

Leitfaden für nachvollziehbare Schritte

Autor: Andreas Schau

1. Kurze Darstellung des Problembereichs / Aufriss des Themas

1.1 Inhaltlich

Kern der Untersuchung

Explorative Datenanalyse – kann ein Zusammenhang zwischen Persönlichkeitsprofilen und Drogenkonsum gefunden werden? Kann Machine-Learning dazu verwendet werden, Menschen mit solchen „Risiko“-Persönlichkeiten zu identifizieren, um ihnen gezieltere Aufklärung für die entsprechenden Drogen anzubieten?

Grobziele der Arbeit

Korrelationen und lernbare Zusammenhänge zwischen Persönlichkeitswerten und Drogenkonsum finden.

1.2 Begründung des Themas

Darstellung der Relevanz des Themas?

Warum ist das Thema wichtig und interessant und daher bearbeitungs- und förderungswürdig?

„2022 starben deutschlandweit 1.990 Menschen an den Folgen ihres Drogenkonsums. 749 dieser Menschen starben aufgrund von Heroin und Morphin. Der Anteil von Opiaten und Opioiden hat sich in den letzten Jahren sukzessive erhöht und die meisten Drogentoten sind auf Vergiftungen mit diesen Substanzen zurückführen.“ Quelle: de.statista.com

An diesem Zitat erkennt man, dass möglicherweise durch die Verwendung von Schmerzmitteln bei Operationen, die Zahlen für Opioid-Tode in den letzten Jahren gestiegen sind. Eventuell gibt es einen Zusammenhang zwischen den Persönlichkeitsprofilen von Patienten und bestimmten Drogengruppen, die diese anfälliger dafür macht, bestimmte Drogen eher zu verwenden, da ihnen die Wirkung oder ähnliches mehr zusagt, als anderen.

Wenn so ein Zusammenhang festgestellt werden kann, könnte man diesen nutzen um gezieltere Drogenaufklärungsangebote anzubieten.

Darstellung eines persönlichen Erkenntnisinteresses.

Ein trainiertes Model, und ein psychologisch entwickelter Fragebogen könnten dazu verwendet werden einen online Service anzubieten, mit dem Nutzer mehr über ihre eigenen Anfälligkeiten in Bezug auf Drogen erfahren könnten. Dieses Portal könnte dann auch gleich weiterführende Aufklärungs-Informationen

anbieten.

So könnten unsichere Patienten vor einer großen Operation überprüfen, wie anfällig sie wahrscheinlich für Opiode sind und sich besser darauf vorbereiten. Oder jugendliche die mit Zigaretten liebäugeln, könnten es sich vielleicht noch einmal anders überlegen.

2. Nachvollziehbare Schritte

2.1 Der Stand der Forschung / Auswertung der vorhandenen Literatur / Tutorials ...

Wurde das Problem früher bereits untersucht?

Ja, das Problem wurde bereits auf verschiedene Arten untersucht.

Welche Aspekte wurden untersucht und welche nicht?

Es wurden bereits Persönlichkeitsmerkmale, Psychologische Faktoren, Sozioökonomische Faktoren und psychische Gesundheit und Drogenkonsum in verschiedenen Studien untersucht.

Welche Kontroversen gab es und welche Methoden standen bis jetzt im Vordergrund?

Kontroversen:

Einfluss des psychosozialen Umfelds: Eine der Hauptkontroversen betrifft die Rolle des psychosozialen Umfelds bei der Drogenabhängigkeit. Es gibt Diskussionen darüber, ob der Unterschied im Drogenkonsum zwischen Personen, die in derselben Umgebung leben, zufällig ist oder ob bestimmte Persönlichkeitsmerkmale Menschen helfen, Drogenabhängigkeit zu verhindern.

Persönlichkeitsmerkmale als Risikofaktoren: Eine weitere Kontroverse dreht sich um die Frage, ob bestimmte Persönlichkeitsmerkmale, wie Feindseligkeit und Wettbewerbsorientierung, tatsächlich Risikofaktoren für den Drogenkonsum sind. Es gibt auch Diskussionen darüber, ob diese Persönlichkeitsmerkmale als Schutzfaktoren gegen Drogenkonsum dienen, insbesondere bei Alkohol und Nikotin.

Methoden:

Quantitative und qualitative Datenauswertung: Die Untersuchungen zum Zusammenhang zwischen Persönlichkeitsprofilen und dem Drogenkonsum haben sowohl quantitative als auch qualitative Methoden verwendet. Quantitative Methoden umfassen die Verwendung von statistischen Werkzeugen zur Analyse von Daten, während qualitative Methoden die Analyse von Daten durch Beobachtung, Interviews oder die Inhaltsanalyse von Texten beinhalten.

Wichtigste (verwendete) wissenschaftliche Positionen zum ausgewählten Thema?

(Z.B. **Tutorials ...**)

Das Thema wurde auf kaggle schon von einigen anderen untersucht, jedoch bisher hauptsächlich mit Hilfe von Python. Es gibt ebenfalls ein Beispiel, wie man in Python die gegebenen Daten etwas aufbereiten kann, damit sie intuitiver verständlich sind.

2.2 Fragestellung

Kann mithilfe von KNIME und Machine-Learning vorausgesagt werden, welche Persönlichkeits-Typen eher zum Konsum bestimmter Drogen neigen?

2.3 Stand der Forschung

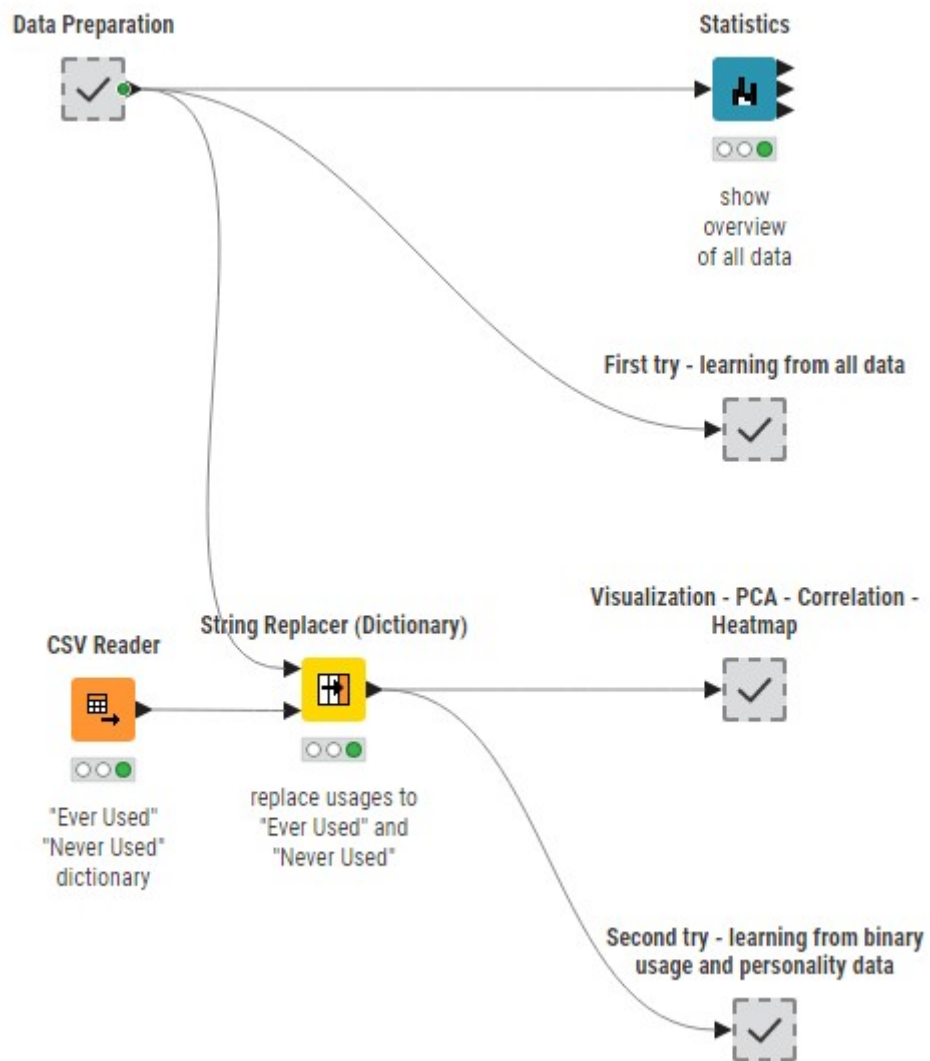
Die aktuelle Forschung zu Persönlichkeit und Drogenkonsum zeigt, dass Persönlichkeitsmerkmale wie Extraversion und Neurotizismus mit einem höheren Risiko für den Drogenkonsum verbunden sind. Psychologische Faktoren, wie Stress und psychische Gesundheitsprobleme, sowie Sozioökonomische Faktoren, wie soziale Isolation, können ebenfalls das Risiko für den Drogenkonsum erhöhen. Aktuelle Forschungsrichtungen konzentrieren sich auf Interventionen und Prävention sowie auf neurowissenschaftliche Erkenntnisse, um die Mechanismen des Drogenkonsums besser zu verstehen und effektivere Behandlungsansätze zu entwickeln.

2.4 Wissenslücke

Es gibt noch Wissenslücken in der Forschung zum Zusammenhang zwischen Persönlichkeit und Drogenkonsum. Insbesondere gibt es Unklarheiten über die genauen Mechanismen, wie Persönlichkeitsmerkmale und psychologische Faktoren den Drogenkonsum beeinflussen, sowie über die Effektivität verschiedener Interventionen und Präventionsprogramme.

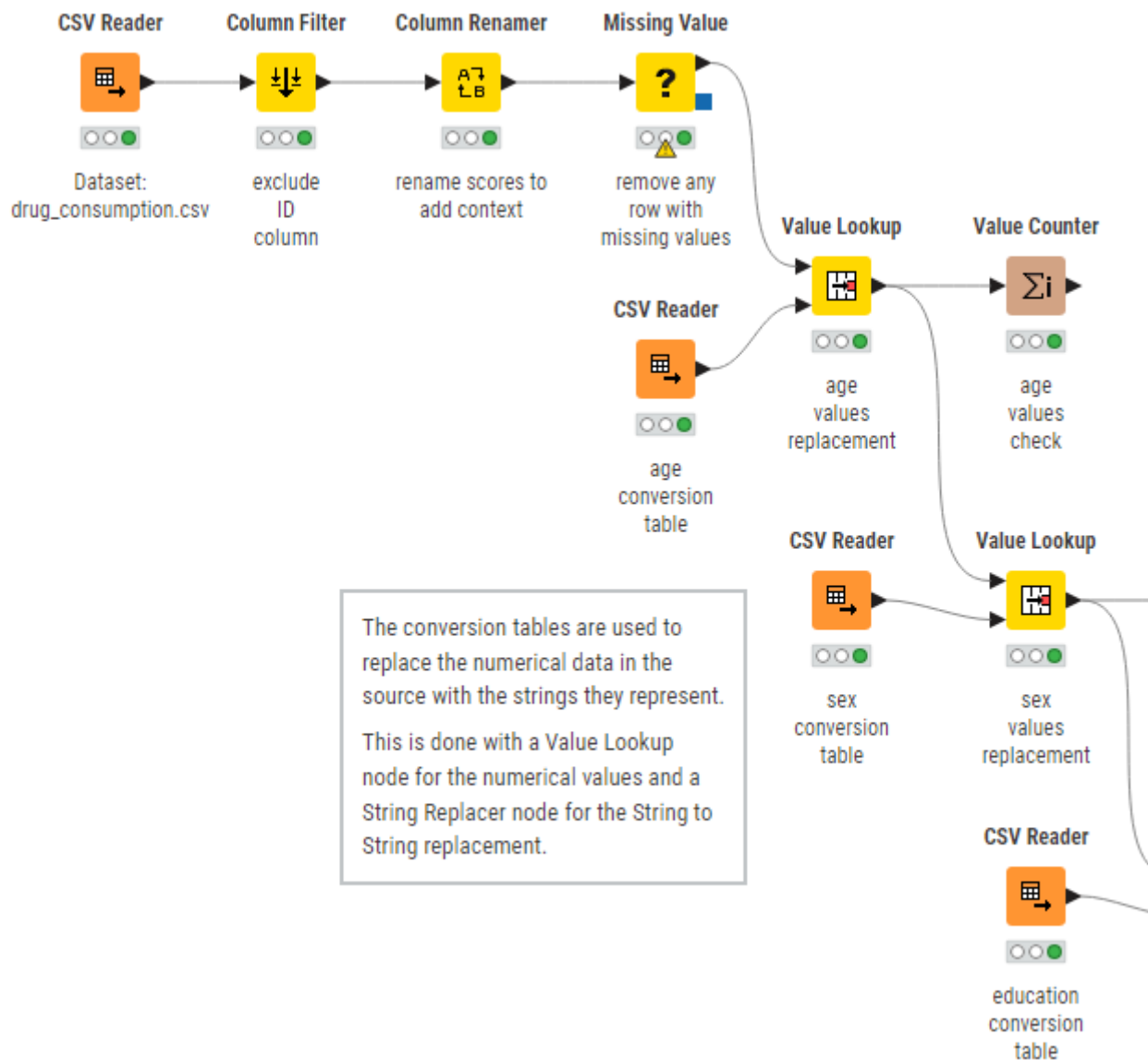
2.5 Methode

Öffnet man das KNIME-Projekt, so sieht man als erstes diese Ansicht:



Projekt Überblick

Gehen wir in dieser Ansicht in den ersten Schritt die „Data Preparation“. Hier sehen wir unter anderem diesen Ausschnitt:

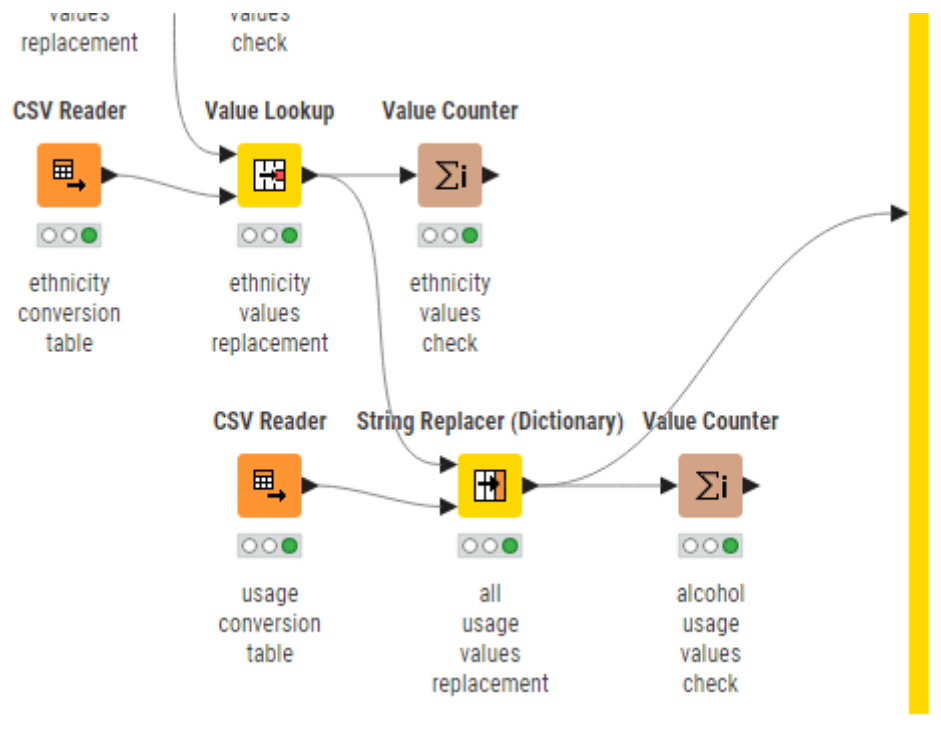


Hier werden die Daten wie auf [kaggle](#) vorgeschlagen vorbereitet:

- „CSV Reader“ zum Laden der Daten aus der „drug_consumption.csv“ Datei.
- „Column Filter“ um die ID Spalte zu entfernen, da diese keinen Lerngehalt hat.
- „Column Renamer“ um den psychologischen Bewertungen „scores“ etwas mehr Kontext zu geben.
- „Missing Value“ um fehlende Werte zu entfernen. (Wäre in diesem Datensatz nicht notwendig gewesen, aber es rundet die Vorverarbeitung ab.)

Darauf hin folgt eine Kette von „Value Lookup“-Knoten die jeweils die schon veränderten Daten und eine weitere Lookup-Tabelle via „CSV Reader“ nehmen um die numerischen Werte in den Spalten, die keine Psychologischen Bewertungen sind, in deren repräsentative Zeichenketten umwandeln. Zum Beispiel werden in der Spalte „Country“ die Zahlen, die wahrscheinlich schon Gaussian-Z-Normiert wurden, durch „USA“, „UK“ und die anderen Ländernamen ersetzt.

Diese Kette setzt sich so fort, bis alle entsprechenden Spalten bedeutsame Inhalte enthalten.



Darauf hin kommt dann noch ein ähnlicher Schritt, in dem die Konsum-Zeitraum-Klassen in den Spalten der unterschiedlichen Drogen ebenfalls von ihrer Kennung durch ihre Bedeutung ersetzt werden. Beispiel: „CL0“ wird „Never Used“.

Der lange gelbe Balken ist die Verbindung des Meta-Knotens nach außen, sein Ausgang, der dann die Tabelle mit den aufbereiteten Daten zur Verfügung stellt.

Gehen wir aus der Komponente heraus, sind wir wieder in der Ansicht für das gesamte Projekt, dem Projekt Überblick.


Dort kann man sich die vorbereiteten Daten nun via dem „Statistics“ Knoten näher anschauen. (Via Rechtsklick auf den Knoten → „Open View“ oder der F10-Taste auf der Tastatur.)



- shows that more young than older people took part
- shows equal gender/ses participation
- mostly from UK and USA
- mostly white ethnicity
- usage values are often dominated by one category but not the same for all drugs

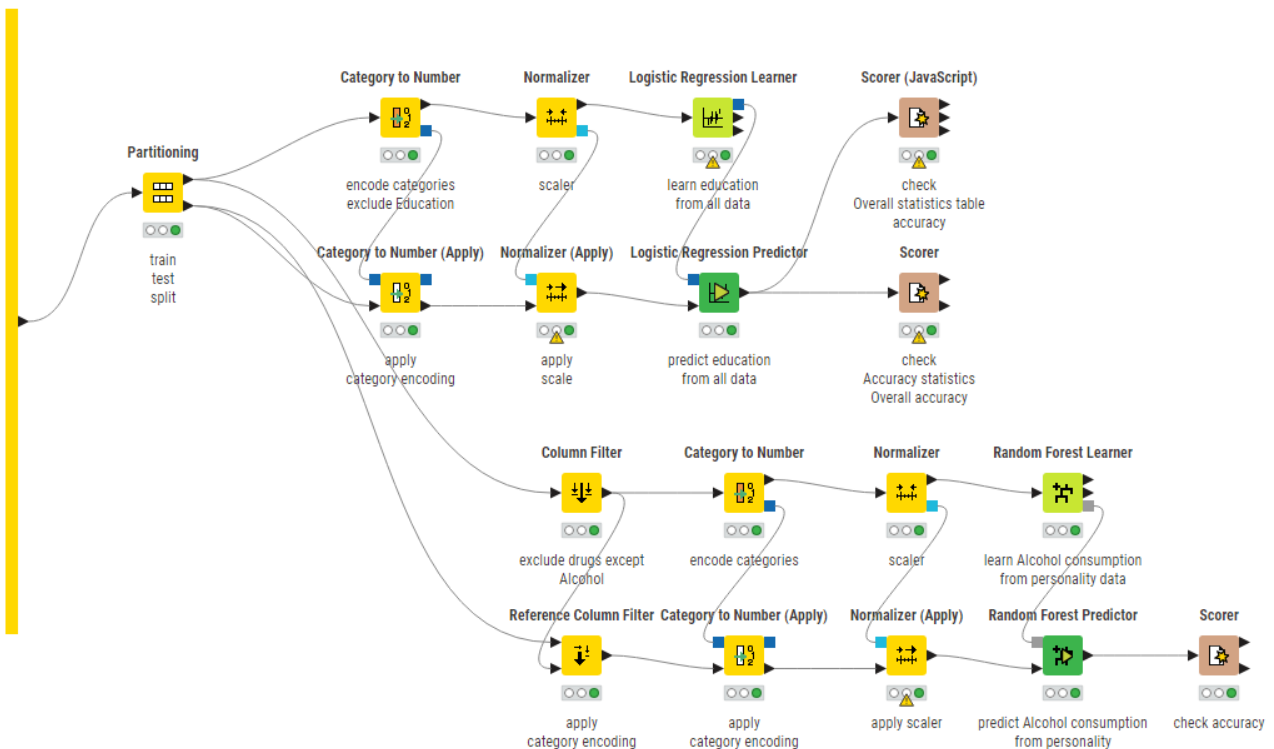
Numeric Nominal Top/bottom				
Age	Gender	Education	Country	Ethnicity
No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0
Top 20: 18 - 24 : 643 25 - 34 : 481 35 - 44 : 356 45 - 54 : 294 55 - 64 : 93 65 + : 18	Top 20: Male : 943 Female : 942	Top 20: Some College,No Certificate Or Degree : 506 University Degree : 480 Masters Degree : 283 Professional Certificate/ Diploma : 270 Left School at 18 years : 100 Left School at 16 years : 99 Doctorate Degree : 89 Left School at 17 years : 30 Left School Before 16 years : 28	Top 20: UK : 1044 USA : 557 Other : 118 Canada : 87 Australia : 54 Republic of Ireland : 20 New Zealand : 5	Top 20: White : 1720 Other : 63 Black : 33 Asian : 26 Mixed-White/Asian : 20 Mixed-White/Black : 20 Mixed-Black/Asian : 3

In dieser Ansicht kann man schnell erkennen, wie die Daten in den verschiedenen Spalten verteilt sind. Ein paar dieser Verteilungs-Merkmale wurden in einer Workflow-Bemerkung neben dem Statistik-Knoten aufgeführt. Zum Beispiel, dass eher jüngere Leute an der Statistik beteiligt sind.

First try - learning from all data


Shows quite low prediction power without further data optimiation.

Im nächsten Schritt wurden zwei einfache Machine-Learning Modelle ausprobiert.



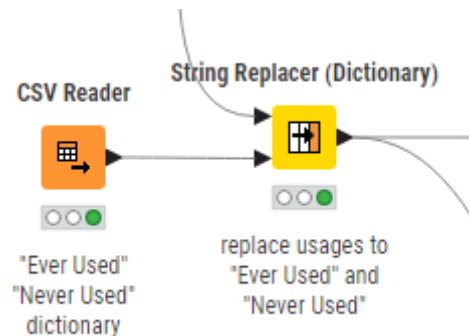
In diesem Meta-Knoten kommen vom rechten gelben Balken die Daten aus der Übersicht herein und werden zuerst einmal durch den „Partitioning“-Knoten in Trainings- und Testdaten aufgeteilt und an die beiden Modelle, in deren Trainings- und Test-Sektion geleitet.

Im oberen Modell, welches aus den oberen zwei Knotenzeilen besteht, wurde untersucht, ob man aus den Daten, so wie sie sind lernen kann, welcher Bildungsstand in Abhängigkeit der Persönlichkeits-Bewertungen und Drogen-Konsum-Muster erreicht wurde. Dabei wurde unter Verwendung eines „Logistic Regression Learners/Predictors“ jedoch nur eine Genauigkeit von ca. 30% erreicht. Es scheint also einen Zusammenhang zu geben, dieser ist jedoch nicht sonderlich stark.

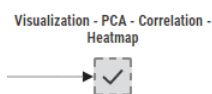
Im unteren Modell wurden die Daten weiter gefiltert indem im „Column Filter“-Knoten alle Drogen und deren Konsumverhalten, bis auf Alkohol entfernt wurden. Dann kam ein „Random Forest Learner/Predictor“ Paar zum Einsatz um aus den Persönlichkeits- und Lebenssituations-Daten zu lernen, welchen Alkoholkonsum solch eine Person am wahrscheinlichsten haben würde. Jedoch konnten auch hier nur Genauigkeiten von um die 30% erreicht werden.

Schaut man sich an, was das Modell lernen und bestimmen soll, so fällt einem auf, dass es wohl daran liegt, dass nicht nur gelernt werden muss, welche Arten von Charakteren zu welchem Drogenkonsum neigen, sondern auch in welchen vergangenen Zeiträumen sie dies getan haben müssen. Diese Zeitkomponente bringt sehr wahrscheinlich einiges an Streuung in das Modell, denn ein Erwachsener fortgeschrittenen Alters, mit einer sehr offenen Persönlichkeit mag vor sehr vielen Jahren mal verschiedene Dinge probiert haben, während jemand, der gerade erwachsen geworden ist möglicherweise vor sehr kurzer Zeit erst mit verschiedenen Dingen in Berührung gekommen ist.

Es könnte sich also lohnen die Konsum-Daten um ihre temporale Dimension einzukürzen und auf „Hat je konsumiert“ und „Hat nie konsumiert“ zu reduzieren. Dadurch würde man besser beleuchten können, ob bestimmte Persönlichkeiten für bestimmten Drogenkonsum anfälliger sind als andere.



Um diesen Ansatz verfolgen zu können wurden bei allen Drogen die verschiedenen Klassen, die Zeiträume des Konsums beinhalteten so verändert, dass sie nur noch eine Aussage darüber geben, ob eine Droge je oder nie verwendet wurde. Dafür kam wieder der „String Replacer“-Knoten und ein „CSV Reader“ mit der Übersetzungs-Tabelle zum Einsatz.



To demonstrate how to do a PCA and correlation heatmap.

Heatmap shows correlations of usages of drugs with similar reported effects - for example:

- Mushrooms and LSD and Cannabis - are psychedelics and seem to be experimented with by similar people and seem to be correlated with openness.
- Crack and Heroin and Coke - are stimulants and also seem to be experimented with by the same kind of people and seem to be correlated with impulsiveness.

(The author is not a professional in this topic so these are just educated guesses on quick research.)

Die so aufbereiteten Daten wurden dann in den Visualisierungs Meta-Knoten gegeben um einen PCA Scatter Plot und eine Korrelations-Heatmap anzufertigen.

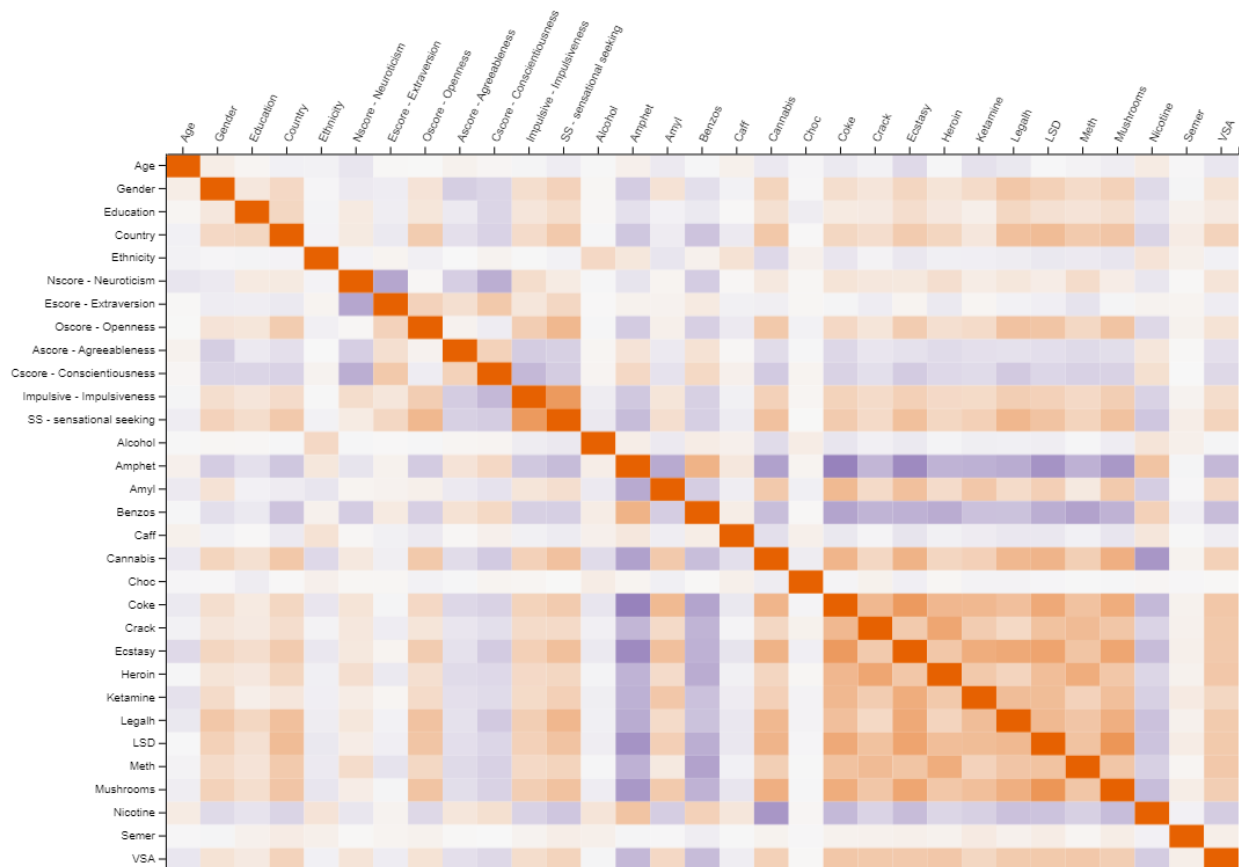
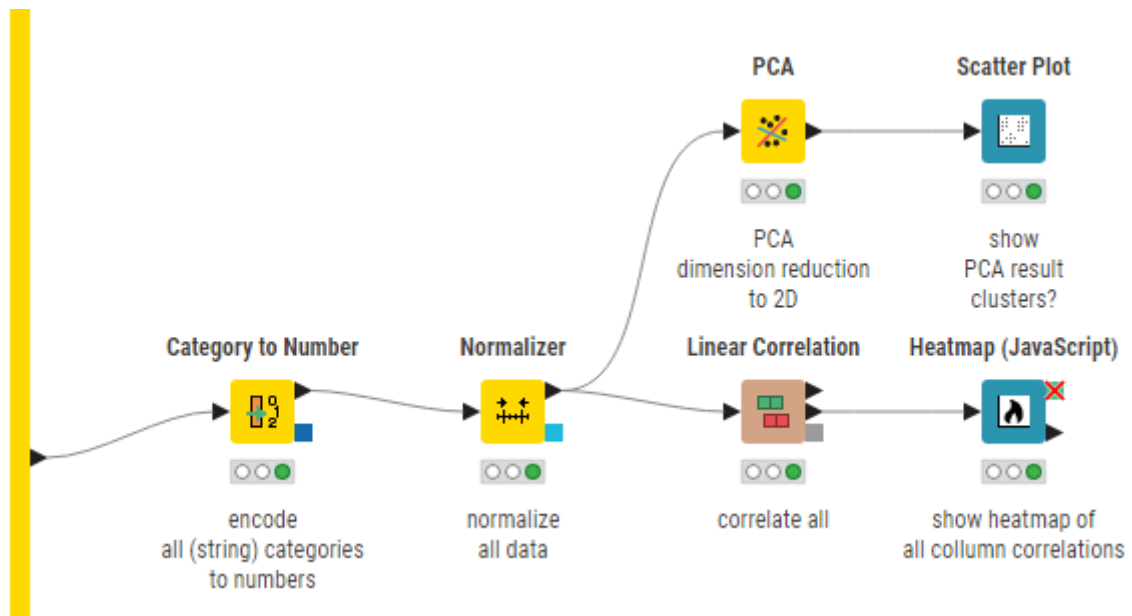
In diesen Grafiken kann man einige Dinge recht gut erkennen, die in einer Workflow-Notiz zusammengefasst wurden.

Der PCA Plot zeigt keine auffälligen Cluster und wurde daher nicht weiter verfolgt.

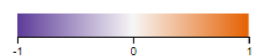
Die Heatmap ist jedoch sehr aufschlussreich. Man muss nur die Zeilen der einzelnen Persönlichkeits-Bewertungen entlang gehen und kann sehen, dass Offenheit mit einigen Drogen positiv korreliert (Cannabis, Pilze und LSD) und mit anderen negativ (Coke und Crack) und bei Impulsivität ist es wiederum fast genau anders herum.

Verträglichkeit wiederum ist negativ mit fast allen typischerweise illegalen Drogen korreliert aber etwas positiv mit Nikotin, diese Art von Charakter scheint also wenn dann mit legalen Drogen zu experimentieren.

(Falschangaben zum Selbstschutz des Bildes von einem selbst wären jedoch bei all diesen Angaben auch nicht ausgeschlossen.)



Showing 1 to 31 of 31 entries



Second try - learning from binary usage and personality data



When reducing the usage categories to a binary choice between "Ever Used" and "Never Used" it is easier to get good prediction results.

For example to predict the LSD consumption - if an individual has ever tried it or not - the other drug data is disregarded and only the personality descriptors and the life circumstances like country, education etc. are used and can yield a prediction accuracy of 70%. This is much better than the 30% accuracy when it would also have to predict the correct timeframe of consumption.

Basically the question:

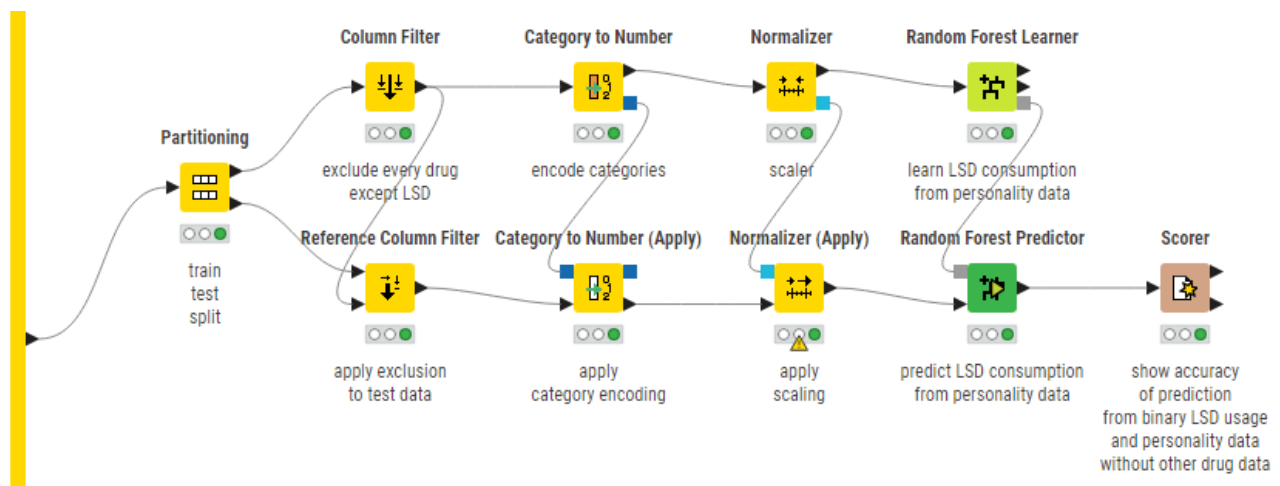
"Has a person with this profile consumed this drug in this or that timeframe?"

becomes the much simpler question:

"Will a person with this profile have consumed this drug by the time of the survey?"

(or: "What personality profile is more likely to ever consume/seek to experience this kind of drug.")

Nach den Visualisierungen und neuen Einsichten wurde dann noch einmal ein Machine-Learning Modell mit den Binären Konsum-Daten ausprobiert. Dessen Einsichten ebenfalls in einer Workflow-Notiz festgehalten wurden. Der Inhalt des Knotens ist folgender.



Hier werden wie in den anderen Modellen, von außen, durch den gelben Balken rechts, die Daten mit binärem Konsumverhalten entgegengenommen und in diesem Fall durch den „Column Filter“ alle anderen Drogen, bis auf LSD heraus gefiltert. Dann werden die Daten von ihren Kategorischen Zeichenketten Repräsentationen durch „Category to Number“ in Zahlen umgewandelt und via „Normalizer“-Knoten zwischen Null und Eins normalisiert.

Folgend wurde ein „Random Forest Learner/Predictor“ Paar verwendet um nun zu trainieren aus den Persönlichkeits- und Lebenssituations-Daten vorhersagen darüber treffen zu können, ob ein Mensch mit diesem Profil in seinem Leben je LSD konsumieren wird.

2.6 Ergebnisse

Confusion Matrix - 6:88:66 - Scorer (show accuracy)		
File Hilite		
LSD \ Predi...	Never Used	Ever Used
Never Used	180	37
Ever Used	43	117
Correct classified: 297		Wrong classified: 80
Accuracy: 78,78%		Error: 21,22%
Cohen's kappa (κ): 0,564%		

Mit dem „Scorer“ können wir diesmal eine Vorhersage-Genauigkeit von ca. 78% feststellen, was deutlich genauer ist, als in den anderen Versuchen. (Man beachte, dass diese Werte je nach Ausführung und zufälliger Partitionierung der Trainings-Test-Daten leicht variieren.)

2.7 Ausblick

Diese Genauigkeit suggeriert auch, dass es einen Zusammenhang zwischen Persönlichkeit und Drogenkonsum zu geben scheint, den man mit weiteren Analysen noch genauer untersuchen könnte. Außerdem könnte man ein so trainiertes Model nun verwenden um zum Beispiel auf einer Webseite einen Umfragebogen anzubieten, der die Persönlichkeit zu bewerten versucht und dann gezielt Präventions-Materialien auf die jeweilige Person abgestimmt anbietet. Eventuell kann so besser gewarnt werden oder im Falle von Opioiden bei Operationen Tipps zur Nachbehandlung nach der OP gegeben werden.

Es sei noch einmal angemerkt, dass der Autor kein Experte auf diesem Gebiet ist und alle hier angestellten Vermutungen nur auf gesundem Menschenverstand basieren.