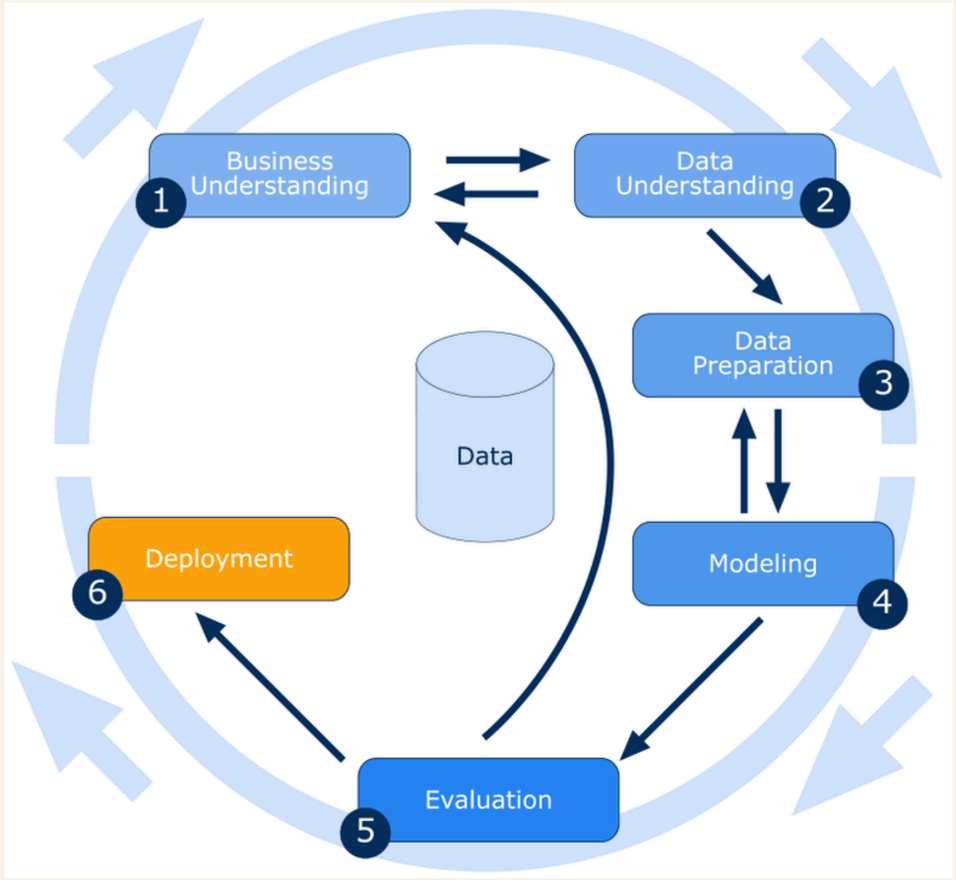
ML-Klassifikation von Spam-Buchungsanfragen

auf einem Immobilienportal

Irina Ukhanova

IKT SS2024, Semesterprojekt



CRISP DM

01 - Business Understanding

02 - Data Understanding

03 - Data Preparation

04 - Modeling

05 - Evaluation

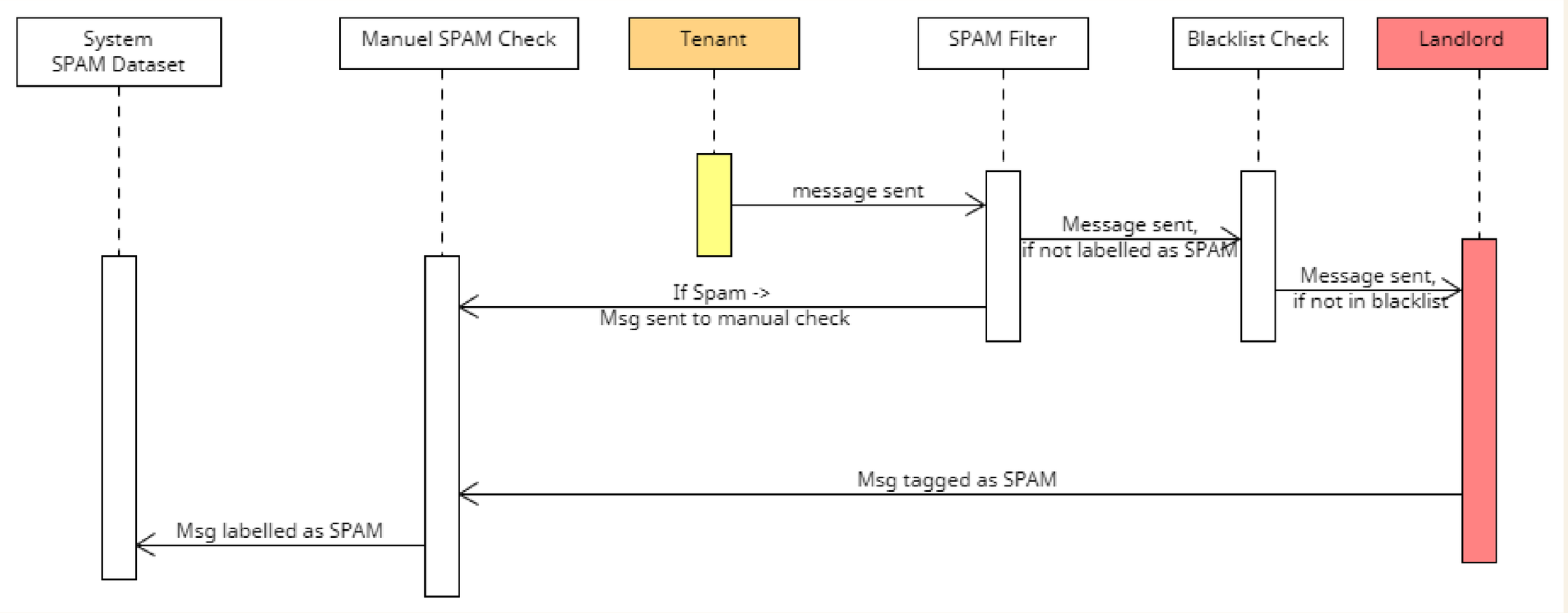
01 - Business Understanding : Zielsetzung und Anforderungen

- Problemdefinierung
  - Das Buchungsanfragenformular auf der Plattform für die kurzfristige Vermietung von Unterkünften wird leider für verschiedene Arten von Spam missbraucht.

- Ziel
  - Die Sicherheit von Buchungsanfragen erhöhen. Dafür wollen wir anfällige und gefährliche Anfragen automatisch durch ML-Techniken erkennen und vom System blockieren.

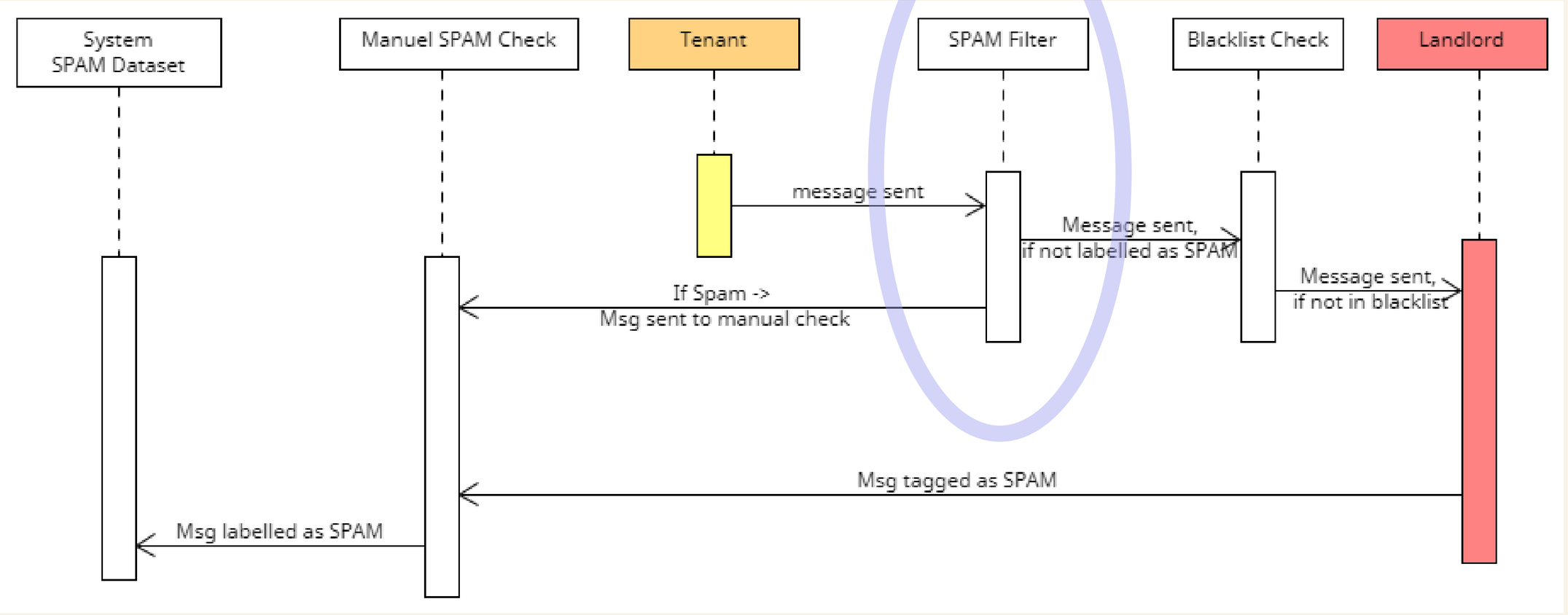
01 - Business Understanding : Zielsetzung und Anforderungen

- Status Quo Prozesse



01 - Business Understanding : Zielsetzung und Anforderungen

- Status Quo Prozesse

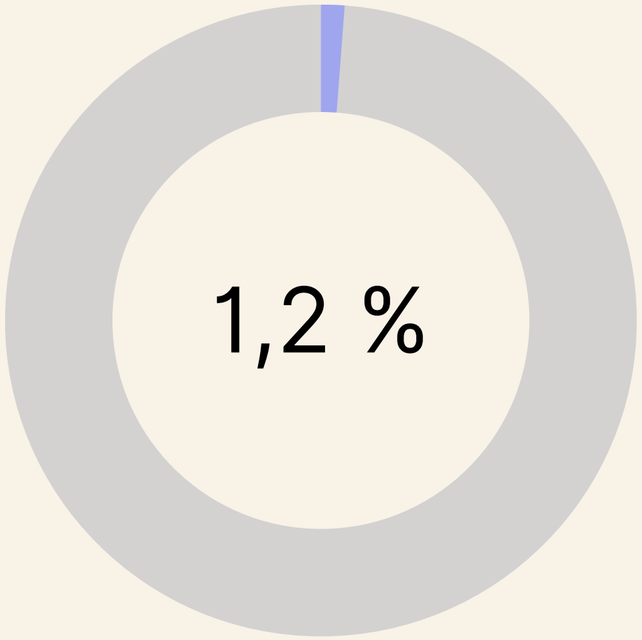


02, 03 - Data Understanding&Preparation

Data Understanding: Daten sammeln und analysieren

- Die Daten stammen aus der historischen Unternehmensdatenbank.
- Die relevantesten Informationen für Problemlösung - Nachricht
- Der Datensatz ist mit 0 für Nicht-Spam und mit 1 für Spam gekennzeichnet.

	Datenanzahl	%
Nicht Spam	5416	98,8%
Spam	68	1,2%



Data Preparation: Daten aufbereiten und bereinigen

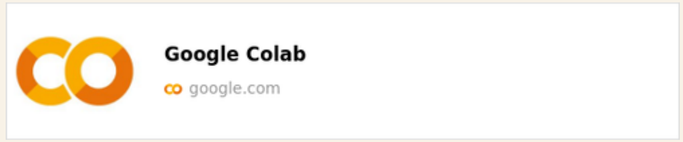
- Dubletten und nicht benötigte Daten gelöscht

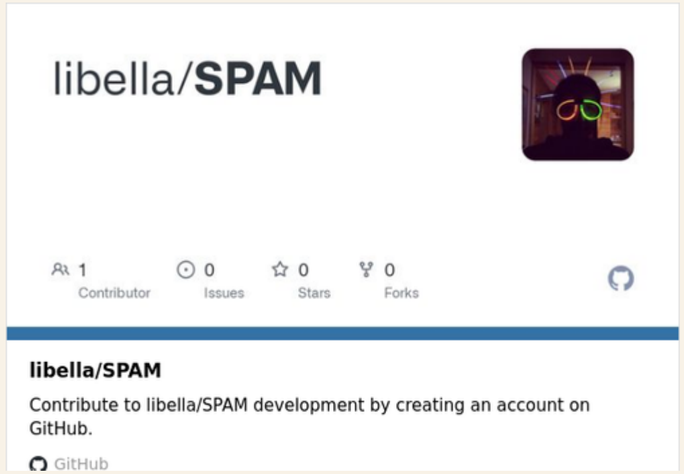
04 Modeling: Modelle auswählen und anwenden.

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

- Datensplit (Trainings- und Testdatensatz):
  - Sklearn.model_selection -> train_test_split
- Auswahl des Modells:
  - sklearn.linear_model -> Logistische Regression

Trainingssetgröße:4387
Testsetgröße:1097

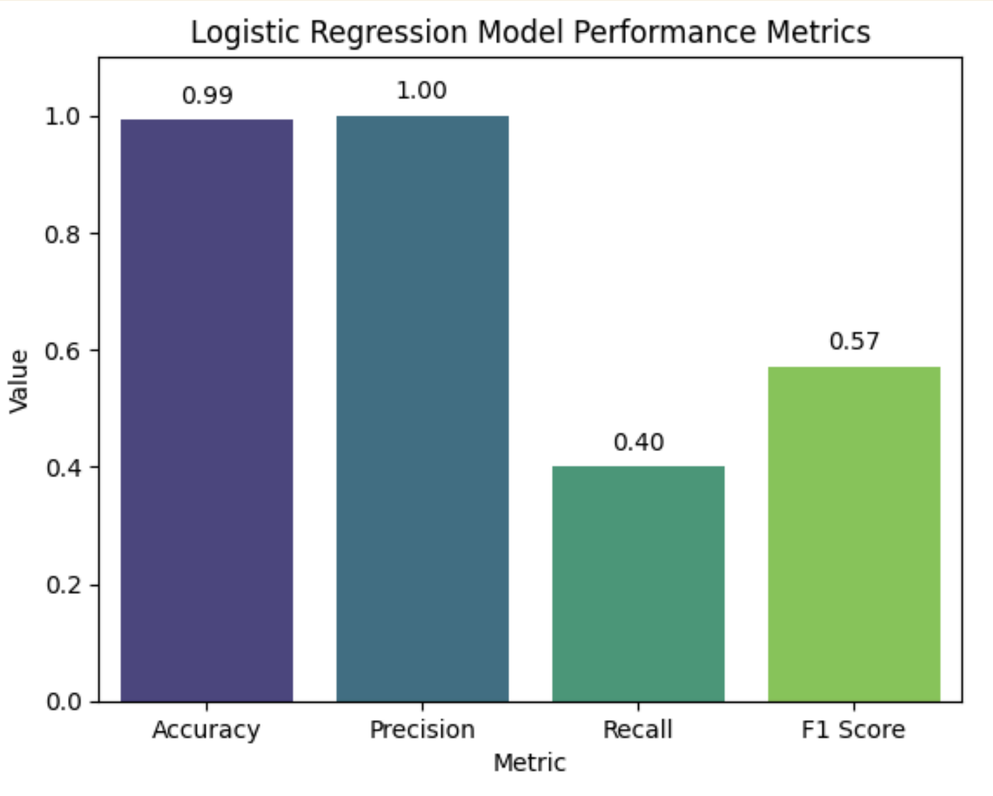


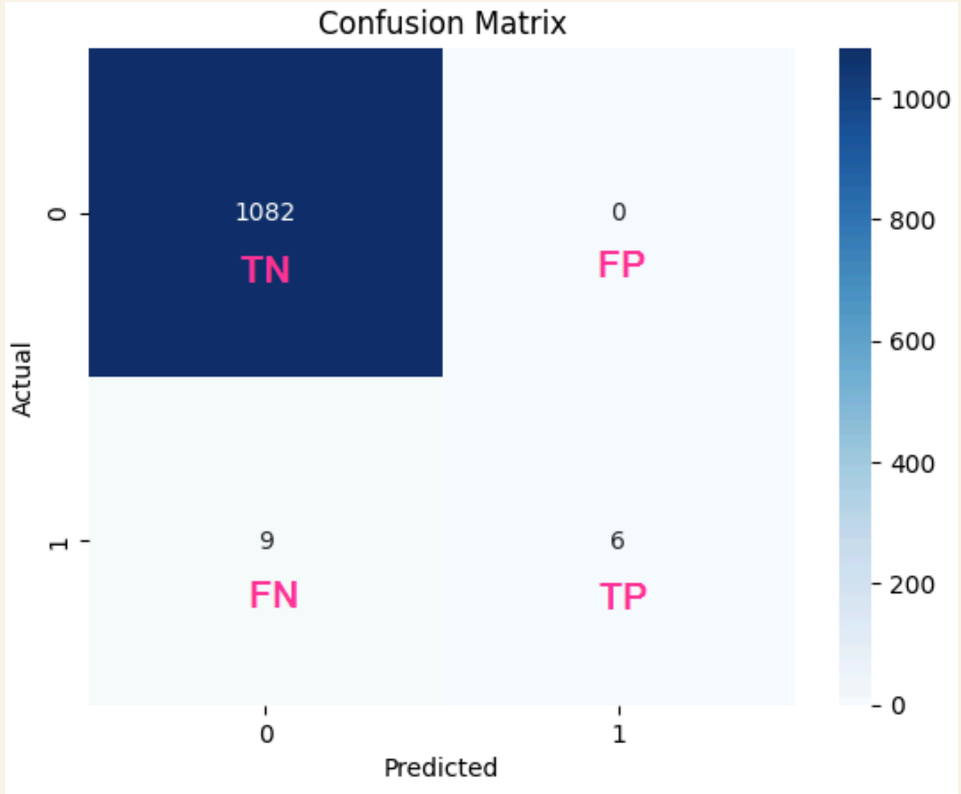


05 Evaluation: Modellleistung bewerten, 1.Iteration

Erfolgsmetriken:

- Ganauigkeit
- Prezision
- Recall
- F1-Score

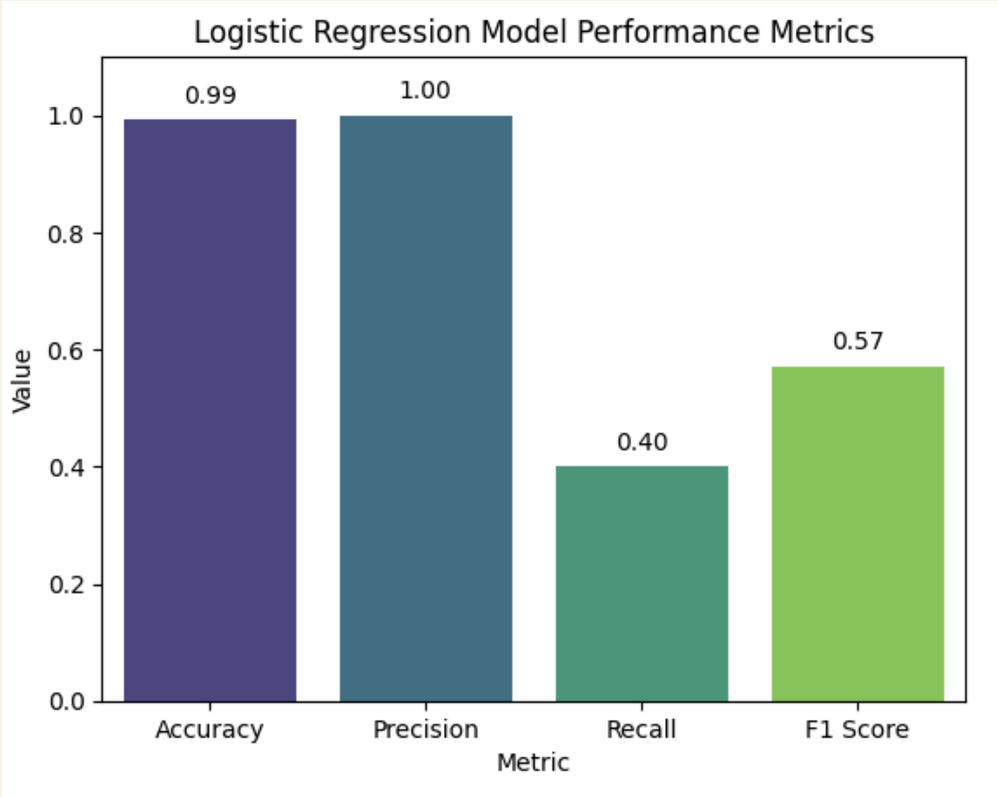




05 Evaluation: Modellleistung bewerten

Erkenntnisse zur Verbesserung des Modells:

- **Genauigkeit/Accuracy = 0.99 ,** sehr hoch, kann aber irreführend bei ungewogenen Daten sein
- **Precision = 1,** ideal
- **Recall = 0.4,** relativ niedrig und weist auf viele falsch negative Vorhersagen hin.
- **F1-Score = 0.57,** deutet darauf hin, dass das Modell in Bezug auf die True Positives und False Negatives verbessert werden muss.



Ausblick

- Krossvalidation
- Balance der Klassen des Datensatzes überprüfen (Über-sampling)