Zusammenfassung

- Das Praxislabor bietet einen praxisnahen Einstieg in die Arbeit mit Orange Data Mining mit besonderem Schwerpunkt auf der Anwendung in der historischen Forschung.
- Vorkenntnisse in der Programmierung sind nicht erforderlich.
- Das ist nämich das Schöne an Orange Data Mining: Das Programmieren ist in Widgets und Dialogfenstern versteckt. Die Methode nennt man 'visual programming'.
- Das Angebot richtet sich an Studierende, Doktoranden und Forscher, die einen leicht zugänglichen Einstieg in datengestützte Analysemethoden suchen.

Zusammenfassung – 2

- Nach einer kurzen Einführung in die Grundlagen der visuellen Programmierung erkunden wir die Benutzeroberfläche und erstellen eigenständig einige Workflows.
- Dabei lernen wir, Daten zu laden, aufzubereiten, zu visualisieren und mit einfachen Verfahren des maschinellen Lernens auszuwerten.
- 90 Minuten ist nicht viel Zeit – aber dennoch möchte ich versuchen, die drei großen Bereiche der Arbeit mit historischen Daten kurz zu thematisieren: das sind: Data, Text und Bild.
- Für die selbständige Vertiefung steht eine Vielzahl an YouTube Tutorials und eine ziemlich umfassende Dokumentation auf der Website zur Verfügung.
- Gut zu Wissen, ist das Orange Data Mining in erster Linie in der Bioinformatik entwickelt wurde, sich aber immer weiterentwickelt und zunehmend auch DH Bereiche mit abdeckt. Dennoch: nicht dezidiert für historisches Arbeiten gedacht.
- Auch wichtig: Das Tool ist in erster Lineie ein Instrument zum Lernen und Testen. Die Erfahrung zeigt, dass Studierende nach einer Einführung irgendwann auf die unmittelbare Nutzung von python wechseln.

Aufbau des Praxislabors

ERSTE SCHRITTE

DATEN

TEXT

BILDER

Einführung in die Analyse historischer Daten

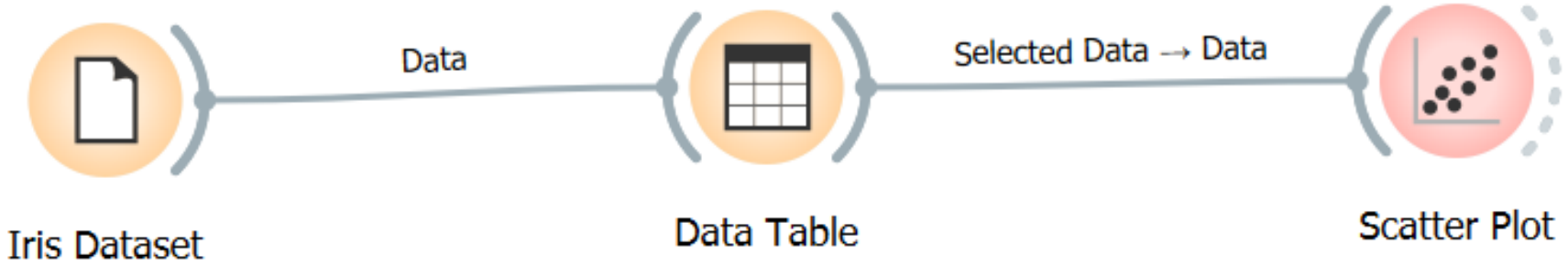
EINHEIT 1: ERSTE SCHRITTE

Quelle: www.orangedatamining.com

Kurzer Blick auf die Oberfläche (canvas)

- Über eine visuelle Oberfläche lassen sich komplexe analytische Workflows per Drag-and-Drop zusammenstellen.
- Die Bausteine sind gruppiert in Kategorien
- Einzelne Bausteine – sogenannte Widgets – werden miteinander verbunden und bilden so transparente, nachvollziehbare Verarbeitungsketten.
- Das Spektrum reicht von klassischen Data-Mining-Verfahren über die Analyse von Textkorpora bis hin zur Arbeit mit Image Embeddings.
- Add-Ons, die zum Teil von externen Anbietern entwickelt wurden, bieten weitere Funktionen. Zum Beispiel: Textable – ein Add-On für das Arbeiten mit TEI; oder 'geo' – ein Add-On für Geovisualisierung.
- Wir brauchen später die Add-Ons 'Text Mining' und 'Image Analytics'

Beispiel für einen Workflow, der aus drei Widgets besteht



Datei-Widget
Umbenannt (F2)
Dient zum Laden der Daten, mit denen Sie arbeiten möchten

Datentabellen-Widget
Nimmt die geladenen Daten entgegen
Zeigt die Daten in tabellarischer Form an
Manuelle Auswahl von Instanzen (Beobachtungen, Zeilen) ist möglich

Visualisierung der ausgewählten Daten als Streudiagramm (eine beliebte Diagrammart)

Was haben wir gelernt?

- Canvas ist die Grundlage; Standardauswahl an Widgets (Prozesse) kann mit Add-Ons erweitert werden
- Daten werden geladen – nicht hochgeladen
- Widgets werden konfiguriert
- Widgets bearbeiten Daten und leiten sie weiter
- Die Weiterleitung kann auch konfiguriert werden
- Das Ergebnis kann visualisiert werden
- Verschiedene Fenster können nebeneinander geöffnet sein und kommunizieren miteinander

Einführung in die Analyse historischer Daten

EINHEIT 2: DATEN
KAPITEL 2: GRUNDLEGENDE TRANSFORMATIONEN

Methoden

- Spalten auswählen
- Zeilen auswählen
- Pivot-Tabelle

- Bearbeitung anhand des Beispiels "African Name Database" – eine CSV Datei mit ca. 67000 Zeilen

Unser Beispiel: Die Datenbank für afrikanische Namen

- „In den letzten 60 Jahren des transatlantischen Sklavenhandels verurteilten Gerichte rund um den Atlantik über zweitausend Schiffe wegen Beteiligung an diesem Handel und hielten die Details der an Bord gefundenen Gefangenen fest, einschließlich ihrer afrikanischen Namen.“
- „Die Datenbank für afrikanische Namen wurde auf der Grundlage dieser Aufzeichnungen erstellt, die sich heute in den Registern der befreiten Afrikaner im Nationalarchiv von Sierra Leone in Freetown sowie in den Serien FO84, FO313, CO247 und CO267 im britischen Nationalarchiv in London befinden.“
- „Es werden Links zu den Schiffen in der ‚Voyages Database‘ bereitgestellt, von denen die befreiten Afrikaner gerettet wurden, sowie zur Website ‚African Origins‘, auf der Nutzer die Aussprache der Namen anhören und uns dabei helfen können, die Sprachen zu identifizieren, in denen die Namen ihrer Meinung nach verwendet werden.“

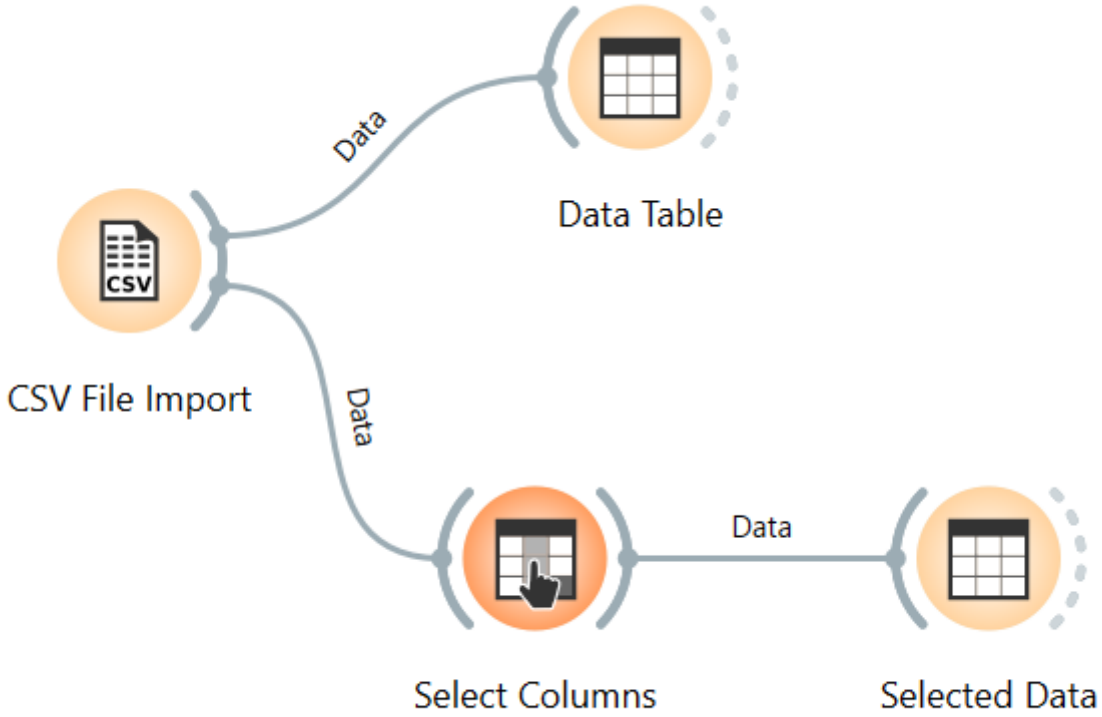
Wichtigste Merkmale der Datenbank

- 90.493 Einträge
- 16 Spalten:
  - Afrikanische ID
  - Namen
  - Moderner Name
  - Sprachgruppe
  - Alter
  - Geschlecht
  - Größe (in.)
  - Reise-ID
  - Schiffsname
  - Einschiffungshafen
  - Ausschiffungshafen
  - Ankunft
  - Geplanter Ausschiffungshafen
  - Schicksal des Gefangenen
  - Standort nach der Ausschiffung
  - Schicksal des Schiffes
- Zwei zusätzliche Spalten, die die Datenbank um Tonaufnahmen der Namen erweitern

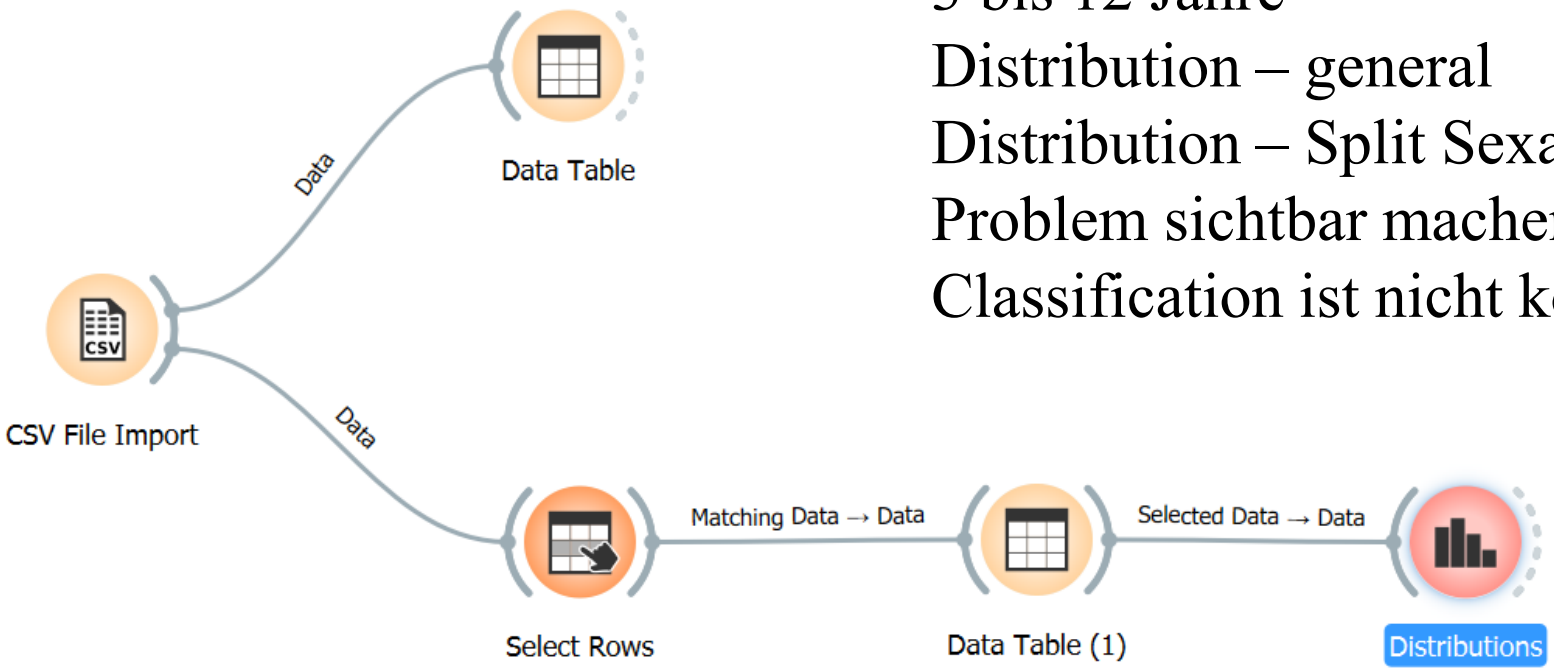
Was ist CSV?

- CSV = Comma Separated Values
- Das Dateiformat ist plattformunabhängig – Sie benötigen keine spezielle Software, um die Datei zu öffnen und zu verwenden.
- In einer CSV-Datei werden alle Werte durch Kommas (oder Semikolons) getrennt und jeder Wert wird in einfache oder doppelte Anführungszeichen gesetzt.
- Die erste Zeile der Datei enthält die Spaltenüberschriften. Diese stellen die *Attribute* der Daten dar.
- Die folgenden Zeilen enthalten den Inhalt – eine Zeile pro Eintrag. Es handelt sich dabei um die *Instanzen* der Daten.

Füge das Widget „Spalten auswählen“ zur Arbeitsfläche hinzu



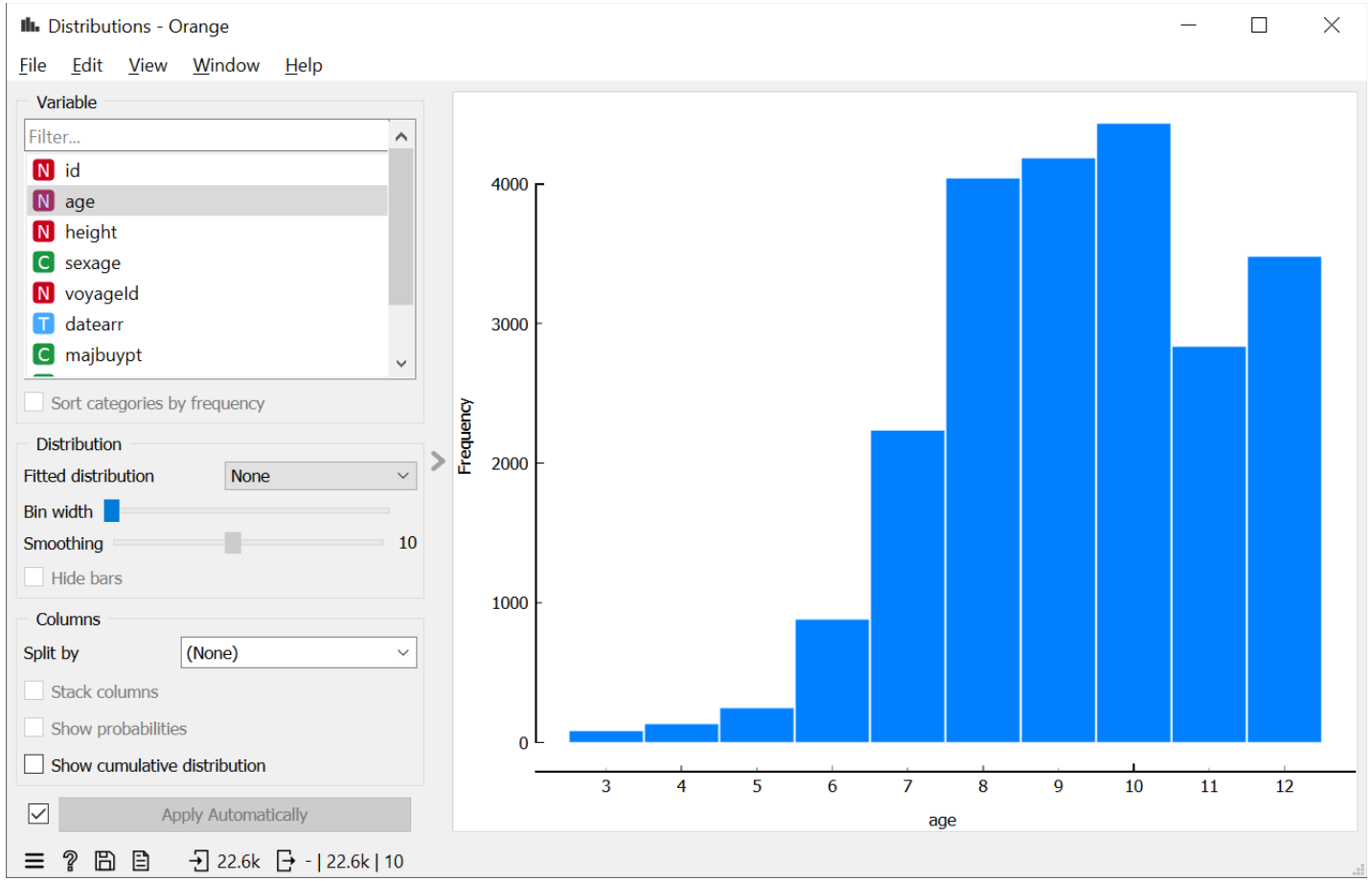
Wir können diese Auswahl für weitere Untersuchungen nutzen, z. B. um die Altersverteilung versklavter Kinder zu untersuchen



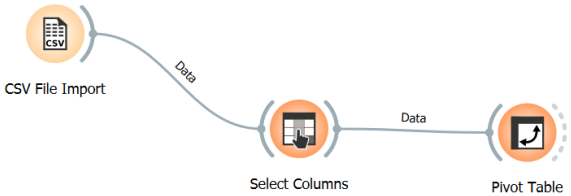
3 bis 12 Jahre
Distribution – general
Distribution – Split Sexage =>
Problem sichtbar machen: Die Classification ist nicht konsistent

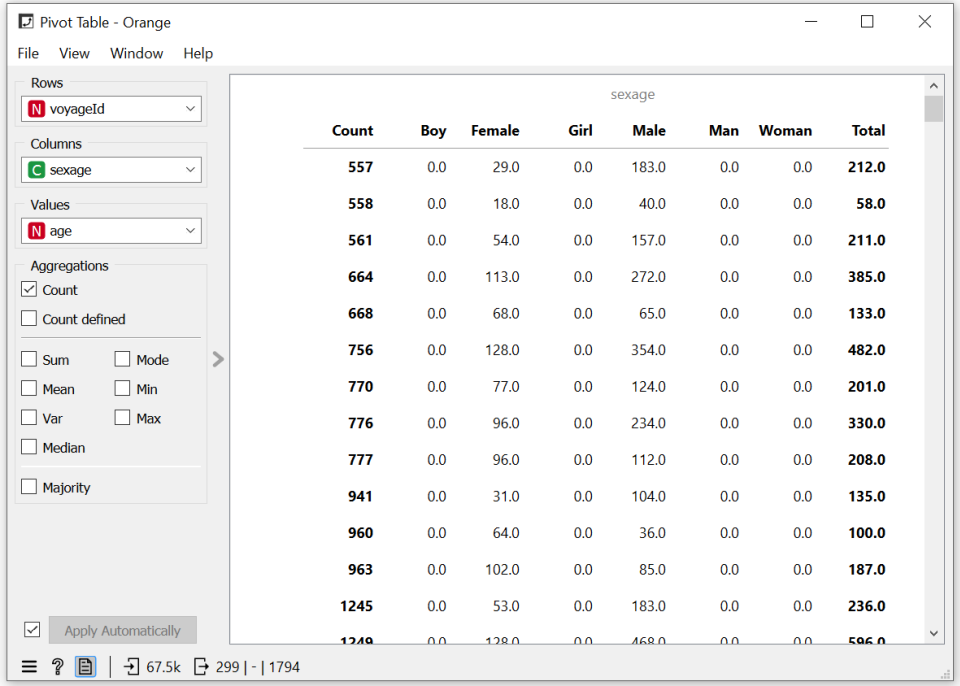
Schau dir das Ergebnis an

Verwenden Sie die „Bin-Breite“, um die Anzahl der Balken in der Verteilung zu ändern. Eine höhere „Bin-Breite“ führt zu größeren Gruppen => probieren Sie es aus, um sich ein Bild davon zu machen!



Die Pivot-Tabelle ist eine leistungsstarke Methode, um die Beziehungen zwischen mehreren Variablen in einem Datensatz zu untersuchen



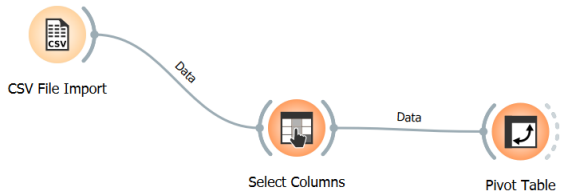


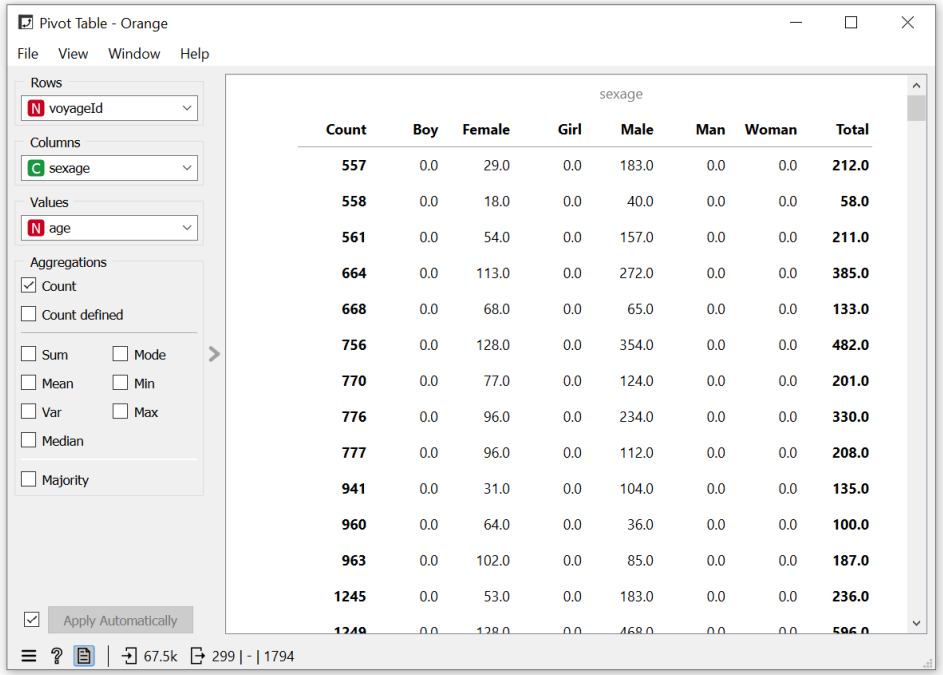
sexage

Count	Boy	Female	Girl	Male	Man	Woman	Total
557	0.0	29.0	0.0	183.0	0.0	0.0	212.0
558	0.0	18.0	0.0	40.0	0.0	0.0	58.0
561	0.0	54.0	0.0	157.0	0.0	0.0	211.0
664	0.0	113.0	0.0	272.0	0.0	0.0	385.0
668	0.0	68.0	0.0	65.0	0.0	0.0	133.0
756	0.0	128.0	0.0	354.0	0.0	0.0	482.0
770	0.0	77.0	0.0	124.0	0.0	0.0	201.0
776	0.0	96.0	0.0	234.0	0.0	0.0	330.0
777	0.0	96.0	0.0	112.0	0.0	0.0	208.0
941	0.0	31.0	0.0	104.0	0.0	0.0	135.0
960	0.0	64.0	0.0	36.0	0.0	0.0	100.0
963	0.0	102.0	0.0	85.0	0.0	0.0	187.0
1245	0.0	53.0	0.0	183.0	0.0	0.0	236.0
1249	0.0	128.0	0.0	468.0	0.0	0.0	596.0

Die Werte der ersten Variablen werden zu den Zeilenüberschriften in der Pivot-Tabelle.
Die Werte der zweiten Variablen werden zu den Spaltenüberschriften in der Pivot-Tabelle.
Die Werte der dritten Variablen werden verwendet, um eine Aggregation (z. B. Summe, Mittelwert, Anzahl usw.) der ausgewählten Variablen zu erstellen.

Die Pivot-Tabelle ist eine leistungsstarke Methode, um die Beziehungen zwischen Variablen in einem Datensatz zu untersuchen



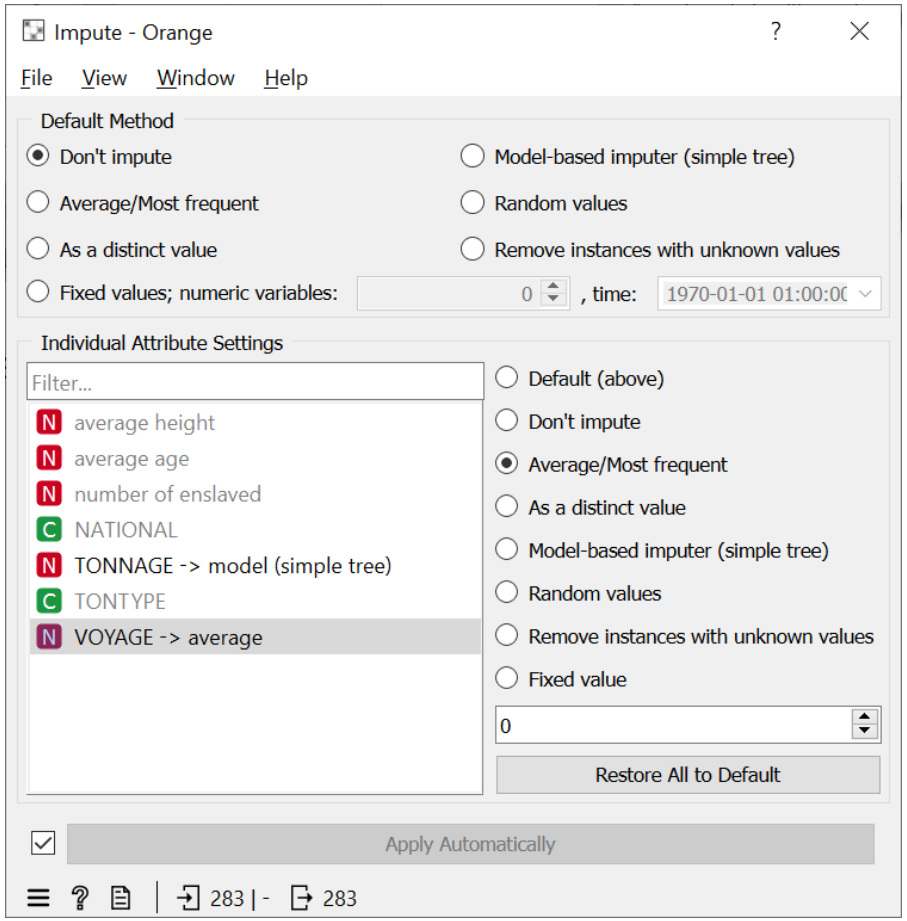


sexage

Count	Boy	Female	Girl	Male	Man	Woman	Total
557	0.0	29.0	0.0	183.0	0.0	0.0	212.0
558	0.0	18.0	0.0	40.0	0.0	0.0	58.0
561	0.0	54.0	0.0	157.0	0.0	0.0	211.0
664	0.0	113.0	0.0	272.0	0.0	0.0	385.0
668	0.0	68.0	0.0	65.0	0.0	0.0	133.0
756	0.0	128.0	0.0	354.0	0.0	0.0	482.0
770	0.0	77.0	0.0	124.0	0.0	0.0	201.0
776	0.0	96.0	0.0	234.0	0.0	0.0	330.0
777	0.0	96.0	0.0	112.0	0.0	0.0	208.0
941	0.0	31.0	0.0	104.0	0.0	0.0	135.0
960	0.0	64.0	0.0	36.0	0.0	0.0	100.0
963	0.0	102.0	0.0	85.0	0.0	0.0	187.0
1245	0.0	53.0	0.0	183.0	0.0	0.0	236.0
1249	0.0	128.0	0.0	468.0	0.0	0.0	596.0

Die Pivot-Tabelle sieht wie folgt aus:
„Auf der Reise mit der ID-Nummer 557 befanden sich 29 Frauen und 183 Männer an Bord. Die Gesamtzahl der Personen an Bord betrug 212.“

Mehr als 20 % der Werte fehlen, insbesondere in den Spalten „Tonnage“ und „Reise“. Verbinden Sie das Impute-Widget mit der Datentabelle, um diese leeren Werte zu ergänzen



Es gibt verschiedene Möglichkeiten, Werte zu imputieren. Ich habe für „Tonnage“ den modellbasierten Imputierer „Simple Tree“ und für „Voyage“ den Imputierer „Average/Most Frequent“ gewählt

Ein modellbasierter Imputator erstellt ein Modell zur Vorhersage des fehlenden Werts auf der Grundlage der Werte anderer Attribute; für jedes Attribut wird ein separates Modell erstellt. Das Standardmodell ist ein 1-NN-Lernverfahren, das den Wert aus dem ähnlichsten Beispiel übernimmt (dies wird manchmal als „Hot-Deck-Imputation“ bezeichnet).

Beachten Sie, dass an den Stellen, an denen zuvor ein „?“ stand, nun Werte eingefügt wurden. Beachten Sie die Unterschiede zwischen der komplexeren „Simple-Tree“-Methode und der einfachen „Average“-Methode zur Imputation von Werten.

	voyageId	SHIPNAME	average height	average age	number of enslavec	NATIONAL	TONNAGE	TONTYPE	VOYAGE
1	557	Orestes	59.8216	24.82550	212	Spain	128.77	?	48
2	558	Fingal	51.1373	17.89660	58	Spain	93.67	?	33.36
3	561	Nuevo Campea...	56.6095	21.36970	211	Spain	113	Spanish	33.36
4	664	Xerxes	55.2545	20.67790	385	Spain	138	Spanish	33.36
5	668	Intrepido	53.0690	19.13530	133	?	151	English 1	33.36
6	756	Firme	56.2854	21.65980	482	Spain	273.56	?	33.36
7	770	Josefa (a) Fortu...	55.0529	20.31840	201	Spain	97.15	?	33.36
8	776	Voladora (a) M...	57.0578	24.04240	330	Spain	240	Spanish	33.36
9	777	Midas	52.8836	17.35100	208	Spain	89.10	?	33.36
10	941	Gallito	56.4655	29.17780	135	?	59	English 1	33.36
11	960	Santiago	52.2989	19.38000	100	Spain	43	Spanish	33.36
12	963	Emilio (a) Caesar	52.7988	20.08650	187	Spain	85	Spanish	33.36
13	1245	Planeta	55.8947	22.22030	236	Spain	97.35	?	33.36
14	1249	Aguila	52.2981	17.34400	596	?	300	English 1	33.36
15	1250	Indagadora	60.4107	23.08210	134	Spain	129.08	?	33.36
16	1266	Negrito	57.8443	19.68760	477	Spain	273.56	?	33.36
17	1295	Joaquina	57.1367	21.44160	317	Spain	101	Spanish	33.36
18	1298	Manuelita	59.0653	19.70860	477	Spain	273.56	?	33.36
19	1307	Rosa	54.5896	16.39100	289	Spain	83.20	?	33.36
20	1338	Carlota	56.9695	19.07450	163	Spain	84.87	?	33.36
21	1355	Maria	55.7897	20.20290	340	Spain	154.89	?	33.36
22	1361	Julita	56.3689	17.62800	336	Spain	128.40	?	33.36
23	1367	Joven Reina	52.8063	16.52760	254	Spain	93.83	?	33.36
24	1368	Chubasco	54.4455	21.38260	230	Spain	84	Spanish	33.36
25	1372	Marte	54.7509	19.39690	326	Spain	128.40	?	33.36
26	1383	Tita	55.8232	17.41330	392	Spain	159.99	?	33.36
27	1396	Amália	49.6768	13.66000	200	Spain	64.88	?	33.36

Einführung in die Analyse historischer Daten

EINHEIT 3: TEXT
KAPITEL 1: EINLEITUNG

Was ist ein Text?

- Ein Text ist eine Abfolge von Wörtern
- Ein Korpus ist eine Sammlung von Texten
- Ein Text besteht aus Wörtern, Sätzen, Absätzen, Abschnitten usw.

- Text repräsentiert Wissen => Grundlage der Wissensverarbeitung

Merkmale des Textes

- Natürliche Sprache(n)
- Länge
- Redundanz und Häufigkeit von Wörtern
- Auffälligkeit von Wörtern (z. B. seltene Wörter; auffällige Wörter; bemerkenswerte Wörter)
- Wortverteilung (wie viele verschiedene Wörter verwendet werden)
- Häufiges Vorkommen von Wörtern

Verarbeitung natürlicher Sprache

- Die grundlegenden Eigenschaften von Text können für Berechnungen (statistische Methoden) herangezogen werden.
- Bei den meisten Berechnungen wird davon ausgegangen, dass die sprachliche Struktur des Textes berücksichtigt wird. Dies ist beispielsweise notwendig, um deklinierte und konjugierte Wortformen auf ihre Grundform zurückzuführen oder Lesezeichen aus den Berechnungen auszuschließen.
- Um die sprachliche Struktur eines Textes zu berücksichtigen, stehen verschiedene algorithmische Verfahren zur Verfügung, die als Verarbeitung natürlicher Sprache (kurz NLP) bezeichnet werden.
- NLP reichert Texte mit statistischen Merkmalen (Features) an, die anschließend zur Analyse von Texten und Textkorpora verwendet werden können.

Text Mining

- Die Anwendung regelbasierter, statistischer oder neuronaler Methoden auf Texte wird als Text Mining bezeichnet.
  - Regelbasierte Methoden entwickeln und nutzen Regeln, um Muster (wie beispielsweise bestimmte Wortfolgen) in Texten zu erkennen.
  - Statistische Methoden nutzen Häufigkeiten und Wahrscheinlichkeitsverteilungen, um Zusammenhänge zwischen Texten zu erkennen.
  - Im Gegensatz zu regelbasierten und statistischen Methoden sind neuronale Methoden in der Lage, Merkmale in einem Text selbst zu erlernen und diese zur Lösung spezifischer Aufgaben zu nutzen.
- Texte können so anhand bestimmter inhaltlicher oder struktureller Merkmale gruppiert (Text-Clustering) oder klassifiziert (Textklassifizierung) werden.

Was lässt sich mit Text Mining erreichen?

	Makro	Mikro
Beschreibung	• Einblicke in große Textkorpora gewinnen • Nach relevanten Zusammenhängen suchen • Strukturieren Sie große Textsammlungen • Identifizieren Sie Themen und deren Verteilung in einem Textkorpus	• Finden und extrahieren Sie bemerkenswerte Informationen und verknüpfen Sie diese gegebenenfalls mit anderen Texten • Relevante Eigennamen (benannte Entitäten) identifizieren (Personen, Organisationen, Orte, Produkte usw.)
Verwandte Begriffe	• Distant Reading • Clustering • Themenmodellierung	• Erkennung benannter Entitäten

EINHEIT 3: TEXT
KAPITEL 3: TEXT-CLUSTERING
Abschnitt 1: Bag-of-Words-Methode

Einleitung

- Ein aufbereitetes (vorverarbeitetes) Korpus eröffnet weitere Möglichkeiten zur Analyse der Texte im Korpus
- In der Regel basieren die Analysemöglichkeiten auf Berechnungen (statistischen Verfahren), die mit den Wörtern im Korpus durchgeführt werden.
- Eine erste Methode – WordCloud – wurde bereits ausprobiert (siehe Textvorverarbeitung).
- Nun betrachten wir eine weitere Methode, die Berechnungen mit den Wörtern nutzt, um Texte in einem Korpus zu gruppieren. Diese Methode wird als „Text-Clustering“ bezeichnet.

Bag-of-Words – Erläuterung

- Das Text-Clustering baut auf den Ergebnissen der Textvorverarbeitung auf.
- Anstatt jedoch direkt zu einer Visualisierung als WordCloud überzugehen, wird zunächst ein sogenannter „Bag of Words“ erstellt.
- „Bag“ (oder „Multiset“) ist ein Begriff aus der Mathematik. Er bedeutet, dass Elemente in einer Menge mehrmals vorkommen können. Ein „Bag of Words“ ist eine ungeordnete Sammlung aller Wörter eines Korpus und ihrer Häufigkeit.
- Ein einfaches Beispiel aus Wikipedia veranschaulicht den Begriff „Bag of Words“.

Bag of Words – Beispiel

- Quelle: https://en.wikipedia.org/wiki/Bag-of-words_model
- Hier sind zwei einfache Textdokumente:
  - (1) John schaut gerne Filme. Mary mag Filme auch.
  - (2) Mary schaut auch gerne Fußballspiele.
- Auf der Grundlage dieser beiden Textdokumente wird für jedes Dokument eine Liste wie folgt erstellt:
  - „John“, „mag“, „gerne“, „Filme“, „sehen“, „Mary“, „mag“, „Filme“, „auch“ „Mary“, „mag“, „außerdem“, „gerne“, „Fußballspiele“, „sehen“
- Darstellung jedes „Bag-of-Words“ als JSON-Objekt (eine Datenstruktur, die aus Schlüsseln und Werten besteht):
  - BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
  - BoW2 = {"Mary":1,"auch":1,"mag":1,"gerne":1,"Fußball":1,"Spiele":1};
- Jeder Schlüssel ist das Wort, und jeder Wert ist die Anzahl der Vorkommen dieses Wortes im jeweiligen Textdokument.
- Die Reihenfolge der Schlüssel-Wert-Paare ist beliebig.

Bag of Words – Implementierung in Orange

- „Bag of Words“ berücksichtigt in Orange drei Variablen:
  - Termhäufigkeit = Häufigkeit der Token in einem Korpus (Sammlung)
  - Dokumenthäufigkeit = Häufigkeit der Token in den Dokumenten und in der Sammlung insgesamt
  - Regularisierung = Normalisierung statistischer Merkmale; darauf werden wir hier nicht näher eingehen

Bag-of-Words – Variablen: Termhäufigkeit

- Drei Optionen für die Termhäufigkeit (abgekürzt: TF)
  - Anzahl: Anzahl der Vorkommen eines Wortes in einem Dokument
  - Binär: Das Wort kommt im Dokument vor oder nicht (auch 0 oder 1)
  - Sublinear: Logarithmus der Termhäufigkeit (Anzahl)
- Die Termhäufigkeit TF berechnet, wie oft ein Token in einem Dokument vorkommt. (Die Häufigkeit wird in der Regel normalisiert, um Verzerrungen aufgrund der Länge des Dokuments zu beseitigen.)

Bag-of-Words – Variable: Dokumenthäufigkeit

- Drei Auswahlmöglichkeiten für die Dokumenthäufigkeit (abgekürzt: DF):
  - (Keine)
  - IDF: inverse Dokumenthäufigkeit (IDF)
  - Smooth IDF: Addiert eins zur Dokumenthäufigkeit, um eine Division durch Null zu vermeiden.
- Es wird die inverse Dokumenthäufigkeit berechnet, d. h. die Anzahl der Dokumente (Texte) in einem Korpus [= Zähler] geteilt durch die Anzahl der Dokumente, in denen ein bestimmtes Token (Term) vorkommt [= Nenner].
- Je größer der Nenner, desto weniger aussagekräftig ist der Begriff.

Bag-of-Words – TF-IDF

- TF-IDF ist das Produkt aus der Termhäufigkeit und der inversen Dokumenthäufigkeit
- TF-IDF ist ein statistisches Maß, das verwendet wird, um die Relevanz von Begriffen in Dokumenten einer Dokumentensammlung (eines Korpus) zu bewerten.

TF-IDF

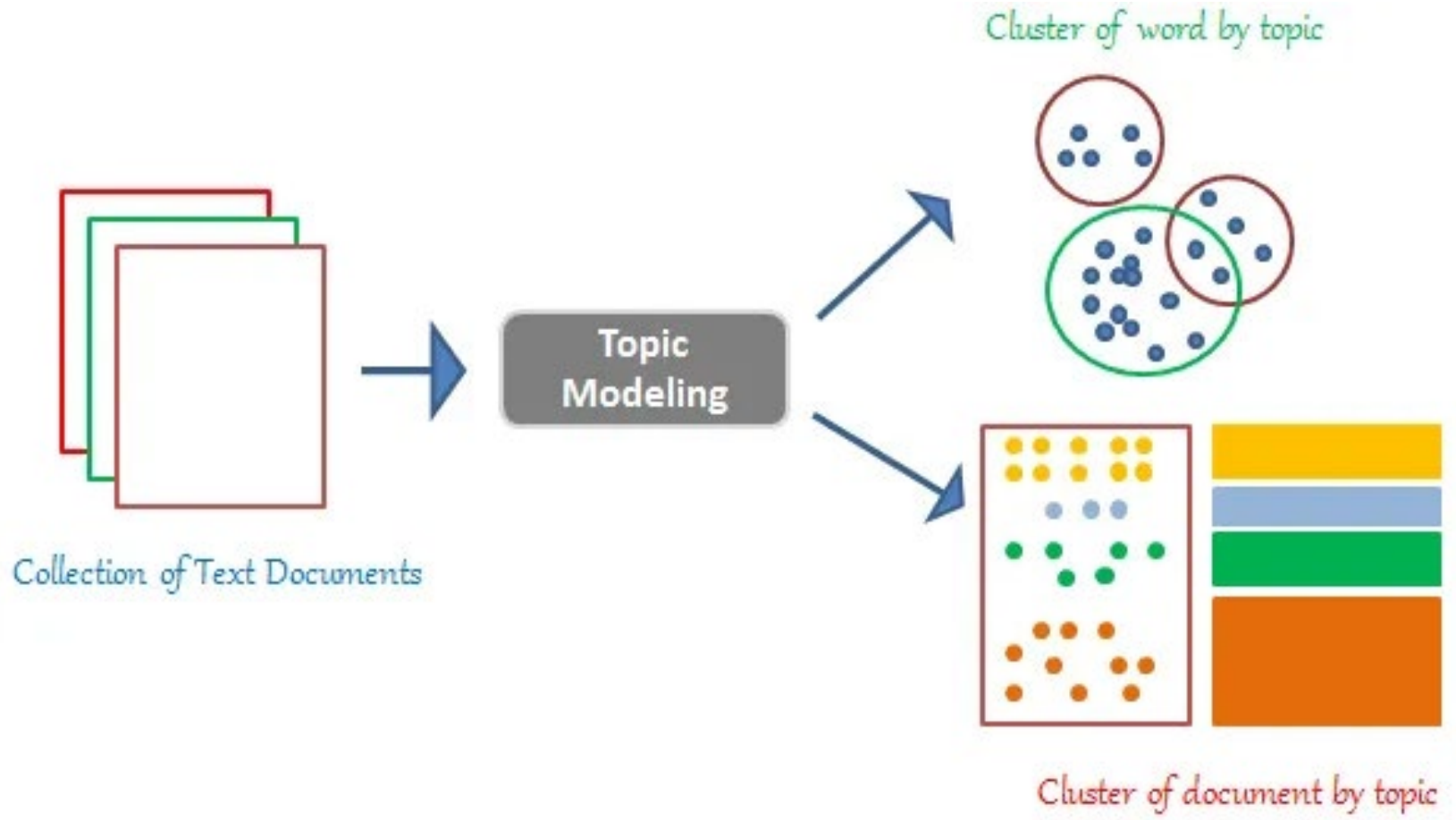
$$TF = \frac{Number\ of\ times\ a\ word\ "X"\ appears\ in\ a\ Document}{Number\ of\ words\ present\ in\ a\ Document}$$

$$IDF = log\left(\frac{Number\ of\ Documents\ present\ in\ a\ Corpus}{Number\ of\ Documents\ where\ word\ "X"\ has\ appeared}\right)$$

$$TF\ IDF\ =\ TF * IDF$$

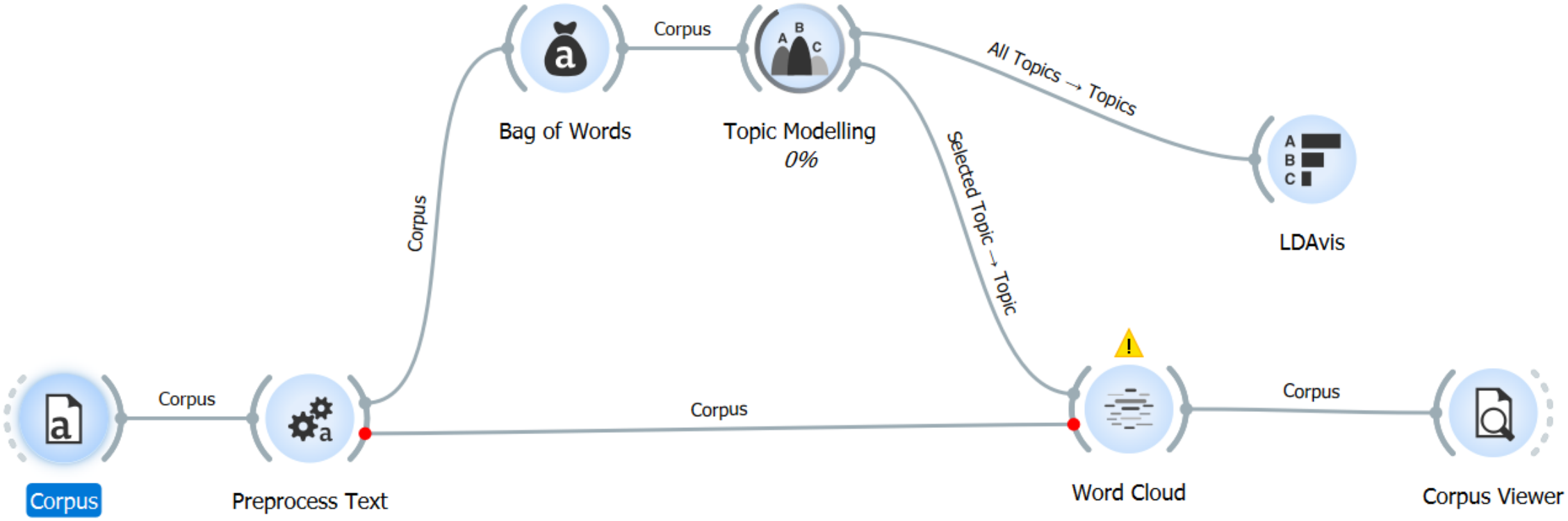
Quelle: https://letsdatascience.com/tf-idf/

LDA



Source: https://medium.com/analytics-vidhya/latent-dirichelt-allocation-1ec8729589d4

Try to apply this topic modelling workflow to the DTGC. Interpret the results using LDAvis.



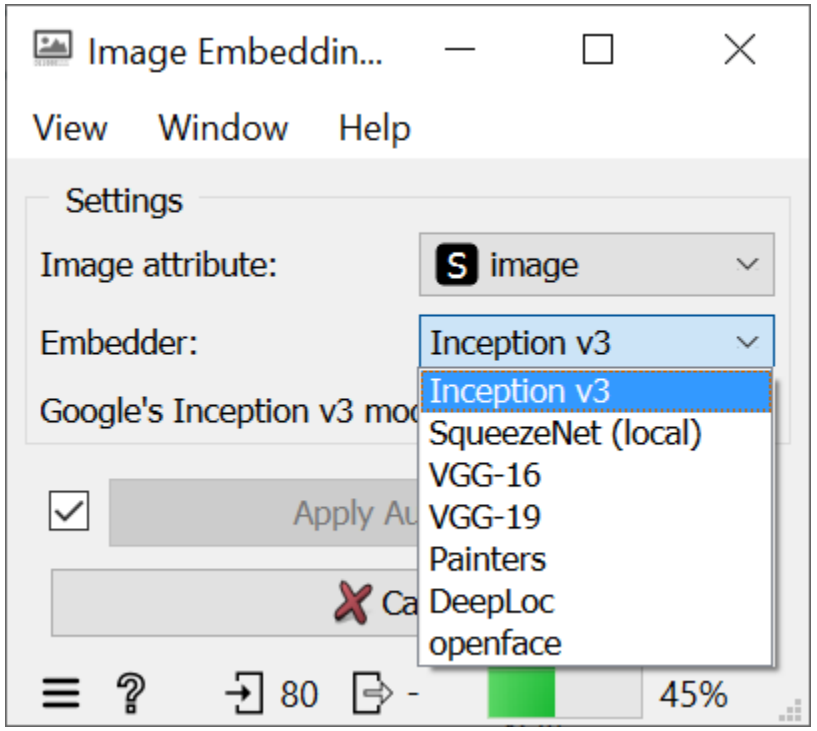
Einführung in die Analyse historischer Daten

EINHEIT 4: BILDER
Abschnitt 2: Bild-Embeddings

Mit Bildern arbeiten – Warum?

- Den Inhalt eines Bildes direkt über das Bild selbst erkunden und nicht anhand einer Metadatenbeschreibung des Bildes;
- Mit digitalen Methoden lassen sich Bilder in großer Zahl analysieren;
- Die Bildanalyse kann auch dazu genutzt werden, bestimmte Bilder aus einer großen Anzahl auszuwählen, wobei die Möglichkeiten hier ebenfalls über die Auswahl auf der Grundlage von Bildbeschreibungen hinausgehen;
- Die Bildanalyse als organisatorischer Schritt in einem komplexen Arbeitsablauf.

Bild-Embedding



„Bei der Bild-Einbettung werden Bilder eingelesen und entweder auf einen Remote-Server hochgeladen oder lokal ausgewertet. Mithilfe von Deep-Learning-Modellen wird für jedes Bild ein Merkmalsvektor berechnet. Das Ergebnis ist eine erweiterte Datentabelle mit zusätzlichen Spalten (Bilddeskriptoren).“

- Es werden mehrere „Embedder“ (tiefe neuronale Netze für die Bildanalyse) vorgeschlagen
- Alle Embedder außer „SqueezeNet“ erfordern einen stabilen Internetzugang
- Jeder Embedder wird für eine bestimmte Aufgabe trainiert. Das Bild wird an einen Remote-Server gesendet, wo es vom Embedder ausgewertet wird. Die Ergebnisse werden als Matrix (große Datentabelle) zurückgegeben

Bild-Embedder

Die meisten Einbettungsmodelle werden anhand von Daten aus ImageNet trainiert, einem Datensatz des Stanford Vision Lab, der mehr als 14 Millionen Bilder enthält.

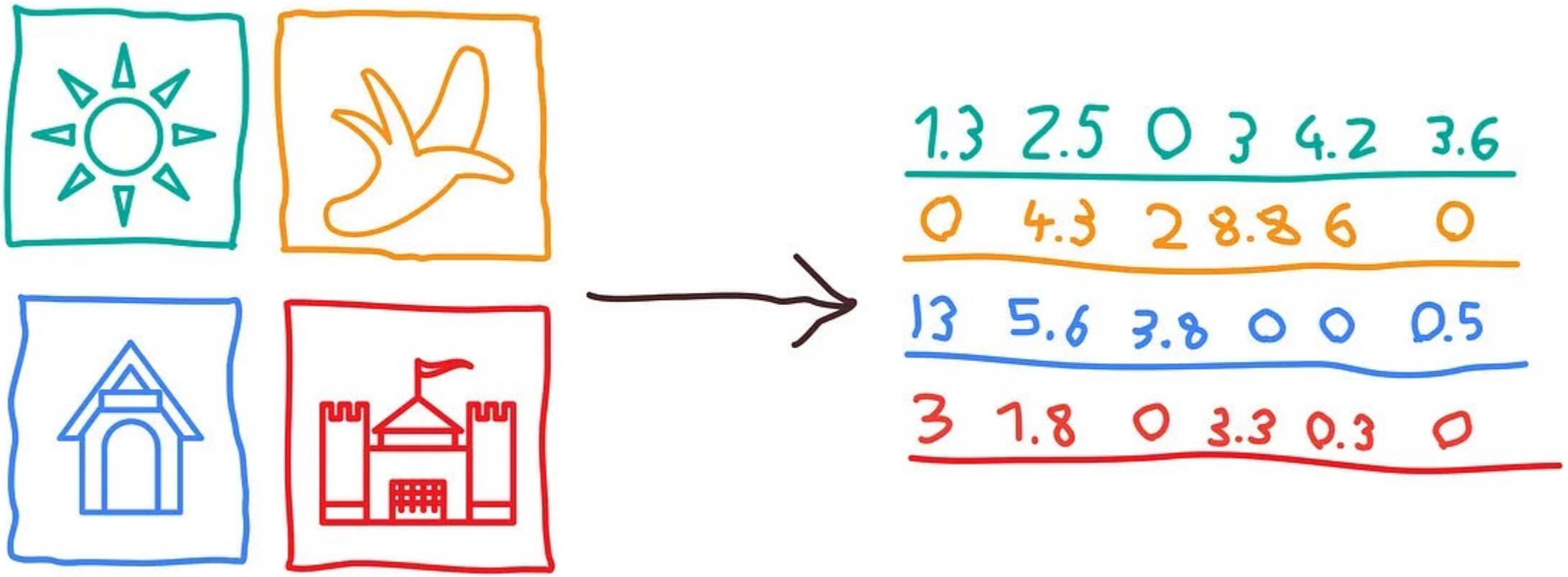
- SqueezeNet: Kleines und schnelles Modell zur Bilderkennung, das auf ImageNet trainiert wurde.
- Inception v3: Googles Inception-v3-Modell, trainiert auf ImageNet.
- VGG-16: Ein 16-schichtiges Bilderkennungsmodell, das auf ImageNet trainiert wurde.
- VGG-19: Ein 19-schichtiges Bilderkennungsmodell, das auf ImageNet trainiert wurde.
- Painters: Ein Modell, das darauf trainiert wurde, anhand von Kunstwerken die jeweiligen Maler zu identifizieren.
- DeepLoc: Ein Modell, das für die Analyse von Hefezellbildern trainiert wurde.
- openface: https://cmusatyalab.github.io/openface/ Kostenlose und quelloffene Gesichtserkennung

Quelle: Orange Data Mining-Dokumentation (enthält auch weitere Informationen zu den einzelnen Embeddern)

Erläuterung

- Um Bilder zu klassifizieren oder zu gruppieren, benötigen wir Zahlen.
- Bild-Embeddings wandeln Bilder in Zahlenfolgen (Vektoren) um.
- Die Berechnung der Vektoren ist für uns nicht nachvollziehbar – die Grundlage bilden Modelle, die anhand von annotierten Bildern trainiert wurden (wie beispielsweise Inception_v3 von Google).
- Die Vektoren sind nur für den Computer verständlich.
- Jedes Bild wird in insgesamt 2048 Zahlen umgewandelt.
- Jede Zahl beschreibt eine bestimmte Eigenschaft des Bildes.
- Ähnliche Motive im Bild führen zu ähnlichen Vektoren.
- Auf dieser Grundlage lassen sich Bilder gruppieren und klassifizieren.

Bild-Embeddings



Quelle: https://towardsdatascience.com/image-analytics-for-everyone-image-embeddings-with-orange-7f0b91fa2ca2

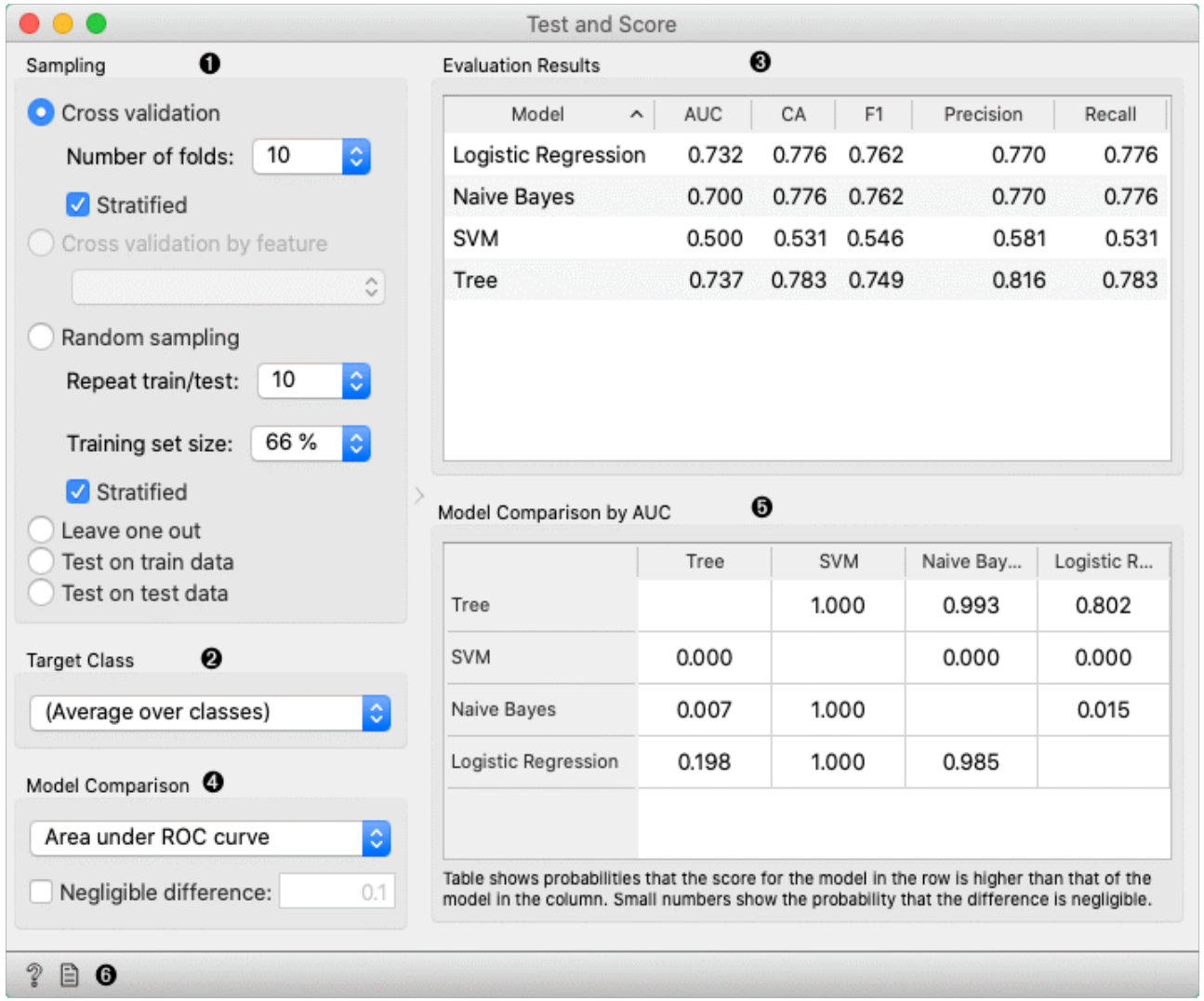
Einführung in die Analyse historischer Daten

EINHEIT 5: Bewertung der Lernergebnisse

Text- und Punktestand-Widget – 1

- Das Test- und Bewertungs-Widget bewertet „Lernende“ (Lernalgorithmen) wie beispielsweise die logistische Regression, die wir bereits für die Textklassifizierung verwendet haben.
- Das Widget erfüllt zwei Aufgaben:
  - 1) Es wendet verschiedene Methoden zur Bewertung eines oder mehrerer Lernalgorithmen an und gibt die Ergebniswerte in tabellarischer Form zurück, sortiert nach Methode;
  - 2) Es stellt die Ergebniswerte für die weitere Verwendung in anderen Widgets zur Verfügung (z. B. im „Confusion Matrix“- oder „ROC-Analyse“-Widget).

Text- und Partitur-Widget – 1 – Dialogfenster



Text- und Partitur-Widget – 2: Stichprobenverfahren – Kreuzvalidierung

- Die Bewertungsergebnisse lassen sich mithilfe der Stichprobenverfahren „Kreuzvalidierung“ und „Zufallsstichprobe“ ermitteln:
- Die Kreuzvalidierung ist eine häufig verwendete Methode zur Berechnung der Zuverlässigkeit von Lernalgorithmen. Bei der einfachen Kreuzvalidierung werden die Daten (Beobachtungen; Instanzen) zunächst gemischt und anschließend in Teilmengen (Folds) aufgeteilt (stratifiziert = die Teilmengen sind nach Möglichkeit gleich groß). Daraus ergeben sich mehrere Verteilungen der Daten in Trainings- und Testdaten. Der Lernalgorithmus wird auf jede Verteilung angewendet und die Fehlerquote berechnet. Das Ergebnis der Kreuzvalidierung ist der Durchschnitt der Fehlerraten für jede Verteilung (d. h. für jeden Fold).

Text- und Partitur-Widget – 2: Stichprobenverfahren – Zufällige Stichproben

- **Zufällige Stichprobe**: Die Daten werden in Trainings- und Testdaten aufgeteilt (z. B. 70 % Trainingsdaten; 30 % Testdaten). Der Lernalgorithmus wird dann mehrmals auf die Daten angewendet. Die Bewertungswerte ergeben sich aus der Wiederholung.

Text- und Punktestand-Widget – 4: Statistische Werte

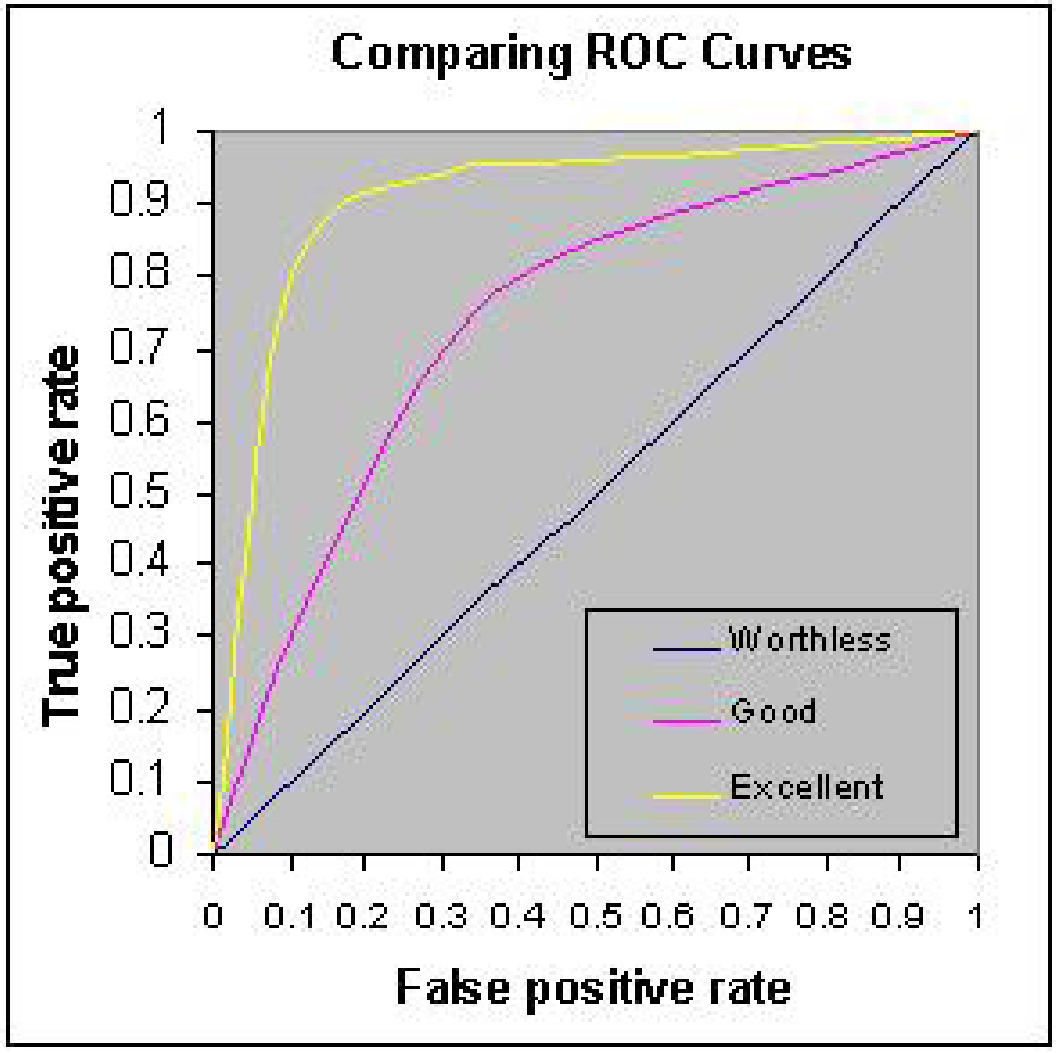
- Die folgenden statistischen Werte werden standardmäßig als Ergebnis des Kreuzvalidierungs- oder Zufallsstichprobenverfahrens für jeden verwendeten Lernalgorithmus ermittelt und angezeigt:
  - AUC
  - CA
  - F1, Präzision und Recall

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.732	0.776	0.762	0.770	0.776
Naive Bayes	0.700	0.776	0.762	0.770	0.776
SVM	0.500	0.531	0.546	0.581	0.531
Tree	0.737	0.783	0.749	0.816	0.783

Text- und Ergebnis-Widget – 4: Fläche unter der Kurve

- AUC = Fläche unter der Kurve, genauer gesagt „unter der Empfänger-Operations-Kurve“ oder ROC.
- Die AUC analysiert die Testergebnisse und vergleicht „echte“ positive Werte mit „falsch“ positiven Werten. Daraus ergibt sich eine Kurve. Die AUC ist der prozentuale Anteil der Fläche unter der Kurve. Je höher dieser Wert ist, desto besser sind die Ergebnisse (= wenige „falsch“ positive Werte).
- Auf der nächsten Folie sehen Sie ein Beispieldiagramm aus: https://darwin.unmc.edu/dxtests/roc3.htm

Text- und Noten-Widget – 4: Fläche unter der Kurve



Die folgende „Faustregel“ ist hilfreich für die Interpretation von AUC-Werten:
0,90–1 = ausgezeichnet (A)
0,80–0,90 = gut (B)
0,70–0,80 = befriedigend (C)
0,60–0,70 = schlecht (D)
0,50–0,60 = ungenügend (F)

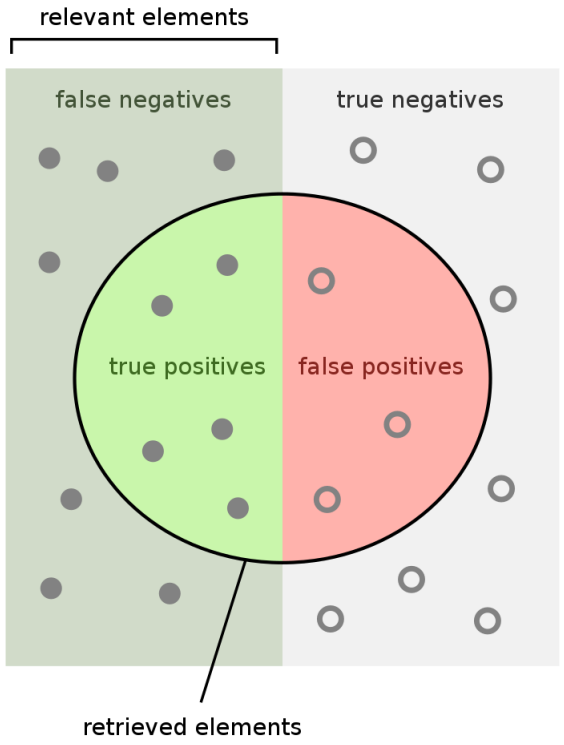
Quelle:
https://darwin.unmc.edu/dxtests/roc3.htm

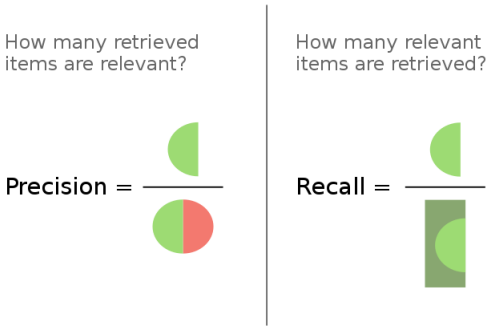
Text- und Punktestand-Widget – 4: Klassifikationsgenauigkeit

- CA = Klassifikationsgenauigkeit, d. h. der Anteil korrekt klassifizierter Beobachtungen (Instanzen)
- Genauigkeit ist ein zentrales Konzept in der Wissenschaft.
- In der Statistik wird die Genauigkeit verwendet, um Aussagen über die Datenqualität zu treffen.

Text- und Punktestand-Widget – 4: Präzision und Erfassungsgrad

- Quelle: https://en.wikipedia.org/wiki/Precision_and_recall
- Auch dies ist ein Maß für die Qualität des Lernalgorithmus
- Genauer gesagt werden zwei Kennzahlen berechnet: Präzision und Recall
- Die Präzision gibt an, wie viele der gefundenen Elemente relevant sind
- Der Recall gibt an, wie viele der relevanten Elemente gefunden wurden“
- Die Präzision kann als Maß für die Qualität und der Recall als Maß für die Quantität angesehen werden.
- Eine höhere Präzision bedeutet, dass ein Algorithmus mehr relevante als irrelevante Ergebnisse liefert, und ein hoher Recall bedeutet, dass ein Algorithmus die meisten relevanten Ergebnisse liefert (unabhängig davon, ob auch irrelevante Ergebnisse zurückgegeben werden).“ (Quelle: Wikipedia)
- F1 = gewichteter Median (Mittelwert) aus Präzision und Recall





Verwechslungsmatrix

- Das Widget „Verwechslungsmatrix“ zeigt die Beziehung zwischen tatsächlichen und vorhergesagten Werten an.
- Es nutzt die Ergebnisse des Lernalgorithmus und des Test- und Bewertungsverfahrens aus dem vorherigen Schritt.
- Die Verwechslungsmatrix zeigt, für welche Klasse die Ergebniswerte besser oder schlechter sind. Dadurch lässt sich das Klassifizierungsverfahren gezielt optimieren.