Petar Petrov
25.05.2022

Zertifizierungsprüfung „Data Scientist“

Vorhersage der Überfallwahrscheinlichkeit in LA auf Basis von LAPD Police Reports

Datenanalyse als Hilfe für Touristen

Der Datensatz

- Quelle: LOS ANGELES OPEN DATA https://data.lacity.org/
- Crime Data from 2010 to 2019 for Los Angeles
- Über 2 Mio Instanzen als Crime Reports
- 28 Features
  - Angaben zum Zeitpunkt des Verbrechens (Date, Time)
  - Angaben zum Area des Verbrechens (Central, Hollywood, etc.)
  - Angaben zum Art des Verbrechens (Rape, Robbery, Assault etc.)
  - Angaben zum Opfer (Age, Sex, Descent)
  - Angaben zum genauen Ort des Verbrechens (Street, Store, Parking, etc.)
  - Usw.

DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NA...	Rpt Dist ...
001307355	2010 Feb 20 ...	2010 Feb 20 ...	1350	13	Newton	1385
011401303	2010 Sep 13 ...	2010 Sep 12 ...	0045	14	Pacific	1485
070309629	2010 Aug 09 ...	2010 Aug 09 ...	1515	13	Newton	1324
090631215	2010 Jan 05 1...	2010 Jan 05 1...	0150	06	Hollywood	0646
100100501	2010 Jan 03 1...	2010 Jan 02 1...	2100	01	Central	0176
100100506	2010 Jan 05 1...	2010 Jan 04 1...	1650	01	Central	0162
100100508	2010 Jan 08 1...	2010 Jan 07 1...	2005	01	Central	0182
100100509	2010 Jan 09 1...	2010 Jan 08 1...	2100	01	Central	0157
100100510	2010 Jan 09 1...	2010 Jan 09 1...	0230	01	Central	0171
100100511	2010 Jan 09 1...	2010 Jan 06 1...	2100	01	Central	0132

Aufgabenstellung

- Identifikation von Opfergruppen anhand von Parametern wie Geschlecht, Alter und Abstammung
- Filterung irrelevanter Instanzen → Relevanz für Touristen
  - Entfernen von Einträgen ohne Opfer
  - Entfernen irrelevanter Verbrechen (Child Abandonment, Conspiracy, Violation of Court Order. etc.)
  - Entfernen irrelevanter Locations (Fire Station, School, Office, etc.)
- Gruppierung der Features in signifikanten Hauptgruppen, um die Varianz zu reduzieren (z.B. Alle Waffentypen in bewaffnet/unbewaffnet unterteilen)
- Visualisierung der Daten
- Trainieren unterschiedlicher Klassifikationsmodelle und die Bewertung deren Akkuranz.
- Optimierung von Hyperparametern
- Auswahl des optimalen Modells für die jeweilige Prädiktionsaufgabe
- Test der Prädiktion mit den Daten von einem Menschen

Data Preprocessing

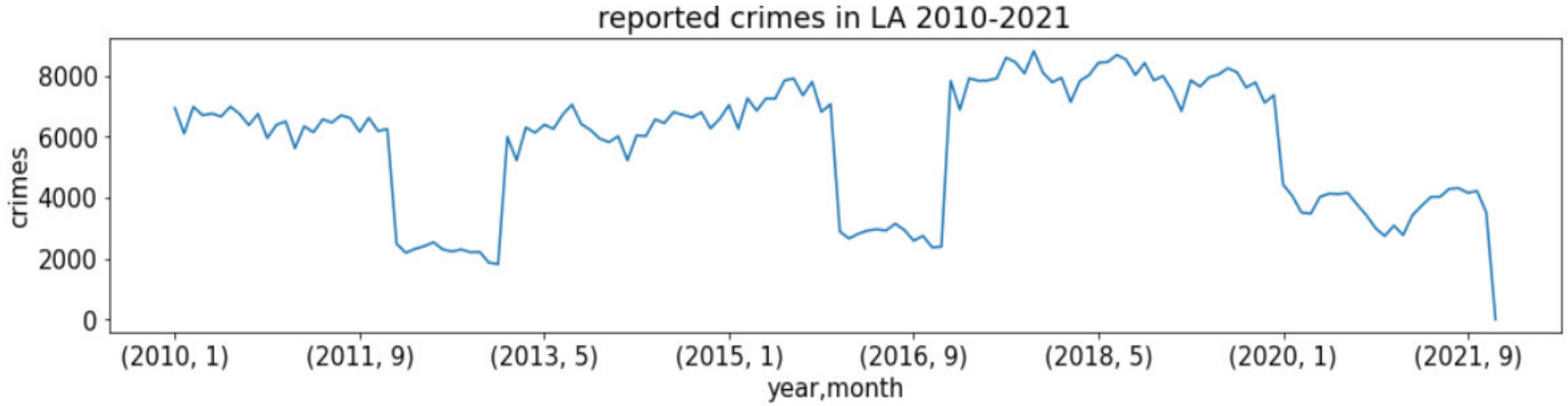
- Zusammenfassung von Features in signifikanten Hauptgruppen
- Die Vielzahl der Elemente pro Feature zerstreut die Daten und führt zu einer zu hoher Varianz → Gruppierung ist notwendig
  - LA_crime.Vict_Age: Opfers Alter in 6 Gruppen
    (<18, 19-30, 31-40, 41-50, 51-60, >60)
  - LA_crime.Vict_Descent: 19 Herkünfte wurden in 5 Gruppen
    (white, black, asian, latin, others)
  - LA_crime.TIME_OCC: Uhrzeit des Verbrechens in 2 Gruppen
    (Tag und Nacht)
  - LA_crime.AREA: 21 Areas in LA wurden in 4 sog. Communities
    (Valley, West, Central, South).
  - LA_crime.Crm_Cd_Desc: 103 Verbrechenstypen in 4 Gruppen
    (theft, assault, vandalism, other)
  - LA_crime.Premis_Desc: 140 Räumlichkeiten in 5 Gruppen
    (outdoors, vehicle_related, publ_transportation, store, others)

Überblick der Daten

Verlauf der Verbrechenberichte über dem gesamten Zeitraum

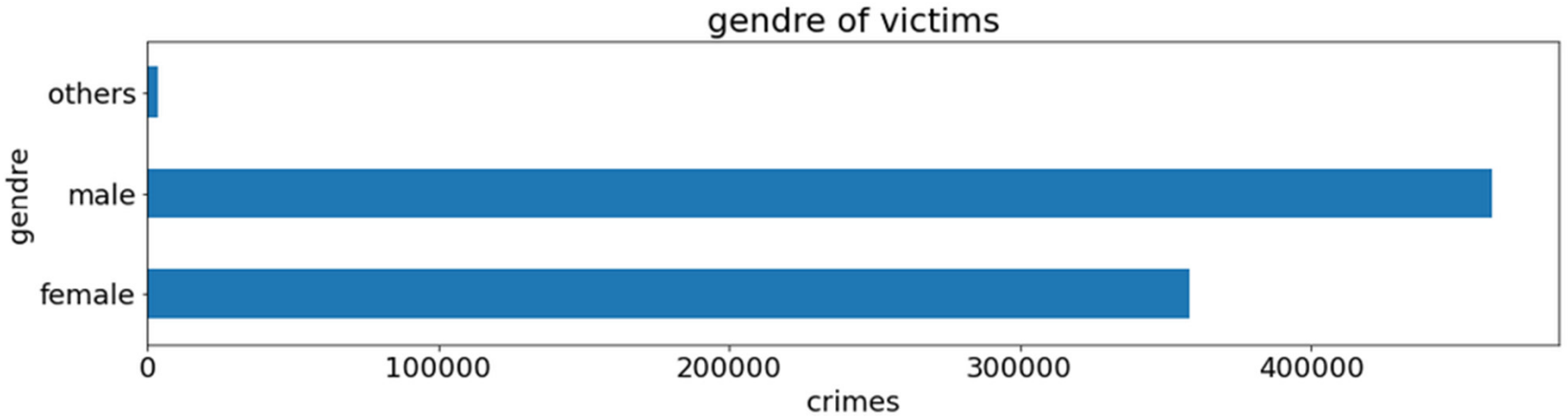
- Möglich fehlende Daten um die Jahre 2012 und 2016 → kein Hindernis für eine Klassifikationsaufgabe
- Offiziell ist der Datensatz für die Zeitspanne von 2009 bis 2019 gültig

  → Unstetiger Verlauf ab 2020 aufgrund Datenunvollständigkeit

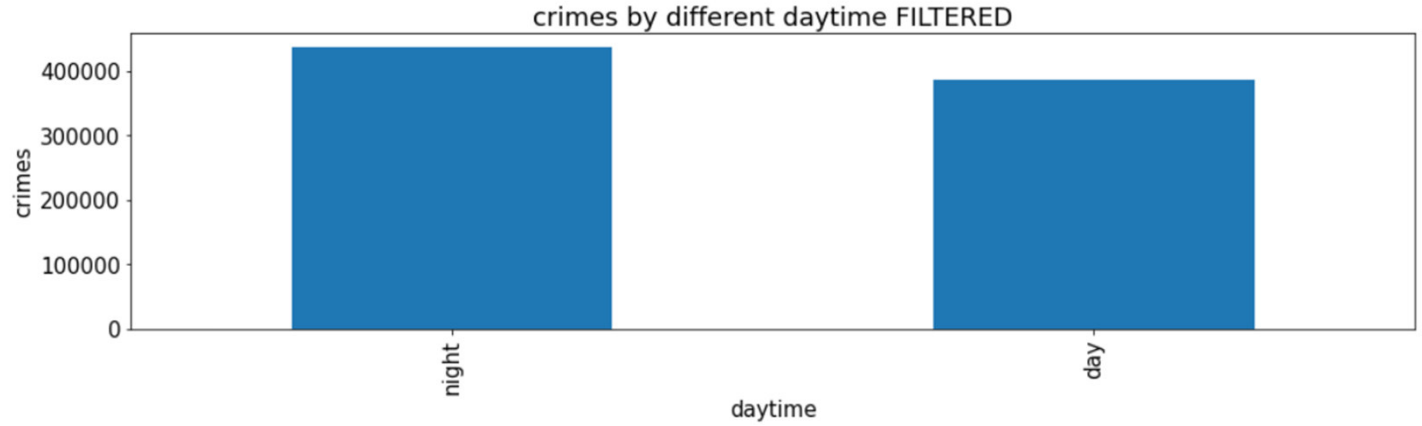


Überblick der Daten

Verteilung der Daten nach Geschlecht nach der Gruppierung

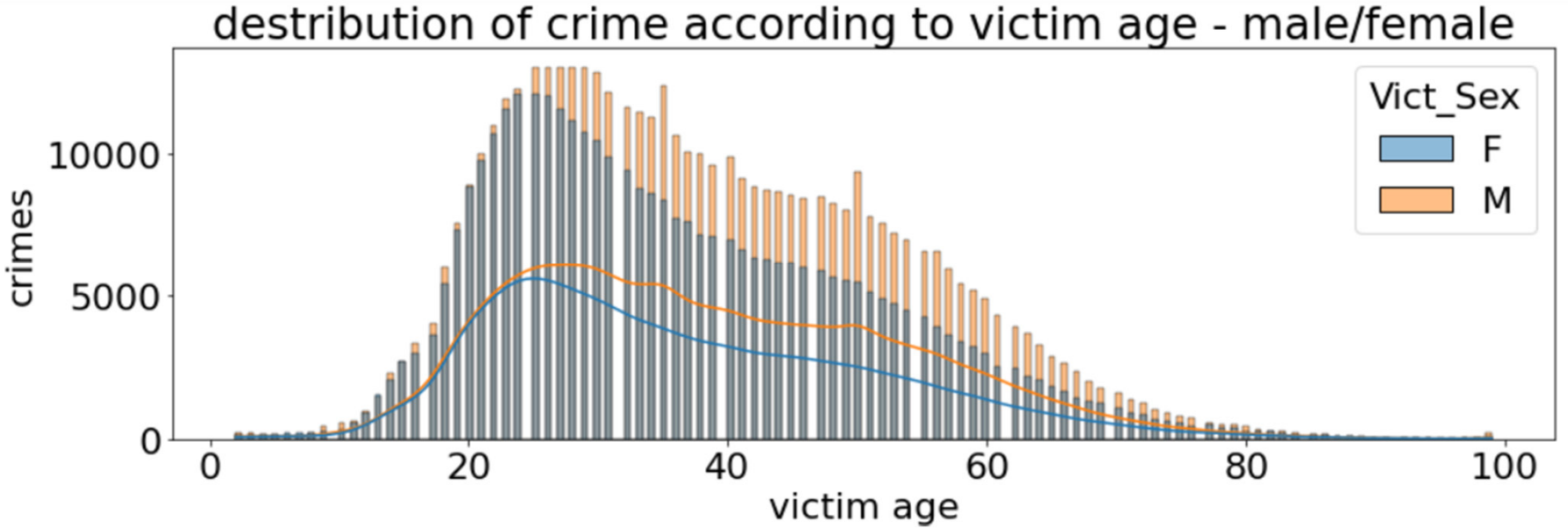


Verteilung der Daten nach Tag/Nacht



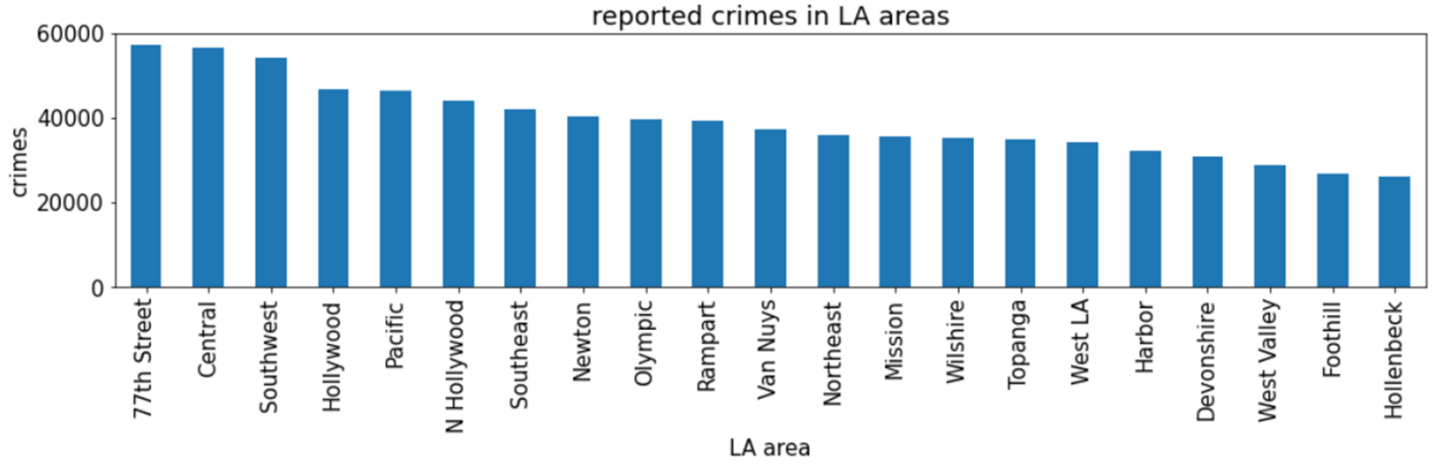
Überblick der Daten

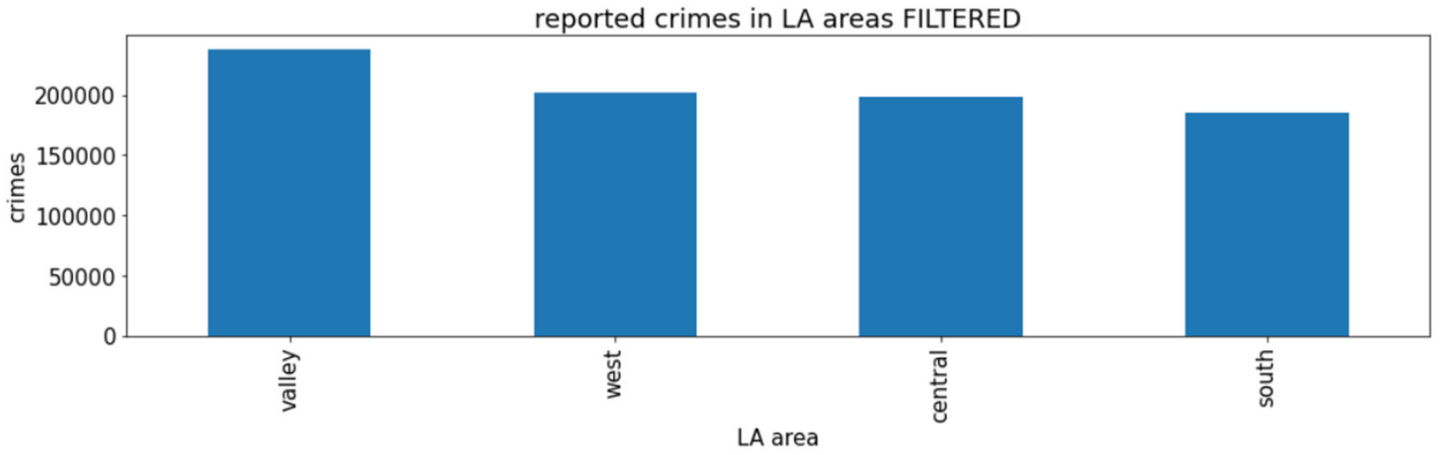
Histogramm der Daten: Verteilung über Alter geteilt in male/female



Überblick der Daten

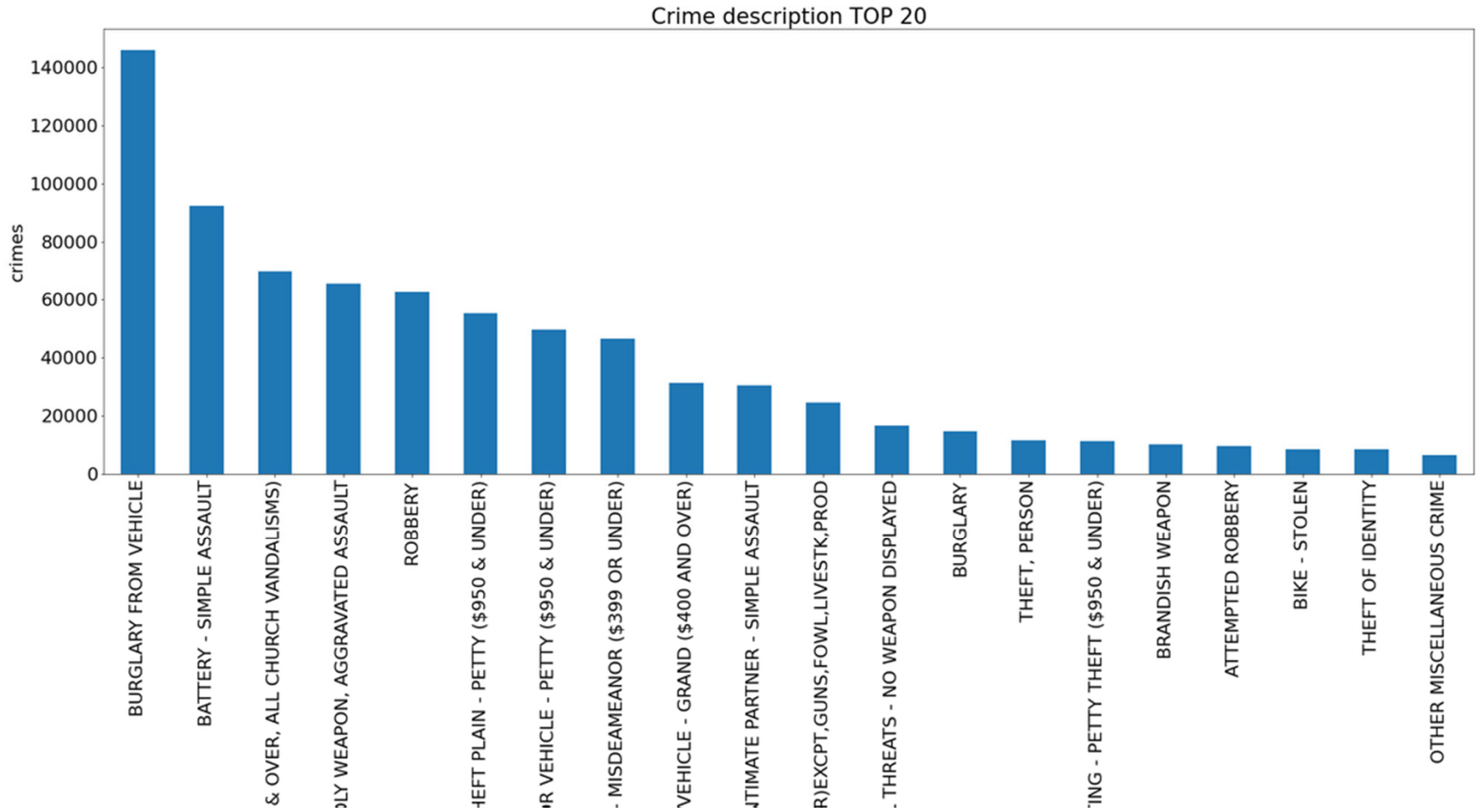
Verteilung der Daten AREA vor und nach der Gruppierung





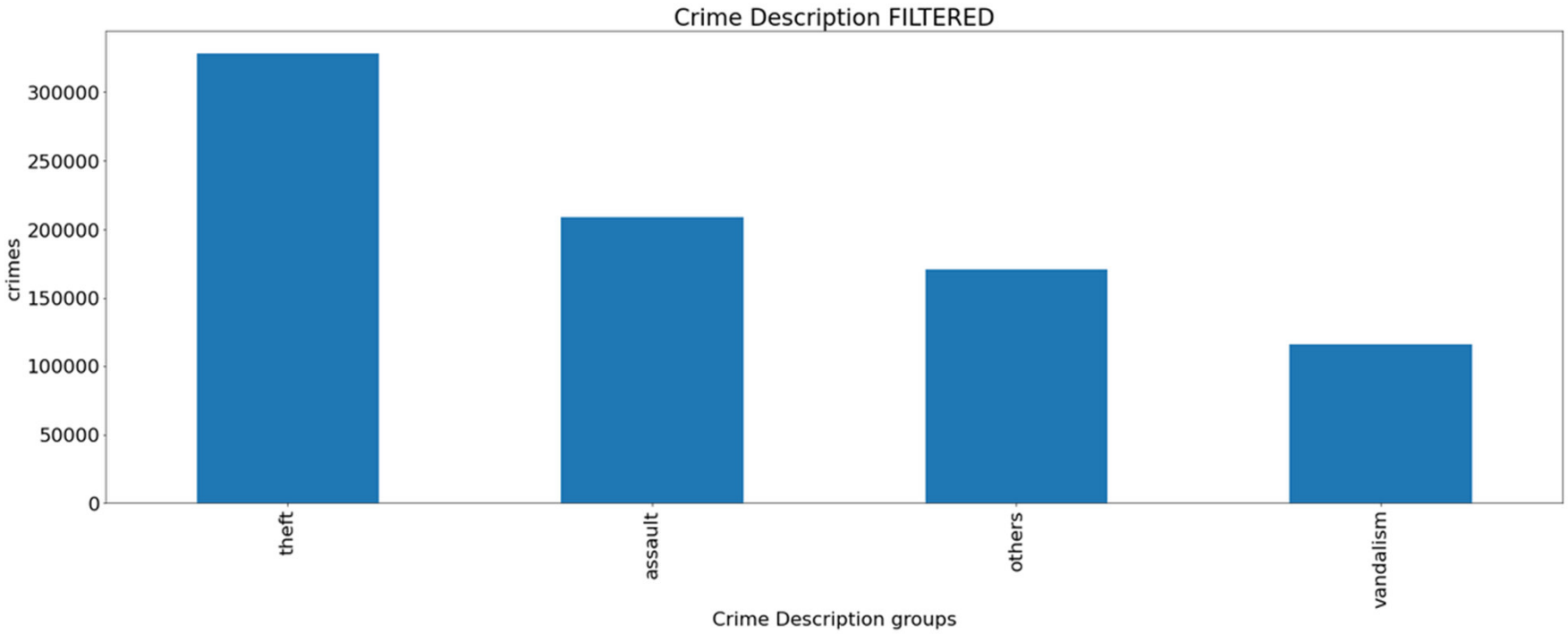
Überblick der Daten

Verteilung der Daten Crime Description vor der Gruppierung



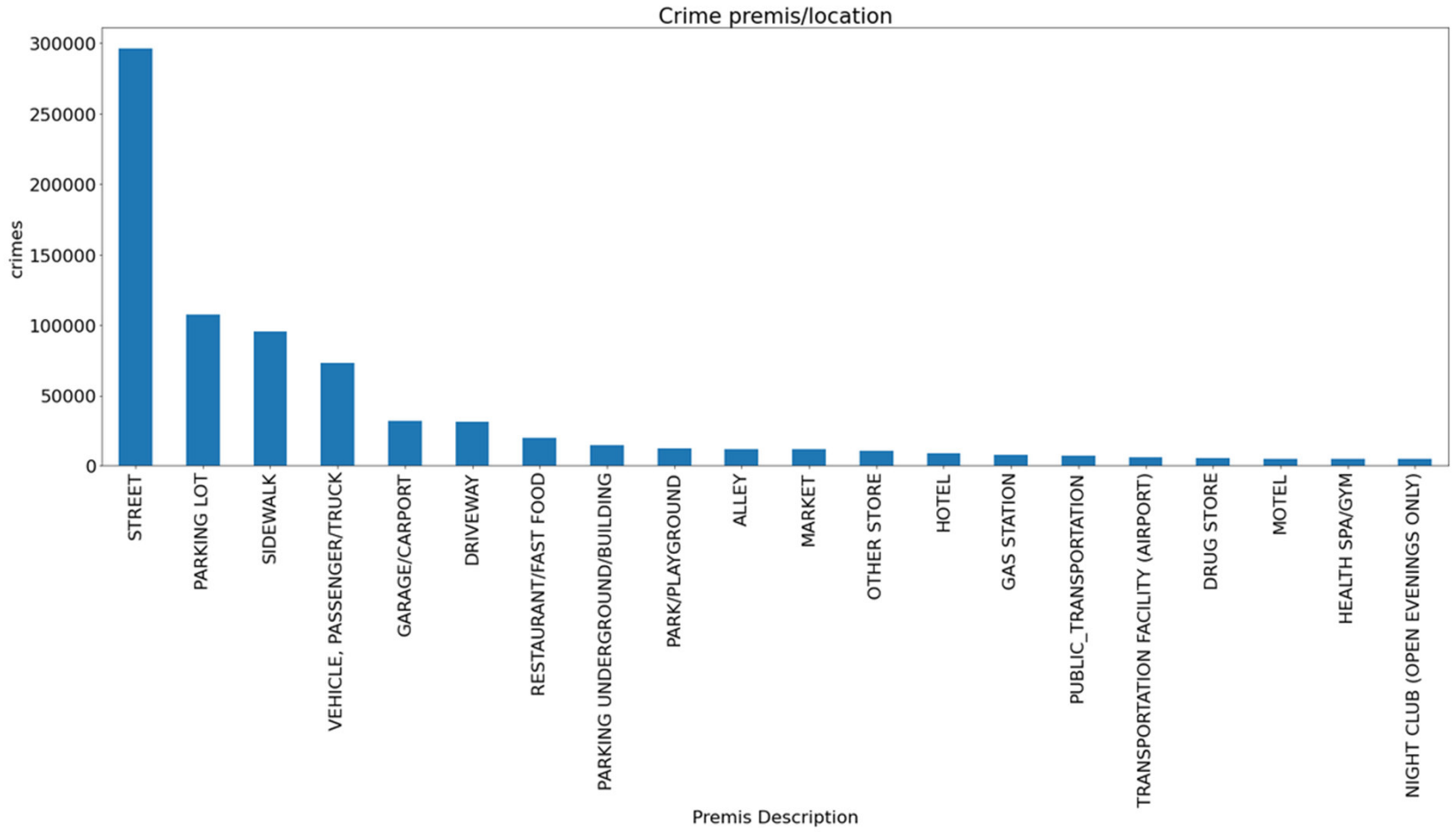
Überblick der Daten

Verteilung der Daten Crime Description nach der Gruppierung



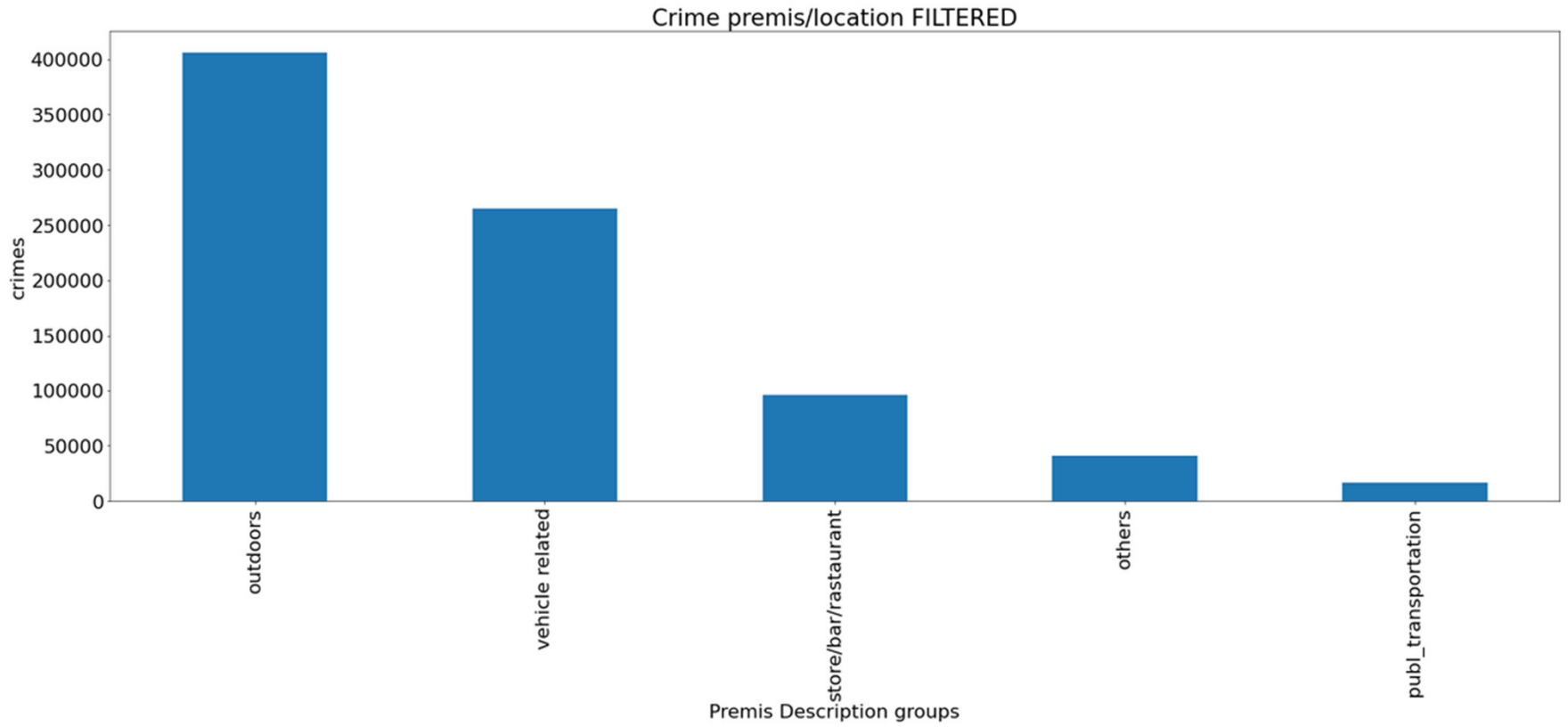
Überblick der Daten

Verteilung der Premis Description (Räumlichkeit) vor der Gruppierung



Überblick der Daten

Verteilung der Premis Description (Räumlichkeit) nach der Gruppierung

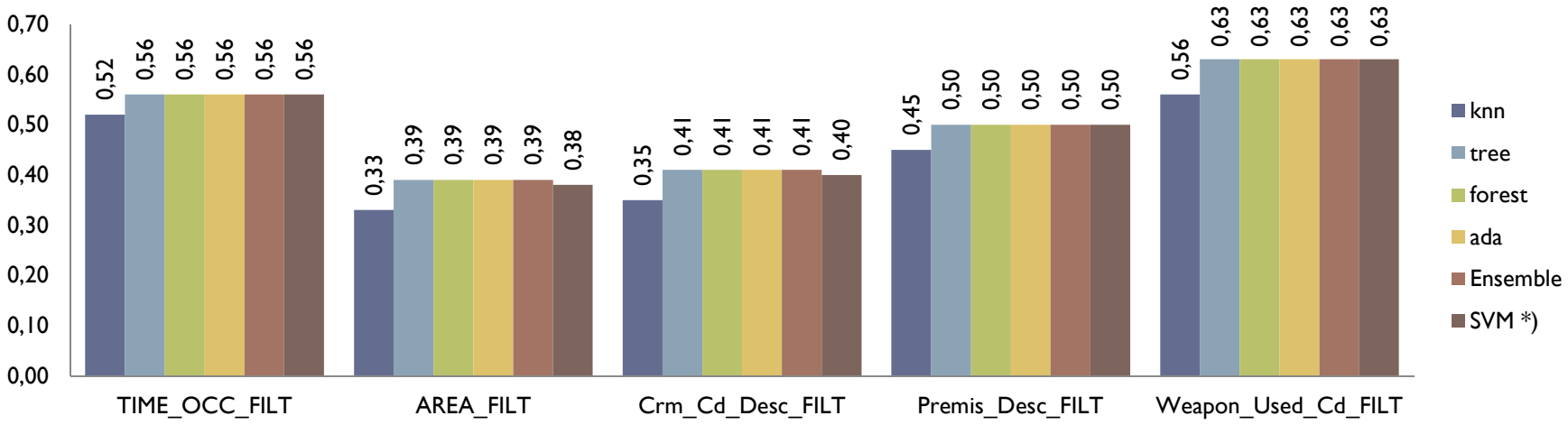


Prädiktionsmodelle für die Klassifikation

- Klassifikationsmethoden aus SKlearn
  - KNeighborsClassifier
  - DecisionTreeClassifier
  - RandomForestClassifier
  - AdaBoostClassifier
  - SVC
  - VotingClassifier als Ensambly Method
- Eingangsfeatures für die Klassifikation (X-Werte)
  - Opfers Alter: Vict_Age_FILT
  - Opfers Geschlecht: Vict_Sex_FILT
  - Opfers Herkunft: Vict_Descent_FILT
- Ausgangsfeatures für die Klassifikation (Y-Werte, Zielwerte)
  - Tageszeit Tag/Nacht: TIME_OCC_FILT
  - Area: AREA_FILT
  - Verbrechensbeschreibung: Crm_Cd_Desc_FILT
  - Location: Premis_Desc_FILT
  - Mit/ohne Waffe: Weapon_Used_Cd_FILT

Modellakkuranz

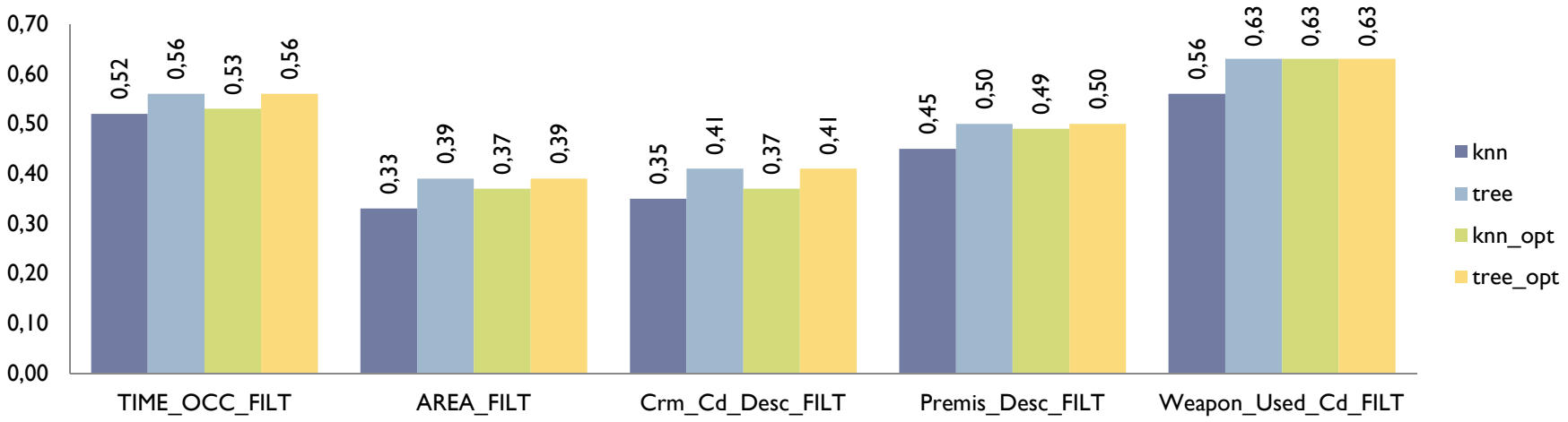
- Trotz Gruppierung der Features in Hauptgruppen ist die Akkuranz relativ gering.
- Mögliche Ursachen sind:
  - Falsche Zuordnung in die Hauptgruppen (fehlendes Fachwissen)
  - Fehlende Informationen im Datensatz (LA-Areas: Wohngebiet, Industriezone, Party-Locations, Immobilienpreise etc.)
  - Datensatz mit geringem Informationsgehalt (wenige bzw. schwache X-Faktoren für die Targetgruppe „Opfer“)
- Eine Anhebung der Akkuranz durch Hyperparameteroptimierung soll durchgeführt werden.



*) SVM mit 30% der Daten

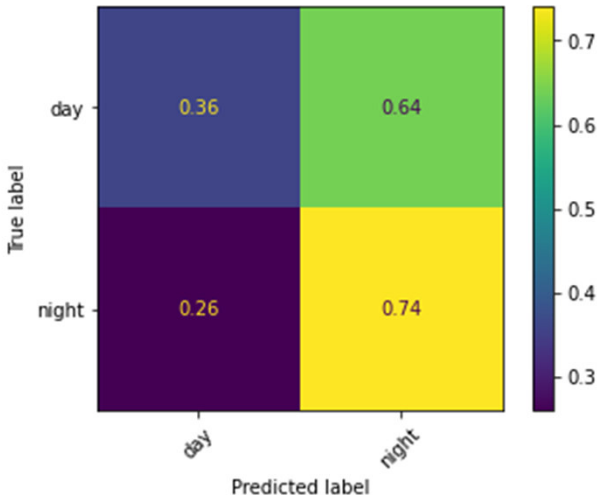
Hyperparameteroptimierung

- Aus Rechenzeitgründen soll die Optimierung nur bei den Methoden KNeighborsClassifier und DecisionTreeClassifier durchgeführt werden
  - KNeighborsClassifier: weights = ['uniform', 'distance‘]
    n_neighbors = list(range(15, 31)
  - DecisionTreeClassifier: criterion = ['gini', 'entropy']
    max_depth = [2, 4, 6, 8, 10, 12]
  - Optimierungsschleifen mit 30% der Daten aufgrund erhöhter Berechnungsdauer

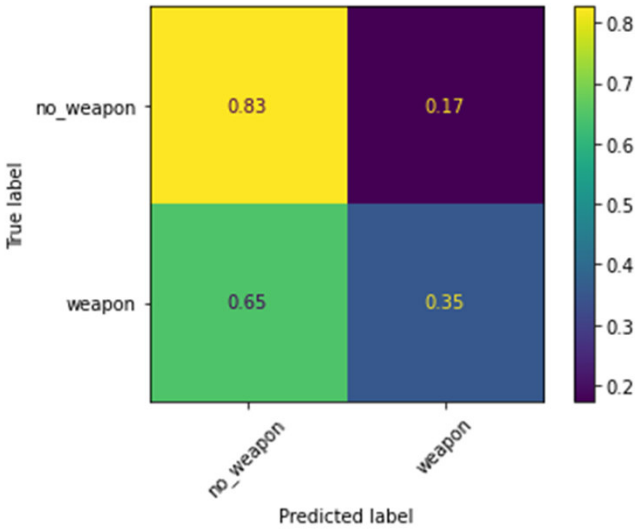


Confusion Matrix für DecisionTreeClassifier

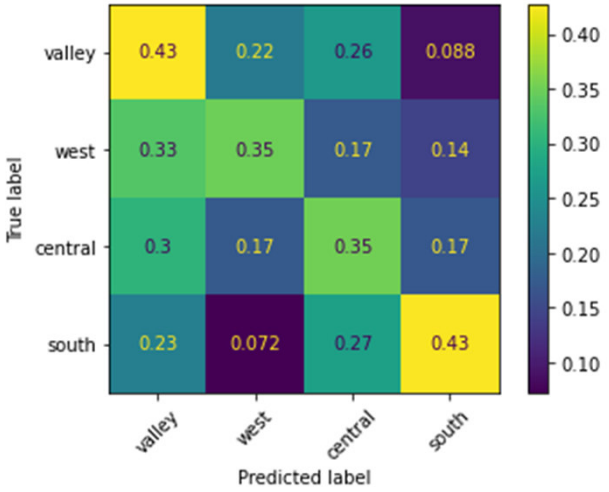
Daytime



Weapon

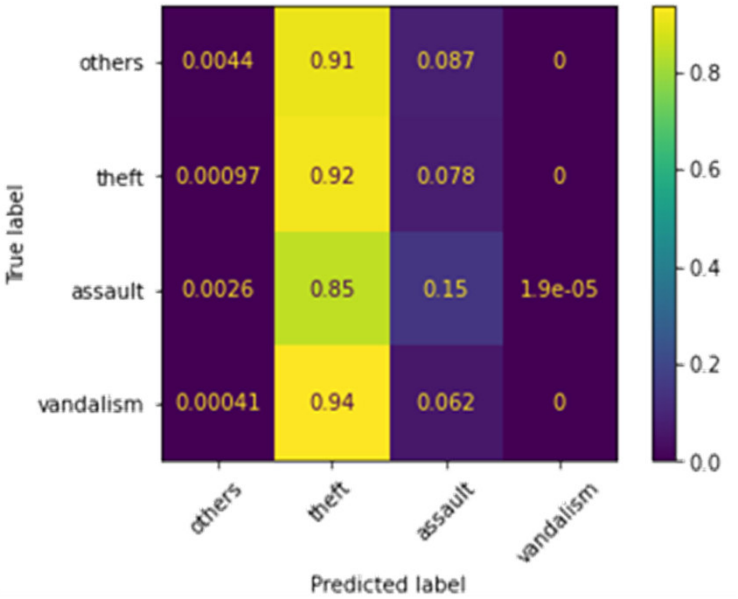


Area

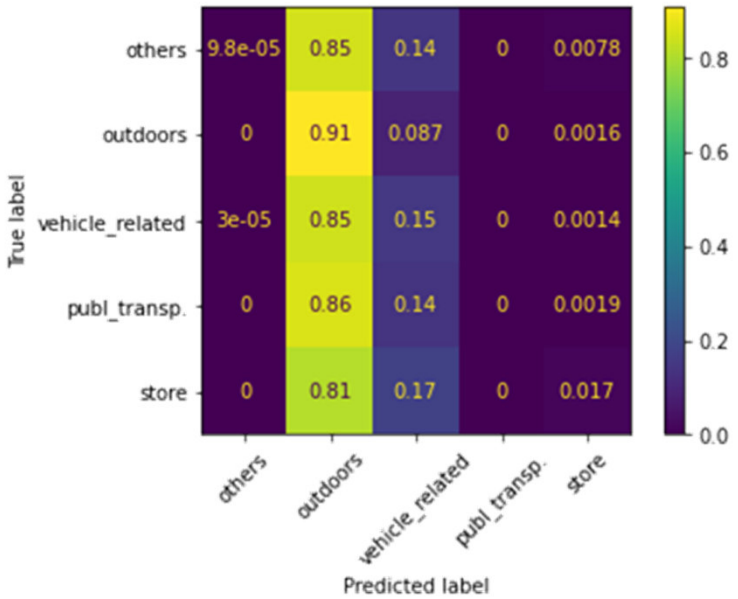


Confusion Matrix für DecisionTreeClassifier

Crime

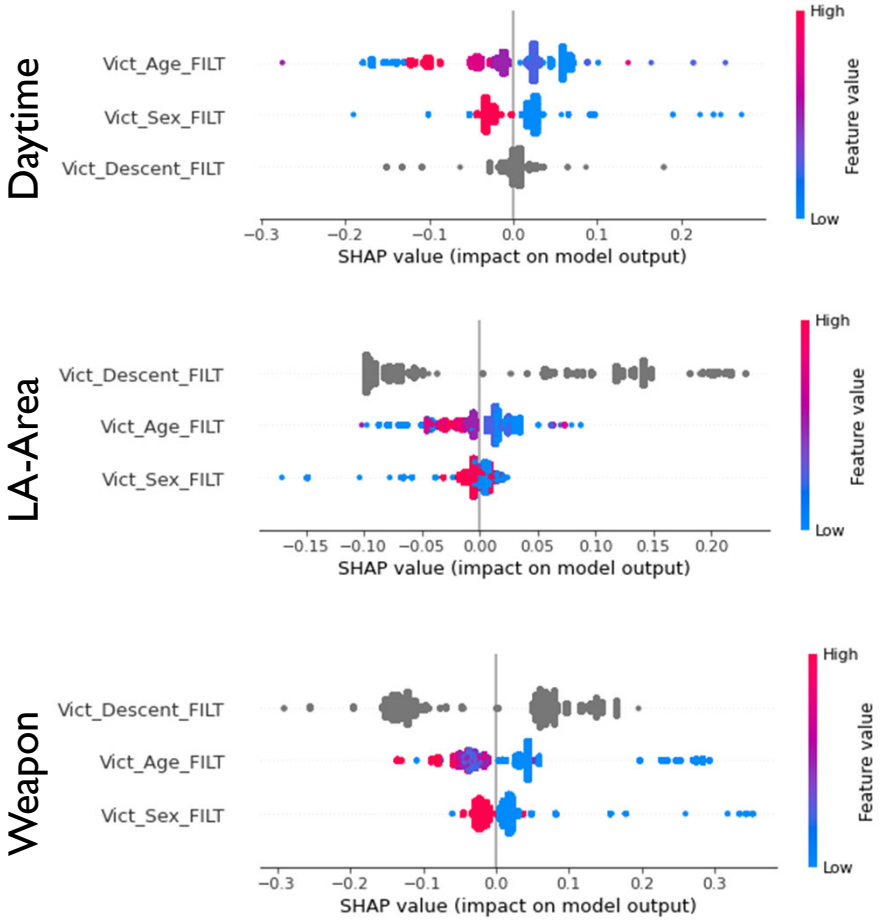


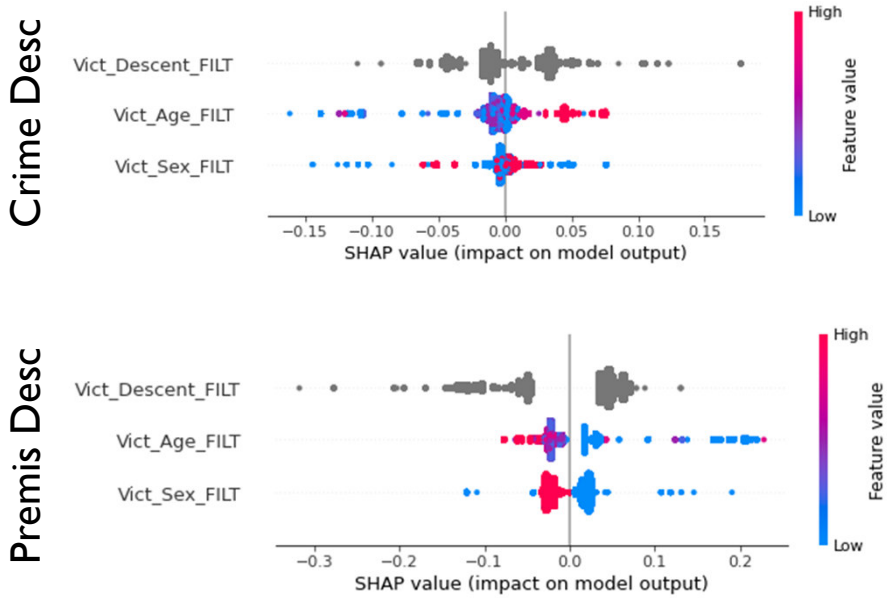
Premis



Zuordnung der meisten Elemente in der größten Klasse aufgrund schwacher X-Werte

SHAP Plots





Crime Desc und Premis Desc mit Anhäufung der Datenpunkte um den Nullpunkt
→ kleinerer Einfluss auf dem Modelloutput

Prädiktion

- Prädiktion mit DecisionTreeClassifier für:
  - Vict_Age = 37
  - Vict_Sex = M
  - Vict_Descent = white
- Ergebnis:
  - TIME_OCC_FILT → night
  - AREA_FILT → west
  - Crm_Cd_Desc_FILT → theft
  - Premis_Desc_FILT → outdoors
  - Weapon_Used_Cd_FILT → no_weapon