





# EduPredictors

Elisabeth Barar und Johanna Unruh



# Ziel

- Finden eines geeigneten Datensatzes
- Trainieren unterschiedlicher Klassifikationsmodelle
- Herausfinden, welches Modell am Besten ist und warum
- Bau einer beispielhaften Applikation



# Die Modelle

1

**kNN**

2

**Decision  
Trees**

3

**Random  
Forest**

4

**Naive  
Bayes**



# Die Modelle

1

**kNN**

2

**Decision  
Trees**

3

**Random  
Forest**

4

**Naive  
Bayes**

Alle Modelle für Klassifizierungsprobleme geeignet

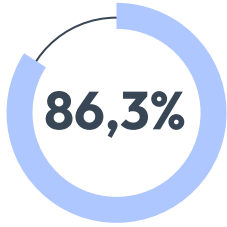
# Datenvorbereitung

1. Entfernen der Zeilen, die NaN Werte haben
2. Korrelationsanalyse durchführen
3. Entfernen der Spalten, die eine geringe Korrelation zu der Zielspalte haben
4. Entfernen der Zeilen, die bei der Zielspalte „Enrolled“ stehen haben

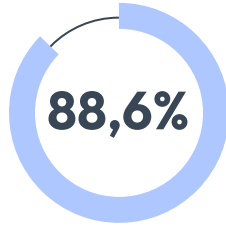
# Kriterien für ein gutes Modell

- ✓ Accuracy Score von über 80%
- ✓ Guter Recall Score
- ✓ AUC nahe 1

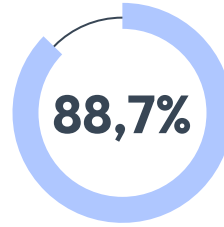
# Vergleich der Modelle



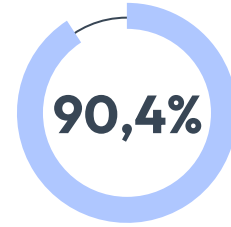
**Naïve Bayes**



**kNN**



**Decision  
Tree**



**Random  
Forest**



# Random Forest Evaluation

Kriterium	Ergebnis
Accuracy Score von über 80%	

# Random Forest Evaluation

Kriterium	Ergebnis
Accuracy Score von über 80%	90,4%

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“		

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“	83%	

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“	83%	✓

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“	83%	✓
AUC Score nahe 1		

# Random Forest Evaluation

Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“	83%	✓
AUC Score nahe 1	0,947	



# Random Forest Evaluation

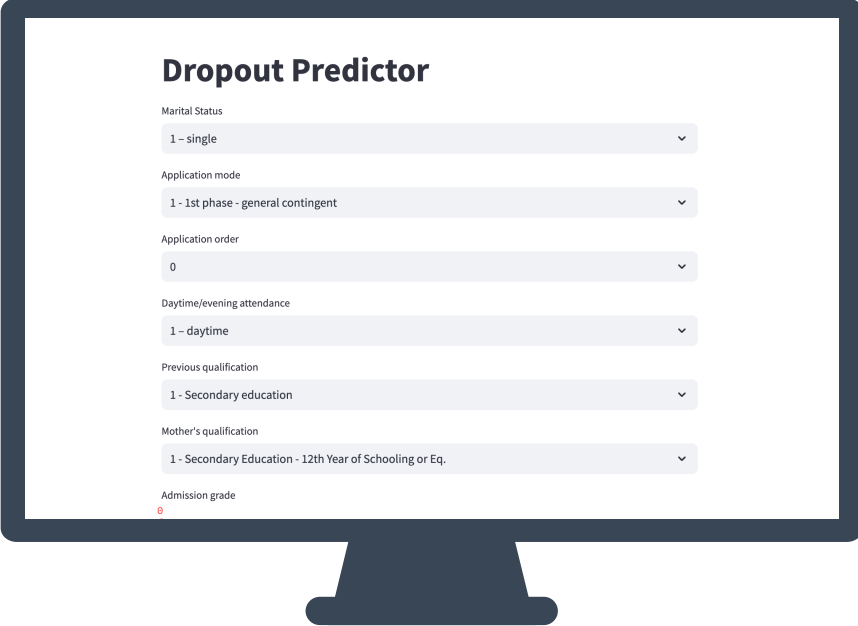
Kriterium	Ergebnis	
Accuracy Score von über 80%	90,4%	✓
Guter Recall Score für „Dropout“	83%	✓
AUC Score nahe 1	0,947	✓

# Warum hat Random Forest eine gute Performance?



1. Ensemble von Decision Trees erlaubt eine Fehlerreduktion
2. Geringeres Risiko zum Overfitten was zu einer niedrigeren Varianz führt
3. Geringe Varianz führt zu einer besseren Generalisierung daher bessere Performance auf einem unbekannten Datensatz

# Demo



**Dropout Predictor**

Marital Status  
1 - single

Application mode  
1 - 1st phase - general contingent

Application order  
0

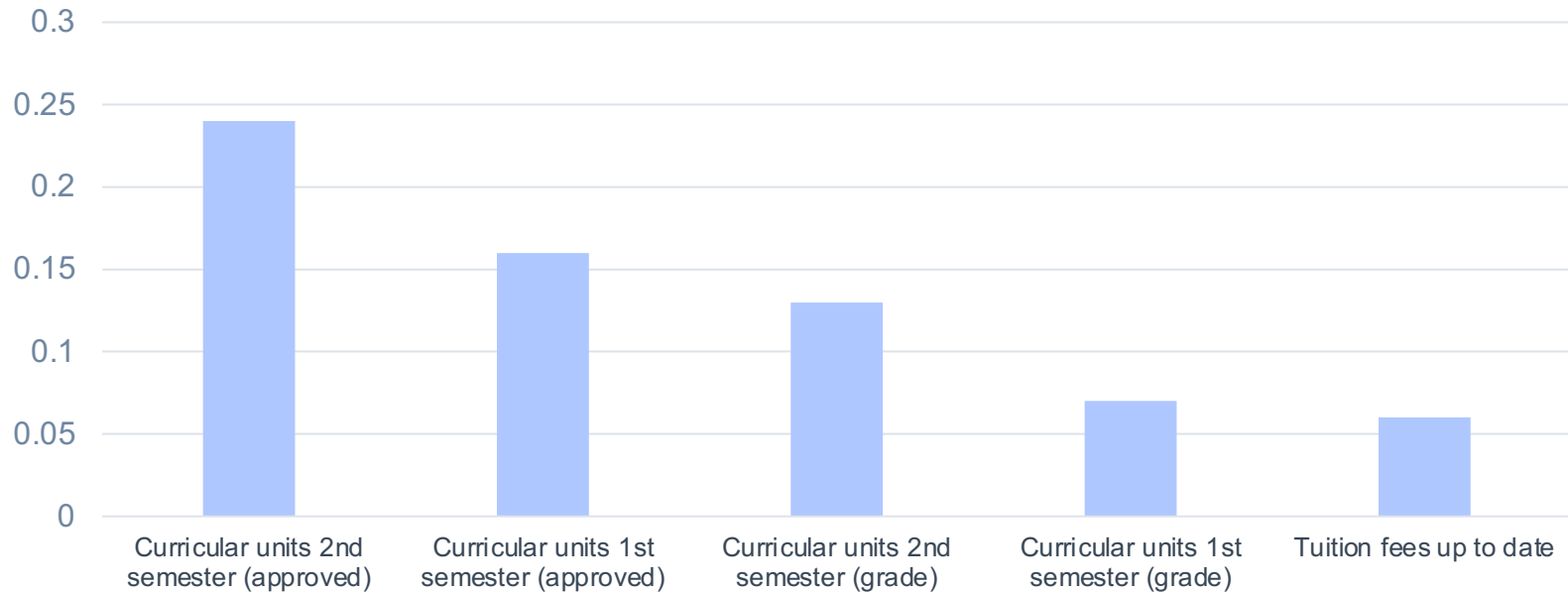
Daytime/evening attendance  
1 - daytime

Previous qualification  
1 - Secondary education

Mother's qualification  
1 - Secondary Education - 12th Year of Schooling or Eq.

Admission grade  
6

# Die wichtigsten Features für die Entscheidungsfindung



# Fazit und Ausblick

- Noch stärkere Konzentration auf wichtige Features
- Datensatz hätte größer sein können, um noch ein genaueres Modell zu bekommen
- Leicht ungleichmäßige Verteilung von „Dropout“ und „Graduate“
- Zurzeit auf das amerikanische Notensystem angepasst, müsste auf das deutsche System übertragen werden

**Danke für Eure  
Aufmerksamkeit!**



# Literatur

Die Präsentation bezieht sich auf unseren Bericht