schulnoten ausgehend von der zehnten bis elften Klasse der Sekundarschule auf die Abschlussnoten bestehet. In Kapitel 5 wurde ebenso festgestellt, dass die Schulnoten aus der ersten und zweiten Klasse, die Abschlussnoten stark beeinflussen. Es ist anzumerken, dass mithilfe dieser Technik keine Voraussage der Schülerleistungen getroffen werden kann.

Die dritte Visualisierungstechnik Stickfigureplot stellt die Datensätze in einem mehrdimensionalen Raum dar. Darüberhinaus werden in dieser Darstellung neben schulbezogener Leistung, soziale und demographische Merkmale in Erwägung gezogen. Die Anwender können Einblicke darüber erhalten, wie die sozialen und demographischen Faktoren in Zusammenhang mit der Abschlussnoten stehen.

Zu den verwendeten Visualisierungstechniken in dieser Projektarbeit können auch mögliche Anpassungen vorgenommen werden. Denkbar ist es, die Daten für Schüler getrennt zu visualisieren, also in männliche und weibliche Schüler. Weiterhin kann auch neben Buttons, Dropdown- und Navbar-Menü, zusätzlich ein Drag and Drop oder auch Flyout-Menü dargestellt werden. Eine Mischung der Elemente - Navbar-, Flyout-Menü und Drag and Drop - zur Zuweisung des Attributs auf der x- und y-Achse ist auch eine interessante Darstellung. Bei der Stickfigureplot-Darstellung ist es denkbar, die Stickfigures so darzustellen, dass, sobald die Anwender eine bestimmte Stickfigure auswählen, die anderen Stickfigures aus der Visualisierung ausgeblendet

werden. Somit werden nur die ausgewählten Stickfigure im Vordergrund stehen und die zugehörigen Attributwerte werden zu den einzelnen Bestandteile der Stickfigure positioniert.
Eine Überlegung ist auch, andere Variablen in den vorgestellten Visualisierungen in Erwägung zu ziehen, zum Beispiel demographische Merkmale wie die Zeit, die die Schüler von zu Hause bis zu der Schule benötigen.

Litterature:

Joos Th., & Schmitz P. (2021): Sicherheit in Oracle-Datenbanken konfigurieren. Security-insider: Vogel Communications Group.URL:https://www.security-insider.de/sicherheit-in-oracle-datenbanken-konfigurieren-a-1031255/, Abruf am 10.03.2022.

RPubs. (2020): Students Performance URL: https://rstudio-pubs-static.s3.amazonaws.com/628286_10dbbdb732b54309b941d2dfb5188efb.html#eda, Abruf am 10.03.2022.
Rausch

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Koler P., (2014): Rausch und Identität Jugendliche in Alkoholszenen. 1. Auflage. URL: https://www.forum-p.it/smartedit/documents/inhaltelements/_published/Rausch&Identität.pdf, Abruf am 15.08.2022.

U.S. Department of Health & Human Services (2016): Why Do Adolescents Drink, What Are the Risks, and How Can Underage Drinking Be Prevented? URL: https://pubs.niaaa.nih.gov/publications/AA67/AA67.pdf, Abruf am 07.09.2022.

Notenumrechnung Portugal. URL: https://www.uni-potsdam.de/fileadmin/projects/international/docs/Notenumrechnung_Länder/Notenumrechnung_Portugal_IO.pdf, Abruf am 09.09.2022.

Doerfel A. Helmholtz-Gemeinschaft (2018): Wenn das Gehirn trunken ist. URL: https://www.helmholtz.de/newsroom/artikel/wenn-das-gehirn-trunken-ist/, Abruf am 09.09.2022.

Yi M. A Complete Guide to Scatter Plots. URL: https://chartio.com/learn/charts/what-is-a-scatter-plot/, Abruf am 09.09.2022.

Last Updated (2022): Decision Tree Algorithm Examples In Data Mining. URL: https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/, Abruf am 09.09.2022.

Yanasagaran S., Tan J., Che J. & Adinda A. (2018): RPubs - R PROJECT 1 (ENVX1002/ DATA1001) URL: https://rpubs.com/Jonchee0903/377825, Abruf am 09.09.2022.

Pickett M. R. & Grinstein G. G. (1988): Iconographic Displays For Visualizing Multidimensional Data. URL: https://www.researchgate.net/publication/3767640, Abruf am 09.09.2022.

Gemignani Z. (2021): Better Know a Visualization: Understanding Parallel Coordinates Charts. URL: https://www.juiceanalytics.com/writing/writing/parallel-coordinates, Abruf am 09.09.2022.

Balsa A. L., Giuliano L.M., & French M T. (2011): The effects of alcohol use on academic achievement in high school. NIH-PA Author Manuscript, Abruf am 09.09.2022.

DeSimone J. & Wolaver A. (2005): Drinking and Academic Performance in High School. Working Paper 11035 http://www.nber.org/papers/w11035, Abruf am 30.09.2022.

Dee Th. S. & Evans W. N. (2003): Teen Drinking and EducationalAttainment: Evidence from Two-Sample Instrumental Variables Estimates. Journal of Labor Economics, vol. 21, no. 1, University of Chicago.

Soyka, M. (2001): Serie - Alkoholismus: Psychische und soziale Folgen chronischen Alkoholismus. URL: https://www.aerzteblatt.de/archiv/29088/Serie-Alkoholismus-Psychische-und-soziale-Folgen-chronischen-Alkoholismus, Abruf am 28.09.2022.

Gonçalves I. A., & de Sousa Carvalho A. A. (2017): Pattern of alcohol consumption by young people from North Eastern Portugal. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC5836530/pdf/med-12-494.pdf, Abruf am 30.08.2022.

Stumpp G., Stauber B., & Reinl H. (2009): Einflussfaktoren, Motivation und Anreize zum Rauschtrinken bei Jugendlichen. URL: https://www.berlin-suchtpraevention.de/wp-content/uploads/ 2016/10/2009_Jugendliche_und_Rauschtrinken_BMG.pdf, , Abruf am 15.08.2022.

FOCUS online, (2017): Studie behauptet: Wer Alkohol trinkt, schreibt danach die besseren Prüfungen. URL: https://www.focus.de/wissen/mensch/wunderlernmittel-fuer-studenten-studie-belegt-wer-alkohol-trinkt-schreibt-die-besseren-klausuren_id_7409530.html, Abruf am 15.08.2022.

Student Alcohol Consumption, update (2016). URL: https://www.kaggle.com/datasets/uciml/ student-alcohol-consumption?select=student-merge.R, Abruf am 01.07.2022.

Anhang: Git-Historie