



Bewertung von ML-Verfahren mit scikit-learn

Francisca Hartmann, Elena Malkin, Look Phanthavong, Tobias Brückner, Daniel Koch

01.März.2021, Informatik-Workshop, WS 2020/2021,
Hochschule Mannheim



Zeitplan

1. Einführung

- 1.1. Machine Learning / Supervised Learning
- 1.2. Python

2. Hauptteil

- 2.1. Scikit-learn
- 2.2. Begriffserklärungen
- 2.3. Klassifikation
- 2.4. Regression
- 2.5. Cross-Validation
- 2.6. Optimierung der Hyperparameter

3. Fragerunde & Diskussion

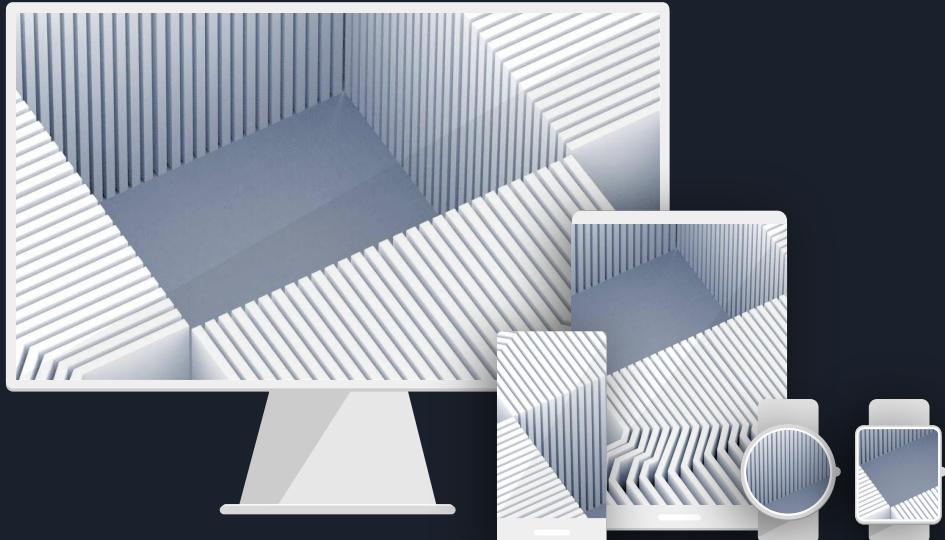
Pausen:

- Nach 2.1: 15 min
- Nach 2.3: 45 min

1.1

Machine Learning / Supervised Learning

Francisca Hartmann

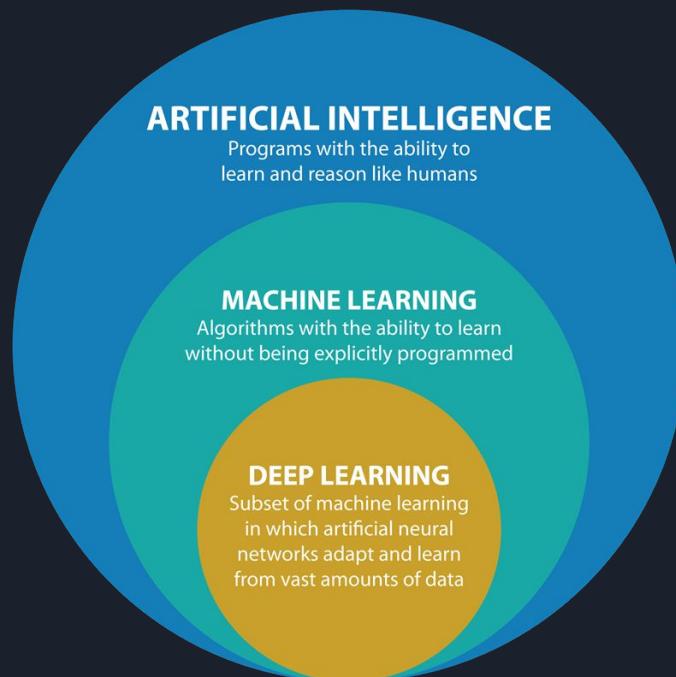




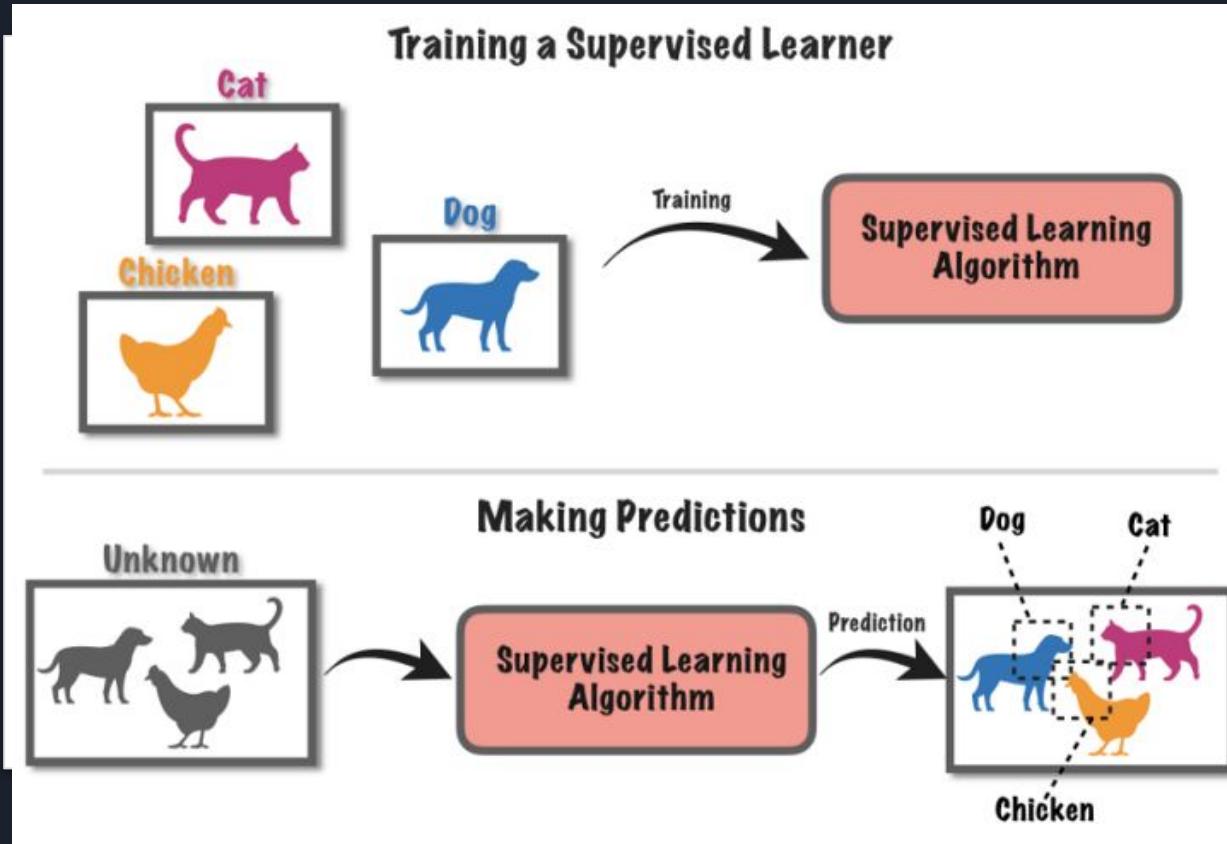
Machine Learning = Künstliche Intelligenz ?

„A computer program is said to learn from experience E with respect to some task T and performance measure P , if its performance on T , as measured by P , improves with experience E .—Tom Mitchell, 1997“

Nein



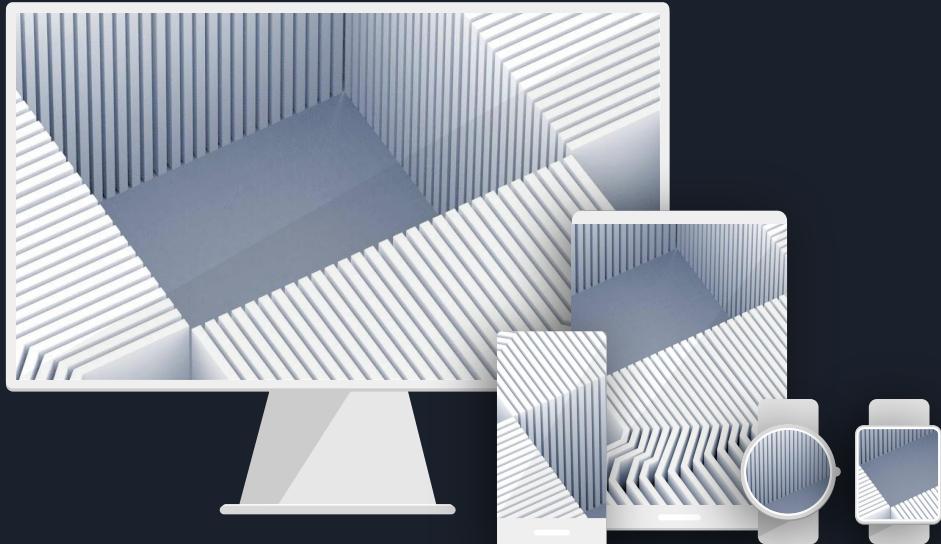
Arten des maschinellen Lernens





1.2 Python

Francisca Hartmann





Basics

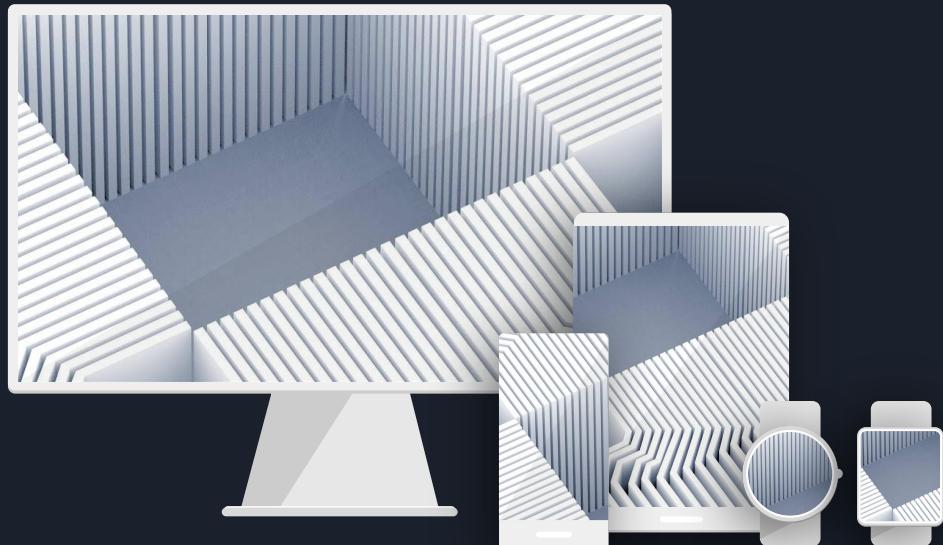
Bitte Python öffnen, wir machen ein paar Übungen zusammen.

Datensatz: Tierbabys

Danke!

2.1 Scikit-learn

Elena Malkin



Scikit-learn - Vorstellung

Scikit-Learn

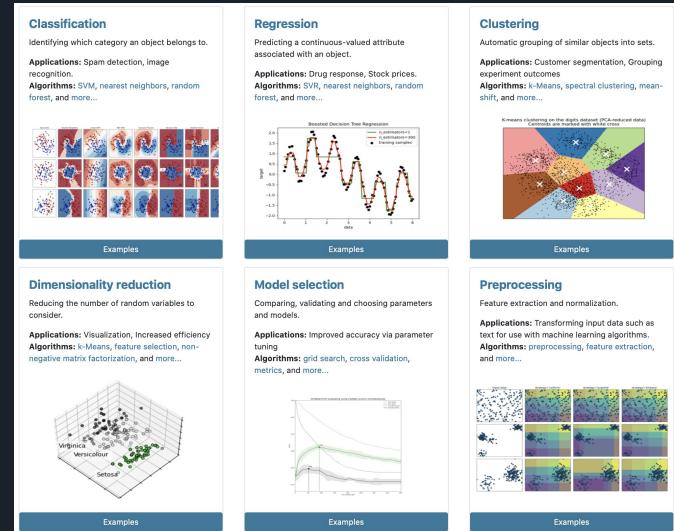
- ist eine Software-Bibliothek zum ML für Python
- open source, verfügbar unter BSD-Lizenz
- basiert auf NumPy, SciPy und matplotlib

Entstehung

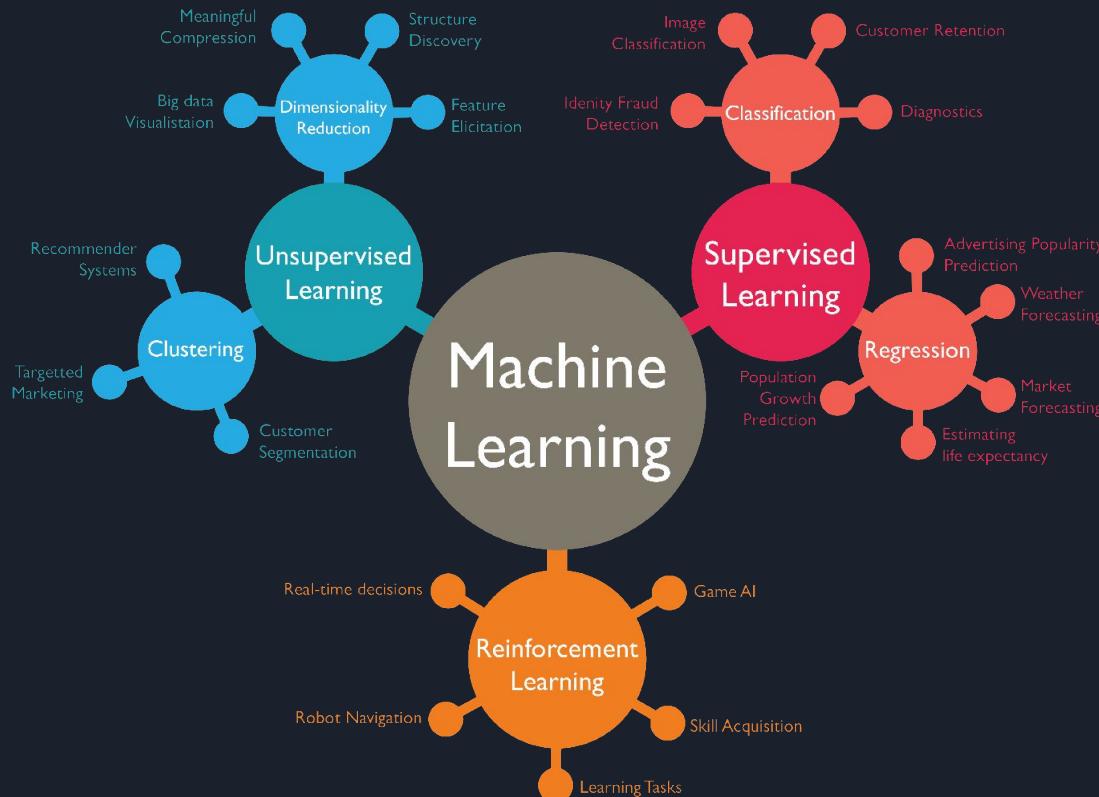
- 2007 Google Summer of Code (David Cournapeau),
Teil der Abschlussarbeit (Matthieu Brucher)
- 2010 Erste Veröffentlichung unter Leitung von INRIA
(Nationalen Forschungsinstitut für Informatik und Automatisierung)

Aktuell

- gehört zu den drei beliebtesten Bibliotheken für maschinelles Lernen für Python auf GitHub
- gute und intuitive Dokumentation, viele Beispiele
- Anfänger geeignet
- umfangreiche Sammlung von ML Algorithmen



Scikit-learn und maschinelles Lernen



Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection

Scikit-learn und maschinelles Lernen



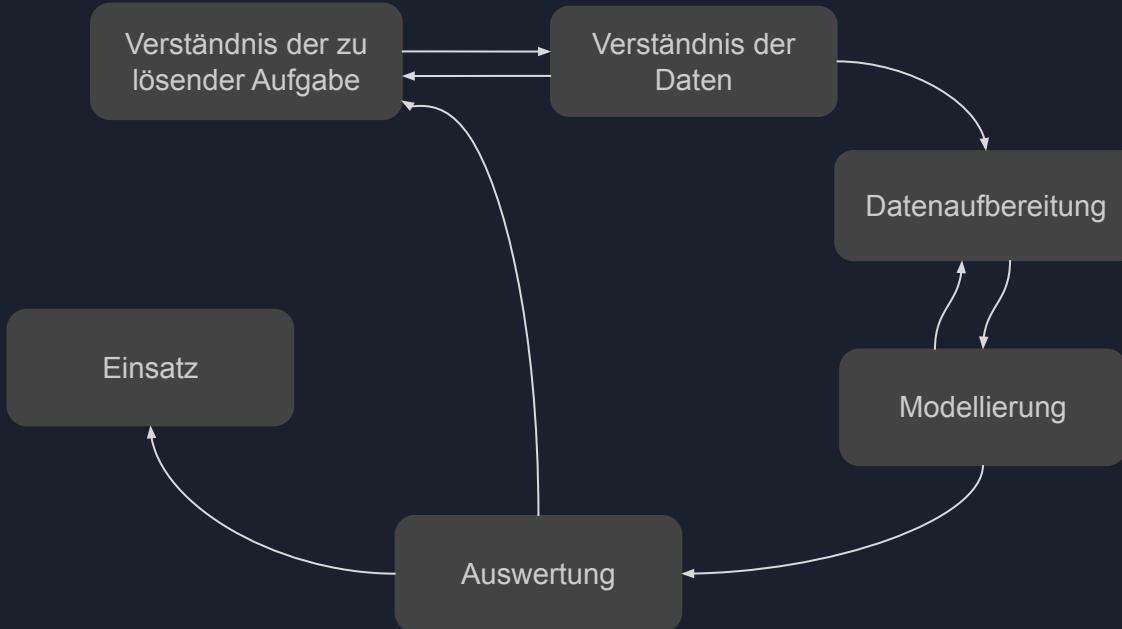
https://scikit-learn.org/stable/modules/semi_supervised.html

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection

Scikit-learn im Data-Mining-Prozess



Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection

CRISP-DM

(Cross-Industry Standard Process for Data Mining)
- branchenübergreifende Standardprozess für das Data Mining

Data-Mining

- automatische Auswertung großer Datens Mengen zur Bestimmung bestimmter Regelmäßigkeiten, Gesetzmäßigkeiten und verborgener Zusammenhänge
Quelle: Duden

Scikit-learn im Data-Mining-Prozess



Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection

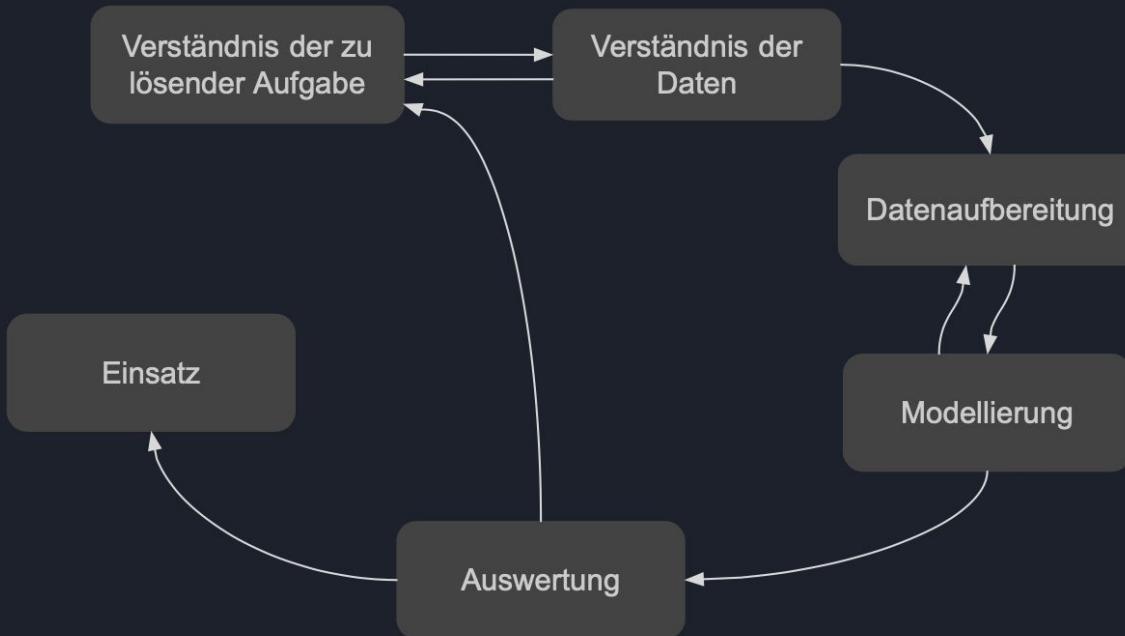
CRISP-DM

(Cross-Industry Standard Process for Data Mining)
- branchenübergreifende Standardprozess für das Data Mining

Data-Mining

- automatische Auswertung großer Datens Mengen zur Bestimmung bestimmter Regelmäßigkeiten, Gesetzmäßigkeiten und verborgener Zusammenhänge
Quelle: Duden

Scikit-learn im Data-Mining-Prozess



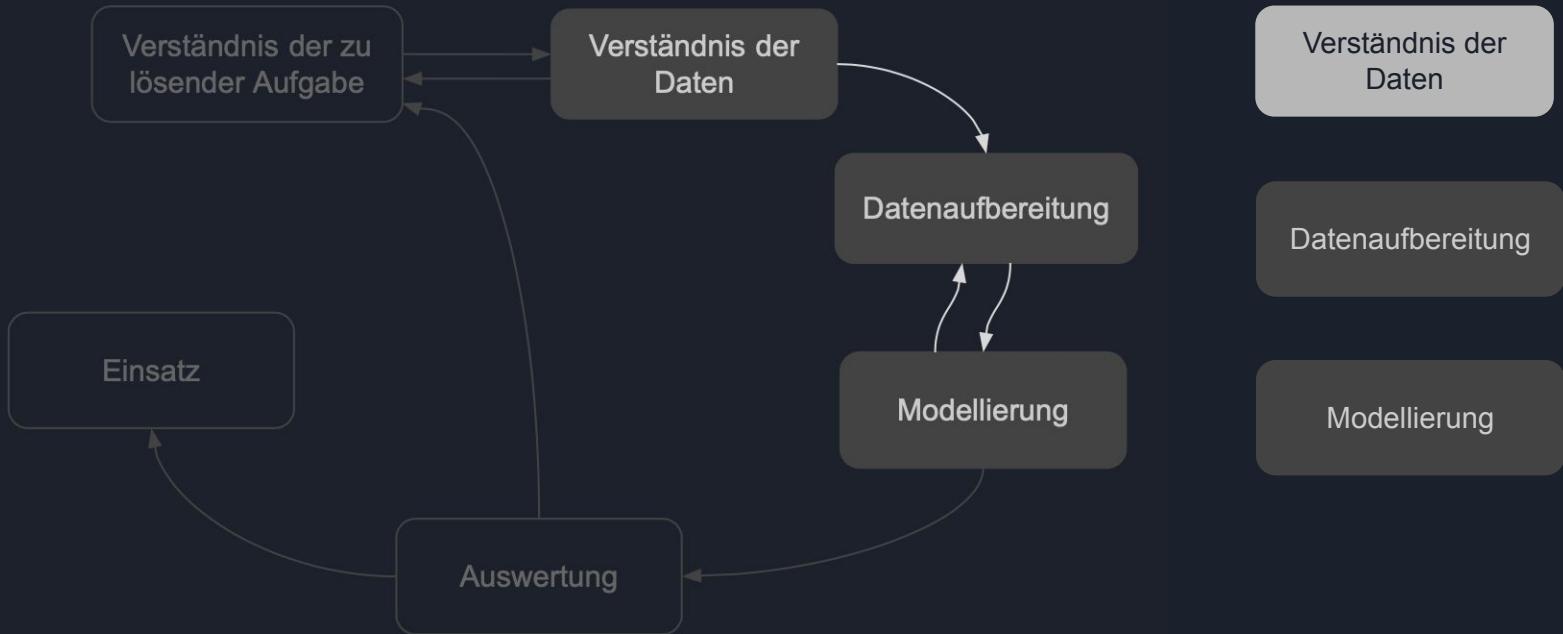
Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection

CRISP-DM
(Cross-Industry Standard Process for Data Mining)
- branchenübergreifende Standardprozess für das Data Mining

Data-Mining
- automatische Auswertung großer Datenmengen zur Bestimmung bestimmter Regelmäßigkeiten, Gesetzmäßigkeiten und verborgener Zusammenhänge
Quelle: Duden

Scikit-learn und Daten



Scikit-learn und Daten

The dataset loaders	Instanzen	Dimensionen
.load_iris	150	4
.load_breast_cancer	569	30
.load_digits	1797	64
...		

The dataset fetchers	Instanzen	Dimensionen
.fetch_20newsgroups	18.846	1
.fetch_kddcup99	4.898.431	41
.fetch_rcv1	804.414	47.236
...		

Verständnis der Daten

Datenaufbereitung

Modellierung

The dataset generation functions

- Klassifikation (`sklearn.datasets.make_classification`,
`sklearn.datasets.make_multilabel_classification`)
- Regression (`sklearn.datasets.make_regression`)
- ...

Andere Formate und Quellen

- Datensätze von „<https://openml.org>“ mit `sklearn.datasets.fetch_openml`
- txt - `datasets.load_files`
- CSV, Excel, JSON and SQL - mit `panda`
- Spaltenorientierte DB in NumPy Array - mit `NumPy`
- ...

Scikit-learn und Datenaufbereitung



Normalisierung
Diskretisierung

...

Scikit-learn Bereiche:

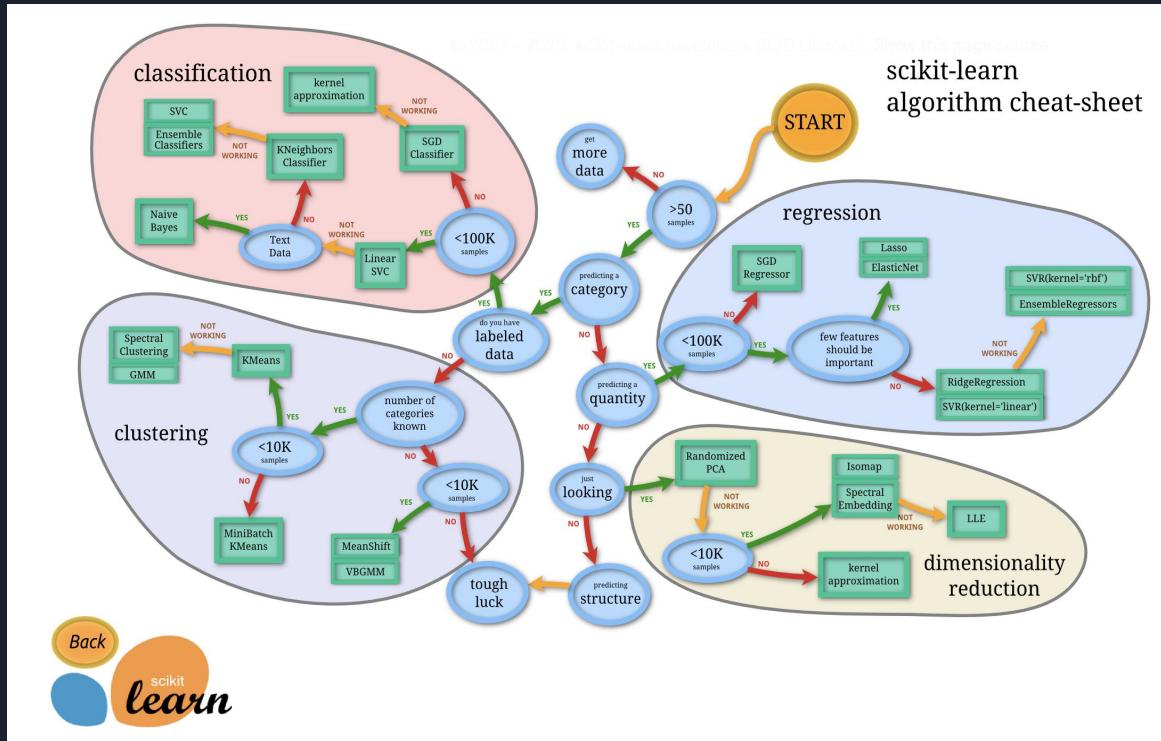
- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing**
- Model selection und evaluation

Verständnis der Daten

Datenaufbereitung

Modellierung

Scikit-learn und Modellierung



Scikit-learn Bereiche:

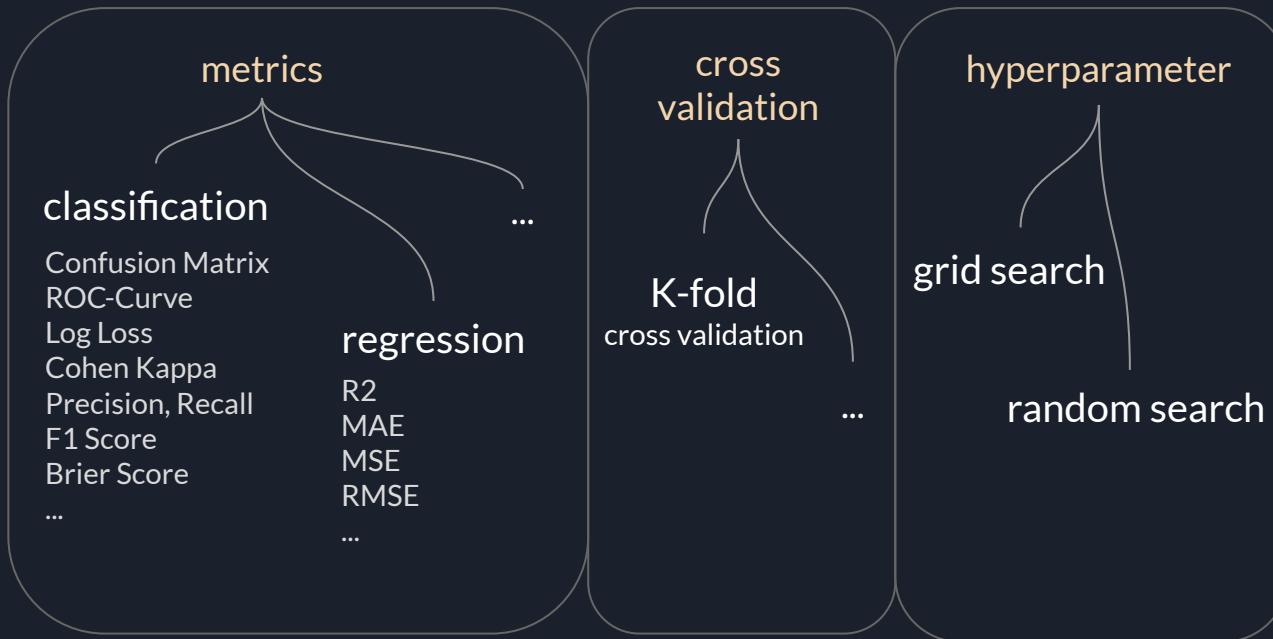
- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- **Model selection and evaluation**

Verständnis der Daten

Datenaufbereitung

Modellierung

Scikit-learn und Evaluierung



Scikit-learn Bereiche:

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Preprocessing
- Model selection and evaluation**

Verständnis der Daten

Datenaufbereitung

Modellierung



Scikit-learn und weitere Bibliotheken

Scikit-Learn basiert auf

- SymPy (Python library for symbolic mathematics) Quelle: <https://www.sympy.org>
- NumPy (fundamental package for scientific computing with Python) Quelle: <https://numpy.org>
- Matplotlib (is a comprehensive library
for creating static, animated, and interactive visualizations in Python) Quelle: <https://matplotlib.org>

Scikit-Learn

- einfach zu verwenden, viele effiziente Algorithmen für ML
- bei kleinen bis mittelgroßen Datensätzen schneller als anderen Bibliotheken* für ML
- Reinforcement Learning, Deep Learning nicht umgesetzt

Deep Learning

- TensorFlow, Keras, PyTorch

*Shogun, PyMVPA, MDP, MLPy, milk, PyBrain



Verwendung

Scikit Learn wird in Wirtschaft und in Forschung eingesetzt

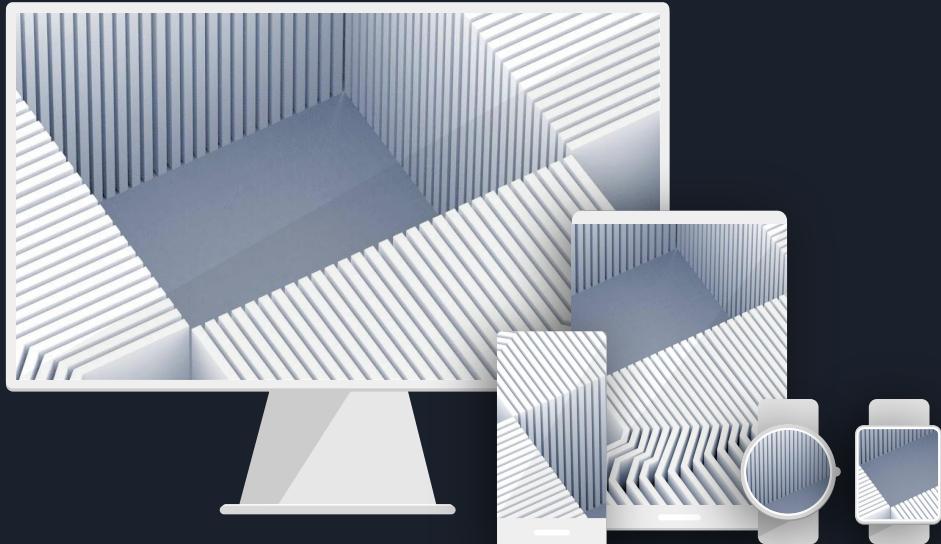
- Kampf gegen Betrug und Spam
- Analyse medizinischer Bilder
- Vorhersage des Nutzerverhaltens
- Optimierung industrieller und logistischer Prozesse.
- Vorhersage des Kaufverhaltens der Benutzer um Produktempfehlungen anzubieten, Trends und missbräuchliches Verhalten zu erkennen.

Sponsorenkonsortium: BCG Gamma, Microsoft, Axa, BNP Paribas Cardif, Intel, Nvidia und Dataiku



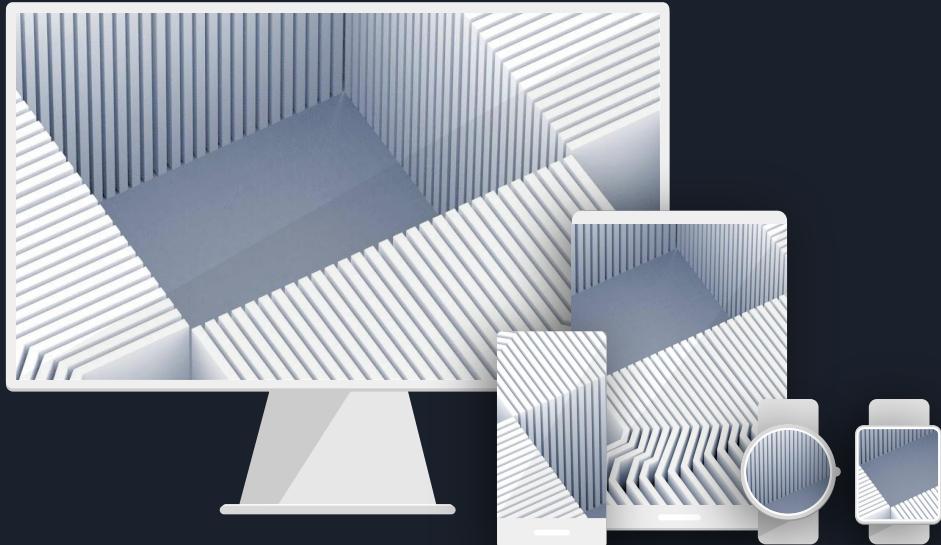
Kleine Pause

15 Minuten



2.2 Begriffserklärungen

Francisca Hartmann





Begriffe

Feature

Eingabe in System

Unabhängige Variablen

Eine Spalte des Datensatzes = ein Feature

Anzahl Features = dimension

Target

kategorisch/kontinuierlich

Unabhängig von Ausgabe der
EingabevARIABLEN

Ausgabe (möchte man
vorhersagen)

Label

Endgültige Ausgabe

Abhängige Variablen

Nicht beim unüberwachten Lernen



Begriffe

Overfitting

Zu viele erklärende Variablen

Nicht vorhandene Muster erkennt

Auswendig lernen

-> kleiner Vorhersagefehler bei
Trainingsdaten, aber hohen bei
Testdaten

Underfitting

Zu wenig relevante Variablen

Keine relevanten Schlüsse

-> hoher Vorhersagefehler bei
Trainings- und Testdaten

Beispiel

Overfitting: Student mit fotografischen Gedächtnis, der nicht abstrahieren kann

Underfitting: Student mit veralteten, fehlerhaften Skript



Begriffe

Regularisierung

Schätzung
Komplexität

Over-Underfitting
vermeiden

Parameter

Konfigurationsvariablen
Modellintern
Aus Trainingsdaten

Hyperparameter

Nicht aus Trainingsdaten
Kombination aus Heuristiken
und Erfahrung

Training

Lernen
Muster finden
Ergebnis: Lernmodell



Begriffe

Trainingsset

Model trainieren

Testset

Model testen

Predict

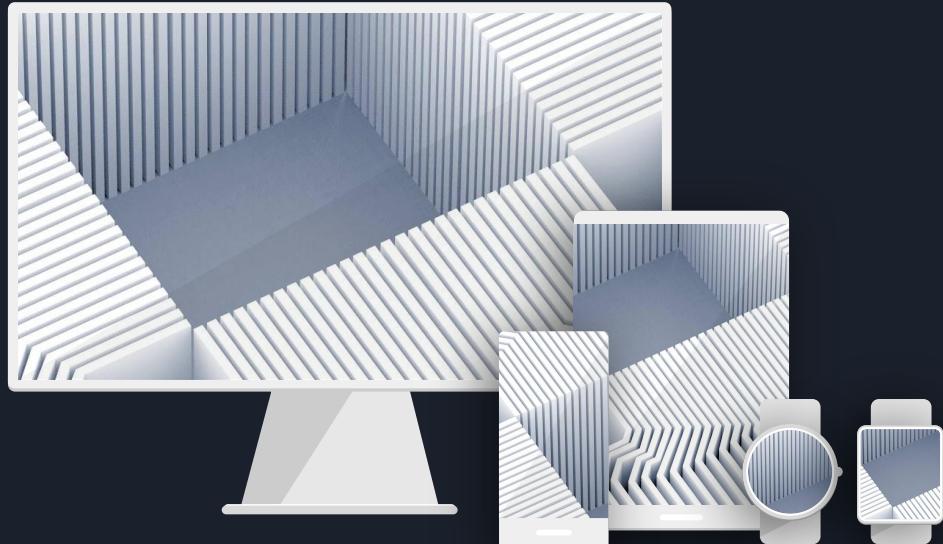
Vorhersagen

Kleiner als Trainingsset

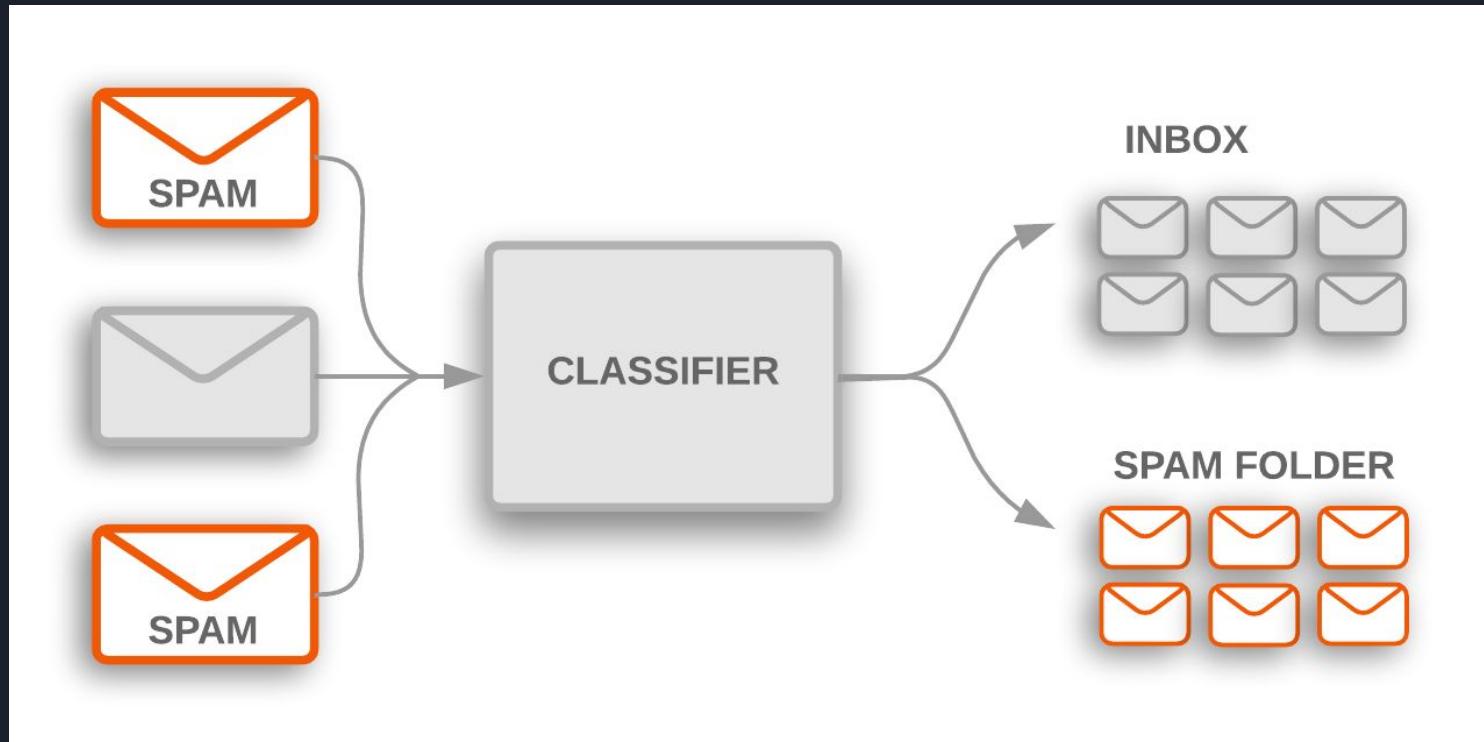
Bezug auf trainierten Algorithmus
angewendet auf neue Daten

2.3 Klassifikation

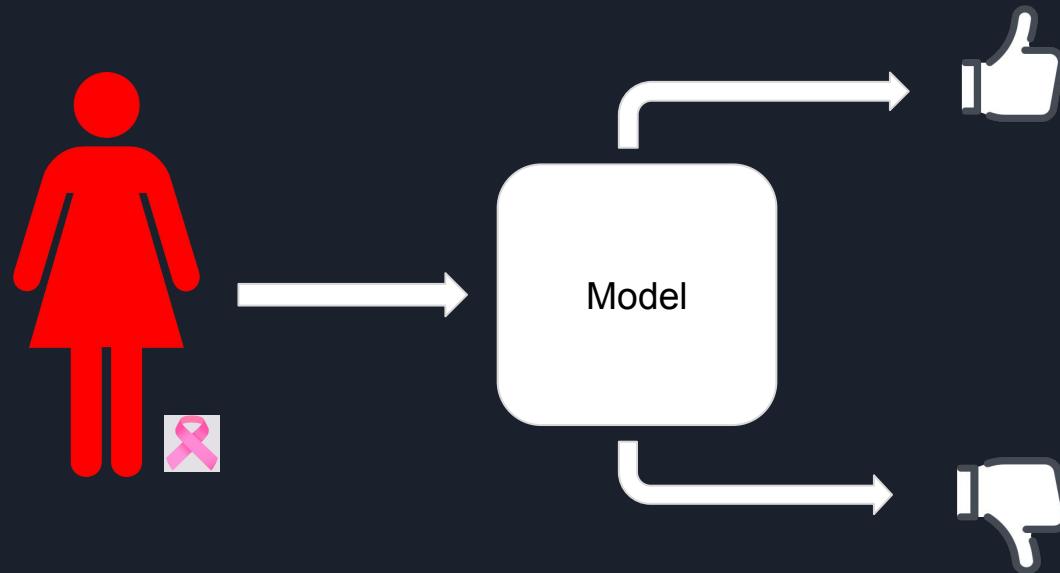
Look Phanthavong, Elena Malkin



2.3 Klassifikation



Anwendungsbeispiel 1 - Breast Cancer



Liste von Bewertungsmetriken

Classification

'accuracy'	<code>metrics.accuracy_score</code>
'balanced_accuracy'	<code>metrics.balanced_accuracy_score</code>
'top_k_accuracy'	<code>metrics.top_k_accuracy_score</code>
'average_precision'	<code>metrics.average_precision_score</code>
'neg_brier_score'	<code>metrics.brier_score_loss</code>
'f1'	<code>metrics.f1_score</code>
'f1_micro'	<code>metrics.f1_score</code>
'f1_macro'	<code>metrics.f1_score</code>
'f1_weighted'	<code>metrics.f1_score</code>
'f1_samples'	<code>metrics.f1_score</code>
'neg_log_loss'	<code>metrics.log_loss</code>
'precision' etc.	<code>metrics.precision_score</code>
'recall' etc.	<code>metrics.recall_score</code>
'jaccard' etc.	<code>metrics.jaccard_score</code>
'roc_auc'	<code>metrics.roc_auc_score</code>
'roc_auc_ovr'	<code>metrics.roc_auc_score</code>
'roc_auc_ovo'	<code>metrics.roc_auc_score</code>
'roc_auc_ovr_weighted'	<code>metrics.roc_auc_score</code>
'roc_auc_ovo_weighted'	<code>metrics.roc_auc_score</code>

Bildquelle:

https://scikit-learn.org/stable/modules/model_evaluation.html#model-evaluation

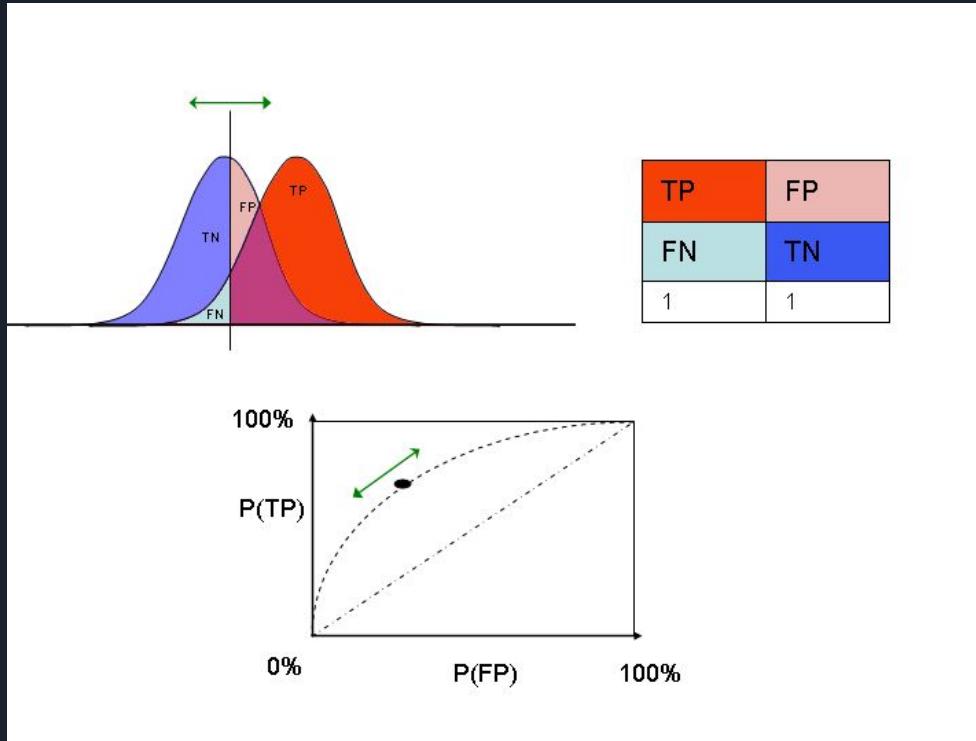
Confusion Matrix

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Bildquelle:

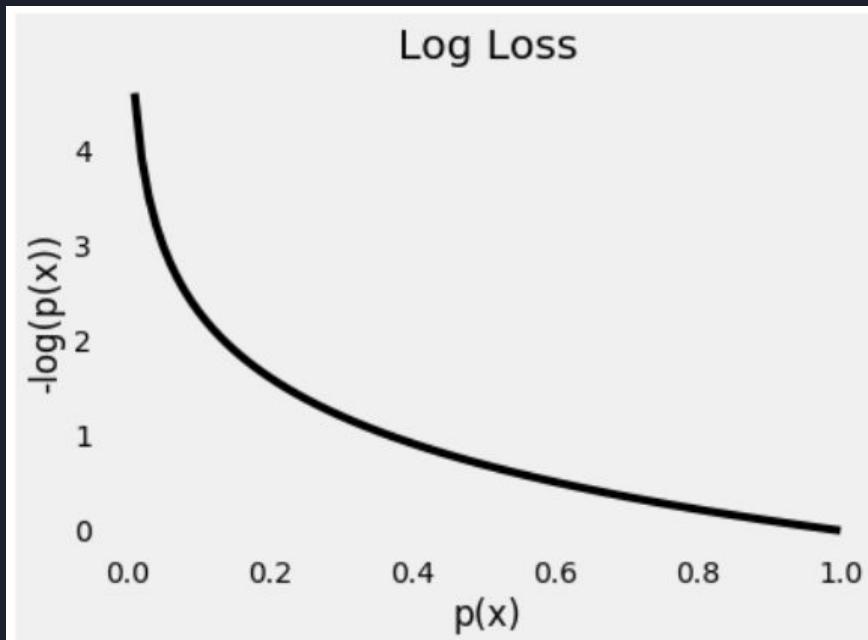
<https://www.kaggle.com/getting-started/175898>

Receiver Operating Characteristic - Curve



Bildquelle:
https://de.wikipedia.org/wiki/ROC-Kurve#/media/Datei:Receiver_Operating_Characteristic.png

Log Loss



- y = Zielvariable
- $p(y)$ = Wahrscheinlichkeit, dass die Instanz zu y gehört
- Hoher Verlust, wenn $p(y)$ "schlecht" ist

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Bildquelle:

<https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>

Cohen's Kappa

- Indikator für die Übereinstimmung von **zwei** Beobachter
- Bewertung der Leistung eines Klassifikationsmodells

Besonderheit

Berücksichtigung der Zufallsübereinstimmung

Wertebereich

-1 und +1, je höher desto besser

Mehr als zwei Beobachter

Kappa paarweise berechnen, Median bilden

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

> 0,75	sehr gut
0,6 - 0,75	gut
0,4 - 0,6	ausreichend

Cohen's Kappa

		Predicted Class			
		Positive	Negative		
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$	
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$	
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$	

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$p_o = \frac{TP + TN}{TP + TN + FP + FN}$$

Cohen's Kappa

		Predicted Class			
		Positive	Negative		
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $b = \frac{TP}{TP + FN}$	
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $d = \frac{TN}{TN + FP}$	
	Precision $a = \frac{TP}{TP + FP}$	Negative Predictive Value $c = \frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$		

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$p_e = \frac{a*b + c*d}{\text{Accuracy}^2}$$

Cohen's Kappa

- Indikator für die Übereinstimmung von **zwei** Beobachter
- Bewertung der Leistung eines Klassifikationsmodells

Besonderheit

Berücksichtigung der Zufallsübereinstimmung

Wertebereich

-1 und +1, je höher desto besser

Mehr als zwei Beobachter

Kappa paarweise berechnen, Median bilden

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

> 0,75	sehr gut
0,6 - 0,75	gut
0,4 - 0,6	ausreichend



Achtung !!!



Achtung:

Ein Modell ist nur so gut wie seine Bewertungsmethode !!!

Frage:

Warum macht die Accuracy bei Imbalanced Datensätzen keinen Sinn?

Anwendungsbeispiel 2 - Bank Marketing



Bildquelle: <https://www.socialtoaster.com/retail-bank-marketing-increase-loyalty/>

Datensatz: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

Confusion für Imbalanced Data

		Predicted Class		(Recall)
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Bildquelle:

<https://www.kaggle.com/getting-started/175898>



Precision und Recall

Precision:

$$\text{TP}/(\text{TP}+\text{FP})$$

Anteil der korrekt klassifizierten
Instanzen, die als positiv
vorhergesagt wurden

Recall:

$$\text{TP}/(\text{TP}+\text{FN})$$

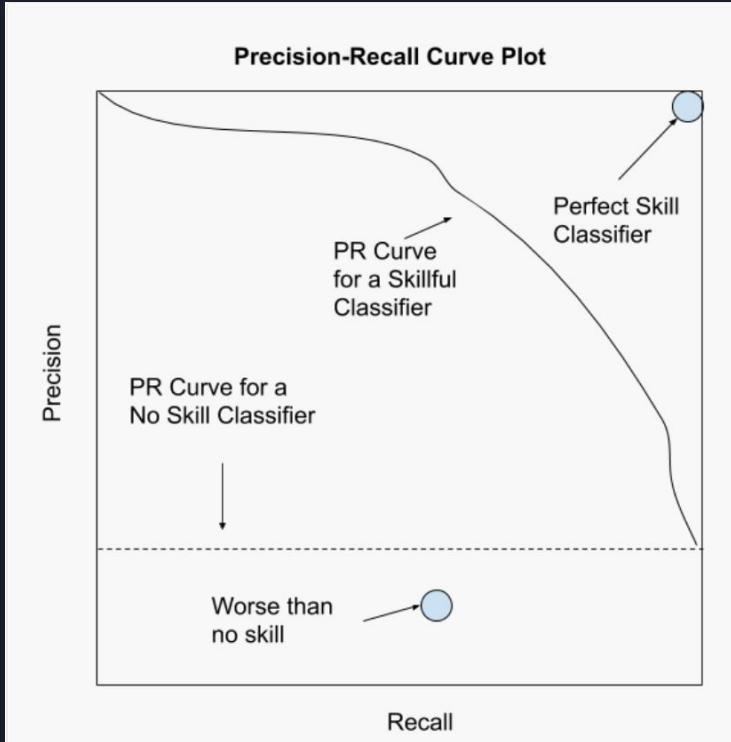
Anteil der korrekt klassifizierten
Instanzen, die wirklich positiv sind



F1 Score

- Berechnet das harmonische Mittel zwischen Precision und Recall
- $F1\text{ Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$
- F1 Score = 0 -> Schlecht
- F1 Score = 1 -> Gut

Precision Recall Curve



- Alternative für ROC
- Fokus auf die Minority Class
- (1,1) -> Perfektes Modell

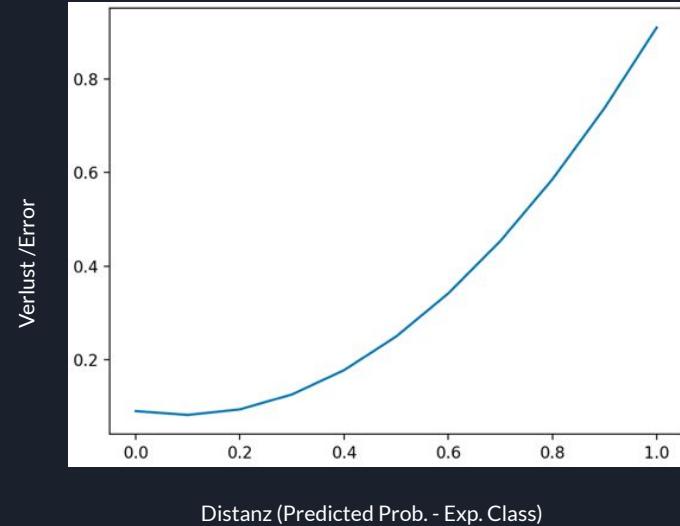
Brier Score

$$BS = \frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2$$

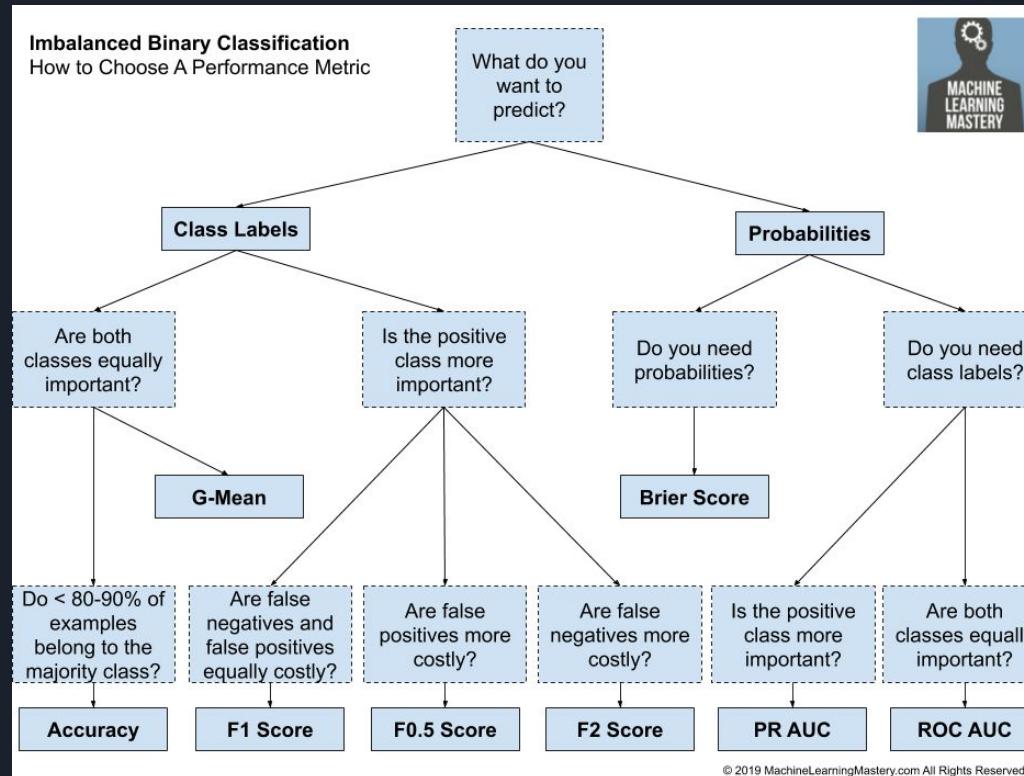
Bildquelle:
<https://www.statisticshowto.com/brier-score/>

- N = the number of items you're calculating a Brier score for.
- f_t is the forecast probability (i.e. 25% chance),
- o_t is the outcome (1 if it happened, 0 if it didn't).
- Σ is the **summation symbol**. It just means to “add up” all of the values.

Kurvenverlauf -> Brier Score



Road Map für Imbalanced Data



© 2019 MachineLearningMastery.com All Rights Reserved.

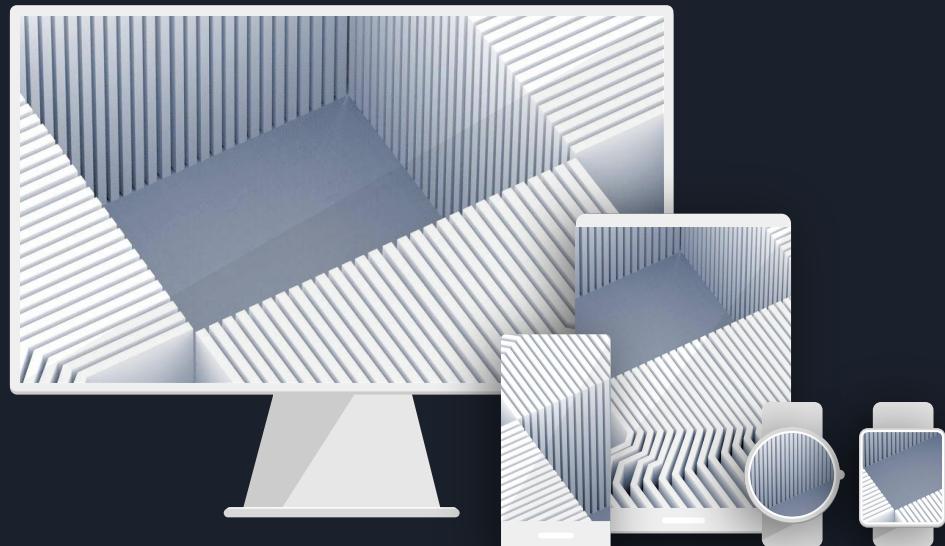


Fazit

- Ein Modell ist nur so gut wie seine Bewertungsmetrik !!!
- Es gibt Balanced und Imbalanced Data
- Es gibt dafür jeweils eigene Bewertungsmetriken
- Außerdem bietet SciKit-Learn eine Menge von diesen Metriken bereits an
- Es ist sogar möglich eigene Metriken zu verwenden
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html?highlight=make_scorer#sklearn.metrics.make_scorer

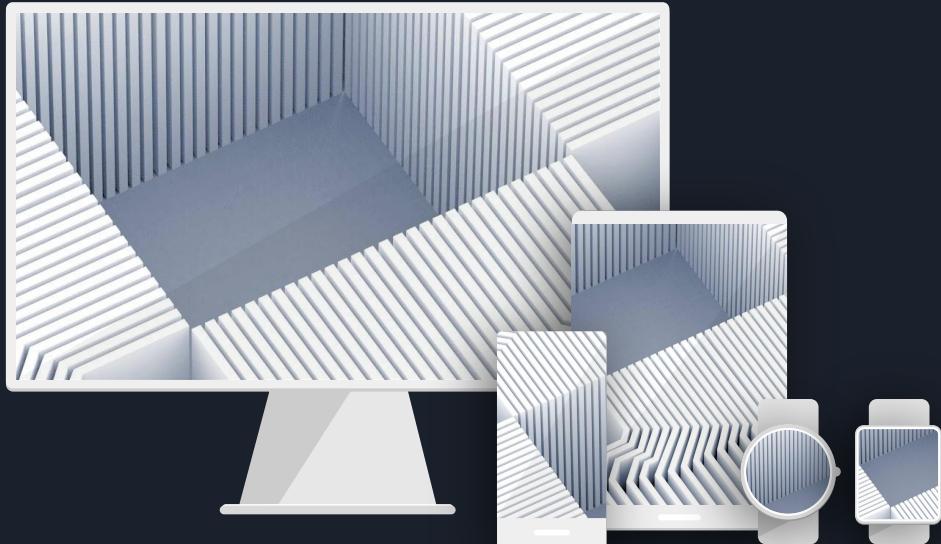
Große Pause

45 Minuten



2.4 Regression

Tobias Brückner



Beispieldaten für Regression

- Sportwagendatensatz
- 9 Features
 - Hersteller
 - Jahr des Fahrzeugberichts
 - Leistung in PS
 - Drehmoment in Nm
 - Hubraum in cm³
 - Anzahl Zylinder
 - Höchstgeschwindigkeit in km/h
 - Leergewicht in kg
 - Preis in €
- Target = Beschleunigung (Zeit benötigt für eine Viertelmeile in Sekunden)

	Modell	Hersteller	Jahr	Leistung_PS	Drehmoment_Nm	\
0	AMG GT Black Series	Mercedes	2020	730.0	800.0	
1	Focus ST	Ford	2020	280.0	420.0	
2	Golf GTI	Volkswagen	2020	245.0	370.0	
3	i30 N	Hyundai	2020	275.0	378.0	
4	Panamera Turbo S	Porsche	2020	630.0	820.0	

	Hubraum_cm3	Zylinder	Höchstgeschwindigkeit_kmh	Leergewicht_kg	\
0	3982.0	8	325.0	1626.0	
1	2261.0	4	250.0	1455.0	
2	1984.0	4	250.0	1408.0	
3	1998.0	4	250.0	1478.0	
4	3996.0	8	310.0	2140.0	

	Preis_euro	BeschleunigungViertelmeile_sek	
0	342780.0	10.71	
1	37432.0	14.05	
2	38144.0	14.60	
3	35676.0	14.38	
4	185982.0	11.00	

Beispieldaten für Regression

Target Variable

	BeschleunigungViertelmeile_sek
count	154.000000
mean	12.473312
std	1.303216
min	9.990000
25%	11.540000
50%	12.290000
75%	13.295000
max	15.440000

- 154 Instanzen
- Im Schnitt eine Beschleunigung von circa 12,5 Sekunden

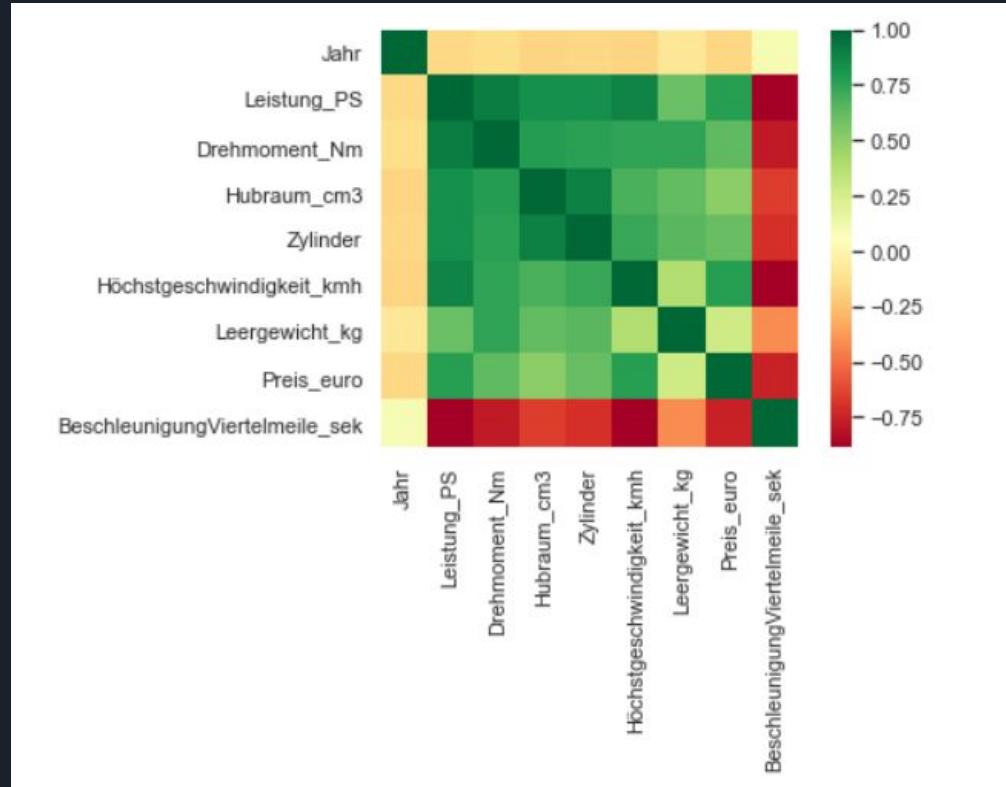
Beschreibung der Feature Variablen

	Jahr	Leistung_PS	Drehmoment_Nm	Hubraum_cm3	Zylinder	\
count	154.000000	154.000000	154.000000	154.000000	154.000000	
mean	2018.915584	453.772727	561.987013	3519.155844	6.441558	
std	1.084359	161.806732	179.725389	1399.665554	2.175283	
min	2016.000000	192.000000	205.000000	1497.000000	3.000000	
25%	2018.000000	306.000000	400.000000	1998.000000	4.000000	
50%	2019.000000	450.000000	535.000000	3528.500000	6.000000	
75%	2020.000000	600.000000	700.000000	4395.000000	8.000000	
max	2020.000000	887.000000	959.000000	6592.000000	12.000000	

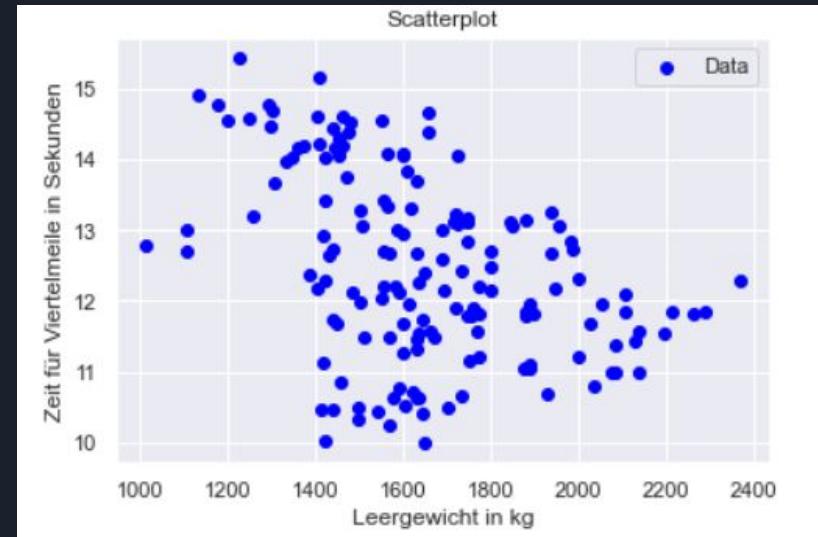
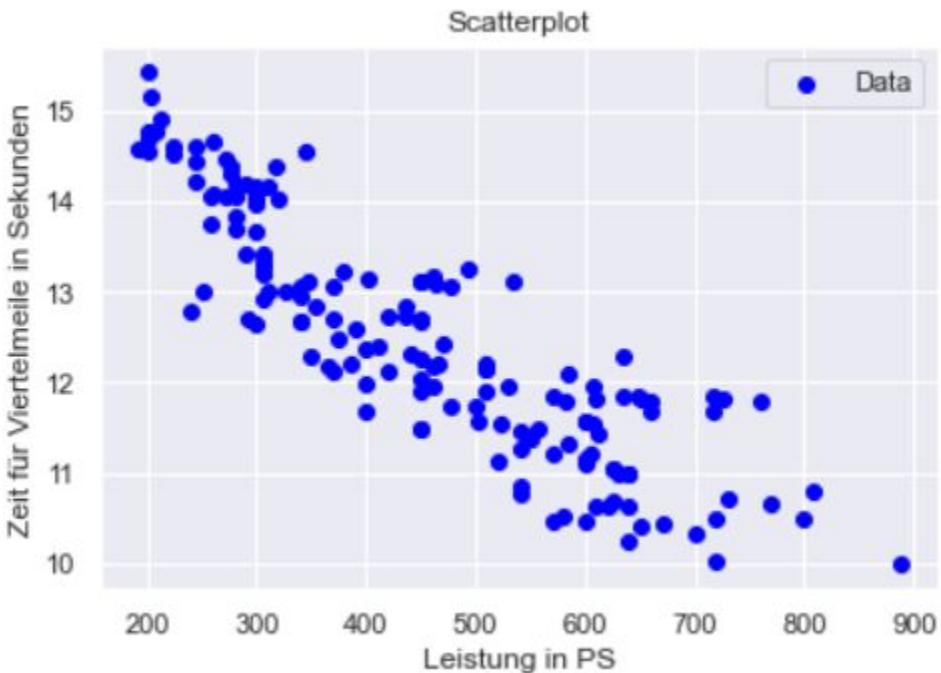
	Höchstgeschwindigkeit_kmh	Leergewicht_kg	Preis_euro	\
count	154.000000	154.000000	154.000000	
mean	282.727273	1650.103896	117236.168831	
std	33.401161	259.144878	91084.420397	
min	225.000000	1016.000000	24600.000000	
25%	250.000000	1460.500000	54970.750000	
50%	280.000000	1629.000000	91676.500000	
75%	313.500000	1777.000000	164564.500000	
max	350.000000	2368.000000	768026.000000	

Beispieldaten für Regression

Korrelationsmatrix



Beispieldaten für Regression



- Negativer Linearer Zusammenhang zwischen Leistung und Beschleunigung
- Eher kein Zusammenhang mit Gewicht

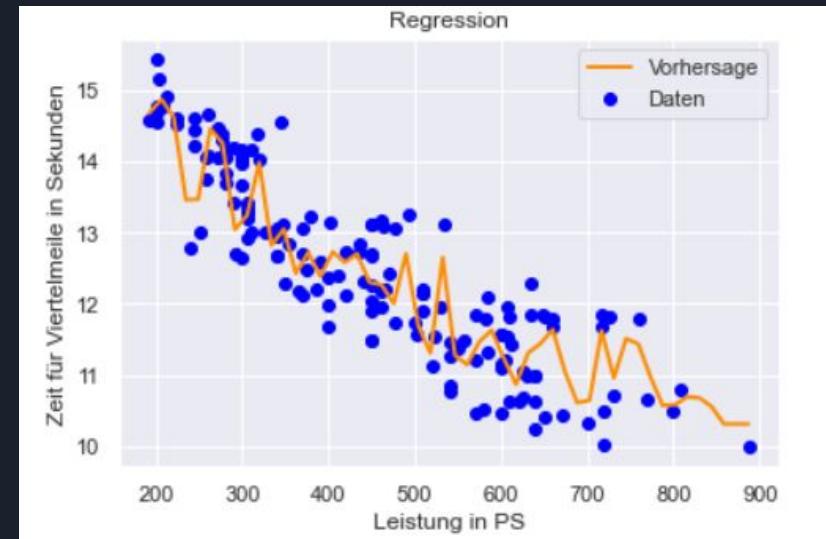
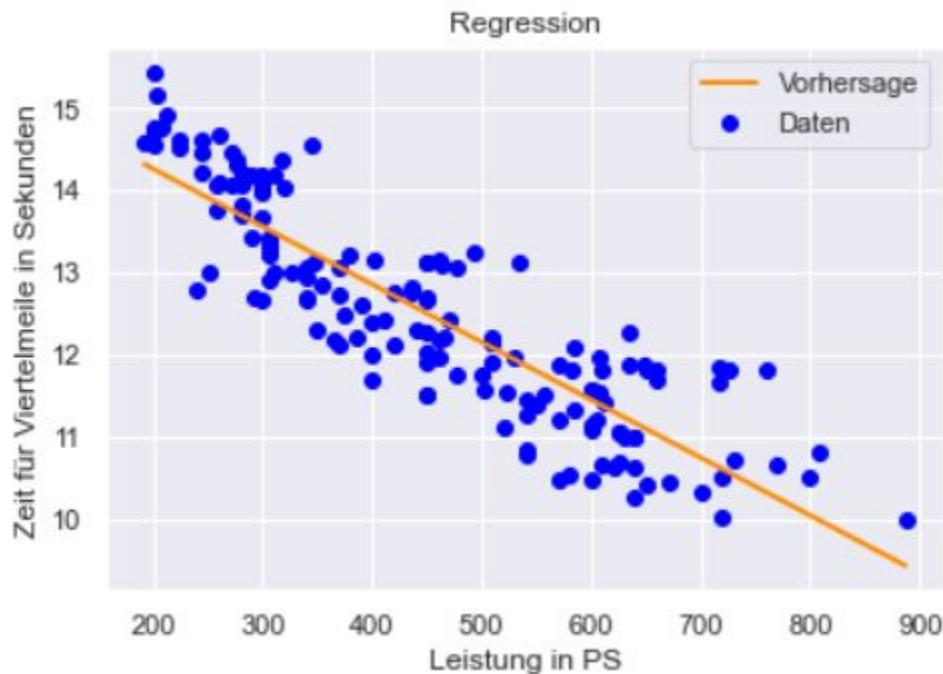
Grundidee Regressionsmodelle

- Gegeben: n Wertepaare mit Messungen der Zielvariablen
- Ziel: Modell finden, dass bei gegebenen Wertekombinationen einen guten Wert für das Target schätzt
- Residuum = Abweichung einer Schätzung von der zugehörigen Messung
- Ansatz: Modell soll Residuenquadrate minimieren

Einfachster Schätzer: Mittelwert



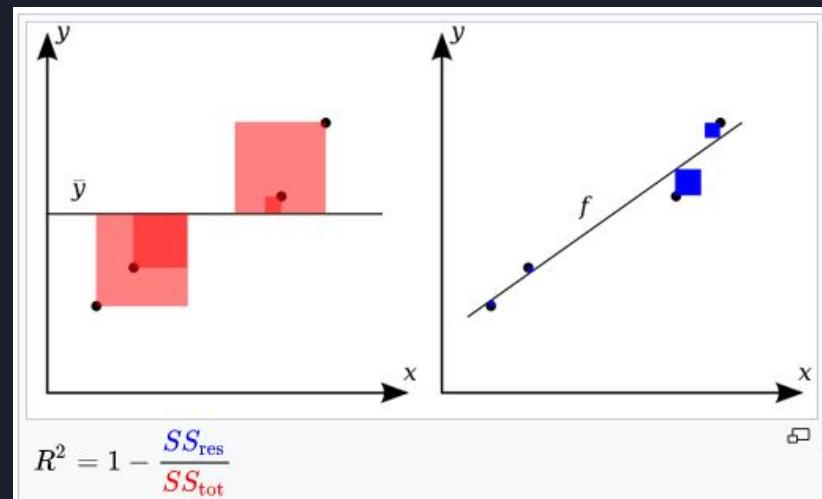
Grundidee Regressionsmodelle



Links: Lineares Modell (Ridge Regression)
Rechts: Random Forest

Bewertung: R² Bestimmtheitsmaß

- Definition: Anteil der durch die Regression erklärten Varianz an der zu erklärenden gesamten Varianz
→ Wie viel Streuung in den Daten kann durch das Modell erklärt werden?
- Dimensionslos, Wert nahe 1 bedeutet hohe Güte der Anpassung



SS_tot = Totale Quadratsumme

SS_res = Residuenquadratsumme

(Quelle: https://en.wikipedia.org/wiki/Coefficient_of_determination)

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$



Bewertung: R² Bestimmtheitsmaß

- Beachte: Durch die Aufnahme zusätzlicher erklärender Variablen kann das Bestimmtheitsmaß nicht sinken
 - Zusätzliche unabhängige Variablen könnten keinen Beitrag zur Erklärungskraft liefern und R² würde trotzdem steigen
- Ausblick: Adjustiertes (freiheitsgradbezogenes) Bestimmtheitsmaß
 - Residuenquadratsumme durch die residualen Freiheitsgrade dividieren, um das mittlere Residuenquadrat zu erhalten

Bewertung: MAE (Mean Absolute Error)

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|.$$

- Durchschnittlicher absoluter Fehler
- Einfachste und intuitive Metrik
- Behandelt alle Datenpunkte gleich
- Jeder Fehler beeinflusst das Ergebnis proportional zum absoluten Wert des Fehlers

Bewertung: MSE (Mean Squared Error)

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

- Mittlerer quadratischer Fehler
→ erwarteter quadratischer Abstand, den ein Schätzer vom wahren Wert hat
- Ausreißer haben großen Einfluss auf Gesamtfehler

Bewertung: RMSE (Root Mean Squared Error)

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- Quadratwurzel der mittleren quadratischen Fehler
- Einfluss jedes Fehlers ist proportional zur Größe des quadratischen Fehlers
→ Größere Fehler haben eine überproportionale Wirkung



Bewertung: Vergleich

- MAE und RMSE einfacher zu interpretieren als MSE, da gleiche Einheit wie Zielvariable
- Vorhersagefehler verschiedener Modelle für einen bestimmten Datensatz können verglichen werden
- Kein Vergleich zwischen unterschiedlichen Datensätzen möglich wegen Abhängigkeit von der Skala
- Ausblick: Normalisierung möglich über Teilen durch Spannweite/Range oder durch Mittelwert



Bewertung: Realisierung in scikit-Learn

- Scikit-Learn hat vordefinierte Werte für Scorer-Objekte
- All diese folgen der Konvention, dass höhere Rückgabewerte besser sind als niedrigere
- Bei Metriken, in denen der Vorhersagefehler in Abhängigkeit von einem Abstand zu den Daten gemessen wird, sind höhere Werte jedoch in der Regel schlechter

Bewertung: Realisierung in scikit-Learn

Trainieren des Modells

```
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.1, random_state = 1234)
ridge = Ridge(alpha = 0.01, normalize = True)
ridge.fit(X_train, y_train)
predictions = ridge.predict(X_test)
```

Bewertung: Realisierung in scikit-Learn

Mean Absolute Error (MAE)

```
from sklearn.metrics import mean_absolute_error

mae1 = np.average(np.abs(y_test - predictions))
print("MAE manuell: {}".format(mae1))

mae2 = mean_absolute_error(y_test, predictions)
print("MAE scikit-learn: {}".format(mae2))
```

MAE manuell: 0.3325739110428545

MAE scikit-learn: 0.3325739110428545

Bewertung: Realisierung in scikit-Learn

Mean Squared Error (MSE)

```
from sklearn.metrics import mean_squared_error

mse1 = np.average((y_test - predictions) ** 2)
print("MSE manuell: {}".format(mse1))

mse2 = mean_squared_error(y_test, predictions)
print("MSE scikit-learn: {}".format(mse2))

MSE manuell: 0.20786958588653123
MSE scikit-learn: 0.20786958588653123
```

Bewertung: Realisierung in scikit-Learn

Root Mean Squared Error
(RMSE)

```
rmse = np.sqrt(mean_squared_error(y_test, predictions))
print("RMSE: {}".format(rmse))
```

RMSE: 0.45592717169141306

Bestimmtheitsmaß R²

```
from sklearn.metrics import r2_score

r2 = r2_score(y_test, predictions)
print("R^2 Score: {}".format(r2))
```

R^2 Score: 0.8481914237247246

Bewertung: Realisierung in scikit-Learn

Vergleich Score auf Testmenge zu Score auf Trainingsmenge

```
test_pred = ridge.predict(X_test)
train_pred = ridge.predict(X_train)

mae_test = mean_absolute_error(y_test, test_pred)
mae_train = mean_absolute_error(y_train, train_pred)

print("MAE on test data: {}".format(mae_test))
print("MAE on training data: {}".format(mae_train))
```

```
MAE on test data: 0.3325739110428545
MAE on training data: 0.24143459296656444
```

Bewertung: Realisierung in scikit-Learn

Vergleich mit Dummy Regressor

```
predictions = ridge.predict(X_test)
mae_one = mean_absolute_error(y_test, predictions)
print("MAE Ridge Regression: {}".format(mae_one))

MAE Ridge Regression: 0.3325739110428545
```

```
from sklearn.dummy import DummyRegressor

dummy = DummyRegressor(strategy = 'mean')
dummy.fit(X_train, y_train)
dummy_pred = dummy.predict(X_test)

mae_two = mean_absolute_error(y_test, dummy_pred)

print("MAE Dummy Estimator Mean: {}".format(mae_two))

MAE Dummy Estimator Mean: 0.9468478260869562
```

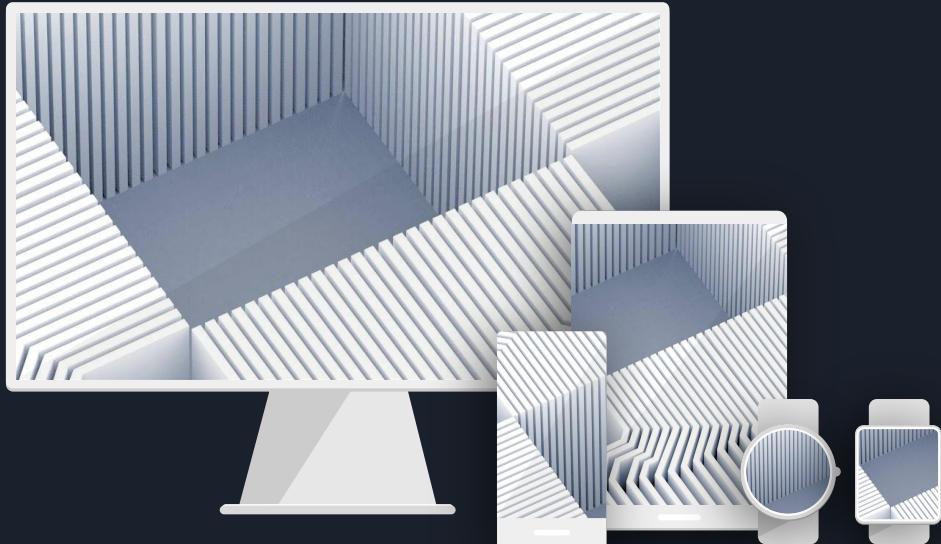


Übungsaufgabe Regression

- Importieren Sie den Sportwagendatensatz
- Erstellen Sie einen RandomForestRegressor (aus `sklearn.ensemble`)
 - Mit 50 Bäumen (`n_estimators`)
 - Mit einer maximalen Tiefe von 10 (`max_depth`)
 - Mit einem Seed von 1234 (`random_state`)
- Teilen Sie die Daten in eine Trainings- und eine Testmenge auf (`train_test_split` aus `sklearn.model_selection`)
 - Mit einem Testanteil von 10 % (`test_size`)
 - Mit einem Seed von 1234
- Lassen Sie Ihr Modell mit den Trainingsdaten lernen
- Lassen Sie Ihr Modell eine Vorhersage für die Testdaten treffen
- Berechnen Sie die folgenden Scores für Ihr Modell (aus `sklearn.metrics`)
 - Den "Mean Absolute Error" (MAE)
 - Den "Root Mean Squared Error" (RMSE)
 - Das Bestimmtheitsmaß R²

2.5 Cross-Validation

Daniel Koch

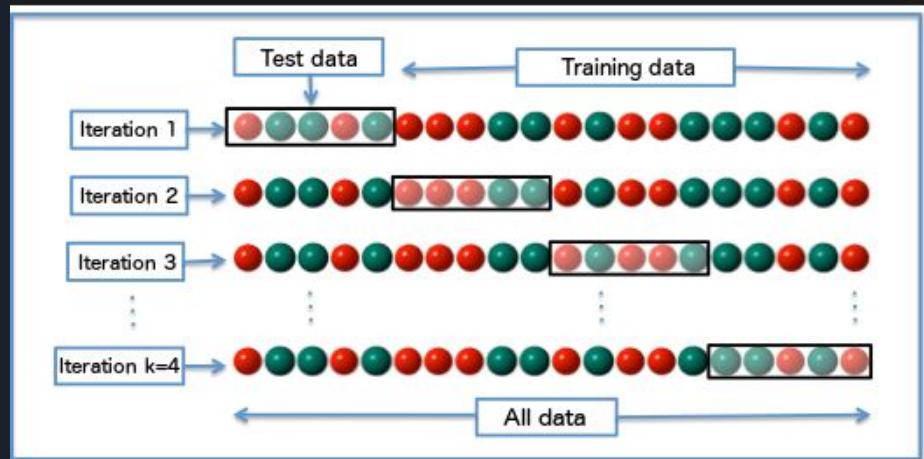


Cross-Validation

Methode, um Modelle anhand von Daten zu bewerten

Ablauf:

- Mischen der Daten
- Daten in k-Gruppen aufteilen
- Für jede Gruppe
 - Test / Training set
 - Fit auf Training set
 - Predict auf Test set
 - Gebe den Score zurück





Varianten von Cross-Validation

- Train/Test/Split: k=2
- LOOCV: k=Anzahl Datensätze

Erweiterungen:

- Repeated CV: Mehrmaliges Ausführen der CV
- Stratified CV: Berücksichtigung der Klassen der Daten



Cross-Validation

- `cross_val_score(clf, X, y, cv=4)`
=> [0.94324, 0.95492, 0.963294, 0.94232] → Bester Fall
=> [0.94324, 0.95492, 0.503294, 0.94232] → Fold 3 hat eine ungünstige Verteilung
=> [0.80324, 0.95492, 0.403294, 0.66232] → Modell oder Daten nicht konsistent
- `numpy.mean(cross_val_score(clf, X, y, cv=4))`
=> 0.95

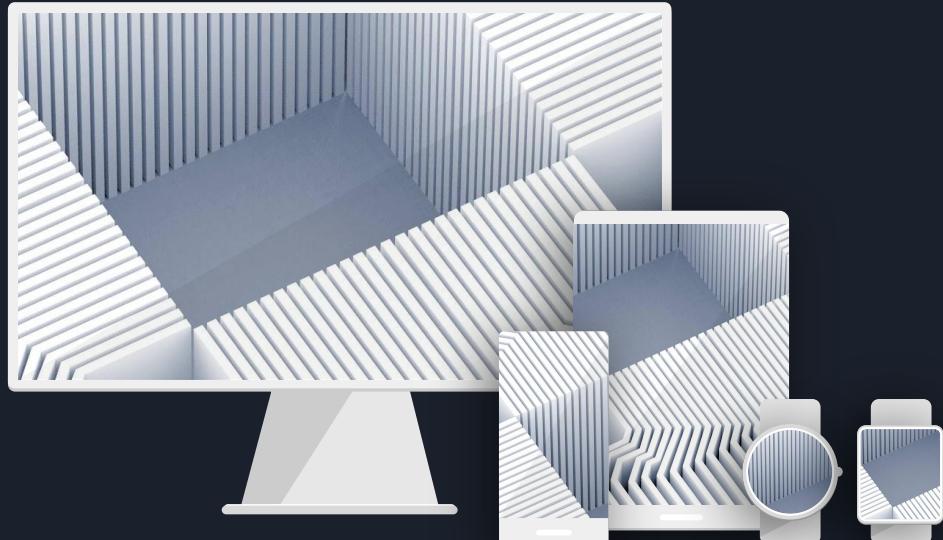


Gründe für Cross-Validation

- Nutzen aller Daten (vgl. `train_test_split`)
- Mehr Metriken
- Arbeiten mit abhängigen Daten
- Hyperparameteroptimierung

2.6 Hyperparameter- optimierung

Daniel Koch





Einstieg

Parameter: Werden trainiert / Hyperparameter: Setting des Models

Estimator benötigen optimierte Parameter, um bestmögliche Ergebnisse zu liefern, z.B:

```
svm.SVC(gamma=0.002)
```

Manuelles Optimieren

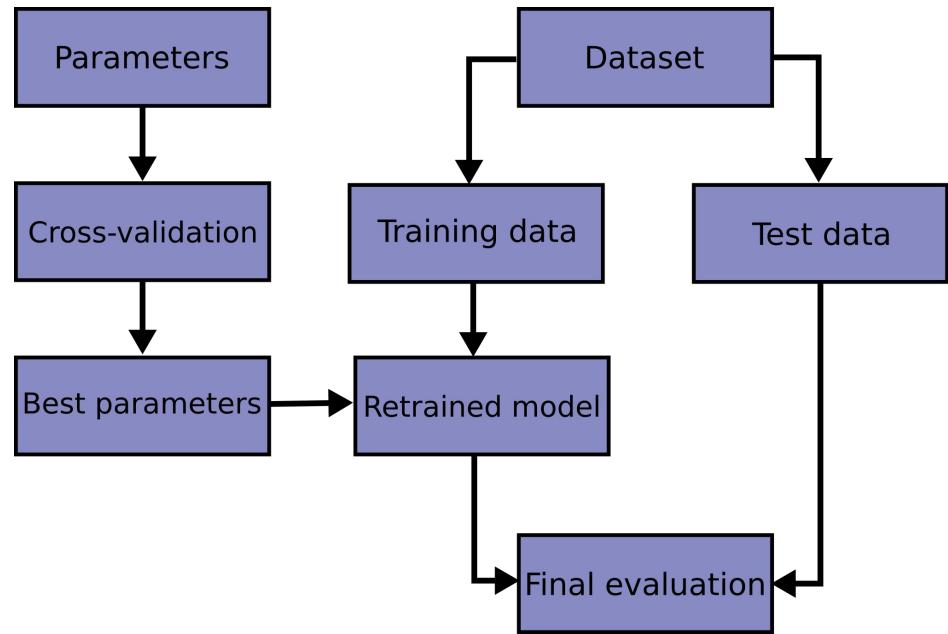
1. Ändern der Parameter
2. Auswertung der Scores mittels `cross_val_score(...)`

→ Sehr zeitaufwendig!

Daher: scikit-learn bietet eine Lösung!

Ablauf

Cross Validation Workflow:
Model Training





Hyperparameteroptimierung

- Bestandteile einer Suche
 1. Parameter Space
 2. Verfahren zur Suche
 3. Estimator
 4. Cross Validation-Schema
 5. Score-Funktion
- Ablauf einer Suche
 1. Generieren von Parametern nach Verfahren
 2. Prüfen, wie gut diese Parameter abschneiden
 3. Speichern der besten Scores / Parameter



Hyperparameteroptimierung

- 1. Parameter Space
 - Als Python Dictionary {}
 - Key: Zu optimierender Parameter
 - Value: Mögliche Optionen
 - Beispiel für Values:
 - Array mit definierten Werten: [True, False]
 - Array mit generieren Werten:
 - `np.linspace(start,stop,length)`
 - `np.logspace(start,stop,length)`
 - Zufallszahlen-Generator: `randrange(a, b+1)`, `uniform(a, b)`

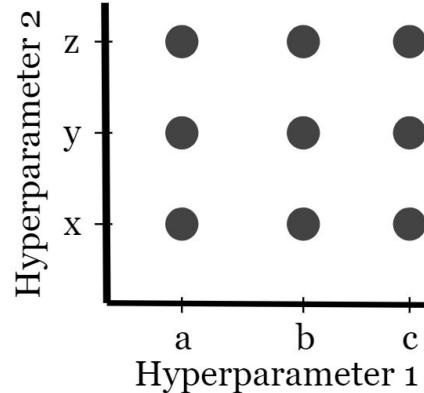
Hyperparameteroptimierung

- 2. Verfahren zur Suche:

Grid Search

Pseudocode

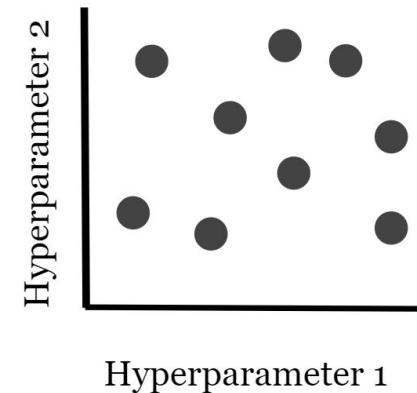
```
Hyperparameter_One = [a, b, c]
Hyperparameter_Two = [x, y, z]
```



Random Search

Pseudocode

```
Hyperparameter_One = random.num(range)
Hyperparameter_Two = random.num(range)
```





Hyperparameteroptimierung

- GridSearchCV
 - Vorteil: "exhaustive search"
 - Nachteil: Aufwendig:
 - 4 Parameter à 20 Werte => 160k Möglichkeiten
 - 5 Fold-CV => 800k Predictions
- RandomizedSearchCV
 - Vorteil: weniger aufwendig
 - Nachteil: Suche ist nicht "exhaustive"
- Fazit:
 - Kleine Datenmengen ⇒ GridSearchCV
 - Große Datenmengen, viele Parameter ⇒ RandomizedSearchCV



2.6 Hyperparameteroptimierung

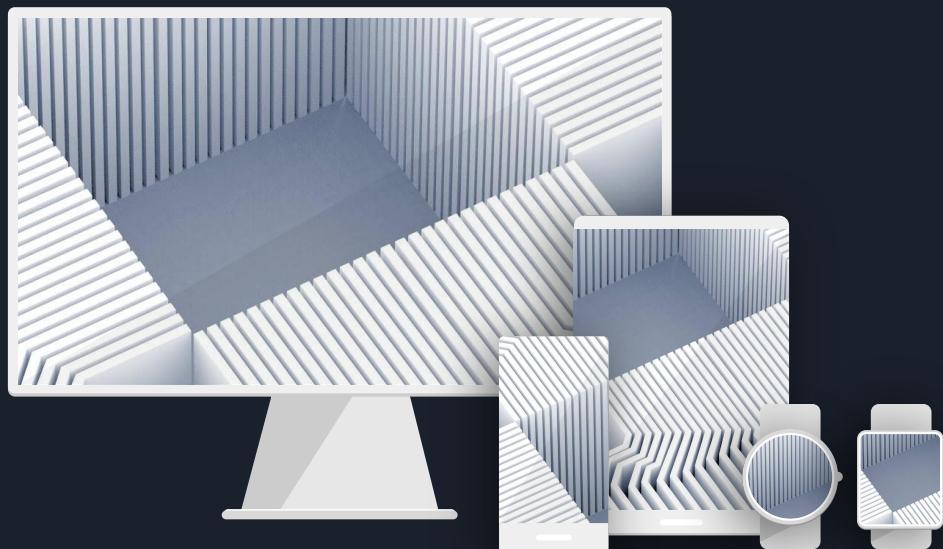
- 3. Estimator
 - Decision Tree Classifier
 - criterion
 - min_samples_leaf
 - max_depth
 - SVC
 - gamma
- 4. Cross Validation-Schema: Anzahl der Folds: cv=5
- 5. Score-Funktion
 - optional
 - Default: Score-Funktion des Estimators



Übungsaufgabe Hyperparameteroptimierung

- Nutzen Sie die gegebene Vorlage, um Ihr Programm zu schreiben. Dort sind auch relevante Links zur Dokumentation.
- Nutzen Sie den KNeighborsClassifier() zusammen mit RandomizedSearchCV, um die besten Hyperparameter zu finden. Nutzen Sie 5-Fold-Cross-Validation.
- Optimieren Sie die folgenden Parameter:
 - n_neighbors
 - leaf_size
 - p
- Geben Sie formatiert auf der Konsole die folgenden Ergebnisse aus:
 - Die besten gefundenen Parameter
 - Der beste gefundene Score

Fragerunde & Diskussion





Overview

Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl.

Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl. Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend.



Overview

Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl.

Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl. Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend.



Understanding the problems

01

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Class aptent taciti sociosqu ad litora et nec torquent per conubia nostra.

02

Amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Class aptent taciti sociosqu ad litora torquent per conubia nostra.

03

Consectetur adipiscing elit. Curabitur eleifend lorem a diam quis suscipit. Class aptent taciti sociosqu ad litora torquent ipsum per conubia nostra.



Project objective



Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl. Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend.



Target audience

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl. Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend.

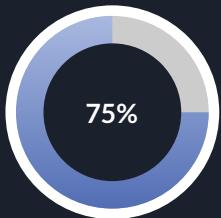




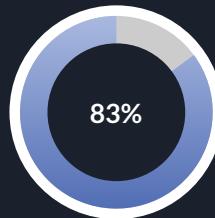
Persona 01

Wendy Writer

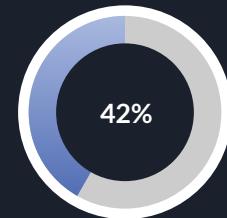
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod.



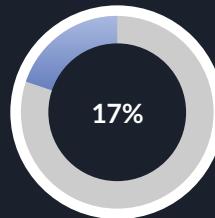
Lorem Ipsum



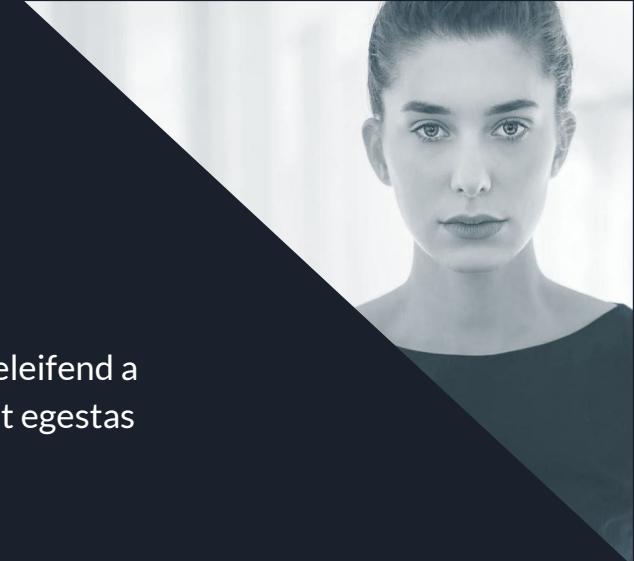
Lorem Ipsum



Lorem Ipsum



Lorem Ipsum

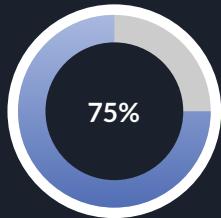




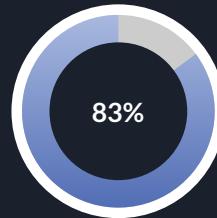
Persona 02

Berry Books

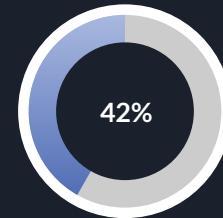
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod.



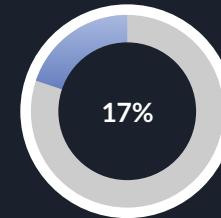
Lorem Ipsum



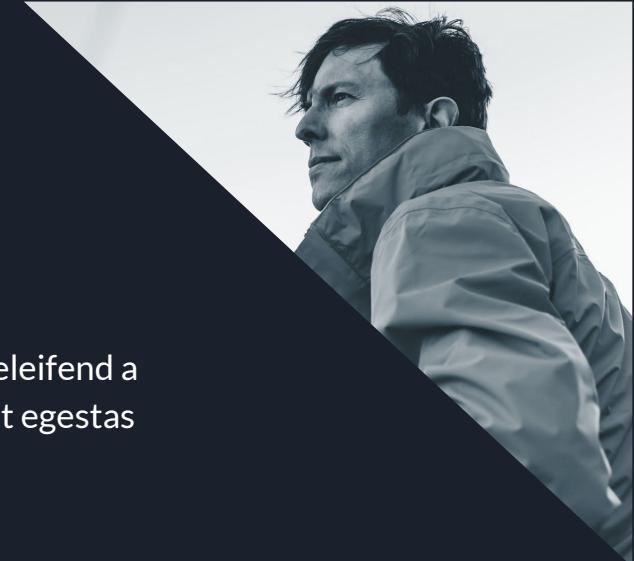
Lorem Ipsum



Lorem Ipsum



Lorem Ipsum



Market trends

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum. Maecenas molestie ultricies euismod. Morbi a rutrum nisl. Vestibulum laoreet enim id sem fermentum, sed aliquam arcu dictum. Donec ultrices diam sagittis nibh pellentesque eleifend.





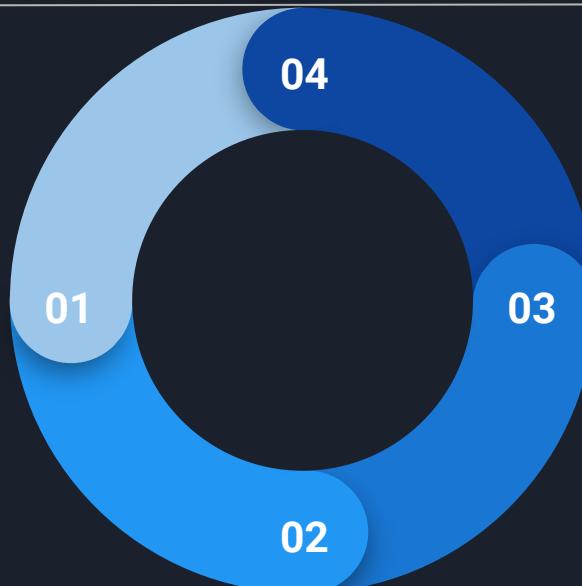
Cycle diagram

Prototype

Lorem ipsum dolor sit amet,
consectetur adipiscing.

Share

Lorem ipsum dolor sit amet,
consectetur adipiscing.



Refine

Lorem ipsum dolor sit amet,
consectetur adipiscing.

Get feedback

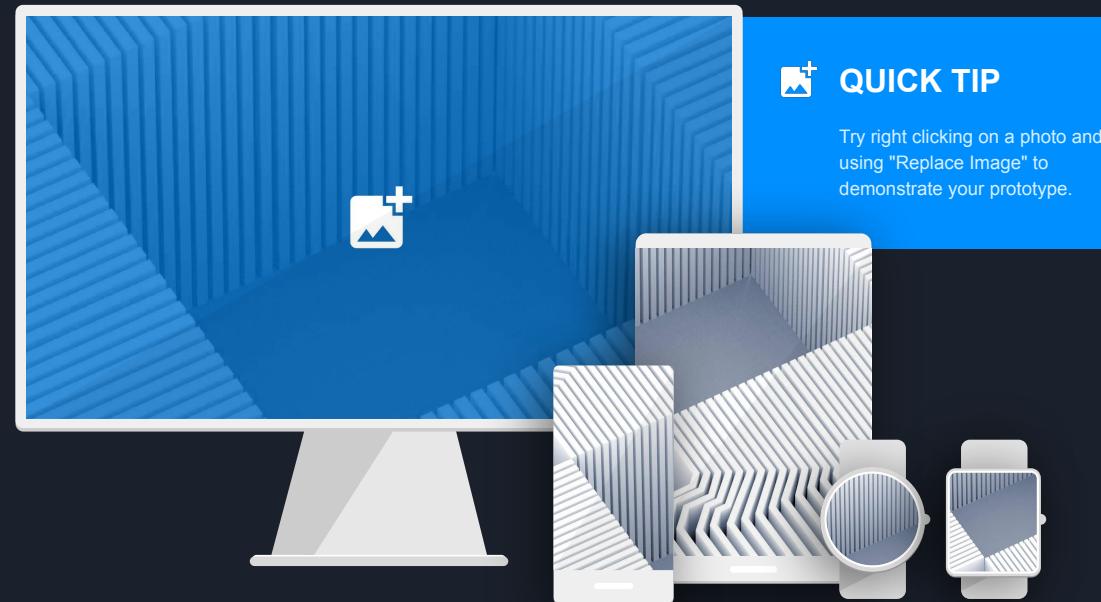
Lorem ipsum dolor sit amet,
consectetur adipiscing.



Introducing: Lorem ipsum

Showcase how your tools work across different devices

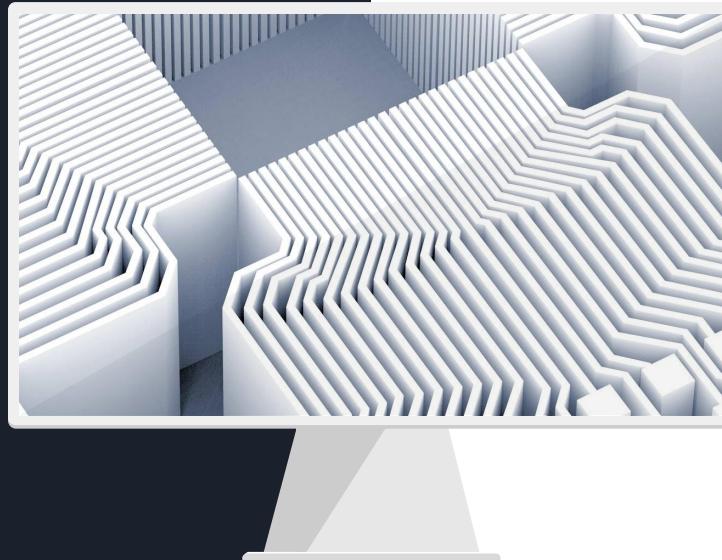
Etiam euismod, **tempor**, **adipiscing** elit. Curabitur eleifend a diam quis suscipit. Fusce venenatis nunc ut lectus convallis, sit amet egestas mi rutrum.





Spotlight on desktop

**Lorem ipsum
dolor sit
consectetur amet
adipiscing donec**



**Lorem ipsum dolor sit amet,
consectetur adipiscing elit.
Curabitur eleifend a diam quis
suscipit. Fusce venenatis nunc ut
lectus convallis, sit amet egestas mi
rutrum. Maecenas molestie
ultricies euismod.**



Spotlight on mobile

**Lorem ipsum dolor sit
consectetur amet
adipiscing donec**



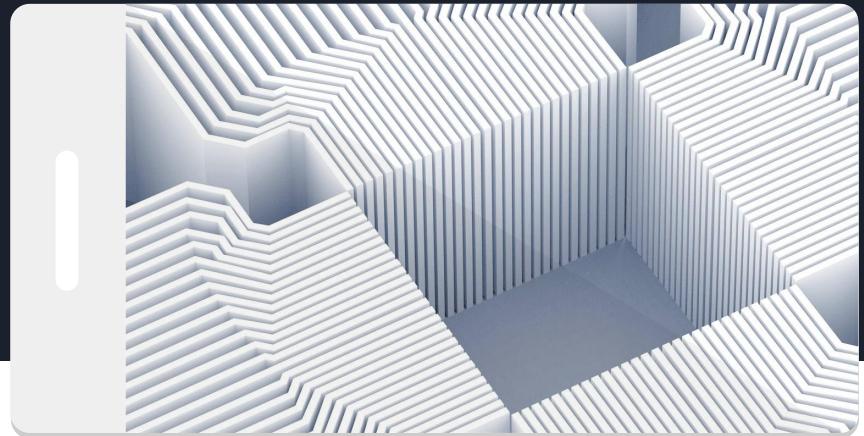
**Lorem ipsum dolor sit amet, consectetur
adipiscing elit. Curabitur eleifend a diam
quis suscipit. Fusce venenatis nunc ut lectus
convallis, sit amet egestas mi rutrum.
Maecenas molestie ultricies euismod.**



Spotlight on landscape view on mobile

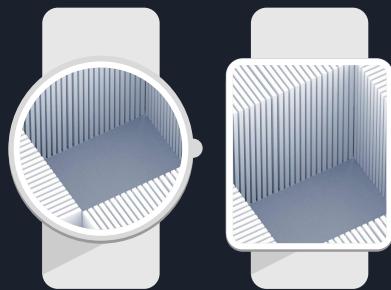
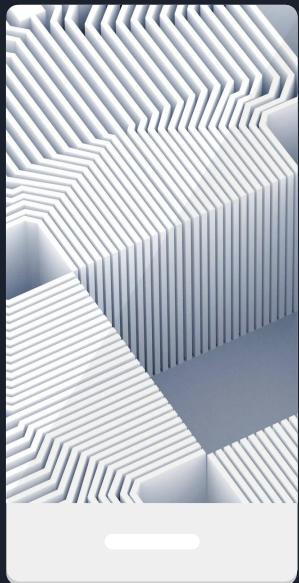
Lorem ipsum dolor sit
consectetur amet
adipiscing donec

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur eleifend a diam quis suscipit. Fusce venenatis
nunc ut lectus convallis, sit amet egestas mi rutrum.
Maecenas molestie ultricies euismod.





Spotlight on wearables



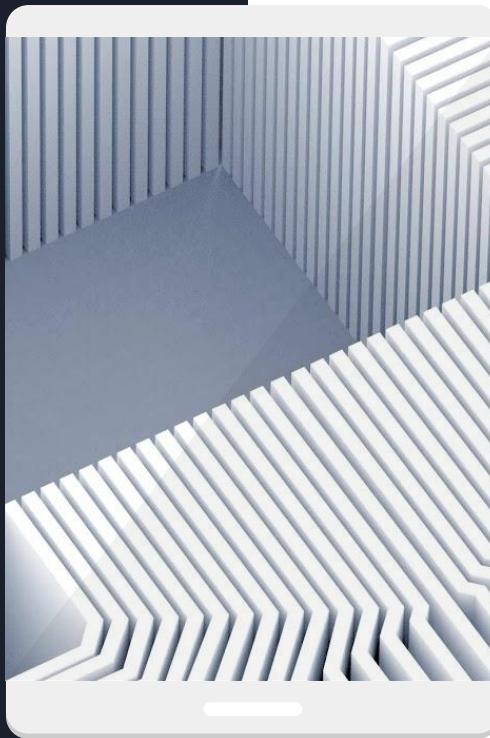
**Lorem ipsum dolor
sit consectetur amet
adipiscing donec**

**Lorem ipsum dolor sit amet, consectetur
adipiscing elit. Curabitur eleifend a diam quis
suscipit. Fusce venenatis nunc ut lectus
convallis, sit amet egestas mi rutrum.
Maecenas molestie ultricies euismod.**



Spotlight on tablet

**Lorem ipsum
dolor sit
consectetur amet
adipiscing donec**



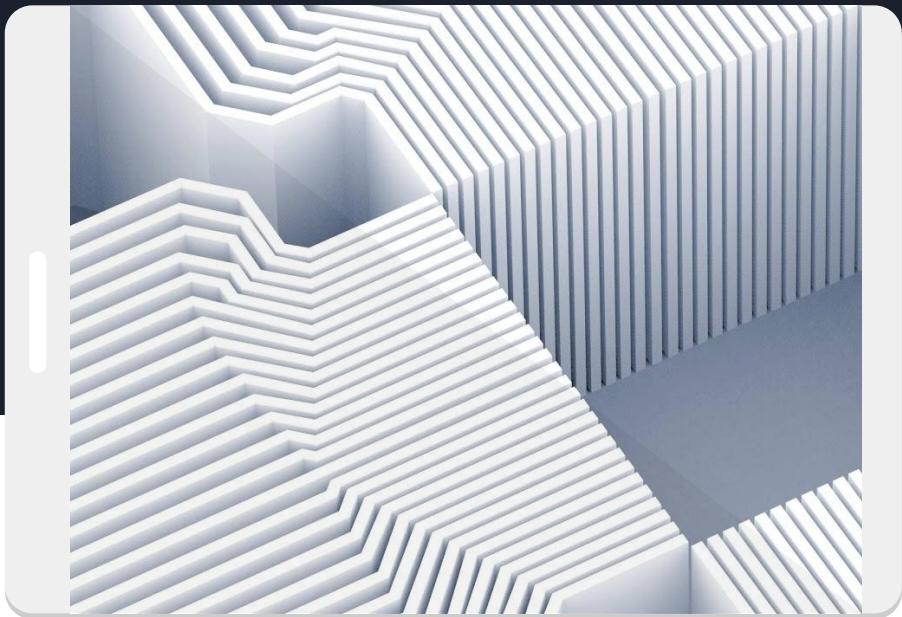
**Lorem ipsum dolor sit amet,
consectetur adipiscing elit.
Curabitur eleifend a diam quis
suscipit. Fusce venenatis nunc ut
lectus convallis, sit amet egestas mi
rutrum. Maecenas molestie
ultricies euismod.**



Spotlight on landscape view on tablet

**Lorem ipsum dolor sit
consectetur amet
adipiscing donec**

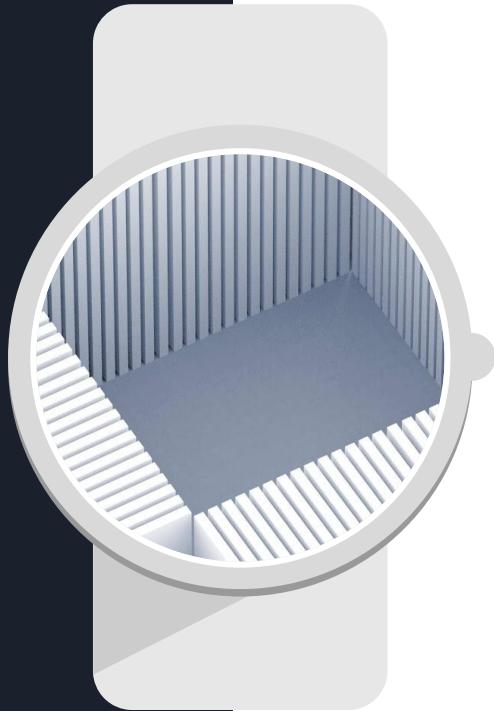
**Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur eleifend a diam quis suscipit. Fusce venenatis
nunc ut lectus convallis, sit amet egestas mi rutrum.
Maecenas molestie ultricies euismod.**





Spotlight on wearables

Lorem ipsum
dolor sit
consectetur amet
adipiscing donec



Lorem ipsum dolor sit amet,
consectetur adipiscing elit.
Curabitur eleifend a diam quis
suscipit. Fusce venenatis nunc ut
lectus convallis, sit amet egestas mi
rutrum. Maecenas molestie
ultricies euismod.



Spotlight on wearables

Lorem ipsum
dolor sit
consectetur amet
adipiscing donec



Lorem ipsum dolor sit amet,
consectetur adipiscing elit.
Curabitur eleifend a diam quis
suscipit. Fusce venenatis nunc ut
lectus convallis, sit amet egestas mi
rutrum. Maecenas molestie
ultricies euismod.



Project timeline

