1. **Predict Accident Severity**

Model: **ML Type:** Classification  
**Models:** Scikit-learn (Random Forest & Gradient Boosting)

**Key Takeaways**

* **Random Forest slightly struggles with false negatives but still detects some Severe accidents correctly.**
* **Gradient Boosting almost never predicts Severe accidents correctly, leading to high false negatives (Severe misclassified as Slight).**
* **If Severe accident detection is important, Random Forest might be the better choice, but if Slight accident classification is the priority, Gradient Boosting performs better.**

**Improvement to model:**  
- Balance Classes: Since the model is struggling with Serious/Fatal accidents, it may be biased toward Slight accidents (class imbalance). Using SMOTE (Synthetic Minority Over-sampling Technique):  
This will increase Serious/Fatal cases synthetically and improve classification

**High-Level Insights**

* Random Forest is more balanced but needs improvements in identifying Severe accidents.
* Gradient Boosting is highly skewed toward Slight cases, meaning it ignores Severe accidents entirely.

**Model: XGBoost model & SVM model**

**SMOTE helped balance the dataset, but additional fine-tuning is needed to improve classification for Severe accidents: Use XGBoost model**

**Overall model performance comparison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Performance Comparison** | | |  |  |
| **Metric** | **Random Forest** | **XGBoost** | **SVM** | **Key Insights** |
| **Accuracy** | **67%** | **57%** | **68%** | SVM slightly outperforms Random Forest; XGBoost lags behind. |
| **Precision (Slight)** | **0.87** | **0.87** | **0.87** | All models classify "Slight" cases with high precision. |
| **Recall (Slight)** | **0.72** | **0.57** | **0.73** | SVM has the best recall for "Slight" cases; XGBoost struggles. |
| **F1-score (Slight)** | **0.79** | **0.69** | **0.8** | SVM performs best for "Slight" case balance. |
| **Precision (Severe)** | **0.18** | **0.17** | **0.19** | All models struggle to correctly classify "Severe" cases. |
| **Recall (Severe)** | **0.36** | **0.52** | **0.35** | XGBoost captures more "Severe" cases, but at a cost of accuracy. |
| **F1-score (Severe)** | **0.24** | **0.26** | **0.24** | XGBoost has the best balance for "Severe" cases, but all models need improvement. |
| **Weighted F1-score** | **0.71** | **0.63** | **0.72** | SVM slightly outperforms Random Forest; XGBoost is less balanced. |

**Insights & Takeaways**

* **SVM edges out Random Forest slightly in overall accuracy (68% vs. 67%).**
* **SVM is strongest at predicting "Slight" cases** with the highest **F1-score (0.80)**.
* **XGBoost is better at capturing "Severe" cases** due to its **higher recall (0.52 vs. 0.35 for SVM and 0.36 for Random Forest)**.
* **Random Forest remains the most balanced model**, performing reasonably well across both classes without extreme trade-offs.
* **None of the models are strong in Severe case detection**, though XGBoost offers slightly better recall.

**Final Recommendation**

* If **overall balanced classification** is needed, **Random Forest** is your best bet.
* If **capturing more Severe cases** is the goal, **XGBoost** performs better, but at the cost of accuracy.
* If **prioritizing correct Slight case predictions**, **SVM delivers the highest recall and F1-score** for that class.

**Conclusion: Best Overall Model**

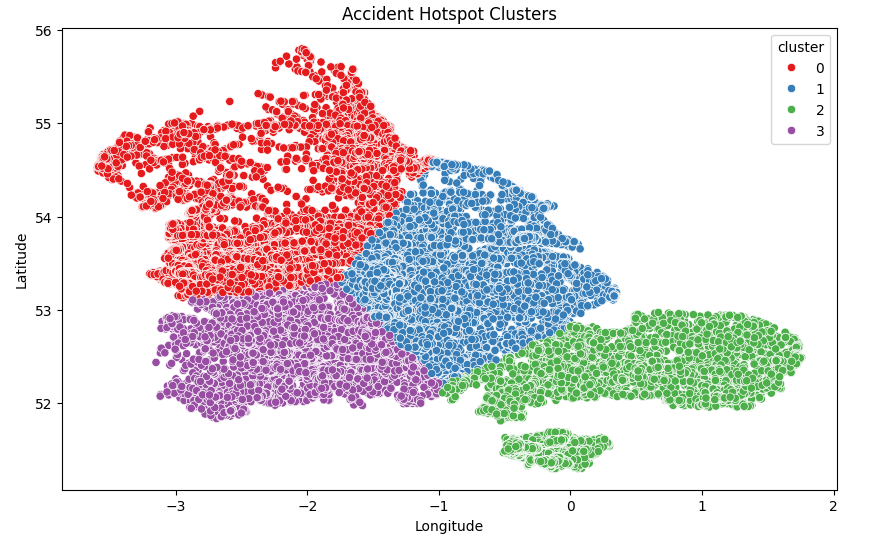
* **XGBoost** is the best overall model as our priority is detecting severe accidents (highest recall).  
  It trades a few more false positives (slight misclassified as severe) to catch more actual severe accidents, which is often a worthwhile tradeoff in safety-critical scenarios.

**2. Identify High-Risk Accident Zones**

**‘A) Using K-Means (Unsupervised Learning model)**

* ML Type: Clustering
* Models: K-Means **with 4 clusters**

Since dataset includes latitude and longitude, we'll primarily focus on spatial clustering.



**Accident Hotspot Clusters** scatter plot provides valuable insights into **high-risk zones** where accidents frequently occur

**Interpretation of the Clusters**

**✔ Cluster 0 (Red) → This region has high accident density, likely indicating urban congestion, busy intersections, or poor visibility.  
✔ Cluster 1 (Blue) → These accidents may be spread across major roads or highways, suggesting speed-related or lane-merging risks.  
✔ Cluster 2 (Green) → This cluster might show localized accident hotspots, possibly near residential or commercial areas.  
✔ Cluster 3 (Purple) → Typically found in less densely populated regions, meaning accidents here may be influenced by weather conditions or rural road designs.**

B**) Using DBSCAN forms:**

**Result:** DBSCAN forms clusters based on density, detecting naturally occurring high-risk areas.

**What This Signifies:**

This image likely represents a **clustered map visualization** of accident locations based on **DBSCAN clustering**, where different markers indicate accident severity. Here’s a breakdown of its insights:

**1. Identification of High-Risk Accident Zones**

* The **clusters** in the map show areas where accidents are concentrated.
* DBSCAN effectively highlights **dense accident regions** where intervention may be necessary.
* Larger clusters suggest **frequent accident occurrences**, making them priority zones for safety measures.

**2. Severity-Based Marker Representation**

* **Color-coded markers** differentiate between accident severity:
  + **Red:** High-risk severe accidents.
  + **Green:** Slight accidents.
* This allows easy identification of hazardous areas.

**3. Geographic Patterns & Implications**

* **Accident hotspots near city centers or intersections** may indicate risky zones for drivers.
* **Sprawled clusters** suggest wider regional risks, potentially linked to infrastructure issues or weather conditions.
* **Dense clusters** could align with high-traffic routes requiring safety enhancements.

**2 (C) Unsupervised learning**  
  
Overlaying **Accident Severity** onto the cluster map will help prioritize high-risk zones

**Key Observations:**

**DBSCAN**, which detects clusters based on density

Extracts relevant accident data (latitude, longitude, severity).

**🚦 Dense Clustered Zones & Severity-Based Mapping:**

* Fatal accident clusters are marked in red, emphasizing high-risk areas.
* Serious accident zones (orange) show potential danger hotspots needing intervention.

**🚦 Strategic Safety Insights:**

* High-risk urban areas likely require traffic calming measures, enhanced enforcement, or infrastructure upgrades.
* Remote accident sites may benefit from better emergency response planning and road condition improvements.

**Result**: **It is analyzing patterns & identifying existing high-risk zones.**