Master Thesis Report

Comparative Analysis of Machine Learning Models for Classification of Traffic Accident Severity in the Netherlands

Author: Jade Paus (2671276)

First Supervisor: Dr. Rikkert Hindriks
Second Reader: Prof. Dr. Rob van der Mei
Organizational Supervisor: Xander van Uffelen

Vrije Universiteit Amsterdam
Faculty of Science

February 2026

Abstract

Traffic accident severity prediction is challenged by complex nonlinear relationships and extreme class imbalance, particularly for severe and fatal outcomes. While previous research has demonstrated the potential of machine learning for accident severity analysis, comprehensive studies using long-term, national-scale data remain limited, especially for the Netherlands. This thesis investigates the extent to which supervised classification models can predict traffic accident severity in the Netherlands using large-scale national accident data from 2003–2024. Accident severity is modeled as a three-class classification problem, distinguishing material-damage-only accidents, accidents with serious or lethal injuries, and fatal accidents.

Baseline models Multinomial Logistic Regression and Decision Tree are compared with gradient boosting ensemble methods, including LightGBM, CatBoost, and XGBoost, with class imbalance addressed through weighted learning. Model performance is evaluated using stratified train–validation–test splits, and interpretability is achieved through SHAP-based feature attribution. The results show that gradient boosting models substantially outperform simpler classifiers, achieving meaningful recall for severe accident classes despite their rarity. XGBoost provides the strongest overall balanced performance, while LightGBM and CatBoost show slightly higher recall for the most severe outcomes.

Key influential predictors include the number of involved parties, speed limits, involvement of vulnerable road users and seniors, and vehicle type, with temporal factors playing a more prominent role in XGBoost. Overall, the findings demonstrate that interpretable ensemble models are well suited for severity prediction under extreme class imbalance and can provide actionable insights into the factors associated with severe traffic accidents, while also highlighting that accurately predicting fatal accidents remains inherently difficult due to their extreme rarity and many different underlying factors.

Contents

1 Introduction

Traffic accidents represent a persistent societal challenge, with significant human, economic, and social consequences. Despite ongoing investments in road safety, the number and severity of accidents remain a major concern worldwide [36]. Although the Netherlands performs generally well within the EU, compared to the EU average, specifically the distribution of fatalities shows a relatively high proportion of cyclist and people aged 65 and over[13]. Understanding the factors that contribute to accident severity - whether an incident results in material damage, injury, or fatality - can support policymakers, municipalities, and transport authorities in developing targeted interventions and preventive measures.

The *Bestand geRegistreerde Ongevallen Nederland (BRON)* dataset, provided by Rijkswaterstaat, offers a comprehensive record of road traffic accidents across the Netherlands of the past two decades, including details on location, time, environmental conditions, involved parties and vehicle. This dataset provides a valuable opportunity to explore the complex interplay between these variables and the severity of accidents through data-driven modeling.

This research focuses on developing and evaluating machine learning models that can predict the severity of a traffic accident based on locational, temporal, and situational features. As mentioned, the target variable accident severity can fall in one of the three categories: material damage only (UMS), lethal damage (LET) or a fatal outcome (DOD). The central research question guiding this study is:

"To what extent can the severity of traffic accidents in the Netherlands be predicted using classification models?"

To address this question, various supervised classification algorithms are applied and compared, including simple baseline models as well as various more advanced boosting-based ensemble classification models. The study also investigates the relative importance of different predictive features, to gain insights into which factors are most influential in determining accident outcomes. These insights can be used as journalistic novelties to publish in news articles of *De Volkskrant*.

The problem underlying this research lies in the complexity and heterogeneity of traffic accident data. Many factors interact nonlinearly, and traditional statistical methods often struggle to capture these relationships. Furthermore, the imbalance between accident severity classes - with most incidents resulting in material damage only and far fewer in injuries or fatalities - poses an additional modeling challenge. By employing modern machine learning techniques and careful feature engineering, this research aims to develop a reliable predictive framework that can support data-informed road safety strategies and contribute to the prevention of severe traffic accidents in the Netherlands.

2 Review of related literature

Predicting and understanding the severity of road traffic accidents has become an important research topic in transportation safety and data science. Traditional approaches to accident modelling relied mainly on statistical methods such as logistic regression, ordered probit/logit, or count-based Poisson and negative-binomial regressions. These models offered interpretability and theoretical grounding, but their linear structure limited their ability to capture complex nonlinear interactions between driver, vehicle, environmental, and temporal variables [22, 38].

Over the past decade, the emergence of machine learning (ML) methods has enabled more flexible and accurate modelling of accident outcomes. Unlike classical regression, ML models can automatically learn nonlinear relationships and high-order interactions from data without strong distributional assumptions [32]. Among these, ensemble tree-based models such as Random Forest, Gradient Boosted Decision Trees (GBDT), XGBoost, and LightGBM have shown outstanding predictive performance for traffic accident severity classification [23, 35, 12, 2, 26]. These methods efficiently handle high-dimensional, heterogeneous data and provide robustness against multicollinearity and missing values.

In particular, recent studies combining boosting-based models with explainable AI techniques, such as SHAP (SHapley Additive exPlanations), have enabled both strong predictive performance and interpretability. For instance, Wen et al. [35] demonstrated that LightGBM combined with SHAP can quantify and compare the marginal effects of key risk factors on crash types, while Dong et al. [12] applied SHAP-based interpretation to LightGBM models for injury severity prediction, identifying the most influential spatial and temporal features. Similarly, Parsa et al. [23] showed that XGBoost integrated with SHAP can accurately classify crash severity and simultaneously provide meaningful feature attributions suitable for policy design. These approaches represent an important step toward "explainable machine learning" in transport safety research.

Besides boosting-based models, other ML approaches such as Support Vector Machines, k-Nearest Neighbours, and Neural Networks have also been employed for crash severity or frequency prediction [26, 2]. Deep learning models, including convolutional and hybrid neural networks, have been used to exploit complex spatio-temporal dependencies in large-scale crash data. Rahim and Hassan [29] proposed a deep learning framework with a customized loss function to handle class imbalance, showing improved performance on injury severity prediction tasks. However, deep learning methods require large, high-quality datasets and often lack interpretability.

Some studies have focused on the spatial dimension of accident prediction. Li and Luo [17] developed a model integrating spatial and visual semantic features for road risk prediction, while Shahi et al. [33] conducted a spatial crash analysis for Rotterdam, emphasizing geographic and user-based factors. Similarly, Bakış et al. [5] and Pourroostaei et al. [26] demonstrated the benefits of ML models for spatially explicit crash trend prediction in Turkey and China, respectively. However, these studies often focus on regional or short-term datasets, limiting the generalizability of their conclusions.

Several comparative analyses underline that no single ML model dominates across all contexts. Random forests tend to offer good baseline performance and stability, while boosting models (XGBoost, LightGBM, CatBoost) usually achieve higher accuracy and better handling of complex, structured datasets [32, 12]. Deep learning can outperform traditional ML when abundant, high-resolution data (e.g., imagery, trajectory data) are available, but at the cost of transparency and higher computational demands [34]. Therefore, an appropriate modelling strategy involves using interpretable ML approaches such as gradient boosting in combination with feature importance or SHAP analysis, benchmarked against simpler statistical baselines such as logistic regression.

An important methodological distinction in crash research is between frequency modelling (predicting how often accidents occur) and severity classification (predicting how serious an accident is likely to be). Frequency models, often based on Poisson or negative-binomial regression, are useful for hotspot analysis and exposure assessment [22]. Variables on traffic intensity of some sort is crucial for this type of modelling in order to normalize the data, which is not included in the dataset on which this research is based. Classification models, by contrast, directly model the categorical severity levels (e.g. fatal, injury, property damage only) assuming an accident has happened [2, 23, 35]. These models are also perfect for analyzing which factors contribute to accident severity. For large datasets spanning multiple years and regions, including many information on situational and temporal factors, as in this research, classification models provide the flexibility and performance needed to handle heterogeneous, imbalanced, and nonlinear accident data.

Finally, while international research on crash severity prediction is abundant, the existing literature for the Netherlands remains limited. Most Dutch studies are either very outdated, spatial analyses focusing on specific municipalities (e.g., Rotterdam) or use smaller datasets covering only a few years [33, 25]. To date, no comprehensive study has applied modern classification models to predict accident severity across the entire Netherlands over two decades of national-level data. Addressing this gap can provide both methodological and policy contributions by benchmarking interpretable ML classifiers, such as LightGBM and XGBoost, on a uniquely rich and longitudinal Dutch dataset.

3 Data Preprocessing

3.1 Data Collection

The main raw datasets used in this research are *Verkeersongevallen - Bestand geRegistreerde Ongevallen Nederland* from Rijkswaterstaat [31] for the years 2003-2024. These contain all the registered accidents that happened in the Netherlands during that period. Each year contains four folders, the folder *Netwerkgegevens* contains data on the specific locations of the accidents and *Ongevallengegevens* contains data on the circumstances during the accidents and the involved parties. Party data is publicly unavailable for anonymity purposes, however through request the vehicle type and age group of the involved parties were obtained and added to the dataset. The other two folders *ReferentiebestandenOngevallen* and *ReferentiebestandenNetwerk* contain explaining reference data for variable ID values. All four files are combined into a single dataset through linked variable ID's, for all available years. All provided variable names are in Dutch, so for clarity purposes all newly created variables names are also in Dutch.
The dataset *Wijk- en buurtkaart 2025* from the Dutch Institution of Statistics (CBS) [8], which contains geodata on the borders of all municipalities in the Netherlands, is also obtained and merged for visualisation purposes.

The data used consists of 2,597,357 accidents. Each accident can involve multiple parties, thus the total number of involved parties in this dataset is 4,652,947. All examined variables are on accident-level, except for the vehicle type and age group, this is party-level data.

3.2 Data Cleaning

Many irrelevant columns with over 80% of missing values are dropped, such as house number or police district name. Variables on the number of casualties and kinds of casualties (hospitalized, emergency care etc.) are also dropped since these are a direct indicator of the target variable accident severity. Remaining interesting variables and their definition are presented below.

Accident-related variables

- "VKL_NUMMER" - Unique accident ID.
- "AP3_CODE" - Accident severity (material, lethal, fatal).
- "ANTL_PTJ" - Number of involved parties.

Party-related variables

- "OTE_OMS" - Vehicle or object type (tree, lantern etc.).
- "LKE_OMS" - Age group of the driver/pedestrian.

Location-related variables

- "GME_NAAM" - Municipality name of where the accident happened.
- "PVE_CODE" - Province name.
- "WPSNAAM" - City name.
- "X_COORD" - X coordinate based on Rijksdriehoeksmeting.
- "Y_COORD" - Y coordinate based on Rijksdriehoeksmeting.
- "BEBKOM" - Rural or urban area.

- "MAXSNELHD" - Maximum allowed speed limit.
- "WVG_OMS" - Road paving.
- "WSE_OMS" - Road situation (curve, roundabout, exit lane etc.)
- "RIJRICHTNG" - Driving direction.
- "BST_CODE" - Road segment (main road, parallel road, etc.)

Temporal variables

- "UUR" - Hour of the day during which the accident happened (0-23).
- "DDL_ID" - Part of the day (7h-9h, 9h-12h, 12h-16h, 16h-18h, 18h-22h, 22h-7h)
- "JAAR_VKL" - Year (2003-2024).
- "DAG_CODE" - Day of the week (MA-ZO).
- "MAAND" - Month number (1-12).
- "WEEKNR" - Week number (1-52).

Circumstantial variables

- "WDK_OMS" - Road surface coverage (snow, wet, etc.) during the accident.
- "WVL_OMS" - Road lighting (on/off/not present).
- "LGD_OMS" - Daylight, darkness or twilight/dawn.
- "WGD_CODE_1" - Weather conditions.
- "AOL_OMS" - Nature of the accident (rear-end, head-on etc.)

3.2.1 Missing values

Provided in Table 1, are those variables that contain missing values, the percentage of accidents with no data for these variables and the applied imputation method.

The instances for hour (UUR) and part of day (DDL_ID) are dropped since this percentage is neglectable and impossible to impute. Missing values for weather (WGD_CODE_1), road surface coverage (WDK_OMS) and the nature of the accident (AOL_OMS) are set to unknown because this cannot be imputed sensibly either. The same holds for vehicle type (OTE_OMS) and age group (LKE_OMS). All instances with missing vehicle types and age groups are material accidents, thus for all lethal and fatal accidents the party data is known. Good to mention is that this 12% of missing values is based on party-level.
The driving direction (RIJRICHTNG) and road segment (BST_CODE) are only known for accidents that happened on road sections, not for those that happened on junctions. Therefore these are set to unknown as well as these variables are inapplicable for junction-instances, which is about one-third of the data.

Missing values for the remaining variables are imputed based on rules.
Imputation for the maximum allowed speed limit (MAXSNELHD) first looks to find the speed limit from other accidents that happened on the same exact coordinate-location of the regarding road segment or junction, in the same year. If there is no recorded accident with this location for the same year, it takes the speed limit from the same location in nearest year, picking the speed of the lowest year whenever nearest years are tied.
The missing values indicating if the accident happened on a rural or urban road (BEBKOM) are the imputed

Variable name	Missing values %	Imputation
BEBKOM	34	Rule-based
MAXSNELHD	23	Rule-based
WSE_OMS	30	Rule-based
WVL_OMS	33	Rule-based
WVG_OMS	32	Rule-based
LGD_OMS	19	Rule-based
BST_CODE	30	Set to unknown
RIJRICHTNG	30	Set to unknown
WDK_OMS	32	Set to unknown
WGD_CODE_1	32	Set to unknown
AOL_OMS	37	Set to unknown
OTE_OMS	12	Set to unknown
LKE_OMS	12	Set to unknown
UUR	0.08	Instances are dropped
DDL_ID	0.08	Instances are dropped

Table 1: Variables with percentage of missing values and imputation method.

based on speed limit, imputing urban road for speed limits 15, 30 and 50, and rural road for all others.

Missing values for road situation (WSE_OMS) and road paving (WVG_OMS) are imputed based on accidents that happened on the same location, and if this is not available using a mapping of the same road segment or junction ID. Missing values for daylight (LGD_OMS) are imputed using a mapping table of the combination of hour and month. Road lightning (WVG_OMS) is imputed using a mapping table based on the combination of daylight, location and part of day.

3.2.2 Outliers

Very few variables in the dataset contain outliers. For the number of involved parties (ANTL_PTJ), all counts of 5 parties or more are merged into the same group. The variables for road segment (BST_CODE), vehicle type (OTE_OMS), road situation (WSE_OMS) and age group (LKE_ID_FIJN) have a large number of unique values or values that occur in very few instances, therefore some values with similar characteristics are are merged together so that the value range for these variables are smaller and more meaningful (e.g. e-bike and bicycle are merged, entry and exit lanes are merged, etc.).

3.3 Exploratory Data Analysis

3.3.1 Distribution of the data

The majority of the dataset concerns material-accidents, namely 83.6%, then 15.9% of the data regards lethal accidents and only 0.5% are on fatal accidents.
Figure 1 shows the yearly distribution of the total dataset, as well as broken down to each value of the target variable. A spokesperson of the Rijkswaterstaat with knowledge on this dataset has explained that during 2009-2012 their only data source to register accidents was the police, and that the police decided on a major reorganization in 2008 after which they significantly registered less accidents, especially for material accidents. At the urging of the Ministry of Infrastructure and Water Management, among others, and many users of the BRON dataset, the police gradually (taking several years) switched back to the standard recording of material accidents. This is highly likely to explain the large decrease visible for material accidents in 2009, and a steady increase visible from 2014 again. The decrease for lethal accidents between 2010-2014 might very well be influenced by this as well. The decrease around 2019-2020 can highly likely be explained by the Covid pandemic.
Figure 1 clearly shows that the distribution of the total number of accidents is highly similar to that of material accidents, since this is the largest proportion of the data. Therefor from here on, visualizations will be provided for each separate target value, but not for the dataset as a whole.

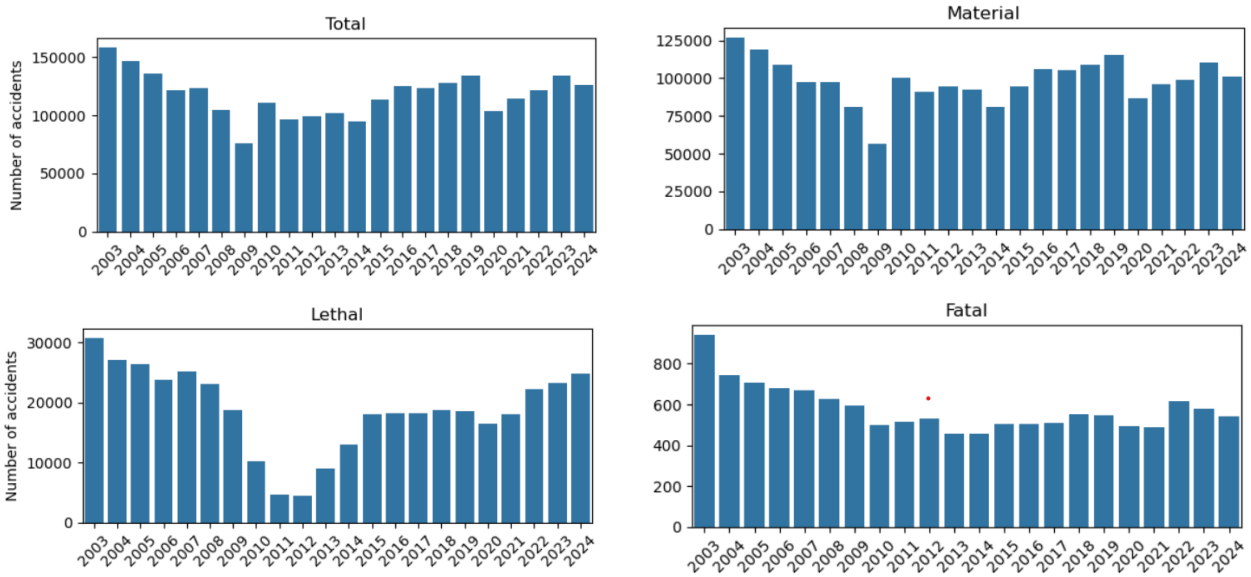


Figure 1: Distribution of accident severity over the years by severity.

3.3.2 Temporal features

Hour

Figure 2 shows the distribution of accidents during the day per hour the accident happened in, per severity outcome. All severity outcomes show that the afternoon rush period between 14.00-18.00 is most likely for accidents to happen in, and for material and lethal accidents the morning rush hour between 8.00-9.00 also stands out. Evenings and nights hours 20.00-7.00 have shown to be the least dangerous periods, especially for material and lethal accidents.

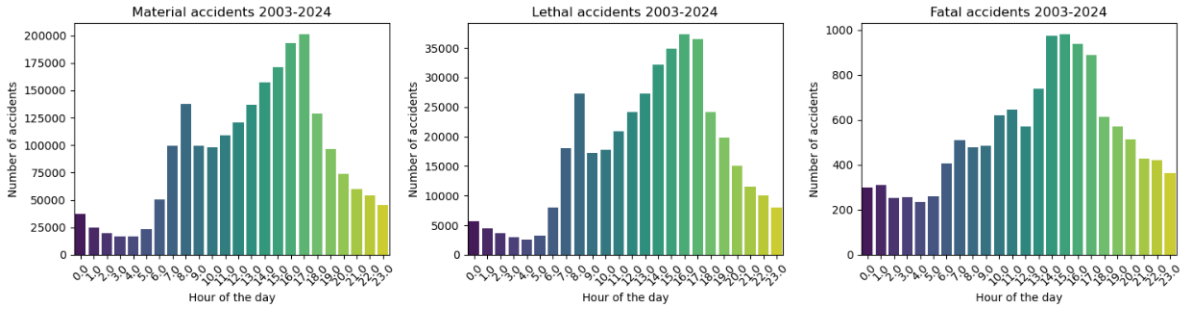


Figure 2: Total number of accidents by hour of the day per severity category

Day of week

Figure 3 shows the distribution of accidents across days of the week per accident outcome. Material and lethal accidents happened most often on weekdays, specifically on Fridays. Fatal accidents on the other happened more evenly distributed during both weekdays and weekends, peaking on Saturdays, which has the second to lowest accident count for material and lethal accidents remarkably.

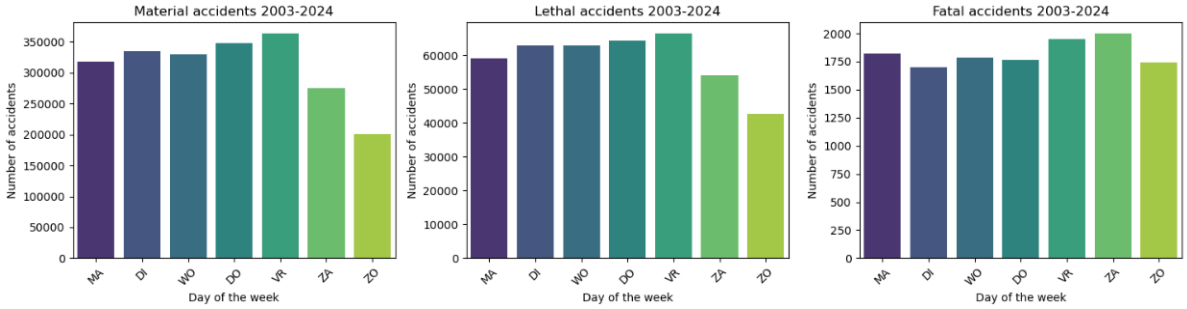


Figure 3: Total number of accidents by day of the week per severity category

Month

Figure9 shows the total number of accidents per month for the entire dataset, as well as categorized per severity category. It can be seen that material accidents happened most often during the colder/darker months November through January, and least in the summer months July and August. As for lethal accidents on the other hand, the numbers are lowest during those colder/darker months of December through February, and highest in June and September, the months before and after summer break. The fatal accidents count is also lowest in January and February and peaking in September.

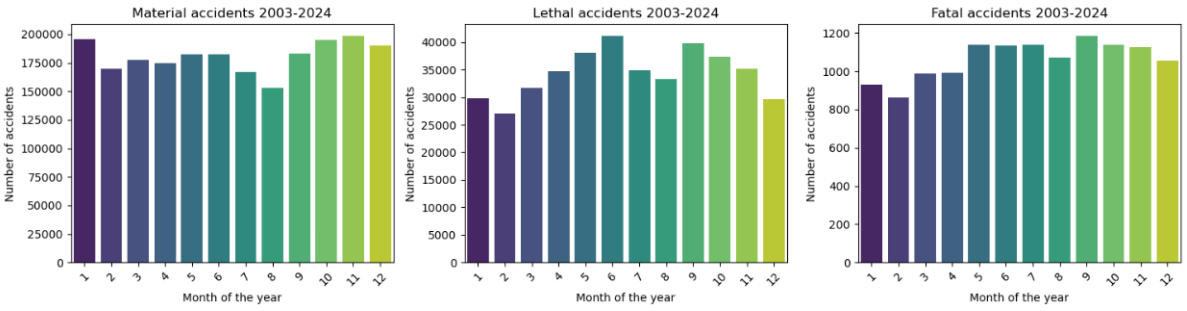


Figure 4: Total number of accidents by month per severity category

In Figure5 heatmaps for each severity category are shown for the frequency of accidents per combination of the hour of the day and month of the year, in order to investigate interactions between variables. It can be seen that material accidents happened clearly most often in one specific hour, namely during 17:00-18:00 in the month November, see also [24] in 8. Lethal and fatal accidents also happened very often during this hour, but lethal accident frequency peaking during 16:00-17:00 for the months June and September. As for fatal accidents, it can be observed that the morning rush hour is barely no indicator for these types of accidents as opposed to lethal and material ones, and is more spread around the entire period during 14:00-17:00 for the months April through October.

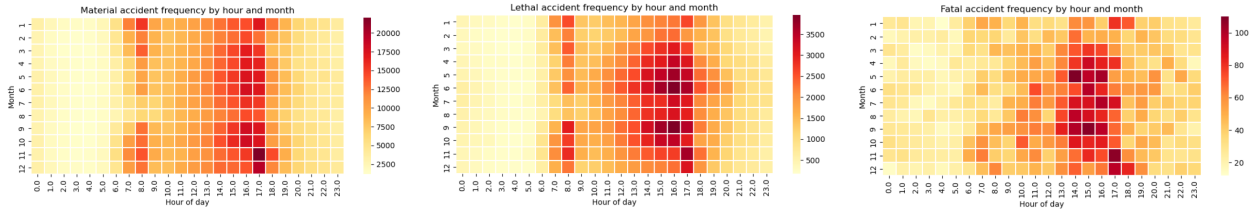


Figure 5: Heatmaps of hour of day by month of year per severity category.

3.3.3 Location features

Figure 6 indicates in which municipalities accident happened most often during the past two decades, and Figure 7 visualizes accident density across the Netherlands. It can be seen that on municipality-level, accident frequency is not spread out at all, but that a few municipalities clearly have highest accident count, such as Rotterdam, Amsterdam, the Hague, Groningen, Eindhoven and Enschede, which are also part of the municipalities with most inhabitants in the Netherlands. Looking at the accident density in the hexbin plot, the main highroads and largest cities are clearly highlighted as having the highest accident density.

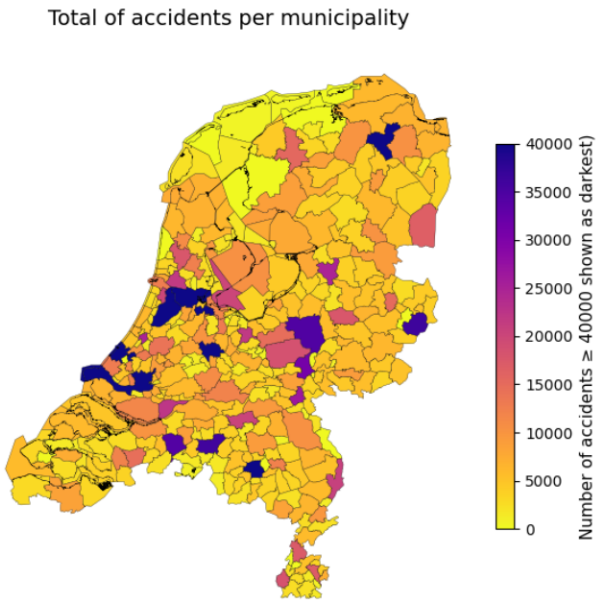


Figure 6: Heatmap for total number of accidents per municipality in the Netherlands

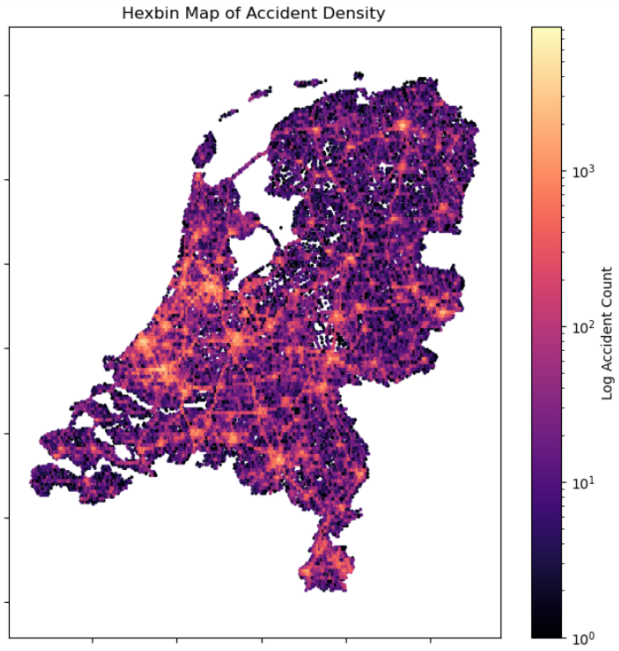


Figure 7: Hexbin map of accident density in the Netherlands

Junction VS Rural area

Figure 8 shows heatmaps for each severity category for the combination of whether accidents happened on a junction or not and in rural or urban areas. About 30% of accidents in the data happened on junctions. It can be seen that the frequency of accidents for this combination differ greatly per severity category. For material accidents the strongest indicator is they happened most often on road sections, not on cross-roads,

whereas for lethal accidents urban areas shows the biggest relation to this category, more specifically on cross roads. Fatal accidents very clearly happened almost always in rural areas and not on junctions.

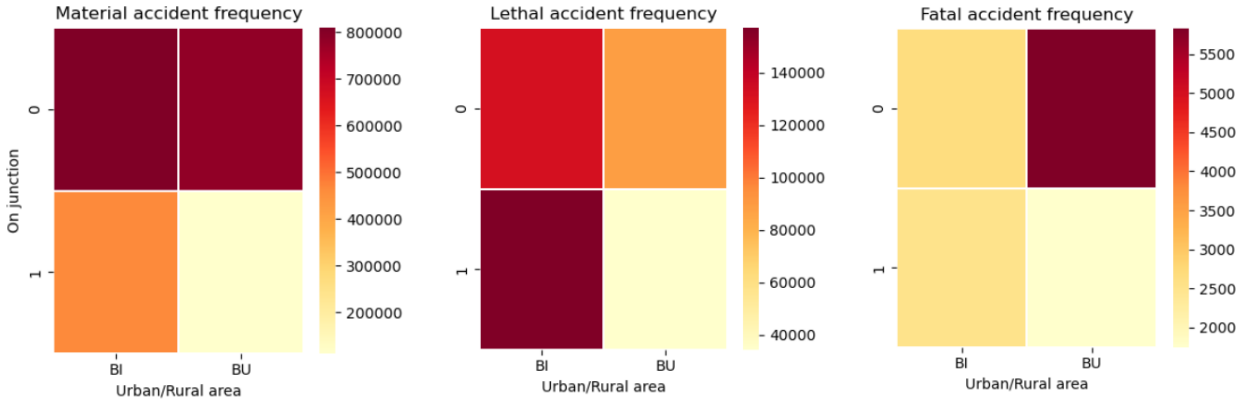


Figure 8: Heatmaps of junction or road section VS rural or urban located accidents per severity category.

3.3.4 Situational features

Speed limit

Across the three severity categories, there is a visible upward shift in the median and distribution peak of the maximum allowed speed limits on the locations of where the accidents happened. Material accidents show the strongest concentration around 50 km/h, typical of urban streets. Lethal injuries still peak around 50 km/h, but with noticeably more accidents in the 70–100 km/h range. Fatal accidents shift substantially upward, with the highest density around 80 km/h besides 50 km/h, and a larger share of accidents occurring at or above 100 km/h. This demonstrates a clear trend, namely the higher the mandated speed limit of the road, the greater the probability that an accident results in casualties. The fatal accident distribution (DOD) is much more spread out compared to the other categories however. This broader range suggests that while fatal accidents are more common at higher speed limits, they are not exclusive to them.

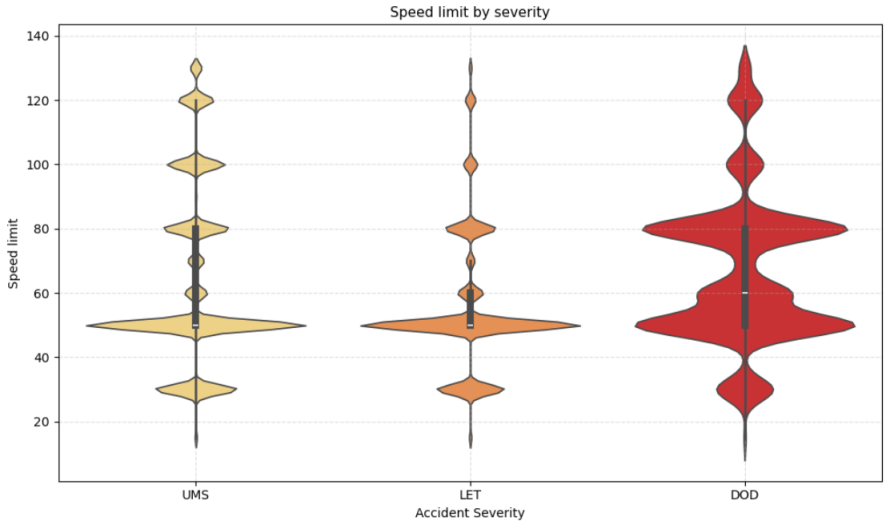


Figure 9: Violin-plot for speed limit by severity

Nature of the accident

From Figure 10 it can be observed that for material accidents the nature of the accident often was not recorded. Besides this, accidents where one party was hit from the side were most often the nature of material the accidents, just as for lethal and fatal accidents. In second place material accidents happened most often due to a rear-end collision. Lethal accidents were often a result of an individual party getting in trouble, like slipping of the road or getting a flat tire for instance. Fatal accidents remarkably happened often due to a collision with a solid object like a tree or lantern.

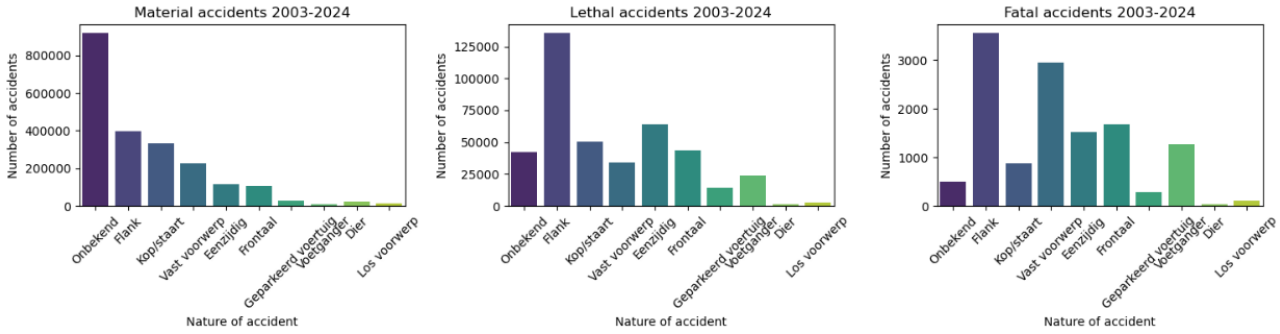


Figure 10: Most common nature of the accident per severity category

Number of involved parties

Figure 11 presents the proportion of involved parties for each accident severity category. It can be seen that a substantial share of material-damage accidents contains no recorded information on the number of involved parties. Furthermore, single-party accidents result almost equally frequently in both lethal and material outcomes. A notable observation is that accidents involving more than two parties show a relatively high share of fatal cases, almost 22% of fatal accidents involve three parties or more. Overall, the majority of accidents involve two parties.

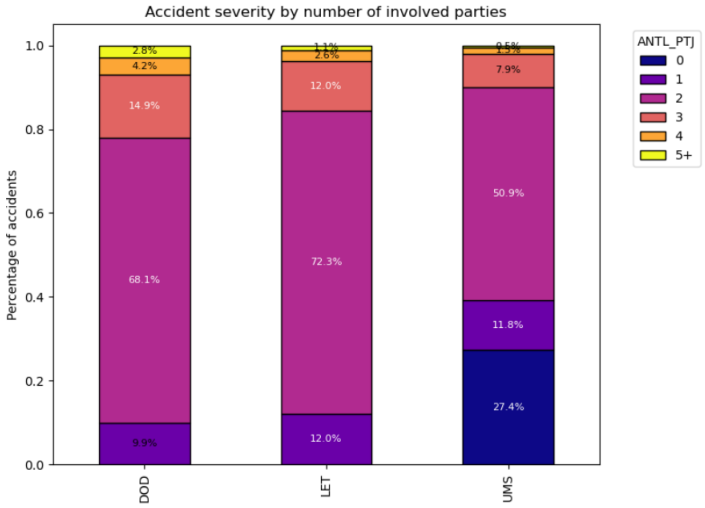
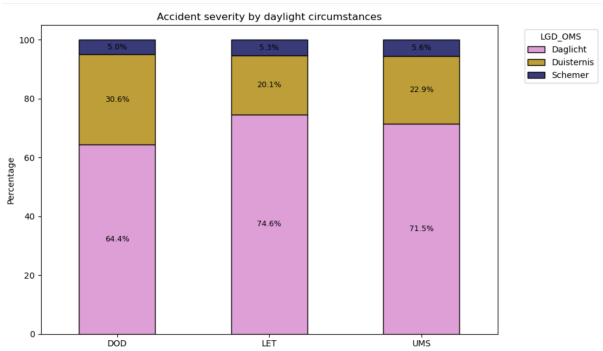


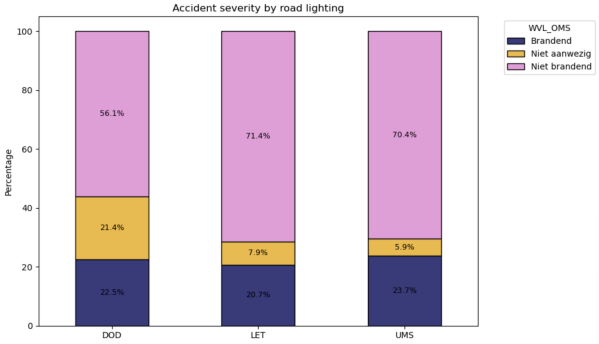
Figure 11: Stacked bar plot of percentage of number of involved parties per accident severity

Daylight circumstances & Road lighting

Figure 12a shows the proportion of accidents that happened during daylight, dusk or darkness/night for each accident severity category. It can be observed that lethal and material accidents show similar distributions, while accidents with a fatal outcome happen most frequently during the night/darkness of the three severity categories.
Figure 12b shows the proportion of accidents that happened on locations without road lighting, and with road lighting on or off. It can be seen that again lethal and material accidents show similar distributions, while fatal accidents happen more often on locations where no road lighting is present comparatively. The percentages of accidents happening during daylight and those during times when road lighting is off, show somewhat similar percentages that align with their definition.



(a) Stacked bar plot of percentage of accidents by daylight circumstances per accident severity



(b) Stacked bar plot of percentage of accidents by road lighting per accident severity

Road situation

Figure 13 shows the proportion of accidents that happened on certain road situations. It can be observed that of the three severity types, material accidents happen most often on straight roads, while lethal accidents happen more often on crossroads, and fatal ones on roundabouts.

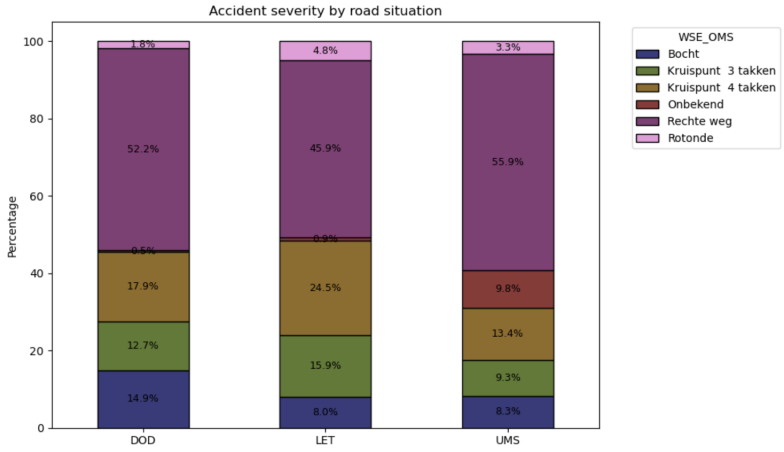


Figure 13: Stacked bar plot of percentage of accidents by road situation per accident severity

3.3.5 Party-level features

Vehicle types

Because of anonymity purposes, it is unknown for lethal and fatal accidents which party experienced the damage, thus it is not possible to link age groups and vehicle types directly to the casualties. Table 14 shows the 10 most occurring vehicles in the totality of accidents, as well as for each severity category the five most common combinations of vehicles. It can be seen that passenger cars are involved most often overall. As mentioned before, for lethal and fatal accidents all party-level variables are known, for material accidents however a large proportion does not contain information on the vehicle types and age groups of the parties. This is clearly reflected by the presence of "Onbekend" in second place in the total top 10 and material top 5. Other than that, the involved combinations or single vehicles differ quite a lot per severity category. Lethal accidents most often of all severity types involve cyclists, and fatal accidents stand out by the involvement of trees, pedestrians and motor cycles compared to lethal and material accidents.

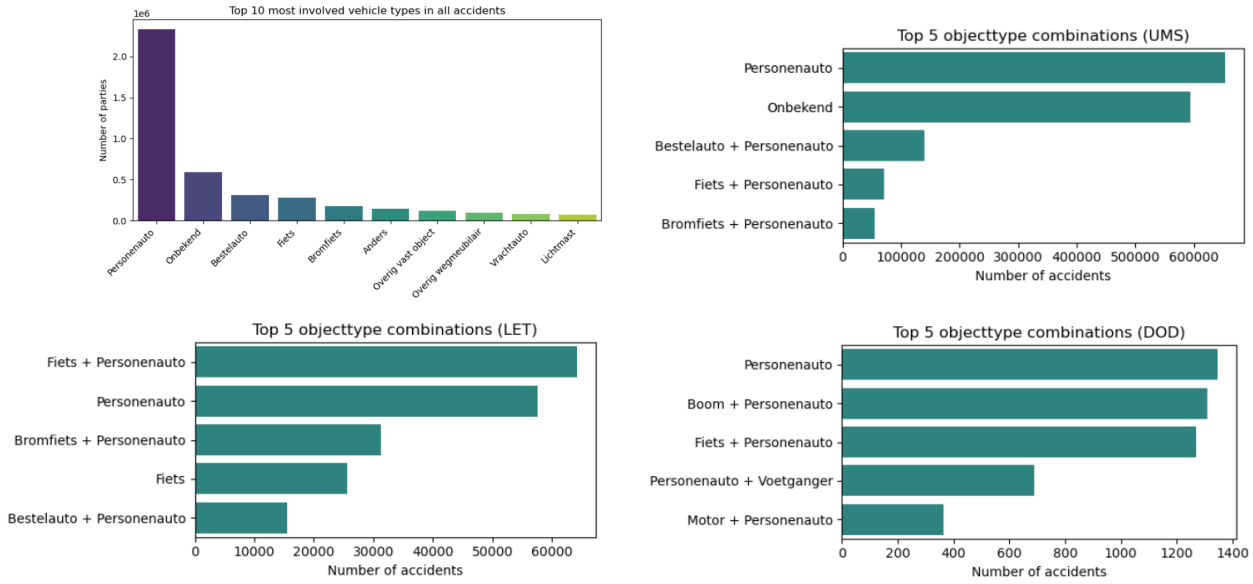


Figure 14: Most occurring vehicles in accidents, and most common vehicle combinations involved per severity category

About 25 different vehicle types are documented in the data, including many minority vehicles. In order to make more meaningful analyses on involved vehicle types, certain road users are grouped together, also explained in the feature engineering section 3.4. Figure 15 shows the vehicle types categorized into five road user groups. It can be seen that both vulnerable road users (pedestrians and cyclists) and smaller vehicles (scooter, moped, moped car, motor, mobility scooter) are much more often involved in lethal and fatal accidents compared to material ones. Also notable, is that accidents caused by drivers running into objects (tree, lantern, animal, road furniture, other solid objects, loose object) most often results in death. Accidents involving heavy vehicles (car, truck, agricultural vehicles, public transport) make up the largest proportion of accidents in general, and most often lead to material-damage only.

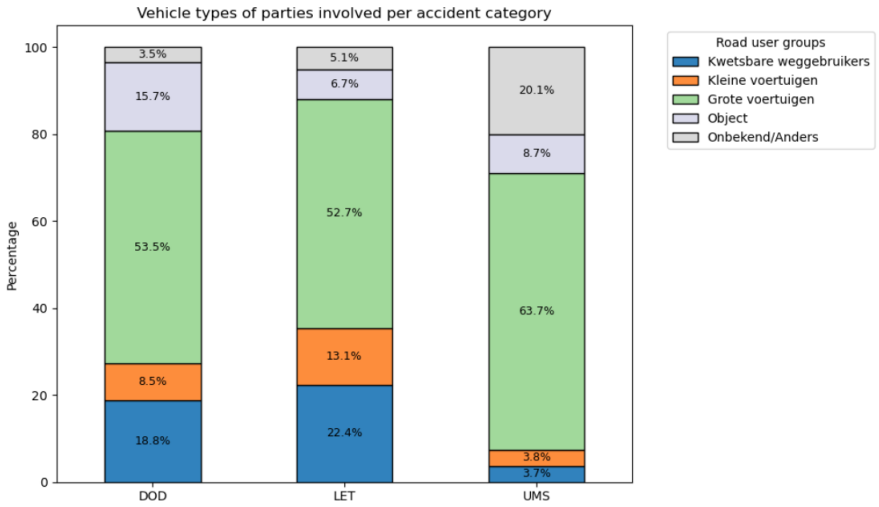


Figure 15: Stacked bar plot of percentage of road user groups per accident severity

Age groups

Figure 16 presents the number of parties involved in accidents for each age group in the dataset. It should be noted that the y-axis reflects the number of parties rather than the number of accidents. The age bins provided in the original data are relatively narrow, which makes differences in age representation across accident severity categories difficult to observe directly. Therefore, additional variables have been created, as described in the feature engineering section 3.4, by aggregating the age bins into four broader population groups: children (0-17), young drivers (18-29), mature drivers (30-64) and seniors (65+). Figure 17 displays the proportion of each population group involved in each accident severity category. Again, these proportions do not indicate that individuals in a given age group experience more injuries or fatalities, as the role of each party in the accident is unknown. Nonetheless, the figure shows that accidents resulting in death involve a higher share of seniors - approximately 10% more than in material accidents -, while accidents resulting in injury involve a substantially larger share of children - more than twice than observed in fatal and material accidents. In contrast, the proportion of young and mature drivers is highest in material-damage accidents.

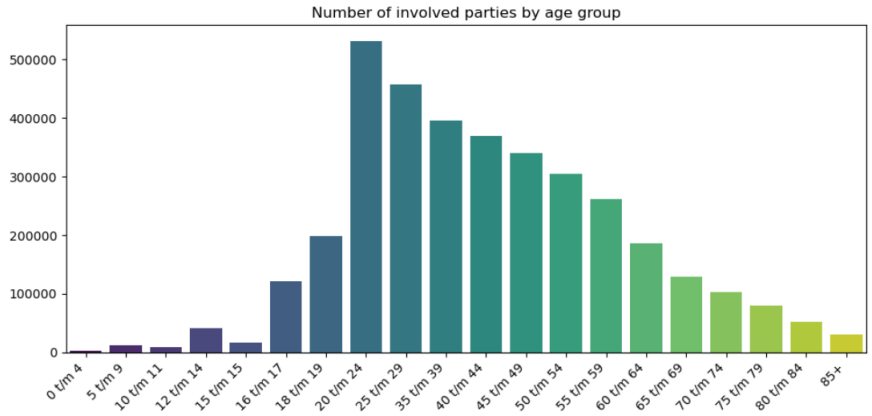


Figure 16: Histogram of age groups of parties involved in all accidents

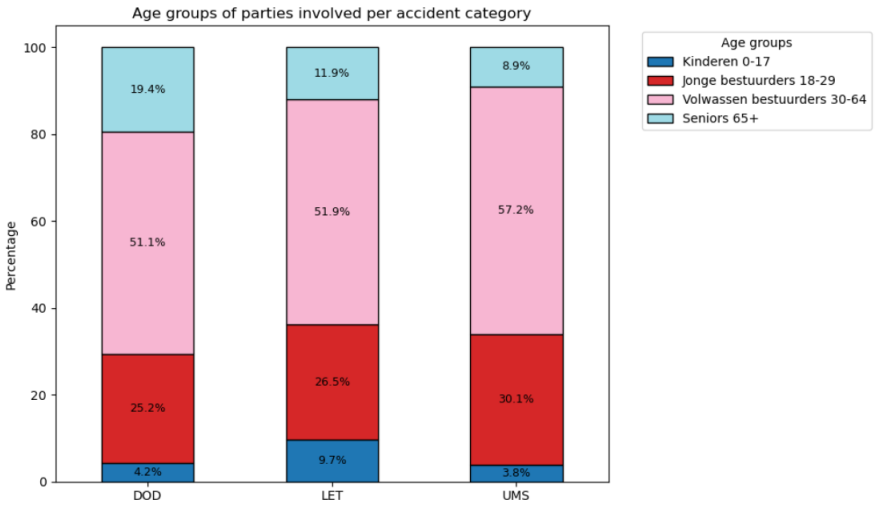


Figure 17: Stacked bar plot of percentage of population groups per accident severity

3.3.6 Correlations

Figure 18 shows the correlation matrix of all relevant variables that are explored in the data analysis, as well as newly added features that are presented in section 3.4. The accident-level dataset is used, therefor the correlations for age group and vehicle type might not be fully accurate, since if more than one party is involved in an accident their age group and vehicle is not included in this dataset.
Overall, most feature pairs show very weak correlation, indicating that the dataset does not exhibit severe multicollinearity. Expected correlations are observed between temporal features, such as month, week number, and day of year, reflecting their inherent cyclical structure. Weather conditions (WVG_ID_1) and road surface coverage (WDL_OMS) also show an expected correlation (wet road when it is raining). Some engineered features, such as (tijdens_spits) and (tijdens_werkuren), show moderate correlation with the hour of day and day of the week, which aligns with their definition.

The target variable of (AP3_CODE) demonstrates only weak correlations with most individual predictors, highlighting the complexity of accident severity and the need for non-linear models.
Highest positive correlations for severity can be observed with road surface coverage (WDK_OMS, 0.25), the nature of the accident (AOL_OMS, 0.21) and weather conditions (WGD_CODE_1, 0.20) and highest negative correlations with age group (LKE_ID_FIJN, -0.23), number of involved parties (ANTL_PTJ, -0.23) and junction-located (KRUISPUNT, -0.15). Road surface coverage and weather conditions might very well score high because of the large proportion of unknown data for mostly material accidents.
Some other remarkable correlations can be seen between age group, number of involved parties, nature of the accident and weather conditions.

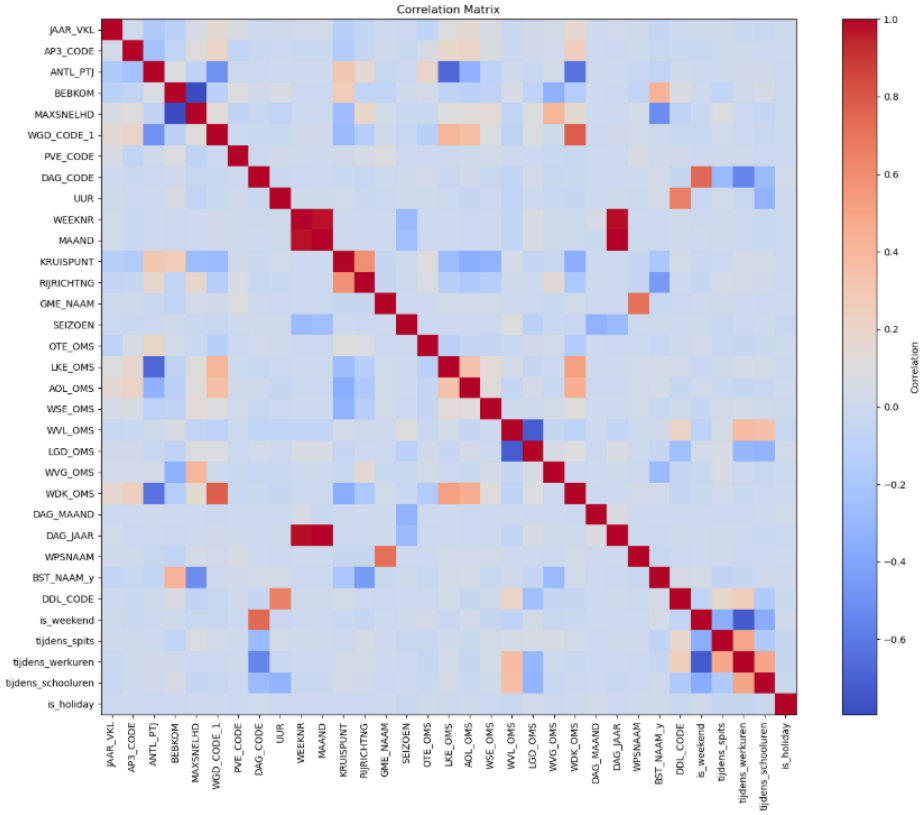


Figure 18: Correlation matrix

3.3.7 Inferential statistics

While the exploratory analysis reveals apparent differences in accident severity across several categorical variables, these observations are descriptive in nature. To assess whether the observed differences are statistically meaningful, chi-square tests of independence are conducted for selected variables. Furthermore, interaction effects identified during exploratory analysis are formally assessed using multinomial logistic regression models including interaction terms in order to test whether these observations are inferential.

Pearson's chi-squared test

First of all the Pearson chi-square test of independence is applied. Given a contingency table with observed frequencies O_{ij} and expected frequencies E_{ij} under the assumption of independence, the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

which follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom [21]. The associated p-value represents the probability of observing such a deviation from independence under the null hypothesis. Due to the very large sample size (approximately 2.6 million observations), p-values are expected to be extremely small even for weak associations; therefore, statistical significance alone is insufficient to assess practical relevance. To quantify the strength of association independently of sample size, Cramér's V is computed as an effect size measure:

$$\mathrm{V} = \sqrt{\frac{\chi^2}{n \cdot (\min(r, c) - 1)}}, \tag{2}$$

where n denotes the total number of observations [10]. Cramér's V ranges between 0 and 1, with values below 0.05 indicating negligible association, values between 0.05 and 0.10 weak association, values between 0.10 and 0.20 moderate association, and values above 0.20 strong association.

To assess whether two explanatory variables jointly interact with accident severity, the same chi-square framework is extended to interaction effects by constructing a contingency table for the combined variable formed by the Cartesian product of two categorical variables. Let variables A and B have r and c categories respectively, and let Y denote the outcome variable. The interaction variable $A \times B$ then has $r \cdot c$ categories, and the chi-square statistic is computed as

$$\chi^2_{\text{int}} = \sum_{k=1}^{r \cdot c} \sum_{j=1}^{h} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}, \tag{3}$$

where O_{kj} and E_{kj} denote the observed and expected frequencies for the interaction category k and outcome category j, and h is the number of outcome categories [1]. Under the null hypothesis that the joint distribution of A and B is independent of Y, the statistic follows a chi-square distribution with $(r \cdot c - 1)(h - 1)$ degrees of freedom. As with single-variable tests, the corresponding effect size is quantified using Cramér's V:

$$\mathrm{V}_{\text{int}} = \sqrt{\frac{\chi^2_{int}}{n \cdot (\min(r \cdot c, h) - 1)}}, \tag{4}$$

which provides a standardized measure of the strength of interaction effects. This approach allows interaction effects to be evaluated consistently with single-variable associations while accounting for the increased dimensionality introduced by combining categorical predictors.

Results Cramér's V

Indeed, all tested single variables yield very small p-values ($p < 0.0001$), indicating statistically significant associations with accident severity. However, given the large sample size, these results primarily confirm the presence of deviations from independence rather than their practical relevance. Table 2 therefore reports Cramér's V values to assess the strength of these associations. Overall, most variables exhibit negligible to weak associations with accident severity, with Cramér's V values below 0.10. Temporal variables such as DAG_CODE, UUR, MAAND, and WEEKNR show particularly weak associations, suggesting that time-related factors alone have limited explanatory power for accident severity. Moderate associations are observed for infrastructural and contextual variables, including KRUISPUNT, ANTL_PTJ, and MAXSNELHD. The strongest associations are found for AOL_OMS and OTE_OMS, both exceeding a Cramér's V of 0.20, indicating a relatively strong relationship with accident severity.

The variables WVG_OMS, BST_NAAM, RIJRICHTNG, WGD_CODE_1 and WDK_OMS are not presented in the exploratory data analysis. This is because for these features, the majority of missing values occur within the material damage only accident category, and the observed distributions of the remaining values of these variables are highly similar across severity levels. Therefore, the corresponding association measures of Cramér's V are likely driven by the disproportionate presence of missing values in a single target class rather than by substantive differences between severity outcomes.

Table 3 presents Cramér's V values for pairwise interactions between the most sensible combinations of variables. Interactions involving only temporal variables (DAG_CODE, MAAND and UUR) exhibit weak associations, comparable in magnitude to their individual effects, suggesting limited interaction strength within purely temporal dimensions. In contrast, interactions combining temporal variables with infrastructural or contextual factors generally yield higher Cramér's V values. Notably, interactions involving LKE_OMS, ANTL_PTJ, MAXSNELHD, and KRUISPUNT consistently reach moderate association levels (Cramér's V ≈ 0.15–0.20), indicating that the relationship between these features and accident severity depends partially on time-related conditions.

The strongest interaction effects are observed for combinations involving AOL_OMS and OTE_OMS. Interactions such as AOL_OMS & OTE_OMS, OTE_OMS & ANTL_PTJ, and OTE_OMS combined with temporal or infrastructural variables all exceed a Cramér's V of 0.25, with AOL_OMS & OTE_OMS reaching a peak of 0.311. These results suggest that accident severity is most strongly associated with complex, multi-factor configurations involving accident location characteristics or contextual conditions, rather than isolated features.
Taken together, the single-variable and interaction analyses indicate that mainly involvement of AOL_OMS and OTE_OMS and thus the combination of the nature of the accident or vehicle type characteristics with contextual factors yields substantially stronger associations with accident severity.

A more detailed illustrative example is provided in the published article shown in Figure 25 (see Section 8). This figure presents an analysis of lethal and fatal accidents involving at least one cyclist in the Netherlands between 2018 and 2024. The results highlight, for instance, that in The Hague, one-sided bicycle accidents occur at a substantially higher rate compared to other municipalities and accident types. This pattern demonstrates how the interaction between location, vehicle type, and the nature of the accident can affect accident severity: the same combination of factors is not observed to the same extent among accidents resulting in material damage only. Such an example supports the broader finding that combinations of contextual and crash characteristics can be more informative for understanding severity outcomes than marginal associations alone.

Feature	Cramér's V
DAG_CODE	0.0140
LGD_OMS	0.0205
UUR	0.0248
MAAND	0.0342
WEEKNR	0.0353
WVL_OMS	0.0410
PVE_CODE	0.0692
WVG_OMS	0.0801
BST_NAAM	0.0865
BEBKOM	0.0903
LKE_OMS	0.0953
GME_NAAM	0.0994
MAXSNELHD	0.106
RIJRICHTNG	0.125
WSE_OMS	0.129
KRUISPUNT	0.158
ANTL_PTJ	0.176
WGD_CODE_1	0.190
WDK_OMS	0.193
AOL_OMS	0.241
OTE_OMS	0.271

Table 2: Cramér's V of tested variables

Features	Cramér's V
DAG_CODE & MAAND	0.0385
DAG_CODE & UUR	0.0339
UUR & MAAND	0.0444
MAXSNELHD & UUR	0.111
BEBKOM & KRUISPUNT	0.119
MAXSNELHD & PVE_CODE	0.132
DAG_CODE & LKE_OMS	0.163
MAAND & LKE_OMS	0.165
UUR & LKE_OMS	0.167
KRUISPUNT & LKE_OMS	0.170
LKE_OMS & ANTL_PTJ	0.175
DAG_CODE & ANTL_PTJ	0.178
PVE_CODE & LKE_OMS	0.180
UUR & ANTL_PTJ	0.181
WGD_CODE_1 & LKE_OMS	0.182
LKE_OMS & MAXSNELHD	0.182
WSE_OMS & ANT_PTJ	0.190
WGD_CODE_1 & UUR	0.192
WGD_CODE_1 & MAAND	0.193
WDK_OMS & WSE_OMS	0.199
WGD_CODE_1 & ANTL_PTJ	0.201
MAXSNELHD & ANTL_PTJ	0.201
AOL_OMS & LKE_OMS	0.238
AOL_OMS & MAAND	0.244
AOL_OMS & UUR	0.245
AOL_OMS & PVE_CODE	0.254
AOL_OMS & KRUISPUNT	0.255
AOL_OMS & WSE_OMS	0.260
AOL_OMS & WDK_OMS	0.262
AOL_OMS & MAXSNELHD	0.266
AOL_OMS & ANTL_PTJ	0.269
MAAND & OTE_OMS	0.272
DAG_CODE & OTE_OMS	0.272
UUR & OTE_OMS	0.274
KRUISPUNT & OTE_OMS	0.277
LKE_OMS & OTE_OMS	0.279
WGD_CODE_1 & OTE_OMS	0.281
PVE_CODE & OTE_OMS	0.282
WSE_OMS & OTE_OMS	0.283
MAXSNELHD & OTE_OMS	0.286
OTE_OMS & ANTL_PTJ	0.290
AOL_OMS & OTE_OMS	0.311

Table 3: Cramér's V of tested interactions between variables

3.4 Feature engineering

3.4.1 Feature Creation

After exploration of the dataset, the date and year variables and x and y coordinates are dropped, since they do not directly encode interpretable or causal factors related to accident severity and could lead to overfitting.

Additional temporal features that have proven to be insightful in the related literature are created and added to the dataset, which are presented below. Furthermore, in order to capture the party-level data even better, the top 10 most occurring vehicle combinations in accidents for each severity outcome are added as new features. Also, certain risk groups are created for grouping similar vehicle types together, such as vulnerable road users or larger vehicles, and additional variables of larger age groups are created.

Created variables

- "SEIZOEN" - Season in which the accident happened.
- "KRUISPUNT" - Accident happened on a junction.
- "is_weekend" - Accident happened during the weekend.
- "tijdens_spits" - Accident happened during rush hour (7.00-9.00 or 16.00-19.00).
- "tijdens_schooluren" - Accident happened during school hours (8.00-15.00 on a weekday).
- "tijdens_werkuren" - Accident happened during work hours (7.00-20.00 on a weekday).
- "is_feestdag" - Accident happened on a Dutch holiday (New Years Eve, New Years Day, Good Friday, Easter Sunday and Monday, Kingsday, Liberation Day, Ascension Day, Pentecost, Saint Nicholas, Christmas).
- "fiets_auto" - Collision between bicycle and passenger car.
- "bromfiets_auto" - Collision between scooter and passenger car.
- "bestelauto_auto" - Collision between van and passenger car.
- "fiets_fiets" - Collision between bicycles.
- "auto_auto" - Collision between passenger cars.
- "fiets_voetganger" - Collision between bicycle and pedestrian.
- "auto_voetganger" - Collision between passenger car and pedestrian.
- "boom/lichtmast_auto" - Collision between tree/lantern and passenger car.
- "motor_auto" - Collision between motorcycle and passenger car.
- "vastobject_auto" - Collision between solid object (other than tree or lantern) and passenger car.
- "wegmeubilair_auto" - Collision between other road furniture than lanterns/trees and passenger car.
- "fiets_bestelauto" - Collision between bicycle and van.
- "snorbromscoot_auto" - Collision between moped/moped car/mobility scooter and passenger car.
- "fiets_vrachtauto" - Collision between bicycle and truck.

- "auto_vrachtauto" - Collision between passenger car and truck.
- "kwetsbare_weggebruikers" - Accident involves at least on of the following: bicycle, e-bike or pedestrian.
- "kleine_voertuigen" - Accident involves at least on of the following: scooter, modped, moped car, mobility scooter, motorcycle.
- "grote_voertuigen" - Accident involves at least on of the following: passenger car, van, truck, public transport or agricultural vehicle.
- "object" - Accident involves at least on of the following: tree, lantern, animal, other road furniture, other solid object or loose object.
- "kind" - Accident involves party in one of the following age groups; 0-4, 5-9, 10-11, 12-14, 15-15, 16-17.
- "jonge_bestuurder" - Accident involves party in one of the following age groups; 18-19, 20-24, 25-29.
- "volwassen_bestuurder" - Accident involves party in one of the following age groups; 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64.
- "senior" - Accident involves party in one of the following age groups; 65-69, 70-74, 75-79, 80-84, 85+.

3.4.2 Feature Encoding & Scaling

After feature creation, feature encoding is applied. Frequency encoding is applied for the municipalities and cities as these have hundreds of unique values. Vehicle type and age group are encoded using count encoding in order to capture the information if for instance two cars were involved, which would be lost when using one-hot encoding. This way the dataset is also automatically transformed from party-level to accident-level. All temporal features are kept numeric, except for day of week and season. These are one-hot encoded, as well as all other remaining categorical variables.

3.4.3 Feature Selection

After feature encoding, the final modelling dataset consists of 2,595,343 accidents with 139 numerical features. Feature selection is conducted in two stages. First, a variance threshold of 0.01 is applied to remove near-constant predictors, reducing the set to 115 features. Second, a Random Forest classifier (70/30 train–test split, 200 trees, random state 42) is used to estimate feature importance. Random Forest is a robust ensemble method that performs well on high-dimensional, mixed-type datasets and can naturally capture nonlinear relationships and complex interactions between variables. Unlike linear models, it does not assume monotonicity, handles multicollinearity well, is resilient to noise, and provides stable, interpretable estimates of feature importance through impurity reduction[7].

Features with an importance score above half of the mean importance are retained for model development. This results in a final set of 58 predictive features. The 20 features with highest scoring importance are presented in Figure 19. The Random Forest feature importance results show a clear hierarchy among predictors. Five features have relatively high importance scores (0.30–0.65), indicating that they contribute substantially more to the model than all other variables. Scoring highest is the vulnerable road users indicator, then come the three major temporal features week number, hour and month, and fifth if at least one bicycle is involved in the accident. Another group of four features falls in the medium-importance range (0.20–0.30), which is still considered meaningful in tree-based models. These are the variables indicating a small vehicle is involved, the number of involved parties, the part of day and the speed limit. The remaining predictors have relatively low importance values (≤ 0.20), suggesting a limited marginal contribution. Such a pattern, with a small subset of dominant predictors and many weaker ones, is common in large, heterogeneous datasets and indicates that accident severity is driven primarily by a few key factors rather than uniformly across all available variables.

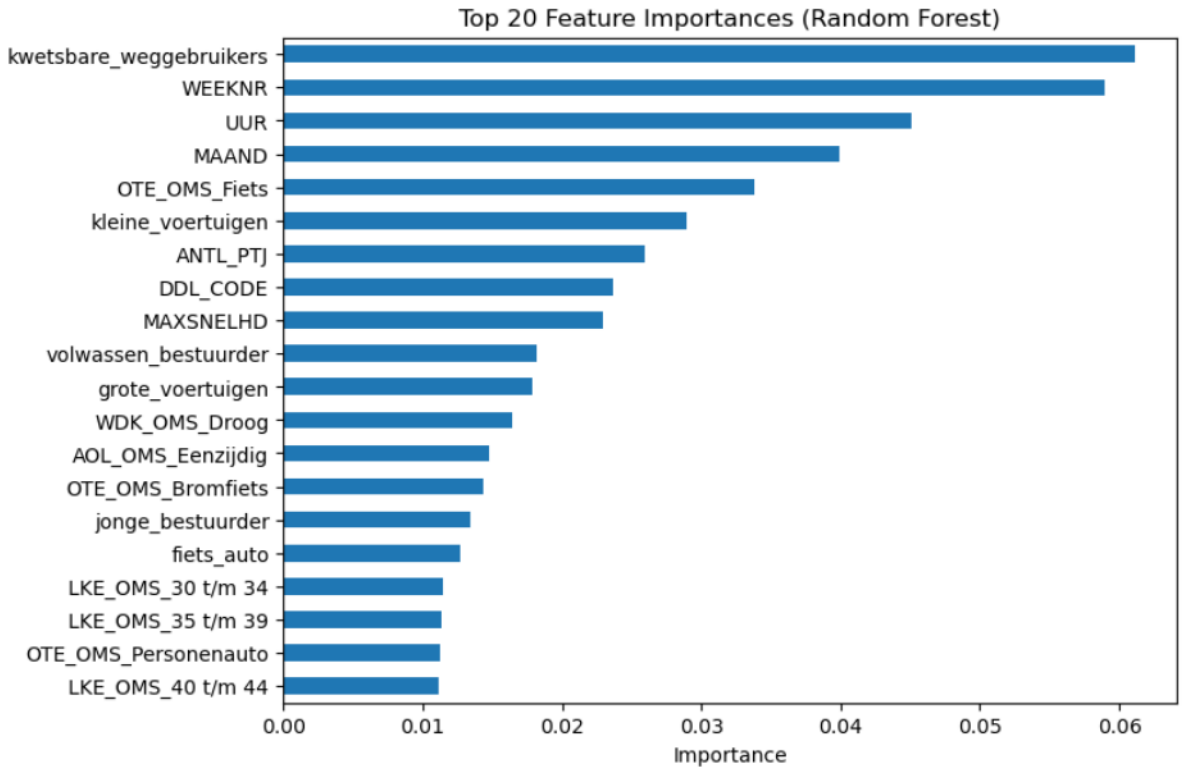


Figure 19: Feature importance scores of top 20 features using Random Forest Classifier feature selection

4 Methods

4.1 Class Imbalance

To address the severe class imbalance present in the traffic accident dataset as mentioned in section 3.3.1 (UMS ≈ 83%, LET ≈ 16%, DOD ≈0.5%), all models are trained using class weighting. Class weights compensate for underrepresented classes by increasing their contribution to the loss function. A common and model-agnostic formulation is the inverse-frequency balancing rule:

$$w_c = \frac{n_{\text{samples}}}{k \cdot n_c} \tag{5}$$

where w_c is the weight for class c, n_{samples} is the total number of instances, k is the number of classes, and n_c is the number of instances in class c [11]. This weighting scheme ensures that minority classes receive proportionally more influence during model training.
Scikit-learn classifiers (e.g. logistic regression, decision trees, random forests) implement this formula directly when *class_weight="balanced"* is specified. Gradient boosting libraries such as XGBoost, LightGBM, and CatBoost use conceptually equivalent mechanisms, either through explicit class weights or through per-sample weights. Thus, the same theoretical rationale applies across all models used in this study.

4.2 Hyperparameter tuning

Hyperparameter tuning is applied to improve predictive performance by systematically searching the hyperparameter space of a model and balancing exploration with exploitation using automated strategies. Modern surveys summarize that Bayesian, model-based, and multi-fidelity methods offer efficient alternatives to exhaustive search, particularly for complex models and large search spaces [6]. Optuna is chosen as the optimization framework for LightGBM, CatBoost and XGBoost because it provides a flexible define-by-run API, efficient samplers and integrated pruning of unpromising trials. This enables scalable optimisation on large datasets as in this research [3]. Hyperparameter tuning is not applied to the baseline MLR and DT models, to preserve their role as simple, interpretable reference models against which tuned ensemble methods are compared.

4.3 Baseline Classification Models

4.3.1 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is used as a first baseline model to predict accident severity (classified as UMS - material damage, LET - serious injury, and DOD - fatal) using the selected features from the preprocessed dataset.
Multinomial Logistic Regression models the *log-odds* of each class c relative to a reference class as a linear combination of the input features:

$$\log \frac{P(Y = c \mid X)}{P(Y = C \mid X)} = \beta_{0c} + \sum_{j=1}^{p} \beta_{jc} X_j, \tag{6}$$

where $P(Y = c \mid X)$ is the probability of class c given the feature vector X, β_{0c} is the intercept for class c, and β_{jc} are the coefficients for feature j in class c. This formulation allows the estimation of class probabilities using the softmax function [15].
Multinomial Logistic Regression is suitable as a baseline model because it is interpretable, computationally efficient, and provides probabilistic outputs for each class, enabling comparison with more complex models. The model is implemented using the `saga` solver with a maximum of 500 iterations to ensure convergence given the large dataset (approximately 2.6 million instances) and the presence of multiple numerical and binary features.

4.3.2 Decision Tree Classifier

Decision Trees (DT) partition the feature space into increasingly homogeneous subsets by recursively selecting the feature and split point that provides the highest reduction in impurity. The Gini impurity is used as the splitting criterion and is defined as:

$$G = 1 - \sum_{c=1}^{k} p_c^2, \tag{7}$$

where p_c is the proportion of instances of class c in a given node, and k is the number of classes. At each split, the feature and threshold that minimize the weighted Gini impurity of the child nodes are selected, thereby maximizing node purity [28].
Decision Trees are well suited as an interpretable baseline model, particularly in datasets such as this one, which contain a mixture of binary, categorical, and integer-valued features without requiring any form of scaling. The model is trained using the `Gini` impurity criterion and a maximum depth left unconstrained (`max_depth=None`) to obtain a fully grown tree, providing a transparent reference point before introducing more regularized or ensemble-based methods. The default settings of `min_samples_split=2` and `min_samples_leaf=1` are used to allow for a simple reference model against which more advanced and tuned models can be meaningfully compared.

4.4 Tree-Based Ensemble Classifiers

4.4.1 LightGBM

LightGBM is a gradient boosting framework that builds decision trees using a leaf-wise growth strategy with depth constraints and relies on techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce computational cost while maintaining predictive accuracy [16]. The method has been shown to achieve substantial efficiency gains by retaining data instances with large gradients during training and by compressing mutually exclusive sparse features into compact bundles without information loss. Formally, the model optimizes an additive objective function of the form

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\Big(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\Big) + \Omega(f_t) \tag{8}$$

where $\mathcal{L}^{(t)}$ is the objective at iteration t, $l(\cdot)$ is the differentiable loss function, $\hat{y}_i^{(t-1)}$ denotes the model prediction from the previous iteration, f_t is the tree added at iteration t, and $\Omega(f_t)$ is the regularization term controlling the complexity of the tree. LightGBM constructs each tree by selecting splits that maximize the reduction in this objective, but differs from traditional level-wise algorithms by growing the leaf that yields the largest decrease, allowing the model to focus on the most informative regions of the feature space.
LightGBM is well suited for this research as it efficiently handles large datasets, captures nonlinear relationships and complex feature interactions, and supports categorical features and sparse inputs. It also provides built-in handling of class imbalance, trains quickly, and uses less memory compared to XGBoost.

4.4.2 CatBoost

CatBoost is a gradient boosting framework that enhances the standard boosting paradigm with mechanisms designed to handle categorical features and to avoid target-leakage biases common in conventional implementations. The model iteratively builds an ensemble of decision trees, optimizing at boosting iteration t the additive objective as presented in equation 8 [27]. CatBoost's key innovations are ordered boosting - which randomly permutes the dataset and estimates gradients in an online fashion to prevent prediction shift - and a specialized treatment of categorical features. These allow CatBoost to efficiently handle high-cardinality categorical variables, avoid overfitting linked to naive target-statistics encoding, and maintain strong generalization performance. This makes it a suitable model for classifying accident severity with heterogeneous and partly categorical features. Often CatBoost is more robust compared to XGBoost and LightGBM in terms of hanlding class imbalance, however it is slower as well.

4.4.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that constructs an ensemble of decision trees in a stage-wise manner, optimizing an additive objective function at each iteration. The objective combines a differentiable loss function with a regularization term to control model complexity, as presented in equation 8 [9]. XGBoost differs from standard gradient boosting by incorporating second-order Taylor approximation of the loss, shrinkage, column subsampling, and a sparsity-aware split finding algorithm, enabling faster training and better handling of sparse and high-dimensional datasets.
XGBoost is well suited for modelling traffic accident severity because it efficiently captures nonlinear relationships and complex feature interactions, handles sparse data natively, and provides high predictive performance even with large datasets containing a mix of numerical and categorical features. Compared to LightGBM, the model is heavier and slower on multi-million row datasets and memory usage may be high.

4.5 Evaluation metrics

Accuracy

Accuracy expresses the proportion of correctly classified instances across all classes. In order to measure this, the following occurrences are counted:

- TP_i: number of true positives (instances of class i correctly predicted as i);
- FP_i: number of false positives (instances predicted as i, but whose true class is not i);
- FN_i: number of false negatives (instances whose true class is i, but predicted as some other class);
- TN_i: number of true negatives (instances whose true class is not i, and also not predicted as i).

Accuracy then is computed as

$$\text{Accuracy} = \frac{\sum_i (TP_i + TN_i)}{\sum_i (TP_i + FP_i + FN_i + TN_i)}. \tag{9}$$

Although accuracy provides a simple, interpretable indicator of overall model performance, in the context of severe class imbalance accuracy tends to overstate performance because a trivial classifier predicting always the majority class may reach high accuracy while failing to detect rare but critical classes. This limitation of accuracy in imbalanced settings is well documented [4].

Precision and Recall (Sensitivity)

For each class i, precision and recall are defined as

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \qquad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{10}$$

Precision quantifies the fraction of predicted positive instances for class i that are truly of class i. Recall (sensitivity) quantifies the fraction of true instances of class i that are correctly identified [30]. These two metrics capture different aspects of classification quality: precision reflects reliability of positive predictions, while recall reflects completeness of detection. Their use and interpretation are standard in classification research.

F1-score

The F1-score for class i is defined as the harmonic mean of precision and recall:

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} = \frac{2\,TP_i}{2\,TP_i + FP_i + FN_i} \tag{11}$$

Because F1-score balances between precision and recall, it is particularly useful when both false positives and false negatives matter. However, in imbalanced settings its interpretation remains sensitive to class prevalence and may still obscure poor performance on minority classes if aggregated naively [14].

Macro and Weighted Averages

In multi-class classification, aggregated summaries of per-class metrics are often employed. The macro average computes the simple arithmetic mean of the metric over all classes, giving equal weight to each class regardless of class frequency. The weighted average instead weights each class's metric by the number of true instances (support) in the dataset, thereby reflecting class imbalance. Reporting both macro and weighted averages provides a more comprehensive view: macro average shows how the model behaves per class uniformly, while weighted average indicates how it performs on the overall dataset considering class distribution [20].

Confusion Matrix

The confusion matrix - a $C \times C$ table for C classes - displays actual classes against predicted classes. Diagonal entries represent correct predictions; off-diagonal entries correspond to different types of misclassification. The confusion matrix enables detailed analysis of which classes are often confused, reveals systematic biases (e.g. minority classes being misclassified as majority class), and supports calculation of class-specific metrics (precision, recall, F1) and aggregated summaries [19]. In case of predicting three classes, Figure 20 shows which cell in the table indicates the True/False Positives and Negatives

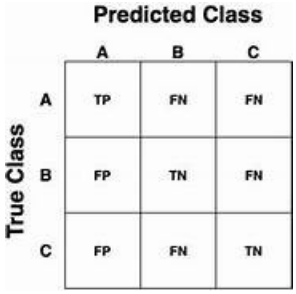


Figure 20: Confusion matrix for predicting three classes

SHAP values

SHAP values (SHapley Additive exPlanations) provide a theoretically grounded method for interpreting complex machine-learning models by quantifying the contribution of each feature to an individual prediction. The approach is based on cooperative game theory, where the predictive model is interpreted as a game and features represent players whose contributions to the final output are assessed through Shapley values. SHAP values satisfy desirable properties such as local accuracy, consistency, and symmetry, which ensures that the sum of all feature contributions equals the model output for any given instance [18].
Formally, for a model f and an input instance x with feature set N, the SHAP value ϕ_i for feature i is defined as

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \big[\, f_{S \cup \{i\}}(x) - f_S(x) \big], \tag{12}$$

where S denotes a subset of features not containing i, $f_S(x)$ is the model's output when only features in S are present (or marginalised over the remaining features), and the combinatorial coefficient ensures an average over all possible inclusion orders of feature i in the coalition. This formulation guarantees a fair attribution of the prediction's deviation from a baseline (typically the expected output) to each feature.

In the context of multi-class classification for traffic accident severity, SHAP values enable identification of how specific features influence the probability of each severity class. Beyond measuring individual feature contributions, SHAP values also facilitate the analysis of interaction effects between variables. For large datasets however, computing SHAP values is computationally expensive. Therefore, interaction values are not computed, and the individual SHAP analyses are performed on a stratified subsample of the test set to ensure computational feasibility while preserving class distributions.

5 Experimental setup

The dataset is divided into three mutually exclusive subsets: a training set, a validation set, and a test set. First, 70% of the data is assigned to the training set, while the remaining 30% is temporarily held out. This held-out portion is subsequently split into a validation set (10% of the total dataset) and a test set (20% of the total dataset). The training set is used to fit the model parameters, while the validation set serves to tune hyperparameters and apply early stopping in order to prevent overfitting. The test set is kept completely separate and is used only once to assess the final predictive performance of each model. Because the dataset contains a strong class imbalance, with severe and fatal accidents occurring much less frequently than material-damage-only accidents, all splits are performed in a stratified manner to preserve the original class distributions across subsets. Given the large size of the dataset, this approach provides stable and reliable performance estimates while avoiding the substantial computational cost associated with cross-validation. Class imbalance is further addressed by applying instance-level weighting during model training, ensuring that minority classes contributes proportionally to the learning process.

LightGBM

For each gradient boosting model, 40 optimization trials are conducted, balancing search efficiency with computational feasibility given the large dataset size. During each trial, models are trained on the training set and evaluated on the validation set, and early stopping is applied at 50 to prevent overfitting and reduce unnecessary training iterations.

The hyperparameters selected for tuning primarily control model complexity and regularization, including the learning rate, number of boosting iterations, tree depth, number of leaves, minimum number of samples per leaf, subsampling ratios for observations and features, and L1 and L2 regularization terms. These hyperparameters are chosen because they have a strong influence on the bias–variance trade-off and are particularly important when modeling complex, non-linear relationships in large and imbalanced datasets [37]. Ranges are chosen based on common best practices for gradient boosting models designed to balance model expressiveness and regularization. These tuning settings together with the obtained optimal hyperparameter values found by Optuna are provided in Table 4. These values will be used to build the model.

Hyperparameter	Tuning settings	Optimal values
learning_rate	[0.01, 0.2]	0.0145
num_leaves	[31, 256]	255
max_depths	[-1, 16]	16
min_child_samples	[20, 200]	28
subsample	[0.6, 1]	0.673
colsample_bytree	[0.6, 1]	0.745
L1	[0, 5]	0.485
L2	[0, 5]	3.93
n_estimators	[300, 2000]	1039

Table 4: Hyperparameter tuning settings and optimal searched values for LightGBM model

CatBoost

The hyperparameters selected for optimization are learning rate, tree depth, L2 regularization, bagging temperature (control for randomness in sampling), random strength (adding noise to split selection) and number of trees in order to improve generalization and model complexity. The ranges of these hyperparameters are chosen so that they are wide enough to explore but narrow enough to remain computationally feasible, with lower bounds reflecting conservative defaults and upper bounds practical limits for large datasets. The tuning settings as well as the optimal obtained hyperparameters by Optuna for the CatBoost model can be found in Table 5.
Due to CatBoost being a computationally expensive model, the stratified sample for which the SHAP values are computed is 10 times as small as the samples for LightGBM and XGBoost, in order for the kernel not to die during computation. Also, Pool is used instead of TreeExplainer. This might influence the results of the CatBoost SHAP values and comparison to the other models.

Hyperparameter	Tuning settings	Optimal values
learning_rate	[0.01, 0.2]	0.0789
depth	[4, 10]	10
l2_leaf_reg	[4, 10]	6.30
bagging_temperature	[0.0, 1.0]	0.955
random_strength	[0.0, 1.0]	0.159
iterations	[300, 2000]	748

Table 5: Hyperparameter tuning settings and optimal searched values for CatBoost model

XGBoost

The hyperparameters that are tuned for XGBoost include the learning rate, maximum tree depth, minimum instance weight in a leaf, the subsample, feature subsampling, minimum loss reduction, L1 and L2 regularization and the number of boosting rounds. These parameters ensure increasing model complexity and reducing overfitting. Both the Optuna settings and the searched optimal values are provided in Table 6.

Hyperparameter	Tuning settings	Optimal values
learning_rate	[0.02, 0.15]	
max_depth	[5, 10]	
min_child_weight	[1.0, 10.0]	
subsample	[0.6, 1.0]	
colsample_bytree	[0.6, 1.0]	
gamma	[0.0, 5.0]	
reg_alpha	[0.0, 5.0]	
reg_lambda	[1.0, 10.0]	
n_estimators	[300,2000]	

Table 6: Hyperparameter tuning settings and optimal searched values for XGBoost model

6 Results

6.1 Multinomial Logistic Regression

The classification report in Table 7 shows that the model performs well for the majority class of material accidents (UMS), achieving high precision (0.94) and recall (0.80). This indicates that most material-damage accidents are correctly predicted, although a notable number are misclassified as LET or DOD, as confirmed by the confusion matrix in Table 8.
For accidents resulting in serious injury (LET), the model achieves moderate performance, with a precision of 0.43 and recall of 0.60. This suggests that while a majority of LET cases are correctly identified, there is significant confusion with the majority UMS class: 21,395 of the true LET accidents are predicted as UMS, which represents roughly 26% of all LET cases. Additionally, 11,287 LET cases are misclassified as DOD, indicating that the model occasionally overestimates the severity for these accidents.
Performance for fatal accidents (DOD) remains poor, reflecting the extreme rarity of this class. Precision is only 0.03, meaning that the vast majority of predictions labeled as DOD are false positives, while recall is 0.33, indicating that only about one-third of actual fatal accidents are correctly identified. The confusion matrix shows that 990 fatal accidents are predicted as LET and 709 as UMS, emphasizing the tendency of the model to misclassify rare fatal events as less severe outcomes.

Overall, the confusion matrix highlights systematic patterns in misclassification; the model reliably identifies UMS accidents, moderately identifies LET accidents, and struggles with DOD. Most misclassifications occur from minority classes (LET and DOD) into the majority class (UMS), which is typical for models trained on highly imbalanced datasets. This explains the discrepancy between high overall accuracy (0.77) and relatively low macro-averaged F1-score (0.47), which reflects poor minority-class performance, while the weighted F1-score (0.80) emphasizes strong performance on the prevalent UMS class.

Class	**Precision**	**Recall**	**F1-score**	**Support**
DOD	0.03	0.33	0.05	2552
LET	0.43	0.60	0.50	82527
UMS	0.94	0.80	0.86	433990
Accuracy			0.77	519069
Macro average	0.46	0.58	0.47	519069
Weighted average	0.85	0.77	0.80	519069

Table 7: Classification report of Multinomial Logistic Regression model

Predicted/True	DOD	LET	UMS
DOD	853	990	709
LET	11287	49845	21395
UMS	20516	65983	347491

Table 8: Confusion matrix of Multinomial Logistic Regression model

6.2 Decision Tree Classifier

Table 9 shows that the Decision Tree achieves an overall accuracy of 0.81 on the test set. Performance varies considerably across the three accident severity classes. For the majority class UMS, the model performs very well, with a recall of 0.90, precision of 0.88, and F1-score of 0.89, indicating that most material-damage accidents are correctly identified. The confusion matrix in Table 10 confirms this, as 389,273 of 433,990 UMS cases are correctly predicted, while the majority of remaining misclassifications are assigned to the LET class.
For accidents resulting in serious injury (LET), the model shows moderate performance, achieving a precision of 0.41 and recall of 0.38. The confusion matrix shows that 50,342 LET cases are misclassified as UMS, and 932 as DOD, while 31,253 are correctly predicted. This pattern indicates that the model tends to underpredict LET cases, misclassifying many as the more prevalent UMS class, which is consistent with the challenges of distinguishing minority classes in imbalanced datasets.
Performance on fatal accidents (DOD) remains extremely limited. Only 127 of 2,552 fatal accidents are correctly predicted, resulting in a precision and recall of 0.05 each and an F1-score of 0.05. Most fatal cases are misclassified as LET (905) or UMS (1,520), highlighting the difficulty of capturing rare events with a simple decision tree, even with full depth and minimal sample restrictions.

Overall, the confusion matrix reveals that the model reliably identifies the majority class while struggling with minority classes, particularly DOD. Most misclassifications occur from DOD and LET into UMS, reflecting both class imbalance and the limited discriminative power of a single-tree model. The macro F1-score of 0.45 illustrates low performance when averaging across classes, while the weighted F1-score of 0.81 emphasizes strong performance on the dominant UMS class. These results position the Decision Tree as a more flexible baseline than MLR, with slight improvement in overall accuracy and material-accident detection, but similar limitations for rare fatal outcomes.

Class	**Precision**	**Recall**	**F1-score**	**Support**
DOD	0.05	0.05	0.05	2552
LET	0.41	0.38	0.40	82527
UMS	0.88	0.90	0.89	433990
Accuracy			0.81	519069
Macro average	0.45	0.44	0.45	519069
Weighted average	0.80	0.81	0.81	519069

Table 9: Classification report of Decision Tree Classifier

Predicted/True	DOD	LET	UMS
DOD	127	905	1520
LET	932	31253	50342
UMS	1431	43286	389273

Table 10: Confusion matrix of Decision Tree Classifier

6.3 LightGBM

The classification results of the LightGBM model indicate heterogeneous performance across accident severity categories. Table 11 shows that the model achieves high precision (0.96) and F1-score (0.86) for the majority class of material-damage accidents (UMS), reflecting strong performance on the prevalent category. Recall for UMS is comparatively lower at 0.78, indicating that a notable portion of material-damage accidents is misclassified as more severe outcomes, either LET or DOD. The confusion matrix in Table 12 confirms this pattern, with 69,674 UMS cases predicted as LET and 326 as DOD, while the majority (337,715) are correctly classified.
For injury-related accidents (LET), the model attains a recall of 0.67 and an F1-score of 0.53, suggesting that a substantial proportion of LET cases is correctly identified, although precision remains moderate at 0.44. The confusion matrix shows that 15,224 LET cases are misclassified as UMS and 12,100 as DOD, indicating a tendency of the model to underpredict and overpredict severity depending on feature patterns.
Performance for fatal accidents (DOD) differs markedly from the other classes. The model achieves a high recall of 0.53, indicating that more than half of fatal accidents are correctly identified. However, precision is extremely low at 0.034, meaning that many non-fatal accidents are incorrectly predicted as DOD. This results in a large number of false positives for the DOD class, as visible in the confusion matrix where 874 LET and 326 UMS cases are misclassified as fatal accidents.

Overall, the macro F1-score of 0.48 highlights the impact of minority-class performance on aggregated metrics, while the weighted F1-score of 0.80 emphasizes strong predictive ability on the dominant UMS class. Misclassifications are most frequent between UMS and LET, reflecting inherent overlap in accident characteristics between these categories. Despite challenges in predicting the rare DOD class, LightGBM shows a strong balance of precision and recall for LET and UMS, demonstrating its ability to better capture non-linear relationships and interactions in the data than simpler models.

Class	**Precision**	**Recall**	**F1-score**	**Support**
DOD	0.034	0.53	0.064	2552
LET	0.44	0.67	0.53	82527
UMS	0.96	0.78	0.86	433990
Accuracy			0.76	519069
Macro average	0.47	0.66	0.48	519069
Weighted average	0.87	0.76	0.80	519069

Table 11: Classification report of LightGBM

Predicted/True	DOD	LET	UMS
DOD	1352	874	326
LET	12100	55203	15224
UMS	266010	69674	337715

Table 12: Confusion matrix of LightGBM

Global feature importance across all classes are obtained by aggregating SHAP values over the class dimension, Figure 21 displays the 20 most influential features of the LightGBM model ranked from top to bottom. In Appendix A, the class-specific SHAP bar and beeswarm plots are presented, which are used to analyze feature effects for each accident severity category individually.
It can be observed that the number of involved parties in an accident (ANTL_PTJ) is outstandingly the most significant predictor of accident severity, with a global feature importance of more than 4.0. In Figures 28, 29, and 30, this feature consistently exhibits markedly higher SHAP values than all other features. For fatal accidents, the SHAP value of nearly 2.00 is the highest among the three target classes. The corresponding

beeswarm plot in Figure 31 clearly indicates that accidents involving a larger number of parties are less likely to be predicted as fatal, whereas accidents involving few or single parties contribute strongly to fatal outcome predictions. The beeswarm plots for lethal outcomes (Figure 32) and material damage only (Figure 33) show the inverse pattern, suggesting that accidents involving more parties are more likely to result in less severe outcomes.
Additional noteworthy observations that can be drawn from the beeswarm plots are that accidents involving cyclists (OTE_OMS_Fiets), pedestrians (OTE_OMS_Voetganger), and small vehicles (kleine_voertuigen) tend to decrease the predicted probability of a fatal outcome while increasing the likelihood of a lethal outcome. The involvement of mature drivers (volwassen_bestuurder) decreases the probability of fatalities, and increases that of injury or material damage. Accidents occurring in the province Zuid-Holland (PVE_CODE_Zuid-Holland) are associated with a higher probability of resulting in injuries, whereas accidents in Gelderland (PVE_CODE_Gelderland) show an increased probability of resulting in death. Accidents involving seniors (senior) increase the predicted probability of material damage, as do accidents occurring in areas with higher speed limits (MAXSNELHD). Furthermore, accidents involving seniors or children (kind) increase the predicted probability of injury, as do one-sided accidents (AOL_OMS_Eenzijdig) and accidents involving a scooter (OTE_OMS_Bromfiets). Accidents involving passenger cars (OTE_OMS_Personenauto) or large vehicles (grote_voertuigen) decrease the likelihood of material damage. Finally, among the temporal features, the only notable observation is that accidents occurring in the first few months of the year (winter) decrease the predicted probability of resulting in material damage.

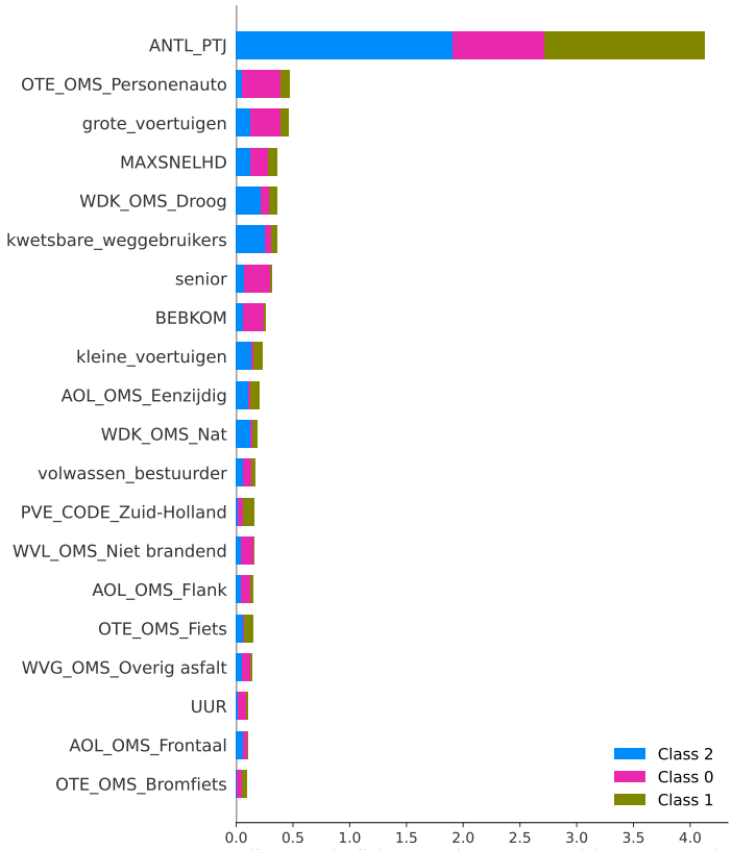


Figure 21: Class-specific SHAP global importance values of LightGBM model (Class 0 = UMS, Class 1 = LET, Class 2 = DOD)

6.4 CatBoost

The classification report in Table 13 and confusion matrix in Table 14 of the CatBoost model shows strong overall performance driven primarily by its ability to correctly classify material damage only accidents (UMS), which represent the majority class. This is reflected in a high precision of 0.95, a recall of 0.78, and an F1-score of 0.86 for UMS, as well as a substantial concentration of correctly classified instances along the diagonal of the confusion matrix. A considerable number of UMS cases are nevertheless misclassified as LET, indicating overlap in the feature space between material and injury-related accidents.
For accidents resulting in serious injury (LET), the model achieves a recall of 0.70, which indicates that a large proportion of these cases are successfully identified. Precision for LET remains moderate at 0.44, suggesting that a non-negligible share of predicted LET cases actually belong to other severity categories, primarily UMS. This trade-off results in an F1-score of 0.54, highlighting reasonable but imperfect discrimination for this intermediate severity class.
Performance for fatal accidents (DOD) remains limited. Although the recall of 0.44 indicates that the model identifies a notable fraction of fatal cases relative to their rarity, precision is very low at 0.04, meaning that most instances predicted as fatal correspond to non-fatal outcomes. This pattern is also visible in the confusion matrix, where a large share of fatal predictions are assigned to UMS or LET cases.

Overall accuracy reaches 0.77, largely influenced by the dominance of the UMS class. The macro-averaged F1-score of 0.49 highlights the imbalance in performance across classes, as it weights all classes equally and therefore reflects the comparatively weak performance on the minority DOD class. In contrast, the weighted F1-score of 0.81 emphasizes the model's strong performance on the majority class.

Class	**Precision**	**Recall**	**F1-score**	**Support**
DOD	0.038	0.44	0.070	2552
LET	0.44	0.70	0.54	82527
UMS	0.95	0.78	0.86	433990
Accuracy			0.77	519069
Macro average	0.48	0.64	0.49	519069
Weighted average	0.87	0.77	0.81	519069

Table 13: Classification report of CatBoost

Predicted/True	DOD	LET	UMS
DOD	1124	1060	368
LET	8509	58125	15893
UMS	19998	74033	339959

Table 14: Confusion matrix of CatBoost

Figure 22 shows the global SHAP bar plot for the CatBoost model. This plot displays the overall importance of each feature for the model's predictions across the entire dataset, aggregated over all accident severity classes. Each bar represents the mean absolute contribution of a feature to the model output, regardless of the predicted class, providing a general overview of which features the model relies on most to predict accident severity.
Notably, the presence of vulnerable road users (kwetsbare_weggebruikers) and seniors above 65 years of age (senior) show stronger predictive influence compared to the LightGBM model. One-sided accidents (AOL_OMS_Eenzijdig) also rank higher in importance. The number of involved parties (ANTL_PTJ) remains the highest scoring feature. Other features with consistently high importance, similar to LightGBM, include dry road surface (WDK_OMS_Droog), involvement of cars (OTE_OMS_Personenauto) or large vehicles (grote_voertuigen), and speed limit (MAXSNELHD).

In contrast to LightGBM and XGBoost, CatBoost employs shared symmetric trees across classes, which prevents direct extraction of class-specific SHAP values. Therefore, class-conditional explanations are derived by grouping SHAP values according to the model's predicted class labels. These values indicate feature contributions specifically for samples predicted as a given class. Figures 34, 35, and 36 in Appendix A show these class-conditional bar plots. Overall, the feature rankings are similar across classes, whereas the class-specific SHAP values bar plots for LightGBM show more distinction between classes.
The involvement of vulnerable road users scores highly across all classes, with SHAP values around 0.35, occupying first or second place. This is notable compared to LightGBM, where vulnerable road users are less influential for the LET and UMS classes. Similarly, involvement of seniors has SHAP values around 0.25 and ranks third or fourth for each class in CatBoost, while for LightGBM this feature is primarily relevant for the UMS class. Minor distinctions between classes can be observed: for DOD predictions, speed limit (MAXSNELHD) is influential, while for LET predictions, the involvement of small vehicles (kleine_voertuigen) plays a stronger role.
Class-conditional SHAP beeswarm plots further illustrate feature influence within each predicted outcome. For accidents predicted as fatal, high values of large vehicle or car involvement and accidents occurring in Gelderland are associated with higher SHAP values, indicating greater contribution to predicted fatal severity. For lethal accidents, the presence of cars, large vehicles, scooters, or bike-car collisions strongly drives predictions of injury outcomes. For material-damage-only predictions, high speed limits, involvement of cars or large vehicles, and accidents occurring during rush hour or on straight roads have the strongest influence on model predictions.

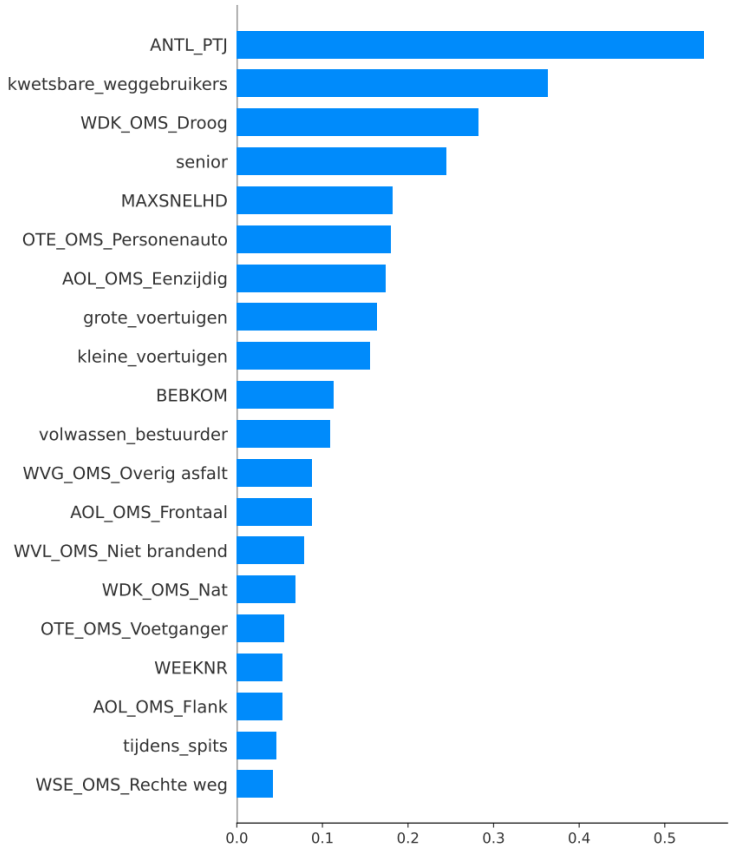


Figure 22: Global SHAP importance values of CatBoost model

6.5 XGBoost

The XGBoost model achieves the highest overall performance among the evaluated models in terms of accuracy and aggregated F1-scores, presented in Table 15. The confusion matrix in Table 16 shows a strong concentration of correct predictions along the diagonal for material damage only accidents (UMS), with a recall of 0.81 and a precision of 0.95, indicating that the majority class is identified reliably. Misclassifications of UMS cases occur primarily as LET, suggesting overlap between material and injury-related accidents, while confusion with fatal outcomes remains limited.
For accidents resulting in injury (LET), the model attains a recall of 0.75, which is higher than for the baseline and other ensemble models, indicating improved identification of injury-related accidents. Precision for LET remains moderate at 0.44, combined with the recall resulting in an F1-score of 0.55, which is the highest observed for the LET class across the evaluated models.
Performance on fatal accidents (DOD) remains challenging, with a relatively high precision at 0.048, but recall of 0.19 is remarkably low, indicating cautious predictions of the model, only predicting a positive class when highly confident, resulting in very few false positives but missing many actual positive cases. The confusion matrix shows that most fatal accidents are misclassified as UMS or LET, which is consistent with the extreme class imbalance. Nevertheless,the F1-score of 0.08 represents a modest improvement compared to other models.

Overall accuracy reaches 0.79, driven largely by strong performance on the dominant UMS class. The macro F1-score of 0.50 highlights the imbalance in class-wise performance but exceeds that of the other models, indicating improved average performance across classes. The weighted F1-score of 0.82 further reflects strong predictive performance when accounting for class prevalence, positioning XGBoost as the best-performing model in terms of aggregate evaluation metrics.

Class	**Precision**	**Recall**	**F1-score**	**Support**
DOD	0.048	0.19	0.077	2552
LET	0.44	0.75	0.55	82527
UMS	0.95	0.81	0.87	433990
Accuracy			0.79	519069
Macro average	0.48	0.58	0.50	519069
Weighted average	0.87	0.79	0.82	519069

Table 15: Classification report of XGBoost

Predicted/True	DOD	LET	UMS
DOD	474	1537	541
LET	2919	62123	17485
UMS	6409	78113	349468

Table 16: Confusion matrix of XGBoost

Figure 23 presents the 20 most influential features of the XGBoost model based on their mean SHAP values. As for the other models, the number of involved parties emerges as the most influential feature, however more strongly driven by the UMS class in contrast to the DOD class for the LightGBM model. The speed limit (MAXSNELHD) ranks second in importance, whereas it appears lower for LightGBM and CatBoost. Other features with consistently high influence include involvement of large vehicles (grote_voertuigen), vulnerable road users (kwetsbare_weggebruikers), elderly (senior), and passenger cars (OTE_OMS_Personenauto).
Temporal features contribute more strongly to the predictive probabilities in the XGBoost model than in the other boosting models. The week in which the accident occurred (WEEKNR) appears prominently among the most influential features, whereas it ranks only 17th in the CatBoost results and not at all for

LightGBM. A similar pattern is observed for the hour of the accident (UUR), which appears only in 18th position for LightGBM and not for CatBoost. Additionally, XGBoost is the only model in which the month of occurrence (MAAND) appears among the most influential features.

Examining the class-specific SHAP values, Figure 40 for the DOD class shows that, besides the number of involved parties, notably, the week number exhibits high mean SHAP importance of nearly 0.4, followed by the speed limit. The beeswarm plot in Figure 43 indicates that higher speed limits increase the predicted probability of a fatal outcome, whereas the beeswarm plot for the LET class in Figure 44 shows that lower speed limits increase the predictive probability of lethal injury. Furthermore, the involvement of seniors, vulnerable road users, and mature drivers (volwassen_bestuurder) increases the predicted probability of fatal accidents, while the involvement of passenger cars decreases it.
For the lethal injury class, the SHAP bar plot in Figure 41, together with the corresponding beeswarm plot in Figure 44, indicates that one-sided accidents (AOL_OMS_Eenzijdig), accidents occurring in the province of Zuid-Holland (PVE_CODE_Zuid-Holland), and bicycle-related accidents (OTE_OMS_Fiets) increase the probability of a lethal outcome. This observation aligns with earlier findings reported in published research on one-sided bicycle accidents in The Hague [**....fietser**]. In addition, the involvement of children (kind), scooters (OTE_OMS_Bromfiets), vulnerable road users, and accidents occurring in the province of Noord-Holland (PVE_CODE_Noord-Holland) are associated with an increased predicted probability of lethal injury.
For material-damage-only accidents, Figure 42 shows that, besides the number of involved parties, the involvement of vulnerable road users and small vehicles (kleine_voertuigen) influences model predictions. The beeswarm plot in Figure 45 indicates that most of these features decrease the predicted probability of the UMS class, while no features exhibit a strong positive contribution to increasing the likelihood of material-damage-only outcomes.

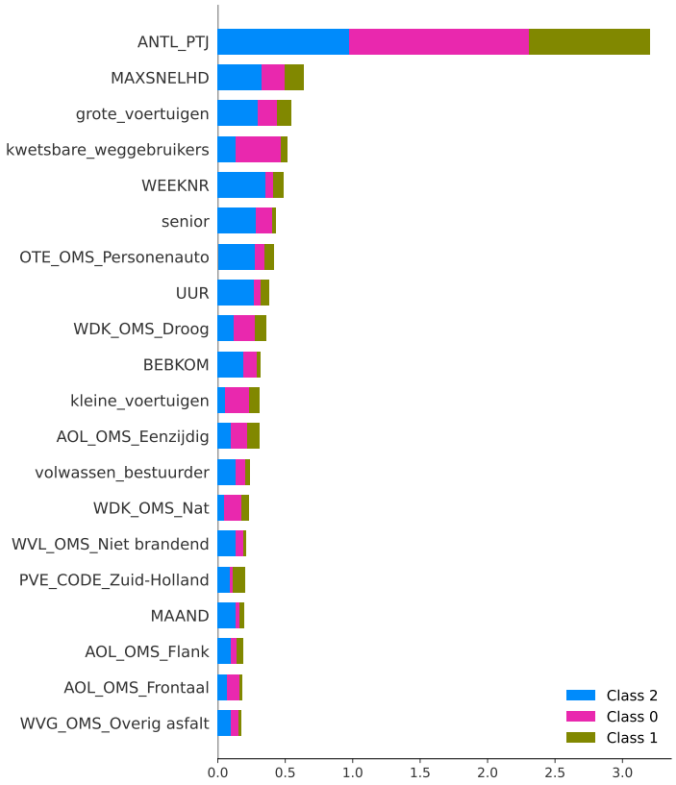


Figure 23: Class-specific SHAP global importance values of XGBoost model (Class 0 = UMS, Class 1 = LET, Class 2 = DOD)

7 Conclusion and Discussion

This thesis investigated to what extent the severity of traffic accidents in the Netherlands can be predicted using supervised classification models based on locational, temporal, and situational characteristics. Accident severity was defined as a three-class outcome, distinguishing material-damage-only accidents (UMS), accidents with serious or lethal injuries (LET), and fatal accidents (DOD).

The results show that ensemble models based on gradient boosting, such as LightGBM, CatBoost, and XGBoost, perform clearly better than simpler baseline models like Multinomial Logistic Regression and single Decision Trees, especially in the presence of extreme class imbalance. While all evaluated models achieve strong performance for the majority class of material-damage-only accidents, the ensemble models are markedly better at identifying severe outcomes. In particular, they attain meaningful recall for lethal accidents and, to a more limited extent, fatal accidents. Although precision for these rare classes remains low, this outcome is expected in highly imbalanced settings and may be acceptable when the goal is to identify high-risk situations rather than to make perfectly accurate classifications.

Among the ensemble approaches, XGBoost delivers the most balanced overall performance, while LightGBM and CatBoost achieve slightly higher recall for the most severe accident categories. Interpretation of the models using SHAP analysis highlights several important factors that consistently influence predictions. The number of involved parties stands out as the most influential feature, with single-party accidents being more strongly associated with fatal outcomes. Other important factors include speed limits, the involvement of vulnerable road users and seniors, and vehicle type. Temporal factors such as the hour of the day, week number, and month generally have smaller effects, but they play a more noticeable role in the XGBoost model, indicating differences in how models capture temporal patterns. Overall, the results show that interpretable ensemble models can combine strong predictive performance with useful insights into the factors that contribute to accident severity.

Some limitations of this study should be noted. First, the analysis is based solely on accident records and does not include traffic exposure information, such as traffic volumes, which restricts causal interpretation. But most importantly, accurately predicting fatal accidents, as well as lethal accidents to a lesser extend, remains very difficult due to their extreme rarity. Future research could address these issues by using spatio-temporal modeling approaches, or focusing on smaller, more targeted datasets to explore relationships between factors in greater detail and enable the estimation of SHAP interaction effects. These extensions could further improve both predictive performance and understanding of the many and complex factors that lead to severe traffic accidents.

8 Published articles

November 4th 2025: November is de maand met de meeste ongevallen, vooral in de avondspits tussen 17.00 en 18.00 uur [24]

deVolkskrant

ımns Opinie Cartoons Podcasts Cultuur

November is de maand met de meeste ongevallen, vooral in de avondspits tussen 17.00 en 18.00 uur

Het gevaarlijkste uur op de Nederlandse wegen staat weer voor de deur. Historisch gezien vinden de meeste ongelukken plaats op novemberdagen tussen 17.00 en 18.00 uur, zo blijkt uit een analyse van *de Volkskrant* op basis van data van Rijkswaterstaat.

De avondspits op een novemberavond. In geen maand ligt het aantal ongelukken zo hoog als in november. Arie Kievit / de Volkskrant

Jade Paus 4 november 2025, 05:00

Figure 24: "Cijfers bij het nieuws" rubric

February 13th 2026: Nederland fietsland? In het verkeer schuilt steeds meer gevaar [24]

Figure 25: Front page news

Figure 26: Most read article of the day

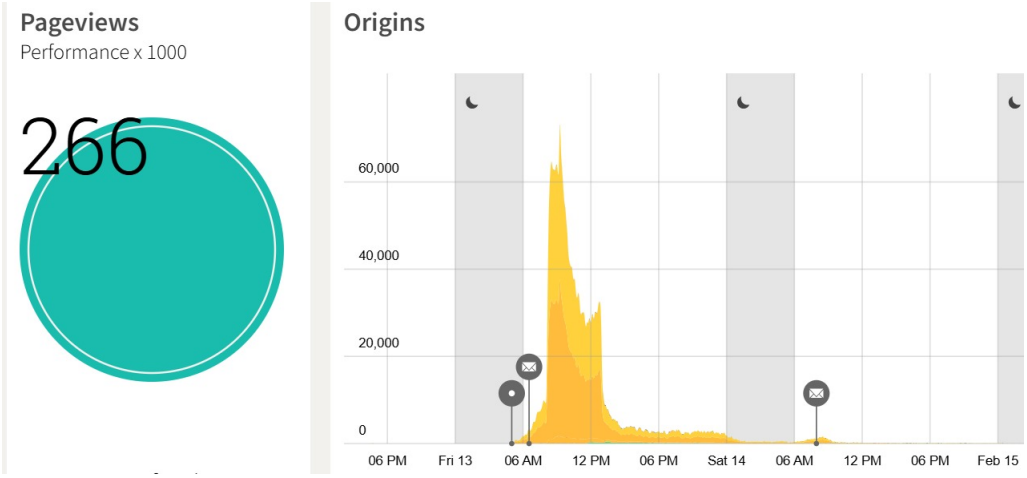


Figure 27: Around 250.000 readers on the day of publishing

References

[1] Alan Agresti. *Categorical Data Analysis*. 2nd. Comprehensive treatment of chi-square tests and interaction analysis for categorical data. Wiley-Interscience, 2002. ISBN: 9780471360920.

[2] Shakil Ahmed et al. "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance". In: *Transportation Research Interdisciplinary Perspectives* 19 (2023), p. 100814. ISSN: 2590-1982. DOI: https://doi.org/10.1016/j.trip.2023.100814. URL: https://www.sciencedirect.com/science/article/pii/S2590198223000611.

[3] Takuya Akiba et al. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701.

[4] Josephine S. Akosa. "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data". In: *Proceedings of the SAS Global Forum 2017* (2017).

[5] Enes Bakış et al. "Prediction of traffic accidents trend with learning methods: a case study for Batman, Turkey". In: *Scientific Reports* 15 (2025), p. 28906. DOI: 10.1038/s41598-025-11835-9.

[6] Bernd Bischl et al. "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges". In: *WIREs Data Mining and Knowledge Discovery* 13.6 (2023), e1484. DOI: 10.1002/widm.1484.

[7] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[8] CBS. *Wijk- en buurtkaart 2025*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2025. Accessed: 2025-09-29. 2025.

[9] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

[10] Harald Cramér. *Mathematical Methods of Statistics*. 1st. Princeton, NJ: Princeton University Press, 1946. ISBN: 9780691051570.

[11] Yin Cui et al. "Class-Balanced Loss Based on Effective Number of Samples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9268–9277. DOI: 10.1109/CVPR.2019.00949.

[12] Sheng Dong et al. "Predicting and Analyzing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models with SHAPley Additive exPlanations". In: *International Journal of Environmental Research and Public Health* 19.5 (2022), p. 2925. DOI: 10.3390/ijerph19052925. URL: https://doi.org/10.3390/ijerph19052925.

[13] European Commission — ERSO (European Road Safety Observatory). *National Road Safety Profile: The Netherlands*. Tech. rep. European Commission, 2023. URL: https://road-safety.transport.ec.europa.eu/system/files/2023-02/erso-country-overview-2023-netherlands_0.pdf.

[14] David J. Hand. "F*: an interpretable transformation of the F-measure". In: *Machine Learning* 110.3 (2021), pp. 569–596.

[15] E. M. Hashimoto et al. "The multinomial logistic regression model for predicting the discharge status after liver transplantation: estimation and diagnostics analysis". In: *Journal of Applied Statistics* 47.12 (2019), pp. 2159–2177. DOI: 10.1080/02664763.2019.1706725.

[16] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems* (2017).

[17] Wenjing Li and Zihao Luo. "Research on Traffic Accident Risk Prediction Method Based on Spatial and Visual Semantics". In: *ISPRS International Journal of Geo-Information* 12 (Dec. 2023), p. 496. DOI: 10.3390/ijgi12120496.

[18] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1 (2020), pp. 252–259. DOI: 10.1038/s42256-019-0138-9.

[19] Amalia Luque et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix". In: *Pattern Recognition* 91 (2019), pp. 216–231. DOI: 10.1016/j.patcog.2019.02.023.

[20] Johan Braet Maria Cristina Hinojosa Lee and Johan Springael. "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores". In: *Special Issue Affective Computing: Technology and Application* (2024).

[21] Mary L. McHugh. "The Chi-square test of independence". In: *Biochemia Medica* 23.2 (2013), pp. 143–149. DOI: 10.11613/BM.2013.018.

[22] Habibollah Nassiri, P. Najaf, and Amir Amiri. "Prediction of roadway accident frequencies: Count regressions versus machine learning models". In: *Scientia Iranica* 21 (2014), pp. 263–275.

[23] Amir Bahador Parsa et al. "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis". In: *Accident Analysis & Prevention* 136 (2020), p. 105405. DOI: 10.1016/j.aap.2019.105405.

[24] Jade Paus. "November is de maand met de meeste ongevallen, vooral in de avondspits tussen 17.00 en 18.00 uur". In: *de Volkskrant* (2025). URL: https://www.volkskrant.nl/binnenland/november-is-de-maand-met-de-meeste-ongevallen-vooral-in-de-avondspits-tussen-17-00-en-18-00-uur~b002762e/.

[25] Wegman F van Petegem JW. "Analyzing road design risk factors for run-off-road crashes in The Netherlands with crash prediction models". In: *J Safety Res* (2014). DOI: 10.1016/j.jsr.2014.03.003.

[26] Saeid Pourroostaei Ardakani et al. "Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis". In: *Sustainability* 15 (2023). DOI: 10.3390/su15075939.

[27] Liudmila Prokhorenkova et al. "CatBoost: unbiased boosting with categorical features". In: *Advances in Neural Information Processing Systems*. 2018.

[28] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (1986), pp. 81–106.

[29] Md Adilur Rahim and Hany M. Hassan. "A deep learning based traffic crash severity prediction framework". In: *Accident Analysis & Prevention* 154 (2021), p. 106090. DOI: 10.1016/j.aap.2021.106090.

[30] O. Rainio et al. "Evaluation metrics and statistical tests for machine learning". In: *PMC (open access)* (2024).

[31] Rijkswaterstaat. *Verkeersongevallen - Bestand geRegistreerde Ongevallen Nederland*. Accessed: 2025-09-16. URL: https://data.overheid.nl/dataset/a516ffaf-fbcc-44bc-88cb-fca799c5cd29.

[32] Daniel Santos and coauthors. "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction". In: *Computers* 10.12 (2021), p. 157. DOI: 10.3390/computers10120157.

[33] Sachita Shahi, Mark Brussel, and Anna Beatriz Grigolon. "Spatial analysis of road traffic crashes and user-based assessment of road safety: A case study of Rotterdam". In: *Traffic Injury Prevention* (2023). DOI: 10.1080/15389588.2023.2234530.

[34] Lars Skaug et al. "Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies". In: *Applied Sciences* 15.13 (2025). ISSN: 2076-3417. DOI: 10.3390/app15137115. URL: https://www.mdpi.com/2076-3417/15/13/7115.

[35] Xiao Wen et al. "Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP". In: *Accident Analysis & Prevention* 159 (2021), p. 106261. DOI: 10.1016/j.aap.2021.106261.

[36] World Health Organization. *Global status report on road safety 2023*. Tech. rep. Geneva: World Health Organization, 2023. URL: https://assets.bbhub.io/dotorg/sites/64/2023/12/WHO-Global-status-report-on-road-safety-2023.pdf.

[37] Ling Yang et al. "A Comparative Machine Learning Study Identifies Light Gradient Boosting Machine (LightGBM) as the Optimal Model for Unveiling the Environmental Drivers of Yellowfin Tuna (*Thunnus albacares*) Distribution Using SHapley Additive exPlanations (SHAP) Analysis". In: *Biology* 14.11 (2025), p. 1567. DOI: 10.3390/biology14111567.

[38] Lai Zheng and Xianghai Meng. "An approach to predict road accident frequencies: Application of fuzzy neural network". In: *3rd International Conference on Road Safety and Simulation* (2011).

A Appendix

LightGBM

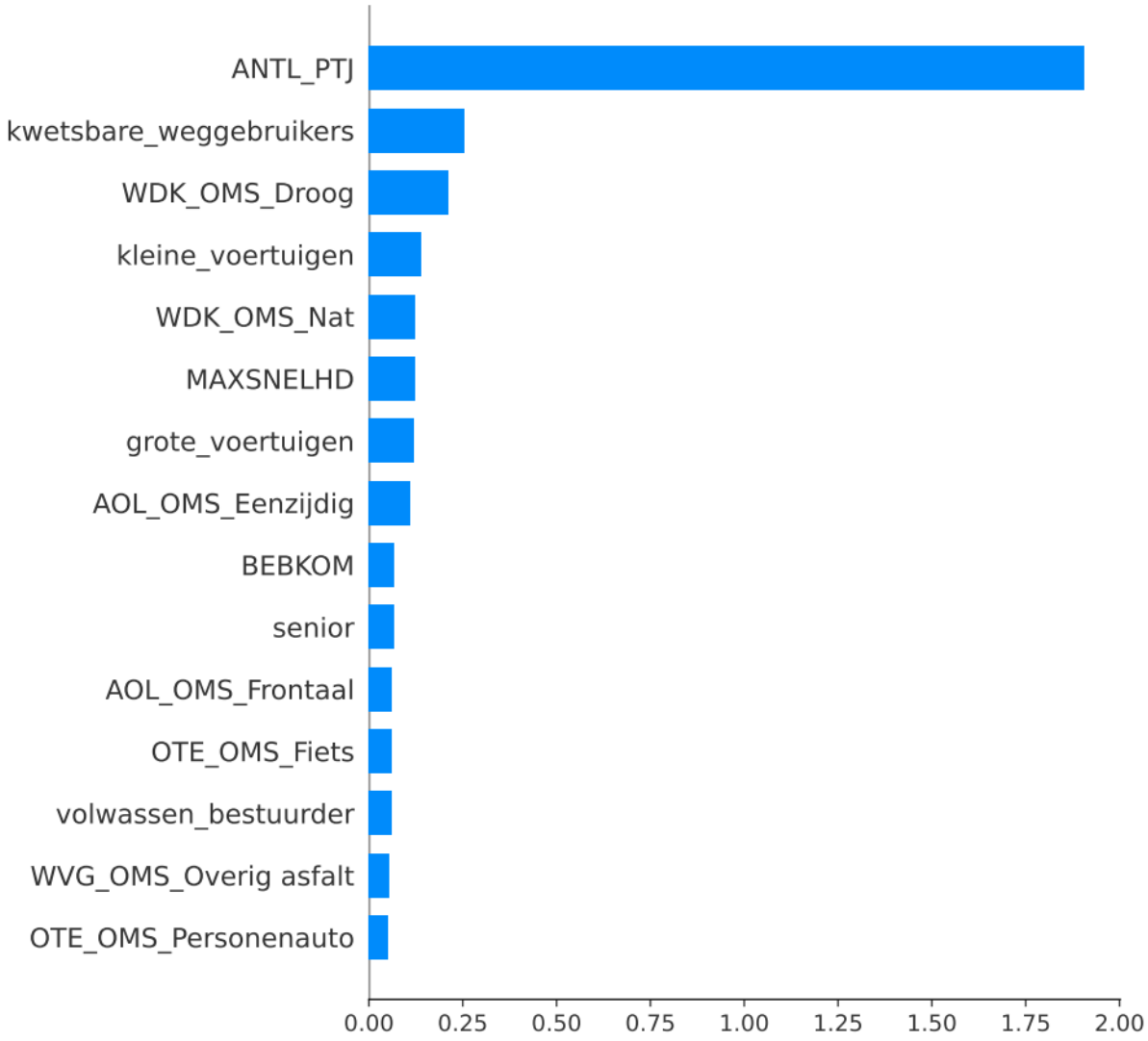


Figure 28: Mean SHAP importance values of LightGBM model for class DOD (fatal outcome)

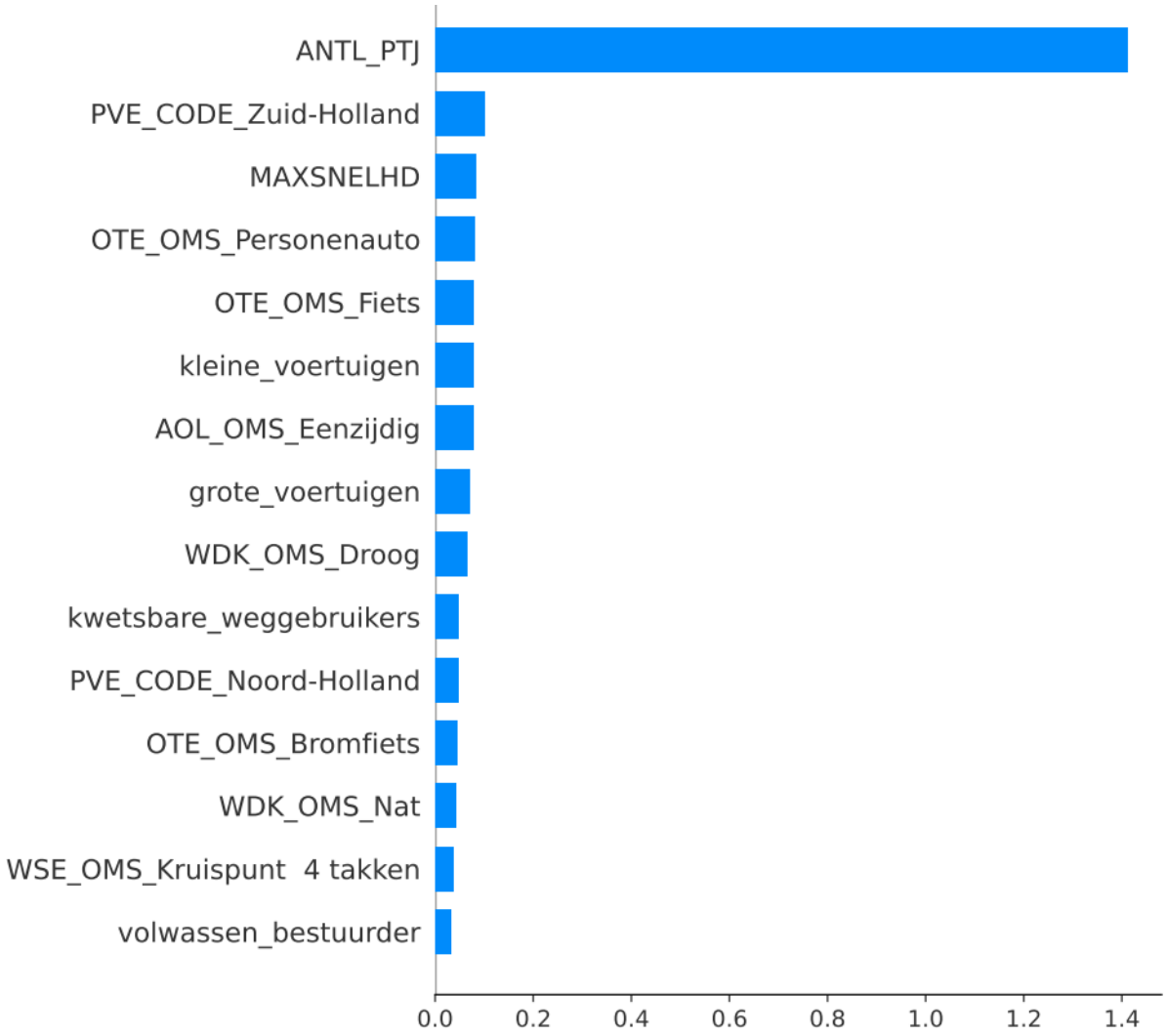


Figure 29: Mean SHAP importance values of LightGBM model for class LET (lethal outcome)

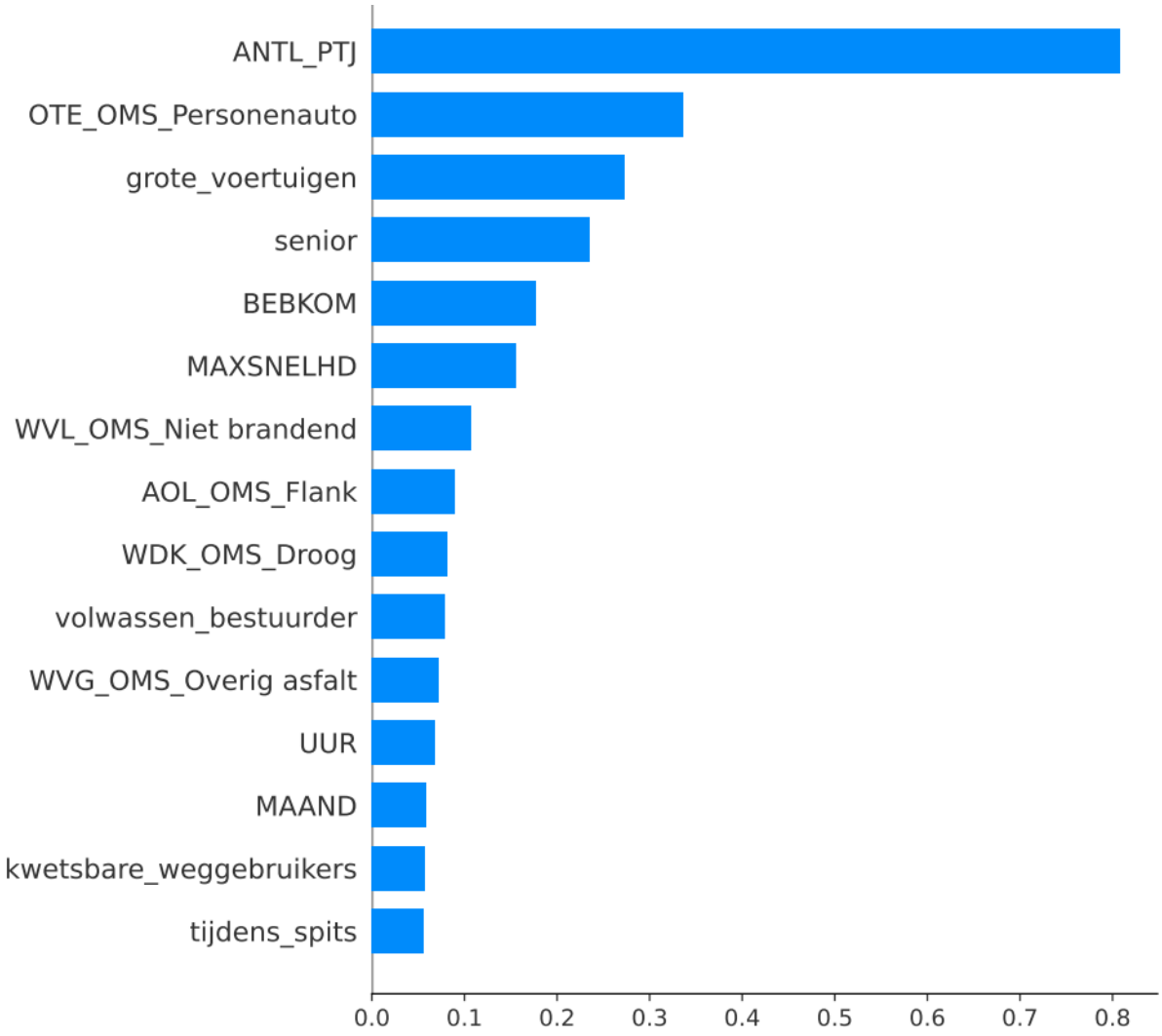


Figure 30: Mean SHAP importance values of LightGBM model for class UMS (material outcome)

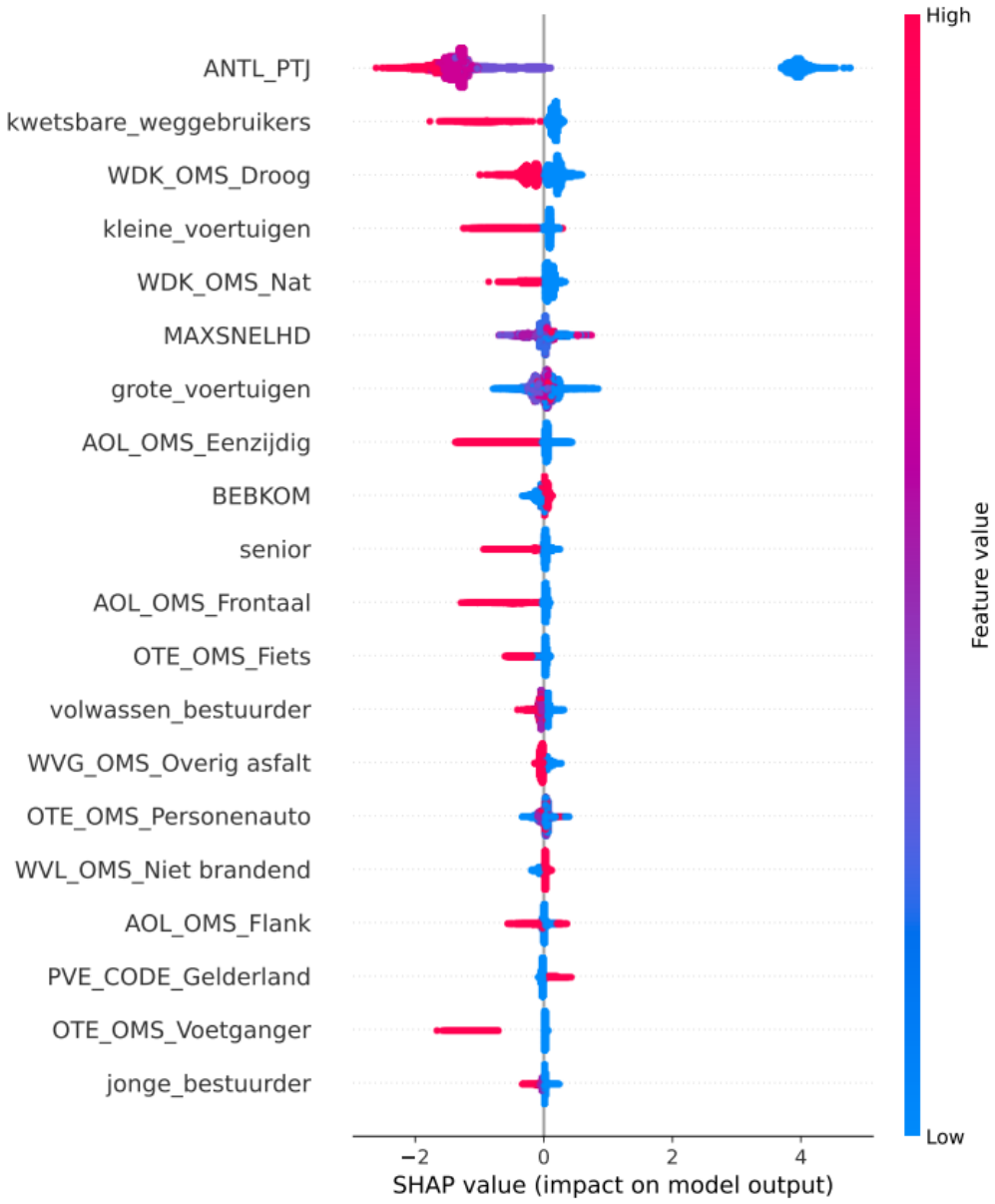


Figure 31: Beeswarm plot of SHAP values of LightGBM model for class DOD (fatal outcome)

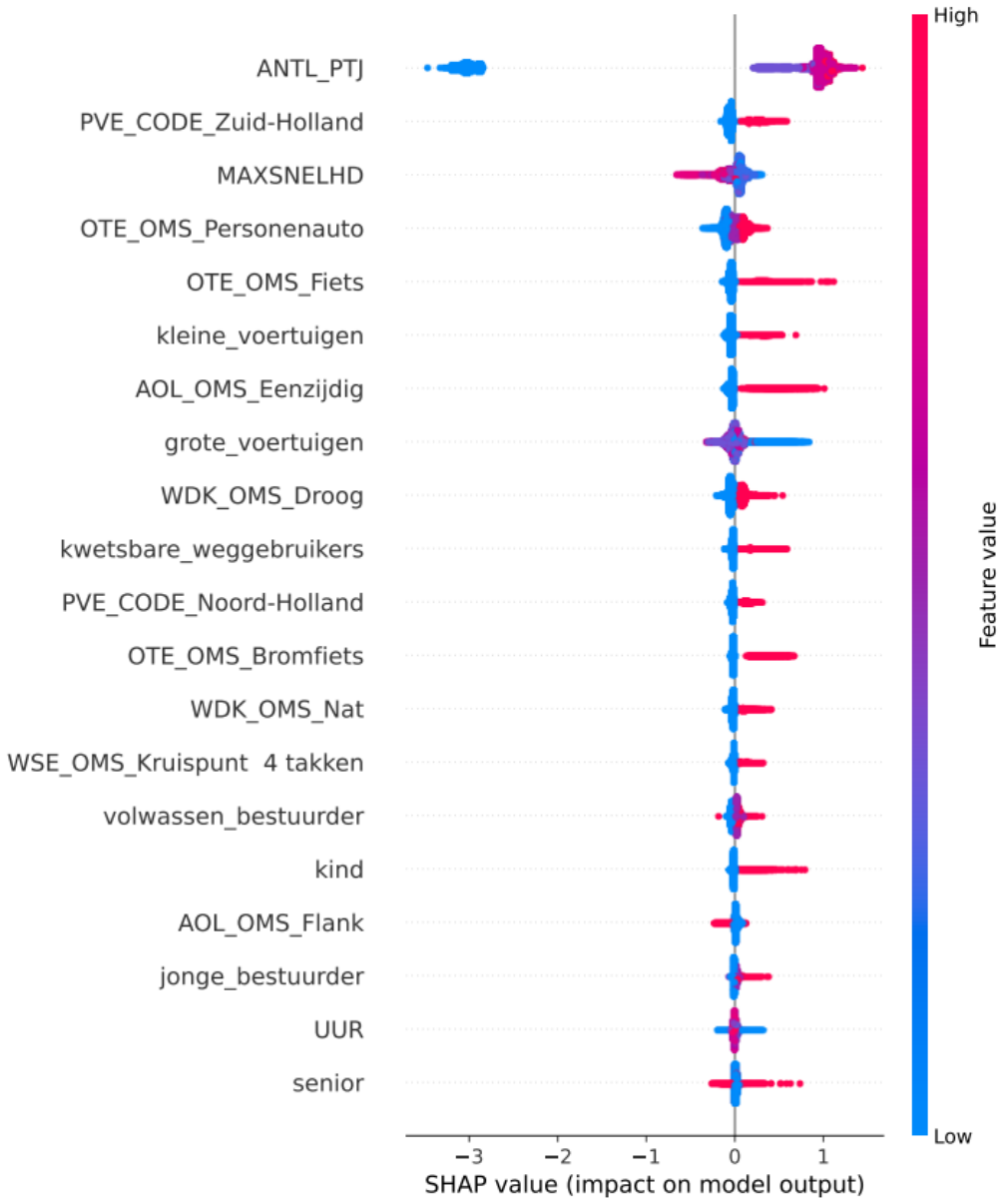


Figure 32: Beeswarm plot of SHAP values of LightGBM model for class LET (lethal outcome)

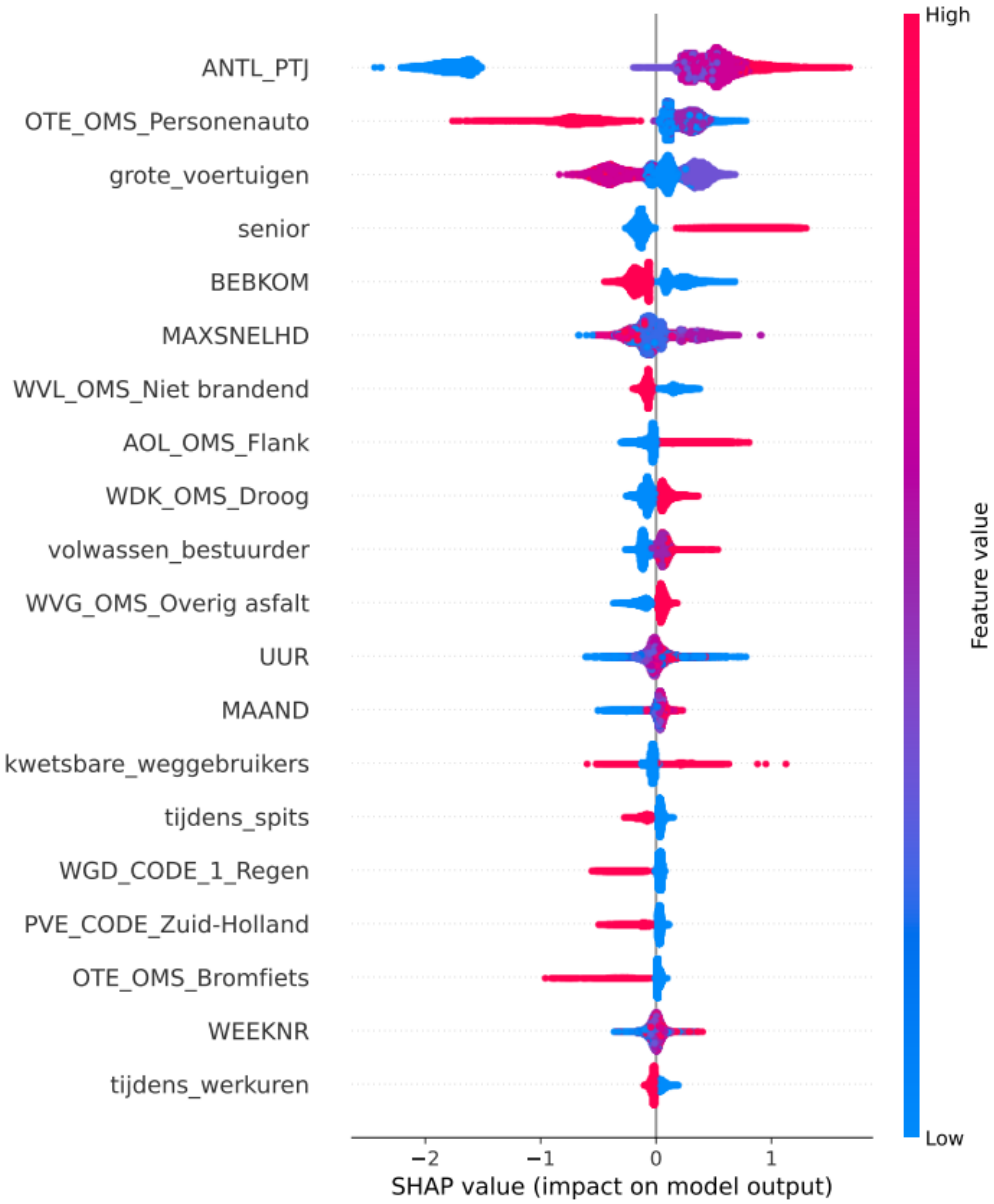


Figure 33: Beeswarm plot of SHAP values of LightGBM model for class UMS (material outcome)

CatBoost

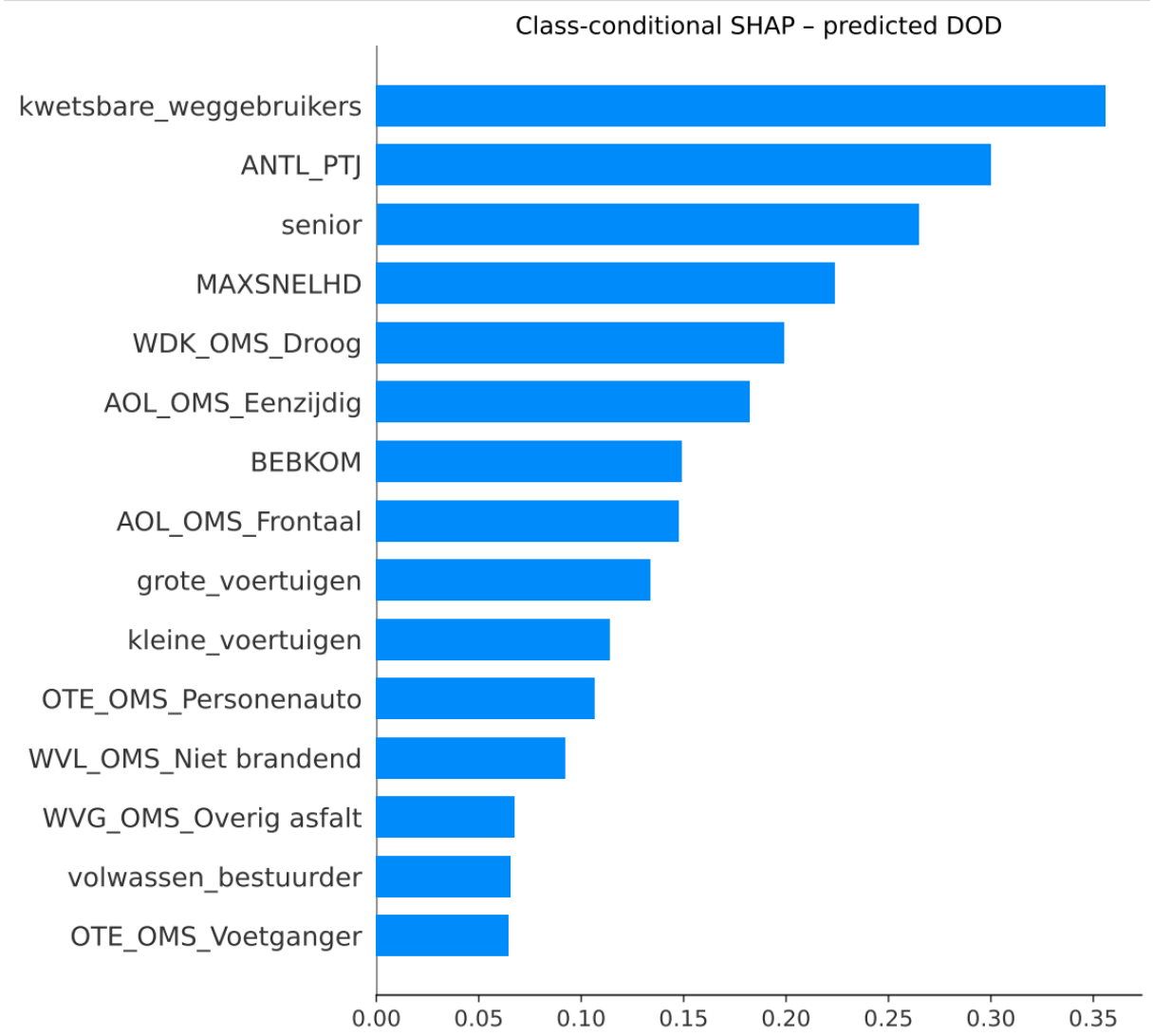


Figure 34: Mean SHAP importance values of CatBoost model for class DOD (fatal outcome)

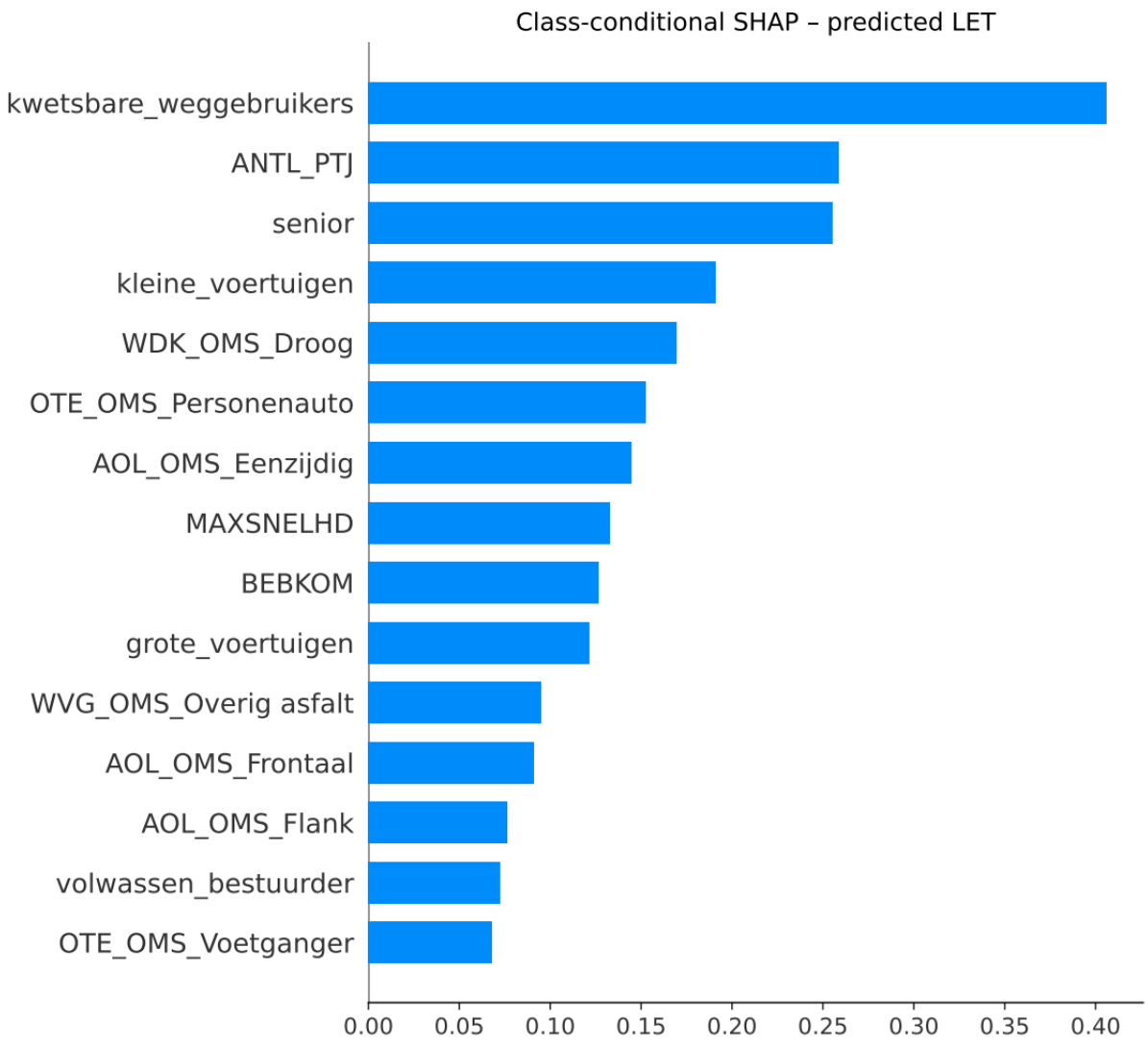


Figure 35: Mean SHAP importance values of CatBoost model for class LET (lethal outcome)

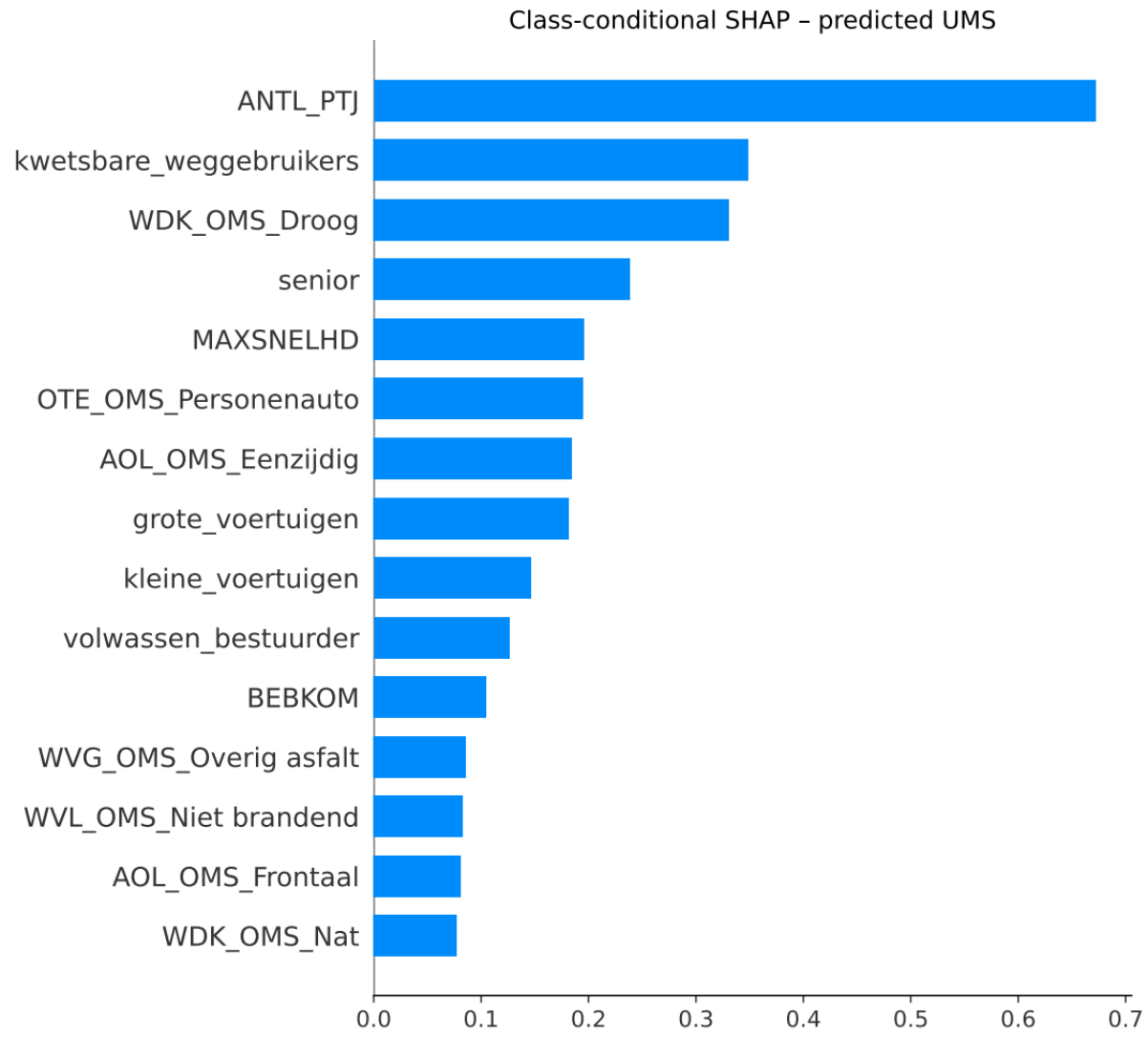


Figure 36: Mean SHAP importance values of CatBoost model for class UMS (material outcome)

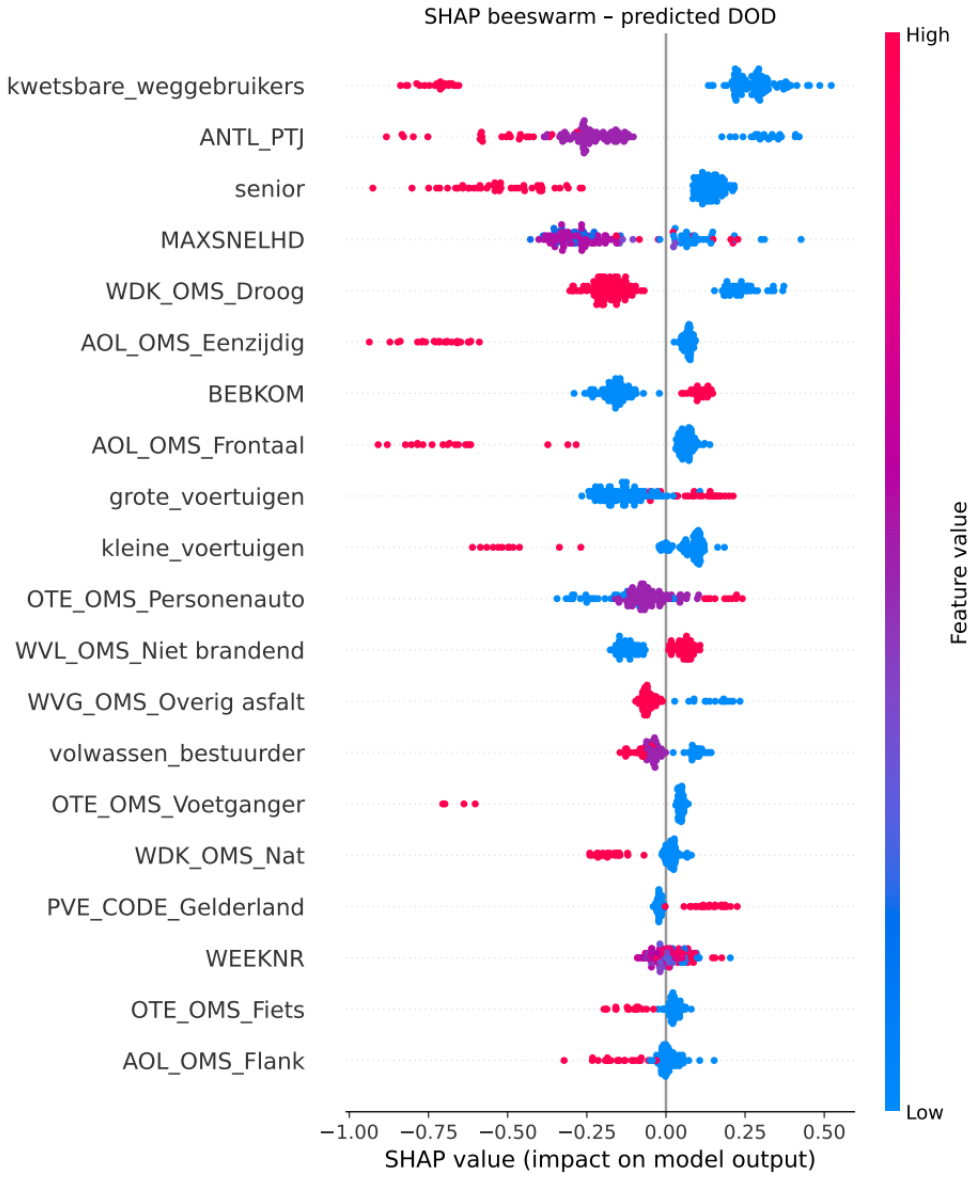


Figure 37: Beeswarm plot of SHAP values of CatBoost model for class DOD (fatal outcome)

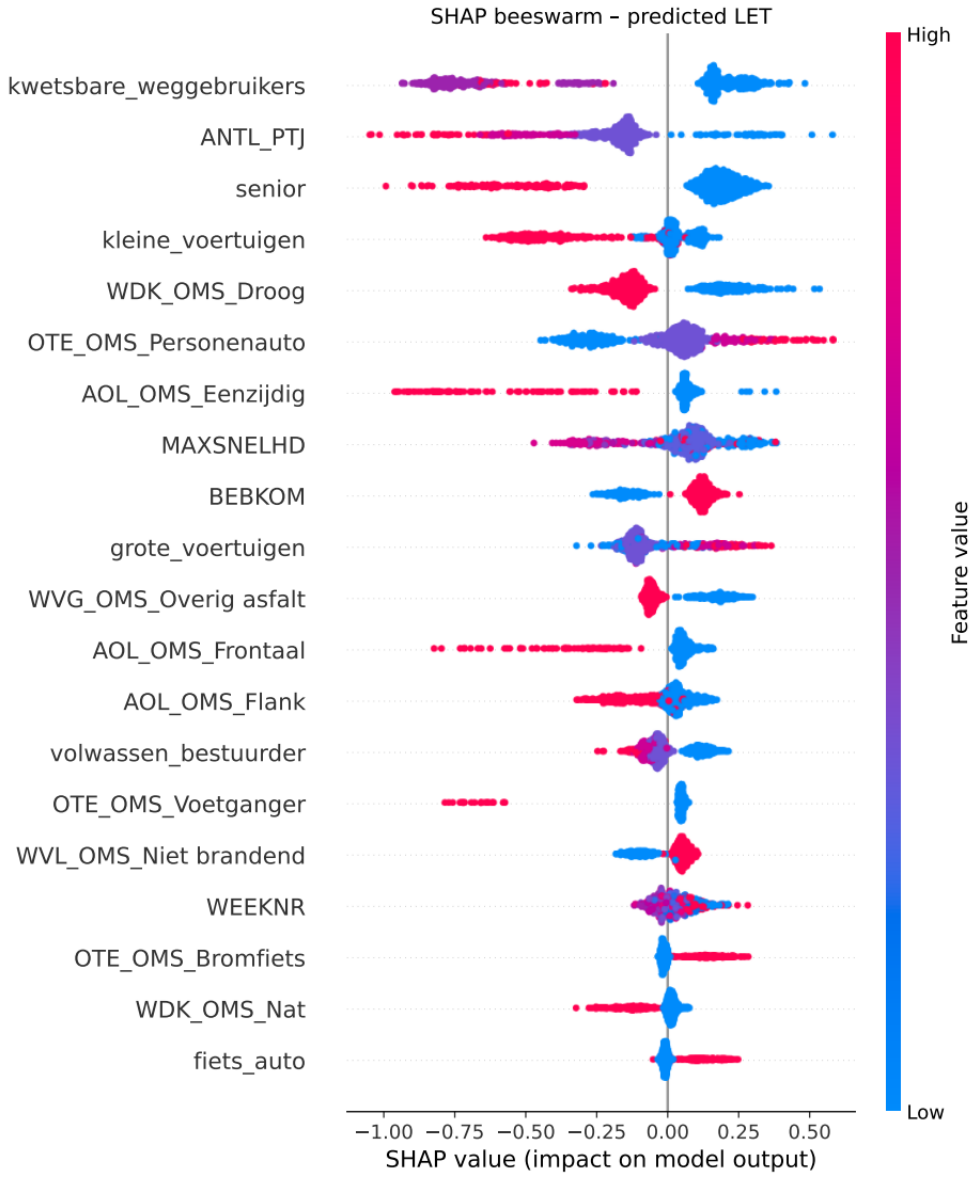


Figure 38: Beeswarm plot of SHAP values of CatBoost model for class LET (lethal outcome)

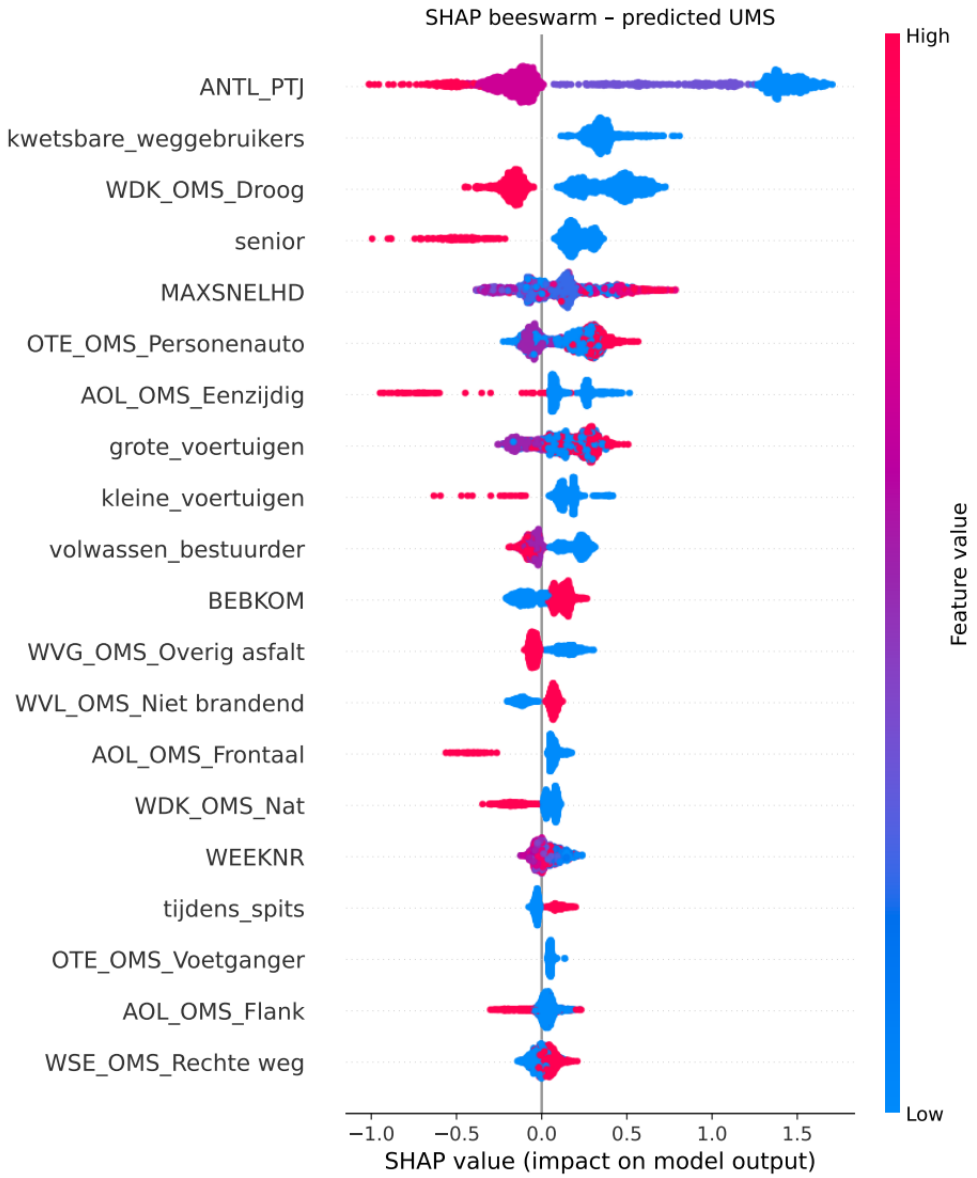


Figure 39: Beeswarm plot of SHAP values of CatBoost model for class UMS (material outcome)

XGBoost

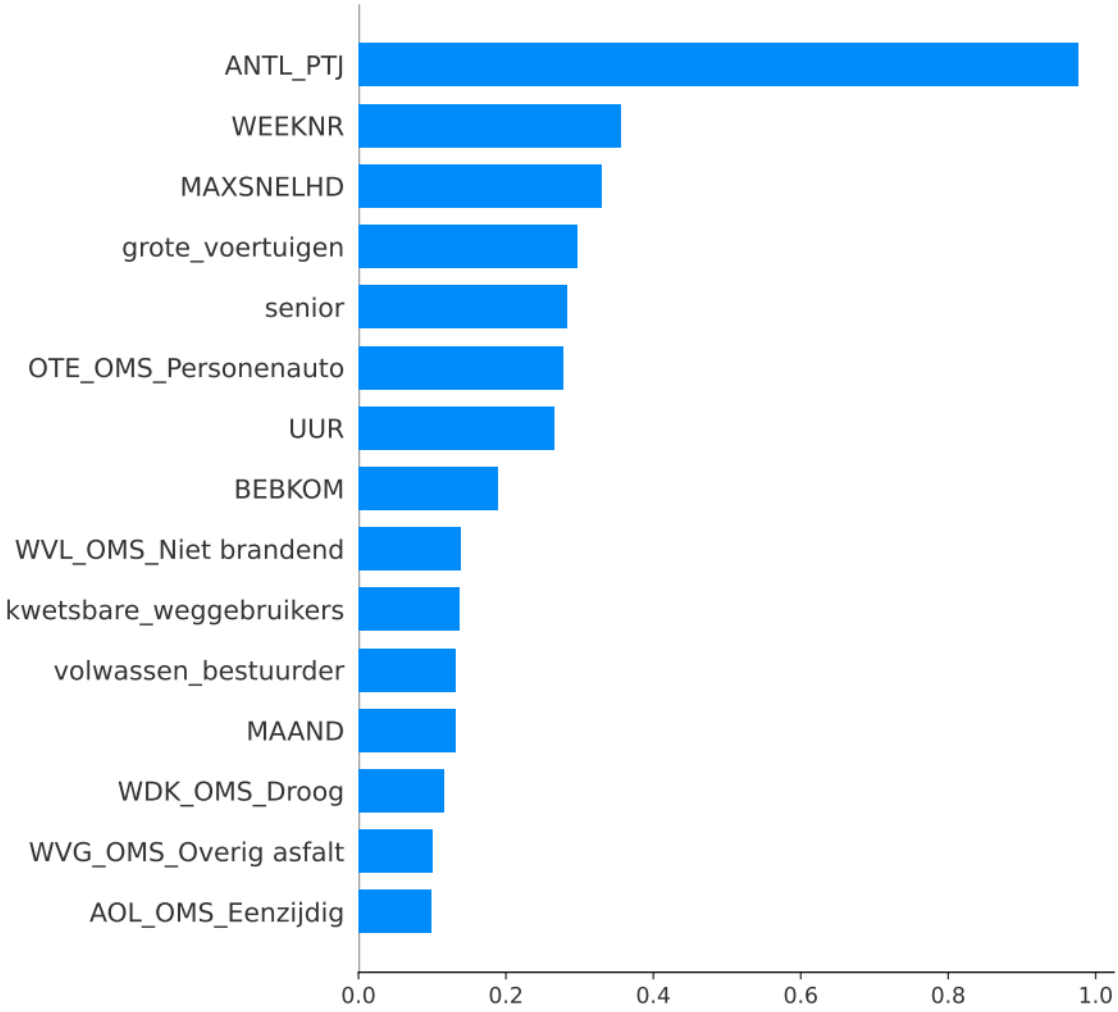


Figure 40: Mean SHAP importance values of XGBoost model for class DOD (fatal outcome)

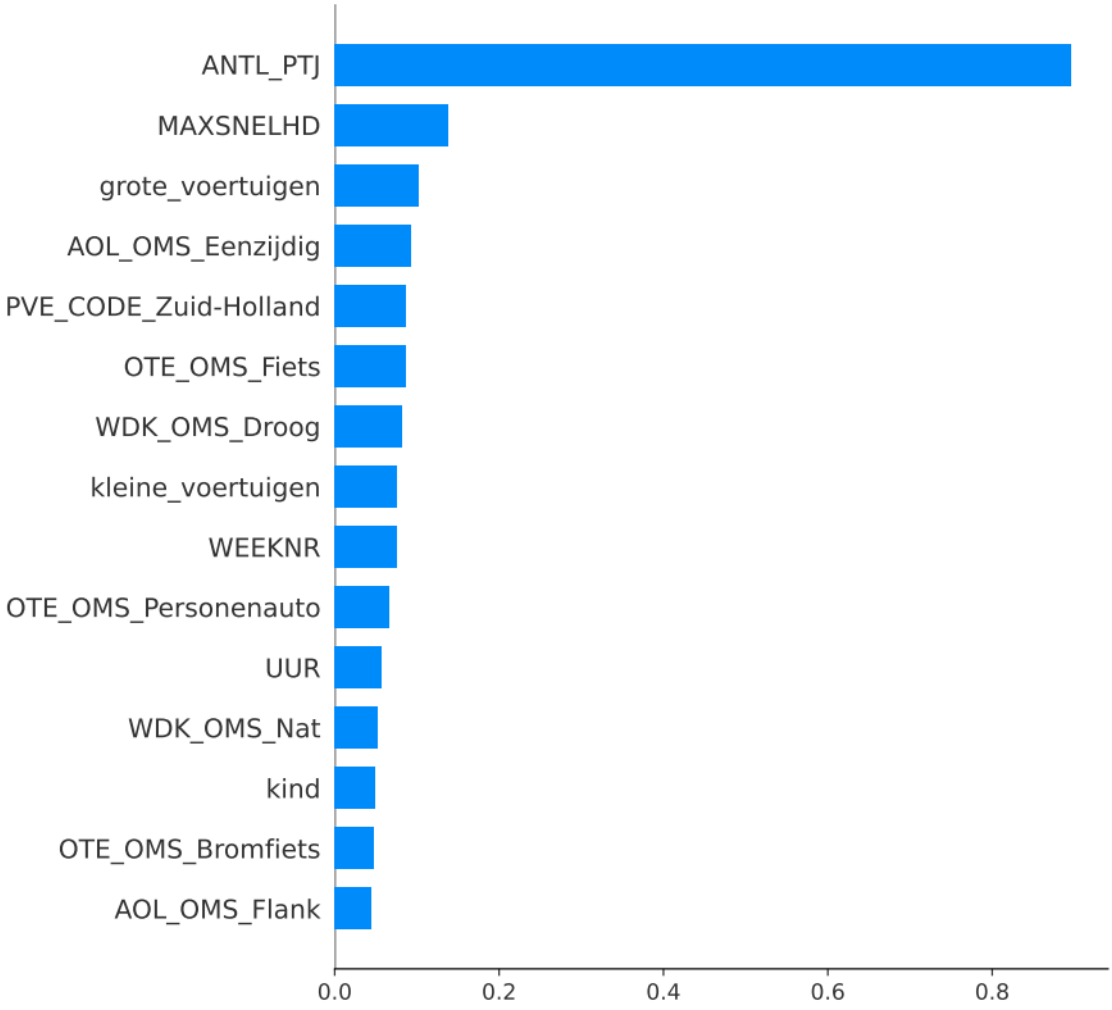


Figure 41: Mean SHAP importance values of XGBoost model for class LET (lethal outcome)

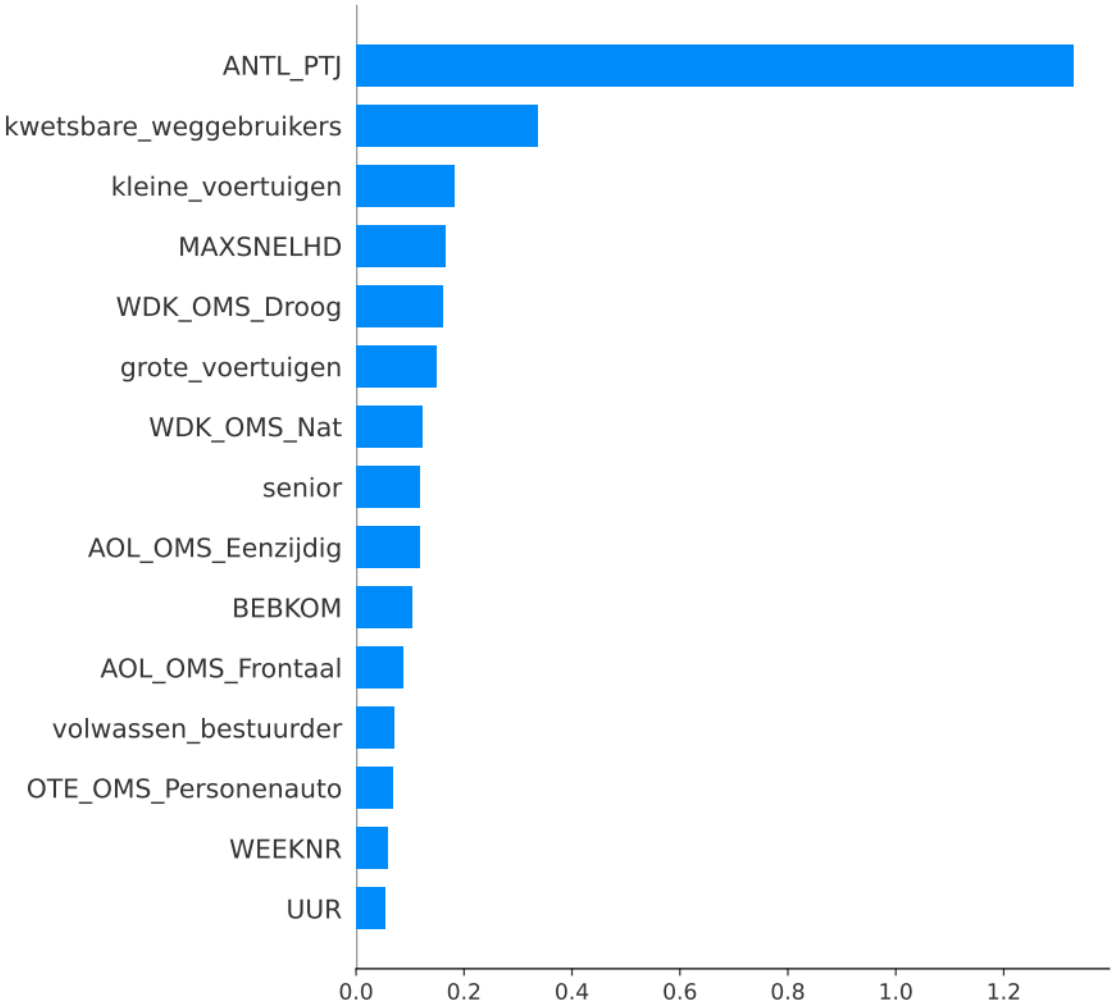


Figure 42: Mean SHAP importance values of XGBoost model for class UMS (material outcome)

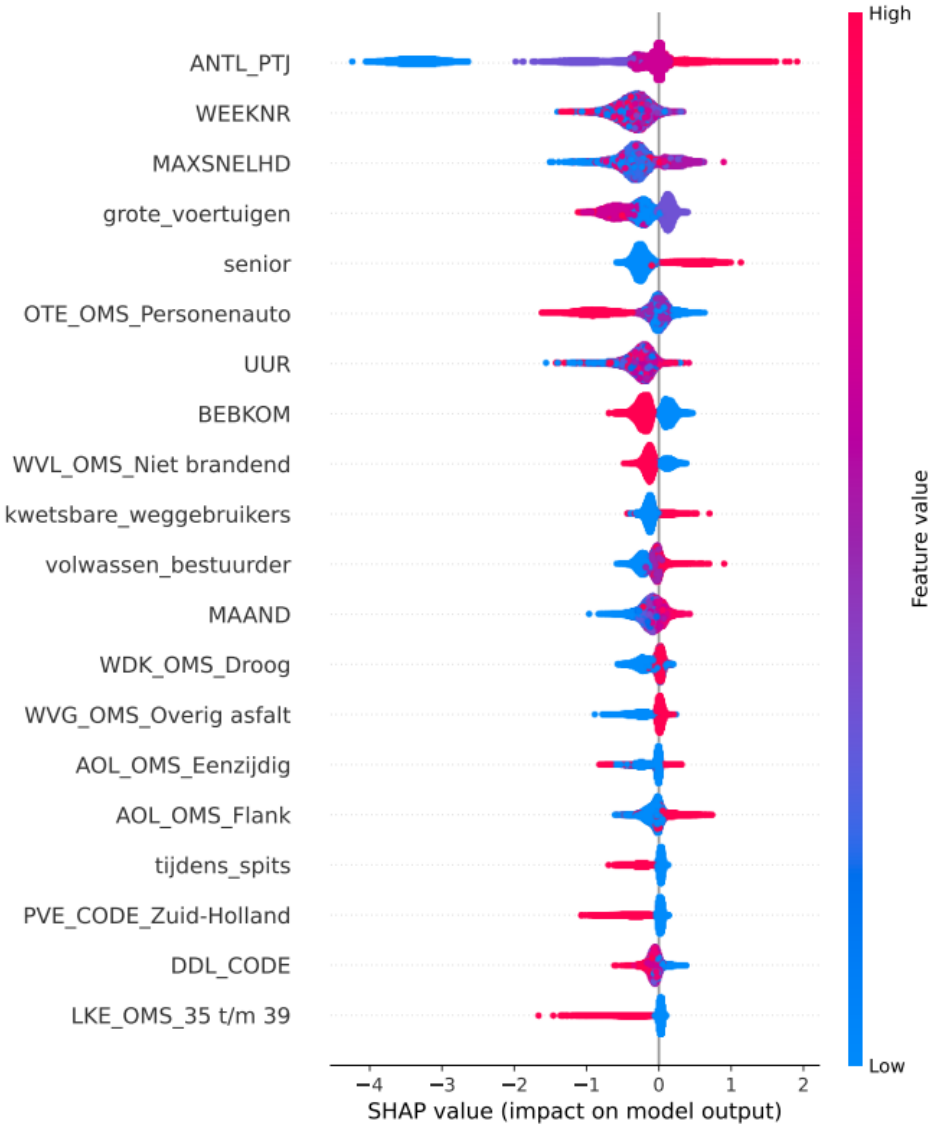


Figure 43: Beeswarm plot of SHAP values of XGBoost model for class DOD (fatal outcome)

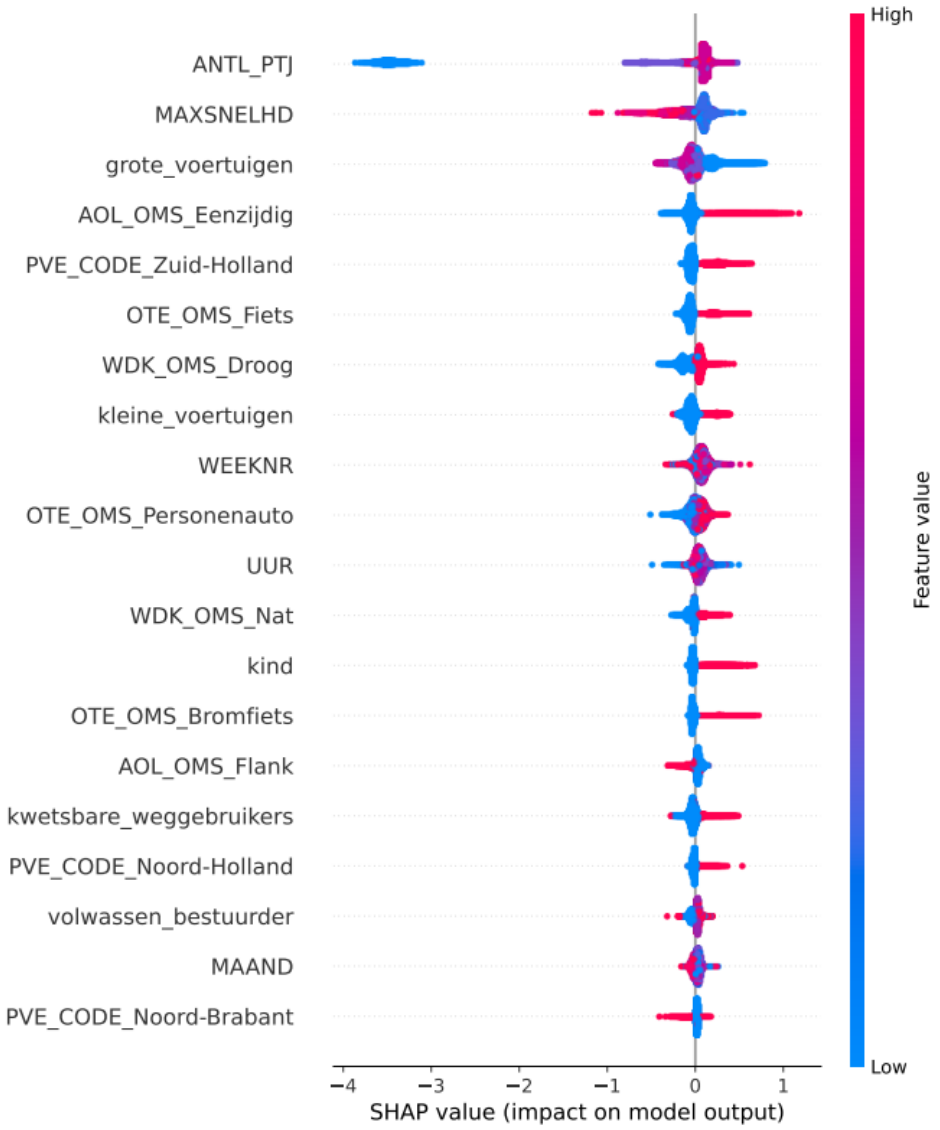


Figure 44: Beeswarm plot of SHAP values of XGBoost model for class LET (lethal outcome)

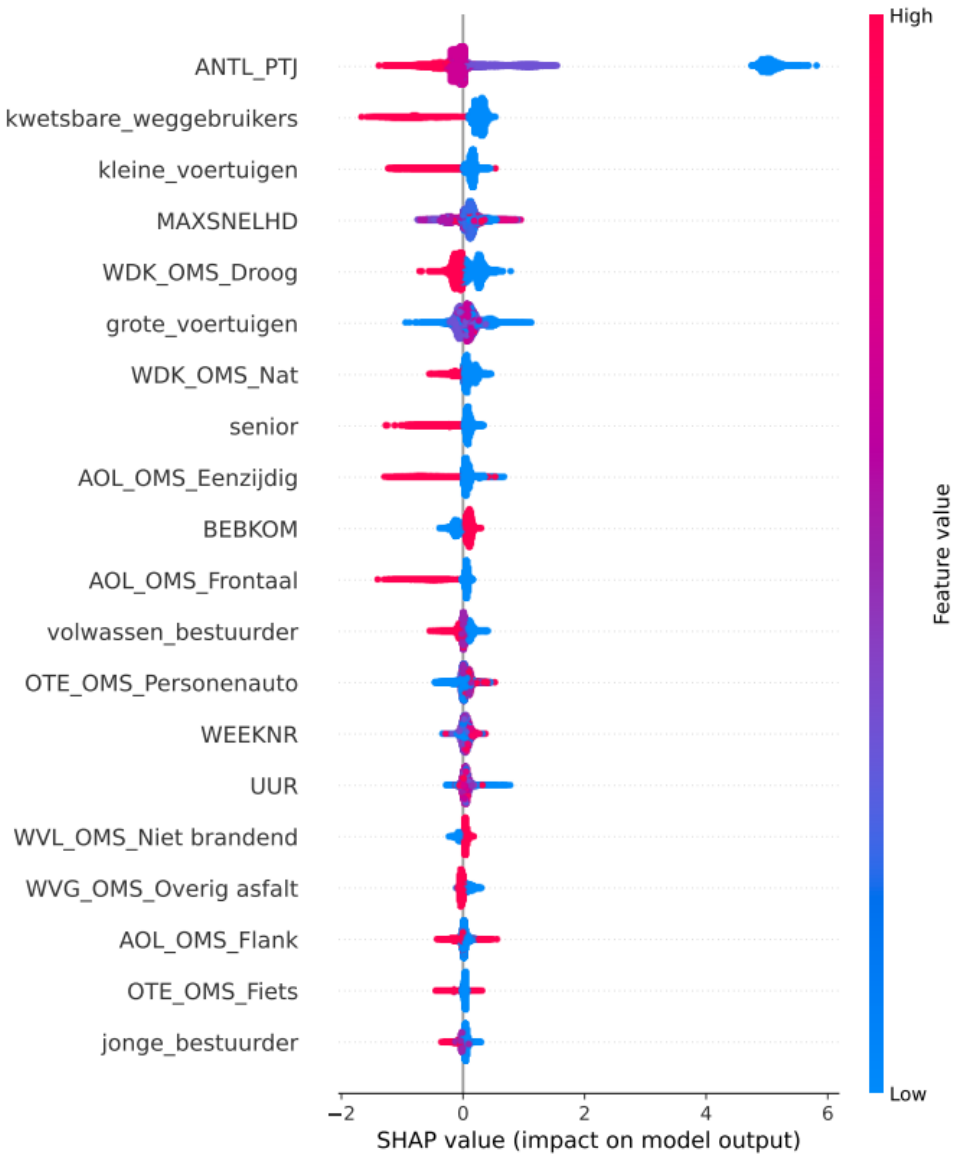


Figure 45: Beeswarm plot of SHAP values of XGBoost model for class UMS (material outcome)