Model Selection and Explainability in Traffic Accident Severity Prediction

CS512 — Machine Learning
Group 14: Başak Çarhacıoğlu Sima Adleyba Ahmet İhsan Gül
Sabancı University
January 2026

1 Introduction

Accurate prediction of traffic crash severity is important for safety analytics, resource allocation, and identifying conditions associated with high-impact outcomes. In this project, we address crash severity prediction using a curated, real-world dataset derived from the Crash Report Sampling System (CRSS, 2016–2023). Our target is a four-class ordered label capturing the maximum injury severity in a crash (0: no injury, 1: minor, 2: severe, 3: fatal). This is a challenging problem due to strong class imbalance with rare fatal crashes and structured missingness where many variables are unknown or not applicable depending on crash context.

A class-weighted Linear SVM and an ordinal XGBoost pipeline form the base of our analysis. Additionally, we analyze the effect of oversampling by comparing oversampled and non-oversampled variants. Finally, we apply three distinct explainability tools to interpret model behavior and validate whether the models rely on plausible signals: SHAP for global feature attribution, LIME for local instance-level explanations, and DiCE for counterfactual explanations that identify minimal feature changes that shift a prediction toward higher severity.

Our results show that both SVM and XGBoost achieve comparable overall performance (approximately 0.59 accuracy and 0.53–0.54 macro-F1 on the oversampled evaluation), but oversampling has a pronounced impact on the rare Fatal class. In particular, the ordinal XGBoost model exhibits a large improvement in fatal-class detection when trained/evaluated under oversampling, increasing fatal-class F1 from roughly 0.26 to 0.59 and raising macro-F1 accordingly, while weighted-F1 changes modestly due to dominance of majority classes. Explainability analyses consistently highlight crash composition and reporting-context variables (including unknown/not-applicable indicators) as influential, underscoring the importance of careful missingness handling. Finally, an explanation-driven feature elimination attempt does not improve generalization, suggesting that SHAP/LIME/DiCE are most reliable as interpretation tools rather than aggressive feature selection mechanisms in this sparse, interaction-heavy setting.

2 Problem Description

The objective of this project is to predict crash severity from real-world crash records using supervised learning. Concretely, given a crash-level feature vector summarizing driver, vehicle, roadway, and environmental factors, we aim to classify the maximum injury severity of the crash into one of four ordered categories: no injury, minor injury, severe injury, or fatal. This task is inherently challenging because the class distribution is highly imbalanced, in that fatal outcomes are rare, and because the data exhibits substantial structured missingness. Many fields are recorded as unknown or not applicable for practical reasons, such as limited information available at the scene, reporting constraints, or contextual irrelevance of certain attributes for particular crash types. As a result, the learning problem requires methods that remain robust under sparsity and imperfect data capture.

Beyond predictive performance, we seek to understand how and why the models make their decisions. We therefore investigate which features most strongly influence predicted severity through global and local explanation analyses, and we perform counterfactual analysis to identify minimal changes in feature values that shift a prediction toward higher severity under the model. Finally, we examine whether explanation-driven feature reduction simplifies the high-dimensional representation and improves generalization, or whether the predictive signal is distributed across many weak features and interactions that are lost under aggressive pruning.

3 Methods

3.1 Dataset

CRSS is a highly distributed relational dataset where a single crash is represented across multiple linked tables (e.g., crash-, vehicle-, and person-level files), and each crash can have multiple associated vehicles and occupants. To enable standard supervised learning, we convert the relational structure into a single crash-level feature vector by joining related tables on the crash identifier and aggregating multi-entity information into a fixed-length representation. Multi-entity information is incorporated using a combination of indexed feature slots (captur-

Table 1: Dataset split sizes and test-set class distribution for `MAX_SEV`.

Dataset	Train	Val	Test
`data_final`	71,332	15,286	15,286
`data_final_oversampled`	73,591	15,769	15,770

Test Split	C0	C1	C2	C3
Non-oversampled	48.87%	37.69%	11.08%	2.37%
Oversampled	47.37%	36.53%	10.74%	5.36%

ing limited per-entity detail) and crash-level aggregates (counts, proportions, and summary statistics) to avoid row explosion while retaining informative signals.

The prediction target is `MAX_SEV`, a four-class ordered label: 0 (no injury), 1 (minor), 2 (severe), 3 (fatal). To address the imbalance we utilize SMOTE oversampling and compare the results for oversampled and non-oversampled variants. Table 1 shows the characteristics of datasets.

Real-world crash data contains substantial missingness that is largely structured rather than random. Many variables are unobserved due to reporting constraints, while others are absent by definition because the corresponding condition or entity type does not occur in a given crash. Rather than discarding these fields, our curated dataset explicitly encodes missingness using dedicated `_UNKNOWN` and `_NOT_APPLICABLE` indicator features, which capture information availability and contextual relevance, respectively. Concretely, a `_UNKNOWN` flag is set to 1 when the underlying attribute is entirely missing for the crash, whereas `_NOT_APPLICABLE`=1 denotes that the attribute is conceptually irrelevant for that crash (e.g., because the relevant entity type is absent).

In addition, we distinguish between the feature representation used for prediction and the representation used for explainability. Model training and evaluation use a leakage-safe feature set aligned with inference-time availability, while some explainability tools require fully specified inputs; therefore, an imputed explanation-only view is constructed (using available imputed columns and simple train-set statistics as fallback) to ensure stable neighborhood perturbations and counterfactual search without altering the predictive feature space.

3.2 Models: Linear SVM and XGBoost

In earlier experiments where we evaluate several classical learning methods, Linear SVM and XGBoost consistently achieve the strongest performance on our task, particularly in terms of Macro-F1; therefore, we select them as the primary models for the main study. Additional preliminary experiments can be viewed in the Appendices section.

Linear SVM: We use a Linear Support Vector Machine as a strong baseline for high-dimensional crash-level features. Class imbalance is handled via `class_weight=balanced`, and the regularization parameter C is tuned on the validation split over a logarithmic grid. Our pipeline applies standard preprocessing where the numeric features are scaled, categorical variables are encoded consistently with the curated schema, and the resulting feature matrix is trained end-to-end using the fixed train/validation/test splits. Given the structured nature of missingness in CRSS, we avoid heavy reliance on generic imputation and instead preserve information availability through the curated `_UNKNOWN` and `_NOT_APPLICABLE` indicators, allowing the model to distinguish between unobserved values and contextually irrelevant attributes.

Ordinal XGBoost: To reflect the ordered nature of severity labels, we implement a cumulative (ordinal) XGBoost formulation. Specifically, we decompose the four-class problem into three binary subproblems defined by severity thresholds ($y \geq 1$, $y \geq 2$, $y \geq 3$). For each threshold, we train a separate XGBoost classifier with early stopping on the validation split, using imbalance-aware weighting via `scale_pos_weight` computed from the corresponding binary label distribution. At inference time, the three predicted cumulative probabilities are transformed into a valid four-class distribution: $P(y=0) = 1-p_1$, $P(y=1) = p_1-p_2$, $P(y=2) = p_2-p_3$, and $P(y=3) = p_3$, where $p_k = P(y \geq k \mid x)$. The final prediction is selected as the class with the largest probability. This ordinal decomposition encourages consistent separation between adjacent severity levels and empirically improves minority-class behavior compared to a single flat multi-class formulation in our setting.

3.3 Explainability Tools

To interpret model behavior beyond aggregate metrics, we use three complementary explainability tools that operate at different levels: global attribution (SHAP), local instance-level explanation (LIME), and counterfactual analysis (DiCE). Our goal is twofold across all analyses: we aim to identify which variables the models rely on when assigning severity levels, and to probe how predictions change under controlled perturbations or minimal hypothetical feature changes, with particular emphasis on transitions toward Fatal outcomes.

Feature representations for prediction vs. explanation. We distinguish between the feature representation used for model training/evaluation and the representation used for explainability. Predictive performance uses a leakage-safe feature set aligned with inference-time availability, excluding dataset artifacts (e.g., `SMOTE`) and imputed `_IM` columns. Since LIME and DiCE require fully specified inputs, we construct an explanation-only representation by filling missing values using available imputed

columns and simple train-set statistics (median/mode). This separation preserves the predictive feature space while ensuring stable explanations.

SHAP (global attributions). SHAP quantifies feature contributions to model outputs in a manner that remains consistent with the trained predictive model. We apply SHAP on the same feature space used for prediction to obtain global importance scores (e.g., mean absolute SHAP values across the evaluation set) and, when needed, class-conditional patterns. This provides a dataset-level view of which variables most strongly influence severity predictions and helps assess whether the learned signals align with domain expectations. Since SHAP explains the model as trained, we avoid introducing imputed-only features that the model does not consume.

LIME (local explanations). LIME explains individual predictions by fitting a simple surrogate model in the local neighborhood of a selected instance. It generates perturbed samples, queries the black-box model, and fits a locally weighted linear model whose coefficients rank the most influential features for that prediction. Because this procedure requires complete inputs, LIME operates on the imputed explanation-only representation. We use LIME primarily for qualitative case studies, including representative correct predictions and informative errors.

DiCE (counterfactual explanations). DiCE performs counterfactual analysis by searching for minimal feature changes that shift a prediction to a specified target class, most notably Fatal. It returns alternative feature configurations that achieve the target while remaining close to the original instance under a distance metric, which allows us to summarize which features most frequently change when pushing predictions toward higher severity. Since counterfactual search requires fully specified inputs, DiCE operates on the imputed explanation-only representation. We interpret these counterfactuals as decision-boundary sensitivity rather than causal mechanisms and use them as diagnostic signals for severity transitions.

Explanation-driven feature reduction. Finally, we investigate whether explanations support dimensionality reduction. We compute feature importance signals using SHAP, LIME, and DiCE and drop features that are consistently assigned negligible influence across methods, then re-evaluate predictive performance on the validation/test splits. This experiment tests whether the predictive signal is concentrated in a small subset of variables or distributed across many weak features and interactions in the curated crash-level representation.

3.4 Overall Experimental Workflow

For both Linear SVM and ordinal XGBoost, we train models on the fixed training split and select configurations using validation performance under both non-oversampled and SMOTE-oversampled settings. We then apply SHAP, LIME, and DiCE to interpret predictions at global, local, and counterfactual levels, and we use these signals to analyze influential variables and identify consistently low-impact features. For XGBoost we aggregate LIME/DiCE results over a fixed evaluation subset to summarize recurring local drivers, while for SVM we report class-conditional explanations to directly interpret class-specific decision scores. Finally, we re-train and re-evaluate the models with the reduced feature set to assess whether explanation-driven pruning simplifies the representation and improves generalization on the held-out test split.

4 Results

4.1 Experimental Questions

Our experiments address three questions:

(i) how Linear SVM and ordinal XGBoost perform under imbalanced, high-dimensional settings; (ii) which variables most strongly drive severity predictions according to complementary explainability tools (SHAP, LIME, DiCE); and (iii) whether explanation-driven feature reduction simplifies the representation and improves generalization, or instead removes distributed predictive signal.

4.2 Initial Predictive Performance (Validation)

The baseline validation performance for the selected models can be seen in Table 2, comparing all methods on validation using Macro-F1 and weighted-F1. For XGBoost, Variants A and B denote oversampled and non-oversampled data, respectively.

Table 2: Validation performance comparison (Macro-F1 and weighted-F1).

System	Macro-F1	Weighted-F1
XGBoost Var A	0.5235	0.5849
XGBoost Var B	0.4321	0.5616
SVM	0.5356	0.5678

Figure 1 shows that all models struggle most on the intermediate classes (Minor and Severe), which frequently confuse with neighboring severities. The oversampled XGBoost setting reduces fatal-class errors relative to Var B, while SVM maintains strong separation for the No Injury class but still confuses Minor and Severe outcomes.

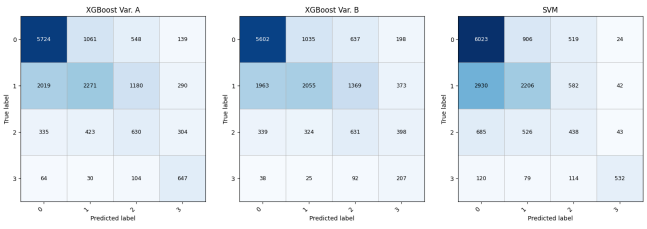

Figure 1: Validation confusion matrices for XGBoost Var A, XGBoost Var B, and SVM (left to right). Rows denote true labels and columns denote predicted labels. Classes: 0 = No Injury, 1 = Minor, 2 = Severe, 3 = Fatal.

4.3 Explainability Analysis: XGBoost

XGBoost decisions are interpreted using three complementary explainability tools. SHAP provides global attributions that remain faithful to the trained predictors, LIME provides instance-level explanations via a local surrogate model, and DiCE summarizes counterfactual "levers" by searching for minimal feature changes that move predictions toward a target class. Attribution magnitudes are omitted in the main text and only the most frequently top-ranked feature names are reported; extended outputs are deferred to the Appendix.

Table 3: XGBoost: Top-5 SHAP-ranked feature names per ordinal threshold model.

Rank	$P(y \geq 1)$	$P(y \geq 2)$	$P(y \geq 3)$
XGBoost Var A			
1	NMACTION_NA	P_CRASH1_2	ALCOHOL
2	VEH_COUNT	TRAV_SP_1	P_CRASH1_2
3	GVWR_AVG_1	VEH_COUNT	DRDISTRACT_1
4	P_CRASH1_2	DRIVR_MALE_P	TRAV_SP_1
5	P_CRASH1_1	P_CRASH1_1	DRIVER_MALE_P
XGBoost Var B			
1	NMACTION_NA	TRAV_SP_2	ALCOHOL
2	TOTAL_COUNT	TRAV_SP_1	TRAV_SP_2
3	GVWR_AVG_1	P_CRASH1_2	TRAV_SP_1
4	REGION	ALCOHOL	DRDISTRACT_1
5	P_CRASH1_1	DRIVR_MALE_P	DRIVER_MALE_P

Table 3 lists the top-5 SHAP-ranked features for each ordinal threshold model. Across both variants, crash configuration (`P_CRASH1_*`), exposure/composition (`VEH_COUNT`, `TOTAL_COUNT`), travel speed (`TRAV_SP_*`), and vehicle-mass proxies (`GVWR_AVG_*`) repeatedly appear, while `NMACTION_NA` reflects the structured missingness/context signals that are informative in CRSS. Differences across columns are expected because each threshold model captures a different severity regime.

Table 4 aggregates LIME explanations over the fixed 89-instance set used for local analyses (stratified correct predictions per class, supplemented with a small number of high-confidence errors and additional minority-class samples). Because LIME fits a locally weighted linear surrogate around each selected crash after generating neighborhood perturbations, the aggregated ranking highlights features that recur across many local decision contexts. The dominant features align closely with SHAP—crash composition and configuration, impairment/distraction cues, roadway/alignment indicators, and `*_UNKNOWN/*_NA` flags—supporting consistency between global and local explanations.

Table 4: XGBoost: Top-10 LIME-ranked feature names (values omitted), aggregated over 89 analyzed instances.

Rank	XGBoost Var A	XGBoost Var B
1	VTRAFWAY_4	TRAV_SP_1
2	PEDESTRIAN_COUNT	P_CRASH1_4
3	P_CRASH1_4	TOTAL_COUNT
4	DRIMPAIR_2	VPROFILE_NA
5	DRDISTRACT_3	TRAV_SP_4
6	P_CRASH1_3	RELJCT1_UNKNOWN
7	MANEUVER_4	VPROFILE_4
8	P_CRASH1_UNKNOWN	DRIVERRF_4
9	VALIGN_4	DRDISTRACT_4
10	MINUTE_UNKNOWN	P_CRASH1_NA

DiCE complements attribution methods by producing counterfactual explanations. A fatal-focused protocol is applied on the test split using an imputed, explanation-only representation to keep the counterfactual search well-defined under missingness, while leaving the predictive feature space unchanged. The subset consists of 20 true Fatal crashes, and three counterfactuals per instance (60 total) are generated with the target set to Severe (class 2), i.e., the analysis emphasizes which feature changes most often move predictions away from Fatal.

Table 5: XGBoost: DiCE fatal-focused counterfactual summary (feature names only). Features are ranked by how frequently they change across 60 counterfactuals generated from 20 true Fatal test instances (3 counterfactuals per instance).

Rank	XGBoost Var A	XGBoost Var B
1	P_CRASH1_2	TRAV_SP_1
2	MANEUVER_1	TRAV_SP_2
3	DRIVER_DRUGS_PCT	TOTAL_COUNT
4	REL_ROAD	VSURCOND_4
5	VSURCOND_1	HOUR
6	P_CRASH1_1	P_CRASH1_4
7	NMCC	DRIVERRF_2
8	VEH_COUNT	P_CRASH1_NOT_APPLICABLE
9	VTRAFWAY_2	NMHELMET_ANY
10	DRIMPAIR_1	REL_ROAD

Table 5 ranks features by counterfactual change frequency, serving as a practical proxy for which variables act as high-leverage decision-boundary "knobs" under the trained model. Var A emphasizes crash configuration/road context and maneuver/pedestrian-position descriptors, whereas Var B is dominated by travel-speed

and vehicle-mass proxies; in both cases, the rankings reflect model sensitivity rather than causal mechanisms.

4.4 Explainability Analysis: SVM

SVM explanations are reported at two complementary granularities. Table 6 lists the top-5 feature names per class under SHAP and LIME, with attribution magnitudes omitted. SHAP ranks features on the same (leakage-safe) predictive feature space used by the trained SVM, producing class-specific global attribution summaries. LIME ranks features by aggregating local surrogate explanations over a fixed 89-instance analysis set (stratified correct predictions across classes, augmented with a small set of informative errors and minority-class samples). Color encodes the average direction of contribution to the corresponding class score (positive in red, negative in blue).

Table 6: SVM: Top-5 SHAP and LIME feature names per class. Red indicates a positive contribution to the class score, while blue indicates a negative contribution (values omitted).

Class	SHAP	LIME
0	AGE_18_30_PCT	AGE_30_45_PCT
0	AGE_0_18_PCT	AGE_18_30_PCT
0	AGE_30_45_PCT	AGE_45_60_PCT
0	TRAV_SP_4	AGE_45_60_PCT
0	GVWR_AVG_UNKNOWN	TOTAL_COUNT
1	MANEUVER_UNKNOWN	PEDESTRIAN_COUNT
1	AGE_0_18_PCT	PEDPOS_CROSSWALK_ANY
1	AGE_18_30_PCT	TYP_INT
1	VNUM_LAN_4	VNUM_LAN_UNKNOWN
1	RELJCT2	DRIVER_COUNT
2	TRAV_SP_4	DRIMPAIR_UNKNOWN
2	AGE_30_45_PCT	TRAV_SP_2
2	AGE_18_30_PCT	TOTAL_COUNT
2	GVWR_AVG_UNKNOWN	PEDPOS_ROADWAY_ANY
2	TRAV_SP_1	AGE_45_60_PCT
3	NMOTHPRE_UNKNOWN	ALCOHOL_UNKNOWN
3	REGION	PEDESTRIAN_COUNT
3	AGE_45_60_PCT	DRIVER_COUNT
3	TRAV_SP_UNKNOWN	REGION
3	TOTAL_COUNT	AGE_60_80_PCT

Counterfactual behavior is summarized in Table 7. For each non-fatal source class, DiCE generates counterfactuals that target Fatal using the imputed explanation-only representation (to keep the search well-defined under missingness), and features are ranked by DiCE importance (frequency × mean absolute change). The table reports the top-5 feature names per source class, indicating which variables most often act as high-leverage "knobs" for moving SVM predictions toward Fatal under this protocol.

Table 7: SVM: DiCE per-source-class summary (source → target Fatal). Top-5 features ranked by DiCE importance (names only).

Rank	Source 0	Source 1	Source 2
1	PEDESTRIAN_COUNT	DRIVERRF_4	PEDESTRIAN_COUNT
2	DRIVERRF_4	PEDESTRIAN_COUNT	DRIVERRF_4
3	VEHICLECC_4	GVWR_AVG_4	VSURCOND_3
4	GVWR_AVG_4	NMIMPAIR	DRDISTRACT_4
5	NMIMPAIR	TRAV_SP_1	AGE_80_100_PCT

4.5 Feature Selection and Final Metrics: XGBoost

Oversampling effect (Var A vs Var B). Table 8 compares XGBoost validation performance under the oversampled (Var A) and non-oversampled (Var B) settings. The main difference concentrates on the rare Fatal class (Class 3), where oversampling yields a substantial F1 gain, while the majority-class metrics change only modestly.

Table 8: Comparison of oversampled (Var A) vs non-oversampled (Var B) XGBoost validation results

Metric	Var A (Over)	Var B (Non)	Diff (A-B)
F1 - Class 0	0.733728	0.734023	-0.000295
F1 - Class 1	0.459861	0.447016	0.012845
F1 - Class 2	0.314968	0.289179	0.025789
F1 - Class 3	0.594908	0.263959	0.330949
Macro F1	0.525866	0.433544	0.092322
Weighted F1	0.581277	0.565455	0.015822

Feature reduction uses a SHAP-gated voting scheme that combines SHAP (global), LIME (local), and DiCE (counterfactual) signals. The bottom `threshold_pct`% features are identified separately by (i) mean SHAP importance (averaged across the three threshold models), (ii) mean absolute LIME weight (only among LIME-observed features), and (iii) DiCE change frequency. Each feature receives a weighted low-importance score (SHAP=2, LIME=1, DiCE=1) and is dropped only if it is low-ranked by SHAP and also low-ranked by at least one of LIME or DiCE (score ≥ 3).

The explainability analysis is rerun on the reduced feature set to check attribution stability, and Table 9 compares the full vs. reduced performance.

Table 9: XGBoost full vs reduced feature evaluation

Metric	Full	Reduced
F1-0	0.726334	0.725452
F1-1	0.422237	0.422542
F1-2	0.274283	0.274321
F1-3	0.514220	0.507269
Macro F1	0.484269	0.482396
Weighted F1	0.555334	0.554659
# Features	163	154

4.6 Feature Selection and Final Metrics: SVM

Impact on validation and final test performance.

Table 10 summarizes the observed trade-off between aggressive explainability-driven reduction (substantial dimensionality drop but slightly lower Macro-F1) versus a minimal cleanup that removes only `RELJCT1`. The final SVM configuration uses the minimal cleanup and is retrained on Train+Validation before reporting test performance.
Table 10: SVM full vs reduced feature evaluation on validation set

Metric	Full	Reduced
F1-0	0.700000	0.690000
F1-1	0.470000	0.460000
F1-2	0.260000	0.250000
F1-3	0.720000	0.730000
Macro F1	0.535644	0.532119
Weighted F1	0.570000	0.560000
# Features	163	55

Table 11 shows the final results for both SVM and XGBoost models.

Table 11: Full vs. reduced feature performance comparison across models (**evaluated on the test set**)

Metric	XGBoost		SVM
	Full	Reduced	Full
F1-0	0.726334	0.725452	0.700000
F1-1	0.422237	0.422542	0.450000
F1-2	0.274283	0.274321	0.220000
F1-3	0.514220	0.507269	0.400000
Macro F1	0.484269	0.482396	0.445644
Weighted F1	0.555334	0.554659	0.540000
# Features	163	154	163

5 Discussion

Across models, performance is strongest on the majority class (No Injury) and weakest on the intermediate severities (Minor vs. Severe), where confusion is frequent and consistent with the ordinal structure of the labels. This indicates that the learned decision boundaries capture coarse severity regimes, but separating adjacent mid-level categories remains difficult under the available tabular descriptors.

For ordinal XGBoost, oversampling primarily improves minority behavior: Variant A yields a large F1 gain on Fatal and a clear Macro-F1 increase, while class-0 and Weighted-F1 change only modestly. This pattern is consistent with class imbalance limiting the model's capacity to represent Fatal cases in the non-oversampled setting.

The explainability tools provide a broadly coherent picture for XGBoost. SHAP (global) and aggregated LIME (local) repeatedly surface crash configuration (`P_CRASH1_*`), exposure/speed signals (`TRAV_SP_*`, counts), and roadway/context variables as dominant drivers. DiCE complements these by highlighting high-leverage decision-boundary "knobs" under a fatal-focused counterfactual protocol; importantly, these reflect model sensitivity (what the model changes to flip a prediction) rather than causal mechanisms.

For SVM, SHAP and LIME overlap on several prominent signals (age-group proportions, counts, speed/missingness indicators) but can disagree in sign and ranking, which is expected given LIME's local surrogate nature and the probability approximation required for `LinearSVC`. The per-source-class DiCE analysis further suggests class-conditional pathways to the Fatal target rather than a single universal feature lever.

Explainability-guided feature reduction does not improve validation performance. XGBoost shows near-parity between full and reduced representations, while SVM exhibits a small Macro-F1 drop after more aggressive pruning. Overall, the results support a "distributed signal" interpretation in which many weak, partially redundant variables collectively stabilize the boundary; removing low-ranked features can therefore simplify the representation without yielding better generalization. Finally, structured missingness markers appear among top-ranked features, implying that information availability in CRSS is itself predictive and should be treated carefully in interpretation.

6 Conclusion

This study compared linear SVM and ordinal XGBoost for four-level crash severity prediction under high-dimensional, imbalanced tabular data. Both models capture broad severity structure but repeatedly confuse the adjacent intermediate classes (Minor vs. Severe), suggesting that the available covariates provide only limited separability for fine-grained outcomes. Oversampling in XGBoost (Variant A) primarily benefits the minority Fatal class and yields a clear Macro-F1 improvement over the non-oversampled variant, while changes in Weighted-F1 are comparatively small, indicating that imbalance handling is a key factor for minority-class performance.

Across SHAP, aggregated LIME, and fatal-focused DiCE, the dominant signals are consistent: crash configuration (`P_CRASH1_*`), exposure/composition counts, travel-speed indicators (`TRAV_SP_*`), and roadway/context descriptors appear as the main drivers, while DiCE highlights which variables most often act as high-leverage

decision-boundary knobs rather than causal mechanisms. Explainability-guided feature reduction does not improve validation results and can slightly degrade performance (notably for SVM), reinforcing the view that predictive signal is distributed across many weak features. Future work should prioritize more explicitly ordinal and cost-sensitive training to reduce mid-class confusion, improved calibration for minority decisions, and richer contextual features to better explain intermediate severity outcomes.

A Team Contributions

- **Başak Çarhacıoğlu:** Implementation of training, test and explainability analysis pipeline for SVM model.
- **Sima Adleyba:** Implementation of training, test and explainability analysis pipeline for XGBoost model, dataset generation and adjustment
- **Ahmet İhsan Gül:** Sanity check and analysis for dataset, XGBoost and SVM Models; creation and maintanence of report and slides.

B Preliminary Method Analysis

An extensive analysis was conducted across several machine learning methods to decide on the best model to dive deep into analysis. The results obtained for these models are presented in this section.

Table 12: Classification report of the logistic regression model on the validation set

Class	Precision	Recall	F1-score	Support
0	0.671	0.6551	0.661	1466
1	0.595	0.347	0.439	1131
2	0.199	0.298	0.239	332
3	0.102	0.606	0.175	71
Macro Avg	0.392	0.476	0.378	3000
Weighted Avg	0.577	0.496	0.519	3000

C Detailed Explainability Results, XGBoost Variant A

Figure 2 provides the full SHAP beeswarm summaries and corresponding mean absolute SHAP importance bar plots for the three cumulative ordinal heads in XGBoost.

To complement the SHAP global attributions, Table 16 summarizes the most influential LIME features by aggregating local explanations over 89 analyzed instances.

DiCE is used in a fatal-focused counterfactual protocol to identify which variables most frequently need to change to move predictions away from Fatal. Table 17 lists the top-20 features ranked by change frequency across the generated counterfactuals.

Table 13: SVM family preliminary results. Linear SVM is reported on the validation set with and without SMOTE, while the Nyström-based SVM result is reported on the test set (as obtained during the milestone phase).

Model / Split	Precision	Recall	F1-score	Support
Linear SVM (No SMOTE)				
Class 0	0.633	0.782	0.700	1466
Class 1	0.605	0.351	0.444	1131
Class 2	0.246	0.178	0.206	332
Class 3	0.116	0.479	0.186	71
Macro Avg	0.400	0.447	0.384	3000
Weighted Avg	0.567	0.545	0.537	3000
Linear SVM (SMOTE)				
Class 0	0.640	0.803	0.712	1421
Class 1	0.617	0.411	0.493	1095
Class 2	0.254	0.280	0.266	322
Class 3	0.395	0.243	0.301	70
Macro Avg	0.477	0.434	0.443	2908
Weighted Avg	0.583	0.584	0.571	2908
Nyström-based SVM				
Class 0	0.646	0.745	0.692	1422
Class 1	0.577	0.405	0.476	1096
Class 2	0.244	0.320	0.277	322
Class 3	0.218	0.246	0.231	69
Macro Avg	0.421	0.429	0.419	2909
Weighted Avg	0.566	0.558	0.554	2909

Table 14: Performance metrics for Random Forest configurations on the test set. NOS = no oversampling; OS = oversampling; Raw = raw features; Imp = imputed features; Unw = unweighted loss; W = weighted loss.

Model	Acc	$F1_{macro}$	$F1_{weighted}$	F1(0)	F1(1)	F1(2)	F1(3)
M1: NOS, Raw, Unw	**0.605**	0.353	**0.561**	0.722	0.520	0.090	0.081
M2: NOS, Raw, W	0.600	0.371	0.555	0.720	0.509	0.055	0.198
M3: NOS, Imp, Unw	0.601	0.355	0.559	0.719	0.513	0.110	0.080
M4: NOS, Imp, W	0.600	**0.396**	0.558	0.715	0.514	0.067	**0.289**
M5: OS, Raw, Unw	0.590	0.323	0.545	0.706	0.508	0.077	0.000
M6: OS, Raw, W	0.598	0.374	0.556	0.713	0.522	0.052	0.209
M7: OS, Imp, Unw	0.592	0.331	0.548	0.708	0.511	0.077	0.028
M8: OS, Imp, W	0.589	0.321	0.534	0.713	0.484	0.012	0.075

Table 15: XGBoost ablation results on test set. Cum: cumulative; OS: oversampled; Imp: imputed; Pen: weighted loss.

Model	Acc	F_1^{mcr}	F_1^{wgt}	$F_1^{(0)}$	$F_1^{(1)}$	$F_1^{(2)}$	$F_1^{(3)}$
M1: Std, No-OS, No-Imp, No-Pen	**0.61**	0.40	**0.58**	**0.73**	0.54	0.13	0.18
M2: Std, No-OS, No-Imp, Pen	0.58	0.42	**0.58**	0.69	**0.56**	0.25	0.20
M3: Cum, No-OS, No-Imp, Pen	0.57	0.44	0.57	0.71	0.49	0.29	0.25
M4: Std, No-OS, Imp, No-Pen	0.60	0.37	0.57	0.72	0.53	0.15	0.10
M5: Std, No-OS, Imp, Pen	0.58	0.41	0.57	0.68	**0.56**	0.25	0.17
M6: Cum, No-OS, Imp, Pen	0.57	0.42	0.56	0.71	0.48	0.29	0.18
M7: Std, OS, No-Imp, No-Pen	0.59	0.37	0.56	0.71	0.52	0.14	0.12
M8: Cum, OS, No-Imp, No-Pen	0.58	**0.45**	0.57	0.70	0.49	**0.34**	0.26
M9: Std, OS, Imp, No-Pen	0.59	0.37	0.56	0.70	0.52	0.17	0.10
M10: Cum, OS, Imp, No-Pen	0.57	0.44	0.56	0.70	0.48	0.31	**0.27**

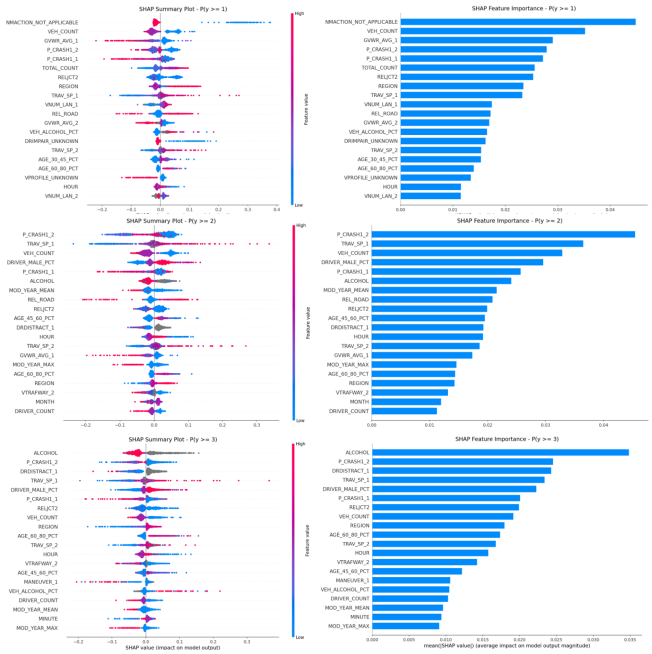

Figure 2: XGBoost SHAP results for (top) $P(y \geq 1)$, (middle) $P(y \geq 2)$, and (bottom) $P(y \geq 3)$. Each panel pairs a SHAP beeswarm summary (left) with mean absolute SHAP feature importance (right).

Table 16: XGBoost: Top-20 LIME features aggregated over 89 instances

Rank	Feature	Mean $\lvert w \rvert$	Count
1	VTRAFWAY_4	0.231394	1
2	PEDESTRIAN_COUNT	0.184694	37
3	P_CRASH1_4	0.180862	3
4	DRIMPAIR_2	0.176630	28
5	DRDISTRACT_3	0.173920	7
6	P_CRASH1_3	0.165148	5
7	MANEUVER_4	0.163224	6
8	P_CRASH1_UNKNOWN	0.162790	1
9	VALIGN_4	0.162504	2
10	MINUTE_UNKNOWN	0.161459	4
11	DRIMPAIR_1	0.160333	17
12	VNUM_LAN_4	0.159103	6
13	VEHICLECC_3	0.155305	4
14	VSURCOND_NOT_APPLICABLE	0.152968	7
15	MANEUVER_3	0.152781	5
16	NMCC_UNKNOWN	0.150103	7
17	VEH_COUNT	0.149971	48
18	VSURCOND_4	0.149148	9
19	NMDISTRACT	0.148349	2
20	DRIVERRF_4	0.148236	5

Table 17: XGBoost (Var A): DiCE fatal-focused summary. Top-20 most frequently changed features across counterfactuals

Rank	Feature	Change Count	Change Rate
1	P_CRASH1_2	52	0.866667
2	MANEUVER_1	14	0.233333
3	DRIVER_DRUGS_PCT	12	0.200000
4	REL_ROAD	11	0.183333
5	VSURCOND_1	11	0.183333
6	P_CRASH1_1	11	0.183333
7	NMCC	10	0.166667
8	VEH_COUNT	10	0.166667
9	VTRAFWAY_2	10	0.166667
10	DRIMPAIR_1	10	0.166667
11	AGE_0_18_PCT	9	0.150000
12	TOTAL_COUNT	9	0.150000
13	LGT_COND	9	0.150000
14	TRAV_SP_4	9	0.150000
15	VEH_ALCOHOL_COUNT	8	0.133333
16	DRIMPAIR_2	8	0.133333
17	PEDPOS_ROADWAY_ANY	8	0.133333
18	DRIVERRF_2	8	0.133333
19	VTRAFWAY_1	7	0.116667
20	HOUR	7	0.116667

D Detailed Explainability Results, XGBoost Full Model

Figure 3 presents the complete SHAP beeswarm summaries and corresponding mean absolute SHAP importance bar plots for the three cumulative ordinal heads in the full-feature XGBoost model.

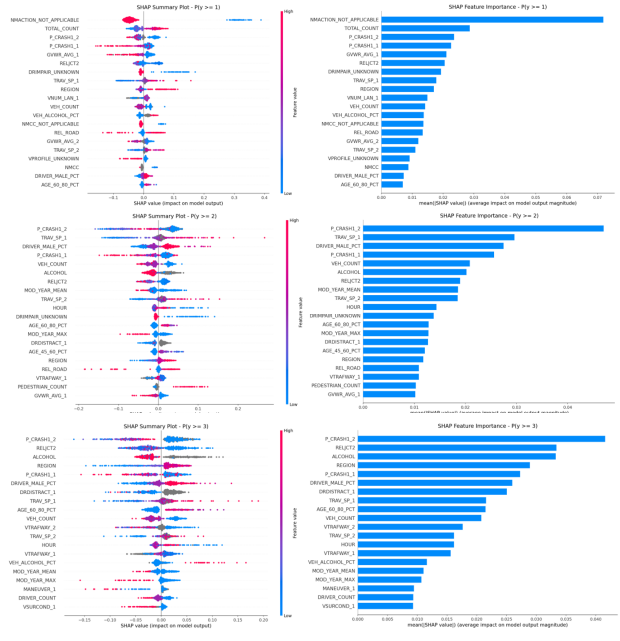

Figure 3: XGBoost (Full) SHAP results for (top) $P(y \geq 1)$, (middle) $P(y \geq 2)$, and (bottom) $P(y \geq 3)$. Each panel pairs a SHAP beeswarm summary (left) with mean absolute SHAP feature importance (right).

Table 18 summarizes the most influential LIME features

by aggregating local explanations over 89 analyzed instances.

Table 18: XGBoost (Full): Top-20 LIME features aggregated over 89 instances

Rank	Feature	Mean $\|w\|$	Count
1	P_CRASH1_3	0.163860	1
2	TRAV_SP_4	0.163819	3
3	VEH_COUNT	0.158879	38
4	RELJCT1_UNKNOWN	0.157395	1
5	P_CRASH1_2	0.154680	70
6	NMIMPAIR	0.154557	1
7	DRDISTRACT_3	0.152819	3
8	DRDISTRACT_NOT_APPLICABLE	0.151034	4
9	VEHICLECC_3	0.147625	3
10	MANEUVER_4	0.146759	1
11	DRIMPAIR_1	0.146134	6
12	MANEUVER_3	0.144506	1
13	DRIMPAIR_NOT_APPLICABLE	0.143508	2
14	P_CRASH1_NOT_APPLICABLE	0.138363	3
15	DRIVER_MALE_PCT	0.136983	16
16	PEDPOS_INTERSECTION_ANY	0.134626	3
17	MANEUVER_NOT_APPLICABLE	0.133528	3
18	VTRAFWAY_4	0.131642	4
19	VSURCOND_4	0.130580	7
20	GVWR_AVG_4	0.129859	2

DiCE is applied in a fatal-focused counterfactual protocol to identify which variables most frequently need to change to move predictions away from Fatal. Table 19 lists the top-20 features ranked by change frequency.

Table 19: XGBoost (Full): DiCE fatal-focused summary. Top-20 most frequently changed features across counterfactuals.

Rank	Feature	Change Count	Change Rate
1	P_CRASH1_2	51	0.850000
2	MANEUVER_1	33	0.550000
3	DRDISTRACT_1	21	0.350000
4	DRIMPAIR_1	21	0.350000
5	AGE_0_18_PCT	18	0.300000
6	PEDPOS_CROSSWALK_ANY	17	0.283333
7	REL_ROAD	15	0.250000
8	VPROFILE_2	15	0.250000
9	DRIVER_DRUGS_PCT	15	0.250000
10	DRIVER_COUNT	14	0.233333
11	DRIVER_MALE_PCT	13	0.216667
12	TOTAL_COUNT	13	0.216667
13	VNUM_LAN_2	13	0.216667
14	VPROFILE_4	12	0.200000
15	GVWR_AVG_2	11	0.183333
16	VEHICLECC_4	11	0.183333
17	VPROFILE_3	11	0.183333
18	GVWR_AVG_1	10	0.166667
19	VTRAFWAY_1	10	0.166667
20	NMDISTRACT_UNKNOWN	9	0.150000

E Detailed Explainability Results, XGBoost Reduced Model

Figure 4 provides the full SHAP beeswarm summaries and corresponding mean absolute SHAP importance bar plots for the three cumulative ordinal heads after explainability-guided feature reduction.

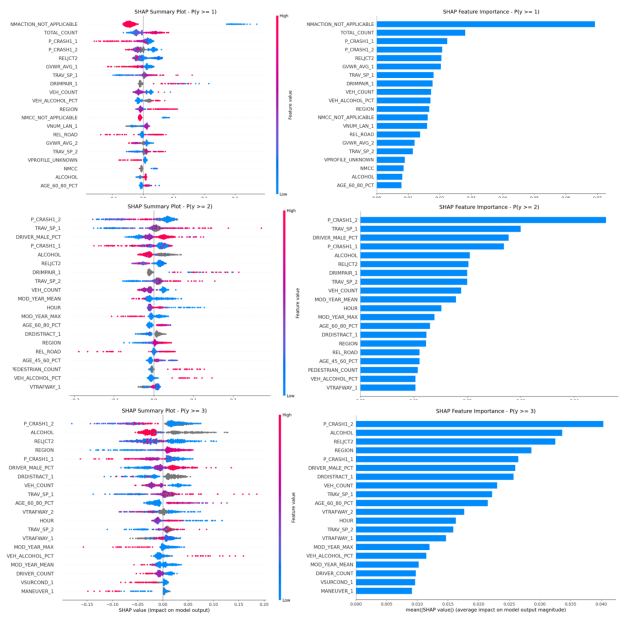

Figure 4: XGBoost (Reduced) SHAP results for (top) $P(y \geq 1)$, (middle) $P(y \geq 2)$, and (bottom) $P(y \geq 3)$. Each panel pairs a SHAP beeswarm summary (left) with mean absolute SHAP feature importance (right).

To complement SHAP, Table 20 reports the most influential LIME features aggregated over 89 analyzed instances.

Table 20: XGBoost (Reduced): Top-20 LIME features aggregated over 89 instances

Rank	Feature	Mean $\|w\|$	Count
1	VEH_COUNT	0.174239	35
2	MINUTE_UNKNOWN	0.164803	2
3	VSURCOND_4	0.158577	4
4	VEHICLECC_2	0.154732	5
5	P_CRASH1_2	0.152273	75
6	VTRAFWAY_NOT_APPLICABLE	0.149987	4
7	VNUM_LAN_NOT_APPLICABLE	0.148733	3
8	VEHICLECC_NOT_APPLICABLE	0.145890	4
9	MANEUVER_4	0.145104	4
10	ALCOHOL	0.144787	1
11	AGE_100_120_PCT	0.143120	2
12	VNUM_LAN_4	0.142995	2
13	PEDPOS_OFFROAD_ANY	0.141097	2
14	TRAV_SP_4	0.139233	5
15	DRDISTRACT_NOT_APPLICABLE	0.136138	2
16	MANEUVER_3	0.135229	3
17	DRIVER_MALE_PCT	0.134830	17
18	VTRAFWAY_4	0.134826	4
19	PEDESTRIAN_COUNT	0.133953	23
20	1.00	0.133638	42

Finally, Table 21 provides the fatal-focused DiCE summary for the reduced model, listing the top-20 most frequently changed features across the generated counterfactuals.

Table 21: XGBoost (Reduced): DiCE fatal-focused summary. Top-20 most frequently changed features across counterfactuals.

RankFeature	Change Count	Change Rate
1P__CRASH1__2	53	0.883333
2MANEUVER__1	25	0.416667
3REL__ROAD	20	0.333333
4DRDISTRACT__1	19	0.316667
5PEDPOS__CROSSWALK__ANY	18	0.300000
6MANEUVER__2	16	0.266667
7DRIVER__COUNT	16	0.266667
8DRIVER__MALE__PCT	15	0.250000
9PEDESTRIAN__COUNT	14	0.233333
10LGT__COND	13	0.216667
11DRDISTRACT__2	13	0.216667
12DRIVERRF__2	12	0.200000
13P__CRASH1__1	11	0.183333
14MOD__YEAR__MEAN	11	0.183333
15VNUM__LAN__2	11	0.183333
16NMCC	11	0.183333
17VTRAFWAY__1	11	0.183333
18DRDISTRACT__3	11	0.183333
19DRIVER__DRUGS__PCT	10	0.166667
20GVWR__AVG__1	10	0.166667

F Detailed Explainability Results, SVM

Figure 5 presents the class-conditional SHAP results for the linear SVM, visualizing the most important features for each class, ranked by mean absolute SHAP value. Figure 6 complements this analysis by illustrating the low-importance tail, highlighting features whose SHAP attributions remain near-zero across samples.

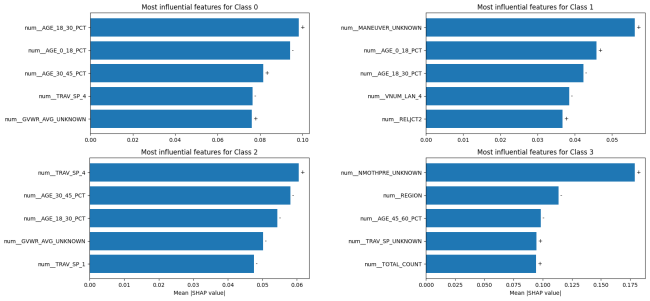

Figure 5: SVM SHAP most important features for classes 0–3. Each panel reports the top-ranked features by mean absolute SHAP value, indicating the dominant contributors to the corresponding class scores.

In addition to SHAP, we report complementary explainability results obtained using LIME and DiCE, focusing exclusively on the most influential features.

Figure 7 visualizes the class-conditional LIME explanations for the linear SVM, highlighting the features that most strongly contribute to each class prediction.

To complement the SHAP attributions, Table 22 reports the top-5 LIME features for each class. Features are ranked by mean absolute LIME weight, but the table reports the signed mean weight ($\overline{w}$), where positive values increase the corresponding class score and negative values decrease it.

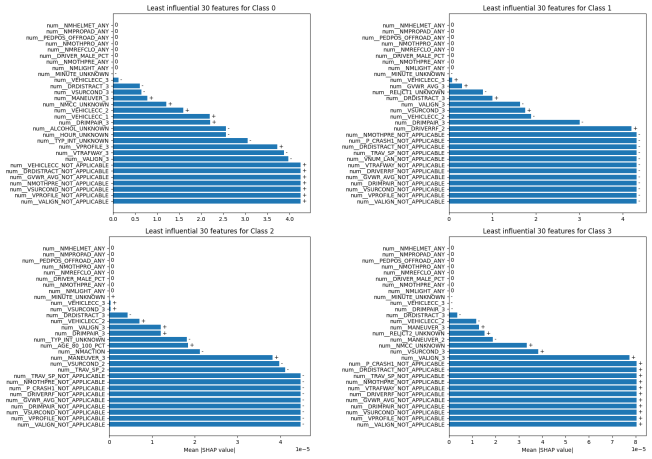

Figure 6: SVM SHAP least important features (low-importance tail) for classes 0–3. The visualizations highlight features with near-zero SHAP attributions, indicating minimal influence on the class predictions.

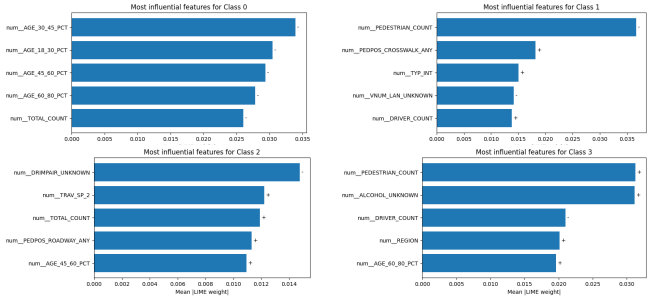

Figure 7: LIME most important features for classes 0–3. Each panel highlights the features with the strongest local contributions to the corresponding class predictions of the linear SVM.

Table 22: SVM: Top-5 LIME features per class (ranked by mean $|w|$, values reported as signed mean weight $\overline{w}$).

Class	Rank	Feature	$\overline{w}$
0	1	num__AGE__30__45__PCT	-0.03393
0	2	num__AGE__18__30__PCT	-0.03048
0	3	num__AGE__45__60__PCT	-0.02939
0	4	num__AGE__60__80__PCT	-0.02785
0	5	num__TOTAL__COUNT	-0.02606
1	1	num__PEDESTRIAN__COUNT	-0.03669
1	2	num__PEDPOS__CROSSWALK__ANY	+0.01817
1	3	num__TYP__INT	+0.01508
1	4	num__VNUM__LAN__UNKNOWN	-0.01419
1	5	num__DRIVER__COUNT	+0.01385
2	1	num__DRIMPAIR__UNKNOWN	-0.01478
2	2	num__TRAV__SP__2	+0.01222
2	3	num__TOTAL__COUNT	+0.01192
2	4	num__PEDPOS__ROADWAY__ANY	+0.01132
2	5	num__AGE__45__60__PCT	+0.01095
3	1	num__PEDESTRIAN__COUNT	+0.03127
3	2	num__ALCOHOL__UNKNOWN	+0.03114
3	3	num__DRIVER__COUNT	-0.02099
3	4	num__REGION	+0.02013
3	5	num__AGE__60__80__PCT	+0.01960

Figure 8 presents the corresponding DiCE-based analysis, summarizing the most important features identified through counterfactual reasoning. DiCE is applied using a fatal-focused counterfactual protocol with target class 3 (Fatal). For each non-fatal source class, counterfactuals are generated toward the target, and features are ranked by DiCE importance (change frequency × mean absolute delta).

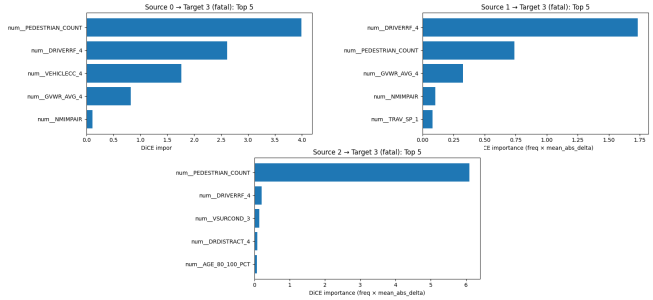

Figure 8: DiCE most important features for classes 0–3. The visualizations summarize the features that most frequently appear in counterfactual explanations, indicating their relevance for changing the predicted class of the linear SVM.

Table 23 summarizes the top-5 features per source class based on this counterfactual analysis.

Table 23: SVM: DiCE fatal-focused summary (source → target 3). Top-5 features per source class ranked by DiCE importance (freq × mean_abs_delta).

Src	R	Feature	Imp.	Freq	MeanAbsΔ
0	1	num__PEDESTRIAN_COUNT	3.362092	0.350	9.605978
0	2	num__DRIVERRF_4	3.006504	0.150	20.043362
0	3	num__TRAV_SP_1	0.254871	0.100	2.548706
0	4	num__NMACTION_UNKNOWN	0.173330	0.100	1.733301
0	5	num__DRIMPAIR_4	0.083021	0.050	1.660421
1	1	num__PEDESTRIAN_COUNT	1.897380	0.250	7.589520
1	2	num__GVWR_AVG_4	0.821643	0.150	5.477621
1	3	num__VSURCOND_3	0.473718	0.100	4.737180
1	4	num__TRAV_SP_1	0.299701	0.100	2.997009
1	5	num__DRDISTRACT_3	0.140664	0.050	2.813285
2	1	num__PEDESTRIAN_COUNT	4.160077	0.450	9.244615
2	2	num__GVWR_AVG_4	0.113536	0.050	2.270719
2	3	num__DRIMPAIR_3	0.097282	0.050	1.945636
2	4	num__VSURCOND_3	0.095976	0.050	1.919520
2	5	num__VPROFILE_4	0.079981	0.050	1.599619